

Imprecise Probabilities in Machine Learning: Structure and Semantics

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
M.Sc. Christian Fröhlich
aus Würzburg

Tübingen
2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:

23.07.2025

Dekan:

Prof. Dr. Thilo Stehle

1. Berichterstatter/-in:

Prof. Robert C. Williamson, PhD

2. Berichterstatter/-in:

Jun.-Prof. Dr. Stephan Eckstein

3. Berichterstatter/-in:

Prof. Dr. Thomas Augustin

Imprecise Probabilities in Machine Learning: Structure and Semantics

ABSTRACT

Motivated by problems in machine learning, this dissertation advances the theory of *imprecise probabilities*, which offers a more flexible framework for representing uncertainty compared to traditional precise probabilities. In this work, we investigate the mathematical structure of imprecise probabilities, while considering a plurality of semantics for them. Although semantics and mathematics may initially appear independent, we observe that the choice of semantics shapes the mathematical framework that emerges.

With the goal of expressing risk aversion and ambiguity aversion, Part I investigates imprecise probabilities primarily in the form of *law invariant coherent risk measures*. In machine learning, these can help to reduce the tail risk in a distribution, as well as to guard against distributional shifts. Drawing on insights from the *rearrangement invariant Banach function spaces* literature, we investigate the structure of law invariant coherent risk measures. In particular, we study in depth the tail sensitivity of such risk measures and show how this yields a stratification of the law invariant coherent risk measures.

In Part II, we move beyond law invariance and conduct a fundamental investigation of imprecise probabilities from a generalized frequentist viewpoint. Here, we challenge the assumption that a precise probability always suffices to capture the aggregate regularity of a data sequence, and show how imprecise probability naturally arises in the general case. To this end, we study imprecision under various *data models*. We furthermore propose a general framework for the evaluation of *imprecise forecasts* under such data models. Specifically, we develop viable notions of *proper scoring rules* and *calibration* for imprecise probabilities, generalizing their traditional correspondences.

Conceptually, our viewpoint on uncertainty may be of broader interest and also yields insights into the precise case. Our focus on two key ingredients, *data models* and *decision problems*, proves fruitful. Moreover, we illustrate how looking to *insurance* can aid a better understanding of uncertainty in general. In Part III, we exemplify this perspective by establishing bridges between fairness concepts in insurance and machine learning.

Imprecise Probabilities in Machine Learning: Structure and Semantics

ZUSAMMENFASSUNG

Motiviert durch Problemstellungen des maschinellen Lernens entwickelt diese Dissertation die Theorie der *unpräzisen Wahrscheinlichkeiten* weiter, welche im Vergleich zu traditionellen präzisen Wahrscheinlichkeiten einen flexibleren Rahmen für die Darstellung von Unsicherheiten bietet. In dieser Arbeit untersuchen wir die mathematische Struktur unpräziser Wahrscheinlichkeiten, während wir verschiedene Semantiken für sie in Betracht ziehen. Obwohl Semantik und Mathematik auf den ersten Blick voneinander unabhängig erscheinen mögen, stellen wir fest, dass die Wahl der Semantik den sich ergebenden mathematischen Rahmen prägt.

Mit dem Ziel, Risikoaversion und Ambiguitätsaversion auszudrücken, untersucht Teil I unpräzise Wahrscheinlichkeiten primär in Form von *verteilungsinvarianten kohärenten Risikomaßen*. Beim maschinellen Lernen können diese helfen, das Tail-Risiko in einer Verteilung zu reduzieren und vor den Konsequenzen aus Verteilungsverschiebungen zu schützen. Auf der Grundlage von Erkenntnissen aus der Literatur über *umordnungsinvariante Banach-Funktionenräume* untersuchen wir die Struktur von verteilungsinvarianten kohärenten Risikomaßen. Insbesondere untersuchen wir eingehend die Tail-Sensitivität solcher Risikomaße und zeigen, wie diese zu einer Stratifizierung der verteilungsinvarianten kohärenten Risikomaße führt.

In Teil II gehen wir über die Verteilungsinvarianz hinaus und führen eine grundlegende Untersuchung von unpräzisen Wahrscheinlichkeiten aus einer verallgemeinerten frequentistischen Sichtweise durch. Hier stellen wir die Annahme, dass eine präzise Wahrscheinlichkeit immer ausreicht, um die gesamte Regelmäßigkeit einer Datenfolge zu erfassen, in Frage, und zeigen, wie unpräzise Wahrscheinlichkeit im allgemeinen Fall entsteht. Zu diesem Zweck untersuchen wir die Unschärfe unter verschiedenen *Datenmodellen*. Darüber hinaus stellen wir einen allgemeinen Rahmen auf für die Bewertung von *unscharfen Vorhersagen* unter solchen Datenmodellen. Insbesondere entwickeln wir praktikable Konzepte von “proper scoring rules” und Kalibration für unscharfe Wahrscheinlichkeiten, welche deren traditionelle Entsprechungen verallgemeinern.

Konzeptionell kann unsere Perspektive zu Problemen im Zusammenhang mit Unsicherheit von allgemeinem Interesse sein und sie liefert auch Einsichten für den präzisen Fall. Unser Fokus auf zwei Schlüsselkomponenten erweist sich als fruchtbar: *Datenmodelle* und

Entscheidungsprobleme. Außerdem veranschaulichen wir, wie der Blick auf *Versicherungen* zu einem besseren Verständnis von Unsicherheit im Allgemeinen beitragen kann. In Teil III exemplifizieren wir diese Perspektive, indem wir eine Brücke zwischen Fairnesskonzepten in Versicherungen und im maschinellen Lernen bauen.

Contents

0	Prologue	1
0	INTRODUCTION	2
0.1	Semantics — A First Look	2
0.2	Syntax	4
0.3	Structure	7
0.4	Semantics — A Second Look	7
0.5	The Connection to Machine Learning	31
0.6	Relating Risk, Ambiguity and Aleatoric Plurality	35
0.7	Appendix	39
1	SUMMARY	42
1.1	Summary of Chapter 2	44
1.2	Summary of Chapter 3	51
1.3	Summary of Chapter 4	55
1.4	Summary of Chapter 5	59
1.5	Summary of Chapter 6	63
1.6	Summary of Chapter 7	67
1.7	Key Contributions	69
I	Risk and Ambiguity — Law Invariance	74
2	RISK MEASURES AND UPPER PROBABILITIES: COHERENCE AND STRATIFICATION	76
2.1	Introduction	78
2.2	Coherent Lower and Upper Previsions	83
2.3	Coherent Risk Measures	89
2.4	Rearrangement Invariant Banach Function Spaces	102
2.5	Creating New Risk Measures from Old	127
2.6	Experiments	137

2.7	Conclusion	146
2.8	Appendix: Proofs	148
2.9	Appendix: (Non)-Expected Utility Theories	170
2.10	Appendix: Experiments	183
3	TAILORING TO THE TAILS: RISK MEASURES FOR FINE-GRAINED TAIL SENSITIVITY	186
3.1	Introduction	188
3.2	Orlicz Norms and Spaces	193
3.3	Orlicz Regret Measures	194
3.4	Rearrangement Invariant Banach Norms and Fundamental Functions	203
3.5	Tail-Specific Marcinkiewicz and Lorentz Norms	210
3.6	Tail-Specific Divergence Risk Measures	218
3.7	Tail-Specific Orlicz Deviation Inequalities	225
3.8	Application	228
3.9	Conclusion	230
3.10	Appendix: Orlicz Regret Measures	231
3.11	Appendix: Rearrangement Invariant Banach Norms and Fundamental Func- tions	240
3.12	Appendix: Tail-Specific Marcinkiewicz and Lorentz Norms	244
3.13	Appendix: Utility-Based Shortfall Risk	248
3.14	Appendix: Plain Language Summary and Diagram	251
II	Aleatoric Plurality — Beyond Law Invariance	254
4	STRICTLY FREQUENTIST IMPRECISE PROBABILITY	256
4.1	Introduction	258
4.2	Unstable Relative Frequencies	265
4.3	From Cluster Points to Sequence	271
4.4	Unstable Conditional Probability	274
4.5	Related Work	280
4.6	Conclusion	283
4.7	Appendix: Proofs	285
4.8	Appendix: Existence of Sequences with Prespecified Relative Frequency Clus- ter Points	289
4.9	Appendix: Unstable Independence	308
4.10	Appendix: Pathological or Normal?	311
5	DATA MODELS WITH TWO MANIFESTATIONS OF IMPRECISION	320

5.1	Introduction	322
5.2	Data Models	325
5.3	Non-Stationary, Locally Precise Data Models	330
5.4	Stationary, Locally Imprecise Data Models	336
5.5	Applications and Estimability	347
5.6	Discussion	357
5.7	Appendix	360
6	SCORING RULES AND CALIBRATION FOR IMPRECISE PROBABILITIES	372
6.1	Introduction	374
6.2	IP Scoring Rules	378
6.3	IP Calibration	387
6.4	Experiments	392
6.5	Related Work	399
6.6	Discussion	400
6.7	Appendix	403
III	An Actuarial Perspective	430
7	INSIGHTS FROM INSURANCE FOR FAIR MACHINE LEARNING	432
7.1	Introduction	434
7.2	A Brief History of Insurance Rationalities	435
7.3	Responsibility	442
7.4	The Link to Fair Machine Learning	445
7.5	Tensions between Aggregate and Individuals	449
7.6	Related Work	452
7.7	Discussion	452
7.8	Appendix	453
IV	Conclusion	460
8	AN OUTLOOK	462
	REFERENCES	464

Acknowledgments

First and foremost, I want to thank Bob for his guidance and advice throughout my PhD. I have learned so much from you, and your approach to research has made a lasting influence on me. It was a great pleasure to work with you.

I want to thank my TAC members Paul Bürkner and Konstantin Genin for their helpful advice. Many thanks also to Stephan Eckstein for reviewing this thesis, and to Philipp Hennig and Charley Wu for participating in the oral examination.

I truly enjoyed the PhD experience, and this is mostly due to the great environment I was embedded in. I am especially grateful to the FMLS group: Armando, Ben, Bob, Charlotte, Laura, Mina, Nan, Rabanus, Raphael, Raysa, and Wenkai. You made the whole journey fun! I don't know how things would have turned out without Rabanus' support, especially in the beginning. I cherish the discussions we had in front of the whiteboard.

Beyond the FMLS group, I was fortunate to be part of the MvL6 environment, surrounded by friendly colleagues. Special thanks to Nina for the coffees we shared and the mutual advice about navigating the PhD process. The "third order optimization" group helped significantly in getting started. In addition to the MvL6, the IMPRS-IS events were valuable, and the bootcamps were a lot of fun. Many thanks to Heike, Leila, Pannapa and Sara for their support.

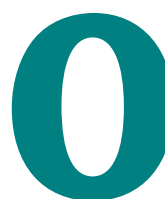
I am beyond grateful for the amazing people who are in my life.

You must understand the whole of life, not just one little part of it. That is why you must read, that is why you must look at the skies, that is why you must sing, and dance, and write poems, and suffer, and understand, for all that is life. [Krishnamurti, 1989]

Doing a PhD is just a little part of life, and I want to thank my friends, family and Lara for all the rest.

Part 0

Prologue



Introduction

This thesis is about *imprecise probabilities*. Such an imprecise probability, from a mathematical point of view, is essentially a set of probabilities.¹ For example, consider a rain forecaster who announces that “the probability for rain tomorrow is between 20% and 30%” — in the case of a single binary event, an imprecise probability is simply a sub-interval of $[0, 1]$.² In some way, this statements intends to express a form of *uncertainty* about the event that it rains. An obvious question comes to mind: what could this statement mean?

0.1 SEMANTICS — A FIRST LOOK

Through the process of my research I arrived at the conclusion that we should consider a *plurality of semantics* for imprecise probabilities, with different useful roles in different modeling contexts. For example, the above rain forecast could be interpreted to mean each of the following:

RS1) Actually we know that the probability is, say, 25%; but because we are *risk-averse*, we pretend it is either 20% or 30%, depending on the directionality of the decision we want to make (do we want to bet *on* or *against* rain?)

¹We will put aside the subtle question of convexity for now, see Section 0.2.

²This relates to convexity, see Footnote 1.

RS2) We do not “know” the probability for rain in the given context (perhaps we are asked about rain in a country that is entirely unfamiliar to us), but can form a rough guess. Perhaps we face unspecific or conflicting evidence, so we are hesitant to pin down the probability to a single real number. Again, an interval enables more cautious decision-making — depending on the decision, we would take the lower or upper number (the same applies to the next cases).

RS3) *What if there is no true probability for rain?* We call this *aleatoric plurality*. This in turn has its own duality of intricately related semantics:

- (a) Perhaps the data sequence (indicating whether it rained or not on each day) has “unstable” relative frequencies, meaning the relative frequencies still exhibit considerable instability even after many trials. This can occur in the context of *heavy tails* [Nair et al., 2022] — think of extreme rainfall events, for example. The character of such events is quite different from highly regular, well-behaved events like throwing dice or gambling in the casino.
- (b) There might be *hidden heterogeneity* in the data. For example, the probability for rain in country A might differ from the probability in country B, and the data at hand is from both countries. Why could this mean that there “is no true probability”? Consider the setting where we want to train a machine learning model on the combined data, but we restrict access to the feature “country origin”, so our model is not allowed to see it. There are multiple good reasons for this, for instance because we only have access to this feature at training time, but not at test time, and our aim is to obtain a model which works well on both countries. This leads to the paradigm of *group distributional robustness*. In this case (we will elaborate on this), there is no single true probability.

This scheme and subsets of it will be re-iterated throughout the thesis. Interestingly, RS3b and RS3a are intricately related; a key contribution of this thesis is to clarify (mathematically) their relation. Imprecise probabilities under these semantics have already been studied by different scientific communities. Under the name of *coherent risk measures* [Artzner et al., 1999], scholars in finance and insurance have focused mostly on RS1. Primarily motivated by a subjectivist view of probability, the community formed under the banner of imprecise probabilities (also using the terminology of *coherent lower and upper previsions*) [Walley, 1991] centers their attention on RS2. In machine learning, programs such as group distributional robustness, federated learning and fair machine learning are closely related to RS3b. Finally, few scholars so far (e.g. [Walley and Fine, 1982, Gorban, 2011]) have considered RS3a. In this thesis, all of these semantics will play a role, with RS1 and RS2 primarily featuring in Part I, and Part II focusing on RS3.

0.2 SYNTAX

From a mathematical viewpoint, imprecise probabilities share a common form. To present it, we now introduce some basic objects. A key formal object throughout this thesis is the possibility set Ω ; we think of an elementary event $\omega \in \Omega$ as capturing all contextually relevant information. Intuitively, $\omega \in \Omega$ is a description of a state of the world. Already here — though this will become much clearer later — we highlight the essential context dependence, which information is included in ω and which is left out. For example, in a machine learning use case, $\Omega = \mathcal{X} \times \mathcal{Y}$ might capture all information about an individual, with features $\mathcal{X}$ and labels $\mathcal{Y}$. The possibility set will be the domain of the functions we consider for doing statistics. Reminiscent of the familiar concept of a random variable, a bounded function $Z : \Omega \rightarrow \mathbb{R}$ is called a *gamble*. We collect these in the set

$$L^\infty := \{Z : \Omega \rightarrow \mathbb{R} : \sup_{\omega \in \Omega} |Z(\omega)| < \infty\}.$$

The codomain $\mathbb{R}$ of a gamble represents loss (i.e. disutility) to a decision-maker. The concept of utility is borrowed from economics and is intended as an abstract unit for general satisfaction or value [Bernoulli, 1738/1954]; its mirror concept, equivalent up to a sign flip, is loss. Throughout the thesis, we work with a loss-based orientation, where positive numbers correspond to undesirable loss, and negative numbers correspond to desirable utility (i.e. gain); this aligns with the orientation of loss functions in machine learning, and also with insurance literature (e.g. [Pichler, 2014]), but contrasts with the convention in the imprecise probability community. Translating results across literatures can thus require sign flips.

Remark 1. In this introduction, we will avoid technical distractions as much as possible and focus on concepts. Each of the chapters has its own local scope, meaning all necessary definitions and technical assumptions are introduced therein, and these are valid within this scope. For example, Part I considers potentially unbounded functions $Z : \Omega \rightarrow \mathbb{R}$, whereas for example Chapter 5 and Chapter 6 only work with finite Ω , so all functions $Z : \Omega \rightarrow \mathbb{R}$ are automatically bounded. In my personal view, these subtle technical distinctions are of minor or no conceptual importance; but for example the distinction of finite vs. countable additivity causes significant mathematical headaches which require detailed treatment. As a research heuristic, I choose to emphasize commonality over difference, in order to bridge seemingly distinct frameworks. In the introduction, we now assume finite Ω to avoid topological complications; this also implies that we need not worry about set systems, and can simply work with the power set of Ω .

We are now ready to define the prime objects of this thesis.

Definition 2. We call any non-empty set of probability measures $\mathcal{P}$ on Ω an *imprecise probability*. Recall that a probability measure on Ω is any measure³ which in addition satisfies $P(\Omega) = 1$.

Note that we do not impose convexity or closedness here; we will elaborate later. We denote by “sup” the supremum over a set (which might be a maximum under suitable technical conditions).

Definition 3. For any imprecise probability $\mathcal{P}$ and any $Z \in L^\infty$, we define the upper expectation as

$$\bar{R}_{\mathcal{P}}(Z) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[Z], \quad (1)$$

and similarly, the conjugate lower expectation as

$$\underline{R}_{\mathcal{P}}(Z) := \inf_{P \in \mathcal{P}} \mathbb{E}_P[Z]. \quad (2)$$

Here, $\mathbb{E}_P[\cdot]$ is the familiar expectation functional.⁴ We always have that $\underline{R}_{\mathcal{P}}(Z) \leq \bar{R}_{\mathcal{P}}(Z)$ and $\underline{R}_{\mathcal{P}}(Z) = -\bar{R}_{\mathcal{P}}(-Z) \forall Z \in L^\infty$. Due to the conjugacy relation, it suffices to focus on the upper expectation without loss of generality.⁵ Why focus on upper *expectations* when our goal is to study imprecise *probabilities*? In the familiar case of precise probabilities, there is a one-to-one relationship between probability measures P and linear expectation functionals $\mathbb{E}_P[\cdot]$,⁶ so there, the choice of which to focus on is inconsequential.

Definition 4. For a set $A \subseteq \Omega$, called “event”, we define the indicator gamble of A as

$$\chi_A(\omega) := \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

We denote the power set of Ω as 2^Ω due to the bijection between events A and functions of signature $f_A : \Omega \rightarrow \{0, 1\}$. By applying an expectation to indicator gambles, we get back the probability, i.e. $\mathbb{E}_P[\chi_A] = P(A)$. In the imprecise case, we can define similar quantities, but the picture is not as simple anymore.

Definition 5. For any imprecise probability $\mathcal{P}$ and any $A \in 2^\Omega$, we define the upper probability

³A measure P , defined on a σ -algebra Σ (in the introduction, $\Sigma := 2^\Omega$), satisfies $\mu(A) \geq 0 \forall A \in \Sigma$, $\mu(\emptyset) = 0$ and the countable additivity property $\mu(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mu(A_i)$ for a pairwise disjoint sequence $(A_1, A_2, \dots)$, where $A_i \in \Sigma$.

⁴In some cases, we move beyond countable additivity and then need to consider finitely additive “expectation-like” functionals. See Chapter 4. In the introduction, we avoid these complications.

⁵It is not a coincidence that we choose to focus on the upper instead of the lower expectation. Our motivation comes from the notion of the *support function* of a convex set [Hiriart-Urruty and Lemaréchal, 2004]; a notion corresponding to the lower quantity does not exist — convexity is more fundamental than concavity.

⁶See e.g. [Walley, 1991, Corollary 3.2.3] for the case of finitely additive probabilities and finitely additive “linear previsions” (expectation-like functionals) on 2^Ω .

as

$$\bar{P}_{\mathcal{P}}(A) := \bar{R}_{\mathcal{P}}(\chi_A) = \sup_{P \in \mathcal{P}} P(A), \quad (3)$$

and similarly, the conjugate lower probability as

$$\underline{P}_{\mathcal{P}}(A) := \underline{R}_{\mathcal{P}}(\chi_A) = \inf_{P \in \mathcal{P}} P(A). \quad (4)$$

We get that $\underline{P}_{\mathcal{P}}(A) \leq \bar{P}_{\mathcal{P}}(A) \forall A \in 2^{\Omega}$ and thus obtain an interval $[\underline{P}_{\mathcal{P}}(A), \bar{P}_{\mathcal{P}}(A)]$; recall the rain example. Further, we have the conjugate relation $\underline{P}_{\mathcal{P}}(A) = 1 - \bar{P}_{\mathcal{P}}(A^C) \forall A \in 2^{\Omega}$. We can thus focus solely on the upper probability.

For the sake of introduction, we have defined the functionals $\bar{R}_{\mathcal{P}}$ and $\bar{P}_{\mathcal{P}}$ directly from a set of probability measures $\mathcal{P}$. To foreshadow: in Section 1.1, we will discuss axiomatic approaches that characterize functionals $\bar{R}$, behaving like $\bar{R}_{\mathcal{P}}$, based on a list of axioms, instead of defining them explicitly from a set of probability measures. A representation of the form as in Equation 1 will then turn out to be a *consequence*, i.e. a theorem, not a definition. Now, how do $\mathcal{P}$, $\bar{R}$ and $\bar{P}$ relate in terms of generality? Any non-empty set of probability measures $\mathcal{P}$, which we have called imprecise probability, induces an upper expectation $\bar{R}_{\mathcal{P}}$; next, there exists a mechanism which allows to recover $\mathcal{P}$ from the upper expectation *up to closed convex hull*. That is, there is a one-to-one relationship between closed convex sets of probability measures and upper expectations — a well-known result [Walley, 1991, Section 3.6].⁷ Furthermore, in contrast to the precise case, it turns out that the upper probability is a weaker object in the sense that it cannot be used in general to recover uniquely the underlying set $\mathcal{P}$ *even up to closed convex hull*. Thus, perhaps surprisingly, we have to focus either on the set $\mathcal{P}$ or the upper expectation $\bar{R}_{\mathcal{P}}$ to study imprecise probabilities. Due to their equivalence, we will switch freely between focusing on $\mathcal{P}$ and $\bar{R}_{\mathcal{P}}$. In contrast, even if we have collected interval assessments $[\underline{P}(A), \bar{P}(A)]$ for all binary events $A \in 2^{\Omega}$, we have not yet uniquely (up to closed convex hull)⁸ characterized a corresponding set $\mathcal{P}$.

Having defined the mathematical objects which are central to the thesis, what is the relation to the previously discussed plurality of semantics? One way of thinking about the mathematical quantities which express imprecise probabilities is at the level of *syntax*. Prima facie, it appears that the mathematics is not tied to any particular interpretation; a mathematical proof holds regardless of the interpretation. However, I have observed throughout the process of investigation that *semantics matter*, and I claim that clear semantical motivations are the best guide to mathematical developments — hopefully the reader

⁷We emphasize once again that in the introduction we are deliberately avoiding topological complications by assuming finite Ω . Then one can work with the Euclidean topology without loss of generality for our purposes, see Chapter 4.8.1.

⁸Observe that it depends on Ω : for $\Omega = \{0, 1\}$, any closed convex $\mathcal{P}$ corresponds simply to a closed subinterval of $[0, 1]$.

will be convinced of this after reading this thesis. A (the?) key problem in research is to figure out which (mathematical) questions to ask; and it has become clear to me that which questions are meaningfully asked about imprecise probabilities crucially depends on the semantics that one has in mind. This brings us to *structure*.

Remark 6. Some may consider this as bad style, but I will often use citations as nouns nevertheless, such as “see [Walley, 1991]”.

Remark 7. In the introduction, I usually write from the first person plural (“we”), but sometimes switch to the first person singular (“I”) to emphasize subjectivity of viewpoint (e.g. in particularly opinionated statements).

0.3 STRUCTURE

A main research objective of this thesis is to understand better the structure of imprecise probabilities, that is, to make out patterns and relationships in the class of imprecise probabilities, to find insightful organizing principles. What can we say about the set of probabilities $\mathcal{P}$? How would we obtain it? Indeed, structural investigations in this thesis will be driven by semantical considerations. This will become visible throughout, but to foreshadow a key concept: some upper expectations have a mathematical property called *law invariance* (which can be stated formally at the level of syntax), which acts as an organizing principle that approximately separates Part I from Part II. This property will turn out to be easily linked to certain semantics, but not to others (although not logically incompatible). What is intriguing about law invariance is that it provides much additional structure. Thus, a key structural endeavor in Part I is to propose a *stratification* of the class of law invariant imprecise probabilities in terms of $\bar{\mathcal{P}}$, that is, to partition it in an insightful way.

In Part II, we move beyond law invariance, and divert our structural attention to another problem. Namely, we study the semantics of aleatoric plurality (RS3) in depth, and investigate the structure of adequately corresponding *data models*. The imprecise probabilities which appear in this part will be structurally more flexible as compared to the law invariant case.

0.4 SEMANTICS — A SECOND LOOK

We have had a tentative first look at the plurality of semantics in Section 0.1. With the now established background of the syntax of imprecise probabilities, we will revisit it in more depth — with this picture in mind, it is easier to localize the chapter-wise summaries in the appropriate context.

As a running example, we will consider a dataset containing car insurance claims.⁹ In this tabular dataset, originating from a US company which offers private passenger insurance, each row (datum) corresponds to an individual claim that has been made and closed. As features, we have the US state where the driver resides, the rating class, age and gender of the driver; the outcome of interest is the amount of the claim in US dollars (paid by the insurer to the policyholder). Since the dataset only contains claims which have been made and closed, it cannot be used to estimate how often a claim is made.

We will consider this dataset now in the contexts of:

- S1) risk and fairness,
- S2) ambiguity,
- S3) and aleatoric plurality.

The distinction of *risk vs. ambiguity* is due to Ellsberg [1961], albeit foreshadowed by Knight [1921]. The term *aleatoric plurality* is new, and serves to unify some previously studied phenomena (under the names of dataset shift, distributional heterogeneity, aleatoric indeterminacy and more).

What characterizes my perspective on uncertainty in general is that I emphasize the roles of *data (model)* and *decision problem*, thereby always combining frequentist with subjective elements; these two elements will appear in each of the following discussions.

Remark 8. In the following discussions, I will refer much more to insurance than to machine learning. Given the title of the thesis, this might seem surprising or confusing. The reason behind this is that I have developed the following research heuristic in the process of my studies:¹⁰

Insurance is a great analogon to machine learning when it comes to matters related to uncertainty and fairness (and in particular their interplay).

I believe that machine learning research can take many insights from the arguably more mature field of insurance. See Section 0.6.1 and Chapter 7.

0.4.1 RISK AND FAIRNESS

Let us first stick to what we actually have: the dataset at hand. In this empirical, finite population view, we can consider Ω as the set of individuals in the dataset; $\omega \in \Omega$ could

⁹We use the AutoClaims dataset from <https://www.kaggle.com/datasets/saisatish09/insuranceclaimsdata?select=AutoClaims.csv> (Accessed on: 2024-10-24), for documentation see <https://cran.r-project.org/web/packages/insuranceData/> (Accessed on: 2024-10-24).

¹⁰I am typically using latex's `quote` environment for quotes from the literature, which are then accompanied by a citation. Sometimes I use it for a "self-quote", and the formatting serves to highlight.

be any unique identifier of an individual, e.g. $\Omega = \{1, \dots, K\}$. The main gamble of interest $\ell : \Omega \rightarrow \mathbb{R}$ would assign the claim paid (the loss from the viewpoint of the insurer) to each individual. We might now ask:

- If we are uncertain about which ω occurs when we “randomly draw from Ω ”, how should we subjectively value a gamble $Z \in L^\infty$?
- In a social choice context, if each $\omega \in \Omega$ represents an individual of a society, how “fair” is an assignment of loss values $Z \in L^\infty$ in the sense of social desirability?

For a precise meaning of “subjectively valuing”, see the concept of *desirable gambles* in Section 1.1. The two questions, related to risk and fairness, respectively, turn out to be highly analogous, with similar mathematics arising from distinct strands of literatures (see Section 2.9.5).

Clearly, these questions are under-determined: what does it mean to randomly draw an individual from Ω ? A classical mental model is that of an urn. For each $\omega \in \Omega$ we place a ball in an urn and then draw from the urn with replacement. Depending on which of the two major philosophical camps one subscribes to, the setting would could now be framed quite differently.

For a *subjectivist* (we avoid the term *Bayesian* here for good reason, see Section 0.4.2), the concept of probability would refer to their subjective state of uncertainty about the outcome ω . In this benign situation, (arguably all sane) subjectivists would agree to assign probability $1/|\Omega|$ to each outcome ω ; we write $\hat{P}(\{\omega\}) := 1/|\Omega| \forall \omega \in \Omega$ (extended by additivity to 2^Ω). A gamble would then be valued by the induced expectation $\mathbb{E}_{\hat{P}}[\cdot]$, which is the familiar average. From the viewpoint of the car insurance company, $\mathbb{E}_{\hat{P}}[\ell]$ would tell us how much loss we have to shoulder, given that a claim was made and closed.

In contrast, for a (hypothetical) *frequentist*, probability would be about *relative frequencies* and *randomness*; essentially, about a *data model*. We will have more to say on the concept of a data model in Chapter 5, but for now we will use the term intuitively. We will set this up in a perhaps somewhat unfamiliar and seemingly unnecessarily complicated way, but this will pay off in Section 0.4.3. In this approach, we introduce another set that expresses an abstract source of randomness. We will take $\mathcal{Z} := [0, 1)$ to be this *set of seeds*. On this set of seeds, we use the Lebesgue measure λ to express randomness. A sequence of outcomes is then $\omega^\infty = (\omega_1, \omega_2, \dots)$, where each $\omega_i \in \Omega, i \in \mathbb{N}$.¹¹

Proposition 9 (The urn model). *Let $\mathcal{Z} := [0, 1)$. There exist $W_i : [0, 1) \rightarrow \Omega, i \in \mathbb{N}$, so that the following holds:*

$$\lambda \{z \in \mathcal{Z} : W_1(z) = \omega_1, W_2(z) = \omega_2, \dots, W_n(z) = \omega_n\} = (1/|\Omega|)^n, \quad \forall n \in \mathbb{N}.$$

¹¹We will use this notation for a sequence in the introduction and in a subset of Part II.

From this model, one can derive, for example, that the limiting relative frequency of drawing some specific outcome ω is $1/|\Omega|$ *almost surely*, where *almost surely* means that the size of the set of seeds which give this result is 1 under the Lebesgue measure λ . In the car insurance example, this model would tell us that in the limit, when sampling individuals “randomly” from the dataset (!), we would get an average loss of $\mathbb{E}_{\hat{P}}[\ell]$ *almost surely*. Observe that in the two-stage way that we stated the familiar *i.i.d.* urn model here, the sense of “probability” which concerns “sampling” the data sequence is separated clearly from the “probability” concerning relative frequencies of events $A \in 2^\Omega$; the first is expressed by the Lebesgue measure λ , whereas the latter appears via probability measures on Ω . In Chapter 5, we highlight this distinction, and refer to the first sense as *typicality* and only use the term “probability” for the latter.

Although the approaches are quite distinct, both subjectivists and frequentists can agree in this setting that probabilities are “known”, and they are represented by $\hat{P}$. There may be other settings where many would agree that probabilities are “known”, such as flipping a fair coin,¹² but I believe urn examples are the most paradigmatic, also featuring in many works which investigate decision-making under risk (e.g. [Levy et al., 2010], to name just one).

Finally, we have arrived at the name of this subsection: “Risk refers to a situation in which we know the precise probability of potential outcomes of each option” [Wu et al., 2021]. Or, some authors add (e.g. [Levy et al., 2010]), if the probability can be estimated. Thus, the first condition is that there *is* a single, “true” probability which we could possibly know, and the second step is that we indeed know it or can estimate it. To unpack this, we would need to know what it means to “know” the probability, and this is not at all an uncontroversial matter. How could we *not know* the probability? To some authors in the subjectivist camp, every uncertainty *is* a risk (e.g. [Savage, 1954]), since they argue that it is always possible to assess subjective probabilities. In interesting contrast, on the reading of de Finetti’s work by Feduzi et al. [2012], we do find a distinction which is grounded in intersubjective agreement; intuitively, when different subjectivists agree on the same probabilities, the situation is one under risk (e.g. in the previous urn example). For example, in a “benign” evidence situation, multiple experts may agree; when experts disagree, caution seems warranted.

From the frequentist perspective, the situation is different again. First, the existence question: in this camp, the standard assumption is that the data model is parameterized by a single, “true” — but potentially unknown — probability, and this assumption is rarely

¹²The example of flipping a coin also nicely shows how careful one must be with claims about “knowing” the probability. It is quite common to assume that the probability that a fair coin yields heads (or tails, respectively), is 50%. This is neither exactly wrong, nor exactly right: Bartoš et al. [2023] show that there is a slight tendency for coins to land on the same side that faced upwards before throwing (with long-run relative frequency of $\approx 50.8\%$). Depending on the experimental setup (e.g. one could balance which side faces upwards before throwing), this tendency might or might not manifest in the data.

questioned (but see e.g. [Hájek, 2009, Ivanenko, 2010, Gorban, 2017]). “We assume the data is *i.i.d.*” — a standard assertion in many, if not most, statistics and machine learning papers. Are there uncertainties which are intrinsically (putting aside the question of knowledge) not risks from a frequentist perspective? Perhaps surprisingly, indeed there exists a variety of data models which cannot be captured under the umbrella of risk and thus call for a generalized frequentism. We will investigate this as *aleatoric plurality* in Part II (which also has intriguing consequences for subjectivist decision-makers under such data models). Bracketing out this additional flexibility for the moment, and still assuming the existence of a single, “true” probability, the question whether we have knowledge of this probability remains.

Directing our attention to the problem of estimation, note that until now we have assumed that Ω contains the actually observed data. This setting is relevant, for example, when we are interested in a known finite population and want to measure the fairness of a loss distribution. A crucial step in doing statistics and machine learning, however, is to move to a hypothetical population: the probability distribution that the data “is sampled from”. In this case, we might assume an uncountable continuum of individuals that we might hypothetically observe; this enables speaking of generalization. We will not go too deep into this intricate step, but point out that the distinction of finite vs. hypothetical population maps quite well on the distinction regarding *risk as known probabilities* and *risk as estimated probabilities* (what Knight [1921] called “statistical probability”).

This begs the question of what it means to be able to estimate a probability. First, this phrasing seems to suggest the existence of the estimand — the probability. In this step of moving from finite to hypothetical population, however, the problem of the *reference class* arises which has caused a great deal of headaches to philosophers [Hájek, 2007]: revisiting the car insurance example, if we think of the dataset as “sampled” from the hypothetical population, what hypothetical population is this supposed to be, exactly? Perhaps the broad reference class of all world-wide car insurance claims, or more specifically just US car insurance claims? Or even more specifically, accounting for how the customer base of this specific company tends to be skewed in particular ways (e.g. depending on targeted advertising)? In this way, it is not clear what the estimand to be estimated is. While mostly leveraged to criticize the frequentist approach, the reference class problem turns out to be challenging also for the subjectivist approach [Hájek, 2007]. I believe that the reference class problem points to a fundamental tension between individual and aggregate that is unavoidable and runs throughout the history of probability and statistics, see Chapter 7.

To pin down what it means to be able to estimate (reliably? accurately? confidently?) a probability is challenging. While in the finite population example, we could leverage the mental model of an urn to explain *risk as known probabilities*, to demarcate better the boundaries of *risk as estimated probabilities* it will help to contrast it with its conceptual

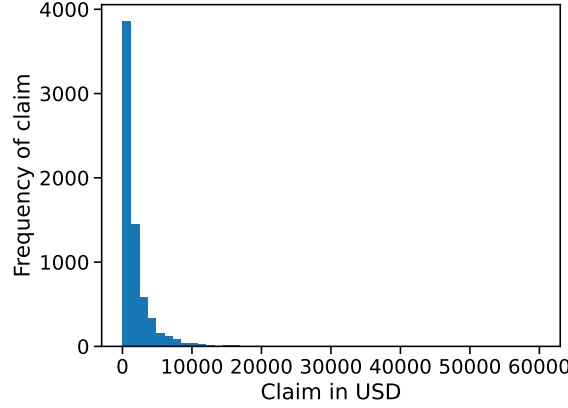


Figure 1: Distribution of claims in the car insurance example. Recall that the dataset only contains claims which have been made and closed, so that data for persons which did not make a claim is not included. There are some very large claims (as the scale of the x -axis indicates), but they are not visible. To give some intuition, in total, there are 6,773 claims, of which 126 claims are greater than 10,000USD and 18 claims are greater than 20,000USD. For the histogram on a log-scale, see Figure 2.

counterpart, *ambiguity* (Section 0.4.2).

Remark 10. A cautionary remark is warranted. We could now dive deep into a philosophical effort to establish clearly an analytical distinction here, between risk, ambiguity and so on. In this style, we could attempt to argue as clearly as possible that this distinction is fundamental, that there are criteria which neatly allow us to divide situations into either category (or perhaps the opposite, that the distinction is entirely meaningless). This is not at all our intention here. A thread that runs (mostly implicitly, with the exception of Chapter 7) throughout the thesis, is taking a perspective on uncertainty in general that I call an *actuarial view of probability* (see Section 0.6.1 and Chapter 7), heavily inspired by insurance. In particular, to me this view entails that the distinction discussed here is intrinsically blurry and context-sensitive, and it is thus hopeless to look for crystal clear separation criteria. In this way, I am very pragmatic about these distinctions, and rather consider them as useful heuristics, in particular for guiding the mathematical developments.

DECISION-MAKING UNDER RISK

Before we contrast risk with ambiguity, however, we introduce an aspect of decision-making into the picture. Essentially, risk is about a form of variability in the individual outcomes, which we assume to be loss-valued; a variability that can be captured by a normalized measure (a probability). Consider for example the distribution of claims in the car insurance example (Figure 1), where the claims in the dataset are presented in the form of a histogram; if Ω represents the finite population in the dataset, Figure 1 depicts the distribution of the claim gamble $\ell : \Omega \rightarrow \mathbb{R}$. The magnitudes of the claims vary widely (mean $\approx 1,853$ USD, median $\approx 1,002$ USD, min = 9.5USD, max = 60,000USD); all of these

individual claims together make up an aggregate, which can, for example, be summarized as a single number using the mean or the median.

How ought a decision-maker behave under risk? Consider a decision-maker who is supposed to bet on the uncertain outcome of the gamble $\ell : \Omega \rightarrow \mathbb{R}$. For example, to bet on the binary event that ℓ will exceed a specific threshold. What is an appropriate betting rate? Or, what is the subjective value of the gamble ℓ to the decision-maker? We are essentially asking about how to appropriately *aggregate* the individual outcomes to arrive at a reasonable summary.

To make this more concrete, we will take the perspective of the insurer. Our (here heavily simplified toy-like) task is to set the *insurance premium* $\pi \in \mathbb{R}$, so that $\ell(\omega) - \pi$ is a “desirable” transaction (foreshadowing the concept of *desirable gambles*, see Section 1.1). For simplicity, we will not differentiate at all and thus the premium will be the same for every policyholder in the insurance pool. For realizations $(\omega_1, \dots, \omega_N)$, if the pool consists of N policyholders, we obtain the final loss value of $L := \sum_{i=1}^N \ell(\omega_i) - N \cdot \pi$. From the viewpoint of the insurer, we would like to have that this term is less than zero; otherwise we are bankrupt. On the other hand, being excessively cautious would mean that we lose business opportunity.

A simple approach would be to set $\pi := \mathbb{E}_p[\ell]$, which is known as the *pure premium* [Rejda, 2001, p. 584]. However, if we now “sample” (which we do by using a pseudo-random number generator) the outcomes $(\omega_1, \dots, \omega_N)$ independently (with replacement) from our dataset, and evaluate L , then we often go bankrupt; due to the central limit theorem, about 50% of the time, with convergence (in theory) to 50% as $N \rightarrow \infty$.

Under these circumstances, a decision-maker like an insurer might want to express an attitude of *risk aversion* for more conservative decision-making. One could say that the study of risk aversion provides systematic answers for deriving safety loadings that we add to π . In simple words, the goal is to go bankrupt less often. The above example appears like a problem involving multiple individual realizations $1..N$, but it can equivalently be viewed as a single realization of a “compound” problem, where we obtain a single realization of L (at the level of the whole insurance pool). This is why one often hears that insurers can afford to be “risk-neutral” due to the *law of large numbers*, as they face many realizations, while individual policyholders might be risk-averse (the distribution of L is narrower, as we increase N , compared to that of ℓ). However, a small degree of risk aversion at the level of individual realization can translate to a substantial risk aversion at the level of the pool; and indeed there are good reasons for insurers to exhibit risk aversion at the level of the pool (for example due to regulatory requirements such as Solvency II in the EU [European Union, 2009]).

The phenomenon of risk (and ambiguity) has by now been thoroughly studied in economic

decision theory (see e.g. [Anand et al., 2009, Etner et al., 2012]) and psychology (see e.g. [Lopes, 1984, Kuhberger, 2002]), with interdisciplinary research agendas emerging.

The classical framework for decision theory under risk is *expected utility (EU) theory*, which comes both in a subjectivist [Savage, 1954] as well as an objectivist [von Neumann and Morgenstern, 1947] flavor. A very brief description of objectivist EU is as follows: let us imagine that we face a lottery with possible outcomes $c \in C$ and a probability measure P on these outcomes. Typically, we would think of $c \in C$ as monetary rewards; assume for now $C := \mathbb{R}$. A decision-maker who follows the postulates of EU theory would have a personal utility function $u : C \rightarrow \mathbb{R}$, and then aggregate (value a lottery) using $\mathbb{E}_P[u(c)]$. Here, the codomain $\mathbb{R}$ of the utility function represents the flip side of loss (equivalent to loss up to a sign flip). If the utility function is concave, the decision-maker is said to be risk-averse, since this implies preferring the expected value of a lottery over playing the lottery itself (see [Denuit et al., 2006, Sections 1.6, 2.2]).

Remark 11. Recall that utility is equivalent to loss, up to a sign flip. Concavity of the utility function translates to convexity of the loss function. Risk aversion is related to the convexity of $\bar{R}$, and to the concavity of $\underline{R}$.

In our car insurance example, we have said that the gamble of interest $\ell : \Omega \rightarrow \mathbb{R}$, which represents the claim payments, is money-valued (e.g. $\ell(\omega) = 1$ means we have to pay out 1 USD), and at the same time we have assumed it is loss-valued already (1 USD = 1 loss). The conjunction of these assumptions means that we have effectively assumed that paying money (in USD) *is* loss to the insurer. That is, there is no space left in this picture for a concave utility function to express risk aversion. One way out would be to add a second layer of utility abstraction anyway, so that we would compose $\tilde{\ell} \circ \ell$ to obtain abstract loss values, which are not in USD anymore, but now live on another scale. In the case of a company, however, it seems less natural than for a single individual to assume a concave utility. Intuitively, the concavity of the utility function models diminishing marginal utility of money: winning 1000\$ is perhaps not considered to be 10 times “as beneficial” as winning 100\$ *to an individual*. Thus, from the viewpoint of the insurer, we might be dissatisfied with how EU theory expresses risk aversion; it cannot anymore be disentangled from diminishing marginal utility. Denuit et al. [2006] write:

Note that in the EU framework, the agents attitude towards risk and the agents attitude towards wealth are forever bonded together (since they are both derived from the characteristics of u): risk aversion and diminishing marginal utility of wealth are synonymous.

I agree with Denuit et al. [2006] and Buchak [2013] that EU does not capture risk aversion *proper*, if we understand risk aversion as a concern about the intrinsic variability (inequality) in the loss distribution. Merely changing the scale of measurement from money to

utility/loss does not address this at all (this is perhaps most salient in the social choice context). Instead, we are looking for systematic ways to obtain a “safer” premium π .

IMPRECISE PROBABILITIES FOR DECISION-MAKING UNDER RISK

As a response, *rank-dependent* competitors to EU theory have been developed [Quiggin, 1982, Yaari, 1987, Schmeidler, 1989, Tversky and Kahneman, 1992]; see also [Quiggin, 2012]. Here, the idea is to explicitly account for the, as Lopes [1984] calls it, “distributional inequality”, i.e. the variability in the loss distribution based on the ranks (with importance shifting from lower to higher losses). With this, finally imprecise probabilities enter the picture: in the rank-dependent approaches, we find (implicitly) sets of probabilities as in Definition 1. In the car insurance example, rank-dependent EU can be used to obtain a risk-averse aggregation functional that specifies a more cautious insurance premium π , with the goal of avoiding bankruptcy. Under risk, rank-dependent EU can express risk aversion; in a social choice context, it can express *inequality aversion* [Buchak, 2017]; the concept of a Lorenz curve is related, see Section 2.6.2. In Chapter 2, we discuss the intimate relation of upper expectations known as *spectral risk measures* and rank-dependent EU and provide new insights into their structural properties. Interestingly, whereas classical EU can express diminishing marginal utility of money via concavity, in rank-dependent formulations we can have a kind of diminishing marginal utility of *probability mass* via concavity, see Chapter 2.

Spectral risk measures form a sub-class of the *coherent risk measures* studied in finance and insurance¹³ [Artzner et al., 1999, Artzner, 1999, Delbaen, 2002, Dhaene et al., 2004, Föllmer and Weber, 2015, Föllmer and Schied, 2016]. Spectral risk measures are special: in Chapter 2, we will highlight their specialness from multiple different angles. However, we will also consider other interesting coherent risk measures in the thesis, for example in Chapter 3, *divergence risk measures* (based on f -divergences) will be at the center. Recall that in our terminology a coherent risk measure is just an upper expectation of the form as in Definition 1 (we will justify this in Section 1.1). As the name suggests, scholarship under the banner of coherent *risk measures* is primarily concerned with risk, although not exclusively.¹⁴ Again, this is not merely of philosophical interest, but can crucially change which mathematical questions are asked, for example regarding estimation.¹⁵

¹³In insurance, coherent risk measures can serve at least two goals: first, as in financial mathematics in general, they can be applied at the level of the whole portfolio for risk management to determine capital requirements [Artzner, 1999]; second, they can be applied at the level of individual contracts as *premium principles* [Wang and Dhaene, 1998, Pichler, 2015]; the discussion by Denuit et al. [2006] about translation equivariance shows that which mathematical properties are desired and how they are justified can depend on which goal one has in mind.

¹⁴Some exceptions, where the term *coherent risk measures* appears in a context of modeling ambiguity, are [Pichler, 2014, Breuer and Csiszár, 2016].

¹⁵When we come up with an estimate, we should have an idea of *what* we want to estimate. That is, as

To sum it up, in decision-making under risk, imprecise probabilities (equivalently: upper expectations, coherent risk measures), and in particular rank-dependent approaches, can be useful to obtain risk-averse aggregation functionals. Intuitively, considering the form of Equation 1, $\bar{R}_{\mathcal{P}}(\ell) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell]$, the known or estimated probability P_0 would be replaced by a more cautious set $\mathcal{P}$, typically constructed in a way that could be described as $\mathcal{P}$ being “centered around” P_0 . Consequently, the weight would be shifted to larger losses.

Remark 12. With simple sign flips, risk (inequality) aversion can become a risk (inequality) seeking attitude. This is equivalent to the upper and lower expectation duality. Throughout the paper, we focus on the aversion (pessimism) attitude to enable more cautious decision-making.

The idea of risk, perhaps, will become clearer when we now look at its counterpart, *ambiguity*.

Remark 13. In prose, I use both the terms “the probability” and “probabilities”. Here, “the probability” can either refer to the real number $P(A)$, associated to a single event $A \subseteq \Omega$, but also to a whole probability *measure* P on Ω , which specifies “the probabilities” on a whole set system. To emphasize the measure-theoretic character, I sometimes explicitly speak of a “probability measure”.

0.4.2 AMBIGUITY

Ellsberg [1961] famously contrasted risk with *ambiguity* and thereby sparked the development of a now substantial body of literature. In this section, we take a firmly subjectivist perspective on decision-making under uncertainty. The community under the umbrella term of imprecise probabilities [Walley, 1991] is mostly motivated by considerations of the kind found here. Broadly speaking, ambiguity refers to situations where the decision-maker does not “know”, or cannot assess in a meaningful way, probabilities — where we do not necessarily dispense of the assumption that true probabilities exist, but the focus shifts towards the *epistemic* state of the decision-maker. In other words, Etner et al. [2012] describes ambiguity as “non-probabilized” uncertainty, contrasting with “probabilized” uncertainty (risk).

Already Knight [1921], in turn foreshadowed by Keynes [1937], famously distinguished risk as “measurable” uncertainty from “unmeasurable” uncertainty, which he considered as “true” uncertainty [Knight, 1921, p. 20; Chapter VII]; the latter is now often called *Knightian uncertainty*, and used approximately as a synonym for ambiguity.¹⁶ Knight [1921] made

$N \rightarrow \infty$, to which quantity should our estimate converge to. The settings of risk and ambiguity may give different answers.

¹⁶See the discussion which follows later in the text.

a tripartite distinction for probabilities: first, probabilities based on *a priori* judgements (e.g. considerations of symmetry, combinatorial arguments); an example would be throwing a fair die or gambling in the casino.¹⁷ Second, statistical probabilities, based on relative frequencies in a reference class of available data (past experience). And third, subjective estimates which are formed without a statistical basis. For example, what is the probability that person A will be the next president of country B? For Knight [1921, p. 233], the first two forms could also be called “objective”, and the third is “subjective”.¹⁸ And it is this third form of allegedly non-measurable uncertainty that Knight is most interested in, and where Knight diverges from firm subjectivists like Savage [1954] in demanding a different treatment than for risk.¹⁹ Interestingly, however, Knight subsumes all three under the umbrella of probability, but it is arguably not clear whether he viewed probability as Kolmogorov-style additive functionals, perhaps he simply did not have an alternative framework in mind [Spohn, 2017].

DECISION-MAKING UNDER AMBIGUITY

It was Ellsberg [1961] who introduced the term *ambiguity* and demonstrated the behavioral significance of Knight’s earlier distinction: decision-makers may behave differently under risk and ambiguity (from both a descriptive as well as arguably from a normative viewpoint). Ellsberg [1961] arrived at this insight by comparing different urn experiments, contrasting “risky” with “ambiguous” urns. To make the example concrete: imagine that in urn I, there are 50 red and 50 black balls. Urn II on the other hand contains 100 balls of either red or black color, but the exact composition is unknown to the decision maker. In this case, subjects usually prefer to place bets on the risky urn as opposed to the ambiguous urn. Such a behavioral disposition is known as *ambiguity aversion*. There is now a vast literature on ambiguity (aversion), from both descriptive, experimental perspectives (see e.g. [Curley et al., 1986, Hogarth and Kunreuther, 1989, Camerer and Weber, 1992, Liu and Colman, 2009, Levy et al., 2010, Dupont-Courtade, 2012, Smithson, 2015, Pushkarskaya et al., 2015, FeldmanHall et al., 2016, Blankenstein et al., 2017, Kocher et al., 2018, Fox et al., 2021]) as well as theoretical perspectives (see e.g. [Schmeidler, 1989, Gilboa and Schmeidler, 1989, Ghirardato and Marinacci, 2002, Etner et al., 2012, Guidolin and Rinaldi, 2013, Ellsberg, 2015]). While ambiguity aversion is often observed, the details in the empirical landscape

¹⁷Recall from Footnote 12 that even seemingly simple settings, where the probability is “obvious”, might not be that simple and obvious after all.

¹⁸I use the terms “subjectivist” and “objectivist” to describe the perspective that a framework is situated in; in contrast, I refer to subjective and objective uncertainties.

¹⁹This means that another way of phrasing the subjectivist approach à la Savage, in contrast to the subjectivist approach of Walley [1991] for example, is that every subjective uncertainty (within the scope of his framework) is measurable. It is no coincidence that non-measurability in the mathematical sense is related to imprecise probabilities, as well, see e.g. [Walley, 1991, Section 2.9.6] on the inner and outer Lebesgue measure and see also [Isaacs et al., 2021].

(e.g. which factors influence ambiguity aversion, or under which conditions people tend to be even ambiguity *seeking*) remain complicated, see [Trautmann and Van De Kuilen, 2015] for an overview and [Kocher et al., 2018].

For our purposes, we are less interested in human behaviour from a descriptive perspective, but rather believe that there are solid (normative) reasons for being ambiguity-averse. In the case of insurance, ambiguity aversion implies adjusting the premium in a pessimistic way (to increase it). There is indeed a clear empirical case for ambiguity aversion from the viewpoint of the insurer [Hogarth and Kunreuther, 1989, Cabantous, 2007, Dietz and Niehörster, 2021] — already Knight [1921, p. 248] observed the “cost” of ambiguity in insurance; intuitively, the rationale behind this is that it is more favorable to quote a higher premium than to go bankrupt because the risk was underestimated, implying that in an ambiguous setting, the premium should be cautiously increased. Revisiting the car insurance example, consider for example the ambiguity arising from the finiteness of the data. The dataset contains 6,773 observations, but we would like to extrapolate from it to the hypothetical population. As discussed in Section 0.4.1, here the fundamental problem of the reference class arises. For example, it might be the case that for some demographic sub-populations, we have only few observations in the data, but we anticipate many more of them in the test data; hence we would like to arrive at cautious estimates.

In the context of machine learning, very similar considerations motivate an ambiguity-averse attitude from the viewpoint of the machine learning engineer, see Section 0.5. For example, machine learning systems are often criticized for a lack of robustness (e.g. [Carlini and Wagner, 2017]) and overconfidence (e.g. [Kristiadi et al., 2020]), not for making overly cautious predictions. Throughout the thesis, we will combine risk, ambiguity and aleatoric plurality with an attitude of *aversion*, although in other settings, the corresponding opposite attitude might be of interest as well.²⁰

Before moving on to the key (type of) argument for the importance of imprecision, with the role of expressing ambiguity, we briefly compare Knight’s [1921] “uncertainty” with “ambiguity”; the terms appear to be used synonymously in contemporary literature. In one sense, the word “ambiguity” has a more gradual connotation (a situation can be more or less ambiguous), whereas Knight’s distinction might appear binary (either probabilities are known/estimable or they are not). Although, on my reading, Knight [1921] himself seemed well-aware of the blurriness and gradualness.²¹ Personally, I thus prefer the term “ambiguity” to emphasize the gradualness, i.e. that we are dealing with a spectrum. In fact, Ellsberg [1961, p. 643] was aware of Knight’s earlier distinction, but criticized that Knight nevertheless wanted to stick to probabilistic treatment [Ellsberg, 1961, p. 653]. In

²⁰For example, the optimistic lower expectations could be used for handling outliers in the data.

²¹See the discussion in Section 0.6.1 for a statement by Knight [1921] on the gradual character of his uncertainty conception.

this way, [Ellsberg \[1961\]](#) goes further than [Knight \[1921\]](#) by highlighting the behavioral significance of the distinction. With the background of [Savage's \[1954\]](#) highly influential work on subjective EU, [Ellsberg \[1961, p. 669\]](#) also raised the normative question; about the subjects who violate [Savage's \[1954\]](#) axioms in the context of his urn experiments, he asked “Are they foolish? [...] Indeed, it seems out of the question summarily to judge their behavior as irrational: I am included among them”.

There is now a substantial body of literature, for example in epistemology as well as generally in the field of imprecise probabilities, which is less concerned with the descriptive, experimental, but rather asks whether or when imprecise beliefs (in this context also known as imprecise *credences*) are rationally permissible (see e.g. [[Elga, 2010](#), [Schoenfield, 2012](#), [Bradley and Steele, 2014](#), [Chandler, 2014](#), [Schoenfield, 2017](#), [Mayo-Wilson and Wheeler, 2016](#), [Konek, 2019](#), [Schoenfield, 2020](#), [Isaacs et al., 2021](#)]); see also [[Walley, 1991](#), Chapter 5] for the “importance of imprecision”. Some of these authors argue against the permissibility of entertaining imprecise beliefs (e.g. [Elga \[2010\]](#)); personally, I have not found any argument in this direction to be convincing, but leave it to the reader to consult the literature and judge for themselves.

Instead, I'd like to present from a bird's eye view what I believe is the most compelling cluster of arguments, or type of argument, to motivate the use of imprecise probabilities under the semantics of ambiguity (broadly construed). The key challenge against the sufficiency of precise probabilities was formulated by [Gilboa et al. \[2009\]](#) as follows:

The Bayesian approach is lacking because it is not rich enough to describe ones degree of confidence in ones assessments.

Intuitively, the basic idea of the criticism is that the Bayesian (i.e. the precise probabilistic, subjectivist) approach cannot distinguish *knowing the probability* from *not knowing the probability*, and that this indeed matters — but what could this even mean to a subjectivist? Let's elaborate. Already [Keynes \[1921, p. 78, emphasis in original\]](#) argued that:

As the relevant evidence at our disposal increases, the magnitude of the probability of the argument may either decrease or increase, according as the new knowledge strengthens the unfavourable or the favourable evidence; but *something* seems to have increased in either case,—we have a more substantial basis upon which to rest our conclusion. I express this by saying that an accession of new evidence increases the *weight* of an argument. New evidence will sometimes decrease the probability of an argument, but it will always increase its ‘weight.’

This leads [Keynes \[1921\]](#) to distinguish *weight* from *balance* (see also the presentation by [Joyce \[2005\]](#)). Standard illustrative examples go roughly as follows. Consider the question

“will person A or person B will become the next president of country C?”. Given that we know nothing whatsoever about the politics of country Z, it seems that the evidence is balanced, but has little or no weight. In contrast, assume we are prompted to bet on the election outcome in country D, where we know that polls indicate that about 50% of voters intend to vote for person A and person B, respectively. Here, the evidence is similarly balanced, but in addition it now has substantial weight. The Bayesian approach à la [de Finetti \[1931\]](#) and [Savage \[1954\]](#) cannot account for this difference. Or consider the car insurance example. From a business perspective it is quite clear that there is a crucial difference between a benign evidence situation, where a lot of data is available, covering the relevant sub-populations well, and a situation where data is scarce. Similarly, in machine learning, it seems desirable to capture the difference between knowing that there is substantial intrinsic variability and a state of ignorance — see the discussion below on *epistemic vs. aleatoric* uncertainty [[Hüllermeier and Waegeman, 2021](#)]. As [Konek \[2015\]](#) puts it: “Unspecific evidence calls for imprecise credence”.

I believe that a key underlying issue here is that even subjectivists usually would not like to sever the connection between belief and “objective chance”,²² or more generally statistical relative frequencies. For example, the influential “principal principle” of [Lewis \[1980\]](#) demands that if a decision-maker knows the chances (e.g. as in urn examples with known composition), then they should adopt them also as their subjective belief (credence). However, this implies directly that the Bayesian approach cannot anymore distinguish “knowing that the chance for the event A is $p\%$ ” from a state of belief, intended to express ignorance, which reduces nevertheless to the number $P(A) = p\%$. I agree with many others that it is important to keep the tight connection between belief and relative frequencies; in my view, a “good” belief is one which is “appropriately matched” to the data (model); this idea will play an important role in Chapter 6.²³ A response might be that these concerns are of little or no relevance for real-world data. While I agree that in cases such as the insurance example it might be much less clear what it would mean that a “true probability” exists, one would think that a general theoretical framework should be able to deal well at least with seemingly simple examples such as urns or coin flipping as a minimum requirement. However, in light of the distinction of weight and balance, we find that the Bayesian approach cannot express the typical (arguably rational) preferences in the urns examples of [Ellsberg \[1961\]](#).

A standard way of representing a lack of knowledge in the Bayesian framework is to adopt a *uniform distribution* over the possible alternatives, derived by applying the principle of *maximum entropy* (a uniform distribution maximizes Shannon entropy) [[Jaynes, 1957](#)]. In

²²Essentially what [Knight \[1921\]](#) called “a priori” probabilities.

²³In simple words, if the chance is actually $p\%$, then my belief should also be $p\%$ to enable successful decision-making under the data in the long-run. In more complex data models than the standard *i.i.d.* model, this has intriguing implications (when generalizing to the imprecise case).

a situation where evidence is symmetric but without weight (e.g. facing a coin with unknown bias), it would seem prudent to assign 50% probability (binary uniform). One problem with this approach is that a uniform distribution is not in general invariant under re-parameterization of the problem (see e.g. [Walley, 1991, Section 5.5]). For an expression of total ignorance, it would seem surprising that a simple bijective re-parameterization could lead to a “gain of information”. Moreover, a uniform belief has strong behavioral implications. Joyce [2010, pp. 283-284; emphasis in original] illustrates this as follows.:

Black/Grey Coins. An urn contains a large number of coins which have been painted black on one side and grey on the other. The coins were made at a factory that can produce coins of any bias $\beta : (1 - \beta)$ where β , the objective chance of the coin coming up black, might have any value in the interval $0 < \beta < 1$. You have *no information* about the proportions with which coins of various biases appear in the urn. If a coin is drawn at random from the urn, how confident should you be that it will come up black when tossed? [...] For example, f_U [the uniform density on β] commits you to thinking that in a hundred independent tosses of the black/grey coin the chances of black coming up fewer than 17 times is *exactly* 17/101, just a smidgen ($= 1/606$) more probable than rolling an ace with a fair die. Do you really think that your evidence justifies such a specific probability assignment?

Perhaps the “least informative” uniform distribution is in fact very informative after all [Joyce, 2010, p. 284]. Still, we will revisit a maximum entropy principle in depth in Chapter 6. In a nutshell, it turns out that even under more general imprecise data models, *if* one takes the actual decision-making problem into account (by choosing a loss function), *then* there is a certain but only partial sense of optimality to a generalized maximum entropy principle, which is adapted to the loss function (where Shannon entropy is merely a special case).

What is essentially behind this, I would argue, is a manifestation of the *law of the excluded middle*. Recall that a probability measure P , due to additivity on disjoint sets, satisfies $P(A) + P(A^c) = 1$ for any $A \in 2^\Omega$, where A^c denotes the set complement relative to Ω . In this way, the total “belief mass” of 1 is distributed between A and A^c , but there is no room for abstaining. What if our evidence is completely silent about both A and A^c ? The Bayesian approach demands that we make a precise distribution among them anyway. In contrast, with imprecise probability, we can have that $\bar{P}(A) > \underline{P}(A)$, and thus the *degree of imprecision* $I(A) := \bar{P}(A) - \underline{P}(A) = 1 - (\underline{P}(A) + \underline{P}(A^c)) \geq 0$ can represent a suspension of judgment [Bradley, 2019]. In other words, with imprecise probabilities, we can drop the law of the excluded middle.

Finally, we remark that beyond representing incomplete/partial evidence, imprecise prob-

abilities may also represent ambiguity arising from *conflict*, e.g. when experts disagree. Empirically, decision-makers tend to display even higher conflict aversion than ambiguity aversion [Smithson, 1999, Cabantous, 2007, Dupont-Courtade, 2012, Smithson, 2015].

IMPRECISE PROBABILITIES FOR DECISION-MAKING UNDER AMBIGUITY

Recall the form of upper expectations and upper probabilities ($Z \in L^\infty$, $A \in 2^\Omega$):

$$\bar{R}_{\mathcal{P}}(Z) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[Z], \quad \bar{P}_{\mathcal{P}}(A) := \sup_{P \in \mathcal{P}} P(A). \quad (5)$$

Here, we interpret $\mathcal{P}$ as an *ambiguity set*, representing the ambiguity of the subjective belief state of the decision-maker. Intuitively, $\mathcal{P}$ contains the set of “plausible” probability measures, and then the decision-maker takes a pessimistic (worst-case) stance towards this set. Throughout the thesis, we will stick to using this pessimistic attitude, expressed by the sup, in decision-making; this principle is also known as Γ -minmax (Γ -maxmin, up to a sign flip) [Gilboa and Schmeidler, 1989].²⁴

With imprecise probabilities, we can thus interpolate along a spectrum from complete knowledge of a precise probability ($\mathcal{P} = \{P\}$) to complete ignorance; in the latter case, $\mathcal{P}$ is the set of all probability measures on 2^Ω , which results in the “vacuous” lower and upper previsions $\bar{R}_{\text{vac}}(Z) = \sup(Z)$, $\underline{R}_{\text{vac}}(Z) = \inf(Z)$,²⁵ for any $Z \in L^\infty$, and in the “vacuous” lower and upper probabilities $[\underline{P}_{\text{vac}}(A), \bar{P}_{\text{vac}}(A)] = [0, 1]$ for any $A \in 2^\Omega$. This, unlike the uniform distribution, can be considered as a model for complete ignorance [Walley, 1991].

Recall the interpretation of Equation 5 under risk, instead. Under risk, we do believe that one $P_0 \in \mathcal{P}$, known to us, represents the “true probability”, but we introduce the set $\mathcal{P}$ in a somewhat artificial way to enable more cautious decision-making. In Chapter 2, we call this *hallucinated ambiguity*.²⁶ In contrast, under ambiguity, we do believe that one of the $P \in \mathcal{P}$ is the “true probability” but we do not know *which*.²⁷

Could we combine risk aversion with ambiguity aversion? Schoenfield [2020] criticizes this as “double counting”. I disagree (although qualified by the acknowledgement that the distinction of risk and ambiguity can be slippery in practice, see Section 0.6.1). Indeed,

²⁴See Section 2.9.2. The main difference to the framework of Gilboa and Schmeidler [1989] is that we suppose that gambles $Z \in L^\infty$ are already given, whereas Gilboa and Schmeidler [1989] jointly axiomatize a set $\mathcal{P}$ and a utility function.

²⁵Here, $\sup(Z)$ refers to $\sup(\{Z(\omega) : \omega \in \Omega\})$, and similarly for $\inf(Z)$.

²⁶In similar spirit, Bassett Jr. et al. [2004, p. 3] wrote: “Admittedly, there is something mildly schizophrenic about decision makers who accept at face value probabilities [...], and then intentionally distort these probabilities in the process of turning them into actions.”

²⁷This statement should be taken with some caution, and is intended to an extent for didactical purpose here. Not everyone who proposes the use of imprecise probabilities for expressing “non-probabilized” epistemic uncertainty would subscribe to the idea that there need be an underlying “true” probability [Walley, 1991].

for example [Pichler \[2014\]](#) proposed an insurance premium principle of the form

$$\bar{R}_{\text{amb}}(Z) := \sup_{P \in \mathcal{P}} \bar{R}_{\text{risk};P}(Z). \quad (6)$$

where $\mathcal{P}$ represents ambiguity, and the upper expectation $\bar{R}_{\text{risk};P}$ represents risk and is therefore “centered” around P .²⁸

Under ambiguity, another way to understand the semantics of $\mathcal{P}$ is to consider an often-made distinction between *aleatoric* and *epistemic uncertainty* [[Hüllermeier and Waegeman, 2021](#)]. Intuitively, *aleatoric* uncertainty relates to the inherent randomness of the data-generating process (related to chance), whereas epistemic uncertainty refers to the lack of knowledge about the data-generating process. Viewing Equation 5 through this lens, we would say that each $P \in \mathcal{P}$ represents aleatoric uncertainty, and $\mathcal{P}$ represents our epistemic uncertainty. In this way, the aleatoric vs. epistemic distinction appears to map reasonably well onto the risk vs. ambiguity distinction; for example, [Fox et al. \[2021\]](#) conceive of risk as purely aleatoric, and ambiguity (which they call “uncertainty”, following [Knight \[1921\]](#)) as “at least partly epistemic in nature”. However, the conceptual situation is rather messy and complicated. The term “aleatoric” seems to suggest that this form of uncertainty, due to “objective chance”, is fundamentally irreducible [[Hüllermeier and Waegeman, 2021](#)]. Consider the example of car insurance in the context of *risk as estimated probabilities*, i.e. as car insurance would work in practice. Here, it seems absurd to claim that the variability in losses could not be further reduced in principle by including additional information. My position, drawing mainly on lessons from insurance, is that while the two distinctions (risk vs. ambiguity, aleatoric vs. epistemic) align quite well, they should both be understood as inherently contextual, and more as an expression of a modeling mindset, as part of a modeling toolbox, rather than as revealing something deep about the nature of the phenomena; we elaborate on the blurriness in Section 0.6.1. In particular, I do not think of aleatoric uncertainty as fundamentally irreducible.

Again — we have already touched on this — a core issue is how belief relates to relative frequencies. [Hacking \[2006\]](#), e.g. p. 12] famously highlighted the “janus-faced” character of probability. One face concerning the subjective (belief), and the other face concerning the objective (chance, relative frequencies). In insurance, both faces historically play a major role (providing valuable insights, see [[Friedman, 2020](#)]), as well as now in machine learning. In the introduction of the second edition (the original book was published in 1975), [Hacking \[2006\]](#) writes “We have not made peace with what appears to be the upshot here, namely two distinct modes of inference. I very much doubt that we ever will”. I am more optimistic: my heuristic is to begin by studying a (mathematical) *data model* (Chapter 5) on the “objective” (or “frequentist”) side, and then proceed to investigate the “reasonableness”

²⁸Formally, $\bar{R}_{\text{risk};P}$ is taken to be *law invariant* relative to P .

of subjective beliefs, or more generally procedures which output beliefs, jointly with concrete decision-making problems *under the given data model*; the “reasonableness” of beliefs referring to a kind of *accuracy-oriented* epistemology [Pettigrew, 2016, Schoenfield, 2017], but going beyond accuracy-oriented epistemology by relativizing to the specific decision problem at hand. Thus, the picture always contains both “faces”, and in particular a decision-making element. While this may sound quite abstract at this point, we will instantiate this style of approach in Chapter 6. In general, I am also motivated by the *game-theoretic account* of probability [Shafer and Vovk, 2001, 2019], which can, I believe, nicely integrate the two faces.

The move from risk to ambiguity, or rather to modeling ambiguity with imprecise probabilities, could be framed as challenging the adequacy of precise probability from a subjectivist perspective. We will now make a similar move from a frequentist perspective, again arriving at imprecise probabilities, yet from a different (although not unrelated) perspective. A usual conception of imprecise probabilities is along the following lines [Bradley, 2019]:

Imprecise probabilities is a theory born of our limitations as reasoning agents, and of limitations in our evidence base. If only we had better evidence, a single probability function would do.

Going beyond this, we now show that there is in fact more to imprecise probabilities; *what if there is no single “true” probability?*

0.4.3 ALEATORIC PLURALITY

Part II advances the theory of imprecise probabilities (primarily) under the semantics of *aleatoric plurality*, which I propose as a new umbrella term. In a nutshell, it refers to the case when the data (model)²⁹ is not adequately described by a single precise probability, but rather by a set of probabilities.

What could it mean that there is no true probability? It is instructive to go back to von Mises’ [1919] foundational account and observe when the probability fails to exist on his account, which inspires our more general theoretical developments. In the spirit of his “strict frequentism”, let us assume we have access to an infinite *data sequence* $\omega^\infty = (\omega_1, \omega_2, \dots)$, where each $\omega_i \in \Omega$; recall that we assume that elementary events $\omega \in \Omega$ capture all relevant information. von Mises [1919] calls such a sequence ω^∞ (which we now fix) a *collective* when two conditions are satisfied:³⁰

²⁹The question (of whether precise probability suffices) regarding the data is not exactly the same as the question regarding the data model, but they are related via *typicality* assumptions, see Chapter 5.

³⁰We here present a simplified version, ignoring issues of the set system; also, we here consider selection rules of type signature $S_j : \mathbb{N} \rightarrow \{0, 1\}$, while von Mises [1919] allowed the selection rule to depend on the previous outcomes in the sequence. See the discussion in Chapter 4.

vM1. Aggregate stability: the limiting relative frequency for each $A \in 2^\Omega$ exists:

$$P(A) := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \chi_A(\omega_i).$$

This limit is called “the probability of A ”.

vM2. Local irregularity: let $\mathcal{J}$ be a set which is finite or countable, and $S_j : \mathbb{N} \rightarrow \{0, 1\}$ a function called “selection rule” for each $j \in \mathcal{J}$. For each $j \in \mathcal{J}$ and $A \in 2^\Omega$, the selection rule S_j preserves limiting relative frequencies:

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \chi_A(\omega_i) \cdot S_j(i)}{\sum_{i=1}^n S_j(i)} = P(A).$$

This is equivalent to computing the limiting relative frequency of the event A along the subsequence “extracted” by S_j , and demanding that it coincides with $P(A)$.

Notably, [von Mises \[1919\]](#) takes these as *axioms*, as conditions for applying the calculus of probabilities, instead of deriving them (we will elaborate later on how these conditions can *typically* follow (or not) under a data model). Let’s consider the meaning of these conditions, and in particular what it would mean when they do not hold — breaking either vM1 or vM2 leads to aleatoric plurality.

AGGREGATE (IN)STABILITY

The condition vM1 demands, or asserts, the stability of relative frequencies in the limit (in the reference class given by ω^∞). Consider the case of flipping a fair coin; many would say the probability (interpreted as *chance*) for “heads” is 0.5; a *consequence* would be to expect a limiting relative frequency of 0.5. Turning the picture on its head, to [von Mises \[1919\]](#), probability *is* limiting relative frequency. The assumption of aggregate stability may seem intuitive in many cases, and it is for example implied by the familiar *i.i.d.* model (*almost surely*) due to the *law of large numbers*. Moving from the infinite to the finite case, [Gorban \[2011\]](#) calls the assumption that gathering more and more data will lead to aggregate stability in the long-run the *hypothesis of statistical stability*; yet [Gorban \[2011\]](#) provided intriguing case studies, demonstrating the existence of phenomena where averages substantially fluctuate even after a high number of trials. Recently, [Lovejoy \[2015\]](#) and [Franzke et al. \[2020\]](#) draw attention to such aggregate *instability* in climate science.

As a mathematical statement, condition vM1 is binary: either the limit exists or it does not. In practice, when we have to deal with the fact that we only have finite data, we move from “limit” to some understanding of “long-run”; we elaborate in Chapter 4 and Chapter 5. How to quantify aggregate (in)stability becomes less clear in the finite case, and it is then a gradual notion. In any case, the move from infinite to finite is intricate (also in

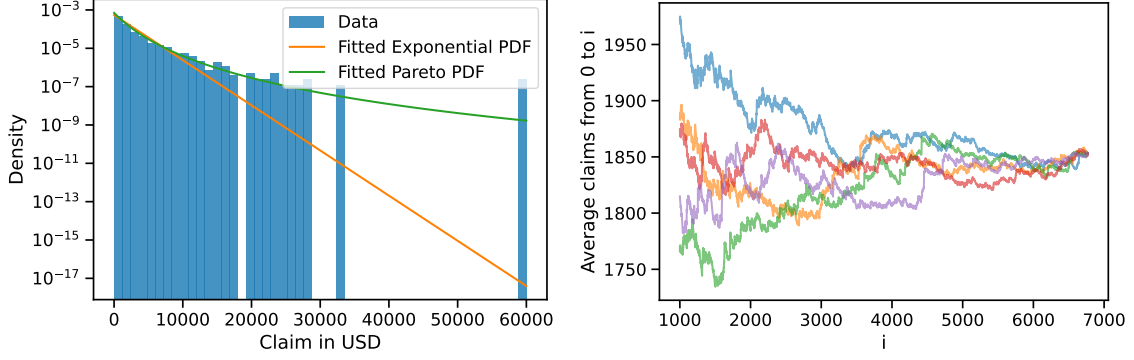


Figure 2: Left: The car insurance example (from Figure 1), now with a vertical log-scale and normalized. In addition, we show the probability densities of an exponential distribution and a Pareto distribution (with a location and scale parameter) obtained from a maximum likelihood fit to the data. We used `scipy.stats.expon.fit(data)` and `scipy.stats.pareto.fit(data)`, respectively, referring to the `scipy` library in Python. Right: we take the claims ($N = 6,773$), and shuffle them 5 times. Each curve shows the averages $\sum_{j=1}^i \ell(\omega_j)$ as i increases along the x -axis; for small i , the variability in averages is large, so we start from $i = 1000$. By construction, all curves meet at $i = 6,773$ at the overall average claim of $\sum_{j=1}^{6,773} \ell(\omega_j) \approx 1,853\text{USD}$.

the classical case of stability). Analogous to the move in the classical case from [von Mises \[1919\]](#) to [Kolmogorov \[1933\]](#), the tension can, at least to some extent, be resolved by moving from data to *data models*; whereas [von Mises \[1919\]](#) considers a fixed data sequence ω^∞ , [Kolmogorov \[1933\]](#) would consider a probability measure on the set of such sequences, i.e. a probabilistic *data model*. In other words, the move is from *strict* frequentism to *hypothetical* frequentism. Intuitively, in this thesis, one of our main contributions is the “von Mises to Kolmogorov move” but for a generalized frequentism based on imprecise probabilities. Thus, in Chapter 5, we study data models which typically yield data sequences with unstable relative frequencies, both in the long run as well as in the limit. However, we leave applied, quantitative investigations in the finite case to future research.

I believe that “breaking” condition vM1 can be a useful modeling tool in the context of *heavy tails* [[Taleb, 2007](#), [Nair et al., 2022](#)] (a similar suggestion was made by [Fierens and Fine \[2001\]](#)). Intuitively, a heavy-tailed probability distribution (a precise probability) is one where there is a relatively large probability of observing large outcomes, much larger than in the case of a Gaussian distribution, for example. Mathematically, it is customary to make a binary distinction based on a comparison to the tails of an exponential distribution (see e.g. [[Nair et al., 2022](#)]), thus separating heavy from light tails. In Chapter 3 we will conduct a general study of tail sensitivity and imprecise probabilities, where we employ more fine-grained technical tools (primarily *Orlicz spaces*) to go beyond the binary distinction.³¹

If the study of heavy-tails is situated in the context of familiar precise probability, why would we consider them under the semantics of aleatoric plurality? Let’s revisit the car

³¹There, we use unbounded probability distributions, which in turn require an infinite Ω .

insurance example — in insurance, it is common that claim distributions are *modelled as* heavy-tailed (e.g. [Ibragimov et al., 2015, Peng and Qi, 2018, Punzo et al., 2018]).³² Figure 2 (left) shows a histogram of the exemplary claims, now with a vertical log-scale and normalized so that the histogram forms a discrete probability distribution. In addition, we show the results of a maximum likelihood fit (including a fitted location and scale parameter) for a light-tailed exponential and a heavy-tailed Pareto distribution. Observe that the density of the exponential distribution is linear on a log-scale, and therefore the probability for extremely large claims decays quickly, as compared to the Pareto fit. The Pareto fit is in this case the more “cautious” one — consider for example a possible scenario of *dataset shift*, where the customer base at test time differs significantly from the customer base at training time. Concretely, the insurer might attract more customers who own very expensive cars at test time, potentially leading to an increased frequency in large claims. This observation already highlights the danger of modeling this data with a *stationary*³³ model such as an *i.i.d.* model, without including informative features; it is highly plausible that there is in fact underlying heterogeneity in the data (see the discussion of condition vM2). Without access to features, a simple approach would be to adopt a stationary, precise probabilistic model with heavy tails, such as an *i.i.d.* model based on the Pareto distribution.³⁴ Except for extremely heavy-tailed probability distributions (e.g. Pareto distributions with a shape parameter $\alpha < 1$), the law of large numbers would still apply, and we would expect convergence in the limit, *in theory*. In contrast, another approach would be to account for the observed aggregate instability directly. For illustration, in Figure 2 (right), we shuffle the claims data multiple times, and for each of the resulting finite data sequences, we plot the finite sequence of claims averages $\sum_{j=0}^i \ell(\omega_j)$, starting from $i = 1000$. Thus, at $i = 6,773$ (the size of the whole dataset), we plot the overall average of the claims (on which all shuffled versions agree). To my eye, what we observe is that there is still substantial instability in the aggregate even after many trials.

I advocate leaving behind the idea of a binary distinction between stability and instability, and instead understanding the phenomenon as *fundamentally gradual* — especially in light of the finiteness of data in practice. What I believe is the benefit of viewing the data in a way that acknowledges potential aggregate instability is that it makes visible something that was previously hidden; in other words, I see this as a modeling mindset, as an attitude towards data processing [Fine, 1970], rather than believing that there is a clear and natural distinction between stable and unstable aggregates. Yet especially in the context of

³²While some phenomena are well-modeled (in a certain non-trivial sense) as heavy-tailed, outcomes in practice will always be bounded (e.g. by the largest number that we can represent in a computer). Why do we need idealizations of infinity? In simple words (and in my subjective opinion): it makes the math nicer. It is no contradiction from a practical perspective to say that e.g. insurance claims can be both heavy-tailed *and* finite.

³³Intuitively, a stationary model has no time evolution.

³⁴Again, note that mathematically we would need infinite Ω here.

phenomena which are commonly conceived of as “heavy-tailed”, substantial aggregate instability will show up. The development of concrete data handling methods which adopt this perspective (for example specifically tailored to assumptions about the magnitude of instability) we leave to further research; however, in Part II we establish the theoretical groundwork for such advancements based on imprecise probabilities.

Our framework shows that imprecise probabilities arise rather naturally when mathematically investigating decision-making under aggregate (in)stability. In this case, the imprecise probability (i.e. the set $\mathcal{P}$ in Equation 1) will correspond to the *set of cluster points of relative frequencies*. In other words, we propose a generalized von Misesian theory (Chapter 4) which does not require that condition vM1 holds. We furthermore investigate corresponding imprecise data models in Chapter 5. One of our main results (Chapter 4, further generalized in Chapter 5) is as follows: under mild technical conditions, we show that, to any sequence ω^∞ , we can associate an imprecise probability $\mathcal{P}$ which captures its limiting aggregate instability (generalizing the convergence to “the probability”), and conversely, to any imprecise probability $\mathcal{P}$ we can construct a sequence ω^∞ so that its limiting aggregate instability is captured by $\mathcal{P}$.

LOCAL (IR)REGULARITY

A highly related further “manifestation” of imprecision (Chapter 5.2) is *local regularity*, which is obtained when condition vM2 does not hold. Consider the meaning of condition vM2. A selection rule S_j , $j \in \mathcal{J}$, is a function that extracts a subsequence from the original data sequence ω^∞ . Condition vM2 demands that the limiting relative frequencies along these subsequences coincide with the overall limiting relative frequencies. Intuitively, this is an assertion of *outcome randomness*: the sequence ω^∞ is then called random with respect to the set of selection rules (see e.g. [Derr and Williamson, 2022]). This form of randomness refers to a kind of stability under additional information: the intuition is that when condition vM2 holds, an adversarial “skeptic” (in the language of contemporary game-theoretic probability, which uses a generalization of the selection rule concept [Shafer and Vovk, 2019]), who can test us on the subsequence induced by S_j , cannot get infinitely rich in the limit by taking us up on our betting commitments (see the notion of *marginally desirable gambles* in Section 1.1). For example, consider flipping a fair coin. It seems obvious that knowing whether the coin was flipped on a Monday does not really help in predicting the outcome, and thus the process is random with respect to a corresponding selection rule. In contrast, knowing which side faced upwards before throwing indeed gives a slight advantage to a skeptic (see Footnote 12). What is intriguing about von Mises’ [1919] definition of randomness through vM2 is that it is explicitly relativized to a set of selection rules, i.e. to a set of testing strategies, without universal pretence — simi-

larly, we will argue that insurance and machine learning (in line with [Derr and Williamson \[2022\]](#)) call for such relativism.

In the language of insurance, a subsequence corresponds to selecting a subset of the insurance pool (population). In the car insurance example, we start to see heterogeneity in the claim distribution when classifying the individuals based on rating class or age, for example. In fact, insurance is fundamentally based on the process of classification, of aggregating, of putting individuals in groups, with the (elusive) goal of achieving homogeneity within groups [[Krippner and Hirschman, 2022](#)]; following the principle of *actuarial fairness* [[Arrow, 1963](#)], the price for insurance is then set in accordance with the average of the group that an individual belongs to.

The question is: what kind of information is visible to the insurer? Consider the feature map $X : \Omega \rightarrow \mathbb{R}$. The insurer can then “see” an individual ω only through $X(\omega)$. What I call *hidden heterogeneity* in Chapter 5 is a heterogeneity (a form of non-randomness; a local regularity) that is in a way invisible to the insurer (decision-maker). Consider another feature A (a piece of information about an individual), which is not included in ω , thus also not in $X(\omega)$. This feature might correlate decisively with the outcome of interest. This puts the insurer at risk of *adverse selection*: the tendency that individuals, with unfavorable values for such features (e.g. reckless driving style), tend to purchase more insurance. In the framework of [von Mises \[1919\]](#), the corresponding selection rule would yield limiting relative frequencies which differ from those of the whole aggregate; I call this “hidden heterogeneity”. In contrast, other heterogeneity might be revealed by the features $X(\omega)$ (e.g. a correlation with age). As another example, consider a “sensitive”³⁵ feature A that might be available in principle, for example the gender or migrant background status of an individual, but the insurer may choose to not feed this feature to the predictor (statistical model), perhaps because legal reasons forbid it. Or as yet another example, consider a feature which is only available at training time, but not at test time, so that the predictor cannot depend on it. Perhaps we would then like to use this information as a “meta”-feature at training time, with the goal of developing a statistical model which is robust to it in terms of performance. In summary, accounting for specific forms of (non)randomness plays a key role in developing “good” (e.g. robust, fair) statistical models.

So what happens when condition vM2 “breaks” for some selection rule(s)? In any case, for each selection rule we can attempt to compute limiting relative frequencies; when the limit does not exist, we might have to deal with a set of cluster points, as described in the discussion on aggregate instability — note that each selection rule induces an aggregate in its own right. Taking the union then yields an imprecise probability $\mathcal{P}$, which captures the non-randomness of the data, relativized to this set of selection rules. Moving from data to

³⁵In the current terminology of fair machine learning, a “sensitive” feature is a protected piece of information about an individual such as race, gender and more.

data *model*, considerations of this kind motivate the following (see Chapter 5.2).³⁶

Proposition 14. *Let $\mathcal{Z} := [0, 1)$ and $(P_1, P_2, \dots)$ be a sequence of probability measures on Ω . Then there exist $W_i : [0, 1) \rightarrow \Omega, i \in \mathbb{N}$, so that the following holds:*

$$\lambda \{z \in \mathcal{Z} : W_1(z) = \omega_1, W_2(z) = \omega_2, \dots, W_n(z) = \omega_n\} = \prod_{i=1}^n P_i(\{\omega_i\}), \quad \forall n \in \mathbb{N}.$$

We call this the *non-stationary, locally precise model*, parameterized by $(P_1, P_2, \dots)$; of course, we are not the first to consider a model which keeps the “independence”, but not the “identical” in *i.i.d.*. Such data models, depending on the sequence $(P_1, P_2, \dots)$, can *almost surely* (with respect to λ) yield data sequences ω^∞ with hidden heterogeneity. Now, it pays off that we have differentiated an abstract set of seeds $\mathcal{Z} = [0, 1]$ from our possibility set Ω in the two-stage setup; this contrasts with the typical approach of simply working with “random variables” of type signature $Z : [0, 1) \rightarrow \mathbb{R}$, where usually the so-called “probability space” is not explicitly mentioned. By introducing an intermediate layer, relevant concepts (process vs. outcome randomness) can be neatly separated in our framework, which helps in Chapter 5. For example, “probability” to us refers to a measure on Ω , not on $\mathcal{Z}$, which is instead the realm of typicality. Also, because our gambles are of type signature $Z : \Omega \rightarrow \mathbb{R}$, we obtain additional control. As a consequence, such a gamble might not be adequately described by a single probability distribution, but rather by a set of probability distributions.

One may wonder what the connection is between local (ir)regularity and aggregate (in)stability under this data model. One of our main results (Chapter 5) asserts that if $\mathcal{P} := \{P_i : i \in \mathbb{N}\}$, then the set of cluster points of relative frequencies will almost surely coincide with some *closed connected subset of the closed convex hull of $\mathcal{P}$* (and the theorem also comes with a corresponding converse direction). To summarize in simple words: the relation between breaking vM1 and vM2 is quite intricate. However, breaking vM2 is a necessary, but not sufficient condition for breaking vM1, see Chapter 5.

To wrap up the discussion of the semantical plurality of imprecise probabilities: while mathematically (syntactically) we are working with the common form of sets of probabilities, there is a plurality of relevant interpretations. What is more, and this seems highly intriguing to me, is that these are related in intricate ways. For example, how does ambiguity relate to aleatoric plurality? In Section 0.6.1 we will shed some light on this, but a conclusive answer to the conceptual interactions remains out of reach (for now).

³⁶This follows directly from a basic result in probability theory. The statement is rather intended as a definition than a proposition, but technically the existence of the W_i with the asserted property needs to be shown.

0.5 THE CONNECTION TO MACHINE LEARNING

Until now, we have said much more about insurance than about machine learning. Recall the claim in Remark 8: “Insurance is a great analogon to machine learning when it comes to matters related to uncertainty and fairness (and in particular their interplay)”. Insurance, as *the* technology for economizing uncertainty [Lehtonen and Van Hoyweghen, 2014], provides a highly instructive analogon for studying uncertainty in general, in a setting where uncertainty may feel more concrete than in general machine learning contexts. Yet uncertainty is pervasive also in machine learning, and our primary concern is to advance the use of imprecise probabilities in this context. Having discussed the plurality of semantics, it seems warranted to illustrate their relevance for machine learning.

RISK AVERSION IN ML

The development of risk-averse machine learning is particularly relevant in high-stakes, safety-critical systems. For example, Green [2022] mention driving, robotic surgery, and finance. In these domains, it is considered highly important to avoid catastrophically bad outcomes. To achieve this, decision-makers might be willing to make sacrifices in terms of average performance (expected loss) to some extent.

Specifically, upper expectations (mostly by the name of “coherent risk measures”) for expressing risk-averse attitudes have gained increasing interest in reinforcement learning (see e.g. [Morimura et al., 2010, Chow and Ghavamzadeh, 2014, Tamar et al., 2015, Lee et al., 2020, Rigter et al., 2021, Vijayan et al., 2021, Urpí et al., 2021, Greenberg et al., 2022]); Rigter et al. [2021] also mentions applications for the corresponding risk-seeking counterparts (lower expectations), for example in learning under noisy labels or curriculum learning.

To express risk aversion in supervised learning, where we have features X , labels Y , a loss function ℓ , and some hypothesis class $\mathcal{H}$, the standard problem of *expected risk minimization* can be replaced by the minimization of a coherent risk measure:

$$\min_{f \in \mathcal{H}} \bar{R}_{\mathcal{P}}(\ell(f(X), Y)). \quad (7)$$

A problem of the same form will be of interest under ambiguity and aleatoric plurality. In this way, risk-averse classification and risk-averse regression can be obtained, where the choice of $\mathcal{P}$ controls the risk aversion. A focus of Part I is to give new insights into how this can be done and what the implications of making a particular choice are. Conceptually, using an upper expectation here is not the same as simply changing the loss function since the concern is about the variability in the loss distribution (recall Section 0.4.1), although I do not claim that a similar effect could not be achieved in practice, for a fixed problem, by

making the loss function steeper (see [Hu et al., 2018]).

In an analogous way, Williamson and Menon [2019] have proposed the objective in Equation 7 to mitigate fairness concerns in machine learning, with the goal of reducing inequality in the loss distribution. This framing of fairness may be of interest when the fairness-specific goal is to achieve a similar performance across sensitive sub-populations.

AMBIGUITY AVERSION IN ML

The problem of ambiguity has received considerable attention already in machine learning research, notably under the banner of *distributionally robust optimization* (DRO) [Kuhn et al., 2019, Namkoong, 2019]. In the context of supervised learning, consider again the problem

$$\min_{f \in \mathcal{H}} \bar{R}_{\mathcal{P}}(\ell(f(X), Y)), \quad (8)$$

now with the semantics that $\mathcal{P}$ expresses the subjective ambiguity of the machine learning engineer. Here, the goal in using upper expectations is to guard against this ambiguity and thus obtain a more “robust” model, i.e. to take a cautious stance (similar to the case of risk aversion) in light of our lack of knowledge about the probability. We note that there are also proposals for distributionally robust reinforcement learning [Derman and Mannor, 2020].

One obvious source for ambiguity in machine learning is that in practice we only have access to the *empirical distribution* $\hat{P}$, i.e. the *training data*, while we aim for optimal performance on the *test data*. Granting for now that we actually face data from an underlying *i.i.d.* model, so there indeed exists a single, “true” probability, we only have partial knowledge of it due to the finiteness of our training data. In this spirit, Kuhn et al. [2019] have proposed the use of Wasserstein risk measures to tackle this form of ambiguity. Related to this, there are recent proposals that employ imprecise probabilities to capture epistemic uncertainty (in contrast to aleatoric uncertainty) [Hofman et al., 2024, Javanmardi et al., 2024]. What differentiates ambiguity from risk here is the observation that following this logic, the ambiguity should disappear as more and more data is acquired — the set $\mathcal{P}$ would be constructed in a way such that it “contracts” to a single P over time (see for example [Duchi et al., 2021]). In contrast, under risk, we consider a set $\mathcal{P}$ even if we perfectly know the “true” probability.

Going beyond the finiteness problem, ambiguity aversion in machine learning can be motivated with the aim of achieving robustness to *distributional shifts* (e.g. [Duchi and Namkoong, 2021]). The typical phrasing goes roughly as follows: while we have access to the training *distribution*, the test *distribution* might differ from it, for numerous possible reasons: an adversary (a skeptic in the language of von Mises [1919]) might be able to perturb

the distribution in an unfavorable way [Dong et al., 2020]. Or, we might face a different *domain* at test time [Koh et al., 2021], for example because our training data is from one geographical region, but at test time the system is deployed in a different region. Or, because at test time we face a *subpopulation* of the population which we trained on [Hashimoto et al., 2018, Koh et al., 2021]. In the car insurance example, we could for instance observe more customers who own luxury cars at test time, or observe more customers with a reckless driving style. Fortunately, with upper expectations, we can express an ambiguity-averse attitude in all of these settings; with potentially different ways of constructing $\mathcal{P}$, however, depending on the problem context.

The above framing suggested viewing the training data (the empirical distribution) as partial evidence for the single, “true”, but unknown test distribution. Another framing would embrace a more dynamic perspective and acknowledge that there simply need not be a single underlying “true” distribution, but instead to expect shifts which continuously unfold over time. This brings us to aleatoric plurality.

ALEATORIC PLURALITY IN ML

Some of the previously discussed cases of distributional shifts are perhaps better understood, I suggest, from the more general perspective of aleatoric plurality. Why *the* test distribution? Instead, we can embrace a more dynamic perspective, and drop the assumption of a single, “true” probability.

A number of problems in machine learning can be framed as concerning hidden heterogeneity, that is, as a violation of condition vM1. The paradigm of *group distributional robustness* [Sagawa et al., 2019], as a sub-case of distributional robustness, aims for a robust performance over a set of groups (subpopulations) of the data. Here, it is explicitly acknowledged that “the” probability can differ among the groups. For example, in a fairness context, we might have access to a sensitive feature S (e.g. gender, race), but choose to not include this in the features X . Then, the relationship between the outcome of interest Y and X might depend on S , but this is not revealed explicitly to the model (although possibly learnable to some extent through proxies). As another example, Sagawa et al. [2019] construct the groups based on “spurious” correlations in the dataset; improving the performance on rarely observed groups (e.g. images of water animals on land) can then robustify the model. To mention one more example, the paradigm of *federated learning*, where the data comes from multiple domains, fits with the setting of group distributional robustness. In any case, we again encounter the problem

$$\min_{f \in \mathcal{H}} \overline{R}_{\mathcal{P}}(\ell(f(X), Y)), \quad (9)$$

where now, the set $\mathcal{P}$ is explicitly constructed based on groups: in practice, we would separate the training data based on the groups, that is, $\mathcal{P} := \{\hat{P}_{S_j} : j \in \mathcal{J}\}$, where $\hat{P}_{S_j}$ is the empirical distribution associated with the selection rule S_j . By this, we mean taking the empirical distribution of the subset of the training data for which $S_j(\omega_i) = 1$, i.e. observing that there is a bijection between the set of groups and a set of selection rules.

The preceding discussion may seem reminiscent of how we motivated ambiguity in relation to distribution shifts. To some extent, I think it is indeed a matter of perspective. What does distinguish the perspectives, however, is whether or not we keep the assumption of a single, “true” probability. For example, Ellsberg’s urn cases clearly belong in the area of ambiguity, with a single but unknown probability, whereas under aleatoric plurality, we would continue to observe data from multiple groups (e.g. multiple urns with distinct compositions). In cases of actual machine learning practice, the distinction may not be that clear-cut, and which perspective we want to take might depend on dynamical considerations. In the context of supervised learning, we have a problem of the common form (Equation 9) anyway.

To differentiate between “standard” DRO and group DRO is, in light of these observations, not entirely simple, and they do not neatly map onto the ambiguity vs. aleatoric plurality distinction (itself a slippery one). “Standard” DRO is sometimes clearly motivated by finite data ambiguity (e.g. [Kuhn et al., 2019]); sometimes attention is drawn to distribution shifts (e.g. [Duchi and Namkoong, 2021]). In any case, this literature has centered its attention on law invariant risk measures such as f -divergence risk measures (see Section 1.1). Intuitively, this means constructing $\mathcal{P}$ by forming a “neighborhood” around some best guess P_0 , usually the empirical distribution (training data). In contrast, group DRO aligns more naturally with hidden heterogeneity (aleatoric plurality), and here the set $\mathcal{P}$ is explicitly constructed based on subpopulations. As a result, law invariance cannot be expected.³⁷

As to the aggregate instability aspect of aleatoric plurality, recall that one of our main results (Theorem 240) implies (within the class of non-stationary, locally precise models) that hidden heterogeneity is a necessary but not sufficient condition for aggregate instability. So far, it appears that aggregate instability has not received significant attention yet in machine learning. However, the prime focus of machine learning is *aggregate* (usually average) performance, and stability is simply assumed without question. If you do not look for instability, you will also not find it; nevertheless, some who *have* looked (e.g. Lovejoy [2015] and Franzke et al. [2020] in climate science), have observed it. I believe the development of methods which explicitly allow for potential aggregate instability is an interesting avenue with relevant use cases in machine learning. For example, distribution shift (also called *dataset shift*) can lead to aggregate instability. We have also already hinted at the possibil-

³⁷One could, of course, interpret the set $\mathcal{P}$ associated with a law invariant risk measure as arising from a set of groups (subpopulations).

ity of modeling heavy-tailed phenomena as unstable aggregates. [Walley and Fine \[1982\]](#) provide the following intuition (you may replace “indeterminacy” by “imprecision”):

Just as ‘randomness’ (chance) is introduced in additive probability models to account for poorly understood (‘accidental’) variation in outcomes, so ‘indeterminacy’ might be introduced in upper and lower probability models to account for poorly understood variations in chance behaviour.

I believe many phenomena which are commonly conceived of as heavy-tailed, such as when we use machine learning to predict extreme rainfall events, are well-described by an underlying hidden heterogeneity (a form of non-stationarity) which results in a “poorly understood” aggregate instability. Thus, in a rather subtle way, imprecision here is also due to a lack of knowledge about the phenomenon.³⁸ While we leave algorithmic developments to future work, I believe this thesis lays the necessary theoretical groundwork for combining the study of heavy tails with imprecise probabilities, and from there derive practical methods.

Finally, note that it is common in *online learning* as a sub-field of machine learning to dispense entirely with the assumption of a “true” probability — in line with the generalized strict frequentism we present in Chapter 4. Then, aggregate instability needs to be accounted for when the focus is on aggregate performance, and indeed I am aware of one machine learning research paper [[Zhao and Ermon, 2021](#)] which uses an evaluative criterion that is connected to imprecise probabilities and aggregate instability.

0.6 RELATING RISK, AMBIGUITY AND ALEATORIC PLURALITY

Once a conceptual edifice has been erected, it is ready to be torn down. We have already seen that risk, ambiguity and aleatoric plurality can be linked to quite similar motivations, such as the desire for deploying “safe” and “robust” machine learning systems. What is going on here? Are we not merely double-counting, as [Schoenfield \[2020\]](#) suggests, but counting thrice? A summary of my position is as follows. I side with [de Finetti \[1974/2017\]](#) in believing that probability is not something objective “out there”. The idea that data is sampled from a distribution is merely an idealization. The concepts of risk, ambiguity and aleatoric plurality are part of a useful imprecise probability modeling toolbox. Distinctions such as epistemic vs. aleatoric should be understood *locally*, that is, with respect to a given

³⁸The reader might remember that I have said aleatoric plurality is not about our lack of knowledge, but more fundamental, and hence might feel uneasy about this statement. Does the imprecision under aleatoric plurality refer to epistemic uncertainty after all? Yes and no! Ultimately I believe that every uncertainty is “epistemic” in the sense that it refers to a lack of knowledge, but I *locally relativize* the aleatoric vs. epistemic distinction. Fortunately, using the two-stage setup with the typicality set $\mathcal{Z}$ and the possibility set Ω allows to rigorously distinguish *within the mathematical model*. But the model is of course not the world.

modeling context only. In practice, a given phenomenon of interest may be more or less amenable to a description from one of these perspectives.

To make this concrete, consider again the car insurance example. Assume we are developing a machine learning model which is supposed to predict the magnitude of a claim from the given features of an individual. Our concern might be that at test time we face very high claims more often (resulting for instance from more customers with expensive cars). This could be described as a risk-averse attitude, since our worry concerns the right tail of the loss distribution. However, consider the pre-image (a subset of Ω) of this tail; we can think of this as a subpopulation. With this phrasing in terms of a distributional shift, it appears that we are ambiguity-averse in the spirit of distributional robustness. In Chapter 2, we call this situation one of “hallucinated ambiguity”, referring to the reduction of risk aversion to ambiguity aversion. In essence, the set $\mathcal{P}$ appearing in the form of an upper expectation intended to express risk aversion can as well be interpreted as an ambiguity set (we are of course not the first to observe this). Moreover, we have already noted that distribution shift could also be couched as aleatoric plurality: perhaps there simply isn’t a single stable aggregate. In this way, I believe that the aleatoric plurality perspective is the most flexible and general, as it dispenses with the idea of a single probability.

Nevertheless I believe there is some virtue to the distinction. In simple settings such as urn examples, it can be relatively clear if the situation is one of risk, ambiguity or aleatoric plurality. These simple casino-like settings can and do provide much intuition for conceptual, mathematical and technical developments. And to some extent, I think they are helpful guides, and for this reason I like to continue to refer to the plurality of semantics and use the distinction as a guidance in research. However, I also believe we must very critically examine the blurriness of the distinction in practice. For example, I think some of the most interesting phenomena like heavy tails are situated at the conceptual boundaries and pose special challenges.

0.6.1 AN INSURANCE PERSPECTIVE ON BLURRY BOUNDARIES

To understand the conceptual demarcations better and how blurry they can be, let us once again look at insurance, which “economizes” uncertainty [Lehtonen and Van Hoyweghen, 2014]. The key mechanism of insurance is aggregation, the grouping of similar cases, to obtain more or less predictable outcomes at the level of the insurance pool. It is most instructive how the prominent insurance scholar [Ewald, 1991, emphasis in original] conceived of risk:

Nothing is a risk in itself; there is no risk in reality. But on the other hand, anything *can* be a risk; it all depends on how one analyzes the danger, considers the event. [...] As a technology of risk, insurance is first and foremost a schema

of rationality, a way of breaking down, rearranging, ordering certain elements of reality. [...] It is the practice of a certain type of rationality: one formalized by the calculus of probabilities. This is why one never insures oneself except against risks [...] He [the insurer] 'produces risks' [...]

What Ewald [1991] expresses very clearly here is the *performative* character of insurance (see Chapter 7): risk is not simply 'out there'; it is rather a matter of co-construction between insurer and world. In line with von Mises' [1919] account, to Ewald [1991], risk is intrinsically collective, or social. Similarly, in machine learning, I believe we should think of risk (and hence also its counterparts ambiguity and aleatoric plurality) as resulting from a *practice* that concerns aggregates.

Yet does this mean anything can be a risk? The limits of risk, separating it from ambiguity and aleatoric plurality, are the limits of *insurability*. Interestingly, yet not coincidentally, criteria for insurability run parallel to Knight' [1921] distinction of risk vs. (unmeasurable) uncertainty (see e.g. [Feduzi et al., 2012]), and insurability is tightly linked to the two conditions of von Mises [1919]. For example, ideally, losses should occur independently (related to randomness, i.e. condition vM2), be of high frequency but low magnitude, so that the insurer can accurately estimate the outcome at the level of the pool (related to aggregate stability, i.e. condition vM1).³⁹ Intuitively, an insurable situation is a well-behaved one.

To improve insurability, insurers usually attempt to classify individuals into groups, that is, sub-aggregates. In von Misesian language, this means revealing as much heterogeneity as possible: *hidden* heterogeneity (aleatoric plurality) threatens insurability. However, heterogeneity (and consequently, instability) is a fact of life that we can ultimately not get rid of. Knight [1921] observed how insurance reduces uncertainty to risk by grouping similar cases, which he called "consolidation", but was well-aware of the fundamental problem of heterogeneity [Knight, 1921, p. 247, emphasis in original]:

Indeed, the very concept of contingency seems to preclude absolute uniqueness (as for that matter there is doubtless nothing absolutely unique in the universe). For to say that a certain event is contingent or "possible" or "may happen" appears to be equivalent to saying that "*such* things" have been known to happen before, and the "such things" manifestly constitute a class of cases formed on some ground or other. The principal subject for investigation is thus the *degree* of assimilability, or the amount of homogeneity of classes securable, or, stated inversely, the *degree of uniqueness* of various kinds of business contingencies. Insurance deals with those which are "fairly" classifiable or show a relatively low degree of uniqueness, but the different branches of insurance show a wide

³⁹Recall the discussion on heavy tails and aggregate instability; it is challenging to insure low-frequency, high-magnitude claims.

range of variation in the accuracy of measurement of probability which they secure.

In a nutshell: an “insurable” aggregate should display aggregate stability. Since heterogeneity potentially causes aggregate instability, we would like to eliminate it by forming sub-aggregates which are themselves homogeneous. But the above passage by Knight [1921] makes it clear that there are limits to this: the idea of “perfect” homogeneity is exactly as elusive as the idea of “individual risk”.

There is another important reason for classification into sub-aggregates. Recall the two conditions of von Mises’ [1919], which can be considered as two key preconditions for insurability. Condition vM2 introduces an element of *testing* into the picture. Recall that it demands stability of relative frequencies under application of a set of selection rules, where the idea is that a “skeptic” can test us through them. That is, the testing strategies which we assume possible are encoded in the set of selection rules, and thus this notion of randomness (homogeneity) is inherently relativized. In contrast, assume that some other selection rule would reveal a hidden heterogeneity by violating condition vM2. From the viewpoint of the insurer, this puts us at the risk of *adverse selection* [Akerlof, 1970], the phenomenon that individuals with high-risk characteristics are more likely to seek insurance coverage. By segmenting the portfolio into sub-aggregates, leading to premium differentiation, the insurer hopes to counter this to some extent. The effect is that insurability is improved because the portfolio is more robust to distribution shifts, in particular to subpopulation shifts — without sufficient differentiation, a shifting customer base could lead to bankruptcy.

In summary, we could say that insurability demands the stability of an aggregate and a robustness to testing strategies in the form of subpopulation shifts: exactly what the conditions vM1 and vM2 assert. In particular, this observation implies a relativization of the risk vs. ambiguity or aleatoric plurality boundary *with respect to the decision situation* which we face, i.e. possible testing strategies. Employing the aleatoric plurality modeling perspective will then turn out to be useful when we face a “poorly understood” [Walley and Fine, 1982] violation of insurability, but we still have a partial handle on it (see Chapter 5). The result is an explicit robustification against the violation — recall the concept of group distributional robustness.

In machine learning, similar considerations apply and are highly relevant, even if current research does not tend to highlight these issues. While the performance of machine learning methods is usually reported in terms of aggregate statistics, we should also carefully look at the performance on subpopulations, as has indeed been acknowledged by the literature on (group) distributional robustness. I believe that paying close attention in particular to boundary phenomena can lead to the development of more robust and flexi-

ble methods that weaken the overly optimistic *i.i.d.* assumption. For example, it is worth pondering to what extent a problem situation is one under risk, or how it can be *made* a risk. Due to competition, insurers are being pushed “into the unknown” [Lehtonen and Van Hoyweghen, 2014], and thus constantly re-negotiate the boundary of risk and ambiguity. Instructive cases are for example the development of kidnap for random insurance and arctic shipping insurance [Baker and Shortland, 2023], where it was challenging to establish conditions of insurability (this serves to highlight the point that risk is constructed). Similarly, in machine learning we should carefully consider how “probabilized” an uncertain situation is, and as we have seen this may depend on the decision situation which we face.

In my view, the now-emerging field of *data-centric* machine learning [Zha et al., 2023] can fruitfully be understood as attempting to improve “insurability conditions” (transposed to the machine learning context) through putting data at the center stage, and thereby shaping which heterogeneity can be revealed. As an example, an ambiguous situation could be reduced to a risky situation by collecting more or better data. Relatedly, by understanding the distinction of aleatoric vs. epistemic uncertainty as local and contextual, we are also prompted to consider the potential behaviour-shaping dispositions that arise from it. This point has been nicely made by Walters et al. [2023], who demonstrate that whether uncertainty about stock market prices is perceived as aleatory or epistemic influences investor behavior;⁴⁰ similarly, in machine learning this may prompt different decisions by the engineer (e.g. regarding data collection).

0.7 APPENDIX

For completeness, the code for generating the plots for the car insurance toy example, intended for didactical purposes, is as follows.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
import seaborn as sns
np.random.seed(42)
colors = sns.color_palette()

# read the data
df = pd.read_csv("AutoClaims.csv")
print(df.head(5))

print(len(df))
```

⁴⁰Investors who perceive it as more aleatory would tend to respond with risk management strategies such as diversification; those who perceive it as more epistemic could respond by consulting additional experts, for example.


```

# claim distribution
fs=1.4
with sns.plotting_context('notebook',font_scale=fs):
    plt.hist(df["PAID"],bins=50);
    plt.xlabel("Claim in USD")
    plt.ylabel("Frequency of claim")
    plt.savefig("autoclaims_claims.pdf",bbox_inches="tight");
    plt.close()

# fit exponential and pareto distributions
# note that this also fits a location and scale parameter
data = df["PAID"].values
locexp, scaleexp = stats.expon.fit(data)
shapepar, locpar, scalepar = stats.pareto.fit(data)

fs=1.15
with sns.plotting_context('notebook',font_scale=fs):
    plt.figure(figsize=(6,4))
    x = np.linspace(9.5, np.max(data), 100)
    pdf = stats.expon.pdf(x, loc=locexp, scale=scaleexp)

    plt.hist(data, bins=50, density=True, alpha=0.8, label='Data')
    plt.plot(x, pdf, c=colors[1], label='Fitted Exponential PDF')

    pdfpar = stats.pareto.pdf(x, shapepar, locpar, scalepar)
    plt.plot(x, pdfpar, c=colors[2], label='Fitted Pareto PDF')

    #plt.title('Exponential Distribution Fit')
    plt.xlabel('Claim in USD')
    plt.ylabel('Density')
    plt.legend()
    plt.yscale("log")
    plt.savefig("claims_pareto_exponential_fit.pdf", bbox_inches="tight")
    plt.show()
    plt.close()

# shuffle data randomly and plot the sequence of averages
# by sequence of average, we mean the running averages which always start from the beginning
N=len(df)
expon_rvs = stats.expon(locexp, scaleexp).rvs(N)
pareto_rvs = stats.pareto(shapepar, locpar, scalepar).rvs(N)

# compute efficiently all averages of a 1d array
def computeAvg(data):
    avgs = []
    oldavg = data[0]
    for n in range(1,len(data)):
        newavg = (n-1)/n * oldavg + 1/n*data[n]
        avgs.append(newavg)
        oldavg = newavg
    return avgs

```

```

expon_avgs = computeAvgs(expon_rvs)
pareto_avgs = computeAvgs(pareto_rvs)

# start to plot only at:
start=1000

fs=1.15
with sns.plotting_context('notebook',font_scale=fs):
    plt.figure(figsize=(6,4))
    reps=5
    for i in range(reps):
        true_claims_shuffled = df["PAID"].values.copy()
        np.random.shuffle(true_claims_shuffled)
        data_avgs = computeAvgs(true_claims_shuffled)
        plt.plot(start+np.array(range(len(data_avgs[start:]))), data_avgs[start:],c=colors[i],alpha=0.6)

    plt.xlabel("i")
    plt.ylabel("Average claims from 0 to i")
    plt.savefig("claims_shuffled_avgs.pdf",bbox_inches='tight')
    plt.show()
    plt.close()

```

1

Summary

This thesis is based on the following publications and manuscripts:

- PP1.** Christian Fröhlich and Robert C. Williamson. Risk measures and upper probabilities: Coherence and stratification. *Journal of Machine Learning Research*, 25, 2024.
- PP2.** Christian Fröhlich and Robert C. Williamson. Tailoring to the tails: Risk measures for fine-grained tail sensitivity. *Transactions on Machine Learning Research*, 2023. URL <https://openreview.net/forum?id=UntUoeLwwu>.
- PP3.** Christian Fröhlich, Rabanus Derr, and Robert C. Williamson. Towards a strictly frequentist theory of imprecise probability. In *International Symposium on Imprecise Probability: Theories and Applications*, pages 230–240, 2023.
- PP4.** Christian Fröhlich, Rabanus Derr, and Robert C. Williamson. Strictly frequentist imprecise probability. *International Journal of Approximate Reasoning*, 168, 2024.
- PP5.** Christian Fröhlich and Robert C. Williamson. Data models with two manifestations of imprecision. *arXiv preprint arXiv:2404.09741 (v2)*, 2025.
- PP6.** Christian Fröhlich and Robert C. Williamson. Scoring rules and calibration for imprecise probabilities. *arXiv preprint arXiv:2410.23001*, 2024.

PP7. Christian Fröhlich and Robert C. Williamson. Insights from insurance for fair machine learning. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 407–421, 2024.

The above order relates to the order of appearance as chapters in this thesis; it almost coincides with the chronological order of writing the papers, except that I decided to move the “insights from insurance” paper to the last position. The respective author contributions are stated at the beginning of each chapter. The conference paper “towards a strictly frequentist theory of imprecise probability.” [Fröhlich et al., 2023] is essentially (approximately) a short version of the journal paper “strictly frequentist imprecise probability” [Fröhlich et al., 2024], and is not included on its own in this thesis.

The papers listed above, with the exception of [Fröhlich et al., 2023], appear in the thesis as a chapter each. Concretely, I have included the LaTeX source of the following variants: For Paper PP1, I used the source of the JMLR publication; for paper PP2, I used the source of the most recent arxiv version [Fröhlich and Williamson, 2022]; for paper PP4, I used the source of the IJAR publication; for papers PP5, PP6 and PP7, I used the most recent arxiv versions [Fröhlich and Williamson, 2025, 2024e,d].

After including the LaTeX source, I have made some very minor modifications. For example, I introduced some line breaks when equations were too long for the layout of the thesis due to the relatively wide margins; in some cases, I have accepted overflows into the margin. Further, I have made minor changes to section levels (e.g. in appendices). However, I have *not* unified notation across papers. I attempted to hide the numbers for equations which are never referred to, using `\mathtoolsset{showonlyrefs}`, but this caused unexpected LaTeX trouble.

I preferred to have a single bibliography, which meant having to merge different .bib files, a somewhat intricate process given the number of references. In few cases, I merged bibliography entries, i.e. when I cited an older version of a paper as well as a newer version. In other cases, e.g. when the citations referred to concrete pages, I did not merge.

We now briefly (indeed) summarize and contextualize each paper. Obviously, the summary contains some elements which are directly from the the papers (such as definitions). Similarly to the introduction, I try to avoid distraction by technical headaches (such as finite vs. countable additivity) here; for a rigorous presentation, the respective paper chapters should be consulted. Notation-wise, I opted for a compromise between a unified notation throughout the summary and matching the notation of the paper. In general, when in doubt because of a tension between principles, I prioritized improving the accessibility of the summary section.

1.1 SUMMARY OF CHAPTER 2

Title: “Risk Measures and Upper Probabilities: Coherence and Stratification”

Abstract: Machine learning typically presupposes classical probability theory which implies that aggregation is built upon expectation. There are now multiple reasons to motivate looking at richer alternatives to classical probability theory as a mathematical foundation for machine learning. We systematically examine a powerful and rich class of alternative aggregation functionals, known variously as spectral risk measures, Choquet integrals or Lorentz norms. We present a range of characterization results, and demonstrate what makes this spectral family so special. In doing so we arrive at a natural stratification of all coherent risk measures in terms of the upper probabilities that they induce by exploiting results from the theory of rearrangement invariant Banach spaces. We empirically demonstrate how this new approach to uncertainty helps tackling practical machine learning problems.

When I started the PhD, my supervisor Bob suggested that I investigate the connection between coherent risk measures [Artzner et al., 1999] and rearrangement invariant Banach function norms [Bennett and Sharpley, 1988].¹ Simultaneously, he recommended that I look at the coherent lower and upper previsions of Walley [1991], which appeared to be similar, so a natural first question was whether these mathematical objects are actually equivalent, and more generally how these approaches relate. The resulting paper therefore studies in depth the conceptual and mathematical relations between coherent risk measures, rearrangement invariant Banach norms and coherent lower and upper previsions (a term used in the community of “imprecise probabilities”). Moreover, we present in an accessible way generalizations of expected utility theory and show how these relate to the previous concepts. Among the coherent risk measures, we focus on the class of *spectral risk measures*, which turn out to be rather special objects when illuminated from multiple different angles. A key objective of this paper is thus to gather diverse characterizations of the distinctiveness of spectral risk measures. To this end, we focus on the *fundamental function* [Bennett and Sharpley, 1988], which is essentially an upper probability in the lingo of imprecise probabilities.

What motivates us here (as already discussed in Chapter 0) is the idea of replacing the expectation in *expected* risk minimization by a functional that can express a risk-averse or ambiguity-averse attitude. Concretely:

$$\operatorname{argmin}_{f \in \mathcal{H}} \mathbb{E}_P[\ell(f(X), Y)] \longrightarrow \operatorname{argmin}_{f \in \mathcal{F}} \bar{R}_{\mathcal{P}}(\ell(f(X), Y)). \quad (1.1)$$

¹For brevity, we will also equivalently write *rearrangement invariant (Banach) norm* or *ri norm*.

But which $\bar{R}_{\mathcal{P}}$ to choose? In this paper (as well as the follow-up project in Chapter 3) we attempt to give some insight in how to make this choice and to demonstrate what the consequences of the choice are. To this end, we propose a *stratification* of the class of (law invariant) coherent risk measures.

Walley's [1991] seminal treatment of imprecise probabilities is *axiomatic*. As in the introduction (Chapter 0), consider a bounded gamble $X : \Omega \rightarrow \mathbb{R}$ on a possibility set Ω . Walley [1991] asks when such a gamble is *desirable* to a decision maker, and comes up with the following list of axioms as minimal rationality demands. If a gamble is desirable, we write $X \in \mathcal{D}$. Then we demand:²

- D1.** $\sup(X) < 0 \Rightarrow X \in \mathcal{D}$
- D2.** $\inf(X) > 0 \Rightarrow X \notin \mathcal{D}$
- D3.** $X \in \mathcal{D}, \lambda \in \mathbb{R}^+ \Rightarrow \lambda X \in \mathcal{D}$
- D4.** $X \in \mathcal{D}, Y \in \mathcal{D} \Rightarrow X + Y \in \mathcal{D}$

For example, if a gamble only ever yields gain ($\sup(X) < 0$), then certainly it should be desirable. A set $\mathcal{D}$ of gambles which satisfies these axioms is called a *coherent set of desirable gambles*. Having access to $\mathcal{D}$, we can use it to specify buying and selling prices for arbitrary gambles. Walley [1991] calls the following an *upper prevision*:³

$$\bar{P}(X) := \inf\{\alpha \in \mathbb{R} : X - \alpha \in \mathcal{D}\}. \quad (1.2)$$

Intuitively, $\bar{P}(X)$ is the smallest amount of certain loss α that, when subtracted from the uncertain loss X , makes $X - \alpha$ desirable. Hence $X - \bar{P}(X)$ is desirable, and a gamble of this form is known as a *marginally desirable gamble*, since it is just barely desirable. This may already remind the reader of insurance (consider $\bar{P}(X)$ as the premium, and X as the uncertain loss), and indeed such marginally desirable gambles will play a key role later on in the thesis (Chapter 6), when we develop a notion of calibration for imprecise probabilities from an insurantal perspective.

When $\bar{P}$ is defined from a coherent set of desirable gambles $\mathcal{D}$, and $\bar{P}$ is defined on a vector space of gambles, it satisfies the following properties:

- P1.** $\bar{P}(X) \leq \sup(X)$ (bounds)
- P2.** $\bar{P}(\lambda X) = \lambda \bar{P}(X) \quad \forall \lambda \in \mathbb{R}^+$ (positive homogeneity)

²Here we assume the obvious vector space structure on the set of gambles.

³We translate results of Walley [1991] by taking the loss/utility sign flip into account. By conjugacy, the lower prevision is: $\underline{P}(X) := -\bar{P}(-X) = \sup\{\alpha \in \mathbb{R} : \alpha - X \in \mathcal{D}\}$.

$$\text{P3. } \bar{P}(X + Y) \leq \bar{P}(X) + \bar{P}(Y) \quad (\text{subadditivity})$$

Conversely, one could simply take these properties as *definitory* for the concept of a *coherent upper prevision*. In this way, coherency of the set of desirable gambles is equivalent to the coherency of the upper prevision.

Remark 15. To match notation of the papers and much of the literature, I choose to use the symbols X and Y for arbitrary loss-valued gambles. This does not fit the role of X and Y in the supervised machine learning setup, however, but which semantics are implied should be clear from the context. In essence, when I use X or Y as an arbitrary gamble, in supervised machine learning it rather corresponds to “ $\ell(f(X), Y)$ ”.

In finance, Artzner et al. [1999], seemingly unaware of Walley’s [1991] work, proposed a highly similar development. They axiomatized “acceptance sets” (analogous to $\mathcal{D}$), and axiomatized “coherent risk measures” (analogous to $\bar{P}$), the main difference being a sign flip. For completeness, the axioms of Artzner et al. [1999] for coherent risk measures are as follows (when translating in accordance with the sign flip):

$$\text{C1. } \bar{R}(\lambda X) = \lambda \bar{R}(X) \quad \forall \lambda \in \mathbb{R}^+ \quad (\text{positive homogeneity})$$

$$\text{C2. } \bar{R}(X + Y) \leq \bar{R}(X) + \bar{R}(Y) \quad (\text{subadditivity})$$

$$\text{C3. } \bar{R}(X + c) = \bar{R}(X) + c \quad \forall c \in \mathbb{R} \quad (\text{translation equivariance})$$

$$\text{C4. } X(\omega) \leq Y(\omega) \quad \forall \omega \Rightarrow \bar{R}(X) \leq \bar{R}(Y) \quad (\text{monotonicity})$$

Indeed, that the concepts of coherent risk measures and coherent lower and upper previsions coincide⁴ was already observed by Vicig [2008]: with a vector space of bounded gambles, P1–P3 hold if and only if C1–C4 hold. Still, the literature in financial mathematics, following Artzner et al. [1999], and the literature of imprecise probabilities following Walley [1991], remain largely separate. In this paper, we strengthen the connection by adding multiple insights. In any case, in both approaches it is well-known that a coherent risk measure (upper prevision) allows a representation of the form:

$$\bar{R}(X) = \sup_{P \in \mathcal{P}} \mathbb{E}_P[X], \quad \forall X \in L^\infty, \quad (1.3)$$

called “envelope representation”. Conversely, the possibility of writing a functional in such a representation is again equivalent to coherency.⁵ From now on we again write $\bar{R}$ (for brevity called upper expectation) to subsume coherent risk measures and coherent

⁴Assuming bounded gambles.

⁵Again, putting technical subtleties aside for now.

upper previsions.⁶ Note that I avoid here a discussion of the issue with finite vs. countable additivity, see instead Chapter 2 and Chapter 4. For now, we stick to countable additivity.

Let us finally introduce an upper expectation of practical interest, the *conditional value at risk* (CVar_α), also known as *expected shortfall*, *tail conditional expectation* or *superquantile* [Laguel et al., 2021]. Fix a base probability measure P_0 . A gamble $X : \Omega \rightarrow \mathbb{R}$ has a right-continuous distribution function F_X (relative to P_0) given by $F_X(z) := P_0(X \leq z)$, $z \in \mathbb{R}$; this has a generalized inverse $F_X^{-1}(q) = \sup\{\lambda \geq 0 : F_X(\lambda) < q\}$ called the *quantile function*. Then the conditional value at risk CVar_α can be defined via an integral over certain quantiles:

$$\text{CVar}_\alpha(X) := \frac{1}{1-\alpha} \int_\alpha^1 F_X^{-1}(q) \, dq. \quad (1.4)$$

If F_X is continuous, this can be further written as

$$\text{CVar}_\alpha(X) = \mathbb{E} \left[X | X \geq F_X^{-1}(\alpha) \right], \quad (1.5)$$

The intuition is that CVar_α generalizes the notion of an expectation, but only integrates over the part where the gamble exceeds the threshold of its α -th quantile, and all values below that are disregarded entirely. Hence, $\text{CVar}_{\alpha=0} = \mathbb{E}$ and “ $\text{CVar}_{\alpha \uparrow 1} = \sup(X)$ ”.⁷ In this way, with CVar_α we can interpolate between risk neutrality and maximal risk aversion. CVar_α is perhaps the most well-known coherent risk measure, and it has been “reinvented” multiple times; recently, in machine learning, it has been proposed as learning with “top-k loss” [Fan et al., 2017], and it has also been used in reinforcement learning (e.g. [Urpí et al., 2021]).

The definition of CVar_α can be generalized and we then arrive at the class of *spectral risk measures* [Acerbi, 2002], or equivalently, *distortion risk measures* [Wang, 1995]. Fix some function $w : [0, 1] \rightarrow \mathbb{R}^+$, called *weighting function*, which is non-negative, non-decreasing (monotonically increasing) and satisfies $\int_0^1 w(q) \, dq = 1$. We then define the spectral risk measure with weighting function w as:

$$R_{(w)}(X) := \int_0^1 F_X^{-1}(q) w(q) \, dq, \quad (1.6)$$

⁶In this thesis, we will almost always stick to *coherent* risk measures instead of investigating more general *convex* risk measures [Föllmer and Weber, 2015]. Thus, when we write “upper expectation” or “risk measure” we typically mean *coherent*. Chapter 3 also provides some insight into a specific class of convex risk measures, the *utility-based shortfall risk*.

⁷Stated informally here due to technical subtleties as α tends to 1. This is again related to the issue of finite vs. countable additivity.

In the special case of CVar_α , $\alpha \in [0, 1)$, we have

$$w_\alpha(q) = \begin{cases} 0 & 0 < q < \alpha, \\ 1/(1 - \alpha) & q \geq \alpha. \end{cases} \quad (1.7)$$

The case of $w_{\alpha=0}(q) = 1 \ \forall q \in [0, 1]$ corresponds to $\mathbb{E}$. Thus, while in the case of CVar_α all values above the α -th quantile receive a constant weight and all values below it are ignored entirely, spectral risk measures can allow for more flexible weightings of the quantiles. The key is, however, that w is increasing, which amounts to obtaining a risk measure that is more pessimistic than the expectation. Indeed, it can be checked that spectral risk measures, and in particular CVar_α , are indeed coherent.

Furthermore, spectral risk measures have an interesting and powerful property called *law invariance*. Fix a base probability measure P_0 . Then a risk measure is law invariant if $\forall X, Y \in L^\infty$:

$$F_X = F_Y \implies \bar{R}(X) = \bar{R}(Y).$$

Intuitively, a law invariant risk measure judges the risk inherent in a gamble relative to a base probability measure P_0 , and treats same loss distributions in exactly the same way. Law invariance is intriguing because it adds a lot of useful mathematical structure to the risk measure, and thus also heavily constrains it. Application-wise, law invariant risk measures are relevant to express risk aversion, or under ambiguity when we have a “best guess” of the probability available. Thus, Part I of the thesis will be mostly concerned with law invariance, while Part II will consider more general risk measures.

Law invariance brings us to the connection with *rearrangement invariant Banach function spaces* [Bennett and Sharpley, 1988], or *symmetric spaces* [Rubshtein et al., 2016]. Under appropriate technical conditions (a form of continuity is needed), coherent risk measures induce Banach spaces (i.e. complete normed vector spaces) by defining a norm as

$$\|X\|_{\mathcal{R}} := \bar{R}(|X|), \quad (1.8)$$

where X is any measurable function of signature $X : \Omega \rightarrow \mathbb{R}$, what is commonly called a *random variable*. Concretely, we assume $\Omega = [0, 1]$. We write $\mathcal{M}$ for the set of such functions. Consequently, the Banach space is $\mathcal{R} := \{X \in \mathcal{M} : \bar{R}(|X|) < \infty\}$ (if it is indeed a Banach space). To make this interesting, we need to allow unbounded functions now, otherwise we trivially end up with the space L^∞ . Further, when the coherent risk measure is law invariant, then the Banach space is *rearrangement invariant*, meaning that if $X \in \mathcal{R}$, then $F_X = F_Y \implies \|X\|_{\mathcal{R}} = \|Y\|_{\mathcal{R}}$.⁸ To the best of our knowledge, we are the first to exploit in depth the connection between coherent risk measures and rearrangement

⁸Note that the monotonicity of $\bar{R}$ is also an ingredient in making the Banach space rearrangement invariant.

invariant (ri) Banach spaces. Indeed, many older results from this literature (e.g. [Kreĭn et al., 1982a, Bennett and Sharpley, 1988]) can be fruitfully transferred, and from there we go on to contribute new insights.

Remark 16. Terminologically, we move freely between statements about law invariant coherent risk measures (our objects of interest) and their respective induced rearrangement invariant norms (with derived interest to us). Observe, however, that the implicit assumption is that the risk measure in fact induces such a legitimate norm, for which a continuity assumption is needed (see Section 2.4).

A key object of interest is the so-called *fundamental function* [Bennett and Sharpley, 1988]. Fix a law invariant coherent risk measure $\bar{R}$. For the fundamental function $\phi_{\bar{R}} : [0, 1] \rightarrow \mathbb{R}^+$, take any measurable subset $E \subseteq \Omega$ of measure $P_0(E) = t$, and set

$$\phi_{\bar{R}}(t) := \|\chi_E\|_{\mathcal{R}} = \bar{R}(\chi_E). \quad (1.9)$$

Due to the non-negativity of the indicator function χ_E , we know that $\|\chi_E\|_{\mathcal{R}} = \bar{R}(\chi_E)$, and due to law invariance, we know that the choice of E does not matter, as long as $P_0(E) = t$. Recall that by applying an upper expectation (coherent risk measure) to indicator functions, we obtain an upper probability, in the lingo of imprecise probabilities. Thanks to the rearrangement invariance, we can summarize the upper probability by the single one-dimensional function $\phi_{\mathcal{R}}$ relative to P_0 . In general, the fundamental function is *quasi-concave*, which means that it satisfies $\phi(0) = 0$, and that $t \mapsto \phi(t)/t$ is decreasing (non-increasing), and it is continuous except perhaps at $t = 0$.

For example, the familiar expectation $\mathbb{E}_{P_0}[\cdot]$ has fundamental function $\phi(t) = t$, which corresponds to a non-distortion of P_0 (risk neutrality). For the conditional value at risk, it is easy to compute that

$$\phi_{\text{CVar}_\alpha}(t) = \min(t/(1 - \alpha), 1).$$

Looking at the weighting function for CVar_α , we observe that $\phi'_{\text{CVar}_\alpha}(t) = w_\alpha(1 - t)$,⁹ and indeed the analogous relation holds for any spectral risk measure. Thus, for a spectral risk measure, the fundamental function is in a one-to-one relationship to the weighting function. Recall that in general, however, the concept of an upper probability is weaker than that of an upper expectation; in principle, many distinct upper expectations can share the same upper probability. Yet, to any legitimate fundamental function (of some ri space), there exists a unique spectral risk measure,¹⁰ and it has the intriguing property of being

⁹The dash denotes the familiar derivative.

¹⁰The statement needs to be qualified slightly. For a spectral risk measure, we require that ϕ be concave, not merely quasi-concave. However, the distinction is not very substantial, since a space with a quasi-concave function can be equivalently (up to norm equivalence) re-normed such that it has concave fundamental function, see the discussion in the paper.

the most pessimistic extension of that fundamental function (Theorem 37). By “extension”, we mean constructing a risk measure whose fundamental function is the specified one. Conversely, the most optimistic extension of a given fundamental function corresponds to a norm which is in a dual relationship to the spectral risk measure. This “dual” norm does not satisfy the axioms of a coherent risk measure,¹¹ as it lacks translation equivariance, but in the paper we show how to construct an equivalent¹² most optimistic coherent risk measure.

So what can we say about the class of coherent risk measures in terms of their fundamental functions? It turns out, a lot. For example, we prove in Theorem 52 that it is exactly (“if and only if”) for the CVar_α -style fundamental functions of form $\phi(t) = \min(t/(1-\alpha), 1)$, where all possible rearrangement invariant norms collapse to a single point, namely CVar_α ; observe also that the space corresponding to CVar_α is $\mathcal{L}^1 := \{X \in \mathcal{M} : \mathbb{E}_{P_0}[|X|] < \infty\}$ for $\alpha \in [0, 1)$, and L^∞ in the limiting (and very special) case of $\alpha \uparrow 1$. Note that two norms need not be equal¹³ but can still induce the same Banach space, i.e. when they are finite on the same set of functions.

Focusing on the concept of the fundamental function turns out to be highly fruitful, and we derive a range of results from it. In particular, in this paper we take the first steps towards a study of risk measures based on their *tail sensitivity*, in turn based on the fundamental function, an idea on which we expand in the follow-up paper PP2 (Chapter 3). Here, we show how the “space” of coherent risk measures can be stratified based on the fundamental function, for example by demonstrating that comparisons of fundamental functions can allow us to judge whether two norms are equivalent or not; two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ are called “equivalent” if $\exists c, c' > 0$ such that $c\|X\|_1 \leq \|X\|_2 \leq c'\|X\|_1 \forall X \in \mathcal{M}$.

Beyond these observations, Section 2.5 (contributed entirely by Bob) demonstrates how to combine two risk measures to obtain a sort of “compromise” risk measure; again, the key concept is that of the fundamental function.

Finally, the paper also contains some toy-like experiments, which illustrate the theory. We make a new suggestion for the evaluation of loss distributions here; instead of merely focusing on average loss, we suggest to plot “ CVar_α curves” of the loss distribution (Section 2.6.1), where $\alpha \in [0, 1]$ varies along the x -axis, and the y -axis shows $\text{CVar}_\alpha(\ell(f(X), Y))$. This allows us to visualize the distributional inequality, and in contrast to Lorenz curves, it also allows us to understand absolute magnitudes.

Remark 17. My aim with the paper summaries is more of a contextualization and giv-

¹¹What does it even mean to say this? Note that for this statement to make sense we need to restrict ourselves to the cone of non-negative measurable functions. This gives some additional headaches, with which we deal in the paper, see Section 2.4.6.

¹²See the definition later in the text.

¹³By equality of norms or coherent risk measures we understand that they yield exactly the same numbers, i.e. $\bar{R}(X) = \bar{R}'(X) \forall X \in \mathcal{M}$. In contrast, *equivalence* of norms is a weaker notion.

ing a brief overview; they are necessarily (highly) non-comprehensive. For the respective key contributions of the papers, still on a rather abstract level, see Section 1.7 as well as contribution statements in the respective papers themselves.

1.2 SUMMARY OF CHAPTER 3

Title: “Tailoring to the Tails: Risk Measures for Fine-Grained Tail Sensitivity”

Abstract: Expected risk minimization (ERM) is at the core of many machine learning systems. This means that the risk inherent in a loss distribution is summarized using a single number – its average. In this paper, we propose a general approach to construct *risk measures* which exhibit a desired tail sensitivity and may replace the expectation operator in ERM. Our method relies on the specification of a reference distribution with a desired tail behaviour, which is in a one-to-one correspondence to a *coherent upper probability*. Any risk measure, which is compatible with this upper probability, displays a tail sensitivity which is finely tuned to the reference distribution. As a concrete example, we focus on *divergence risk measures* based on f -divergence ambiguity sets, which are a widespread tool used to foster distributional robustness of machine learning systems. For instance, we show how ambiguity sets based on the Kullback-Leibler divergence are intimately tied to the class of subexponential random variables. We elaborate the connection between divergence risk measures and *rearrangement invariant Banach norms*.

In the previous paper we have already made some first observations about the tail sensitivity of coherent risk measures, viewed through the lens of the fundamental function, and in this paper we advance in this direction. By tail sensitivity, we mean how strongly a risk measure “reacts” (penalizes) to large loss values; that is, we consider the right tail only. For some intuition, consider $\text{CVar}_\alpha(X)$, which interpolates between $\mathbb{E}[X]$ and $\sup(X)$ as α increases, and is thus increasingly tail-sensitive.

In this paper, spectral risk measures also appear, but the focus lies more on a different class, that of *f-divergence risk measures*, which are also law invariant. This class of risk measures has been relatively popular in machine learning, and they are of the following form:

$$\bar{R}_{f,\varepsilon}(X) := \sup\{\mathbb{E}_Q[X] : d_f(Q, P_0) \leq \varepsilon\}. \quad (1.10)$$

Intuitively, the f -divergence d_f measures how “close” Q is to P through a convex function f that penalizes deviations. In practice, we would typically choose P_0 to be the empirical distribution as our “best guess”, and in this way construct an f -divergence ball of radius

ε around it — motivated by risk aversion or ambiguity aversion. The class of functions f that we allow is specified in the following.¹⁴

Definition 18. A divergence function f is a proper, lower semi-continuous, convex function $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+ \cup \{\infty\}$, which satisfies $f(1) = 0$, $f(0) < \infty$ and the supercoercivity condition $\lim_{x \rightarrow \infty} \frac{f(x)}{x} = \infty$.

We adopted the conditions that $f(0) < \infty$ and $\lim_{x \rightarrow \infty} \frac{f(x)}{x} = \infty$ from [Dommel and Pichler \[2020\]](#); they are non-standard but play a subtle role. Given such an f , we can define the f -divergence.

Definition 19. Given a divergence function f , the f -divergence between probability measures Q and P is defined as:

$$d_f(Q, P) := \int_{\Omega} f\left(\frac{dQ}{dP}\right) dP = \mathbb{E}_P f\left(\frac{dQ}{dP}\right), \quad (1.11)$$

if Q is absolutely continuous with respect to P , and $+\infty$ otherwise.

As a prominent example, the Kullback-Leibler (KL) divergence is obtained from $f_{\text{KL}}(x) := x \log(x) - (x - 1)$. Note that an f -divergence is not in general symmetric. An obvious question is: how to choose an f , and what are the consequences of this choice? In this paper, we demonstrate in detail how the choice of f is directly relatable to a choice of tail sensitivity. For instance, it is highly common to choose the KL divergence without questioning why (e.g. because of mathematical convenience). In this example, our findings show that this entails the assumption that tails are at most as heavy as those of an exponential distribution.

A first step is to connect f -divergence risk measures with certain rearrangement invariant norms known as *Orlicz norms*. Here, we rely on the previous work by [Dommel and Pichler \[2020\]](#). Recall that by taking absolute values, a law invariant coherent risk measure can induce (under appropriate technical conditions) a rearrangement invariant norm. [Dommel and Pichler \[2020\]](#) have demonstrated a tight relationship between f -divergence risk measures and Orlicz norms, and their corresponding *Orlicz spaces* (the corresponding Banach function spaces). We make this connection clearer by introducing an intermediate object, which we call *Orlicz regret measure* that behaves like an “asymmetric Orlicz norm” (asymmetric in the sense that positive and negative values are treated differently). This object fits in the “risk quadrangle” framework proposed by [Rockafellar and Uryasev \[2013\]](#), and the corresponding risk measure is indeed the f -divergence risk measure. Essentially, translating between an f -divergence risk measure and its corresponding Orlicz norm is merely a change of parameterization, where convex conjugates are the key mathematical tool.

¹⁴For the respective definitions, consult a textbook on convex analysis or a search engine.

As a useful example of Orlicz spaces, recall the definition of the Lebesgue spaces, denoted by $\mathcal{L}^p$, $p \in [1, \infty]$, and defined as:

$$\|X\|_p := \begin{cases} \left(\int_0^1 |X|^p dP \right)^{\frac{1}{p}} & 1 \leq p < \infty \\ \text{ess sup}(X) & p = \infty, \end{cases} \quad \mathcal{L}^p := \{X \in \mathcal{M} : \|X\|_p < \infty\}, \quad (1.12)$$

with $\text{ess sup}(X) := \inf \{ \lambda \geq 0 : P_0(\{\omega \in \Omega : |X(\omega)| > \lambda\}) = 0 \}$. For example, CVar_α is finite exactly on $\mathcal{L}^1$. Indeed, $\mathcal{L}^1$ is the largest of all rearrangement invariant spaces,¹⁵ and $\mathcal{L}^\infty$ is the smallest (see Theorem 125).

Against this background, we go on to characterize the tail sensitivity of law invariant coherent risk measures, and in particular f -divergence risk measures. To understand this, consider the following known representation result (see Section 3.4.1).

Proposition 20. *A rearrangement invariant norm allows a representation of the form:*

$$\|X\|_{\mathcal{R}} = \sup \left\{ \int_0^1 F_X^{-1}(q) F_Z^{-1}(q) dq : Z \in \mathcal{M}, Z \geq 0, \|Z\|_{\mathcal{R}'} \leq 1 \right\}, \quad (1.13)$$

where $\|\cdot\|_{\mathcal{R}'}$ is the associate norm to $\|\cdot\|_{\mathcal{R}}$.

Informally, the *associate norm* $\|\cdot\|_{\mathcal{R}'}$ is like the norm which is “dual” to $\|\cdot\|_{\mathcal{R}}$, although technically there is a subtle difference between these concepts (see Section 2.4.1). For example, the well-known $\mathcal{L}^p$ - $\mathcal{L}^q$ duality (with $1/p + 1/q = 1$) is a special case of such a relation; the pathological case of the $\mathcal{L}^1$ - $\mathcal{L}^\infty$ relationship serves as an example where the concepts of associate norm and duality diverge.

What the statement tells us, intuitively, is that a law invariant coherent risk measure (which under suitable conditions induces a rearrangement invariant norm) can always be represented as a supremum over a set of possible quantile re-weightings. Furthermore, if the norm is induced by a coherent risk measure which is in particular translation equivariant, then we know in addition that we can restrict ourselves to Z for which $\mathbb{E}_{P_0}[Z] = 1$, implying that we can view Z as a probability density. If Z is a probability density, then F_Z^{-1} in fact fulfills the demands on a spectral weighting function. In other words, Equation 1.13 could be rewritten as a supremum over a set of spectral risk measures with weighting functions F_Z^{-1} , where the Z are from the unit ball of the associate norm. This again shows how “special” spectral risk measures are: they are exactly those for which the representation of Equation 1.13 reduces to a singleton (a single F_Z^{-1} suffices). In the community of coherent risk measures, a highly similar result is known as the *Kusuoka representation* [Kusuoka, 2001], although we have shown in Chapter 2, that it is pre-dated by the result in the ri

¹⁵A rearrangement invariant (Banach function) space is one which is induced by a rearrangement invariant norm.

Banach function spaces literature.

Now, understanding the tail sensitivity amounts to understanding the relationship between the possible tail behaviours of F_X^{-1} and F_Z^{-1} , i.e. the “primal” and “dual” variables. For instance, we might ask, for a given ri norm $\|\cdot\|_{\mathcal{R}}$, how extreme can the tails of the F_Z^{-1} be? Or conversely, how heavy can the tails of F_X^{-1} be such that $\|X\|_{\mathcal{R}}$ is still finite?

Remark 21. We can freely switch between focusing on X and F_X^{-1} in a context of rearrangement invariance, since these measurable functions are *equimeasurable* in the sense that $F_{F_X^{-1}} = F_X$. It might be somewhat unfamiliar to consider the quantile function as a random variable (gamble) itself, but this is indeed a fruitful perspective. Implicitly, we work on the possibility set $\Omega = [0, 1]$. I emphasize once again that for the technically rigorous presentation, the reader should directly consult the respective paper.

Let us define a function which “bounds” the possible tail behaviour of the dual variables. Assume we have an ri space $\mathcal{R}$ (with norm $\|\cdot\|_{\mathcal{R}}$) with associate ri space $\mathcal{R}'$ (with norm $\|\cdot\|_{\mathcal{R}'}$). Define the *envelope function* of the space $\mathcal{R}'$ as $E_{\mathcal{R}'} : (0, 1] \rightarrow \mathbb{R}^+ \cup \{\infty\}$ by [Haroske, 2006]:

$$E_{\mathcal{R}'}(t) := \sup\{F_Z^{-1}(1-t) : Z \in \mathcal{M}, Z \geq 0, \|Z\|_{\mathcal{R}'} \leq 1\}. \quad (1.14)$$

Due to the change of direction here, the tail-relevant behaviour is $\lim_{t \downarrow 0} E_{\mathcal{R}'}(t)$. Intuitively, if the norm $\|\cdot\|_{\mathcal{R}}$ is highly tail-sensitive, then $E_{\mathcal{R}'}(t)$ will be relatively large for small t ($t \downarrow 0$), meaning that a wide class of quantile reweightings is allowed. An analogous definition works for the *primal* space (induced by $\|\cdot\|_{\mathcal{R}}$). The envelope function turns out to be highly interesting and insightful; studying the tail behaviour of the envelope function allows us to understand the tail sensitivity of the norm. Interestingly, we have the following known relationship.

Proposition 22. *If $\phi_{\mathcal{R}}$ is the fundamental function of $\mathcal{R}$, then:*

$$E_{\mathcal{R}'}(t) = \phi(t)/t, \quad t \in (0, 1]. \quad (1.15)$$

In a nutshell: the fundamental function constrains the tail sensitivity of the norm. Hence, by studying properties of the fundamental function, we can gain insights into the tail sensitivity of law invariant coherent risk measures.

In the paper, we specifically work out the relevant relationships for the case of f -divergence risk measures, showing how to translate between the function f , the envelope function, and the function characterizing the Orlicz space. The results can be used in two directions: first, we can start with a choice of function f (such as the one corresponding to KL divergence), and then ask what this implies about the tail sensitivity. In this case, it turns out that the resulting *primal* envelope function has the tail behaviour of an exponential distribution, implying that the f -divergence risk measure with KL divergence is finite on random

variables with at most exponential tails. In the other direction, we can specify an envelope function (equivalently, fundamental function) with a specific tail behaviour, for example based on a Pareto distribution with parameter α , and then ask what the corresponding f -divergence risk measure and Orlicz space is. In the case of a Pareto distribution, we obtain the Lebesgue spaces.

In summary: in this paper, we have advanced the study of tail sensitivity for law invariant coherent risk measures, and specifically for f -divergence risk measures.

1.3 SUMMARY OF CHAPTER 4

Title: “Strictly Frequentist Imprecise Probability”

Abstract: Strict frequentism defines probability as the limiting relative frequency in an infinite sequence. What if the limit does not exist? We present a broader theory, which is applicable also to data that exhibit diverging relative frequencies. In doing so, we develop a close connection with the theory of imprecise probability: the cluster points of relative frequencies yield a coherent upper prevision. We show that a natural frequentist definition of conditional probability recovers the generalized Bayes rule. Finally, we prove constructively that, for a finite set of elementary events, there exists a sequence for which the cluster points of relative frequencies coincide with a prespecified set which demonstrates the naturalness, and arguably completeness, of our theory.

With this project we enter Part II of the thesis, which conducts a fundamental investigation of imprecise probabilities from a generalized frequentist perspective. Already at the very beginning of the PhD, Bob suggested that I compare the works of [Ivanenko \[2010\]](#) and [Walley \[1991\]](#), since they both end up mathematically with upper expectations, but arrive at them from quite different perspectives; this is the starting point for the present paper. While Part I is mostly motivated by considerations of risk and ambiguity aversion, and therefore focuses on law invariant risk measures, Part II goes further and challenges the adequacy of a precise probabilistic description *even in principle* (that is, not due to epistemic shortcomings, although recall the discussion in Section 0.6).

In this paper, we rid ourselves of (some) metaphysical baggage and, like [von Mises \[1919\]](#), take a *strictly frequentist* perspective. That is, we dispense of any kind of “process randomness” (i.e. “sampling”) regarding individual outcomes, and instead simply consider a data sequence as given and fixed. Concretely, if Ω is our possibility set, by a data sequence we mean an infinite sequence of outcomes $\omega^\infty = (\omega_1, \omega_2, \dots)$ (clearly this implies new baggage from the infinite idealization). Recall [von Mises’ \[1919\]](#) two conditions for a “collective”:

vM1. Aggregate stability: the limiting relative frequency for each $A \in 2^\Omega$ exists:

$$P(A) := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \chi_A(\omega_i).$$

This limit is called “the probability of A ”.

vM2. Local irregularity: let $\mathcal{J}$ be a set which is finite or countable, and $S_j : \mathbb{N} \rightarrow \{0, 1\}$ a function called “selection rule” for each $j \in \mathcal{J}$. For each $j \in \mathcal{J}$ and $A \in 2^\Omega$, the selection rule S_j preserves limiting relative frequencies:

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \chi_A(\omega_i) \cdot S_j(i)}{\sum_{i=1}^n S_j(i)} = P(A).$$

Equivalently, we could call sequences satisfying these two conditions *stochastic*. In this paper we address the question: what happens when condition vM2 does not hold? It turns out that when we leave the realm of precise probability in this way, we naturally arrive at imprecise probabilities to express the aggregate instability of a sequence. Our approach is to combine the development of [Ivanenko \[2010\]](#) with the (arguably more accessible) mathematical language of [Walley \[1991\]](#), in order to construct a “statistical regularity” [[Ivanenko, 2010](#)] to any data sequence (in the “forward” direction of the paper); in the converse direction, we show that to any imprecise probability, we can explicitly construct a data sequence whose statistical regularity coincides with that imprecise probability.

For the forward direction of our theory we proceed in a similar way to [Ivanenko \[2010\]](#), but instead of finitely additive measures we work with so-called “linear previsions” [[Walley, 1991](#)] directly for improved clarity. Fix a data sequence ω^∞ . We are interested in the long-run behaviour of gamble averages; recall that a gamble $X \in L^\infty$ is a *bounded* function of signature $X : \Omega \rightarrow \mathbb{R}$. Throughout this paper, we require the boundedness assumption for technical reasons. Consider the sequence of gamble averages, specified by:

$$\text{avg}(X)(n) := \frac{1}{n} \sum_{i=1}^n X(\omega_i), \quad n \in \mathbb{N}.$$

where the gamble is evaluated at each of the outcomes of the data sequence. As special cases, we can apply to indicator functions to obtain relative frequencies of events, as:

$$r_A(n) = \frac{1}{n} \sum_{i=1}^n \chi_A(\omega_i), \quad n \in \mathbb{N}.$$

In the classical precise probabilistic setting, we would expect convergence of gamble averages to their expectation and convergence of relative frequencies to probabilities. What if this does not happen?

Consider now, with a slight change of type signature, the object:

$$E(n) := X \mapsto \frac{1}{n} \sum_{i=1}^n X(\omega_i), \quad n \in \mathbb{N}, \quad Z \in L^\infty.$$

For any fixed $n \in \mathbb{N}$, $E(n)$ is an additive expectation-like functional, defined for all gambles $X \in L^\infty$. In the language of [Walley \[1991\]](#), it is a *linear prevision*. Such linear previsions can be seen as elements of the dual space $(L^\infty)^*$; concretely, by $(L^\infty)^*$ we mean the dual space to L^∞ as a Banach space with the norm topology induced by $\|X\|_{L^\infty} := \sup_{\omega \in \Omega} X(\omega)$. This dual space consists of the continuous linear functionals $F : L^\infty \rightarrow \mathbb{R}$. Now, if we impose the so-called *weak* topology* on $(L^\infty)^*$ — a convenient choice — then it turns out that the set of linear previsions is a compact subset of L^∞ . In fact, it is the following set:

$$\text{PF}(\Omega) := \{E \in (L^\infty)^* : E(X) \geq 0 \text{ whenever } X \geq 0, E(\chi_\Omega) = 1\}.$$

And thus $E(n) \in \text{PF}(\Omega)$.

Remark 23. It should be clear from context whether we mean by $E(n)$ the n -th element of the sequence or the sequence itself.

Due to the compactness of the set of linear previsions under the weak*-topology on $(L^\infty)^*$, we know that the sequence $E(n)$ has a non-empty closed set of *cluster points*. A cluster point generalizes the concept of a limit. Recall that a point z is a cluster point of a sequence $S : \mathbb{N} \rightarrow \mathcal{T}$, where $\mathcal{T}$ is any topological space, if:

$$\forall N, \text{ where } N \text{ is a neighbourhood of } z \text{ with respect to } \mathcal{T}, \forall n_0 \in \mathbb{N} : \exists n \geq n_0 : S(n) \in N.$$

Intuitively, a cluster point is visited infinitely often by the sequence. However, this does not imply that those cluster points are limits of convergent subsequences — this would require sequential compactness, which would be satisfied under the assumption that $|\Omega| < \infty$. We denote the set of cluster points as $\text{CP}(E(n))$. Based on this, we define an upper prevision as:

$$\overline{R}(X) := \sup \{E(X) : E \in \text{CP}(E(n))\}, \quad X \in L^\infty,$$

and an upper probability as

$$\overline{P}(A) := \sup \{E(\chi_A) : E \in \text{CP}(E(n))\}, \quad A \subseteq \Omega.$$

Note that we do not need to restrict ourselves to a smaller set system than 2^Ω , even if Ω is infinite. This allows to derive the following nice result.

Proposition 24. *The sequence of gamble averages $\text{avg}(X)$ has the set of cluster points (with respect*

to the Euclidean topology on $\mathbb{R}$):

$$\text{CP}(\text{avg}(X)) = \{E(X) : E \in \text{CP}(E(n))\},$$

and therefore

$$\overline{R}(X) = \sup \text{CP}(\text{avg}(X)) = \limsup_{n \rightarrow \infty} \text{avg}(X)(n).$$

$$\text{Similarly, for any } A \subseteq \Omega, \overline{P}(A) = \limsup_{n \rightarrow \infty} r_A(n) = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \chi_A(\omega_i).$$

In summary: the set of cluster points $\text{CP}(E(n))$ of linear previsions indeed deserves the name “statistical regularity”, as it suffices to compute the cluster points for any gambles, in particular for indicator functions. In the paper, we also provide a decision-oriented justification for why the supremum over the cluster points of gamble averages is a relevant object of interest. In this way, $\overline{R}$ and $\overline{P}$ naturally generalize the precise probabilistic concepts of $\mathbb{E}$ and P to the case of potential divergence.

Our paper also provides a “converse” direction. For this, we assume $|\Omega| = k < \infty$ to rid ourselves of topological headaches, so that we are licensed to work with the Euclidean topology (why? See Section 4.8.1). Then there is an easy way to think about the linear previsions $E(n)$: simply compute the relative frequencies $r_{\{\omega^1\}}(n), \dots, r_{\{\omega^k\}}(n)$ for all elementary events $\omega^i \in \Omega, i = 1..k$, and stack them in a vector $r(n) \in \mathbb{R}^k$. In fact, $r(n)$ is a legitimate probability measure, and in a one-to-one correspondence to $E(n)$. Geometrically, we can think of a finite-dimensional probability measure as a point in the Euclidean simplex due to the unit normalization:

$$\Delta^k := \left\{ d = (d_1, \dots, d_k) \in \mathbb{R}^k : \sum_{i=1}^k d_i = 1, d_i \geq 0 \forall i \in \{1, \dots, k\} \right\}.$$

We can then prove the following.

Theorem 25. *Let $|\Omega| < \infty$. Let $\overline{R}$ be a coherent upper prevision on L^∞ . There exists a sequence $\omega^\infty = (\omega_1, \omega_2, \dots)$ such that we can write $\overline{R}$ as:*

$$\overline{R}(X) = \overline{R}_{\omega^\infty}(X) := \sup \{E(X) : E \in \text{CP}(E_{\omega^\infty})\}, \quad \forall X \in L^\infty,$$

where $E_{\omega^\infty}(n) := X \mapsto \frac{1}{n} \sum_{i=1}^n X(\omega_i)$, recalling that $\omega^\infty = (\omega_1, \omega_2, \dots)$.

This result, which is arguably more general than a related result by [Ivanenko \[2010\]](#) (see the discussion in Section 4.3) rests almost entirely on Theorem 194, proven by Bob. It also follows as a special case of the more general Theorem 240 which we prove in Chapter 5, building on some techniques from the proof of Theorem 194.

To put this result into perspective: together with the “forward” direction, we have indeed established a strictly frequentist theory of imprecise probability — the aggregate characteristics of any sequence can be captured by an imprecise probability, and conversely, to any imprecise probability we can construct a corresponding sequence. In essence, we have shown that imprecise probability naturally arises when we deal with sequences whose aggregate regularity is “non-stochastic”, i.e. when condition vM1 is violated.

In addition, the paper also introduces a suitable notion of conditional imprecise probability in the case of aggregate instability, which generalizes the corresponding notion by [von Mises \[1919\]](#). Consequently, a notion of independence can also be introduced (Appendix 4.9). As to relevant previous work, we note that besides the work by [Ivanenko \[2010\]](#), which directly inspired our work, Section 4.3 in the work of [Walley and Fine \[1982\]](#) also contains relevant pieces in this direction but is less general.

1.4 SUMMARY OF CHAPTER 5

Title: “Data Models With Two Manifestations of Imprecision”

Abstract: Motivated by recently emerging problems in machine learning and statistics, we propose data models which relax the familiar *i.i.d.* assumption. In essence, we seek to understand what it means for data to come from a set of probability measures. We show that our frequentist data models, parameterized by such sets, manifest two aspects of imprecision. We characterize the intricate interplay of these manifestations, aggregate (ir)regularity and local (ir)regularity, where a much richer set of behaviours compared to an *i.i.d.* model is possible. In doing so we shed new light on the relationship between non-stationary, locally precise and stationary, locally imprecise data models. We discuss possible applications of these data models in machine learning and how the set of probabilities can be estimated. For the estimation of aggregate irregularity, we provide a negative result but argue that it does not warrant pessimism. Understanding these frequentist aspects of imprecise probabilities paves the way for deriving generalization of proper scoring rules and calibration to the imprecise case, which can then contribute to tackling practical problems.

In this project we directly follow up on the “strictly frequentist” development, but now make a move akin to that from [von Mises \[1919\]](#) to [Kolmogorov \[1933\]](#), moving from *data sequences* to *data models*. In the strictly frequentist framework, the concept of an infinite data sequence is fundamental, and (imprecise) probability is *defined* from its limiting behaviour, where our general theory allows potential divergence; yet we cannot speak of “sampling”

individual outcomes, since there is no notion of process randomness. As a consequence, grappling with the case of finite data is a key challenge.

In the precise case, the remedy is to put a probability measure on the set of sequences, the most important example being the familiar *i.i.d.* model: “assume $X_1, X_2, \dots \sim P_0$ are *i.i.d.*”¹⁶ This allows one to make statements about the randomness of individual outcomes (process randomness, see [Eagle, 2021]). A nice visual terminology, that I learned from Rabanus Derr, distinguishes *vertical* from *horizontal* randomness. Here, vertical randomness refers to the hypothetical randomness that concerns an individual outcome: X_i *could have had* a different value. In contrast, if you imagine a sequence as laid out horizontally, horizontal randomness is the outcome randomness that von Mises [1919] deals with, the actually realized randomness, which concerns properties of the actually realized aggregate. Based on this distinction, the current project can be framed as follows: we construct “vertical” wrappers around the strictly frequentist development of the previous project. The result is a cleanly laid out conceptual and mathematical interaction of the “vertical” and “horizontal” aspects, both involving imprecise probabilities. In this way, we add more assumptions but also end up with a more general picture. For instance, our main theorem here (Theorem 240) yields the main theorem (Theorem 192) of Chapter 4 as a special case. To us, a *data model* is essentially a theoretical construct which supplies a *typicality* notion for data sequences. In simple words, a data model tells us which data sequences we should expect to see. In the familiar framework due to Kolmogorov [1933], the typicality notion is itself a probabilistic one: recall our statement of the *i.i.d.* model in the introduction, where the Lebesgue measure λ on the unit interval (formally a probability measure) expresses typicality. For example, an *almost sure* statement is such a typicality statement — it serves to justify neglecting the complementary event. The probabilistic nature of this typicality notion is well-summarized by La Caze [2016, emphasis in original]:

Importantly, ‘almost sure convergence’ is also given a frequentist interpretation. Almost sure convergence is taken to provide a justification for assuming that the relative frequency of an attribute *would* converge to the probability in actual experiments *were* the experiment to be repeated indefinitely.

For example, consider how *p*-values are commonly explained in terms of indefinitely repeating an experiment under stable conditions.

We believe it is no conceptual necessity that a typicality notion is itself probabilistic, as it serves an entirely different purpose than a probability measure on the possibility set. Thus, we have found inspiration in the work of Galvan [2006], who axiomatically investigated more general typicality notions. In our case, we will also consider imprecise probabilistic

¹⁶By “ $\sim$ ”, we denote equality of probability distributions (with respect to a base measure).

typicality, which yields data models with intriguing interactions with “horizontal” (strictly frequentist) imprecise probability.

To begin with, however, we study a less radical data model, which keeps probabilistic typicality. Recall the non-stationary, locally precise model which we have already introduced in Chapter 0:

Proposition 26. *Let $\mathcal{Z} := [0, 1)$ and $(P_1, P_2, \dots)$ be a sequence of probability measures on Ω . Then there exist $W_i : [0, 1) \rightarrow \Omega, i \in \mathbb{N}$, so that the following holds:*

$$\lambda \{z \in \mathcal{Z} : W_1(z) = \omega_1, W_2(z) = \omega_2, \dots, W_n(z) = \omega_n\} = \prod_{i=1}^n P_i(\{\omega_i\}), \quad \forall n \in \mathbb{N}.$$

We assume finite Ω for simplicity in the paper. Here, $\mathcal{Z}$ is the realm of typicality, meaning λ is our probabilistic *typicality* measure. In contrast, the P_i are *probability* measures on the possibility set Ω (in the paper, we reserve the term *probability* for appropriately normalized measures defined on Ω). In other words, the “vertical” refers to $\mathcal{Z}$, whereas the “horizontal” refers to aggregates of outcomes from Ω . This model keeps the *independence*, but not the *identical* in “i.i.d.”. For example, it fits well with the paradigm of group distributional robustness. As an even simpler example, consider flipping coins with varying biases independently.

Clearly, this model can express violations of condition vM2, that is, we can incorporate hidden heterogeneity. For example, construct the sequence as $(P_1, P_2, P_1, P_2, \dots)$ by alternating and recall that a feature map is $X : \Omega \rightarrow \mathbb{R}$. For example, P_1 could correspond to a majority sub-population, whereas P_2 corresponds to a minority. If P_1 and P_2 , as probabilities on Ω , actually differ, this may be “invisible” to the predictor $f(X)$ (for intuition, consider the case when we have essentially no features, i.e. a constant feature map).

Another question of interest is what kind of aggregate (in)stability can result from such a data model. For example, if we alternate between P_1 and P_2 , we will typically (with Lebesgue measure 1) obtain data sequences $\omega^\infty = (\omega_1, \omega_2, \dots)$ with convergent relative frequencies, specifically with convergence to $\frac{1}{2}P_1 + \frac{1}{2}P_2$. Given a set $\mathcal{P}$ of probabilities on Ω to choose from, i.e. an imprecise probability, could we also construct sequences of probabilities from $\mathcal{P}$ such that the resulting data model typically yields sequences with divergent relative frequencies? The answer is yes, and our main theorem (Theorem 240) fully characterizes the possible relationships between violating vM1 and vM2.

To state our theorem, if $\Omega = \{\omega^1, \dots, \omega^k\}$, we define the relative frequencies of a data sequence $\omega^\infty = (\omega_1, \omega_2, \dots)$ as:

$$r^{\omega^\infty}(n) := \left(\frac{1}{n} \sum_{i=1}^n \chi_{\{\omega^1\}}(\omega_i), \frac{1}{n} \sum_{i=1}^n \chi_{\{\omega^2\}}(\omega_i), \dots, \frac{1}{n} \sum_{i=1}^n \chi_{\{\omega^k\}}(\omega_i) \right)^\top \in \Delta^k, \quad n \in \mathbb{N}.$$

Due to the finite dimensionality, we can identify a probability on Ω with a point in the simplex Δ^k . We write $G(z) = (W_1(z), W_2(z), \dots)$ for the data sequence induced by a seed $z \in \mathcal{Z}$, and we write $(\omega_1, \omega_2, \dots, \omega_n) \subset G(z)$ if $(\omega_1, \omega_2, \dots, \omega_n)$ is a prefix of the sequence $G(z)$. Then:

Theorem 27. *Let $\mathcal{Z} = [0, 1)$ and $\emptyset \neq \mathcal{P} \subseteq \Delta^k$ an imprecise probability. Then there exists a sequence of probability measures $P_1, P_2, \dots \in \mathcal{P}$ with independent product*

$$\lambda \{z \in \mathcal{Z} : (\omega_1, \omega_2, \dots, \omega_n) \subset G(z)\} := \prod_{i=1}^n P_i(\{\omega_i\}), \quad \forall n \in \mathbb{N}, \quad (1.16)$$

so that it holds

$$\lambda \left\{ z \in \mathcal{Z} : \text{CP} \left(n \mapsto r^{G(z)}(n) \right) = \text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n P_i \right) = \mathcal{Q} \right\} = 1 \quad (1.17)$$

if and only if $\emptyset \neq \mathcal{Q} \subseteq \text{co}(\mathcal{P})$ is a closed connected subset of the closed convex hull of $\mathcal{P}$.

We can in addition demand that any $P \in \mathcal{P}$ is a cluster point of the sequence of measures P_i (with respect to the Euclidean metric).

In words: if we form an independent product, using probabilities from a set $\mathcal{P}$, then the resulting aggregate (in)stability (potentially we could have stability) will *typically* correspond to a closed connected subset of the closed convex hull of $\mathcal{P}$. As a consequence, under such a data model, breaking vM2 is a necessary but not sufficient condition for breaking vM1.

Remark 28. In the paper, I have used the terms *aggregate regularity* and *local irregularity* for the case when condition vM1 and condition vM2, respectively, hold; and conversely, *aggregate irregularity* and *local regularity* when they do not. I now believe that the pair *aggregate (in)stability* is an even better description for condition vM1.

In the paper, we go on to investigate *stationary, locally imprecise data models*, building on the work of the line of work by [Walley and Fine \[1982\]](#), [Kumar and Fine \[1985\]](#), [Grize and Fine \[1987\]](#), [Fine \[1988\]](#), [Papamarcou and Fine \[1991a,b\]](#) and [Sadrolhefazi and Fine \[1994\]](#). Here, we move beyond probabilistic typicality to imprecise probabilistic forms of typicality. Interestingly, these more exotic models can be “constructed” by taking the non-stationary, locally precise models as the basic building blocks. Informally, if $(P_1, P_2, \dots)$ is any sequence of probability measures from a set $\mathcal{P}$, and considering the induced independent product measure P^∞ on the set of Ω -valued sequences, we can construct a data model by setting:

$$\overline{P}_\mathcal{P}^\infty(A^\infty) := \sup\{P^\infty(A^\infty) : P^\infty = (P_1, P_2, \dots) : P_i \in \mathcal{P}, i \in \mathbb{N}\},$$

where A^∞ is a set of infinite sequences itself, and we must restrict the domain of definition (see the paper). Hence $\overline{P}_\mathcal{P}^\infty$ is a coherent upper probability, defined on sets of sequences.

Beyond providing a relatively accessible introduction to, and discussion of, previous work on such models, we offer some new insights. In particular, we prove an intriguing impossibility result for the estimation of aggregate instability under such data models (Proposition 263).

1.5 SUMMARY OF CHAPTER 6

Title: “Scoring Rules and Calibration for Imprecise Probabilities”

Abstract: What does it mean to say that, for example, the probability for rain tomorrow is between 20% and 30%? The theory for the evaluation of precise probabilistic forecasts is well-developed and is grounded in the key concepts of *proper scoring rules* and *calibration*. For the case of imprecise probabilistic forecasts (sets of probabilities), such theory is still lacking. In this work, we therefore generalize proper scoring rules and calibration to the imprecise case. We develop these concepts as relative to data models and decision problems. As a consequence, the imprecision is embedded in a clear context. We establish a close link to the paradigm of (*group*) *distributional robustness* and in doing so provide new insights for it. We argue that proper scoring rules and calibration serve two distinct goals, which are aligned in the precise case, but intriguingly are not necessarily aligned in the imprecise case. The concept of decision-theoretic entropy plays a key role for both goals. Finally, we demonstrate the theoretical insights in machine learning practice, in particular we illustrate subtle pitfalls relating to the choice of loss function in distributional robustness.

This paper follows up on the generalized frequentist picture we presented in terms of data models, and develops a foundational framework for the evaluation of *imprecise forecasts* under such data models. To us, an imprecise forecast is simply a prediction $\mathcal{Q}(X)$, based on some features X , which is an imprecise probability on Ω . In light of a given data sequence or data model, we are then interested in evaluating such a forecaster: how “good” is a specific forecast? When is a forecaster A better than a forecaster B ? This would also give us a way of training a machine learning model, say a neural network, to give imprecise probabilistic forecasts. In fact, in this work, we develop a close relation to the paradigm of *group distributional robustness* [Sagawa et al., 2019] and, through both theory as well as experiments, provide new insights for it.

We take inspiration from the evaluation of precise probabilistic forecasts, where the forecaster outputs a precise probability. In the familiar precise case, the two key concepts are *proper scoring rules* and *calibration*. While at the ISIPTA 2023 conference, I learned about

the (until then open) problem of constructing similar proper scoring rules for imprecise probabilities. Previous works [Seidenfeld et al., 2012, Mayo-Wilson and Wheeler, 2016, Schoenfeld, 2017, Konek, 2019, 2023] have mostly yielded *impossibility* rather than possibility results. My subjective analysis of the situation is as follows: previous work was firmly subjectivist, and did not attempt to develop scoring rules concepts with respect to data (model) and decision problem, but these are two key ingredients for the evaluation of forecasts. Thus, in contrast to previous work, I proceed by putting data (model) and decision problem at center stage from the beginning, and use such semantical considerations to guide the mathematical development. The result is a viable (I believe) account of proper scoring rules, as well as a notion of calibration, for imprecise probabilities.

The setup is as follows. Let Ω , with $|\Omega| < \infty$ for simplicity, be the possibility set. To us, the concept of a forecast is intrinsically linked to making decisions, i.e. selecting actions. Hence, assume an *action space* $\mathcal{A}$, with $|\mathcal{A}| < \infty$ (merely to simplify the mathematics). We assume there is an agent who uses a *loss function* $\ell : \mathcal{A} \times \Omega \rightarrow \mathbb{R}$ to evaluate the consequences of taking action $a \in \mathcal{A}$ when the state of the world $\omega \in \Omega$ actually materializes. Often, the loss function would depend on ω only indirectly through some label $Y(\omega)$ and the chosen action. As a simple example, the forecast could be about the event of rain, and the agent needs to decide whether or not to bring an umbrella.

If $X : \Omega \rightarrow \mathcal{X} \subset \mathbb{R}$ are available features, then an *imprecise forecast* is any imprecise probability $\mathcal{Q}(X)$ on Ω , i.e. $\mathcal{Q} : \mathcal{X} \rightarrow \mathcal{IP}$, where $\mathcal{IP}$ is the set of imprecise probabilities (sets of probabilities, not necessarily closed convex) on Ω . The first question to answer is: given an imprecise probabilistic (IP) forecast, how does the agent pick an action? In this paper, we do not allow indeterminism in the picture, but instead assume that the agent picks according to the Γ -MinMax (MaxMin, up to a sign flip) principle [Gilboa and Schmeidler, 1989] — a highly consequential assumption, and a target of potential criticism:

$$a_{\mathcal{Q}(x)}^* \leftarrow \operatorname{argmin}_{a \in \mathcal{A}} \sup_{Q \in \mathcal{Q}(x)} \mathbb{E}_Q[\omega \mapsto \ell(a, \omega)], \quad (1.18)$$

Here, the “ $\leftarrow$ ” operator consistently resolves the problem of tie-breaking. Thus, given an imprecise forecast, the agent picks the action in a cautious (pessimistic) way. Based on this, we can define our IP scoring rule, relativized to the loss function ℓ :

$$S_\ell(\mathcal{Q}(X(\omega)), \omega) := \ell(a_{\mathcal{Q}(X(\omega))}^*, \omega). \quad (1.19)$$

Intuitively, given an IP forecast $\mathcal{Q}(X(\omega))$, we consider the action that the agent picks according to this forecast (specified above), and then look at the resulting consequence in terms of the loss function. For a single ω , this evaluates the quality of the forecast based on the decision-making consequence.

A key question remains: how to aggregate? In practice, we make decisions repeatedly, i.e. over many ω s, and we want a single summary for the quality of a forecaster. To answer this, we add a data model to the picture; the aggregation is then in accordance with the data model. In the paper, we focus on data models which are parameterized by an imprecise probability $\mathcal{P}$, but in fact the conceptual development may also be of relevance under precise data models (e.g. to model ambiguity). As an example, recall the non-stationary, locally precise data model parameterized by some set of probabilities $\mathcal{P}$. Intuitively, this data model asserts that outcomes are drawn independently from a sequence of probabilities $(P_1, P_2, \dots)$, where $P_i \in \mathcal{P}, i \in \mathbb{N}$. In particular, group distributional robustness can be framed in this way, where $\mathcal{P}$ contains the group-wise probabilities. More generally, our development simply requires some data model which is parameterized by an imprecise probability.

We assume that the forecaster intends to achieve a robustness with respect to the set of probabilities, and we formalize this through a second MinMax assumption. For example, in group distributional robustness, this is the familiar goal of deploying a machine learning model with robust performance with respect to the set of groups. As a result, the overall objective of interest is as follows:

$$\text{IP-Score}^*(\ell, \mathcal{P}) := \min_{Q: \mathcal{X} \rightarrow \mathcal{IP}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\omega \mapsto S_\ell(Q(X(\omega)), \omega)], \quad (1.20)$$

That is, we minimize over all possible IP forecasts, and we maximize (take the supremum, if $\mathcal{P}$ is not closed) over the probabilities $\mathcal{P}$, which parameterize the data model. In this way, we perform a cautious (pessimistic) aggregation of the decision-making consequences, where we consider the actions recommended to the agent by the forecaster.

The two MinMax assumption assumptions interact nicely, and as a result we obtain a properness statement for our scoring rules, although the situation is rather intricate (see Section 6.2.1). Furthermore, we can — under a subtle technical condition — establish a close link to the *maximum entropy principle* of Grünwald and Dawid [2004], showing that the maximum entropy forecast is optimal in terms of IP score.

The second part of the paper develops a corresponding notion of calibration for imprecise forecasts, and shows how in the imprecise case the goal of being calibrated may diverge from the goal of optimality in terms of score — an intriguing result, since in the precise case these goals are compatible. Our motivation for calibration is insurantal, and at first sight might not resemble familiar calibration, but they are deeply related (see [Zhao et al., 2021a]). Essentially, we say that a forecaster is calibrated if they can offer *actuarially fair* insurance to an agent with some loss function. By “actuarially fair” insurance [Arrow, 1963], we mean that the forecaster neither profits from offering the insurance, nor goes bankrupt, *under the data model*; the “claims” to be insured are the uncertain partial losses

$\ell(a_{\mathcal{Q}(X(\omega))}^*, \omega)$. Intuitively, a calibrated imprecise forecaster can provide a trustworthiness guarantee to the agent: *if you pick the optimal action in accordance with my forecast, then this is how much loss you have to expect in the worst-case long-run*, where the “worst-case” refers to the cautious aggregation under the data model, i.e. relative to $\mathcal{P}$.

Intriguingly, we show that the two qualities of forecasts are not always aligned in the imprecise case. That is, the optimal IP forecast need not be IP calibrated! This contrasts with the precise case, where the Bayes predictor, optimal in terms of score under any proper scoring rule, is also perfectly calibrated. Perhaps we should have anticipated that these concepts may diverge; conceptually, they appear to be different. The first quality of a forecast, related to scoring rules, is about recommending actions with favorable outcomes to an agent; the second quality, related to calibration, is about providing trustworthy assessments of the consequences of one’s action recommendations. Clearly, one can give bad recommendations, but be aware of how bad they are. Conversely, and more interestingly, in the imprecise case it is possible to give optimal recommendations, but these optimal recommendations can be overconfident in a way, and therefore not trustworthy. We observe that this perhaps puzzling situation is essentially related to the question of updating imprecise probabilities, a controversial topic in the literature [Gong and Meng, 2021].

Finally, in our experiments (Section 6.4) we illustrate the relevance of our theory and the insights it has generated for group distributional robustness. In particular, we highlight that the choice of loss function has to be carefully considered in group DRO, in a way that is fundamentally different to how the loss function matters in expected risk minimization. This fits with our overarching narrative: our scoring rules and calibration concepts are always relativized to data model and decision problem.

1.6 SUMMARY OF CHAPTER 7

Title: “Insights From Insurance for Fair Machine Learning”

Abstract: We argue that insurance can act as an analogon for the social situatedness of machine learning systems, hence allowing machine learning scholars to take insights from the rich and interdisciplinary insurance literature. Tracing the interaction of uncertainty, fairness and responsibility in insurance provides a fresh perspective on fairness in machine learning. We link insurance fairness conceptions to their machine learning relatives, and use this bridge to problematize fairness as calibration. In this process, we bring to the forefront two themes that have been largely overlooked in the machine learning literature: responsibility and aggregate-individual tensions.

We have already met the idea that studying insurance is helpful to better understand uncertainty in general, and specifically in machine learning. In this project, we shift the focus slightly and focus on lessons from insurance for *fair machine learning*. This recently emerging sub-field of machine learning studies the fairness of machine learning systems [Barocas et al., 2019], for example with the goal of avoiding discrimination. In fact, uncertainty again plays a major role in this, hence we might say that we are focusing on the interaction of fairness and uncertainty, but we also broaden the scope in the paper beyond uncertainty. In any case, after working on this project, I was persuaded that the research heuristic of taking insights from insurance for machine learning (or more generally, any statistical technology) is a most fruitful one.

I believe that the paper is relatively accessible, and also short, and thus less in need of a summary. Briefly, the main storyline goes as follows: in insurance, there exists a fundamental tension between the principles of *solidarity* and *actuarial fairness*. Solidarity calls for equal contribution to the insurance pool, whereas actuarial fairness demands to set the insurance premium in accordance with the expected loss for each individual, or in practice, within each class.¹⁷ We show that an analogous tension exists between recently proposed fairness notions in machine learning, and we can thus learn from the social studies of insurance scholars. More specifically, we demonstrate how the tension can be better understood based on notions of causality and control, as well as responsibility and responsabilization.

A key insight, which I want to highlight specifically here, relates to the concept of *performativity*. The paper contains a section on performativity in Appendix 7.8.4; we choose to “hide” it in the appendix due to the page limit and the high level of abstraction, contrasting

¹⁷While the principle of actuarial fairness is typically framed as concerning individuals, in practice insurers classify people, i.e. put them in groups, to be able to apply probabilistic calculus.

with more concrete insights. However, to me personally it is perhaps the most interesting insight from the social studies of insurance.

Performativity is a broad notion, and other insights in the paper could be understood by viewing them through this lens. As a concrete example, we demonstrate that some relevant questions which appear to be descriptive (e.g. is the individual in control of some behaviour?) are in fact implicitly *normative*. Consequently, actuarial practice is not “objective” and “neutral” [Glenn, 2003a]. So what is performativity? See Section 7.8.4 for some ways of approaching this highly subtle concept.

Here, I want to attempt yet another framing. Consider two viewpoints at the end of a spectrum: on the one hand, we could believe that statistics (the actuarial method) is entirely objective and neutral, merely in the business of accurately representing a reality out there; also, the objective probability (e.g. that some individual makes a claim) exists. At the other extreme, one viewpoint could claim that anything is entirely subjective; probabilities are subjective, and how the world is “carved up” is just as subjective as well. What I believe is so intriguing about performativity approaches¹⁸, is that they are situated in-between these two extremes, and put forward very insightful and subtle ways of understanding the in-betweenness (unfortunately, easily misunderstood). A performativity perspective explicitly acknowledges that *representation is practice*,¹⁹ and moreover this perspective can indeed change practice (as pointed out by Mol [2002], *doing* theory is again a practice, which may or may not have an effect).

What does it help to have an understanding of representation as practice? Not only in insurance, but also in machine learning, I think this is a highly relevant perspective: crucially, it directs our attention to how we could have represented differently. In machine learning research as well as practice, we often find beliefs in the objectivity of the statistical technology [Moss, 2021]. Yet, to name just a few examples, the way we choose the categories in the data, what information we collect or we do not collect, which aggregate criteria we use for evaluation — many choices are implicit in representational practice and influence the result. In particular, normative considerations (e.g. are individuals responsible for some behaviour) can be implicit in this but might not be visible. Yet if they are not visible, they are also not open to criticism. The field of data-driven machine learning has now acknowledged the importance of data in the machine learning pipeline — I believe that performative perspectives in this field could significantly improve machine learning practice.

¹⁸By this, I mean various works that loosely fit this broad umbrella term. There is no single account of “performativity”, but rather a cluster of accounts.

¹⁹In the paper, I have said, following Vosselman [2014], that “representation and intervention are entangled”.

1.7 KEY CONTRIBUTIONS

In this section, I provide a concise summary of the key contributions of the papers at a medium to high level of abstraction for the reader's convenience. Many of the passages are directly derived from the original papers,²⁰ while also changes have been made to improve conciseness. Quotations are not used in this section.

KEY CONTRIBUTIONS OF "RISK MEASURES AND UPPER PROBABILITIES: COHERENCE AND STRATIFICATION" (CHAPTER 2)

We study in depth the connections between coherent risk measures, imprecise probabilities and rearrangement invariant Banach function spaces. In this way, results from one literature can be imported into another literature. In particular, we obtain many insights from the work on rearrangement invariant spaces. From this perspective, we derive various novel results for coherent risk measures, typically in terms of their fundamental function, which corresponds to an imprecise probability. For example, we show how the fundamental function allows a comparison of risk measures in terms of their tail sensitivity. In addition, we present some new results regarding the combination of two risk measures, showing relationships to the theory of interpolation of operators.

Throughout the paper, we specifically focus on the class of spectral risk measures. We bring together many different characterizations of them in the paper, which underlines the distinctiveness of the spectral class.

We conduct experiments, which demonstrate that spectral risk measures can encode risk aversion and robustness. For our experiments, we propose a new graphical way (based on " CVar_α curves") for evaluating the tail risk of a loss distribution.

KEY CONTRIBUTIONS OF "TAILORING TO THE TAILS: RISK MEASURES FOR FINE-GRAINED TAIL SENSITIVITY" (CHAPTER 3)

Expanding on some first results of the previous paper, we establish a general framework to study the tail sensitivity of coherent risk measures based on their fundamental function. Again, we leverage insights from the rearrangement invariant Banach function spaces literature to this end. We explicate how the fundamental function is, via its one-to-one relation to the envelope function, intricately linked to the tail sensitivity of a risk measure. We show how one can therefore construct a corresponding Marcinkiewicz norm, Orlicz norm

²⁰Since I have written the respective contribution passages (sometimes in an explicit contribution section, or otherwise contained in the introduction or discussion sections) of the papers, co-authors do not need to be mentioned here, with the exception of one sentence in the contribution section of Paper PP1 which was written by Bob.

(hence, a divergence risk measure) and a Lorentz norm, which share this tail sensitivity. We illuminate the connections and (non)-equivalences between these norms.

Beyond tail sensitivity in general, the major focus of this paper lies on the class of f -divergence risk measures. We define the Orlicz regret measure, which is an asymmetric variant of an Orlicz norm, and show that its “natural extension” yields the f -divergence risk measure. Following [Dommel and Pichler \[2020\]](#), we embed the Orlicz regret measure and the f -divergence risk measure in the mathematical framework of rearrangement invariant Banach function spaces and study their fundamental functions. We show how our general findings regarding tail sensitivity specialize to the case of Orlicz spaces, and hence f -divergence risk measures. Finally, we link our tail sensitivity approach to Orlicz deviation inequalities, of which many well-known concentration inequalities are special cases.

KEY CONTRIBUTIONS OF “STRICTLY FREQUENTIST IMPRECISE PROBABILITY” (CHAPTER 4)

In this paper, we develop a generalized strict frequentism for the case when condition vM1 does not hold, i.e. when the data sequence exhibits aggregate instability.

We follow the approach of [Ivanenko \[2010\]](#), but reformulate his construction of a statistical regularity of a sequence using the mathematical language of [Walley \[1991\]](#). In essence, to each infinite sequence we can naturally associate a set of probability measures, which constitute the statistical regularity that describes the cluster points of relative frequencies and consequently also those of sample averages. Since this works for any sequence and any event, we have thus countered a typical argument against frequentism, namely that the limit may not exist and hence probability is undefined [[Hájek, 2009](#)].

We demonstrate that (for a finite set of outcomes) the converse direction works, too: any coherent upper prevision can be induced in a strictly frequentist way from a sequence, which we can explicitly construct.

Furthermore, we derive in a natural way a conditional upper prevision; this approach recovers the generalized Bayes rule, the arguably most important updating principle in imprecise probability.

KEY CONTRIBUTIONS OF “DATA MODELS WITH TWO MANIFESTATIONS OF IMPRECISION” (CHAPTER 5)

Essentially, here we show how the “[von Mises \[1919\]](#) to [Kolmogorov \[1933\]](#)” move, i.e. from strict to hypothetical frequentism (from data sequences to data models), works for the more general imprecise case.

First, we consider a data model which still assumes precise, probabilistic independence,

but where we allow a set of probability measures instead of a single one. We call this the non-stationary, locally precise data model. Our main theorem characterizes possible manifestations of imprecision in such models as aggregate instability and local regularity. Specifically, we show that the set of cluster points is typically a closed connected subset of the closed convex hull of the set of probabilities. In particular, this theorem recovers as a special case the main theorem of the previous paper.

We discuss stationary, locally imprecise data models and compare them to the highly related work of [Walley and Fine \[1982\]](#) and subsequent papers by Terrence Fine and collaborators, and to generalized laws of large numbers. In doing so, we provide a relatively accessible introduction to previous work. In addition, we make some remarks about estimation. Here, we provide a negative result concerning an estimator proposed by [Walley and Fine \[1982\]](#), but argue that it is not as troubling as it may appear at first sight.

KEY CONTRIBUTIONS OF “SCORING RULES AND CALIBRATION FOR IMPRECISE PROBABILITIES” (CHAPTER 6)

We develop a viable account of proper IP scoring rules, which are relativized to data models and decision problems. We characterize the optimal IP forecast for a fixed loss function based on the maximum entropy principle by [Grünwald and Dawid \[2004\]](#). We demonstrate that imprecision in the forecasts might be required if the loss function is not known a priori. We shed light on the tight connection between our proposal of imprecise scoring rules and the paradigm of (group) distributional robustness [[Hu et al., 2018](#), [Sagawa et al., 2019](#)], where a coherent risk measure of the loss is optimized.

Further, we develop an adequate concept of calibration for imprecise probabilities. As for scoring rules, the concept of a generalized decision-theoretic entropy plays a key role here. We discuss how proper scoring rules and calibration serve two distinct goals of forecasting quality, and show that they do not necessarily align anymore in the imprecise case. In particular, the optimal forecast (in the sense of score) need not be calibrated.

Finally, our experiments illustrate the theoretical insights in machine learning practice. Specifically, we draw attention to the choice of loss function in group distributional robustness.

In addition, we believe that even readers who are only interested in precise probability may profit from reading this paper, as the development we offer here provides new insight into the conceptual interplay between scoring rules and calibration in general.

KEY CONTRIBUTIONS OF “INSIGHTS FROM INSURANCE FOR FAIR MACHINE LEARNING” (CHAPTER 7)

The main claim of our work is that insurance is an insightful analogon for the social situatedness and impact of machine learning systems, specifically with respect to fairness. We hence establish a conceptual bridge between insurance and machine learning, and present it in an accessible way to machine learning scholars.

We offer the following insights. Recent impossibility theorems in the fair machine learning literature [[Kleinberg et al., 2017](#)] are not as surprising when considering them in the light of the old, fundamental tension in insurance between solidarity and actuarial fairness. In essence, this tension is grounded in how individuals are related to the aggregates they form. This relation rests on responsabilization. Responsibility and responsabilization should be conceptually distinguished, even if in the recent mode of personalized insurance (tightly linked to machine learning) the two are increasingly intertwined. For insurance and machine learning purposes, reasoning about responsibility crucially requires reasoning about causality and control. We have emphasized that the relevant control question has a normative flavour, and thus cannot be left to engineers alone. What is under individuals' control is often hotly contested.

Part I

Risk and Ambiguity — Law Invariance

2

Risk Measures and Upper Probabilities: Coherence and Stratification

AUTHOR CONTRIBUTIONS

Robert Williamson (RW) provided the initial idea and relevant literature for relating coherent risk measures (in particular, spectral risk measures) to rearrangement invariant Banach norms, and also suggested to consider the work of [Walley \[1991\]](#). Some observations, e.g. the relevance of the fundamental function, were previously made by RW. Moreover, RW supervised the project. Further, RW is responsible for the entirety of Section 2.5 as well as the proofs associated with this section (found in the Appendix). The remaining content and writing is due to Christian Fröhlich.

ABSTRACT

Machine learning typically presupposes classical probability theory which implies that aggregation is built upon expectation. There are now multiple reasons to motivate looking at richer alternatives to classical probability theory as a mathematical foundation for machine learning. We systematically examine a powerful and rich class of alternative aggregation functionals, known variously as spectral risk measures, Choquet integrals or Lorentz norms. We present a range of characterization results, and demonstrate what makes this spectral family so special. In doing so we arrive at a natural stratification of all coherent risk measures in terms of the upper probabilities that they induce by exploiting results from the theory of rearrangement invariant Banach spaces. We empirically demonstrate how this new approach to uncertainty helps tackling practical machine learning problems.

2.1 INTRODUCTION

Machine learning (ML) typically presupposes classical probability theory. Recently, the assumption of a single stable probability distribution has been problematized, however. Our motivation stems from the following ML problems: in many cases, the empirical distribution of the data is not the true one, so that some degree of distrust is warranted. Under *data set shift*, for instance, the learning method fails to generalize due to different distribution of the test data in the wild as compared to the well-controlled training environment. Furthermore, as machine learning is being increasingly deployed in sensitive domains (e.g. medical problems, robot control), where failure can be catastrophic, demand for *risk-averse* learning methods has arisen. This problem is often framed as aiming for *distributional robustness* [Rahimian and Mehrotra, 2019], where the goal is to perform well with respect to perturbations of the reference distribution (the empirical distribution). A *prima facie* different problem is that of fair machine learning. In this setting, a machine learning system has direct bearing on ethically relevant individuals and we may hence ask for a system that does not discriminate between specified subgroups, consisting of ethically fungible individuals (e.g. based on race, gender). This ethical ML problem of fairly distributing loss values is analogous to the technical ones discussed before and can be described in the same mathematical formalism.¹

The commonality that we identify among these problems is that they require rethinking the presumption that probability is merely about risk, but instead to realize a distinction between risk and (Knightian) uncertainty or ambiguity. Our method of inquiry is to take inspiration from other fields, where similar problems have received much more attention and treatment already. We find that there are numerous convergent strains of research scattered across the literature, with only a subset of their intricate interconnections laid out clearly so far. In particular, we consider ideas from imprecise probability, rational and social choice theory, finance, insurance, distributive justice and the theory of rearrangement invariant Banach spaces. Our main workhorse is the equivalence between coherent risk measures [Artzner et al., 1999] from finance and coherent upper previsions [Walley, 1991], an influential approach to imprecise probability. These functionals can replace the expectation operator in expected risk minimization of a machine learning problem.

The full generality of coherent risk measures is attractive, but we find that zooming in on a particular subclass, the *spectral risk measures*, provides benefits such as clear interpretability. This subclass is particularly relevant, occupies a central place in the theory of coherent risk measures, and has been rediscovered numerous times by different authors with different motivations [Yaari, 1987, Schmeidler, 1989, Wang, 2000, Acerbi, 2002, Quiggin, 2012,

¹This is one of multiple ways of mathematizing fairness, and requires formulating the loss function in a way that expresses the fairness-relevant aspects. See [Williamson and Menon, 2019].

[Buchak, 2013](#)]. We explicate this relevance from various angles. Essentially, spectral risk measures offer a systematic way to interpolate in the risk aversion spectrum.

We place the class of coherent risk measures in the broader framework of rearrangement invariant Banach function spaces [[Bennett and Sharpley, 1988](#)] and find that there, as well, the spectral risk measures occupy a prime position. This new connection also enables us to rederive the well-known Kusuoka representation theorem from a different angle, which states that any coherent risk measure has a representation in terms of spectral risk measures, thereby further underlining their centrality. Moreover, we leverage the theory to derive various characterization results of coherent risk measures in terms of their *fundamental function*. Such a function specifies the underlying imprecise probability associated with a risk measure. To each such function, we characterize the most optimistic and pessimistic extension from the imprecise probability to a risk measure. We explicate that the most significant distinctions of risk measures stem from their behaviour for tail events (an idea on which we have elaborated in a subsequent paper, see [[Fröhlich and Williamson, 2023](#)]).

Finally, we apply coherent and spectral risk measures to practical machine learning problems, and find they lead to more robust and risk-averse solutions. We begin by outlining the risk and uncertainty distinction, which is the conceptual motivation for the following mathematical development.

2.1.1 RISK AND UNCERTAINTY

By uncertain knowledge, let me explain, I do not mean merely to distinguish what is known for certain from what is only probable. The game of roulette is not subject, in this sense, to uncertainty; nor is the prospect of a Victory bond being drawn. Or, again, the expectation of life is only slightly uncertain. Even the weather is only moderately uncertain. The sense in which I am using the term is that in which the prospect of a European war is uncertain, or the price of copper and the rate of interest twenty years hence, or the obsolescence of a new invention, or the position of private wealth-owners in the social system in 1970. About these matters there is no scientific basis on which to form any calculable probability whatever. We simply do not know.

— John Maynard [Keynes \[1937\]](#)

A distinction between risk and uncertainty has been around since Frank Knight’s seminal work *Risk, Uncertainty and Profit* [[Knight, 1921](#)], with precursors going back even to Adam Smith [[1776](#)]. While it has received considerable attention in the economics literature, it has not yet been firmly established in the machine learning community. ‘Risk’ refers to the benign situation, in which probabilities can be meaningfully associated to outcomes

and full knowledge of the distribution is accessible; contrariwise, ‘uncertainty’ refers to outcomes to which probabilities cannot be assigned. The meaning of cannot here is subtle and warrants further discussion. Classical probability theory, based on Kolmogorov’s widely accepted axioms [Kolmogorov, 1950], is well-equipped to deal with the former, but is arguably not an appropriate model for the latter. Along similar lines, Phil Dawid [2017] wrote

If you studied any Probability at school, it will have focused on the behaviour of unbiased coins, well-shuffled packs of cards, perfectly balanced roulette wheels, *etc., etc.* — in short, an excellent training for a life misspent in the Casino. This is the ambit of *Classical Probability*[.] [emphasis in original].

From a frequentist perspective, the crucial (problematic) assumption is that stochastic phenomena display stable relative frequencies in the limit. While often seemingly correct, this does not occur universally [Gorban, 2017]. Frequentist probability is often given a metaphysical interpretation, by imagining an experiment which could be repeated infinitely many times to obtain independent outcomes [Dawid, 2017]. This is unlike the practical setting, where a ML system is deployed in a dynamically unfolding environment. The failure to comply with a single stable probability distribution is then typically theorized using the notion of *data set shift* [Quiñonero-Candela et al., 2008].

On the other hand, Bayesians assert that it is possible to supply a precise probability for any event or sequence of events. Such a precise credence (degree of belief) is then interpreted as your personal fair betting rate on an event [de Finetti, 1974/2017]. However, it is unclear whether you should have a precise betting rate on the event that right now *24 men in Bulgaria are standing on their heads* [Schoenfield, 2012], as there is no evidence on which you could reasonably base your precise belief. Giving up on the insistence that you have a single betting rate, and instead positing that you have lower and upper betting rates, depending on whether you are required to bet for or against the event, yields imprecise probability [Walley, 1991], which we discuss in the next section.

A now classical challenge to probability theory is due to Daniel Ellsberg [1961]. Consider two urns, containing red and black balls. In urn I, there are 50 red and 50 black balls. Urn II contains an unknown proportion of red and black balls, adding up to 100 balls in total. On these four events (IR, IB, IIR, IIB), the subject may place a bet, which delivers \$100 if the ball is of the specified color and \$0 otherwise. Most subjects display the preference $IR \sim IB \succ IIR \sim IIB$, where $\sim$ denotes indifference and $\succ$ preference, so they prefer to bet on the first urn. We might call urn I the risk urn, as probabilities can be precisely assigned as proportions of outcomes². Urn II is an ambiguous urn, as the subject must entertain a

²This holds when adopting the “principal principle” of Lewis [1980], which asserts that knowledge of chances requires that these be taken as subjective probabilities.

whole set of possible urn compositions. The typical preference cannot be reconciled with probability theory and hence expected utility theory (in economic terms). If the subject is indifferent between a bet on red or on black for the ambiguous urn, it means in effect that she assigns the probability 0.5 to each color; but then she cannot strictly prefer betting on the first urn, where the probability is also 0.5. [Ellsberg \[1961\]](#) calls decision makers, which exhibit this paradoxical pattern, *ambiguity-averse*. Here, ambiguity is to be understood as in-between risk and total Knightian uncertainty. After all, the subject supposes that the urn will exhibit stable relative frequencies, as opposed to an unstable real-world process (e.g. a machine learning system in a changing environment).

Ellsbergs urn paradoxes can be taken as purely descriptive, but are often interpreted and defended from a normative perspective: it is rationally permissible for subjects to be ambiguity-averse [[Stefánsson and Bradley, 2019](#)]. An education in classical probability may lead individuals to revise their initial preferences, after the inconsistency has been pointed out, in order to conform with probabilistic reasoning. The challenge for such a response, however, is to give a non-circular justification for classical probability in the first place as the only permissible rational decision theory. An appeal to probability itself in such an argument is pointless. Contrariwise, we take Ellsbergs urns to be a serious challenge with normative appeal. Besides this thought experiment, a wealth of other challenges have been raised against classical probability theory [[Allais, 1953](#), [Walley, 1991](#), [Joyce, 2005](#), [Gilboa et al., 2009](#), [Bradley, 2019](#), [Isaacs et al., 2021](#)]. We do not attempt to summarize the vast literature on this topic.

While the above example may at first sight seem irrelevant to the practical concerns of a machine learning engineer, the challenge of ambiguity has in fact been recognized in the framework of distributionally robust ML [[Rahimian and Mehrotra, 2019](#)]. The typical expected risk minimization problem

$$\operatorname{argmin}_f \mathbb{E}_P \ell(f(X), Y) \quad (2.1)$$

is there replaced by a worst-case attitude with respect to an ambiguity set of probability measures

$$\operatorname{argmin}_f \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \ell(f(X), Y), \quad (2.2)$$

where the ambiguity set $\{Q : d(Q, P) < \epsilon\}$ typically contains all probability measures in a specified ϵ -neighbourhood of the base measure with respect to some divergence measure d (e.g. an f -divergence). One rationale for employing a distributionally robust (DR) approach is to account for the issue that the empirical distribution $\hat{P}_n$ is not the ‘true’ one for finite sample size n . Instead, a whole set of probability distributions is considered and the most pessimistic, ambiguity-averse attitude is adopted by taking the supremum over

the expected risks. We call the essence of this situation *hallucinated ambiguity*: while the decision maker, i.e. the machine learning engineer, is faced with a decision problem under risk, she also has good reason to believe that $\hat{P}_n$ does not coincide exactly with the ‘true’ distribution. Therefore, she decides to introduce artificial ambiguity into the problem. In contrast, in an Ellsberg-like decision problem ambiguity arises naturally.

Distributionally robust optimization has proven to be useful for a range of machine learning problems. For instance, it has been used to counteract the possibility of adversarial attacks [Sinha et al., 2017]. Due to its breadth, the DR framework can also tackle the problem of data set shift [Zhang et al., 2021b], where the training and test distributions differ and which potentially yields diminished generalization performance. This is especially relevant since data set shift has been recognized as one of the most pressing problems in AI safety [Amodei et al., 2016]. For example, Kirschner et al. [2020] proposed a distributionally robust Bayesian optimization to deal with this phenomenon.

The line between ambiguity and risk can be blurry. If full access to the underlying distribution is available, the decision problem is one under risk. However, decision makers may have different rationally permissible attitudes towards such stochastic risk [Buchak, 2013]. When using expected risk minimization, the decision maker takes a neutral stance towards risk and cares merely about the average. Another decision maker might emphasize downside risk, in financial terms. These are losses which exceed the expected loss. This raises the question of how to systematically encode such an attitude. As machine learning is being increasingly deployed in sensitive domains, demand for *risk-averse* learning methods has arisen. In such domains, tail risks, i.e. unlikely events with highly negative impact, pose a threat to the system or even lead to human death. In reinforcement learning, risk-averse methods have been put forward e.g. by Singh et al. [2020], Urpí et al. [2021], Dabney et al. [2018], Tamar et al. [2015], Vijayan et al. [2021]. To this end, these authors have employed coherent risk measures, which we study in this paper. This effectively amounts to a transformation of *risk* to *hallucinated ambiguity*, as we will show. To a first approximation, the mathematical approach of coherent risk measures to handle risk in fact coincides with the mathematics to handle ambiguity with imprecise probabilities.

In this paper, we consider distributional robustness in the general frameworks of risk measures and imprecise probability. This conceptual unification provides novel justification and interpretation and can guide further developments.

In summary, we take there to be a broad epistemic spectrum, where certainty, risk, (hallucinated) ambiguity and Knightian uncertainty lie. In this order, the adequateness of classical probability theory is increasingly challenged. We will argue that spectral risk measures are a distinguished subclass of coherent risk measures because they enable an interpolation between the two ends in a sensible manner.

2.1.2 CONTRIBUTIONS

We elaborate the connection of coherent risk measures and imprecise probability, which has so far received little attention. Thereby, we clarify the relation of risk (aversion) and ambiguity and what bearing this has on machine learning. In particular, we focus on the subclass of spectral risk measures. Our goal is to bring as many different characterizations of them as possible together in one place. On the way, we discover multiple new connections between theories.

We first summarize the existing theories of imprecise probability (Section 2.2) and coherent risk measures (Section 2.3). We embed the theory of coherent risk measures in the broader mathematical framework of rearrangement invariant Banach function spaces, thereby establishing an insightful connection (Section 2.4). This enables the direct import of mathematical results, which are not yet known in the theory of coherent risk measures. Also, we can easily rederive the celebrated Kusuoka representation theorem [Kusuoka, 2001] in this setting and provide an intuitive interpretation of it. From this perspective, we derive various novel characterization results of coherent risk measures, typically in terms of their fundamental function, which corresponds to an imprecise probability. We present some new results regarding the combination of two risk measures, showing relationships to the theory of interpolation of operators (Section 2.5). In particular, we illustrate that one cannot avoid the element of choice in risk measure by appealing to an “objective” combination rule, because the set of legitimate combination rules ends up being essentially as rich as the set of risk measures. We conduct experiments, which demonstrate that spectral risk measures can encode risk aversion and robustness. For our experiments (Section 2.6), we suggest two ways to evaluate the tail risk of a loss distribution. Specifically, we propose a graphical evaluation based on the *conditional value at risk* and we employ Lorenz curves from the study of economic inequality. Throughout the paper, we translate results from different fields to a loss-based formulation, which aids the unification. As a consequence, when checking references, results might appear different from our statement of them.

2.2 COHERENT LOWER AND UPPER PREVISIONS

The umbrella term *imprecise probability* was popularized by Walley [1991], who offered a behavioural account of rational belief, which strictly generalizes probability theory. Walley takes inspiration from the work of the Bayesian de Finetti [1974/2017], who identified probability with personal fair betting rates. In contrast to de Finetti, however, Walley departs from the *dogma of precision* and allows for a divergence of lower and upper betting rate. In this section we outline the basics of Walley’s approach, with its main pillars of *avoiding sure loss*, *coherence* and *natural extension*. While Walley’s theory is formulated in

terms of reward, we use a loss-based formulation throughout the paper, so that different theories can be directly related without tedious translations.

2.2.1 GAMBLES AND PREVISIONS

Consider a possibility space Ω , where $\omega \in \Omega$ represents a state of the world, including all information deemed relevant to the problem at hand. A *gamble* is a bounded function $X : \Omega \rightarrow \mathbb{R}$, yielding an uncertain loss $X(\omega)$ when the state ω is realized. In the ML context, such a gamble corresponds to a bounded loss function; here, we will allow negative loss values, too, which are then interpreted as reward. Gambles carry the obvious vector space structure with scalar multiplication $(\lambda X)(\omega) = \lambda X(\omega)$, $\lambda \in \mathbb{R}$, and addition $(X + Y)(\omega) = X(\omega) + Y(\omega)$. Constant gambles $\alpha(\omega) = \alpha \forall \omega$ are set in lowercase. We assume that a vector space $\mathcal{L}$ of gambles is given.

With simple axioms, we can characterize the set $\mathcal{D}$ of gambles which are desirable to the decision maker (i.e. the ML engineer). A critical assumption is that loss lives on a bipolar linear measurement scale, where 0 separates loss (> 0 , bad) from reward (< 0 , good). Then we can postulate the following structure:

$$\mathbf{D1.} \sup X < 0 \Rightarrow X \in \mathcal{D}$$

$$\mathbf{D2.} \inf X > 0 \Rightarrow X \notin \mathcal{D}$$

$$\mathbf{D3.} X \in \mathcal{D}, \lambda \in \mathbb{R}^+ \Rightarrow \lambda X \in \mathcal{D}$$

$$\mathbf{D4.} X \in \mathcal{D}, Y \in \mathcal{D} \Rightarrow X + Y \in \mathcal{D}$$

We may take these as axioms, but they are explained through the choice of a linear utility scale. As to D1, certainly a gamble which yields only rewards is desirable. Conversely, a gamble which yields only loss is not desirable (D2). Axioms D3 and D4 imply that the set $\mathcal{D}$ forms a convex cone, which due to D1 includes the interior of the negative orthant $\mathcal{L}^-$, and due to D2 excludes the interior of the positive orthant $\mathcal{L}^+$. We call a set $\mathcal{D}$ satisfying D1–D4 a *coherent set of desirable gambles*.

As such, this framework does not yet provide us with an evaluation of gambles which contain a mix of positive and negative outcomes. For this, we define a functional, called *upper prevision* as follows:

$$\bar{P}(X) := \inf\{\alpha \in \mathbb{R} : X - \alpha \in \mathcal{D}\}. \quad (2.3)$$

We interpret $\bar{P}(X)$ as specifying the smallest amount of certain loss α that, when subtracted from the uncertain loss X , makes the resulting gamble desirable. In financial terms, this is

the *certainty equivalent* for X : the decision maker is willing to shoulder the risky position X when offered the reward $-\alpha$ in exchange. Symetrically, we can define a *lower prevision*:

$$\underline{P}(X) := -\overline{P}(-X) \quad (2.4)$$

$$= -\inf\{\alpha \in \mathbb{R} : -X - \alpha \in \mathcal{D}\} \quad (2.5)$$

$$= \sup\{\alpha \in \mathbb{R} : \alpha - X \in \mathcal{D}\}, \quad (2.6)$$

which specifies the largest certain loss α we are willing to shoulder in exchange for giving away the uncertain X . In virtue of their conjugacy relation, we focus on the upper prevision in the following. When an upper prevision is defined from a coherent set of desirable gambles as in Equation 2.3, it can be shown to satisfy the properties [Walley, 1991, p. 65]:

P1. $\overline{P}(X) \leq \sup(X)$ (bounds)

P2. $\overline{P}(\lambda X) = \lambda \overline{P}(X), \forall \lambda \in \mathbb{R}^+$ (positive homogeneity)

P3. $\overline{P}(X + Y) \leq \overline{P}(X) + \overline{P}(Y)$ (subadditivity)

We call a functional satisfying P1-P3 a *coherent upper prevision*. The corresponding coherent set of desirable gambles can be defined as $\mathcal{D} := \{X : \overline{P}(X) \leq 0\}$, a definition which interacts well with Equation 2.3. P2 and P3 together imply

P4. $\overline{P}(\alpha X + (1 - \alpha)Y) \leq \alpha \overline{P}(X) + (1 - \alpha) \overline{P}(Y) \forall \alpha \in [0, 1]$ (CX: convexity)

but the converse is not true. In virtue of P2 and P3, $\overline{P}$ is a sublinear function and hence the support function of a closed convex set, a geometric fact which we will exploit later. Furthermore, P1-P3 imply [Walley, 1991, p. 76]

P5. $\overline{P}(c) = c, \forall c \in \mathbb{R}$ (agreement)

P6. $\overline{P}(X + c) = \overline{P}(X) + c, \forall c \in \mathbb{R}$ (translation equivariance)

P7. $X(\omega) \leq Y(\omega) \forall \omega \in \Omega \Rightarrow \overline{P}(X) \leq \overline{P}(Y)$ (monotonicity)

An upper prevision generalizes the classical notion of the linear expectation $\mathbb{E}(X)$. In Walley's setting, a linear prevision is defined as a prevision which satisfies the self-conjugacy relation $\overline{P}(X) = -\underline{P}(-X)$. For a coherent prevision, it holds that $\underline{P}(X) \leq \overline{P}(X)$ and we may call the width of the interval $[\underline{P}(X), \overline{P}(X)]$ the *degree of imprecision*. As a first simple example of a nonlinear upper prevision, consider the *vacuous prevision* $\overline{P}(X) = \sup(X)$ and correspondingly $\underline{P}(X) = \inf(X)$. This prevision maximizes the degree of imprecision,

while still being coherent. It is a model for complete ignorance, unlike a uniform distribution, which actually expresses precise beliefs [Konek, 2015]. On the other hand, the familiar expectation $\mathbb{E}$ is a precise, linear prevision. However, defining an expectation requires much more structure (a σ -algebra and a probability measure) than Walley imposes.

To understand the structural implications of coherence, we first consider a strictly weaker rationality condition: *avoiding sure loss*. In the Bayesian tradition, a typical justification for probability theory is based on *Dutch book* arguments. A Dutch book is a collection of gambles, each of which is desirable to the decision maker, but the combination of which surely incurs a loss for the decision maker, no matter the outcome ω . If it is not possible to find such a finite combination, we say that the prevision avoids sure loss.

Definition 29. A functional $\bar{P}$ defined on $\mathcal{L}$ avoids sure loss if

$$\forall n \in \mathbb{N} : \forall X_1, \dots, X_n \in \mathcal{L} : \sup_{\omega \in \Omega} \left[\sum_{j=1}^n \bar{P}(X_j) - X_j(\omega) \right] \geq 0. \quad (2.7)$$

Consider what happens if Equation 2.7 fails. Then $\forall \omega \in \Omega : \sum_{j=1}^n \bar{P}(X_j) < \sum_{j=1}^n X_j(\omega)$. This means that our risk assessments $\bar{P}(X_j)$ were too small, whatever the outcome. In the next section, we observe that the concept of avoiding sure loss has an approximate correspondence in the theory of risk measures as *aversity*. A coherent upper prevision always avoids sure loss and is hence immune to Dutch books, but the converse is not generally true. In geometric terms, the above condition is equivalent to requiring that the set of desirable gambles excludes the interior of the positive orthant $\mathcal{L}^+$, where sure loss would occur.

It can be shown [Walley, 1991, p. 134] that any upper prevision which avoids sure loss dominates at least one linear prevision pointwise, so the set $\mathcal{Q} = \{Q : Q(X) \leq \bar{P}(X) \forall X \in \mathcal{L}, Q \text{ is linear prevision}\}$ is non-empty. We call such a set $\mathcal{Q}$ an *envelope*.³ Then we can construct a canonical coherent extension of $\bar{P}$ by forming the supremum over this set

$$\bar{E}(X) = \sup_{Q \in \mathcal{Q}} Q(X). \quad (2.8)$$

This process is the *natural extension* of $\bar{P}$ and yields a coherent upper prevision if and only if $\bar{P}$ avoids sure loss. If $\bar{P}$ was already coherent, then $\bar{E} = \bar{P}$. On the other hand, if $\bar{P}$ merely avoided sure loss, then the natural extension is the least committal extension from a behavioural perspective. This means that for any other coherent upper prevision $\bar{P}'$ which is dominated by $\bar{P}$, meaning $\bar{P}'(X) \leq \bar{P}(X) \forall X \in \mathcal{L}$, the natural extension lies in-between:

³In the literature on imprecise probabilities, the functional $\bar{E}$ in Equation 2.8 is called the envelope of $\mathcal{Q}$, whereas we use the term envelope for the set $\mathcal{Q}$ itself, in line with works such as [Rockafellar and Royset, 2015].

$\bar{P}'(X) \leq \bar{E}(X) \leq \bar{P}(X) \forall X \in \mathcal{L}$. In this sense, the natural extension is the most pessimistic one, as it reduces the risk assessment by $\bar{P}$ just as little as necessary to achieve coherence.

Conversely, every coherent upper prevision can be written in the form of Equation 2.8 for some set $\mathcal{Q}$. Also, any representation of this form is automatically coherent. For the lower prevision, the infimum is taken over the same set. This provides a direct link to the ambiguity sets in DR optimization: any ambiguity set of linear previsions yields a coherent upper prevision. Note that until now, measure theory has not entered the picture, as Walley's theory is more general. Later we will identify linear previsions with $\mathbb{E}_{\mu_Q}[X]$ for some probability measure μ_Q .

In the imprecise probability literature, the envelope $\mathcal{Q}$ is called a *credal set*. Figuratively, each linear prevision in the set corresponds to a member of a 'credal committee' [Joyce, 2010]. Whereas each member holds a precise belief (credence) on the risk of X , their joint decision is based on a worst-case consideration and hence introduces imprecision. The question of desirability of a gamble consists in a unanimous vote of all credal members. By construction, Walley's theory thus encodes a maximally pessimistic attitude with respect to some envelope.

2.2.2 LOWER AND UPPER PROBABILITIES

So far we have focused on upper previsions, i.e. nonlinear expectations, instead of probability. To obtain an imprecise probability on events, the prevision is applied on indicator gambles

$$A \subseteq \Omega : \chi_A(\omega) := \begin{cases} 1 & \omega \in A \\ 0 & \text{otherwise} \end{cases} \quad (2.9)$$

so that $\bar{P}(A) := \bar{P}(\chi_A)$ is an *upper probability* and $\underline{P}(A) := 1 - \bar{P}(\chi_{A^c})$ a *lower probability*, where A^c is the complement of A , i.e. $\Omega \setminus A$. These probabilities can be interpreted as a personal upper and lower betting rate, respectively, on the event that A occurs. To verify coherence, the same criteria as for previsions may be used, but where the gambles are restricted to be indicator gambles. In the following, we assume $\bar{P}$ to be defined on a field⁴ of events. Some consequences of coherence are then [Walley, 1991, p. 84]:

Pa) $0 \leq \underline{P}(A) \leq \bar{P}(A) \leq 1$

Pb) $\underline{P}(\emptyset) = \bar{P}(\emptyset) = 0; \quad \underline{P}(\Omega) = \bar{P}(\Omega) = 1$

Pc) $A \subseteq B \Rightarrow (\underline{P}(A) \leq \underline{P}(B) \text{ and } \bar{P}(A) \leq \bar{P}(B)).$

⁴A field $(\Omega, \mathcal{F})$ consists of a set Ω and a family of subsets $\mathcal{F}$, which is closed under complements, finite unions and finite intersections. This is weaker than the definition of a σ -algebra, where closure under countable unions and intersections is assumed.

Like for previsions, the width of the interval $[\underline{P}(A), \bar{P}(A)]$ is a natural measure for the degree of imprecision. An interesting interpretation for this comes from a comparison to modal logic [Augustin et al., 2014], where the possibility operator $\Diamond$ and the necessity operator $\Box$ stand in a similar conjugacy relation $\Diamond p = \neg \Box \neg p$ and likewise $\Box p = \neg \Diamond \neg p$. When the event A is seen as a proposition, which represents incurring a unit loss, the lower probability quantifies the evidence that is certainly in favor of A and likewise, the upper probability captures the evidence possibly in favor of A . Just as probability theory can be seen as an extension of propositional logic, imprecise probability theory extends modal logic. Similarly, a lower prevision gives the most optimistic (certain) assessment of the risk, whereas the upper prevision gives a more pessimistic (possible) assessment.

In classical probability theory, there is a one-to-one correspondence between probability measures and the expectations they induce via Lebesgue integration. However, coherent upper probabilities in general do not uniquely determine a coherent upper prevision, which is why Walley focuses on previsions. The subclass of spectral risk measures we are particularly interested in, however, is based on upper probabilities in a one-to-one correspondence, which are then naturally extended to an upper prevision.

Similar to previsions, upper probabilities are characterized by the set of additive probabilities which they dominate. Additive probabilities are those which satisfy Kolmogorov's axioms, but with σ -additivity weakened to finite additivity:

$$\mathbf{K1)} \quad P(A) \geq 0$$

$$\mathbf{K2)} \quad P(\Omega) = 1$$

$$\mathbf{K3)} \quad P(A \cup B) = P(A) + P(B), \text{ if } A \cap B = \emptyset.$$

An upper probability avoids sure loss if and only if it dominates an additive probability. It is furthermore coherent if and only if it is the envelope (cf. Equation 2.8) of a set of additive probabilities. Hence, to extend an upper probability to an upper prevision, we may extend the additive probabilities in the envelope to linear previsions. This process, which may be complicated in general, is simplified for *submodular* upper probabilities, which induce the class of spectral risk measures (Section 2.3.6)⁵. Due to their computationally convenient properties, submodular upper probabilities have received much attention in the imprecise probability literature [see e.g. Montes et al., 2018]. They are also called *2-alternating* [Miranda et al., 2003] and the corresponding lower probabilities are *2-monotone*. In the next section, we discuss coherent risk measures and relate the subclass of spectral risk measures to their corresponding submodular upper probabilities.

⁵Technically, this is true if and only if the submodular probabilities are given as the composition of a concave function and a σ -additive probability.

2.3 COHERENT RISK MEASURES

The study of risk measures in financial mathematics aims to establish a systematic approach to the quantification of risk inherent in a portfolio. Such a portfolio, a collection of assets, will yield an uncertain future monetary loss or gain $X(\omega)$ when the state $\omega \in \Omega$ is realized. In this setting, risk is inherently asymmetrical: financial institutions are much more concerned with their downside risk, that is, returns below the expected value. Unexpectedly high gain is not a similar matter of concern. It is customary to view X as a real-valued random variable, that is, a measurable function, on some underlying probability space $(\Omega, \mathcal{F}, P)$. From the viewpoint of a regulating agency, the risk of this uncertain return must be quantified in order to arrive at a sensible capital requirement to prevent insolvency. Shouldering excessive risk without an appropriate capital requirement puts customers and the economy at risk. The failure to quantify risk properly has indeed been linked to the financial crisis, as discussed in *The Turner Review* [Financial Services Authority, 2009]. Hence there has been increasing interest in risk measures which satisfy certain desiderata.

Artzner et al. [1999] initiated the study of *coherent risk measures*. They imposed axioms on *acceptance sets*, which contain acceptable positions — in Walley’s terms, desirable gambles. The structure they imposed led to the corresponding risk functional having the following properties⁶:

- C1.** $R(\lambda X) = \lambda R(X), \forall \lambda \in \mathbb{R}^+$ (positive homogeneity)
- C2.** $R(X + Y) \leq R(X) + R(Y)$ (subadditivity)
- C3.** $R(X + c) = R(X) + c, \forall c \in \mathbb{R}$ (translation equivariance)
- C4.** $X(\omega) \leq Y(\omega) \forall \omega \Rightarrow R(X) \leq R(Y)$ (monotonicity)

Artzner et al. [1999] justified these axioms from a financial perspective. Subadditivity is particularly interesting and much hinges on it. The rationale is that diversification should not be penalized. Intuitively, X and Y could act as a hedge against each other, thereby decreasing total risk. We will later discuss how subadditivity is related to ambiguity aversion. Translation equivariance is motivated as *cash invariance*: adding a certain loss to a financial position X should increase risk by exactly the same amount.

It has been observed [Pelessoni and Vici, 2003] that a risk measure corresponds to an upper prevision when random variables are bounded. This can be directly seen from the axioms of acceptance sets, which are equivalent to those for coherent sets of desirable gambles, but it is also instructive to relate the functional properties.

⁶For consistency, we work again with losses corresponding to positive real values, whereas it is common to work with monetary gains in the literature. In insurance, however, working with losses is common as well.

Theorem 30. [Pelessoni and Vicig, 2003]. Let $\mathcal{L}$ be a linear space of bounded real-valued random variables, containing all constants $c \in \mathbb{R}$. A functional R is a coherent risk measure on $\mathcal{L}$ if and only if it is a coherent upper prevision on $\mathcal{L}$.

Proof Let R a coherent risk measure. We need to show only $R(X) \leq \sup(X)$. Since $X \leq \sup(X)$, we have by monotonicity and translation equivariance that $R(X) \leq R(\sup(X)) = R(0 + \sup(X)) = R(0) + \sup(X) = \sup(X)$. Hence R is a coherent upper prevision. For the converse direction, we refer to [Walley, 1991, p. 76] and [Pelessoni and Vicig, 2003]. ■

As of now, the equivalence (barring the technicality of boundedness) of coherent risk measures and coherent upper previsions is a formal, mathematical observation. We assert, however, that it has profound philosophical consequences, which can be understood with regard to the risk and uncertainty spectrum, discussed in Section 2.3.7.

2.3.1 BOUNDEDNESS AND LAW INVARIANCE

Recall that Walley's approach to imprecise probability does not require an underlying probability space, i.e. a measure space, but instead presupposes boundedness of the gambles. From this, he derives, using the Hahn-Banach theorem, that any coherent upper prevision admits a representation of the form

$$\bar{P}(X) = \sup_{Q \in \mathcal{Q}} Q(X), \quad (2.10)$$

where $\mathcal{Q}$ is a set of linear previsions. Yet these linear previsions are merely finitely additive, instead of countably additive. Moreover, boundedness is inconvenient for theory.

On the other hand, coherent risk measures are typically introduced on an underlying probability space. A common choice for the space of random variables is then $\mathcal{L}^2(\Omega, \mathcal{F}, P)$, which are those random variables with finite second moment. However, we will see in Section 2.4 that there is in fact a more natural space to work with. Then any coherent risk measure admits a representation of the form (Section 2.3.2)

$$R(X) = \sup_{\mu_Q: Q \in \mathcal{Q}} \mathbb{E}_{\mu_Q}[X], \quad (2.11)$$

where the μ_Q are *countably additive* probability measures. To ensure that these are indeed valid probability measures, translation equivariance and monotonicity are key (Section 2.3.2). Note that now the definition of monotonicity is adapted to the measure⁷:

$$X \leq Y \text{ } P\text{-a.s.} \Rightarrow R(X) \leq R(Y), \quad (2.12)$$

⁷Artzner et al. [1999] considered only finite Ω , so they could define monotonicity as holding for all $\omega \in \Omega$.

where P -a.s means almost surely (P -almost everywhere). Mathematically, it is more convenient to work with countably additive probability measures and unbounded random variables. On the other hand, with the additional assumption of a single distinguished base measure this setup is less parsimonious than Walley's framework. We believe, however, that little generality is lost when doing so. Henceforth we will work with an underlying probability space in line with the risk measurement and machine learning community. The strength and usefulness of Walley's theory lies in the additional conceptual interpretations that it provides. For instance, that a coherent risk measure essentially relies on an underlying imprecise probability has not been appreciated widely.

A much more restrictive, yet useful assumption is *law invariance*. An upper prevision (coherent risk measure), defined on a probability space $(\Omega, \mathcal{F}, P)$ is called law invariant if $\bar{P}(X) = \bar{P}(X')$ whenever X and X' share the same distribution with respect to P . Conceptually, this introduces reliance on a distinguished precise probability. For example, the expectation $\mathbb{E}$ is a law invariant coherent upper prevision. Law invariance encodes the idea that the fine structure of Ω does not actually matter: a decision maker cares only about the distribution of risk, not in which specific states ω it occurs. The property of law invariance, which is not even expressible in Walley's general framework, will be especially useful to us to characterize classes of coherent risk measures in Section 2.4.

2.3.2 ENVELOPE REPRESENTATIONS

In virtue of positive homogeneity and subadditivity, a risk measure is a *sublinear* functional and hence, assuming closedness⁸ for technical reasons, the support function of a closed convex set. This geometric viewpoint provides direct insights into the structure of risk measures. We here work with the space $\mathcal{L}^p := \mathcal{L}^p(\Omega, \mathcal{F}, P)$ of random variables with finite p -th moment, $p \in [1, \infty]$. It is paired with the space $\mathcal{L}^q$, $\frac{1}{p} + \frac{1}{q} = 1$, and the pairing is

$$\langle X, Y \rangle = \int_{\Omega} X(\omega)Y(\omega) \, dP(\omega), \quad \forall X \in \mathcal{L}^p, Y \in \mathcal{L}^q. \quad (2.13)$$

In the case of $1 \leq p < \infty$, $\mathcal{L}^q$ coincides with the dual space of $\mathcal{L}^p$. The case of $p = \infty$ is complicated [see [Schönherr and Schuricht, 2017](#)], but common practice is to pair it with $\mathcal{L}^1$.

A standard result in convex analysis is that a canonical bijection between support functions R and their supported sets $\mathcal{Q}$ is then given by

$$R(X) = \sup_{Q \in \mathcal{Q}} \langle X, Q \rangle, \quad \mathcal{Q} = \{Q \in \mathcal{L}^q : \langle X, Q \rangle \leq R(X) \, \forall X \in \mathcal{L}^p\}. \quad (2.14)$$

⁸A function R is closed if all sublevel sets $\{x \in \text{dom}(R) : R(x) \leq c\}$, $c \in \mathbb{R}$, are closed sets, i.e. contain all limit points.

The following correspondences are known in the literature [Rockafellar and Uryasev, 2013, Shapiro, 2013, Föllmer and Schied, 2016, Liu, 2019]:

- E1. R is monotone iff $\mathcal{Q} \subseteq \mathcal{L}_+^q$, where $\mathcal{L}_+^q = \{Q \in \mathcal{L}^q : Q \geq 0 \text{ } P\text{-a.s.}\}$
- E2. R is translation equivariant iff $\mathcal{Q} \subseteq \mathcal{E}_1$, where $\mathcal{E}_1 = \{Q \in \mathcal{L}^q : \mathbb{E}(Q) = 1\}$
- E3. $R(X) \geq \mathbb{E}(X) \forall X \in \mathcal{L}^p$ iff $1 \in \mathcal{Q}$, where $1(\omega) = 1 \forall \omega \in \Omega$
- E4. If R is a law invariant coherent risk measure, its envelope $\mathcal{Q}$ is invariant under measure-preserving transformations.

For the technicalities regarding E4 see Shapiro [2013]. Intuitively, E4 means that under law invariance, if Q and Q' have the same distribution under the measure P , then either both are in the envelope or none.

Thus the envelope of a coherent risk measure satisfies $\mathcal{Q} \subseteq \mathcal{L}_+^q \cap \mathcal{E}_1$. Each $Q \in \mathcal{Q}$ defines a measure as

$$\mu_Q(A) := \int_A Q(\omega) \, dP(\omega) = \mathbb{E}_P[\chi_A Q] \quad \forall A \in \mathcal{F}. \quad (2.15)$$

Due to $Q \in \mathcal{L}_+^q$, $\mu_Q(A) \geq 0$ and due to $Q \in \mathcal{E}_1$, we have $\mu_Q(\Omega) = 1$. Hence μ_Q is a probability measure and we can equivalently write the risk measure as

$$R(X) = \sup_{\mu_Q: Q \in \mathcal{Q}} \mathbb{E}_{\mu_Q}[X]. \quad (2.16)$$

If X is bounded, then from this representation it is clear that $R(X) \leq \sup(X)$ and therefore R is a coherent upper prevision. Also, this representation provides the rationale for viewing a risk measure as a worst-case “vote” with respect to a set of probabilities. This is equivalent to Walley’s formulation, where the set consists of linear previsions, which we can identify here as $\mathbb{E}_{\mu_Q}[\cdot]$. Whereas the expectation with respect to the base measure is $\mathbb{E}_P[\cdot]$, represented by the singleton envelope $\{1\}$, each probability measure μ_Q defines a legitimate linear prevision. The envelope, in Walley’s terms, consists of the linear previsions which are dominated by R . Natural extension entails finding those linear previsions and taking the supremum over them, hence automatically enforcing both monotonicity and translation equivariance of the functional. Essentially, this relies on the fact that a closed sublinear function equals the supremum of the linear functions minorizing it [Hiriart-Urruty and Lemaréchal, 2004].

2.3.3 THE FUNDAMENTAL COHERENT RISK QUADRANGLE

Rockafellar and Uryasev [2013] put the developments in the theory of risk measures in an even broader perspective by introducing the *fundamental risk quadrangle*, depicted in

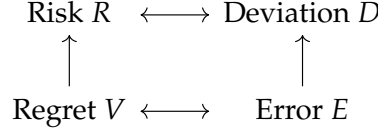


Figure 2.1: The fundamental risk quadrangle [Rockafellar and Uryasev, 2013].

Fig. 2.1. The authors made technical assumptions about certain limits, which were shown to be superfluous by Rockafellar and Royset [2015], and which we therefore drop.

Rockafellar and Uryasev [2013] generally consider *convex risk measures* [Föllmer and Schied, 2016] on $\mathcal{L}^2(\Omega, \mathcal{F}, P)$, where coherence is weakened by dropping subadditivity and positive homogeneity and only assuming convexity in its place. As a consequence, the acceptance set is then a convex set, but not in general a cone anymore. Since we are interested in coherence, we simplify their definitions and theorems to the coherent case. They further demand *aversity*

$$\forall c \in \mathbb{R} : R(c) = c, \text{ but } R(X) > \mathbb{E}[X] \text{ for nonconstant } X, \text{ i.e. } P(\{X = c\}) < 1 \forall c \in \mathbb{R}. \quad (2.17)$$

For reference, we collect properties of averse coherent risk measures in the quadrangle:

- A1.** $R(\lambda X) = \lambda R(X), \forall \lambda \in \mathbb{R}^+$ (positive homogeneity)
- A2.** $R(X + Y) \leq R(X) + R(Y)$ (subadditivity)
- A3.** $R(X + c) = R(X) + c \quad \forall c \in \mathbb{R}$ (translation equivariance)
- A4.** $X \leq Y \text{ } P\text{-a.s.} \Rightarrow R(X) \leq R(Y)$ (monotonicity)
- A5.** $R(c) = c \quad \forall c \in \mathbb{R}$, but $R(X) > \mathbb{E}[X]$ for nonconstant X
- A6.** R is closed, i.e. it has closed sublevel sets $\{X \in \text{dom}(R) : R(X) \leq c\}, c \in \mathbb{R}$. Here, $\text{dom}(R) := \{X : R(X) < \infty\}$.

In the top part of the quadrangle, there is a one-to-one correspondence between coherent risk measures and coherent deviation measures, given by the relation $R(X) = \mathbb{E}(X) + D(X)$. Such deviation measures are positively homogeneous, subadditive and closed and satisfy

$$D(c) = 0 \quad \forall c \in \mathbb{R}, \text{ but } D(X) > 0 \text{ for nonconstant } X \quad (2.18)$$

$$D(X) \leq \text{ess sup}(X) - \mathbb{E}[X], \quad \text{ess sup}(X) := \inf \{\lambda \in \mathbb{R} : P(X > \lambda) = 0\} \quad (2.19)$$

$$D(X + c) = D(X) \quad \forall c \in \mathbb{R} \quad (\text{translation invariance}) \quad (2.20)$$

In practice, the variance is often employed to measure the deviation from the mean in a distribution. In the context of finance, this is the classical *mean-variance analysis* [Markowitz,

1952]. However, the variance is not a coherent deviation measure, as it fails to be sub-additive. Conceptually, the shortcoming is that the variance penalizes variability in both directions, but due to the loss/gain asymmetry, we like to emphasize the importance of losses exceeding the expectation.

In the bottom part of the quadrangle, there is a one-to-one correspondence between coherent regret measures V and coherent error measures E , given by the relationship $V(X) = \mathbb{E}(X) + E(X)$. By coherent regret measure, we mean a functional which is positively homogeneous, subadditive, monotonic, closed and averse in the sense that

$$V(0) = 0, \text{ but } V(X) > \mathbb{E}[X] \text{ for nonzero } X, \text{ i.e. } P(X = 0) < 1. \quad (2.21)$$

According to Rockafellar and Uryasev [2013], “the role of a measure of regret, V , is to quantify the displeasure associated with the mixture of potential positive, zero and negative outcomes of a random variable X that stands for an uncertain cost or loss.”. A coherent regret measure lacks only translation equivariance as compared to a coherent risk measure. A coherent error measure quantifies the nonzeroness of X , is positively homogeneous, sub-additive, closed and averse in the sense that

$$E(0) = 0, \text{ but } E(X) > 0 \text{ for nonzero } X. \quad (2.22)$$

Furthermore, we require

$$E(X) \leq \mathbb{E}[-X] \text{ for } X \leq 0, \quad (2.23)$$

which is equivalent to the monotonicity of the corresponding regret measure. A coherent error measure is hence fundamentally asymmetrical. The bottom part of the quadrangle projects to the top part via the operations

$$R(X) = \inf_{c \in \mathbb{R}} \{V(X - c) + c\}, \quad D(X) = \inf_{c \in \mathbb{R}} \{E(X - c)\}. \quad (2.24)$$

For a coherent regret and error measure, respectively, the result will be a coherent risk and deviation measure. Note, however, that the backwards direction is not unique: one can find an infinity of regret/error measures which project to the same risk/deviation measure.

The projections can be understood as infimal convolution. Let $\sigma_Q(X) := \sup_{Q \in \mathcal{Q}} \langle X, Q \rangle$ be the support function of the set $\mathcal{Q} = \{Q \in \mathcal{L}^q : \langle X, Q \rangle \leq \sigma_Q(X) \forall X \in \mathcal{L}^p\}$.

Theorem 31. [Sun et al., 2020]. Let $V = \sigma_Q$ be a positively homogeneous, subadditive, monotonic and closed functional, i.e. a coherent regret measure, and $\mathcal{Q}$ be its supported set. Suppose $\mathcal{Q}' := \mathcal{Q} \cap \mathcal{E}_1 \neq \emptyset$. Then $R := \sigma_{\mathcal{Q}'}$ is a coherent risk measure and $R(X) = \inf_{c \in \mathbb{R}} V(X - c) + c$.

This process can also be understood from Walley’s perspective, where the projection from

V to R is the natural extension (recall Section 2.2.1):

Theorem 32. *A coherent regret measure V avoids sure loss. Its natural extension coincides with $R(X) = \inf_{c \in \mathbb{R}} V(X - c) + c$.*

Proof Due to aversity of V

$$\forall X_1, \dots, X_n \in \mathcal{L}^p : \sup_{\omega \in \Omega} \left[\sum_{j=1}^n V(X_j) - X_j(\omega) \right] \geq \sup_{\omega \in \Omega} \left[\sum_{j=1}^n \mathbb{E}[X_j] - X_j(\omega) \right] \geq 0, \quad (2.25)$$

and therefore V avoids sure loss. The converse implication (avoiding sure loss $\Rightarrow$ aversity) does not in general hold. Due to monotonicity and subadditivity and positive homogeneity we know that V is the support function of some set $\mathcal{Q}$ and each $Q \in \mathcal{Q}$ is nonnegative almost everywhere. Computing the natural extension entails finding those linear previsions which are dominated by V and forming the envelope of them, but these correspond to expectations $\mathbb{E}_{\mu_Q}[\cdot]$ induced by the set

$$\{Q \in \mathcal{Q} : \mathbb{E}[Q] = 1\} = \mathcal{Q} \cap \mathcal{E}_1, \quad (2.26)$$

which is the supported set of $R(X) = \inf_{c \in \mathbb{R}} V(X - c) + c$. ■

The achievement of [Rockafellar and Uryasev \[2013\]](#) is to put risk in a broad conceptual framework and to establish a link between optimization (R, V) and estimation (D, E) . Consider the archetypical regression problem, framed in terms of a coherent error measure:

$$\text{minimize } E(Y - f(X_1, \dots, X_n)) \text{ over } f \in \mathcal{H} \quad (2.27)$$

for random variables $X_1, \dots, X_n$, outcomes Y and some hypothesis class $\mathcal{H}$. [Rockafellar et al. \[2008\]](#) proved that under a mild technical assumption this problem can be equivalently phrased as

$$\text{minimize } D(Y - f(X_1, \dots, X_n)) \text{ over } f \in \mathcal{H} \text{ s.t. } 0 \in \operatorname{argmin}_{c \in \mathbb{R}} \{E(Y - f(X_1, \dots, X_n) - c)\}. \quad (2.28)$$

See also [\[Rockafellar and Royset, 2015\]](#). This provides a new perspective on regression, where customized risk aversion is directly built in. In this paper, we mainly focus on coherent risk measures rather than coherent error measures, since risk measures applied to a loss random variable are not constrained by a dependence on the $Y - f$ difference. However, our results in Section 2.4 are also linked to coherent regret (and thus error) measures.

2.3.4 THE CONDITIONAL VALUE AT RISK

We now examine a coherent quadrangle of particular interest, that of the *conditional value at risk* CVar_α , with parameter $\alpha \in [0, 1)$. CVar_α is a special case of the larger class of spectral risk measures. In fact, we will see in Section 2.4.7 that the CVar_α are the basic building blocks not only of the spectral risk measures, but of *all* law invariant coherent risk measures. Define the positive part of a random variable as $X^+ := \max(X, 0)$ and the negative part as $X^- := \max(0, -X)$. For each $\alpha \in (0, 1)$, a coherent quadrangle is given by:

$$R(X) = \text{CVar}_\alpha(X), \quad D(X) = \text{CVar}_\alpha(X - \mathbb{E}(X)) \quad (2.29)$$

$$V(X) := \frac{1}{1-\alpha} \mathbb{E}[X^+], \quad E(X) = \mathbb{E} \left[\frac{1}{1-\alpha} X^+ + X^- \right]. \quad (2.30)$$

According to the projection from regret, $\text{CVar}_\alpha(X) = \min_c \{ \frac{1}{1-\alpha} \mathbb{E}((X - c)^+) + c \}$. We also define $\text{CVar}_{\alpha=0} := \mathbb{E}$ in the same way, but this is only “weakly” averse in the degenerate sense that $\mathbb{E} \geq \mathbb{E}$. The random variable X has a right-continuous distribution function F_X with generalized inverse⁹ $F_X^{-1}(q) = \sup\{\lambda \geq 0 : F_X(\lambda) < q\}$. Then CVar_α can be equivalently expressed as an integral over quantiles

$$\text{CVar}_\alpha(X) = \frac{1}{1-\alpha} \int_\alpha^1 F_X^{-1}(q) \, dq. \quad (2.31)$$

If F_X is continuous, this can be further written as

$$\text{CVar}_\alpha(X) = \mathbb{E} \left[X | X \geq F_X^{-1}(\alpha) \right], \quad (2.32)$$

and is also called *expected shortfall*, *tail conditional expectation* or *superquantile* [Laguel et al., 2021]. Intuitively, CVar_α takes the average of the $(1 - \alpha)$ -fraction of the worst outcomes and neglects the more fortunate outcomes completely. In one extreme, $\text{CVar}_{\alpha=0}$ corresponds to the expectation; in the other, $\text{CVar}_{\alpha \rightarrow 1} := \lim_{\alpha \rightarrow 1} \text{CVar}_\alpha$ gives the essential supremum (worst-case) of X .

The envelope of CVar_α is known to be $\mathcal{Q} = \{Q : 0 \leq Q \leq \frac{1}{1-\alpha}, \mathbb{E}[Q] = 1\}$, so coherence can be directly verified from the envelope. We can interpret the elements of the envelope as reweightings of the original distribution, where a reweighting of up to $1/(1 - \alpha)$ is allowed. As a consequence, the supremum is achieved when that reweighting is fully concentrated on the $(1 - \alpha)$ -fraction of the largest losses. For $\alpha = 0$, the supremum is clearly attained at $Q = 1$, which corresponds to the expectation. On the other hand, for $\alpha \rightarrow 1$, the reweighting may be arbitrarily large and hence the worst-case will receive all

⁹For consistency with Section 2.4, we choose to work with the lower instead of the upper quantile.

of the weight (but the supremum will not be attained in general). To see that the above set is indeed the envelope of CVar_α , consider the regret $V = \frac{1}{1-\alpha} \mathbb{E}[X^+]$. It is not hard to see that its envelope is the set $\{Q : 0 \leq Q \leq 1/(1-\alpha)\}$. The projection of V to CVar_α , i.e. the natural extension, entails intersecting this set with the constraint $\mathbb{E}[Q] = 1$.

2.3.5 SPECTRAL RISK MEASURES

CVar_α belongs to the family of spectral risk measures [Acerbi, 2002]. We here work on $\mathcal{L}^2(\Omega, \mathcal{F}, P)$, but in Section 2.4 we show that the natural space to work on is in fact more subtle. Observe that a convex combination of coherent risk measures again yields a coherent risk measure. Given a probability measure λ on $[0, 1]$, this can be generalized to the form

$$R^\lambda(X) := \int_0^1 \text{CVar}_\alpha(X) \, d\lambda(\alpha), \quad (2.33)$$

which yields a coherent risk measure. We assume that the measure λ does not have an atom at 1, i.e. $\lambda(\{1\}) = 0$. By expanding CVar_α as its integral representation and using Fubini-Tonelli, this can be rewritten as

$$R^\lambda(X) = R_{(w)}(X) := \int_0^1 F_X^{-1}(q) w(q) \, dq, \quad (2.34)$$

with a spectral weighting function $w : [0, 1] \rightarrow \mathbb{R}^+$. This generates a coherent risk measure if and only if w is nonnegative, monotonically increasing and $\int_0^1 w(q) \, dq = 1$ [Acerbi, 2002]. These properties are automatically satisfied when w is induced by a probability measure λ from Equation 2.33. The spectrum w has a clear interpretation as a risk aversion profile. A monotonically increasing w puts more weight on worse (highly positive) outcomes as a penalty. In the special case of CVar_α , $\alpha \in [0, 1)$, we have

$$w(q) = \begin{cases} 0 & 0 < q < \alpha \\ 1/(1-\alpha) & q \geq \alpha \end{cases} \quad (2.35)$$

hence all values below the α -th quantile are ignored, and values above it receive the constant weight $1/(1-\alpha) \geq 1$. Note that by demanding that λ does not have an atom at 1, we have excluded the supremum risk measure $\text{CVar}_{\alpha \rightarrow 1}$, which is represented by the Dirac measure at 1. The corresponding weight function would be 0 everywhere, rendering a representation of the form Equation 2.34 impossible. The supremum risk measure, while being in the “closure” of the family of spectral risk measures, cannot be considered a proper member due to its pathological properties. It will lead to additional technical complications in Section 2.4.

For an arbitrary spectral risk measure $R_{(w)}$ with spectrum w , the envelope representation

is [Pflug, 2006]:

$$R_{(w)}(X) = \sup_Q \{ \langle X, Q \rangle : Q = w(U) \}, \text{ where } U \text{ is uniformly distributed on } [0, 1] \text{ wrt. } P, \quad (2.36)$$

which requires that Ω is rich enough to support a uniform distribution. While this result may seem somewhat mysterious, we will obtain a different perspective on it in Section 2.4, which also supplies intuition.

2.3.6 DISTORTION RISK MEASURES

Assume again the space $\mathcal{L}^2(\Omega, \mathcal{F}, P)$. Equivalent to spectral risk measures are *distortion risk measures* with concave distortions, which originate from distortion premium principles in actuarial science [Wang et al., 1997, Wang, 2000]. In insurance, the key challenge is to price a contingent claim. That is, from the viewpoint of the insurer, a random variable X represents an uncertain loss that corresponds to a claim made by a policyholder. Given a probability model under which X has distribution F_X , the question is how much should the insurer charge in exchange for shouldering the risk? This is called the insurance premium. Consider an equivalent definition of the usual expectation:

$$\mathbb{E}[X] = - \int_{-\infty}^0 F_X(x) \, dx + \int_0^{\infty} (1 - F_X(x)) \, dx \quad (2.37)$$

$$= \int_{-\infty}^0 [S_X(x) - 1] \, dx + \int_0^{\infty} S_X(x) \, dx, \quad (2.38)$$

$$(2.39)$$

where we used the *survival function* $S_X(x) := 1 - F_X(x) = P(X > x)$. If the insurer simply charged the expectation as the premium, they could not make any profit and could face bankruptcy due to model misspecification. What if the specified probability does not accurately reflect the real risk? The idea of a distortion premium is to model risk aversion by instead calculating the expectation with respect to another distribution, given by the *Choquet integral*:

$$R_{\phi}(X) := \int_{-\infty}^0 [\phi(S_X(x)) - 1] \, dx + \int_0^{\infty} \phi(S_X(x)) \, dx, \quad (2.40)$$

where $\phi : [0, 1] \rightarrow [0, 1]$ is a monotonically increasing concave function satisfying $\phi(0) = 0$ and $\phi(1) = 1$. In addition, we assume additionally that ϕ is continuous at 0, which excludes the supremum risk measure but avoids technical issues. These boundary conditions ensure that $R(X)$ can be viewed as an expectation with respect to a valid (distorted) probability distribution. Concavity of ϕ models risk aversion in the sense that the higher

the loss level, the higher the increase in the premium. Furthermore, the resulting functional R is a coherent risk measure if and only if ϕ is concave [Gzyl and Mayoral, 2008]. In the special case of $\phi(t) = t$, we obtain the expectation and for all other distortion risk measures we have $R_\phi(X) \geq \mathbb{E}(X)$. The difference $D(X) = R_\phi(X) - \mathbb{E}[X]$, Rockafellar and Uryasev's [2013] deviation measure, is also known as the *risk premium*. For coherence the critical property is subadditivity of R , which corresponds to the concavity of the distortion. If instead ϕ is convex, the functional is superadditive¹⁰. In Appendix 2.9.3, we consider functionals of the form Equation 2.40, Choquet integrals, in more depth. Here we observe that distortion risk measures are equivalent to spectral risk measures.

Theorem 33. [Gzyl and Mayoral, 2008, Ridaoui and Grabisch, 2016]. *For any distortion risk measure R_ϕ with concave distortion ϕ , with $\phi(0) = 0$ and $\phi(1) = 1$, there is an identical spectral risk measure $R_{(w)} = R_\phi$, with $\phi'(t) = w(1 - t)$.*

The proof is in Appendix 2.8.1. For example, in the case of CVar_α we have $\phi(t) = \int_0^t w(1 - u) \, du = \min\{t/(1 - \alpha), 1\}$.

From a given base probability measure P we obtain a distorted probability $\phi(P)$. We can interpret this as an upper probability in Walley's framework. Setting $\bar{\mu}(A) := \phi(P(A)) \, \forall A \in \mathcal{F}$ defines a *capacity* on events. A capacity on $(\Omega, \mathcal{F})$ is a set function $\bar{\mu} : \mathcal{F} \rightarrow \mathbb{R}$ with the normalization $\bar{\mu}(\emptyset) = 0$, $\bar{\mu}(\Omega) = 1$ and the monotonicity property $A \subseteq B \Rightarrow \bar{\mu}(A) \leq \bar{\mu}(B)$. In our case this is satisfied because $\phi(0) = 0$ and $\phi(1) = 1$ and ϕ is monotonically increasing. A *submodular capacity*, sometimes called concave capacity, is a capacity which satisfies the inequality

$$\bar{\mu}(A \cup B) + \bar{\mu}(A \cap B) \leq \bar{\mu}(A) + \bar{\mu}(B) \quad \forall A, B \in \mathcal{F}. \quad (2.41)$$

If and only if the function ϕ is concave and $\phi(0) = 0$, $\bar{\mu}$ is a submodular capacity [Bednarski, 1981, Föllmer and Schied, 2016, Prop. 4.7.5]. Submodularity is not only convenient from a mathematical point of view, but as we will show in Section 2.9.3 it has a rich interpretation in terms of systematic risk aversion. It is known that a monotone submodular capacity is always coherent¹¹, due to its envelope being non-empty:

$$\text{core}(\bar{\mu}) = \{P : P(A) \leq \bar{\mu}(A) \, \forall A \in \mathcal{F}, P \text{ probability measure}\} \quad (2.42)$$

$$\bar{\mu}(A) = \sup_{P \in \text{core}(\bar{\mu})} P(A) \quad (2.43)$$

In the general case of a capacity, the envelope is also called the 'core' and consists of finitely additive probability measures. However, we have defined our capacity on a σ -algebra

¹⁰A functional R is superadditive if $R(X + Y) \geq R(X) + R(Y)$.

¹¹The Choquet integral is convex if and only if the capacity is submodular [Alfonsi, 2015]. Furthermore, it is monotone and translation equivariant. Restricting the Choquet integral to events hence yields a coherent upper probability, which coincides with the submodular capacity on events.

and therefore the core consists of countably additive measures. Employing the core, the Choquet integral Equation 2.40 for a submodular capacity is equivalently given by

$$R_\phi(X) = \sup_{P \in \text{core}(\bar{\mu})} \left\{ \int_{-\infty}^{\infty} X \, dP \right\}. \quad (2.44)$$

Thus, for a submodular capacity, Walley's natural extension coincides with the Choquet integral. This can be understood by viewing Equation 2.44 as forming the linear extensions via integration of the additive probabilities which are dominated by the capacity. This is yet another argument that demonstrates the specialness of distortion/spectral risk measures: they are natural extensions of coherent upper probabilities.

To a given submodular capacity $\bar{\mu}$, we can define a corresponding dual capacity $\underline{\mu}(A) := 1 - \bar{\mu}(A^c) \, \forall A \in \mathcal{F}$. This capacity is supermodular:

$$\underline{\mu}(A \cup B) + \underline{\mu}(A \cap B) \geq \underline{\mu}(A) + \underline{\mu}(B) \quad \forall A, B \in \mathcal{F} \quad (2.45)$$

and is a coherent lower probability. In our case, we can identify it as $\underline{\mu} = \phi \circ P$, with the convex function $\phi(t) = 1 - \phi(1 - t)$. The upper and lower distribution functions are then defined as follows [Walley, 1991, p. 130]:

$$\bar{F}_X(x) := \bar{P}(X \leq x) = 1 - \underline{P}(X > x) \quad (2.46)$$

$$\underline{F}_X(x) := \underline{P}(X \leq x) = 1 - \bar{P}(X > x). \quad (2.47)$$

Hence we can compute upper and lower densities as $\bar{f}_X = \bar{F}'_X$ and $\underline{f}_X = \underline{F}'_X$. This terminology is, however, somewhat unfortunate: the upper distribution function owes its name to the fact that it lies above the lower distribution function, but the upper distribution function is obtained from the lower survival function. Figure 2.2 gives an intuition about the lower and upper probabilities, survival functions and densities for an exemplary distortion. To compute the distortion risk of X , i.e. the upper prevision, one computes the expectation with respect to the upper survival function $\phi(S_X)$ or the lower density $\underline{f}$.

We subsequently use the term *distortion risk measure* to refer to a distortion risk measure with a concave distortion ϕ . Hence we may use the term interchangeably with *spectral risk measure*.

2.3.7 COHERENT MEASURES OF RISK OR UNCERTAINTY?

In the finance context, the theory of coherent risk measures has been advanced as a theory about risk, as the name suggests. There is still a conceptual reliance on a single “true” probability measure and the goal is to embody a risk-averse attitude by specifying a more conservative (pessimistic) summary of a distribution. Due to the envelope representation,

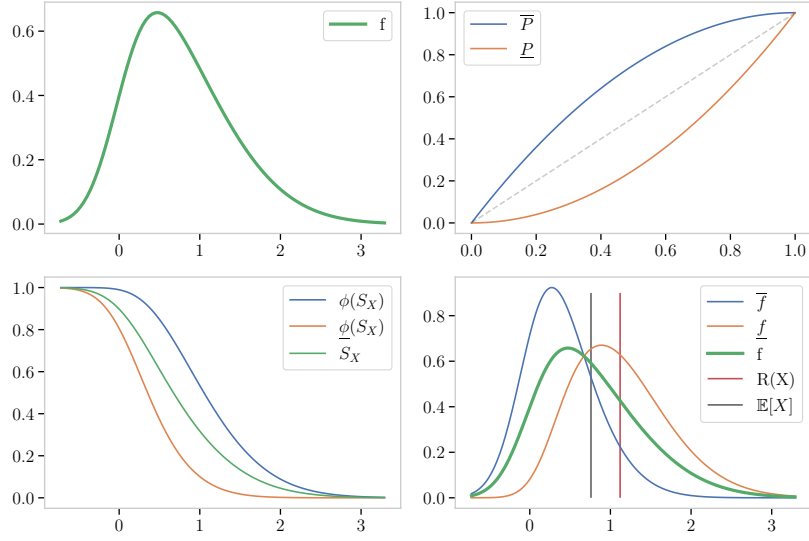


Figure 2.2: Top left: the density of an exemplary skew-normal distribution, belonging to some random variable X . Top right: lower and upper probabilities with distortion function $\phi(t) = 1 - (1 - t)^2$. Bottom left: lower and upper distortion of the survival function, corresponding to the exemplary distribution. Bottom right: lower and upper densities, resulting from the distortion. The vertical lines indicate the expectation and the distortion risk. Note that $R_\phi(X)$ is substantially greater than $\mathbb{E}[X]$.

we can however interpret a risk measure as taking the worst-case decision with respect to a set of probability measures. This amounts to introducing artificial “hallucinated” ambiguity into a decision under risk. Hence a connection to Walley’s theory of imprecise probability is established and the mathematical equivalence is given an interpretation. A key conceptual difference is whether a distinguished base measure can still be identified, as in the case of risk measures, or whether one deals with a credal set consisting of various linear previsions, as in Walley’s case.

A decision maker who uses a law invariant coherent risk measure has an underlying probability measure, but discounts her own belief in it. As an important example of this line of thinking, following the financial crisis, “the Turner Review points to an excessive reliance on a single probabilistic model P derived from past observations” [Föllmer and Weber, 2015]. As a response, coherent risk measures have received increasing attention. Using such a risk measure, a decision maker transforms the risky situation into an ambiguous situation by considering other similar probability measures, as well. If she further employs a spectral risk measure which is based on a distortion of the original distribution, we can conclude that she is coherent if and only if she assigns bigger weights to worse cases [Acerbi, 2002]. Thus she exhibits a systematic risk aversion attitude, which is encoded in the spectrum (or equivalently, the distortion function).

The coincidence of the coherence concept in imprecise probability and the finance literature on risk measures (Theorem 30) is particularly interesting because the axioms are moti-

vated in different fashion. Walley [1991] provides a behavioral justification for coherence, tailored to the situation in which a decision maker finds herself when facing uncertainty. Walley’s [1991] goal is to provide a guide to rational decision making. In finance, the agent is an institution or a regulator. For instance, subadditivity is then motivated as encouraging diversification; translation equivariance, in this context called “cash-invariance”, is motivated by requiring that adding a certain amount of cash (negative loss) should decrease risk by exactly that amount. That the extensions of these two coherence concepts coincides is remarkable and serves as a corroboration of their groundedness.

Coherent risk measures are also intimately connected with generalized utility theories in rational choice theory, situated in the context of economics. These theories offer formal axiomatic bases for rational decision making under uncertainty. In Appendix 2.9, we explore the connections between risk measures and (non)-expected utility theories. In particular, the class of spectral risk measures has been reinvented in this setting as *Choquet expected utility* or, more precisely, as *rank dependent expected utility*.

In machine learning, the “excessive reliance on a single probabilistic model P derived from past observations”, in the words of Föllmer and Weber [2015], is problematized in the context of data set shift and more generally, it is problematic due to having only a finite amount of training data, from which the “true” distribution can only be approximated. By putting *true* in quotes, we wish to emphasize that the assumption of a single underlying probability measure is itself a questionable one, although it has received little attention yet. Data from the real world may exhibit unstable relative frequencies over time [Gorban, 2017] and hence, at least from a frequentist perspective, cannot be based on a single probability distribution (see Fröhlich et al. [2023]). Furthermore, predictions can even influence the outcomes they aim to predict – a phenomenon known as *performative prediction* [Perdomo et al., 2020]. Our goal is to contribute to tackling such problems by demonstrating how coherent risk measures, in particular spectral risk measure, can be helpful as a generalized theory of uncertainty.

2.4 REARRANGEMENT INVARIANT BANACH FUNCTION SPACES

In this section, we show that coherent risk measures are an incarnation of *rearrangement invariant Banach function norms* and are hence embedded in a rich mathematical literature. This connection is, to the best of our knowledge, previously unknown and enables us to obtain novel characterization results. While some authors have studied norms related to risk measures (e.g. Pichler [2013], Mafusalov and Uryasev [2016] and Gotoh and Uryasev [2016]), we here present a broader picture. We follow mainly the technical setup of Bennett and Sharpley [1988]. For a more accessible introduction we refer to [Rubshtein et al., 2016], who use the term “symmetric spaces” instead. Throughout, we work with the probability

space $\Omega = [0, 1]$ with the Lebesgue measure μ , so $\mu(\Omega) = 1$. This space is a standard probability space and all such non-atomic standard spaces are Borel isomorphic [see e.g. [Bäuerle and Müller, 2006](#)]. Hence this introduces no loss of generality for our setting but allows for a cleaner exposition. Let $\mathcal{M}$ denote the class of Lebesgue measurable functions from Ω to $\mathbb{R}$ and $\mathcal{M}^+$ the subset of Lebesgue measurable functions with values in $[0, \infty]$, i.e. nonnegative random variables. (In)equalities between elements of $\mathcal{M}$ are to be understood as holding μ -almost everywhere. Often, we will drop writing $X \in \mathcal{M}$ for brevity, since we have no concern for non-measurable functions throughout.

Definition 34. A functional $R : \mathcal{M}^+ \rightarrow [0, \infty]$ is called *Banach function norm* if the following conditions hold for all $X_n, X \in \mathcal{M}^+$ and measurable $E \subseteq \Omega$:

$$\mathbf{R1.} \quad R(X) = 0 \Leftrightarrow X = 0; \quad R(\lambda X) = \lambda R(X) \quad \forall \lambda \geq 0; \quad R(X + Y) \leq R(X) + R(Y)$$

$$\mathbf{R2.} \quad 0 \leq X \leq Y \Rightarrow R(X) \leq R(Y)$$

$$\mathbf{R3.} \quad 0 \leq X_n \uparrow X \text{ } \mu\text{-a.e.} \Rightarrow R(X_n) \uparrow R(X)$$

$$\mathbf{R4.} \quad R(\chi_E) < \infty; \quad \int_E X \, d\mu < c_E R(X) \text{ for some } 0 < c_E < \infty \text{ depending only on } E \text{ and } R.$$

Since we work with a finite measure space, we also impose $R(\chi_\Omega) = 1$ without loss of generality throughout the paper. Due to positive homogeneity, $R(\chi_\Omega) = c$ would simply correspond to a scaling of the function norm. Observe that a function norm R is defined only on the positive cone of measurable functions. However, it induces a norm on the space $\mathcal{R} = \{X : R(|X|) < \infty\}$ by setting

$$\|X\|_{\mathcal{R}} := R(|X|). \tag{2.48}$$

Then it can be shown that the pair $(\mathcal{R}, \|\cdot\|_{\mathcal{R}})$ forms a *Banach space*, i.e. a complete normed vector space. For completeness of the space, the key axiom is the Fatou property R3. In the context of risk measures, however, it is undesirable to extend a function norm from the positive cone to the whole space by stipulating Equation 2.48. The reason is that we want to treat negative values (gain) as different from positive values (loss). Hence we restrict ourselves to nonnegative random variables $\mathcal{M}^+$ in the following discussion. Then, in virtue of R1 and R2, a coherent risk measure can be viewed as a valid Banach function norm, if it also satisfies the mild technical axioms R3 and R4. A coherent risk measure further satisfies translation equivariance, however. We discuss the subtle role of (non)negativity and translation equivariance in Section 2.4.6 below.

We are specifically interested in *rearrangement invariance* of norms, which corresponds to the law-invariance property of risk measures. The idea is that such a norm only attends to the distribution of a function and hence respects the base measure μ in a suitable way,

thereby disregarding the order in which the values are arranged. To this end, one defines the *distribution function* $\mu_X : \mathbb{R}^+ \rightarrow [0, 1]$ of $X \in \mathcal{M}$ as

$$\mu_X(\lambda) := \mu \{ \omega \in \Omega : |X(\omega)| > \lambda \}. \quad (2.49)$$

For nonnegative random variables, this decreasing (non-increasing) and right-continuous function is just the survival function $S_X = 1 - F_X$. Two functions X and Y are called *equimeasurable* if their distribution functions coincide, i.e. $\mu_X(\lambda) = \mu_Y(\lambda) \forall \lambda \geq 0$.

Definition 35. A Banach function norm R is called *rearrangement invariant* if $R(X) = R(Y)$ for every equimeasurable X, Y . The space $\mathcal{R}$ is then called a *rearrangement invariant Banach space*.

From now on we abbreviate rearrangement invariant as *ri* and call $\mathcal{R}$ an *ri space*. For each $X \in \mathcal{M}$, we obtain a canonical equimeasurable function $X^* : [0, 1] \rightarrow \mathbb{R}^+$ as the generalized inverse of its distribution function:

$$X^*(\omega) := \inf\{\lambda \geq 0 : \mu_X(\lambda) \leq \omega\}, \quad \omega \in [0, 1] \quad (2.50)$$

$$X^*(1) := \lim_{\omega \uparrow 1} X^*(\omega). \quad (2.51)$$

X^* is called the *decreasing rearrangement* of X , as it arranges the (absolute) values of X in decreasing order. It is therefore the continuous analog of sorting a list in descending order. X^* is clearly decreasing and right-continuous. In the context of standard probability theory, this corresponds to the lower “backwards” quantile of $|X|$:

$$X^*(\omega) = \inf\{\lambda \geq 0 : \mu_X(\lambda) \leq \omega\} \quad (2.52)$$

$$= \sup\{\lambda \geq 0 : \mu_X(\lambda) > \omega\} \quad (2.53)$$

$$= \sup\{\lambda \geq 0 : 1 - \mu_X(\lambda) < 1 - \omega\} \quad (2.54)$$

$$= \sup\{\lambda \geq 0 : F_{|X|}(\lambda) < 1 - \omega\} \quad (2.55)$$

$$= F_{|X|}^{-1}(1 - \omega). \quad (2.56)$$

The rationale for working with a decreasing, instead of an increasing rearrangement, is that the ri Banach space theory generally considers spaces of potentially infinite measure; hence a plot of an increasing rearrangement might not show anything interesting until $+\infty$. For an ri function norm R , we have in particular $R(X) = R(X^*)$. A law invariant coherent risk measure induces an ri function norm, which is furthermore translation equivariant.

2.4.1 DUALITY AND THE ASSOCIATE SPACE

Banach spaces have an interesting duality aspect, which we will connect to the envelope representation. The *dual space* $\mathcal{R}^*$ of a Banach space $\mathcal{R}$ consists of all linear, continuous and

bounded functionals $u : \mathcal{R} \rightarrow \mathbb{R}$, and is equipped with the norm [Rubshtein et al., 2016, p. 83]

$$\|u\|_{\mathcal{R}^*} = \sup \{|u(X)| : \|X\|_{\mathcal{R}} \leq 1\} < \infty. \quad (2.57)$$

There exists a close relationship between the dual space and the *associate space* to a function norm R , which is of more practical interest. For an ri norm R , the associate (function) norm is defined by

$$R'(X) := \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega : R(Y) \leq 1, Y \in \mathcal{M}^+ \right\} \quad (2.58)$$

$$\|X\|_{\mathcal{R}'} := \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega : \|Y\|_{\mathcal{R}} \leq 1, Y \in \mathcal{R} \right\}. \quad (2.59)$$

With this pairing, the associate space $\mathcal{R}'$ is canonically isometrically isomorphic to a closed “norm-fundamental” subspace of $\mathcal{R}^*$ [Bennett and Sharpley, 1988, p. 13]. For our purposes, we may ignore the subtle distinction between $\mathcal{R}'$ and $\mathcal{R}^*$. An important aspect of the associate pairing is that Hölder’s inequality holds. If $X \in \mathcal{R}$ and $Y \in \mathcal{R}'$ then

$$\int_{\Omega} |XY| \, d\mu \leq \|X\|_{\mathcal{R}} \|Y\|_{\mathcal{R}'}. \quad (2.60)$$

Also, we have that $\mathcal{R} = (\mathcal{R}')'$ under the assumption of the Fatou property R3.

The prime example of ri spaces are the Lebesgue spaces $\mathcal{L}^p$. A family of function norms is defined as

$$R^p(X) := \begin{cases} \left(\int_0^1 X^p \, d\mu \right)^{\frac{1}{p}} & 1 \leq p < \infty \\ \text{ess sup}(X) & p = \infty, \end{cases} \quad (2.61)$$

where $\text{ess sup}(X) := \inf \{\lambda \geq 0 : \mu_X(\lambda) = 0\}$. We label the space induced by R^p as $\mathcal{L}^p$. The associate space of $\mathcal{L}^p$ is $\mathcal{L}^q$, where $\frac{1}{p} + \frac{1}{q} = 1$. For example, $R^1 = \mathbb{E}[\cdot]$ is paired with its associate $R_{\infty} = \text{ess sup}$. On the other hand, the associate of R^{∞} is R^1 , but the dual space is more subtle and in this case, the canonical embedding of $\mathcal{L}^1$ into $(\mathcal{L}^{\infty})^*$ is strict. For a systematic treatment of this dual space, see [Schönherr and Schuricht, 2017].

2.4.2 THE EMBEDDING THEOREM

Given ri spaces $\mathcal{R}$ and $\mathcal{S}$, where $\mathcal{S} \subseteq \mathcal{R}$, there exists a constant c such that [Bennett and Sharpley, 1988, p. 7]

$$\|X\|_{\mathcal{R}} \leq c \|X\|_{\mathcal{S}} \quad \forall X \in \mathcal{S}. \quad (2.62)$$

In this case, $\mathcal{S}$ continuously embeds into $\mathcal{R}$, which we denote as $\mathcal{S} \hookrightarrow \mathcal{R}$ and refer to a feasible c as embedding constant (not unique). Let $\mathcal{R}$ be any ri space. The following is

known [Bennett and Sharpley, 1988, p. 77, specialized to $\mu(\Omega) = 1$]:

$$\mathcal{L}^\infty \hookrightarrow \mathcal{R} \hookrightarrow \mathcal{L}^1, \quad (2.63)$$

and 1 is a feasible embedding constant:

$$\|X\|_{\mathcal{L}^1} \leq \|X\|_{\mathcal{R}} \quad \forall X \in \mathcal{R}, \quad \|X\|_{\mathcal{R}} \leq \|X\|_{\mathcal{L}^\infty} \quad \forall X \in \mathcal{L}^\infty. \quad (2.64)$$

Therefore, $\mathcal{L}^1$ and $\mathcal{L}^\infty$ are special as they are the extremes of all ri spaces. This implies in particular that any law invariant coherent risk measure “lives between” the expectation and the essential supremum, which stand in an associate relationship. This distinguished status is also visible from their envelope representations (Section 2.3.2): the envelope of the expectation is the singleton $\{1\}$ (a singleton envelope c yields a constant multiple of the expectation¹²), whereas the envelope of the worst-case risk measure consists of all probability measures.

Note also that $\mathbb{E}[\cdot] = \text{CVar}_{\alpha=0}$ and $\text{ess sup} = \text{CVar}_{\alpha=1}$, hence CVar_α in a sense interpolates between the smallest and largest ri function norms. We will later state a more refined embedding theorem, which situates any law invariant coherent risk measure between the spectral risk measure corresponding to its upper probability and the *Marcinkiewicz norm*.

2.4.3 HARDY-LITTLEWOOD’S INEQUALITY

For nonnegative real sequences (x_i) and (y_i) , Hardy-Littlewood’s inequality asserts that

$$\sum_{i=1}^n x_i y_i \leq \sum_{i=1}^n x_i^* y_i^*, \quad (2.65)$$

where (x_i^*) and (y_i^*) are the sequences where the elements of (x_i) , respectively (y_i) , are arranged in decreasing order. This inequality carries over to the continuous case. If X and Y are finite μ -almost everywhere, then [Bennett and Sharpley, 1988, p. 44]:

$$\int_{\Omega} |XY| \, d\mu \leq \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega. \quad (2.66)$$

While this inequality has been employed in the study of Kusuoka representations and envelopes [see e.g. Pichler, 2015], the connection to the general theory of ri spaces has not yet been made.

¹²However, requiring that $R(1_\Omega) = 1$ precludes such ri norms for constants $c \neq 1$.

When Y is the indicator of a measurable set E with $\mu(E) = t > 0$, this specializes to

$$\frac{1}{t} \int_E |X| \, d\mu \leq \frac{1}{t} \int_0^t X^*(\omega) \, d\omega. \quad (2.67)$$

This suggests the definition of the *maximal function* [Bennett and Sharpley, 1988, pp. 52-53]:

$$X^{**}(t) := \frac{1}{t} \int_0^t X^*(\omega) \, d\omega = \frac{1}{t} \sup \left\{ \int_E |X| \, d\mu : \mu(E) = t \right\}, \quad t > 0, \quad (2.68)$$

where the latter equality is here stated without proof. The maximal function achieves the highest average of the function X over sets of measure t . This is done by integrating quantiles backwards:

$$\forall t \in (0, 1] : X^{**}(t) = \frac{1}{t} \int_0^t X^*(\omega) \, d\omega \quad (2.69)$$

$$= \frac{1}{t} \int_0^t F_{|X|}^{-1}(1 - \omega) \, d\omega \quad (2.70)$$

$$= \frac{1}{1-t} \int_{1-t}^1 F_{|X|}^{-1}(\omega) \, d\omega \quad (2.71)$$

$$= \text{CVar}_{1-t}(|X|). \quad (2.72)$$

The special behaviour of CVar_α is due to the fact that it integrates the function only in its $1 - \alpha$ tail, where the function values are highest (recall that X^* is decreasing). We observe the following remarkable fact [Bennett and Sharpley, 1988, p. 61]

$$(\forall \alpha \in [0, 1] : \text{CVar}_\alpha(|X|) \leq \text{CVar}_\alpha(|Y|)) \implies \|X\|_{\mathcal{R}} \leq \|Y\|_{\mathcal{R}} \quad (2.73)$$

for any ri norm $\|\cdot\|_{\mathcal{R}}$. We will later see that the special behaviour of CVar_α is in some sense shared by the wider class of spectral risk measures (Theorem 47). To this end, we need to introduce the *fundamental function* of an ri space.

2.4.4 THE FUNDAMENTAL FUNCTION

Definition 36. Let $\mathcal{R}$ be an ri space with function norm R . For each measurable subset $E \subseteq \Omega$ with measure $\mu(E) = t$, we define the *fundamental function* $\phi_{\mathcal{R}} : [0, 1] \rightarrow \mathbb{R}^+$ as

$$\phi_{\mathcal{R}}(t) := \|\chi_E\|_{\mathcal{R}} = R(\chi_E), \quad (2.74)$$

where the latter equality comes from the nonnegativity of indicator functions.

When the space is clear from the context, we drop the subscript. Due to the ri property, the choice of the set E does not matter. If R is a law invariant coherent risk measure,

i.e. an upper prevision, $\phi(t)$ specifies a coherent upper probability and then¹³ $\phi(1) = 1$. Since we stipulated $R(\chi_\Omega) = 1$ for any ri function norm, it always holds that $\phi(1) = 1$. The value of t corresponds to the underlying probability with respect to the base measure, $\mu(E)$, which is then distorted through R . For example, the expectation has the fundamental function $\phi_{\mathcal{L}^1}(t) = t$, whereas the ess sup has fundamental function $\phi_{\mathcal{L}^\infty}(t) = \chi_{(0,1]}$, so that $\phi_{\mathcal{L}^\infty}(0) = 0$ and $\phi_{\mathcal{L}^\infty}(t) = 1$ otherwise. For any ri space, the fundamental function is *quasiconcave*, that is, it satisfies [Bennett and Sharpley, 1988, p. 67]:

$$\phi \text{ is non-decreasing and } \phi(0) = 0 \quad (2.75)$$

$$t \mapsto \phi(t)/t \text{ is decreasing} \quad (2.76)$$

$$\phi \text{ is continuous except perhaps at the origin.} \quad (2.77)$$

However, we focus on concave fundamental functions. Every concave function is also quasiconcave, but the converse is not necessarily true. The conceptual reason for our restriction is that the fundamental function models risk aversion on events: the indicator function χ_E represents the uncertain unit loss with probability $\mu(E)$ and 0 loss with probability $1 - \mu(E)$. Then $\phi(t) = \phi(\mu(E))$ is our risk assessment for this simple random variable. We argued before that a reasonable risk aversion profile is always concave, as it then puts more weight on worse outcomes in a systematic way. Hence $\phi \circ \mu$ defines a submodular capacity on events.

Mathematically, the restriction to concave fundamental functions is also not significant since it can be shown that an ri space with quasiconcave fundamental function ϕ can always be equivalently renormed to have a concave fundamental function, the least concave majorant of ϕ [Bennett and Sharpley, 1988, p. 71]. Henceforth we always assume ϕ to be concave. We denote the class of concave functions $\phi : [0, 1] \rightarrow [0, 1]$ with $\phi(0) = 0$ and $\phi(1) = 1$ as Φ . The right limit at 0 is $\phi(0+)$. If the additional condition of continuity at 0 is also satisfied, i.e. $\phi(0+) = 0$ and $\phi \in \Phi$, we write $\phi \in \Phi_{0+}$.

Since ϕ only encodes the behavior of R on events, i.e. an upper probability, there is some freedom left in specifying a corresponding risk measure. However, we will show that ϕ still imposes significant structure (Section 2.4.8, 2.4.9). As an example, we consider $\phi(t) = 1 - (1 - t)^2 = 2t - t^2$. Two different risk measures, which share this fundamental function, are *MaxVar* [Cherny and Madan, 2009] (MaxV) and the *Dutch risk measure* [Van Heerwaarden and Kaas, 1992] (Du):

$$\text{MaxV}(X) := \mathbb{E}[\max(X_1, X_2)], \quad X_1, X_2 \stackrel{\text{ind}}{\sim} X \quad \forall X \in \mathcal{M}^+ \quad (2.78)$$

$$\text{Du}(X) := \mathbb{E}[\max(X, \mathbb{E}[X])] \quad \forall X \in \mathcal{M}^+, \quad (2.79)$$

¹³A coherent risk measure satisfies translation equivariance, which also implies $R(c) = c$, hence $R(\chi_\Omega) = \phi(1) = 1$.

where $X_1, X_2 \stackrel{\text{ind}}{\sim} X$ means that the random variables are independent and share the same distribution. MaxV is indeed the spectral risk measure corresponding to the distortion ϕ . Let $X \in \mathcal{M}^+$:

$$\text{MaxV}(X) = \int_0^\infty 1 - (1 - S_X(\omega))^2 \, d\omega = \int_0^\infty 1 - (1 - (1 - F_X(\omega)))^2 \, d\omega \quad (2.80)$$

$$= \int_0^\infty 1 - F_X^2(\omega) \, d\omega. \quad (2.81)$$

This is just the expectation of a random variable with distribution function $F_Y = F_X^2$, i.e. $Y = \max(X_1, X_2)$, $X_1, X_2 \stackrel{\text{ind}}{\sim} X$. The Dutch risk measure, on the other hand, is also law invariant but not spectral. It is also easy to see that $\mathbb{E}[X] \leq \text{Du}(X) \leq \text{MaxV}(X) \, \forall X \in \mathcal{M}^+$ by applying Jensen's inequality in Equation 2.78. In this way, the Dutch risk measure is more optimistic than MaxVar on general random variables, even if they share the same risk aversion profile on indicator functions. This is no coincidence: we now show that, given an arbitrary concave fundamental function, spectral risk measures correspond to the most pessimistic extension of ϕ to all (nonnegative) random variables. In contrast, we observe in Theorem 50 that the Dutch risk measure is the most optimistic extension for its specific ϕ .

2.4.5 THE LORENTZ AND MARCINKIEWICZ NORMS

Given any concave fundamental function $\phi \in \Phi$, the *Lorentz norm* of $X \in \mathcal{M}$ is defined as

$$\|X\|_{\Lambda_\phi} := \int_0^1 X^*(\omega) \, d\phi(\omega) \quad (2.82)$$

$$= X^*(0)\phi(0+) + \int_0^1 X^*(\omega)\phi'(\omega) \, d\omega \quad (2.83)$$

$$= X^*(0)\phi(0+) + \int_0^1 F_{|X|}^{-1}(1-\omega)\phi'(\omega) \, d\omega, \quad (2.84)$$

where we immediately recognize the correspondence to the distortion (spectral) risk measure Equation 2.176 on the positive cone with distortion ϕ , if $\phi \in \Phi_{0+}$, i.e. if ϕ is continuous at 0. To the best of our knowledge, this connection has not been reported yet. It is easy to check that the Lorentz norm indeed has fundamental function ϕ . In particular, $\|\cdot\|_{\mathcal{L}^1}$ and $\|\cdot\|_{\mathcal{L}^\infty}$ are both Lorentz norms for their respective fundamental functions. While $\phi_{\mathcal{L}^1} \in \Phi_{0+}$, we have $\phi_{\mathcal{L}^\infty} \notin \Phi_{0+}$. The effect of $\phi(0+) > 0$ is to put a fixed weight on the supremum, meaning that further decreasing its probability would not further decrease its weight. In the extreme case of $\phi_{\mathcal{L}^\infty}$, all the weight is put on $X^*(0) = \text{ess sup}(|X|)$. In practice, we see little motivation for choosing a $\phi \in \Phi \setminus \Phi_{0+}$.

Another important norm, the *Marcinkiewicz norm* of $X \in \mathcal{M}$, is defined as

$$\|X\|_{M_\phi} = \sup_{0 < t \leq 1} \{\phi(t) X^{**}(t)\} \quad (2.85)$$

$$= \sup_{0 < t \leq 1} \left\{ \phi(t) \frac{1}{t} \int_0^t X^*(\omega) \, d\omega \right\} \quad (2.86)$$

$$= \sup_{0 < t \leq 1} \{\phi(t) \text{CVar}_{1-t}(|X|)\} \quad (2.87)$$

and also has fundamental function ϕ . It is clear that both the Lorentz and the Marcinkiewicz norms are rearrangement invariant, as they are defined in terms of X^* . For the proof that they are indeed valid ri norms, we refer to [Rubshtein et al., 2016, pp. 116, 143].

Theorem 37. [Bennett and Sharpley, 1988, p. 72]. Let Λ_ϕ (resp. M_ϕ) be the ri spaces of the functions for which the Lorentz (resp. Marcinkiewicz) norm is finite. For any other ri space $\mathcal{R}$ with fundamental function $\phi \in \Phi$ we have the embedding

$$\Lambda_\phi \hookrightarrow \mathcal{R} \hookrightarrow M_\phi \quad (2.88)$$

and 1 is a feasible embedding constant:

$$\|X\|_{M_\phi} \leq \|X\|_{\mathcal{R}} \quad \forall X \in \mathcal{R}, \quad \|X\|_{\mathcal{R}} \leq \|X\|_{\Lambda_\phi} \quad \forall X \in \Lambda_\phi. \quad (2.89)$$

Since an ri space consists of those functions for which the norm is finite, the largest norm yields the smallest space and vice versa. We here state the theorem without proof. In Section 2.4.8 we provide a novel proof, which also gives an intuition for the *why* behind the result. This “sandwiching” result justifies the name *fundamental function*: it indeed captures a fundamental aspect of an ri norm and confines all coherent risk measures with a given fundamental function to live between the Marcinkiewicz and the Lorentz norm of that fundamental function. From this it follows, for example, that $\text{Du}(X) \leq \text{MaxV}(X) \, \forall X \in \mathcal{M}^+$, as the MaxV is the Lorentz norm and they have the same fundamental function. This result has direct behavioural implications for a decision maker: given a law invariant coherent upper probability, the natural extension, which coincides with its spectral risk measure (Section 2.3.6), hence the Lorentz norm, is the most pessimistic in the sense that it assigns the highest risk to random variables, while being compatible with the specified upper probability. On the other hand, the Marcinkiewicz norm is its most optimistic extension. However, in contrast to the Lorentz norm, the Marcinkiewicz norm is not in general translation equivariant (see Section 2.4.6, 2.4.8) and thus not in general a coherent risk measure (on $\mathcal{M}^+$).

In fact, the Lorentz and the Marcinkiewicz norm stand in a dual relationship. The dual fundamental function to ϕ is $\phi^*(t) := t/\phi(t)$ and can be shown to be the fundamental

function of the associate space¹⁴. Then we can write the Lorentz norm as

$$\|X\|_{\Lambda_\phi} = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) d\omega : \|Y\|_{M_{\phi^*}} \leq 1, Y \in \mathcal{M}^+ \right\}, \quad (2.90)$$

using the Marcinkiewicz norm with the dual fundamental function as its associate norm. The other direction is more complicated: if ϕ is concave, ϕ^* might in general only be quasiconcave. However, the Lorentz norm is only a norm for concave fundamental functions. It can be shown that the dual of the Marcinkiewicz norm then is the Lorentz norm with respect to the least concave majorant of ϕ^* [Rubshtein et al., 2016, p. 147]. For example, the associate relationship of $\|\cdot\|_{\mathcal{L}^\infty}$ and $\|\cdot\|_{\mathcal{L}^1}$ is due to the Marcinkiewicz-Lorentz duality.

It has been observed [Rubshtein et al., 2016, p. 157] that in some special cases $\|\cdot\|_{\Lambda_\phi} = \|\cdot\|_{M_\phi}$, that is, a coincidence of the Lorentz and the Marcinkiewicz norm for the same fundamental function. As a consequence, the space of all ri norms collapses to a point due to the embedding theorem: there is then only a single ri norm with the given fundamental function. For instance, this holds true for $\mathcal{L}^1$: Let $\phi(t) = t$. Then $\|X\|_{\Lambda_\phi} = \|X\|_{\mathcal{L}^1}$. Also,

$$\|X\|_{M_\phi} = \sup_{0 < t \leq 1} \{\phi(t) X^{**}(t)\} = \sup_{0 < t \leq 1} \{t \cdot X^{**}(t)\} \quad (2.91)$$

$$= \sup_{0 < t \leq 1} \left\{ \int_0^t X^*(\omega) d\omega \right\} = \int_0^1 X^*(\omega) d\omega = \|X\|_{\mathcal{L}^1}. \quad (2.92)$$

Similarly, one easily checks that the coincidence also holds for $\mathcal{L}^\infty$. Hence $\mathcal{L}^1$ and $\mathcal{L}^\infty$ are distinguished spaces as they allow only a single ri norm. We prove a novel result in Section 2.4.8 (Theorem 52): the Lorentz and Marcinkiewicz norm coincide if and only if the fundamental function is of the form $\phi(t) = \min(t/(1-\alpha), 1)$ for some $\alpha \in [0, 1)$ or $\alpha \rightarrow 1$, i.e. for CVar-type fundamental functions. This further underlines the particularity of CVar, as it is the single coherent risk measure with this fundamental function (upper probability).

2.4.6 NONNEGATIVITY AND TRANSLATION EQUIVARIANCE

In the literature on ri spaces, a function norm is only defined on functions taking values in $[0, \infty]$. Recall that to obtain a valid Banach space, this is then extended to a norm by $\|X\| = R(|X|)$ using the absolute value. In our context, this is undesirable, as we want to distinguish loss from gain. Furthermore, we are interested in translation equivariant functionals. One possibility to resolve this tension is to postulate that all random variables are bounded from below — in the context of machine learning, losses are often bounded from below by 0. If the lower bound is negative, we can compute in the presence of translation

¹⁴This holds true generally [Rubshtein et al., 2016, p. 135], not restricted to the Lorentz/Marcinkiewicz duality. Note that if ϕ is concave, the dual fundamental function might only be quasiconcave.

equivariance:

$$R(X) = R(X + c) - c, \quad (2.93)$$

for some constant c so that $\text{ess inf}(X + c) \geq 0$. It is then sufficient to define the norm only for nonnegative random variables. However, the definition of translation equivariance itself requires dealing with potentially negative random variables. We instead propose the following restricted definition of positive translation equivariance (PTE).

Definition 38. *An ri function norm R is called PTE if*

$$\forall X \in \mathcal{M}^+, c \in \mathbb{R} \text{ s.t. } X + c \geq 0 : R(X + c) = R(X) + c. \quad (2.94)$$

We also call an ri norm PTE if it is induced by an ri function norm which is PTE, and similarly we call an ri space PTE if it carries an ri norm which is PTE.

The constant c can potentially be negative but we require that $X + c$ is nonnegative. We now show that PTE is equivalent to the possibility of reducing the representation of a function norm via its associate to dual variables with $\mathbb{E}[Y] = R'(Y) = 1$. Recall that *any* ri function norm admits a representation of the form:

$$R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega : R'(Y) \leq 1, Y \in \mathcal{M}^+ \right\}. \quad (2.95)$$

Theorem 39. *An ri function norm R can be represented in the following reduced form if and only if it is positive translation equivariant (PTE):*

$$R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+ \right\}. \quad (2.96)$$

We call this the positive translation equivariant representation of R . The proof is in Appendix 2.8.2.

Remark 40. When risk measures R are defined on the whole $\mathcal{M}$, translation equivariance *requires* that for any Y in any envelope representation $\mathcal{Y}$ of R , it holds $\mathbb{E}[Y] = 1$. Standard proofs for this (see Section 2.3.2) rely on negative values. With the restriction to the positive cone, we are only able to make the weaker statement that R *allows* such a representation, not that any representation needs to be of this form.

Example 41. Consider $R^1 = \mathbb{E}$. When it is defined on the whole space, the only envelope representation of R is the singleton $\{1\}$. When R is restricted to the positive cone $\mathcal{M}^+$, the set $\{Y : R^\infty(Y) \leq 1\}$ is also a valid representation. To see that this set is not a valid envelope on the whole space, consider the case of negative X and $Y = 0$.

Example 42. It is easy to see that the Lorentz norm $\|\cdot\|_{\Lambda_\phi}$ is PTE for any $\phi \in \Phi$. In contrast, the Marcinkiewicz norm $\|\cdot\|_{M_\phi}$ is PTE if and only if $\phi(t) = \min\{t/(1-\alpha), 1\}$ for some $\alpha \in [0, 1)$ or for $\alpha \rightarrow 1$, $\phi(t) = \chi_{(0,1]}$ (Theorem 52).

As an example of the above, consider the function norm R^1 , which can be written as

$$R^1(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) d\omega : R^\infty(Y) \leq 1, Y \in \mathcal{M}^+ \right\} \quad \forall X \in \mathcal{M}^+. \quad (2.97)$$

For nonnegative functions $X \in \mathcal{M}^+$, this is the expectation $\mathbb{E}[X]$. However, if we want to extend the above definition to work on general $X \in \mathcal{M}$, we must use its positive translation equivariant representation

$$R^1(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) d\omega : \mathbb{E}[Y] = R^\infty(Y) = 1, Y \in \mathcal{M}^+ \right\} \quad \forall X \in \mathcal{M}^+. \quad (2.98)$$

In fact, the singleton $\{1\}$ is sufficient to represent this function norm. The two representations are equivalent for nonnegative X , since in this case, the supremum will always be attained for the constant $Y = 1$. The second representation, but not the first, can easily be extended to work on potentially negative $X \in \mathcal{M}$. We define the generalized distribution function $\mu_X^- : \mathbb{R} \rightarrow [0, 1]$ and the generalized decreasing rearrangement $X^{*-} : [0, 1] \rightarrow \mathbb{R}$ as (cf. Equation 2.49, Equation 2.50)

$$\mu_X^-(\lambda) := \mu \{ \omega \in \Omega : X(\omega) > \lambda \} \quad (2.99)$$

$$X^{*-}(\omega) := \inf \{ \lambda \in \mathbb{R} : \mu_X^-(\lambda) \leq \omega \}. \quad (2.100)$$

Clearly, $X^{*-}(\omega) = F_X^{-1}(1 - \omega)$.

Theorem 43. *Let R be an ri function norm which has a positive translation equivariant representation. Then the extended functional*

$$R^-(X) := \sup \left\{ \int_0^1 X^{*-}(\omega) Y^*(\omega) d\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+ \right\} \quad \forall X \in \mathcal{M} \quad (2.101)$$

is a coherent law invariant risk measure, which coincides with R on nonnegative random variables and is translation equivariant (for all $c \in \mathbb{R}$).

Proof For $X \in \mathcal{M}^+$, the coincidence is obvious, since then $X^{*-} = X^*$. Law invariance is obvious since the definition is only in terms of the distribution of X . The other properties of a coherent risk measure are easily checked, but also follow from Kusuoka's theorem discussed in the next section. ■

Therefore, there is no real loss in generality when restricting ourselves to nonnegative functions in the following discussion. All comparison results obtained for positive translation

equivariant ri norms defined on nonnegative functions, which are essentially bounded from below, carry directly over to the extended functionals. Note, however, that for instance the Marcinkiewicz norm is not in general positive translation equivariant, hence cannot be extended to a translation equivariant functional. But see Theorem 50 for the construction of a PTE norm related to the Marcinkiewicz norm.

2.4.7 KUSUOKA REPRESENTATIONS

The celebrated Kusuoka representation theorem [Kusuoka, 2001] states that CVar's are the basic building blocks of any law invariant coherent risk measure. Kusuoka [2001] proved the theorem on $\mathcal{L}^\infty$ for law invariant coherent risk measures satisfying the Fatou property (akin to R3) and it has been subsequently extended to $\mathcal{L}^p$ spaces [see e.g. Pflug and Romisch, 2007]. In general, Kusuoka representations require an *atomless*¹⁵ probability space. We continue to work on the atomless standard probability space $[0, 1]$ with the Lebesgue measure and restrict ourselves to the positive cone to draw the connection to ri function norms.

Theorem 44. *Every ri function norm which is PTE admits a representation on the form*

$$R(X) = \sup_{\lambda \in \mathfrak{M}} \left\{ \int_0^1 \text{CVar}_\alpha(X) \, d\lambda(\alpha) \right\} \quad \forall X \in \mathcal{M}^+ \quad (2.102)$$

for some set $\mathfrak{M}$ of probability measures on $[0, 1]$.

Proof We observe that in the framework of ri spaces, the Kusuoka representation is a direct corollary of the representation of a norm via its associate norm:

$$R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega : R'(Y) \leq 1, Y \in \mathcal{M}^+ \right\} \quad \forall X \in \mathcal{M}^+. \quad (2.103)$$

This is a particular instantiation of the result in convex analysis that a norm is the support function of the unit ball of its dual norm. Here, the Y^* are nonnegative and decreasing. If R is a coherent risk measure, it is positive translation equivariant. Therefore it suffices to restrict ourselves to the subset of dual variables $\mathcal{Y}_1 := \{Y^* : R'(Y) = \mathbb{E}[Y] = 1\}$ (Theorem 39), that is $\int_0^1 Y^*(\omega) \, d\omega = 1$. Then we can write this as ($Y \in \mathcal{M}^+$)

$$R(X) = R_{\mathcal{Y}_1}(X) = \sup \left\{ \int_0^1 F_X^{-1}(1 - \omega) Y^*(\omega) \, d\omega : R'(Y) = \mathbb{E}[Y] = 1 \right\} \quad \forall X \in \mathcal{M}^+. \quad (2.104)$$

We recognize a supremum over a set of spectral risk measures, as each $w(\omega) := Y^*(1 - \omega)$ is a legitimate spectral weighting function (Section 2.3.5). Thus we can also write

¹⁵A set $B \subseteq \mathcal{F}$ on $(\Omega, \mathcal{F}, P)$ is an *atom* if $P(B) > 0$ and $A \subsetneq B \Rightarrow P(A) = 0$. A probability space is *atomless* if it has no atoms.

this as a supremum over distortion risk measures with concave distortions $\mathcal{Z} := \{t \mapsto \int_0^t Y^*(\omega) d\omega : Y \in \mathcal{Y}_1\}$. For each $Z \in \mathcal{Z}$, $Z : [0, 1] \rightarrow [0, 1]$, we have $Z(0) = 0$, $Z(1) = 1$ (due to PTE), since $Z(1) = \int_0^1 Y^*(\omega) d\omega$. Therefore we obtain a representation as the supremum over Choquet integrals:

$$R_{\mathcal{Y}_1}(X) = R_{\mathcal{Z}}(X) = \sup \left\{ \int_0^\infty Z(\mu_X(\omega)) d\omega : Z \in \mathcal{Z} \right\} \quad \forall X \in \mathcal{M}^+. \quad (2.105)$$

We call either of the sets $\mathcal{Z}$ or $\mathcal{Y}_1$ a *Kusuoka set* of R , since either fully characterizes the risk measure; in general, we notate dual variables as Y and the integrals of their decreasing rearrangements as Z . In subsequent discussions, we shall also use the term “Kusuoka set” when PTE is not satisfied and the representation therefore describes a general ri function norm, without the constraint that $Z(1) = 1 \Leftrightarrow \int_0^1 Y^*(\omega) d\omega = 1$. Finally, each spectral weighting function $w(\omega) = Y^*(1 - \omega)$ can be associated with a probability measure λ_w on $[0, 1]$, via the relationship [see e.g. [Pichler, 2015](#)]

$$\lambda_w(E) := w(0)\delta_0(E) + \int_E 1 - \alpha dw(\alpha) \quad E \text{ measurable}, \quad (2.106)$$

where δ_0 is the Dirac measure at 0. With this family of measures $\mathfrak{M} := \{\lambda_w : w(\omega) = Y^*(1 - \omega), Y^* \in \mathcal{Y}_1\}$, the representation Equation 2.102 is recovered. We remark that Kusuoka representations need not be unique in general. Conversely, we can specify a functional $R_{\mathcal{Y}}$ directly as

$$R_{\mathcal{Y}}(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) d\omega : Y \in \mathcal{Y} \right\} \quad \forall X \in \mathcal{M}^+. \quad (2.107)$$

For any set of nonnegative decreasing functions $\mathcal{Y}$, we are guaranteed to obtain a valid ri norm¹⁶, since the supremum over a set of Lorentz norms preserves the relevant properties (Lemma 90). However, $\mathcal{Y}$ need not be the maximal envelope.

When PTE is satisfied, the domain of R can be extended from $\mathcal{M}^+$ to $\mathcal{M}$ to yield a coherent risk measure via the extension in Section 2.4.6, so that the original Kusuoka representation on the whole space is recovered. Observe that if only a single Y^* suffices to represent the risk measure (of course, Y^* which never obtain the supremum may be added to an envelope representation) and the normalization $\int_0^1 Y^*(\omega) d\omega = 1$ holds, then Equation 2.104 reduces to the definition of a Lorentz norm with $\phi(0+) = 0$, i.e. a spectral risk measure. This relates to the observation made by [Pichler and Shapiro \[2012\]](#) that if the Kusuoka set is generated by a single element (modulo equimeasurability), then the risk measure is spectral. On the other hand, when starting with a single measure λ on $[0, 1]$, the situation is technically more subtle. If λ does not have an atom at 1, i.e. $\lambda(\{1\}) = 0$, then it is equiv-

¹⁶The normalization $R_{\mathcal{Y}}(1) = 1$ may not hold in general when no constraints on $\mathcal{Y}$ are imposed.

alent to a concave distortion, which is continuous at 0, and hence a spectral risk measure. Consider for instance the supremum risk measure R^∞ , represented by the Dirac measure at 1. This cannot be expressed as a single distortion function, when demanding continuity at 0. However, it can be represented as a supremum over a family of such functions. ■

This way of arriving at the Kusuoka representation provides new insights as compared to standard proofs. First, it reveals that the natural space to work on with coherent risk measures is not in general an $\mathcal{L}^p$ space, but rather a specific ri space. In the case of a spectral risk measure, this is a Lorentz space. A similar observation has been made by Pichler [2013], but the author did not establish the link to the general theory of ri spaces. Moreover, it reveals that the Kusuoka representation is nothing more than the representation of a norm via its dual norm in the presence of the ri property. In general, a Banach function norm which is not ri can be written as

$$R(X) = \sup \left\{ \int_0^1 X(\omega)Y(\omega) d\omega : R(Y) \leq 1 \right\} \quad \forall X \in \mathcal{M}^+. \quad (2.108)$$

Under the ri property, X and Y can be replaced by any distributionally equivalent choice. The Hardy-Littlewood inequality Equation 2.66 tells us that the supremum is achieved for X^* and Y^* . Another statement is in terms of *Fréchet bounds*. Let H be a bivariate distribution function with marginals F and G , where “distribution function” is in the classical probabilistic sense. Then it holds (see e.g. Pflug and Ruszczynski [2001, Section 1.2.2]):

$$\max\{F(x) + G(y) - 1, 0\} \leq H(x, y) \leq \min\{F(x), G(y)\}, \quad \forall x, y \in \mathbb{R}. \quad (2.109)$$

Let X have distribution F and Y have distribution G . The lower bound is achieved when X and Y are *antimonotone*; the upper bound is achieved when they are *comonotone*. Comonotonicity means that any of the following equivalent conditions hold:

- C1)** $H(x, y) = \mu(X \leq x, Y \leq y) = \min(F(x), G(y))$
- C2)** $(X(\omega) - X(\omega'))(Y(\omega) - Y(\omega')) \geq 0 \quad \forall \omega, \omega' \in \Omega$
- C3)** $\exists Z \in \mathcal{M}$, non-decreasing functions f, g such that $X = f(Z), Y = g(Z)$.

Comonotone X and Y have perfect rank correlation. The definition of antimonotonicity is the opposite, perfect negative rank correlation¹⁷. It is known that

$$\mathbb{E}[XY] \leq \mathbb{E}[\tilde{X}\tilde{Y}], \quad (2.110)$$

where $\tilde{X}$ and $\tilde{Y}$ are coupled in a comonotone way, but with the same marginals as X and Y , respectively. Note that for any pair of random variables, X^* and Y^* are comonotone (C2

¹⁷Antimonotonicity means that $(X(\omega) - X(\omega'))(Y(\omega) - Y(\omega')) \leq 0 \quad \forall \omega, \omega' \in \Omega$.

obviously holds). Therefore we recover Hardy-Littlewoods inequality Equation 2.66.

The concept of comonotonicity is relevant both from a financial as well as from a purely uncertainty-motivated perspective. In finance, a desirable property for a risk measure is additivity for comonotone risks. That is, if X and Y are comonotone:

$$R(X + Y) = R(X) + R(Y). \quad (2.111)$$

The rationale is that in the presence of perfect rank correlation, X cannot work as a hedge against Y and vice versa. Essentially, comonotone gambles are bets on the same event in the world (due to C3). A decision maker under uncertainty may reason in an analogous way that when adding two comonotone uncertain losses, no hedge against uncertainty is possible.

Theorem 45. [*Kusuoka, 2001*] on $\mathcal{L}^\infty$. *A law invariant coherent risk measure with the Fatou property (akin to R3) is comonotonically additive if and only if it has a Kusuoka representation by a single probability measure on $[0, 1]$.*

That is, only spectral risk measures and its pathological neighbours (e.g. the supremum risk measure) are comonotonically additive.

2.4.8 FAMILIES OF FUNDAMENTAL FUNCTIONS

The Kusuoka representation suggests a new way to characterize an ri norm fully by a family of fundamental functions. Recall that any ri norm has the representation

$$R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y(\omega) \, d\omega : Y \in \mathcal{Y} \right\}, \quad \mathcal{Y} = \{Y^* : Y \in \mathcal{M}^+, R'(Y) \leq 1\} \quad (2.112)$$

for a set $\mathcal{Y}$ of nonnegative decreasing functions Y . Let $E \subseteq \Omega$ be a measurable subset with $\mu(E) = t$. The fundamental function induced by R is

$$\phi(t) = R(\chi_E) = \sup \left\{ \int_0^t Y(\omega) \, d\omega : Y \in \mathcal{Y} \right\} = \sup_{Z \in \mathcal{Z}} Z(t), \quad (2.113)$$

where $\mathcal{Z} = \{t \mapsto \int_0^t Y(\omega) \, d\omega : Y \in \mathcal{Y}\}$. Depending on the context, we may call either $\mathcal{Y}$ or $\mathcal{Z}$ a Kusuoka set, as either fully describes the representation¹⁸. The fundamental function $\phi \in \Phi$ can be expressed as a supremum of concave distortions Z , each of which can be seen as the fundamental function of a spectral risk measure. Conceptually, we may say that ambiguity about the risk aversion spectrum exhausts the whole space of

¹⁸We typically use $\mathcal{Y}$ as the primary objects, as they appear directly in the ri space version of the Kusuoka representation. When starting with $\mathcal{Z}$, we need that their derivatives be defined almost everywhere. If R is PTE, then the set of probability measures $\mathfrak{M}$ offers yet another representation.

coherent risk measures. From this angle, it is possible to derive intuitive and instructive proofs, for instance for the extremal status of the Marcinkiewicz and the Lorentz norm. Furthermore, we construct the smallest translation equivariant norm, given an arbitrary concave fundamental function. First, we need a technical lemma, which we specialize to the domain $(0, 1]$.

Lemma 46. *Hardy's lemma [Bennett and Sharpley, 1988]. Let Y_1 and Y_2 be nonnegative measurable functions on $(0, 1]$ and*

$$\int_0^t Y_1(\omega) \, d\omega \leq \int_0^t Y_2(\omega) \, d\omega \quad \forall t \in (0, 1]. \quad (2.114)$$

If η is any nonnegative decreasing function on $(0, 1]$, then

$$\int_0^1 \eta(\omega) Y_1(\omega) \, d\omega \leq \int_0^1 \eta(\omega) Y_2(\omega) \, d\omega. \quad (2.115)$$

As an example, in our context this implies that if a concave distortion $Z_1(t) = \int_0^t Y_1(\omega) \, d\omega$ majorizes another Z_2 pointwise ($Z_1, Z_2 \in \Phi_{0+}$), then the Lorentz norm corresponding to Z_1 majorizes the one corresponding to Z_2 for all random variables (set $\eta = X^*$). This does not apply to Lorentz norms, where the distortion is not continuous at 0, however, as such a distortion cannot be represented by an integral of the form $Z_1(t) = \int_0^t Y_1(\omega) \, d\omega$.

Theorem 47. [Bennett and Sharpley, 1988, p. 72]. *Given any ri norm R with fundamental function $\phi \in \Phi_{0+}$, we have $R(X) \leq \|X\|_{\Lambda_\phi} \, \forall X \in \mathcal{M}^+$.*

Proof Let $R(X) = \sup_{Y \in \mathcal{Y}} \{ \int_0^1 X^*(\omega) Y(\omega) \, d\omega \}$. We know that $\phi(t) = \sup_{Y \in \mathcal{Y}} \{ \int_0^t Y(\omega) \, d\omega \}$. Recall that $\|X\|_{\Lambda_\phi} = \int_0^1 X^*(\omega) \phi'(\omega) \, d\omega$ if $\phi \in \Phi_{0+}$. But since ϕ majorizes all elements of $\{t \mapsto \int_0^t Y(\omega) \, d\omega : Y \in \mathcal{Y}\}$ pointwise and $\phi \in \Phi_{0+}$, we immediately obtain from the qualification in Hardy's lemma

$$\int_0^t \phi'(\omega) \, d\omega \geq \int_0^t Y(\omega) \, d\omega \quad \forall 0 < t \leq 1 \quad \forall Y \in \mathcal{Y} \quad (2.116)$$

that it holds

$$R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y(\omega) \, d\omega : Y \in \mathcal{Y} \right\} \quad (2.117)$$

$$\leq \sup \left\{ \int_0^1 X^*(\omega) \phi'(\omega) \, d\omega : Y \in \mathcal{Y} \right\} = \|X\|_{\Lambda_\phi}. \quad (2.118)$$

The statement also holds true if $\phi \in \Phi \setminus \Phi_{0+}$ [Rubshtein et al., 2016]. ■

Theorem 48. [Bennett and Sharpley, 1988, p. 70]. Given any ri norm R with fundamental function $\phi \in \Phi$, we have $\|X\|_{M_\phi} \leq R(X) \forall X \in \mathcal{R}$.

Proof Write

$$R(X) = \sup_{Z_\gamma \in \mathcal{Z}} \left\{ \int_0^1 X^*(\omega) Z'_\gamma(\omega) d\omega \right\}, \quad \phi(t) = \sup_{Z_\gamma \in \mathcal{Z}} Z_\gamma(t) \quad (2.119)$$

for some Kusuoka set $\mathcal{Z}$ of R . Each Z_γ is concave since it is the integral of a nonnegative decreasing function. Hence, Z_γ pointwise majorizes all of the piecewise linear functions¹⁹

$$\forall t \in (0, 1] : Z_{\gamma,t}(x) := \begin{cases} Z_\gamma(t) \frac{x}{t} & , x \leq t \\ Z_\gamma(t) & , x > t. \end{cases} \quad (2.120)$$

The $Z_{\gamma,t}$ are constructed as the integrals of the functions²⁰

$$\forall t \in (0, 1] : Z'_{\gamma,t} := \begin{cases} \frac{Z_\gamma(t)}{t} & , x \leq t \\ 0 & , x > t. \end{cases} \quad (2.121)$$

Applying Hardy's lemma yields

$$R(X) = \sup_{Z_\gamma \in \mathcal{Z}} \left\{ \int_0^1 X^*(\omega) Z'_\gamma(\omega) d\omega \right\} \quad (2.122)$$

$$\geq \sup_{Z_{\gamma,t}} \left\{ \int_0^1 X^*(\omega) Z'_{\gamma,t}(\omega) d\omega \right\} \quad (2.123)$$

$$= \sup_{Z_{\gamma,t}} \left\{ \frac{Z_\gamma(t)}{t} \int_0^t X^*(\omega) d\omega \right\} \quad (2.124)$$

$$\geq \sup_{0 < t \leq 1} \left\{ \sup \left\{ \frac{Z_\gamma(t)}{t} : Z_\gamma \in \mathcal{Z} \right\} \int_0^t X^*(\omega) d\omega \right\} \quad (2.125)$$

$$= \sup_{0 < t \leq 1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega) d\omega \right\} = \|X\|_{M_\phi}, \quad (2.126)$$

since we have $\phi(t) = \sup_{Z_\gamma} Z_\gamma(t) = \sup_{Z_\gamma} Z_\gamma(t) \frac{t}{t} = \sup_{Z_\gamma} Z_{\gamma,t}(t)$ ■

Remark 49. Throughout the paper, we use shorthand notation to avoid explicitly writing

¹⁹Let $Z_\gamma : [0, 1] \rightarrow \mathbb{R}^+$ be concave, i.e. $\forall \alpha \in [0, 1] : Z_\gamma(\alpha t + (1 - \alpha)x) \geq \alpha Z_\gamma(t) + (1 - \alpha)Z_\gamma(x)$. Choosing $x = 0$ yields $Z_\gamma(\alpha t) \geq \alpha Z_\gamma(t)$. For any $x \leq t$ hence $x = \alpha t$ for some $\alpha \in [0, 1]$, we obtain $Z_\gamma(x) \geq \frac{x}{t} Z_\gamma(t)$. For $x > t$ the statement is obvious, as the concave Z_i has nonnegative derivative, whereas $Z_{\gamma,t}$ has zero derivative.

²⁰The “derivatives” $Z'_{\gamma,t}$ are here the primary objects. The derivative of $Z_{\gamma,t}$ at a kink may not exist, but we have defined $Z'_{\gamma,t}$ as prior to $Z_{\gamma,t}$. This enables us to apply Hardy's lemma, which works with the integrals of $Z'_{\gamma,t}$ directly.

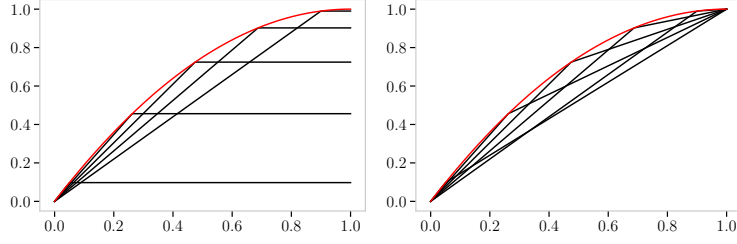


Figure 2.3: The red curve is the fundamental function $\phi(t) = 1 - (1 - t)^2$. Left: the black lines correspond to five selected ϕ_t in the Marcinkiewicz norm construction. Right: the black lines correspond to five selected ϕ_t in the positive translation equivariant Marcinkiewicz norm construction. In this particular case, the latter yields the Dutch risk measure. Due to PTE, the ϕ_t need to reach 1 at $t = 1$. In both cases, the supremum over the (infinite) family of black lines recovers the red line, i.e. the fundamental function ϕ .

the set over which a supremum ranges, when it is clear from context. For instance, the notation ' $\sup_{Z_{\gamma,t}} f(Z_{\gamma,t}, t)$ ' means ' $\sup\{f(Z_{\gamma,t}, t) : Z_{\gamma} \in \mathcal{Z}, t \in (0, 1]\}$ ', i.e. implicitly γ and t vary in the expression. By contrast, as an example where t is already bound, ' $h(t) = \sup_{Z_{\gamma,t}} f(Z_{\gamma,t}, t)$ ' means ' $h(t) = \sup\{f(Z_{\gamma,t}, t) : Z_{\gamma} \in \mathcal{Z}\}$ '.

We constructed the functions $Z_{\gamma,t}$ so that they are linear up to t and then constant. This yields the Marcinkiewicz norm. A slight extension, where the functions are piecewise linear and reach $Z_{\gamma,t}(1) = 1$ yields the smallest positive translation equivariant ri norm.

Theorem 50. *Given any concave fundamental function $\phi \in \Phi$, we can construct the smallest positive translation equivariant ri norm as:*

$$\|X\|_{TM_{\phi}} = \sup_{0 < t < 1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega) d\omega + \frac{\phi(t) - 1}{t - 1} \int_t^1 X^*(\omega) d\omega \right\}. \quad (2.127)$$

For any other PTE ri norm R with fundamental function ϕ we have $\|X\|_{M_{\phi}} \leq \|X\|_{TM_{\phi}} \leq R(X) \forall X \in \mathcal{R}$. We call TM_{ϕ} the positive translation equivariant Marcinkiewicz norm.

Example 51. Recall that both the Dutch risk measure and the spectral MaxVar share the fundamental function $\phi(t) = 2t - t^2$. Then: $\|X\|_{TM_{\phi}} = \text{Du}(|X|)$. This result implies that given this fundamental function, the Dutch risk measure is the most optimistic coherent risk measure, whereas MaxVar is the most pessimistic one.

The proof is in Appendix 2.8.3. Next, we show when equality of all ri norms for a given fundamental function holds.

Theorem 52. *Given any concave fundamental function $\phi \in \Phi$, it holds that*

$$\|X\|_{M_{\phi}} = R(X) = \text{CVar}_{\alpha}(X) = \|X\|_{\Lambda_{\phi}} \quad \forall X \in \mathcal{M}^+ \quad (2.128)$$

for all ri function norms R with fundamental function ϕ if and only if $\phi(t) = \min\{t/(1 - \alpha), 1\}$ for some $\alpha \in [0, 1)$ or for $\alpha \rightarrow 1$, $\phi(t) = \phi_{\infty}(t) := \chi_{(0,1]}(t)$. For $\alpha \rightarrow 1$, $R(X) = \|X\|_{\mathcal{L}^{\infty}}$.

The proof is in Appendix 2.8.3. This result implies that for CVar_α -type fundamental functions (including $\mathcal{L}^1$ and $\mathcal{L}^\infty$ as special cases²¹), there is only a *single* ri norm and hence a single law invariant coherent risk measure. However, there is another interesting class of fundamental functions, for which all law invariant coherent risk measures coincide, but not all ri norms.

Theorem 53. *Let $\phi(t) = \beta t + (1 - \beta) \min(1, t/(1 - \alpha))$ for any $\alpha, \beta \in [0, 1]$. Then the Lorentz norm coincides with the positive translation equivariant Marcinkiewicz norm*

$$\|X\|_{\Lambda_\phi} = \|X\|_{TM_\phi} = \beta \mathbb{E}[X] + (1 - \beta) \text{CVar}_\alpha(X) =: \text{RIM}_{\alpha, \beta}(X) \quad \forall X \in \mathcal{M}^+. \quad (2.129)$$

Proof The Lorentz norm is easily computed

$$\|X\|_{\Lambda_\phi} = \int_0^{1-\alpha} X^*(\omega) \left(\beta + (1 - \beta) \frac{1}{1 - \alpha} \right) d\omega + \int_{1-\alpha}^1 X^*(\omega) \beta d\omega \quad (2.130)$$

$$= \beta \mathbb{E}[|X|] + (1 - \beta) \text{CVar}_\alpha(|X|). \quad (2.131)$$

A Kusuoka set of concave functions for the PTE Marcinkiewicz norm is

$$\forall t \in (0, 1) : \phi_{TM, t}(x) := \begin{cases} \phi(t) \frac{x}{t} & , x \leq t \\ \frac{1 - \phi(t)}{1 - t} x + \frac{\phi(t) - t}{1 - t} & , x > t \end{cases} \quad (2.132)$$

Observe that ϕ is piecewise linear with a kink at $t = 1 - \alpha$, irrespective of the value of β , which adjusts the slope. Choose $t = 1 - \alpha$. Tedious calculation reveals what is obvious, that $\phi(x) = \phi_{TM, 1-\alpha}(x)$. Therefore this $\phi_{TM, 1-\alpha}$ dominates all other $\phi_{TM, t}$ and the supremum in the Kusuoka representation is in fact attained. But then $\|X\|_{TM_\phi} = \int_0^1 X^*(\omega) \phi'_{TM, 1-\alpha}(\omega) d\omega = \|X\|_{\Lambda_\phi}$. ■

As a consequence, for this family of fundamental functions, the space of law invariant coherent risk measures collapses to a point. Moreover, the result is a spectral risk measure which is useful in practice as it can be easily computed. This function norm is called the *risk measure for integrated risk measurement* [Pflug and Ruszczynski, 2001]. The parameters α, β are intuitive knobs to adjust the tradeoff of tail-sensitivity (risk aversion) and globality (taking the full range of risk into account). Note also the close relation to the Dutch risk measure. From the representation

$$\text{Du}(X) = \sup_{0 < \beta < 1} \{ \beta \mathbb{E}[X] + (1 - \beta) \cdot \text{CVar}_\beta(X) \} \quad \forall X \in \mathcal{M}^+ \quad (2.133)$$

²¹More precisely, $\mathcal{L}^1$ and $\mathcal{L}^\infty$ are in fact the only two spaces for which the Marcinkiewicz and Lorentz norm coincide. This is due to the fact that for $\alpha \in [0, 1]$, the space induced by CVar_α is $\mathcal{L}^1$, whereas $\alpha \rightarrow 1$ yields the $\mathcal{L}^\infty$ space. See Section 2.4.9.

we observe that the Dutch risk measure can be seen as an ambiguity set over $\text{RIM}_{\alpha,\beta}$, where $\alpha = \beta$. This can be interpreted as a combination of risk and ambiguity aversion. To understand the relationship, consider the fundamental function $\phi(t) = 2t - t^2$ of the Dutch risk measure and construct the corresponding $\phi_{TM,t}$. Each such $\phi_{TM,t}$ can be written as a $\phi_{TM,t}(x) = \beta x + (1 - \beta) \min(1, x/(1 - \alpha))$, where $t = 1 - \alpha = 1 - \beta$. More generally, let $\phi(x) = 1 - (1 - x)^n$ for some natural number $n \geq 2$. As n increases, risk aversion increases. Then:

$$\|X\|_{TM_\phi} = \sup_{0 < \beta < 1} \beta^{n-1} \mathbb{E}[|X|] + (1 - \beta^{n-1}) \cdot \text{CVar}_\beta(|X|) \quad \forall X \in \mathcal{M}. \quad (2.134)$$

The corresponding Lorentz norm is [Cherny and Madan, 2009]

$$\|X\|_{\Lambda_\phi} = \mathbb{E}[\max(X_1, \dots, X_n)], \quad X_1, \dots, X_n \stackrel{\text{ind}}{\sim} |X| \quad \forall X \in \mathcal{M}. \quad (2.135)$$

We can generalize this further to find that the PTE Marcinkiewicz norm has a family of RIMs as its basic building blocks.

Theorem 54. *Let $\phi \in \Phi_{0+}$ a fundamental function. Then*

$$\|X\|_{TM_\phi} = \sup_{0 < t < 1} \frac{1 - \phi(1 - t)}{t} \mathbb{E}[|X|] + \left(1 - \frac{1 - \phi(1 - t)}{t}\right) \cdot \text{CVar}_t(|X|) \quad (2.136)$$

$$= \sup_{0 < t < 1} \text{RIM}_{\alpha(t), \beta(t)}(|X|), \quad \text{where } \alpha(t) = t, \beta(t) = \frac{1 - \phi(1 - t)}{t}. \quad (2.137)$$

Proof Set $\phi_{TM,1-\alpha}(x) = \phi_{\text{RIM}_{\alpha,\beta}}(x) = \beta x + (1 - \beta) \min\{1, x/(1 - \alpha)\}$. Both $\phi_{TM,1-\alpha}$ and $\phi_{\text{RIM}_{\alpha,\beta}}$ are piecewise linear with a kink at $1 - \alpha$. A piecewise calculation shows that $\beta = \frac{1 - \phi(1 - \alpha)}{\alpha}$ is a solution for both pieces. Then $\|X\|_{TM_\phi}$ has a Kusuoka representation in terms of spectral risk measures corresponding to the family of $\phi_{\text{RIM}_{\alpha,\beta}}$, but these are just the $\text{RIM}_{\alpha,\beta}$. ■

We remark that $\text{RIM}_{\alpha,\beta}$ admits the following variational representation:

$$\text{RIM}_{\alpha,\beta}(X) = \inf_{\mu \in \mathbb{R}} \mu + \mathbb{E}v(X - \mu) \quad \forall X \in \mathcal{M}, \quad (2.138)$$

where the regret function v is given by the piecewise linear function

$$v(t) = \begin{cases} \beta t & t \leq 0 \\ \frac{\beta\alpha - 1}{\alpha - 1} t & t > 0. \end{cases} \quad (2.139)$$

For the proof, see Appendix 2.8.3, which translates a result from Pflug and Ruszczynski [2001]. Note that $V(X) = \mathbb{E}v(X)$ fulfills the requirements of a coherent regret measure in

the quadrangle (Figure 2.1).

2.4.9 NORM EQUIVALENCES AND TAIL RISK

The fundamental function ϕ , which corresponds to a coherent upper probability, imposes substantial structure on the compatible norms, which have this ϕ as their fundamental function. In this section, we expand on this claim by proving several (non)-equivalence results based on the derivative of ϕ at the origin. Recall that two norms $\|\cdot\|_{\mathcal{R}_1}$ and $\|\cdot\|_{\mathcal{R}_2}$ are said to be equivalent if $\exists c_1, c_2 > 0 : c_1 \|X\|_{\mathcal{R}_1} \leq \|X\|_{\mathcal{R}_2} \leq c_2 \cdot \|X\|_{\mathcal{R}_1} \forall X \in \mathcal{R}_1 = \mathcal{R}_2$. While a theoretical norm equivalence does not imply equivalence from a practical standpoint, it is nevertheless interesting how much of the norm behaviour is controlled by $\phi'(0)$ already. From our findings we conclude that the theoretically most essential differences between ri norms concern their behaviour with regard to tails of the random variables, an observation which we further develop in [Fröhlich and Williamson, 2023].

We take inspiration from a result for coherent risk measures by Pichler [2013]. Here we restate it in terms of ri norms.

Theorem 55. *Let $\|\cdot\|_{\mathcal{R}_1}$ be an ri norm with Kusuoka set $\mathcal{Z}_1 = \{\phi_{1\gamma}\}$ and $\|\cdot\|_{\mathcal{R}_2}$ another ri norm with Kusuoka set $\mathcal{Z}_2 = \{\phi_{2\zeta}\}$, where γ and ζ are from some arbitrary index sets. Denote the corresponding Banach spaces of functions, on which the norms are finite, as $\mathcal{R}_1$ and $\mathcal{R}_2$. Then if the constant*

$$C := \sup_{\phi_{2\zeta} \in \mathcal{Z}_2} \inf_{\phi_{1\gamma} \in \mathcal{Z}_1} \sup_{0 < \alpha \leq 1} \frac{\phi_{2\zeta}(\alpha)}{\phi_{1\gamma}(\alpha)} \quad (2.140)$$

is finite, we have the relationship

$$\|X\|_{\mathcal{R}_2} \leq C \cdot \|X\|_{\mathcal{R}_1} \quad \forall X \in \mathcal{R}_1 \quad (2.141)$$

and $\mathcal{R}_1 \subseteq \mathcal{R}_2$, therefore $\mathcal{R}_1 \hookrightarrow \mathcal{R}_2$. If furthermore $\exists c > 0 : c \cdot \|X\|_{\mathcal{R}_1} \leq \|X\|_{\mathcal{R}_2} \forall Y \in \mathcal{R}_2$, then $\mathcal{R}_1 = \mathcal{R}_2$ and we say that the norms $\|\cdot\|_{\mathcal{R}_1}$ and $\|\cdot\|_{\mathcal{R}_2}$ are equivalent.

The proof is in Appendix 2.8.4, where we also discuss a subtle issue with the original result. In contrast, if $C = \infty$, we cannot make a statement for general ri norms (possibly, $\mathcal{R}_1 \subseteq \mathcal{R}_2$ or $\mathcal{R}_1 \not\subseteq \mathcal{R}_2$). At first sight one might conjecture that $C = \infty$ implies nonequivalence, but we provide a counterexample in Theorem 62. However, we can state the following slightly refined result.

Theorem 56. *Let the quantities $\|\cdot\|_{\mathcal{R}_1}$, $\|\cdot\|_{\mathcal{R}_2}$, $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{Z}_1 = \{\phi_{1\gamma}\}$, $\mathcal{Z}_2 = \{\phi_{2\zeta}\}$, be defined as in Theorem 55, so that $\phi_1(t) = \sup_{\phi_{1\gamma} \in \mathcal{Z}_1} \phi_{1\gamma}(t)$ and $\phi_2(t) = \sup_{\phi_{2\gamma} \in \mathcal{Z}_2} \phi_{2\gamma}(t)$ are the respective*

fundamental functions. If the constant

$$C' := \sup_{\alpha \rightarrow 0} \sup_{\phi_{2\zeta} \in \mathcal{Z}_2} \inf_{\phi_{1\gamma} \in \mathcal{Z}_1} \frac{\phi_{2\zeta}(\alpha)}{\phi_{1\gamma}(\alpha)} \quad (2.142)$$

is infinite, then the norms are not equivalent; we have $\mathcal{R}_1 \not\subseteq \mathcal{R}_2$ and

$$\nexists c : \|X\|_{\mathcal{R}_2} \leq c \cdot \|X\|_{\mathcal{R}_1} \quad \forall X \in \mathcal{R}_1. \quad (2.143)$$

Proof Suppose $C' = \infty$. The norm of the identity embedding $\mathcal{R}_1 \hookrightarrow \mathcal{R}_2$ is

$$\|\text{id}\| = \sup \left\{ \frac{\|X\|_{\mathcal{R}_2}}{\|X\|_{\mathcal{R}_1}} : X \in \mathcal{R}_1 \right\}. \quad (2.144)$$

We restrict the supremum to measurable indicator functions and obtain:

$$\|\text{id}\| \geq \sup_{\chi_A} \frac{\|\chi_A\|_{\mathcal{R}_2}}{\|\chi_A\|_{\mathcal{R}_1}} = \sup_{\alpha \rightarrow 0} \sup_{\phi_{2\zeta} \in \mathcal{Z}_2} \frac{\phi_{2\zeta}(\alpha)}{\sup_{\phi_{1\gamma} \in \mathcal{Z}_1} \phi_{1\gamma}(\alpha)} = \sup_{\alpha \rightarrow 0} \sup_{\phi_{2\zeta} \in \mathcal{Z}_2} \inf_{\phi_{1\gamma} \in \mathcal{Z}_1} \frac{\phi_{2\zeta}(\alpha)}{\phi_{1\gamma}(\alpha)} = C' = \infty. \quad (2.145)$$

Since $\|\text{id}\|$ is unbounded, the norms are not equivalent. ■

Note, however, that this criterion is not useful to test for non-equivalence of two norms with the same fundamental function, since in this case $C' = 1$.

Throughout this section, we focus on those fundamental functions with $\phi(0+) = 0$ since otherwise both the Marcinkiewicz M_ϕ and the Lorentz space Λ_ϕ are equal to $\mathcal{L}^\infty$ [Rubshstein et al., 2016, p. 164]. We now give various characterization results in terms of $\phi'(0)$. To intuitively understand why this particular value is of interest, consider the Lorentz norm ($\phi \in \Phi_{0+}$):

$$\|X\|_{\Lambda_\phi} = \int_0^1 X^*(\omega) \phi'(\omega) \, d\omega, \quad (2.146)$$

and recall that these are the basic building blocks of any ri norm (Section 2.4.7). Since X^* are the backwards quantiles, $\phi'(0)$ is the highest weight which the most extreme loss receives. Due to concavity of ϕ , its derivative ϕ' is nonnegative and decreasing. Risk measures fundamentally differ with respect to their tail behaviour: for instance, the expectation is maximally *insensitive* to tails, as all quantiles receive constant weight 1. On the other hand, for CVar_α we have $\phi'(0) = 1/(1-\alpha)$, which for $\alpha \rightarrow 1$ may grow arbitrarily large. In general, the most benign situation occurs when $\phi'(0)$ is finite.

Theorem 57. *Let $\phi \in \Phi_{0+}$. If the derivative of ϕ at 0, i.e. $\phi'(0)$, is finite, then there exists a constant K such that*

$$\|X\|_{\Lambda_\phi} \leq K \cdot \|X\|_{M_\phi} \quad \forall X \in M_\phi. \quad (2.147)$$

In view of the embedding theorem, we then have $\Lambda_\phi = M_\phi$, i.e. equivalence of the Marcinkiewicz

and the Lorentz norm. This implies in particular that given such a fundamental function, all law invariant coherent risk measures are equivalent. Moreover, a feasible constant is $K = 1/(\phi(\frac{1}{\phi'(0)}))$.

The proof is in Appendix 2.8.4. Depending on the value of $\phi'(0)$, the constant K can be relatively small: as an example, for the fundamental function $\phi(t) = 1 - (1 - t)^2$ of the Dutch risk measure and MaxVar, $\phi'(0) = 2$, the constant is only $K = \frac{4}{3}$, implying that

$$\|X\|_{M_\phi} \leq \text{Du}(|X|) \leq \text{MaxV}(|X|) \leq \frac{4}{3} \|X\|_{M_\phi} \quad \forall X \in M_\phi. \quad (2.148)$$

The smallest K , however is achieved for CVar_α : $K = 1/\phi(\frac{1}{\phi'(0)}) = 1 \quad \forall \alpha \in [0, 1)$, a sanity check for Theorem 52.

Remark 58. Assume R_1 and R_2 are coherent risk measures. Then $R_1(X) \leq K \cdot R_2(X) \quad \forall X \in \mathcal{M}$ implies $K = 1$ necessarily due to translation equivariance [Pichler, 2017]. Therefore, to obtain interesting and useful comparisons, we must restrict ourselves to the positive cone $\mathcal{M}^+$. Working with the norm $\|\cdot\|$ instead of the function norm has this effect.

Theorem 59. Let $\phi \in \Phi_{0+}$ with $\phi'(0) = \infty$. Then the Marcinkiewicz and the Lorentz norm are not equivalent. We have

$$\|X\|_{M_\phi} \leq \|X\|_{\Lambda_\phi} \quad \forall X \in \Lambda_\phi \quad \text{but} \quad \nexists K : \|X\|_{\Lambda_\phi} \leq K \cdot \|X\|_{M_\phi} \quad \forall X \in M_\phi. \quad (2.149)$$

The proof is in Appendix 2.8.4.

Theorem 60. Given any two ri norms $\|\cdot\|_{\mathcal{R}_1}, \|\cdot\|_{\mathcal{R}_2}$ with possibly different fundamental functions $\phi_1, \phi_2 \in \Phi_{0+}$. If $\phi'_1(0)$ and $\phi'_2(0)$ are finite, then the norms are equivalent. In particular, all such norms are equivalent to the $\mathcal{L}^1$ norm, i.e. the expectation of a nonnegative random variable.

The proof is in Appendix 2.8.4.

Corollary 61. Given any two ri norms $\|\cdot\|_{\mathcal{R}_1}, \|\cdot\|_{\mathcal{R}_2}$ with possibly different fundamental functions $\phi_1, \phi_2 \in \Phi_{0+}$. If $\phi'_1(0) = \infty$ but $\phi'_2(0)$ is finite, then they are not equivalent.

Proof We use the following result from [Rubshtein et al., 2016, p. 164]: If $\phi'_1(0) = \infty$ then $M_{\phi_1} \subsetneq \mathcal{L}^1$. On the other hand, we have shown before that all norms with $\phi'_2(0)$ finite are equivalent to $\mathcal{L}^1$. Altogether, using the embedding theorem, we have

$$\mathcal{R}_1 \subseteq M_{\phi_1} \subsetneq \mathcal{L}^1 = \mathcal{R}_2. \quad (2.150)$$

If the spaces do not coincide, the norms cannot be equivalent [Bennett and Sharpley, 1988, p. 7].

■

Theorem 62. *Given any $\phi \in \Phi_{0+}$ with $\phi'(0) = \infty$. Then the Marcinkiewicz norm $\|\cdot\|_{M_\phi}$ is equivalent to the positive translation equivariant Marcinkiewicz norm $\|\cdot\|_{TM_\phi}$, even though $C = \infty$.*

The proof is in Appendix 2.8.4.

Corollary 63. *Let $\phi \in \Phi_{0+}$ with $\phi'(0) = \infty$. In view of the embedding theorem, Theorem 59 and Theorem 62, we have the embeddings:*

$$\Lambda_\phi \subsetneq TM_\phi = M_\phi \quad (2.151)$$

for the spaces induced by the Lorentz, PTE Marcinkiewicz and the Marcinkiewicz norm. This means that if $\phi'(0) = \infty$ not all coherent risk measures are equivalent; however, the smallest ri norm is equivalent to the smallest coherent risk measure. On the other hand, we have shown that if $\phi'(0) < \infty$, all of these are equivalent.

We have seen that $\phi'(0)$ plays an important role. If $\phi'(0)$ is bounded, such as for the fundamental function of the Dutch risk measure and MaxVar, the space of compatible law invariant coherent risk measures is very “small”: from a theoretical perspective, they are all equivalent. We may summarize the role of $\phi'(0)$ by stating that *it’s all about the tails*. On a coarse level, $\phi(0+)$ controls how much weight is given to the most extreme event (the supremum), hence a risk measure with $\phi(0+) > 0$ mimics the supremum (or if $\phi(0+) = 1$, it is the supremum). On a more fine grained level, $\phi'(0)$ is the weight that the extreme tails receive in the Lorentz norm. For an arbitrary ri norm, the interpretation of $\phi'(0)$ is more subtle due to the involved supremum in the Kusuoka representation.

2.4.10 REARRANGEMENT INVARIANT NORMS AND RISK

We have seen that, on the positive cone, a law invariant coherent risk measure can be seen as a rearrangement invariant Banach function norm with the additional property of positive translation equivariance. Philosophically, this implies agreement about a base probability measure; however, decision makers may disagree about their risk aversion attitudes or they might want to introduce ‘hallucinated’ ambiguity to account for a degree of distrust in the base measure. This is the specification of a fundamental function, a coherent upper probability. After a fundamental function is specified, there exist in general many different compatible norms. Among them, the Marcinkiewicz norm and the positive translation equivariant Marcinkiewicz norm are distinguished as the most optimistic extensions (subject to a constraint of requiring positive translation equivariance or not). Diametrically opposed, the Lorentz norms (spectral risk measures) are the most pessimistic extensions. In virtue of the Kusuoka representation, any other ri norm can be understood as being

formed from an ambiguity set over spectral risk measures. Hence ambiguity about risk aversion exhausts the whole space of coherent risk measures. When there is no specific motivation for such a construction, the Lorentz norm is however the natural extension (indeed, also in Walley’s terms) of the fundamental function: it is the only ri norm with a singleton Kusuoka set (if $\phi \in \Phi_{0+}$) and therefore fully described by its risk aversion profile.

We remark that ri function norms, which are not positive translation equivariant, are candidates for regret measures in the risk quadrangle [Rockafellar and Uryasev, 2013]. First, the ri function norm needs to be extended to the whole space, including potentially negative functions. Given an arbitrary ri norm V , a similar extension to Equation 2.101 can be constructed:

$$V^-(X) := \sup \left\{ \int_0^1 X^{*-}(\omega) Y^*(\omega) \, d\omega : V'(Y) \leq 1, Y \in \mathcal{M}^+ \right\} \quad \forall X \in \mathcal{M} \quad (2.152)$$

It is easy to check that this fulfills the desiderata of a coherent regret measure except perhaps aversity²². The corresponding coherent risk measure can then be obtained by infimal convolution (Theorem 31). We believe that this opens up room for future research concerning risk-averse regression in the quadrangle, where the fundamental function offers fine control over the degree and shape of risk aversion.

2.5 CREATING NEW RISK MEASURES FROM OLD

In this section we investigate how one can combine several risk measures (ri function norms) to create new ones. Our motivation is two-fold. First, one can develop a better understanding of a “thing” by understanding the various transformations of the thing — a heuristic known as Grothendieck’s relative method. We shall see, for example, that by considering the result of combining two risk measures reinforces the importance of the fundamental function. Second, risk measures can be not only used to encode risk aversion attitudes, but also fairness requirements [Williamson and Menon, 2019], and since people will sometimes disagree on the right notion of fairness for a given situation (fairness being a prototypical example of an “essentially contested concept” [Gallie, 1955]), a means is needed to reach a compromise between two distinct views on fairness, as codified by choices of risk measures. The same argument applies sans fairness where two people have different risk aversion attitudes.

²²However, a “weak aversity” condition $V^-(X) \geq \mathbb{E}[X]$ follows from the embedding theorem (essentially from law invariance). Note that we presupposed $R(1_\Omega) = 1$ for any ri norm R . Many regret measures will not satisfy this. Monotonicity holds since $Y \in \mathcal{M}^+$, cf. Section 2.3.2. Positive homogeneity and subadditivity are easily checked.

2.5.1 PROPERTIES OF QUASICONCAVE FUNCTIONS

We will first present some elementary results concerning quasiconcave functions and the effect of various combinations of ri Banach function norms on the corresponding fundamental functions.

Lemma 64. *Suppose $\phi: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ and $\phi(1) = 1$. If ϕ is quasiconcave then*

$$\forall t \geq 0, \quad 1 \wedge t \leq \phi(t) \leq 1 \vee t. \quad (2.153)$$

The proof is in Appendix 2.8.5.

Lemma 65. *[Rubshtein et al., 2016, p. 127]. The function $\phi: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ is quasiconcave if and only if $t \mapsto t/\phi(t)$ is quasiconcave.*

Lemma 65 has a natural interpretation in terms of fundamental functions:

Lemma 66. *[Kreĭn et al., 1982a, p. 106]. If ϕ is the fundamental function of an ri space $\mathcal{X}$, then $t \mapsto t/\phi(t)$ is the fundamental function of the associate space $\mathcal{X}'$.*

Quasiconcavity is preserved under pointwise minima and maxima:

Lemma 67. *Suppose $\phi_i: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ are quasiconcave, $i \in [n]$. Then $\bigwedge_{i \in [n]} \phi_i$ and $\bigvee_{i \in [n]} \phi_i$ are quasiconcave.*

The proof is in Appendix 2.8.5. Lemma 67 suggests the question as to what other combinations of quasiconcave functions are guaranteed to be quasiconcave. We now show that quasiconcavity is preserved under a range of binary operations induced by another quasiconcave function.

Definition 68. *Suppose $\psi: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ is quasiconcave and $\psi(1) = 1$. The perspective of ψ is the function*

$$\check{\psi}: \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \ni (x, y) \mapsto y\psi(x/y).$$

Let $\mathcal{P}$ denote the set of functions $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ which are positively homogeneous, non-zero (except at $(0, 0)$) and nondecreasing (in both arguments), and let $\mathcal{Q}$ denote the set of quasiconcave functions on $\mathbb{R}_{\geq 0}$.

Lemma 69. *Suppose $\psi: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$. The perspective $\check{\psi} \in \mathcal{P}$ if and only if $\psi \in \mathcal{Q}$.*

The proof is in Appendix 2.8.5. The following lemma shows that combining two quasiconcave functions using $\check{\psi}$ is guaranteed to result in a quasiconcave function, and that this is the only way to ensure such a preservation of quasiconcavity.

Lemma 70. Suppose $\phi_0, \phi_1, \psi: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$, and let $f_{\phi_0, \phi_1}(t) := \check{\psi}(\phi_0(t), \phi_1(t))$, $t \geq 0$. Then $[f_{\phi_0, \phi_1} \in \mathcal{Q}, \forall \phi_1, \phi_2 \in \mathcal{Q}]$ if and only if $\psi \in \mathcal{Q}$.

The proof is in Appendix 2.8.5. Observe that if $\psi(1) = 1$, and $\phi_1 = \phi_0$, then

$$\check{\psi}(\phi_0(t), \phi_1(t)) = \phi_0(t)\psi(\phi_0(t)/\phi_0(t)) = \phi_0(t). \quad (2.154)$$

2.5.2 INTERPOLATION SPACES

The creation of new ri norms from given norms can be viewed as the construction of an “interpolation space” [Bennett and Sharpley, 1988, pp. 99ff]. Given two ri spaces $\mathcal{X}_0$ and $\mathcal{X}_1$, embedded in some separable linear topological space, let $\Delta(\mathcal{X}_0, \mathcal{X}_1) := \mathcal{X}_0 \cap \mathcal{X}_1$ and $\Sigma(\mathcal{X}_0, \mathcal{X}_1) := \mathcal{X}_0 + \mathcal{X}_1$ with the corresponding norms

$$\|f\|_{\Delta(\mathcal{X}_0, \mathcal{X}_1)} := \|f\|_{\mathcal{X}_0} \vee \|f\|_{\mathcal{X}_1} \quad (2.155)$$

$$\|f\|_{\Sigma(\mathcal{X}_0, \mathcal{X}_1)} := \inf\{\|f_0\|_{\mathcal{X}_0} + \|f_1\|_{\mathcal{X}_1} : f = f_0 + f_1, f_0 \in \mathcal{X}_0, f_1 \in \mathcal{X}_1\}. \quad (2.156)$$

The spaces $\Delta(\mathcal{X}_0, \mathcal{X}_1)$ (resp. $\Sigma(\mathcal{X}_0, \mathcal{X}_1)$) are the smallest (resp. largest) intermediate spaces between $\mathcal{X}_0$ and $\mathcal{X}_1$ in the sense that any *intermediate space* $\mathcal{X}$ is continuously embedded between them:

$$\Delta(\mathcal{X}_0, \mathcal{X}_1) \hookrightarrow \mathcal{X} \hookrightarrow \Sigma(\mathcal{X}_0, \mathcal{X}_1).$$

(This serves as a definition of intermediate space). If 1 is a feasible embedding constant, which we notate by $\xrightarrow{1}$, and which can always be ensured by simple scaling, for any intermediate space $\mathcal{X}$, for all $f \in \mathcal{X}_0 + \mathcal{X}_1$,

$$\|f\|_{\Sigma(\mathcal{X}_0, \mathcal{X}_1)} \leq \|f\|_{\mathcal{X}} \leq \|f\|_{\Delta(\mathcal{X}_0, \mathcal{X}_1)}.$$

In order to appeal to results in the literature, we need to make some assumptions regarding the measure spaces (Ω, μ) upon which our ri spaces are defined. We can restrict ourselves to finite measures spaces $\mu(\Omega) < \infty$, and in fact will assume $\mu(\Omega) = 1$; for example, $\Omega = [0, 1]$ with the Lebesgue measure as in Section 2.4. All of the results below then hold for any measure space that is purely non-atomic, or completely atomic which all atoms having equal measure. (This is a consequence of [Bennett and Sharpley, 1988, Theorem II.2.7].) Recall that we denote the associate space of $\mathcal{X}$ as $\mathcal{X}'$.

Lemma 71. [Bergh and Löfström, 1976, Theorem 2.7.1]. Suppose $\mathcal{X}_0$ and $\mathcal{X}_1$ are ri spaces and $\Delta(\mathcal{X}_0, \mathcal{X}_1)$ is dense in both $\mathcal{X}_0$ and $\mathcal{X}_1$. Then $\Delta(\mathcal{X}_0, \mathcal{X}_1)' = \Sigma(\mathcal{X}_0', \mathcal{X}_1')$ and $\Sigma(\mathcal{X}_0, \mathcal{X}_1)' = \Delta(\mathcal{X}_0', \mathcal{X}_1')$.

We subsequently have the following analog of Lemma 67 in terms of fundamental func-

tions.

Lemma 72. Suppose $\mathcal{X}_0$ and $\mathcal{X}_1$ are ri spaces over a measure space (Ω, μ) with corresponding fundamental functions $\phi_{\mathcal{X}_0}$ and $\phi_{\mathcal{X}_1}$. Then $\phi_{\Delta(\mathcal{X}_0, \mathcal{X}_1)} = \phi_{\mathcal{X}_0} \vee \phi_{\mathcal{X}_1}$ and $\phi_{\Sigma(\mathcal{X}_0, \mathcal{X}_1)} = \phi_{\mathcal{X}_0} \wedge \phi_{\mathcal{X}_1}$.

Proof For $t \geq 0$, let $E_t \subset \Omega$ be such that $\mu(E_t) = t$. Equation 2.155 implies that for all $t \geq 0$,

$$\phi_{\Delta(\mathcal{X}_0, \mathcal{X}_1)}(t) = \|\chi_{E_t}\|_{\Delta(\mathcal{X}_0, \mathcal{X}_1)} = \|\chi_{E_t}\|_{\mathcal{X}_0} \vee \|\chi_{E_t}\|_{\mathcal{X}_1} = \phi_{\mathcal{X}_0}(t) \vee \phi_{\mathcal{X}_1}(t).$$

Lemmas 66 and 71 together imply that for all $t \geq 0$,

$$\phi_{\Sigma(\mathcal{X}_0, \mathcal{X}_1)}(t) = \frac{t}{\phi_{\Sigma(\mathcal{X}_0, \mathcal{X}_1)'}(t)} = \frac{t}{\phi_{\Delta(\mathcal{X}_0', \mathcal{X}_1')}(t)} = \frac{t}{\phi_{\mathcal{X}_0'}(t) \vee \phi_{\mathcal{X}_1'}(t)} \quad (2.157)$$

$$= \frac{t}{\frac{t}{\phi_{\mathcal{X}_0}(t)} \vee \frac{t}{\phi_{\mathcal{X}_1}(t)}} = \phi_{\mathcal{X}_0}(t) \wedge \phi_{\mathcal{X}_1}(t). \quad (2.158)$$

■

Since $\check{\psi}(x, 1) = \psi(x)$, we see that the ψ functions from Lemma 70 corresponding to max and min are $\psi_{\max}(x) = \max(x, 1)$ and $\psi_{\min}(x) = \min(x, 1)$, which are indeed both quasi-concave.

2.5.3 INTERPOLATION FUNCTORS AND THEIR FUNDAMENTAL FUNCTIONS

We make use of a number of definitions and results of Brudnyi et al. [1986]. Given the pair $(\mathcal{X}_0, \mathcal{X}_1)$ and an intermediate space $\mathcal{X}$ for this pair, the triple $((\mathcal{X}_0, \mathcal{X}_1); \mathcal{X})$ is called an *interpolation triple*. The triple $((\mathcal{X}_0, \mathcal{X}_1); \mathcal{X})$ is called an *interpolation triple relative to the triple* $((\mathcal{Y}_0, \mathcal{Y}_1); \mathcal{Y})$ if any bounded linear operator from the pair $(\mathcal{X}_0, \mathcal{X}_1)$ to $(\mathcal{Y}_0, \mathcal{Y}_1)$ maps $\mathcal{X}$ into $\mathcal{Y}$. When that occurs, there exists $c > 0$ such that for any linear operator $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $\|T\|_{\mathcal{X} \rightarrow \mathcal{Y}} \leq c \|T\|_{(\mathcal{X}_0, \mathcal{X}_1) \rightarrow (\mathcal{Y}_0, \mathcal{Y}_1)}$, where the operator norm $\|T\|_{\mathcal{X} \rightarrow \mathcal{Y}} = \sup\{\|TX\|_{\mathcal{Y}} : X \in \mathcal{X} \text{ and } \|X\|_{\mathcal{X}} \leq 1\}$. If $c \leq 1$ then $((\mathcal{X}_0, \mathcal{X}_1); \mathcal{X})$ is called a *normal interpolation triple relative to the triple* $((\mathcal{Y}_0, \mathcal{Y}_1); \mathcal{Y})$. Let $\mathfrak{B}$ denote the category of Banach spaces and $\mathfrak{B}$ denote the category of Banach pairs (for what follows it suffices to just consider these as sets).

Definition 73. An interpolation functor is a functor $\mathcal{F} : \mathfrak{B} \rightarrow \mathfrak{B}$ which assigns to each Banach pair $\vec{\mathcal{X}} = (\mathcal{X}_0, \mathcal{X}_1)$ a Banach space $\mathcal{F}(\vec{\mathcal{X}})$ intermediate between $\mathcal{X}_0$ and $\mathcal{X}_1$, and to each operator $T \in \mathcal{L}(\vec{\mathcal{X}}, \vec{\mathcal{Y}})$ it assigns the restriction to the space $\mathcal{F}(\vec{\mathcal{X}})$.

The triples $(\vec{\mathcal{X}}; \mathcal{F}(\vec{\mathcal{X}}))$ and $(\vec{\mathcal{Y}}; \mathcal{F}(\vec{\mathcal{Y}}))$ are interpolation triples relative to each other. If for any pairs $\vec{\mathcal{X}}$ and $\vec{\mathcal{Y}}$ the resulting triples are normalised then $\mathcal{F}$ is said to be a *normalised interpolation functor*. The functors Δ and Σ introduced in (2.155) and (2.156) are both normalised interpolation functors.

For $\alpha > 0$, the space $\alpha\mathbb{R}$ is the set $\mathbb{R}$ along with norm given by $\|x\|_{\alpha\mathbb{R}} = \alpha|x|$, for $x \in \mathbb{R}$. Suppose $\alpha, \beta > 0$. Given an interpolation functor $\mathcal{F}$, if we apply it to the Banach pair $(\alpha\mathbb{R}, \beta\mathbb{R})$ we obtain $\mathcal{F}(\alpha\mathbb{R}, \beta\mathbb{R}) = \phi_{\mathcal{F}}(\alpha, \beta)\mathbb{R}$, where the constant $\phi_{\mathcal{F}}(\alpha, \beta)$ is known as the *fundamental function of the functor $\mathcal{F}$* [Brudnyi and Krugljak, 1991]. (Sometimes $\phi_{\mathcal{F}}$ is called the *characteristic function of the functor $\mathcal{F}$* [Brudnyi et al., 1986], but such terminology conflicts with the characteristic function χ_E of a set E which we make considerable use of.) For any functor $\mathcal{F}$, $(\alpha, \beta) \mapsto \phi_{\mathcal{F}}(\alpha, \beta)$ is positive, positively homogeneous, and nondecreasing in each argument. The *dual fundamental function of the functor $\mathcal{F}$* is given by $\phi_{\mathcal{F}}^*(\alpha, \beta) = \frac{1}{\phi_{\mathcal{F}}(1/\alpha, 1/\beta)}$. If $\mathcal{F}$ is normalised, $\phi_{\mathcal{F}}(1, 1) = 1$. Normalised interpolation functors, when restricted to the Banach pair $(\alpha\mathbb{R}, \beta\mathbb{R})$ are characterised by their fundamental function $\phi_{\mathcal{F}}$.

Given a Banach pair $\vec{\mathcal{X}} = (\mathcal{X}_0, \mathcal{X}_1)$, the *K-functional* is defined as

$$K(s_0, s_1, X, \vec{\mathcal{X}}) := \inf\{s_0\|X_0\|_{\mathcal{X}_0} + s_1\|X_1\|_{\mathcal{X}_1} : X_0 \in \mathcal{X}_0, X_1 \in \mathcal{X}_1 \text{ s.t. } X = X_0 + X_1\}, \quad s_0, s_1 \geq 0$$

Pick an arbitrary function $\phi \in \mathcal{P}$ (recall definition 68), and let $\mathcal{F}$ denote the interpolation functor on $(\alpha\mathbb{R}, \beta\mathbb{R})$ with fundamental function $\phi_{\mathcal{F}} = \phi$. On one dimensional spaces, the interpolation functor is entirely determined by its fundamental function; taking ϕ as given, then $\mathcal{F}\phi$ is given as $\mathcal{F}(\alpha\mathbb{R}, \beta\mathbb{R}) = \phi(\alpha, \beta)$. If one defines an interpolation functor on one dimensional spaces, then it can be extended in many ways to arbitrary pairs of spaces. It turns out [Brudnyi et al., 1986, Section 1.16] that there is a lower $\underline{\mathcal{F}}$ and upper extension $\overline{\mathcal{F}}$ such that for all pairs of Banach spaces $\vec{\mathcal{X}} = (\mathcal{X}_0, \mathcal{X}_1)$ and all interpolation functors $\mathcal{F}$,

$$\underline{\mathcal{F}}(\vec{\mathcal{X}}) \xhookrightarrow{1} \mathcal{F}(\vec{\mathcal{X}}) \xhookrightarrow{1} \overline{\mathcal{F}}(\vec{\mathcal{X}}).$$

The lower and upper extensions are characterised by the following [Brudnyi et al., 1986, Section 1.17]:

Lemma 74. *Let $\vec{\mathcal{X}} = (\mathcal{X}_0, \mathcal{X}_1)$ be an arbitrary Banach pair. The lower and upper extensions $\underline{\mathcal{F}}(\vec{\mathcal{X}}) = \Lambda_{\phi}(\vec{\mathcal{X}})$ and $\overline{\mathcal{F}}(\vec{\mathcal{X}}) = M_{\phi}(\vec{\mathcal{X}})$ correspond to the space of all elements of $\mathcal{X}_0 + \mathcal{X}_1$ with (respectively) finite norms*

$$\|X\|_{\Lambda_{\phi}(\vec{\mathcal{X}})} := \inf \sum_k \phi(\|X_k\|_{\mathcal{X}_0}, \|X_k\|_{\mathcal{X}_1}), \quad (2.159)$$

where the infimum is taken over all representations of X of the form $X = \sum_k X_k$, with $X_k \in \mathcal{X}_0 + \mathcal{X}_1$ for all k ; and

$$\|X\|_{M_{\phi}(\vec{\mathcal{X}})} := \sup_{s_0, s_1} \frac{K(s_0, s_1, X, \vec{\mathcal{X}})}{\phi^*(s_0, s_1)}. \quad (2.160)$$

The spaces $\Lambda_{\phi}(\vec{\mathcal{X}})$ and $M_{\phi}(\vec{\mathcal{X}})$ are called the *abstract Lorentz space* and *abstract Marcinkie-*

wicz space respectively²³, and the functors Λ_ϕ and M_ϕ are known as the *lower and upper extensions* of the functor $\mathcal{F}_\phi$ defined on $(\alpha\mathbb{R}, \beta\mathbb{R})$ in terms of the fundamental function $\phi \in \mathcal{P}$ because they bound the behaviour of interpolation functors with a given fundamental function:

Lemma 75. [*Brudnyi et al., 1986, Section 1.17*]. $\bar{\mathcal{X}} = (\mathcal{X}_0, \mathcal{X}_1)$ be an arbitrary pair of Banach spaces. Let $\mathcal{X}$ be a normal interpolation space between $\mathcal{X}_0$ and $\mathcal{X}_1$, and let $\mathcal{F}$ be some normalised interpolation functor for which $\mathcal{F}(\bar{\mathcal{X}}) = \mathcal{X}$ with fundamental function ϕ . Then

$$\Lambda_\phi(\bar{\mathcal{X}}) \xhookrightarrow{1} \mathcal{X} \xhookrightarrow{1} M_\phi(\bar{\mathcal{X}}), \quad (2.161)$$

and thus

$$\|X\|_{M_\phi(\bar{\mathcal{X}})} \leq \|X\|_{\mathcal{X}} \leq \|X\|_{\Lambda_\phi(\bar{\mathcal{X}})}. \quad (2.162)$$

Recall that the Lorentz space Λ_ϕ is always positive translation equivariant (PTE) (Example 42). We might thus conjecture that the interpolation functor $\Lambda_\phi: \bar{\mathfrak{B}} \rightarrow \mathfrak{B}$ would preserve PTE; indeed that is the case as the lemma below shows.

Lemma 76. Suppose $\mathcal{X}_0, \mathcal{X}_1$ are PTE, and $\phi \in \mathcal{Q}$, then $\Lambda_\phi(\mathcal{X}_0, \mathcal{X}_1)$ is PTE.

The proof is in Appendix 2.8.6. Lemma 72 implies that if ϕ is the fundamental function of $\mathcal{X}$, an intermediate space between $\mathcal{X}_0$ and $\mathcal{X}_1$, we have for all $t \geq 0$,

$$\phi_0(t) \wedge \phi_1(t) \leq \phi(t) \leq \phi_0(t) \vee \phi_1(t). \quad (2.163)$$

Observe that if $\phi_0 = \phi_1$, this means if $\mathcal{X}$ is an intermediate space between $\mathcal{X}_0$ and $\mathcal{X}_1$ with fundamental function ϕ , then $\phi_0 \leq \phi \leq \phi_0$ and hence $\phi = \phi_0$. We formalise this observation as follows.

²³That these define norms is obvious enough except perhaps for the convexity of $\|\cdot\|_{\Lambda_\phi(\bar{\mathcal{X}})}$. Since ϕ is positively homogeneous and thus obviously $\|\cdot\|_{\Lambda_\phi(\bar{\mathcal{X}})}$ is, it suffices to demonstrate subadditivity, namely that (writing $Z = X + Y$),

$$\begin{aligned} & \|X\|_{\Lambda_\phi(\bar{\mathcal{X}})} + \|Y\|_{\Lambda_\phi(\bar{\mathcal{X}})} \geq \|Z\|_{\Lambda_\phi(\bar{\mathcal{X}})} \\ \Leftrightarrow & \inf_{\sum_{k_1} X_{k_1} = X} \sum_{k_1} \phi(\|X_{k_1}\|_{\mathcal{X}_0}, \|X_{k_1}\|_{\mathcal{X}_1}) + \inf_{\sum_{k_2} Y_{k_2} = Y} \sum_{k_2} \phi(\|Y_{k_2}\|_{\mathcal{X}_0}, \|Y_{k_2}\|_{\mathcal{X}_1}) \geq \inf_{\sum_k Z_k = Z} \sum_k \phi(\|Z_k\|_{\mathcal{X}_0}, \|Z_k\|_{\mathcal{X}_1}) \\ \Leftrightarrow & \inf_{\substack{\sum_{k_1} X_{k_1} = X \\ \sum_{k_2} Y_{k_2} = Y}} \left(\sum_{k_1} \phi(\|X_{k_1}\|_{\mathcal{X}_0}, \|X_{k_1}\|_{\mathcal{X}_1}) + \sum_{k_2} \phi(\|Y_{k_2}\|_{\mathcal{X}_0}, \|Y_{k_2}\|_{\mathcal{X}_1}) \right) \geq \inf_{\sum_k Z_k = Z} \sum_k \phi(\|Z_k\|_{\mathcal{X}_0}, \|Z_k\|_{\mathcal{X}_1}), \end{aligned}$$

which holds since $\sum_{k_1} X_{k_1} + \sum_{k_2} Y_{k_2} = X + Y = Z$ and the infimum on the left is taken over a smaller set since the X_{k_1} s have to sum to X and separately the Y_{k_2} s have to sum to Y , but on the right this choice is also available plus additional ones where no subset of the Z_k are constrained to sum to X , and thus its infimum is less than or equal to that on the left.

Lemma 77. Suppose $\bar{\mathcal{X}} = (\mathcal{X}_0, \mathcal{X}_1)$ is an arbitrary pair of ri spaces, that $\mathcal{X}_0$ and $\mathcal{X}_1$ have the same fundamental function ϕ , and that $\mathcal{X}$ is an intermediate space of $\bar{\mathcal{X}}$ with feasible embedding constant 1: $\mathcal{X}_0 \xrightarrow{1} \mathcal{X} \xrightarrow{1} \mathcal{X}_1$. Then $\phi_{\mathcal{X}} = \phi$.

Proof The embedding ensures that for all $X \in \mathcal{X}$, we have $\|X\|_{\mathcal{X}_1} \leq \|X\|_{\mathcal{X}} \leq \|X\|_{\mathcal{X}_0}$ and thus choosing $X = \chi_{[0,t]}$ for some arbitrary $t > 0$ we obtain

$$\phi(t) = \|\chi_{[0,t]}\|_{\mathcal{X}_1} \leq \|\chi_{[0,t]}\|_{\mathcal{X}} = \phi_{\mathcal{X}}(t) \leq \|\chi_{[0,t]}\|_{\mathcal{X}_0} = \phi(t).$$

Thus $\phi_{\mathcal{X}}(t) = \phi(t)$ for all $t > 0$. ■

This illustrates the significance of the fundamental function and justifies its name: *interpolation between two spaces with the same fundamental function does not change the fundamental function*. Thus the fundamental function provides a natural stratification of all possible ri spaces and their associated norms. When the fundamental functions of $\mathcal{X}_0$ and $\mathcal{X}_1$ differ, Lemma 72 shows a simple functional dependence of the fundamental functions of $\Sigma(\mathcal{X}_0, \mathcal{X}_1)$ and $\Delta(\mathcal{X}_0, \mathcal{X}_1)$ on the fundamental functions of $\mathcal{X}_0$ and $\mathcal{X}_1$. We now develop a general result along these lines that appears to be new. We need some additional lemmas first.

Lemma 78. Suppose $\bar{\mathcal{X}} = (\mathcal{X}_0, \mathcal{X}_1)$ is an arbitrary pair of ri spaces over a measure space (Ω, μ) , with corresponding fundamental functions ϕ_0 and ϕ_1 . Let $\bar{\phi} \in \mathcal{P}$. Then the fundamental function of $M_{\bar{\phi}}(\bar{\mathcal{X}})$ satisfies

$$\phi_{M_{\bar{\phi}}(\bar{\mathcal{X}})}(t) = \bar{\phi}(\phi_0(t), \phi_1(t)), \quad t > 0.$$

The proof is in Appendix 2.8.6. We have an analogous (one-sided) result for $\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})$:

Lemma 79. Suppose $\bar{\mathcal{X}} = (\mathcal{X}_0, \mathcal{X}_1)$ is an arbitrary pair of ri spaces with corresponding fundamental functions ϕ_0 and ϕ_1 . Let $\bar{\phi} \in \mathcal{P}$. Then the fundamental function of $\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})$ satisfies

$$\phi_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})}(t) \leq \bar{\phi}(\phi_0(t), \phi_1(t)), \quad t > 0.$$

The proof is in Appendix 2.8.6. Lemmas 74, 78, and 79 combined with (2.162) from Lemma 75 imply that for all $t > 0$,

$$\bar{\phi}(\phi_0(t), \phi_1(t)) = \phi_{M_{\bar{\phi}}(\bar{\mathcal{X}})} = \|\chi_{E_t}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} \leq \|\chi_{E_t}\|_{\mathcal{X}} = \phi_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})}(t) \leq \|\chi_{E_t}\|_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})} \leq \bar{\phi}(\phi_0(t), \phi_1(t)), \quad (2.164)$$

and thus by Lemma 77, we have for all $t > 0$,

$$\phi_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})}(t) = \bar{\phi}(\phi_0(t), \phi_1(t)). \quad (2.165)$$

We have thus proved:

Theorem 80. Suppose $(\mathcal{X}_0, \mathcal{X}_1)$ is an arbitrary pair of ri spaces and that the fundamental functions of $\mathcal{X}_0$ and $\mathcal{X}_1$ are ϕ_0 and ϕ_1 respectively. Suppose $\mathcal{F}$ is a normalized interpolation functor with fundamental function $\phi_{\mathcal{F}}$. Then $\mathcal{F}(\mathcal{X}_0, \mathcal{X}_1)$ is an intermediate space for $(\mathcal{X}_0, \mathcal{X}_1)$ and its fundamental function satisfies

$$\phi_{\mathcal{F}(\mathcal{X}_0, \mathcal{X}_1)} = \phi_{\mathcal{F}}(\phi_0, \phi_1). \quad (2.166)$$

2.5.4 IMPLICATIONS AND EXAMPLES

The “fundamental function” $\phi_{\mathcal{X}}$ of an ri space $\mathcal{X}$ is justified in name from a mathematical viewpoint, as a risk aversion profile, and from an ethical perspective. Mathematically, we have seen that combining two ri norms with the same fundamental function will always result in another norm with the same fundamental function. Thus the function $\phi_{\mathcal{X}}$ really does pick out something fundamental. From the risk aversion perspective, it captures the broad brush features of a decision maker’s risk aversion. From an ethical perspective, if we conceive of our function $X: \Omega \rightarrow \mathbb{R}$ as representing some ‘bad’ over a population of people Ω , then the fundamental function $t \mapsto \phi(t)$ of a norm $\|\cdot\|$ captures a coarse aspect of the ethical implications of our choice: it tells us what value we ascribe to assigning a bad of 1 to fraction t of the population and a no bad to the rest. (Recall we are consistently adopting the loss perspective in this paper, where larger values are worse.) The extreme cases of $\|\cdot\|_{\mathcal{L}^1}$ and $\|\cdot\|_{\mathcal{L}^\infty}$ then correspond to the ethical choices of John Harsanyi and John Rawls respectively; see [Williamson and Menon, 2019] for an elaboration of this.

Since the choice of fundamental function is a personal choice (risk aversion, or ethical), different designers will likely make different choices. This begs the question of how a compromise between different choices can be made. An obvious approach is to interpolate between the two choices using an exact interpolation functor. But which functor? The results and arguments above show that the choice of functor has just as wide a scope as the original choice of ri norm. Essentially, the choice in both cases is as large as the set of quasiconcave functions. Thus there is no easy mechanical method to achieve a compromise between two distinct ethical positions (as encoded by two fundamental functions ϕ_1 and ϕ_2) because the result of interpolation between the two norms is dependent upon the choice of the interpolation functor, which is stratified by precisely the same class of functions as the original norms²⁴.

This perspective is further strengthened by noting the fact that *every ri space is an exact interpolation space between $\mathcal{L}^1$ and $\mathcal{L}^\infty$* [Bennett and Sharpley, 1988, Theorem III.2.2]. Notwith-

²⁴The perspective developed here can be compared to that of Semmes, who considered geodesics in the Banach space of all Banach spaces [Semmes, 1988] and argued that we should not restrict our thinking about interpolation between normed spaces to the notion of interpolation of operators. In finite dimensional spaces, the question of interpolation can be posed in terms of constructing families of centre symmetric convex bodies (i.e. norm balls) “in-between” two given norm balls. Similarly, the “interpolation” between two or more proper losses is essentially controlled by another proper loss [Williamson and Cranko, 2023].

standing the previous somewhat negative conclusion, the use of interpolation functors to create new ri norms is valuable from a practical and computational perspective in offering a wider choice of explicitly parametrised and easily computable norms. Some examples of special cases of Theorem 80 are given below.

Example 81. When $\phi_{\mathcal{X}_0} = \phi_{\mathcal{X}_1}$, Theorem 80 implies $\phi_{\mathcal{X}}(t) = \phi(\phi_{\mathcal{X}_1}(t), \phi_{\mathcal{X}_1}(t)) = \phi_{\mathcal{X}_1}(t)$ using (2.154) and the fact that $\phi(1, 1) = 1$ since $\mathcal{F}$ is a *normalised* interpolation functor. This agrees with the observation made earlier following (2.163).

Example 82. Consider two interpolation functors from Lemma 72: $\Delta(\mathcal{X}_0, \mathcal{X}_1) = \mathcal{X}_0 \cap \mathcal{X}_1$ and $\Sigma(\mathcal{X}_0, \mathcal{X}_1) = \mathcal{X}_0 + \mathcal{X}_1$. For $\alpha, \beta > 0$, we have $\|X\|_{\Delta(\alpha\mathbb{R}, \beta\mathbb{R})} = \alpha|X| \vee \beta|X| = (\alpha \vee \beta)|X|$ and thus $\phi_{\Delta}(\alpha, \beta) = \alpha \vee \beta$. Similarly, we have $\|X\|_{\Sigma(\alpha\mathbb{R}, \beta\mathbb{R})} = \inf\{\alpha|X_0| + \beta|X_1| : X_0 + X_1 = X\}$. The infimum is achieved by choosing $X_0 = X$ and $X_1 = 0$ when $\alpha \leq \beta$ and $X_0 = 0$ and $X_1 = X$ when $\alpha \geq \beta$, in which case $\|X\|_{\Sigma(\alpha, \beta)} = (\alpha \wedge \beta)|X|$ and thus $\phi_{\Sigma}(\alpha, \beta) = \alpha \wedge \beta$. Both cases correspond to the elementary results of Lemma 72.

Example 83. Simple mean: $\|\cdot\|_{\mathcal{X}_0 \tilde{+} \mathcal{X}_1} := \frac{1}{2}(\|\cdot\|_{\mathcal{X}_0} + \|\cdot\|_{\mathcal{X}_1})$. It is immediate that $\mathcal{X}_0 \tilde{+} \mathcal{X}_1$ is rearrangement invariant if both $\mathcal{X}_0$ and $\mathcal{X}_1$ are. We have $\|f\|_{\mathcal{X}_0 + \mathcal{X}_1} \leq \|f\|_{\mathcal{X}_0 \tilde{+} \mathcal{X}_1} \leq \|f\|_{\mathcal{X}_0 \cap \mathcal{X}_1}$, where the first inequality follows from the fact that the formula for the norm $\|f\|_{\mathcal{X}_0 + \mathcal{X}_1}$ takes the infimum over all additive decompositions $f = f_0 + f_1$ but that for $\|f\|_{\mathcal{X}_0 \tilde{+} \mathcal{X}_1}$ chooses $f_1 = f_2 = f/2$. The second inequality is a consequence of the mean of two numbers being no greater than their maximum. Thus $\mathcal{X}_0 \tilde{+} \mathcal{X}_1$ is an intermediate space between $\mathcal{X}_0$ and $\mathcal{X}_1$. It is immediate then that for all $t \geq 0$, we have

$$\phi_{\mathcal{X}_0 \tilde{+} \mathcal{X}_1}(t) = \frac{1}{2}(\phi_{\mathcal{X}_0}(t) + \phi_{\mathcal{X}_1}(t)).$$

Example 84. More generally, let ρ be a norm on $\mathbb{R}^2$ normalised such that $\rho(1, 1) = 1$, and define $\|\cdot\|_{\rho(\mathcal{X}_0, \mathcal{X}_1)} := \rho(\|\cdot\|_{\mathcal{X}_0}, \|\cdot\|_{\mathcal{X}_1})$. Obviously $\rho(\mathcal{X}_0, \mathcal{X}_1)$ is rearrangement invariant if both $\mathcal{X}_0$ and $\mathcal{X}_1$ are. It is standard that $\alpha + \beta \leq \rho(\alpha, \beta) \leq \alpha \vee \beta$ and it immediately follows from the definition that for all $t \geq 0$,

$$\phi_{\rho(\mathcal{X}_0, \mathcal{X}_1)}(t) = \rho(\phi_{\mathcal{X}_0}(t), \phi_{\mathcal{X}_1}(t)).$$

Example 85. A special case of theorem 80 (albeit stated for the interpolation of N distinct ri spaces, and not just 2) is presented by Cobos and Fernández-Cabrera [2017], who considered a particular family of interpolation functors $\mathcal{F}^\theta$ “of exponent θ ”, $\theta \in (0, 1)$. They showed that for all $t \geq 0$,

$$\phi_{\mathcal{F}^\theta(\mathcal{X}_0, \mathcal{X}_1)}(t) = \phi_{\mathcal{X}_0}^{1-\theta}(t) \phi_{\mathcal{X}_1}^\theta(t),$$

which can be seen to be of the form of (2.166). An analogous result is shown in [Fernández-Cabrera, 2017] for a related but more complex interpolation method. A related result (for

“envelopes,” which are inversely related to fundamental functions [Haroske, 2006, section 3.3]) was presented in Haroske [2007].

Example 86. Consider two Lorentz norms Λ_{ϕ_0} and Λ_{ϕ_1} and an arbitrary interpolation functor $\mathcal{F}$. A natural question to ask is when (if ever) is $\mathcal{F}(\Lambda_{\phi_0}, \Lambda_{\phi_1})$ a Lorentz space Λ_{ϕ} , and when it is, is there some nice formula expressing $\phi = \psi(\phi_0, \phi_1)$? We conjecture that when $\mathcal{F}$ corresponds to the abstract Lorentz norm this is true.

We do know that if $\mathcal{F} = \Lambda_{\bar{\phi}}(\cdot, \cdot)$ then $\phi_{\mathcal{F}} = \bar{\phi}$, since $\Lambda_{\bar{\phi}}$ is the extension of the functor defined on one dimensional spaces with fundamental function $\bar{\phi}$. We also know by (2.165) that $\phi_{\Lambda_{\bar{\phi}}(\mathcal{X}_0, \mathcal{X}_1)} = \bar{\phi}(\phi_0, \phi_1)$, where $\phi_0 = \phi_{\mathcal{X}_0}$ and $\phi_1 = \phi_{\mathcal{X}_1}$. By 37, we thus have for all X ,

$$\|X\|_{\Lambda_{\bar{\phi}}(\mathcal{X}_0, \mathcal{X}_1)} \leq \|X\|_{\Lambda_{\bar{\phi}(\phi_0, \phi_1)}}. \quad (2.167)$$

We conjecture, but do not know, that the above inequality is in fact an equality.

Regardless of whether our conjecture is true, it does suggest a simple means of interpolating between a pair of Lorentz spaces (i.e. spectral risk measures) Λ_{ϕ_0} and Λ_{ϕ_1} by choosing $\psi \in \mathcal{Q}$ and letting $\bar{\phi} = \check{\psi}$ and then constructing the space $\Lambda_{\bar{\phi}(\phi_0, \phi_1)}$ — all one needs to do is to combine the fundamental functions ϕ_0 and ϕ_1 via $\phi(t) = \bar{\phi}(\phi_0(t), \phi_1(t))$. In combining two spectral risk measures in this fashion, one may wish to ensure that $\bar{\phi}$ is symmetric (for equity reasons, so that the order in which the spaces are provided will not affect the outcome). Fortunately this has a simple characterisation. In order that $\check{\psi}(x, y) = \check{\psi}(y, x)$ for all $x, y > 0$, it is necessary and sufficient that $\psi^\circ(t) = \psi(t)$ for all $t \in (0, 1]$, where $\psi^\circ(t) := t\psi(1/t)$ is the Csiszár conjugate of ψ . One can thus readily construct symmetric $\bar{\phi} = \check{\psi}$ by choosing an arbitrary quasiconcave function ψ on $[0, 1]$ and extending it to $[1, \infty)$ via $\psi(t) = t\psi(1/t)$ for $t \geq 1$, the resulting ψ is then guaranteed to satisfy the Csiszár conjugate condition and thus the induced perspective $\check{\psi}$ is guaranteed symmetric.

Observe that in contrast to the method in Example 84, the present method enables the construction of an interpolated norm $\|\cdot\|$ from $\|\cdot\|_{\mathcal{X}_0}$ and $\|\cdot\|_{\mathcal{X}_1}$ that can give finite values to $\|X\|$ even when one of the values of $\|X\|_{\mathcal{X}_0}$ or $\|X\|_{\mathcal{X}_1}$ is infinite (because of the tail behaviour of X).

Interpolation of certain (classical) Lorentz spaces was considered in [Cobos and Martín, 2005, Section 5] and an analogous question for Marcinkiewicz spaces in [Fernández-Cabrera, 2017, Section 5], however the form of results is different to those which we sought here.

Example 87. In order to provide some insight into (2.166), especially for its use in Example 86, in Figure 2.4 we illustrate the interpolation between two given fundamental functions, and show how the choice of the functor (in particular *its* fundamental function) affects the interpolation.

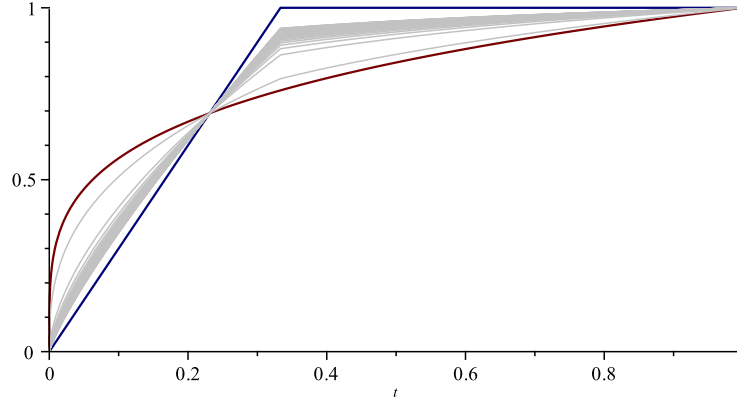


Figure 2.4: Illustration of the interpolation between two quasiconcave fundamental functions. The graph shows $\phi_{\text{red}}(t) = t^{1/4}$ (in red) and $\phi_{\text{blue}}(t) = 3t \wedge 1$ (in blue). The grey curves are obtained via $\phi(t) = \check{\phi}_a(\phi_{\text{red}}(t), \phi_{\text{blue}}(t))$ where $\check{\phi}_a$ is the perspective of $\phi_a(t) = t^{1/a}$, with $a = \alpha^{1/4}$ and α ranges from 2 to 400 in steps of 10. Small values of a result in ϕ being closer to ϕ_{red} and larger values result in ϕ_a being closer to ϕ_{blue} . Observe that at the three points where ϕ_{red} and ϕ_{blue} agree, so too does ϕ_a .

2.6 EXPERIMENTS

Coherent risk measures have already been successfully used in machine learning. For instance, [Williamson and Menon \[2019\]](#) have employed them in a fairness context and demonstrated that using CVar_α on subgroup losses leads to them being more commensurate. In the context of machine learning, CVar_α has also been reinvented as “average top-k loss” [[Fan et al., 2017](#)]. [Curi et al. \[2020\]](#) have proposed an adaptive sampling method for optimizing CVar_α in a batch setting. [Takeda and Sugiyama \[2008\]](#) have established a close relation of CVar_α and the ν -support vector machine. In reinforcement learning, coherent risk measures have been used e.g. by [Singh et al. \[2020\]](#), [Urpí et al. \[2021\]](#), [Dabney et al. \[2018\]](#), [Tamar et al. \[2015\]](#), [Vijayan et al. \[2021\]](#). Furthermore, distributionally robust optimization approaches based on f -divergence or Wasserstein ambiguity sets, which have been used extensively in machine learning, are subsumed in the framework of coherent risk measures [[Rahimian and Mehrotra, 2019](#)].

In our experiments, we aim to illustrate how spectral risk measures can lead to more robust solutions and attenuate inequality in the loss distribution. We focus on two kinds of problems: first, a coherent risk measure can act directly on the individual losses. We then have the risk minimization problem

$$\operatorname{argmin}_f R(\ell(f(X), Y)) \quad (2.168)$$

for a risk measure R , a function f from some hypothesis space, a loss function ℓ , input X and ground truth labels Y . For its empirical counterpart, we use the empirical distribution

of training losses. The risk measure R then aggregates the observed losses, where each datum has a corresponding individual loss. Here, replacing the expectation $\mathbb{E}$ by a coherent risk measure R , in particular a spectral risk measure, has the effect of emphasizing large individual losses. As a consequence, a distribution of individual losses with less extreme losses (tail risk) will be preferred. There appears to be a fundamental trade-off between optimizing average loss versus reducing inequality. The precise nature of this trade-off is encoded in the choice of the fundamental function. In general, this setup is attractive in situations where relevant subgroups are not known or when a regulating agency disallows making decisions about people based on divisions into subgroups.

Second, we can apply coherent risk measures on subgroup losses. In a fairness context, we may wish to divide our data into ethically salient subgroups (e.g. based on gender or race) and then ask for commensurate subgroup losses. In a technical context, for instance in multiclass classification, we may wish to achieve good performance not only on average, but also good performance for underrepresented classes in the training data. This is especially relevant if the distribution of the number of instances per class is heavy-tailed, as for example in natural species classification [Van Horn et al., 2018].

Typically, the performance of a machine learning system is summarized by the average error on a test set. However, we think that this is a poor way of describing its performance, as it neglects the tail risk. In some settings, heavy-tailed risk must be avoided. This raises the question of a better performance representation which is sensitive to tail risks. We put forward two proposals.

2.6.1 CVAR CURVES

In light of the Kusuoka representation, the family CVar_α is the fundamental building block of all coherent risk measures. In a sense, CVar_α can be seen as measuring tail risks in purest form, as it merely integrates the $1 - \alpha$ tail. We also have the property [Bennett and Sharpley, 1988, p. 61] which is clear from the Kusuoka representation:

$$(\forall \alpha \in [0, 1) : \text{CVar}_\alpha(X) \leq \text{CVar}_\alpha(Y)) \Rightarrow R(X) \leq R(Y) \quad (2.169)$$

for any ri function norm R and $X, Y \in \mathcal{M}^+$. Interestingly, the condition that $\forall \alpha \in [0, 1) : \text{CVar}_\alpha(X) \leq \text{CVar}_\alpha(Y)$ is equivalent to saying that X is dominated by Y in the *second stochastic order* [Ding, 2023, Bäuerle and Müller, 2006]. Then Equation 2.169 states that any ri function norm is consistent with the second stochastic order.²⁵

Due to this characterization, we propose to measure the performance by CVar_α loss curves. The x -axis of this visualization corresponds to $\alpha \in [0, 1)$ and the y -axis shows $\text{CVar}_\alpha(X)$

²⁵Note that the assumption that the measure space is *resonant* is crucial here [Bennett and Sharpley, 1988, p. 61]; for a thorough discussion regarding atomic probability spaces see [Bäuerle and Müller, 2006].

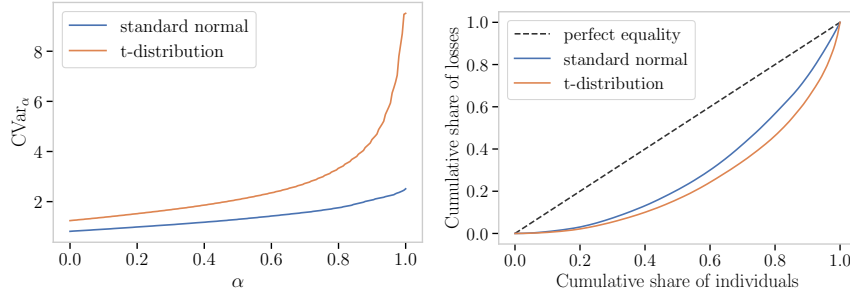


Figure 2.5: Left: $CVar_\alpha$ curves for 500 randomly drawn samples from a standard normal and a t -distribution with 2 degrees of freedom, respectively. Only nonnegative samples were kept. Both have approximately the same mean $CVar_{\alpha=0}$, but the t -distribution has substantially more weight in the tails. Right: empirical Lorenz curves for the same samples. Here, the curve of the standard normal is closer to the diagonal. The diagonal corresponds to perfect equality. The t -distribution exhibits higher inequality as compared to the standard normal.

for some X . As an example, we draw 500 samples from a standard normal distribution and a t -distribution with 2 degrees of freedom. We keep only the nonnegative samples and interpret them as losses. The empirical $CVar_\alpha$ curves are shown in Figure 2.5. Standard risk minimization takes only the value $CVar_{\alpha=0}$ into account, whereas we assert that the whole curve is relevant for measuring the performance. In theory, the outcomes are unbounded for both distributions. In practice, however, we only ever observe finite values which enables plotting the $CVar_\alpha$ curves from samples. Another possibility is to introduce a cutoff value, so only values until for example $\alpha = 0.99$ are plotted.

2.6.2 LORENZ CURVES

A second possibility is to plot *Lorenz curves*, which are widely used in economics to visualize inequality. The Lorenz curve of a random variable X with quantiles F_X^{-1} is defined as [Gastwirth, 1971]:

$$L_X(q) := \frac{1}{\mathbb{E}[X]} \int_0^q F_X^{-1}(p) \, dp, \quad 0 < q \leq 1; \quad L_X(0) = 0. \quad (2.170)$$

The empirical counterpart is defined in the obvious manner. For perfect equality, i.e. $X(\omega) = c \, \forall \omega \in \Omega$, the Lorenz curve is the diagonal. Intuitively, $L(q)$ corresponds to the share of total loss which the individuals with the lowest q -percent of losses have. For an example, look at Figure 2.5. At $q = 0.9$, the empirical Lorenz curve of the t -distribution lies substantially below the curve of the standard normal. This means that the bottom 90% of the t -distribution, i.e. the individuals with the 90% smallest losses, have a smaller share of the total loss than for the standard normal. This implies that for the t -distribution, a larger share of losses is in the 10% tail. Hence the t -distribution has higher tail risk; given the same mean, a risk-averse decision maker would favor the standard normal.

One advantage of Lorenz curves is that even for unbounded random variables X , we can plot the full theoretical Lorenz curve, where CVar_α curves approach infinity. On the other hand, we find that in the context of losses, the interpretation is somewhat unintuitive. In economics, it is undesirable to belong to the bottom; in the context of losses, the bottom is constituted by the well-off. Another disadvantage is that absolute levels of loss are disregarded in this representation, whereas the CVar_α curves also give cardinal information. Therefore we find the CVar_α curves more useful to express the tradeoff between controlling average risk and tail risk.

Like the CVar_α curve, the Lorenz curve is also tightly linked to the second stochastic order [Muliere and Scarsini, 1989]: given $X, Y \in \mathcal{M}$ with $\mathbb{E}[X] = \mathbb{E}[Y]$, it is easy to see that:

$$(\forall \alpha \in [0, 1) : \text{CVar}_\alpha(X) \geq \text{CVar}_\alpha(Y)) \Leftrightarrow (\forall q \in (0, 1] : L_X(q) \leq L_Y(q)).$$

For a visualization, see Figure 2.5. To concisely summarize performance across multiple runs of an experiment, we suggest using the *Gini coefficient*, a single-number measure of inequality. The Gini coefficient of a Lorenz curve is defined as the ratio of the area between the diagonal (perfect equality) and the Lorenz curve over the total area under the diagonal. Therefore, if a distribution is perfectly equal, the Gini coefficient is 0; for a perfectly unequal distribution, where a single ω receives the total loss, it is 1.

2.6.3 SPECTRAL RISK MEASURES ON INDIVIDUAL LOSSES

Throughout, we focus on spectral risk measures in our experiments. First, we apply them to individual losses, i.e. subgroups of size 1. This setup is useful when we do not know relevant classes or care about individual loss in general. Consider, for instance, a self-driving car, which was mostly trained in snow-free environments. When the car is then deployed in a snowy environment, for which training data is scarce, its performance may be diminished. Since it can be a priori hard to partition data into fixed classes, we may employ a risk measure on the individual losses to account for risk aversion. Thus difficult training examples are emphasized. We illustrate this with a simple variant of principal component analysis (PCA), which could of course be replaced with a more sophisticated non-linear method. Since we focus on the risk measures, not the models themselves, we use simple methods for better interpretability of our results.

DATA

We use the MNIST²⁶ and the adult²⁷ data set. MNIST is a standard classification task. The problem is typically to classify grayscale images of handwritten digits, with 28×28 pixels, into the classes 0-9. However, we view it as a dimensionality reduction task, where the goal is to compress the images. MNIST has 60,000 training images and 10,000 test images.

The adult data set contains census data of 48,842 persons with 14 attributes and a binary target attribute, which specifies whether a person earns more or less than 50,000\$ per year. We disregard the binary target attribute and use the other 14 attributes.

To preprocess both MNIST and adult, we apply a MinMaxScaler with the feature range $[-1, 1]$. We split the data into training and test sets. For MNIST, we use 6000 training images and test on the remaining 54,000 images. For adult we use 10,000 data points as the training set and the remaining 38,842 as the test set.

METHOD

Standard PCA can be solved using singular value decomposition. Recall that PCA minimizes the least squares reconstruction error. Assume our n data points $x_i \in \mathbb{R}^m$ with m features are centered, i.e. features have zero mean. Then the PCA objective is

$$\min_{V_k} \sum_{i=1}^n \|x_i - V_k^T V_k x_i\|_{\mathcal{L}^2}^2 \quad \text{so that } V_k V_k^T = I_k \quad (2.171)$$

where $V_k \in \mathbb{R}^{k \times m}$ and I_k is the identity matrix of size k . In contrast to the typical formulation in terms of an eigendecomposition or singular value decomposition, this formulation makes the individual losses explicit. Instead of only the expectation, we use different risk measures on the empirical distribution of reconstruction errors. Hence we obtain a variant of PCA (precisely, of a linear autoencoder) which is sensitive to large individual losses. We replace the objective by:

$$\min_{V_k} R(\|X - V_k^T V_k X\|_{\mathcal{L}^2}^2) \quad (2.172)$$

for a risk measure R and where the distribution of the random loss variable follows the empirical distribution $\hat{P}_n$. We dropped the orthonormality constraint, which does not essentially alter the solution (cf. [Plaut \[2018\]](#)). We label this variant of PCA, where risk measures are employed, as PCA*.

We use the pytorch library to implement our experiments and the Adam optimizer with a learning rate of 0.001. We initialize the matrix V_k with the classical PCA solution, obtained from `sklearn.decomposition.PCA`. For MNIST we use $k = 50$ components and for adult

²⁶<http://yann.lecun.com/exdb/mnist/>

²⁷<https://archive.ics.uci.edu/ml/datasets/Adult>

$k = 5$. We train for 2000 epochs with a learning rate of 0.001. To avoid additional challenges from the stochasticity of mini-batches (see Section 2.6.5), we use the full data for each epoch. We repeat the experiments over 25 independent runs with random training and test splits.

RISK MEASURES

We compare the results when using CVar_α , where $\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$. Recall that $\text{CVar}_{0.0} = \mathbb{E}$. As α increases, sensitivity to tail risk increases. Moreover, we compare variations of the risk measure for integrated risk management ($\text{RIM}_{\alpha,\beta}$), where α is fixed at 0.7 and $\beta \in \{0.2, \dots, 0.8, 1.0\}$. Recall that

$$\text{RIM}_{\alpha,\beta}(X) := \beta \mathbb{E}[X] + (1 - \beta) \text{CVar}_\alpha(X) \quad (2.173)$$

Hence $\beta = 1$ yields the expectation, whereas $\beta = 0$ would yield $\text{CVar}_{\alpha=0.7}$. As β increases, sensitivity to tail risk decreases. In `pytorch`, this risk measure can be easily implemented by combining the `topk` and `mean` function, incurring virtually no computational overhead.

RESULTS

We here show CVar_α curves and Lorenz curves for the test losses on MNIST under the different risk measures. The curves are averaged over the independent runs. The results on `adult` are in Appendix 2.10. In Appendix 2.10, we also show boxplots of the Gini coefficients over the 25 independent runs. From the CVar_α curves, we observe that employing CVar_α or $\text{RIM}_{\alpha,\beta}$ as a risk measure leads to lower tail risks as compared to the expectation. For moderate choices of α and β , we find that performance on average is hardly diminished, but there is substantial gain in tail performance. When choosing a high α such as $\alpha = 0.8$, however, we clearly incur a cost in terms of average performance. From the Lorenz curves and Gini coefficient boxplots we observe that the expectation leads to the highest inequality of loss. As expected, increasing α for CVar_α and decreasing β for $\text{RIM}_{\alpha,\beta}$ (with fixed α) gradually achieves a more equal distribution. This is most clearly visible in the Gini boxplots (Appendix 2.10).

2.6.4 SPECTRAL RISK MEASURES ON SUBGROUP LOSSES

In this experiment, we use spectral risk measures on the aggregated losses of pre-specified subgroups. Denote the random variable which indicates the subgroup belonging to S . The objective in standard expected risk minimization can be written in a two-stage manner as

$$\mathbb{E}[\ell(f(X), Y)] = \mathbb{E}_S [\mathbb{E}_{X,Y|S}[\ell(f(X), Y)]] \quad (2.174)$$

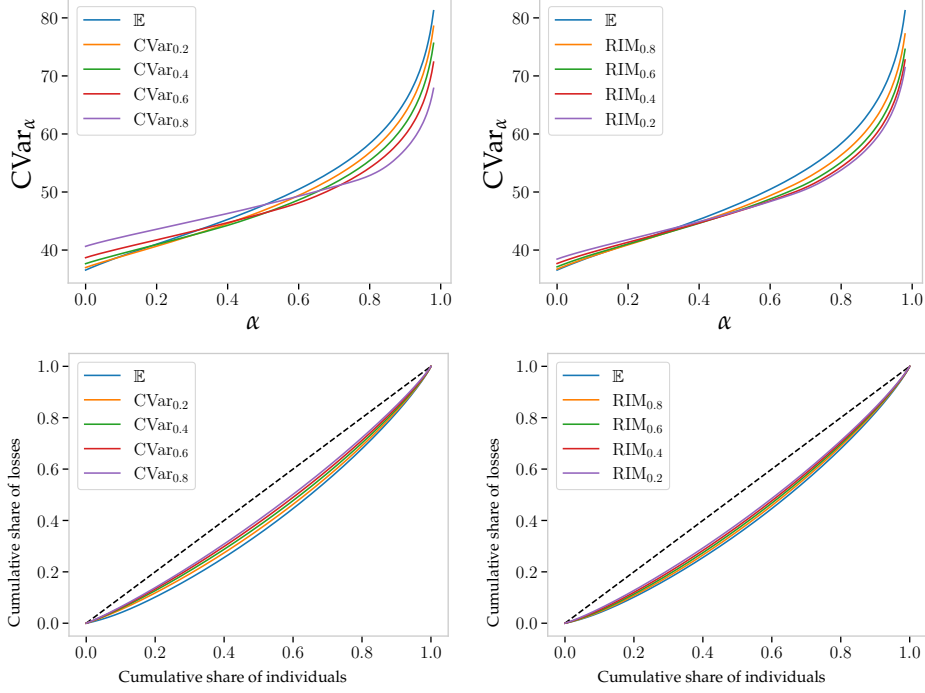


Figure 2.6: PCA* results on MNIST. Top row: CVar_α curves of test losses for CVar_α risk measures (left) with different α and RIMs risk measures (right) with $\alpha = 0.7$ and different β , indicated by subscript. For better visibility of the differences, we cut off α at 0.98. Bottom row: Lorenz curves of test losses for CVar_α (left) and RIMs (right) with $\alpha = 0.7$ and different β .

using a conditional expectation. Our approach is to replace the outer expectation with a risk measure R , as in [Williamson and Menon \[2019\]](#). Within each subgroup, individuals are then treated as fungible and are identified with the subgroup aggregate, since the risk neutral expectation is used. Yet differences between subgroups are considered and punished in a risk-averse fashion, thereby favoring less spread in the distribution of subgroup losses. In our experiments, we employ the respective empirical versions of R and $\mathbb{E}$, i.e. with respect to the empirical distribution.

DATA

In this experiment we perform multi-class classification on MNIST (see Section 2.6.3) and linear regression with the loss $\ell_1(x, y) = |x - y|$ on winequality. On MNIST, we create imbalance in the training data to simulate a setting of *label shift*. We use the following number of random samples (without replacement) for a class with index $s \in [0, 9]$:

$$N(s) = 5000 \cdot \exp\left(\frac{-2s}{10}\right) \quad (2.175)$$

where we round to the nearest integer. The resulting distribution of class frequencies has a moderate imbalance, with frequencies $\{5000, 4098, 3351, 2744, 2246, 1839, 1505, 1232, 1009, 826\}$,

see Figure 2.11 in Appendix 2.10. In the test data, we leave frequencies unchanged, so that the test images are approximately balanced with regard to class.

The task in *winequality* is to predict the perceived quality of a wine (on a numeric scale with integers from 3 to 9) by physiochemical properties (e.g. fixed acidity and alcohol content). In total, there are 11 input attributes. Here, we consider the two subgroups of red and white wine. In contrast to MNIST, we here purposefully rebalance the data set. The frequency of red and white examples is then the same. This experiment serves to illustrate that balancing data does not necessarily solve the problem of disparate subgroup losses. We use 80% of the red wines (1279 examples) in the training set and correspondingly, 1279 examples of white wine. The test set consists of 320 red wines and 3619 white wines. We preprocess both MNIST and *winequality* using a `MinMaxScaler` as in Section 2.6.3.

METHOD

For MNIST, we use a simple multiclass logistic regression, i.e. a cross entropy loss after a single linear layer. We pretrain for 2000 epochs using the expectation as the risk measure and the Adam optimizer with a learning rate of 0.01. Using this initialization, we then train for each risk measure for 5000 epochs with a learning rate of 0.001.

For *winequality* we use a simple feedforward network (one hidden layer of 24 units followed by a nonlinear *ReLU* activation) with the ℓ_1 loss. We train for 3000 epochs using the Adam optimizer with a learning rate of 0.01. Again, we use the full data in each epoch to avoid additional challenges due to stochastic mini-batches (see Section 2.6.5). We compare the same risk measures as in the first experiment (Section 2.6.3). For both data sets we conduct 50 independent runs. On MNIST, for each of these runs we randomly shuffle the assignment of the imbalanced frequencies to the classes.

RESULTS

For MNIST, we report the average subgroup test accuracies and Gini coefficients of subgroup accuracies in Figure 2.7. Due to the data set shift setting, we find that the risk measures even lead to better *average* subgroup performance on the test set. Furthermore, as is visible from the Gini coefficient boxplot, the inequality of subgroup accuracies is reduced.

For *winequality* we show the average of the two subgroup means (red, white) and the absolute difference of the subgroup error means across the 50 runs (Figure 2.8). In general, predictions for white wines incur a higher error than for red wines on average, even though the data is balanced. In this setting, the risk measures yield slightly higher errors on average. However, they reduce the difference between the two subgroup means.

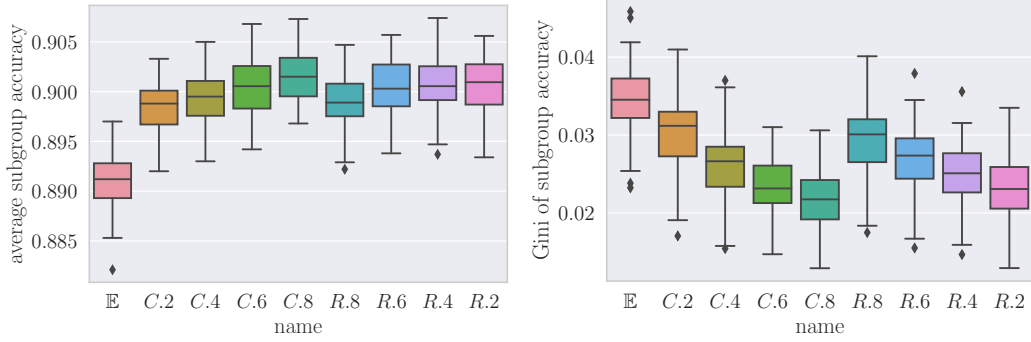


Figure 2.7: Results of logistic regression on MNIST across 50 independent runs. Left: average subgroup accuracies. Right: Gini coefficients of subgroup accuracies. We abbreviate $\text{CVar}_{\alpha=0,f}$ as $C.f$ and $\text{RIM}_{\alpha=0.7,\beta=0,f}$ as $R.f$.

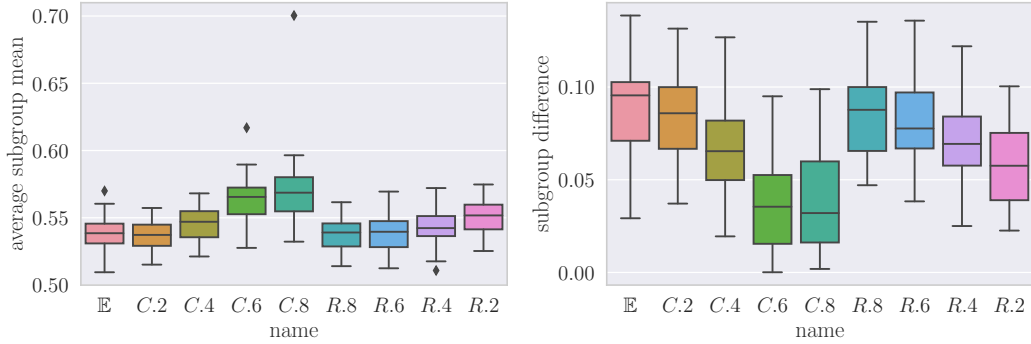


Figure 2.8: Results of linear regression on winequality across 50 independent runs. Left: averages of the two subgroup error means on test data. Right: subgroup absolute error differences on test data across 50 independent runs. We abbreviate $\text{CVar}_{\alpha=0,f}$ as $C.f$ and $\text{RIM}_{\alpha=0.7,\beta=0,f}$ as $R.f$.

2.6.5 DISCUSSION

We have seen that simple spectral risk measures can substantially improve tail performance and hence reduce inequality in the loss distribution. On the other hand, there is a natural trade-off between average and tail performance. In many settings, accounting for tail risk will imply some reduction in average performance. However, this is specific to the train-test data relationship. In the label shift setting on MNIST, we have seen that the increase in robustness by using a spectral risk measure can even lead to better average performance on the test set. Emphasizing “difficult” training examples (associated with high loss) in the optimization process guards against possible data set shift scenarios. For a spectral risk measure, the exact nature of this trade-off is directly encoded in the choice of the fundamental function.

However, there is yet another trade-off: that between *robustness* and *estimability*. We conjecture that it generally holds that higher tail sensitivity of a risk measure is accompanied with higher difficulty in estimating it from empirical samples. For CVar_{α} this is intuitive,

since only a $1 - \alpha$ fraction of the sample is actually used in the estimation. How this trade-off depends exactly on the fundamental function is, to our knowledge, as of now unclear. Since estimating tail-sensitive risk measures may lead to highly variable estimates, using risk measures in a mini-batch setting is problematic; but see [Curi et al. \[2020\]](#) for an approach to optimize CVar_α in a batch setting. See also the recent review by [Laguel et al. \[2021\]](#) on the role of CVar_α in machine learning. The use of other spectral risk measures in practice, which do not admit a simple representation as CVar_α or $\text{RIM}_{\alpha,\beta}$, also raises challenges. To tackle this, [Mehta et al. \[2023\]](#) have recently proposed a practical stochastic gradient-based method for optimizing spectral risk measures and f -divergence risk measures. [Leqi et al. \[2022b\]](#) have obtained uniform convergence results that justify the optimization of a wide class of risk measures (including the spectrals) on the empirical distribution.

2.7 CONCLUSION

In this paper, we have questioned the assumption that the expectation is the only sensible functional to aggregate losses. Instead, we have considered a wide family of possible replacements, the coherent risk measures. These can be used to encode robustness, risk aversion and even fairness. The choice of risk measure is an additional choice to make for the ML engineer and it is orthogonal to the choice of loss function. Therefore we have aimed to stratify the space of possible risk measures. The fundamental function provides such a natural stratification. Depending on the application, it can be interpreted as an imprecise probability, a risk aversion profile or an inequality aversion profile. We have also seen that the fundamental function plays a major role in the combination of different risk measures which further justifies the appellation “fundamental.”

Specifically, we have focused on the subclass of spectral risk measures which, as we have shown, are extremal risk measures with a given fundamental function, and are particularly convenient to work with. These occupy a prime position in the theory of coherent risk measures, can be motivated in different fashions and have been rediscovered multiple times. We assert that this convergence signals that they form a well-founded and important class.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy - EXC number 2064/1 - Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christian Fröhlich. Thanks to Rabanus Derr for many

helpful discussions and comments.

2.8 APPENDIX: PROOFS

2.8.1 PROOF OF THEOREM 33

Proof [Gzyl and Mayoral, 2008, Ridaoui and Grabisch, 2016]:

$$R_\phi(X) = \int_{-\infty}^0 [\phi(S_X(x)) - 1] \, dx + \int_0^\infty \phi(S_X(x)) \, dx \quad (2.176)$$

$$= \int_{-\infty}^0 [\phi(1 - F_X(x)) - 1] \, dx + \int_0^\infty \phi(1 - F_X(x)) \, dx \quad (2.177)$$

$$\stackrel{t=F_X(x)}{=} \int_0^{F_X(0)} [\phi(1 - t) - 1] (F_X^{-1})'(t) \, dt + \int_{F_X(0)}^1 \phi(1 - t) (F_X^{-1})'(t) \, dt \quad (2.178)$$

$$= \int_0^{F_X(0)} \phi'(1 - t) F_X^{-1}(t) \, dt + \int_{F_X(0)}^1 \phi'(1 - t) F_X^{-1}(t) \, dt \quad (2.179)$$

$$= \int_0^1 F_X^{-1}(1 - t) \phi'(t) \, dt = R_{(w)}(X), \quad (2.180)$$

$$(2.181)$$

where the step Equation 2.179 comes from partial integration and F_X is assumed to be continuous. Hence we have the equivalence $\phi'(t) = w(1 - t)$, where w is the spectral weighting function. ■

2.8.2 PROOF OF THEOREM 39

Proof First observe that it is always true that $\mathbb{E}[Y] \leq R'(Y)$, due to the $\mathcal{L}^1$ norm being the smallest of all ri norms (recall that we assume $R(\chi_\Omega) = 1$, implying also $R'(\chi_\Omega) = 1$ due to the associate relationship). Hence we never have any $\mathbb{E}[Y] > 1$ in the unit ball of the associate norm. We label the logical proposition of allowing a representation in the form Equation 2.96 as Terep, that is, the possibility of a representation of the form:

$$R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+ \right\}. \quad (2.182)$$

That Terep implies PTE is trivial:

$$R(X + c) = \sup \left\{ \int_0^1 (X + c)^*(\omega) Y^*(\omega) \, d\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+ \right\} \quad (2.183)$$

$$= \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega + c \int_0^1 Y^*(\omega) \, d\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+ \right\} \quad (2.184)$$

$$= R(X) + c, \quad (2.185)$$

where we used the fact that $(X + c)^* = X^* + c$ holds even if $c < 0$ as long as $X \geq 0$ and $X + c \geq 0$. The other direction is more involved. We show $\text{PTE} \Rightarrow \text{Terep}$ by showing $\neg \text{Terep} \Rightarrow \neg \text{PTE}$. To show this, we need a technical lemma. Define the statement A as: $\exists Z \in \mathcal{M}^+, Z > \epsilon$, for some $\epsilon > 0$, so that:

$$R(Z) = \sup \left\{ \int_0^1 Z^*(\omega) Y^*(\omega) \, d\omega : \mathbb{E}[Y] < 1, R'(Y) \leq 1, Y \in \mathcal{M}^+ \right\}, \quad (2.186)$$

that is, the supremum is not decreased when taking it only over the subset $\{\mathbb{E}[Y] < 1, R'(Y) \leq 1\}$. However, the supremum need not be actually attained.

Lemma 88. $\neg A \wedge \text{PTE} \Rightarrow \text{Terep}$, which is logically equivalent to $\neg \text{Terep} \Rightarrow A \vee \neg \text{PTE}$.

Proof So now assume $\neg A \wedge \text{PTE}$. This means that R is positive translation equivariant and that $\forall Z \in \mathcal{M}^+, Z > \epsilon$, for some $\epsilon > 0$, it holds

$$R(Z) = \sup \left\{ \int_0^1 Z^*(\omega) Y^*(\omega) \, d\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+ \right\}. \quad (2.187)$$

To see that this is indeed the negation of A , observe that $\neg A$ means $\forall Z \in \mathcal{M}^+, Z > \epsilon$, for some $\epsilon > 0$

$$R(Z) > \sup \left\{ \int_0^1 Z^*(\omega) Y^*(\omega) \, d\omega : \mathbb{E}[Y] < 1, R'(Y) \leq 1, Y \in \mathcal{M}^+ \right\}. \quad (2.188)$$

Negating the equality in Equation 2.186 here must yield “strictly greater”, since the set on the righthand side in Equation 2.188 is a subset of the full envelope $\{Y : R'(Y) \leq 1\}$, which is implicit in the lefthand side of Equation 2.188. Now we continue to analyze the statement Equation 2.188. Formally, we can write it as $\sup_C f(Y) > \sup_{C \setminus B} f(Y)$, where $B \subseteq C$. Here, $C := \{Y : R'(Y) \leq 1\}$ and $B := \{Y : \mathbb{E}[Y] = R'(Y) = 1\}$. The function $f(Y) := \int_0^1 Z^*(\omega) Y^*(\omega) \, d\omega$ runs over the respective set in the subscript. But then $\sup_C f(Y) = \sup_B f(Y)$ holds, because $\sup_C f(Y) = \sup_{(C \setminus B) \cup B} f(Y) = \max(\sup_{C \setminus B} f(Y), \sup_B f(Y)) = \sup_B f(Y)$, where we used that by assumption $\sup_C f(Y) > \sup_{C \setminus B} f(Y)$. It is legitimate to “decompose” the supremum over the union into a maximum over the two suprema, cf.

for instance [Hiriart-Urruty and Lemaréchal, 2004, p. 3]. We conclude therefore that Equation 2.187 is the negation of A . Intuitively, $\neg A$ is a slightly weakened form of Terep. But we now show that when it is combined with PTE, we can strengthen it and obtain Terep. Then for any $X \in \mathcal{M}^+$ (for brevity, we drop explicitly writing $Y \in \mathcal{M}^+$):

$$R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega : R'(Y) \leq 1 \right\} \quad (\text{general representation for any } R) \quad (2.189)$$

$$\Rightarrow \forall \epsilon > 0 : R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega + \epsilon \int_0^1 Y^*(\omega) \, d\omega - \epsilon \int_0^1 Y^*(\omega) \, d\omega : R'(Y) \leq 1 \right\}, \quad (2.190)$$

from which it follows due to PTE that:

$$\forall \epsilon > 0 : R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega + \epsilon \int_0^1 Y^*(\omega) \, d\omega : \mathbb{E}[Y] = R'(Y) = 1 \right\} - \epsilon.$$

Since due to PTE: $R(X + \epsilon - \epsilon) = R(X + \epsilon) - \epsilon$ since $(X + \epsilon) - \epsilon \geq 0$. Taking the supremum over $\{Y : \mathbb{E}[Y] = R'(Y) = 1\}$ suffices due to $\neg A$. Then:

$$\forall \epsilon > 0 : R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega + \epsilon \int_0^1 Y^*(\omega) \, d\omega - \epsilon \int_0^1 Y^*(\omega) \, d\omega : \mathbb{E}[Y] = R'(Y) = 1 \right\} \quad (2.191)$$

$$\Rightarrow R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, d\omega : \mathbb{E}[Y] = R'(Y) = 1 \right\}. \quad (2.192)$$

$$(2.193)$$

Hence R has a positive translation equivariant representation, i.e. Terep holds. We have thus shown that $\neg A \wedge \text{PTE} \Rightarrow \text{Terep}$, which is equivalent to $\neg \text{Terep} \Rightarrow A \vee \neg \text{PTE}$. ■

Recall that our goal is to show that $\neg \text{Terep} \Rightarrow \neg \text{PTE}$. Using the lemma 88, that $\neg \text{Terep} \Rightarrow A \vee \neg \text{PTE}$, it only remains to show that $A \Rightarrow \neg \text{PTE}$. So now assume A . This means that $\exists Z \in \mathcal{M}^+, Z > \epsilon$, for some $\epsilon > 0$, so that

$$R(Z) = \sup \left\{ \int_0^1 Z^*(\omega) Y^*(\omega) \, d\omega : \mathbb{E}[Y] < 1, R'(Y) \leq 1, Y \in \mathcal{M}^+ \right\}. \quad (2.194)$$

Write now $X + \epsilon = Z$, which is possible by assumption. Then $X \in \mathcal{M}^+$ and:

$$R(X + \epsilon) = \sup \left\{ \int_0^1 (X + \epsilon)^*(\omega) Y^*(\omega) \, d\omega : \mathbb{E}[Y] < 1, R'(Y) \leq 1, Y \in \mathcal{M}^+ \right\} \quad (2.195)$$

$$\leq R(X) + \epsilon \sup \left\{ \int_0^1 Y^*(\omega) \, d\omega : \mathbb{E}[Y] < 1, R'(Y) \leq 1, Y \in \mathcal{M}^+ \right\} \quad (\text{subadditivity}) \quad (2.196)$$

$$< R(X) + \epsilon. \quad (2.197)$$

Therefore R is not PTE and the proof is complete. ■

2.8.3 FAMILIES OF FUNDAMENTAL FUNCTIONS

PROOF OF THEOREM 50.

Proof Suppose R is an ri norm which is PTE and has fundamental function ϕ . Write

$$R(X) = \sup_{Z \in \mathcal{Z}} \left\{ \int_0^1 X^*(\omega) Z'(\omega) \, d\omega \right\}, \quad \phi(t) = \sup_{Z \in \mathcal{Z}} Z(t). \quad (2.198)$$

for some Kusuoka set $\mathcal{Z}$ of R . And define

$$\forall t \in (0, 1] : Z_t(x) := \begin{cases} \phi(t) \frac{x}{t} & , x \leq t \\ \frac{1-\phi(t)}{1-t} x + \frac{\phi(t)-t}{1-t} & , x > t \end{cases} \quad (2.199)$$

Then the TM norm can be written as:

$$\|X\|_{TM_\phi} = \sup \left\{ \int_0^1 X^*(\omega) Z'_t(\omega) \, d\omega : t \in (0, 1] \right\}. \quad (2.200)$$

Lemma 89. *Using the above definitions, it holds that:*

$$\forall t \in (0, 1] : R(X) \geq \int_0^1 X^*(\omega) Z'_t(\omega) \, d\omega. \quad (2.201)$$

Proof [of the Lemma] Consider some fixed $t \in (0, 1]$. Since $\phi(t) = \sup_{Z \in \mathcal{Z}} Z(t)$, we can, to any $\epsilon > 0$, find some $Z_\epsilon \in \mathcal{Z}$ so that $0 \leq \phi(t) - Z_\epsilon(t) < \epsilon$. Let $\epsilon_n \downarrow 0$ be an arbitrary sequence converging to zero and denote by Z_{ϵ_n} a corresponding sequence of selected concave functions from the Kusuoka set $\mathcal{Z}$. Next, define

$$h_{\epsilon_n}(x) = \begin{cases} Z_{\epsilon_n}(t) \frac{x}{t} & , x \leq t \\ \frac{1-Z_{\epsilon_n}(t)}{1-t} x + \frac{Z_{\epsilon_n}(t)-t}{1-t} & , x > t \end{cases}, \quad h'_{\epsilon_n}(x) := \begin{cases} \frac{Z_{\epsilon_n}(t)}{t} & , x \leq t \\ \frac{Z_{\epsilon_n}(t)-1}{t-1} & , x > t \end{cases} \quad (2.202)$$

where the derivative is defined first so that $h_{\varepsilon_n}(x) := \int_0^x h'_{\varepsilon_n}(\omega) \, d\omega$. ■

We observe that by construction $Z_{\varepsilon_n} \geq h_{\varepsilon_n}$; the condition for Hardy's lemma is fulfilled, i.e. $\int_0^x Z'_{\varepsilon_n}(\omega) \, d\omega \geq \int_0^x h'_{\varepsilon_n}(\omega) \, d\omega$. Therefore

$$I(\varepsilon_n) := \int_0^1 Z'_{\varepsilon_n}(\omega) X^*(\omega) \, d\omega \geq \int_0^1 h'_{\varepsilon_n}(\omega) X^*(\omega) \, d\omega =: \Pi(\varepsilon_n).$$

By definition of h'_{ε_n} ,

$$\Pi(\varepsilon_n) = \frac{Z_{\varepsilon_n}(t)}{t} \int_0^t X^*(\omega) \, d\omega + \frac{Z_{\varepsilon_n}(t) - 1}{t - 1} \int_t^1 X^*(\omega) \, d\omega.$$

Now consider $I(\varepsilon_n)$. We know that $R(X) \geq I(\varepsilon_n)$ by our choice of the Z_{ε_n} . Hence

$$R(X) \geq \sup_{n \in \mathbb{N}} \left\{ \int_0^1 Z'_{\varepsilon_n}(\omega) X^*(\omega) \, d\omega \right\} = \limsup_{\varepsilon_n \downarrow 0} I(\varepsilon_n) \geq \limsup_{\varepsilon_n \downarrow 0} \Pi(\varepsilon_n),$$

since $0 \leq a_n \leq b_n$ implies $\limsup a_n \leq \limsup b_n$.

We find that

$$\lim_{n \rightarrow \infty} \frac{Z_{\varepsilon_n}(t)}{t} \int_0^t X^*(\omega) \, d\omega = \frac{\phi(t)}{t} \int_0^t X^*(\omega) \, d\omega,$$

since the integral term is constant, and as $n \rightarrow \infty$, $Z_{\varepsilon_n}(t) \rightarrow \phi(t)$ by construction (note that t is fixed). Similarly

$$\lim_{n \rightarrow \infty} \frac{Z_{\varepsilon_n}(t) - 1}{t - 1} \int_t^1 X^*(\omega) \, d\omega = \frac{\phi(t) - 1}{t - 1} \int_t^1 X^*(\omega) \, d\omega.$$

Since both limits exist, the limit of their sum exists:

$$\lim_{n \rightarrow \infty} \Pi(\varepsilon_n) = \frac{\phi(t)}{t} \int_0^t X^*(\omega) \, d\omega + \frac{\phi(t) - 1}{t - 1} \int_t^1 X^*(\omega) \, d\omega = \int_0^1 Z'_t(\omega) X^*(\omega) \, d\omega.$$

Therefore $R(X) \geq \int_0^1 Z'_t(\omega) X^*(\omega) \, d\omega$ for fixed t . Since this holds for all $t \in (0, 1]$, we also get $R(X) \geq \sup_{t \in (0, 1]} \left\{ \int_0^1 X^*(\omega) Z'_t(\omega) \, d\omega \right\} = \|X\|_{TM_\phi}$.

Thus, taking the embedding theorem into account, for any PTE ri norm ('coherent risk measure') R :

$$\|X\|_{M_\phi} \leq \|X\|_{TM_\phi} \leq R(X) \leq \|X\|_{\Lambda_\phi} \quad \forall X \in \Lambda_\phi. \quad (2.203)$$

We show explicitly that $\|\cdot\|_{TM_\phi}$ is positive translation equivariant. Let $X \in \mathcal{M}^+$ and $c \in \mathbb{R}$

so that $X + c \geq 0$:

$$\begin{aligned}
& \|X + c\|_{TM_\phi} \\
&= \sup_{0 < t < 1} \left\{ \frac{\phi(t)}{t} \int_0^t (X + c)^*(\omega) \, d\omega + \frac{\phi(t) - 1}{t - 1} \int_t^1 (X + c)^*(\omega) \, d\omega \right\} \\
&= \sup_{0 < t < 1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega) \, d\omega + \frac{\phi(t) - 1}{t - 1} \int_t^1 X^*(\omega) \, d\omega + \frac{\phi(t)}{t} \int_0^t c \, d\omega + \frac{\phi(t) - 1}{t - 1} \int_t^1 c \, d\omega \right\} \\
&= \sup_{0 < t < 1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega) \, d\omega + \frac{\phi(t) - 1}{t - 1} \int_t^1 X^*(\omega) \, d\omega + \frac{\phi(t)}{t} ct + \frac{\phi(t) - 1}{t - 1} (c - ct) \right\} \\
&= \sup_{0 < t < 1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega) \, d\omega + \frac{\phi(t) - 1}{t - 1} \int_t^1 X^*(\omega) \, d\omega + c \right\} \\
&= \|X\|_{TM_\phi} + c
\end{aligned}$$

and therefore the norm is PTE.

Recall that both the Dutch risk measure and the spectral MaxVar share the fundamental function $\phi(t) = 2t - t^2$. We show that the Dutch risk measure is a special case of this TM_ϕ norm. In this case, we have $\phi(t)/t = 2 - t$ and $(\phi(t) - 1)/(t - 1) = 1 - t$, $t \neq 1$. Therefore

$$\|X\|_{TM_\phi} = \sup_{0 < t < 1} \left\{ (2 - t) \int_0^t X^*(\omega) \, d\omega + (1 - t) \int_t^1 X^*(\omega) \, d\omega \right\} \quad (2.204)$$

$$= \sup_{0 < t < 1} \left\{ 2 \int_0^t X^*(\omega) \, d\omega - t \int_0^1 X^*(\omega) \, d\omega + \int_t^1 X^*(\omega) \, d\omega - t \int_t^1 X^*(\omega) \, d\omega \right\} \quad (2.205)$$

$$= \sup_{0 < t < 1} \left\{ (1 - t) \int_0^1 X^*(\omega) \, d\omega + \int_0^t X^*(\omega) \, d\omega \right\} \quad (2.206)$$

$$= \sup_{0 < t < 1} \{ (1 - t) \mathbb{E}[|X|] + t \cdot \text{CVar}_{1-t}(|X|) \} = \text{Du}(|X|). \quad (2.207)$$

According to [Pichler and Shapiro \[2012, Corollary 5.1\]](#), the last expression is equal to the Dutch risk measure.

We still need to show that $\|\cdot\|_{TM_\phi}$ is indeed a valid ri norm. First, consider it on the positive cone as an ri function norm. In Equation 2.200, we have it expressed as a supremum over a set of Lorentz norms. Indeed a supremum over a non-empty but otherwise arbitrary set of ri function norms is a valid ri function norm (see Lemma 90). Recall that we assume $R(\chi_\Omega) = 1$ throughout the paper.

Lemma 90. *Let $\{R_i : i \in \mathcal{I}\}$ be a non-empty family of ri function norms (Definitions 34,35). Then $R(X) := \sup\{R_i(X) : i \in \mathcal{I}\}$ is an ri function norm.*

Proof [of the Lemma] Recall that we globally assume that any ri function norm $\tilde{R}$ satisfies $\tilde{R}(\chi_\Omega) = 1$ (which clearly implies $R(\chi_\Omega) = 1$ here). Properties R1 and R2 are easy to check.

For R3 we want to show that

$$(\forall i \in \mathcal{I} : 0 \leq X_n \uparrow X \text{ } \mu\text{-a.e.} \Rightarrow R_i(X_n) \uparrow R_i(X)) \implies (0 \leq X_n \uparrow X \text{ } \mu\text{-a.e.} \Rightarrow R(X_n) \uparrow R(X)).$$

So let us assume $\forall i \in \mathcal{I} : 0 \leq X_n \uparrow X \text{ } \mu\text{-a.e.} \Rightarrow R_i(X_n) \uparrow R_i(X)$. Then $R(X) = \sup_{i \in \mathcal{I}} R_i(X) = \sup_{i \in \mathcal{I}} \lim_{n \rightarrow \infty} R_i(X_n)$. First, assume $R(X) < \infty$. From the definition of $R(X)$ as the sup, we know that $\forall \varepsilon > 0 : \exists i(\varepsilon) \in \mathcal{I} : R_i(X) > R(X) - \varepsilon/2$. Second, we know that

$$\forall \delta > 0 : \forall i \in \mathcal{I} : \exists n_i \in \mathbb{N} : \forall n \geq n_i : R_i(X_n) > R(X) - \delta/2.$$

Choosing $\varepsilon := \delta$ and taking the corresponding $i(\delta)$ yields:

$$\forall \delta > 0 : \exists i(\delta) : \exists n_i \in \mathbb{N} : \forall n \geq n_i : R_i(X_n) > R_i(X) - \delta/2 > R(X) - \delta/2 \quad (2.208)$$

$$\Leftrightarrow \forall \delta > 0 : \exists n_i \in \mathbb{N} : \forall n \geq n_i : \sup_{i \in \mathcal{I}} R_i(X_n) > R(X) - \delta \quad (2.209)$$

$$\Leftrightarrow \lim_{n \rightarrow \infty} \sup_{i \in \mathcal{I}} R_i(X_n) \geq R(X). \quad (2.210)$$

The statement $\lim_{n \rightarrow \infty} R(X_n) = R(X)$ can be written as

$$\lim_{n \rightarrow \infty} \sup_{i \in \mathcal{I}} R_i(X_n) = \sup_{i \in \mathcal{I}} \lim_{n \rightarrow \infty} R_i(X_n).$$

Since $\lim_{n \rightarrow \infty} \sup_{i \in \mathcal{I}} R_i(X_n) \leq \sup_{i \in \mathcal{I}} \lim_{n \rightarrow \infty} R_i(X_n)$ is obvious, we have, taking this together with Equation 2.210, shown both inequalities and thus equality, i.e. $\lim_{n \rightarrow \infty} R(X_n) = R(X)$. That the convergence is from below is clear; hence $R(X_n) \uparrow R(X)$. Finally, if $R(X) = \infty$ then it is obvious that $R(X_n) \uparrow \infty$.

As to R4, note that $R_i(\chi_\Omega) = 1 \ \forall i \in \mathcal{I}$ implies $R(\chi_\Omega) = 1$, which by monotonicity implies $R(\chi_E) \leq 1$ for measurable E . Also, $\int_E X \, d\mu \leq \int_\Omega X \, d\mu = \mathbb{E}[X] \leq R_i(X) \forall i \in \mathcal{I}$ by assumption that $R_i(\chi_\Omega) = 1$ and the embedding theorem. Hence choosing e.g. $c = 2$ gives $\int_E X \, d\mu < cR(X)$ for any measurable E . Finally, the ri property is obvious. ■

The theorem is thus proved. ■

PROOF OF THEOREM 52

Proof Let $\phi(t) = \min \{t/(1-\alpha), 1\}$ for some $\alpha \in [0, 1)$. Hence $\phi \in \Phi_{0+}$. We show that $\|X\|_{M_\phi} = \text{CVar}_\alpha(|X|) = \|X\|_{\Lambda_\phi}$. Clearly, the Lorentz norm for such ϕ is CVar_α , as

$$\|X\|_{\Lambda_\phi} = \frac{1}{1-\alpha} \int_0^{1-\alpha} X^*(\omega) \, d\omega = \text{CVar}_\alpha(|X|). \quad (2.211)$$

The Marcinkiewicz norm is

$$\|X\|_{M_\phi} = \sup_{0 < t \leq 1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega) \, d\omega \right\} \quad (2.212)$$

We claim that the supremum is reached at $t = 1 - \alpha$. Then:

$$\|X\|_{M_\phi} = \frac{1}{1 - \alpha} \int_0^{1-\alpha} X^*(\omega) \, d\omega = \|X\|_{\Lambda_\phi} = \text{CVar}_\alpha(|X|). \quad (2.213)$$

Hence it remains to show that the supremum is indeed reached at $t = 1 - \alpha$. Let $t = 1 - \alpha + \epsilon$, $\epsilon > 0$. Then

$$\frac{\phi(t)}{t} \int_0^t X^*(\omega) \, d\omega = \phi(1 - \alpha + \epsilon) X^{**}(1 - \alpha + \epsilon) = 1 \cdot \text{CVar}_{\alpha-\epsilon}(|X|) < \text{CVar}_\alpha(|X|). \quad (2.214)$$

Let $t < 1 - \alpha$. Then

$$\frac{\phi(t)}{t} \int_0^t X^*(\omega) \, d\omega = \frac{1}{1 - \alpha} \int_0^t X^*(\omega) \, d\omega < \frac{1}{1 - \alpha} \int_0^{1-\alpha} X^*(\omega) \, d\omega = \text{CVar}_\alpha(|X|), \quad (2.215)$$

since X^* is nonnegative. For $\alpha \rightarrow 1$, the Marcinkiewicz and the Lorentz norm both coincide with the $\mathcal{L}^\infty$ norm. If $\alpha \rightarrow 1$, then $\phi(t) = \chi_{(0,1]} \notin \Phi_{0+}$.

$$\|X\|_{\Lambda_\phi} = \int_0^1 X^*(\omega) \phi'(\omega) \, d\omega + X^*(0) \phi(0+) = X^*(0) = \|X\|_{\mathcal{L}^\infty} \quad (2.216)$$

$$\|X\|_{M_\phi} = \sup_{0 < t \leq 1} \left\{ 1 \cdot \frac{1}{t} \int_0^t X^*(\omega) \, d\omega \right\} = X^*(0) = \|X\|_{\mathcal{L}^\infty}. \quad (2.217)$$

Therefore we have established that the coincidence of Marcinkiewicz and Lorentz norm holds, implying that there is then only a single coherent risk measure CVar_α with the given fundamental function. It remains to show the converse direction, $\|X\|_{M_\phi} = \|X\|_{\Lambda_\phi}$ only if ϕ is of CVar_α -type, i.e. $\phi(t) = \min \{t/(1 - \alpha), 1\}$ for some $\alpha \in [0, 1)$ or $\phi(t) = \phi_\infty(t) = \lim_{\alpha \rightarrow 1} \min \{t/(1 - \alpha), 1\}$. We show that the Marcinkiewicz norm is only positive translation equivariant if ϕ is of that type. Since the Lorentz norm is always positive translation equivariant, the norms can only then coincide. Let $\{\phi_t\}$ be the Kusuoka set of the Marcinkiewicz norm constructed as before.

$$\forall t \in (0, 1] : \phi_t(x) := \begin{cases} \phi(t) \frac{x}{t} & , x \leq t \\ \phi(t) & , x > t. \end{cases} \quad (2.218)$$

If the norm is PTE, we can reduce it (Theorem 39) to a representation consisting only of those $\phi_{t'}$ with $\phi_{t'}(1) = 1$. While the collection $\{\phi_t\}$ need not be the maximal Kusuoka set

$\{t \mapsto \int_0^t Y^*(\omega) \, d\omega : \|Y\|_{M'_\phi} \leq 1\}$, the proof of Theorem 39 is agnostic to the used Kusuoka representation; uniqueness is not assumed. But then, these $\phi_{t'}$ by their definition are

$$\phi_{t'}(x) := \begin{cases} \frac{x}{t'} & , x \leq t' \\ 1 & , x > t' \end{cases} = \min \{x/(1-\alpha), 1\}, \quad \alpha = 1 - t'. \quad (2.219)$$

The fundamental function is then $\phi(x) = \sup_{t'} \phi_{t'}(x) = \sup_{t'} \min \{x/t', 1\}$, and therefore ϕ is of CVar_α type. More specifically, either a single $t' = 1 - \alpha$ suffices in the representation or $\alpha \rightarrow 1$, for which an uncountable infinity of t' is needed, i.e. $\phi_{t'}(x) = \sup_{t' \rightarrow 0} \min \{\frac{x}{t'}, 1\}$, thus $\phi = \chi_{(0,1]}$. ■

VARIATIONAL REPRESENTATION OF $\text{RIM}_{\alpha,\beta}$

Recall that (Theorem 53):

$$\text{RIM}_{\alpha,\beta}(X) := \beta \mathbb{E}[X] + (1 - \beta) \text{CVar}_\alpha(X). \quad (2.220)$$

We show that $\text{RIM}_{\alpha,\beta}$ admits the following variational representation:

$$\text{RIM}_{\alpha,\beta}(X) = \inf_{\mu \in \mathbb{R}} \mu + \mathbb{E}v(X - \mu) \quad \forall X \in \mathcal{M}, \quad (2.221)$$

where the regret function v is given by the piecewise linear

$$v(t) = \begin{cases} \beta t & t \leq 0 \\ \frac{\beta\alpha-1}{\alpha-1} t & t > 0. \end{cases} \quad (2.222)$$

Proof We here translate a result from [Pflug and Ruszczyński, 2001] to a loss-based formulation. Let $0 < \lambda_1 < 1 < \lambda_2$. Consider the function:

$$f_\mu(x) = \mu + \lambda_2(x - \mu)^+ - \lambda_1(x - \mu)^- \quad (2.223)$$

$$= \mu + (\lambda_2 - \lambda_1)(x - \mu)^+ + \lambda_1(x - \mu) \quad (2.224)$$

$$= \mu + (\lambda_2 - \lambda_1)(x - \mu)^+ + \lambda_1 x - \lambda_1 \mu \quad (2.225)$$

$$= (\lambda_2 - \lambda_1)(x - \mu)^+ + \lambda_1 x + (1 - \lambda_1)\mu \quad (2.226)$$

$$= \lambda_1 x + (1 - \lambda_1) \left(\mu + \frac{\lambda_2 - \lambda_1}{1 - \lambda_1} (x - \mu)^+ \right). \quad (2.227)$$

Now,

$$\text{RIM}_{\alpha,\beta}(X) = \inf_{\mu \in \mathbb{R}} \mathbb{E}[f_{\mu}(X)] \quad (2.228)$$

$$= \lambda_1 \mathbb{E}[X] + (1 - \lambda_1) \inf_{\mu \in \mathbb{R}} \left(\mu + \frac{\lambda_2 - \lambda_1}{1 - \lambda_1} \mathbb{E}(X - \mu)^+ \right) \quad (2.229)$$

$$= \beta \mathbb{E}[X] + (1 - \beta) \text{CVar}_{\alpha}(X), \quad (2.230)$$

where $\lambda_1 = \beta$ and $\frac{1}{1-\alpha} = \frac{\lambda_2 - \lambda_1}{1 - \lambda_1}$, hence $\lambda_2 = \frac{\beta\alpha - 1}{\alpha - 1}$. ■

2.8.4 NORM EQUIVALENCES AND TAIL RISK

PROOF OF THEOREM 55

Proof The proof in [Pichler, 2013] relies on translation equivariance, i.e. that $\phi_{1\gamma}(1) = 1 \forall \gamma$ and $\phi_{2\zeta} = 1 \forall \zeta$ (cf. Section 2.4.7). To enable comparison for not necessarily translation equivariant ri norms, as well, we give a new and shorter proof. First, assume that $\|\cdot\|_{\mathcal{R}_1}, \|\cdot\|_{\mathcal{R}_2}$ have singleton Kusuoka sets $\mathcal{Z}_1 = \{\phi_1\}$ and $\mathcal{Z}_2 = \{\phi_2\}$, where ϕ_1, ϕ_2 are concave functions which are not required to satisfy $\phi_1(1) = 1, \phi_2(1) = 1$. Then let

$$K := \sup_{0 < \alpha \leq 1} \frac{\phi_2(\alpha)}{\phi_1(\alpha)}. \quad (2.231)$$

Thus $\forall \alpha \in (0, 1]: \phi_2(\alpha) \leq K \cdot \phi_1(\alpha)$. Then it directly follows from Hardy's lemma that

$$\|X\|_{\mathcal{R}_2} = \int_0^1 X^*(\omega) \phi_2'(\omega) d\omega \leq K \cdot \int_0^1 X^*(\omega) \phi_1'(\omega) d\omega = K \cdot \|X\|_{\mathcal{R}_1}. \quad (2.232)$$

Note that the use of the formal derivatives ϕ_1', ϕ_2' is unproblematic even with kinks, since the Kusuoka sets are constructed as integrals of nonnegative decreasing functions. That is, we use the dash symbol here not as a differentiation operator, but as a mapping which assigns to a ϕ_1 the function from which it was constructed as the integral of. If $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are not singletons, the constant C according to Equation 2.140 is equal to [Pichler, 2013]

$$C = \inf\{c > 0 : \forall \phi_{2\zeta} \in \mathcal{Z}_2 : \exists \phi_{1\gamma} \in \mathcal{Z}_1 : \forall \alpha \in (0, 1] : \phi_{2\zeta}(\alpha) \leq c \cdot \phi_{1\gamma}(\alpha)\},$$

which ensures that $\forall \epsilon > 0 : \forall \phi_{2\zeta} \in \mathcal{Z}_2 : \exists \phi_{1\gamma} \in \mathcal{Z}_1 : \forall \alpha \in (0, 1] : \phi_{2\zeta}(\alpha) \leq (C + \epsilon) \cdot \phi_{1\gamma}(\alpha)$. But then,

$$\|X\|_{\mathcal{R}_2} = \sup \left\{ \int_0^1 X^*(\omega) \phi'_{2\zeta}(\omega) d\omega : \phi_{2\zeta} \in \mathcal{Z}_2 \right\} \quad (2.233)$$

$$\leq (C + \epsilon) \cdot \sup \left\{ \int_0^1 X^*(\omega) \phi'_{1\gamma}(\omega) d\omega : \phi_{1\gamma} \in \mathcal{Z}_1 \right\} = (C + \epsilon) \cdot \|X\|_{\mathcal{R}_1}, \quad (2.234)$$

where we applied Hardy's lemma to each of the replacements of $\phi_{2\zeta} \rightarrow \phi_{1\gamma}$ such that the above inequality holds. Let $\epsilon \downarrow 0$ to obtain $\|X\|_{\mathcal{R}_2} \leq C \cdot \|X\|_{\mathcal{R}_1}$. $\blacksquare$

The converse direction: [Pichler \[2013\]](#) stated that $C = \infty$ Equation 2.140 implies non-equivalence of the norms. However, no proof was provided for the statement. In Theorem 62 we provide a counterexample. We raise the following point: our counterexample involves the Marcinkiewicz norm, which is not positive translation equivariant. In this specific case, the constant C cannot be bounded because of the behaviour for large values of α . This can only happen since the Kusuoka set does not satisfy $\forall \phi_1 : \phi_1(\alpha) \geq \alpha$, which would be the case when $\|\cdot\|_{\mathcal{R}_1}$ is PTE. In contrast, when both involved norms are PTE, $C = \infty$ can hold only because $\lim_{\alpha \rightarrow 0}$ grows unbounded. This then concerns the tail behaviour of the norms. We conjecture that when both norms are PTE, the original statement holds, i.e. $C = \infty$ implies non-equivalence. However, we have been unable to prove this statement.

PROOF OF THEOREM 57

Proof Denote by ϕ_t the family of functions

$$\forall t \in (0, 1] : \phi_t(x) := \begin{cases} \phi(t) \frac{x}{t} & , x \leq t \\ \phi(t) & , x > t \end{cases} \quad (2.235)$$

which generate a Kusuoka set of the Marcinkiewicz norm with fundamental function ϕ . Then the problem of finding a C as in Theorem 55 reduces to

$$C = \inf_{\phi_t} \sup_{0 < \alpha \leq 1} \frac{\phi(\alpha)}{\phi_t(\alpha)} = \inf_{\phi_t} \sup_{0 < \alpha \leq 1} \begin{cases} \frac{\phi(\alpha)}{\alpha} \frac{t}{\phi(t)} & , \alpha \leq t \\ \frac{\phi(\alpha)}{\phi(t)} & , \alpha > t. \end{cases} \quad (2.236)$$

Fix any t . Then certainly

$$C \leq \sup_{0 < \alpha \leq 1} \frac{\phi(\alpha)}{\phi_t(\alpha)} = \sup_{0 < \alpha \leq 1} \begin{cases} \frac{\phi(\alpha)}{\alpha} \frac{t}{\phi(t)} & , \alpha \leq t \\ \frac{\phi(\alpha)}{\phi(t)} & , \alpha > t. \end{cases} \quad (2.237)$$

The supremum over the second term (for $\alpha > t$) is bounded, since t is fixed and ϕ is bounded by $[0, 1]$. As to the first term ($\alpha \leq t$), since $\frac{\phi(\alpha)}{\alpha}$ is decreasing in α (due to the quasiconcavity of ϕ), the supremum must occur²⁸ for $\alpha \rightarrow 0$, or otherwise it occurs in the second term. Observe that $\phi'(0) = \lim_{\alpha \rightarrow 0} \frac{\phi(\alpha)}{\alpha}$ by definition of the difference quotient since $\phi(0) = 0$. Therefore:

$$\lim_{\alpha \rightarrow 0} \frac{\phi(\alpha)}{\alpha} \frac{t}{\phi(t)} = \frac{\phi'(0)}{1} \frac{t}{\phi(t)} < \infty. \quad (2.238)$$

Therefore C is finite. As an example of this statement, with the choice $\phi(t) = 2t - t^2$, we obtain equivalence of the Dutch risk measure and MaxVar.

It remains to show that $K = 1/(\phi(\frac{1}{\phi'(0)}))$ is a feasible constant. Obviously, this K is finite under the assumption that $\phi'(0) < \infty$. Consider a linear function with slope $\phi'(0)$. It reaches 1 at $t = 1/\phi'(0)$. With this choice of t , we have

$$\phi_t(\alpha) = \begin{cases} \frac{\phi(1/\phi'(0))}{1/\phi'(0)} \alpha & , \alpha \leq t \\ \phi(1/\phi'(0)) & , \alpha > t. \end{cases} \quad (2.239)$$

Then

$$\phi(\alpha) \leq K \cdot \phi_{1/\phi'(0)} = \begin{cases} \phi'(0) \cdot \alpha & , \alpha \leq 1/\phi'(0) \\ 1 & , \alpha > 1/\phi'(0). \end{cases} \quad (2.240)$$

To see that this holds, observe that $\forall \alpha \in (0, 1] \frac{\phi(\alpha)}{\alpha} \leq \phi'(0) = \lim_{t \rightarrow 0} \frac{\phi(t)}{t}$ due to quasiconcavity and also $\phi(\alpha) \leq 1$. Indeed K is the smallest constant such that the statement $\exists t' \in (0, 1] : \forall \alpha \in (0, 1] \phi(\alpha) \leq K' \cdot \phi_{t'}(\alpha)$ holds. Assume $t' < t = 1/\phi'(0)$. For $\alpha > t'$ we require $K' \cdot \phi(t') \geq 1$, otherwise the majorization does not hold as $\phi(\alpha)$ approaches 1. But since $\phi(t') < \phi(t)$, this implies $K' > K$. Now assume $t' > t$. Clearly, we must require $(K' \cdot \phi_{t'})'(0) \geq \phi'(0)$ for the majorization to hold in the limit as $\alpha \rightarrow 0$ (this comes from letting $\alpha \rightarrow 0$ in the condition $K' \cdot \frac{\phi(t')}{t'} \alpha \geq \phi(\alpha)$). That is, $(K' \cdot \phi_{t'})'(0) = K' \cdot \frac{\phi(t')}{t'}$. By design, we have $K \cdot \phi_t'(0) = K \cdot \frac{\phi(t)}{t} = \phi'(0)$. Due to quasiconcavity and $t' > t$ we have $\frac{\phi(t')}{t'} > \frac{\phi(t)}{t}$. Hence $(K' \cdot \phi_{t'})'(0) \geq \phi'(0)$ implies $K' > K$.

However, we note that K need not be the smallest constant such that $\|X\|_{\Lambda_\phi} \leq K \cdot \|X\|_{M_\phi}$ holds, but it is the smallest constant so that $\exists t' \in (0, 1] : \forall \alpha \in (0, 1] \phi(\alpha) \leq K' \cdot \phi_{t'}(\alpha)$, which guarantees the previous statement to hold. ■

²⁸When writing *occur*, we do not mean to imply that a supremum is actually attained.

PROOF OF THEOREM 59

Proof The result can be easily derived by combining two statements from [Rubshtein et al. \[2016\]](#)²⁹. According to [[Rubshtein et al., 2016](#), p. 164], a Lorentz space is separable if and only if $\phi(0+) = 0$ ($\phi \in \Phi_{0+}$). On the other hand, if $\phi(0+) = 0$ and $\phi'(0) = \infty$, the Marcinkiewicz space is not separable. Hence the two spaces do not coincide if $\phi'(0) = \infty$ and the norms are therefore not equivalent. This result implies that for a ϕ with $\phi'(0) = \infty$, not all ri norms are equivalent.

As a sanity check, we show that also $C = \infty$, which is a necessary condition for non-equivalence. Assume by contradiction that $C < \infty$, that is:

$$C = \inf \{c > 0 : \exists \phi_t : \forall \alpha \in (0, 1] : \phi(\alpha) \leq c \cdot \phi_t(\alpha)\} < \infty. \quad (2.241)$$

Then $\forall \epsilon > 0 : \exists t : \forall \alpha \in (0, 1] : \phi(\alpha) \leq (C + \epsilon) \cdot \phi_t(\alpha)$. Fix some ϵ . With this choice of t ,

$$1 \geq \sup_{0 < \alpha \leq 1} \begin{cases} \frac{\phi(\alpha)}{\alpha} \frac{t}{(C+\epsilon) \cdot \phi(t)} & , \alpha \leq t \\ \frac{\phi(\alpha)}{(C+\epsilon) \cdot \phi(t)} & , \alpha > t. \end{cases} \quad (2.242)$$

If the supremum occurs in the first term, it occurs as $\alpha \rightarrow 0$ due to the quasiconcavity of ϕ . Recall that $\phi'(0) = \lim_{\alpha \rightarrow 0} \frac{\phi(\alpha)}{\alpha}$ by definition of the difference quotient. Therefore $1 \geq \sup_{0 < \alpha \leq t} \frac{\phi(\alpha)}{\alpha} \frac{t}{(C+\epsilon) \cdot \phi(t)} = \lim_{\alpha \rightarrow 0} \frac{\phi(\alpha)}{\alpha} \frac{t}{(C+\epsilon) \cdot \phi(t)} = \phi'(0) \frac{t}{(C+\epsilon) \cdot \phi(t)} = \infty$, a contradiction. Thus no finite C exists. Finally, the inequality $\|X\|_{M_\phi} \leq \|X\|_{\Lambda_\phi}$ stems from the embedding theorem. ■

PROOF OF THEOREM 60

Proof Assume that $\|\cdot\|_{\mathcal{R}_1} = \mathcal{L}^1$. Let $\mathcal{Z}_2 = \{\phi_{2\zeta}\}$ be the Kusuoka set of $\|\cdot\|_{\mathcal{R}_2}$. Then, since $\{\alpha \mapsto \alpha\}$ is a Kusuoka set for the $\mathcal{L}^1$ norm:

$$C = \sup_{\phi_{2\zeta}} \sup_{0 < \alpha \leq 1} \frac{\phi_{2\zeta}(\alpha)}{\alpha}. \quad (2.243)$$

²⁹Note that [Rubshtein et al. \[2016\]](#) denote the Marcinkiewicz space with fundamental function V as M_{V^*} . With this notation, the associate space to the Lorentz space Λ_V is M_V . In our notation, however, the associate relationship is $\Lambda'_V = M_{V^*}$, that is, the subscript indicates the fundamental function.

Due to the concavity of $\phi_{2\zeta}$, the fraction is decreasing in α and hence

$$C = \sup_{\phi_{2\zeta}} \lim_{\alpha \rightarrow 0} \frac{\phi_{2\zeta}(\alpha)}{\alpha} = \sup_{\phi_{2\zeta}} \phi'_{2\zeta}(0). \quad (2.244)$$

It remains to show that $\phi'_2(0) < \infty \Rightarrow \sup_{\phi_{2\zeta}} \phi'_{2\zeta}(0) < \infty \quad \forall \phi_{2\zeta} \in \mathcal{Z}_2$. We have

$$\phi_2(\alpha) = \sup_{\mathcal{Z}_2} \{\phi_{2\zeta}(\alpha)\}. \quad (2.245)$$

We know that

$$\frac{\phi_2(t)}{t} \geq \frac{\phi_{2\zeta}(t)}{t} \quad \forall t \quad \forall \phi_{2\zeta}, \quad (2.246)$$

since we know the limit of the left hand side exists, we have

$$\infty > \phi'_2(0) = \lim_{h \rightarrow 0} \frac{\phi_2(t)}{t} \geq \lim_{h \rightarrow 0} \frac{\phi_{2\zeta}(t)}{t} = \phi'_{2\zeta}(0) \quad \forall \phi_{2\zeta}. \quad (2.247)$$

Therefore C is finite and we obtain $\|X\|_{\mathcal{L}^1} \leq \|X\|_{\mathcal{R}_1} \leq C \cdot \|X\|_{\mathcal{L}^1} \quad \forall X \in \mathcal{L}^1$. As a consequence, any two ri norms with finite $\phi'_1(0), \phi'_2(0)$ are equivalent. ■

PROOF OF THEOREM 62

Proof First we show that $C = \infty$. Denote by $\phi_{TM,t}, \phi_{M,t}$ the respective Kusuoka sets, constructed as:

$$\forall t \in (0, 1] : \phi_{TM,t}(x) := \begin{cases} \phi(t)^{\frac{x}{t}} & , x \leq t \\ \frac{1-\phi(t)}{1-t}x + \frac{\phi(t)-t}{1-t} & , x > t \end{cases}, \quad \phi_{M,t}(x) := \begin{cases} \phi(t)^{\frac{x}{t}} & , x \leq t \\ \phi(t) & , x > t. \end{cases} \quad (2.248)$$

The desired constant is

$$C = \sup_{\phi_{TM,t}} \inf_{\phi_{M,t'}} \sup_{0 < \alpha \leq 1} \frac{\phi_{TM,t}(\alpha)}{\phi_{M,t'}(\alpha)} \quad (2.249)$$

$$= \inf \{c > 0 : \forall \phi_{TM,t} \exists \phi_{M,t'} : \forall \alpha \in (0, 1] : \phi_{TM,t}(\alpha) \leq c \cdot \phi_{M,t'}(\alpha)\}. \quad (2.250)$$

We show that no such finite constant can exist. Let some t be given. We wish to find t' such that $\forall \alpha \in (0, 1] : \phi_{TM,t}(\alpha) \leq K \cdot \phi_{M,t'}(\alpha)$. The argument in Theorem 57 shows that the smallest feasible K with the Marcinkiewicz norm on the right hand side is $K = 1/\phi_{TM,t}(1/\phi'_{TM,t}(0))$. Since $\phi'_{TM,t}(0) = \frac{\phi(t)}{t}$, $K = 1/\phi_{TM,t}(\frac{t}{\phi(t)})$. Now $\frac{t}{\phi(t)} > t$ unless $t = 1$

(or we have $\phi(t) = 1$) so that by definition of $\phi_{TM,t}$

$$\frac{1}{K} = \frac{1 - \phi(t)}{1 - t} \frac{t}{\phi(t)} + \frac{\phi(t) - t}{1 - t}. \quad (2.251)$$

For each fixed t , this is the best feasible constant (Theorem 57) and it is finite. However, as $t \rightarrow 0$, we find that $\lim_{t \rightarrow 0} K = \frac{\phi(t)}{t} = \phi'(0) = \infty$. As $t \rightarrow 0$, we need not consider the case of $\phi(t) = 1$, since in the limit this condition does not hold (noting that $\phi \in \Phi_{0+}$). We conclude that $C = \infty$.

We now show that, despite $C = \infty$, the Marcinkiewicz and PTE Marcinkiewicz norm are equivalent. Note that

$$\|X\|_{TM_\phi} = \sup_{0 < t < 1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega) \, d\omega + \frac{\phi(t) - 1}{t - 1} \int_t^1 X^*(\omega) \, d\omega \right\} \quad (2.252)$$

$$\leq \|X\|_{M_\phi} + \sup_{0 < t < 1} \left\{ \frac{\phi(t) - 1}{t - 1} \int_t^1 X^*(\omega) \, d\omega \right\}. \quad (2.253)$$

Knowing that $\|X\|_{M_\phi} \leq \|X\|_{TM_\phi}$, the norms could only *not* be equivalent if $\|X\|_{M_\phi} < \infty$ for some X , while $\|X\|_{TM_\phi} = \infty$. Such an X cannot exist, since the second term in Equation 2.253 behaves nicely: $\int_t^1 X^*(\omega) \, d\omega \leq \int_0^1 X^*(\omega) \, d\omega < \infty$ due to $\mathcal{L}^1$ being the largest ri space, in which any Marcinkiewicz space is embedded. Furthermore, the factor $\frac{\phi(t)-1}{t-1}$ does not exhibit pathological behaviour. Noting that ϕ is bounded from below and above, we only have to check the limits as $t \rightarrow 0$ and $t \rightarrow 1$:

$$\lim_{t \rightarrow 0} \frac{\phi(t) - 1}{t - 1} = 1, \quad \lim_{t \rightarrow 1} \frac{\phi(t) - 1}{t - 1} = \phi'(1) < \infty. \quad (2.254)$$

As $t \rightarrow 0$, we can apply the quotient rule, since both limits exist. As $t \rightarrow 1$, we use L'Hôpital's rule. Hence we conclude that the sets of functions, for which the Marcinkiewicz and PTE Marcinkiewicz norms are finite, coincide. Therefore the norms are equivalent [Bennett and Sharpley, 1988, p. 7]. ■

2.8.5 PROPERTIES OF QUASICONCAVE FUNCTIONS

PROOF OF LEMMA 64

Proof By quasiconcavity, and assumption, we have $\phi(0) = 0$, $\phi(1) = 1$ and

$$0 \leq t_0 \leq t_1 \Rightarrow \phi(t_0) \leq \phi(t_1) \text{ and } \phi(t_0)/t_0 \geq \phi(t_1)/t_1.$$

Taking $t_1 = 1$ we have $\phi(t_0) \leq \phi(t_1) = 1$. Taking $1 = t_0 \leq t_1$ implies $1 = \phi(1) \leq \phi(t_1)$. Furthermore, $t_0 \leq t_1 = 1$ implies $\phi(t_0)/t_0 \geq 1$ which implies $\phi(t_0) \geq t_0$. Additionally, $1 = t_0 \leq t_1$ implies $1 \geq \phi(t_1)/t_1$ and thus $\phi(t_1) \leq t_1$. Combining all these facts we have shown

$$\begin{aligned} t \leq 1 &\Rightarrow t \leq \phi(t) \leq 1 \\ t \geq 1 &\Rightarrow 1 \leq \phi(t) \leq t. \end{aligned}$$

For $0 \leq t \leq 1$, $t = 1 \wedge t$ and $1 = 1 \vee t$. For $t \geq 1$, $1 = 1 \wedge t$ and $t = 1 \vee t$. Hence Lemma 2.153 holds. ■

PROOF OF LEMMA 67

Proof Since max and min are continuous, and the composition of continuous functions is continuous, we have that $t \mapsto \bigwedge_{i \in [n]} \phi_i(t)$ and $t \mapsto \bigvee_{i \in [n]} \phi_i(t)$ are continuous for all $t > 0$. Furthermore, min and max are increasing (i.e. non-decreasing) in each argument, and the composition of increasing functions is increasing, and thus $t \mapsto \bigwedge_{i \in [n]} \phi_i(t)$ and $t \mapsto \bigvee_{i \in [n]} \phi_i(t)$ are increasing. Suppose $t_0 \leq t_1$. Let $i^* = \arg \max_i \phi_i(t_1)/t_1$. Then

$$\frac{\bigvee_{i \in [n]} \phi_i(t_0)}{t_0} \geq \frac{\phi_{i^*}(t_0)}{t_0} \geq \frac{\phi_{i^*}(t_1)}{t_1} = \frac{\bigvee_{i \in [n]} \phi_i(t_1)}{t_1},$$

and thus $t \mapsto \left(\bigvee_{i \in [n]} \phi_i(t) \right) / t$ is decreasing. A similar argument holds for $\bigwedge_i \phi_i$. ■

PROOF OF LEMMA 69

Proof (If): For $\alpha > 0$, $\check{\psi}(\alpha x, \alpha y) = \alpha y \psi(\alpha x / (\alpha y)) = \alpha \check{\psi}(x, y)$, and thus $\check{\psi}$ is positively homogeneous. Furthermore, $\check{\psi}(x, y)$ is nondecreasing in each argument as we now show. Let $y \in \mathbb{R}_{\geq 0}$ be arbitrary but fixed and consider $x \mapsto \check{\psi}(x, y) = y \psi(x/y)$. This is nondecreasing since ψ is nondecreasing. Now let $x \in \mathbb{R}_{\geq 0}$ be arbitrary but fixed and let $g(y) := y \psi(x/y)$. Observe that $z \mapsto g(1/z) = \psi(xz)/z$ which is nonincreasing (since ψ is quasiconcave). Hence $y \mapsto g(y)$ is nondecreasing. Thus $\check{\psi}$ is nondecreasing in both of its arguments concluding the demonstration that $\check{\psi} \in \mathcal{P}$.

(Only if): If $x \leq y$ then $\psi(x) = \check{\psi}(x, 1) \leq \check{\psi}(y, 1) = \psi(y)$ and thus ψ is nondecreasing.

Since $\check{\psi}$ is positively homogeneous and nonzero, $\psi(s) \neq 0$ unless $s = 0$. Finally, for $x \leq y$,

$$\frac{\psi(x)}{x} = \frac{\check{\psi}(x, 1)}{x} = \check{\psi}\left(1, \frac{1}{x}\right) \geq \check{\psi}\left(1, \frac{1}{y}\right) = \frac{\check{\psi}(y, 1)}{y} = \frac{\psi(y)}{y},$$

which shows that $x \mapsto \psi(x)/x$ is nonincreasing, demonstrating that ψ is quasiconcave. ■

PROOF OF LEMMA 70

Proof (If): Since ψ is quasiconcave, it is continuous everywhere except at the origin and thus so is $t \mapsto \phi_1(t)\psi(\phi_0(t)/\phi_1(t))$. Furthermore, since ϕ_0 and ϕ_1 are nondecreasing, and $\check{\psi}$ is nondecreasing in each argument, then $f_{\phi_0, \phi_1}(t)$ is nondecreasing in t . Finally, we need to show that

$$t_0 \leq t_1 \Rightarrow \frac{\check{\psi}(\phi_0(t_0), \phi_1(t_0))}{t_0} \geq \frac{\check{\psi}(\phi_0(t_1), \phi_1(t_1))}{t_1}. \quad (2.255)$$

Since $\check{\psi}$ is positively homogeneous, Equation 2.255 is equivalent to

$$t_0 \leq t_1 \Rightarrow \check{\psi}\left(\frac{\phi_0(t_0)}{t_0}, \frac{\phi_1(t_0)}{t_0}\right) \geq \check{\psi}\left(\frac{\phi_0(t_1)}{t_1}, \frac{\phi_1(t_1)}{t_1}\right).$$

But by assumption on ϕ_0 and ϕ_1 we have

$$t_0 \leq t_1 \Rightarrow \frac{\phi_0(t_0)}{t_0} \geq \frac{\phi_0(t_1)}{t_1} \quad \text{and} \quad \frac{\phi_1(t_0)}{t_0} \geq \frac{\phi_1(t_1)}{t_1}.$$

Since $\check{\psi}$ is nondecreasing in each argument, we can thus conclude that Equation 2.255 holds, thus demonstrating the final property needed to show quasiconcavity of f_{ϕ_0, ϕ_1} .

(Only if): Using the definition of $\check{\psi}$ we have $f_{\phi_0, \phi_1}(t) = \phi_1(t)\psi\left(\frac{\phi_0(t)}{\phi_1(t)}\right)$. We need to show that $[\forall \phi_0, \phi_1 \in \mathcal{Q}, f_{\phi_0, \phi_1} \in \mathcal{Q}] \Rightarrow \psi \in \mathcal{Q}$. Now $f_{\phi_0, \phi_1} \in \mathcal{Q}$ means that

1. $t_0 \leq t_1 \Rightarrow \phi_1(t_0)\psi\left(\frac{\phi_0(t_0)}{\phi_1(t_0)}\right) \leq \phi_1(t_1)\psi\left(\frac{\phi_0(t_1)}{\phi_1(t_1)}\right)$.
2. $t_0 \leq t_1 \Rightarrow \frac{\phi_1(t_0)}{t_0}\psi\left(\frac{\phi_0(t_0)}{\phi_1(t_0)}\right) \geq \frac{\phi_1(t_1)}{t_1}\psi\left(\frac{\phi_0(t_1)}{\phi_1(t_1)}\right)$.
3. $f_{\phi_0, \phi_1}(t) = 0 \Leftrightarrow t = 0$.

Choose $\phi_1(t) = t$ and $\phi_0 \in \mathcal{Q}$. Then condition 2 above requires that

$$t_0 \leq t_1 \Rightarrow \psi\left(\frac{\phi_0(t_0)}{t_0}\right) \geq \psi\left(\frac{\phi_0(t_1)}{t_1}\right).$$

But since $\phi_0 \in \mathcal{Q}$, $t_0 \leq t_1 \Rightarrow \frac{\phi_0(t_0)}{t_0} \geq \frac{\phi_0(t_1)}{t_1}$ and thus ψ must be nondecreasing.

Now choose $\phi_1(t) = 1$ and $\phi_0(t) = t$. Then the second condition implies

$$t_0 \leq t_1 \Rightarrow \frac{\psi(t_0)}{t_0} \geq \frac{\psi(t_1)}{t_1}.$$

Furthermore, with the same choice for ϕ_0 and ϕ_1 , we have $f_{\phi_0, \phi_1}(t) = 0 \Leftrightarrow t = 0$ which implies that $\psi(t) = 0 \Leftrightarrow t = 0$. We have thus shown all the required properties of quasi-concavity for ψ . $\blacksquare$

2.8.6 INTERPOLATION FUNCTORS AND THEIR FUNDAMENTAL FUNCTIONS

PROOF OF LEMMA 76

Proof Recalling the definition of PTE (2.94), we show that for any $X \in \mathcal{M}^+$ and $c \in \mathbb{R}$ such that $X + c \geq 0$ that $\|X + c\|_{\Lambda_{\phi(\bar{\mathcal{X}})}} = \|X\|_{\Lambda_{\phi(\bar{\mathcal{X}})}} + c$, where $\bar{\mathcal{X}} = (\mathcal{X}_0, \mathcal{X}_1)$. First consider the case that $c \geq 0$. Writing $\|\cdot\|_{\Lambda} = \|\cdot\|_{\Lambda_{\phi(\bar{\mathcal{X}})}}$, $\|\cdot\|_0 = \|\cdot\|_{\mathcal{X}_0}$ and $\|\cdot\|_1 = \|\cdot\|_{\mathcal{X}_1}$ for brevity we have

$$\begin{aligned} \|X\|_{\Lambda} + c &= \inf \left\{ \sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) + c : X_k \in \mathcal{X}_0 + \mathcal{X}_1, K \in \mathbb{N}, X = \sum_k X_k \right\} \\ &= \inf \left\{ \sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) + \phi(c, c) : X_k \in \mathcal{X}_0 + \mathcal{X}_1, K \in \mathbb{N}, X = \sum_k X_k \right\} \end{aligned}$$

since $\phi(1, 1) = 1$ and ϕ is positively homogeneous,

$$\begin{aligned} &= \inf \left\{ \sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) + \sum_k \phi\left(\frac{c}{K}, \frac{c}{K}\right) : X_k \in \mathcal{X}_0 + \mathcal{X}_1, K \in \mathbb{N}, X = \sum_k X_k \right\} \\ &= \inf \left\{ \sum_k^K [\phi(\|X_k\|_0, \|X_k\|_1) + \phi\left(\frac{c}{K}, \frac{c}{K}\right)] : X_k \in \mathcal{X}_0 + \mathcal{X}_1, K \in \mathbb{N}, X = \sum_k X_k \right\} \\ &\leq \inf \left\{ \sum_k^K \phi(\|X_k\|_0 + \frac{c}{K}, \|X_k\|_1 + \frac{c}{K}) : X_k \in \mathcal{X}_0 + \mathcal{X}_1, K \in \mathbb{N}, X = \sum_k X_k \right\} \end{aligned}$$

since $\phi(z_1) + \phi(z_2) \leq \phi(z_1 + z_2)$ for arbitrary $z_1, z_2 \in \mathbb{R}_+^2$ because ϕ is a concave gauge function [Barbara and Crouzeix, 1994, Proposition 2.1]

$$= \inf \left\{ \sum_k^K \phi(\|X_k + \frac{c}{K}\|_0, \|X_k + \frac{c}{K}\|_1) : X_k \in \mathcal{X}_0 + \mathcal{X}_1, K \in \mathbb{N}, X = \sum_k X_k \right\}$$

since $\|\cdot\|_0$ and $\|\cdot\|_1$ are positive translation equivariant. Now let $X'_k = X_k + \frac{c}{K}$, for k and thus $X_k = X'_k - \frac{c}{K}$, k . Thus

$$\begin{aligned}
&= \inf \left\{ \sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) : X'_k \in \mathcal{X}_0 + \mathcal{X}_1, K \in \mathbb{N}, X = \sum_k (X'_k - \frac{c}{K}) \right\} \\
&= \inf \left\{ \sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) : X'_k \in \mathcal{X}_0 + \mathcal{X}_1, K \in \mathbb{N}, X = \left(\sum_k X'_k \right) - c \right\} \\
&= \inf \left\{ \sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) : X'_k \in \mathcal{X}_0 + \mathcal{X}_1, K \in \mathbb{N}, X + c = \left(\sum_k X'_k \right) \right\} \\
&= \|X + c\|_\Lambda.
\end{aligned}$$

Since $\|\cdot\|_\Lambda$ is a norm it satisfies the triangle inequality $\|X + c\|_\Lambda \leq \|X\|_\Lambda + \|c\|_\Lambda = \|X\|_\Lambda + c$. Thus combining with the above we have

$$\|X\|_\Lambda + c \leq \|X + c\|_\Lambda \leq \|X\|_\Lambda + c$$

and thus $\|X + c\|_\Lambda = \|X\|_\Lambda + c$ as required.

If instead we have $c < 0$ but $X + c \geq 0$, then there exists some $c_0 \geq 0$ such that $c_0 \geq -c$ and $X = X_0 + c_0$, with $X_0 \in \mathcal{M}_+$. Consequently,

$$\|X + c\|_\Lambda = \|X_0 + c_0 + c\|_\Lambda = \|X_0 + (c_0 + c)\|_\Lambda \stackrel{(*)}{=} \|X_0\|_\Lambda + (c_0 + c) \stackrel{(**)}{=} \|X_0 + c_0\|_\Lambda + c = \|X\|_\Lambda + c,$$

where $(*)$ holds from the case already shown, since $X_0 \geq 0$ and $(c_0 + c) \geq 0$, and $(**)$ holds similarly (since $c_0 \geq 0$). ■

PROOF OF LEMMA 78

Proof For $t > 0$, we need to compute

$$\phi_{M_{\bar{\phi}}(\bar{\mathcal{X}})}(t) = \|\chi_{[0,t]}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} = \sup_{s_0, s_1 \geq 0} \frac{K(s_0, s_1, \chi_{[0,t]}, \bar{\mathcal{X}})}{\bar{\phi}^*(s_0, s_1)}.$$

Thus for arbitrary s_0, s_1 we need to determine

$$K(s_0, s_1, \chi_{E_t}, \bar{\mathcal{X}}) = \inf(s_0 \|X_0\|_{\mathcal{X}_0} + s_1 \|X_1\|_{\mathcal{X}_1}),$$

with the infimum taken over all X_0, X_1 such $X_0 + X_1 = \chi_{E_t}$, where E_t is chosen such that $\mu(E_t) = t$. Since by Theorem 37, the Marcinkiewicz norm minorises any ri norm with fundamental function $\phi_{\mathcal{X}}$: $\|X_i\|_{M_{\phi}} \leq \|X\|_{\mathcal{X}}$, we have

$$K(s_0, s_1, \chi_{E_t}, \bar{\mathcal{X}}) \geq \inf_{X_0+X_1=\chi_{E_t}} s_0 \|X_0\|_{M_{\phi_0}} + s_1 \|X_1\|_{M_{\phi_1}},$$

and since $\|X\|_{M_{\phi}} = \sup_{0 < r < \infty} X^{**}(r)\bar{\phi}(r)$, we have

$$\begin{aligned} K(s_0, s_1, \chi_{E_t}, \bar{\mathcal{X}}) &\geq \inf_{X_0+X_1=\chi_{E_t}} s_0 \left(\sup_{0 < r < \infty} X_0^{**}(r)\phi_0(r) \right) + s_1 \left(\sup_{0 < r < \infty} X_1^{**}(r)\phi_1(r) \right) \\ &\geq \inf_{X_0+X_1=\chi_{E_t}} \sup_{0 < r < \infty} (s_0\phi_0(r)X_0^{**}(r) + s_1\phi_1(r)X_1^{**}(r)) \end{aligned}$$

and by choosing $r = t$ we obtain

$$K(s_0, s_1, \chi_{E_t}, \bar{\mathcal{X}}) \geq \inf_{X_0+X_1=\chi_{E_t}} (c_0 X_0^{**}(t) + c_1 X_1^{**}(t)),$$

where $c_0 = s_0\phi_0(t)$ and $c_1 = s_1\phi_1(t)$ are constants (since t is fixed),

$$= \inf_{X_0+X_1=\chi_{E_t}} ((c_0 X_0)^{**}(t) + (c_1 X_1)^{**}(t)).$$

But $(f + g)^{**}(t) \leq f^{**}(t) + g^{**}(t)$ for all $t > 0$ and any f, g , and so

$$K(s_0, s_1, \chi_{E_t}, \bar{\mathcal{X}}) \geq \inf_{X_0+X_1=\chi_{E_t}} (c_0 X_0 + c_1 X_1)^{**}(t).$$

Since for any f we have $f^*(t) \leq f^{**}(t)$, for all t , we have

$$\begin{aligned} K(s_0, s_1, \chi_{E_t}, \bar{\mathcal{X}}) &\geq \inf_{X_0+X_1=\chi_{E_t}} (c_0 X_0 + c_1 X_1)^*(t) \\ &= \inf_{X_0+X_1=\chi_{E_t}} \inf\{\lambda: \mu_{c_0 X_0 + c_1 X_1}(\lambda) \leq t\} \\ &= \inf_{X_0+X_1=\chi_{E_t}} \inf\{\lambda: \mu\{s \in \mathbb{R}: (c_0 X_0 + c_1 X_1)(s) > \lambda\} \leq t\}. \end{aligned}$$

Now let $A_\lambda := \mu\{s \in \mathbb{R}: (c_0 X_0 + c_1 X_1)(s) > \lambda\}$ and $B_\lambda := \mu\{s \in \mathbb{R}: (c_0 \wedge c_1)(X_0 + X_1)(s) > \lambda\}$. Since $(c_0 \wedge c_1)(X_0 + X_1) = (c_0 \wedge c_1)X_0 + (c_0 \wedge c_1)X_1 \leq c_0 X_0 + c_1 X_1$ we have that $B_\lambda \leq A_\lambda$ for all λ . Furthermore, $\lambda \mapsto A_\lambda$ and $\lambda \mapsto B_\lambda$ are nonincreasing and thus $\inf\{\lambda: A_\lambda \leq t\} \geq \inf\{\lambda: B_\lambda \leq t\}$, and hence

$$\begin{aligned} K(s_0, s_1, \chi_{E_t}, \bar{\mathcal{X}}) &\geq \inf_{X_0 + X_1 = \chi_{E_t}} \inf\{\lambda: \mu\{s \in \mathbb{R}: (c_0 \wedge c_1)(X_0 + X_1)(s) > \lambda\} \leq t\} \\ &= \inf_{X_0 + X_1 = \chi_{E_t}} (c_0 \wedge c_1)(X_0 + X_1)^*(t) \\ &= (c_0 \wedge c_1)\chi_{[0,t]}(t) \\ &= c_0 \wedge c_1 \\ &= s_0 \phi_0(t) \wedge s_1 \phi_1(t). \end{aligned}$$

The infimum in the definition of K is in fact attained by choosing $X_0 = \alpha \chi_{[0,t]}$ and $X_1 = (1 - \alpha) \chi_{[0,t]}$ for some $\alpha \in [0, 1]$.

In this case we have

$$\begin{aligned} &\inf_{\alpha \in [0,1]} s_0 \|\alpha \chi_{E_t}\|_{\mathcal{X}_0} + s_1 \|(1 - \alpha) \chi_{E_t}\|_{\mathcal{X}_1} \\ &= \inf_{\alpha \in [0,1]} s_0 \alpha \phi_0(t) + s_1 (1 - \alpha) \phi_1(t) \\ &= s_0 \phi_0(t) \wedge s_1 \phi_1(t). \end{aligned}$$

Thus $K(s_0, s_1, \chi_{[0,t]}, \bar{\mathcal{X}}) = s_0 \phi_0(t) \wedge s_1 \phi_1(t)$.

Consequently

$$\|\chi_{[0,t]}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} = \sup_{s_0, s_1 \geq 0} \frac{s_0 \phi_0(t) \wedge s_1 \phi_1(t)}{\bar{\phi}^*(s_0, s_1)}.$$

Noting that both numerator and denominator are positively homogeneous in (s_0, s_1) it suffices to enforce $s_0 + s_1 = 1$ and thus by setting $s_0 = s$ and $s_1 = (1 - s)$ for $s \in [0, 1]$,

$$\|\chi_{[0,t]}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} = \sup_{s \in [0,1]} \frac{s \phi_0(t) \wedge (1 - s) \phi_1(t)}{\bar{\phi}^*(s, 1 - s)}.$$

Now $\bar{\phi}^*(\alpha, \beta) = \beta \bar{\psi}(\alpha/\beta)$ for some $\bar{\psi} \in \mathcal{Q}$ and so $\bar{\phi}^*(s, 1 - s) = (1 - s) \bar{\psi}(s/(1 - s))$. Furthermore $s \phi_0(t) \wedge (1 - s) \phi_1(t)$ can be written as

$$\begin{aligned} &s \phi_0(t) \quad \text{if } s \phi_0(t) \leq (1 - s) \phi_1(t) \\ &(1 - s) \phi_1(t) \quad \text{if } s \phi_0(t) \geq (1 - s) \phi_1(t). \end{aligned}$$

Setting $\gamma := \phi_1(t)/\phi_0(t)$, we have $s \phi_0(t) \leq (1 - s) \phi_1(t) \Leftrightarrow s/(1 - s) \leq \gamma \Leftrightarrow s \leq \gamma/(1 + \gamma)$.

Hence

$$s\phi_0(t) \wedge (1-s)\phi_1(t) = \begin{cases} s\phi_0(t), & s \leq \frac{\gamma}{1+\gamma} \\ (1-s)\phi_1(t), & s \geq \frac{\gamma}{1+\gamma} \end{cases}.$$

Hence

$$\phi_{M_{\bar{\phi}}(\bar{\mathcal{X}})}(t) = \|\chi_{[0,t]}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} = \min \left(\sup_{s \leq \frac{\gamma}{1+\gamma}} \frac{s\phi_0(t)}{(1-s)\bar{\psi}\left(\frac{s}{1-s}\right)}, \sup_{s \geq \frac{\gamma}{1+\gamma}} \frac{(1-s)\phi_1(t)}{(1-s)\bar{\psi}\left(\frac{s}{1-s}\right)} \right).$$

Since $\phi_0(t), \phi_1(t) \geq 0$, we only need to determine

$$a := \sup_{0 \leq s \leq \frac{\gamma}{1+\gamma}} f(s) \text{ and } b := \sup_{\frac{\gamma}{1+\gamma} \leq s < \infty} g(s),$$

where $f(s) = \frac{s}{(1-s)\bar{\psi}(s/(1-s))}$ and $g(s) = \frac{1}{\bar{\psi}(s/(1-s))}$, and $\phi_{M_{\bar{\phi}}(\bar{\mathcal{X}})}(t) = a\phi_0(t) \wedge b\phi_1(t)$. Considering f first, and setting $t := s/(1-s)$ and so $s = t/(1+t)$ we have

$$a = \sup_{t \in [0, \gamma]} \frac{t}{\bar{\psi}(t)}.$$

But $\bar{\psi}$ is quasiconcave and thus $t \mapsto \bar{\psi}(t)/t$ is positive and nonincreasing and so $t \mapsto t/\bar{\psi}(t)$ is nondecreasing and the supremum is attained at $t = \gamma$ and $a = \gamma/\bar{\psi}(\gamma)$. Similarly for g , we have

$$b = \sup_{t \geq \gamma} \frac{1}{\bar{\psi}(t)}.$$

Since $\bar{\psi}$ is quasiconcave, it is positive and nondecreasing and so $t \mapsto 1/\bar{\psi}(t)$ is nonincreasing and the supremum is attained at $t = \gamma$ and $b = 1/\bar{\psi}(\gamma)$. Recalling $\gamma = \phi_1(t)/\phi_0(t)$, we have

$$\|\chi_{[0,t]}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} = \min \left(\frac{\gamma\phi_0(t)}{\bar{\psi}(\gamma)}, \frac{\phi_1(t)}{\bar{\psi}(\gamma)} \right) = \min \left(\frac{\phi_1(t)}{\bar{\psi}(\gamma)}, \frac{\phi_1(t)}{\bar{\psi}(\gamma)} \right) \quad (2.256)$$

$$= \frac{\phi_1(t)}{\bar{\psi}(\phi_1(t)/\phi_0(t))} = \frac{\phi_1(t)\phi_0(t)}{\bar{\phi}^*(\phi_1(t), \phi_0(t))}. \quad (2.257)$$

But $\bar{\phi}^*$ is positively homogeneous, and thus

$$\|\chi_{[0,t]}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} = \frac{\phi_1(t)\phi_0(t)}{\bar{\phi}^*(\phi_1(t), \phi_0(t))} = \phi_1(t)\phi_0(t)\bar{\phi} \left(\frac{1}{\phi_1(t)}, \frac{1}{\phi_0(t)} \right) \quad (2.258)$$

$$= \bar{\phi} \left(\frac{\phi_1(t)\phi_0(t)}{\phi_1(t)}, \frac{\phi_1(t)\phi_0(t)}{\phi_0(t)} \right) = \bar{\phi}(\phi_0(t), \phi_1(t)). \quad (2.259)$$

■

PROOF OF LEMMA 79

Proof We have

$$\phi_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})}(t) = \|\chi_{E_t}\|_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})},$$

where E_t is such that $\mu(E_t) = t$,

$$= \inf_{(X_k)_k: \chi_{E_t} = \sum_k X_k} \sum_k \bar{\phi}(\|X_k\|_{\mathcal{X}_0}, \|X_k\|_{\mathcal{X}_1}).$$

Taking the particular choice $X_1 = \chi_{[0,t]}$ and $X_k = 0$ for $k > 1$, gives an upper bound on the infimum:

$$\begin{aligned} \phi_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})}(t) &\leq \bar{\phi}(\|\chi_{E_t}\|_{\mathcal{X}_0}, \|\chi_{E_t}\|_{\mathcal{X}_1}) \\ &= \bar{\phi}(\phi_0(t), \phi_1(t)). \end{aligned}$$

■

2.9 APPENDIX: (NON)-EXPECTED UTILITY THEORIES

Coherent risk measures in finance are intimately connected with generalized utility theories in rational choice theory, situated in the context of economics. These theories offer formal axiomatic bases for rational decision making under uncertainty. Our motivation is that we take the following two ideas seriously: empirical risk minimization (ERM) in machine learning is a decision problem not only under *risk*, but also under *ambiguity*, and a loss function is an outcome-contingent *disutility* [Berger, 1985].

In the ERM problem, the decision maker, i.e. the machine learning engineer, faces the problem of summarizing the loss distribution in a single number, which is then employed in a minimization routine. This summary is typically the expectation, reducing to summation under the empirical distribution. Loss is a disutility in the sense that the decision maker wants to have as little as possible of it. Therefore, modulo a sign flip, loss minimization can be described in the framework of expected utility theory, where the aim is to maximize utility in an uncertain setting.

First, if we knew the ‘true’ probability distribution, risk minimization is indeed a decision problem under risk. We can model a risky situation with a probability distribution. In an economics context, the analogy is a choice for the decision maker between different *lotteries* with known probabilities. Think for instance of a coin flip. Probability theory

was historically developed to handle such decisions under risk, where probabilities are “well-behaved”: relative frequencies are stable and can be known, e.g. by combinatorial arguments [Hacking, 1990].

In contrast to risk is the challenge of ambiguity or, in the extreme, Knightian uncertainty. These are “non-probabilized” forms of uncertainty [Etner et al., 2012] and cannot be captured by a single probability distribution. In *empirical* risk minimization, we have no good reason to believe that the observed data perfectly represents the ‘true’ distribution. Moreover, in a dynamically changing environment, such a stable distribution may not exist [Gorban, 2017]. Hence it may be better to assume a whole set of probability distributions to represent the belief of the decision maker, from a subjectivist view, or to represent the behavior of the loss sequence, from a frequentist view³⁰.

In the presence of risk and ambiguity, different attitudes are conceivable: a decision maker might be risk loving, risk-neutral, risk-averse and ambiguity-loving, ambiguity-neutral, ambiguity-averse. We shall focus on a risk-averse and ambiguity-averse attitude. First, we frame the standard ERM problem in the framework of expected utility theory. Throughout, we make the translation to a loss-based formulation. Our aim is to demonstrate the limitations of the classical approach and illuminate attractive alternatives. In particular, we will find yet more ways to arrive at the classes of coherent and spectral risk measures. These new perspectives offer additional motivation for why employing a coherent or spectral risk measure in place of the expectation is normatively permissible and motivated.

2.9.1 EXPECTED UTILITY

Classical expected utility theory comes in two flavors: objective and subjective. In the objective setting developed by von Neumann and Morgenstern [1947], the decision maker chooses between lotteries, which yield specified losses/rewards with known probabilities. In contrast, in the subjective setting of Savage [1954], the decision maker does not know the probability measure a priori. The two formulations differ mainly in interpretation. Mathematically they are closely related. For ease of exposition, we focus on von Neumann’s framework. We refer to Föllmer and Schied [2016] for a detailed account. Denote by C a set of possible consequences. Typically, $C = \mathbb{R}$ and we interpret the elements as monetary outcomes. Assume some σ -algebra $\mathcal{F}$ is given on C . Let $\mathcal{P}$ denote the set of probability distributions over C with finite support.

$$\mathcal{P} = \{P : \mathcal{F} \rightarrow [0, 1] : P(\{c\}) \neq 0 \text{ only for finitely many } c \in C, P(C) = 1\}. \quad (2.260)$$

³⁰For a frequentist interpretation of coherent upper probabilities, see [Walley and Fine, 1982, Fröhlich et al., 2023]. Whereas precise probabilities model converging sequences of relative frequencies, coherent upper probabilities can be linked to sequences whose relative frequencies diverge within an interval, whose boundaries are given by the lower and upper probability.

An element $P \in \mathcal{P}$ is called a *lottery*. We characterize a decision maker by her preference relation $\succsim$ on $\mathcal{P}$. The meaning of $X \succsim Y$ is that the lottery P is preferred over the lottery Q . Similarly, $P \sim Q$ denotes indifference and $P \succ Q$ strict preference. We say that a preference relation $\succsim$ is represented by a functional $R : \mathcal{P} \rightarrow \mathbb{R}$ if

$$P \succsim Q \iff R(P) \leq R(Q), \quad (2.261)$$

Assume that $\succsim$ satisfies the following axioms:

- N1.** $\forall P, Q \in \mathcal{P} : P \succsim Q \text{ or } Q \succsim P \text{ or both.}$ (completeness)
- N2.** $\forall P, Q, S \in \mathcal{P} : P \succsim Q, Q \succsim S \Rightarrow P \succsim S$ (transitivity)
- N3.** $\forall P, Q, S \in \mathcal{P} : \exists \alpha, \beta \in (0, 1) : \alpha P + (1 - \alpha)S \succ Q \succ \beta P + (1 - \beta)S$ (Archimedean)
- N4.** $\forall P, Q, S \in \mathcal{P} : \forall \alpha \in (0, 1] : P \succ Q \Rightarrow \alpha P + (1 - \alpha)S \succ \alpha Q + (1 - \alpha)S$ (independence)

While the first three axioms are relatively uncontroversial and common to different theories of rational choice, it is the independence axiom which characterizes the theory. The independence axiom, essentially equivalent to the *sure thing principle* in Savage [1954], is an additivity principle. Intuitively, it means that the common component $(1 - \alpha)S$ does not matter for the ranking [Al-Najjar and De Castro, 2010]. Another way to express it is that preferences must be separable across mutually exclusive events [Denuit et al., 2006].

Theorem 91. [von Neumann and Morgenstern, 1947]. *If and only if N1-N4 are satisfied, the preference relation $\succsim$ allows an affine representation with a loss function $\ell : C \rightarrow \mathbb{R}$, unique up to an affine transformation:*

$$P \succsim Q \iff \int_C \ell(\omega) dP(\omega) \leq \int_C \ell(\omega) dQ(\omega) \iff \mathbb{E}_P[\ell] \leq \mathbb{E}_Q[\ell]. \quad (2.262)$$

Instead of a utility function u , which is typically used in the literature, we have expressed the theorem using its mirror image, the loss function $\ell(\omega) = -u(-\omega)$.³¹ Instead of expected utility maximization, our decision maker aims for expected loss minimization.

The celebrated theories of von Neumann and Morgenstern [1947] and Savage [1954] have become deeply entrenched in economics and spread into other disciplines in the course of the 20th century. The crucial ingredient is the independence axiom, which corresponds to additivity of the representation. The structure of such a representation implies a strict separation of *belief* and *action* (or *taste*, in the language of Al-Najjar and De Castro [2010]).

³¹This would correspond to also flipping the gain vs. loss orientation of the ω (see e.g. Rockafellar and Uryasev [2013]). This presupposes that C supports a “ $-$ ” operation. An alternative would be to set $\ell(\omega) = -u(\omega)$, which would keep the orientation.

Belief is embodied by the probability distribution; action relates to the choice of the loss function, which specifies the attitude of a decision maker towards outcomes. These two separate domains are then conjoined using the expectation operator. We remark that this separability is at the basis of challenges which have been raised against classical expected utility, such as the Ellsberg's urns [Ellsberg, 1961] and Allais' paradox [Allais, 1953]. The non-expected utility theories which we will consider refrain from making the separation to this extent.

Related is the issue of risk aversion: in expected utility theory, attitudes toward wealth (outcomes $c \in C$) and probabilities are forever bound together. The standard definition of (weak) risk aversion is that

$$\forall P \in \mathcal{P} : \mathbb{E}_P[id] \succsim P, \quad (2.263)$$

$$\Leftrightarrow \ell \left(\int_C \omega \, dP(\omega) \right) \leq \int_C \ell(\omega) \, dP(\omega), \quad (2.264)$$

where id is the identity function and $\mathbb{E}_P[id]$ is a constant lottery which yields the loss $\mathbb{E}_P[id]$ with probability 1. It is a classical result that risk aversity holds if and only if ℓ is a convex function, that is, when u is concave. For an expected utility decision maker, risk aversity is synonymous with diminishing marginal utility of wealth [Denuit et al., 2006], expressed via the utility function. Diminishing marginal utility is the phenomenon that an increase at a higher wealth level is valued less than the same increase at a lower wealth level. This is modelled by a concave utility function, which has a convex loss function as its mirror image. In machine learning, it is customary to employ convex loss functions (with respect to the predictions, not necessarily the parameters) such as the squared loss. This captures the wish to increase punishment the farther the prediction is from the ground truth. Thus, one has in a sense automatically implemented this form of risk aversion. Since there is no sensible unit of 'wealth' in machine learning to establish the analogy to economics, we instead fix a loss function a priori. Then we consider the loss values as making up the space of consequences C and apply expected utility theory with the identity id as a "loss function". In this way, classical expected utility yields the familiar problem of expected risk minimization.

We find this kind of risk aversion too weak; it does not actually capture a risk-averse attitude [Buchak, 2013]. Aversion to risk in the sense of unpredictability seems *prima facie* different from diminishing marginal utility towards wealth, yet in classical expected utility theory they are conflated. A decision maker who has a diminishing attitude towards the amount of some commodity even under certainty seems *prima facie* rational [see Buchak, 2013, for this line of argument]. When then uncertainty enters the picture, the decision maker might display additional risk aversion, a disinclination to take a risky bet on the commodity of interest, which is not exhausted by the concave utility function. To us, risk

aversion amounts to encoding an attitude towards the *probability* itself: a decision maker might prefer a distribution which is less spread-out over a distribution with higher spread, even given that they have the same mean. Risk aversion is asymmetric, however: unexpected high gain is not as problematic as unexpected high loss. It is not clear why the only reason for this preference should arise from diminishing marginal utility instead of from an aversion to the inherent risk.

Another criticism of classical expected utility is *overprecision*, the “excessive faith that you know the truth” [Moore et al., 2015]. We alluded to this problem in Section 2.1.1 and Section 2.2. Using a single probability measure expresses precise belief, when instead sometimes a degree of ignorance is warranted by the available evidence. Gilboa et al. [2009] write: “The Bayesian approach is lacking because it is not rich enough to describe ones degree of confidence in ones assessments”. Here, the Bayesian approach refers to Savage’s axiomatization. Along similar lines, Keynes observes that “new evidence will sometimes decrease the probability of an argument, but it will always increase its ‘weight’” [Keynes, 1921, p. 78]. A decision maker following Savage’s axioms has a precise belief concerning the probability that right now *24 men in Bulgaria are standing on their heads* [Schoenfield, 2012], down to arbitrary precision. Furthermore, she would be willing to take bets both on and against this event, where the betting rate is the specified precise probability. In contrast to this behaviour, a lack of knowledge rather warrants *ambiguity aversion*, a certain pessimism in the face of non-probabilized uncertainty. Hence we now turn to *maxmin expected utility*, closely related to imprecise probability.

2.9.2 MAXMIN EXPECTED UTILITY

An influential generalization of expected utility, *maxmin expected utility*, has been put forward by Gilboa and Schmeidler [1989]. Following Anscombe and Aumann [1963], they work with a two-stage model, comprising both objective and subjective probabilities. Let Ω denote a set of outcomes. Let S denote a set comprising the *states of nature* and let $\mathcal{F}$ an algebra of subsets of S closed under finite intersections. By $\mathcal{P}$, we denote the set of probability distributions on Ω with finite support, i.e. *lotteries* with objective probabilities:

$$\mathcal{P} = \{P : \Omega \rightarrow [0, 1] : P(\omega) \neq 0 \text{ only for finitely many } \omega, \sum_{\omega \in \Omega} P(\omega) = 1\}. \quad (2.265)$$

An *act* is a function $X : S \rightarrow \mathcal{P}$ belonging to some specified convex set of acts $\mathcal{L}$, which includes constant functions. We denote the set of constant acts as $\mathcal{L}_c$. Convex combinations of acts are performed pointwise: let $X, Y \in \mathcal{L}$. Then $\alpha X + (1 - \alpha)Y = \omega \mapsto \alpha X(\omega) + (1 - \alpha)Y(\omega)$. The goal is to obtain a subjective probability about acts, sometimes called *horse lotteries* by leveraging the objective probabilities through the preference relation. Gilboa

and [Schmeidler \[1989\]](#) impose the following axioms:

- M1.** $\forall X, Y \in \mathcal{L} : X \succsim Y$ or $X \precsim Y$ or both. (completeness)
- M2.** $\forall X, Y, Z \in \mathcal{L} : X \succsim Y, Y \succsim Z \Rightarrow X \succsim Z$ (transitivity)
- M3.** $\forall X, Y, Z \in \mathcal{L}$: if $X \succ Y$ and $Y \succ Z$ then $\exists \alpha, \beta \in (0, 1) :$
 $\alpha X + (1 - \alpha)Z \succ Y$ and $Y \succ \beta X + (1 - \beta)Z$ (continuity)
- M4.** If $\forall \omega \in \Omega : \omega' \mapsto X(\omega) \succsim \omega' \mapsto Y(\omega)$ then $X \succsim Y$ (monotonicity) [sic]
- M5.** $\forall X, Y \in \mathcal{L}, c \in \mathcal{L}_c, \alpha \in (0, 1) : X \succ Y \Rightarrow \alpha X + (1 - \alpha)c \succ \alpha Y + (1 - \alpha)c$ (c-independence)
- M6.** $\forall X, Y \in \mathcal{L}, \alpha \in (0, 1) : X \sim Y \Rightarrow \alpha X + (1 - \alpha)Y \succsim X$ (ambiguity aversion)
- M7.** not for all $X, Y \in \mathcal{L} : X \succsim Y$ (non-degeneracy)

where $X \sim Y$ denotes the indifference relation, i.e. $X \succsim Y$ and $Y \succsim X$, and $\succ$ is the strict part of the relation. Certainty independence (c-independence) is strictly weaker than independence; it only requires the separability with respect to constants. As a consequence of this axiom, [Gilboa and Schmeidler \[1989\]](#) obtained the following representation.

Theorem 92. [[Gilboa and Schmeidler, 1989](#)]. *If and only if the preference relation $\succsim$ satisfies M1-M7 then it allows a representation of the form*

$$X \succsim Y \iff \max_{Q \in \mathcal{Q}} \int \ell \circ X \, dQ \leq \max_{Q \in \mathcal{Q}} \int \ell \circ Y \, dQ \quad (2.266)$$

for a loss function $\ell : \mathcal{P} \rightarrow \mathbb{R}$ defined at the level of lotteries and a non-empty closed convex set $\mathcal{Q}$ of finitely additivity probability measures on $\mathcal{F}$.

Because of the translation to losses, it would be more appropriate to call it *minmax* expected loss in our context. Observe that this is essentially nothing but a two-stage formulation of Walley's upper previsions (coherent risk measures) and with a loss function entering the picture. When the loss function is the identity and the functional $R(X) = \max_{Q \in \mathcal{Q}} \int \ell \circ X \, dQ$ is applied only to acts which yield degenerate constant lotteries, i.e. $X(s) = c_s \in \mathcal{P}$, we recover an upper prevision (coherent risk measure). Maxmin expected utility has also recently been formalized in a single-stage subjective setting [[Al-Najjar and De Castro, 2010](#)]. We have chosen to present the two-stage formulation since the axiomatization is simpler there.

A decision maker who adopts the maxmin axioms takes a worst-case stance towards a set of probability measures considered as relevant candidates. For this, the crucial axiom is *ambiguity aversion*. The intuition behind it is that, in financial terms, hedging against

ambiguity is preferred. Consider some $X \sim Y$, which are both ambiguous, i.e. objective probabilities are not known. Ambiguity aversion states that then a convex combination $\alpha X + (1 - \alpha)Y$ is weakly preferred over X or Y . Possibly, X acts as a hedge against Y or vice versa, that is, X tends to yield losses for those states where Y tends to yields gains. In extreme cases, a convex combination of such acts can even reduce the ambiguous situation to a risky one with known probabilities (see e.g. [Föllmer and Weber, 2015] or [Etner et al., 2012] for examples). On the other hand, no hedging is possible when X and Y are comonotone, since they then share the same rank ordering of outcomes³². Ambiguity aversion states that, irrespective of the concrete X and Y , hedging can at least never be strictly worse for the decision maker. The next theory, a close cousin of maxmin expected utility, takes the idea that comonotonicity prevents hedging seriously.

2.9.3 CHOQUET EXPECTED UTILITY

While maxmin expected utility is closely related to coherent risk measures and upper previsions, *Choquet expected utility* contains as important special cases the class of spectral risk measures. The theory was originally developed by Schmeidler [1989] and like maxmin expected utility was set in the two stage model of Anscombe and Aumann [1963]. However, for easier exposition we present the single-stage version of Chateauneuf [1994], translated to losses. This is in contrast to maxmin expected utility, where the single-stage version is significantly more complicated than the two-stage version. Consider a space Ω and a σ -algebra $\mathcal{F}$. The preference relation $\succsim$ is defined on the set $\mathcal{L}$ of bounded, real-valued measurable functions on Ω . Chateauneuf [1994] proposes the following axioms (translated to losses):

CH1. Completeness, transitivity and non-degeneracy

CH2. If $\forall \omega \in \Omega : Y(\omega) \geq X(\omega) \Rightarrow X \succsim Y$ (monotonicity)

CH3. Continuity with respect to monotone uniform convergence, see [Chateauneuf, 1994].

CH4. $\forall X, Y, Z \in \mathcal{L}$: If X and Z are comonotone, Y and Z are comonotone and $X \sim Y$, then $X + Z \sim Y + Z$ (comonotonic independence),

Compared to maxmin expected utility, certainty independence has here been strengthened to comonotonic independence and uncertainty aversion has been dropped. Recall again that two functions X and Y are comonotone if

$$(X(\omega) - X(\omega'))(Y(\omega) - Y(\omega')) \geq 0 \quad \forall \omega, \omega' \in \Omega \quad (2.267)$$

³²We here rely on the intuitive understanding of perfect rank correlation. For the definition of comonotonicity when outcomes are lotteries see [Schmeidler, 1989].

A constant function is comonotone with any other function and therefore comonotonic independence implies certainty independence. In the words of [Chateauneuf \[1994\]](#), “comonotonic independence requires the direction of preference to be retained under adding payments, provided hedging is not involved”. When X and Y are comonotone, neither can work as a hedge against the other due to perfect rank correlation. As a consequence, the ambiguity cannot be reduced in favor of risk. [Chateauneuf \[1994\]](#) obtains the following representation result. Recall that a capacity is a set function with $\bar{\mu}(\emptyset) = 0$ and $\bar{\mu}(\Omega) = 1$,³³ which is monotone.

Theorem 93. [[Chateauneuf, 1994](#)]. *If and only if $\succsim$ satisfies the above axioms, there exists a capacity $\bar{\mu}$ on $\mathcal{F}$ such that*

$$X \succsim Y \iff \int X \, d\bar{\mu} \leq \int Y \, d\bar{\mu}, \quad (2.268)$$

where the Choquet integral with respect to the capacity $\bar{\mu}$ is defined as

$$\int X \, d\bar{\mu} := \int_{-\infty}^0 [\bar{\mu}(\{X \geq x\}) - 1] \, dx + \int_0^{\infty} \bar{\mu}(\{X \geq x\}) \, dx. \quad (2.269)$$

Remark 94. In this single-stage formulation, no loss/utility function has entered the picture. Typically, Choquet expected utility refers to representations of the form $\int u \circ X \, d\bar{\mu}$ with a utility function u . For instance, cf. the axiomatization of [Schmeidler \[1989\]](#). To us, this is not a relevant difference since in our machine learning setup the random variable X directly represents a loss.

Compare this to Equation 2.40, where the capacity is given as the composition of a concave function and a probability measure. The capacity then determines whether the decision maker is ambiguity-averse, -neutral or -loving. Consider the uncertainty aversion axiom

CH5. $\forall X, Y, Z \in \mathcal{L}$: If $X \sim Y$ and Y and Z are comonotone, then $X + Z \succsim Y + Z$

The intuition is that Z cannot act as a hedge against Y , but it could possibly hedge against X , so the direction of preference turns at least weakly in favor of $X + Z$. This axiom is in some sense a combination of comonotonic independence and uncertainty aversion.

Theorem 95. [[Chateauneuf, 1994](#)]. *If and only if $\succsim$ satisfies CH1-CH3 and CH5, then the representation of Theorem 93 holds and the capacity $\bar{\mu}$ is furthermore submodular, that is:*

$$\bar{\mu}(A \cup B) + \bar{\mu}(A \cap B) \leq \bar{\mu}(A) + \bar{\mu}(B). \quad (2.270)$$

To get an intuition for the Choquet integral, let us consider a finite space $\Omega = \{\omega_1, \dots, \omega_n\}$. Assume X is a step function which takes on the values $x_1 \leq x_2 \leq \dots \leq x_n$. Let $x_0 = 0$. Then

³³The normalization $\bar{\mu}(\emptyset) = 0$ is required for any capacity. Capacities with $\bar{\mu}(\Omega) = 1$ are also called *normalized capacities*. We impose $\bar{\mu}(\Omega) = 1$ throughout, however, and simply call it a capacity.

the Choquet integral can be written as

$$\int X \, d\bar{\mu} = \sum_i^n (x_i - x_{i-1}) \bar{\mu}(\{X \geq x_i\}). \quad (2.271)$$

If the capacity is a probability measure, this reduces to the usual expectation. The decision maker starts with the lowest loss value x_1 and then successively adds up the increments $x_i - x_{i-1}$, but weighted with the capacity. In particular a capacity need not be additive for disjoint events, which allows to model interaction effects such as hedging against ambiguity.

Consider the special case of a submodular capacity, which represents uncertainty aversion. For finite Ω , submodularity is equivalent to this property of *diminishing marginal returns*:

$$\forall A \subseteq B \subset \Omega, c \notin B : \bar{\mu}(A \cup \{c\}) - \bar{\mu}(A) \geq \bar{\mu}(B \cup \{c\}) - \bar{\mu}(B). \quad (2.272)$$

This expresses that adding an element to a smaller set results in a greater increase in decision weight. Consequently, large losses (where $\bar{\mu}(\{X \geq x\})$ is small) are emphasized. Whereas risk aversion is expressed by a concave utility function in [von Neumann and Morgenstern, 1947], ambiguity aversion is a submodular attitude towards *probability* itself. Submodular capacities are also called concave, since they exhibit a similar diminishing marginal returns property as concave functions. Furthermore, recall that if the capacity is given as the composition of an increasing function and a probability measure, the capacity is submodular if and only if the function is concave (Section 2.3.6).

Choquet expected utility is closely related to maxmin expected utility. If and only if the capacity is submodular, then the Choquet integral is convex [Alfonsi, 2015] and the representation takes a maxmin form (minmax, in loss-based formulation), where the envelope is given by the *core* of the capacity

$$\text{core}(\bar{\mu}) = \{P : P(A) \leq \bar{\mu}(A) \, \forall A \in \mathcal{F}, P \text{ finitely additive probability measure}\} \quad (2.273)$$

$$\int X \, d\bar{\mu} = \sup_{P \in \text{core}(\bar{\mu})} \left\{ \int_{-\infty}^{\infty} X \, dP \right\}. \quad (2.274)$$

The ambiguity aversion is directly related to the convexity of the functional. The close relationship between maxmin expected utility (MMEU) and Choquet expected utility (CEU) has been concisely summarized by Klibanoff [2001]:

Fundamentally, CEU decision makers view uncertainty in terms of (roughly) how states are ordered by an act's utility payoff. Given a set of acts which all induce the same ordering, a CEU decision maker acts exactly like an expected utility (and thus uncertainty neutral) decision maker. MMEU decision mak-

ers, in contrast, may view uncertainty not only in terms of ordering of states, but also in terms of how much better the payoff is in one state as opposed to another.

Both MMEU and CEU are theories about uncertainty in the sense of ambiguity. A capacity in CEU contains both a component of *belief* and *action* [Diecidue and Wakker, 2001], where *action* refers to a decision attitude. However, for a general capacity, these components cannot be separated. This intertwining empowers Choquet expected utility to tackle problems of ambiguity in a broad sense; yet it also renders it somewhat impractical. Revisiting Ellsberg’s urns (Section 2.1.1), CEU can indeed describe the ambiguity-averse preferences which most decision makers exhibit in this scenario [Schmeidler, 1989]. Ellsberg’s urns are challenging because they not only violate expected utility, but also *probabilistic sophistication* [Etner et al., 2012, Machina and Schmeidler, 1992]. A probabilistically sophisticated decision maker acts in accordance with a belief which can be captured by a probability measure, but uses it in a manner that can extend beyond classical expected utility. Hence belief and action are still intertwined to some degree. For instance, a risk-averse decision maker might express beliefs with an underlying probability measure but decides in a way so as to put more weight on worse outcomes. In the setting of CEU, probabilistic sophistication implies that the capacity is given by a composition $\bar{\mu} = \phi \circ P$ of an increasing function and a probability measure. Indeed, we may equate probabilistic sophistication with law invariance (rearrangement invariance). If a probabilistically sophisticated CEU decision maker satisfies CH5, then ϕ is concave and the Choquet integral is therefore a spectral risk measure. However, the typical preference behaviour in Ellsberg’s urns cannot be modelled by such a functional [Schmeidler, 1989]. Some authors therefore identify probabilistic sophistication with ambiguity neutrality [Epstein, 1999]. We think that this goes too far: for instance, a law-invariant spectral risk measure expresses risk aversion by aversion to hallucinated ambiguity. There is still the assumption of a base measure, on which belief rests, but the action component constructs an ambiguity set around this base measure. This amounts to blurring the line between risk aversion and ambiguity aversion. Under law invariance, their mathematical form is equivalent and we may interpret risk aversion as a form of ambiguity aversion with respect to an artificially constructed (‘hallucinated’) ambiguity set. We emphasize that we do not claim that risk and ambiguity are equivalent, but rather that risk aversion can be modelled via aversion to hallucinated ambiguity. Furthermore, in light of the Kusuoka representation, any coherent risk measure is a combination of the two, since it can be described as an ambiguity set over a risk spectrum. In summary, we advocate thinking of a direct relation between risk aversion and ambiguity: at one extreme of the spectrum, where the supremum risk measure embodies maximal risk aversion, it has the corresponding interpretation of the maximal ambiguity set, consist-

ing of all³⁴ probability measures. At the other extreme, the expectation is risk neutral and is represented by the singleton ambiguity set $\{1\}$. Hence, whether a law invariant coherent risk measure should be seen as modelling risk or ambiguity depends on the context and the modelling intentions of the decision maker. Therefore we will now examine Choquet expected utility under probabilistic sophistication (law invariance), where the capacity can be decomposed into a belief and action attitude. In the concave case, this yields the class of spectral risk measures.

2.9.4 RANK DEPENDENT EXPECTED UTILITY

The crucial difference between CEU and *rank dependent expected utility* (RDEU) is the additional requirement of law invariance (probabilistic sophistication, rearrangement invariance). Therefore, RDEU is typically viewed as CEU under risk. We pointed out, however, that this can also be viewed as theory of hallucinated ambiguity. Different authors have arrived at variants of RDEU [Yaari, 1987, Wang, 1995, Quiggin, 2012, Buchak, 2013], which turned out to approximately coincide. RDEU represents preferences by law invariant Choquet integrals:

$$X \succcurlyeq Y \iff \int \ell \circ X \, d(\phi \circ P) \leq \int \ell \circ Y \, d(\phi \circ P), \quad (2.275)$$

with a loss function ℓ and where the capacity is specialized as the composition of an increasing function ϕ and a probability measure P . Hence, risk aversion (submodularity) of the capacity is equivalent to the concavity of ϕ (cf. Section 2.3.6). If the loss function is the identity, as we take it in machine learning³⁵, then we recover exactly the class of spectral risk measures.

RDEU is *rank dependent*, since the weight of a certain outcome in the decision not only depends on its probability via P , but also on how it is ranked with respect to other outcomes. This enables the decision maker to express a desire for *distributional objectives* [Lopes, 1984]. Given a fixed mean, decision makers may prefer a less spread-out distribution as compared to a more spread-out one. The exact nature of this tradeoff is encoded in the function ϕ , which can be considered a risk aversion profile. In the context of machine learning, it allows us to emphasize the largest losses to increase robustness.

Of particular interest to us is the rank dependent account of Buchak [2013], which is called *risk weighted expected utility*. Buchak [2013] aims to provide argumentative ground for why risk attitudes via rank dependence are normatively permissible, instead of only empirically adequate. Other authors are less clear on this issue or take a different stance. For instance, a slight variant of RDEU, *prospect theory*, is only defended as a descriptive theory [Tversky and Kahneman, 1992]. Furthermore, Buchak [2017] has also considered the theory in the

³⁴Probability measures which are absolutely continuous with respect to a base measure.

³⁵Note again that in our machine learning setup $X(\omega)$ already corresponds to a loss value.

setting of social choice, which is relevant to fair machine learning. The possible application of rank dependence in this context has been hinted at by other authors [Schmeidler, 1989, Quiggin, 2012], but not elaborated.

2.9.5 RATIONAL AND SOCIAL CHOICE WITH SPECTRAL RISK MEASURES

Rational choice is about an individual decision maker in a context where the decision affects only that individual. This can be modelled with a state space Ω , where each $\omega \in \Omega$ represents a possible state of the world. A gamble $X : \Omega \rightarrow \mathbb{R}$ assigns to each state a resulting loss to the decision maker, given that this state is realized. In classical probability, such a gamble is evaluated via the expectation. This is the standard ML problem, where the engineer aims to minimize loss. By contrast, social choice concerns collective decision by a combination of individual preferences. This is closer to the model for a fair ML problem, where the individuals are salient subgroups. However, in ML the engineer chooses the loss function for everyone, whereas in the “real world” setting, individuals might have different loss functions.

The structural analogy is that a state ω in a rational choice problem corresponds to an individual (subgroup) in social choice [Buchak, 2017] and a gamble then describes a social arrangement (“who gets what”). Expected utility theories ask the question *how should an individual value a gamble?* and the classical theory gives the expectation as the unique answer, whereas we have demonstrated that there exists a variety of interesting alternatives. Social choice theory asks: which social arrangements are to be preferred (or *fair*)? Due to the structural analogy, it is not surprising that similar answers have been given. Most prominently, expected utility theory in rational choice has average utilitarianism as its social counterpart [Buchak, 2017]. The analogy also yields an interesting interpretation for probability: an individual considers its possible “future selves”, which would result from each outcome, which makes the question of how to value the gamble equivalent to the problem of finding a fair distribution among those future selves.

Relevant to this discussion is the distinction between *aggregate* (or *groupist*) and *individual* risk [Dawid, 2017]. The former is what we know well from probability theory: statements about relative frequencies are aggregate statements. When tossing a fair coin, on what basis do we assign the probability $p = 0.5$ that it will land heads? Typically, the reasoning proceeds from the aggregate to the individual here. A frequentist explanation is that we have observed many coin tosses and the relative frequency of heads stabilized around 0.5 (although such a statement can only be made in the limit of infinitely many tosses, which itself is problematic). A Bayesian may appeal to a symmetry principle, because there is no reason to favour either heads or tails for a fair coin, one should assign the degree of belief 0.5 that it will land heads. Such a notion might strike one as individualistic. In practice,

however, Bayesian inference is typically with respect to an exchangeable information base — an aggregate. The Bayesian might have flipped the coin many times, considered the sequence exchangeable [de Finetti, 1974/2017], updated their beliefs accordingly and hence arrived at a probability of 0.5. Similarly, if the Bayesian used prior knowledge from other fair coins, which they had experience with, this has an aggregate flavour. Dawid [2017] concludes by stating that the group to individual inference direction remains problematic and elusive.

While the above example of a coin seems innocent, it is problematized in ethical contexts, where the individual coin toss is replaced by an event that concerns a human. A fair ML problem can be phrased as distributing loss (in the ML sense) over individuals or subgroups of ethically fungible individuals. Such a subgroup (e.g. men, women), according to the designer, is then viewed as an individual in the given context. One possibility of expressing imperfect fairness in this context is that we demand that subgroup losses are commensurate, i.e. they should not differ much. Under the assumption of mutual disinterest, an individual (subgroup) is concerned only with its own risk. When the aggregate risk is low, i.e. the average individual risk is low, this is no consolation for any individual, who does not care for the average. When an inference is based on an aggregate, how can we control individual risk? Classical probability, firmly based on an aggregate conception due to its casino origin, is of no help.

As a corollary, we find that the concept of individual risk has mirror images in rational and social choice. On the one hand, inequality aversion can be understood as risk aversion, as it is a focus on the worse outcomes. On the other hand, a risk-averse individual is one that is inequality-averse with respect to future selves. In essence, this means that the individual is concerned with its own individual risk, instead of merely its aggregate risk of future selves. A decision maker who uses the expectation is risk-neutral and cares only about aggregate risk; here, the aggregate is to be understood as formed from the possible outcomes for that single individual. This is reasonable, when an experiment is repeated under stable conditions indefinitely and the possibility of a catastrophic event (e.g. going bankrupt) is excluded. However, in real ML problems, this is not the case and often individuals only have a single shot, for instance at getting a loan. When choosing between a sure gain of c or a lottery which yields $2c$ with probability 0.5 and 0 otherwise, almost all individuals choose the sure gain [Cappelen et al., 2013]. Our interpretation of this robust pattern is that they care about their individual risk and adopt a pessimistic attitude. Standard expected utility theory would model this via a concave attitude towards wealth. But even if c would already be in units of loss, as is the case in ML, we are inclined to think that it is still preferable to have less spread: because we care about individual risk.

Spectral risk measures, provide a partial resolution to this conflict between aggregate and individual, between average utilitarianism and subgroup fairness. The extreme points of

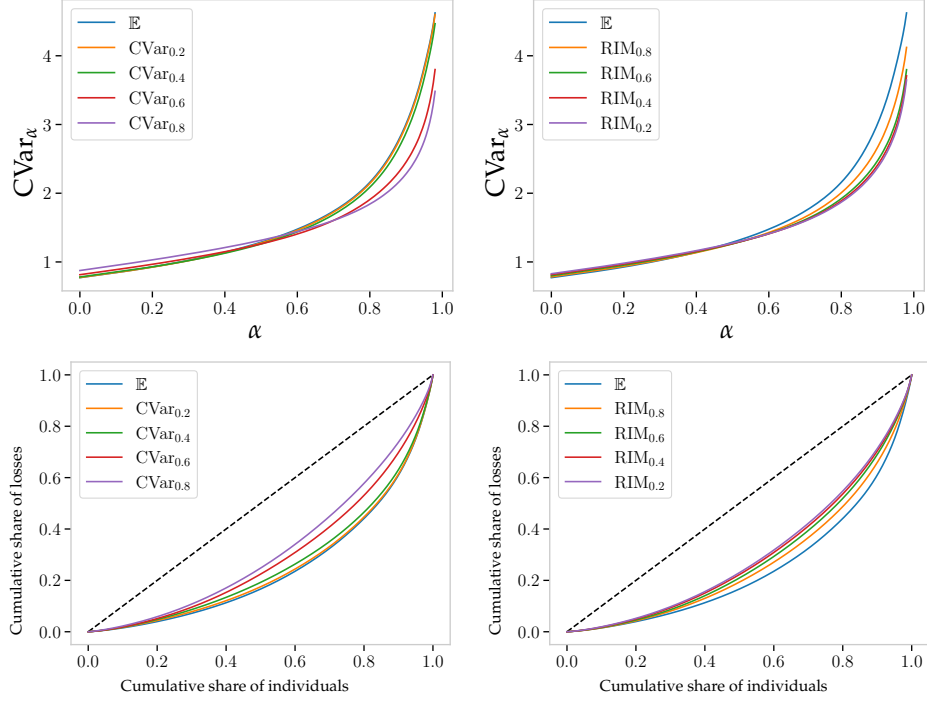


Figure 2.9: PCA* results on `adult`. Top row: CVar_α curves of test losses for CVar_α risk measures (left) with different α and RIMs risk measures (right) with different β , indicated by subscript, where $\alpha = 0.7$. For better visibility of the differences, we cut off α at 0.98. Bottom row: Lorenz curves of test losses for CVar_α (left) and RIMs (right) with $\alpha = 0.7$ and different β .

the family of spectral risk measures are CVar_α . Here, $\alpha = 0$ recovers the risk (inequality)-neutral expectation. On the other hand, $\alpha = 1$ embodies the maximally risk (inequality)-averse attitude. In distributive justice, the corresponding theory is the Rawlsian maximin principle [Rawls, 1971], where only the position of the worst-off counts. The parameter α offers a smooth interpolation between these two ends of the spectrum. The behaviour CVar_α is extreme in the sense that it neglects all outcomes below the $1 - \alpha$ tail of losses. Finer control is possible by employing any spectral risk measure, where the tradeoff aggregate vs. individual is encoded in the shape of the fundamental function ϕ .

2.10 APPENDIX: EXPERIMENTS

Here we report additional results, not shown in the main paper. See Figures 2.9, 2.10, 2.11.

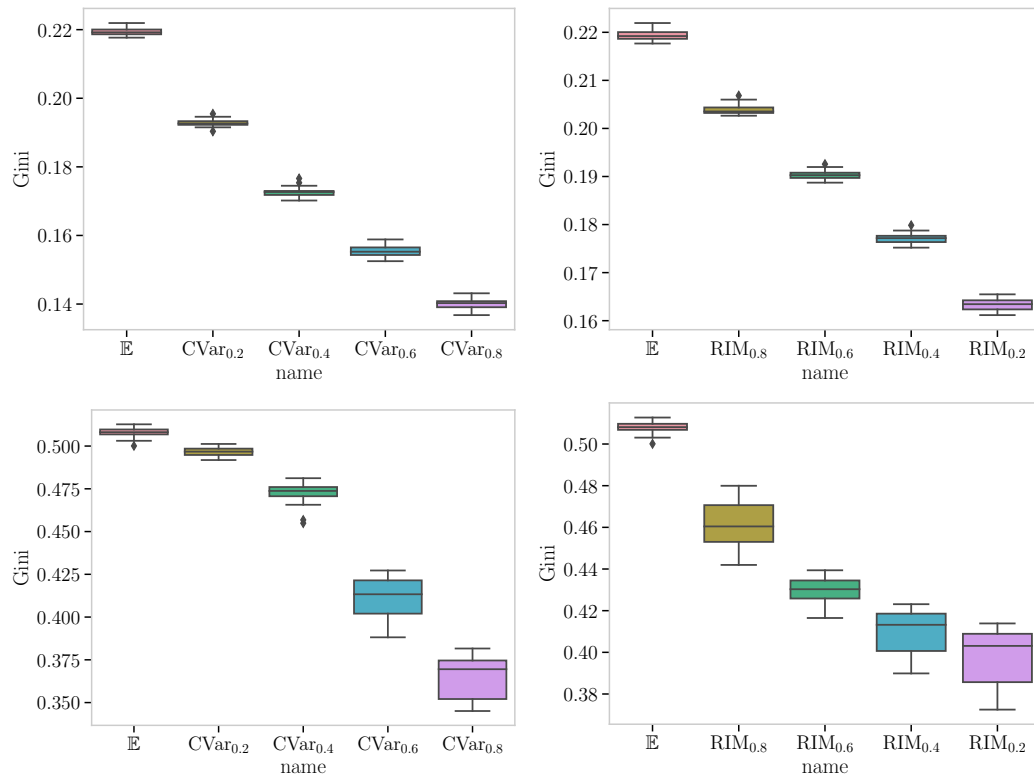


Figure 2.10: PCA* Gini coefficients for MNIST (top) and adult (bottom) over 25 runs.

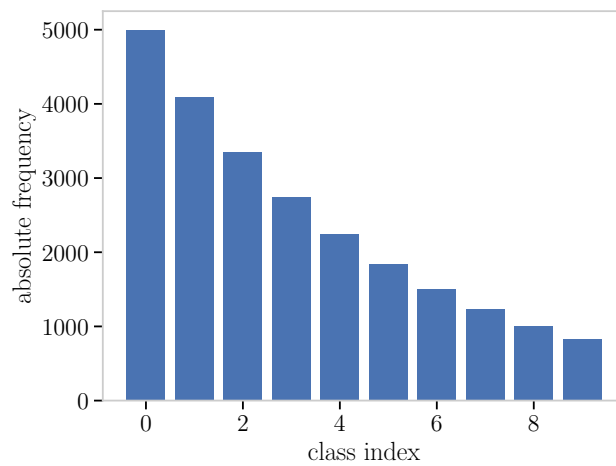


Figure 2.11: Class frequencies of imbalanced MNIST. Before each iteration, the assignments of digit class to frequency are randomly shuffled.

3

Tailoring to the Tails: Risk Measures for Fine-Grained Tail Sensitivity

AUTHOR CONTRIBUTIONS

Robert Williamson contributed through supervision of the project, e.g. through discussions and proof-reading. The remaining work is due to Christian Fröhlich.

ABSTRACT

Expected risk minimization (ERM) is at the core of many machine learning systems. This means that the risk inherent in a loss distribution is summarized using a single number – its average. In this paper, we propose a general approach to construct *risk measures* which exhibit a desired tail sensitivity and may replace the expectation operator in ERM. Our method relies on the specification of a reference distribution with a desired tail behaviour, which is in a one-to-one correspondence to a *coherent upper probability*. Any risk measure, which is compatible with this upper probability, displays a tail sensitivity which is finely tuned to the reference distribution. As a concrete example, we focus on *divergence risk measures* based on f -divergence ambiguity sets, which are a widespread tool used to foster distributional robustness of machine learning systems. For instance, we show how ambiguity sets based on the Kullback-Leibler divergence are intimately tied to the class of subexponential random variables. We elaborate the connection between divergence risk measures and *rearrangement invariant Banach norms*.

3.1 INTRODUCTION

A plain language summary of our work and a diagram showing key conceptual elements are in Appendix 3.14.

Expected risk minimization is at the core of machine learning systems. This means that the risk inherent in a loss distribution is summarized using a single number – its average. Yet this does not capture the *tail behaviour* of the distribution. Since the expectation is insensitive to such behaviour, two distributions may have the same expectation, although they exhibit very different tails. In the context of machine learning, this can be problematic: extreme losses correspond to extremely bad predictions, or, for instance in the context of reinforcement learning, even lead to disastrous consequences. In many settings, there appear to be legitimate motivations for sacrificing some average performance in order to avoid rare but extreme losses. A decision maker with this attitude is said to be *risk averse*.

Another issue to consider is that the empirical distribution perfectly corresponds to the ‘true’ distribution only in the limit of infinite data. With finite data, the observed tails could be much lighter than the ‘true’ tail of the loss distribution, since extreme events are by definition infrequently observed. This suggests that a cautious approach is to pretend that the tail is heavier than actually observed. This approach has been studied under the research theme of *distributional robustness*, where expected risk minimization has been replaced by the distributionally robust f -divergence risk measure objective:

$$R(X) := \sup\{\mathbb{E}_Q[X] : d_f(Q, P) \leq \varepsilon\}, \quad (3.1)$$

where X is the random loss variable, the Q are probability measures and the f -divergence with divergence function f is defined as

$$d_f(Q, P) := \int_{\Omega} f\left(\frac{dQ}{dP}\right) dP, \quad (3.2)$$

if Q is absolutely continuous with respect to P . Here, the decision maker takes a worst-case attitude with respect to a set of probability measures, which are contained in an f -divergence ball of radius ε , centered at the base distribution P . These probability measures are said to constitute an *ambiguity set*. A decision maker, who takes a worst-case stance towards an ambiguity set, is *ambiguity averse*. In practice, we take P to be the empirical distribution, that is, the training data. Intuitively, the function f controls how much reweighting $\frac{dQ}{dP}$ is allowed. As a consequence, we guard against possibly different tail behaviour of the ‘true’ distribution.

The goal of this paper is to construct tail-sensitive *coherent risk measures* (Section 3.3.4), where the tail sensitivity can be controlled by the machine learning engineer. Such risk

measures are then potential replacements for the expectation in ERM and can model both risk and ambiguity aversion, depending on the decision maker’s intention. For instance, the f -divergence risk measure R in Equation 3.1 is a coherent risk measure, and we demonstrate how the function f corresponds exactly to a choice of tail sensitivity. Consequently, the choice of f determines the structure of the divergence ball, that is, the tail behaviour of the reweightings it contains. Due to their popularity as a means to achieve distributional robustness in machine learning, we focus on f -divergence risk measures. However, we also explicate its relation to other coherent risk measures with the same *fundamental function*, which corresponds (to some extent) to the same tail sensitivity. To investigate these relations, we view coherent risk measures from the perspective of *rearrangement invariant Banach norms*.

For each class of specified tail behaviour (for instance, exponential tails, Gaussian tails etc.), subject to a suitable equivalence relation, we show how the following objects can be derived:

- An f -divergence risk measure as in Equation 3.1, which can be employed to replace the expectation in the risk minimization loop.
- A Lorentz and Marcinkiewicz norm and related coherent risk measures, which are further such replacement candidates.
- A *utility-based shortfall risk measure*, which is a *convex* risk measure, with potential applications for instance in reinforcement learning.

The main guiding thread is the family of *Orlicz norms*, which offer fine-grained control over tail risk via the specification of a convex function. The class of such functions is approximately as rich as the class of f -divergences. The f -divergence risk measure Equation 3.1 then arises as the *natural extension* of what we call an *Orlicz regret measure*. The theory of Orlicz spaces is well-developed, and not only underlies divergence risk measures but also offers a general approach to obtaining concentration inequalities and maximal inequalities [Edgar and Sucheston, 1992], but it appears that the machine learning community is not aware of this link and thus cannot exploit the rich mathematical literature on the subject.

3.1.1 RELATED WORK

We weave together ideas from different literatures, from coherent risk measures [Artzner et al., 1999, Pflug and Romisch, 2007] and rearrangement invariant Banach spaces [Kreĭn et al., 1982b, Bennett and Sharpley, 1988, Rubshtein et al., 2016] to distributional robustness [Rahimian and Mehrotra, 2019].

Our goal is to construct risk measures which are sensitive to the heaviness of the tails. The importance of heavy-tailed phenomena has been popularized by Taleb [2007] and they

have been studied generally by [Nair et al. \[2022\]](#) and [Resnick \[2007\]](#). In terms of applications, heavy-tailed phenomena have been investigated in economics [[Nordhaus, 2012](#), [Ibragimov et al., 2015](#)], earth sciences [[Cavanaugh et al., 2015](#), [Merz et al., 2022](#)], neuroscience [[Roberts et al., 2015](#)] and many other areas. Within machine learning, [Gurbuzbalaban et al. \[2021\]](#) have identified heavy tails in the iterates of stochastic gradient descent. Empirical risk minimization in the context of a heavy-tailed loss distribution has been studied by [Brownlees et al. \[2015\]](#) and [Hsu and Sabato \[2016\]](#). [Vladimirova et al. \[2019\]](#) have demonstrated the heavy-tailed nature of prior distributions in Bayesian deep learning.

We are inspired by the work of [Dommel and Pichler \[2020\]](#), themselves building on [Ahmadi-Javid \[2012\]](#) and [Ahmadi-Javid and Pichler \[2017\]](#), who connected f -divergence risk measures to Orlicz norms. Ambiguity sets based on f -divergences have received much attention in the machine learning literature so far, see for instance [[Ben-Tal and Nemirovski, 1999](#)], [[Bayraksan and Love, 2015](#)], [[Namkoong and Duchi, 2016](#)], [[Shapiro, 2017](#)], [[Namkoong and Duchi, 2017](#)], [[Hu et al., 2018](#)] and [[Duchi et al., 2021](#)]. In mathematical finance, the special case of the *coherent entropic risk measure*, which corresponds to the choice of Kullback-Leibler divergence for d_f in Equation 3.1, was studied by [Föllmer and Knispel \[2011\]](#). Closely related is the *convex entropic risk measure*, an empirical variant of which has been proposed by [Li et al. \[2021\]](#) in machine learning. Standard material on Orlicz spaces can be found in [[Krasnoselskii and Rutickii, 1961](#)], [[Bennett and Sharpley, 1988](#)], [[Edgar and Sucheston, 1992](#)], [[Kosmol and Müller-Wichards, 2011](#)], [[Pick et al., 2013](#)] and [[Rubshtein et al., 2016](#)].

3.1.2 CONTRIBUTIONS

We define the *Orlicz regret measure*, which is an asymmetric variant of an Orlicz norm, and show that its *natural extension* yields the f -divergence risk measure. We embed the Orlicz regret measure and the divergence risk measure in the mathematical framework of rearrangement invariant Banach function spaces and study their fundamental functions. Subsequently we explicate how the fundamental function is, via its one-to-one relation to the *envelope function*, intricately linked to the tail sensitivity of a risk measure. We show how one can therefore construct a corresponding Marcinkiewicz norm, Orlicz norm (hence, a divergence risk measure) and a Lorentz norm, which share this tail sensitivity. We illuminate the connections and (non)-equivalences between these norms. Overall, we do not attempt to establish that one among these three families of norms is ‘superior’, but rather to analyze their relations and implications. Finally, we link our approach to tail sensitivity to *Orlicz deviation inequalities*, of which many well-known concentration inequalities are special cases.

3.1.3 USE IN MACHINE LEARNING

In a supervised learning problem, we may wish to replace the expectation operator in ERM by a *risk measure*:

$$\operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}[\ell(f(X), Y)] \longrightarrow \operatorname{argmin}_{f \in \mathcal{F}} R(\ell(f(X), Y)) \quad (3.3)$$

for some risk measure R , a function f from some hypothesis space $\mathcal{F}$, a loss function ℓ , input X and ground truth labels Y . Similarly, in reinforcement learning problems, a risk measure may replace the expectation operator in the maximization of *expected* discounted sum of returns. Correspondingly, in *empirical* risk minimization, the risk measure can be applied to the empirical loss distribution.

Possible motivations for replacing the expectation with a risk measure are manifold: in machine learning, risk measures have been employed to express a risk-averse attitude [Tamar et al., 2015, Dabney et al., 2018, Singh et al., 2020, Urpí et al., 2021, Vijayan et al., 2021] or to achieve distributional robustness [Namkoong and Duchi, 2016, Duchi et al., 2021, Zhang et al., 2021b]. Moreover, consider the case where the individual losses correspond to individual humans: in this context, it has been argued that risk measures can promote subgroup fairness [Hashimoto et al., 2018, Williamson and Menon, 2019].

We restrict ourselves to theoretical analysis as there is ample experimental evidence for the suitability of risk measures in different settings in the aforementioned works. In addition, Chouzenoux et al. [2019] and Fröhlich and Williamson [2024a] provide general treatments on risk measures in machine learning, including experimental evidence. Our focus is therefore on drawing connections and deepening our theoretical understanding. For aspects regarding practical computation see Section 3.8.

A question remains: which risk measure to choose in a given problem setting? The focus of this paper is to give the reader guidance and intuition to answer this question for themselves, relative to the problem at hand. Specifically, our focus is the *tail sensitivity* of risk measures. In the previously mentioned problem settings – aiming for risk aversion, distributional robustness or fairness – the tail behaviour of the loss random variable deserves special attention. For instance, a tail event in a fairness context is an individual who receives an extreme loss. Our goal is therefore *tailoring to the tails* instead of *pandering to the masses*¹.

¹That “pandering to the masses” is the opposite of “tailoring to the tails” was suggested to us by Peter Kootsookos.

3.1.4 NOTATION AND ASSUMPTIONS

Throughout, we assume an atomless probability space $(\Omega, \mathcal{F}, P)$, where P is the Lebesgue measure and we take $\Omega = [0, 1]$. We denote the set of probability measures Q which are absolutely continuous with respect to P as $\mathcal{Q}$. Let $\mathcal{M}$ denote the class of Lebesgue measurable functions from Ω to the reals $\mathbb{R}$, i.e. random variables; however, we have no regard for non-measurable functions throughout and thus often drop writing $X \in \mathcal{M}$. By writing P -a.e. we mean “almost everywhere with respect to the measure P ”. We denote by $\mathcal{L}^p$ the Lebesgue spaces, i.e. the set of functions with finite p -th moment:

$$\|X\|_p := \begin{cases} \left(\int_0^1 |X|^p dP \right)^{\frac{1}{p}} & 1 \leq p < \infty \\ \text{ess sup}(X) & p = \infty, \end{cases} \quad \mathcal{L}^p := \{X \in \mathcal{M} : \|X\|_p < \infty\}, \quad (3.4)$$

where $\text{ess sup}(X) := \inf \{\lambda \geq 0 : P(\{\omega \in \Omega : |X(\omega)| > \lambda\}) = 0\}$. We write $\mathbb{R}^+ = [0, \infty)$ for the nonnegative real line with 0 included and $\mathbb{R}^- = (-\infty, 0)$. By $X^+ := \max(0, X)$ we denote the positive part of a random variable. We write Radon-Nikodym derivatives as $\frac{dQ}{dP}$. For a random variable X with cumulative distribution function $F_X(x) := P(\{\omega \in \Omega : X(\omega) \leq x\})$, we define the quantile function F_X^{-1} as the generalized inverse of the distribution function: $F_X^{-1}(q) := \sup\{\lambda \geq 0 : F_X(\lambda) < q\}$. For expectations $\mathbb{E}_P[X]$ with respect to the base measure P , we often drop the subscript and simply write $\mathbb{E}[X]$. We write indicator functions as χ_A and convex indicator functions as i_A :

$$\chi_A(\omega) := \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A, \end{cases} \quad i_A(\omega) := \begin{cases} 0 & \omega \in A \\ \infty & \omega \notin A \end{cases}. \quad (3.5)$$

By an increasing function we mean a non-decreasing function; similarly, a decreasing function is non-increasing. We write $\phi(0+)$ for the right-hand limit given by $\lim_{x \downarrow 0} \phi(x) = \phi(0+)$. When defining a function $f : \mathbb{R}^+ \rightarrow \mathbb{R} \cup \{\infty\}$, we abuse notation and implicitly take this to imply that $f(x) = \infty$ for $x < 0$. The convex conjugate of a function $f : \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$ is the function $g : \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$:

$$g(y) := \sup_{x \in \mathbb{R}} xy - f(x). \quad (3.6)$$

Throughout, f and g will denote a conjugate pair of divergence functions (Definition 100) and Φ and Ψ will denote a conjugate pair of Young functions (Definition 96).

3.2 ORLICZ NORMS AND SPACES

The motivation behind Orlicz norms is to measure the size of a function in a way which is sensitive to tail behaviour. Often they are introduced as generalizations of $\mathcal{L}^p$ spaces. As a first example, the variance indeed is sensitive to tails and therefore does not exist (is infinite) for many random variables.

Definition 96. A function $\Phi : \mathbb{R}^+ \rightarrow \mathbb{R}^+ \cup \{\infty\}$ is called *Young function* if it is left-continuous, increasing, convex, satisfies $\Phi(0) = 0$ and is nontrivial (not identical to the constant zero or infinity function).

We emphasize that here *increasing* means *non-decreasing*. Since Φ is convex, continuity is guaranteed on the interior of its effective domain and therefore the left-continuity concerns only the point

$$b_\Phi := \sup\{\Phi < \infty\}. \quad (3.7)$$

Note that it is possible that $\Phi(x) = 0$ for $x > 0$. We denote the convex conjugate of Φ as $\Psi : \mathbb{R}^+ \rightarrow \mathbb{R}^+ \cup \{\infty\}$, that is:

$$\Psi(y) := \sup_{x \geq 0} xy - \Phi(x), \quad (3.8)$$

where the restriction to the nonnegative half line is a consequence of our assumption that Φ takes on $+\infty$ on the negative half line. If Φ is a Young function, then so is Ψ . In the following, Φ and Ψ will always refer to a pair of conjugate Young functions.

Definition 97. The *Luxemburg norm* of the random variable X is defined as:

$$\|X\|_\Phi^L := \inf \left\{ \lambda > 0 : \mathbb{E} \Phi \left[\frac{|X|}{\lambda} \right] \leq 1 \right\} \quad \forall X \in \mathcal{M}, \quad (3.9)$$

which is the gauge of the set $\{Z \in \mathcal{M} : \mathbb{E} \Phi(|Z|) \leq 1\}$.

Definition 98. The *Orlicz norm* of the random variable X is defined as:

$$\|X\|_\Phi^O := \sup \{ \mathbb{E}[XZ] : Z \in \mathcal{L}^1, \mathbb{E}[\Psi(|Z|)] \leq 1 \} \quad \forall X \in \mathcal{M}. \quad (3.10)$$

Observe that the conjugate Ψ appears here. Confusingly, the Luxemburg norm [Luxemburg, 1955] is sometimes simply called Orlicz norm in the literature. In fact, these norms are equivalent by a factor of 2:

$$\|X\|_\Phi^L \leq \|X\|_\Phi^O \leq 2\|X\|_\Phi^L \quad (3.11)$$

and thus they are finite for the same set of functions. Therefore it is inconsequential whether one takes the Luxemburg or Orlicz norm in the following definition.

Definition 99. *The Orlicz space $\mathcal{L}^\Phi$ is given by:*

$$\mathcal{L}^\Phi := \{X \in \mathcal{M} : \|X\|_\Phi^O < \infty\}. \quad (3.12)$$

For any Young function Φ , the Orlicz space $\mathcal{L}^\Phi$ is a Banach space, i.e. a complete normed vector space. It is furthermore a Banach lattice, that is, it is monotone in the sense that:

$$|X| \leq |Y| \text{ P-a.e.} \implies \|X\|_\Phi^O \leq \|Y\|_\Phi^O. \quad (3.13)$$

Important in the following development will be an identical expression for the Orlicz norm, which is also called the *Amemiya norm* [Hudzik and Maligranda, 2000]:

$$\|X\|_\Phi^O = \inf_{t>0} t \left(1 + \mathbb{E} \Phi \left(\frac{|X|}{t} \right) \right). \quad (3.14)$$

This representation is the key to compute Orlicz norms in practice.

3.3 ORLICZ REGRET MEASURES

The problem with Orlicz norms in our setting is that, in virtue of satisfying the properties of a norm, they treat losses and gains in the same way. We follow the machine learning convention that losses correspond to positive values, hence gains are negative values. Since we aim for an asymmetric setup, we introduce Orlicz regret measures as the asymmetric variants of Orlicz norms and link them to f -divergence risk measures.

3.3.1 f -DIVERGENCES

First, we define the function class which we consider in the following and their associated f -divergences.

Definition 100. *A divergence function f is a proper, lower semi-continuous, convex function $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+ \cup \{\infty\}$, which satisfies $f(1) = 0$, $f(0) < \infty$ and the supercoercivity condition $\lim_{x \rightarrow \infty} \frac{f(x)}{x} = \infty$.*

The conditions that $f(0) < \infty$ and $\lim_{x \rightarrow \infty} \frac{f(x)}{x} = \infty$ play an important role. They are non-standard restrictions that we impose here and will be further discussed.

We denote the convex conjugate of f as g throughout the paper. It has the following properties.

Proposition 101. *The conjugate g is finite on $\mathbb{R}$, convex and increasing.*

Proof The statement, which relies on the supercoercivity of f , is well known and a proof is included in Appendix 3.10.1. ■

Definition 102. *Given a divergence function f , the f -divergence between probability measures Q and P is defined as:*

$$d_f(Q, P) := \int_{\Omega} f\left(\frac{dQ}{dP}\right) dP = \mathbb{E}_P f\left(\frac{dQ}{dP}\right), \quad (3.15)$$

if Q is absolutely continuous with respect to P , and $+\infty$ otherwise.

A simple shift of perspective, which will be important in the following, is to view $\frac{dQ}{dP}(\omega)$ as a single random variable rather than as a fraction of two densities.

We write $\mathfrak{F}$ for the set of divergence functions f and $\mathfrak{F}_Y$ for the subset of divergence functions f , which are also valid Young functions. We call elements of $\mathfrak{F}_Y$ *Young divergence functions*. Since we assumed nonnegativity, the only difference to a Young function is that a general divergence function f might be positive and decreasing for $0 \leq x < 1$. To a given divergence function f , we can associate a canonical Young function $\bar{f}$ by setting:

$$\bar{f}(x) := \begin{cases} 0 & 0 \leq x \leq 1 \\ f(x) & x > 1, \end{cases} \quad (3.16)$$

as in [Dommel and Pichler, 2020]. We write $\bar{g}$ for the convex conjugate of $\bar{f}$. We will see (Proposition 110) that it stands in a close relationship to the original f . For various levels of risk aversion² $\varepsilon > 0$, we define the shorthand notation $f_{\varepsilon}(x) = \frac{1}{\varepsilon}f(x)$ and $g_{\varepsilon}(y) = \frac{1}{\varepsilon}g(\varepsilon y)$, which is the convex conjugate of f_{ε} . The value of ε is not of theoretical importance: if f is a divergence function, so is f_{ε} .

Example 103. A prominent example is the Kullback-Leibler divergence, where $f_{\text{KL}}(x) := x \log(x) - (x - 1)$. It can also be obtained as $f(x) = x \log(x)$, but the affine shift guarantees nonnegativity without altering the divergence. We complete its definition implicitly by setting $f_{\text{KL}}(0) := \lim_{x \rightarrow 0} f_{\text{KL}}(x) = 1 < \infty$. The induced f -divergence is

$$d_{\text{KL}}(Q, P) = \int_{\Omega} \log\left(\frac{dQ}{dP}\right) dQ. \quad (3.17)$$

The conjugate of f_{KL} is $g_{\text{KL}}(y) = \exp(y) - 1$.

²For simplicity, we shall use the term “risk aversion level” throughout. However, a divergence risk measure may also capture the arguably distinct phenomenon of ambiguity aversion.

Example 104. The χ^2 divergence is given by $f_{\chi^2}(x) = (x - 1)^2$. Then the divergence is

$$d_{\chi^2}(Q, P) = \int_{\Omega} \left(\frac{dQ}{dP} \right)^2 dP - 1. \quad (3.18)$$

Note that the conjugate of f_{χ^2} is $g_{\chi^2}(y) = \max(-2, y) + \max(-2, y)^2/4$, since we take f to be infinite on the negative half line. In some tables, the reader will find that $g(y) = y + y^2/4$, which is the conjugate if the effective domain of $f(x) = (x - 1)^2$ is $\mathbb{R}$. The difference is crucial in our asymmetric setting.

Example 105. Consider the pathological divergence function

$$f_{\mathbb{E}}(x) := i_{[0,1]}(x) = \begin{cases} 0 & 0 \leq x \leq 1 \\ \infty & x > 1. \end{cases} \quad (3.19)$$

We later observe that the corresponding divergence risk measure in Equation 3.1 is simply the expectation $\mathbb{E}$.

Example 106 (Shapiro [2017]). Given some $\alpha \in [0, 1)$, we slightly modify the previous example as:

$$f_{\text{CVar}_{\alpha}} := i_{[0,1/(1-\alpha)]}. \quad (3.20)$$

The corresponding divergence risk measure in Equation 3.1 is the *conditional value at risk* CVar_{α} , which is the expectation in the upper $(1 - \alpha)$ -tail:

$$\text{CVar}_{\alpha}(X) := \frac{1}{1 - \alpha} \int_{\alpha}^1 F_X^{-1}(q) dq. \quad (3.21)$$

For continuous F_X , this can be written in the more suggestive form (“tail-conditional expectation”):

$$\text{CVar}_{\alpha}(X) = \mathbb{E} \left[X | X \geq F_X^{-1}(\alpha) \right]. \quad (3.22)$$

In the machine learning community, CVar_{α} has recently received significant attention, see for instance [Takeda and Sugiyama, 2008], [Fan et al., 2017] and [Curi et al., 2020].

3.3.2 DEFINITION AND FUNCTIONAL PROPERTIES

We now define the *Orlicz regret measure*, which has similar properties to a norm, yet is fundamentally asymmetric in that it treats losses and gains differently. We will find that it has a very close relation to the divergence risk measure in Equation 3.1.

Definition 107. Given a divergence function $f \in \mathfrak{F}$ with conjugate g , the Orlicz regret measure

of $X \in \mathcal{M}$ is defined as:

$$V_g(X) := \inf_{t>0} t \left(1 + \mathbb{E}g \left(\frac{X}{t} \right) \right). \quad (3.23)$$

Intuitively, it is g here that plays the role of the Young function of an Orlicz norm. In contrast to the Orlicz norm in Equation 3.14, there is no use of an absolute value here; hence the behaviour of g on $\mathbb{R}^-$ ($\mathbb{R}^+$) determines the treatment of gains (losses).

Proposition 108. *For any divergence function $f \in \mathfrak{F}$, the regret V_g has the following properties $\forall X, Y \in \mathcal{M}$:*

V1. $V_g(\lambda X) = \lambda V_g(X) \quad \forall \lambda \in \mathbb{R}^+ \quad (\text{positive homogeneity})$

V2. $V_g(X + Y) \leq V_g(X) + V_g(Y) \quad (\text{subadditivity})$

V3. $X \leq Y \text{ P-a.e.} \implies V_g(X) \leq V_g(Y) \quad (\text{monotonicity})$

V4. $V_g(X) \geq \mathbb{E}[X] \quad (\text{aversity})$

V5. *If $X \geq 0$, then $V_g(X) = 0$ if and only if $X = 0$. (nonnegative point-separating)*

Proof The proof is included in Appendix 3.10.2. ■

It will prove convenient to define Orlicz regret measures for varying levels of risk aversion ε . By a simple scaling, this parameter can be hidden in the function itself. Thus we can express the regret either by hiding the dependence on ε or by making it explicit.

Proposition 109. *With $f_\varepsilon(x) = \frac{1}{\varepsilon}f(x)$ and its conjugate $g_\varepsilon(y) = \frac{1}{\varepsilon}g(\varepsilon y)$, we have*

$$V_{g_\varepsilon}(X) = \inf_{t>0} t \left(1 + \mathbb{E}g_\varepsilon \left(\frac{X}{t} \right) \right) = \inf_{\tilde{t}>0} \tilde{t} \left(\varepsilon + \mathbb{E}g \left(\frac{X}{\tilde{t}} \right) \right). \quad (3.24)$$

Proof See Appendix 3.10.3. ■

Recall that $\bar{g}$ is the convex conjugate of $\bar{f}$, where $\bar{f}$ is the Young function canonically associated to f .

Proposition 110. *Let $f \in \mathfrak{F}$ a divergence function with conjugate g . Then it holds that $\bar{g} : \mathbb{R} \rightarrow \mathbb{R}^+ \cup \{\infty\}$, defined as:*

$$\bar{g}(y) := \begin{cases} 0 & y \leq 0 \\ g(y) & y > 0, \end{cases} \quad (3.25)$$

is the conjugate of $\bar{f}$ and it is a valid Young function when restricted to the domain $\mathbb{R}^+$.

Proof The proof is in Appendix 3.10.4. ■

We use $\|\cdot\|_{\bar{g}}^O$ to denote the Orlicz norm, where the Young function is the restriction of g to the nonnegative domain $\mathbb{R}^+$. A coherent regret measure (we define the term generally in Section 3.3.4) canonically induces a norm [Pichler, 2013].

Proposition 111. *Let $f \in \mathfrak{F}$ a divergence function. Then setting*

$$\|X\|_{V_g} := V_g(|X|) \quad (3.26)$$

defines a norm on the space $\mathcal{V}_g := \{X \in \mathcal{M} : \|X\|_{V_g} < \infty\}$. In fact it holds that $\|X\|_{V_g} = \|X\|_{\bar{g}}^O$ and $\mathcal{V}_g = \mathcal{L}^{\bar{g}}$. Furthermore, V_g is finite on the space $\mathcal{V}_g$.

Proof

$$\|X\|_{V_g} = V_g(|X|) = \inf_{t>0} t \left(1 + \mathbb{E}g \left(\frac{|X|}{t} \right) \right). \quad (3.27)$$

This is the Amemiya expression of the Orlicz norm with Young function $\bar{g}$, as g is only ever evaluated on the nonnegative domain. Therefore it holds that $\|X\|_{V_g} = \|X\|_{\bar{g}}^O$. The finiteness of V_g on $\mathcal{V}_g$, that is, $X \in \mathcal{V}_g \implies V_g(X) < \infty$, follows from [Pichler, 2013, Prop. 5], essentially from monotonicity of V_g . ■

Corollary 112. *Let $f \in \mathfrak{F}$ a divergence function. Then the norms $\|\cdot\|_{V_g}$ and $\|\cdot\|_{V_{\bar{g}}}$ are identical.*

This is due to the inherent asymmetry in our story: we care about *loss tails*, that is, right tails. Consider the meaning of $f(\frac{dQ}{dP})$: the behaviour of $f(x)$ for $0 \leq x < 1$ concerns the situation where Q underestimates with respect to the base measure. In contrast, $f(x)$ for $x \geq 1$ specifies the penalty that overestimation incurs. Due to the absolute value, the Orlicz norm is based only on the treatment of losses. Since V_g and $V_{\bar{g}}$ treat losses the same way, the potentially different behaviour for $0 \leq x < 1$ disappears in the norm-perspective. To capture this desired view, we have imposed the constraint that $f(0) < \infty$: underestimation may only be punished “finitely”, hence we can (up to equivalence) neglect underestimation in our approach.

Furthermore, the chosen risk aversion level of the regret measure is not essential, as the resulting norms are equivalent for any choice of risk aversion level.

Proposition 113. *For any risk aversion level, the norms induced by the regret measures are equivalent. Let $0 < \varepsilon_1 < \varepsilon_2$. We have*

$$\|X\|_{V_{g\varepsilon_1}} \leq \|X\|_{V_{g\varepsilon_2}} \leq \frac{\varepsilon_2}{\varepsilon_1} \|X\|_{V_{g\varepsilon_1}} \quad \forall X \in \mathcal{M}. \quad (3.28)$$

Proof The proof is in Appendix 3.10.5. ■

Remark 114. $\mathcal{V}_g$ is indeed the largest vector space, on which V_g is finite. However, this does not mean that $X \notin \mathcal{V}_g \implies V_g(X) = \infty$. This is due to the gain-loss asymmetry: an X with a heavy left tail might have $V_g(X) < \infty$, but $V_g(|X|) = \infty$ and therefore $X \notin \mathcal{V}$. Since we are concerned with right tails, this subtle distinction is inconsequential and we can focus on the natural domain $\mathcal{V}_g$.

3.3.3 ENVELOPE REPRESENTATION

Due to subadditivity and positive homogeneity, the Orlicz regret measure V_g possesses an *envelope representation* (or *dual representation*). This insightful representation will, in our case, illuminate the nature of an underlying tail reweighting mechanism in Section 3.4. As a prelude to the full dual representation, we first consider nonnegative random variables. We denote by $\mathcal{L}_+^{\bar{g}}$ the cone of nonnegative random variables in an Orlicz space $\mathcal{L}^{\bar{g}}$.

Proposition 115. Let $f \in \mathfrak{F}$ a divergence function and $\bar{g}$ be the conjugate of the associated Young divergence function $\bar{f}$. For nonnegative X , we have:

$$V_g(X) = \sup \left\{ \mathbb{E}[XZ] : Z \in \mathcal{L}^{\bar{f}}, Z \geq 0, \mathbb{E}[\bar{f}(Z)] \leq 1 \right\} \quad \forall X \in \mathcal{L}_+^{\bar{g}}. \quad (3.29)$$

Proof From Proposition 111, we know that for nonnegative X , we have $V_g(X) = \|X\|_g^O$. Then the statement is simply the definition of the Orlicz norm $\|X\|_g^O$. ■

The general envelope representation is similar.

Proposition 116. Let $f \in \mathfrak{F}$ a divergence function. Then the following envelope representation holds

$$V_g(X) = \sup \{ \mathbb{E}[XZ] : Z \in \mathcal{L}^{\bar{f}}, Z \geq 0, \mathbb{E}[f(Z)] \leq 1 \} \quad \forall X \in \mathcal{L}^{\bar{g}}. \quad (3.30)$$

Proof The proof is in Appendix 3.10.6. ■

We refer to the Z over which the supremum ranges as *dual variables*. While the dual variables are from $\mathcal{L}^{\bar{f}}$, the *primal* random variables, for which V_g is finite, are from $\mathcal{L}^{\bar{g}}$. To understand the meaning of envelope representations more generally, we turn to the general theory of coherent regret and risk measures.

3.3.4 COHERENT REGRET AND RISK MEASURES

The framework of coherent risk measures [Artzner et al., 1999, Delbaen, 2002, Föllmer and Weber, 2015] aims to provide tail-sensitive risk control. The goal is to summarize the risk

inherent in a financial position, conceptualized as a probability distribution over losses and gains. Here, there is a fundamental asymmetry: exceeding the expected loss is worse than the converse, hence the focus is on the right tail of the distribution.

In the search for possible replacements of the expectation as the tool to measure the risk of a distribution, the question of which properties a suitable replacement must satisfy arises. Artzner et al. [1999] proposed the following influential axioms for coherent risk measures³: $\forall X, Y$

- C1.** $R(\lambda X) = \lambda R(X) \quad \forall \lambda \in \mathbb{R}^+ \quad (\text{positive homogeneity})$
- C2.** $R(X + Y) \leq R(X) + R(Y) \quad (\text{subadditivity})$
- C3.** $X \leq Y \text{ P-a.e.} \implies R(X) \leq R(Y) \quad (\text{monotonicity})$
- C4.** $R(X + c) = R(X) + c \quad \forall c \in \mathbb{R} \quad (\text{translation equivariance})$

These axioms can be motivated from a financial viewpoint, but also from broader considerations regarding decision making under uncertainty; closely related are the works of Gilboa and Schmeidler [1989] in economics and Walley [1991] in imprecise probability.

The divergence risk measure in Equation 3.1 satisfies all of these and is therefore a coherent risk measure. In contrast, the Orlicz regret measure satisfies C1, C2 and C3, but not translation equivariance (C4). The Orlicz regret measure can be thought of as an “asymmetric norm” (hence *not* a norm), where translation equivariance is undesirable. The class of such functionals has been captured by Rockafellar and Uryasev [2013] as the *coherent regret measures*⁴.

Coherent regret measures and coherent risk measures are directly characterized by their envelope representations. C1 and C2, together with an appropriate closedness assumption, imply that such an R is a support function. Let $f \in \mathfrak{F}$ and g its conjugate.

Proposition 117 (Biagini and Frittelli [2009, Corollary 28], Arai [2010, Theorem 1]). *A proper functional $R : \mathcal{L}^{\mathfrak{F}} \rightarrow \mathbb{R}^+ \cup \{\infty\}$ which satisfies C1, C2, C3 and is order lower-semicontinuous⁵ allows the envelope representation*

$$R(X) = \sup_{Z \in \mathcal{Z}} \mathbb{E}[XZ], \quad \mathcal{Z} \subseteq \mathcal{L}_+^{\bar{f}}, \text{ where } \mathcal{L}_+^{\bar{f}} := \left\{ Z \in \mathcal{L}^{\bar{f}} : Z \geq 0 \text{ P-a.e.} \right\}, \quad (3.31)$$

³We translate the axioms to a loss-based orientation. Also, Artzner et al. [1999] worked with finite Ω , hence they stated C3 with the quantification $\forall \omega \in \Omega$ as opposed to P-a.e.

⁴Rockafellar and Uryasev [2013] further demanded the aversity condition $V(X) > \mathbb{E}[X]$ for $X \neq 0$. For consistency, we weaken aversity to a weak inequality. Then it is in fact a consequence of “rearrangement invariance”, see Proposition 125. We only consider rearrangement invariant coherent regret and risk measures throughout the paper. We remark that Rockafellar and Uryasev [2013] do not use the term “coherent regret measure” itself, but instead define regular measures of regret; when adding positive homogeneity, we call it *coherent* by analogy to coherent risk measures

⁵See [Biagini and Frittelli, 2009] for this technical property. We also refer to [Cheridito and Li, 2009] for a discussion of risk measures on Orlicz hearts, which are particularly interesting subspaces of Orlicz spaces.

with some set $\mathcal{Z}$, called the envelope of R , of nonnegative random variables.

If R is furthermore translation equivariant (C4), then $\mathcal{Z} \subseteq \{Z \in \mathcal{L}_+^{\bar{f}} : \mathbb{E}(Z) = 1\}$.

Thus the monotonicity of the Orlicz regret measure can be directly observed from its envelope representation. It is, however, lacking translation equivariance. For the use of coherent regret measures in risk-sensitive regression, see [Rockafellar and Uryasev, 2013].

3.3.5 FROM REGRET TO RISK

An Orlicz regret measure is in some sense like a norm, but one which distinguishes positive and negative orientation. However, it lacks the property of translation equivariance, which is desirable for a replacement of the expectation as an aggregation operator. Yet there is a simple way to canonically construct a coherent risk measure from a coherent regret measure. In this way, the f -divergence risk measure arises from the Orlicz regret measure. The f -divergence risk measure is defined as:

$$R_{g,\varepsilon}(X) := \sup \{ \mathbb{E}_Q[X] : Q \in \mathcal{Q}, d_f(Q, P) \leq \varepsilon \}. \quad (3.32)$$

For reasons of consistency, we use the conjugate of f , which is g , in the subscript. Indeed, the conjugate is the appropriate object here, as the following proposition shows. A dual variable Z which satisfies $Z \geq 0$ and $\mathbb{E}[Z] = 1$ is a valid probability density, hence we introduce

$$Q_Z(A) := \int_A Z(\omega) \, dP(\omega), \quad (3.33)$$

which is guaranteed to be a probability measure.

Proposition 118. *The divergence risk measure $R_{g,\varepsilon}(X)$ is given by an infimal convolution of the Orlicz regret measure V_{g_ε} . Let $X \in \mathcal{L}^{\bar{g}}$. Then:*

$$R_{g,\varepsilon}(X) = \inf_{\mu \in \mathbb{R}} \mu + V_{g_\varepsilon}(X - \mu) \quad (3.34)$$

$$= \inf_{\mu \in \mathbb{R}, t > 0} t \left(1 + \mu + \mathbb{E} g_\varepsilon \left(\frac{X}{t} - \mu \right) \right) \quad (3.35)$$

$$= \inf_{\tilde{\mu} \in \mathbb{R}, \tilde{t} > 0} \tilde{t} \left(\varepsilon + \tilde{\mu} + \mathbb{E} g \left(\frac{X}{\tilde{t}} - \tilde{\mu} \right) \right). \quad (3.36)$$

Its envelope representation is:

$$R_{g,\varepsilon}(X) = \sup \left\{ \mathbb{E}[XZ] : Z \in \mathcal{L}^{\bar{f}}, Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}f(Z) \leq \varepsilon \right\} \quad (3.37)$$

$$= \sup \left\{ \mathbb{E}_{Q_Z}[X] : Z \in \mathcal{L}^{\bar{f}}, Q_Z \in \mathcal{Q}, d_f(Q_Z, P) \leq \varepsilon \right\} \quad (3.38)$$

$$= \sup \left\{ \mathbb{E}_{Q_Z}[X] : Z \in \mathcal{L}^{\bar{f}}, Q_Z \in \mathcal{Q}, \mathbb{E}_{Q_Z}[Y] \leq V_{g_\varepsilon}(Y) \quad \forall Y \in \mathcal{L}^{\bar{g}} \right\}. \quad (3.39)$$

Proof The proof is in Appendix 3.10.7. ■

We typically write R_g with $\varepsilon = 1$ in the following. The infimal convolution in Equation 3.34 ensures translation equivariance of the resulting functional. The combination of monotonicity and translation equivariance then guarantees that the dual variables are legitimate probability measures, that is, $Z \geq 0$ and $\mathbb{E}[Z] = 1$ (see for instance [Fröhlich and Williamson, 2024a]). An interesting perspective is also provided by Equation 3.39: the envelope is the set of linear risk measures, i.e. expectations, which are $L^{\bar{g}}$ -pointwise dominated by the regret measure. In the language of the *imprecise probability* literature, this process of projecting from V_g to R_g by forming a supremum over the set of dominated expectations is known as *natural extension* [Walley, 1991, Troffaes and De Cooman, 2014]. This provides the rationale for interpreting the envelope as an ambiguity set and embeds the divergence risk measure in the field of imprecise probability.

The infimum-based representation in Equation 3.36, under various technical assumptions, has been around in the literature [Ben-Tal and Nemirovski, 1999, Bayraksan and Love, 2015, Shapiro, 2017, Dommel and Pichler, 2020]. However, our derivation from the Orlicz regret measure provides a new perspective and intuition for the result; we remark that Dommel and Pichler [2020], themselves inspired by Ahmadi-Javid [2012] and Ahmadi-Javid and Pichler [2017], have investigated the divergence risk measure in the context of Orlicz spaces, but without reference to an underlying Orlicz regret measure.

Example 119 (Bayraksan and Love [2015]). With $f_{\text{CVar}_\alpha} := i_{[0,1/(1-\alpha)]}$ we obtain the well-known infimal convolution expression of the conditional value at risk:

$$\text{CVar}_\alpha(X) = \inf_{\mu \in \mathbb{R}} \mu + \frac{1}{1-\alpha} \mathbb{E}[(X - \mu)^+], \quad (3.40)$$

where the special case $\alpha = 0$ recovers the expectation.

Similar to the Orlicz regret measure, the divergence risk induces a norm by setting $\|X\|_{R_g} := R_g(|X|)$, with the induced space $\mathcal{R}_g = \{X \in \mathcal{M} : R_g(|X|) < \infty\}$. Then the following holds:

Proposition 120. *The norms $\|\cdot\|_{R_g}$ and $\|\cdot\|_{V_g}$ are equivalent. This implies that the spaces $\mathcal{V}_g$ and $\mathcal{R}_g$ are identical.*

Proof Since $\|\cdot\|_{V_g} = \|X\|_{\bar{g}}^O$, Theorem 4.5 in [Dommel and Pichler, 2020] is applicable. The authors did assume finiteness of f , but this is not required in the proof of the statement. ■

As a consequence, the divergence risk measure is finite on $\mathcal{V}_g = \mathcal{R}_g$ and this is the largest vector space, on which it is finite (cf. Remark 114).

Corollary 121 (Dommel and Pichler [2020]). *Let $f \in \mathfrak{F}$. Then the divergence risk measure*

is equivalent to its corresponding Young divergence risk measure, that is, $\|\cdot\|_{R_g}$ is equivalent to $\|\cdot\|_{R_g^*}$.

Proof The statement holds since $\|\cdot\|_{R_g}$ and $\|\cdot\|_{V_g}$ are equivalent and $\|\cdot\|_{V_g} = \|\cdot\|_{V_g^*}$ (Corollary 112). ■

Corollary 122 (Dommel and Pichler [2020]). Consider $f_{\varepsilon_1}, f_{\varepsilon_2} \in \mathfrak{F}$, that is, variants of the same underlying divergence function at different risk aversion thresholds. Then $\|\cdot\|_{R_{g_{\varepsilon_1}}}$ and $\|\cdot\|_{R_{g_{\varepsilon_2}}}$ are equivalent.

Proof This readily follows from the equivalence of the underlying Orlicz regret measures. ■

3.4 REARRANGEMENT INVARIANT BANACH NORMS AND FUNDAMENTAL FUNCTIONS

We now embed the Orlicz regret measure and the divergence risk measure in the framework of *rearrangement invariant Banach function spaces*, which provides more insight into the envelope representation and allows us to investigate equivalence relationships of the induced spaces.

A desirable property of a coherent regret or risk measure is *rearrangement invariance*, which is known as *law invariance* in the risk measure literature. The characterizing property is that a rearrangement invariant regret (risk) measure depends only on the distribution of the random variable, not on the order in which outcomes are associated with elementary events ω . To capture this notion, we define the *distribution function*⁶ $\mu_X : \mathbb{R}^+ \rightarrow [0, 1]$ of a random variable $X \in \mathcal{M}$ as

$$\mu_X(\lambda) := P(\{\omega \in \Omega : |X(\omega)| > \lambda\}). \quad (3.41)$$

For nonnegative random variables, this decreasing (non-increasing) and right-continuous function is just the survival function $1 - F_X$. We say that random variables X and Y are *equimeasurable* if their distribution functions are the same, that is, $\mu_X(\lambda) = \mu_Y(\lambda) \forall \lambda \geq 0$. For each $X \in \mathcal{M}$, the right-continuous generalized inverse of its distribution function $X^* : [0, 1] \rightarrow \mathbb{R}^+$,

$$X^*(\omega) := \inf\{\lambda \geq 0 : \mu_X(\lambda) \leq \omega\}, \quad (3.42)$$

⁶This is the standard terminology in the rearrangement invariant Banach space literature; it is not to be confused with the *cumulative* distribution function F_X .

is equimeasurable with X itself. We call X^* the *decreasing rearrangement* of X . In the language of probability theory, X^* is just the quantile function backwards: $X^*(t) = F_{|X|}^{-1}(1-t)$. As an intuition, X^* is the continuous analogue of sorting a list of values in decreasing order.

3.4.1 A PRIMER ON REARRANGEMENT INVARIANT SPACES

Definition 123. A Banach space $(\mathcal{X}, \|\cdot\|)$ is called *rearrangement invariant* if [Rubshtein et al., 2016]

R1. If $|X| \leq |Y|$ and $Y \in \mathcal{X}$, then $X \in \mathcal{X}$ and $\|X\| \leq \|Y\|$.

R2. If X and Y are equimeasurable and $Y \in \mathcal{X}$, then $X \in \mathcal{X}$ and $\|X\| = \|Y\|$.

We call such a norm $\|\cdot\|$ *rearrangement invariant (r.i.)*.

From their definitions, it is clear that $\|\cdot\|_{V_g}$ and $\|\cdot\|_{R_g}$ are r.i. norms and therefore $\mathcal{V}_g$ and $\mathcal{R}_g$ are r.i. spaces. In general, any coherent regret measure or risk measure, which satisfies R2, induces a legitimate r.i. norm by taking absolute values.

Example 124. The Lebesgue spaces $\mathcal{L}^p$, $1 \leq p \leq \infty$, are r.i. spaces, which are in particular Orlicz spaces.

Interestingly, any r.i. space is in between $\mathcal{L}^1$ and $\mathcal{L}^\infty$ due to the following embedding theorem.

Proposition 125 (Bennett and Sharpley [1988, p. 77]). Let $\mathcal{X}$ be any ri space. Then:

$$\mathcal{L}^\infty \subseteq \mathcal{X} \subseteq \mathcal{L}^1, \quad (3.43)$$

and the norms are in the relationship⁷:

$$\|X\|_{\mathcal{L}^1} \leq \|X\|_{\mathcal{X}} \leq \|X\|_{\mathcal{L}^\infty} \quad \forall X \in \mathcal{M}. \quad (3.44)$$

Finally, we need the concept of the *associate norm*, which is linked to envelope representations.

Definition 126. Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ a Banach space, where $\|\cdot\|_{\mathcal{X}}$ satisfies the monotonicity condition R1. Its associate norm $\|\cdot\|_{\mathcal{X}'}$ is defined as:

$$\|X\|_{\mathcal{X}'} := \sup \left\{ \int_0^1 |X(\omega)Z(\omega)| \, d\omega : \|Z\|_{\mathcal{X}} \leq 1, Z \in \mathcal{X} \right\}. \quad (3.45)$$

The associate space is $\mathcal{X}' := \{X \in \mathcal{M} : \|X\|_{\mathcal{X}'} < \infty\}$.

⁷Under the assumption that $\|1\|_{\mathcal{X}} = 1$, otherwise renorm first with the multiplicative factor $1/\|1\|_{\mathcal{X}}$.

If $\|\cdot\|_{\mathcal{X}}$ is furthermore an r.i. norm (satisfies R2), then the associate norm can be expressed in terms of decreasing rearrangements.

Proposition 127 (Bennett and Sharpley [1988, p. 60]). *Let $\|\cdot\|_{\mathcal{X}}$ be an r.i. norm. Then its associate norm is:*

$$\|X\|_{\mathcal{X}'} = \sup \left\{ \int_0^1 X^*(\omega) Z^*(\omega) \, d\omega : \|Z\|_{\mathcal{X}} \leq 1, Z \in \mathcal{X} \right\}. \quad (3.46)$$

This remarkable result asserts that for an r.i. norm, the associate relationship can be understood in terms of certain admissible quantile reweightings; recall $Z^*(\omega) = 1 - F_{|Z|}^{-1}(\omega)$.

Example 128 (Bennett and Sharpley [1988, p. 275]). For a pair of conjugate Young functions Φ and Ψ , consider $\mathcal{L}^{\Phi}$ with the Luxemburg norm $\|\cdot\|_{\Phi}^L$. Its associate space is $\mathcal{L}^{\Psi}$ with the Orlicz norm $\|\cdot\|_{\Psi}$ as the associate norm.

Example 129 (Bennett and Sharpley [1988, p. 10]). The Lebesgue spaces $\mathcal{L}^p$: let $1 \leq p \leq \infty$ and the norm $\|\cdot\|_{\mathcal{L}^p}$. Its associate space is $\mathcal{L}^q$, where $1/p + 1/q = 1$, and the associate norm is $\|\cdot\|_{\mathcal{L}^q}$. Note that $p = 1$ is paired with $q = \infty$.

3.4.2 FUNDAMENTAL FUNCTIONS OF ORLICZ REGRETS AND RISK MEASURES

Rearrangement invariant norms have widely varying tail sensitivity. For instance, the $\mathcal{L}^1$ norm (expectation) is tail-*insensitive*, whereas the $\mathcal{L}^2$ norm (variance) is rather tail-sensitive. To “stratify” the space of all rearrangement invariant norms, Bennett and Sharpley [1988] and Fröhlich and Williamson [2024a] emphasize the importance of the *fundamental function*, a one-dimensional function, which characterizes indeed *fundamental* properties of the norm. For instance, its derivative at the origin can be used to characterize norm equivalences.

Definition 130. *Let $\mathcal{X}$ be an r.i. space with r.i. norm $\|\cdot\|_{\mathcal{X}}$. Let $A_t \subseteq \Omega$ be an arbitrary Lebesgue measurable set with measure $P(A_t) = t$. The fundamental function $\phi_{\mathcal{X}} : [0, 1] \rightarrow \mathbb{R}^+$ is defined as:*

$$\phi_{\mathcal{X}}(t) := \|\chi_{A_t}\|_{\mathcal{X}}. \quad (3.47)$$

Due to rearrangement invariance, the choice of A_t is arbitrary. We will also speak of the fundamental function of a coherent regret (risk) measure, which due to nonnegativity of indicator functions just coincides with the fundamental function of the induced norm.

Proposition 131 (Bennett and Sharpley [1988, p. 67]). *For any r.i. space $\mathcal{X}$, $\phi_{\mathcal{X}}$ satisfies:*

$$\phi_{\mathcal{X}} \text{ is increasing.} \quad (3.48)$$

$$t \mapsto \phi_{\mathcal{X}}(t)/t \text{ is decreasing.} \quad (3.49)$$

$$\phi_{\mathcal{X}} \text{ is continuous except perhaps at 0 and } \phi_{\mathcal{X}}(0) = 0. \quad (3.50)$$

A function ϕ_χ satisfying these properties is called *quasiconcave*. Every increasing concave function, which vanishes only at the origin, is quasiconcave, but a quasiconcave function need not be concave in general.

Example 132. The $\|\cdot\|_{\mathcal{L}^1}$ norm has fundamental function $\phi_{\mathcal{L}^1}(t) = t$. In contrast, the $\|\cdot\|_{\mathcal{L}^\infty}$ norm (supremum) has $\phi_{\mathcal{L}^\infty}(t) = \chi_{(0,1]}$, which immediately jumps to 1, hence it is not continuous at 0. It holds that any fundamental function lies pointwise in between those of $\mathcal{L}^1$ and $\mathcal{L}^\infty$ (cf. also Proposition 125).

We denote the fundamental function of the Orlicz regret as $\phi_{V_g}(t) = V_g(\chi_{A_t}) = \|\chi_{A_t}\|_{V_g}$ and similarly, the fundamental function of the divergence risk measure as $\phi_{R_g}(t) = R_g(\chi_{A_t}) = \|\chi_{A_t}\|_{R_g}$. We have that $\phi_{R_g}(1) = 1$ due to translation equivariance of R_g , which implies $R_g(c) = c \ \forall c \in \mathbb{R}$ [Rockafellar and Uryasev, 2013]. We begin by characterizing the fundamental function of the Orlicz regret measure. We use the fact that a left-continuous function f possesses a generalized left-continuous inverse $f^{-1}(y) := \sup\{x : f(x) < y\}$.

Proposition 133 (Kosmol and Müller-Wichards [2011]). *Let $f \in \mathfrak{F}$. Then:*

$$\phi_{V_g}(t) = tf^{-1}\left(\frac{1}{t}\right) \quad (3.51)$$

is the fundamental function of the Orlicz regret measure and it is concave.

Proof The proof is in Appendix 3.11.1. ■

Proposition 134. *Let $f \in \mathfrak{F}$. Then ϕ_{V_g} and ϕ_{R_g} both have the following properties (denote by ϕ either of them):*

FF1. $\phi(0+) := \lim_{t \downarrow 0} \phi(t) = 0$.

FF2. *If f is finite on $\mathbb{R}^+$, then $\phi'(0) = \lim_{t \downarrow 0} \frac{\phi(t)}{t} = \infty$. Otherwise, $\phi'(0) < \infty$.*

Proof The proof is in Appendix 3.11.2. ■

These properties have direct consequences for norm equivalence relationships.

Corollary 135. *Let $f \in \mathfrak{F}$. Then $\mathcal{L}^\infty \subsetneq \mathcal{V}_g = \mathcal{R}_g$, where g is the conjugate of f .*

Proof As the smallest rearrangement invariant space, $\mathcal{L}^\infty \subseteq \mathcal{V}_g = \mathcal{R}_g$ holds. But it is indeed a strict subset: from [Haroske, 2006, Prop. 3.4 iii)] we know that $\mathcal{R}_g \subseteq \mathcal{L}^\infty$ holds if and only if $\lim_{t \downarrow 0} \frac{1}{\phi(t)} < \infty$. But $\lim_{t \downarrow 0} \phi(t) = 0$ and therefore $\lim_{t \downarrow 0} \frac{1}{\phi(t)} = \infty$. ■

Corollary 136. *Let $f \in \mathfrak{F}$. If f is finite, then $\mathcal{V}_g = \mathcal{R}_g \subsetneq \mathcal{L}^1$. In contrast, if f is not finite, then $\mathcal{V}_g = \mathcal{R}_g = \mathcal{L}^1$.*

Proof As the largest rearrangement invariant space, $\mathcal{V}_g = \mathcal{R}_g \subset \mathcal{L}^1$ holds. However, we have from Theorem 31 in [Fröhlich and Williamson, 2024a], that all rearrangement invariant norms with $\phi'(0) < \infty$ are equivalent to $\mathcal{L}^1$. In contrast, if $\phi'(0) = \infty$, the norm cannot be equivalent to $\mathcal{L}^1$ [Fröhlich and Williamson, 2024a, Corollary 32] and thus $\mathcal{R}_g \subsetneq \mathcal{L}^1$. ■

These corollaries are well-known in Orlicz space theory, but it is instructive to see how they easily follow from properties of the fundamental function.

A general formula for the fundamental function ϕ_{R_g} of the divergence risk measure seems infeasible⁸. However, we observe that in the case of a Young divergence function, it has a simple expression.

Proposition 137. *Let $f \in \mathfrak{F}_Y$. Then the fundamental function of the divergence risk measure is*

$$\phi_{R_g}(t) = \min \left\{ 1, t f^{-1} \left(\frac{1}{t} \right) \right\}, \quad (3.52)$$

that is, a capped version of the fundamental function of the corresponding regret measure.

Proof The proof is in Appendix 3.11.3. ■

Example 138. For the Young KL divergence, where $\bar{f}_{\text{KL}}(x) = (x \log(x) - x + 1) \cdot \chi_{[1, \infty)}$, the fundamental function is

$$\phi_{\text{KL}}(t) = \frac{t(1/t - 1)}{W\left(\frac{(1/t - 1)}{e}\right)}, \quad (3.53)$$

where W is the Lambert W function.

Example 139. For the Young χ^2 divergence, where $\bar{f}_{\chi^2}(x) = (x - 1)^2 \cdot \chi_{[1, \infty)}$, the fundamental function is

$$\phi_{\chi^2}(t) = \min \left\{ 1, t + \sqrt{t} \right\}. \quad (3.54)$$

Example 140. For CVar_α , the fundamental function is $\phi_{\text{CVar}_\alpha}(t) = \min \left\{ 1, \frac{t}{1-\alpha} \right\}$. For $\alpha = 0$, $\phi_{\text{E}}(t) = t$.

We illustrate these fundamental functions in Figure 3.1.

The fundamental function characterizes important aspects of an r.i. norm. Of particular interest is $\phi'(0)$, as is exemplified in Corollary 136. Furthermore, we have the following result which will be useful in Section 3.6.

⁸It is easy to write down an equation, but one which is typically difficult to solve.

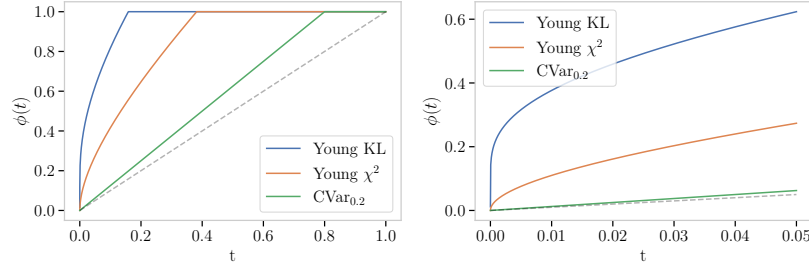


Figure 3.1: The fundamental functions of the Young KL, the Young χ^2 and the $\text{CVar}_{\alpha=0.2}$ divergence risk measures, all with $\varepsilon = 1$. In practice, this will be an extreme choice of ε . The grey dotted line is the identity. Left: the full fundamental function. Right: the fundamental function for $0 \leq x \leq 0.05$. Here, the difference in tail behaviour is better visible: the fundamental function of the Young KL divergence risk measure is extremely steep near the origin, whereas that of $\text{CVar}_{\alpha=0.2}$ is close to the identity, which corresponds to the expectation. All of these fundamental functions are continuous at the origin.

Proposition 141. *Let $\phi : [0, 1] \rightarrow \mathbb{R}^+$ be an increasing concave function with $\phi(0) = 0$ and $\phi(0+) = 0$. Then there exists an Orlicz norm $\|\cdot\|_{\Psi}^{\text{O}}$ which has ϕ as its fundamental function and a supercoercive Young function Φ as the conjugate of Ψ .*

Proof The proof is in Appendix 3.11.4. ■

3.4.3 EQUIVALENCES OF ORLICZ NORMS

Through the framework of r.i. Banach spaces, we can associate each divergence risk measure with its natural space, on which it is finite. However, many non-identical divergence risk measures may induce the same space. Then, however, they are equivalent. Recall that two r.i. norms $\|\cdot\|_{\mathcal{X}_1}$ and $\|\cdot\|_{\mathcal{X}_2}$ are said to be equivalent if

$$\exists c, c' > 0 : \quad c \cdot \|X\|_{\mathcal{X}_1} \leq \|X\|_{\mathcal{X}_2} \leq c' \cdot \|X\|_{\mathcal{X}_1}, \quad \forall X \in \mathcal{M}. \quad (3.55)$$

Two r.i. norms are equivalent if and only if their respective spaces coincide [Bennett and Sharpley, 1988, p. 7]. As an example, we have seen that the parameter ε does not matter up to equivalence (Corollary 122).

Why are we interested in norm (non)-equivalences? Whether two norms are equivalent or not *prima facie* seems like a very coarse criterion for considering them similar. After all, the binary equivalence notion does not tell us how large the multiplicative constants are. Equivalence means that the norms are finite on the same set of functions: yet in practice, when computing the norm of an empirical loss distribution, we will never encounter infinities, since all real-world samples are bounded (from $\mathcal{L}^\infty$). In this way, unbounded tails are a useful, asymptotic idealization used to model the frequency of extreme events.

In a sense, those divergence risk measures with the same space share the same tail sensi-

tivity; although we will later argue that sharing the same fundamental function provides a coarser and simpler way to capture tail sensitivity. Consider what the risk assessment $R(X) = \infty$ means: the decision maker considers the tail risk in X to be *unbearable*, that is, there is no finite certainty equivalent (gain), which in exchange would make the risk bearable. For example, irrespective of the choice of $\varepsilon > 0$, a decision maker who uses the KL divergence risk measure would assess $R(X) = \infty$ for a random variable X which is not subexponential, (see Example 156) and thus prefer any subexponential random variable over it. Thus, investigating equivalence relationships between spaces reveals different attitudes towards tail risk. In general, the more tail-sensitive a risk measure, the smaller the induced space.

We briefly consider equivalence of Orlicz norms, as we require these results in what follows.

Definition 142 (Rubshtein et al. [2016, p. 208]). *Let g_1, g_2 be Young functions. We write $g_1 \succ g_2$ when*

$$g_2(x) \leq b g_1(ax), \quad \forall x \geq 0, \quad (3.56)$$

for some $b > 0, a > 0$.

We write $\sim$ when both $g_1 \succ g_2$ and $g_2 \succ g_1$ holds. We see immediately that $f \sim f_\varepsilon$ for any $\varepsilon > 0$ if $f \in \mathfrak{F}_Y$.

Proposition 143 (Rubshtein et al. [2016, p. 209]). *Let g_1, g_2 be Young functions. The following statements are equivalent:*

E1. $g_1 \succ g_2$.

E2. $\mathcal{L}^{g_1} \subseteq \mathcal{L}^{g_2}$.

E3. $\exists c > 0: \phi_{V_{g_2}}(t) \leq c \phi_{V_{g_1}}(t), \quad \forall t \in [0, 1]$.

E4. $\exists c > 0: \|X\|_{V_{g_2}} \leq c \|X\|_{V_{g_1}}$.

Example 144. The norm induced by the (Young) χ^2 divergence risk measure is equivalent to the $\mathcal{L}^2$ norm, which has fundamental function $\phi_{\mathcal{L}^2}(t) = \sqrt{t}$. Observe that

$$\sqrt{t} \leq t(1 + \sqrt{1/t}) \leq 2\sqrt{t}, \quad \forall t \in (0, 1] \quad (3.57)$$

and $\phi(t) = t(1 + \sqrt{1/t})$ is the fundamental function of the χ^2 Orlicz regret measure. Therefore, the natural r.i. space for the χ^2 divergence risk measure is that of random variables X for which $|X|$ has a variance.

Example 145. CVar_α is equivalent to the expectation for all $\alpha \in [0, 1)$. As $\alpha \rightarrow 1$, it yields $\mathcal{L}^\infty$, however, and the risk measure is the essential supremum.

Example 146. Let h be a Young function, which need not be divergence function, e.g. $h(x) = x^2$. Then $f(x) := (h(x) - h(1)) \cdot \chi_{[1,\infty)}$ is valid Young divergence function which is furthermore equivalent to h . As an example, $h(x) = x^2$ is equivalent to $f(x) = ((x - 1)^2) \cdot \chi_{[1,\infty)}$, the χ^2 Young divergence function.

Proof Obviously f is a valid Young divergence function. Consider the definition of the Luxemburg norm

$$\|X\|_f^L = \inf \left\{ \lambda > 0 : \mathbb{E} \left(h \left[\frac{|X|}{\lambda} \right] - h(1) \right) \cdot \chi_{[1,\infty)} \leq 1 \right\}. \quad (3.58)$$

to see that $\mathcal{L}^h = \mathcal{L}^f$. ■

3.5 TAIL-SPECIFIC MARCINKIEWICZ AND LORENTZ NORMS

Our overarching goal is to show how r.i. norms (and hence, coherent risk measures) can be constructed which exhibit a desired tail sensitivity; thereby yielding natural spaces of “tail classes”, on which the norm is finite, for instance the space of subexponential random variables. In this section we advance on this idea by showing how tail sensitivity is connected to the fundamental function and the *envelope function*. Intuitively, a desired tail sensitivity is in a very close correspondence with the fundamental function. For each thus constructed fundamental function, we consider the *smallest* and *largest* compatible coherent risk measure. In the next section, we derive the corresponding divergence risk measure via the same approach, which is then situated in between these extremes or coincides with either of them.

In the language of the imprecise probability literature, the fundamental function specifies a *coherent upper probability* [Walley, 1991], which is the restriction of the risk measure to indicator functions. In classical precise probability theory, one can take either the expectation operator or a probability measure as the primitive notion, since applying an expectation to indicator functions yields a probability and vice versa, an expectation is uniquely defined via integration. In the more general theory of imprecise probability, this one-to-one correspondence does not hold anymore: a fundamental function can in general be extended in multiple ways to a coherent risk measure⁹. However, the fundamental function still characterizes *fundamental* properties of a norm. For the relation of imprecise probability and coherent risk measures we refer to [Vicig, 2008] and [Fröhlich and Williamson, 2024a]. Our approach is to specify a tail sensitivity via the *envelope function*, which is in a one-to-one relationship with the fundamental function. The specification of an envelope function is a

⁹In expectational cases, there is only a single extension, see Fröhlich and Williamson [2024a, Section 4].

least upper bound on the tail behaviour of the random variables, which are contained in the space. In some cases, the envelope function is itself in the space; in others it is a natural limit object, which is “just outside” the space.

Recall that for an r.i. norm, we have the associate relationship:

$$\|X\|_{\mathcal{X}} = \sup \left\{ \int_0^1 X^*(\omega) Z^*(\omega) \, d\omega : \|Z\|_{\mathcal{X}'} \leq 1, Z \in \mathcal{X}' \right\}. \quad (3.59)$$

By specifying a set $\mathcal{Z} := \{Z^* : \|Z\|_{\mathcal{X}'} \leq 1\}$, we have fully specified the norm in terms of which quantile reweightings are admissible. Conversely, any set of nonnegative and decreasing Z^* leads to a legitimate r.i. norm [see e.g. Fröhlich and Williamson, 2024a, Section 4.7]. Thus, an r.i. norm is fully identified by its set of nonnegative and decreasing functions Z^* on $[0, 1]$. However, the fundamental function, which is a far simpler object than a whole set of Z^* , already stratifies the space of r.i. norms in a useful way regarding tail sensitivity. To see this more clearly, we need to introduce the concept of *envelope functions* [Haroske, 2006].

Definition 147 (Haroske [2006]). *Let $\mathcal{X}$ be any r.i. space. Define $E_{\mathcal{X}} : (0, 1] \rightarrow \mathbb{R}^+ \cup \{\infty\}$ by:*

$$E_{\mathcal{X}}(t) := \sup\{X^*(t) : \|X\|_{\mathcal{X}} \leq 1\}. \quad (3.60)$$

$E_{\mathcal{X}}$ is right-continuous and decreasing, hence $E_{\mathcal{X}}^* = E_{\mathcal{X}}$.

Recall that since $X^*(t) = F_{|\mathcal{X}|}^{-1}(1 - t)$, the tail behaviour is encoded in the behaviour as $t \downarrow 0$.

Proposition 148 (Haroske [2006, p. 53]). *The envelope function is in a one-to-one relation with the fundamental function via*

$$E_{\mathcal{X}}(t) = \frac{1}{\phi_{\mathcal{X}}(t)}, \quad \forall t \in (0, 1]. \quad (3.61)$$

Since multiplicative factors do not matter up to norm equivalence, we shall be concerned with equivalence classes of envelope functions and fundamental functions. That is, similar to E3, we call two envelope functions equivalent if

$$\exists c, c' > 0 : \quad c \cdot E_{\mathcal{X}_1}(t) \leq E_{\mathcal{X}_2}(t) \leq c' \cdot E_{\mathcal{X}_1}(t), \quad \forall t \in (0, 1]. \quad (3.62)$$

and likewise for fundamental functions. Intuitively, the envelope function constrains the tail behaviour of random variables in the space. It is however more instructive to change perspective and consider the associate space. We need the following technical result.

Proposition 149 (Rubshtein et al. [2016, p. 135]). *Let $\mathcal{X}$ be an r.i. space with fundamental function $\phi_{\mathcal{X}}$. Then the fundamental function of the associate space $\mathcal{X}'$ is $\phi_{\mathcal{X}'}(t) = \frac{t}{\phi_{\mathcal{X}}(t)}$.*

We may refer to $\phi_{\mathcal{X}'}$ as the *associate* fundamental function. Consequently, the envelope function in the associate space is:

$$E_{\mathcal{X}'}(t) = \frac{1}{\phi_{\mathcal{X}'}(t)} = \frac{\phi_{\mathcal{X}}(t)}{t} = \sup\{Z^*(t) : \|Z\|_{\mathcal{X}'} \leq 1\}. \quad (3.63)$$

Thus, the envelope function constrains the tail behaviour of the dual variables Z^* in Equation 3.59 and it is the least such upper bound, the supremum. All in all, we have one-to-one relations between $\phi_{\mathcal{X}}, \phi_{\mathcal{X}'}, E_{\mathcal{X}}$ and $E_{\mathcal{X}'}$; which object is easiest to specify depends on motivation and perspective. In what follows, we often write simply ϕ instead of $\phi_{\mathcal{X}}$ for the primal fundamental function to prevent excessive subscripts. Essentially, the choice of either of these objects is a choice of a coherent upper probability underlying an r.i. norm. Also observe that

$$\phi'(0) = \lim_{t \downarrow 0} \frac{\phi_{\mathcal{X}}(t)}{t} = \lim_{t \downarrow 0} E_{\mathcal{X}'}(t). \quad (3.64)$$

The derivative of ϕ at the origin has been used to characterize norm equivalences (see for instance Corollaries 135, 136 and [Fröhlich and Williamson, 2024a, Section 4.9]). Via the envelope function perspective we see that it corresponds to a constraint on the tail behaviour of the dual variables.

The envelope function naturally suggests the following approach to the construction of tail-sensitive r.i. norms: specify a constraint on the tail behaviour of dual or primal variables by specifying an envelope function ($E_{\mathcal{X}}$ or $E_{\mathcal{X}'}$) as the decreasing rearrangement of a reference distribution (e.g. as the decreasing rearrangement of an arbitrary random variable with exponential distribution); then construct an r.i. norm with the corresponding fundamental function. For example, when constraining the quantiles of primal variables X to be bounded from above by the quantiles of an exponential distribution, we will arrive at the class of subexponential variables and the KL divergence risk measure.

For any concave fundamental function, we may consider the smallest and the largest r.i. norm, with the Orlicz norm in between. The largest norm is known as the *Lorentz norm* and the smallest is the *Marcinkiewicz norm*.

3.5.1 THE MARCINKIEWICZ NORM

Assume we have specified a legitimate $E_{\mathcal{X}}$ and thus a quasiconcave $\phi(t) = 1/E_{\mathcal{X}}(t)$, since a fundamental functions must always be quasiconcave (Proposition 131). A first (unsuccessful) attempt to define a norm, which captures this constrained tail behaviour, is to define the following.

Definition 150 (Kreĭn et al. [1982b]). *Let $\phi(t) = 1/E_{\mathcal{X}}(t)$ be a quasiconcave fundamental*

function. We define the Marcinkiewicz quasi-norm as:

$$|X|_{M_\phi} := \sup_{0 < t \leq 1} \phi(t) X^*(t). \quad (3.65)$$

This naturally expresses the constraint on the quantile behaviour of X , compared to the reference $E_{\mathcal{X}}$. Intuitively:

$$|X|_{M_\phi} < \infty \iff \exists c : X^*(t) \leq c \cdot E_{\mathcal{X}}(t), \quad t \in (0, 1]. \quad (3.66)$$

Here, $E_{\mathcal{X}}$ is the decreasing rearrangement of our reference distribution, e.g. exponential. Unfortunately, $|\cdot|_{M_\phi}$ is not a norm, as it does not satisfy the triangle inequality. However, we can define the closely related Marcinkiewicz norm. By integrating the decreasing rearrangement, we obtain the *maximal function* [Bennett and Sharpley, 1988, pp. 52-53]:

$$X^{**}(t) := \frac{1}{t} \int_0^t X^*(\omega) \, d\omega, \quad t > 0. \quad (3.67)$$

Since $X^*(t) = F_{|X|}^{-1}(1-t)$, we find that $X^{**}(t) = \text{CVar}_{1-t}(|X|)$; thus $X^{**}(0)$ is the essential supremum of $|X|$ and $X^{**}(1) = \mathbb{E}[|X|]$.

Definition 151. Given any quasiconcave fundamental function ϕ , the Marcinkiewicz norm $\|\cdot\|_{M_\phi}$ is defined as:

$$\|X\|_{M_\phi} := \sup_{0 < t \leq 1} \{\phi(t) X^{**}(t)\} \quad (3.68)$$

$$= \sup_{0 < t \leq 1} \{\phi(t) \text{CVar}_{1-t}(|X|)\}. \quad (3.69)$$

We denote the Marcinkiewicz space as $M_\phi := \{X \in \mathcal{M} : \|X\|_{M_\phi} < \infty\}$.

This norm also has fundamental function ϕ . Intuitively, the Marcinkiewicz norm measures the growth of the CVar's, but weighted with the fundamental function. In the risk measure community, the Marcinkiewicz norm has been unknowingly re-discovered by Pichler [2013, Section 5]. The difference between $|\cdot|_{M_\phi}$ and $\|\cdot\|_{M_\phi}$ is the use of the decreasing rearrangement (backwards quantiles) vs. the maximal function (backwards CVar's). Subadditivity of the maximal function yields a legitimate norm in the latter case. However, the quasi-norm and the norm are in fact equivalent in many cases of interest. Recall that $E_{\mathcal{X}} = E_{\mathcal{X}}^*$, since $E_{\mathcal{X}}$ is decreasing, and the maximal function is $E_{\mathcal{X}}^{**}$.

Proposition 152 (Kreĭn et al. [1982b]). Let ϕ be quasiconcave. The following are equivalent:

$$\mathbf{M1.} \quad \forall s > 1: \sup_{0 < t \leq 1} \frac{\phi_{\mathcal{X}'}(t/s)}{\phi_{\mathcal{X}'}(t)} < 1.$$

$$\mathbf{M2.} \quad |\cdot|_{M_\phi} \text{ and } \|\cdot\|_{M_\phi} \text{ are equivalent: } \exists K > 0 : |X|_{M_\phi} \leq \|X\|_{M_\phi} \leq K \cdot |X|_{M_\phi}.$$

M3. $\|\phi_{\mathcal{X}'}(t)/t\|_{M_\phi} = \|E_{\mathcal{X}}\|_{M_\phi} < \infty$.

M4. $E_{\mathcal{X}}$ and $E_{\mathcal{X}}^{**}$ are equivalent: $\exists K > 0 : E_{\mathcal{X}}(t) \leq E_{\mathcal{X}}^{**}(t) \leq K \cdot E_{\mathcal{X}}(t), \quad \forall t \in (0, 1]$.

M5. Let $\phi_{\text{CVar}}(t) = 1/E_{\mathcal{X}}^{**}(t)$. $\|\cdot\|_{M_\phi}$ and $\|\cdot\|_{M_{\phi_{\text{CVar}}}}$ are equivalent.

Proof The proof is in Appendix 3.12.1. ■

A necessary, but not sufficient criterion for these conditions to hold is that $\phi_{\mathcal{X}'}$ be strictly increasing. Intuitively, the conditions amount to requiring that the CVar's of the envelope function do not grow too rapidly.

The Marcinkiewicz norm naturally expresses a tail sensitivity based on CVar's. If any of the conditions in Proposition 152 are fulfilled, it does not matter up to equivalence whether the approach to specify a tail sensitivity is based on CVar's or quantiles of a reference distribution: the same space results.

To summarize, our approach is as follows: specify an envelope function $E_{\mathcal{X}}$ as the decreasing rearrangement of a reference distribution with a desired tail behaviour. This amounts to a choice of a coherent upper probability in form of the fundamental function $\phi_{\mathcal{X}}$. Then consider the corresponding Marcinkiewicz quasi-norm, which is typically equivalent to the Marcinkiewicz norm (we later demonstrate that the equivalence “often” holds). In any case, the Marcinkiewicz norm occupies a special status among all r.i. norms, which are compatible with a specified fundamental function. To state the result, we need to define its associate norm, the *Lorentz norm*.

3.5.2 THE LORENTZ NORM

Definition 153. Given any concave fundamental function ϕ , the Lorentz norm $\|\cdot\|_{\Lambda_\phi}$ is defined as:

$$\|X\|_{\Lambda_\phi} := \int_0^1 X^*(\omega) \, d\phi(\omega) \quad (3.70)$$

$$= X^*(0)\phi(0+) + \int_0^1 X^*(\omega)\phi'(\omega) \, d\omega. \quad (3.71)$$

We denote the Lorentz space as $\Lambda_\phi := \{X \in \mathcal{M} : \|X\|_{\Lambda_\phi} < \infty\}$.

The Lorentz norm is fully specified by its concave fundamental function, which indeed is ϕ . It has an intuitive interpretation: the derivative of the fundamental function is used to reweight the decreasing rearrangement. Since ϕ is concave, ϕ' is decreasing and thus more weight is put on worse outcomes. Furthermore, if ϕ is not continuous at the origin, i.e. $\phi(0+) > 0$, then the essential supremum $X^*(0)$ receives the fixed weight $\phi(0+)$. We remark that in finance and insurance, the Lorentz norm appears as the norm induced by a

spectral risk measure [Acerbi, 2002], also called a *distortion risk measure* [Wang, 2000], which plays a major role in financial risk assessment.

In fact, the Marcinkiewicz norm $\|\cdot\|_{M_{\phi_{\mathcal{X}'}}}$, where $\phi_{\mathcal{X}'}(t) = t/\phi(t)$, is the associate norm to the Lorentz norm Λ_ϕ . The converse associate relation also holds, but with a technical subtlety regarding (quasi)concavity (see Section 3.5.3).

Example 154. $\mathcal{L}^1$ is a Lorentz space with $\phi_{\mathcal{L}^1}(t) = t$. Diametrically opposed, $\mathcal{L}^\infty$ is also a Lorentz space with $\phi_{\mathcal{L}^\infty}(t) = \chi_{(0,1]}$. In the case of $\mathcal{L}^1$ and $\mathcal{L}^\infty$, the Marcinkiewicz, Orlicz and Lorentz norm are in fact identical [Fröhlich and Williamson, 2024a, Section 4]. In contrast, for instance, $\mathcal{L}^2$ is *not* a Lorentz space, but an Orlicz space.

Given a concave fundamental function, the Marcinkiewicz norm yields the smallest (most optimistic) and the Lorentz norm the largest (most pessimistic) risk assessment.

Theorem 155 (Bennett and Sharpley [1988, p. 72]). *Let $\mathcal{X}$ be an r.i. space with concave fundamental function ϕ . Then the following embedding holds:*

$$\Lambda_\phi \subseteq \mathcal{X} \subseteq M_\phi \quad (3.72)$$

and the norms are in the relationship:

$$\|X\|_{M_\phi} \leq \|X\|_{\mathcal{X}} \leq \|X\|_{\Lambda_\phi} \quad \forall X \in \mathcal{M}. \quad (3.73)$$

We offer the following intuition for the result: both the Marcinkiewicz and the Lorentz norm are in a special, parsimonious one-to-one correspondence to the fundamental function ϕ . Let ϕ be concave and continuous at the origin. Then the Lorentz norm is defined using only the single dual variable ϕ' , which controls the reweighting (and it is the *only* r.i. norm for which a single dual variable suffices). As to the Marcinkiewicz norm, we have the extremal assessment $\|\phi'\|_{M_{\phi_{\mathcal{X}'}}} = 1$. In this sense, the Marcinkiewicz norm is exactly tuned to the associate fundamental function. Thus, the Marcinkiewicz and the Lorentz norm demarcate the space of r.i. norms, which are compatible with a specified fundamental function.

We build intuition for the use of the Marcinkiewicz and Lorentz norms with the following example of the spaces $L_{\exp}$ and $L_{\log L}$. In this example we consider the class of *subexponential* random variables and its “dual” tail class, the random variables with finite differential entropy.

Example 156 (Bennett and Sharpley [1988, pp. 243-254]). $L_{\exp}$, $L_{\log L}$: Consider the envelope function that expresses subexponential tails. The decreasing rearrangement of an exponential distribution, which is shifted by 1, is $Y^*(t) = 1 - \log(t)$; the reason for the shift will become clear later. We take this to be the envelope function $E_{\mathcal{X}}(t) = 1 - \log(t)$

and define the Marcinkiewicz quasi-norm:

$$|X|_{L_{\exp}} := \sup_{0 < t \leq 1} \frac{1}{1 - \log(t)} X^*(t). \quad (3.74)$$

Note that $\phi(t) = 1/(1 - \log(t))$ is quasiconcave. In this case, it is easy to verify that condition M1 in Proposition 152 holds, and therefore we may equivalently consider the Marcinkiewicz norm:

$$||X||_{L_{\exp}} := \sup_{0 < t \leq 1} \frac{1}{1 - \log(t)} X^{**}(t). \quad (3.75)$$

Observe that the CVar's of an exponential distribution (not shifted) are in fact

$$\text{CVar}_{\alpha}(|X|) = 1 - \log(1 - \alpha) \quad \forall \alpha \in [0, 1). \quad (3.76)$$

We call a random variable *subexponential* if $||X||_{L_{\exp}} < \infty$ and denote the space $L_{\exp}$. It turns out, as we observe in the next section, that an equivalent approach to capture this tail class is via an Orlicz norm, which then yields the following definition: a random variable X is subexponential if $\exists \lambda > 0$ such that:

$$\int_{\Omega} \exp(\lambda |X(\omega)|) d\omega = \mathbb{E}[\exp(\lambda |X|)] < \infty, \quad (3.77)$$

that is, if the moment-generating function of $|X|$ is finite for some $\lambda > 0$. This is a typical textbook definition of subexponential tails [Vershynin, 2020]. In this case, we have equivalence of the two definitions of $L_{\exp}$. We will later observe that this is due to the fact that the Marcinkiewicz space here coincides with the Orlicz space with the same fundamental function.

The space $L_{\exp}$ is in fact the associate space of the Lorentz space $L \log L$, defined by:

$$||X||_{L \log L} := \int_0^1 X^*(\omega) d\phi_{L \log L}(\omega), \quad \phi_{L \log L}(t) := t + t \log(1/t). \quad (3.78)$$

This is a distinguished space: the space $L \log L$ consists of those random variables whose CVar's are integrable:

$$||X||_{L \log L} < \infty \iff \int_0^1 \text{CVar}_{\alpha}(|X|) d\alpha < \infty, \quad (3.79)$$

which is a very weak condition, hence the space is very “large”. In fact, we have the remarkable embedding situation:

$$\mathcal{L}^{\infty} \subsetneq L_{\exp} \subsetneq \mathcal{L}^p \subsetneq L \log L \subsetneq \mathcal{L}^1, \quad \forall p \in (1, \infty). \quad (3.80)$$

Thus, L_{exp} is smaller than any $\mathcal{L}^p$ space, whereas its associate LlogL is larger than any $\mathcal{L}^p$ space, $p \in (1, \infty)$. This behaviour can be intuited from the fundamental functions as $x \downarrow 0$: the fundamental function of LlogL is very moderate, while that of L_{exp} is extremely steep near the origin (see Appendix 3.12.2).

Furthermore, LlogL consists of those random variables with finite differential entropy¹⁰:

$$\|X\|_{\text{LlogL}} < \infty \iff \int_{\Omega} |X(\omega)| \log(|X(\omega)|) \, dP(\omega) < \infty, \quad (3.81)$$

which, if $X = \frac{dQ}{dP}$ is a legitimate density, is just the KL divergence of Q to the base measure P . Hence we observe that the KL divergence risk measure is finite on L_{exp} (a very “small” space) and the feasible dual variables are from LlogL (a very “large” space). This means that almost arbitrary tail reweightings are allowed, which entails an extremely tail-sensitive behaviour of the KL divergence risk measure.

We remark that, confusingly, according to some authors a subexponential random variable is one which has *heavier* tails than an exponential random variable. Such terminology appears counter-intuitive to us and conflicts with the definition of a sub-Gaussian random variable.

3.5.3 QUASICONCAVITY AND CONCAVITY

The question remains whether any random variable Y is a valid reference distribution as an envelope function, with the identification $\phi(t) = 1/Y^*(t)$ for $0 < t \leq 1$. We always set $\phi(0) = 0$. Then ϕ is continuous at 0 if and only if $\lim_{t \downarrow 0} Y^*(t)$ is unbounded. Throughout, we shall only consider such Y^* (with the instructive exception of Example 163). The critical issue is that ϕ must be quasiconcave. For this, we need that ϕ is nonnegative and increasing, which is satisfied for any Y , as Y^* is nonnegative and decreasing. Also, we require that $t \mapsto \phi(t)/t$ be decreasing:

$$\frac{\phi(t)}{t} = \frac{1}{Y^*(t) \cdot t} = \frac{1}{t/\phi(t)} = \frac{1}{\phi_{\mathcal{X}'}(t)}, \quad (3.82)$$

which is decreasing if and only if $t \mapsto Y^*(t) \cdot t$ is increasing. Hence, any nonzero Y which ensures that $t \mapsto Y^*(t) \cdot t$ is increasing yields a legitimate envelope function. Observe that since $\phi_{\mathcal{X}'}(t) = Y^*(t) \cdot t$, this is equivalent to requiring that $\phi_{\mathcal{X}'}$ be quasiconcave (note that $t \mapsto \phi_{\mathcal{X}'}(t)/t = Y^*(t)$ is decreasing). Therefore it is also equivalent to requiring that the envelope function of the associate space $\mathcal{X}'$ is decreasing, which of course is necessary for any legitimate r.i. norm. We make an observation, which relates to the embedding

¹⁰Observe the content in [Bennett and Sharpley, 1988, p. 243-245] and combine it with the equivalence of the Orlicz functions $(x \log(x))\chi_{[1, \infty)}$ and $x \log(x) - (x - 1)$.

theorem.

Proposition 157. *Assume Y^* is differentiable on $(0, 1)$. Then for $t \mapsto \phi_{\mathcal{X}'}(t) = Y^*(t) \cdot t$ to be increasing it is necessary that $Y^*(t) \leq Y^*(1) \cdot \frac{1}{t}$.*

Proof The proof is in Appendix 3.12.3. ■

Recall that $\mathcal{L}^1$ is the largest of all r.i. spaces and has envelope function $E_{\mathcal{L}^1}(t) = \frac{1}{t}$. Since the constraint $Y^*(1) = 1$ (equivalently, $\phi(1) = 1$) can be imposed without loss of generality, we see that unsurprisingly, the envelope function of $\mathcal{L}^1$ is the largest of all legitimate envelope functions.

Working with the maximal function is even simpler: by setting $\phi_{\text{CVar}}(t) = 1/Y^{**}(t)$ for any nonnegative and nonzero Y , the resulting ϕ is guaranteed to be quasiconcave, since $t \mapsto Y^{**}(t) \cdot t = \int_0^t Y^*(\omega) \, d\omega$ is increasing.

Example 158. *Taking the exponential distribution, $Y^*(t) = -\log(t)$, fails since $t \mapsto -\log(t) \cdot t$ is not increasing on $(0, 1]$; it violates the condition in Proposition 157. However, when shifting the distribution to have minimum value 1, i.e. $Y^*(t) = 1 - \log(t)$, we get $t \mapsto (1 - \log(t)) \cdot t$, which is increasing on $(0, 1]$. Of course, such a shift is insignificant for tail behaviour.*

With regard to the quasiconcavity vs. concavity distinction, we remark that it is not of essential importance. For instance, the Lorentz norm requires a concave ϕ in its definition. In this case, one can use the least concave majorant $\tilde{\phi}$ of ϕ , since a quasiconcave function ϕ is always equivalent to its least concave majorant $\tilde{\phi}$: $\frac{1}{2}\tilde{\phi} \leq \phi \leq \tilde{\phi}$ [Rubshtein et al., 2016, p. 131]. In fact, any r.i. space with a quasiconcave fundamental function can be equivalently renormed to have its least concave majorant as the fundamental function [Bennett and Sharpley, 1988, p. 71], [Pick et al., 2013, p. 268], so the quasiconcave–concave distinction is not essential.

Finally, we note that even if ϕ is concave, it may be that $\phi_{\mathcal{X}'}$ is only quasi-concave. The associate relationship of the Marcinkiewicz and Lorentz space is preserved using the least concave majorant then [Rubshtein et al., 2016, p. 147].

3.6 TAIL-SPECIFIC DIVERGENCE RISK MEASURES

In the previous section, we have constructed Marcinkiewicz and Lorentz norms, which capture a desired tail behaviour. In this section, we proceed with the construction of tail-specific Orlicz norms, which yield corresponding divergence risk measures. We focus on the construction of Orlicz norms for a specific tail sensitivity; from this, equivalent Orlicz regret and divergence risk measures can easily be derived (see Example 146 and recall Corollary 120).

Remark 159 (Rubshtein et al. [2016, p. 162, p. 209]). Let ϕ and $\tilde{\phi}$ be equivalent concave fundamental functions. Then Λ_ϕ and $\Lambda_{\tilde{\phi}}$ are equivalent; M_ϕ and $M_{\tilde{\phi}}$ are equivalent; and the Orlicz norms with fundamental functions ϕ and $\tilde{\phi}$ are equivalent.

We follow the same logic as in the previous section, beginning with the specification of a fundamental function or, via its one-to-one correspondence, of an envelope function. In this section, we denote by Φ and Ψ a conjugate pair of Young functions, which need not be divergence functions. Intuitively, they play the role of f and g , respectively. The corresponding fundamental functions of the Orlicz spaces are $\phi_\Phi^L(t)$ and $\phi_\Psi(t) = t/\phi_\Phi(t)$, where we use the L to emphasize that it corresponds to the Luxemburg norm in $\mathcal{L}^\Phi$. Then:

$$\|X\|_\Psi^O = \sup \left\{ \int_0^1 X^*(\omega) Z^*(\omega) d\omega : Z \in \mathcal{L}^\Phi, Z \geq 0, \mathbb{E}[\Phi(Z)] \leq 1 \right\} \quad \forall X \in \mathcal{L}^\Psi. \quad (3.83)$$

In fact, the Orlicz norm is the associate norm to the Luxemburg norm, that is:

$$\|X\|_\Psi^O = \sup \left\{ \int_0^1 X^*(\omega) Z^*(\omega) d\omega : Z \in \mathcal{L}^\Phi, Z \geq 0, \|Z\|_\Phi^L \leq 1 \right\} \quad \forall X \in \mathcal{L}^\Psi. \quad (3.84)$$

To achieve a desired tail sensitivity, we aim to constrain the tail behaviour of dual variables Z^* in an Orlicz norm (a divergence risk measure, respectively). For example, for the LlogL divergence risk measure, we allow dual variables to be from L_{exp} . Conversely, for the L_{exp} (KL) divergence risk measure, we allow dual variables from LlogL, which merely need to have finite differential entropy. We denote the envelope function in $\mathcal{L}^\Phi$ of the Luxemburg norm, indexed by the Young function Φ , as:

$$E_\Phi^L(t) := \sup \{ Z^*(t) : \|Z\|_\Phi^L \leq 1 \}. \quad (3.85)$$

By specifying this envelope function based on the decreasing rearrangement of some reference distribution Y^* , setting $E_\Phi^L = Y^*$, we have the least upper bound on the tail behaviour of the dual variables in Equation 3.84 in terms of Y^* and thereby delimit possible quantile reweightings. The more extreme our envelope function, the higher the reweighting can be. We assume that the envelope function is finite for all $t \in (0, 1]$ and unbounded for $t = 0$ (otherwise there would be no tails).

Proposition 160. Choose $E_\Phi^L : (0, 1] \rightarrow \mathbb{R}^+$ as the decreasing rearrangement of some reference distribution, for which $E_\Phi^L(0+) = \infty$, $\phi_\Psi(t) := E_\Phi^L(t) \cdot t$ is increasing and concave on $(0, 1]$ and which satisfies $\phi_\Psi(0+) = 0$. We set $\phi_\Psi(0) = 0$. Then there exists an Orlicz norm $\|\cdot\|_\Psi^O$ which has fundamental function ϕ_Ψ . The conjugate Young function of Ψ , which is Φ , induces $\mathcal{L}^\Phi$ with envelope function E_Φ^L . Furthermore, Φ is finite and supercoercive and therefore Ψ is finite.

Proof Let ϕ_Ψ be the fundamental function of $\|\cdot\|_\Psi^O$, that is, $\phi_\Psi(t) = t\Phi^{-1}(1/t)$. Denote by $\phi_\Phi^L(t) = t/\phi_\Psi(t)$ the fundamental function of the Luxemburg norm with Young function

Φ . We know that

$$E_{\Phi}^L(t) = \frac{1}{\phi_{\Phi}^L(t)} = \frac{1}{t/\phi_{\Psi}(t)} = \frac{\phi_{\Psi}(t)}{t}, \quad (3.86)$$

by Proposition 147 and Proposition 149; recall that the Orlicz norm $\|\cdot\|_{\Psi}^{\mathcal{O}}$ is in an associate relationship to $\|\cdot\|_{\Phi}^L$. Therefore:

$$\phi_{\Psi}(t) = tE_{\Phi}^L(t) = t\Phi^{-1}(1/t), \quad \forall t \in (0, 1] \quad (3.87)$$

$$\Leftrightarrow E_{\Phi}^L(1/x) = \Phi^{-1}(x), \quad \forall x \in [1, \infty). \quad (3.88)$$

By specifying a legitimate $E_{\Phi}^L(t)$, we thus obtain $\Phi^{-1}(x)$ for $x \geq \Phi^{-1}(1)$. Of course, the behaviour of Φ on $0 \leq x < \Phi^{-1}(1)$ is immaterial for tail sensitivity. As we have seen before, Orlicz norms have *concave* fundamental functions (Proposition 133), hence we have to ensure that $\phi_{\Psi}(t) = E_{\Phi}^L(t) \cdot t$ is a valid concave fundamental function. Given that we specified some E_{Φ}^L which yields a concave $\phi(t) = E_{\Phi}^L(t) \cdot t$ and thereby specifies the behaviour of f on $x \geq \Phi^{-1}(1)$, can we always find a compatible Φ , so that Φ is indeed a legitimate Young function? The answer is affirmative and a constructive approach is in the proof of Proposition 141. For the supercoercivity of Φ , see the proof of FF1 in Proposition 134. For the finiteness of Φ , observe that $E_{\Phi}^L(0+) = \infty \implies \phi'_{\Psi}(0) = \infty$ and then see FF2 in Proposition 134. ■

Of course, the primal approach is also feasible, where the envelope of the Orlicz norm in $\mathcal{L}^{\Psi}$ is specified as $E_{\Psi}^{\mathcal{O}}(t) := \sup\{Z^*(t) : \|Z\|_{\Psi}^{\mathcal{O}} \leq 1\} = 1/\phi_{\Psi}(t)$, hence we then demand that $\phi_{\Psi}(t)$ is concave and $\phi_{\Psi}(0+) = 0$. If we at least have a quasiconcave ϕ , that is, if $t \mapsto \phi_{\Psi}(t) := E_{\Phi}^L(t) \cdot t$ is increasing, we may take its equivalent least concave majorant (see Section 3.5.3). Thus, to achieve a tail sensitivity we merely need that $t \mapsto \phi_{\Psi}(t) := E_{\Phi}^L(t) \cdot t$ is increasing, since $t \mapsto \phi_{\Psi}(t)/t = E_{\Phi}^L(t) = Y^*(t)$ is decreasing for any Y .

We have seen that specifying a least upper bound on the tail behaviour of primal or dual variables yields a corresponding Orlicz space. Assume that we have therefore obtained a Young function Φ and an equivalent Young divergence function $f \in \mathfrak{F}$ via Example 146. To further motivate our approach, observe that if $Z \notin \mathcal{L}^{\Phi}$, no finite ε will suffice for Z to satisfy the f -divergence constraint in the envelope representation of the divergence risk measure. The converse direction is more subtle due to the Δ_2 condition (see Definition 181). Thus, our approach is about the *structure* of the divergence ball, essentially given by the tail constraint, not its particular size, which of course is also significant in practical implementations.

Example 161. $\mathcal{L}^{\Psi} = \text{LlogL}$: we choose the dual envelope $E_{\Phi}^L(t)$ to be the decreasing rearrangement of a random variable with exponential distribution, shifted by 1, that is, $E_{\Phi}^L(t) = 1 - \log(t)$. As desired, $t \mapsto \phi_{\Psi}(t) = E_{\Phi}^L(t) \cdot t$ is increasing and concave and $\phi_{\Psi}(0+) = 0$. Then $E_{\Phi}^L(1/x) = 1 - \log(1/x) = \Phi^{-1}(x)$, $x \geq 1$. This is satisfied by the

Young function

$$\Phi(x) = \begin{cases} x & 0 \leq x < 1 \\ \exp(x-1) & x \geq 1, \end{cases} \quad (3.89)$$

which is in fact equivalent (in the sense of Section 3.4.3) to the function $\Phi(x) = \exp(x) - 1$. Therefore we obtain the LlogL Orlicz norm as $\|\cdot\|_{\Psi}^O$.

Example 162. $\mathcal{L}^{\Psi} = \mathcal{L}_{\exp}$: assume that we want to construct an Orlicz norm where the primal variables X are constrained to be subexponential. This can simply be obtained from the previous example by duality. To provide more intuition, we derive it in another way. When choosing the primal envelope function as $\phi_{\Psi}(t) = 1/E_{\Psi}^O(t) = 1/(1 - \log(t))$ based on a shifted exponential distribution, we face the problem that this is only quasiconcave. However, we have the equivalence to the concave function $t \mapsto \frac{1}{\log(1/t+1)}$:

$$\frac{1}{1 - \log(t)} \leq \frac{1}{\log(1/t+1)} \leq 2 \cdot \frac{1}{1 - \log(t)}, \quad \forall t \in (0, 1]. \quad (3.90)$$

Setting $1/E_{\Psi}^O(t) = \phi_{\Psi}(t) = t\Phi^{-1}(1/t)$ yields an unwieldy form for f involving the Lambert W function instead of $\Phi(x) = x \log(x) - (x-1)$, albeit equivalent. Instead, we consider the Luxemburg norm $\|\cdot\|_{\Psi}^L$ with envelope function $E_{\Psi}^L(t) = 1/\phi_{\Psi}^L(t)$ and choose $E_{\Psi}^L(t) = \log(1/t+1)$, where the fundamental function is [Rubshtein et al., 2016, p. 177]:

$$\phi_{\Psi}^L(t) = \frac{1}{\Psi^{-1}(1/t)} \implies \Psi^{-1}(1/t) = \log(1/t+1) = E_{\Psi}^L(t), \quad \forall t \in (0, 1]. \quad (3.91)$$

Thus we find that $\Psi(y) = \exp(y) - 1$ satisfies the equation. Recall that the Luxemburg and Orlicz norms with the same Young function are equivalent. It follows:

$$\|X\|_{\Psi}^O < \infty \iff \exists \lambda > 0 : \mathbb{E}[\exp(\lambda|X|)] - 1 < \infty, \quad (3.92)$$

which is a standard definition of a subexponential random variable [Vershynin, 2020], cf. Equation 3.77.

Example 163. $\mathcal{L}^{\Psi} = \mathcal{L}^1$; Expectation and CVar_{α} : in contrast to the assumption of Proposition 160 that $E_{\Phi}^L(0+) = \infty$, assume we want to allow only bounded quantile reweightings, i.e. dual variables from $\mathcal{L}^{\infty}$. For instance, let $E_{\Phi}^L(t) = 1, t \in (0, 1]$, so that all quantiles of the dual variables are bounded from above by 1. Therefore $1 = \Phi^{-1}(1/t), \forall t \in (0, 1]$, whence it follows up to equivalence that

$$\Phi(x) := \begin{cases} 0 & 0 \leq x \leq 1 \\ \infty & x > 1, \end{cases} \quad (3.93)$$

which is also a divergence function. Let $g(x) := \Psi(x)$ for $x \geq 0$ and $g(x) := 0$ for $x < 0$, which is the convex conjugate of Φ . In fact, $g(x) = \max(0, x)$. Then $\|X\|_{V_g} = \mathbb{E}[|X|]$ and $V_g(X) = \mathbb{E}[X^+]$. This is nothing more but the $\mathcal{L}^1 - \mathcal{L}^\infty$ associate relation: the function Φ is the Young function whose Orlicz space is $\mathcal{L}^\infty$, thus its associate space is $\mathcal{L}^1$.

Similarly, when allowing reweightings up to $1/(1 - \alpha)$, i.e. $E_\Phi^L(t) = 1/(1 - \alpha)$, we have $V_g(X) = \frac{1}{1-\alpha}\mathbb{E}[X^+]$ and $R_g(X) = \text{CVar}_\alpha(X)$ via the infimal convolution in Section 3.3.5. This does not change the primal space, which is still $\mathcal{L}^1$, whose associate space is $\mathcal{L}^\infty$. The situation is clarified by the envelope representation

$$\text{CVar}_\alpha(X) = \sup \left\{ \int_0^1 X^*(\omega) Z^*(\omega) d\omega : \mathbb{E}[Z] = 1, 0 \leq Z \leq 1/(1 - \alpha) \right\}. \quad (3.94)$$

Example 164. Pareto, χ^2 : The decreasing rearrangement of a Pareto distribution with shape parameter p (typically denoted α) is $Y^*(t) = t^{-\frac{1}{p}}$. Set the dual envelope function $E_\Phi^L(t) = t^{-\frac{1}{p}}$. Then $\Phi(x) = x^p$. For instance, $p = 2$ yields the χ^2 divergence risk measure (up to equivalence, see Example 144). Of course, $L^\Phi = \mathcal{L}^p$ and L^Ψ is then just $\mathcal{L}^q$, where q is the dual exponent. Since 2 is self dual, that is, $1/2 + 1/2 = 1$, we have that the χ^2 risk measure is finite on $\mathcal{L}^2$. The same procedure naturally works for any $1 < p \leq \infty$, as $t \mapsto t^{-\frac{1}{p}} \cdot t$ is increasing, giving the space $\mathcal{L}^p$.

An interesting case is $p = 1$: Then $E_\Phi^L(t) = 1/t$, but $t \mapsto E_\Phi^L(t) \cdot t$ is *not* continuous at 0. Then $\Phi^{-1}(x) = x$ and $\Phi(x) = x$, which is not supercoercive. In fact, $\Phi(x) = x$ is the Young function whose Orlicz norm is the expectation $\mathbb{E}[|\cdot|]$. By duality, we would obtain the essential supremum as $\|\cdot\|_\Psi^O$, where

$$\Psi(x) = \begin{cases} 0 & 0 \leq x \leq 1 \\ \infty & x > 1. \end{cases} \quad (3.95)$$

Due to involved technical complications, however, we have excluded such norms (risk measures, respectively) from our framework and assumed supercoercivity. In contrast, as $p \rightarrow \infty$, we obtain $E_\Phi^L(t) = 1$ and we recover the expectation ($\mathcal{L}^\Psi = \mathcal{L}^1$) as in the previous example 163.

A curious phenomenon can be observed: when constraining the quantiles of dual variables by a Pareto distribution with $p = 2$, we obtain the space $\mathcal{L}^2$. However, the Pareto distribution itself only has a finite variance when $p > 2$. Similarly, $p = 3$ yields the $\mathcal{L}^3$ space, but the Pareto distribution only has a Kurtosis for $p > 3$ (and so on). This means that the random variable Y whose decreasing rearrangement we used to construct the Orlicz space via the envelope function $E_\Phi^L = Y^*$ is itself *not* in the space $\mathcal{L}^\Phi$, although it is the least upper bound on the decreasing rearrangements $\{t \mapsto Z^*(t) : \|Z\|_\Phi^L \leq 1\}$ (recall Definition 147). Thus, the envelope function is a natural limit object. For example, we might say informally

that the “limit” of those distributions which have a finite variance is the Pareto $p = 2$ distribution. An interesting special case is the pathological $p = 1$: the Pareto distribution only has an $\mathcal{L}^1$ norm (an expectation) for $p > 1$. In this way, the extremely heavy-tailed Pareto $p = 1$ distribution with envelope function $t \mapsto 1/t$ is as extreme as it gets, as the corresponding Orlicz space is the largest of all r.i. spaces (recall Proposition 125).

The question of when the reference distribution is itself in the space, i.e. $E_\Phi^L = Y^* \in \mathcal{L}^\Phi$ has a surprisingly simple and insightful answer, which is related to M3 in Proposition 152.

Proposition 165. *Let ϕ_Ψ (resp. ϕ_Φ) be the fundamental function of $\mathcal{L}^\Psi$ (resp. $\mathcal{L}^\Phi$) and $\phi_\Psi(t)/t = E_\Phi^L$. Then*

$$\|t \mapsto \phi_\Psi(t)/t\|_\Phi^L = \|E_\Phi^L\|_\Phi^L < \infty \text{ if and only if } M_{\phi_\Phi} = \mathcal{L}^\Phi. \quad (3.96)$$

Proof

$$\|t \mapsto \phi_\Psi(t)/t\|_\Phi^L = \|t \mapsto \Phi^{-1}(1/t)\|_\Phi^L = \inf \left\{ \lambda > 0 : \mathbb{E}\Phi \left[t \mapsto \frac{\Phi^{-1}(1/t)}{\lambda} \right] \leq 1 \right\}. \quad (3.97)$$

From [Pick et al. \[2013, p. 412\]](#) we have that if the last expression is finite, then $M_{\phi_\Phi} = \mathcal{L}^\Phi$. Conversely, assume $M_{\phi_\Phi} = \mathcal{L}^\Phi$. Then we obtain from M3 in Proposition 152 together with the embedding theorem 155 that $E_\Phi^L = Y^* \in \mathcal{L}^\Phi$. ■

To put this into words, the random variable whose decreasing rearrangement (backwards quantiles) we used to specify the envelope function E_Φ^L is itself in the space $\mathcal{L}^\Phi$ if and only if $\mathcal{L}^\Phi$ coincides with its corresponding Marcinkiewicz space, that is, with the shared fundamental function ϕ_Φ . On the other hand, we know from Proposition 152 that $t \mapsto E_\Phi^L = \phi_\Psi(t)/t \in M_{\phi_\Phi}$ if and only if the envelope approach based on quantiles is equivalent to that based on CVar's; formally, if the Marcinkiewicz quasi-norm $|\cdot|_{M_{\phi_\Phi}}$ and the Marcinkiewicz norm $\|\cdot\|_{M_{\phi_\Phi}}$ coincide (we use the “dual” of Proposition 152). Due to the embedding theorem, $t \mapsto \phi_\Psi(t)/t \in M_{\phi_\Phi}$ is a necessary, but not sufficient condition for $t \mapsto \phi_\Psi(t)/t \in \mathcal{L}^\Phi$. We clarify this with some examples.

Example 166. $\mathcal{L}^\Phi = \text{L}_{\exp} : \Phi(y) = \exp(y) - 1, \Phi^{-1}(x) = \log(x + 1)$. Then:

$$\exists \lambda > 0 : \|t \mapsto \phi_\Psi(t)/t\|_\Phi^L \leq \int_0^1 \exp(\lambda \log(1/x + 1)) - 1 < \infty \quad (3.98)$$

is finite, for instance with $\lambda = 1/e$. Therefore $E_\Phi^L \in \text{L}_{\exp}$, that is, an exponential distribution is itself in the space of subexponential random variables.

Example 167. $\mathcal{L}^\Phi = \text{LlogL} : \Phi(x) = (x \log(x) - x + 1) \cdot \chi_{[1, \infty)}$. We have that $\phi_\Psi(t) =$

$1/(1 - \log(t))$ up to equivalence, since $L^\Psi = L_{\exp}$. Then condition M1 is (here $\phi = \phi_\Psi$):

$$\forall s > 1 : \sup_{0 < t \leq 1} \frac{1 - \log(t)}{1 - \log(t/s)} < 1. \quad (3.99)$$

But for any $s > 1$, we obtain that $\lim_{t \downarrow 0} \frac{1 - \log(t)}{1 - \log(t/s)} = 1$. Hence, $t \mapsto \phi_\Psi(t)/t \notin \text{LlogL}$. Consequently, LlogL does not coincide with the Marcinkiewicz space M_{ϕ_Ψ} and moreover the envelope function E_Φ^L grows so rapidly as $t \downarrow 0$ that the Marcinkiewicz quasi-norm does not coincide with the Marcinkiewicz norm.

Example 168. $L^\Phi = \mathcal{L}^p$: let $1 < p < \infty$ and $\Phi(x) = x^p$, as in the Pareto example 164. Then $\|t \mapsto \phi_\Psi(t)/t\|_\Phi^L = \infty$ is easily checked by using Equation 3.97. Thus, for instance the χ^2 divergence risk measure is *not* finite on the Marcinkiewicz space with fundamental function $\phi(t) = \sqrt{t}$ (here $L^\Phi = L^\Psi = \mathcal{L}^2$). However, we have that $\|t \mapsto \phi(t)/t\|_{M_{\phi_\Phi}} < \infty$, that is, the conditions of Proposition 152 apply. Therefore the Marcinkiewicz quasi-norm and the Marcinkiewicz norm are equivalent, but their induced space is strictly larger than the Orlicz space (the $\mathcal{L}^p$ space). In fact, this larger space is known as the “*weak $\mathcal{L}^p$ space*”, which is used in the theory of interpolation of operators. Thus, we find that the Pareto distribution of parameter p , even if it does not have a finite $\mathcal{L}^p$ norm, is the natural representative and member of the weak $\mathcal{L}^p$ space, as it is the least upper bound on the quantiles of random variables in $\mathcal{L}^p$. For $p = 1$ and $p \rightarrow \infty$, the Marcinkiewicz and the Orlicz space coincide¹¹.

In summary, in this section we have proposed a way to construct an Orlicz norm with a desired tail sensitivity based on the specification of the least upper bound on the quantile behaviour of random variables in the space. Due to the equivalence relation in Example 146, this approach can then be used to construct a corresponding divergence risk measure, where the tail sensitivity is finely tuned to the reference distribution, which is the envelope function.

3.6.1 UTILITY-BASED SHORTFALL RISK

Until now, we have always considered *coherent* risk measures to encode a specific tail sensitivity. However, there is another related class, the *convex* risk measures which have gained much attention recently in financial risk measurement. A potential problem is that due to positive homogeneity, $R(\lambda X) = \lambda R(X)$, the risk assessment scales just like the loss variable itself. This might be undesirable due to hazards arising from very large risks. Thus it may be desirable to construct risk measures which are not positively homogeneous, so that the risk assessment can grow faster in the presence of large risks. A convex risk measure,

¹¹This is due to the special situation that for $\mathcal{L}^1$ and $\mathcal{L}^\infty$ all r.i. norms are equivalent.

which has received much attention in finance particularly as a tool to measure heavy-tailed risk [Giesecke et al., 2008], is the *utility-based shortfall risk* (UBSR) [Föllmer and Schied, 2002]. In the context of machine learning it has not seem widespread use yet; but, for example, Shen et al. [2014] apply it as a tool for risk-sensitive reinforcement learning. Furthermore, Li et al. [2021] apply a special case of the UBSR¹² to practical machine learning problems and show that it can flexibly tune the impact of individual losses. In Appendix 3.13, we discuss the UBSR, which has been linked to Orlicz norms. We make this connection yet more explicit; consequently, the insights from the previous section can then be directly applied to construct a utility-based shortfall risk measure with a specific tail sensitivity. For instance, the choice in Li et al. [2021] corresponds to a subexponential tail sensitivity.

3.7 TAIL-SPECIFIC ORLICZ DEVIATION INEQUALITIES

Few papers in the machine learning literature have made explicit use of Orlicz space theory; see for instance [Andoni et al., 2018], [Song et al., 2019] and [Chamakh et al., 2020]. In contrast, *concentration inequalities* are widely used in machine learning in the context of statistical learning. Many concentration inequalities are in fact merely special cases of a general *Orlicz deviation inequality*; the standard textbook Boucheron et al. [2013, p. 45] on concentration inequalities only mentions Orlicz spaces in passing. We refer to [van der Vaart and Wellner, 1996, Part 2.2] for the general approach; concrete cases of tail sensitivity which have been studied in the literature are the sub-Gaussian and the sub-Exponential norm (see e.g. [Vershynin, 2020]), the sub-Weibull norm [Vladimirova et al., 2020], the Bernstein-Orlicz norm [van de Geer and Lederer, 2013], the Bennett-Orlicz norm [Wellner, 2017] and more [Chamakh et al., 2021]. Consequently, all of these give rise to equivalent divergence risk measures.

We briefly present the Orlicz deviation inequality, since we believe that this general view-point may be helpful to the machine learning community. Furthermore, we show how it naturally fits within our approach to tail sensitivity based on the envelope function (hence, the fundamental function).

Assume a finite Young function Ψ and $X \geq 0$. Then, applying Markov's inequality to the nonnegative and increasing function $x \mapsto \Psi(x/\|X\|_{\Psi}^L)$ yields [van der Vaart and Wellner,

¹²For positive values of their parameter t , the objective in [Li et al., 2021] is the convex entropic risk measure [Föllmer and Knispel, 2011] on the empirical distribution. Note that the convex entropic risk measure is a special case of the UBSR [Giesecke et al., 2008].

1996, Part 2.2]:

$$P(X \geq x) \leq P\left(\Psi(X/||X||_{\Psi}^L) \geq \Psi(x/||X||_{\Psi}^L)\right), \quad \forall x > 0 \quad (3.100)$$

$$\leq \frac{\mathbb{E}[\Psi(X/||X||_{\Psi}^L)]}{\Psi(x/||X||_{\Psi}^L)}, \quad \forall x > 0 \quad (3.101)$$

$$\leq \frac{1}{\Psi(x/||X||_{\Psi}^L)}, \quad \forall x > 0, \quad (3.102)$$

which follows from the definition of the Luxemburg norm. In summary, for a finite Young function Ψ we obtain:

$$P(X \geq x) \leq \frac{1}{\Psi(x/||X||_{\Psi}^L)}, \quad \forall x > 0. \quad (3.103)$$

For each finite Young function Ψ , we therefore obtain a deviation inequality. In fact, we observe that it bounds the tail behaviour by the distribution function of the envelope function. Recall that the distribution function is $\mu_Y(\lambda) := P(\{\omega \in \Omega : |Y(\omega)| > \lambda\})$ and that its generalized inverse is Y^* . For simplicity, we shall assume that μ_Y is continuous and invertible. Of course, $\mu_Y = \mu_{Y^*}$ holds. Since confusion could arise, we distinguish the fundamental function of the Orlicz norm ϕ_{Φ}^O from the fundamental function of the Luxemburg norm ϕ_{Φ}^L , where the subscript indicates the Young function.

Proposition 169 (Rubshtein et al. [2016, p. 177]). *The fundamental function of the Luxemburg norm $|| \cdot ||_{\Psi}^L$ is given by:*

$$\phi_{\Psi}^L(t) = \frac{1}{\Psi^{-1}(1/t)} = \frac{t}{\phi_{\Phi}^O(t)}, \quad \forall t \in (0, 1]. \quad (3.104)$$

For a given Luxemburg norm, we can directly relate the deviation inequality to the envelope function.

Proposition 170. *With the envelope function $Y^*(t) := E_{\Psi}^L(t) = 1/\phi_{\Psi}^L(t)$, where we assume $Y^*(1) = 1$ (without loss of generality) and finite Ψ , we obtain the deviation inequality:*

$$P(X \geq x) \leq \mu_{Y^*}(x/||X||_{\Psi}^L), \quad x \geq ||X||_{\Psi}^L, \quad X \geq 0. \quad (3.105)$$

Proof Just observe that

$$\phi_{\Psi}^L(t) = \frac{1}{\Psi^{-1}(1/t)} = \frac{1}{E_{\Psi}^L(t)} = \frac{1}{Y^*(t)} \implies \Psi(x) = \frac{1}{\mu_{Y^*}(x)}, \quad \forall x \geq 1, \quad (3.106)$$

and use Equation 3.103. ■

To intuit this result, recall that the envelope function constrains the quantile behaviour of

random variables in an Orlicz space. Thus, it is unsurprising that the corresponding distribution function plays an analogous role. Conversely, we offer the following constructive approach to obtain a deviation inequality with a desired tail sensitivity based on a reference distribution.

Proposition 171. Choose $E_\Psi^L : (0, 1] \rightarrow \mathbb{R}^+$ as the decreasing rearrangement of some reference distribution, where $E_\Psi^L(0+) = \infty$, $E_\Psi^L(1) = 1$, and for which $t \mapsto \phi_\Phi^O(t) = E_\Psi^L(t) \cdot t$ is increasing and concave on $(0, 1]$ and which satisfies $\phi_\Phi^O(0+) = 0$. We set $\phi_\Phi^O(0) = 0$. With the envelope function $Y^*(t) := E_\Psi^L(t) = 1/\phi_\Psi^L(t)$, the following Orlicz deviation inequality holds:

$$P(X \geq x) \leq \mu_{Y^*}(x/\|X\|_\Psi^L), \quad x \geq \|X\|_\Psi^L, X \geq 0. \quad (3.107)$$

Proof We apply the “dual” of Proposition 160 and obtain an Orlicz norm $\|\cdot\|_\Phi^O$ with fundamental function ϕ_Φ^O . Observe that by the associate relationship

$$\phi_\Phi^O(t) = \frac{t}{\phi_\Psi^L(t)} = E_\Psi^L(t) \cdot t. \quad (3.108)$$

The associate of the Orlicz norm $\|\cdot\|_\Phi^O$ is the Luxemburg norm $\|\cdot\|_\Psi^L$ with envelope E_Ψ^L . Also, we get from Proposition 160 that Ψ is a finite Young function and therefore the deviation inequality holds due to Proposition 170. ■

Again, if $\phi_\Phi^O(t) := E_\Psi^L(t) \cdot t$ is at least quasiconcave, for which it suffices that $\phi_\Phi^O(t) := E_\Psi^L(t) \cdot t$ is increasing (cf. Section 3.5.3), we may use the equivalent least concave majorant in the construction to obtain a deviation inequality, which is equivalent up to a constant multiplicative factor.

Example 172. The sub-Exponential deviation inequality: let $E_\Psi^L(t) = 1 - \log(t)$ (Example 156). Then:

$$P(X \geq x) \leq \exp(-x/\|X\|_\Psi^L + 1), \quad x \geq \|X\|_\Psi^L, X \geq 0. \quad (3.109)$$

where the Young function $\Psi(x) = x$ for $x \in (0, 1]$ and $\Psi(x) = \exp(x - 1)$ for $x \geq 1$ satisfies the equation $\Psi^{-1}(1/t) = E_\Psi^L(t)$ for $t \in (0, 1]$.

Example 173. The Pareto deviation inequality: let $E_\Psi^L(t) = t^{-1/p}$, $1 < p < \infty$. Then $\Psi(x) = x^p$ and:

$$P(X \geq x) \leq \frac{1}{(x/\|X\|_\Psi^L)^p}, \quad x \geq \|X\|_\Psi^L, X \geq 0. \quad (3.110)$$

For $p = 2$, this gives Chebyshev’s inequality.

3.8 APPLICATION

We have demonstrated how, given a specific tail-sensitivity in the form of a fundamental function, one can canonically construct three r.i. norms: the extremal Marcinkiewicz and Lorentz norms, and the Orlicz norm in between. In some cases, the Orlicz norm is equivalent to one or even both of them. This raises the question of whether one of these norms is superior. To give some practical advice, we consider corresponding risk measures and how to compute with them. These are then candidates to replace the expectation in expected risk minimization, in order to obtain a specified tail sensitivity.

The downside of the Marcinkiewicz norm for such practical use is that it lacks the property of translation equivariance (C4), even when restricted to nonnegative random variables and constants only. However, [Fröhlich and Williamson \[2024a\]](#) recently defined an equivalent norm, which is induced by a translation equivariant risk measure¹³:

Proposition 174. [[Fröhlich and Williamson, 2024a](#), Theorem 4.16] *Given any concave fundamental function ϕ , which is continuous at 0, the positive translation equivariant Marcinkiewicz norm is defined as:*

$$\|X\|_{TM_\phi} := \sup_{0 < t < 1} \frac{1 - \phi(1 - t)}{t} \cdot \mathbb{E}[|X|] + \left(1 - \frac{1 - \phi(1 - t)}{t}\right) \cdot \text{CVar}_t(|X|). \quad (3.111)$$

It holds that $\|\cdot\|_{M_\phi}$ and $\|\cdot\|_{TM_\phi}$ are equivalent. A corresponding coherent risk measure, which induces this norm, is:

$$TM_\phi(X) := \sup_{0 < t < 1} \frac{1 - \phi(1 - t)}{t} \cdot \mathbb{E}[X] + \left(1 - \frac{1 - \phi(1 - t)}{t}\right) \cdot \text{CVar}_t(X). \quad (3.112)$$

The $\|\cdot\|_{TM_\phi}$ norm is the smallest r.i. norm with fundamental function ϕ , which is induced by a coherent risk measure.

Proof We observe that this norm (risk measure, respectively) is a supremum over convex combinations of the expectation and CVar_α , where the weightings depend on the fundamental function to achieve the desired tail-sensitivity. The class of risk measures is closed under convex combinations and taking suprema, hence TM_ϕ is a legitimate coherent risk measure, which is obviously r.i. by definition. By taking absolute values, the norm $\|\cdot\|_{TM_\phi}$ is induced. For the remaining statements, see [Fröhlich and Williamson \[2024a, Theorem 4.16\]](#). ■

¹³An r.i. norm can *never* be translation equivariant due to absolute homogeneity. Therefore [Fröhlich and Williamson \[2024a\]](#) have defined the related property of positive translation equivariance (PTE) for a norm: a norm is called PTE if $\forall X \geq 0, c \in \mathbb{R}$ such that $X + c \geq 0$: $\|X + c\| = \|X\| + c$. Indeed, the norm in Proposition 174 turns out to be the smallest PTE r.i. norm with a given fundamental function [[Fröhlich and Williamson, 2024a](#), Theorem 4.16].

Thus, the TM_ϕ risk measure is distinguished in the sense that the induced norm is the smallest of all r.i. norms which are induced by risk measures of fundamental function ϕ , thereby providing the smallest (most optimistic) risk assessment.

A spectral risk measure, which induces a Lorentz norm, extends the fundamental function in the most cautious (pessimistic) possible way. The Lorentz norm $\|\cdot\|_{\Lambda_\phi}$ corresponds to the spectral risk measure R_ϕ :

$$R_\phi(X) := \int_0^1 F_X^{-1}(1 - \omega) d\phi(\omega). \quad (3.113)$$

Thus, having specified a desired tail sensitivity, using the coherent risk measures TM_ϕ and R_ϕ are the extremal choices and they hence provide a maximal “model uncertainty” regarding the choice of risk measure.

Example 175. [Fröhlich and Williamson, 2024a] Let $\phi(t) = 1 - (1 - t)^2$. Then consider

$$R_\phi(X) = \text{MaxVar}(X) = \mathbb{E}_{X_1, X_2 \sim X} [\max(X_1, X_2)] \quad (3.114)$$

$$TM_\phi(X) = \text{Dutch}(X) = \mathbb{E} [\max(X, \mathbb{E}[X])], \quad (3.115)$$

where $X_1, X_2 \sim X$ means that X_1 and X_1 are independent and identically distributed. These risk measures are known as MaxVar¹⁴ [Cherny and Madan, 2009] and the Dutch risk measure [Van Heerwaarden and Kaas, 1992].

The Orlicz norm (the Orlicz regret and risk measure, respectively) is situated in between these two ends of optimism and pessimism. However, we refrain from making a general recommendation as to which of the three candidates (or even another risk measure with the given fundamental function) to use in practice. We make some remarks regarding the choice.

- Spectral risk measures are embedded in a rich literature, from economics [Quiggin, 2012], insurance [Wang, 2000] to philosophy [Buchak, 2013], and can be motivated from various perspectives as a tool for rational decision making. However, the computation of spectral risk measures is non-trivial. We refer to [Pandey et al., 2021] for the estimation of spectral risk measures from samples and to [Holland and Haess, 2022], [Leqi et al., 2022a] and [Mehta et al., 2022] for learning schemes, which may be used in practice to replace ERM with spectral risk measures. Bäuerle and Glauner [2021] study the minimization of spectral risk measures in the context of Markov decision processes. Some spectral risk measures, like $\mathbb{E}$, CVar_α or convex combinations thereof are easily computed and optimized; a useful form is that of Example 119. To

¹⁴The name MaxVar is due to our loss-based orientation here; the authors originally called it MinVar due to a sign flip.

improve the stability of optimizing CVar_α in a batch setting, [Curi et al. \[2020\]](#) have proposed an adaptive sampling method; see also [\[Soma and Yoshida, 2020\]](#) and [\[Holand and Haress, 2021\]](#).

- The TM_ϕ risk measure, related to the Marcinkiewicz norm, is perhaps not as intuitive. However, it is easy to compute in a risk minimization loop. As of now, we see little motivation for employing it in practice.
- The f -divergence risk measure is situated in between, albeit in many cases equivalent to one of the above, as we observed in Section 3.6. Specifically, a highly tail-sensitive divergence risk measure is typically equivalent to the corresponding Marcinkiewicz norm. An advantage is that a divergence risk measure naturally comes with a corresponding deviation inequality, as we established in Section 3.7. Also, some prominent spectral risk measures like $\mathbb{E}$ or CVar_α are divergence risk measures. The infimum-based representation Equation 3.36 allows practical computation. For instance, when training a neural network, a single line suffices to implement Equation 3.36 in a framework such as PyTorch. We also refer to [Chouzenoux et al. \[2019\]](#) who propose a suitable accelerated projected gradient algorithm.

On the other hand, the utility-based shortfall risk has different use cases such as reinforcement learning, as it lacks positive homogeneity and is thus not suitable in a standard supervised learning risk minimization loop. Finally, use cases for the Orlicz regret measure may appear in risk-sensitive regression; we refer to [Rockafellar and Uryasev \[2013\]](#) and [Rockafellar et al. \[2008\]](#) for regression based on *error measures* E , which are based on regret measures V via the unique correspondence $E(X) = V(X) - \mathbb{E}[X]$. For instance, quantile regression is associated with the CVar_α Orlicz regret measure.

3.9 CONCLUSION

In this work, we have brought together concepts such as the fundamental function and envelope function from the theory of r.i. Banach function spaces with the framework of coherent risk measures, in particular divergence risk measures. As a consequence, we have obtained a general approach to fine-tune the tail sensitivity of a risk measure, which then functions as a more cautious replacement for the expectation in ERM, or more generally, can be used to assess the risk inherent in a loss distribution.

We have demonstrated how tail sensitivity is related to the fundamental function. To a first approximation, those risk measures with the same fundamental function share the same tail sensitivity; on a finer level, however, it might be the case that the spaces on which different such risk measures are finite do not coincide. Nonetheless, the fundamental function offers a useful and simple stratification of the space of coherent risk measures. In this

way, the choice of fundamental function indeed appears more *fundamental* than the exact choice of risk measure; Marcinkiewicz, Orlicz, Lorentz or a still different one.

Abundant opportunities for future work arise. For instance, a quantitative comparison of the (positive translation equivariant) Marcinkiewicz, Orlicz and Lorentz norm and their corresponding risk measures on practical problems has not been conducted yet. Also, estimatability from finite samples is an open problem: we conjecture that the more tail-sensitive a risk measure, the harder it is to estimate it from finite samples. For instance, can one obtain convergence rates based on properties of the fundamental function?

3.10 APPENDIX: ORLICZ REGRET MEASURES

3.10.1 PROOF OF PROPOSITION 101

Proof The convex conjugate of a (not necessarily convex) function is always convex. To see that g is finite-valued on $\mathbb{R}$ due to the supercoercivity of f (a well-known statement, see [Edgar and Sucheston, 1992, Dommel and Pichler, 2020, Föllmer and Schied, 2002]), assume by contradiction that there exists a $y \in \mathbb{R}$ such that

$$\infty = g(y) = \sup_{x \geq 0} xy - f(x). \quad (3.116)$$

Since f is proper, i.e. never takes on $-\infty$, the supremum cannot be attained for some $x \in \mathbb{R}$. Therefore we need to investigate only the limit as $x \rightarrow \infty$:

$$g(y) = \lim_{x \rightarrow \infty} xy - f(x) = \lim_{x \rightarrow \infty} x \left(y - \frac{f(x)}{x} \right). \quad (3.117)$$

But since y is fixed, this could only yield ∞ if $\lim_{x \rightarrow \infty} \left(y - \frac{f(x)}{x} \right) \geq 0$, which contradicts the supercoercivity assumption $\lim_{x \rightarrow \infty} \frac{f(x)}{x} = \infty$.

To see that g is increasing (non-decreasing), observe that if $y_1 \leq y_2$, then

$$g(y_1) = \sup_{z \geq 0} y_1 z - f(z) \leq \sup_{z \geq 0} y_2 z - f(z) = g(y_2). \quad (3.118)$$

■

3.10.2 PROOF OF PROPOSITION 108

Proof V1: Positive homogeneity: let $\lambda > 0$.

$$V_g(\lambda X) = \inf_{t>0} t \left(1 + \mathbb{E}g \left(\frac{\lambda X}{t} \right) \right) \quad (3.119)$$

$$= \inf_{\tilde{t}>0} \tilde{t} \lambda \left(1 + \mathbb{E}g \left(\frac{X}{\tilde{t}} \right) \right), \quad \tilde{t} = \frac{t}{\lambda} \quad (3.120)$$

$$= \lambda V_g(X). \quad (3.121)$$

V2: For subadditivity, we present the proof from [Cui et al., 2008] for subadditivity of an Orlicz norm, which also works in our setting of an Orlicz regret measure. First observe that $X \mapsto 1 + \mathbb{E}g(X)$ is convex due to the convexity of g . We also know that for an arbitrary $\epsilon > 0$ there exist some $t_1, t_2 > 0$ so that

$$\frac{1}{t_1}(1 + \mathbb{E}g(t_1 X)) \leq V_g(X) + \frac{\epsilon}{2}, \quad (3.122)$$

$$\frac{1}{t_2}(1 + \mathbb{E}g(t_2 Y)) \leq V_g(Y) + \frac{\epsilon}{2}. \quad (3.123)$$

Let $t = \frac{t_1 t_2}{t_1 + t_2}$. Then it holds

$$V_g(X + Y) \leq \frac{1}{t}(1 + \mathbb{E}g(t(X + Y))) \quad (3.124)$$

$$= \frac{t_1 + t_2}{t_1 t_2} \left(1 + \mathbb{E}g \left(\frac{t_2}{t_1 + t_2} t_1 X + \frac{t_1}{t_1 + t_2} t_2 Y \right) \right) \quad (3.125)$$

$$\leq \frac{1}{t_1}(1 + \mathbb{E}g(t_1 X)) + \frac{1}{t_2}(1 + \mathbb{E}g(t_2 Y)) \leq V_g(X) + V_g(Y) + \epsilon. \quad (3.126)$$

V3: Monotonicity directly follows from monotonicity of g (Lemma 101). ■

V4: We want to show $V_g(X) \geq \mathbb{E}[X] \forall X \in \mathcal{M}$. Let $f \in \mathfrak{F}$ be a divergence function, i.e. we have $f(1) = 0$. Consequently, we have that

$$g(y) = \sup_{x \geq 0} xy - f(x) \geq y - f(1) = y \quad \forall y \in \mathbb{R}. \quad (3.127)$$

But then:

$$\mathbb{E}[X] = t \mathbb{E} \left[\frac{X}{t} \right] \leq \inf_{t>0} t + t \mathbb{E} \left[\frac{X}{t} \right] \quad (3.128)$$

$$\leq \inf_{t>0} t \left(1 + \mathbb{E}g \left(\frac{X}{t} \right) \right) = V_g(X). \quad (3.129)$$

V5: To see that $X = 0 \Rightarrow V_g(0) = 0$, observe that

$$V_g(0) = \inf_{t>0} t(1 + \mathbb{E}g(0)) = \lim_{t \rightarrow 0} t(1 + \mathbb{E}g(0)) = 0, \quad (3.130)$$

due to

$$g(0) = \sup_{x \geq 0} 0 \cdot x - f(x) < \infty. \quad (3.131)$$

Also note that since $f(1) = 0$, we have $g(0) \geq 0$.

For the converse direction, assume $V_g(X) = 0$ for $X \geq 0$. We want to show that it follows that $X = 0$. By contradiction, assume $X \neq 0$ (P-a.e.). But then since $X \geq 0$, we have $\mathbb{E}[X] > 0$. Due to aversity (V4) we know $V_g(X) \geq \mathbb{E}[X]$, therefore $V_g(X) > 0$.

3.10.3 PROOF OF PROPOSITION 109

Proof

$$V_{g_\varepsilon}(X) = \inf_{t>0} t \left(1 + \mathbb{E}g_\varepsilon \left(\frac{X}{t} \right) \right) \quad (3.132)$$

$$= \inf_{t>0} t \left(1 + \frac{1}{\varepsilon} \mathbb{E}g \left(\frac{\varepsilon X}{t} \right) \right) \quad (3.133)$$

$$\stackrel{\tilde{t} = \frac{t}{\varepsilon}}{=} \inf_{\tilde{t}>0} \tilde{t} \varepsilon \left(1 + \frac{1}{\varepsilon} \mathbb{E}g \left(\frac{X}{\tilde{t}} \right) \right) \quad (3.134)$$

$$= \inf_{\tilde{t}>0} \tilde{t} \left(\varepsilon + \mathbb{E}g \left(\frac{X}{\tilde{t}} \right) \right). \quad (3.135)$$

■

3.10.4 PROOF OF PROPOSITION 110

Proof That $\bar{g}$ is a valid Young function is obvious since $\bar{g}(0) = 0$ and it inherits the properties of g (increasing, convex etc.) on the nonnegative domain. Recall that $\bar{f} = f \cdot \chi_{[1,\infty)}$. Its conjugate is therefore given by

$$h(y) := \sup_{x \geq 0} xy - \bar{f}(x) = \max \left\{ \sup_{0 \leq x < 1} xy, \sup_{x \geq 1} xy - \bar{f}(x) \right\}. \quad (3.136)$$

For negative y , we have $h(y) = 0$, since the maximum is attained in the left term with value 0 (here the supremum is attained at $x = 0$), since the right term will be negative. Thus $h(y) = \bar{g}(y) = 0$ for $y < 0$. It remains to show that for positive y we have $h(y) = g(y)$. For

this, we show that for positive y , only the second term of g suffices:

$$g(y) = \max \left\{ \sup_{0 \leq x < 1} xy - f(x), \sup_{x \geq 1} xy - f(x) \right\} = \sup_{x \geq 1} xy - f(x), \quad y > 0. \quad (3.137)$$

Assume by contradiction that y is positive and the maximum is attained in the first term. Note that $\sup_{x \geq 1} xy - f(x) \geq y$ since $f(1) = 0$. But:

$$\sup_{0 \leq x < 1} xy - f(x) \leq y, \quad (3.138)$$

since $xy \leq y$ and $f(x) \geq 0$. Hence the maximum is attained in the second term. Thus we have shown that $\bar{g} = h$ and therefore $\bar{g}$ is the conjugate of $\bar{f}$. ■

3.10.5 PROOF OF PROPOSITION 113

Proof The first inequality is obvious. Second,

$$\|X\|_{V_{g_{\varepsilon_2}}} = \inf_{t > 0} t \left(\varepsilon_2 + \mathbb{E}g \left(\frac{X}{t} \right) \right) \leq \inf_{t > 0} t \left(\varepsilon_2 + \frac{\varepsilon_2}{\varepsilon_1} \mathbb{E}g \left(\frac{X}{t} \right) \right) \quad (3.139)$$

$$= \frac{\varepsilon_2}{\varepsilon_1} \inf_{t > 0} t \left(\varepsilon_1 + \mathbb{E}g \left(\frac{X}{t} \right) \right) = \frac{\varepsilon_2}{\varepsilon_1} \|X\|_{V_{g_{\varepsilon_1}}}, \quad (3.140)$$

which is true since $\frac{\varepsilon_2}{\varepsilon_1} > 1$. ■

3.10.6 PROOF OF PROPOSITION 116

To establish the dual representation

$$V_g(X) = \sup \{ \mathbb{E}[XZ] : Z \in \mathcal{L}^{\bar{f}}, Z \geq 0, \mathbb{E}[f(Z)] \leq 1 \} \quad \forall X \in \mathcal{L}^{\bar{g}}, \quad (3.141)$$

we follow roughly [Kosmol and Müller-Wichards, 2011, pp. 238-239]. First note that V_g is finite on $\mathcal{L}^{\bar{g}}$ (Proposition 111). We begin with dual variables $Z \in \mathcal{L}^1$, but later show that the restriction to $\mathcal{L}^{\bar{f}}$ is possible. The crucial point is that the functionals $X \mapsto \mathbb{E}[g(X)]$ and $Z \mapsto \mathbb{E}[f(Z)]$ stand in a conjugate relationship due to the conjugacy of g and f :

$$\mathbb{E}[g(X)] = \sup_{Z \in \mathcal{L}^1, Z \geq 0} \mathbb{E}[XZ] - \mathbb{E}[f(Z)] \quad (3.142)$$

This result is from [Rockafellar, 1968]. It follows from the fact that g is a lower semicontinuous proper convex function with interior points in its domain and therefore, $G(\omega, x) = g(x)$ is a normal convex integrand [Rockafellar, 1968, Lemma 2]. The conjugacy then follows from [Rockafellar, 1968, Corollary in Theorem 2]. Also note that the restriction to dual variables with $Z \geq 0$ is possible due to $f(z) = \infty$ for $z < 0$ (this is essentially due to the monotonicity of V_g ¹⁵). Then the Orlicz regret measure can be formulated as

$$V_g(X) = \inf_{t>0} t \left(1 + \mathbb{E}g\left(\frac{X}{t}\right) \right) = \inf_{t>0} t \left(1 + \sup_{Z \in \mathcal{L}^1, Z \geq 0} \left(\mathbb{E}\left[\frac{X}{t}Z\right] - \mathbb{E}[f(Z)] \right) \right) \quad (3.143)$$

$$= \inf_{t>0} \sup_{Z \in \mathcal{L}^1, Z \geq 0} \{ \mathbb{E}[XZ] - t(\mathbb{E}[f(Z)] - 1) \}. \quad (3.144)$$

Now consider the dual representation:

$$DR(X) := \sup \{ \mathbb{E}[XZ] : Z \in \mathcal{L}^1, Z \geq 0, \mathbb{E}[f(Z)] \leq 1 \} \quad \forall X \in \mathcal{L}^1. \quad (3.145)$$

Observe that $Z = 1$ is an interior point of the constraint set due to the fact that $f(1) = 0$. We therefore get strong Lagrangian duality since Slater's condition holds, as observed for instance by Shapiro [2017]:

$$DR(X) = \inf_{t>0} \sup_{Z \in \mathcal{L}^1, Z \geq 0} \{ \mathbb{E}[XZ] - t(\mathbb{E}[f(Z)] - 1) \} = V_g(X), \quad (3.146)$$

which establishes the desired equality. Furthermore, it is sufficient to restrict ourselves to dual variables $Z \in \mathcal{L}^{\bar{f}}$. Assume by contraposition $Z \notin \mathcal{L}^{\bar{f}}$. Hence $\|Z\|_{\bar{f}}^L = \infty$. But $\|Z\|_{\bar{f}}^L \leq \mathbb{E}[\bar{f}(Z)]$ holds, see Rubshtein et al. [2016, Proposition 14.2.1]. But then $\mathbb{E}[\bar{f}(Z)] = \infty$, and since $\mathbb{E}[\bar{f}(Z)] \leq \mathbb{E}[f(Z)]$ such Z is not feasible.

3.10.7 PROOF OF PROPOSITION 118

Proof We begin to show that the dual representation holds:

$$R_{g,\varepsilon}(X) = \sup \{ \mathbb{E}[XZ] : Z \in \mathcal{L}^{\bar{f}}, Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}[f(Z)] \leq \varepsilon \} \quad \forall X \in \mathcal{L}^{\bar{g}}. \quad (3.147)$$

Consider how we defined the divergence risk measure in the first place:

$$R_{g,\varepsilon}(X) := \sup_{Q: d_f(Q,P) \leq \varepsilon} \mathbb{E}_Q[X]. \quad (3.148)$$

¹⁵The following statement is well-known under different technical setups: a regret measure is monotone if and only if all dual variables are nonnegative. See for instance [Rockafellar and Uryasev, 2013] for the $\mathcal{L}^2$ case.

Here, the supremum is taken over the set of all valid probability measures Q which are absolutely continuous with respect to P , and further satisfy the f -divergence constraint. The necessary shift in perspective is to view Q not as a probability measure, but instead take the following route:

$$Q_Z(A) = \int_A Z(\omega) \, dP(\omega), \quad (3.149)$$

for the Radon-Nikodym derivative Z (which can be more suggestively written as $\frac{dQ}{dP}$). Q is a measure, i.e. a set function, whereas Z is a random variable. Then observe that

$$d_f(Q, P) \leq \varepsilon \iff \mathbb{E}f(Z) \leq \varepsilon, \quad (3.150)$$

where the expectation is with respect to the base measure P . Furthermore, imposing the constraint that Z yields a valid probability measure via Equation 3.33 amounts exactly to requiring that $Z \geq 0$ and $\mathbb{E}[Z] = 1$.

Observe that the envelope representation Equation 3.147 is the envelope representation of the corresponding Orlicz regret measure, but intersected with the constraint that $\mathbb{E}[Z] = 1$. If we worked on a reflexive Banach space (where the bidual coincides with the space itself), together with an appropriate closedness condition, we could use the following result. For instance, consider for a moment the dual pair of $\mathcal{L}^2$ and $\mathcal{L}^2$ (the exponent 2 has dual exponent 2 since $1/2 + 1/2 = 1$).

Denote the support function of the set $\mathcal{Q}$ as $\sigma_{\mathcal{Q}}$:

$$\sigma_{\mathcal{Q}}(X) = \sup_{Q \in \mathcal{Q}} \langle X, Q \rangle, \quad \mathcal{Q} = \{Q \in \mathcal{L}^2 : \langle X, Q \rangle \leq \sigma_{\mathcal{Q}}(X) \, \forall X \in \mathcal{L}^2\}, \quad (3.151)$$

and define $\mathcal{E}_1 := \{Z \in \mathcal{L}^2 : \mathbb{E}[Z] = 1\}$. The following is well-known (see for instance [Fröhlich and Williamson \[2024a\]](#) for a statement on $\mathcal{L}^2$):

Proposition 176. *On $\mathcal{L}^2$: let $V = \sigma_{\mathcal{Q}}$ be a positively homogeneous, subadditive and monotonic functional, i.e. a coherent regret measure, and $\mathcal{Q}$ be its supported set. Let $\mathcal{Q}' := \mathcal{Q} \cap \mathcal{E}_1 \neq \emptyset$. Then $R := \sigma_{\mathcal{Q}'}$ is a coherent risk measure and $R(X) = \inf_{c \in \mathbb{R}} V(X - c) + c$.*

The proof uses reflexivity of $\mathcal{L}^2$. Unfortunately, Orlicz spaces are reflexive if and only if they satisfy the Δ_2 condition, which limits the speed of growth of the Young function; see for instance [\[Kosmol and Müller-Wichards, 2011, p. 197, p. 234\]](#). Since we do not want to assume Δ_2 throughout (this would for instance exclude the KL divergence risk measure), we cannot rely on the statement, but we are guided by it for intuition.

Our aim is to show that $R_{g,\varepsilon}$ is given by the infimal convolution

$$R_{g,\varepsilon}(X) = \inf_{\mu \in \mathbb{R}} \mu + V_{g,\varepsilon}(X - \mu). \quad (3.152)$$

The proof works analogously to 3.10.6 and hence [Kosmol and Müller-Wichards, 2011]; confer also [Dommel and Pichler, 2020, Theorem 5.1]. Start with the envelope representation

$$R_g(X) := \sup \left\{ \mathbb{E}[XZ] : Z \in \mathcal{L}^{\bar{f}}, Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}[f_\varepsilon(Z)] \leq 1 \right\} \quad \forall X \in \mathcal{L}^1. \quad (3.153)$$

Regarding the restriction to $Z \in \mathcal{L}^{\bar{f}}$, the same remark as in 3.10.6 applies. Again, since $Z = 1$ is an interior point of the constraint set, observe that by Lagrange duality:

$$R_g(X) = \inf_{\tilde{\mu} \in \mathbb{R}, t > 0} \sup_{Z \in \mathcal{L}^1, Z \geq 0} \{ \mathbb{E}[XZ] - t(\mathbb{E}[f_\varepsilon(Z)] - 1) - \tilde{\mu}(\mathbb{E}[Z] - 1) \} \quad (3.154)$$

$$= \inf_{\tilde{\mu} \in \mathbb{R}, t > 0} t \left(1 + \frac{\tilde{\mu}}{t} + \sup_{Z \in \mathcal{L}^1, Z \geq 0} \mathbb{E} \left[\left(\frac{X - \tilde{\mu}}{t} \right) Z \right] - \mathbb{E}[f_\varepsilon(Z)] \right) \quad (3.155)$$

$$\stackrel{\mu = \frac{\tilde{\mu}}{t}}{=} \inf_{\mu \in \mathbb{R}, t > 0} t \left(1 + \mu + \sup_{Z \in \mathcal{L}^1, Z \geq 0} \mathbb{E} \left[\left(\frac{X}{t} - \mu \right) Z \right] - \mathbb{E}[f_\varepsilon(Z)] \right) \quad (3.156)$$

$$= \inf_{\mu \in \mathbb{R}, t > 0} t \left(1 + \mu + \mathbb{E} g_\varepsilon \left(\frac{X}{t} - \mu \right) \right), \quad (3.157)$$

where we applied Equation 3.142 to the conjugacy relation of f_ε and g_ε . We started with $X \in \mathcal{L}^1$, but the subsequent Proposition 120 shows that R_g is finite on $\mathcal{L}^{\bar{g}}$, which is its natural domain [Pichler, 2013]. Henceforth we restrict X to $\mathcal{L}^{\bar{g}}$.

Now we explicitly compute the infimal convolution of V_{g_ε} :

$$R_{g,\varepsilon}(X) = \inf_{\mu \in \mathbb{R}} \mu + V_{g_\varepsilon}(X - \mu) \quad (3.158)$$

$$= \inf_{\mu \in \mathbb{R}, \tilde{t} > 0} \mu + \tilde{t} \left(\varepsilon + \mathbb{E} g \left(\frac{X - \mu}{\tilde{t}} \right) \right) \quad (3.159)$$

$$\stackrel{\mu = \tilde{t}\tilde{\mu}}{=} \inf_{\tilde{\mu} \in \mathbb{R}, \tilde{t} > 0} \tilde{t}\tilde{\mu} + \tilde{t} \left(\varepsilon + \mathbb{E} g \left(\frac{X - \tilde{t}\tilde{\mu}}{\tilde{t}} \right) \right) \quad (3.160)$$

$$= \inf_{\tilde{\mu} \in \mathbb{R}, \tilde{t} > 0} \tilde{t} \left(\varepsilon + \tilde{\mu} + \mathbb{E} g \left(\frac{X}{\tilde{t}} - \tilde{\mu} \right) \right), \quad (3.161)$$

where we used the form of V_{g_ε} with explicit ε (Proposition 109). This is the representation obtained by Dommel and Pichler [2020]. But we can also hide ε in the function g_ε :

$$R_{g,\varepsilon}(X) = \inf_{\tilde{\mu} \in \mathbb{R}, \tilde{t} > 0} \tilde{t} \left(\varepsilon + \tilde{\mu} + \mathbb{E}g \left(\frac{X}{\tilde{t}} - \tilde{\mu} \right) \right) \quad (3.162)$$

$$\stackrel{\tilde{\mu} = \mu\varepsilon}{=} \inf_{\mu \in \mathbb{R}, \tilde{t} > 0} \tilde{t}\varepsilon \left(1 + \mu + \frac{1}{\varepsilon} \mathbb{E}g \left(\frac{X}{\tilde{t}} - \mu\varepsilon \right) \right) \quad (3.163)$$

$$\stackrel{\tilde{t} = t}{=} \inf_{\mu \in \mathbb{R}, t > 0} t \left(1 + \mu + \frac{1}{\varepsilon} \mathbb{E}g \left(\varepsilon \left(\frac{X}{t} - \mu \right) \right) \right) \quad (3.164)$$

$$= \inf_{\mu \in \mathbb{R}, t > 0} t \left(1 + \mu + \mathbb{E}g_\varepsilon \left(\frac{X}{t} - \mu \right) \right), \quad (3.165)$$

which coincides with Equation 3.157 and therefore we have established that $R_{g,\varepsilon}(X) = \inf_{\mu \in \mathbb{R}} \mu + V_{g_\varepsilon}(X - \mu)$.

The “natural extension” representation: Finally, we establish that the following envelope representation also holds:

$$R_{g,\varepsilon}(X) = \sup \left\{ E_{Q_Z}[X] : Q_Z \in \mathcal{Q}, \mathbb{E}_{Q_Z}[Y] \leq V_{g_\varepsilon}(Y) \quad \forall Y \in \mathcal{L}^{\bar{g}} \right\} \quad \forall X \in \mathcal{L}^{\bar{g}}. \quad (3.166)$$

When one can work with support functions, this representation is easy to show¹⁶. However, in our Orlicz setting we have to prove it. For this, consider the representation (the validity of which we show later):

$$R_{g,\varepsilon}(X) = \sup \left\{ E[XZ] - \alpha(Z) : Z \in \mathcal{L}^{\bar{f}}, Z \geq 0, \mathbb{E}[Z] = 1 \right\} \quad \forall X \in \mathcal{L}^{\bar{g}}, \quad (3.167)$$

where

$$\alpha(Z) := \sup \left\{ \mathbb{E}[XZ] : X \in \mathcal{L}^{\bar{g}}, R_{g,\varepsilon}(X) \leq 0 \right\}, \quad (3.168)$$

which is the envelope representation of a convex risk measure. Since in our case, we have a *coherent* risk measure, it is standard that in fact the “penalty function” α can be taken to be the convex indicator $i_{\mathcal{Z}}$ of the envelope $\mathcal{Z}$, where $\mathcal{Z} := \{Z \in \mathcal{L}^{\bar{f}} : Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}f_\varepsilon(Z) \leq 1\}$. Then:

$$R_{g,\varepsilon}(X) = \sup \left\{ E[XZ] - i_{\mathcal{Z}}(Z) : Z \in \mathcal{L}^{\bar{f}}, Z \geq 0, \mathbb{E}[Z] = 1 \right\}. \quad (3.169)$$

The *acceptance set* is $\mathcal{A} = \{X \in \mathcal{L}^{\bar{g}} : R_{g,\varepsilon}(X) \leq 0\}$. We have:

$$\alpha(Z) = i_{\mathcal{Z}}(Z) = \sup \left\{ \mathbb{E}[XZ] : X \in \mathcal{L}^{\bar{g}}, R_{g,\varepsilon}(X) \leq 0 \right\}. \quad (3.170)$$

Recall that the convex indicator function is $i_{\mathcal{Z}}(Z) = 0$ if $Z \in \mathcal{Z}$ and $+\infty$ otherwise. Thus

¹⁶Just consider the definition of the supported set of a support function and then use standard results to obtain the constraints $Z \geq 0$ and $\mathbb{E}[Z] = 1$ from monotonicity and translation equivariance, which imply $Q_Z \in \mathcal{Q}$.

we find that

$$\mathcal{Z} = \left\{ Z \in \mathcal{L}^{\bar{f}} : Z \geq 0, \mathbb{E}[Z] = 1, 0 = \sup\{\mathbb{E}[XZ] : X \in \mathcal{L}^{\bar{g}}, R_{g_\varepsilon}(X) \leq 0\} \right\} \quad (3.171)$$

$$= \left\{ Z \in \mathcal{L}^{\bar{f}} : Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}[XZ] \leq 0 \quad \forall X \in \mathcal{L}^{\bar{g}} \text{ with } R_{g_\varepsilon}(X) \leq 0 \right\}, \quad (3.172)$$

This follows since $R_{g_\varepsilon}(0) = 0$, i.e. 0 is always in the acceptance set, hence the inner supremum in Equation 3.171 can never be < 0 . Further, with similar reasoning as in [Bellini and Gianin \[2008, Proposition 17\]](#):

$$\mathcal{Z} = \left\{ Z \in \mathcal{L}^{\bar{f}} : Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}[XZ] \leq 0 \quad \forall X \in \mathcal{L}^{\bar{g}} \text{ with } R_{g_\varepsilon}(X) \leq 0 \right\} \quad (3.173)$$

$$= \left\{ Z \in \mathcal{L}^{\bar{f}} : Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}[XZ] \leq R_{g_\varepsilon}(X) \quad \forall X \in \mathcal{L}^{\bar{g}} \right\} =: \mathcal{Z}^+. \quad (3.174)$$

$$(3.175)$$

That $\mathcal{Z}^+ \subseteq \mathcal{Z}$ is obvious from the definition of R_{g_ε} . Conversely, we show $\mathcal{Z} \subseteq \mathcal{Z}^+$. Let $Z \in \mathcal{Z}$, i.e. $Z \geq 0$, $\mathbb{E}[Z] = 1$ and $\mathbb{E}[XZ] \leq 0 \quad \forall X \in \mathcal{L}^{\bar{g}} \text{ with } R_{g_\varepsilon}(X) \leq 0$. Then consider $Y := X - R_{g_\varepsilon}(X)$ for arbitrary $X \in \mathcal{L}^{\bar{g}}$. Then $Y \in \mathcal{L}^{\bar{g}}$ since an Orlicz space is closed under subtraction of constants (consider the definition of the Luxemburg norm). Due to translation equivariance, $R_{g_\varepsilon}(Y) = R_{g_\varepsilon}(X - R_{g_\varepsilon}(X)) = R_{g_\varepsilon}(X) - R_{g_\varepsilon}(X) = 0 \leq 0$, hence Y is acceptable and therefore $\mathbb{E}[YZ] \leq 0$ by assumption that $Z \in \mathcal{Z}$. Then we know that

$$\mathbb{E}[YZ] \leq 0 \Leftrightarrow \mathbb{E}[(X - R_{g_\varepsilon}(X))Z] \leq 0 \Leftrightarrow \mathbb{E}[XZ] - R_{g_\varepsilon}(X) \leq 0 \Leftrightarrow \mathbb{E}[XZ] \leq R_{g_\varepsilon}(X), \quad (3.176)$$

noting that for a constant c , it holds $\mathbb{E}[(X - c)Z] = \mathbb{E}[XZ] - c$ since $\mathbb{E}[Z] = 1$. Since X was arbitrary in $\mathcal{L}^{\bar{g}}$, we conclude that $Z \in \mathcal{Z}^+$, and since Z was arbitrary, $\mathcal{Z} \subseteq \mathcal{Z}^+$. Having established $\mathcal{Z} = \mathcal{Z}^+$, we continue:

$$\mathcal{Z} = \mathcal{Z}^+ = \left\{ Z \in \mathcal{L}^{\bar{f}} : Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}[XZ] \leq \inf_{\mu \in \mathbb{R}} \mu + V_{g_\varepsilon}(X - \mu) \quad \forall X \in \mathcal{L}^{\bar{g}} \right\} \quad (3.177)$$

$$= \left\{ Z \in \mathcal{L}^{\bar{f}} : Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}[XZ] \leq \mu + V_{g_\varepsilon}(X - \mu) \quad \forall X \in \mathcal{L}^{\bar{g}}, \forall \mu \in \mathbb{R} \right\} \quad (3.178)$$

$$= \left\{ Z \in \mathcal{L}^{\bar{f}} : Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}[XZ] - \mu \leq V_{g_\varepsilon}(X - \mu) \quad \forall X \in \mathcal{L}^{\bar{g}}, \forall \mu \in \mathbb{R} \right\} \quad (3.179)$$

$$= \left\{ Z \in \mathcal{L}^{\bar{f}} : Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}[(X - \mu)Z] \leq V_{g_\varepsilon}(X - \mu) \quad \forall X \in \mathcal{L}^{\bar{g}}, \forall \mu \in \mathbb{R} \right\} \quad (3.180)$$

$$= \left\{ Z \in \mathcal{L}^{\bar{f}} : Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}[YZ] \leq V_{g_\varepsilon}(Y) \quad \forall Y \in \mathcal{L}^{\bar{g}} \right\}, \quad (3.181)$$

$$(3.182)$$

again noting that $\mathbb{E}[(X - \mu)Z] = \mathbb{E}[XZ] - \mu$ since $\mathbb{E}[Z] = 1$. Also, we may substitute $Y = X - \mu$, where $\mu = 0$.

We still need to establish that the representation in Equation 3.167 holds. It follows from [Arai \[2010, Theorem 1\]](#); note that their strictness assumption on the Young function is not needed for this. To apply the result, it remains to show that $R_{g,\varepsilon}$ has the “order lower-semicontinuous” property [\[Biagini and Frittelli, 2009\]](#). But this follows from the fact that we have before established the envelope representation Equation 3.147, which expresses $R_{g,\varepsilon}$ as a pointwise supremum over a family of order lower-semicontinuous functionals. Thus it is itself order lower-semicontinuous [\[Biagini and Frittelli, 2009, Section 3.1\]](#); note that all Orlicz Banach lattices are order separable [\[Biagini and Frittelli, 2009, Remark 16\]](#). ■

3.11 APPENDIX: REARRANGEMENT INVARIANT BANACH NORMS AND FUNDAMENTAL FUNCTIONS

3.11.1 PROOF OF PROPOSITION 133

Proof Due to nonnegativity of indicator functions, the fundamental function coincides with the fundamental function of the Orlicz norm $\|\cdot\|_g^O$ (due to Proposition 111). It is known [\[Kosmol and Müller-Wichards, 2011, Lemma 7.4.4\]](#) that this fundamental function is $\phi(t) = t\bar{f}^{-1}(1/t)$; since $1/t \geq 1$, the choice of $\bar{f}$ or f does not matter; note that we defined $f^{-1}(y) := \sup\{x : f(x) < y\}$. To intuit the result, we give a non-rigorous “computation”. Let E be an arbitrary set of measure t :

$$\phi_{V_g}(t) = V_g(\chi_E) = \sup \left\{ \int_0^1 \chi_E(\omega) Z^*(\omega) \, d\omega : \mathbb{E}f(Z) \leq 1 \right\} \quad (3.183)$$

$$= \sup \{ t \cdot z_1 : t \cdot f(z_1) + (1-t) \cdot f(z_0) \leq 1 \}. \quad (3.184)$$

That it is possible to decompose Z^* into z_1 and z_0 is intuitive, but requires formal justification. Due to the nonnegativity of f , clearly $z_0 = 0$ is the best choice. For $t \cdot z_1$ to be as large as possible, we then need $t \cdot f(z_1) = 1$ (recall the left-continuity of f), hence $z_1 = f^{-1}(1/t)$. Thus $\phi_{V_g}(t) = t f^{-1}(1/t)$. Noting that $(\frac{1}{\varepsilon} f(x))^{-1} = f^{-1}(\varepsilon x)$, the fundamental function at different risk aversion levels can be obtained easily.

To see that ϕ_{V_g} is concave, note that since $\bar{f}$ is convex and increasing (non-decreasing), $\bar{f}^{-1}$ is concave [\[Rubshtein et al., 2016, p. 172\]](#). Let $a_{\bar{f}} := \sup\{x : \bar{f}(x) = 0\}$. Note that $\bar{f}$ is strictly increasing on $[a_{\bar{f}}, \infty)$ and $\bar{f}^{-1}$ can only have a jump at 0. Then $p \mapsto p(-\bar{f}^{-1}(1/p))$ is convex, due to the perspective transform of a convex function being convex in both arguments, whence it follows that $p \mapsto p\bar{f}^{-1}(1/p)$ is concave on $(0, 1]$. The next proposition, 3.11.2 shows that it is furthermore continuous at 0 and therefore concave on $[0, 1]$. ■

3.11.2 PROOF OF PROPOSITION 134

Proof FF1: Consider $\phi_{V_g}(t) = t\bar{f}^{-1}\left(\frac{1}{t}\right)$. Hence

$$\lim_{t \downarrow 0} \phi_{V_g}(t) = \lim_{t \downarrow 0} t\bar{f}^{-1}\left(\frac{1}{t}\right) \stackrel{x=1/t}{=} \lim_{x \rightarrow \infty} \frac{\bar{f}^{-1}(x)}{x}. \quad (3.185)$$

By the supercoercivity assumption on f (equivalently, on $\bar{f}$), we have $\lim_{x \rightarrow \infty} \frac{\bar{f}(x)}{x} = \infty$. But then, its inverse f^{-1} (resp. $\bar{f}^{-1}$) satisfies the mirror property of “anti-supercoercivity”: $\lim_{x \rightarrow \infty} \frac{\bar{f}^{-1}(x)}{x} = 0$. For the case when f is finite, see [Kosmol and Müller-Wichards, 2011, Lemma 6.1.26]. If f is not finite everywhere, then its inverse is bounded; therefore $\lim_{x \rightarrow \infty} \frac{\bar{f}^{-1}(x)}{x} = 0$.

It is easy to see that this holds also in the other direction. Assume that $\lim_{t \downarrow 0} \phi_{V_g}(t) = 0$. Then $\bar{f}$ is supercoercive.

It remains to establish that the same property holds for ϕ_{R_g} . This readily follows from the equivalence of risk and regret: we can apply Theorem 4.22 in Fröhlich and Williamson [2024a] and consider the fraction:

$$K'' = \lim_{t \downarrow 0} \frac{\phi_{R_g}(t)}{\phi_{V_g}(t)}. \quad (3.186)$$

If ϕ_{V_g} was continuous at 0, but ϕ_{R_g} was not, then $K'' = \infty$ and then it follows that the norms would not be equivalent.

FF2: We need to show that if f is finite then $\phi'(0) = \lim_{t \downarrow 0} \frac{\phi(t)}{t} = \infty$ and otherwise $\phi'(0) < \infty$. We begin again by considering ϕ_{V_g} . From the definition of the difference quotient we get:

$$\phi'_{V_g}(0) = \lim_{t \downarrow 0} \frac{\phi_{V_g}(t)}{t}, \quad (3.187)$$

and further

$$\lim_{t \downarrow 0} \frac{\phi_{V_g}(t)}{t} = \lim_{t \downarrow 0} \bar{f}^{-1}\left(\frac{1}{t}\right) = \lim_{x \rightarrow \infty} \bar{f}^{-1}(x). \quad (3.188)$$

If f is finite, then so is $\bar{f}$ and its inverse is unbounded and therefore we obtain $\phi'_{V_g}(0) = \infty$. If $f(x) = \infty$ for some $x \in \mathbb{R}^+$, then its inverse is bounded; consequently, $\phi'_{V_g}(0) < \infty$.

We also have $\phi'_{V_g}(0) = \infty \Leftrightarrow \phi'_{R_g}(0) = \infty$, that is, the fundamental function of the risk measure shares the same property. This readily follows by applying Fröhlich and Williamson [2024a, Corollary 4.27].

■

3.11.3 PROOF OF PROPOSITION 137

Proof First, we observe that there exists $t_0 < 1$ such that $\phi_{V_g}(t_0) = t_0 f^{-1}(1/t_0) = 1$, unless f is the pathological divergence function corresponding to the expectation. This is due to $f^{-1}(1) > 1$ (unless $\lim_{x \downarrow 1} f(x) = \infty$). Since the fundamental function is increasing and continuous, it follows that there exists $t_0 < 1$ such that $\phi_{V_g}(t_0) = t_0 f^{-1}(1/t_0) = 1$.

Let E be an arbitrary measurable set of measure t . The fundamental function is then:

$$\phi_{R_g}(t) = R_g(\chi_E) = \sup \left\{ \int_0^1 \chi_E(\omega) Z^*(\omega) d\omega : Z \geq 0, \mathbb{E}[Z] = 1, \mathbb{E}f(Z) \leq 1 \right\}. \quad (3.189)$$

Next, observe that for $t \leq t_0$ the following dual variable $f^{-1}(1/t)$ is feasible:

$$Z(\omega) = \begin{cases} 1/t & \omega \in E \\ \frac{1 - t f^{-1}(1/t)}{1 - t} & \omega \notin E. \end{cases} \quad (3.190)$$

Since $Z \geq 0$, $\mathbb{E}[Z] = 1$ (by construction) and furthermore, noting the left-continuity of f^{-1} :

$$\mathbb{E}f(Z) = t \cdot f(f^{-1}(1/t)) + (1 - t) \cdot f\left(\frac{1 - t f^{-1}(1/t)}{1 - t}\right) \leq 1. \quad (3.191)$$

This holds since if $0 < t < 1$ and $f^{-1}(1/t) > 1$ and $t f^{-1}(1/t) \leq 1$ (since $t \leq t_0$), we have

$$0 \leq \underbrace{\frac{1 - t f^{-1}(1/t)}{1 - t}}_{=: \star} \leq 1, \quad (3.192)$$

and because f is by assumption a Young divergence, $f(\star) = 0$. Thus we have shown that this Z is a feasible dual variable. With this Z , we obtain that $\phi_{R_g}(t) \geq t f^{-1}(1/t)$ for $t \leq t_0$. But this is just the fundamental function of the corresponding regret measure. Since it holds that $R_g(X) \leq V_g(X)$ due to the infimal convolution (this is also clear from the dual representations), we find that $\phi_{R_g}(t) = t f^{-1}(1/t)$ for $t \leq t_0$.

Due to translation equivariance it holds that $\phi_{R_g}(t) \leq 1 \forall t$ and $\phi_{R_g}(1) = 1$. Thus we obtain the statement $\phi_{R_g}(t) = \min\{1, t f^{-1}(1/t)\}$.

Finally, in the special case of the expectation, $R_g(X) = \mathbb{E}[X]$, where

$$f(x) = \begin{cases} 0 & 0 \leq x \leq 1 \\ \infty & x > 1, \end{cases} \quad (3.193)$$

we obtain $f^{-1}(1/t) = 1$ for $0 < t \leq 1$ and hence the formula holds and reduces to $\phi_E(t) = t$. ■

3.11.4 PROOF OF PROPOSITION 141

Proof The fundamental function of an Orlicz norm $\|\cdot\|_{\Psi}^O$ with conjugate Young function Φ is [Kosmol and Müller-Wichards, 2011, Lemma 7.4.4]:

$$\phi(t) = t\Phi^{-1}(1/t), \quad t \in (0, 1]. \quad (3.194)$$

Due to properties of the perspective transform, it is guaranteed that the resulting Φ^{-1} is concave and increasing on $[\Phi^{-1}(1), \infty)$: observe that

$$\Phi^{-1}(x) = x\phi(1/x), \quad x \geq 1. \quad (3.195)$$

If ϕ is concave, $-\phi$ is convex and then its perspective transform $(s, x) \mapsto x(-\phi)(s/x)$ is convex in both arguments (this is well known). By setting $s = 1$ we see that $x \mapsto -\Phi^{-1}(x)$ is convex, hence Φ^{-1} is concave.

To see that Φ^{-1} is increasing on $[\Phi^{-1}(1), \infty)$, note that $t \mapsto \phi(t)/t$ is decreasing due to quasiconcavity, hence $t \mapsto \Phi^{-1}(1/t)$ is decreasing and therefore $x \mapsto \Phi(x)$ is increasing.

The condition that ϕ is continuous at 0 is indeed necessary for Φ to be supercoercive, cf. 3.11.2.

Also, if $\phi'(0)$ is unbounded, Φ is finite and the behaviour of Φ is fully specified for all $x \geq \Phi^{-1}(1)$. If in contrast $\phi'(0) < \infty$, we have a bounded Φ^{-1} and can therefore identify the behaviour of Φ for $\Phi^{-1}(1) \leq x \leq b_\Phi$, where $b_\Phi = \sup\{\Phi < \infty\} = \phi'(0)$. Thus such Φ is infinite for all $x > b_\Phi$.

However, for $0 \leq x < \Phi^{-1}(1)$, we cannot identify the behaviour of f . Consequently, any left-continuous increasing convex “completion” of f in this interval will be compatible with the fundamental function. To see that at least one such choice of f must exist, take ϕ and extend it to the whole nonnegative domain by:

$$\phi^\dagger(t) = \begin{cases} \phi(t) & 0 \leq t \leq 1 \\ h(t) & t > 1, \end{cases} \quad (3.196)$$

in such a manner that $\phi^\dagger : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is concave; such a $\phi^\dagger$ obviously exists: for instance choose a linear continuation with slope $\Phi'(1)$, shifted by $\Phi(1)$, i.e. $h(t) = \Phi(1) + \Phi'(1)(t - 1)$

1). Then using the equality

$$t\Phi^{-1}(1/t) = \phi^\dagger(t), \quad (3.197)$$

we can construct an Orlicz norm with the desired properties. Repeating the above arguments then yields that Φ is a legitimate Young function, where we explicitly set $\Phi(0) = 0$. ■

3.12 APPENDIX: TAIL-SPECIFIC MARCINKIEWICZ AND LORENTZ NORMS

3.12.1 PROOF OF PROPOSITION 152

Proof From [Kreĭn et al., 1982b, Lemma 5.3, Theorem 5.3]¹⁷. we know that M1 to M3 are equivalent for the case of an infinite measure space, i.e. $P(\Omega) = \infty$. First assume that we knew that M1 to M3 are also equivalent for $P(\Omega) = 1$. Then, $M3 \implies M4$: let $\|t \mapsto \phi_{\mathcal{X}'}(t)/t\|_{M_\phi} < \infty$. Consequently:

$$\|t \mapsto \phi_{\mathcal{X}'}(t)/t\|_{M_\phi} = \|E_{\mathcal{X}}(t)\|_{M_\phi} = \sup_{0 < t \leq 1} \frac{1}{E_{\mathcal{X}}(t)} E_{\mathcal{X}}^{**}(t) < \infty \implies \exists K : E_{\mathcal{X}}^{**}(t) \leq K \cdot E_{\mathcal{X}}(t) \quad (3.198)$$

Also, as a general fact, the decreasing rearrangement is dominated by the maximal function, i.e. $E_{\mathcal{X}}(t) \leq E_{\mathcal{X}}^{**}(t)$, $\forall t \in (0, 1]$. Conversely, if $E_{\mathcal{X}}$ and $E_{\mathcal{X}}^{**}$ are equivalent, M5 holds [Rubshtein et al., 2016, p. 162]. Finally, if M5 holds, then M3 obviously holds by definition of $\|\cdot\|_{M_{\phi_{\text{CVar}}}}$:

$$\|E_{\mathcal{X}}\|_{M_\phi} \leq K \cdot \|E_{\mathcal{X}}\|_{M_{\phi_{\text{CVar}}}} = K \cdot \sup_{0 < t \leq 1} \frac{1}{E_{\mathcal{X}}^{**}(t)} E_{\mathcal{X}}^{**}(t) = 1 < \infty. \quad (3.199)$$

Observing that $\phi_{\mathcal{X}'}(t)/t = E_{\mathcal{X}}(t)$, the equivalence of M2 and M3, for the case of both $P(\Omega) = 1$ or $P(\Omega) = \infty$, can be obtained from Pick et al. [2013, p. 265]. In summary, we have shown that the implications form a circle.

We shall now show that M1 to M3 are also equivalent in the case of $P(\Omega) = 1$. Assume without loss of generality that $\phi_{\mathcal{X}}(1) = \phi_{\mathcal{X}'}(1) = 1$ (this corresponds to simply rescaling one norm and then using the associate relation of the fundamental functions). Consider

¹⁷To prevent confusion, we remark that Kreĭn et al. [1982b] use the symbol M_ϕ to denote what in our notation is $M_{\phi_{\mathcal{X}'}}$, that is, we index the Marcinkiewicz norm by its fundamental function, whereas Kreĭn et al. [1982b] index it by the fundamental function of its associate space.

the following extension of $\phi_{\mathcal{X}'} : [0, \infty) \rightarrow [0, \infty)$ to the whole nonnegative domain:

$$\tilde{\phi}_{\mathcal{X}'}(t) = \begin{cases} \phi_{\mathcal{X}'}(t) & 0 \leq t \leq 1 \\ 1 + \phi'_{\mathcal{X}'}(1)(t-1) & 1 < t < \infty, \end{cases} \quad (3.200)$$

which extends $\phi_{\mathcal{X}'}$ as a straight line and where the slope is chosen so as to guarantee concavity (assume without loss of generality that $\phi_{\mathcal{X}}$ is concave, otherwise first use the least concave majorant to concavify the quasi-concave $\phi_{\mathcal{X}'}$, thereby preserving equivalence of Marcinkiewicz norms). Let $\tilde{\phi}(t) = t/\tilde{\phi}_{\mathcal{X}'}(t)$. We now show the following:

Lemma 177. $|\cdot|_{M_\phi}$ and $\|\cdot\|_{M_\phi}$ are equivalent on $\Omega = [0, 1]$ if and only if $|\cdot|_{M_{\tilde{\phi}}}$ and $\|\cdot\|_{M_{\tilde{\phi}}}$ are equivalent on $\Omega = [0, \infty)$.

The Marcinkiewicz norm in the case of $P(\Omega) = \infty$ is defined in the obvious way, as

$$\|X\|_{M_{\tilde{\phi}}} := \sup_{0 < t < \infty} \{\tilde{\phi}(t)X^{**}(t)\} \quad (3.201)$$

and $X^* : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ and $X^{**} : (0, \infty) \rightarrow \mathbb{R}^+$ are defined in the obvious way.

Proof of Lemma: First assume that $|\cdot|_{M_{\tilde{\phi}}}$ and $\|\cdot\|_{M_{\tilde{\phi}}}$ are equivalent on $\Omega = [0, \infty)$. Then it is trivial that $|\cdot|_{M_\phi}$ and $\|\cdot\|_{M_\phi}$ are equivalent on $\Omega = [0, 1]$, by simply extending a random variable $X \in M_\phi$ from $\Omega = [0, 1]$ to $\Omega = [0, \infty)$ as $X^*(t) = 0$ for $t > 1$. For the converse direction, assume that $|\cdot|_{M_\phi}$ and $\|\cdot\|_{M_\phi}$ are equivalent on $\Omega = [0, 1]$. Then we want to show that $|\cdot|_{M_{\tilde{\phi}}}$ and $\|\cdot\|_{M_{\tilde{\phi}}}$ are equivalent by demonstrating that $t \mapsto \tilde{\phi}_{\mathcal{X}'}(t)/t = 1/\tilde{\phi}(t) \in M_{\tilde{\phi}}$, which is equivalent to the desired statement due to [Pick et al. \[2013, p. 265\]](#). Then:

$$\|t \mapsto \tilde{\phi}_{\mathcal{X}'}(t)/t\|_{M_{\tilde{\phi}}} = \sup_{0 < t < \infty} \tilde{\phi}(t) \frac{1}{t} \int_0^t 1/\tilde{\phi}_{\mathcal{X}'}(\omega) \, d\omega \stackrel{!}{<} \infty. \quad (3.202)$$

By assumption, we have that $t \mapsto \phi_{\mathcal{X}'}(t)/t \in M_\phi$ and therefore only have to investigate $t > 1$. But:

$$\star := \sup_{1 < t < \infty} \frac{t}{1 + \phi'_{\mathcal{X}'}(1)(t-1)} \frac{1}{t} \left(\int_0^1 1/\phi(\omega) \, d\omega + \int_1^t 1/\tilde{\phi}(\omega) \, d\omega \right). \quad (3.203)$$

Since $t \mapsto \phi_{\mathcal{X}'}(t)/t \in M_\phi$, we have $c := \int_0^1 1/\phi(\omega) \, d\omega \leq \sup_{0 < t \leq 1} \phi(t) \frac{1}{t} \int_0^t 1/\phi(\omega) \, d\omega <$

∞ ; essentially this follows from $\mathcal{L}^1$ being the largest of all r.i. spaces. Furthermore:

$$\star = \sup_{1 < t < \infty} \frac{t}{1 + \phi'_{\mathcal{X}'}(1)(t-1)} \frac{1}{t} \left(c + \int_1^t \frac{1 + \phi'_{\mathcal{X}'}(1)(\omega-1)}{\omega} d\omega \right) \quad (3.204)$$

$$= \sup_{1 < t < \infty} \frac{t}{1 + \phi'_{\mathcal{X}'}(1)(t-1)} \frac{1}{t} (c + \phi'_{\mathcal{X}'}(1)t - \phi'_{\mathcal{X}'}(1) \log(t) - \phi'_{\mathcal{X}'}(1) + \log(t)) \stackrel{!}{<} \infty. \quad (3.205)$$

$$(3.206)$$

To ensure that this is finite, we need to check only the limit as $t \rightarrow \infty$:

$$\lim_{t \rightarrow \infty} \frac{\phi'_{\mathcal{X}'}(1)t - \phi'_{\mathcal{X}'}(1) \log(t) - \phi'_{\mathcal{X}'}(1) + \log(t)}{1 + \phi'_{\mathcal{X}'}(1)(t-1)} = 1. \quad (3.207)$$

Thus we have that $t \mapsto \tilde{\phi}_{\mathcal{X}'}(t)/t = 1/\tilde{\phi} \in M_{\tilde{\phi}}$, which completes the proof of the lemma. Finally, we need to show the following:

Lemma 178. $\forall s > 1: \sup_{0 < t \leq 1} \frac{\phi_{\mathcal{X}'}(t/s)}{\phi_{\mathcal{X}'}(t)} < 1$ if and only if $\forall s > 1: \sup_{0 < t < \infty} \frac{\tilde{\phi}_{\mathcal{X}'}(t/s)}{\tilde{\phi}_{\mathcal{X}'}(t)} < 1$.

Proof of Lemma: That the condition on $\tilde{\phi}_{\mathcal{X}'}$ implies the condition of $\phi_{\mathcal{X}'}$ is obvious. Conversely, assume the condition holds for $\phi_{\mathcal{X}'}$ on $\Omega = [0, 1]$. Then we need to show only that it also holds for $\tilde{\phi}_{\mathcal{X}'}$ when $1 < t < \infty$, as they coincide for $0 \leq t \leq 1$. Let $s > 1$ be arbitrary. For any fixed $1 < t$, we obviously have that the fraction is < 1 and thus need to investigate only the limit as $t \rightarrow \infty$. Then:

$$\lim_{t \rightarrow \infty} \frac{1 + \phi'_{\mathcal{X}'}(1)(t/s-1)}{1 + \phi'_{\mathcal{X}'}(1)(t-1)} = \frac{1}{s} < 1. \quad (3.208)$$

Chaining all equivalences together, we obtain that M1 to M5 are equivalent for both finite and infinite measure spaces. ■

3.12.2 FUNDAMENTAL FUNCTIONS OF $L_{\exp}$ AND $L_{\log L}$

The fundamental function of $L_{\exp}$, given by $\phi_{L_{\exp}}(t) = 1/(1 - \log(t))$, is only quasiconcave, not concave. In contrast, the fundamental function $\phi_{L_{\log L}}(t) = t + t \log(1/t) = t/\phi_{L_{\exp}}(t)$ is concave; see Figure 3.2.

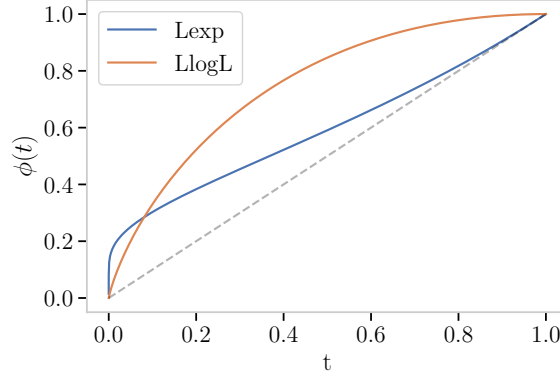


Figure 3.2: Fundamental functions of $L_{\exp}$ and $L_{\log L}$.

3.12.3 PROOF OF PROPOSITION 157

Assume Y^* is differentiable on $(0, 1)$. Then the condition that $\phi_{\mathcal{X}'}(t) = Y^*(t) \cdot t = F_X^{-1}(1 - t) \cdot t$ is increasing becomes

$$\frac{d}{dt}(F_X^{-1}(1 - t) \cdot t) > 0 \Leftrightarrow \frac{d}{ds}F_X^{-1}(s) \leq \frac{1}{1-s}F_X^{-1}(s) \quad \forall s \in (0, 1). \quad (3.209)$$

We use a Gronwall-type inequality.

Proposition 179 (adapted from [Bainov and Simeonov \[2013, Lemma 1.1\]](#)). *Let $f : [0, \infty)$ be continuous and differentiable and $g : [0, \infty)$ be continuous; assume $f(0) \leq f_0$ and*

$$\frac{d}{ds}f(s) \leq g(s)f(s), \quad \forall s \geq 0. \quad (3.210)$$

Then:

$$f(s) \leq f_0 \exp\left(\int_0^s g(r) \, dr\right) \quad \forall s \geq 0. \quad (3.211)$$

Applying this statement with $f(s) = F_X^{-1}(s)$, $f_0 := F_X^{-1}(0)$ and $g(s) = 1/(1 - s)$ and constraining the domain yields

$$F_X^{-1}(s) \leq F_X^{-1}(0) \exp\left(\int_0^s \frac{1}{1-r} \, dr\right) \quad \forall s \in (0, 1) \quad (3.212)$$

$$\Leftrightarrow F_X^{-1}(s) \leq F_X^{-1}(0) \exp(-\log(1 - s)) \quad \forall s \in (0, 1) \quad (3.213)$$

$$\Leftrightarrow F_X^{-1}(s) \leq F_X^{-1}(0) \frac{1}{1-s} \quad \forall s \in (0, 1) \quad (3.214)$$

$$\Leftrightarrow Y^*(t) \leq Y^*(1) \frac{1}{t} \quad \forall s \in (0, 1). \quad (3.215)$$

3.13 APPENDIX: UTILITY-BASED SHORTFALL RISK

Definition 180 (Föllmer and Schied [2002, 2016]). A convex risk measure on some space $\mathcal{X}$ is a functional $R : \mathcal{X} \rightarrow \mathbb{R} \cup \{\infty\}$, which is monotone (C3), translation equivariant (C4) and convex:

$$R(\alpha X + (1 - \alpha)Y) \leq \alpha R(X) + (1 - \alpha)R(Y). \quad (3.216)$$

Typically, also $R(0) = 0$ is demanded. Positive homogeneity and subadditivity together imply convexity; but convexity on its own is a weaker condition. Thus a coherent risk measure is also a convex risk measure.

In this section, we define a loss function $\ell : \mathbb{R} \rightarrow \mathbb{R}^+$ to be a function which is a Young function on $\mathbb{R}^+$ and further satisfies $\ell(x) = 0$ for $x \leq 0$. Consider the *acceptance set*:

$$\mathcal{A} = \{X : \mathbb{E}[\ell(X)] \leq x_0\}. \quad (3.217)$$

The intuition is that a decision maker who has a loss function and some decision threshold x_0 considers a loss variable X “acceptable” if $\mathbb{E}[\ell(X)] \leq x_0$. Subsequently, let $x_0 = 1$. We assumed $\ell(x) = 0$ for $x \leq 0$, which is non-standard, but allows for a cleaner connection to Orlicz space theory. Intuitively, this implies that acceptable random variables are characterized by their losses only, whereas gains (negative values) are neglected in the acceptance set. This also fits with our overall focus on right tails only. However, the trade-off between losses and gains naturally arises by enforcing translation equivariance. A convex risk measure, the UBSR, can be constructed as follows:

$$\text{UBSR}_\ell(X) = \inf\{m \in \mathbb{R} : X - m \in \mathcal{A}\}. \quad (3.218)$$

Hence, UBSR_ℓ specifies the smallest amount that must be subtracted from the risky position X so as to make it acceptable. This construction automatically guarantees translation equivariance and since ℓ is increasing and convex, UBSR_ℓ is a convex risk measure [Föllmer and Schied, 2002].

To study the envelope representation of UBSR_ℓ , one must choose a pair of Banach spaces. Föllmer and Schied [2002] studied the UBSR on $\mathcal{L}^\infty$, which is an overly conservative choice. In fact, the natural setting is the framework of Orlicz spaces, as shown by Biagini and Frittelli [2009] and Arai [2011]. We denote the conjugate of ℓ as ℓ^* . Observe that ℓ^* is also a legitimate Young function¹⁸ on $\mathbb{R}^+$, which is supercoercive¹⁹ and $\ell^*(x) = \infty$ for $x < 0$. To state the envelope representation of UBSR_ℓ on its natural space, we need to define the

¹⁸For this, our assumption that $\ell(y) = 0$ for $y \leq 0$ is important, then ℓ^* is increasing on $[0, \infty)$. Confer [Föllmer and Schied, 2002, Lemma 11].

¹⁹Supercoercive means that $\lim_{y \rightarrow \infty} \ell^*(y)/y = \infty$, cf. Definition 100 and see [Föllmer and Schied, 2002].

Orlicz class.

Definition 181. Given a Young function Φ , the Orlicz class P^Φ is defined as

$$P^\Phi := \{X \in \mathcal{M} : \mathbb{E}[\Phi(|X|)] < \infty\}. \quad (3.219)$$

By definition of the Luxemburg norm, we obviously have $P^\Phi \subseteq \mathcal{L}^\Phi$. We remark that P^Φ need not be a linear subspace of the Orlicz space $\mathcal{L}^\Phi$, but it does contain a linear subspace itself, called the Orlicz heart M^Φ . However, if Φ satisfies the Δ_2 condition [Kosmol and Müller-Wichards, 2011, p. 197, p. 234], then in fact $M^\Phi = P^\Phi = \mathcal{L}^\Phi$. Intuitively, such Φ cannot grow too fast. For instance, the rapidly growing $\Phi(x) = (\exp(x) - 1)\chi_{[1,\infty)}$ does not satisfy Δ_2 .

Indeed, the UBSR is finite on the Orlicz class P^ℓ (consider $m = 0$); here, understand ℓ as its restriction to $\mathbb{R}^+$, where it is a Young function.

Proposition 182 (Föllmer and Schied [2002], Biagini and Frittelli [2009]). The following envelope representation holds for the utility-based shortfall risk measure $\text{UBSR}_\ell : P^\ell \rightarrow \mathbb{R}$:

$$\text{UBSR}_\ell(X) = \sup\{\mathbb{E}[XZ] - \alpha(Z) : Z \in \mathcal{L}_+^{\ell^*}, Z \geq 0, \mathbb{E}[Z] = 1\}, \quad (3.220)$$

where the penalty function $\alpha : \mathcal{L}_+^{\ell^*} \rightarrow \mathbb{R}$ is:

$$\alpha(Z) = \sup\{\mathbb{E}[XZ] : X \in \mathcal{L}^\ell, X \in \mathcal{A}\} = \sup\{\mathbb{E}[XZ] : X \in \mathcal{L}^\ell, X \geq 0, \mathbb{E}[\ell(X)] \leq 1\}. \quad (3.221)$$

In fact, we observe that $\alpha(Z) = \|Z\|_{\ell^*}^O$ as $Z \geq 0$, i.e. the penalty function is an Orlicz norm, and therefore:

$$\text{UBSR}_\ell(X) = \sup\left\{\mathbb{E}[XZ] - \inf_{t>0} t(1 + \mathbb{E}[\ell^*(Z/t)]) : Z \in \mathcal{L}^{\ell^*}, Z \geq 0, \mathbb{E}[Z] = 1\right\}. \quad (3.222)$$

Proof We adapt Proposition 32 in [Biagini and Frittelli, 2009] to our loss-based orientation and to the additional assumption that $\ell(x) = 0$ for $x \leq 0$. By definition, $\alpha(Z) = \|Z\|_{\ell^*}^O$ is finite on $\mathcal{L}^{\ell^*}$, but $\alpha(Z)$ is only evaluated for $Z \geq 0$, hence we take $\mathcal{L}_+^{\ell^*}$ as its domain to emphasize this. The restriction to $X \geq 0$ in Equation 3.221 is possible since $Z \geq 0$ and $\ell(x) = 0$ for $x \leq 0$, hence a negative part of X is neglected. ■

Föllmer and Schied [2002], who established a variant Equation 3.222 in the $\mathcal{L}^\infty$ setting, remarked in passing the similarity to the Amemiya formulation of an Orlicz norm, but did not further go into it.

Among the convex risk measures, the utility-based shortfall risk measure has particularly desirable properties, such as invariance under randomization [Giesecke et al., 2008] and elicibility [Bellini and Bignozzi, 2015]. Most important in the present context, however,

is that since the penalty function is an Orlicz norm, we can immediately apply our method of Section 3.6 to achieve a desired tail sensitivity. In contrast to a coherent risk measure, a convex risk measure considers *all* alternative probability measures Q_Z from an entire space, but they are considered more or less relevant due to the penalty function; note that $Z \geq 0$ and $\mathbb{E}[Z] = 1$, hence the Q_Z induced via Equation 3.33 are valid probability measures. If the penalty function only takes on the values 0 and $+\infty$, the risk measure is coherent. In this sense, a convex risk measure acts in a smoother fashion than a coherent risk measure, where only a subset of probability measures is considered at all. By specifying the tail sensitivity of the penalty function $\|Z\|_{\ell^*}^O$, we specify how alternative probability measures, conceptualized as density ratios with respect to the base measure, are punished for their reweighting in a tail-sensitive way.

In the context of machine learning, the UBSR seems particularly useful to handle tail risk in reinforcement learning applications, as exemplified by [Shen et al. \[2014\]](#). More generally, any context which calls for a risk measure without positive homogeneity is a possible application for the UBSR.

Example 183. [[Föllmer and Schied, 2002](#)] Let $1 < p \leq \infty$ and $\ell(x) = (\frac{1}{p}x^p) \cdot \chi_{[0,\infty)}$. Then

$$\ell^*(y) = \begin{cases} \infty & y < 0 \\ (\frac{1}{q}y^q) & y \geq 0, \end{cases} \quad (3.223)$$

and

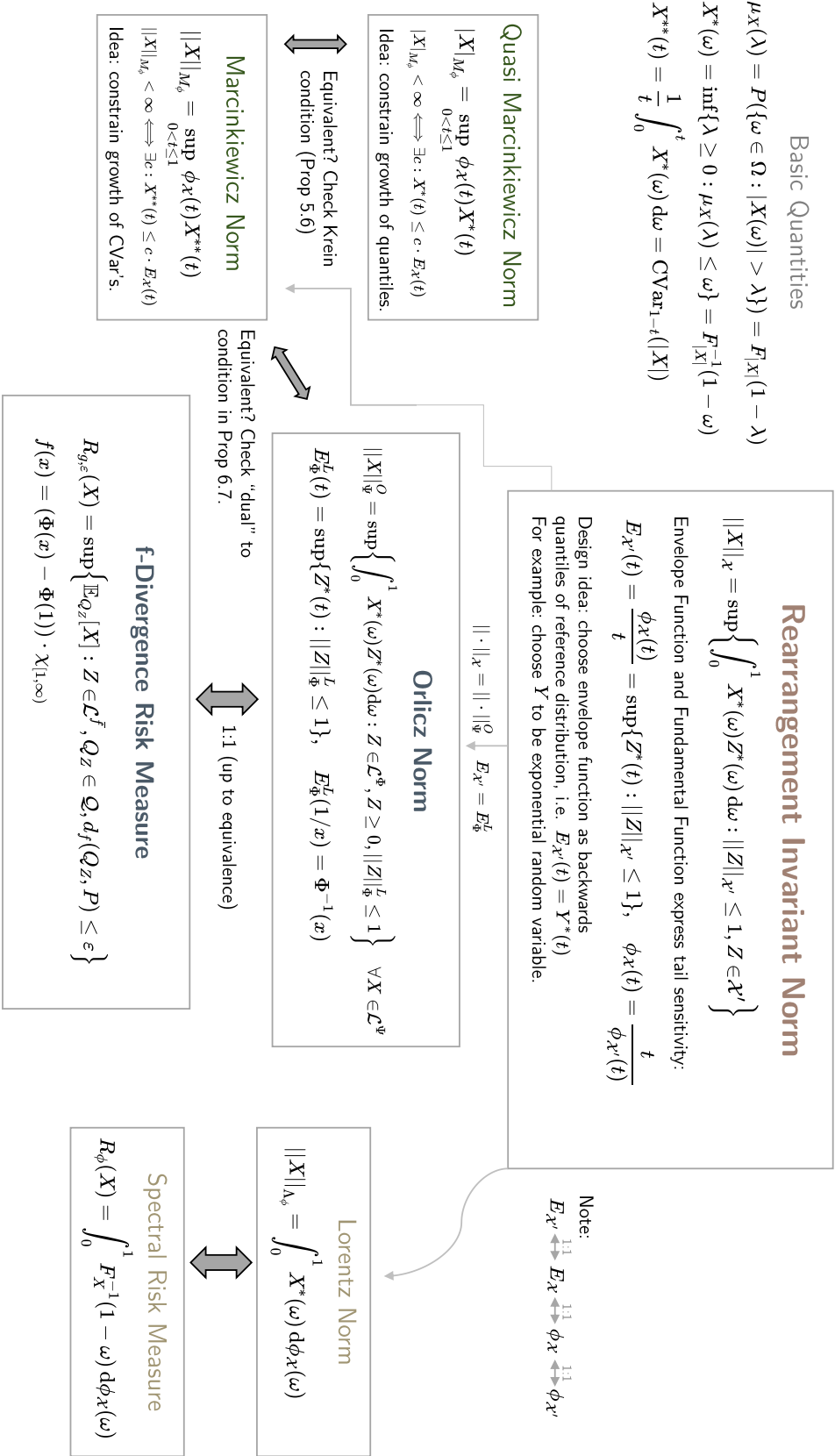
$$\alpha(Z) = p^{1/p} (\mathbb{E}[Z^q])^{1/q}. \quad (3.224)$$

Consider $p \rightarrow \infty$. Then $\ell^*(y) = i_{[0,1]}$ and therefore $\text{UBSR}_\ell(X) = \mathbb{E}[X]$, since any Z which has $Z(\omega) > 1$ for some ω is infinitely punished; this is just the $\mathcal{L}^1 - \mathcal{L}^\infty$ associate relationship.

3.14 APPENDIX: PLAIN LANGUAGE SUMMARY AND DIAGRAM

When training machine learning systems with data, engineers typically try to minimize average prediction error. Yet the justification for why the average is the best summary of those errors is weak: some errors could be very high, some very low. When a prediction error corresponds to a human individual (for example when we predict who gets a loan), this is problematic, as some individuals then suffer a much higher error than others. In such a situation, we would like to have a summary which also penalizes inequality of prediction errors. To this end, we can minimize a “risk measure” of the prediction errors. A risk measure is like an average, but puts more weight on the higher prediction errors. Many researchers have studied risk measures and proposed many different families of risk measures. But one important aspect, which has not been studied systematically yet, is how sensitive risk measures react to very large errors. This is called the tail sensitivity and it is the focus of our work. We study risk measures in a unified framework, which was proposed by mathematicians. In this framework, we find that there exists a simple way of characterizing how tail sensitive a risk measure is. An outcome of our work is a method to generate risk measures which have exactly a desired tail sensitivity. This can be used by a machine learning engineer to flexibly tune the influence of very high prediction errors.

Figure 3.3: Summary of key elements in the paper.



Part II

Aleatoric Plurality — Beyond Law Invariance

4

Strictly Frequentist Imprecise Probability

AUTHOR CONTRIBUTIONS

The following author contribution statement is approximately from [Fröhlich et al., 2023]: The comparison of Ivanenko's and Walley's theories was suggested by Robert Williamson (RW), who supervised the project and is responsible for the majority ($\approx 95\%$) of the proof of Theorem 192 (specifically, RW proved Theorem 194). Further, RW contributed the entirety of Appendix 4.10. Rabanur Derr contributed through many discussions and helpful suggestions and made minor contributions to writing. Christian Fröhlich is responsible for most of the content and writing.

The CRediT authorship contribution statement (taken from [Fröhlich et al., 2024]):

- Christian Fröhlich: Writing — review & editing, Writing — original draft, Methodology, Conceptualization.
- Rabanus Derr: Writing — review & editing, Methodology, Conceptualization.
- Robert C. Williamson: Writing — review & editing, Writing — original draft, Supervision, Methodology, Conceptualization.

ABSTRACT

Strict frequentism defines probability as the limiting relative frequency in an infinite sequence. What if the limit does not exist? We present a broader theory, which is applicable also to data that exhibit diverging relative frequencies. In doing so, we develop a close connection with the theory of imprecise probability: the cluster points of relative frequencies yield a coherent upper prevision. We show that a natural frequentist definition of conditional probability recovers the generalized Bayes rule. Finally, we prove constructively that, for a finite set of elementary events, there exists a sequence for which the cluster points of relative frequencies coincide with a prespecified set which demonstrates the naturalness, and arguably completeness, of our theory.

4.1 INTRODUCTION

Do other statistical properties, that can not be reduced to stochasticity, exist? This question did not attract any attention until the applications of the probability theory concerned only natural sciences. The situation is definitely [changing], when one studies social phenomena: the stochasticity gets broken as soon as we [deal] with deliberate activity of people. — Victor Ivanenko and Valery Labkovsky [1993]

It is now (almost surely) universally acknowledged that probability theory ought to be based on Kolmogorov's [1933] mathematical axiomatization (translated in [Kolmogorov, 1956]).¹ However, if probability is defined in this purely measure-theoretic fashion, what warrants its application to real-world problems of decision making under uncertainty? To those in the so-called *frequentist* camp, the justification is essentially due to the *law of large numbers*, which comes in both an empirical and a theoretical flavour. Our motivation for the present paper comes from questioning both of these.

By the empirical version of the law of large numbers (LLN), we mean not a “law” which can be proven to hold, but the following hypothesis, which seems to guide many scientific endeavours. Assume we have obtained data $x_1, \dots, x_n$ as the outcomes of some experiment, which has been performed n times under “statistically identical” conditions. Of course, conditions in the real-world can never truly be identical — otherwise the outcomes would be constant, at least under the assumption of a deterministic universe. Thus, “identical” in this context must be a weaker notion, that all factors which we have judged as relevant to the problem at hand have been kept constant over the repetitions.² The empirical “law” of large numbers, which Gorban [2017] calls the *hypothesis of (perfect) statistical stability* then asserts that in the long-run, relative frequencies of events and sample averages converge. These limits are then conceived of as the *probability* of an event and the *expectation*, respectively. Thus, even if relative frequencies can fluctuate in the finite data setting, we expect that they stabilize as more and more data is acquired. Crucially, this hypothesis of perfect statistical stability is not amenable to falsification, since we can never refute it in the finite data setting. It is a matter of faith to assume convergence of relative frequencies; confer [King and Kay, 2020].

On the other hand, there is now ample experimental evidence that relative frequencies can fail to stabilize even under very long observation intervals [Gorban, 2017, Part II]. We say that such phenomena display *unstable (diverging)* relative frequencies. Rather than refuting the stability hypothesis, which is impossible, we question its adequateness as an idealized

¹An important exception is quantum probability [Gudder, 1979, Khrennikov, 2016].

²In fact, we do not need that conditions stay exactly constant, but that they change merely in a way which is so benign that the relative frequencies converge. That is, in the limit we should obtain a stable statistical aggregate.

modeling assumption: we view convergence as the idealization of approximate stability in the finite case, whereas divergence idealizes instability. Thus, if probability is understood as limiting relative frequency, then the applicability of Kolmogorov’s theory to empirical phenomena is limited to those which are statistically stable; the founder himself remarked:

Generally speaking there is no ground to believe that a random phenomenon should possess any definite probability [Kolmogorov, 1983].

Building on the works of von Mises and Geiringer [1964], Walley and Fine [1982] and Ivaenko [2010], our goal is to establish a broader theory, which is also applicable to phenomena which are outside of the scope of Kolmogorov’s theory by exhibiting unstable relative frequencies.

One attempt to “prove” (or justify) the empirical law of large numbers, which in our view is doomed to fail, is to invoke the theoretical law of large numbers, which is a purely formal, mathematical statement. The strong law of large numbers states that if $X_1, X_2, \dots$ is a sequence of independent and identically distributed (i.i.d.) random variables with finite expectation $\mathbb{E}[X] := \mathbb{E}[X_1] = \mathbb{E}[X_2] = \dots$, then the sample average $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$ converges almost surely to the expectation:

$$P\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mathbb{E}[X]\right) = 1,$$

where P is the underlying probability measure in the sense of Kolmogorov. To interpret this statement correctly, some care is needed. It asserts that P assigns measure 1 to the set of sequences for which the sample mean converges, but not that this happens *for all* sequences. Thus one would need justification for identifying “set of measure 0” with “is negligible” (“certainly does not happen”), which in particular requires a justification for P . With respect to a different measure, this set might not be negligible at all [Schnorr, 2007, p. 8]; see also [Calude and Zamfirescu, 1999, Seidenfeld et al., 2017] for critical arguments. Moreover, the examples in [Gorban, 2017, Part II] show that sequences with seemingly non-converging relative frequencies (fluctuating substantially even for long observation intervals) are not “rare” in practice. In Appendix 4.10 we examine the question of how pathological or normal such sequences are in more depth.

Conceptually, the underlying problem is that the probability measure P , which is used to measure the event $\{\lim_{n \rightarrow \infty} \bar{X}_n = \mathbb{E}[X]\}$ has no clear meaning. Of course, in the subjectivist spirit, one could interpret it as assigning a *belief* in the statement that convergence takes place. But it is unclear what a frequentist interpretation of P would look like. As La Caze [2016] observed:

Importantly, almost sure convergence is also given a frequentist interpretation. Almost sure convergence is taken to provide a justification for assuming that

the relative frequency of an attribute *would* converge to the probability in actual experiments *were* the experiment to be repeated indefinitely [emphasis in original].

But again, it is unclear on what ground P can be given this interpretation and according to Hájek [2009] this leads to a regress to mysterious “meta-probabilities”. Furthermore, the theoretical LLN requires that P be countably additive, which is problematic under a frequency interpretation [Hájek, 2009, pp. 229–230].

Given these complications, we opt for a different approach, namely a *strictly frequentist* one. Reaching back to Richard von Mises’ [1919] foundational work, a strictly frequentist theory explicitly defines probability in terms of limiting relative frequencies in a sequence. Importantly, we here *do not assume* that the elements of the sequence are random variables with respect to an abstract, countably additive probability measure. Instead, like von Mises, we actually take the notion of a sequence as the primitive entity in the theory. As a consequence, countable additivity does not naturally arise in this setting, and hence we do not subscribe to the frequentist interpretation of the classical strong LLN.

The core motivation for our work is to drop the assumption of perfect statistical stability and instead to explicitly model the possibility of unstable (diverging) relative frequencies. Rather than merely conceding that the “probability” might vary over time [Borel, 1963, pp. 27ff.] (which begs the question what such “probabilities” mean) we follow the approach of Ivanenko [2010], reformulate his construction of a *statistical regularity* of a sequence, and discover that it is closely connected to the subjectivist theory of *imprecise probability*, more specifically, to the theory of lower and upper previsions. In essence, to each sequence we can naturally associate a *set* of probability measures, which constitute the statistical regularity that describes the cluster points of relative frequencies and consequently also those of sample averages. Since this works for *any* sequence and *any* event, we have thus countered a typical argument against frequentism, namely that the limit may not exist and hence probability is undefined [Hájek, 2009]. On an arbitrary (possibly infinite) set of outcomes, the relative frequencies induce a coherent upper probability and the sample averages induce a coherent upper prevision in the sense of Walley [1991]. In the convergent case, this reduces to a precise, finitely additive probability and a linear prevision, respectively. Furthermore, we derive in a natural way a conditional upper prevision; remarkably, this approach recovers the *generalized Bayes rule*, the arguably most important updating principle in imprecise probability.

Furthermore, we demonstrate that (for a finite set of outcomes) the reverse direction works, too: given a set of probability measures, we can explicitly construct a sequence, which corresponds to this set in the sense that its relative frequencies have this set of cluster points. Thereby we establish strictly frequentist *semantics* for imprecise probability: a subjective

decision maker who uses a set of probability measures to represent their belief can also be understood as assuming an implicit underlying sequence and reasoning in a frequentist way thereon.

We emphasize that we leave the question of randomness in this broader picture for future research (see the discussion in Section 4.6). As described in Section 4.1.1, von Mises [1919] modelled randomness with what he called *selection rules*. In the present paper, we disregard the role of selection rules and instead direct our attention towards von Mises' axiom vM1 of converging relative frequencies. Thus our work should only partially be viewed as a generalization of von Mises' approach. Note that throughout the paper we work with a loss-based orientation (see Section 4.1.2).

4.1.1 VON MISES — THE FREQUENTIST PERSPECTIVE

Our approach is inspired by Richard von Mises' [1919] axiomatization of probability theory (refined and summarized in [von Mises and Geiringer, 1964]). In contrast to the subjectivist camp, von Mises' concern was to develop a theory for repetitive events; which gives rise to a theory of probability that is mathematical, but which can also be used to reason about the physical world.

The calculus of probability, i.e. the theory of probabilities, in so far as they are numerically representable, is the theory of definite observable phenomena, repetitive or mass events. Examples are found in games of chance, population statistics, Brownian motion etc. [von Mises, 1981, p. 102].

Hence, von Mises is not concerned with the probability of single events, which he deems meaningless, but instead always views an event as part of a larger *reference class*. Such a reference class is captured by what he terms a *collective*, a disorderly sequence which exhibits both global regularity and local irregularity. For the definition of a collective, we need a possibility set Ω of elementary outcomes $\omega \in \Omega$, together with a set system of events $\mathcal{A} \subseteq 2^\Omega$.

Definition 184. Consider a tuple $(\Omega, \vec{\Omega}, \mathcal{A}, \mathcal{S})$ with the following data:

1. a sequence $\vec{\Omega}: \mathbb{N} \rightarrow \Omega$;
2. a set of selection rules $\mathcal{S} := \{\vec{S}_j: j \in \mathcal{J}\}$, where for each j in a countable index set $\mathcal{J}$, $\vec{S}_j: \mathbb{N} \rightarrow \{0, 1\}$ and $\vec{S}_j(i) = 1$ for infinitely many $i \in \mathbb{N}$;
3. a non-empty set system $\mathcal{A} \subseteq 2^\Omega$, where for simplicity we assume $|\mathcal{A}| < \infty$.³

This tuple forms a *collective* if the following two axioms hold.

³In fact, $\mathcal{A}$ does not necessarily has to be finite. Since an infinite domain of probabilities does not contribute a lot to a better understanding of the frequentist definition at this point, we restrict ourselves to the finite case here. The reader can find details in [von Mises and Geiringer, 1964].

vM1. The limiting relative frequency for $A \in \mathcal{A} \subseteq 2^\Omega$ exists:

$$P(A) := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \chi_A(\vec{\Omega}(i)).^4$$

We call this limit the probability of A .

vM2. For each $j \in \mathcal{J}$, the selection rule $\vec{S}_j$ does not change limiting relative frequencies:⁵

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \chi_A(\vec{\Omega}(i)) \cdot \vec{S}_j(i)}{\sum_{i=1}^n \vec{S}_j(i)} = P(A) \quad \forall A \in \mathcal{A}.$$

Here, we view $\vec{\Omega}$ as a sequence of elementary outcomes $\omega \in \Omega$, for some possibility space Ω on which we have a set system of events $\mathcal{A}$. Axiom vM1 explicitly defines the probability of an event in terms of the limit of its relative frequency. Demanding that this limit exists is non-trivial, since this need not be the case for an arbitrary sequence. Intuitively, vM1 expresses the hypothesis of statistical stability, which captures a global regularity of the sequence.

In contrast, vM2 captures a sense of *randomness* or local irregularity. Note it actually comprises *two claims*: 1) the limit *exists* and 2) it is the *same* as the limit in vM1. It is best understood by viewing a selection rule $\vec{S}_j: \mathbb{N} \rightarrow \{0, 1\}$ as selecting a subsequence of the original sequence $\vec{\Omega}$ and then demanding that the limiting relative frequencies thereof coincide with those of the original sequence. Such a selection rule is called *admissible*, whereas a selection rule which would give rise to different limiting relative frequencies for at least one $A \in \mathcal{A}$ would be *inadmissible*. Why do we need axiom vM2? Von Mises calls this the “law of the excluded gambling system” and it is the key to capture the notion of randomness in his framework. Intuitively, if a selection rule is inadmissible, an adversary could use this knowledge to strategically offer a bet on the next outcome and thereby make long-run profit, at the expense of our fictional decision maker. A *random* sequence, however, is one for which there does not exist such a betting strategy. It turns out, that this statement cannot hold in its totality. A sequence cannot be random with respect to *all* selection rules except in trivial cases (cf. Kamke’s critique of von Mises’ notion of randomness, nicely summarized in [van Lambalgen, 1987]). Thus, von Mises explicitly relativizes randomness with respect to a problem-specific set of selection rules [von Mises and Geiringer, 1964, p. 12].⁶ A sequence which forms a collective (“is random with respect to”) one set of selection

⁴The function χ_A denotes the indicator gamble for a set $A \subseteq \Omega$, i.e. $\chi_A(\omega) := 1$ if $\omega \in A$ and $\chi_A(\omega) := 0$ otherwise.

⁵To be precise, a selection rule in the sense of von Mises is a map from the set of finite Ω -valued strings to $\{0, 1\}$, i.e. a selection rule is able to “see” all previous elements when deciding whether or not to select the next one. Our formulation is more restrictive to avoid notational overhead, but when a sequence is fixed, the two formulations are equivalent.

⁶This class of selection rules necessarily must be specified in advance; confer [Shen, 2009]. A prominent

rules, might not form a collective with respect to another set.

In our view, the role of the randomness axiom vM2 is similar to the role of more familiar randomness assumptions like the standard *i.i.d.* assumption: to empower inference from finite data. In this work, however, we will be exclusively concerned with the idealized case of infinite data, since our focus is the axiom (or hypothesis) of statistical stability. In this spirit, we put aside the question of randomness here.

We are motivated by the following question. What happens to von Mises approach when axiom vM1 breaks down? That is, when relative frequencies of at least some events do not converge. Our answer leads to a confluence with a theory that is thoroughly grounded in the subjectivist camp: the theory of *imprecise probability*. In summary, we establish a strictly frequentist theory of imprecise probability.

4.1.2 IMPRECISE PROBABILITY — THE SUBJECTIVIST PERSPECTIVE

We briefly introduce the *prima facie* unrelated, subjectivist theory of imprecise probability, or more specifically, the theory of *lower and upper previsions* as put forward by Walley [1991]. Orthodox Bayesianism models belief via the assignment of precise probabilities to propositions, or equivalently, via a linear expectation functional. In contrast, in Walley’s theory, belief is interval-valued and the linear expectation is replaced by a pair of a lower and upper expectations. Hence, the theory is strictly more expressive than orthodox Bayesianism, which can be recovered as a special case.

We assume an underlying possibility set Ω , where $\omega \in \Omega$ is an elementary event, which includes all relevant information. We call a function $X: \Omega \rightarrow \mathbb{R}$, which is bounded, i.e. $\sup_{\omega \in \Omega} |X(\omega)| < \infty$, a *gamble* and collect all such functions in the set L^∞ . The set of gambles L^∞ carries a vector space structure with scalar multiplication $(\lambda X)(\omega) = \lambda X(\omega)$, $\lambda \in \mathbb{R}$, and addition $(X + Y)(\omega) = X(\omega) + Y(\omega)$. For a constant gamble $c(\omega) = c \forall \omega$ we write simply c . Note that Walley’s theory in the general case does not require that a vector space of gambles is given, but definitions and results simplify significantly in this case.

We interpret a gamble as assigning an uncertain loss $X(\omega)$ to each elementary event, that is, in line with the convention in machine learning, we take positive values to represent loss and negative values to represent reward.⁷ We imagine a decision maker who is faced with the question of how to value a gamble X ; the orthodox answer would be the expectation $\mathbb{E}[X]$ with respect to a subjective probability measure.

Walley [1991] proposed a betting interpretation of imprecise probability, which is inspired

line of work aspires to fix the set of selection rules as all partially computable selection rules [Church, 1940], but there is no compelling reason to elevate this to a universal choice; cf. [Derr and Williamson, 2022] for an elaborated critique.

⁷Unfortunately, this introduces tedious sign flips when comparing results to Walley [1991].

by [de Finetti \[1974/2017\]](#), who identifies probability with fair betting rates. The goal is to axiomatize a functional $\bar{R}: L^\infty \rightarrow \mathbb{R}$, which assigns to a gamble the smallest number $\bar{R}(X)$ so that $X - \bar{R}(X)$ is a desirable transaction to our decision maker, where she incurs the loss $X(\omega)$ but in exchange gets the reward $-\bar{R}(X)$. Formally:

$$\bar{R}(X) := \inf\{\alpha \in \mathbb{R} : X - \alpha \in \mathcal{D}\},$$

where $\mathcal{D}$ is a set of desirable gambles. [Walley \[1991, Section 2.5\]](#) argued for a criterion of coherence, which any reasonable functional $\bar{R}$ should satisfy, and consequently obtained the following characterization [[Walley, 1991, Theorem 2.5.5](#)], which we shall take here as an axiomatic definition instead.⁸

Definition 185. A functional $\bar{R}: L^\infty \rightarrow \mathbb{R}$ is a coherent upper prevision if it satisfies $\forall X, Y \in L^\infty$:

- UP1. $\bar{R}(X) \leq \sup(X)$ (bounded)
- UP2. $\bar{R}(\lambda X) = \lambda \bar{R}(X), \forall \lambda \in \mathbb{R}^+$ (positive homogeneity)
- UP3. $\bar{R}(X + Y) \leq \bar{R}(X) + \bar{R}(Y)$ (subadditivity)

Together, these properties also imply $\forall X, Y \in L^\infty$ [[Walley, 1991, p. 76](#)]:

- UP4. $\bar{R}(X + c) = \bar{R}(X) + c, \forall c \in \mathbb{R}$ (translation equivariance)
- UP5. $X(\omega) \leq Y(\omega) \forall \omega \in \Omega \Rightarrow \bar{R}(X) \leq \bar{R}(Y)$ (monotonicity)

To a coherent upper prevision, we can define its conjugate lower prevision by:

$$\begin{aligned} \underline{R}(X) &:= -\bar{R}(-X) \\ &= -\inf\{\alpha \in \mathbb{R} : -X - \alpha \in \mathcal{D}\} \\ &= \sup\{\alpha \in \mathbb{R} : \alpha - X \in \mathcal{D}\}, \end{aligned}$$

which specifies the highest certain loss α that the decision maker is willing to shoulder in exchange for giving away the loss $X(\omega)$, i.e. receiving the reward $-X(\omega)$. Due to the conjugacy, it suffices to focus on the upper prevision throughout. In general, we have that $\underline{R}(X) \leq \bar{R}(X)$ for any $X \in L^\infty$. If $\underline{R}(X) = \bar{R}(X) \forall X \in L^\infty$, we say that $R := \bar{R} = \underline{R}$ is a linear prevision, a definition which aligns with [de Finetti \[1974/2017\]](#).

By applying an upper prevision to indicator gambles, we obtain an upper probability $\bar{P}(A) := \bar{R}(\chi_A)$, where $A \subseteq \Omega$. Correspondingly, the lower probability is $\underline{P}(A) := 1 - \bar{P}(A^c) = \underline{R}(\chi_A)$. In the precise case, there is a unique relationship between (finitely) additive probabilities and linear previsions; however, upper previsions are more expressive than upper probabilities. Finally, we remark that via the so-called *natural extension*, a coherent upper

⁸Here, we need the vector space assumption on the set of gambles. We also note that [Walley \[1991, pp. 64–65\]](#) himself made a similar definition, but then proposed the more general coherence concept.

probability which is defined on some subsets of events $\mathcal{A} \subseteq 2^\Omega$ can be extended to a coherent upper prevision $\text{NatExt}(\bar{P})$ on L^∞ , which is compatible with $\bar{P}$ in the sense that $\text{NatExt}(\bar{P})(\chi_A) = \bar{P}(A) \forall A \in \mathcal{A}$ (cf. [Walley, 1991, Section 3.1]).

4.2 UNSTABLE RELATIVE FREQUENCIES

Assume that we have some fixed sequence $\vec{\Omega}: \mathbb{N} \rightarrow \Omega$ on a possibility set Ω of elementary events, but that for some events $A \in \mathcal{A}$, where $\mathcal{A} \subseteq 2^\Omega$, the limiting relative frequencies do not exist. What can we do then? In a series of papers [Ivanenko and Labkovskii, 1986a,b, 1990, 1993, Ivanenko and Munier, 2000, Ivanenko and Labkovskii, 2015, Ivanenko and Pasichnichenko, 2017] and a monograph [Ivanenko, 2010], Ivanenko and collaborators have developed a strictly frequentist theory of “hyper-random phenomena” based on “statistical regularities”. In essence, they tackle mass decision making in the context of sequences with possibly divergent relative frequencies. Like von Mises, they take the notion of a sequence as the primitive, that is, without assuming an a priori probability and then invoking the law of large numbers. They explicitly recognize that “stochasticity gets broken as soon as we deal with deliberate activity of people” [Ivanenko and Labkovskii, 1993]⁹. The presentation of Ivanenko’s theory is obscured somewhat by the great generality with which it is presented (they work with general nets, rather than just sequences). We build heavily upon their work but entirely restrict ourselves to working with sequences. While in some sense this is a weakening, our converse result (see Section 4.3) is actually stronger as we show that one can achieve any “statistical regularity” by taking relative frequencies of only sequences. For simplicity, we will dispense with integrals with respect to finitely additive measures in our presentation, so that there are less mathematical dependencies involved;¹⁰ instead, we work with linear previsions. Moreover, we establish¹¹ the connection to imprecise probability and give a different justification for the construction. The contribution in this section may be viewed as unifying ideas from Ivanenko with Walley’s framework.

4.2.1 IVANENKO’S ARGUMENT — INFORMALLY

We begin by providing an informal summary of Ivanenko’s construction of *statistical regularities* on sequences. Assume we are given a fixed sequence $\vec{\Omega}: \mathbb{N} \rightarrow \Omega$ of elementary

⁹This not only occurs because of non-equilibrium effects, but also from feedback loops, what has become known as “performativity” [MacKenzie et al., 2008] or “reflexivity” [Soros, 2009]. See the epigraph at the beginning of the present paper.

¹⁰The two well-known accounts of the theory of integrals with finitely additive measures are [Rao and Rao, 1983] and [Dunford and Schwartz, 1988]. The theory of linear previsions as laid out in [Walley, 1991] appears to be an easier approach for our purposes.

¹¹Ivanenko and Labkovskii [2015] mention in passing that sets of probabilities also appear in [Walley, 1991].

events $\vec{\Omega}(1), \vec{\Omega}(2), \dots$, where we may intuitively think of $\mathbb{N}$ as representing time. In contrast to von Mises, who demands the existence of relative frequency limits to define probabilities, we ask for something like a probability for *all* events $A \subseteq \Omega$, even when the relative frequencies have no limit. To this end, we exploit that sequences of relative frequencies always have a non-empty set of cluster points, each of which is a finitely additive probability. Hence, a decision maker can use this set of probabilities to represent the global statistical properties of the sequence. In fact, we will see that this induces a coherent upper probability. Also, our decision maker is interested in assessing a value for each gamble $X: \Omega \rightarrow \mathbb{R}$, which is evaluated infinitely often over time. Here, the sequence of averages $n \mapsto \frac{1}{n} \sum_{i=1}^n X(\vec{\Omega}(i))$ is the object of interest. In the case of convergent relative frequencies, a decision maker would use the expectation to assess the risk in the limit, whereas in the general case of possible non-convergence, a different object is needed. This object turns out to be a coherent upper prevision. We provide a justification for using this upper prevision to assess the value of a gamble, which links it to imprecise probability.

4.2.2 IVANENKO'S ARGUMENT — FORMALLY

Let Ω be an arbitrary (finite, countably infinite or uncountably infinite) set of outcomes and fix $\vec{\Omega}: \mathbb{N} \rightarrow \Omega$, an Ω -valued sequence. We define a gamble $X: \Omega \rightarrow \mathbb{R}$ as a bounded function from Ω to $\mathbb{R}$, i.e. $\exists K \in \mathbb{R}: |X(\omega)| \leq K \forall \omega \in \Omega$ and collect all such gambles in the set L^∞ . We assume the vector space structure on L^∞ as in Section 4.1.2.

The set L^∞ becomes a Banach space, i.e. a complete normed vector space, under the supremum norm $\|X\|_{L^\infty} := \sup_{\omega \in \Omega} |X(\omega)|$. We denote the topological dual space of L^∞ by $(L^\infty)^*$. Recall that it consists exactly of the continuous linear functionals $E: L^\infty \rightarrow \mathbb{R}$. We endow $(L^\infty)^*$ with the *weak*-topology*, which is the weakest topology (i.e. with the fewest open sets) that makes all evaluation functionals of the form $X^*: (L^\infty)^* \rightarrow \mathbb{R}, X^*(E) := E(X)$ for any $X \in L^\infty$ and $E \in (L^\infty)^*$ continuous. Consider the following subset of $(L^\infty)^*$:

$$\text{PF}(\Omega) := \{E \in (L^\infty)^*: E(X) \geq 0 \text{ whenever } X \geq 0, E(\chi_\Omega) = 1\}.$$

Due to the Alaoglu-Bourbaki theorem, this set is compact under the weak* topology, see Appendix 4.7.1.

A finitely additive probability $P: \mathcal{A} \rightarrow [0, 1]$ on some set system $\mathcal{A}$, where $\Omega \in \mathcal{A}$, is a function such that:

PF1. $P(\Omega) = 1$.

PF2. $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$ and $A, B \in \mathcal{A}$.

We induce a sequence of finitely additive probabilities $\vec{P}$ where $\vec{P}(n) := A \mapsto \frac{1}{n} \sum_{i=1}^n \chi_A(\vec{\Omega}(i))$ for each $n \in \mathbb{N}$. It is easy to check that indeed $\vec{P}(n)$ is a finitely additive probability on

the whole powerset 2^Ω for any $n \in \mathbb{N}$. We shall call $\vec{P}$ the *sequence of empirical probabilities*. Due to [Walley, 1991, Corollary 3.2.3], a finitely additive probability defined on 2^Ω can be uniquely extended (via natural extension) to a linear prevision $E_P: L^\infty \rightarrow \mathbb{R}$, so that $E_P(\chi_A) = P(A) \forall A \subseteq \Omega$. Furthermore, we know from [Walley, 1991, Corollary 2.8.5], that there is a one-to-one correspondence between elements of $\text{PF}(\Omega)$ and linear previsions $E_P: L^\infty \rightarrow \mathbb{R}$. Hence, we associate to each empirical probability $\vec{P}(n)$ an *empirical linear prevision* $\vec{E}(n) := X \mapsto \text{NatExt}(\vec{P}(n))(X)$, where $X \in L^\infty$ and we denote the natural extension by NatExt . We thus obtain a sequence $\vec{E}: \mathbb{N} \rightarrow \text{PF}(\Omega)$.

On the other hand, each gamble $X \in L^\infty$ induces a sequence of evaluations as $\vec{X}: \mathbb{N} \rightarrow \mathbb{R}$, where $\vec{X}(n) := X(\vec{\Omega}(n))$. For $X \in L^\infty$, we define the sequence of averages of the gamble over time as $\vec{\Sigma X}: \mathbb{N} \rightarrow \mathbb{R}$, where $\vec{\Sigma X}(n) := \frac{1}{n} \sum_{i=1}^n X(\vec{\Omega}(i))$. For each finite n , we can also view the average as a function in X , i.e. $X \mapsto \frac{1}{n} \sum_{i=1}^n X(\vec{\Omega}(i))$. Observe that this is a coherent linear prevision and by applying it to indicator gambles χ_A , we obtain $\vec{P}(n)$. Hence, we know from [Walley, 1991, Corollary 3.2.3] that this linear prevision is in fact the natural extension of $\vec{P}(n)$, i.e. $\vec{E}(n) = X \mapsto \frac{1}{n} \sum_{i=1}^n X(\vec{\Omega}(i)) = X \mapsto \vec{\Sigma X}(n)$. This concludes the technical setup; we now begin reproducing Ivanenko's argument.

Since $\text{PF}(\Omega)$ is a compact topological space under the subspace topology induced by the weak*-topology on $(L^\infty)^*$, we know that any sequence $\vec{E}: \mathbb{N} \rightarrow (L^\infty)^*$ has a non-empty closed set of cluster points. Recall that a point z is a *cluster point* of a sequence $\vec{S}: \mathbb{N} \rightarrow \mathcal{T}$, where $\mathcal{T}$ is any topological space, if:

$\forall N$, where N is any neighbourhood of z with respect to $\mathcal{T}$, $\forall n_0 \in \mathbb{N} : \exists n \geq n_0 : \vec{S}(n) \in N$.

We remark that this *does not* imply that those cluster points are limits of convergent subsequences.¹² We denote the *set of cluster points* as $\text{CP}(\vec{E})$. Equivalently, by applying these linear previsions to indicator gambles, we obtain the set of finitely additive probabilities $\mathcal{P} := \left\{ A \mapsto E(\chi_A) : E \in \text{CP}(\vec{E}) \right\}$. Due to the one-to-one relationship, we might work with either $\text{CP}(\vec{E})$ or $\mathcal{P}$. Following Ivanenko, we call $\mathcal{P}$ the *statistical regularity* of the sequence $\vec{\Omega}$; in the language of imprecise probability, it is a *credal set*.

We further define

$$\bar{R}(X) := \sup \left\{ E(X) : E \in \text{CP}(\vec{E}) \right\} = \sup \left\{ E_P(X) : P \in \mathcal{P} \right\}, \quad X \in L^\infty,$$

where $E_P := \text{NatExt}(P)$, and

$$\bar{P}(A) := \sup \left\{ E(\chi_A) : E \in \text{CP}(\vec{E}) \right\} = \sup \left\{ P(A) : P \in \mathcal{P} \right\}, \quad A \subseteq \Omega.$$

¹²This would hold under sequential compactness, which is not fulfilled here in general, but it is for finite Ω .

Observe that $\bar{R}$ is defined on all $X \in L^\infty$ and $\bar{P}$ is defined on *all* subsets of Ω , even if Ω is uncountably infinite, since each $P \in \mathcal{P}$ is a finitely additive probability on 2^Ω . We further observe that $\bar{R}$ is a coherent upper prevision on L^∞ or equivalently, a coherent risk measure in the sense of Artzner et al. [1999].¹³ Correspondingly, $\bar{P}$ is a coherent upper probability on 2^Ω , which is obtained by applying $\bar{R}$ to indicator functions. This follows directly from the *envelope theorem* in [Walley, 1991, Theorem 3.3.3].

So far, the definition of $\bar{R}$ and $\bar{P}$ may seem unmotivated. Yet they play a special role, as we now show.

Proposition 186. *The sequence of averages $\vec{\Sigma X}$ has the set of cluster points*

$$\text{CP}(\vec{\Sigma X}) = \{E(X) : E \in \text{CP}(\vec{E})\} = \{E_P(X) : P \in \mathcal{P}\},$$

and therefore

$$\bar{R}(X) = \sup \text{CP}(\vec{\Sigma X}) = \limsup_{n \rightarrow \infty} \vec{\Sigma X}(n).$$

Proof First observe that

$$\vec{E}(n)(X) = \vec{\Sigma X}(n).$$

We use the following result from [Ivanenko and Pasichnichenko, 2017, Lemma 3].¹⁴

Lemma 187. *Let $f : Y \rightarrow \mathbb{R}$ be a continuous function on a compact space Y and $\vec{y}$ a Y -valued sequence. Then $\text{CP}(n \mapsto f(\vec{y}(n))) = f(\text{CP}(\vec{y}))$.*

On the right side, the application of f is to be understood as applying f to each element in the set $\text{CP}(\vec{y})$. Consider now the evaluation functional $X^* : \text{PF}(\Omega) \rightarrow \mathbb{R}$, $X^*(E) := E(X)$, which is continuous under the weak*-topology. The application of the lemma with $f = X^*$, $Y = \text{PF}(\Omega)$, $\vec{y} = \vec{E}$ hence gives:

$$\text{CP}(n \mapsto X^*(\vec{E}(n))) = X^*(\text{CP}(\vec{E})).$$

But since $X^*(\vec{E}(n)) = \vec{\Sigma X}(n)$, we obtain that

$$\text{CP}(\vec{\Sigma X}) = X^*(\text{CP}(\vec{E})) = \{E(X) : E \in \text{CP}(\vec{E})\}.$$

■

¹³For the close connection of coherent upper previsions and coherent risk measures we refer to [Fröhlich and Williamson, 2024a].

¹⁴A subtle point in the argument, which Ivanenko and Pasichnichenko [2017] do not make visible, is the sequential compactness of $\mathbb{R}$, which means that for any cluster point of an $\mathbb{R}$ -valued sequence we can find a subsequence converging to it.

A similar statement holds for the coherent upper probability.

Corollary 188. For any $A \subseteq \Omega$, $\bar{P}(A) = \limsup_{n \rightarrow \infty} \left(\vec{P}(n)(A) \right) = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \chi_A \left(\vec{\Omega}(i) \right)$.

Proof Just observe that $\vec{P}(n)(A) = \overrightarrow{\Sigma \chi_A}(n)$ and apply the previous result. ■

Thus the limes superior of the sequence of relative frequencies induces a coherent upper probability on 2^Ω ; similarly, the limes superior of the sequence of a gamble's averages induces a coherent upper prevision on L^∞ . By conjugacy, we have that the lower prevision and lower probability are :

$$\begin{aligned} \underline{R}(X) &= \inf \left\{ E(X) : E \in \text{CP} \left(\vec{E} \right) \right\} = \liminf_{n \rightarrow \infty} \overrightarrow{\Sigma X}(n), \quad \forall X \in L^\infty, \\ \underline{P}(A) &= \inf \{ P(A) : P \in \mathcal{P} \} = \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \chi_A \left(\vec{\Omega}(i) \right), \quad A \subseteq \Omega, \end{aligned}$$

which are obtained in a similar way using the limes inferior. Finally, when an event is *precise* in the sense that $\bar{P}(A) = \underline{P}(A)$ (and thus the lim inf equals the lim sup and hence the limit exists), we denote the upper (lower) probability as $P(A)$ and say that the precise probability of A exists.

Example 189. von Mises and Geiringer [1964, p. 11] considered a binary sequence ($\Omega = \{0, 1\}$) given by

$$\vec{\Omega} = \left\langle 0^{[1]}, 1^{[1]}, 0^{[2]}, 1^{[2]}, 0^{[4]}, 1^{[4]}, \dots, 0^{[2^i]}, 1^{[2^i]}, \dots \right\rangle,$$

for $i \rightarrow \infty$. The notation $i^{[j]}$ here means j repetitions of i and $\langle \cdot \rangle$ forms a finite or infinite sequence. It is a straightforward calculation to show that the induced relative frequencies of ones have all elements of $\left[\frac{1}{3}, \frac{1}{2} \right]$ as cluster points.

Example 190. Let $\Omega = \{0, 1, 2, 3\}$ and

$$\vec{\Omega} = \left\langle \langle 3, 0 \rangle^{[1]}, \langle 1, 2 \rangle^{[1]}, \langle 3, 0 \rangle^{[2]}, \langle 1, 2 \rangle^{[2]}, \langle 3, 0 \rangle^{[4]}, \langle 1, 2 \rangle^{[4]}, \dots, \langle 3, 0 \rangle^{[2^i]}, \langle 1, 2 \rangle^{[2^i]}, \dots \right\rangle,$$

for $i \rightarrow \infty$. Similarly, the notation $\langle a, b \rangle^{[j]}$ means j repetitions of the tuple $\langle a, b \rangle$, e.g. $\langle a, b \rangle^{[2]} = a, b, a, b$. It is easy to see that the events $\{0, 2\}$ and $\{1, 3\}$ each have the limiting relative frequency 0.5. But for the elementary events the cluster points of the relative frequencies are: $\left[\frac{1}{4}, \frac{1}{3} \right]$ for $\omega = 3$ and $\omega = 0$; $\left[\frac{1}{6}, \frac{1}{4} \right]$ for $\omega = 1$ and $\omega = 2$. This illustrates that the set system of events for which precise probabilities exist need not form a *field*, but in fact a *pre-Dynkin system* (see Derr and Williamson [2023]).

Example 191. The arising of finite additivity is not a feature of the divergence, but is due to the strictly frequentist setting. Consider for instance $\Omega = \mathbb{N}$, $\vec{\Omega}(i) = i$. Then the relative

frequencies for each elementary converge to 0. Hence $\bar{P}(\{\omega\}) = \underline{P}(\{\omega\}) = 0 \forall \omega \in \Omega$ and $\bar{P} = \underline{P}$ is a finitely additive probability, but $\bar{P}(\Omega) = 1$ in violation of countable additivity.

4.2.3 THE INDUCED UPPER PREVISION

We have seen that the upper prevision $\bar{R}$, as we have just defined it, has the property that it is induced by the statistical regularity of the sequence, and at the same time corresponds to taking the supremum over the cluster points of the sequence of averages of a gamble over time. The set of cluster points is in general a complicated object, hence it is unclear why one should take the supremum to reduce it to a single number in a decision making context. Our goal in this section is to provide some intuition *why* it is reasonable to use $\bar{R}$, as we defined it, to assess the risk inherent in a sequence. [Ivanenko and Munier \[2000\]](#) argued that $\bar{R}$ is the unique object which satisfies a certain list of axioms, which are similar to those for an upper prevision, but including a so-called “principle of guaranteed result”, which appears rather mysterious to us.

Our setup is as follows. We imagine an individual decision maker, who is faced with a fixed sequence $\vec{\Omega}: \mathbb{N} \rightarrow \Omega$ and various gambles $X: \Omega \rightarrow \mathbb{R}$. The question to the decision maker is how to value this gamble in light of the sequence, i.e. imagining that the gamble is evaluated at each $\vec{\Omega}(1), \vec{\Omega}(2), \dots$, infinitely often. Here, $X(\vec{\Omega}(i))$ represents a loss for positive values, and a gain for negative values. We can think of the $\vec{\Omega}(i)$ as the states of nature, and the sequence determines which are realized and how often. Importantly, we view our decision maker as facing a *mass decision*, i.e. the gamble will not only be evaluated once, but instead infinitely often.

In the imprecise probability literature, a key concept for decision making is *desirability*. A coherent set of almost-desirable gambles is a set $\mathcal{D} \subset L^\infty$, satisfying the following properties [[Walley, 1991](#), p. 152]:

- D1.** If $\inf(X) > 0$ then $X \notin \mathcal{D}$ (avoiding sure loss)
- D2.** If $\sup(X) < 0$ then $X \in \mathcal{D}$ (accepting sure gains)
- D3.** If $X \in \mathcal{D}$ and $\lambda > 0$, then $\lambda X \in \mathcal{D}$ (positive homogeneity)
- D4.** If $X, Y \in \mathcal{D}$ then $X + Y \in \mathcal{D}$ (addition)
- D5.** If $X - \varepsilon \in \mathcal{D}$ for any $\varepsilon > 0$, then $X \in \mathcal{D}$ (closure).

Almost-desirable gambles are those which the decision maker would accept even without any reward for it. We remark the existence of a subtle controversy about the role of 0 that distinguishes “almost” from “strict” desirability [[Walley, 1991](#), Section 3.7], see also [[Couso and Moral, 2011](#)]. In this paper, we focus on almost-desirability, and from now on drop writing “almost”. Which gambles are *desirable* to our decision maker? We argue that

an appropriate set of desirable gambles is given by:

$$\mathcal{D}_{\vec{\Omega}} := \left\{ X \in L^\infty : \limsup_{n \rightarrow \infty} \vec{\Sigma} X \leq 0 \right\}.$$

It is easy to check that this satisfies D1–D5. Consider what $X \in \mathcal{D}_{\vec{\Omega}}$ means. If the limes superior of the gamble’s average sequence, i.e. the growth rate of the accumulated loss, is negative or zero, then we are guaranteed that there is no strictly positive accumulated loss which we will face infinitely often. The choice of the average as the aggregation functional is justified from the mass decision character of the setting, since we assume that our decision maker does not care about individual outcomes, but only about long-run capital. Now, given this set of desirable gambles, we seek a functional $\bar{R}(X): L^\infty \rightarrow \mathbb{R}$, so that when at each time step i , the transaction $X(\vec{\Omega}(i)) - \bar{R}(X)$ takes place, this results in a desirable gamble for our decision maker. Our decision maker shoulders the loss $X(\vec{\Omega}(i))$, while at the same time asking for $-\bar{R}(X)$ in advance. Intuitively, $\bar{R}(X)$ is supposed to be the certainty equivalent of the “uncertain” loss X , in the sense that $X(\vec{\Omega}(i))$ will vary over time. Therefore we define, in correspondence with $\mathcal{D}_{\vec{\Omega}}$, the upper and lower previsions ($\forall X \in L^\infty$):

$$\begin{aligned} \bar{R}(X) &:= \inf \{ \alpha \in \mathbb{R} : X - \alpha \in \mathcal{D}_{\vec{\Omega}} \} \\ \underline{R}(X) &:= -\bar{R}(-X) = \sup \{ \alpha \in \mathbb{R} : \alpha - X \in \mathcal{D}_{\vec{\Omega}} \}. \end{aligned} \tag{4.1}$$

When a set of desirable gambles and an upper (lower) prevision are in this correspondence, it holds that $X - \bar{R}(X) \in \mathcal{D}_{\vec{\Omega}}$ and furthermore $\bar{R}(X)$ is the least (smallest) functional for which this holds. We can now observe that in fact $\bar{R}(X) = \limsup_{n \rightarrow \infty} \vec{\Sigma} X(n)$, since

$$\mathcal{D}_{\vec{\Omega}} = \{ X \in L^\infty : \bar{R}(X) \leq 0 \}$$

is the general correspondence for a set of desirable gambles and a coherent upper prevision. Hence, we have motivated the definition of $\bar{R}(X)$ in Section 4.2.2. It is easy to see explicitly (cf. Appendix 4.7.2) that $X - \bar{R}(X) \in \mathcal{D}_{\vec{\Omega}}$ and that $\bar{R}(X) = \limsup_{n \rightarrow \infty} \vec{\Sigma} X(n)$ is in fact the smallest number such that the relation in Equation 4.1 holds.

4.3 FROM CLUSTER POINTS TO SEQUENCE

In the previous section, we have shown how from a given sequence we can construct a coherent upper prevision from the set of cluster points $\text{CP}(\vec{E})$. In this section, we show the converse, thus “closing the loop”: given an arbitrary coherent upper prevision, we construct a sequence $\vec{\Omega}$ such that the induced upper prevision is just the specified one. We take this to be an argument for the well-groundedness of our approach. For simplicity, we

assume a finite possibility space Ω .

Theorem 192. *Let $|\Omega| < \infty$. Let $\bar{R}$ be a coherent upper prevision on L^∞ . There exists a sequence $\vec{\Omega}$ such that we can write $\bar{R}$ as:*

$$\bar{R}(X) = \bar{R}_{\vec{\Omega}}(X) := \sup \{E(X) : E \in \mathcal{E}_{\vec{\Omega}}\}, \quad \mathcal{E}_{\vec{\Omega}} := \text{CP}(\vec{E}_{\vec{\Omega}}) \quad \forall X \in L^\infty,$$

where we now make the dependence on the sequence $\vec{\Omega}$ explicit in the notation, i.e. $\vec{E}_{\vec{\Omega}}(n) := X \mapsto \frac{1}{n} \sum_{i=1}^n X(\vec{\Omega}(i))$.

The proof is constructive and deterministically yields a sequence. The significance of this result is that it establishes strictly frequentist *semantics* for imprecise probability. It shows that to any decision maker who, in the subjectivist fashion, uses a coherent upper prevision, we can associate a sequence, which would yield the same upper prevision in a strictly frequentist way. We interpret this result as evidence for the naturalness, and arguably completeness, of our theory.

The key to prove this is Theorem 194, for which we introduce some convenient notation. For $k \in \mathbb{N}$, let $[k] := \{1, \dots, k\}$ and define the $(k-1)$ -simplex as

$$\Delta^k := \left\{ d = (d_1, \dots, d_k) \in \mathbb{R}^k : \sum_{i=1}^k d_i = 1, d_i \geq 0 \forall i \in [k] \right\}.$$

It is also helpful to have a dual notation for sequences $\vec{x} : \mathbb{N} \rightarrow [k]$, whereby we write either $\vec{x}(i)$ or $\vec{x}_i$ to mean the same thing.

Definition 193. *Suppose $k \in \mathbb{N}$ and $\vec{x} : \mathbb{N} \rightarrow [k]$. For any $n \in \mathbb{N}$ define the relative frequency of $\vec{x}$ with respect to i at n , $\vec{r}_i^{\vec{x}} : \mathbb{N} \rightarrow [0, 1]$ via*

$$\vec{r}_i^{\vec{x}}(n) := \frac{1}{n} |\{j \in [n] : \vec{x}_j = i\}|$$

and the relative frequency of $\vec{x}$ at n , $\vec{r}^{\vec{x}} : \mathbb{N} \rightarrow \Delta^k$ as

$$\vec{r}^{\vec{x}}(n) := \vec{r}_{[k]}^{\vec{x}}(n) = \begin{pmatrix} \vec{r}_1^{\vec{x}}(n) \\ \vdots \\ \vec{r}_k^{\vec{x}}(n) \end{pmatrix}. \quad (4.2)$$

Theorem 194. *Suppose $k \in \mathbb{N}$ and C is a rectifiable closed curve in Δ^k . There exists $\vec{x} : \mathbb{N} \rightarrow [k]$ such that $\text{CP}(\vec{r}^{\vec{x}}) = C$.*

The proof (which is constructive) is in Appendix 4.8 along with an example. From this, we obtain the following Corollary (proven in Appendix 4.8.10). Denote the topological boundary of a set D as ∂D .

Corollary 195. Suppose $k \in \mathbb{N}$ and $D \subseteq \Delta^k$ is a non-empty convex set. There exists $\vec{x}: \mathbb{N} \rightarrow [k]$ such that $\text{CP}(\vec{r}^{\vec{x}}) = \partial D$.

Since we have a finite possibility space Ω , we can identify each linear prevision E with a point in the simplex, by assigning coordinates to its underlying finitely additive probability; in the case of $\vec{E}_{\vec{\Omega}}(n)$, this is the relative frequency $\vec{r}^{\vec{\Omega}}(n)$. This is formalized in the following.

Proposition 196. Let $\vec{E}(n): \mathbb{N} \rightarrow \text{PF}(\Omega)$ be a sequence of linear previsions with underlying probabilities $\vec{P}(n) := A \mapsto \vec{E}(n)(A)$. Then $E \in \text{CP}(\vec{E}(n))$ with respect to the weak* topology if and only if the sequence $\vec{D}: \mathbb{N} \rightarrow \Delta^k$, $\vec{D}(n) := (\vec{P}(n)(\omega_1), \dots, \vec{P}(n)(\omega_k))$ has as cluster point $d_E = (E(\chi_{\{\omega_1\}}), \dots, E(\chi_{\{\omega_k\}}))$ with respect to the topology induced by the Euclidean norm on $\mathbb{R}^k$.

For the proof, see Appendix 4.8.1. Here, the assumption of finite Ω is crucial; we leave it to future research to (if possible) extend Theorem 192 to infinite Ω . Combining Corollary 195 and Proposition 196 allows us to now prove Theorem 192.

Proof [Proof of Theorem 192] Let $\Omega = [k]$. If $\bar{R}$ is a coherent upper prevision on L^∞ , we can write it as [Walley, 1991, Theorem 3.6.1]:

$$\bar{R}(X) = \sup \{E(X): E \in \mathcal{E}\}, \quad \forall X \in L^\infty,$$

for some weak* compact and convex set $\mathcal{E} \subseteq \text{PF}(\Omega)$. From [Walley, 1991, Theorem 3.6.2] we further know that

$$\bar{R}(X) = \sup \{E(X): E \in \mathcal{E}\} = \sup \{E(X): E \in \text{ext } \mathcal{E}\}, \quad \forall X \in L^\infty,$$

where ext denotes the set of extreme points of $\mathcal{E}$.¹⁵ Then:

$$\begin{aligned} \bar{R}(X) &= \sup \{E(X): E \in \mathcal{E}\} \\ &= \sup \{E(X): E \in \text{ext } \mathcal{E}\} \\ &\leq \sup \{E(X): E \in \partial \mathcal{E}\} \\ &\leq \sup \{E(X): E \in \mathcal{E}\} = \bar{R}(X), \end{aligned}$$

since $\text{ext } \mathcal{E} \subseteq \partial \mathcal{E}$ and $\partial \mathcal{E} \subseteq \mathcal{E}$; note that $\mathcal{E}$ is closed. In summary, $\bar{R}(X) = \sup \{E(X): E \in \partial \mathcal{E}\}$.

Now choose $D := \{(E(\chi_{\omega_1}), \dots, E(\chi_{\omega_k})) : E \in \mathcal{E}\}$, which is a non-empty convex set in Δ^k . We then obtain from Corollary 195 a sequence $\vec{\Omega}: \mathbb{N} \rightarrow [k]$ with $\text{CP}(\vec{r}^{\vec{\Omega}}) = \partial D$. But then it follows from Proposition 196 that the sequence $\vec{E}_{\vec{\Omega}}$ has cluster points $\text{CP}(\vec{E}_{\vec{\Omega}}) = \partial \mathcal{E}$.

¹⁵A point $E \in \mathcal{E}$ is an extreme point of $\mathcal{E}$ if it cannot be written as a convex combination of any other elements in $\mathcal{E}$.

Thus

$$\bar{R}(X) = \sup \left\{ E(X) : E \in \text{CP} \left(\vec{E}_{\vec{\Omega}} \right) \right\}, \quad \forall X \in L^\infty,$$

which concludes the proof. ■

Ivanenko [2010] offers a somewhat similar result to Theorem 192 by generalizing from sequences to *sampling nets*. Ivanenko's [2010] main result states that "any sampling directedness has a regularity, and any regularity is the regularity of some sampling directedness." [Ivanenko, 2010, Theorem 4.2]. Our result is more parsimonious in the sense that it relies only on sequences, which are arguably more intuitive objects than such sampling nets.

Our result should also be compared to Theorem 4.2 in [Walley and Fine, 1982] and Theorem 2.2 in [Paparmarcou and Fine, 1991b]. On the one hand, our result is stronger since it holds for upper previsions, whereas Theorem 4.2 in [Walley and Fine, 1982] and Theorem 2.2 in [Paparmarcou and Fine, 1991b] hold for upper probabilities only; note that upper previsions are more expressive than upper probabilities.¹⁶ On the other hand, Theorem 2.2 in [Paparmarcou and Fine, 1991b] is stronger in the sense that it guarantees that the same upper probability is induced when applying selection rules.

We observe that two sequences $\vec{\Omega}_1, \vec{\Omega}_2$, might have different sets of cluster points $\text{CP} \left(\vec{E}_{\vec{\Omega}_1} \right)$, $\text{CP} \left(\vec{E}_{\vec{\Omega}_2} \right)$, but when their convex hull coincides, the same upper probability and prevision is induced.¹⁷ Thus, in light of the argument in Section 4.2.3, *for the purpose of mass decision making*, we may consider these sequences equivalent. While in the classical case, relative frequencies are the relevant description of a sequence, the statistical regularity provides an analogous description in the general case; moreover, we differentiate only "up to the same convex hull" for decision making.

4.4 UNSTABLE CONDITIONAL PROBABILITY

An interesting aspect of the strictly frequentist approach is that there is a natural way of introducing conditional probability for events $A, B \subseteq \Omega$, which is the same for the case of converging or diverging relative frequencies. Furthermore, this approach generalizes directly to gambles. We will observe that this, perhaps surprisingly, yields the *generalized Bayes rule*. In the precise case, the standard Bayes rule is recovered.

Recall that for a countably or finitely additive probability Q , we can define conditional

¹⁶Indeed, the proof of Theorem 4.2 in [Walley and Fine, 1982] exploits this simplification by assuming that the credal set has a finite number of extreme points.

¹⁷Assume $\bar{R}(X) := \sup \{E(X) : E \in \mathcal{E}\}$. Then indeed $\bar{R}(X) = \sup \{E(X) : E \in \overline{\text{co}}\mathcal{E}\}$, where $\overline{\text{co}}$ denotes the weak* closure of the convex hull; cf. [Walley, 1991, Section 3.6].

probability as:

$$Q(A|B) := \frac{Q(A \cap B)}{Q(B)}, \quad A, B \subseteq \Omega, \text{ if } Q(B) > 0. \quad (4.3)$$

Important here is the condition that $Q(B) > 0$. Conditioning on events of measure zero may create trouble. Kolmogorov then allows the conditional probability to be arbitrary. This is rather unfortunate, as there arguably are settings where one would like to condition on events of measure zero.

As a prerequisite, given a linear prevision $E \in \text{PF}(\Omega)$, we define the conditional linear prevision as:

$$E(X|B) := \frac{E(\chi_B X)}{E(\chi_B)}, \quad \text{if } E(\chi_B) > 0, B \subseteq \Omega. \quad (4.4)$$

The application to indicator gambles then recovers conditional probability. As long as $E(\chi_B) > 0$, it is insignificant whether we condition the linear prevision, or instead condition on the level of its underlying probability and then naturally extend it; confer [Walley, 1991, Corollary 3.2.3].

Nearly in line with Kolmogorov's conditional probability, von Mises started from the following intuitive, frequentist view: the probability of an event A conditioned on an event B is the frequency of the occurrence of the event A given that B happens. In what follows, we build upon this idea, which von Mises called "partition operation" [von Mises and Geiringer, 1964, p. 22]. Walley and Fine [1982, Section 4.3] have extended this definition to the divergent case of conditional probability on a finite possibility space; we further extend it to conditional upper previsions on arbitrary possibility spaces and link them to the generalized Bayes rule. As a technical preliminary, we define a wrapper function $\Psi: \text{PF}(\Omega) \cup \{\perp\} \rightarrow \text{PF}(\Omega)$ as:

$$\Psi(P) := \begin{cases} P_0 & \text{if } P = \perp, \\ P & \text{otherwise,} \end{cases}$$

where P_0 is an arbitrary finitely additive probability on 2^Ω , and $\perp$ represents "undefined".

4.4.1 CONDITIONAL PROBABILITY

Recall our sequence of unconditional finitely additive probabilities $\vec{P}(n) := A \mapsto \frac{1}{n} \sum_{i=1}^n \chi_A(\vec{\Omega}(i))$. We want to define a similar sequence of *conditional* finitely additive probabilities. A very natural approach is the following: let $A, B \subseteq \Omega$ be such that $\vec{\Omega}(i) \in B$ for at least one $i \in \mathbb{N}$. We write $2_{1+}^{\vec{\Omega}}$ for the set of such events, i.e. events which occur at least once in the sequence.

Define a sequence of conditional probabilities $\vec{P}(\cdot|B): \mathbb{N} \rightarrow \text{PF}(\Omega)$ by

$$\vec{P}(\cdot|B)(n) := \Psi \left(A \mapsto \frac{\sum_{i=1}^n (\chi_A \cdot \chi_B) (\vec{\Omega}(i))}{\sum_{i=1}^n \chi_B (\vec{\Omega}(i))} \right), \quad (4.5)$$

where we consider only those $\vec{\Omega}(i)$ which lie in B , and hence we adapt the relative frequencies to the occurrence of B . Informally, this is simply counting $|A \text{ and } B \text{ occurred}| / |B \text{ occurred}|$. Until B occurs for the first time, the denominator will be 0 and thus the mapping undefined (returning the falsum $\perp$). Throughout, we demand that the event B on which we condition is in $2_{1+}^{\vec{\Omega}}$, i.e. occurs *at least once* in the sequence. Note that this is a much weaker condition than demanding that $P(B) > 0$, if B is precise. Denote by n_B the smallest index so that $\vec{\Omega}(n_B) \in B$. Note that $\vec{P}(A|B)(n) = \vec{P}(n)(A \cap B) / \vec{P}(n)(B)$ for $n \geq n_B$.

Proposition 197. *Assume $B \in 2_{1+}^{\vec{\Omega}}$. Then $\vec{P}(\cdot|B)$ is a sequence of finitely additive probabilities.*

Proof For $n < n_B$, this is clear due to Ψ . Now let $n \geq n_B$.

PF1: $\vec{P}(\Omega|B)(n) = 1$: obvious.

PF2: If $A, C \subseteq \Omega$, $A \cap C = \emptyset$, then we show that $\vec{P}(A \cup C|B)(n) = \vec{P}(A|B)(n) + \vec{P}(C|B)(n)$.

$$\begin{aligned} \vec{P}(A \cup C|B)(n) &= \frac{\sum_{i=1}^n (\chi_{A \cup C} \cdot \chi_B) (\vec{\Omega}(i))}{\sum_{i=1}^n \chi_B (\vec{\Omega}(i))} \\ &= \frac{\sum_{i=1}^n (\chi_A \cdot \chi_B) (\vec{\Omega}(i)) + \sum_{i=1}^n (\chi_C \cdot \chi_B) (\vec{\Omega}(i))}{\sum_{i=1}^n \chi_B (\vec{\Omega}(i))} \\ &= \vec{P}(n)(A \cap B) / \vec{P}(n)(B) + \vec{P}(n)(C \cap B) / \vec{P}(n)(B) \\ &= \vec{P}(A|B)(n) + \vec{P}(C|B)(n). \end{aligned}$$

Noting that since A and C are disjoint, $\vec{\Omega}(i)$ cannot lie in both at the same time for any i . ■

Even though the probability is conditional, we deal with a sequence of finitely additive probabilities again. Hence, we can now essentially repeat the argument from Section 4.2.2. To each $\vec{P}(\cdot|B)(n)$, associate its uniquely corresponding linear prevision $\vec{E}(\cdot|B)(n)$, which is of course given by ($\forall X \in L^\infty$, $n \geq n_B$):

$$\vec{E}(\cdot|B)(n) = \vec{\Sigma X|B}(n) := X \mapsto \frac{\sum_{i=1}^n (X \cdot \chi_B) (\vec{\Omega}(i))}{\sum_{i=1}^n \chi_B (\vec{\Omega}(i))}.$$

It is easy to check that $\vec{E}(\cdot|B)(n)$ is coherent. For $n < n_B$, set $\vec{E}(\cdot|B)(n) = \text{NatExt}(P_0)$. From the weak* compactness of $\text{PF}(\Omega)$, we obtain a non-empty closed set of cluster points

$\text{CP}(\vec{E}(\cdot|B))$.

Definition 198. If $B \in 2_{1+}^{\vec{\Omega}}$, we define the conditional upper prevision and the conditional upper probability as:

$$\bar{R}(X|B) := \sup \left\{ \tilde{E}(X) : \tilde{E} \in \text{CP} \left(\vec{E}(\cdot|B) \right) \right\} \quad (4.6)$$

$$\bar{P}(A|B) := \sup \left\{ Q(A) : Q \in \text{CP} \left(\vec{P}(\cdot|B) \right) \right\}, \quad A \subseteq \Omega. \quad (4.7)$$

Since they are expressed via an envelope representation,¹⁸ $\bar{R}$ and $\bar{P}$ are automatically coherent [Walley, 1991, Theorem 3.3.3]. By similar reasoning as in Section 4.2.2, we get the following representation.

Proposition 199. Assume $B \in 2_{1+}^{\vec{\Omega}}$. The conditional upper prevision (probability) can be represented as:

$$\bar{R}(X|B) = \limsup_{n \rightarrow \infty} \bar{\Sigma} \vec{X}|B(n), \quad X \in L^\infty; \quad \bar{P}(A|B) = \limsup_{n \rightarrow \infty} \vec{P}(A|B)(n), \quad A \subseteq \Omega.$$

Also, we obtain the corresponding lower quantities $\bar{R}(X|B) = \liminf_{n \rightarrow \infty} \bar{\Sigma} \vec{X}|B(n)$ and $\underline{P}(A|B) = \liminf_{n \rightarrow \infty} \vec{P}(A|B)(n)$. Note that these definitions also have reasonable frequentist semantics even when B occurs only finitely often; then the sequence $\vec{P}(\cdot|B)$ is eventually constant and we have $\vec{P}(A|B) = |A \text{ and } B \text{ occurred}| / |B \text{ occurred}|$. For instance, if A and B occur just once, but simultaneously, then $\bar{P}(A|B) = \underline{P}(A|B) = 1$. This is an advantage over Kolmogorov's approach, where conditioning on events of measure zero is not meaningfully defined.

We now further analyze the conditional upper probability and the conditional upper prevision. As a warm-up, we consider the case of precise probabilities. If for some event $A \subseteq \Omega$, we have $\bar{P}(A|B) = \underline{P}(A|B)$, we write $\tilde{P}(A|B) := \lim_{n \rightarrow \infty} \vec{P}(A|B)(n)$.

Proposition 200. Assume $\vec{\Omega}$ is such that $P(B), P(A \cap B)$ exist for some $A, B \subseteq \Omega$ and $P(B) > 0$, that is, relative frequencies converge to those values. Then it holds that $\tilde{P}(A|B) = P(A|B)$, where $P(\cdot|B)$ is the conditional probability in the sense of Equation 4.3.

¹⁸An envelope representation expresses a coherent upper prevision as a supremum over a set of linear previsions.

Proof

$$\begin{aligned}
\tilde{P}(A|B) &= \lim_{n \rightarrow \infty} \vec{P}(A|B)(n) \\
&= \lim_{n \rightarrow \infty} \frac{\frac{1}{n} \sum_{i=1}^n (\chi_A \cdot \chi_B) \left(\vec{\Omega}(i) \right)}{\frac{1}{n} \sum_{i=1}^n \chi_B \left(\vec{\Omega}(i) \right)} \\
&\stackrel{(1)}{=} \frac{\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\chi_A \cdot \chi_B) \left(\vec{\Omega}(i) \right)}{\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \chi_B \left(\vec{\Omega}(i) \right)} \\
&= \frac{P(A \cap B)}{P(B)} \\
&\stackrel{(2)}{=} P(A|B).
\end{aligned}$$

(1) The limits exist by assumption and the denominator is > 0 .

(2) In the sense of Equation 4.3. ■

Thus, when the relative frequencies of B and $A \cap B$ converge, we reproduce the classical definition of conditional probability. Now what happens under non-convergence?

4.4.2 THE GENERALIZED BAYES RULE

We now relax the assumptions of Proposition 200 and only demand that $\underline{P}(B) > 0$.¹⁹ Then we observe that the conditional upper prevision coincides with the *generalized Bayes rule*, which is an important updating principle in imprecise probability (see e.g. [Miranda and Cooman, 2014]). The unconditional set of desirable gambles is:

$$\mathcal{D}_{\vec{\Omega}} := \left\{ X \in L^\infty : \limsup_{n \rightarrow \infty} \overrightarrow{\Sigma X} \leq 0 \right\} = \left\{ X \in L^\infty : \overline{R}(X) \leq 0 \right\}.$$

Definition 201. For $\underline{P}(B) > 0$, we define the conditional set of desirable gambles as:

$$\mathcal{D}_{\vec{\Omega}|B} := \left\{ X \in L^\infty : X\chi_B \in \mathcal{D}_{\vec{\Omega}} \right\} = \left\{ X \in L^\infty : \limsup_{n \rightarrow \infty} \overrightarrow{\Sigma(X\chi_B)} \leq 0 \right\},$$

¹⁹This condition is indispensable in order to make the connection to the generalized Bayes rule.

and a corresponding upper prevision, which we call the generalized Bayes rule, as:

$$\begin{aligned}\text{GBR}(X|B) &:= \inf \left\{ \alpha \in \mathbb{R} : X - \alpha \in \mathcal{D}_{\bar{\Omega}|B} \right\} \\ &= \inf \left\{ \alpha \in \mathbb{R} : \chi_B(X - \alpha) \in \mathcal{D}_{\bar{\Omega}} \right\} \\ &= \inf \left\{ \alpha \in \mathbb{R} : \bar{R}(\chi_B(X - \alpha)) \leq 0 \right\}.\end{aligned}\tag{4.8}$$

Remark 202. In fact, [Walley \[1991, Section 6.4\]](#) defines the generalized Bayes rule as the solution of $\bar{R}(\chi_B(X - \alpha)) = 0$ for α . It can be checked that this solution coincides with Definition 201,²⁰ see Appendix 4.7.3.

Proposition 203. Let $\underline{P}(B) > 0$. It holds that $\bar{R}(X|B) = \text{GBR}(X|B)$.

Proof It is not hard to check that $\bar{R}(\cdot|B)$ is a coherent upper prevision on L^∞ , hence we can represent it as [[Walley, 1991, Theorem 3.8.1](#)]:

$$\bar{R}(X|B) = \inf \left\{ \alpha \in \mathbb{R} : X - \alpha \in \mathcal{D}_{\bar{R}(\cdot|B)} \right\}, \quad \text{where } \mathcal{D}_{\bar{R}(\cdot|B)} := \{X \in L^\infty : \bar{R}(X|B) \leq 0\}.$$

We show that $\bar{R}(X|B) = \text{GBR}(X|B)$ by showing that $\mathcal{D}_{\bar{R}(\cdot|B)} = \mathcal{D}_{\bar{\Omega}|B}$.

Let $X \in L^\infty$. On the one hand, we know

$$\begin{aligned}X \in \mathcal{D}_{\bar{\Omega}|B} &\iff X\chi_B \in \mathcal{D}_{\bar{\Omega}} \iff \bar{R}(X\chi_B) \leq 0 \\ &\iff \limsup_{n \rightarrow \infty} \sum_{i=1}^n \frac{(X\chi_B)(\bar{\Omega}(i))}{n} \leq 0.\end{aligned}\tag{4.9}$$

On the other hand,

$$X \in \mathcal{D}_{\bar{R}(\cdot|B)} \iff \bar{R}(X|B) \leq 0 \iff \limsup_{n \rightarrow \infty} \frac{\frac{1}{n} \sum_{i=1}^n (X\chi_B)(\bar{\Omega}(i))}{\frac{1}{n} \sum_{i=1}^n \chi_B(\bar{\Omega}(i))} \leq 0.\tag{4.10}$$

It remains to show that the two limit statements (Equation 4.9 and Equation 4.10) are equivalent. Due to the limit operation we can neglect the terms $n = 1, \dots, n_B - 1$. Furthermore, we know that $\vec{b}(n) \in (0, 1]$, $n \geq n_B$, and also $0 < \liminf_{n \rightarrow \infty} \vec{b}(n)$. Thus, defining $\vec{a}(n) := \frac{1}{n} \sum_{i=1}^n (X\chi_B)(\bar{\Omega}(i))$ and $\vec{b}(n) := \frac{1}{n} \sum_{i=1}^n \chi_B(\bar{\Omega}(i))$, we can leverage Lemma 207,²¹ included in Appendix 4.7.4 to show:

$$\limsup_{n \rightarrow \infty} \vec{a}(n) \leq 0 \iff \limsup_{n \rightarrow \infty} \frac{\vec{a}(n)}{\vec{b}(n)} \leq 0.$$

²⁰The conditional set of desirable gambles is considered for instance in [[Augustin et al., 2014](#)] and [[Wheeler, 2021](#)], but there the link to the generalized Bayes rule is not made technically clear.

²¹To rigorously apply the Lemma, we would again introduce a wrapper for the sequence $\vec{b}(n)$ to ensure strict positivity, since finitely many terms $i = 1, \dots, n_B - 1$ might be zero.

■

Remark 204. Note that $X \mapsto \limsup_{n \rightarrow \infty} \overrightarrow{\Sigma(X\chi_B)} = \overline{R}(X\chi_B)$ is *not* in general a coherent upper prevision on L^∞ , as it can violate UP1; see Appendix 4.7.5. In general, we have $\text{GBR}(X|B) \neq \overline{R}(X\chi_B)$.

As a consequence, we can apply the classical representation result for the generalized Bayes rule.

Corollary 205. *If $\underline{P}(B) > 0$, the conditional upper prevision can be obtained by updating each linear prevision in the set of cluster points, that is:*

$$\overline{R}(X|B) = \sup \left\{ E(X|B) : E \in \text{CP} \left(\vec{E} \right) \right\},$$

where conditioning of the linear previsions is in the sense of Definition 4.4.

This follows from [Walley, 1991, Theorem 6.4.2]. Intuitively, it makes no difference whether we consider the cluster points of the sequence of conditional probabilities or whether we condition all probabilities in the set of cluster points in the classical sense.

Closely related to conditional probability is the concept of *statistical independence*, which plays a central role not only in Kolmogorov’s [Durrett, 2019, p. 37], but more generally in most probability theories (Levin [1980]; Fine [1973, Sections IIF, IIIG and VH]). In Appendix 4.9 we offer an independence concept for the case of possibly diverging relative frequencies and discuss how it relates to the classical independence notion in Kolmogorov’s framework.

4.5 RELATED WORK

In this section we examine previous research at the intersection of frequentism and imprecise probability. While divergence of relative frequencies has been linked to imprecise probability before, this has almost exclusively been done in settings which are not *strictly* frequentist. Fine [1970] was one of the first authors to critically evaluate the hypothesis of statistical stability. Fine [1970] observed that this widespread hypothesis is regarded as a “striking instance of order in chaos” in the statistics community, and sought to challenge its nature as an empirical “fact”. In contrast to our approach, Fine [1970] was concerned with finite sequences and the question what it means for such a sequence to be random. While Fine did mention von Mises, Fine [1970] opted for a randomness definition based on computational complexity. Intuitively, one can consider a sequence random if it cannot be generated by a short computer program (i.e. Turing machine). Fine then showed that

statistical stability (“apparent convergence”) occurs *because of*, and not in spite of, high randomness of the sequence. In contrast, a sequence for which relative frequencies diverge has low computational complexity. We consider these findings surprising, and believe that an interesting avenue for future research with respect to statistical stability lies in the comparison of the computational complexity approach to von Mises randomness notion based on selection rules. We agree with Fine [1970] that apparent convergence is not some law of nature, but rather a consequence of data handling.

The previously mentioned paper may be seen as a predecessor to a long line of work by Terrence Fine and collaborators, [Fine, 1976, Walley and Fine, 1982, Kumar and Fine, 1985, Grize and Fine, 1987, Fine, 1988, Papamarcou and Fine, 1991a,b, Sadrolhefazi and Fine, 1994, Fierens et al., 2009]; see also [Fine, 2016] for an introduction. A central motivation behind this work was to develop a frequentist model for the puzzling case of stationary, unstable phenomena with bounded time averages. What differentiates this work from ours is that we take a *strictly frequentist* approach: we explicitly define the upper probability and upper prevision from a given sequence. In contrast, the above works (with the exceptions of Section 4.3 in [Walley and Fine, 1982], [Papamarcou and Fine, 1991b] and [Fierens et al., 2009]) use an imprecise probability to represent a single trial in a sequence of unlinked repetitions of an experiment, and then induce an imprecise probability via an infinite product space. This is in the spirit of, and can be understood as a generalization of, the standard frequentist approach, where one would assume that $X_1, X_2, \dots$ form an i.i.d. sequence of random variables; here, there is both an “individual P ,” as well as an induced “aggregate P ” on the infinite product space, which can be used to measure an event such as convergence or divergence of relative frequencies.

When a single trial is assumed to be governed by an imprecise probability, how can this be interpreted? And what is the interpretation of the mysterious “aggregate imprecise probability”? This model falls prey to similar criticisms as we outlined in the Introduction (Section 4.1) concerning the theoretical law of large numbers. In fact, Walley and Fine [1982] subscribed to a frequency-propensity interpretation (specifically, they were inspired by Giere [1973]), where the imprecise probability of a single trial represents its propensity, that is, its tendency or disposition to produce a certain outcome. Consequently, one obtains a propensity for compound trials in terms of an imprecise probability and thus one can ascribe a lower and upper probability to events such as divergence of relative frequencies. To us, the meaning of such a propensity is unclear. While we are not against a propensity interpretation as such, our motivation was to work with a parsimonious set of assumptions. To this end, we took the sequence as the primitive entity, without relying on an underlying “individual” (imprecise) probability.

Closely related to our work is [Papamarcou and Fine, 1991b], who were also inspired by von Mises. The authors proved that, for any set of probability measures $\mathcal{P}$ on $(\Omega, 2^\Omega)$,

$|\Omega| < \infty$, and any countable set of place selection rules $\mathcal{S}$,²² the existence of a sequence $\vec{\Omega}: \mathbb{N} \rightarrow \Omega$ with the following property can be guaranteed [Papamarcou and Fine, 1991b, Theorem 2.2]:

$$\forall \vec{S}_j \in \mathcal{S} : \forall A \subseteq \Omega : \limsup_{n \rightarrow \infty} \frac{\sum_{i=1}^n \chi_A(\vec{\Omega}(i)) \cdot \vec{S}_j(i)}{\sum_{i=1}^n \vec{S}_j(i)} = \sup\{P(A) : P \in \mathcal{P}\}.$$

That is, the sequence has the specified upper probability (take $\vec{S}_j(i) = 1 \forall i \in \mathbb{N}$) and this property is stable under subselection. Note that this claim is in one sense weaker than our Proposition 192, where we construct a sequence for which the set of cluster points is exactly a prespecified one — coherent upper previsions are more expressive than coherent upper probabilities; on the other hand it is stronger, since the property holds also when applying selection rules.

Within the setup of [Walley and Fine, 1982], Cozman and Chrisman [1997, Theorem 1, Theorem 2] proposed an estimator for the underlying imprecise probability of the sequence. Specifically, they computed relative frequencies along a set of selection rules (however without referring to von Mises) and then took their minimum to obtain a lower probability; in a specific technical sense, this estimation succeeds. What motivated the authors to do this is an assumption on the data-generating process: at each trial, “nature” may select a different distribution from a set of probability measures; the trials are then independent but not identically distributed. This viewpoint also motivated Fierens et al. [2009], who restricted themselves to finite sequences. They offered the metaphor of an *analytical microscope*. With more and more complex selection rules (“powerful lenses”), along which relative frequencies are computed, more and more structure of the set of probabilities comes to light. The authors also proposed a way to simulate data from a set of probability measures. Cattaneo [2017] investigated an empirical, frequentist interpretation of imprecise probability in a similar setting, where $X_1, X_2, \dots$ is a sequence of precise Bernoulli random variables, but $p_i := P(X_i = 1)$ is chosen by nature and may differ from trial to trial, hence $p_i \in [p_i, \bar{p}_i]$. The author drew the sobering conclusion that “imprecise probabilities do not have a generally valid, clear empirical meaning, in the sense discussed in this paper”.

Works which proposed extensions (modifications) of the law of large numbers to imprecise probabilities include [Marinacci, 1999, Maccheroni and Marinacci, 2005, De Cooman and Miranda, 2008, Chen et al., 2013, Peng, 2019].

In order to access a more powerful toolbox, De Cooman and De Bock [2022] and Persiau et al. [2022] more recently studied the interplay of imprecise probability and randomness in the game-theoretic setup with references to a frequentist perspective, see e.g. [De Cooman

²²See [Papamarcou and Fine, 1991b] for the definition. Intuitively, a place selection rule is causal, i.e. depends only on past values.

and De Bock, 2022, Corollary 28]. While we believe there is potential for establishing relations between our approach and theirs, the differences in technical setup make it challenging to do so straightforwardly.

Separate from the imprecise probability literature, Gorban [2017] studied the phenomenon of statistical stability and its violations in depth, including theory and experimental studies. Similarly, the work of Ivanenko [2010], a major motivation for our work, does not appear to be known in the imprecise probability literature.

4.6 CONCLUSION

In this work, we have extended strict frequentism to the case of possibly divergent relative frequencies and sample averages, tying together threads from [von Mises, 1919], [Ivanenko, 2010] and [Walley, 1991]. In particular, we have recovered the generalized Bayes rule from a strictly frequentist perspective. Furthermore, we have established strictly frequentist *semantics* for imprecise probability, by demonstrating that (under the mild assumption that $|\Omega| < \infty$) we can explicitly construct a sequence for which the relative frequencies have a prespecified set of cluster points, corresponding to the coherent upper prevision. Previous results only covered the more restrictive case of upper probabilities [Walley and Fine, 1982, Papamarcou and Fine, 1991b].

Statisticians exclusively assume that their data is part of a stable sequence, but the hypothesis of perfect statistical stability is just a *hypothesis*; see Appendix 4.10 for an elaboration of this point. Importantly, when one blindly *assumes* convergence of relative frequencies, one will not notice when it is violated — in the practical case, when only a finite sequence is given, such a violation amounts to instability of relative frequencies even for long observation intervals [Gorban, 2017]. In this work, we have rejected the assumption of stability; furthermore, in contrast to other related work, we have aimed to weaken the set of assumptions by taking the concept of a sequence as the primitive. However, this gives rise to the critique that no finite part of a sequence has any bearing on what the limit is, as has been pointed out by other authors whose studies attempted a frequentist understanding of imprecise probability (e.g. [Cattaneo, 2017]). So what is the empirical content of our theory, what are its practical implications?

The reader may wonder why we have introduced von Mises' frequentist account but not further used selection rules afterwards. In von Mises' framework, the set of selection rules expresses randomness assumptions about the sequence, similar to what the i.i.d. assumption achieves in the standard picture. In our view, randomness assumptions are *the* key to empower generalization in the finite data setting. Hence, to supplement our theory with empirical content, the introduction of selection rules is needed. However, multiple directions can be pursued here. For instance, Papamarcou and Fine [1991b] have defined

the concept of an *unstable collective*, where divergence remains unchanged when applying selection rules. By contrast, we could introduce a set of selection rules and assume that relative frequencies converge *within each* selection rule, but to potentially different limits across selection rules.²³ Hence, we view this paper as only the first step of a larger research agenda. The next step is to incorporate randomness assumptions into the picture and explore the connections between various possible approaches, specifically how different ways of relaxing vM1 and vM2 are related. As a consequence, we would obtain potentially multiple ways of bridging the finite and the infinite case, which enables practical applications²⁴.

Finally, we remark that an interesting avenue for future research may investigate the use of *nets*, which generalize the concept of a sequence. Indeed, *fraction-of-time probability* [Gardner, 1986, Leśkow and Napolitano, 2006, Napolitano and Gardner, 2022, Gardner, 2022] is a theory of probability with remarkable parallels to von Mises' [1919]. Instead of sequences, this theory is based on continuous time, hence a net $\vec{\Omega}: \mathbb{R} \rightarrow \Omega$.²⁵ Sample averages are then given by integration instead of summation. In essence, this amounts to using a different *relative measure* than the counting measure, which is implicit in the work of von Mises [1919]. However, fraction-of-time probability was so far developed only for the convergent case; we expect that a similar construction as in Section 4.2.2 could be used to extend it to the case of divergence.

ACKNOWLEDGMENTS

Robert Williamson thanks Jingni Yang for a series of discussions over several years about Ivanenko's work which provided much inspiration for the present work. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy - EXC number 2064/1 - Project number 390727645. Thanks to the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christian Fröhlich and Rabanus Derr.

²³As demonstrated by Examples 2 and 3 in [Cozman and Chrisman, 1997], converging relative frequencies within selection rules can lead to both overall convergence or divergence (on the whole sequence).

²⁴Note that precise probability in the framework of von Mises also faces this bridging problem, which is a special case of ours; a point explicitly recognised by Kolmogorov long ago: "Here I insist on the view, expressed by Mr. von Mises himself [von Mises, 1931, pp. 21–26] that 'collectives' are finite (though very large) in real practice" [Kolmogorov, 1939]. Kolmogorov reiterated von Mises' view that the infinite sample case is indeed a "mathematical idealization" of the practical situation. Approximately a century of effort has produced some reasonable refinements of that idealization in the finite *stable* case. We hope it will not take quite that long in the finite *unstable* case, but recognise the many difficulties that will need addressing.

²⁵We remark that the notion of a sampling net in [Ivanenko, 2010] is a different one, that is, fraction-of-time probability does not fit this concept, although it is based on the usage of a net.

4.7 APPENDIX: PROOFS

4.7.1 WEAK* COMPACTNESS OF $\text{PF}(\Omega)$

[Walley \[1991, Appendix D4\]](#) states that the set of linear previsions, $\text{PF}(\Omega)$, is compact due to the Alaoglu-Bourbaki theorem, but does not explain how this follows. For completeness, we provide an argument. First, we can observe, like [Walley \[1991\]](#), that the set is weak* closed. We use the following well known Lemma, see e.g. [\[Deitmar, 2016, Lemma 12.3.4.\]](#).

Lemma 206. *Let $\mathcal{X}$ be a topological space, $K \subseteq \mathcal{X}$ be compact and $L \subseteq K$ be closed. Then L is compact.*

Thus we will show that $\text{PF}(\Omega)$ is a subset of some weak* compact set in $(L^\infty)^*$, hence it is in fact weak* compact. From the Alaoglu-Bourbaki theorem (see e.g. [\[Holmes, 1975, p. 70\]](#)), we know that the unit ball of the dual norm is weak* compact in $(L^\infty)^*$. By definition of the dual norm $\|\cdot\|^*$ of the L^∞ norm $\|X\| := \sup_{\omega \in \Omega} \{|X(\omega)|\}$, this is the following set:

$$\begin{aligned} B &:= \{X^* \in (L^\infty)^*: \|X^*\|^* \leq 1\} \\ &= \{X^* \in (L^\infty)^*: \sup \{|X^*(X)|: X \in L^\infty \text{ for which } \|X\| \leq 1\} \leq 1\} \\ &= \left\{ X^* \in (L^\infty)^*: \sup \left\{ |X^*(X)|: X \in L^\infty \text{ for which } \sup_{\omega \in \Omega} \{|X(\omega)|\} \leq 1 \right\} \leq 1 \right\}. \end{aligned}$$

Thus, to show that $\text{PF}(\Omega)$ is weak* compact, it suffices to show that $\text{PF}(\Omega) \subseteq B$. That is, given some $E \in \text{PF}(\Omega)$, we show that if $X \in L^\infty$ is such that $\sup_{\omega \in \Omega} \{|X(\omega)|\} \leq 1$, then $|E(X)| \leq 1$, since then the supremum over all such X is also ≤ 1 . Thus it suffices to show that $|E(X)| \leq \sup_{\omega \in \Omega} \{|X(\omega)|\}$. We know that $E(|X|) \leq \sup_{\omega \in \Omega} \{|X(\omega)|\}$, since E is a coherent upper prevision, cf. [Walley \[1991, 2.6.1a\]](#). But we also have that $|E(X)| \leq E(|X|)$ from monotonicity and $E(0) = 0$, see e.g. [Pichler \[2013, Proposition 5\]](#), and thus $|E(X)| \leq E(|X|) \leq \sup_{\omega \in \Omega} \{|X(\omega)|\} \leq 1$, which concludes the proof.

4.7.2 PROPERTIES OF THE INDUCED UPPER PREVISION

To see that $X - \bar{R}(X) \in \mathcal{D}_{\vec{\Omega}}$:

$$\begin{aligned} X - \bar{R}(X) \in \mathcal{D}_{\vec{\Omega}} &\iff \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left((X - \bar{R}(X)) \left(\vec{\Omega}(i) \right) \right) \leq 0 \\ &\iff \limsup_{n \rightarrow \infty} \bar{\Sigma} \vec{X} - \bar{R}(X) \leq 0, \quad \text{since } \bar{R}(X) \text{ is constant} \\ &\iff \limsup_{n \rightarrow \infty} \bar{\Sigma} \vec{X}(n) - \limsup_{n \rightarrow \infty} \bar{\Sigma} \vec{X}(n) = 0 \leq 0. \end{aligned}$$

Now we show that $\bar{R}(X) = \limsup_{n \rightarrow \infty} \bar{\Sigma} \vec{X}(n)$ is in fact the smallest number such that the relation in Equation 4.1 holds. Suppose there exists $\varepsilon > 0$ such that $\bar{R}(X) - \varepsilon$ (with $\bar{R}$ defined as before) makes $X - (\bar{R}(X) - \varepsilon)$ desirable, that is

$$\limsup_{n \rightarrow \infty} \bar{\Sigma} \vec{X} - \bar{R}(X) + \varepsilon \leq 0,$$

which is a contradiction due to our choice of $\bar{R}$.

4.7.3 PROOF OF REMARK 202

In the literature, the generalized Bayes rule is defined as the solution α^* of $\bar{R}(\chi_B(X - \alpha)) = 0$. We show that $\alpha^* = \inf \{\alpha \in \mathbb{R} : \bar{R}(\chi_B(X - \alpha)) \leq 0\}$. Of course, we get for $\alpha := \alpha^*$ that equality holds ($= 0$). We just have to exclude the possibility that there exists $\tilde{\alpha} < \alpha^*$ so that $\bar{R}(\chi_B(X - \tilde{\alpha})) \leq 0$.

Assume such an $\tilde{\alpha}$ exists, so $\bar{R}(\chi_B X - \chi_B \tilde{\alpha}) \leq 0$. Write $\tilde{\alpha} + \varepsilon \leq \alpha^*$ for some $\varepsilon > 0$. Since $\chi_B X - \chi_B \tilde{\alpha} - \chi_B \varepsilon \geq \chi_B X - \chi_B \alpha^*$, it follows from monotonicity that $\bar{R}(\chi_B X - \chi_B \tilde{\alpha} - \chi_B \varepsilon) \geq \bar{R}(\chi_B X - \chi_B \alpha^*) = 0$. But from subadditivity, $\bar{R}(\chi_B X - \chi_B \tilde{\alpha} - \chi_B \varepsilon) \leq \bar{R}(\chi_B X - \chi_B \tilde{\alpha}) + \bar{R}(-\chi_B \varepsilon)$ and since $\varepsilon \bar{R}(-\chi_B) < 0$ due to coherence and $\underline{P}(B) > 0$, we have $\bar{R}(\chi_B X - \chi_B \tilde{\alpha} - \chi_B \varepsilon) < \bar{R}(\chi_B X - \chi_B \tilde{\alpha})$. Taking this together, we obtain $\bar{R}(\chi_B X - \chi_B \tilde{\alpha}) > 0$, a contradiction to the assumption.

Thus we have shown that $\alpha^* = \inf \{\alpha \in \mathbb{R} : \bar{R}(\chi_B(X - \alpha)) \leq 0\}$. The other expressions in Definition 201 follow by simple manipulations.

4.7.4 SUPPLEMENT FOR PROOF OF PROPOSITION 203

Lemma 207. Let $\vec{a} : \mathbb{N} \rightarrow \mathbb{R}$ be a sequence and $\vec{b} : \mathbb{N} \rightarrow (0, 1]$ be a nonnegative sequence such that $\liminf_{n \rightarrow \infty} \vec{b}(n) > 0$. Then:

$$\limsup_{n \rightarrow \infty} \vec{a}(n) \leq 0 \iff \limsup_{n \rightarrow \infty} \frac{\vec{a}(n)}{\vec{b}(n)} \leq 0.$$

Proof For brevity we simply write a_n and b_n for the sequences $\vec{a}(n)$ and $\vec{b}(n)$.

$$\begin{aligned} \limsup_{n \rightarrow \infty} a_n \leq 0 &\iff \limsup_{n \rightarrow \infty} \frac{a_n}{b_n} \leq 0 \\ \lim_{n \rightarrow \infty} \left(\sup_{k \geq n} a_k \right) \leq 0 &\iff \lim_{n \rightarrow \infty} \left(\sup_{k \geq n} \frac{a_k}{b_k} \right) \leq 0. \end{aligned}$$

where we know that $b_n \in (0, 1]$ and furthermore $0 < \liminf_{n \rightarrow \infty} b_n \leq \limsup_{n \rightarrow \infty} b_n$. If the sequence b_n would actually converge, then the statement is clearly true, since we can then

pull out the limit of b_n (this is allowed).

We begin by showing that $LHS \leq 0 \implies RHS \leq 0$. Our assumption is that

$$\forall \varepsilon > 0 : \exists n_0 \in \mathbb{N} : \forall n \geq n_0 : \sup_{k \geq n} a_k < \varepsilon. \quad (4.11)$$

Our aim is to show that

$$\forall \varepsilon' > 0 : \exists n'_0 \in \mathbb{N} : \forall n \geq n'_0 : \sup_{k \geq n} \frac{a_k}{b_k} < \varepsilon'.$$

So let some $\varepsilon' > 0$ be given and fixed. We have to exhibit some n'_0 such that the above statement holds. Choose $\varepsilon := \varepsilon' \cdot \lim_{n \rightarrow \infty} \inf_{k \geq n} b_k \cdot \frac{1}{\kappa}$, for an arbitrary $\kappa > 1$. Then $\varepsilon > 0$ by our assumption that $\lim_{n \rightarrow \infty} \inf_{k \geq n} b_k > 0$, i.e. that $\underline{P}(B) > 0$. Note that $\varepsilon \leq \varepsilon'$. Then, we know that $\exists n_0(\varepsilon)$ such that $\forall n \geq n_0(\varepsilon) \sup_{k \geq n} a_k < \varepsilon$.

Also, we know that $\forall \kappa' > 1 : \exists n''_0 \in \mathbb{N} : \forall n \geq n''_0$:

$$\frac{\lim_{n \rightarrow \infty} \inf_{k \geq n} b_k}{\inf_{k \geq n} b_k} \leq \kappa'. \quad (4.12)$$

The numerator is the limit of the denominator (which exists) and furthermore $\inf_{k \geq n} b_k$ is monotone increasing in n , that is, $\forall n \in \mathbb{N} : \lim_{n \rightarrow \infty} \inf_{k \geq n} b_k \geq \inf_{k \geq n} b_k$. Thus, the ratio approaches 1 from above for large n .

Now choose $\kappa' := \kappa$ and $n'_0 := \max(n_0(\varepsilon), n''_0(\kappa'))$. That is, we know that then both Equation 4.12 and Equation 4.11 hold. Then we want to show:

$$\sup_{k \geq n} \frac{a_k}{b_k} = \max \left(\sup_{k \geq n, a_k \geq 0} \frac{a_k}{b_k}, \sup_{k \geq n, a_k < 0} \frac{a_k}{b_k} \right) \stackrel{!}{<} \varepsilon',$$

which is a legitimate decomposition of the supremum into the “negative” and “nonnegative” subsequences. But look at the second term ($a_k < 0$) and observe that since $b_k > 0$, clearly $\sup_{k \geq n, a_k < 0} \frac{a_k}{b_k} \leq 0 < \varepsilon'$. Thus we only have to consider the first term. Further observe that

$$\sup_{k \geq n, a_k \geq 0} \frac{a_k}{b_k} \leq \sup_{k \geq n, a_k \geq 0} a_k \cdot \sup_{k \geq n, a_k \geq 0} \frac{1}{b_k} = \sup_{k \geq n, a_k \geq 0} a_k \cdot \frac{1}{\inf_{k \geq n, a_k \geq 0} b_k},$$

due to nonnegativity of the $a_k \geq 0$ and a general rule for the supremum/infimum, which applies since b_k is strictly positive. Now by assumption,

$$\sup_{k \geq n, a_k \geq 0} a_k \cdot \frac{1}{\inf_{k \geq n, a_k \geq 0} b_k} < \varepsilon' \lim_{n \rightarrow \infty} \inf_{k \geq n} b_k \underbrace{\frac{1}{\kappa \inf_{k \geq n, a_k \geq 0} b_k}}_{\leq \kappa} = \varepsilon' \cdot \frac{\lim_{n \rightarrow \infty} \inf_{k \geq n} b_k}{\inf_{k \geq n, a_k \geq 0} b_k} \cdot \frac{1}{\kappa} \leq \varepsilon'.$$

Noting that $\inf_{k \geq n} b_k \leq \inf_{k \geq n, a_k \geq 0} b_k$ and therefore

$$\frac{\lim_{n \rightarrow \infty} \inf_{k \geq n} b_k}{\inf_{k \geq n, a_k \geq 0} b_k} \leq \frac{\lim_{n \rightarrow \infty} \inf_{k \geq n} b_k}{\inf_{k \geq n} b_k} \leq \kappa'.$$

Altogether, we have shown that

$$\sup_{k \geq n} \frac{a_k}{b_k} < \varepsilon',$$

and therefore $LHS \leq 0 \implies RHS \leq 0$.

It remains to show that $RHS \leq 0 \implies LHS \leq 0$. Our assumption is that

$$\forall \varepsilon' > 0 : \exists n'_0 \in \mathbb{N} : \forall n \geq n'_0 : \sup_{k \geq n} \frac{a_k}{b_k} < \varepsilon. \quad (4.13)$$

and our aim is to show that then

$$\forall \varepsilon > 0 : \exists n_0 \in \mathbb{N} : \forall n \geq n_0 : \sup_{k \geq n} a_k < \varepsilon.$$

So let $\varepsilon > 0$ be fixed. Choose $\varepsilon' := \varepsilon$ and set $n_0 := n'_0$. Then we want to show that $\forall n \geq n_0$:

$$\sup_{k \geq n} a_k = \max \left(\sup_{k \geq n, a_k \geq 0} a_k, \sup_{k \geq n, a_k < 0} a_k \right) \stackrel{!}{<} \varepsilon.$$

As to the second term, it is obviously negative, in particular $\sup_{k \geq n, a_k < 0} a_k < \varepsilon$. For the first term, where the a_k are nonnegative, observe that then $a_k \leq \frac{a_k}{b_k}$ since $b_k \in (0, 1]$, consequently we have $\forall n \geq n_0$:

$$\sup_{k \geq n, a_k \geq 0} a_k \leq \sup_{k \geq n, a_k \geq 0} \frac{a_k}{b_k} \leq \sup_{k \geq n} \frac{a_k}{b_k} < \varepsilon = \varepsilon'.$$

by our assumption Equation 4.13. And thus we have shown that $RHS \leq 0 \implies LHS \leq 0$.

■

4.7.5 PROOF OF REMARK 204

Take for example $X(\omega) = -1$ for a $B \subseteq \Omega$ where $\underline{P}(B) < 1$. Then $\sup X = -1$, but

$$\overline{R}(X\chi_B) = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (X\chi_B) \left(\vec{\Omega}(i) \right) = \limsup_{n \rightarrow \infty} -\frac{1}{n} \sum_{i=1}^n \chi_B \left(\vec{\Omega}(i) \right) \quad (4.14)$$

$$= -\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \chi_B \left(\vec{\Omega}(i) \right) = -\underline{P}(B), \quad (4.15)$$

and $\sup X = -1 < -\underline{P}(B)$, hence UP1 does not hold. Thus $X \mapsto \overline{R}(X\chi_B)$ is not a coherent upper prevision on L^∞ in general (it is of course for $B = \Omega$).

4.8 APPENDIX: EXISTENCE OF SEQUENCES WITH PRESPECIFIED RELATIVE FREQUENCY CLUSTER POINTS

In this section we prove Theorem 194 and thus demonstrate the existence of sequences $\vec{x}: \mathbb{N} \rightarrow [k]$ whose corresponding relative frequencies $\vec{r}^{\vec{x}}: \mathbb{N} \rightarrow \Delta^k$ have the property that their set of cluster points $\text{CP}(\vec{r}^{\vec{x}}) = C$, where C is an arbitrary closed rectifiable curve in Δ^k . We do so constructively by providing an explicit procedure which takes a chosen C and constructs a suitable $\vec{x}$. We illustrate our method with some examples. The question of how common sequences with non-convergent relative frequencies are is addressed in Section 4.10.

4.8.1 SUFFICIENT TO WORK WITH TOPOLOGY INDUCED BY EUCLIDEAN NORM ON THE SIMPLEX

Let $|\Omega| = k < \infty$. A linear prevision E on L^∞ is in a one-to-one correspondence with a finitely additive probability P , which we can represent as a point in the $(k-1)$ -simplex (see Lemma 209 below). It is convenient to then consider the cluster points of a sequence of such probabilities in the $(k-1)$ -simplex, with respect to the topology induced by the Euclidean norm on $\mathbb{R}^k$. We show that the notion of such a cluster point coincides with a cluster point in $(L^\infty)^*$ with respect to the weak* topology. This is the goal of this section; in particular, we prove Proposition 196.

Since $|\Omega| = k < \infty$, we can represent any $X \in L^\infty$ as:

$$X(\omega) = c_1\chi_{\{\omega_1\}} + \cdots + c_k\chi_{\{\omega_k\}},$$

where $c_i = X(\omega_i)$.

Similarly, any $Z \in (L^\infty)^*$ can be represented as:

$$\begin{aligned} Z(X) &= Z(c_1\chi_{\{\omega_1\}} + \cdots + c_k\chi_{\{\omega_k\}}) \\ &= c_1Z(\chi_{\{\omega_1\}}) + \cdots + c_kZ(\chi_{\{\omega_k\}}), \end{aligned}$$

since the $Z \in (L^\infty)^*$ are linear functionals and the coefficients c_i depend on X . Intuitively, $Z(\chi_{\{\omega_i\}}) = P(\omega_i)$ if $Z \in \text{PF}(\Omega)$. Define $d_i := Z(\chi_{\{\omega_i\}}) \forall i \in 1, \dots, k$ and consequently define

$$\|Z\| := \sqrt{d_1^2 + \cdots + d_k^2}.$$

For a given $Z \in (L^\infty)^*$, call $d_Z := (d_1, \dots, d_k) \in \mathbb{R}^k$ the coordinate representation of Z .

Lemma 208. $\|\cdot\|$ is a norm on $(L^\infty)^*$.

Proof *Point-separating property:* if and only if $Z = 0$, where $0 \in (L^\infty)^*$ is given by $0(X) = 0$, $\forall X \in L^\infty$. But this is easily observed, due to the similar property holding for the Euclidean norm: $\|Z\| = 0$ if and only if $d_i = 0 \forall i = 1, \dots, k$. Since $d_i = Z(\chi_{\{\omega_i\}})$, this is the case exactly if $Z = 0$.

Subadditivity: $\|Y + Z\| \leq \|Y\| + \|Z\|$. For the input $Y + Z$ we get $d_i = (Y + Z)(\chi_{\{\omega_i\}}) = Y(\chi_{\{\omega_i\}}) + Z(\chi_{\{\omega_i\}})$ and then subadditivity follows from the similar property for the Euclidean norm.

Absolute homogeneity: $\|\lambda Z\| = |\lambda| \|Z\|$, $\forall \lambda \in \mathbb{R}$. This follows easily.

With these properties, $\|\cdot\|$ is a valid norm on $(L^\infty)^*$. ■

We now show that the $(k - 1)$ -simplex is in a one-to-one correspondence with the set of linear previsions $\text{PF}(\Omega)$ via the coordinate representation.

Lemma 209. Let $d \in \Delta^k$. Then $Z(X) := c_1 d_1 + \dots + c_k d_k \in \text{PF}(\Omega)$. Conversely, let $Z \in \text{PF}(\Omega)$. Then the corresponding $d_Z \in \Delta^k$.

Proof Let $d \in \Delta^k$ and $Z(X) := c_1 d_1 + \dots + c_k d_k$. Since $\sum_{i=1}^k d_i = 1$ we have immediately that $Z(\chi_\Omega) = 1$, noting that $\chi_\Omega = 1\chi_{\{\omega_1\}} + \dots + 1\chi_{\{\omega_k\}}$. Also, if $X \geq 0$, i.e. $c_i \geq 0$, $\forall i \in [k]$, then $Z(X) \geq 0$ since $d_i \geq 0 \forall i$. Thus, $Z \in \text{PF}(\Omega)$.

Conversely, let Z be a linear prevision, i.e. $Z(\chi_\Omega) = 1$ and $Z(X) \geq 0$ if $X \geq 0$. From $Z(\chi_\Omega) = 1$ we can deduce that $\sum_{i=1}^k d_i = 1$. If $X \geq 0$, we know that $c_i \geq 0 \forall i \in [k]$, hence $Z(X) \geq 0$ can only be true if all $d_i \geq 0$. Thus $d_Z \in \Delta^k$. ■

We here restate Proposition 196 for convenience.

Proposition 210. Let $\vec{E}(n) : \mathbb{N} \rightarrow \text{PF}(\Omega)$ be a sequence of linear previsions with underlying probabilities $\vec{P}(n) := A \mapsto \vec{E}(n)(A)$. Then $E \in \text{CP}(\vec{E}(n))$ with respect to the weak* topology if and only if the sequence $\vec{D} : \mathbb{N} \rightarrow \Delta^k$, $\vec{D}(n) := (\vec{P}(n)(\omega_1), \dots, \vec{P}(n)(\omega_k))$ has as cluster point $d_E = (E(\chi_{\{\omega_1\}}), \dots, E(\chi_{\{\omega_k\}}))$ with respect to the topology induced by the Euclidean norm on $\mathbb{R}^k$.

First note that if $E \in \text{PF}(\Omega)$, then $d_E = (E(\chi_{\{\omega_1\}}), \dots, E(\chi_{\{\omega_k\}})) = (P(\omega_1), \dots, P(\omega_k))$, where P is the underlying probability of E , and hence $\|E\| = \sqrt{P(\omega_1)^2 + \dots + P(\omega_k)^2}$. To complete the proof, we need some further statements first.

Definition 211. A vector space $\mathcal{X}$ is called *topological vector space* if the topology on $\mathcal{X}$ is such that $(x, y) \mapsto x + y$ is continuous with respect to the product topology on $\mathcal{X} \times \mathcal{X}$ and $(\lambda, x) \mapsto \lambda x$ is continuous with respect to the product topology on $\mathbb{R} \times \mathcal{X}$. We call a topology which makes $\mathcal{X}$ a topological vector space a *linear topology*.

Remark 212. The weak* topology makes $(L^\infty)^*$ a topological vector space with a Hausdorff topology,²⁶ where vector addition and scalar product are defined pointwise: $Y + Z := X \mapsto Y(X) + Z(X) \forall X \in L^\infty$, $\lambda Z := X \mapsto \lambda Z(X) \forall X \in L^\infty$.

Remark 213 (well-known). A vector space whose topology is induced by a norm is a topological vector space.

Proposition 214. On every finite dimensional vector space X there is a unique topological vector space structure. In other words, any two Hausdorff linear topologies on X coincide [[Nagy, 2007](#)].

Now Proposition 210 directly follows.

Proof [Proof of Proposition 210] Our norm $\|\cdot\|$ makes $(L^\infty)^*$ a topological vector space due to Remark 213, and any topology induced by a norm is Hausdorff; but the weak* topology also makes $(L^\infty)^*$ a topological vector space, and the weak* topology is Hausdorff. Hence we can conclude from Proposition 214 that they coincide. But then the two notions of what a cluster point is of course also coincide, since this depends just on the topology. ■

Thus, for the proof of Theorem 192, we will work exclusively with the topology induced by the Euclidean metric restricted to Δ^k . For $z \in \Delta^k$ and $\epsilon > 0$ define the ϵ -neighbourhood (ball)

$$N_\epsilon(z) := \left\{ p \in \Delta^k : \|p - z\| < \epsilon \right\},$$

where $\|\cdot\|$ is the Euclidean norm (restricted to the simplex). Then from [[Schechter, 1997](#), p. 430] we have an equivalent definition of a cluster point:

Definition 215. Say that $z \in \Delta^k$ is a cluster point of a sequence $\vec{x} : \mathbb{N} \rightarrow \Delta^k$ (and denote by $CP(\vec{x})$ the set of all cluster points of $\vec{x}$) if for all $\epsilon > 0$, $|\{n \in \mathbb{N} : \vec{x}(n) \in N_\epsilon(z)\}| = \aleph_0$, where $\aleph_0 := |\mathbb{N}|$ is the cardinality of the natural numbers.

4.8.2 FURTHER NOTATION

We will work solely with the topology on Δ^k induced by the Euclidean metric; by the argument in the previous subsection the cluster points w.r.t. this topology coincide with those w.r.t. the weak* topology.

²⁶See for instance Exercise 13 and 14 in [Tao \[2009\]](#).

We introduce further notation to assist in stating our algorithm. For terser notation, we drop the $\vec{\cdot}$ symbol for sequences $\vec{x} : \mathbb{N} \rightarrow [k]$ throughout this appendix and simply write x . The i th canonical unit vector in Δ^k is denoted $e_i := (0, \dots, 1, \dots, 0)$, where the 1 is in the i th position. The boundary of the simplex is

$$\partial\Delta^k := \{(z_1, \dots, z_k) : z_1, \dots, z_k \geq 0, z_1 + \dots + z_k = 1\}.$$

If $p_1, p_2 \in \Delta^k$ then $l(p_1, p_2) := \{\lambda p_1 + (1 - \lambda)p_2 : \lambda \in [0, 1]\}$ is the *line segment connecting p_1 and p_2* . If $C \subset \Delta^k$ is a rectifiable closed curve parametrised by $c : [0, 1] \rightarrow \Delta^k$, its length is $\text{length}(C) = \int_0^1 |c'(t)| dt$. For $y \in \mathbb{R}$, $\lfloor y \rfloor$ is the nearest integer to y : $\lfloor y \rfloor := \lfloor y + \frac{1}{2} \rfloor$. We apply certain operations T elementwise. For example, if $z = \langle z_1, \dots, z_k \rangle \in \Delta^k$, then $\lfloor Tz \rfloor := \langle \lfloor Tz_1 \rfloor, \dots, \lfloor Tz_k \rfloor \rangle$ and (overloading notation) for $T \in \mathbb{N}$ and $\iota \in \mathbb{N}^k$, $\iota/T \in \mathbb{R}^n$ is simply $\langle \iota_1/T, \dots, \iota_k/T \rangle$. If $i < j \in \mathbb{N}$ the “interval” is $[i, j] := \{m \in \mathbb{N} : i \leq m \leq j\}$.

To avoid confusion, we will reserve “sequence” for the infinitely long $x : \mathbb{N} \rightarrow [k]$ and use “segment” to denote finite length strings $z : [n] \rightarrow [k]$ which we will write explicitly as $\langle z_1, \dots, z_n \rangle$. We construct the sequence x attaining the desired behavior of r^x by iteratively appending a series of segments. We denote the empty segment as $\langle \rangle$. If x^1 and x^2 are two finite segments of lengths ℓ_1 and ℓ_2 then their concatenation is the length $\ell_1 + \ell_2$ segment $x^1 x^2 := \langle x_1^1, \dots, x_{\ell_1}^1, x_1^2, \dots, x_{\ell_2}^2 \rangle$. We extend the $i^{[j]}$ notation to segments: if $z = \langle z_1, \dots, z_\ell \rangle$, then $z^{[i]} := \langle z, z, \dots, z \rangle$ is the length ℓi segment formed by concatenating i copies of z . Given $n \in \mathbb{N}$ and a sequence $x : \mathbb{N} \rightarrow [k]$, the shifted sequence $x^{+n} : \mathbb{N} \rightarrow [k]$ is defined via $x^{+n}(i) := x(i + n)$ for $i \in \mathbb{N}$.

4.8.3 PROPERTIES OF RELATIVE FREQUENCY SEQUENCES

Our construction of x relies upon the following elementary property of relative frequency sequences.

Lemma 216. *Suppose $k, n, m \in \mathbb{N}$, $x : \mathbb{N} \rightarrow [k]$. Then*

$$r^x(n + m) = \frac{n}{n + m} r^x(n) + \frac{m}{n + m} r^{x^{+n}}(m). \quad (4.16)$$

Proof For any $i \in [k]$ we have

$$\begin{aligned} r_i^x(n + m) &= \frac{1}{n + m} |\{j \in [n + m] : x(j) = i\}| \\ &= \frac{1}{n + m} (|\{j \in [n] : x(j) = i\}| + |\{j \in [n + m] \setminus [n] : x(j) = i\}|) \\ &= \frac{1}{n + m} \frac{n}{n} |\{j \in [n] : x(j) = i\}| + \frac{1}{n + m} \frac{m}{m} |\{t \in [m] : x(n + t) = i\}| \\ &= \frac{n}{n + m} r_i^x(n) + \frac{m}{n + m} \frac{1}{m} |\{t \in [m] : x^{+n}(t) = i\}| \\ &= \frac{n}{n + m} r_i^x(n) + \frac{m}{n + m} r_i^{x^{+n}}(m). \end{aligned}$$

Since this holds for all $i \in [k]$ we obtain Equation 4.16. ■

Observe that (4.16) also holds when $x: [n+m] \rightarrow [k]$ is a segment, in which case $x^{+n} = \langle x_{n+1}, \dots, x_{n+m} \rangle$. Furthermore note that (4.16) is a convex combination of the two points $r^x(n)$ and $r^{x^{+n}}(m)$ in Δ^k since $\frac{n}{n+m} + \frac{m}{n+m} = 1$ and both coefficients are positive. These two points are (respectively) the relative frequency of x at n , and the relative frequency of x^{+n} at m . This latter sequence will be the piece “added on” at each stage of our construction and forms the basis of our piecewise linear construction of r^x such that its cluster points are a given $C \subset \Delta^k$.

The set of cluster points of any sequence is closed. In addition, we have

Lemma 217. *For any $k \in \mathbb{N}$ and $x: \mathbb{N} \rightarrow [k]$, $\text{CP}(r^x)$ is a connected set.*

This follows immediately from [Bauschke et al., 2015, Lemma 2.6] upon observing that $\lim_{n \rightarrow \infty} \|r^x(n) - r^x(n+1)\| = 0$ since $\|r^x(n) - r^x(n+1)\| = \|r^x(n) - \frac{n}{n+1}r^x(n) - \frac{1}{n+1}e_{x(n+1)}\| = \frac{1}{n+1}\|r^x(n) - e_{x(n+1)}\| \leq \frac{2}{n+1}$. The boundedness of r^x is essential for this to hold — for unbounded sequences the set of cluster points need not be connected [Ašić and Adamović, 1970].

4.8.4 LOGIC OF THE CONSTRUCTION

The idea of our construction is as follows (see Figure 4.1 below for a visual aid). In order to satisfy the definition of cluster points, we need to return to each neighbourhood of each point in C infinitely often. To that end we iterate through an infinite sequence of generations indexed by $g \in \mathbb{N}$. For each g , we approximate C by a polygonal approximation C^g comprising V^g -many separate segments. We choose the sequence $(V^g)_{g \in \mathbb{N}}$ so that C^g approaches C in an appropriate sense. Then for generation g we append elements to x to ensure the sequence of relative frequencies makes another cycle approximately following C^g . We control the approximation error of this process and ensure its error is of a size that also decreases with increasing g .

We now describe the construction of a single generation. Thus suppose g is now fixed and suppose the current partial sequence (segment) x has length n . We suppose (and will argue this is ok later) that $r^x(n)$ is close to one of the vertices of C^{g-1} . We then choose a finer approximation C^g of C (since $V^g > V^{g-1}$).

For each vertex $p_v^g, v \in [V^g]$ we append elements to x resulting in a segment of length n' . We do this in a manner such that we move the relative frequency from $r^x(n)$ to $r^x(n') \approx p_v^g$. We do so by appending multiple copies of a vector z to x where $r^z(m)$ points in the same direction as the direction one needs to go from p_{old} to p_{new} . This can only be done approximately because with a finite length segment, the set of directions one can move the relative frequencies is quantized. We choose the fineness of the quantization to be fine

Algorithm 1 Construction of x such that $\text{CP}(r^x) = C$

Require: $C \subset \Delta^k$, a rectifiable closed curve parametrized as $c: [0, 1] \rightarrow \Delta^k$
Require: $V: \mathbb{N} \rightarrow \mathbb{N}$ $\triangleright$ Number of segments at generation g ; as a function of n
Require: $T: \mathbb{N} \rightarrow \mathbb{N}$ $\triangleright$ Controls quantization of angle; needs to be increasing

- 1: $x \leftarrow \langle 1 \rangle$ $\triangleright$ Arbitrary initialization $x_1 = 1$
- 2: $p_{\text{old}} \leftarrow e_1$ $\triangleright p_{\text{old}} = r^{(1)}(1) = e_1$
- 3: $n \leftarrow 1$ $\triangleright n$ is always updated to correspond to the current length of x
- 4: $g \leftarrow 1$
- 5: **while** true **do** $\triangleright$ Iterate over repeated generations g ; V is chosen at start of generation
- 6: $V \leftarrow V(g)$ $\triangleright$ Choose V for generation g
- 7: $p_v \leftarrow c(v/V)$ for $v = 0, \dots, V$ $\triangleright$ Vertices of $C^g := \bigcup_{v \in [V]} l(p_{v-1}, p_v)$
- 8: $v \leftarrow 0$
- 9: **while** $v \leq V$ **do** $\triangleright$ For all vertices of C^g
- 10: $T \leftarrow T(n)$ $\triangleright$ Quantization of angle; chosen per segment
- 11: $p_{\text{new}} \leftarrow p_{v+1}$ $\triangleright$ The next vertex of C^g
- 12: $\gamma_i \leftarrow p_{\text{old},i} / (p_{\text{old},i} - p_{\text{new},i})$ for $i \in [k]$ $\triangleright$ Will have $p_{\text{old}} \approx p_v$
- 13: $\gamma \leftarrow \min\{\gamma_i : i \in [k], \gamma_i > 0\}$ $\triangleright$ See (4.19)
- 14: $p^* \leftarrow \gamma(p_{\text{new}} - p_{\text{old}}) + p_{\text{old}}$ $\triangleright$ Determine $p^* \in \partial\Delta^k$
- 15: $\iota \leftarrow \lfloor Tp^* \rfloor$ $\triangleright$ Elementwise; $\iota = (\iota_1, \dots, \iota_k)$
- 16: $\tilde{p}^* \leftarrow \iota/T$ $\triangleright$ Elementwise; quantized version of p^*
- 17: $\tilde{T} \leftarrow \sum_{i=1}^k \iota_i$ $\triangleright$ Will have $\tilde{T} \approx T$
- 18: $y \leftarrow \langle 1^{\lfloor \iota_1 \rfloor}, \dots, k^{\lfloor \iota_k \rfloor} \rangle$ $\triangleright$ The string y is thus of length $\tilde{T}$
- 19: $\tilde{\ell} \leftarrow \lceil \frac{n}{T(\gamma-1)} \rceil$ $\triangleright$ Integer number of repetitions of y needed
- 20: $x \leftarrow x y^{\lfloor \tilde{\ell} \rfloor}$ $\triangleright$ Construct new x by appending z , comprising $\tilde{\ell}$ copies of y
- 21: $n \leftarrow n + \tilde{\ell}\tilde{T}$ $\triangleright$ Length of x now
- 22: $p_{\text{old}} \leftarrow r^x(n)$ $\triangleright$ Relative frequency at current n
- 23: $v \leftarrow v + 1$ $\triangleright$ Move onto next vertex of C^g
- 24: **end while**
- 25: $g \leftarrow g + 1$ $\triangleright$ Move onto next generation of the construction
- 26: **end while** $\triangleright$ Procedure never terminates

enough to achieve the accuracy we need. That is governed by the parameter $T \in \mathbb{N}$. We then append $\tilde{\ell}$ copies of z to x where $\tilde{\ell}$ is the integer closest to the real number ℓ that would be the ideal number of steps needed to get to the desired point p_{new} . We also control the error incurred by approximating ℓ by $\tilde{\ell}$. The upshot of this is that with the resulting extension to x we have $r^x(n')$ is sufficiently close to p_{new} . We then repeat this operation for all the vertices p_v^g for $v \in [V^g]$. This completes generation g . We show below that for each generation g , all the points in the relative frequency sequence are adequately close to C^g , where “adequately close” is quantified and increases in accuracy as g increases.

We consistently use the following terminology in describing our algorithm:

generation These are indexed by g and entail an entire pass around the curve C , or more precisely its polygonal approximation $C^g := \bigcup_{v \in [V^g]} l(p_{v-1}, p_v)$

segment Corresponds to a single line segment $l(p_{v-1}, p_v)$ of the g th polygonal approximation.

piece Corresponds to appending $z = \langle 1^{[l_1]}, \dots, k^{[l_k]} \rangle$ to x , which results in moving $r^x(n)$ in the direction $\hat{p}^*$.

step The appending of a single element of z , which will always move $r^x(n)$ towards one of the vertices of the simplex e_i ($i \in [k]$).

The end result is that we have constructed a procedure (Algorithm 1) which runs indefinitely (g increases without bound), and which has the property that for any choice of $\epsilon > 0$, if one waits long enough, there will be a sufficiently large g such that all the relative frequencies associated with generated x are within ϵ_g of C , and $(\epsilon_g)_{g \in \mathbb{N}}$ is a null sequence. We will thus conclude that $\text{CP}(r^x) \supseteq C$. We will also argue that $\text{CP}(r^x) \subseteq C$ completing the proof.

4.8.5 CONSTRUCTION OF p^* AND ITS APPROXIMATION $\tilde{p}^*$

The basic idea of the construction is to exploit Lemma 216. Suppose $n \in \mathbb{N}$ (and suppose it is “large”) and fix $m = 1$ in (4.16) to obtain

$$r^x(n+1) = \frac{n}{n+1}r^x(n) + \frac{1}{n+1}r^{x^{+n}}(1). \quad (4.17)$$

Now $r^{x^{+n}}(1) = e_{x(n+1)}$ and so $r^x(n+1) = \frac{n}{n+1}r^x(n) + \frac{1}{n+1}e_{x(n+1)}$. When n is large $\frac{n}{n+1} \approx 1$ and $\frac{1}{n+1}$ is small, and so this says that appending $x(n+1)$ to the length n segment $x([n])$ moves the relative frequency r^x from $r^x(n)$ in the direction of $e_{x(n+1)}$ by a small amount. Observe that the *only* directions which the point $r^x(n)$ can be moved is towards one of the vertices of the k -simplex, $e_1, \dots, e_k$. Thus if we had, for a fixed n that $r^x(n) = p_{\text{old}}$ and we wished to append m additional elements z to x to produce xz such that $r^{xz}(n+m) = p_{\text{new}}$, we need to figure out a way of heading in the direction $d = p_{\text{new}} - p_{\text{old}}$ when at each step we are constrained to move a small amount to one of the vertices. The solution is to approximate the direction d by a quantized choice that can be obtained by an integer number of elements of $[k]$.

Given arbitrary $p_{\text{old}} \neq p_{\text{new}} \in \text{relint } \Delta^k$, we define p^* to be the intercept by $\partial\Delta^k$ of the line segment starting at p_{old} and passing through p_{new} . (If $p_{\text{new}} \in \partial\Delta^k$ set $p^* = p_{\text{new}}$.) The intercept on the boundary of Δ^k is denoted p^* and is given by

$$p^* := \gamma(p_{\text{new}} - p_{\text{old}}) + p_{\text{old}} \quad (4.18)$$

for some $\gamma > 0$. We can determine γ as follows. The choice of γ can not take p^* outside the simplex. Thus let γ_i ($i \in [k]$) satisfy $\gamma_i(p_{\text{new}_i} - p_{\text{old}_i}) + p_{\text{old}_i} = 0$. Thus $\gamma_i = \frac{p_{\text{old}_i}}{p_{\text{old}_i} - p_{\text{new}_i}}$.

Any $\gamma_i < 0$ points in the wrong direction and so we choose

$$\gamma := \min\{\gamma_i : i \in [k] \text{ and } \gamma_i > 0\}. \quad (4.19)$$

Such a choice of γ guarantees that $p^* \in \partial\Delta^k$. Observe that the requirement that $\gamma_i > 0$ means the denominator in the definition of γ_i is positive and less than the numerator, and thus all γ_i which are positive exceed 1, and consequently $\gamma > 1$.

We can now take p^* to be the direction we would like to move $r^x(n)$ towards. However our only control action is to choose a sequence $z \in [k]^m$. To that end we suppose we quantize the vector p^* so that it has rational components with denominator $T \in \mathbb{N}$ (which will be strategically chosen henceforth). As we will shortly show, this will allow us to move (approximately) towards p^* . Thus let $\iota_i := \lfloor Tp_i^* \rfloor$ for $i \in [k]$ and set

$$\tilde{p}^* := \left(\frac{\iota_1}{T}, \dots, \frac{\iota_k}{T} \right). \quad (4.20)$$

Observe that $\tilde{p}^*$ is not guaranteed to be in Δ^k because there is no guarantee that $\sum_{i=1}^k \tilde{p}_i^* = 1$.

We have

Lemma 218. *Let p^* and $\tilde{p}^*$ be defined by (4.18) and (4.20) respectively. Then for all $i \in [k]$, $|\tilde{p}_i^* - p_i^*| \leq \frac{1}{2T}$.*

Proof $|\tilde{p}_i^* - p_i^*| = \frac{1}{T} |T\tilde{p}_i^* - Tp_i^*| = \frac{1}{T} |\lfloor Tp_i^* \rfloor - Tp_i^*| \leq \frac{1}{2T}$, by definition of the rounding operator $\lfloor \cdot \rfloor$. ■

4.8.6 DETERMINING THE NUMBER OF STEPS TO TAKE

Observe that (4.16) can be written as

$$r^x(n+m) = (1-\alpha)r^x(n) + \alpha r^{x+n}(m). \quad (4.21)$$

where $\alpha = \frac{m}{n+m}$ and thus $1-\alpha = \frac{n}{n+m}$. This suggests that we can engineer the construction of x by requiring a suitable α such that

$$(1-\alpha)p_{\text{old}} + \alpha p^* = p_{\text{new}}.$$

Recall we want the sequence r^x to move from p_{old} to p_{new} which can be achieved by a taking a suitable convex combination of p_{old} and p^* , which corresponds to appending a suitable number of copies of z to x , where z is chosen to move $r^x(n)$ in the direction of p^* .

If we substitute the definition of p^* from (4.18) we obtain the problem:

$$\begin{aligned} & \text{Find } \alpha \text{ such that } (1 - \alpha)p_{\text{old}} + \alpha[\gamma(p_{\text{new}} - p_{\text{old}}) + p_{\text{old}}] = p_{\text{new}} \\ \Leftrightarrow & \text{Find } \alpha \text{ such that } (1 - \alpha)p_{\text{old}} + \alpha\gamma p_{\text{new}} - \alpha\gamma p_{\text{old}} + \alpha p_{\text{old}} - p_{\text{new}} = 0_k \in \mathbb{R}^k \\ \Leftrightarrow & \text{Find } \alpha \text{ such that } p_{\text{old}}(1 - \alpha\gamma) + p_{\text{new}}(\alpha\gamma - 1) = 0_k, \end{aligned}$$

which is only true when either $p_{\text{old}} = p_{\text{new}}$ (which is a trivial case) or when $1 - \alpha\gamma = 0$ and thus $\alpha := 1/\gamma$, which we take as a definition. Since $\gamma > 1$ this implies $\alpha < 1$, which is consistent with our original motivation for taking convex combinations.

We will append the segment z to x , where $z = y^{[\ell]}$ and $y = \langle 1^{[l_1]}, \dots, k^{[l_k]} \rangle$. Now if each y is of length T and we notionally made ℓ repetitions, we would have $m = \ell T$. From the definition of α this means

$$\alpha = \frac{\ell T}{n + \ell T}. \quad (4.22)$$

We presume n is given (at a particular stage of construction) and $T \in \mathbb{N}$ is a fixed design parameter. We can thus solve for ℓ to obtain

$$\ell = \frac{\alpha n}{T(1 - \alpha)} = \frac{(1/\gamma)n}{T(1 - 1/\gamma)} = \frac{n}{T(\gamma - 1)}.$$

Observe that ℓ is not guaranteed to be an integer, a complication we will deal with later. If it was an integer, we would create z by concatenating ℓ copies of y which is of length T . The vector y moves p_{old} towards p^* . By appending ℓ copies we should move r^x to p_{new} as desired.

However we do not head exactly in the direction of p^* , since we worked with a quantized version $\tilde{p}^*$ instead, and we can not always take ℓ copies because ℓ is not guaranteed to be an integer; instead we will take

$$\tilde{\ell} := \left\lceil \frac{\alpha n}{T(1 - \alpha)} \right\rceil = \left\lceil \frac{n}{T(\gamma - 1)} \right\rceil$$

copies of y which will move $r^x(n)$ towards $\hat{p}^*$ instead of p^* . We now proceed to analyse the effect of these approximations on our construction.

4.8.7 ANALYZING THE EFFECT OF APPROXIMATIONS

The parameter $T \in \mathbb{N}$ is a design variable. Our construction will utilize

$$\tilde{T} := \sum_{j=1}^k l_j, \quad (4.23)$$

where $\iota_j = \lfloor Tp_j^* \rfloor$ for $j \in [k]$. Then $\tilde{T} \approx T$, a claim which we quantify below.

Lemma 219. *Suppose $T \in \mathbb{N}$, and $\tilde{T}$ is defined as above. Then $T - \frac{k}{2} \leq \tilde{T} \leq T + \frac{k}{2}$.*

Proof By definition of the rounding operator $\lfloor \cdot \rfloor$, we have that

$$\left| \iota_j - \lfloor Tp_j^* \rfloor \right| \leq \frac{1}{2} \quad \forall j \in [k].$$

Thus

$$\begin{aligned} Tp_j^* - \frac{1}{2} &\leq \iota_j \leq Tp_j^* + \frac{1}{2} \quad \forall j \in [k] \\ \Rightarrow \sum_{j=1}^k \left(Tp_j^* - \frac{1}{2} \right) &\leq \sum_{j=1}^k \iota_j \leq \sum_{j=1}^k \left(Tp_j^* + \frac{1}{2} \right) \\ \Rightarrow T - \frac{k}{2} &\leq \tilde{T} \leq T + \frac{k}{2}. \end{aligned}$$

■

Ideally we move $r^x(n)$ to

$$p_{\text{new}} = (1 - \alpha)p_{\text{old}} + \alpha p^*. \quad (4.24)$$

by appending z (i.e. we hope that $r^{xz}(n + m) = p_{\text{new}}$). But in fact the segment z which we will append to x will move r^x from p_{old} instead to

$$\hat{p}_{\text{new}} := (1 - \tilde{\alpha})p_{\text{old}} + \tilde{\alpha} \hat{p}^*, \quad (4.25)$$

where

$$\tilde{\alpha} := \frac{\tilde{\ell}T}{\tilde{\ell}T + m}. \quad (4.26)$$

and

$$\hat{p}^* := \left(\frac{\iota_1}{\tilde{T}}, \dots, \frac{\iota_k}{\tilde{T}} \right). \quad (4.27)$$

We now determine the error incurred from these approximations. We first need the following Lemma:

Lemma 220. *Suppose $n, T \in \mathbb{N}$, α is defined by (4.22) and $\tilde{\alpha}$ is defined by (4.26). Then*

$$|\alpha - \tilde{\alpha}| \leq \frac{T}{n}.$$

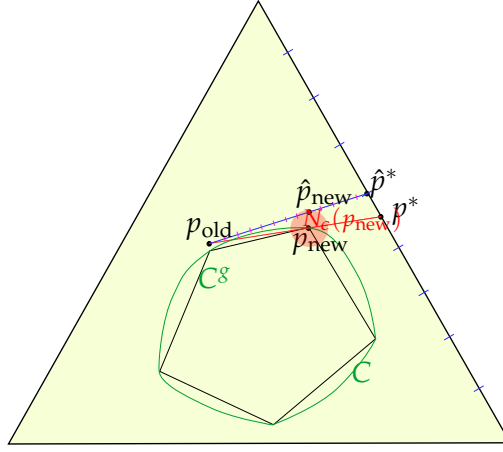


Figure 4.1: Illustration of construction of the sequence x . The figure shows how a single segment is created. We have the desired curve C (in dark green) and a polygonal approximation C^g using 5 segments (thus $V^g = 5$). We assume that we have already constructed the first n elements of x and thus we can compute $r^x(n)$. Denote this by p_{old} , an element of the simplex, itself shown in pale green. We hope to append a segment y of length m to x such $r^{xy}(n+m) = p_{\text{new}}$. However we can only choose elements of y from $[k]$ and that means that at each step we move towards one of the vertices of the simplex. We note that ideally we would move from p_{old} towards $p^* \in \partial\Delta^k$. In order to deal with the restrictions on directions we can head, we quantize p^* as $\hat{p}^* = \lfloor Tp^* \rfloor / T$ where in this example we have chosen $T = 9$. As argued in the main text this means we will thus be restricted to heading towards one of a fixed set of points on the boundary (marked in blue). Observe $\hat{p}^*$ is located at one of these points in the diagram, but in general it might not even be on the boundary of the simplex. We then construct y to move r^x towards $\hat{p}^*$ and take sufficient steps to move to $\hat{p}_{\text{new}}$. This too is done in repeated steps by setting $y = z^{[\tilde{\ell}]}$ where z is a shorter segment (marked by purple ticks on the line segment $l(p_{\text{old}}, \hat{p}^*)$) which will move towards $\hat{p}^*$ by a small amount. The end result is that we get $r^{xy}(n+m) = \hat{p}_{\text{new}}$ which is contained within $N_\epsilon(p_{\text{new}})$, an ϵ -ball centered at p_{new} .

Proof By definition of $\tilde{\alpha}$ we have

$$\begin{aligned} |\alpha - \tilde{\alpha}| &= \left| \frac{\tilde{\ell}T}{\tilde{\ell}T + n} - \frac{\ell T}{\ell T + n} \right| \\ &= \left| \frac{\tilde{\ell}T(\ell T + n) - \ell T(\tilde{\ell}T + n)}{(\tilde{\ell}T + n)(\ell T + n)} \right| \\ &= \left| \frac{\tilde{\ell}Tn - \ell Tn}{(\tilde{\ell}T + n)(\ell T + n)} \right|. \end{aligned}$$

Since $(\tilde{\ell}T + n)(\ell T + n) > 0$ and $\tilde{\ell} = \lceil \ell \rceil \geq \ell$, we have

$$(\tilde{\ell}T + n)(\ell T + n) \geq (\ell T + n)(\ell T + n) \geq n^2$$

and since $|\tilde{\ell} - \ell| \leq 1$,

$$|\alpha - \tilde{\alpha}| = \frac{|\tilde{\ell} - \ell| \cdot Tn}{n^2} \leq \frac{T}{n}.$$

■

Our construction does not move $r^x(n)$ towards $\tilde{p}^*$ but an approximation of it, namely $\hat{p}^*$ defined in (4.27). We exploit the fact that repeating a segment does not change its relative

frequencies, which we state formally as

Lemma 221. *Let $z = \langle 1^{[l_1]}, \dots, k^{[l_k]} \rangle$. Then $r^z(\tilde{T}) = \hat{p}^* = r^{z^{[\tilde{\ell}]}}(\tilde{\ell}\tilde{T})$.*

Proof For any $i \in [k]$ we have $r_i^z(\tilde{T}) = \frac{1}{\tilde{T}} |\{j \in [\tilde{T}] : z_j = i\}| = \frac{1}{\tilde{T}} l_i$. The first equality is immediate. For the second, similarly we have $r_i^{z^{[\tilde{\ell}]}}(\tilde{\ell}\tilde{T}) = \frac{1}{\tilde{\ell}\tilde{T}} |\{j \in [\tilde{\ell}\tilde{T}] : z_j^{[\tilde{\ell}]} = i\}| = \frac{1}{\tilde{\ell}\tilde{T}} \cdot \tilde{\ell} l_i = \frac{1}{\tilde{T}} l_i$ by definition of $z^{[\tilde{\ell}]}$. ■

Since $\tilde{p}^* = (\frac{l_1}{\tilde{T}}, \dots, \frac{l_k}{\tilde{T}})$ we have that $\tilde{p}^* = \frac{\tilde{T}}{T} \hat{p}^*$. This allows us to show:

Lemma 222. *Suppose $T \in \mathbb{N}$ and $\hat{p}^*$ is defined via (4.27). Then $\|\hat{p}^* - \tilde{p}^*\| \leq \frac{k}{2T}$.*

Proof We have $\|\hat{p}^* - \tilde{p}^*\| = \left\| \hat{p}^* - \frac{\tilde{T}}{T} \hat{p}^* \right\| = \left| 1 - \frac{\tilde{T}}{T} \right| \cdot \|\hat{p}^*\| \leq \left| 1 - \frac{\tilde{T}}{T} \right|$. Suppose $\tilde{T} < T$, then $1 - \frac{\tilde{T}}{T} > 0$ and $\left| 1 - \frac{\tilde{T}}{T} \right| = 1 - \frac{\tilde{T}}{T} \leq 1 - \frac{T-k/2}{T} = \frac{k}{2T}$ by Lemma 219. Similarly if $\tilde{T} > T$, then $1 - \frac{\tilde{T}}{T} < 0$ and $\left| 1 - \frac{\tilde{T}}{T} \right| = \frac{\tilde{T}}{T} - 1 \leq \frac{T+k/2}{T} - 1 = \frac{k}{2T}$ completing the proof. ■

Lemma 223. *Suppose $k, T \in \mathbb{N}$ and p_{new} and $\hat{p}_{\text{new}}$ are defined as above. Then*

$$\|\hat{p}_{\text{new}} - p_{\text{new}}\| \leq \frac{4T}{n} + \frac{k}{T}. \quad (4.28)$$

Proof From (4.24) and (4.25) we have

$$\begin{aligned} \|\hat{p}_{\text{new}} - p_{\text{new}}\| &= \|(1 - \tilde{\alpha})p_{\text{old}} + \tilde{\alpha}\hat{p}^* - (1 - \alpha)p_{\text{old}} - \alpha p^*\| \\ &= \|[(1 - \tilde{\alpha}) - (1 - \alpha)]p_{\text{old}} + (\tilde{\alpha}\hat{p}^* - \alpha p^*)\| \\ &\leq \|(\alpha - \tilde{\alpha})p_{\text{old}}\| + \|\tilde{\alpha}\hat{p}^* - \alpha p^*\| \\ &\leq \sqrt{2}|\tilde{\alpha} - \alpha| + \|\tilde{\alpha}\hat{p}^* - \alpha p^*\|. \end{aligned} \quad (4.29)$$

The second term in (4.29) can be bounded as follows:

$$\begin{aligned} \|\tilde{\alpha}\hat{p}^* - \alpha p^*\| &= \|(\tilde{\alpha} - \alpha + \alpha)\hat{p}^* - \alpha p^*\| \\ &= \|(\tilde{\alpha} - \alpha)\hat{p}^* + (\alpha\hat{p}^* - \alpha p^*)\| \\ &\leq \|(\tilde{\alpha} - \alpha)\hat{p}^*\| + \|\alpha\hat{p}^* - \alpha p^*\| \\ &= |\tilde{\alpha} - \alpha| \cdot \|\hat{p}^*\| + \alpha \|\hat{p}^* - p^*\| \\ &= |\tilde{\alpha} - \alpha| \cdot \|\hat{p}^*\| + \alpha \|(\hat{p}^* - \tilde{p}^*) + (\tilde{p}^* - p^*)\| \\ &\leq |\tilde{\alpha} - \alpha| \sqrt{2} + \alpha \|\hat{p}^* - \tilde{p}^*\| + \alpha \|\tilde{p}^* - p^*\| \\ &\leq \sqrt{2}|\tilde{\alpha} - \alpha| + \frac{k}{2T} + \alpha \left(\sum_{i=1}^k (\tilde{p}_i^* - p_i^*)^2 \right)^{1/2}, \end{aligned}$$

by Lemma 222 and the fact that $\|\hat{p}^*\| \leq 1$,

$$\begin{aligned} &\leq \sqrt{2}|\tilde{\alpha} - \alpha| + \frac{k}{2T} + \alpha \left(\sum_{i=1}^k \left(\frac{1}{2T} \right)^2 \right)^{1/2} \\ &= \sqrt{2}|\tilde{\alpha} - \alpha| + \frac{k}{2T} + \alpha \frac{\sqrt{k}}{2T}, \end{aligned} \tag{4.30}$$

where we used Lemma 218 in the penultimate step. Since $\alpha \leq 1$, combining (4.29) and (4.30) we have

$$\|\hat{p}_{\text{new}} - p_{\text{new}}\| \leq 2\sqrt{2}|\tilde{\alpha} - \alpha| + \frac{k}{2T} + \frac{\sqrt{k}}{2T} \leq 4|\tilde{\alpha} - \alpha| + \frac{k}{T}.$$

Appealing to Lemma 220 gives us (4.28). ■

The above arguments control the errors at the end of a *piece* (and thus in a *segment*). But for later purposes we need control at each *step*. This follows immediately by the fact that we make small steps:

Lemma 224. *For $n \in \mathbb{N}$ and $m \in [\tilde{T}]$ and any $x: \mathbb{N} \rightarrow [k]$,*

$$\|r^x(n) - r^x(n+m)\| \leq \frac{2T+k}{n}.$$

Proof By Lemma 216,

$$\begin{aligned} \|r^x(n) - r^x(n+m)\| &= \left\| r^x(n) - \frac{n}{n+m}r^x(n) - \frac{m}{n+m}r^{x^{+n}}(m) \right\| \\ &= \left\| \left(1 - \frac{n}{n+m}\right)r^x(n) - \frac{m}{n+m}r^{x^{+n}}(m) \right\| \\ &= \frac{m}{n+m} \left\| r^x(n) - r^{x^{+n}}(m) \right\| \\ &\leq \frac{2m}{n+m} \\ &\leq \frac{2m}{n} \\ &\leq \frac{2T+k}{n}, \end{aligned}$$

where the first inequality holds since $\|r^x(n)\|, \|r^{x^{+n}}(m)\| \leq 1$ and the last step follows from Lemma 219. ■

4.8.8 COMPLETING THE PROOF OF THEOREM 194

The remaining piece of the argument concerns the piecewise linear approximation of C by C^g . For $A, B \subseteq \Delta^k$ and $a \in \Delta^k$ define $d(a, B) := \min\{\|a - b\| : b \in B\}$ and the Hausdorff distance

$$d(A, B) := \max\{d(a, B) : a \in A\}. \quad (4.31)$$

Let us write V^g (the number of vertices in the piecewise linear approximation at generation g) in functional form as $V(g)$. Let $n(g)$ denote the length of the segment of x that has been constructed at the beginning of generation g , and let $g(n) := \inf\{g \in \mathbb{N} : n \leq n(g)\}$ denote its quasi-inverse. Clearly $n(g)$ is strictly increasing in g and $g(n)$ is increasing, but often constant. With these definitions, we have $V = V(g) = V(g(n))$.

Denote by $\tilde{C}(V)$ the best piecewise linear approximation of C with V vertices, in the sense of minimizing $\psi_C(V) := d(C, \tilde{C}(V))$. Since every rectifiable curve C has a Lipschitz continuous parametrisation, we have that $V \mapsto \psi_C(V)$ is decreasing in V and $\lim_{V \rightarrow \infty} \psi_C(V) = 0$. Thus $\lim_{g \rightarrow \infty} \psi_C(V(g)) = 0$ and $\lim_{n \rightarrow \infty} \psi_C(V(g(n))) = 0$, although the convergence could be very slow (in n) and its speed will depend on the choice of C . Denote by $\bar{C}(n) := \tilde{C}(V(g(n)))$ the sequence of best possible piecewise linear approximations of C indexed by n , and let $\bar{\psi}_C(n) := d(C, \bar{C}(n))$. We have thus shown:

Lemma 225. *Let $C \subset \Delta^k$ be a rectifiable curve. Then*

$$\lim_{n \rightarrow \infty} \bar{\psi}_C(n) = 0.$$

We summarize what we know so far.

1. For all generations g , $\|\hat{p}_{\text{new}} - p_{\text{new}}\| \leq \frac{4T}{n} + \frac{k}{T}$, where $p_{\text{new}} = p_v$ for $v \in [V^g]$ (Lemma 223). This means the following. Suppose at the beginning of segment v in generation g we have $n = \text{length}(x)$. By definition, we have $r^x(n) = p_{\text{old}}$ and $r^{xz^{[\tilde{\ell}]}}(n + \tilde{\ell}\tilde{T}) = \hat{p}_{\text{new}}$. Furthermore, for $m = i\tilde{T}$, $i \in [\tilde{\ell}]$ we have $r^{xz^{[i]}}(n + i\tilde{T}) \in l(p_{\text{old}}, \hat{p}_{\text{new}})$.
2. Furthermore, (by Lemma 224) for all $j \in [\tilde{\ell}\tilde{T}]$, $d\left(r^{xz^{[j]}}(j), l(p_{\text{old}}, \hat{p}_{\text{new}})\right) \leq \frac{2T+k}{n}$ — the relative frequencies for all points in the segment are close to the line segment $l(p_{\text{old}}, \hat{p}_{\text{new}})$.

Combining these facts, and appealing to the triangle inequality, we conclude that the sequence x constructed by Algorithm 1 satisfies

$$d(r^x(n), C) \leq \frac{4T}{n} + \frac{k}{T} + \frac{2T+k}{n} + \bar{\psi}_C(n) \quad \forall n \in \mathbb{N}. \quad (4.32)$$

We now choose $T = T(n)$ and $V = V(g(n))$ appropriately. One choice is to choose $T(n) =$

$\sqrt{n}$. Equation 4.32 then implies

$$d(r^x(n), C) \leq \frac{4\sqrt{n}}{n} + \frac{k}{\sqrt{n}} + \frac{2\sqrt{n}+k}{n} + \bar{\psi}_C(n) \quad \forall n \in \mathbb{N}$$

which implies that for all $n \in \mathbb{N}$ with $\sqrt{n} > k$,

$$d(r^x(n), C) \leq \frac{4}{\sqrt{n}} + \frac{k}{\sqrt{n}} + \frac{4}{\sqrt{n}} + \bar{\psi}_C(n) = \frac{8+k}{\sqrt{n}} + \bar{\psi}_C(n),$$

and thus Lemma 225 implies $\lim_{n \rightarrow \infty} d(r^x(n), C) = 0$. By definition of the Hausdorff distance this means that for any $\epsilon > 0$ and any point $y \in C$, there exists an n such that $\|r^x(n) - y\| \leq \epsilon$. Furthermore, by the generational nature of our construction, if $r^x(n)$ is ϵ -close to y then for each subsequent generation g , there exists n^g such that $r^x(n^g)$ is also ϵ -close to y . Since there are an infinite number of generations g , the sequence r^x visits an ϵ -neighbourhood of y infinitely often. Since ϵ was arbitrary, y is thus a cluster point of r^x . Since $y \in C$ was arbitrary, every element of C is a cluster point of r^x .

Finally, since in each generation the bounds above constrain r^x more and more tightly, there cannot exist cluster points that are not in C ; that is, $\text{CP}(r) \subseteq C$. We have thus proved Theorem 194.

4.8.9 REMARKS ON THE CONSTRUCTION

We make a few remarks on the construction.

1. By the definition of $\tilde{\ell}$ we are guaranteed that $\tilde{\ell} \geq 1$ for each segment and each generation. For the construction to approximate well, we need $\tilde{\ell} \gg 1$, which it will inevitably be when n gets large enough.
2. Recall $n(g)$ is the length of x at the beginning of generation g . Let $L_C^g := \text{length}(C^g) > 0$. Since each step which r^x moves is of size less than $1/n(g)$ (see Equation 4.17), we require at least $L_C^g \cdot n(g)$ steps for r^x to traverse the whole of C in generation g . Thus at the end of generation g and the beginning of generation $g+1$ we have

$$n(g+1) \geq n(g) + L_C^g \cdot n(g) = \lambda_C^g \cdot n(g).$$

Furthermore, L_C^g is increasing in g and approaches $\text{length}(C)$. Thus for all g , $\lambda_C^g > 1$ (and is in fact increasing in g). Hence $n(g)$ grows exponentially with g .

3. The growth of the length of x is controlled in a complex manner by the nature of the curve C . In particular if C is very complex, then $\bar{\psi}_C$ must decay slowly. Furthermore, if C has parts close to $\partial\Delta^k$, in particular if some vertices v of the piecewise linear approximation C^g are close to $\partial\Delta^k$, then $\tilde{\ell}$ can end up very large for that segment,

meaning that the length of the sequence x grows more rapidly. See Lemma 226 for an illustration of this observation.

Finally note that since (by Lemma 217) $\text{CP}(r^x)$ must always be connected, we have in Theorem 194 what appears to be the most general result possible (under the restriction that x takes values only in a finite set $[k]$). We do not know what the appropriate generalization is to sequences x that can take values in an infinite (or uncountable) set²⁷.

4.8.10 FROM BOUNDARY TO CURVES

Proof [Proof of Corollary 195]

[Bronshteyn and Ivanov \[1975\]](#) show that if D is a convex set contained in the unit ball (w.r.t. the Euclidean norm) in $\mathbb{R}^n$ and $\epsilon < 10^{-3}$, then there exists a set of at most $K_\epsilon := 3\sqrt{n}(9/\epsilon)^{(n-1)/2}$ points whose convex hull is at most ϵ away from D . Thus for any convex $D \subseteq \Delta^k$, and any $\epsilon < 10^{-3}$ there is a polyhedron D_ϵ comprising the convex hull of $K_\epsilon < \infty$ points $Q_\epsilon := \{q_i \in \Delta^k : i \in [K_\epsilon]\}$ such that $d(D, D_\epsilon) < \epsilon$, where d is the Hausdorff distance (4.31). Hence for any $\epsilon < 10^{-3}$ there exists a closed rectifiable curve C_ϵ (constructed by linearly connecting successive points in Q_ϵ) such that $d(\text{co } C_\epsilon, D) < \epsilon$. One can then construct a sequence x such that $\text{CP}(r^x) = \partial D$ as follows. Start with $\epsilon_0 = 10^{-3}$. Pick and construct C_{ϵ_0} as above. Construct x according to the previous procedure (Appendix 4.8.4–4.8.9) to get an entire generation of r^x within ϵ_0 of C_{ϵ_0} . Then let $\epsilon_1 = \epsilon_0/2$ and repeat the procedure, appending the constructed sequence. Continue iterating (dividing the ϵ in half each phase) and one achieves that in the limit $\text{CP}(r^x) = \partial D$. ■

4.8.11 ILLUSTRATION

We illustrate our construction for $k = 3$ for two (identical) generations and thus a single piecewise linear (in fact polygonal) approximation C^g of C . We take for C (a scaled version of) the lemniscate of Bernoulli [[Lockwood, 1961](#), Chapter 12], [[Lawrence, 1972](#), Section 5.3]

²⁷Although we do not pursue this in any detail, we remark that one could design an algorithm to construct x such that $\text{CP}(r^x)$ is *any* subset $D \subseteq \Delta^k$ by using our algorithm as a subroutine. The idea would be to construct a space filling curve that fills D , each generation of which is a rectifiable curve. One would appeal to our algorithm for each generation, and then once within a suitable tolerance, change the target to be the next generation of the space filling curve. A suitable method would be to simply intersect an extant families of closed space filling curves for k dimensional cubes with the $(k-1)$ -simplex (e.g. generalisations of the Moore curve), attaching joins on ∂D where necessary. Since (see the main body of the paper) it ends up being only the convex hull of $\text{CP}(r^x)$ that matters, such an exotic construction is of little direct interest.

mapped onto the 2-simplex as the space curve $\{(z_1(t), z_2(t), z_3(t)) : t \in [0, 2\pi]\}$, where

$$\begin{aligned} z_1(t) &= \frac{1}{3} + \frac{1}{12} \frac{2 \cos(t)}{1 + \sin^2(t)} \\ z_2(t) &= \frac{1}{3} + \frac{1}{12} \frac{2 \sin(t) \cos(t)}{1 + \sin^2(t)} \\ z_3(t) &= 1 - z_1(t) - z_2(t). \end{aligned}$$

We set $T = 12$ and $V = 30$ (number of nodes in the polygonal approximation C^g for both $g = 1, 2$ to make the figure less cluttered) and we iterated long enough to go around the lemniscate twice, which resulted in a sequence of length 85677. As the construction proceeded, $\tilde{T}$ went from 9 up to 164 and $\tilde{\ell}$ went from 1 or 2 for the first few segments up to 69 for the last (with the largest value being 104). The results can be seen in Figure 4.2 which plots the achieved relative frequencies at different zoom levels. The small red squares are the vertices of C^g .

4.8.12 CONSTRUCTION OF x SUCH THAT $\text{co CP}(r^x) = \Delta^k$

How much of the simplex can we cover with $\text{CP}(r^x)$? This question is poorly posed as (it seems) we can only ever construct one dimensional sets that are the set of cluster points of r^x . However, for inducing an upper prevision, all that matters is the convex hull of $\text{CP}(r^x)$. We now show the convex hull of $\text{CP}(r^x)$ can be made as large as conceivable with simpler and more explicit construction:

Lemma 226. *Suppose $k \in \mathbb{N}$. There exists a sequence $x : \mathbb{N} \rightarrow [k]$ such that $\text{co}(\text{CP}(r^x)) = \Delta^k$.*

Proof As before, our proof is constructive. Recall $e_1, \dots, e_k$ are the vertices (and extreme points) of the $(k-1)$ -simplex, and $\text{co}\{e_1, \dots, e_k\} = \Delta^k$. We will construct a sequence x such that for all $\epsilon > 0$, and all $i \in [k]$, r^x visits $N_\epsilon(e_i)$ infinitely often. Since $\lim_{\epsilon \rightarrow 0} \text{co} \bigcup_{i \in [k]} N_\epsilon(e_i) = \Delta^k$ we will have achieved the desired result.

We again make use of (4.16). We will construct a sequence x by adding segments (s) such that for each successive s we drive r^x closer and closer towards one of the vertices e_i ($i \in [k]$). In order to do this, at each s we append m copies of i to the current x . Specifically, we construct x as follows:

We need to make m large enough so that the convex combination coefficient $\frac{m}{n+m}$ approaches 1. To that end, consider an increasing function $\phi : \mathbb{N} \rightarrow \mathbb{N}$ which will further restrict later. The role of ϕ is to control n as a function of segment number s ; that is $n = \phi(s)$ and thus $m = \phi(s+1) - \phi(s)$. With this choice, we have

$$\frac{n}{n+m} = \frac{\phi(s)}{\phi(s+1)} \quad \text{and} \quad \frac{m}{n+m} = 1 - \frac{\phi(s)}{\phi(s+1)}.$$

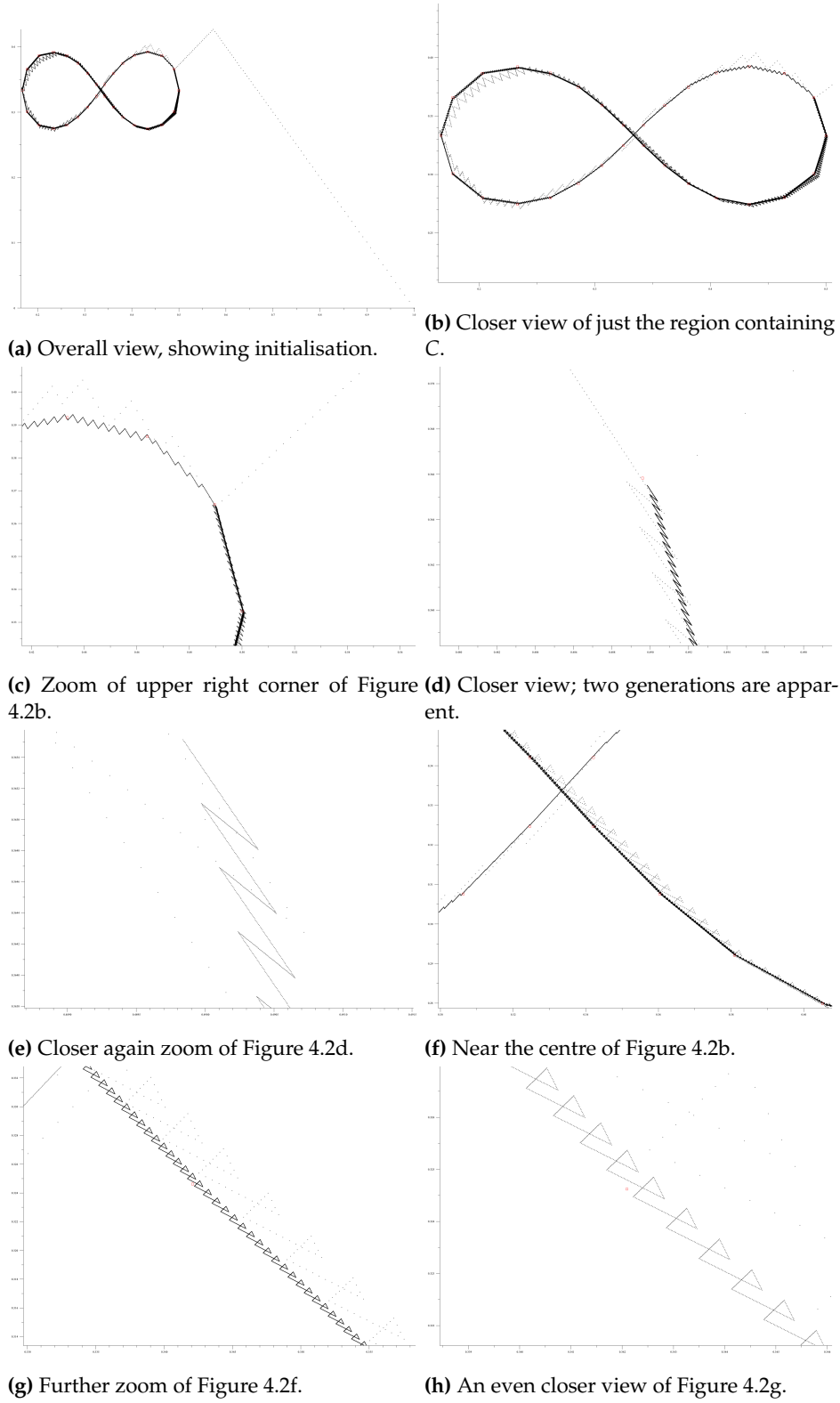


Figure 4.2: Illustration of approximation of the polygonal curve C^8 by the relative frequencies of the sequence constructed according to Algorithm 1. Two generations were used. Red squares are vertices of C^8 . See Section 4.8.11 for details.

Require: $\phi: \mathbb{N} \rightarrow \mathbb{N}$

1: $x \leftarrow \langle \rangle$

2: $s \leftarrow 1$

3: **while** true **do**

4: $i \leftarrow s \bmod k$

▷ Cycle around the vertices of the simplex

5: $m \leftarrow \phi(s+1) - \phi(s)$

6: $x \leftarrow x i^m$

▷ Append m copies of i to x

7: $s \leftarrow s + 1$

8: **end while**

and thus for all $s \in \mathbb{N}$

$$r^x(\phi(s+1)) = \frac{\phi(s)}{\phi(s+1)} r^x(\phi(s)) + \left(1 - \frac{\phi(s)}{\phi(s+1)}\right) r^{x+\phi(s)}(\phi(s+1) - \phi(s)). \quad (4.33)$$

We demand that $\lim_{s \rightarrow \infty} \frac{\phi(s)}{\phi(s+1)} = 0$ so that as s increases, the second term in (4.33) dominates. For any $s \in \mathbb{N}$, $r^x(\phi(s)) \in \Delta^k$ and thus

$$\begin{aligned} \|e_{s \bmod k} - r^x(\phi(s+1))\| &= \left\| e_{s \bmod k} - \frac{\phi(s)}{\phi(s+1)} r^x(\phi(s)) - \left(1 - \frac{\phi(s)}{\phi(s+1)}\right) r^{x+\phi(s)}(\phi(s+1) - \phi(s)) \right\| \\ &= \left\| \frac{\phi(s)}{\phi(s+1)} e_{s \bmod k} - \frac{\phi(s)}{\phi(s+1)} r^x(\phi(s)) \right\| \\ &= \frac{\phi(s)}{\phi(s+1)} \|e_{s \bmod k} - r^x(\phi(s))\| \\ &\leq \frac{\phi(s)}{\phi(s+1)} \cdot 2, \end{aligned}$$

where the second line follows from the fact that we constructed x such that $r^{x+\phi(s)}(\phi(s+1) - \phi(s)) = e_{s \bmod k}$. But by assumption, $\lim_{s \rightarrow \infty} \frac{\phi(s)}{\phi(s+1)} = 0$ and hence for any $\epsilon > 0$ there exists s_ϵ such that for all $i \in [k]$,

$$|\{s \in \mathbb{N} : s > s_\epsilon, s \bmod k = i, \|e_i - r^x(\phi(s+1))\| \leq \epsilon\}| = \aleph_0.$$

That is, for each $i \in [k]$, each ϵ -neighbourhood of e_i is visited infinitely often by the sequence r^x and hence $\{e_1, \dots, e_k\} \subseteq \text{CP}(r^x)$. But since $r^x(n) \in \Delta^k$ for all $n \in \mathbb{N}$ we conclude that indeed $\text{co}(\text{CP}(r^x)) = \text{co}(\{e_1, \dots, e_k\}) = \Delta^k$ as required. ■

A suitable choice of ϕ is $\phi(s) = \lceil \exp(s^\alpha) \rceil$ for some $\alpha > 1$, in which case $\frac{\phi(s)}{\phi(s+1)} \approx \exp(-\alpha s^{\alpha-1})$. An argument as in the proof of Theorem 194 would show that $S_k := \bigcup_{i \in [k]} l(e_i, e_{(i+1) \bmod k}) \subseteq \text{CP}(r^x)$. Observe that when $k = 3$, $S_k = \partial \Delta^k$, but for $k > 4$, that is not true, even though $\text{co}(S_k) = \partial \Delta^k$.

4.9 APPENDIX: UNSTABLE INDEPENDENCE

Closely related to conditional probability is the concept of *statistical independence*. Independence plays a central role not only in Kolmogorov's [Durrett, 2019, p. 37], but more generally in most probability theories (Levin [1980]; Fine [1973, Section IIF, IIIG and VH]). Already de Moivre [1738/1967, Introduction, p. 6] nicely summarized a pre-theoretical, probabilistic notion of real-world independence:

Two events are independent, when they have no connexion one with the other, and that the happening of one neither forwards nor obstructs the happening of the other.

This intuitive conception was then formalized by Kolmogorov [1933] (translated in [Kolmogorov, 1956]) in the following classical definition.

Definition 227. Let $(\Omega, \mathcal{F}, P)$ be a probability space.²⁸ We call events $A \in \mathcal{F}$ and $B \in \mathcal{F}$ classically independent if $P(A \cap B) = P(A)P(B)$.

If $P(B) > 0$, then we can equivalently express this condition as $P(A|B) = P(A)$ by using the definition of conditional probability.

Kolmogorov's definition is formal and it has been questioned whether it is an adequate expression of what we mean by independence in a statistical context [Von Collani, 2006]. As it is stated in purely measure-theoretic terms, it is unclear whether it has reasonable frequentist semantics. In our framework, we construct an intuitive definition of independence, where the independence of events is based on an independence notion of *processes* (cf. [von Mises and Geiringer, 1964, p. 35-39]). Therefore, our definition is thoroughly grounded in the frequentist setting. Furthermore, we shall generalize the independence concept to the case of possible divergence, where new subtleties come into play. We will then consider how our definitions relate to the classical case when relative frequencies converge. Assume that a sequence $\vec{\Omega}$ is given and we have constructed an upper probability $\bar{P}$ as in Section 4.2.2.

Definition 228. We call an event $B \in 2_{1+}^{\vec{\Omega}}$ irrelevant to another event $A \subseteq \Omega$ if:

$$\bar{P}(A|B) = \bar{P}(A).$$

This definition captures the concept of *epistemic irrelevance* in the imprecise probability literature [Miranda, 2008]. Why does this definition possess reasonable frequentist semantics? Consider what $\bar{P}(A|B)$ means (see Section 4.4.1): we are considering a subsequence,

²⁸Here, Ω is the possibility space, $\mathcal{F}$ is a σ -algebra and P is a countably additive probability measure.

induced by the indicator gamble χ_B , that is, we condition (in an intuitive sense) on the occurrence of B ; and on this subsequence, we then consider an unconditional upper probability. If this then coincides with the original upper probability, our decision maker values A just the same whether B occurs or not. Thus B is irrelevant for putting a value on A . In contrast to the classical, precise case, irrelevance is not necessarily symmetric. Hence, we define independence as follows.

Definition 229. Let $A, B \in 2_{1+}^{\bar{\Omega}}$. We call A and B independent if $\bar{P}(A|B) = \bar{P}(A)$ and $\bar{P}(B|A) = \bar{P}(B)$.

Thus, we have obtained a grounded concept of independence for events. We note that Definition 229 is similar to a condition proposed by Walley and Fine [1982] for independence of joint experiments²⁹; they did not propose an independence concept for gambles.

How can we extend this to an irrelevance and independence concept for gambles? First, we briefly recall how this is done in the classical case.

Definition 230. Let $(\Omega, \mathcal{F}, P)$ be a probability space and fix the Borel σ -algebra $\mathcal{B}$ on $\mathbb{R}$. Given two gambles $X, Y: \Omega \rightarrow \mathbb{R}$, we say that they are classically independent if:

$$P(A \cap B) = P(A)P(B) \quad \forall A \in \sigma(X), B \in \sigma(Y),$$

where the σ -algebra generated from a gamble X , $\sigma(X)$, is defined as the smallest σ -algebra which X is measurable with respect to:

$$\sigma(X) := \sigma\left(X^{-1}(\mathcal{B})\right),$$

and $\sigma(\mathcal{H})$ is the smallest σ -algebra containing all sets $H \in \mathcal{H}$, $\mathcal{H} \subseteq 2^{\Omega}$.

Thus independence of gambles is reduced to independence of events. But note that this definition inherently depends on the choice of the Borel σ -algebra on $\mathbb{R}$. In our case, this is similar: to define irrelevance and independence on gambles, we need to fix a set system on $\mathbb{R}$, but we leave the choice open in general.

Definition 231. Assume a set system $\mathcal{H} \subseteq 2^{\mathbb{R}}$ and two gambles $X, Y: \Omega \rightarrow \mathbb{R}$ are given. We call Y irrelevant to X with respect to $\mathcal{H}$ if

$$\bar{P}(X^{-1}(A)|Y^{-1}(B)) = \bar{P}(X^{-1}(A)) \quad \forall A, B \in \mathcal{H} \text{ if } Y^{-1}(B) \in 2_{1+}^{\bar{\Omega}}.$$

Similarly, we call them independent when both directions hold.

²⁹Walley and Fine [1982] considered outcomes of “joint experiments” in $\Omega \times \Omega$. They furthermore demanded that lower limits of relative frequencies factorize; to us it is not clear from a strictly frequentist perspective why this condition should be introduced.

Observe that if $\mathcal{H} = \mathcal{B}$ and $\bar{P}$ was actually a precise P on $\sigma(X)$ and $\sigma(Y)$, this definition would be equivalent to Definition 230 (modulo the subtlety regarding conditioning on measure zero events), due to the following.

Lemma 232. *Given set systems $\mathcal{A}, \mathcal{B} \subseteq \Omega$, in the precise case, the following statements are equivalent.*

PI1. $P(A|B) = P(A) \forall A \in \mathcal{A}, B \in \mathcal{B} \text{ and } P(B) > 0.$

PI2. $P(A \cap B) = P(A)P(B) \forall A \in \mathcal{A}, B \in \mathcal{B}.$

Proof Obviously PI2 implies PI1 by the definition of conditional probability. One only has to check that when PI1 holds, that PI2 holds even if $P(B) = 0$. But if $P(B) = 0$, then also $P(A \cap B) = 0$ due to monotonicity of P in the sense of a capacity. ■

Example 233. Choose $\mathcal{H} := \{(-\infty, a] : a \in \mathbb{R}\}$ in Definition 231. Such an $\mathcal{H}$ is called a Π -system, which is a non-empty set system that is closed under finite intersections. This particular Π -system can in fact be used to define independence in the classical case, which is done in terms of the joint cumulative distribution function.

4.10 APPENDIX: PATHOLOGICAL OR NORMAL?

Much of the confusion about probability arises because the true depth of the law of large numbers as an extremely hard analytical assertion is not appreciated at all. — Detlef Dürr and Stefan Teufel [2009, p. 62]

When one looks at *finite* sequences $x: [n] \rightarrow [2]$, there is a simple counting argument using the binomial theorem that illustrates that the vast majority of the 2^n possible sequences have roughly equal numbers of elements with values of 1 and 2. If one *assumes* that an infinite sequence $x: \mathbb{N} \rightarrow [2]$ is generated i.i.d. then this argument can be used to prove the law of large numbers, which ensures “most” sequences have relative frequencies which converge.

Hence the construction, as illustrated in the present paper, of sequences $x: \mathbb{N} \rightarrow [k]$ with *divergent* relative frequencies naturally raises the question of how contrived they are. That is, are we examining a rare pathology, or something “normal” that we might actually encounter in the world? We will refer to sequences whose relative frequencies converge as “stochastic sequences” and sequences whose relative frequencies do not converge as “non-stochastic sequences”³⁰.

The classical law of large numbers suggests that indeed “almost all” sequences are stochastic, and therefore, by such reasoning, the non-stochastic sequences with which we have concerned ourselves in the present paper are indeed pathological exceptions. In this appendix we will argue:

1. This very much depends upon what one means by “rare” or “almost all” and there are many choices, and the only real argument in favour of the usual ones (which declare non-stochastic sequences rare) is familiarity — different notions of “typicality” (for that is what is at issue) lead to very different conclusions. Specifically, there are choices (arguably just as “natural” as the familiar ones) which imply that rather than non-stochastic sequences being rare, they are in fact the norm in a very strong sense.
2. Nevertheless, none of the mathematical nuances of the previous point allow one to conclude *anything* about the empirical prevalence of stochastic or non-stochastic sequences in the world. Indeed, no purely mathematical reasoning allows one to draw such conclusions, unless one wishes to appeal to some conception of a Kantian “synthetic a priori.”

We will first explore what can be said from a purely mathematical perspective, illustrating that there is a surprising amount of freedom of choice in precisely posing the problem,

³⁰ This dichotomy would appear clear-cut, but there is a subtlety: there exist sequences such that $(x_i)_{i \in \mathbb{N}}$ and $(y_i)_{i \in \mathbb{N}}$ are both stochastic, but the *joint* sequence $((x_i, y_i))_{i \in \mathbb{N}}$ is *non-stochastic* [Rivas, 2019]; that is, the marginal relative frequencies converge, but the joint relative frequencies do not!

and that the choices are consequential. Then in Subsection 4.10.3 we examine the question of prevalence of non-stochastic sequences actually in the world.

4.10.1 THE MATHEMATICAL ARGUMENT — THE CHOICES TO BE MADE

The classical Law of Large Numbers says “almost all sequences” are stochastic. But the “almost all” claim comes from the mapping of sequences to real numbers in $[0, 1]$ and then making a claim that “almost all” numbers correspond to stochastic sequences. Thus there are at least three choices being made here:

Mapping from Sequences to Real Numbers The choice of mapping from sequences to real numbers, to enable to use of some notion of typicality on $[0, 1]$ to gauge how common stochastic sequences are.

Notion of Typicality The notion of typicality to be used (e.g. Cardinality, Hausdorff dimension, Category or Measure).

Specific Index of Typicality Within the above choice of notion of typicality, the particular choice of typicality index, e.g. the measure or topology that underpins the notion of typicality.

The choices for the classical law of large numbers are 1) k -ary positional representation; 2) a σ -additive measure on $[0, 1]$; 3) The Lebesgue measure. As we shall summarize below, each of these three choices substantially affects the theoretical preponderance of non-stochastic sequences.

That there are alternate choices that lead to the unusual conclusion that non-stochastic sequences are “typical” has been known for some time: “This result may be interpreted to mean that the category analogue of the strong law of large numbers is false” [Oxtoby, 1980, p. 85]; see also [Méndez, 1981]. The significance of this fact has been stressed recently [Seidenfeld et al., 2017, Cisewski et al., 2018]. And it has been observed that the introduction of alternate topologies can change whether sequences are stochastic [Khrennikov, 2013]. However, the strongest results arise in number theory, motivated by the notion of a “normal number.”

4.10.2 NOTIONS OF TYPICALITY — CARDINALITY, DIMENSION, COMEAGRENESS, AND MEASURE

Let $\mathcal{S}$ (resp. $\mathcal{N}$) denote the set of stochastic (resp. non-stochastic) sequences $\mathbb{N} \rightarrow [k]$. That is, $\mathcal{S} := \{x: \mathbb{N} \rightarrow [k]: \lim_{n \rightarrow \infty} r^x(n) \text{ exists}\}$ and $\mathcal{N} := [k]^{\mathbb{N}} \setminus \mathcal{S}$. (For simplicity, and alignment with Appendix 4.8, we restrict ourselves to sequences whose domain is $[k]$.)

In order to make a claim regarding the relative preponderance of stochastic versus non-

stochastic sequences, they are often mapped onto the unit interval.³¹ In such cases, the question of relative preponderance of classes of sequences is reduced to that of a question concerning the relative preponderance of classes of subsets of $[0, 1]$. The question then arises of how to measure the size of such subsets. Unlike in the finite case mentioned above, merely counting (i.e. determining the cardinality of the respective subsets) is hardly adequate, as it is easy to argue that $|\mathcal{S}| = |\mathcal{N}| = \aleph_1$. There are three notions that have been used to compare the size of $\mathcal{S}$ and $\mathcal{N}$:

Measure A countably additive measure, usually the Lebesgue measure on $[0, 1]$.

Meagre / Comeagre A subset S of a topological space X is meagre if it is a countable union of nowhere dense sets (i.e. sets whose closure has empty interior). A set S is comeagre (residual) if $X \setminus S$ is meagre.

Dimension A variety of fractal dimensions, such as the Hausdorff dimension, have also been used to judge the size of non-stochastic sequences (and the numbers they induce); however for space considerations we omit discussion of these results³².

Some of the results obtained in the literature are summarized below. The object is not to state them in an entirely formal manner, or even to describe them in their full generality. Rather we simply wish to show the diversity of conclusions available by tweaking the three choices enumerated above. If no representation is mentioned, the usual k -ary positional representation is used, whereby $\tilde{x} \in [0, 1]$ is constructed from $x: \mathbb{N} \rightarrow [k]$ via $\tilde{x} := \sum_{i \in \mathbb{N}} (x_i - 1)k^{-i}$ (the $(x_i - 1)$ term is required because our sequences map to $[k] = \{1, \dots, k\}$). Obviously every $x \in [k]^{\mathbb{N}}$ maps to some $\tilde{x} \in [0, 1]$; and every $z \in [0, 1]$ corresponds to at least one $x \in [k]^{\mathbb{N}}$ (recalling we have to handle the situation that, when $k = 10$ for example, $0.4\bar{9} = 0.5\bar{0}$, where $\bar{i}$ means that i is repeated infinitely, and thus there are two sequences $x_1, x_2 \in [k]^{\mathbb{N}}$ such that $1/2 = \tilde{x}_1 = \tilde{x}_2$).³³ Let $\tilde{\mathcal{S}} := \{\tilde{x} \in [0, 1]: x \in \mathcal{S}\}$ and $\tilde{\mathcal{N}} := \{\tilde{x} \in [0, 1]: x \in \mathcal{N}\}$.

Most (Lebesgue measure) sequences are stochastic This is the classical strong law of large numbers. If μ_{leb} denotes the Lebesgue measure on $[0, 1]$, then the claim is that $\mu_{\text{leb}}(\tilde{\mathcal{S}}) = 1$.

Most (comeagre) sequences are non-stochastic Let $X_i = [2]$ and $X = \prod_{i \in \mathbb{N}} X_i$ equipped with the product topology. As a set $X \cong [2]^{\mathbb{N}}$. Then $\mathcal{N} \subset X$ is comeagre [Oxtoby, 1980].

Most (comeagre) sequences are stochastic With different choices of topology, the opposite conclusion holds — there are topologies such that $\mathcal{S}$ is comeagre [Calude et al.,

³¹Attempts known to us to judge the relative sizes of $\mathcal{S}$ and $\mathcal{N}$ which do not rely on such a mapping are described in the first, third and sixth of the cases listed below, and rely upon imposing a topology directly on the set of sequences $[k]^{\mathbb{N}}$.

³²See for example [Eggleston, 1949, Olsen, 2004, Gu and Lutz, 2011, Bishop and Peres, 2017, Albeverio et al., 2017].

³³This non-uniqueness of the representation will not affect the results below because $\{\tilde{x}_1 = \tilde{x}_2 \in [0, 1]: x_1 \neq x_2\} = \mathbb{Q} \cap [0, 1]$ and is of cardinality $\aleph_0$, whereas $|[k]^{\mathbb{N}}| = |[0, 1]| = \aleph_1$.

2003].

Most (comeagre) sequences are extremely non-stochastic Let $\tilde{\mathcal{N}}^*$ denote the subset of $[0, 1]$ of $\tilde{x}$ corresponding to $x \in [k]^\mathbb{N}$ which satisfy $\forall i \in [k], \liminf_{n \rightarrow \infty} r_i^x(n) = 0$ and $\limsup_{n \rightarrow \infty} r_i^x(n) = 1$. These sequences are (justifiably) called extremely non-stochastic; the sequence constructed in Subsection 4.8.12 is an example. Then the set $\tilde{\mathcal{N}}^*$ is comeagre in the usual topology of real numbers [Calude and Zamfirescu, 1999]; confer [Calude, 2002, section 7.3].

Most (comeagre) sequences are perversely non-stochastic Denote the set of perversely non-stochastic sequences $\mathcal{N}^{**} := \{x \in [k]^\mathbb{N} : \text{CP}(r^x) = \Delta^k\}$. Observe $\mathcal{N}^{**} \subset \mathcal{N}^*$. Let $\tilde{\mathcal{N}}^{**} := \{\tilde{x} \in [0, 1] : x \in \mathcal{N}^{**}\}$ (what Olsen [2004] calls “extremely non-normal numbers”, but we use “extremely” for the larger set $\tilde{\mathcal{N}}^*$). Then $\tilde{\mathcal{N}}^{**}$ is comeagre (in the usual topology of real numbers) [Aveni and Leonetti, 2022, Olsen, 2004]. An even stronger result holds. Let A denote a (not necessarily uniform) finite averaging operator and let $\mathcal{N}^{***} := \{x \in [k]^\mathbb{N} : \text{CP}(A(r^x)) = \Delta^k\}$ and $\tilde{\mathcal{N}}^{***} := \{\tilde{x} \in [0, 1] : x \in \mathcal{N}^{***}\}$. Observe $\mathcal{N}^{***} \subset \mathcal{N}^{**}$ and $\tilde{\mathcal{N}}^{***} \subset \tilde{\mathcal{N}}^{**}$. Then $\tilde{\mathcal{N}}^{***}$ is also comeagre [Stylianou, 2020]!

Most (Lebesgue measure) sequences are non-stochastic There exist a range of representations of real numbers called Q^* -representations (Q^* is a $k \times \infty$ matrix valued parameter of the representation); see [Albeverio et al., 2005, Section 4] for details. Let $\tilde{x}^{Q^*}$ denote the Q^* -representation of a sequence $x \in [k]^\mathbb{N}$, and $\tilde{\mathcal{S}}^{Q^*} := \{\tilde{x}^{Q^*} : x \in \mathcal{S}\}$ and $\tilde{\mathcal{N}}^{Q^*} := \{\tilde{x}^{Q^*} : x \in \mathcal{N}\}$. Then there exist Q^* such that $\mu_{\text{leb}}(\tilde{\mathcal{N}}^{Q^*}) = 1$. [Albeverio et al., 2005, p. 627]. Thus if the size of $\mathcal{N}$ is judged via certain Q^* representations, Lebesgue almost all sequences are non-stochastic!

An obvious conclusion to draw from the above examples is that in answering the question of the preponderance of non-stochastic sequences, one can get essentially whatever answer one wants by choosing a range of different precise formulations of the question. At the very least, this should make us skeptical of any purely mathematical attempts to reason whether one might expect to encounter non-stochastic sequences in practice — the topic to which we now turn.

4.10.3 TYPICAL REAL SEQUENCES

The laws of large numbers cannot be applied for describing the statistical stabilization of frequencies in sampling experiments. — Andrei Khrennikov [2009, p. 20]

What do the above points imply about the likelihood one will encounter stochastic or non-stochastic sequences when performing real measurements?

Nothing.

This is not to say that in actuality we will often encounter non-stochastic sequences. Rather our point is that no amount of purely theoretical reasoning will be able to tell us in advance how “likely” it is to do so. What is at issue is whether stochastic sequences are in fact “typical” in our world.

Perhaps the most surprising thing about the mathematical results summarized above is the extent to which different notions of typicality affect the conclusions. This raises the question of whether some notions of typicality are more justified when wishing to consider real sequences that have been measured in the world. In the study of physics (especially aspects of physics that are apparently intrinsically statistical) such questions have been raised, and below we briefly summarize what is known.

Traditionally, “probability” is considered as a primitive, and notions of typicality are derived from that in terms of their “probability” of occurring. And the above examples illustrate that attempts to argue for the Lebesgue measure having a privileged role as the “right” notion of typicality are barking up the wrong tree; confer [Pitowsky, 2012]. But this will not do for our question. Typicality is a more fundamental notion [Galvan, 2006, 2007] — arguably the “mother of all” notions of probability [Goldstein, 2012]. Typicality is at the core of questions of non-stochastic randomness in physics, thus (consistent with the perspective of the present paper) leading to non-additive measures of typicality [Galvan, 2022] (essentially defining a measure of typicality inspired by a coherent upper probability) which allows the extension to notions of mutual typicality necessary to reason about situations such as that referred to in footnote 30.

In fact, typicality plays an even stronger role than answering questions regarding the preponderance of non-stochastic sequences. As Dürr and Struyve [2021, p. 36] observe “the notion of typicality is necessary to understand what the statistical predictions of a physical theory really mean.” They note that the usual appeal to the law of large numbers misses the point because while its conclusion is true (convergence of relative frequencies) *if* one sees typical sequences, but “What needs to be explained is why we only see typical sequences! That’s actually the deep question underlying the meaning of probability theory from its very beginning ...” [Dürr and Struyve, 2021, p. 37]. In classical mechanics, appeal to Liouville’s theorem suggests an “invariant measure” as being a natural choice; in the quantum realm, there is an analogous choice (invariant to Bohmian flow) [Dürr and Struyve, 2021, p. 41]. But these situations are rather special from the perspective of a statistician. The situation is well summarized by Dürr [2001, p. 130]: “What is typicality? It is a notion for defining the smallness of sets of (mathematically inevitable) exceptions and thus permitting the formulation of law of large numbers type statements. Smallness is usually defined in terms of a measure. What determines the measure? In physics, the physical theory.” Confer [Dürr and Teufel, 2009, Chapter 4] who observe that from a *scientific* per-

spective (where one wants to make claims about the world) establishing the pre-conditions for the law of large numbers to hold is “exceedingly difficult”³⁴.

Very well one might say, but the arguments in favor of typicality of non-stochastic sequences given above all rely on topological arguments, or unusual encodings of sequences to numbers. What is the justification for topological notions of typicality when considering sequences of measurements obtained from the world? Sklar [2000, p. 270] has actually argued that the topological perspective might offer a foundational perspective with *fewer* opportunities for claims of arbitrariness than measure theoretical approaches. See also [Sklar, 1995, p. 185] and the discussion in [Guttmann, 1999, Chapter 4] which reframes the problem away from typicality to viewing the whole question from an approximation perspective where the notion of smallness of sets is naturally one of meagreness. Our point is that even within the restricted realm of physics, there are compelling arguments at least not to take the measure-based notion of typicality for granted. Once that is accepted, non-stochastic sequences seem less unusual.

4.10.4 VIOLATIONS OF THE LAW OF LARGE NUMBERS

A typical universe is in equilibrium; but “our universe is atypical or in non-equilibrium” [Dürr and Teufel, 2009, p. 81] and “what renders knowledge at all possible is nonequilibrium” [Dürr et al., 1992, p. 886] so we should not be surprised if it is not “typical”. And indeed that is what we see as long as we look: “The so-called law of large numbers is also invalid for social systems with finite elements during transition” [Chen, 1991]. Gorban [2011, 2017, 2018] has documented many examples of real phenomena failing to be statistically stable. Such failures are held to explain departures from “normal” distributions [Philip and Watson, 1987]. But more importantly, they mean we should not expect even convergence of relative frequencies in non-equilibrium situations.

Such was the conclusion of Prigogine in his ground-breaking studies of non-equilibrium thermodynamics where he spoke of a “breakdown of the ‘law of large numbers’” [Nicolis and Prigogine, 1977, p. 9 and 228]; see also [Prigogine, 1978, p. 781], [Prigogine and Stengers, 1985, p. 180] and [Prigogine, 1980, p. 131]. And more recently, studies of the use of machine learning systems “in the wild” have recognized that non-stochasticity is not so exotic after all [Katsikopoulos et al., 2021]. Thus perhaps its time to downgrade this “law” of nature.

³⁴The example they give is for the Galton board, or quincunx, a device often appealed to in order to teach the reality of the central limit theorem — an even stronger claim than the law of large numbers. The irony is that it is rarely checked empirically. And when it has been, it has been found to be untrue! [Bagnold, 1983, Figure 8].

4.10.5 REPEAL OF THE LAW OF LARGE NUMBERS

A typical universe shows statistical regularities as we perceive them in a long run of coin tosses. It looks as if objective chance is at work, while in truth it is not. There is no chance. That is the basis of mathematical probability theory. — Detlef Dürr and Stefan Teufel [2009, p. 64].

Desrosières [1998] in his history of statistical reasoning has observed the awe with which stable frequencies were viewed when they were first encountered; the effect been interpreted as a hidden divine order³⁵. And indeed in many practical situations, stable frequencies do arise. But that does not mean we should take such situations as the only ones that can occur. We may well legitimately call them “normal.” But we can better understand the normal by studying the pathological [Canguilhem, 1978, p. 19–20]. Ironically in his attempt to clarify the notion of “normal” Canguilhem [1978, p. 103] considered whether “normal” was simply “average” and concluded “the concepts of norm and average must be considered as two different concepts”. As we have seen, averages can indeed be far from normal, and potentially quite often.

Perhaps we have been misled by the strange name given to the famous theorem we are considering: by calling it a “law” we are inheriting a lot of baggage as to what we mean by that, baggage that has been traced to notions of divine origin of lawfulness [Zisel, 1942].³⁶ And we hanker after lawfulness:

We ...naturally hope that the world is orderly. We like it that way... All of us ...find this idea sustaining. It controls confusion, it makes the world seem more intelligible. But suppose the world should happen in fact to be not very intelligible? Or suppose merely that we do not know it to be so? Might it not then be our duty to admit these distressing facts? [Midgley, 2013, p. 199]

Perhaps the theory of imprecise probabilities presented in this paper which we have grounded in the instability of relative frequencies may help us to admit this “distressing fact.” It does

³⁵“I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the ‘Law of Frequency of Error.’ The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.” [Galton, 1889, p. 66]. See also [Rose, 2016] for a recent discussion on statistical normality.

³⁶The contrary views regarding the historical origin of the notion of a scientific law [Milton, 1981, Ruby, 1986, Weinert, 1995] do not contradict our point. Ironically, Zisel was convinced that the law of large numbers *was* a natural law (indeed the most fundamental of natural laws!) [Zisel, 1916], and in his PhD thesis attempted to argue the case for this philosophically, although after errors were pointed out, he renounced the argument and made no further reference to it [Lenhard and Krohn, 2022, p. 125].

suggest to us that the law of large numbers, while a fine and true theorem, as a “law” might be in need of repealing.

5

Data Models With Two Manifestations of Imprecision

AUTHOR CONTRIBUTIONS

Robert Williamson contributed through supervision of the project, e.g. through discussions (specifically on typicality) and proof-reading. The remaining work is due to Christian Fröhlich.

ABSTRACT

Motivated by recently emerging problems in machine learning and statistics, we propose data models which relax the familiar *i.i.d.* assumption. In essence, we seek to understand what it means for data to come from a set of probability measures. We show that our frequentist data models, parameterized by such sets, manifest two aspects of imprecision. We characterize the intricate interplay of these manifestations, aggregate (ir)regularity and local (ir)regularity, where a much richer set of behaviours compared to an *i.i.d.* model is possible. In doing so we shed new light on the relationship between non-stationary, locally precise and stationary, locally imprecise data models. We discuss possible applications of these data models in machine learning and how the set of probabilities can be estimated. For the estimation of aggregate irregularity, we provide a negative result but argue that it does not warrant pessimism. Understanding these frequentist aspects of imprecise probabilities paves the way for deriving generalization of proper scoring rules and calibration to the imprecise case, which can then contribute to tackling practical problems.

5.1 INTRODUCTION

In machine learning and statistics, the assumption that data is drawn *i.i.d.* is ubiquitous, so firmly entrenched to the extent that it is seldom questioned. Recently, however, various challenges to the *i.i.d.* assumption have surfaced and received increasing attention, among them the problem settings of *data corruptions* (including *dataset shift*) [Iacovissi et al., 2023], *fairness desiderata*, *federated learning*, *multi-source adaptation* and others. These settings call for a relaxation of the *i.i.d.* assumption to the weaker assumption that the data comes from a *set of probabilities* instead of a single one.

Under the banner of *imprecise probabilities* (IP; Walley [1991]) scholars have gathered to systematically study more flexible models, essentially equivalent to sets of probability measures. With relatively few exceptions, work on IP has been rooted in a subjectivist approach, taking a set of probability measures to represent a state of *belief*, which justifies the name *credal set* for such a set of probability measures. As a connection to the frequentist side, an often-used interpretation for imprecise probabilities is that data is drawn *i.i.d.*, but since the true precise probability is not known to the agent, the imprecision supposedly models the *epistemic uncertainty* arising from the lack of knowledge (see e.g. [Hüllermeier et al., 2022]). This is in line with some works in the paradigm of *distributional robustness*, which assume a true but unknown precise probability and aim to construct an ambiguity set of probability measures to guard against the “estimation” error [Kuhn et al., 2019]. In this view, knowledge of the true probability eliminates imprecision. In contrast, the problems listed above which motivate us (data corruptions etc.) are not appropriately modelled as *i.i.d.*, even when putting epistemic concerns aside. Few works in the IP literature [Walley and Fine, 1982] have taken an interpretation of *aleatoric indeterminacy* or what is also called an *ontic interpretation* [Augustin, 2022], which grounds the imprecision not in a lack of knowledge of the data-generating process, but in frequentist aspects of the data itself: the data is drawn from a set of probabilities — or rather, it is modelled in this way.

Taking this perspective, in this paper we relax the *i.i.d.* assumption. First, we consider a data model which still assumes precise, probabilistic independence, but where we allow a set of probability measures instead of a single one. We call this the *non-stationary, locally precise* data model.¹ It is specified by an equation of the form (stated rigorously in Section 5.3)

$$\lambda \{ \omega^\infty = (\omega_1, \omega_2, \dots) \} = \prod_{i=1}^{\infty} p_i(\{\omega_i\}),$$

where ω^∞ is an infinite sequence of elementary events, and the p_i are probability measures. This model is locally precise in the sense that for each $i \in \mathbb{N}$, a precise probability p_i gov-

¹Of course we are not the first ones to use such a basic model, but we do study the aspects of aggregate (ir)regularity and local (ir)regularity of these models in detail.

erns the data generation. The imprecision is thus, in contrast, a global feature of the model. We are then interested in studying the implications of such a *data model with underlying (global) imprecision*. In particular, we focus on two aspects that the familiar *i.i.d.* model exhibits, *aggregate regularity* and *local irregularity* (randomness), and our main theorem fully characterizes how these aspects are transformed and interact in such more general data models (Theorem 240). In summary:

- Under an *i.i.d.* model, relative frequencies of events converge *almost surely* to a limit. In our data models with underlying imprecision, this might break down. Here, the natural generalization is to consider the *set of cluster points* (accumulation points) of relative frequencies.
- Under an *i.i.d.* model, outcome sequences are *almost surely* “perfectly random” (to be explained later). In our data models with underlying imprecision, we may find hidden heterogeneity, that is, subsequences with relative frequencies that even in the limit do not coincide with the aggregate regularity. For example, in a fairness context, this would correspond to subpopulations for which the probability is distinct from the population probability.

Previous frequentist work in the literature can be broadly distinguished by which of these two *manifestations* of imprecision they focus on. [Walley and Fine \[1982\]](#) and [Ivanenko \[2010\]](#) have focused on the first, whereas a later line of work by Fine and colleagues (e.g. [Fierens et al. \[2009\]](#)) studies the second manifestation. These works can be, in our opinion, challenging to access. In our paper, we aim to provide an easily accessible account of frequentist imprecision, which also for the first time explicates the relation between these two manifestations and thereby unifies previous works.

In the process of establishing relations to a line of work initiated by [Walley and Fine \[1982\]](#) (Section 5.4.1), we go beyond non-stationarity and local precision. By using certain non-stationary, locally precise data models as basic building blocks, we arrive at more general data models, which are stationary and where the imprecision is now already a local feature, and sometimes a global feature. This puts us in a position to also compare our approach (Section 5.4.2) to a literature on generalized laws of large numbers for imprecise probabilities and coherent risk measures, a strain of literature which is actively being developed (e.g. [\[Hu et al., 2016, Peng, 2019, Zhang et al., 2024\]](#)). Here, the distinction of *risk* and *ambiguity* is a prominent motivation. This comparison also illustrates how data models relate to subjectivist approaches more generally.

Finally, to the possible implications of this work. We hope that our paper can provide a starting point for addressing practical problems outlined above (and others) in a more principled manner by providing a conceptual foundation. In the classical picture, the *i.i.d.* data

model is paired with *proper scoring rules* and *calibration* to obtain a general framework for obtaining and evaluating probabilistic forecasts. We believe that introducing generalized proper scoring rules and calibration, adapted to the imprecise case, is required, and can then lead to the development of practically useful methods. Our position is that these concepts need to be considered as relative to a data model, and in this way our work makes first steps in this direction. Developing these concepts would also serve a related goal: in the philosophical literature on rational decision making there is a heated debate about whether imprecise belief states are ever warranted (e.g. [Elga, 2010, Bradley and Steele, 2014, Schoenfield, 2017]); for example, when might a rational decision maker justifiably believe that the probability of rain tomorrow is between 20% and 40%? Our approach makes it plausible that one part of the controversy could be resolved: at least under a data model with underlying imprecision, imprecise belief states are warranted. Still, in our opinion, definite answers will require conceptually developing proper scoring rules and calibration for such imprecise data models. Also, this reasoning remains silent on the case for imprecise belief states when the data model is precise and the goal is to capture epistemic uncertainty with imprecision, as in the paradigm of distributional robustness. While developing proper scoring rules and calibration for imprecise data models is out of scope for the present paper, we show how the two manifestations of imprecision can (and *cannot*) in principle be estimated.

The paper is organized as follows. We begin by a brief, general treatment of *data models* to set the stage for going beyond the *i.i.d.* model (Section 5.2). We highlight two aspects of the *i.i.d.* model: aggregate regularity and local irregularity. We discuss a further crucial ingredient in any data model, *typicality*. We introduce the non-stationary, locally precise data model based on a set of probabilities (Section 5.3). Our main theorem characterizes possible manifestations of imprecision in such models as aggregate (ir)regularity and local (ir)regularity (Theorem 240). We introduce stationary, locally imprecise data models in Section 5.4.1, compare them to the highly related work of Walley and Fine [1982] and subsequent papers, and to generalized laws of large numbers (Section 5.4.2). We discuss possible applications for our data models (Section 5.5.1) and make some general remarks about estimation. Here, we provide a negative result concerning an estimator proposed by Walley and Fine [1982], but argue that it is not as troubling as it may appear at first sight. We conclude by motivating the need for imprecise scoring rules and imprecise calibration, relativized to such data models.

Who should read this paper? We believe this paper might be of relevance to machine learning and statistics scholars interested in establishing conceptual foundations for the above practical (and similar) problems, which escape being modelled as *i.i.d.*, as well as scholars generally interested in frequentist aspects of imprecise probabilities and coherent risk measures. In particular, those who have found previous works on such frequentist aspects

hard to access (e.g. [Walley and Fine, 1982]) may find our exposition helpful.

General notation. For a set A , we denote its powerset as 2^A . We write $\text{c}\bar{o}(A)$ for the closed convex hull in the appropriate topology. By χ_A we denote the indicator function of the set A , defined as $\chi_A(\omega) := 1$ if $\omega \in A$, 0 otherwise. We write the set of A -valued sequences as A^∞ . An element of A^∞ is some infinite sequence $a^\infty = (a_1, a_2, \dots) \in A^\infty$. For a finite $n \in \mathbb{N}$, we define $A^n := \{a = (a_1, \dots, a_n) : a_i \in A\}$. If $a = (a_1, \dots, a_u) \in A^u$ and $b = (b_1, \dots, b_v) \in A^v$, if $u \leq v$, we write $a \subseteq b$ to mean that a is a prefix of b , and likewise if b is an infinite sequence. We use the terms *probability* and *probability measure* synonymously, but sometimes choose the latter to stress the measure-theoretic character.

5.2 DATA MODELS

Assume a finite *possibility set* $\Omega = \{\omega^1, \dots, \omega^k\}$, where an elementary event $\omega^i \in \Omega$ captures all relevant aspects of the world given a decision-making context. Throughout the paper, we will assume finite Ω and $1 < |\Omega| = k$. In line with the IP literature, we call bounded functions $X : \Omega \rightarrow \mathbb{R}$ *gambles* and collect them in the set L^∞ . For example, $\Omega = \mathcal{X} \times \mathcal{Y}$ might be a joint space containing features and labels, which are then extracted by $X, Y \in L^\infty$; since $\mathbb{R}$ and $\mathbb{R}^d$ have the same cardinality, this allows handling vector-valued features. Given a loss function $\ell : \mathcal{A} \times \Omega \rightarrow \mathbb{R}$, we are typically interested in the gamble $\ell(a, \cdot)$ for a chosen action $a \in \mathcal{A}$.²

By *data* (or *data sequences*) we understand a finite or infinite sequence of elementary events $\omega^\infty = (\omega_1, \omega_2, \dots)$. In the following, we restrict ourselves to infinite sequences and denote the set of such sequences as Ω^∞ . Data is that which is actually fixed. In contrast, a *data model* is a description of a data-generating process, for specifying what kind of data we expect to observe. For a data model we require a *generator*, which is a function $G : \mathcal{Z} \rightarrow \Omega^\infty$ which takes in a *seed* $z \in \mathcal{Z}$ and returns data. Such a seed can be understood as an abstract source of *process randomness* (which we contrast below with *outcome randomness*). A standard choice is to use $\mathcal{Z} = [0, 1)$, since a real number from the unit interval is via its binary expansion identified with an infinite binary sequence [Williams, 1991]. To make this more concrete, consider the familiar *i.i.d.* model, though perhaps expressed here in an unfamiliar way. Denote the Lebesgue measure as λ .

Definition 234 (The *i.i.d.* model). *Let $\mathcal{Z} := [0, 1)$. Given a probability measure p on Ω , it is possible to consistently define $W_i : [0, 1) \rightarrow \Omega$, $i \in \mathbb{N}$ so that the following holds:*

$$\lambda \{z \in \mathcal{Z} : W_1(z) = \omega_1, W_2(z) = \omega_2, \dots, W_n(z) = \omega_n\} := \prod_{i=1}^n p(\{\omega_i\}), \quad \forall n \in \mathbb{N},$$

²The IP literature usually works with a gain orientation, where positive values correspond to gain (utility), whereas we interpret positive values as losses (disutility). No difference results beyond simple sign flips.

Define $G : [0, 1) \rightarrow \Omega^\infty$ as $G(z) := (W_1(z), W_2(z), \dots)$.

Note that the W_i are the quantities that are being defined here, not p or λ . For a proof that the W_i can be defined so that Equation 234 holds, see [Spreij, Theorem 3.16]. The *i.i.d.* model forms the departure point for our more general data models and it is thus worth highlighting some of its features. We do so in a way that is grounded in Richard von Mises [1919] foundational account of probability.

Aggregate regularity, also called *statistical stability* [Gorban, 2017]. Under an *i.i.d.* model it holds that

$$\lambda \left\{ z \in \mathcal{Z} : G(z) = (\omega_1, \omega_2, \dots) : \forall A \subseteq \Omega : \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \chi_A(\omega_i) = p(A) \right\} = 1. \quad (5.1)$$

We call $r_n(A) := \frac{1}{n} \sum_{i=1}^n \chi_A(\omega_i)$ the relative frequency of the event A . In words, “almost surely, relative frequencies converge to their *probability*”. While the above statement, the *strong law of large numbers*, is of formal, mathematical nature, there exists an associated empirical sibling, what Gorban [2017] calls the *hypothesis of statistical stability*. This refers to the statistician’s in principle unfalsifiable belief that the data at hand is the prefix of a statistically stable sequence, that is, gathering more and more data will lead to stabilization of relative frequencies and gamble averages. Note that this belief is not logically implied by the theoretical law of large numbers, which is a formal property of the *i.i.d.* data model. Already de Finetti [1974/2017, p. 175] has noted

“In any case, the force of the stability of frequencies as a probabilistic or statistical principle is completely illusory, and without solid foundation.”

In fact, Gorban [2017] has demonstrated empirically that one does not always observe stable relative frequencies even for very long observation intervals. Recent studies in climate science also observe this [Lovejoy, 2015, Franzke et al., 2020]. Consequently, it seems prudent to explore more flexible data models. We also refer the reader to [Fröhlich et al., 2024] for a discussion of statistical stability.

Local irregularity, also called *randomness*. Under an *i.i.d.* model, the regularity emerges only at the aggregate level, while the individual level remains fully unpredictable (random), relative to the probability p .³ One way of mathematizing this intuition is with von Mises [1919] concept of *selection rules*. For our purposes, a selection rule is a function $S : \mathbb{N} \rightarrow \{0, 1\}$, which either selects or does not select at a given index $i \in \mathbb{N}$. Such a selection rule effectively extracts a subsequence. Then, for a countable set of selection rules $\mathcal{S}$,

³Intuitively, the best forecast when being judged with a proper scoring rule that a forecaster could announce is the true probability.

under the *i.i.d.* model ($\forall A \subseteq \Omega$):

$$\lambda \left\{ z \in \mathcal{Z} : G(z) = (\omega_1, \omega_2, \dots) : \forall S \in \mathcal{S} : \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \chi_A(\omega_i) S(i)}{\sum_{i=1}^n S(i)} = p(A) \right\} = 1. \quad (5.2)$$

That limiting relative frequencies along a subsequence coincide with the probability as specified by p is a form of randomness desideratum. Intuitively, a skeptic who uses a selection rule S to decide whether or not to place a bet of the form $\omega_i - p(A)$ on the next outcome cannot get infinitely rich, when accumulating capital as $i \rightarrow \infty$. In light of the above assertion, we can expect “perfect randomness” under an *i.i.d.* model. By contrast, assume that Equation 5.2 would *not* hold, meaning some $S \in \mathcal{S}$ reveals what we might call *local regularity*, *non-randomness* or *hidden heterogeneity*. In a fairness context, for example, this would mean that the probability associated to a subpopulation does not coincide with the probability associated to the whole population (see [de Finetti, 1974/2017, p. 175]). In the context of machine learning, optimizing only an aggregate criterion (e.g. average loss) then implies risking unfairness (non-robustness) [Williamson and Menon, 2019]. The term *hidden heterogeneity* highlights that since all information about a single datum is captured by the elementary event ω (recall that the feature map is $X : \Omega \rightarrow \mathbb{R}$), a predictor $f(X)$ cannot, if there are distinct probabilities on Ω , be “aware” of this heterogeneity. This also illustrates that whether there is hidden heterogeneity depends on the modelling, the choice of features. If for example, in a fairness context, the sensitive feature is added to $\mathcal{X}$, the heterogeneity is “revealed” and a predictor can make use of it. In this way, hidden heterogeneity refers to a non-randomness that is not captured by the probabilistic randomness of the features.

We remark that, in this paper, we brush aside a subtle distinction: in von Mises framework, a selection rule is a function that can depend on all previous outcomes (meaning a skeptic could take previous outcomes into account when deciding whether to place a bet or not), whereas in our setting it may only depend on the index $i \in \mathbb{N}$. For a fixed sequence, the formulations coincide. Mathematically, our setting is significantly simpler, however.

While the two features above are specific to the *i.i.d.* model (we will later “break” them), another ingredient of any data model in our view is a notion of **typicality**. Some data sequences possess specific properties. For instance, we might be interested in sequences with converging relative frequencies. Such a property can then be identified with a subset of seeds:

$$\mathcal{Z}_A := \{z \in \mathcal{Z} : G(z) \text{ has property } A\} \subseteq \mathcal{Z}.$$

Both statements Equation 5.1 and Equation 5.2 hold with the qualification that the Lebesgue measure of the set of seeds which generate sequences with the property of interest is 1; not that it holds *for all* sequences. We think the essence of such an assertion lies in asserting

typicality: it means that *typically* we will observe sequences with converging relative frequencies; other sequences are so rare that they are considered negligible. An advantage of framing this in terms of typicality rather than probability, is that it circumvents the following problematic move by frequentists, discussed by La Caze [2016]:

“Importantly, ‘almost sure convergence’ is also given a frequentist interpretation. Almost sure convergence is taken to provide a justification for assuming that the relative frequency of an attribute *would* converge to the probability in actual experiments *were* the experiment to be repeated indefinitely” [emphasis in original].

That is, there are actually two measures at play, but one expresses typicality, whereas the other expresses probability; they are conjoined, indeed, but only *typically*. For our purposes, we do not require interpreting the typicality statement as a frequentist statement itself, we simply accept the more parsimonious notion that typicality is somehow defined and can be used instrumentally (that it has some justification). When we speak of *probability*, we mean a measure on Ω which is appropriately normalized. In contrast, the “randomness” in the data generation instead is tied to typicality and should be conceptually distinguished from what we have above called local irregularity, a property of a fixed data sequence; one way of phrasing this distinction is in terms of *process* randomness (concerning typicality) vs. *outcome* (or *product*) randomness (concerning local irregularity) [Eagle, 2021]. Thus, in our setup of the *i.i.d.* model, the Lebesgue measure expresses typicality, we may call it *probabilistic typicality* [Galvan, 2006] since the Lebesgue measure on $[0, 1)$ formally is a probability measure in the sense of measure theory, but the semantics are distinct from that of a probability p on Ω . Conceptually distinguishing typicality from probability is valuable because it opens up room for *nonprobabilistic typicality* notions which have been axiomatically introduced and studied by Galvan [2006]. In our case, we will use a *coherent lower probability* as the typicality notion for our data models.

Galvan [2006] is, to our knowledge, the first to initiate an axiomatic study of typicality beyond probabilistic typicality. To this end, the author axiomatically defines the notion of a *typicality distance* between sets of $\mathcal{Z}$ (in our context), giving rise also to a *relative* and *absolute typicality measure*. We refer the reader to Galvan [2006] for details, and here simply remark that in the probabilistic case, the author suggests as an absolute typicality measure (adapted to our context):

$$T_a(Z) := \lambda(\Omega \setminus Z), \quad \forall Z \in \mathcal{B}([0, 1)),$$

where $\mathcal{B}([0, 1))$ is the Borel σ -algebra on $[0, 1)$. Galvan [2006] defines a *typical set* as a set satisfying $T_a(Z) \approx 0$; to make this formally rigorous, we could demand that $T_a(Z) = 0$, which suffices when we only consider the asymptotic case. Thus, under an *i.i.d.* model, the

set of sequences for which relative frequencies converge to p forms a typical set. Note that different sets can be typical: for instance, the set of sequences for which relative frequencies converge to p is a strict superset of the set of sequences for which relative frequencies converge to p and relative frequencies along a countable set of subsequences also converge to the same limit. Both are typical sets under an *i.i.d.* model.⁴ Which typical set we need to consider depends on what and how we try to estimate.

The notion of a typical set is crucial for estimation. An *i.i.d.* model which is parameterized by a true precise probability p typically generates sequences where limiting relative frequencies coincide with p . This means that relative frequencies provide a means to estimate the data model's parameterization under the assumption that our data sequence belongs to a typical set. In the precise case, minimizing a proper scoring rule guarantees that we recover the true probability in the limit for typical sequences (see [Gneiting and Raftery, 2007]). We would like to extend this to the imprecise case, so that we find estimation procedures, which for typical sequences recover the underlying set of measures.

This concludes setting the stage for moving to sets of probabilities, where the associated typicality notion will also be imprecise.

5.2.1 BACKGROUND ON IMPRECISE PROBABILITIES

We keep the background on imprecise probabilities at a bare minimum and then refer the reader to [Walley, 1991] for a comprehensive account, and to [Augustin et al., 2014] for a more accessible introduction. For our purposes, an imprecise probability is essentially given as a set of probability measures. Assume a possibility set Ω with a set system $\mathcal{A} \subseteq 2^\Omega$, so that at least $\emptyset, \Omega \in \mathcal{A}$; often, $\mathcal{A} = 2^\Omega$. In this introduction, to reduce the burden of technicalities, we assume $|\Omega| < \infty$. In Section 5.4, we will need to consider imprecise probabilities on infinite possibility sets (namely a set of infinite sequences); we caution the reader that the statement below then hold only when replacing “probability measures” with finitely additive expectation functionals.⁵

Definition 235. We say that a set function $\underline{P} : \mathcal{A} \rightarrow [0, 1]$ is a coherent lower probability if

$$\underline{P}(A) = \inf\{\mu(A) : \mu \in \mathcal{M}\}, \quad A \subseteq \Omega,$$

for some set $\mathcal{M}$ of probability measures on Ω .

⁴In Galvan's [2006] terminology, this idea can be expressed as relative typicality.

⁵To be precise, with linear previsions $E \in (L^\infty)^*$, which are elements of the dual space of L^∞ endowed with the weak* topology, which satisfy $E(\chi_\Omega) = 1$ and $E(X) \geq 0$ when $X \geq 0$. See [Walley, 1991]. If p is a countably additive probability measure, $\mathbb{E}_p[\cdot]$ is a finitely additive expectation, since countable additive implies finite additivity. In fact, all models considered in Section 5.4 arise as the envelopes of countably additive probability measures.

A representation of this form is called *envelope representation*.⁶ The conjugate coherent upper probability is given as $\bar{P}(A) := 1 - \underline{P}(A^c) = \sup\{\mu(A) : \mu \in \mathcal{M}\}$. Clearly, for a conjugate pair of coherent lower and upper probability, $\underline{P}(A) \leq \bar{P}(A) \forall A \subseteq \Omega$, in short $\underline{P} \leq \bar{P}$. Such a pair satisfies the following properties [Walley, 1991, Section 2.7.4] ($A, B \in \mathcal{A}$):

- P1. $\underline{P}(\emptyset) = \bar{P}(\emptyset) = 0, \underline{P}(\Omega) = \bar{P}(\Omega) = 1$.
- P2. If $A \subseteq B$, then $\underline{P}(A) \leq \underline{P}(B)$ and $\bar{P}(A) \leq \bar{P}(B)$.
- P3. $\bar{P}(A \cup B) \leq \bar{P}(A) + \bar{P}(B)$.
- P4. If $A \cap B = \emptyset$, then $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B)$.

However, these properties are not sufficient for coherence.

In the precise case, a probability p and an expectation functional $\mathbb{E}_p[\cdot]$ are in a one-to-one correspondence. We remark that in the imprecise case, the generalized expectation functional, called *coherent lower (upper) prevision*, turns out to be more general,⁷ and is therefore the focus of study in the literature (e.g. [Miranda and Cooman, 2014]): a coherent lower prevision is in a one-to-one correspondence to a closed, convex set of probabilities. For a coherent lower probability, the relation is not one-to-one [Walley, 1991, Section 2.7.3]. Therefore we will mostly work directly with sets of probabilities.

Definition 236. We say that functional $\underline{R} : L^\infty \rightarrow \mathbb{R}$ is a coherent lower prevision if

$$\underline{R}(X) = \inf\{\mathbb{E}_\mu[X] : \mu \in \mathcal{M}\}, \quad X \in L^\infty,$$

for some set $\mathcal{M}$ of probability measures on Ω . Similarly, $\bar{R}(X) = -\underline{R}(-X)$ is the conjugate coherent upper prevision.

Due to the conjugacy, it suffices to consider either $\underline{R}$ or $\bar{R}$ (respectively, $\underline{P}$ or $\bar{P}$). The set $\mathcal{M}$ is unique up to closed convex hull. We may sometimes write $\underline{R}_\mathcal{M}$ and $\bar{R}_\mathcal{M}$ to make the dependence on the set of measures explicit. We remark that a coherent upper prevision $\bar{R}$ is equivalently a *coherent risk measure* in the sense of Artzner et al. [1999].

5.3 NON-STATIONARY, LOCALLY PRECISE DATA MODELS

Since we work with a finite possibility set Ω , we can identify each probability measure p on Ω with a point in the simplex $\Delta^k := \{(p_1, p_2, \dots, p_k) : \sum_{j=1}^k p_j = 1\}$; it induces a linear expectation functional $\mathbb{E}_p[\cdot]$ in the familiar way. We abuse notation and write $p(A)$ when $p \in \Delta^k, A \subseteq \Omega$. Due to the finite dimensionality, we assume the topology induced by the Euclidean metric on Δ^k without loss of generality (see [Fröhlich et al., 2024, Appendix B.1])

⁶Here, $\underline{P}$ is the *envelope* of $\mathcal{M}$. Other authors call the set $\mathcal{M}$ itself the *envelope*, e.g. Rockafellar and Royset [2015].

⁷Slightly more general even are *sets of desirable gambles* [Augustin et al., 2014], which can handle conditioning on events of lower probability zero.

for details). Fix a set $\emptyset \neq \mathcal{M} \subseteq \Delta^k$. For any specified sequence of probability measures $m^\infty = (m_1, m_2, \dots) \in \mathcal{M}^\infty$, $m_i \in \mathcal{M}$, we can define a generator G so that

$$\lambda \{z \in \mathcal{Z} : (\omega_1, \omega_2, \dots, \omega_n) \subset G(z)\} := \prod_{i=1}^n m_i(\{\omega_i\}), \quad \forall n \in \mathbb{N}.$$

for some $W_i : [0, 1) \rightarrow \Omega$, $i = 1.. \infty$ and $G : [0, 1) \rightarrow \Omega^\infty$ with $G(z) := (W_1(z), W_2(z), \dots)$. In this model, we keep probabilistic independence but allow the probability to vary “over time” (when conceiving of the index i as time), under the constraint that the probabilities belong to $\mathcal{M}$.

It is clear that such a model can exhibit local regularity when $\mathcal{M}$ is not a singleton. More intricate is the question of the aggregate (ir)regularity of such a model. We illustrate this first with a simple example of two coins, taken from [Cozman and Chrisman \[1997\]](#). As a prerequisite, we define the relative frequencies of a data sequence $\omega^\infty = (\omega_1, \omega_2, \dots)$ as:

$$r_n^{\omega^\infty} := \left(\frac{1}{n} \sum_{i=1}^n \chi_{\{\omega^1\}}(\omega_i), \frac{1}{n} \sum_{i=1}^n \chi_{\{\omega^2\}}(\omega_i), \dots, \frac{1}{n} \sum_{i=1}^n \chi_{\{\omega^k\}}(\omega_i) \right)^\top \in \Delta^k, \quad n \in \mathbb{N},$$

where we stack relative frequencies of each elementary event ω^j , $j = 1..k$, in a vector; similarly for the finite case. We write $r_n^{\omega^\infty}(A)$ for $\mathbb{E}_{r_n^{\omega^\infty}}[\chi_A]$, $A \subseteq \Omega$, i.e. the relative frequency of the event A .

The next examples show that both aggregate regularity and aggregate irregularity can result from such a model [[Cozman and Chrisman, 1997](#)].

Example 237. Consider two coins, associated with probabilities $1/3$ and $2/3$, respectively, for the event “heads”. Construct a sequence of probabilities by choosing them in an alternating fashion. Under this independent product, almost surely, the limiting relative frequency of the event “heads” is $1/2$. That is, we obtain a stable aggregate, even though there is a hidden heterogeneity, which would be revealed by the selection rule $S(i) = 1$ if i is odd, 0 otherwise and similarly for “even”.

Example 238. Using the coins as probabilities, construct the sequence as follows. For the i -th flip, if the second most significant bit of the binary expansion of i is 0 , take the coin with probability $1/3$ for “heads”. Otherwise, take the coin with probability $2/3$ for “heads”. As a result, almost surely there exists no limiting relative frequency for the event “heads”.

When the limit of relative frequencies does not exist, the natural generalization to consider is a set of cluster points (see [[Fröhlich et al., 2024](#), Section 2.3] for an argument). Recall that $c \in \mathbb{R}^k$ is a cluster point of a sequence $a^\infty = (a_1, a_2, \dots) \in (\mathbb{R}^k)^\infty$ if

$$\forall \varepsilon > 0 : \forall n_0 \in \mathbb{N} : \exists n \geq n_0 : a_n \in B_\varepsilon(c), \quad B_\varepsilon(c) := \{x \in \mathbb{R}^k : d(x, c) < \varepsilon\},$$

where $d(\cdot, \cdot)$ is the Euclidean metric. In Example 238, the set of cluster points of relative frequencies for the event “heads” is the interval $[4/9, 1/2]$. When a topological space is sequentially compact, which Δ^k with the Euclidean topology is, then a point is a cluster point of a sequence if and only if it is the limit of a convergent subsequence.

Indeed, the focus on relative frequencies of elementary events is justified due to the following.

Proposition 239 ([Ivanenko and Pasichnichenko, 2017, Theorem 1], [Fröhlich et al., 2024, Proposition 2.1]). *Let $\omega^\infty = (\omega_1, \omega_2, \dots)$ a data sequence and $\ell \in L^\infty$ any gamble. Then it holds:*

$$\text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n \ell(\omega_i) \right) = \left\{ \mathbb{E}_p[\ell] : p \in \text{CP}(n \mapsto r_n^{\omega^\infty}) \right\}$$

Here, the operator $\text{CP}(\cdot)$ gives the cluster points of a sequence which is $\mathbb{R}$ -valued (for cluster points of gamble averages) or $\mathbb{R}^k$ -valued (for cluster points of relative frequencies) with respect to the appropriate Euclidean metric in both cases. If Ω is infinite, a similar statement would hold true when using the weak* topology on the dual space of L^∞ but then sequential compactness does not hold (see [Walley, 1991, Appendix D4]). In words, knowing the cluster points of relative frequencies suffices to obtain the cluster points of gamble averages for any gamble $\ell \in L^\infty$. When a decision maker is interested in long-run loss but a limit does not exist, considering the cluster points provides a natural generalization that may be used for aggregate decision-making.

Our main theorem characterizes exactly which aggregate (ir)regularities can result by choosing from a set of measures to form an infinite independent product.

Theorem 240. *Let $\mathcal{Z} = [0, 1)$ and $\emptyset \neq \mathcal{M} \subseteq \Delta^k$. Then there exists a sequence of probability measures $m_1, m_2, \dots \in \mathcal{M}$ with independent product*

$$\lambda \{z \in \mathcal{Z} : (\omega_1, \omega_2, \dots, \omega_n) \in G(z)\} := \prod_{i=1}^n m_i(\{\omega_i\}), \quad \forall n \in \mathbb{N}, \quad (5.3)$$

so that it holds

$$\lambda \left\{ z \in \mathcal{Z} : \text{CP} \left(n \mapsto r_n^{G(z)} \right) = \text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right) = \mathcal{N} \right\} = 1 \quad (5.4)$$

if and only if $\emptyset \neq \mathcal{N} \subseteq \text{co}(\mathcal{M})$ is a closed connected subset of the closed convex hull of $\mathcal{M}$.

We can in addition demand that any $m \in \mathcal{M}$ is a cluster point of the sequence of measures m_i (with respect to the Euclidean metric).

For the proof see Section 5.7.2. The additional demand that any $m \in \mathcal{M}$ is a cluster point of the m_i sequence only strengthens the statement, meaning intuitively even if any $m \in$

$\mathcal{M}$ approximately appears infinitely often in the m_i sequence, the set of cluster points of relative frequencies can still be made to coincide with the potentially smaller $\mathcal{N}$.

Definition 241. For brevity, we define $\mathcal{C}(\mathcal{M})$ as the set of closed connected subsets of the closed convex hull of some $\mathcal{O} \neq \mathcal{M} \subseteq \Delta^k$. Hence $\mathcal{N} \in \mathcal{C}(\mathcal{M})$ in the above statement.

Theorem 240 can be viewed as constructing a data model with a typicality notion, which in this case is indeed “probabilistic” due to the Lebesgue measure λ . We define a data model from the converse direction.

Definition 242. Let $\mathcal{Z} = [0, 1)$ and $m^\infty = (m_1, m_2, \dots) \in (\Delta^k)^\infty$. Set

$$\mathcal{M} := \{m_1, m_2, \dots\} \quad \text{and} \quad \mathcal{N} := \text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right) \in \mathcal{C}(\mathcal{M}).$$

Consider the resulting independent product as in Theorem 240, specified by $G(z)$. We call it the non-stationary, locally precise (NSLP) data model with $(\mathcal{M}, \mathcal{N})$ -imprecision and probabilities m^∞ . The Lebesgue measure λ acts as a typicality notion; formally, $T_a(Z) := \lambda(\mathcal{Z} \setminus Z)$, $Z \in \mathcal{B}([0, 1))$, is an absolute typicality measure.

Here, $\mathcal{M}$ reflects local regularity, whereas $\mathcal{N}$ embodies aggregate irregularity. For full rigour we should write “potentially non-stationary”; note that the label “non-stationary” is misleading when m^∞ is an infinite *i.i.d.* product; of course here we are hardly interested in the *i.i.d.* models as extreme cases. Note that we do not make further assumptions on m^∞ here. For instance, a probability measure m need not be a cluster point of m^∞ , implying that it cannot be “revealed” asymptotically by a selection rule. As an independent product of precise probability measures, but with potential non-stationarity, we believe this data model has many applications in machine learning and statistics (see Section 5.5.1). This view is only a mild departure from the familiar *i.i.d.* model, since at each “timestep” $i \in \mathbb{N}$, data is still modelled as drawn from a precise probability, but the probability varies over “time”. Hence imprecision cannot be considered a local characteristic of the data model. In other words, the imprecision manifests itself through non-stationarity in the forms of aggregate (ir)regularity and local (ir)regularity.

Under this data model, parameterized by a sequence of measures, diverging relative frequencies with cluster points $\mathcal{N}$ are typical. Observe that to obtain such typical divergence of relative frequencies, the sequence m^∞ cannot itself converge, since then $\mathcal{N}$ must be a singleton that coincides with this limit: in other words, perfect local irregularity (as in the *i.i.d.* model) cannot yield aggregate irregularity.

As a trivial corollary of Theorem 240, we can characterize the limiting behaviour of subsequences extracted by selection rules. Denote by ω_S^∞ the subsequence of ω^∞ extracted by some selection rule S (keeping outcomes at those indices for which $S(i) = 1$).

Corollary 243. Let $m^\infty = (m_1, m_2, \dots) \in (\Delta^k)^\infty$ as in Definition 242 and the corresponding non-stationary, locally precise data model, and $S : \mathbb{N} \rightarrow \{0, 1\}$ a selection rule. Then there exists a unique $\mathcal{N} \in \mathcal{C}(\mathcal{M})$ so that

$$\lambda \left\{ z \in \mathcal{Z} : \omega^\infty := G(z) : \text{CP}(n \mapsto r_n^{\omega^\infty}) = \mathcal{N} \right\} = 1.$$

The implication: even if the manifestation we are interested in is local (ir)regularity, the result on aggregate (ir)regularity is useful since it tells us what behaviour we can expect along selection rules: asymptotically, relative frequencies along a selection rule never end up outside the closed convex hull of $\mathcal{M}$. Intuitively, any hidden heterogeneity is confined to lie therein. Since a selection rule induces an aggregate in its own right, a subsequence, it can itself exhibit aggregate (ir)regularity — indeed it is even possible to have overall aggregate regularity, but that a selection rule yields aggregate irregularity.⁸ Furthermore, if some $m \in \mathcal{M}$ is a cluster point of the sequence of measures, there exists a selection rule which “reveals” it.

Proposition 244. Let $m^\infty = (m_1, m_2, \dots) \in (\Delta^k)^\infty$ so that $m \in \text{CP}(i \mapsto m_i)$. Then there exists a selection rule $S : \mathbb{N} \rightarrow \{0, 1\}$ so that

$$\lambda \left\{ z \in \mathcal{Z} : \omega^\infty := G(z) : \text{CP}(n \mapsto r_n^{\omega^\infty}) = m \right\} = 1.$$

Proof Take any sequence $\varepsilon^\infty = (\varepsilon_1, \dots) \downarrow 0$. Since $m \in \text{CP}(i \mapsto m_i)$, we can, to any ε_i find some i^* so that $d(m_{i^*}, m) < \varepsilon_i$ and that the sequence of the i^* is strictly increasing. Choose these i^* to construct a selection rule, i.e. $S(j) = 1$ if $j = i^*$ for some ε_i , 0 otherwise. That is, this selection rule selects a subsequence of measures which converges to m , but then also the sequence of averages of those measures converges to m . Applying Theorem 240 on this subsequence then gives the result. ■

Observe that in general, if $p \in \mathcal{N}$ is *not* in $\text{CP}(i \mapsto m_i)$ but in $\text{CP}(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i)$, then clearly there need not exist a selection rule that reveals it as local irregularity in the above sense.

As a special case of Theorem 240, we obtain Theorem 3.3 in [Fröhlich et al., 2024].

Corollary 245. Take $\mathcal{M} = \{e_1, \dots, e_k\}$ where $e_i = (0, \dots, 1, \dots, 0)$ with the 1 at the i -th position, that is, the vertices of the simplex. Then Theorem 240 asserts that if and only if $\mathcal{N} \subseteq \Delta^k$ is closed and connected, there exists a sequence ω^∞ so that $\text{CP}(n \mapsto r_n^{\omega^\infty}) = \mathcal{N}$.

In contrast to us, Fröhlich et al. [2024] have worked in a *strictly frequentist setting*, meaning that they have taken the sequence ω^∞ as the primitive entity, and defined a coherent lower

⁸Consider the infinite binary sequence $\omega^\infty = (0, 1, 0, 1, \dots)$, with aggregate regularity. Acknowledging the existence of binary sequences with diverging relative frequencies implies that we can find a selection rule that yields aggregate irregularity.

(upper) prevision based on the set of cluster points of $r_n^{\omega_\infty}$. In this way, they follow up on the work of [Ivanenko \[2010\]](#). In their approach, data models do not appear. Put differently, process randomness and typicality have no place in their framework, only outcome randomness.

5.3.1 A COMPARISON TO THE FIERENS-FINE MODEL

It is also pertinent to compare the data model of Definition 242 to the works of [Fierens and Fine \[2001, 2003\]](#), [Fierens \[2003\]](#), [Rêgo and Fine \[2005\]](#), [Fierens \[2009\]](#), [Fierens et al. \[2009\]](#). In the early paper [[Fierens and Fine, 2001](#)] the authors focus on the manifestation of aggregate (ir)regularity, as they study “persistent oscillations of relative frequencies”, following up on a previous line of work which we discuss in Section 5.4.1. However the paper [[Fierens and Fine, 2003](#)] marks a conceptual turning point:⁹ from there on the work focuses on the manifestation local (ir)regularity. Why? Our understanding of their motivation for disregarding aggregate (ir)regularity is that they wanted “to ensure that there was no further exploitable structure in the time evolution” [[Fierens and Fine, 2003](#)]. Yet as [Fierens \[2003, p. 43\]](#) points out in the language of game-theoretic probability, (in our reading) a skeptic who is allowed to take up betting commitments from a decision maker whose estimation method is focused on aggregate (ir)regularity can exploit local (ir)regularity to get infinitely rich. This can be illustrated with Example 238: here, the underlying probabilities are the set $\{1/3, 2/3\}$, and the cluster points of relative frequencies are the interval $[4/9, 1/2]$. If a skeptic knows the pattern for which coin is chosen based on $i \in \mathbb{N}$, and the decision maker offers fair bets based on the interval $[4/9, 1/2]$,¹⁰ skeptic can get infinitely rich when gambling infinitely often since $[4/9, 1/2] \subsetneq [1/3, 2/3]$. We think this is hardly surprising, and serves to highlight that whether a focus on aggregate (ir)regularity or local (ir)regularity is warranted depends on the decision-making context, see Section 5.5.1. In [[Fierens and Fine, 2003](#)] and subsequent papers, the authors then consider a data model similar to the NSLP model, and compute relative frequencies along selection rules. Their model is more expressive, however, since they allow the choice of probabilities p_i to depend on previously observed outcomes. In this paper, for simplicity we focus on the case where the sequence of probabilities is fixed a priori. We believe that Theorem 273 can be extended to their setting, but leave the details to the interested reader.¹¹ We also note that

⁹[Fierens and Fine \[2003\]](#) write “We subsequently judged this approach to be inadequate, in part after considering the performance of martingale betting systems on such time series as advocated by the then newly-published [book by] Shafer and Vovk [..]”; they refer here to the game-theoretic account of probability put forward by [Shafer and Vovk \[2001\]](#).

¹⁰Such a bet is “fair” when it consists in playing a *marginally desirable gamble* [[Walley, 1991](#)]. For example, the gamble $X - \bar{P}(X)$ and $\underline{P}(X) - X$ are marginally desirable for a decision maker with coherent upper prevision $\bar{P}$. In the precise case, this is equivalent to $\mathbb{E}_p[X - \mathbb{E}_p[X]] = 0$, i.e. $\mathbb{E}_p$ determines betting rates.

¹¹To this end, consider the following variant of the strong law of large numbers given by Loève (1978, as cited in [[Papamarcou and Fine, 1991b](#)]): For an increasing sequence of σ -algebras $\mathcal{B}_i$ and a uniformly bounded

the earlier paper by [Cozman and Chrisman \[1997\]](#) represents a “bridge” between the two manifestations, since they consider both subsequences of the relative frequency sequence as well as subsequences of the data sequence.

Starting with the non-stationary, locally precise data model, more exotic data models can be obtained, where the imprecision is already inherent to each “timestep” itself, meaning it is a local feature. As we will see, imprecision can therefore manifest even under *stationarity*. Investigating such models is also of interest since it explicates the relation of Theorem 240 and the present data model to the works of [Walley and Fine \[1982\]](#) and subsequent papers [[Kumar and Fine, 1985](#), [Grize and Fine, 1987](#), [Fine, 1988](#), [Papamarcou and Fine, 1991a,b](#), [Sadrolhefazi and Fine, 1994](#)]. One of our goals is to provide an accessible introduction to this line of work, therefore we now consider such more general models.

5.4 STATIONARY, LOCALLY IMPRECISE DATA MODELS

Previously we have fixed a specific sequence of probability measures from $\mathcal{M}$. Like [Kumar and Fine \[1985, Section 5\]](#), we now construct data models which subsume whole sets of sequences of measures. The resulting models are highly related to those of [Walley and Fine \[1982\]](#). Let $\emptyset \neq \mathcal{M} \subseteq \Delta^k$ be a set of probability measures. We write $\underline{P}_{\mathcal{M}}(A) := \inf\{m(A) : m \in \mathcal{M}\}$, $A \subseteq \Omega$, and recall that we denote by $\mathcal{M}^\infty$ the set of sequences of probability measures from $\mathcal{M}$.

Given a subset $\emptyset \neq \tilde{\mathcal{M}}^\infty \subseteq \mathcal{M}^\infty$ we construct a data model based on independent products. Define the infinite product space of Ω and the associated product σ -algebra as:

$$\Omega^\infty := \prod_{i=0}^{\infty} \Omega, \quad \prod_{i=0}^{\infty} 2^\Omega := \sigma \left(\bigcup_{i=1}^{\infty} \sigma(\pi_i) \right),$$

where the projections are $\pi_i : \Omega^\infty \rightarrow \Omega$, $\pi_i(\omega^\infty = (\omega_1, ..)) := \omega_i$. This product σ -algebra coincides with the Borel σ -algebra $\mathcal{B}(\Omega^\infty)$ when assuming the product topology induced by the discrete topology on Ω . For terse notation, we therefore use $\mathcal{B}(\Omega^\infty)$. A set $A^\infty \in \mathcal{B}(\Omega^\infty)$ is a set of infinite sequences; but note that $\mathcal{B}(\Omega^\infty) \subsetneq 2^{\Omega^\infty}$. Fix any sequence of probability measures $m^\infty = (m_1, m_2, ..) \in \tilde{\mathcal{M}}^\infty$. Ionescu-Tulcea’s theorem [[Klenke, 2013, Section 14.3](#)] guarantees the existence of a unique measure $\text{IT}(m^\infty)$ on $(\Omega^\infty, \mathcal{B}(\Omega^\infty))$ which satisfies

$$\text{IT}(m^\infty) \left(A_1 \times .. \times A_n \times \prod_{i=n+1}^{\infty} \Omega \right) := \prod_{i=1}^n m_i(A_i), \quad \forall n \in \mathbb{N}_0. \quad (5.5)$$

and adapted sequence of measurable functions Y_i , it holds that $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (Y_{i+1} - \mathbb{E}[Y_{i+1} | \mathcal{B}_i]) = 0$ *almost surely*.

In particular, this implies

$$\text{IT}(m^\infty)(\{\omega^\infty = (\omega_1, \omega_2, \dots, \omega_n)\}) := \prod_{i=1}^n m_i(\{\omega_i\}), \quad \forall n \in \mathbb{N}_0.$$

For each such sequence $m^\infty \in \tilde{\mathcal{M}}^\infty$, we obtain a measure $\text{IT}(m^\infty)$ on $(\Omega^\infty, \mathcal{B}(\Omega^\infty))$ from Ionescu-Tulcea's theorem. Since these measures are all defined on the same probability space, we can define a coherent upper probability from them via an envelope representation.

Depending on how we constrain the set $\tilde{\mathcal{M}}^\infty$, we can obtain different data models. For instance, based on a set of probability measures $\mathcal{M}$, we could consider *all* sequences of measures from $\mathcal{M}$, yielding the following.

Definition 246. Let $\emptyset \neq \mathcal{M} \subseteq \Delta^k$. We define the stationary, locally imprecise (SLI) data model with $\mathcal{M}$ -imprecision via a coherent lower probability:

$$\underline{P}_{\mathcal{M}}^\infty(A^\infty) := \inf \{\text{IT}(m^\infty)(A^\infty) : m^\infty \in \mathcal{M}^\infty\}, \quad \forall A^\infty \in \mathcal{B}(\Omega^\infty),$$

with conjugate coherent upper probability $\bar{P}_{\mathcal{M}}^\infty(A^\infty) = 1 - \underline{P}_{\mathcal{M}}^\infty((A^\infty)^c)$.

This model is essentially (see the discussion below) equivalent to the *i.i.d.* model of [Walley and Fine \[1982\]](#). We have here expressed a notion of typicality on $\mathcal{B}(\Omega^\infty)$ through a coherent lower probability: intuitively, $\underline{P}_{\mathcal{M}}^\infty(A^\infty) = 1$ means that A^∞ is a typical set, since this implies that all measures in the envelope assign measure 1 to the set. We can formalize this in the framework of [Galvan \[2006\]](#).

Proposition 247. Let $\bar{P}$ be a coherent upper probability, defined on some possibility set Λ and set system $\mathcal{A}$, which is closed under complementation and $\emptyset \in \mathcal{A}$. Then $d(A, B) := \bar{P}(A \triangle B)$ satisfies the axioms of a typicality distance as in [\[Galvan, 2006\]](#). Thus $(\Omega, \mathcal{A}, d)$ forms a typicality space [\[Galvan, 2006\]](#). The induced absolute typicality operator is $T_a(A) = 1 - \underline{P}(A)$, where $\underline{P}$ is conjugate to $\bar{P}$.

Corollary 248. If $d(A^\infty, B^\infty) := \bar{P}_{\mathcal{M}}^\infty(A^\infty \triangle B^\infty)$, then $(\Omega^\infty, \mathcal{B}(\Omega^\infty), d)$ forms a typicality space.

For the definitions of the involved concepts and the proof, see Appendix 5.7.1 Thus, $T_a(A^\infty) = 0 \Leftrightarrow \underline{P}_{\mathcal{M}}^\infty(A^\infty) = 1$ means that A^∞ is a typical set. To better fit the notion of an abstract random seed, we could have also defined it on $[0, 1)$ instead.¹²

In the above model, we have allowed arbitrary sequences of probability measures from $\mathcal{M}$. Another possibility is to constrain the set by a specific set of cluster points. Consider the

¹²One possibility would be embedding Ω into $\mathbb{R}$, so that the m_i would share the same finite support, and then exploit that $[0, 1)$ and $\mathbb{R}^\infty$ are Borel-isomorph.

subset

$$\mathcal{M}_{\rightarrow \mathcal{N}}^\infty := \left\{ m^\infty = (m_1, m_2, \dots) \in \mathcal{M}^\infty : \text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right) = \mathcal{N} \right\},$$

for some $\mathcal{N} \in \mathcal{C}(\mathcal{M})$, i.e. a closed connected subset of the closed convex hull of $\mathcal{M}$. In this set, all sequences of probability measures from $\mathcal{M}$ are collected which yield a specific set $\mathcal{N}$ as cluster points. This leads to the following more constrained model.

Definition 249. Let $\emptyset \neq \mathcal{M} \subseteq \Delta^k$ and $\mathcal{N} \in \mathcal{C}(\mathcal{M})$. We define the stationary, locally imprecise data model with $(\mathcal{M}, \mathcal{N})$ -imprecision through a notion of typicality expressed by a coherent lower probability:

$$\underline{P}_{\mathcal{M}_{\rightarrow \mathcal{N}}^\infty}^\infty(A^\infty) := \inf \{ \text{IT}(m^\infty)(A^\infty) : m^\infty \in \mathcal{M}_{\rightarrow \mathcal{N}}^\infty \}, \quad \forall A^\infty \in \mathcal{B}(\Omega^\infty).$$

Clearly, cluster points of relative frequencies will typically coincide with $\mathcal{N}$ in this data model.

Proposition 250. Consider the event $\text{CP}=\mathcal{N} := \{ \omega^\infty : \text{CP}(n \mapsto r_n^{\omega^\infty}) = \mathcal{N} \} \in \mathcal{B}(\Omega^\infty)$. Then $\underline{P}_{\mathcal{M}_{\rightarrow \mathcal{N}}^\infty}^\infty(\text{CP}=\mathcal{N}) = \bar{P}_{\mathcal{M}_{\rightarrow \mathcal{N}}^\infty}^\infty(\text{CP}=\mathcal{N}) = 1$, that is, $\text{CP}=\mathcal{N}$ is a typical set.

Proof First observe that indeed $\text{CP}=\mathcal{N} \in \mathcal{B}(\Omega^\infty)$. Next, by construction, if $m^\infty \in \mathcal{M}_{\rightarrow \mathcal{N}}^\infty$, then $\text{IT}(m^\infty)(\text{CP}=\mathcal{N}) = 1$. But then taking the infimum gives $\underline{P}_{\mathcal{M}_{\rightarrow \mathcal{N}}^\infty}^\infty(\text{CP}=\mathcal{N}) = 1$. ■

5.4.1 A COMPARISON TO WALLEY AND FINE'S [1982] AND FINE ET AL.'S LINE OF WORK

It is instructive to compare our models to the seminal work of Walley and Fine [1982] and the subsequent line of work by Fine and collaborators (starting with [Kumar and Fine, 1985]). We begin with a comparison to Walley and Fine [1982], which yields important insight into the behaviour of these models. We consider their imprecise *i.i.d.* model, which is also what they focus on. In contrast to our coherent lower probability, theirs is defined on the set system

$$\mathcal{A}^\infty := \bigcup_{n \in \mathbb{N}} \mathcal{A}^n, \quad \text{where } \mathcal{A}^n := \sigma \left(\bigcup_{i=1}^n \sigma(\pi_i) \right).$$

Recall that the projections are $\pi_i : \Omega^\infty \rightarrow \Omega$, $\pi_i(\omega^\infty = (\omega_1, \dots)) := \omega_i$. Note that $\mathcal{A}^\infty$ is an algebra (also called field), but not a σ -algebra. Since $\sigma(\mathcal{A}^\infty) = \mathcal{B}(\Omega^\infty)$, their set system is a strict subset of ours. Our model partially agrees with the *i.i.d.* model of Walley and Fine [1982] in the following sense. We denote by $\underline{P}_{\mathcal{M}, \text{WF}}^\infty : \mathcal{A}^\infty \rightarrow [0, 1]$ their *i.i.d.* model, which is also parameterized by a set of measures $\mathcal{M}$. Intuitively, this model corresponds to $\underline{P}_{\mathcal{M}}^\infty$, since it is built from independent lower probabilities, each of which is the envelope of $\mathcal{M}$;

for the detailed construction, we refer the reader to [Walley and Fine \[1982\]](#). However, for events which depend only on a finite number of outcomes, the model agrees with both $\underline{P}_{\mathcal{M}^\infty \rightarrow \mathcal{N}}^\infty$ and $\underline{P}_{\mathcal{M}}^\infty$

Proposition 251. *Let $A^\infty := A_1 \times \dots \times A_n \times \bigtimes_{i=n+1}^\infty \Omega$ be a rectangle in $\mathcal{B}(\Omega^\infty)$, $A_i \in 2^\Omega$. It holds that*

$$\underline{P}_{\mathcal{M}^\infty \rightarrow \mathcal{N}}^\infty(A^\infty) = \underline{P}_{\mathcal{M}}^\infty(A^\infty) = \underline{P}_{\mathcal{M}, \text{WF}}^\infty(A^\infty) = \inf \left\{ \prod_{i=1}^n m_i(A_i) : m_i \in \mathcal{M} \right\}, \quad \forall n \in \mathbb{N}_0.$$

Proof The part of the statement regarding $\underline{P}_{\mathcal{M}, \text{WF}}^\infty(A^\infty)$ follows from its definition, see [Walley and Fine \[1982\]](#). The statement regarding $\underline{P}_{\mathcal{M}}^\infty$ is clear from its definition. For $\underline{P}_{\mathcal{M}^\infty \rightarrow \mathcal{N}}^\infty$, we know that

$$\underline{P}_{\mathcal{M}^\infty \rightarrow \mathcal{N}}^\infty \left(A_1 \times \dots \times A_n \times \bigtimes_{i=n+1}^\infty \Omega \right) = \inf \left\{ \prod_{i=1}^n m_i(A_i) : m^\infty = (m_1, m_2, \dots) \in \mathcal{M}_{\rightarrow \mathcal{N}}^\infty \right\}, \quad \forall n \in \mathbb{N}_0.$$

But since the cluster points of m^∞ remain unchanged when prepending an arbitrary prefix to the sequence of measures, the statement follows obviously: any chosen prefix $(m_1, \dots, m_n) \in \mathcal{M}^n$ can be extended to lie in $\mathcal{M}_{\rightarrow \mathcal{N}}^\infty$. $\blacksquare$

The algebra that is generated by sets of the above form, the rectangles, is $\mathcal{A}^\infty$, which implies the following.

Corollary 252. *Let $A^\infty \in \mathcal{A}^\infty \subsetneq \mathcal{B}(\Omega^\infty)$. Then*

$$\underline{P}_{\mathcal{M}^\infty \rightarrow \mathcal{N}}^\infty(A^\infty) = \underline{P}_{\mathcal{M}}^\infty(A^\infty) = \underline{P}_{\mathcal{M}, \text{WF}}^\infty(A^\infty).$$

Proof Rectangles as above form a π -system, i.e. a set system closed under finite intersections, and hence the measures in the envelope, which agree pairwise on this π -system, extend uniquely to algebra generated by it (according to the π - λ -theorem), which is $\mathcal{A}^\infty$. By pairwise agreement, we mean that for any measure in the envelope of $\underline{P}_{\mathcal{M}}$ ($\underline{P}_{\mathcal{M}^\infty \rightarrow \mathcal{N}}$, respectively), we can find a measure in the envelope of $\underline{P}_{\mathcal{M}, \text{WF}}^\infty$ so that these agree on the π -system, and conversely. $\blacksquare$

In contrast, for events $A^\infty \in \mathcal{B}(\Omega^\infty) \setminus \mathcal{A}^\infty$, the assessments of the models can be starkly different. Define the tail σ -algebra induced by the projections as

$$\mathcal{T} := \bigcap_{n \in \mathbb{N}} \sigma(\pi_{n+1}, \pi_{n+2}, \dots), \quad \mathcal{T} \subsetneq \mathcal{B}(\Omega^\infty).$$

Intuitively, $\mathcal{T}$ is the sub- σ -algebra of $\mathcal{B}(\Omega^\infty)$ which does not depend on the outcomes of finitely many trials. It is clear that $A^\infty \in \mathcal{T} \implies A^\infty \notin \mathcal{A}^\infty$; however note that $\mathcal{B}(\Omega^\infty) \setminus \mathcal{T} \neq \mathcal{A}^\infty$, implying that $\mathcal{T} \subsetneq \mathcal{B}(\Omega^\infty) \setminus \mathcal{A}^\infty$. For illustration, consider the event $\{\omega^\infty : \omega^\infty = \omega_0^\infty\}$,

asserting that the sequence is exactly a specified sequence ω_0^∞ . This event is neither in $\mathcal{T}$ nor in $\mathcal{A}^\infty$.

In a sense, $\underline{P}_\mathcal{M}^\infty$ can be viewed as an extension of the Walley-Fine *i.i.d.* model to all of $\mathcal{B}(\Omega^\infty)$, so that the event of divergence is a measurable tail event: $CP=\mathcal{N} \in \mathcal{T} \subsetneq \mathcal{B}(\Omega^\infty)$.

Proposition 253. *Let $|\mathcal{M}| > 1$. It holds that:*

$$(253.1) \quad \underline{P}_{\mathcal{M} \rightarrow \mathcal{N}}^\infty(CP=\mathcal{N}) = \overline{P}_{\mathcal{M} \rightarrow \mathcal{N}}^\infty(CP=\mathcal{N}) = 1.$$

$$(253.2) \quad \underline{P}_\mathcal{M}^\infty(CP=\mathcal{N}) = 0 \text{ but } \overline{P}_\mathcal{M}^\infty(CP=\mathcal{N}) = 1.$$

Proof Property (253.1) is just a restatement of Proposition 250. To see that $\underline{P}_\mathcal{M}^\infty(CP=\mathcal{N}) = 0$, assume first that $\mathcal{N} \subsetneq \text{c}\bar{\text{o}}(\mathcal{M})$. Then consider the sequence $m^\infty = (m_0, m_0, \dots)$ for some $m_0 \in \mathcal{M} \cap \mathcal{N}^c$. Then $IT(m^\infty)(CP=\mathcal{N}) = 0$. On the other hand, if $\mathcal{N} = \text{c}\bar{\text{o}}(\mathcal{M})$, then take any $m_0 \in \mathcal{M}$, yielding $IT(m^\infty)(CP=\mathcal{N}) = 0$. Finally, $\overline{P}_\mathcal{M}^\infty(CP=\mathcal{N}) = 1$ follows from Theorem 240. ■

We emphasize again that throughout this section we assume that $\mathcal{N}$ is any set that satisfies the appropriate relation to $\mathcal{M}$, i.e. $\mathcal{N} \in \mathcal{C}(\mathcal{M})$. Intuitively, the Walley-Fine *i.i.d.* model is completely silent about the set of cluster points, about convergence or divergence. Note that in general $[\underline{P}_{\mathcal{M} \rightarrow \mathcal{N}}^\infty(A^\infty), \overline{P}_{\mathcal{M} \rightarrow \mathcal{N}}^\infty(A^\infty)] \subseteq [\underline{P}_\mathcal{M}^\infty(A^\infty), \overline{P}_\mathcal{M}^\infty(A^\infty)]$ since $\mathcal{M}_{\rightarrow \mathcal{N}}^\infty \subsetneq \mathcal{M}^\infty$ for any $\mathcal{N} \in \mathcal{C}(\mathcal{M})$.

Example 254. Take $\mathcal{M} = \{e_1, \dots, e_k\}$ where $e_i = (0, \dots, 1, \dots, 0)$ with the 1 at the i -th position. Then a suitable $\mathcal{N} \in \mathcal{C}(\mathcal{M})$ is any closed connected subset of Δ^k . Each infinite sequence for which cluster points of relative frequencies coincide with $\mathcal{N}$ then corresponds to an infinite product measures in $\mathcal{M}_{\rightarrow \mathcal{N}}^\infty$. Hence $CP=\mathcal{N}$ is a typical set. Similarly, $\underline{P}_\mathcal{M}$ is a fully vacuous model which is silent on all “nontrivial” events, since it is the envelope of all Ω^∞ , as represented by Dirac-measures.

While $\underline{P}_\mathcal{M}^\infty$ is a reasonable extension of the model to $\mathcal{B}(\Omega^\infty)$, the original proposal by Walley and Fine [1982] made the architectural choice to work on $\mathcal{A}^\infty$ as opposed to $\mathcal{B}(\Omega^\infty)$. One reason is that the scope of Walley and Fine [1982] also includes non-*i.i.d.* models in principle (dependent coherent lower probabilities), which are built up from a consistent sequence of coherent lower probabilities on $\mathcal{A}^1, \mathcal{A}^2, \dots$; but the focus of Walley and Fine [1982] is strongly on the *i.i.d.* model. The restriction to $\mathcal{A}^\infty$ means that “strong law”-related statements¹³ of the above form cannot be made. Instead, Walley and Fine [1982] have studied the event of *apparent convergence* as

$$C_n(A; k, \varepsilon) := \bigcap_{j=k}^n \left\{ \omega^\infty : |r_j^{\omega^\infty}(A) - r_n^{\omega^\infty}(A)| < \varepsilon \right\}.$$

¹³In the terminology of [Kumar and Fine, 1985] and meaning roughly *in relation to a generalization of the strong law of large numbers*.

In words, “the event that $r_1(A), \dots, r_n(A)$ apparently converges (k, ε) ”. Considering all events simultaneously, they define $C_n(k, \varepsilon) := \bigcap_{A \in 2^\Omega} C_n(A; k, \varepsilon)$. Similarly, the event of apparent divergence is $D_n(k, \varepsilon) := C_n(k, \varepsilon)^c$. Note that this event is in $\mathcal{A}^\infty$, and thus our model agrees with that of [Walley and Fine \[1982\]](#); we then state results in terms of $\underline{P}_\mathcal{M}^\infty$. We emphasize that the nature of the event $D_n(k, \varepsilon) \in \mathcal{A}^n$ is fundamentally different to $\text{CP-}\mathcal{N} \in \mathcal{T} \cap (\mathcal{A}^\infty)^c$; intuitively, $D_n(k, \varepsilon)$ is a “weak law”-related statement. Recall that the weak law of large numbers in the precise *i.i.d.* case asserts the following:

$$\forall \varepsilon > 0 : \lim_{n \rightarrow \infty} P^\infty \left(|r_n^{\omega^\infty}(A) - P(A)| > \varepsilon \right) = 0, \quad P^\infty = \text{IT}((P, P, \dots)).$$

In the imprecise case, the situation is somewhat unsettling. Recall that we write $\underline{P}_\mathcal{M}(A) := \inf\{m(A) : m \in \mathcal{M}\}$, $A \subseteq \Omega$, for the envelope of $\mathcal{M}$. Intuitively, this is the marginal of $\underline{P}_\mathcal{M}^\infty$.

Proposition 255 ([\[Walley and Fine, 1982, Theorem 4.1\]](#)). *Let $\tilde{k} : \mathbb{N} \rightarrow \mathbb{N}$, $\tilde{k}(n) \rightarrow \infty$ as $n \rightarrow \infty$. For any $\varepsilon > 0$,*

$$(255.1) \quad \lim_{n \rightarrow \infty} \underline{P}_{\mathcal{M}, \text{WF}}^\infty(D_n(\tilde{k}(n), \varepsilon)) = 0.$$

$$(255.2) \quad \text{If } |\mathcal{M}| > 1, \limsup_{n \rightarrow \infty} \tilde{k}(n)/n = 0, \text{ and } \exists A \subseteq \Omega : \varepsilon < \bar{P}_\mathcal{M}(A) - \underline{P}_\mathcal{M}(A), \text{ then}$$

$$\lim_{n \rightarrow \infty} \bar{P}_\mathcal{M}^\infty(D_n(\tilde{k}(n), \varepsilon)) = 1.$$

From (255.1), it follows that $\lim_{n \rightarrow \infty} \underline{P}_{\mathcal{M}_{\rightarrow \mathcal{N}}}^\infty(D_n(\tilde{k}(n), \varepsilon)) = 0$. The situation is a curious one: while a “strong law”-type statement holds, that divergence with a set of cluster points $\mathcal{N}$ is typical under $\underline{P}_{\mathcal{M}_{\rightarrow \mathcal{N}}}^\infty$, a corresponding “weak law” fails. To address this, [Walley and Fine \[1982\]](#) coined a substantially weaker notion of typicality, called *asymptotic favorability*, so that $D_n(\tilde{k}(n), \varepsilon)$ is found to be asymptotically favorable; the statement (255.2) is similar to Theorem 4.1.d in [Walley and Fine \[1982\]](#), which expresses asymptotic favorability of $D_n(\tilde{k}(n), \varepsilon)$: a sequence of events A_n is called asymptotically favorable if $\underline{P}^\infty(A_n^c)/\underline{P}^\infty(A_n) \rightarrow 0$, which implies $\bar{P}^\infty(A_n) \rightarrow 1$. The intuitive reason for why a “weak law”-type statement must fail in this model is that for any rate of divergence, there exists some other product measure which yields even slower divergence. Hence statements regarding apparent divergence *must* be vacuous (for appropriate ε). In contrast, to anticipate confusion for when we later contextualize our results and compare to existing weak laws for imprecise probabilities in the literature, we give a slightly modified restatement of [\[Walley and Fine, 1982, Theorem 4.1.a\]](#).

Proposition 256. *For any $\varepsilon > 0$,*

$$\lim_{n \rightarrow \infty} \underline{P}_\mathcal{M}^\infty(\{\omega^\infty : \forall A \subseteq \Omega : \underline{P}_\mathcal{M}(A) - \varepsilon < r_n^{\omega^\infty}(A) < \bar{P}_\mathcal{M}(A) + \varepsilon\}) = 1.$$

A similar proposition holds for gamble averages by replacing the lower/upper probability with the lower/upper prevision. Intuitively, this means that while “weak law” statements about divergence fail, i.e. the model is silent on them, the precise probabilistic weak law¹⁴ does hold for all of the product measures in the set of measures, and therefore we find that with increasing lower probability, relative frequencies must lie between the lower and upper probability. This should be rather unsurprising. Thus, in this section and in contrast to Section 5.4.2, when we speak of “strong law” or “weak law”-type statements, we refer to statements that concern events of divergence or convergence, instead of merely a confinement to the interval determined by the marginal $\underline{P}_{\mathcal{M}}$.

The work of [Walley and Fine \[1982\]](#) sparked a long and difficult to access line of work by Fine and collaborators, [[Kumar and Fine, 1985](#), [Grize and Fine, 1987](#), [Fine, 1988](#), [Papamarcou and Fine, 1991a,b](#), [Sadrolhefazi and Fine, 1994](#)], which took the above phenomenon, the lack of typical divergence, as the starting point. [Fine \[1988\]](#) provides a relatively accessible overview to the philosophy and approach of this line of work. These works also moved from the setup $(\Omega^\infty, \mathcal{A}^\infty)$ to $(\Omega^\infty, \mathcal{B}(\Omega^\infty))$, like ours. The authors were motivated by the study of physical systems which seemingly display the puzzling combination of stationarity and unstable averages. [Kumar and Fine \[1985\]](#) define stationarity as follows.

Definition 257. For $k \in \mathbb{N}$, $T^k : \Omega^\infty \rightarrow \Omega^\infty$ is the left-shift-by- k operator defined by

$$T^k(\omega^\infty = (\omega_1, \omega_2, \dots)) := (\omega_{k+1}, \omega_{k+2}, \dots),$$

which acts on a set $A^\infty \in \mathcal{B}(\Omega^\infty)$ as $T^k(A^\infty) := \{T^k(\omega^\infty) : \omega^\infty \in A^\infty\}$. Conversely, T^{-k} is defined by $T^{-k}(A^\infty) := \{\omega^\infty : T^k(\omega^\infty) \in A^\infty\}$, $A^\infty \in \mathcal{B}(\Omega^\infty)$.

Definition 258. A set function $\underline{P}$ on $(\Omega^\infty, \mathcal{B}(\Omega^\infty))$ is called stationary when

$$\forall A^\infty \in \mathcal{B}(\Omega^\infty) : \forall k \in \mathbb{N} : \underline{P}(T^{-k}(A^\infty)) = \underline{P}(A^\infty).$$

Indeed, the models presented in this section are stationary in this sense (compare Example 5.1 of [Kumar and Fine \[1985\]](#)).

Proposition 259. $\underline{P}_{\mathcal{M}}^\infty$ and $\underline{P}_{\mathcal{M}^\infty \rightarrow \mathcal{N}}^\infty$ are stationary.

Proof We first show stationarity of $\underline{P}_{\mathcal{M}}$. Consider an arbitrary $m^\infty = (m_1, m_2, \dots) \in \mathcal{M}^\infty$. We show that if $\text{IT}(m^\infty)(A^\infty) = a$, there is some other $\tilde{m}^\infty \in \mathcal{M}^\infty$ so that $\text{IT}(\tilde{m}^\infty)(T^{-k}(A)) = a$. From this we then conclude $\underline{P}_{\mathcal{M}}(T^{-k}(A)) \leq \underline{P}_{\mathcal{M}}(A)$. Take

$$\tilde{m}^\infty = (\underbrace{m_0, \dots, m_0}_{k \text{ times}}, m_1, m_2, \dots)$$

¹⁴A version of the weak law for independent but non-*i.i.d.* measurable functions.

for arbitrary $m_0 \in \mathcal{M}$. Consider rectangles of the form

$$\mathcal{R}_n := \left\{ \underbrace{\Omega \times \dots \times \Omega}_{k \text{ times}} \times A_1 \times \dots \times A_n \times \bigtimes_{k+n+1}^{\infty} \Omega : A_i \in 2^\Omega \right\}, \quad n \in \mathbb{N}.$$

These rectangles $\{\mathcal{R}_n : n \in \mathbb{N}\}$ generate the σ -algebra $\mathcal{A}^{k+1} := \sigma(\bigcup_{i=k+1}^{\infty} \sigma(\pi_i))$. It is clear that for these rectangles, $R \mapsto \text{IT}(\tilde{m}^\infty)(R)$ coincides with $R \mapsto m_0(\Omega) \cdot m_0(\Omega) \cdot \text{IT}(T^k(R))$ due to Equation 5.5, and $m_0(\Omega) = 1$. Since these rectangles form a π -system, these measures also coincide on the generated σ -algebra $\mathcal{A}^{k+1}$. Then, noting that $T^{-k}(A^\infty) \in \mathcal{A}^{k+1}$, we have

$$\text{IT}(\tilde{m}^\infty)(T^{-k}(A^\infty)) = \text{IT}(m^\infty)(T^k(T^{-k}(A^\infty))),$$

which since $T^k(T^{-k}(A^\infty)) = A^\infty$ implies $\text{IT}(\tilde{m}^\infty)(T^{-k}(A^\infty)) = \text{IT}(m^\infty)(A^\infty)$; we caution the reader that $T^{-k}(T^k(A^\infty)) \supseteq A^\infty$.

For the converse direction we want to show that $\underline{P}_{\mathcal{M}}(A^\infty) \leq \underline{P}_{\mathcal{M}}(T^{-k}(A))$ for $A^\infty \in \mathcal{B}(\Omega^\infty)$. For this we exhibit to any $m^\infty = (m_1, m_2, \dots) \in \mathcal{M}^\infty$ with $\text{IT } m^\infty(T^{-k}(A^\infty)) = a$ some $\tilde{m}^\infty \in \mathcal{M}^\infty$ with $\tilde{m}^\infty(A^\infty) = a$. But here it is clear that $\tilde{m}^\infty := (m_{k+1}, m_{k+2}, \dots)$ works.

Finally, for the stationarity of $\underline{P}_{\mathcal{M}_{\rightarrow \mathcal{N}}}$, observe that the logic of the proof works as well with the constraint $m^\infty \in \mathcal{M}_{\rightarrow \mathcal{N}}^\infty$ (compare the proof of Proposition 251). ■

We need some preliminaries to state the main result of [Kumar and Fine \[1985\]](#). The authors define a lower/upper probability as a conjugate pair $\underline{P}(A) = 1 - \overline{P}(A)$ of set functions, satisfying a list of axioms (implying UP1- UP4), which are weaker than the coherence condition of [Walley \[1991, Section 2.5\]](#). To a lower probability, associate the set

$$\mathcal{M}_{\underline{P}} := \{\mu : \mu \in \Delta^k \text{ and } \mu(A) \geq \underline{P}(A) \ \forall A \subseteq \Omega\}.$$

When this set is nonempty, call $\underline{P}$ a *dominated* lower probability (equivalent to the *avoiding sure loss* condition in [Walley \[1991, Section 2.4\]](#)); when it is empty, call $\underline{P}$ *undominated*. Any coherent lower probability (Definition 235) is dominated, but being dominated does not automatically imply coherence (cf. [Walley \[1991, Section 2.4, Section 2.6.3\]](#)).

Write $D := \{\omega^\infty : \exists A \subset \Omega : \limsup_{n \rightarrow \infty} r_n^{\omega^\infty}(A) > \liminf_{n \rightarrow \infty} r_n^{\omega^\infty}(A)\}$ for the event of divergence for at least one $A \subset \Omega$. [Kumar and Fine \[1985\]](#) have shown the following (here presented with adapted notation¹⁵).

Proposition 260 ([\[Kumar and Fine, 1985, Corollary 4.4\]](#)). *If $\underline{P}^\infty$ is monotone set function on*

¹⁵[Kumar and Fine \[1985\]](#) use the letter $\mathcal{A}^\infty$ for what is $\mathcal{B}(\Omega^\infty)$ in our notation, i.e. the σ -algebra generated by all projection maps, or equivalently by the rectangles or cylinder sets.

$(\Omega^\infty, \mathcal{B}(\Omega^\infty))$ which is dominated, continuous from below on $\mathcal{A}^{\infty 16}$ and stationary, then $\underline{P}^\infty(D) = 0$.

This result is intimately linked to the continuity condition. The reason for introducing this condition is that it is directly linked to the possibility of obtaining “weak law”-type statements, which allow a bridge from finite observations to divergence or convergence. Indeed, the authors give examples very similar to (special cases of) our Definition 249 of stationary coherent lower probabilities, hence dominated. In light of this result, however, such a model cannot be continuous from below. Intuitively, this prohibits any form of estimating cluster points from finite observations. As a conclusion, [Kumar and Fine \[1985\]](#) suggest that set functions which satisfy the desired properties must be sought in the area of *undominated* upper and lower probabilities, undominated implying *incoherent* in the sense of [Walley \[1991\]](#). The subsequent paper of [Papamarcou and Fine \[1991a\]](#) provided an affirmative answer to the problem of finding a lower probability which is monotonely continuous along $\mathcal{A}^{\infty 17}$, stationary, and has $\underline{P}(D) = 1$ (a partial answer was provided in [\[Grize and Fine, 1987\]](#)). In light of the above result, it is however undominated, hence incoherent. The final paper of [Sadrolhefazi and Fine \[1994\]](#) added to the result of [Papamarcou and Fine \[1991a\]](#) by introducing an additional constraint, and proving that the problem is still feasible: the constraint that the lower probability should, on finite dimensional cylinder sets, behave approximately like an additive probability.

Philosophically, [Fine \[1988\]](#) subscribe to a propensity interpretation of the lower probabilities in these models. Due to the assumption of stationarity, the whole set of probabilities is considered to be “at play” for each “timestep” — each “timestep” is characterized by an inherently imprecise propensity. To us, it remains unclear how to interpret such a propensity in the context of a data generation process. This clearly contrast this with the non-stationary, locally precise model in Section 5.3, where each “timestep” is linked to a single precise probability, and the imprecision manifests from a perspective of non-stationarity as aggregate (ir)regularity and local (ir)regularity. While we believe there might be applications for such more exotic data models like $\underline{P}_{\mathcal{M}_{\rightarrow \mathcal{N}}^\infty}$ (e.g. for certain physical phenomena such as flicker noise; see also Section 5.5.1), we are mainly motivated by problems that are better described by non-stationarity. By going beyond stationarity, we can keep coherence and continuity, and still have models which typically yield divergence. Still there might be value in stationary models to describe a non-stationary process, since the stationary model might be more parsimonious, see [\[Persiau et al., 2021\]](#) for this argument. For example, non-stationary precise Markov chains can sometimes be efficiently modelled by stationary imprecise Markov chains [\[De Cooman et al., 2009, Tjoens et al., 2019\]](#).

¹⁶Meaning if $\forall i \in \mathbb{N} : A_i^\infty \in \mathcal{A}^\infty$ and $A_i \uparrow A^\infty \in \sigma(\mathcal{A}^\infty)$, then $\underline{P}^\infty(A_i^\infty) \uparrow \underline{P}^\infty(A)$. Here, $\uparrow$ signifies “non-decreasing” and convergence of sets refers to the set-theoretic limit.

¹⁷Similar to continuity from below, see [\[Papamarcou and Fine, 1991a\]](#)

5.4.2 A COMPARISON TO GENERALIZED LAWS OF LARGE NUMBERS AND THE SUBJECTIVIST PERSPECTIVE

It is insightful to contextualize these data models in another vein of literature on generalized *laws of large numbers* for various non-additive set functionals, going under the names of capacities or imprecise probabilities, and for the generalized expectation functionals called coherent lower (upper) previsions, coherent risk measures, or more recently, G-expectations [Dow and Werlang, 1993, Marinacci, 1999, Epstein and Schneider, 2003, Maccheroni and Marinacci, 2005, Peng, 2007, De Cooman and Miranda, 2008, Rébillé, 2009, Cozman, 2010, Chen et al., 2013, Terán, 2014, Hu et al., 2016, Peng, 2019, Zhang and Lan, 2020, Zhang et al., 2024]. We refer the reader to [Fröhlich and Williamson, 2024a] for an account that explicates relations of capacities, coherent lower (upper) previsions and coherent risk measures. We first describe the character of the mathematical results in these works, independent of interpretation. Then, since these authors primarily take a subjectivist perspective, it is also instructive to compare the interpretation of the mathematical objects, which may fulfill entirely different conceptual roles.

These results often take a certain form, which we informally describe to emphasize commonality over differences. Let $\underline{P}^\infty$ be a set functional which satisfies certain desirable properties, e.g. is a convex capacity and continuous, and $\underline{R}, \bar{R}$ be a corresponding pair of lower and upper generalized expectation¹⁸. For concreteness, let $\underline{P}$ a coherent lower probability, obtained by applying the coherent lower prevision $\underline{R}$ to indicator functions, i.e. $\underline{P}(A) = \underline{R}(\chi_A)$. Let $X_i \sim X$ be a sequence of “independent” (in an appropriate *imprecise* sense) gambles, then

$$\underline{P}^\infty \left(\left\{ \omega^\infty : \underline{R}(X) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i(\omega^\infty) \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i(\omega^\infty) \leq \bar{R}(X) \right\} \right) = 1.$$

Such a result is then called a *strong law*. Similarly, a *weak law* is roughly of the form

$$\lim_{n \rightarrow \infty} \underline{P}^n \left(\left\{ \omega^\infty : \underline{R}(X) - \varepsilon \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i(\omega^\infty) \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i(\omega^\infty) \leq \bar{R}(X) + \varepsilon \right\} \right) = 1.$$

For example, $X_i := X \circ \pi_i$. Hence these statements are highly similar to what we obtained in Section 5.4.1. As an example for an independence notion, consider Proposition 251, expressing a factorization on rectangles for $\underline{P}_{\mathcal{M}_{\rightarrow \mathcal{N}}^\infty}$; in the literature, a variety of independence notions has been considered.¹⁹ Note however that such statements are silent on divergence or convergence within the interval (recall the discussion after Proposition 256), so that this

¹⁸Often the Choquet integral is considered as the corresponding notion of generalized expectation.

¹⁹The reader may consult the works cited above on generalized laws of large numbers to learn about the concrete independence concepts that they employ.

kind of weak law is not in contradiction to the previously discussed impossibility of obtaining “weak law”-type statements *concerning apparent divergence* [Kumar and Fine, 1985] for stationary, appropriately continuous and dominated lower probabilities.

While some authors emphasize the neutrality of their mathematical results [De Cooman and Miranda, 2008], an often used interpretation in this literature is a subjectivist (generalized Bayesian) one, in the sense of Gilboa and Schmeidler [1989] and Walley [1991], which can be viewed as a generalization of the approach by de Finetti [1974/2017]. Here, imprecision is taken to represent *ambiguity*, sometimes called *Knightian uncertainty*. Ambiguity refers to a situation where a decision maker has no knowledge of objective probabilities and is unwilling to make precise, probabilistic assessments [Trautmann and Van De Kuilen, 2015, Bühren et al., 2023]. We illustrate this with an *Ellsberg example* [Ellsberg, 1961], following the distinction of Dow and Werlang [1993]. Assume a decision maker who is *uncertain* about the composition of urns, from which balls are drawn at random. It is known that the proportion of black balls is between 30% and 40%, and the rest are red balls. Not intending to commit further, the decision maker might assign the lower $\underline{P}(B) = 0.3$ and upper probability $\bar{P}(B) = 0.4$ for drawing a black ball, and similarly $\underline{P}(R) = 0.6$ and $\bar{P}(R) = 0.7$. That is, the decision maker entertains a non-additive belief, which can be represented as a set $\mathcal{M}$ of probability measures. Dow and Werlang [1993] distinguishes two long-run scenarios:

- U1. Balls are drawn independently (with replacement) from the same urn.
- U2. Balls are drawn independently from a sequence of urn, with possibly different compositions, but satisfying the above constraint on the proportions. There is no further evidence about the urn’s compositions, implying that the urns are viewed as *indistinguishable* by the decision maker [Epstein and Schneider, 2003].

An appropriate subjective model for U1 is an extension to $\mathcal{B}(\Omega^\infty)$ by a model proposed by Walley and Fine [1982]:

$$\underline{P}_{\mathcal{M},\text{iid}}^\infty(A^\infty) := \inf\{\text{IT}((m, m, m, \dots))(A^\infty) : m \in \mathcal{M}\}$$

That is, this model is the envelope of all *i.i.d.* product measures, but where any $m \in \mathcal{M}$ is taken into account as a possibility. In contrast, an appropriate subjective model for U2 is given by our $\underline{P}_{\mathcal{M}}^\infty$, which embodies a notion of independence, but leaves open the possibility that the compositions of the urns vary within $\mathcal{M}$. Note that the above discussed strong laws, due to their independence assumptions, can be viewed as referring to situation U2 (see the Ellsberg example in [Chen et al., 2013]) rather than U1. The indistinguishability is transformed into a formal stationarity of the subjective model, even though the decision maker acknowledges that the true data generating process is non-stationary. Note that $\underline{P}_{\mathcal{M},\text{iid}}^\infty$ does not factorize in the same way as $\underline{P}_{\mathcal{M}}$, thus it is not suitable for an imprecise independence concept (intuitively, according to $\underline{P}_{\mathcal{M},\text{iid}}^\infty$, if a probability measure $m \in \mathcal{M}$

operates on the first trial, it also has to operate on all other trials).

This way of re-interpreting $\underline{P}_{\mathcal{M}}$ from a subjectivist perspective makes clear the different conceptual role that it plays there. In situation U2, the actual data generating process is described by a non-stationary, locally precise data model of the form of Definition 242. But since the decision maker is uncertain about the data model, they represent their uncertainty using a set of such data models, yielding $\underline{P}_{\mathcal{M}}$. What is expressed is *belief about the data*. Hence the models considered throughout Section 5.4 are suitable for different scenarios when looking through frequentist or subjectivist glasses.

Yet another way of interpreting the $\underline{P}_{\mathcal{M} \rightarrow \mathcal{N}}$ family of data models is this. Since it is constructed as the envelope of all NSLP data models with aggregate (ir)regularity $\mathcal{N}$, the typicality $\underline{P}_{\mathcal{M} \rightarrow \mathcal{N}}(A^\infty) = 1$ expresses that some event A^∞ is typical *with respect to any* of the respective NSLP data models. This can then formally be used to justify that some estimation procedure (earning method) works for any such data model. In this sense, the typicality statement acts like a “ $\forall$ ” quantifier, instead of aiming to describe the actual data generating process. This viewpoint however is in contrast to our attitude towards data models, which we interpret as describing the data generating process, untouched from any epistemic uncertainty about the data generating process itself. Compare this to how the standard *i.i.d.* model is employed: in practice, we do not know the underlying probability P , but we use the *i.i.d.* data model with respect to a single, but arbitrary P as an abstract tool to justify our procedures, and we like to show that they work for any P .

5.5 APPLICATIONS AND ESTIMABILITY

We have shown that, from a frequentist perspective, imprecision manifests in two distinct, yet interrelated, ways: as aggregate irregularity and as local regularity. Which of these manifestations we are interested in depends on context, essentially the decision problem at hand. Relatedly, these manifestations are linked to distinct ways of estimation (and thus learning problems). Before we discuss estimation, however, we believe it is insightful to first go into possible applications for these data models. The various, distinct such application contexts then shed light on which kind of estimation questions need to be asked and how they might be answered.

5.5.1 APPLICATIONS

The non-stationary, locally precise (NSLP) model keeps the familiar probabilistic concept, but departs from the *identical* assumption. As such, we believe this model is an appropriate description of data-generating processes in many contexts. In light of the discussion in Section 5.4.2, the models $\underline{P}_{\mathcal{M}}$ and $\underline{P}_{\mathcal{M} \rightarrow \mathcal{N}}$ can describe subjective beliefs in such situations,

for a decision maker who is uncertain about the data-generating process; the stationarity of the model would then describe this uncertainty. Non-stationary application scenarios may include general data corruptions [Iacovissi et al., 2023], in particular dataset shift and learning with outliers, collaborative PAC learning, multi-source adaptation, fair machine learning and performativity. These problems share the common feature that they can be modelled by a set of distributions instead of a single one. Such multi-distribution settings have been receiving increasing attention in mainstream machine learning literature, see e.g. [Haghtalab et al., 2022]. For these problems, assume we work on a joint space $\Omega = \mathcal{X} \times \mathcal{Y}$, where $X : \Omega \rightarrow \mathcal{X}$ corresponds to the features of the learning problem, and $Y : \Omega \rightarrow \mathcal{Y}$ to the labels. We then consider a set of probability measures on Ω , and seek to obtain a predictor $f(X)$ for Y .

- A1.** In dataset shift, we want to model the possibility that the training and test distribution can differ: as machine learning systems are deployed in a dynamically unfolding environment, we can hardly expect stationarity. For safety reasons, it is therefore of great importance to robustify such systems against shifts in the underlying distribution.
- A2.** In the field of robust statistics [Huber, 1981], neighborhood models have been widely studied to robustify statistical methods to outliers. For example, as a frequentist version of the ε -contamination model, assume that the data-generating process is described by an NSLP data model parameterized by a sequence of probability measures $m^\infty = (P_0, \dots, P_0, \dots, P, \dots, P_0, \dots, P, \dots, P_0)$, where $P \in \Delta^k$ is arbitrary, and P_0 occurs with frequency at least $(1 - \varepsilon)$ in the sequence in the sense that $\liminf_{n \rightarrow \infty} \chi_{\{j: P_j = P_0\}}(i) \geq 1 - \varepsilon$. Intuitively, P_0 represents “clean” data and P represents outliers (“corrupted” data). A subjectivist decision maker who is uncertain about P might then use the set $\mathcal{M} = \{(1 - \varepsilon)P_0 + \varepsilon P : P \in \Delta^k\}$.
- A3.** In collaborative PAC learning [Blum et al., 2017] and federated learning [McMahan et al., 2017, Zhang et al., 2021a], the goal is to find a predictor that performs well on multiple distributions, each corresponding to the data distribution of one client.
- A4.** In fair machine learning, the literature describes multiple competing goals [Barocas et al., 2019]. One among them is to obtain predictors which have equal predictive performance over multiple *sensitive (ethically salient) subgroups*, determined for instance by gender, race etc. [Williamson and Menon, 2019]. Each subgroup can be understood as corresponding to one probability measure on $\mathcal{X} \times \mathcal{Y}$, when the sensitive feature itself is not contained in Ω .
- A5.** Under performativity [Perdomo et al., 2020], the machine learning model itself supposedly acts on the data-generating process, yielding dataset shift in response to model deployment and therefore non-stationarity.

The focus in these scenarios is often on the manifestation of local (ir)regularity (non-randomness,

hidden heterogeneity), in the sense that the goal is often to obtain robust performance over the set of probabilities $\mathcal{M}$. For instance, in the fair ML problem, hidden heterogeneity corresponds to the different ethically salient subgroups and the goal is not simply good performance on the whole population, but also on each of these subgroups. Sometimes, however, aggregate performance becomes the focus: for instance in outlier robustness, dataset shift or performativity we aim for overall good performance (on the whole aggregate) but acknowledge that a single, precise probability is insufficient as a data model.

The non-stationary, locally precise data model is couched in the language of precise probability. However, we view it also as inherently imprecise from the perspective of $\mathcal{M}$ as a summary of the model. Is there any virtue to this view, or in different words, why would we want to work with imprecise probabilities in such settings? By this, we mean for example introducing predictors which provide imprecise probabilistic forecasts as opposed to precise ones. We think there are legitimate reasons for doing so. [Walley and Fine \[1982\]](#) have offered the following intuition for the usefulness of imprecise modelling:

“Just as ‘randomness’ (chance) is introduced in additive probability models to account for poorly understood (‘accidental’) variation in outcomes, so ‘indeterminacy’ might be introduced in upper and lower probability models to account for poorly understood variations in chance behaviour.”

In this way, imprecision (here called “indeterminacy”) relates to a second-order modelling level, distinct from first-order randomness. In the standard *i.i.d.* picture, the goal would be to approximate the ideal Bayes predictor $f(X) = \mathbb{E}_P[Y|X]$ with respect to the precise underlying probability P . Intuitively, all information in X should be used by the predictor. However, the above problems make plausible that this is not always the case: in collaborative PAC learning, we know which distribution corresponds to which agent, but we might not want our predictor f to depend on this. Similarly, in a fairness context, it might be undesirable or indeed prohibited to include the sensitive feature in $\mathcal{X}$. In this way, the above problems share a *hidden heterogeneity* which cannot or should not be captured in $\mathcal{X}$ (which would reduce the problem to a single, precise distribution), but with robustness desiderata with respect to the hidden heterogeneity. We believe that the main use of imprecision for learning under an NSLP data model is to account for “poorly understood variations in chance behaviour” as claimed above, but still understood *to some extent*. That is, there must be some handle available to get to this variation (see the discussion on estimability below).

Besides its subjectivist interpretation, what are possible applications for $\underline{P}_{\mathcal{M}}$ and $\underline{P}_{\mathcal{M} \rightarrow \mathcal{N}}$ as data models? We have already noted that they might serve as more parsimonious stationary models for precise, non-stationary phenomena. For example, when there is no understanding *at all* of such variation in chance, in contrast to above. However their actual

character would prove to be relevant when we assume indeterminacy in chance *even under stationarity*. Applications under this philosophy remain more speculative; Fine [1988] have given flicker noise as an example for a process which is hypothesized to be stationary but exhibits aggregate irregularity. We believe the prime area of application might be *heavy-tailed phenomena*, which are recently receiving increasing attention [Taleb, 2007, Resnick, 2007, Nair et al., 2022] — already Fierens and Fine [2001], who focus on aggregate irregularity, suggested “[..] that these models will be appropriate when dealing with such long sequences as might arise from heavy-tailed waiting times.”. For heavy-tailed phenomena, an often-used model is that of a stationary Pareto distribution with parameter α , which produces relative frequencies that converge on much slower rates than for light-tailed distributions for $\alpha > 1$; for $\alpha < 1$, relative frequencies diverge. However, Fröhlich et al. [2024] have argued that divergence of relative frequencies in the infinite limit is merely an idealization of unstable relative frequencies (“apparent divergence” as in Section 5.4.1) for finite n , in the same way that convergence idealizes apparent stability; in this way, a data model which typically yields divergence may also be appropriate for $\alpha > 1$. We leave open the question of whether (or which) such heavy-tailed phenomena are better understood and modelled as non-stationary, locally precise or stationary, locally imprecise. In fact, recent results in imprecise game-theoretic probability support the claim that there is a fundamental equivalence relationship between non-stationary, locally precise and stationary, locally imprecise models [Persiau et al., 2021, De Cooman and De Bock, 2022]. From a practical point of view, a relevant area of applications would be machine learning and statistics with heavy-tailed outcomes. Such use cases arise for instance in (re)insurance uses cases, where claims (losses) can exhibit heavy-tails [Ibragimov et al., 2015, Peng and Qi, 2018, Punzo et al., 2018]. As another interesting research avenue, recent publications in the field of climate science [Lovejoy, 2015, Franzke et al., 2020] uncover aggregate irregularity in climate time series. We take this to suggest the development of a foundational account which links together and explicates the relationships between the concepts of power laws, heavy-tails, (non)-stationarity and aggregate irregularity.

Given the manifold nature of these application scenarios are, we shy away from constructing more concrete, constrained data models in the present paper. Rather, our aim is to offer a firm foundation on which such subsequent work can take place. We do however want to make some general remarks on estimation in such data models, which then needs to be tailored to the specific scenarios.

5.5.2 ESTIMATION OF AGGREGATE (IR)REGULARITY

Assume a fixed sequence $\omega^\infty = (\omega_1, \omega_2, \dots)$. Can we estimate its cluster points of relative frequencies, i.e. $\text{CP}(n \mapsto r_n^{\omega^\infty})$? In fact, there seems hardly a reason for attempting to

do this. This set could potentially be uncountable, which would make for a daunting task. Also observe that knowing $\underline{P}(\{\omega^j\})$, $\bar{P}(\{\omega^j\})$ for each elementary event ω^j is not sufficient for knowing whole set of cluster points, since a coherent lower (upper) probability is not in a one-to-one correspondence to a set of measures.²⁰

Further, in practice we do not observe elementary events, but rather the values that specific gambles take. In a decision problem, we are rather interested in estimating quantities such as

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ell(\omega_i) \quad \text{and} \quad \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ell(\omega_i)$$

for a loss gamble $\ell : \Omega \rightarrow \mathbb{R}$ (e.g. a loss function partially applied to an action). This is a much simpler problem, and indeed can be solved to some extend. Note that due to connectedness, we have

$$\text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n \ell(\omega_i) \right) = \left[\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ell(\omega_i), \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ell(\omega_i) \right]$$

Thus, for gamble averages, an estimator for the $\liminf$ would due to conjugacy suffice to estimate the whole interval of cluster points. First, we observe that this problem is in principle solvable for a fixed sequence.

Proposition 261. *For any ω^∞ , there exists functions g_n of the form $g_n(\ell(\omega_1), \dots, \frac{1}{n} \sum_{i=1}^n \ell(\omega_i))$ so that*

$$\lim_{n \rightarrow \infty} g_n(\cdot) = \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ell(\omega_i).$$

Proof $\mathbb{R}$ with the Euclidean topology is sequentially compact, and thus each cluster point of the sequence of gamble averages corresponds to the limit point of a convergent subsequence. Hence there exists a selection rule $S : \mathbb{N} \rightarrow \{0, 1\}$ which extracts this subsequence . ■

Not only does this “estimator” depend on the concrete sequence ω^∞ , it is completely useless, since the selection rule is unknown in practice. Using “all” selection rules, on the other hand, would not meaningfully work since in the finite case this would imply the “vacuous” estimate $\{\min_{1 \leq j \leq n} \frac{1}{j} \sum_{i=1}^j \ell(\omega_i), \max_{1 \leq j \leq n} \frac{1}{j} \sum_{i=1}^j \ell(\omega_i)\}$. We therefore desire an estimator which works on all sequences ω^∞ . [Walley and Fine \[1982\]](#) have proposed the

²⁰Recall that a coherent lower (upper) prevision is in a one-to-one correspondence to a closed convex set of (finitely additive) probabilities.

following estimator:²¹

$$\underline{\text{avg}}_{\ell,n} = \min \left\{ \frac{1}{j} \sum_{i=1}^j \ell(\omega_i) : \tilde{k}(n) \leq j \leq n \right\}, \quad (5.6)$$

where the function $\tilde{k} : \mathbb{N} \rightarrow \mathbb{N}$ is such that $\lim_{n \rightarrow \infty} \tilde{k}(n) = \infty$ and $\lim_{n \rightarrow \infty} \tilde{k}(n) \leq n$. For example, $\tilde{k}(n) = \sqrt{n}$. Similarly, $\overline{\text{avg}}_{\ell,n}$ is defined using max instead of min.

Proposition 262. *This estimator succeeds in the sense that:*

$$\liminf_{n \rightarrow \infty} \underline{\text{avg}}_{\ell,n} = \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ell(\omega_i) = \min \left\{ \mathbb{E}_p[\ell] : p \in \text{CP}(n \mapsto r_n^{\omega_\infty}) \right\}.$$

Proof The second equality of the statement is due to Proposition 239. The first equality follows straightforwardly by adapting the proof of Theorem 4.2 by [Walley and Fine \[1982\]](#). For completeness: for fixed $m \in \mathbb{N}$, $\inf \{ \underline{\text{avg}}_{\ell,n} : n \geq m \} = \inf \{ \frac{1}{n} \sum_{i=1}^n \ell(\omega_i) : n \geq \min_{n \geq m} \tilde{k}(n) \} =: I(m)$. Since $\tilde{k}(n) \rightarrow \infty$ we then get that $\lim_{m \rightarrow \infty} I(m) = \lim_{m \rightarrow \infty} \inf \{ \frac{1}{n} \sum_{i=1}^n \ell(\omega_i) : n \geq m \} = \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ell(\omega_i)$. ■

An analogous statement holds for the conjugate upper quantity. In this sense, the estimation succeeds asymptotically, but in a very weak sense only: we only get that the $\liminf$ of the estimator gives the desired estimand. This result can hardly be considered satisfying; also note that $\tilde{k}(n) = n$ in the above simply gives the familiar estimator $\underline{\text{avg}}_{\ell,n} = \frac{1}{n} \sum_{i=1}^n \ell(\omega_i)$, which has the same guarantee due to Proposition 262. The following suggests that *no choice* of $\tilde{k}(n)$ will work in general in the stronger sense that the limit of the estimator would coincide with the estimand.

Proposition 263. *For any choice of $\tilde{k}(n)$ which satisfies $\lim_{n \rightarrow \infty} \tilde{k}(n) = \infty$ and $\tilde{k}(n) \leq n$ and set of probability measures $\mathcal{M} \subseteq \Delta^k$, $|\mathcal{M}| > 1$, and any $\mathcal{N} \in \mathcal{C}(\mathcal{M})$, there exists a sequence of measures $m^\infty \in \mathcal{M}^\infty$ so that*

$$\text{CP} \left(n \mapsto \min \left(\frac{1}{j} \sum_{i=1}^j m_i : \tilde{k}(n) \leq j \leq n \right) \right) = \mathcal{N}. \quad (5.7)$$

Proof Note that it is clear that $\text{CP} \left(n \mapsto \min \left(\frac{1}{j} \sum_{i=1}^j m_i : \tilde{k}(n) \leq j \leq n \right) \right) \subseteq \text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right)$. We now want to show that the sequence m^∞ can be constructed so that the set on the left side is not a singleton, i.e. the limit does not exist. Proposition 274, shouldering the main work of the proof of Theorem 240, has shown the existence of some m^∞ with the property that $\text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right) = \mathcal{N}$ for any $\mathcal{N} \in \mathcal{C}(\mathcal{M})$. We now show that the proof of Propo-

²¹[Walley and Fine \[1982\]](#) proposed the estimator for relative frequencies of events, but the generalization to gamble averages is obvious. Originally, they demanded that $\lim_{n \rightarrow \infty} \tilde{k}(n)/n = 0$, but the proof of this statement in fact works if $\tilde{k}(n) \leq n$.

sition 274 can be slightly modified (or more precisely, can be concretely instantiated) so that Equation 5.7 holds in addition. Intuitively, the average $\frac{1}{n} \sum_{i=1}^n m_i$ has to change extremely slow relative to $\tilde{k}(n)$, so that $\{\frac{1}{j} \sum_{i=1}^j m_i : \tilde{k}(n) \leq j \leq n\}$ is approximately constant. The proof of Proposition 274 contains two relevant “gaps”, where it is only asserted that there exists a finite $\tilde{k} \in \mathbb{N}$ will make a condition hold, but no concrete $\tilde{k}$ is chosen:

1. Equation 5.27: $\exists k \in \mathbb{N} : r(m \oplus b^k) \in B_\kappa(\tilde{c}_2)$,
2. Equation 5.32: $\exists k \in \mathbb{N} : r(m \oplus b^k) \in B_\kappa(q)$.

The proof therefore needs to be instantiated with a concrete choice of k in these statements. Note that if either condition holds for some k , it also holds for all $l \geq k$ due to Lemma 276. It is guaranteed that there exists always a choice of k so that the proof yields $\text{CP}(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i) = \mathcal{N}$. But further, if we choose k large enough, we can in addition guarantee that Equation 5.7 will hold for the resulting sequence m^∞ : we need to find some k^* so that from $j = \tilde{k}(k^*)..k^*$, $r(m \oplus b^j) \in B_\kappa(\tilde{c}_2)$ (analogously for Equation 5.32), which is clearly possible. This implies that $\min\left(\frac{1}{j} \sum_{i=1}^j m_i : \tilde{k}(k^*) \leq j \leq k^*\right) \in B_\kappa(\tilde{c}_2)$. The logic of the proof of Proposition 274 then implies Equation 5.7. ■

We conclude by applying Proposition 270 that with this estimator, not even the cluster points of relative frequencies can be estimated, dispensing with the hope to estimate gamble averages. We now write $\underline{\text{avg}}_{\ell,n}^{\omega^\infty}$ to highlight the dependence on the sequence ω^∞ .

Corollary 264. *Let $\emptyset \subsetneq A \subsetneq \Omega$. For any choice of $\tilde{k}(n)$ with $\lim_{n \rightarrow \infty} \tilde{k}(n) = \infty$ and $\tilde{k}(n) \leq n$, there exists some sequence ω^∞ so that $\lim_{n \rightarrow \infty} \underline{\text{avg}}_{\chi_A,n}^{\omega^\infty}$ does not exist.*

Proof

Choose probability measures $\mathcal{M} := \{m_1, m_2\}$ so that $m_1(A) \neq m_2(A)$. Proposition 270 yields that the sequence $m^\infty \in \mathcal{M}^\infty$ gives $\text{IT}(m^\infty)(\{\omega^\infty : \lim_{n \rightarrow \infty} \underline{\text{avg}}_{\chi_A,n}^{\omega^\infty} \text{ does not exist}\}) = 1$, therefore one such sequence exists. ■

Until now we considered arbitrary sequences ω^∞ . To contextualize estimation in our data models, we emphasize that estimation is inextricably tied to typicality: only for typical sequences can we hope to recover the underlying parameterization; for example, even a familiar *i.i.d.* model can generate sequences with diverging relative frequencies, but only *atypically*. The previous result suggests that this estimator will not *typically* work in our data models. First, however, we look at a positive result. [Walley and Fine \[1982\]](#) have shown that, if $\tilde{k}(n)$ satisfies a specific growth behaviour, estimation works at least in the sense of asymptotic favorability. Recall that $A_1, A_2, ..$ are asymptotically favored under $\underline{P}$ when $\lim_{n \rightarrow \infty} \underline{P}((A_n)^c) / \underline{P}(A_n) = 0$, implying $\lim_{n \rightarrow \infty} \bar{P}(A_n) = 1$. The result is stated in terms of relative frequencies instead of gamble averages.

Proposition 265. [Walley and Fine, 1982, Theorem 5.3] Assume $\tilde{k}(n)/n \rightarrow 0$ and $\tilde{k}(n)/\log(n) \rightarrow \infty$. Let $\bar{P}(A) < 1$ or $\underline{P}(A) = 1$. Then

$$\forall \varepsilon > 0 : \{\omega^\infty : |\underline{\text{avg}}_{\chi_{A,n}}^{\omega^\infty} - \underline{P}_{\mathcal{M}}(A)| < \varepsilon\} \text{ asymptotically favored under } \underline{P}_{\mathcal{M}}^\infty.$$

Since they agree on $\mathcal{A}^\infty$, the same holds for $\underline{P}_{\mathcal{M}^\infty_{\rightarrow \mathcal{N}}}^\infty$. On the other hand, it is easily seen that if $0 < \varepsilon < \bar{P}_{\mathcal{M}}(A) - \underline{P}_{\mathcal{M}}(A)$ then $\underline{P}_{\mathcal{M}}^\infty(\{\omega^\infty : |\underline{\text{avg}}_{\chi_{A,n}}^{\omega^\infty} - \underline{P}_{\mathcal{M}}(A)| < \varepsilon\}) = 0$ and therefore a “weak law”-type statement for estimation fails. However, from Proposition 263 we obtain:

Proposition 266. Let $\emptyset \neq A \subsetneq \Omega$ an imprecise event, that is, $\bar{P}(A) > \underline{P}(A)$. For any choice of $\tilde{k}(n)$ with $\lim_{n \rightarrow \infty} \tilde{k}(n) = \infty$ and $\tilde{k}(n) \leq n$, the Walley-Fine estimator fails in the sense that

$$\underline{P}_{\mathcal{M}^\infty_{\rightarrow \mathcal{N}}} \left(\left\{ \omega^\infty : \lim_{n \rightarrow \infty} \underline{\text{avg}}_{\chi_{A,n}}^{\omega^\infty} \text{ exists } \right\} \right) = 0,$$

and similarly for $\underline{P}_{\mathcal{M}}^\infty$, since $\underline{P}_{\mathcal{M}}^\infty \leq \underline{P}_{\mathcal{M}^\infty_{\rightarrow \mathcal{N}}}^\infty$.

Proof Observe that the sequence $m^\infty = (m_1, m_2, \dots)$ of Proposition 263 lies in $\mathcal{M}_{\rightarrow \mathcal{N}}^\infty$. ■

This result seems unsettling, and motivates us to conjecture the following.

Conjecture 267. There is no “reasonable” (without access to an “oracle”) estimator so that

$$\underline{P}_{\mathcal{M}^\infty_{\rightarrow \mathcal{N}}}(\{\text{estimation of cluster points of relative frequencies succeeds}\}) = 1.$$

On the other hand, for *some* sequences of probabilities, the estimator *will* work typically. Kumar and Fine [1985, p. 13], with a similar motivation, have therefore reversed the logic and proposed to construct the model exactly in a way such that the estimator works in a “strong law” sense (by collecting all sequences of probabilities for which the estimator with a specified $\tilde{k}$ succeeds almost surely and forming the envelope). We believe this is indeed the way forward: in our view, $\underline{P}_{\mathcal{M}}$ and $\underline{P}_{\mathcal{M}^\infty_{\rightarrow \mathcal{N}}}$ only represent the starting point for developing data models which are finely tailored to specific assumptions about rates of convergence (divergence).

On a broader level, do we even *want* to estimate aggregate irregularity? Our position is that when considering estimation we cannot ignore the relevant decision-making context, as it determines which questions are meaningful to be asked. We believe that an explicit estimation of aggregate irregularity is only called for when the decision problem is with respect to aggregate criteria *and* context demands stationary (constant) forecasts. Recall the criticism by Fierens and Fine [2003] discussed in Section 5.3.1: if a skeptic can probe for local regularity, a forecast based on aggregate irregularity may perform poorly. Intuitively, the problem of the estimator Equation 5.6 is that relative frequencies are always computed from the very beginning; therefore the estimator adapts extremely slowly to

changing *local* relative frequencies in the data sequence, e.g. as computed by averages over moving windows, whereas a non-stationary forecast could quickly adapt to them. This shows that if evaluation is based on local (ir)regularity, estimation should also focus on this manifestation. On the other hand, when the evaluation of our forecasts is based on the whole aggregate *and* there is a demand for stationary (constant) forecasts, estimating aggregate (ir)regularity becomes meaningful. For instance, in the example in Section 5.5.1 of machine learning with heavy-tailed insurance claims, our evaluation is on the aggregate, and we might have no reason to suspect a time evolution and thus may want a stationary model. Finally, when evaluation is based on aggregate criteria, but there is no demand for stationary (constant) forecasts, a wider space opens up: an intriguing recent work by [Zhao and Ermon \[2021\]](#), which exhibits remarkable parallels to imprecise probabilities, has provided an online learning method that, based on non-stationary forecasts, satisfies a strong performance guarantee even under aggregate irregularity. In other words, their scheme passes a test in terms of aggregate-level performance even under aggregate irregularity without directly aiming to estimate aggregate irregularity; likewise for other online learning methods [[Vovk et al., 2005](#)].

To summarize, what is important is arguably not to estimate aggregate irregularity per se, but to have learning methods which have performance guarantees even under aggregate irregularity, however this is achieved.

5.5.3 ESTIMATION OF LOCAL (IR)REGULARITY

First, assume a non-stationary, locally precise model parameterized by the sequence $m^\infty = (m_1, m_2, \dots)$ so that $\mathcal{M} := \{m_1, m_2, \dots\}$. In problem contexts where the focus is on guarding against hidden heterogeneity, we may wish to estimate $\mathcal{M}$. For this, assume we have access to a fixed, finite set of selection rules $\mathcal{S}$. For a selection rule $S \in \mathcal{S}$, define its induced subsequence as m_S^∞ as the subsequence of m^∞ extracted by the condition that $S(i) = 1$. The theoretical mean associated to a selection rule is:

$$\mu_{S,n}(\{\omega^j\}) := \frac{\frac{1}{n} \sum_{i=1}^n m_i(\{\omega^j\}) S(i)}{\frac{1}{n} \sum_{i=1}^n S(i)}, \quad \omega^j \in \Omega,$$

(well-defined beginning with some $n_0 \in \mathbb{N}$ for which $S(n_0) = 1$) which extends to a probability measure $\mu_{S,n}$ on Δ^k in the obvious way by additivity. On the other hand, for a fixed sequence ω^∞ , the empirical relative frequencies induced by a selection rule are as follows:

$$r_{S,n}^{\omega^\infty}(\{\omega^j\}) := \frac{\frac{1}{n} \sum_{i=1}^n \chi_{\{\omega^j\}}(\omega_i) S(i)}{\frac{1}{n} \sum_{i=1}^n S(i)}, \quad \omega^j \in \Omega,$$

(again, well-defined for all $n \geq n_0 \in \mathbb{N}$) which similar extends to a probability measure on Δ^k . [Fierens et al. \[2009, Theorem 1\]](#) have shown the following.

Proposition 268. *Let $D(p, q) := \max_{\omega \in \Omega} |p(\{\omega\}) - q(\{\omega\})|$. Let $0 < m \leq n$. Then:*

$$\lambda \left\{ \omega^\infty := G(z) : \max_{S \in \mathcal{S}} \left(D(\mu_{S,n}, r_{S,n}^{\omega^\infty}) : \frac{1}{n} \sum_{i=1}^n S(i) \geq m \right) \geq \varepsilon \right\} \leq 2k|\mathcal{S}| \exp(-\varepsilon^2 m^2 / 2n)$$

Intuitively, the familiar weak law holds for each selected subsequence, and we can therefore avoid extracting “arbitrary patterns”, but we must consider only those selection rules which have selected sufficiently many indices until n . For example, if m^∞ truly was an *i.i.d.* model, i.e. $m^\infty = (m_0, m_0, \dots)$, the above statement reduces to a weak law, asserting that relative frequencies converge to m_0 along all selection rules. On the other hand, if there is underlying imprecision in m^∞ , manifested as non-stationarity, then selection rules might converge to different limits, or indeed they might not converge at all; Theorem 240 implies that $\text{CP}(n \mapsto \mu_{S,n}) \in \mathcal{C}(\mathcal{M})$.

A set of selection rules $\mathcal{S}$ therefore provides a natural estimate for $\mathcal{M}$:

$$\hat{M}_n^{\omega^\infty} := \bigcup_{S \in \mathcal{S}} r_{S,n}^{\omega^\infty}$$

For the purpose of computing lower (upper) expectations, the closed convex hull might be used equivalently. The implication is that for large enough n , we do not (approximately) over-estimate the set $\mathcal{M}$, that is, our estimate is (approximately) contained in $\mathcal{M}$. On the other hand, we have no guarantee that we will recover *all of* $\mathcal{M}$. Access to selection rules of the form of Proposition 244 would at least guarantee that we asymptotically recover all $m \in \text{CP}(i \mapsto m_i)$. But assuming access to such a set seems highly impractical. Motivated by the goal of recovering the whole $\mathcal{M}$, [Fierens et al. \[2009\]](#) impose computability requirements, to argue that they can recover $\mathcal{M}$ “universally”.

The situation can be clarified by taking a step back and reflecting about how such data models may actually be used in practice. In practice, we do not have access to an oracle that gives us the “true” data model (we put aside the question of the ontological nature of such data models) but rather we start with an approach to data handling/processing. Concretely, we may construct a set of selection rules, guided by the application context, i.e. starting with assumptions about how the hidden heterogeneity might be structured. Then we invoke a data model to argue, or justify, the sensibility of our approach, by showing that it has desirable properties under this data model. This line of reasoning sheds light also on the familiar *i.i.d.* model: it is not because real world data is drawn *i.i.d.* that the *i.i.d.* model is so popular; rather, because it is customary to evaluate based on average performance and ignore the heterogeneity of subpopulations that the *i.i.d.* model serves

to justify such an approach. For a more general example, if in some context the relevant properties of our decision problems depend only on the aggregate (ir)regularity, there is no difference for whether we would use $\underline{P}_{\mathcal{M} \rightarrow \mathcal{N}}$ as a data model or an NSLP data model with aggregate (ir)regularity $\mathcal{N}$. The choice of data model (crucially, its typicality notion) should reflect our data processing, in particular with respect to the relevant characteristics of the decision problem at hand.

On the other hand, there are of course limits to the “self-fulfilling” prophecy of assuming a data model. The data we have actually at hand might be *atypical* with respect to our assumed data model. For example, we may have assumed an *i.i.d.* model, but investigating with selection rules might uncover local regularity in the data, atypical for an *i.i.d.* model.²² This should motivate us to then reject the *i.i.d.* model and to look for a data model which is consistent with the data at hand — this is also the logic of a frequentist hypothesis test.

For $\underline{P}_{\mathcal{M}}$ we note that if $|\mathcal{M}| > 1$, the model is vacuous about the existence of local (ir)regularity, since the corresponding set of measures also contains *i.i.d.* product measures. It seems hardly feasible to obtain interesting statements of the local (ir)regularity associated with $\underline{P}_{\mathcal{M} \rightarrow \mathcal{N}}$. If $\mathcal{N}$ is not a singleton, no *i.i.d.* product measure can be in the corresponding set of measures, implying that each such product measure has *some* local regularity. However, since the set of measures is only constrained by their divergent behaviour, it appears impossible that they “agree” on the structure of the local regularities — otherwise the model could not be stationary.

5.6 DISCUSSION

Fine [1970] had already noted about aggregate regularity:

Apparent convergence of the relative frequency occurs because of, and not in spite of, the high irregularity (randomness) of the data sequence. [emphasis in original]

Our main theorem implies that we only partially agree with this finding: indeed, local regularity (non-randomness) is a prerequisite for aggregate irregularity (divergence). But the existence of local regularity as a hidden heterogeneity need not imply aggregate irregularity. While we believe we have thus fully characterized the relation between these two manifestations of imprecision, and explicated how they manifest under different data models, many new questions have surfaced in this process. One line of future research would develop more constrained data models with imprecision, tailored also to specific application contexts and provide methods for estimation. For instance, can we craft a notion of

²²Here we are using typicality in an approximate, non-asymptotic sense. For small, finite n , even an *i.i.d.* model could yield local regularity, but we expect it to disappear for large enough n .

typicality directly in terms of rate of convergence (divergence)? In light of Section 5.4.1, such models would necessarily be non-stationary, if they are to be coherent.

Often, imprecision is taken to be synonymous with subjective ambiguity, where the idea is that it reflects merely epistemic uncertainty. We have departed from this picture by introducing data models which cannot be described by a single probability, irrespective of our epistemic considerations. This suggests that there is a role for imprecise forecasts in this picture, where a predictor would be of the type $f(X = x) \in 2^{\Delta^k}$, outputting for each $X = x$ a set of probabilities. Under an imprecise data model like NSLP or SLI parameterized by $\mathcal{M}$, it is plausible that forecasting the set $\mathcal{M}$ itself is a “reasonable” stationary forecast; but how to evaluate IP forecasts? A rigorous development, in our opinion, would introduce and study generalizations of proper scoring rules and calibration to the imprecise case. Impossibility results in the literature [Seidenfeld et al., 2012, Mayo-Wilson and Wheeler, 2016, Schoenfield, 2017], directed against IP scoring rules, *prima facie* threaten this endeavour. We do not share the pessimism — see also [Konek, 2019] for a discussion — and highlight that these results are concerned with *strict propriety*. Intuitively, this is the demand that there cannot, to any imprecise forecast, exist a precise forecast which is at least as accurate as measured by the scoring rule. In line with Mayo-Wilson and Wheeler [2016] we think that the strictness is an overly stringent desideratum. Preliminary findings which we have obtained indeed indicate the feasibility of developing weakly proper IP scoring rules and IP calibration. In our view, this requires relativizing these concepts to the assumed data model; in previous works on IP scoring rules which follow a subjectivist approach, data models have not appeared. The present paper provides the foundation for proceeding further in this direction.

Another follow-up question is how to represent epistemic uncertainty, when we use imprecision to represent aleatoric indeterminacy. In an estimation process like that of Section 5.5.3 the estimate, which is a set of probabilities, is itself subject to epistemic uncertainty, which calls for a principled, disentangled representation of the distinct sources of uncertainty.

Finally, in the present paper we have not rigorously linked and compared these frequentist data models to recent work on imprecise game-theoretic probability [Persiau et al., 2021, 2022, De Cooman and De Bock, 2022]. These authors also take inspiration from the works of Walley and Fine [1982] and subsequent papers by Fine and collaborators, but operate in the language of game-theoretic probability, rendering a translation of results non-trivial.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy EXC number 2064/1 Project number

390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christian Fröhlich. Many thanks to Rabanus Derr and Marvin Pförtner for helpful discussions.

5.7 APPENDIX

5.7.1 PROOF OF PROPOSITION 247

Proof A typicality space in the sense of Galvan [2006] is a triplet $(\Lambda, \mathcal{A}, d)$, where Λ is a “sample space” (not further defined), $\mathcal{A} \subseteq 2^\Lambda$ is a set system closed under complementation and which contains the emptyset, and d is a *typicality distance*. We show that $d(A, B) := \bar{P}(A \triangle B)$, $A, B \in \mathcal{A}$, satisfies the axioms of a typicality distance as in [Galvan, 2006], which are as follows:

T1. $A \subseteq A' \Rightarrow d(A, \emptyset) \leq d(A', \emptyset)$.

T2. $d(\Lambda, \emptyset) = 1$.

T3. $d(A, \Lambda) = d(A^c, \emptyset)$.

T4. If $A_1 \cap A_2 \in \mathcal{A}$, then $d(A_1 \cap A_2, B) \leq d(A_1, B) + d(A_2, B)$.

T5. If $A_1 \cap A_2, B_1 \cap B_2 \in \mathcal{A}$, then $d(A_1 \cap A_2, B_1 \cap B_2) \leq d(A_1, B_1) + d(A_2, B_2)$.

We note that if P is a probability measure, then $\tilde{d}(A, B) := P(A \triangle B)$ satisfies these axioms. Since $\bar{P}$ is sub-additive in the sense that $\bar{P}(A \cup B) \leq \bar{P}(A) + \bar{P}(B)$, we find that the relevant properties are preserved.

T1 follows from monotonicity of $\bar{P}$; note that $d(A, \emptyset) = \bar{P}(A)$. T2 follows since $\bar{P}(\Lambda) = 1$. For T3, note that $d(A, \Lambda) = \bar{P}(A \triangle \Lambda) = \bar{P}(A^c)$ and $d(A^c, \emptyset) = \bar{P}(A^c)$. To see that T4 and T5 hold, it is easiest so draw Venn diagrams. For completeness we provide the computations.

For T4, observe that $d(A_1 \cap A_2, B) = \bar{P}(I)$ with

$$I := (A_1 \cap A_2) \triangle B = (A_1 \cap A_2 \cap B^c) \cup (A_1^c \cup A_2^c) \cap B \quad (5.8)$$

$$\subseteq (A_1^c \cap B) \cup (A_1 \cap B^c) \cup (A_2^c \cap B) \cup (A_2 \cap B^c) =: II. \quad (5.9)$$

From monotonicity of $\bar{P}$ we get that $\bar{P}(I) \leq \bar{P}(II)$, and from subadditivity

$$\bar{P}(II) \leq \bar{P}((A_1^c \cap B) \cup (A_1 \cap B^c)) + \bar{P}((A_2^c \cap B) \cup (A_2 \cap B^c)) = d(A_1, B) + d(A_2, B).$$

For T5, $d(A_1 \cap A_2, B_1 \cap B_2) = \bar{P}(I)$, where $I := (A_1 \cap A_2 \cap (B_1 \cap B_2)^c) \cup ((A_1 \cap A_2)^c \cap (B_1 \cap B_2))$. Now,

$$I \subseteq (A_1 \cap B_1^c) \cup (A_1^c \cap B_1) \cup (A_2 \cap B_2^c) \cup (A_2^c \cap B_2) =: II, \quad (5.10)$$

because

$$(A_1^c \cup A_2^c) \cap B_1 \cap B_2 = (A_1^c \cap B_1 \cap B_2) \cup (A_2^c \cap B_1 \cap B_2) \subseteq (A_1^c \cap B_1) \cup (A_2^c \cap B_2)$$

and

$$A_1 \cap A_2 \cap (B_1^c \cup B_2^c) = (A_1 \cap A_2 \cap B_1^c) \cup (A_1 \cap A_2 \cap B_2^c) \subseteq (A_1 \cap B_1^c) \cup (A_2 \cap B_2^c).$$

From $I \subseteq II$, we conclude by monotonicity and subadditivity that

$$\bar{P}(I) \leq \bar{P}((A_1 \cap B_1^c) \cup (A_1^c \cap B_1)) + \bar{P}((A_2 \cap B_2^c) \cup (A_2^c \cap B_2)) = d(A_1, B_1) + d(A_2, B_2).$$

Based on a typicality distance, [Galvan \[2006\]](#) defines a *relative typicality measure* as

$$T_r(A|B) := \inf_{A' \subseteq A, B' \supseteq B} \frac{d(A', B')}{d(B, \emptyset)}, \quad A, B \in \mathcal{A},$$

and the *absolute typicality measure* as

$$T_a(A) := T_r(A|\Lambda) = d(A^c, \emptyset), \quad A \in \mathcal{A}.$$

In our case, $T_a(A) = \bar{P}(A^c \triangle \emptyset) = \bar{P}(A^c) = 1 - \underline{P}(A)$. ■

5.7.2 PROOF OF THE MAIN THEOREM 240

Throughout this section, we work on the measure space $([0, 1], \mathcal{B}([0, 1]), \lambda)$. We here prove our main theorem, restated for convenience:

Theorem 269. *Let $\mathcal{Z} = [0, 1]$ and $\emptyset \neq \mathcal{M} \subseteq \Delta^k$. Then there exists a sequence of probability measures $m_1, m_2, \dots \in \mathcal{M}$ with independent product*

$$\lambda \{z \in \mathcal{Z} : (\omega_1, \omega_2, \dots, \omega_n) \subset G(z)\} := \prod_{i=1}^n m_i(\{\omega_i\}), \quad \forall n \in \mathbb{N}, \quad (5.11)$$

so that it holds

$$\lambda \left\{ z \in \mathcal{Z} : \text{CP} \left(n \mapsto r_n^{G(z)} \right) = \text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right) = \mathcal{N} \right\} = 1 \quad (5.12)$$

if and only if $\emptyset \neq \mathcal{N} \subseteq \bar{\text{co}}(\mathcal{M})$ is a closed connected subset of the closed convex hull of $\mathcal{M}$.

We can in addition demand that any $m \in \mathcal{M}$ is a cluster point of the sequence of measures m_i (with respect to the Euclidean metric).

We build this up from some helper propositions.

Proposition 270. Let $p^\infty = (p_1, p_2, \dots)$ be a sequence of probability measures in Δ^k , with independent product

$$\lambda \{z \in \mathcal{Z} : (\omega_1, \omega_2, \dots, \omega_n) \subset G(z)\} := \prod_{i=1}^n p_i(\{\omega_i\}), \quad \forall n \in \mathbb{N}.$$

Then

$$\lambda \left\{ z \in \mathcal{Z} : \text{CP} \left(n \mapsto r_n^{G(z)} \right) = \text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n p_i \right) \right\} = 1.$$

In words, almost surely the cluster points of relative frequencies will coincide exactly with the cluster points of the sequence of measures (with respect to the Euclidean topology on Δ^k).

This follows by applying a variant of the strong law of large numbers, specialized to our context:

Theorem 271 ([Feller, 1991, Section 10.7]). Let $X_1, X_2, \dots$ be independent but not necessarily identically distributed random variables. Define $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$. If they satisfy

$$\text{Var}(X_k) < \infty \quad \forall k \in \mathbb{N} \quad \text{and} \quad \sum_{k=1}^{\infty} \frac{\text{Var}(X_k)}{k^2} < \infty,$$

then it holds

$$\lambda \left\{ \lim_{n \rightarrow \infty} \bar{X}_n - \mathbb{E}[\bar{X}_n] = 0 \right\} = 1.$$

In words, the sample average converges almost surely to its expectation.

Remark 272. The statement can be extended to vector-valued random variables, i.e. $X_i \in \mathbb{R}^d$, when we check the conditions for each component. Then we get d -many almost sure events (almost sure convergence in each dimension), and intersecting finitely many almost sure events yields an almost sure event.

We now prove Proposition 270.

Proof Under our independent product, the projections $W_i : [0, 1) \rightarrow \Omega$ are measurable. Consider the sequence of $\{0, 1\}^k$ -valued random variables $\chi_1, \chi_2, \dots$ where

$$\chi_i(z) := \left(\chi_{\{\omega^1\}}(W_i(z)), \dots, \chi_{\{\omega^k\}}(W_i(z)) \right)^\top.$$

This sequence satisfies the condition of Theorem 271:

First, we check that each component $\chi_i^j(z)$, $j = 1..k$, of $\chi_i(z)$ has finite variance; but its variance is simply $\text{Var}(\chi_i^j) = p_i(\omega^j)(1 - p_i(\omega^j)) \leq 0.25 < \infty$ (Bernoulli variance). But then

it is easy to see that for any component j

$$\sum_{i=1}^{\infty} \frac{\text{Var}(\chi_i^j)}{i^2} < \infty.$$

Applying Theorem 271 therefore gives

$$\lambda \left\{ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \chi_i(z) - \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \chi_i(z) \right] = 0 \right\} = 1 \quad (5.13)$$

$$\Leftrightarrow \lambda \left\{ \lim_{n \rightarrow \infty} r_n^{G(z)} - \mathbb{E} \left[r_n^{G(z)} \right] = 0 \right\} = 1 \quad (5.14)$$

$$\Leftrightarrow \lambda \left\{ \lim_{n \rightarrow \infty} r_n^{G(z)} - \frac{1}{n} \sum_{i=1}^n p_i = 0 \right\} = 1. \quad (5.15)$$

It remains to observe that the last condition implies

$$\lambda \left\{ z \in \mathcal{Z} : \text{CP} \left(n \mapsto r_n^{G(z)} \right) = \text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n p_i \right) \right\} = 1.$$

But it is clear that if for two $\mathbb{R}^k$ valued sequences $a^\infty = (a_1, ..), b^\infty = (b_1, ..)$, we have $\lim_{n \rightarrow \infty} a_n - b_n = 0$, then $\text{CP}(a^\infty) = \text{CP}(b^\infty)$. We briefly demonstrate this. Assume c is a cluster point of a^∞ , which is equivalent to

$$\forall \varepsilon > 0 : \forall n_0 \in \mathbb{N} : \exists n \geq n_0 : a_n \in B_\varepsilon(c).$$

We want to show that then c is also a cluster point of b^∞ , equivalent to $\forall \varepsilon' > 0 : \forall n'_0 \in \mathbb{N} : \exists n \geq n'_0 : b_n \in B_{\varepsilon'}(c)$. But since $\lim_{n \rightarrow \infty} a_n - b_n = 0$, we can find $\varepsilon + \kappa < \varepsilon'$ so that $\exists n''_0 \in \mathbb{N} : \forall n \geq n''_0 : d(a_n, b_n) < \kappa$. Setting $n_0 := n'_0 := n''_0$ then gives us some $n \geq n_0$ for which $a_n \in B_\varepsilon(c)$ and $d(a_n, b_n) < \kappa$, implying that $d(c, b_n) < \varepsilon + \kappa < \varepsilon'$. ■

Thus, in light of Proposition 270, in order to prove Theorem 240, it remains to show the following.

Proposition 273. *Let $\mathcal{O} \neq \mathcal{M} \subseteq \Delta^k$. Then there exists a sequence of measures $m^\infty = (m_1, m_2, ..)$, $m_i \in \mathcal{M}$, so that*

$$\text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right) = \mathcal{N},$$

if and only if $\mathcal{O} \neq \mathcal{N} \subseteq \bar{\text{co}}(\mathcal{M})$ is a closed connected subset of the closed convex hull of $\mathcal{M}$. We can in addition demand that any point $m \in \mathcal{M}$ is a cluster point of the sequence of measures m_i (with respect to the Euclidean metric).

For the forward direction, the key proposition of interest is the following.

Proposition 274. *Let $\emptyset \neq \mathcal{M} \subseteq \Delta^k$ and $\emptyset \neq \mathcal{N} \subseteq \bar{\text{co}}(\mathcal{M})$ a closed connected subset of the closed convex hull of $\mathcal{M}$. Then there exists a sequence of measures $m^\infty = (m_1, m_2, \dots)$, $m_i \in \mathcal{M}$, so that*

$$\text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right) = \mathcal{N}.$$

In addition, it can be guaranteed that any point $n \in \mathcal{N}$ is a cluster point of the sequence of measures m_i .

We defer its proof to later. The forward direction is concluded by showing in addition:

Proposition 275. *Let $\emptyset \neq \mathcal{M} \subseteq \Delta^k$ and $m^\infty = (m_1, m_2, \dots)$ a sequence of measures from $\mathcal{M}$. Then there exists a sequence of measure $\tilde{m}^\infty = (\tilde{m}_1, \tilde{m}_2, \dots)$ from $\mathcal{M}$ so that any $m \in \mathcal{M}$ is a cluster point of $\tilde{m}^\infty$ and*

$$\text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right) = \text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n \tilde{m}_i \right).$$

Proof Note that $\bar{\text{co}}(\mathcal{M})$ itself is closed and connected, hence satisfies the assumptions of Proposition 274. Thus we can obtain a sequence $m^{\infty,2}$ for which every measure in $\mathcal{M}$ is a cluster point (indeed any measure from $\bar{\text{co}}(\mathcal{M})$). Now construct a selection rule $S : \mathbb{N} \rightarrow \{0, 1\}$, with “rate going to zero”, meaning $\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n S(i)}{n} = 0$. We use this selection rule to “interleave” the sequences $m^{\infty,2}$ and m^∞ in the following way. The new sequence of measures is

$$\tilde{m}_i := \begin{cases} m_j^2 : j := \sum_{j=1}^i S(j) & \text{if } S(i) = 1 \\ m_j : j := i - \sum_{j=1}^i S(j) & \text{if } S(i) = 0. \end{cases}$$

From this we obtain

$$\frac{1}{n} \sum_{i=1}^n \tilde{m}_i = \frac{1}{n} \sum_{i=1}^{\sum_{j=1}^n S(j)} m_i^2 + \frac{1}{n} \sum_{i=1}^{n - \sum_{j=1}^n S(j)} m_j.$$

But since $\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n S(i)}{n} = 0$ and all measures in the Δ^k simplex are bounded, the first term tends to zero as $n \rightarrow \infty$. We hence find that $\text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right) = \text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n \tilde{m}_i \right)$. ■

PROOF OF PROPOSITION 274

We now prove Proposition 274, for which we require some helpful notation and helper lemmas. Our proof takes some inspiration from [Fröhlich et al., 2024, Appendix B].

We define some notation first. Recall that for a set A and some $n \in \mathbb{N}$, we have defined $A^n := \{a = (a_1, \dots, a_n) : a_i \in A\}$. If $a = (a_1, \dots, a_u) \in A^u$ and $b = (b_1, \dots, b_v) \in A^v$, we write

$a \oplus b := (a_1, \dots, a_u, b_1, \dots, b_v)$. For $a \in A^u$ and $k \in \mathbb{N}$, we write $a^k := a \oplus a \oplus \dots \oplus a \in A^{k \cdot u}$. Our notation uses different letters for points in the simplex p, q (like relative frequencies), infinite sequences of probability measures a^∞, b^∞ and finite sequences of probability measures a, b (of which the average is in the simplex). For $a = (a_1, \dots, a_u) \in (\Delta^k)^u$, we denote its average as $r(a) := \frac{1}{u} \sum_{i=1}^u a_i$, $r(a) \in \Delta^k$. This average corresponds to the expectation of an independent product of the a_i . We denote the Euclidean norm by $\|\cdot\|$. As a useful fact, any $p \in \Delta^k$ has $\|p\| \leq 1$. Thus $\|p - q\| \leq 2$ if p, q are in the simplex. For a point $p \in \Delta^k$ and a line segment $[q, r] \subset \Delta^k$ we say that p is at most δ away from $[q, r]$ iff $\exists x \in [q, r] : d(p, x) \leq \delta$.

Lemma 276 ([Fröhlich et al., 2024, Lemma B.9]). *Let $a = (a_1, \dots, a_u) \in (\Delta^k)^u$, $b = (b_1, \dots, b_v) \in (\Delta^k)^v$. Then*

$$r(a \oplus b) = \frac{u}{u+v} r(a) + \frac{v}{u+v} r(b) = (1-\alpha)r(a) + \alpha r(b),$$

where $\alpha := v/(u+v)$ and $1-\alpha = u/(u+v)$.

Proof Just compute

$$r(a \oplus b) = \frac{1}{u+v} \left(\sum_{i=1}^u a_i + \sum_{i=1}^v b_i \right) \quad (5.16)$$

$$= \frac{u}{u+v} \frac{1}{u} \sum_{i=1}^u a_i + \frac{v}{u+v} \frac{1}{v} \sum_{i=1}^v b_i \quad (5.17)$$

$$= \frac{u}{u+v} r(a) + \frac{v}{u+v} r(b). \quad (5.18)$$

■

The next lemma will formalize the intuition: if $a \in (\Delta^k)^u$ is a “long” finite sequence of measures and $b \in (\Delta^k)^v$ is relatively “short”, then the relative frequencies $r(a \oplus b)$ will be close to $r(a)$.

Lemma 277 (similar to [Fröhlich et al., 2024, Lemma B.17]). *Let $a \in (\Delta^k)^u$ and $b = (b_1, \dots, b_v) \in (\Delta^k)^v$. Then*

$$d(r(a), r(a \oplus (b_1, \dots, b_i))) \leq \frac{2v}{u+v}, \quad 1 \leq i \leq v.$$

Proof Consider first the extreme case $i = v$. We use the previous Lemma 276. Then

$$d(r(a), r(a \oplus b)) = \|r(a) - r(a \oplus b)\| \quad (5.19)$$

$$= \|(1 - u/(u+v)r(a) - v/(u+v)r(b)\| \quad (5.20)$$

$$= v/(u+v) \|r(a) - r(b)\| \quad (5.21)$$

$$\leq 2v/(u+v), \quad (5.22)$$

since $\|r(a) - r(b)\| \leq 2$ in the simplex, as any $\|p\| \leq 1$ for $p \in \Delta^k$. For $i < v$ we have that $2i/(u+i) \leq 2v(u+v)$ and thus the bound also holds. $\blacksquare$

We will later want to approximate measures $q \in \text{co}(\mathcal{M})$ by finite sequences of measures from $\mathcal{M}$. That this is possible is formalized in the following; the approximation quality of course depends on the length of the finite sequence.

Lemma 278 (Approximation lemma). *Let $\emptyset \neq \mathcal{M} \subseteq \Delta^k$ and $q \in \text{co}(\mathcal{M})$. Let $v \in \mathbb{N}$ be given. Then there exists some $b \in \mathcal{M}^v$ such that $d(q, r_b) \leq 4(k+1)/v$.*

Proof First note that due to Carathéodory's theorem, we need at most $k+1$ points from $\mathcal{M}$ to express $q \in \text{co}(\mathcal{M})$, that is, we can write $q = \sum_{i=1}^w \lambda_i m_i$ where $w \leq k+1$ and $m_i \in \mathcal{M}$, $\sum_{i=1}^w \lambda_i = 1$. Write

$$p := \sum_{i=1}^w \frac{\lfloor v\lambda_i \rfloor}{v} m_i + \xi p_0, \quad \xi := 1 - \sum_{i=1}^w \frac{\lfloor v\lambda_i \rfloor}{v},$$

for some $p_0 \in \mathcal{M}$. Note that $p \in \Delta^k$ and it is expressed as a convex combination of measures from $\mathcal{M}$, where all coefficients are rational with common denominator v . Implicitly here we constructed a sequence of measures

$$b := \underbrace{m_1 \oplus \dots \oplus m_1}_{\lfloor v\lambda_1 \rfloor \text{ many times}} \oplus \dots \oplus \underbrace{m_w \oplus \dots \oplus m_w}_{\lfloor v\lambda_w \rfloor \text{ many times}} \oplus \underbrace{p_0}_{\xi v \text{ many times}}.$$

Note that b is a finite sequence of measures from $\mathcal{M}$ of length v and $r_b \in \Delta^k$. It holds that $p = r_b$.

First, note that if $v \in \mathbb{N}$, then $0 \leq \lambda_i - \lfloor v\lambda_i \rfloor / v \leq 1/v$ since $0 \leq v\lambda_i - \lfloor v\lambda_i \rfloor \leq 1$. We then find that

$$d\left(\sum_{i=1}^w \frac{\lfloor v\lambda_i \rfloor}{v} m_i, \sum_{i=1}^w \lambda_i m_i\right) = \left\| \sum_{i=1}^w \left(\lambda_i - \frac{\lfloor v\lambda_i \rfloor}{v}\right) m_i \right\| \quad (5.23)$$

$$\leq \sum_{i=1}^w \left(\lambda_i - \frac{\lfloor v\lambda_i \rfloor}{v}\right) \|m_i\| \quad (5.24)$$

$$\leq \frac{2w}{v}. \quad (5.25)$$

Furthermore, $\xi = 1 - \sum_{i=1}^w \frac{\lfloor v\lambda_i \rfloor}{v} \leq w/v$, which holds because we have seen that $\lambda_i - \frac{\lfloor v\lambda_i \rfloor}{v} \leq 1/v$ and $\sum_{i=1}^w \lambda_i = 1$. Therefore

$$d\left(\sum_{i=1}^w \frac{\lfloor v\lambda_i \rfloor}{v} m_i, \sum_{i=1}^w \frac{\lfloor v\lambda_i \rfloor}{v} m_i + \xi p_0\right) \leq \xi \|p_0\| \leq 2\xi \leq \frac{2w}{v}. \quad (5.26)$$

Taking this together we obtain by the triangle inequality

$$d\left(q, \sum_{i=1}^w \frac{\lfloor v\lambda_i \rfloor}{v} m_i + \xi p_0\right) \leq 4(k+1)/v.$$

■

The next lemma formalizes the following: if we have already accumulated a finite sequence of measures $a \in (\Delta^k)^u$, so that our current average $r(a)$ is at most δ far away from some measure p for sufficiently large δ , then we can find some new measures $(d_1, \dots, d_v)$ from $\mathcal{M}$ so that the new average $r(a \oplus (d_1, \dots, d_v))$ is again at most δ far away from p . The smaller δ is, the more stringent the condition, so v needs to be smaller. The purpose of this lemma is to “approximately stay where we are” but grow the length of the finite sequence of measures. This in turn then allows for higher quality approximations as in the previous lemma.

Lemma 279 (The “stay and grow” lemma). *Let $p \in \text{co}(\mathcal{M})$, $q \in B_\delta(p)$, $q = r_a$, $a \in (\Delta^k)^u$, and assume $\delta > 0$ is so that it satisfies*

$$\exists v \in \mathbb{N} : \frac{2v}{u+v} \leq \delta \text{ and } 4(k+1)/v \leq \delta.$$

Then $\exists b_1, \dots, b_v \in \mathcal{M}$ so that $r(a \oplus (b_1, \dots, b_v)) \in B_\delta(p)$, as well as

$$d(r(a \oplus (b_1, \dots, b_i)), p) < 2\delta \quad \forall 0 \leq i \leq v.$$

Proof Take the above $v \in \mathbb{N}$ and use the approximation lemma 278 to obtain some $b \in \mathcal{M}^v$ -approximation of p , call $\tilde{p} := r(b)$, so that $d(p, \tilde{p}) \leq \delta$. Then the half-open segment $[q, \tilde{p})$ lies fully in $B_\delta(p)$, since q itself is in $B_\delta(p)$. We can then construct $a \oplus (b_1, \dots, b_v)$ and it is guaranteed that $r(a \oplus (b_1, \dots, b_v)) \in [q, \tilde{p}) \subset B_\delta(p)$. And since $2v/(u+v) \leq \delta$ by assumption, we have

$$d(r(a \oplus (b_1, \dots, b_i)), p) \leq d(r(a \oplus (b_1, \dots, b_i)), q) + d(q, p) < 2v/(u+v) + \delta \leq 2\delta$$

due to Lemma 277. ■

Lemma 280. *Let $\mathcal{N} \subseteq \Delta^k$. Then there exists an ε -cover of $\mathcal{N}$ using finitely many points in $\mathcal{N}$. Formally, $\forall \varepsilon > 0 : \exists n_1, \dots, n_o \in \mathcal{N} : \mathcal{N}_\varepsilon := \{n_1, \dots, n_o\} : \forall n \in \mathcal{N} : \exists n_i \in \mathcal{N}_\varepsilon : d(n, n_i) < \varepsilon$. We call the n_i the ε -centers.*

Proof Any subset of Δ^k is precompact, meaning its closure is compact. Then see e.g. Remark 4.2.3 in [Vershynin, 2020] for the result. ■

Finally, we need to exploit the connectedness of $\mathcal{N}$.

Lemma 281. Let $\mathcal{N} \subseteq \Delta^k$ be connected. Then $\forall \varepsilon > 0$ and any $a, b \in \mathcal{N}$ there exists an ε -chain between a and b , that is, a finite set of points $n_1, \dots, n_o \in \mathcal{N}$ such that $n_1 = a$, $n_o = b$ and $d(n_i, n_{i+1}) < \varepsilon$.

Proof A connected subset of Δ^k is a connected metric space by inheriting the Euclidean metric. Any connected metric space is ε -chainable, see e.g. [Joshi, 1983, p. 148]. ■

From this is the next lemma readily follows.

Lemma 282. Let $\mathcal{N} \subseteq \text{co}(\mathcal{M}) \subseteq \Delta^k$ be connected. Then $\forall \varepsilon > 0$ there exists a finite ε -cover of $\mathcal{N}$, which also forms an ε -chain, where all ε -centers are in $\text{co}(\mathcal{M})$ (possibly outside of $\mathcal{N}$), but at most ζ -far away from $\mathcal{N}$. Formally, $\forall \varepsilon > 0 : \exists p_1, \dots, p_o \in \text{co}(\mathcal{M}) : \forall n \in \mathcal{N} : \exists p_i : d(n, p_i) < \varepsilon$; also, $d(p_i, p_{i+1}) < \varepsilon$; and $\forall p_i : \exists n \in \mathcal{N} : d(p_i, n) < \zeta$.

Proof Let $\varepsilon > 0$ be arbitrary. Take some $\varepsilon' + 2\zeta < \varepsilon$ and use the previous two Lemmata to obtain points $n_1, \dots, n_o \in \mathcal{N}$ which form an ε' -cover of $\mathcal{N}$ and also an ε' -chain. Now to each n_i associate some $p_i \in \text{co} \mathcal{M}$ so that $d(n_i, p_i) < \zeta$, which is possible by assumption that $\mathcal{N} \subseteq \text{co} \mathcal{M}$. By the triangle inequality, these p_i satisfy all desiderata. ■

We are now ready to prove Proposition 273.

Proof Recall that we work in the simplex Δ^k . We begin by setting up some infinite sequences of parameterization data, which we then use for a constructive method. Let $\varepsilon^\infty = (2, \varepsilon_2, \dots)$ and $\delta^\infty = (4(k+1), \delta_2, \dots)$ and $\zeta^\infty = (\zeta_1, \dots)$ be decreasing and strictly positive sequences satisfying $\varepsilon^\infty \downarrow 0$, $\delta^\infty \downarrow 0$ and $\zeta^\infty \downarrow 0$, to be interpreted as tolerance parameters. For each $\delta_i > 0$, we compute

$$v_i := \lceil 4(k+1)/\delta_i \rceil + 1 \in \mathbb{N},$$

which guarantees that $4(k+1)/v_i < \delta_i$. Intuitively, this will be the length with which we invoke the approximation lemma at iteration i . However, to move “safely” we need to ensure the finite sequence is long enough already. For this, we compute a minimally required length $l_i \in \mathbb{N}$ as the smallest natural number that satisfies $0 < 2v_i/(l_i + v_i) \leq \delta_i$. Clearly, by making l_i large enough this can always be guaranteed. For example, for $\delta_1 = 4(k+1)$ we find that $v_1 = 2$ and $l_1 = 1$. In addition, for each $i \in \mathbb{N}$, from invoking Lemma 282 with ε_i and ζ_i , we obtain a finite set of points $E_i = (c_1^i, \dots, c_{o(i)}^i) \in \text{co}(\mathcal{M})^{o(i)}$ which form an ε_i -cover of $\mathcal{N}$ and are also an ε_i -chain, with some finite length $o(i) \in \mathbb{N}$ depending on i . For each of these points $c_j \in E_i$, we have $\exists n \in \mathcal{N} : d(c_j, n) < \zeta_i$.

We now describe how a sequence of measures can be constructed that yields the desired $\mathcal{N}$ as the set of cluster points. The procedure is given as a countable infinity of subsequent iterations, each parameterized by $(\varepsilon_i, \delta_i, \zeta_i, v_i, l_i, E_i)$.

Iteration $i = 1$. For the first iteration $i = 1$, we begin by initializing a finite sequence of

measures of length 1 as $m := (m_0)$ for some arbitrary measure m_0 from $\mathcal{M}$.

We now describe how some iteration $i \in \{2, \dots\}$, of our algorithm works. We assume that we have already accumulated a finite sequence of measures $m = (m_1, \dots, m_n)$ from $\mathcal{M}$. Intuitively, we now want to “get very close to” all measures in E_i .

Iteration i . At the beginning of iteration i we make two key inductive assumptions:

1. The current average $r(m)$ lies in $B_{\delta_i}(c_1^i)$.
2. The length $|m| = n$ satisfies $n \geq l_i$.

From the ε_i -centers together with the first point of the ε_{i+1} -centers, create a list of pairs of the form

$$C = (c_1^i, c_2^i), \dots, (c_{o(i)-1}^i, c_{o(i)}^i), (c_{o(i)}^i, x_1), \dots, (x_t, c_1^{i+1}),$$

where we also connected $c_{o(i)}^i$ and c_1^{i+1} with an ε_i -chain $c_{o(i)}^i, x_1, \dots, x_t, c_1^{i+1}$, where all points satisfy that their distance to $\mathcal{N}$ is at most ζ_i (possible by the reasoning of Lemma 282). Note that all such points are in $\text{co}(\mathcal{M})$, but possibly outside of $\mathcal{N}$.

The Pair Subroutine. We now consider each pair $(c_1, c_2) \in C$ with the above order. For each pair, at the beginning of our construction, the current average satisfies $r(m) \in B_{\delta_i}(c_1)$, and we want to append finitely many measures $\tilde{m} = (\dots)$ from $\mathcal{M}$ to m so that the new average $r(m \oplus \tilde{m})$ is in $B_{\delta_i}(c_2)$ at the end of the pair subroutine. To now get closer to c_2 we first approximate c_2 by $\tilde{c}_2 = r_b$ as in the approximation lemma 278, using the length v_i for some $b \in (\mathcal{M})^{v_i}$, and we get that $d(c_2, \tilde{c}_2) \leq 4(k+1)/v_i < \delta_i$. Noting that there must exist some $k \in \mathbb{N}$ such that

$$r(m \oplus b^k) \in B_\kappa(\tilde{c}_2), \quad (5.27)$$

where $\kappa > 0$ is chosen such that $\kappa + d(c_2, \tilde{c}_2) < \delta_i$, which is possible. The existence of such $k \in \mathbb{N}$ for arbitrarily small κ follows from Lemma 276. Thus we can append b often enough to end up in the δ_i -ball around $\tilde{c}_2$, since $d(r(m \oplus b^k), c_2) \leq d(r(m \oplus b^k), \tilde{c}_2) + d(\tilde{c}_2, c_2) \leq \delta_i$. Now set $\tilde{m} := m \oplus b^k$.

We now observe that the maximum distance we might have been away from $\mathcal{N}$ during this procedure is bounded by $\varepsilon_i + 2\delta_i + \zeta_i$: first, we have moved on the $[r(m), \tilde{c}_2]$ line, and potentially went as far as δ_i away from it during appending but no further. Why?:

$$d(m \oplus b^{k'}, m \oplus b^{k'} \oplus (b_1, \dots, b_j)) \leq 2v_i / (l_i + v_i) \leq \delta, \quad \forall k' \in \{0, \dots, k-1\} \quad \forall j \in \{1, \dots, v_i\},$$

noting that $2v_i / (\tilde{l}_i + v_i) < 2v_i / (l_i + v_i)$ for any $\tilde{l}_i > l_i$.

Second, each point e on $[r(m), \tilde{c}_2]$ is at most δ_i away from $[c_1, c_2]$, i.e. there exists a point $c \in [c_1, c_2]$ such that $d(e, c) < \delta_i$. Why? Write $e = (1 - \alpha)r(m) + \alpha\tilde{c}_2$, $\alpha \in [0, 1]$. Consider

the point $f = (1 - \alpha)c_1 + \alpha c_2$. Then

$$d(e, f) = \|(1 - \alpha)r(m) + \alpha\tilde{c}_2 - (1 - \alpha)c_1 + \alpha c_2\| \quad (5.28)$$

$$\leq (1 - \alpha)\|r(m) - c_1\| + \alpha\|\tilde{c}_2 - c_2\| \quad (5.29)$$

$$\leq (1 - \alpha)\delta_i + \alpha\delta_i \quad (5.30)$$

$$= \delta_i, \quad (5.31)$$

since $d(r(m), c_1) < \delta_i$ and $d(\tilde{c}_2, c_2) < \delta_i$. Now, since each point on $[c_1, c_2]$ is at most $\varepsilon_i + \zeta_i$ away from $\mathcal{N}$ due to the ε_i -chain property, we conclude by the triangle inequality that the new averages in this iteration i were never further than $\varepsilon_i + 2\delta_i + \zeta_i$ away from $\mathcal{N}$. Now set $m := \tilde{m}$ and proceed with the next pair.

[End of Pair Subroutine]

After executing the pair subroutine for all pairs, we have accumulated a finite sequence of measures m such that $r(m) \in B_{\delta_i}(c_1^{i+1})$. Now we would like to move from a δ_i -ball around it to some $\delta_{i+1} < \delta_i$ ball around it, to satisfy the inductive assumption of the next iteration $i + 1$. For this, we invoke the “stay and grow” lemma 279 with $v = v_i$ to end up again within the δ_i -ball around c_1^{i+1} and grow the length of the finite sequence until it has reached at least l_{i+1} : applying the lemma once yields some $b \in \mathcal{M}^{v_i}$ so that

$$r(m \oplus b) \in B_{\delta_i}(c_1^{i+1}) \quad \text{and} \quad d(r(m \oplus (b_1, \dots, b_j)), c_1^{i+1}) < 2\delta_i, \quad \forall j \in \{1, \dots, v_i\}.$$

In words, the lemma guarantees that when making these appendings, the average is never further than $2\delta_i$ away from c_1^{i+1} . Invoking the lemma r many times, for some finite $r \in \mathbb{N}$, set $m := m \oplus b^r$, guarantees that we reach length $|m| = l_{i+1}$.

We now show that this allows us to safely move to a δ_{i+1} -ball around c_1^{i+1} . Our current location in the simplex is $p = r(m)$. We invoke the approximation lemma with $v = v_{i+1}$ around c_1^{i+1} to obtain a $q = r_b$, $b \in \mathcal{M}^{v_i+1}$, with $d(c_1^{i+1}, q) < \delta_{i+1}$. But then there exists some $\kappa > 0$ so that $\kappa + d(c_1^{i+1}, q) < \delta_{i+1}$. Now we can move on the line $[p, q]$ until we are in a κ -ball around q , which in turn means we are in a δ_{i+1} -ball around c_1^{i+1} . Formally:

$$\exists k \in \mathbb{N} : r(m \oplus b^k) \in B_\kappa(q). \quad (5.32)$$

How “safe” are these iterative appendings? Since c_1^{i+1} is at most ζ_i -far away from $\mathcal{N}$, $r(m) \in B_\delta(c_1^{i+1})$ and for appending b we have the δ_{i+1} -bound,

$$d(m \oplus b^{k'}, m \oplus b^{k'} \oplus (b_1, \dots, b_j)) < \delta_{i+1}, \quad \forall k' \in \{0, \dots, k-1\} \quad \forall j \in \{1, \dots, v_i\},$$

we have been at most $\delta_i + \delta_{i+1} + \zeta_i < 2\delta_i + \zeta_i$ away from $\mathcal{N}$. Now set $m := m \oplus b^k$.

To conclude the logic of an iteration i , observe that all new averages during this iteration

have been at most $\varepsilon_i + 2\delta_i + \zeta_i$ away from $\mathcal{N}$, and also we have been $\varepsilon_i + \delta_i + \zeta_i$ -close to any point on $\mathcal{N}$.

It remains to argue that the set of cluster points coincides exactly with $\mathcal{N}$. First, obviously the above procedure guarantees that any $n \in \mathcal{N}$ is a cluster point: as $\varepsilon^\infty \downarrow 0$, $\delta^\infty \downarrow 0$ and $\zeta^\infty \downarrow 0$, we enter any neighborhood around any point on $\mathcal{N}$ infinitely often. To see that no other point (outside of $\mathcal{N}$) can be a cluster point, just observe that the averages generated by the above construction get arbitrarily close to $\mathcal{N}$, and $\mathcal{N}$ is already closed by assumption. So for any point $p \notin \mathcal{N}$ we can find an open neighborhood U around it so that $U \cap \mathcal{N} = \emptyset$, which is sufficiently small so that our algorithm never enters this neighborhood, hence the point cannot be a cluster point. ■

CONCLUDING THE PROOF OF PROPOSITION 240

It remains only to show the backward direction for Proposition 273.

Proposition 283. *Let $\emptyset \neq \mathcal{M} \subseteq \Delta^k$ and $m^\infty = (m_1, m_2, \dots)$ a sequence of measures from $\mathcal{M}$. Then the set $\text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right)$ must be closed, connected and a subset of $\text{co}(\mathcal{M})$.*

Proof It is clear in general that a set of cluster points must be closed in the respective topology. For connectedness of $\text{CP} \left(n \mapsto \frac{1}{n} \sum_{i=1}^n m_i \right)$, see [Fröhlich et al., 2024, Lemma B.10], stated for the case where $\mathcal{M}$ is the set of vertices of Δ^k , but the argument applies for any $\mathcal{M} \subseteq \Delta^k$. Finally, assume by contradiction that some point $p \notin \text{co}(\mathcal{M})$ is a cluster point of $n \mapsto \frac{1}{n} \sum_{i=1}^n m_i$. But for any $n \in \mathbb{N}$, $\left(\frac{1}{n} \sum_{i=1}^n m_i \right) \in \text{co}(\mathcal{M})$, and since $p \notin \text{co}(\mathcal{M})$ we can find an open neighborhood U around it so that $U \cap \text{co}(\mathcal{M}) = \emptyset$. But then the sequence cannot ever enter U , and hence p cannot be a cluster point. ■

6

Scoring Rules and Calibration for Imprecise Probabilities

AUTHOR CONTRIBUTIONS

Robert Williamson (RW) contributed through supervision of the project, e.g. through discussions and proof-reading. Importantly, RW pointed to the paper of [Grünwald and Dawid \[2004\]](#). The remaining work is due to CF.

ABSTRACT

What does it mean to say that, for example, the probability for rain tomorrow is between 20% and 30%? The theory for the evaluation of precise probabilistic forecasts is well-developed and is grounded in the key concepts of *proper scoring rules* and *calibration*. For the case of imprecise probabilistic forecasts (sets of probabilities), such theory is still lacking. In this work, we therefore generalize proper scoring rules and calibration to the imprecise case. We develop these concepts as relative to data models and decision problems. As a consequence, the imprecision is embedded in a clear context. We establish a close link to the paradigm of *(group) distributional robustness* and in doing so provide new insights for it. We argue that proper scoring rules and calibration serve two distinct goals, which are aligned in the precise case, but intriguingly are not necessarily aligned in the imprecise case. The concept of decision-theoretic entropy plays a key role for both goals. Finally, we demonstrate the theoretical insights in machine learning practice, in particular we illustrate subtle pitfalls relating to the choice of loss function in distributional robustness.

6.1 INTRODUCTION

The field of machine learning is inextricably intertwined with probability theory. In particular, a range of machine learning methods output probabilistic predictions (or *forecasts*). In the case of classification, given an input datum $X = x$, such a predictor $f(x)$ returns a probabilistic forecast for the associated label Y . While it is often assumed that probability theory is *the* comprehensive framework for mathematizing uncertainty, various challenges have been raised that focus on its shortcomings, both from subjectivist [Levi, 1980, Walley, 1991, Gilboa et al., 2009, Joyce, 2010] as well as generalized frequentist perspectives [Fine, 1988, Fröhlich and Williamson, 2024a, Fröhlich et al., 2024] (see Section 6.1.1). In one cluster of responses, scholars have argued that relevant shortcomings may be addressed by moving from precise probabilities to *imprecise* probabilities (IP), which are essentially equivalent to sets of probabilities. In the case of a binary event (“will it rain tomorrow?”), such an imprecise probability is simply an interval $[p, q] \subseteq [0, 1]$.

Following this line, our goal is to construct machine learning methods which output *imprecise* forecasts. However, this necessitates the development of a principled theory for the *evaluation* of such forecasts which is non-existent so far: in light of training data, how “good” was a given IP forecast? Access to such an evaluation procedure would also provide guidance on how to obtain such forecasts, which allows the training of machine learning models. In similar spirit, in recent years there have been attempts to develop a theory for the evaluation of imprecise *beliefs* from a subjectivist perspective [Seidenfeld et al., 2012, Konek, 2015, 2019, 2023].¹ To the best of our knowledge, there is no proposal which claims to offer a satisfying positive development of a theory for the evaluation of imprecise forecasts, and indeed some authors provide impossibility theorems, which at first sight may seem to threaten the whole endeavour. We are not discouraged by the negative results (see the discussion in Section 6.2.1) and offer the following diagnosis for what has been missing in previous approaches so far: *data models* and *decision problems*. Our philosophy differs from the subjectivist paradigm in that we emphasize the primacy of data and data model in forecasting; we do not think that imprecise forecasts can be evaluated “in the void”, but only relative to data (model) and decision problem. Consequently, attention centers on imprecision *in the data model*, and indeed our findings support the position that this kind of imprecision is to some extent even more fundamental than imprecision *in forecasts*.

We summarize our paper and its contributions as follows: we develop a viable account of proper IP scoring rules, which are relativized to data models and decision problems. We characterize the optimal IP forecast *for a fixed loss function* based on the maximum entropy principle by Grünwald and Dawid [2004]. We demonstrate that imprecision in the fore-

¹While Seidenfeld et al. [2012] also use the term “forecast”, their approach is subjectivist and thus aligned with Konek [2015], who uses the term “credences”. To us, the term “forecast” carries a (generalized) frequentist flavor and implies evaluation against a data sequence or a data model.

casts might be required if the loss function is not known a priori. We shed light on the tight connection between our proposal of imprecise scoring rules and the paradigm of (*group*) *distributional robustness* [Hu et al., 2018, Sagawa et al., 2019], where a *coherent risk measure* of the loss is optimized. Further, we develop an adequate concept of *calibration* for imprecise probabilities. As for scoring rules, the concept of a generalized decision-theoretic *entropy* plays a key role here. We discuss how proper scoring rules and calibration serve two distinct goals of forecasting quality, and show that they do not necessarily align anymore in the imprecise case. In particular, the optimal forecast (in the sense of score) need not be calibrated. Finally, in Section 6.4, we demonstrate the benefits of imprecise forecasts in machine learning practice. In addition, we believe that even readers who are only interested in precise probability may profit from reading this paper, as the development we offer here provides new insight into the conceptual interplay between scoring rules and calibration in general.

To set the stage, by an *imprecise probability* [Walley, 1991, Augustin et al., 2014], we here mean any non-empty set of probability measures $\mathcal{P}$. Associated to this set is a generalized non-linear expectation functional

$$\bar{R}_{\mathcal{P}}(X) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[X].$$

This functional is known as *coherent upper prevision* [Walley, 1991] in the IP community and equivalently as *coherent risk measure* [Artzner et al., 1999] in finance and insurance.² In this paper, we choose the perhaps more neutral term *upper expectation*. Similarly, the *upper probability* of an event A is $\bar{P}_{\mathcal{P}}(A) := \sup_{P \in \mathcal{P}} P(A)$. The conjugate lower quantities are $\underline{R}(X) := -\bar{R}(-X)$ and $\underline{P}(A) := 1 - \bar{P}(A^C)$. Hence for an event we obtain an interval $[\underline{P}(A), \bar{P}(A)] = [\inf_{P \in \mathcal{P}} P(A), \sup_{P \in \mathcal{P}} P(A)]$, where the width is the degree of imprecision, and similarly for $\underline{R}$ and $\bar{R}$.

6.1.1 DATA MODELS AND IMPRECISION

In contrast to previous literature with the goal of developing scoring rules for imprecise forecasts, we put *data models* [Fröhlich and Williamson, 2024b] at the center of the stage. Intuitively, we understand by data model a mathematical formalization and idealization of a data-generating process, therefore with an inevitable aspect of subjectivity as in any modelling process. As an example, consider the familiar *i.i.d.* model, which, at the level of abstraction where we use the term, is one data model. Our logic is that we should first seek to understand what possible semantics imprecision can have with respect to data models. This enables us to subsequently embed imprecise forecasts in a meaningful context of data

²We refer the reader to [Fröhlich and Williamson, 2024a] for an explanation of the subtle conceptual relations.

model and decision problem. The situation is however complicated, since imprecision has various different semantics in the literature, which are moreover not conceptually clear-cut. We therefore briefly survey these semantics now (non-comprehensively), noting that there is a rough hierarchical progression among them.

First, in the setting of objective *risk*, the assumption is that the data model is parameterized by a single, precise probability, and that it is known to the decision maker. For example, consider flipping a fair coin. This process is well-described by an *i.i.d.* model with a probability $p = 0.5$ for the event “heads”. In this setting there is still uncertainty about the outcome, often called *aleatoric uncertainty*. Following expected utility (EU) theory [von Neumann and Morgenstern, 1947], a decision maker would use the expectation to assess the risk of a gamble (a random variable). However, a plenitude of criticism has been raised against EU theory [Allais, 1953, Buchak, 2013]. Importantly, it is unclear whether EU theory can adequately represent the phenomenon of risk aversion [Buchak, 2013]. As a response, *rank-dependent* EU [Wang, 1995, Quiggin, 2012] has been developed, closely related to particular upper expectations known as *spectral risk measures* [Acerbi, 2002, Wang, 2000]. In this way, imprecise probabilities can serve to express risk aversion. Note that here imprecision manifests not in the (precise) data model but in the non-linear *aggregation* of consequences.

Risk is often contrasted with *ambiguity* [Ellsberg, 1961].³ Assume now that the hypothesized “true” data model is parameterized by a single, precise probability, but it is unknown to the decision maker. For instance, consider flipping a coin with unknown bias. Here, the model of an underlying precise probability still appears adequate. The problem now has a component of *epistemic uncertainty*, although aleatoric uncertainty is additionally still present.⁴ At this level, imprecise probabilities, appearing also under the banner of *distributional robustness* [Kuhn et al., 2019, Rahimian and Mehrotra, 2019] in machine learning, can be employed to capture the attitude of *ambiguity aversion* towards epistemic uncertainty (see also [Fox et al., 2021]). This setting describes a typical machine learning problem, where we only have access to the empirical train distribution (the training data), but not to the hypothesized “true” distribution. The goal of distributionally robust optimization (DRO) in this context is then to robustify against this lack of knowledge [Kuhn et al., 2019].

Until now, the data model was parameterized by a precise probability, and imprecision expressed an attitude towards either risk (related to aleatoric uncertainty) or ambiguity

³We remark that this distinction is not without controversy. Some authors define risk as the situation where probabilities are either known *or* they can be reliably estimated from experience (e.g. [Levy et al., 2010]). Risk in the context of insurance (e.g. for workplace accidents) would fit this definition, for example. Other times, the term risk is used in a strong objective sense, where the probabilities are exactly known, as in the lotteries of von Neumann and Morgenstern’s [1947] expected utility theory. To us, the line appears blurry.

⁴See [Hüllermeier and Waegeman, 2021] for the distinction of aleatoric versus epistemic uncertainty.

(related to epistemic uncertainty, plus aleatoric uncertainty). As another setting, intensifying further, we introduce the data model which will be the focus of this paper, where the data comes from a set of probabilities. Let Ω with $|\Omega| < \infty$ be a possibility set, which captures all relevant information. Intuitively, $\omega \in \Omega$ is a state of the world. For example, $\Omega = \mathcal{X} \times \mathcal{Y}$. Then $\omega^\infty = (\omega_1, \omega_2, \dots)$ is a data sequence. The *non-stationary, locally precise (NSLP) data model* [Fröhlich and Williamson, 2024b] is defined using independent draws from a sequence of probabilities. Fix a sequence $P^\infty = (P_1, P_2, \dots)$ of probability measures on Ω , and demand that

$$\lambda \{ \omega^\infty = (\omega_1, \omega_2, \dots) \} = \prod_{i=1}^{\infty} P_i(\{\omega_i\}),$$

where λ is the Lebesgue measure.⁵ This model generalizes the *i.i.d.* model to a set of probabilities, an imprecise probability $\mathcal{P} := \{P_i : i \in \mathbb{N}\}$, and models the data as independent with respect to a sequence of probabilities from this set. Intuitively, consider a sequence of coin flips with varying biases. This model, in our view, can be seen as a rather mild departure from an *i.i.d.* model, as it keeps independence and the drawing of each datum is still precise probabilistic. Note, however, that under this model, imprecision is inherent to the data model, and not reflective of our epistemic uncertainty (the predictor only has access to information in ω).⁶ How to adequately model epistemic uncertainty in this picture in addition is an open problem to the best of our knowledge, which we put aside here. This model appears (implicitly) in many applications in machine learning, for example in the context of fairness [Buolamwini and Gebru, 2018], federated learning [McMahan et al., 2017] and more. The paradigm of *group distributional robustness*⁷ [Sagawa et al., 2019] fits well with this data model, where the assumption is that each group $S = s$ corresponds to a distinct probability $P_{S=s}$. Group DRO is relevant in practice, for example, in a fairness context when the “sensitive attribute”⁸ is only available at training time, but we aim to deploy a single model at test time, with the aim of robust performance with respect to the sensitive attribute; or, when we know of a spurious correlation in the training data, which can be used to define a set of groups [Sagawa et al., 2019]. Throughout the paper

⁵We implicitly work on the measure space $([0, 1], \mathcal{B}([0, 1]), \lambda)$, but suppress this in the notation. See [Fröhlich and Williamson, 2024b] for the rigorous formulation involving maps of the signature $W_i : [0, 1] \rightarrow \Omega$.

⁶It should be highlighted that we think of the epistemic vs. aleatoric distinction as essentially contextual: adding additional information to ω , for instance by collecting a new feature, can change the picture.

⁷The distinction between *distributional robustness* (e.g. [Duchi and Namkoong, 2021]) and *group distributional robustness* (e.g. [Sagawa et al., 2019]) is, in our eyes, to some extent blurry, but can be related to how the set of probabilities is constructed. Distributional robustness would typically employ *law-invariant coherent risk measures* [Föllmer and Schied, 2016], for example based on an f -divergence ball around the empirical distribution, whereas in group distributional robustness, the set of probabilities corresponds directly to a set of identified subpopulations in the data. Due to envelope representations [Pflug and Romisch, 2007], both approaches can be interpreted in terms of subpopulations, however.

⁸In fair machine learning, a sensitive attribute is a sensitive piece of information about a person, such as race or gender. Such an attribute canonically induces a set of groups.

we are guided by this data model for inspiration, as the imprecision here is fundamental to the data model, implying that the imprecision is not a matter of our lack of knowing the model. The mathematical development here is not tied to this data model though, but interpretations may change.

Finally, we mention in passing that there exists a whole zoo of stationary data models which are parameterized by an imprecise probability [Fine, 1988, Fröhlich and Williamson, 2024b], but we believe the clearest in terms of applications is the NSLP model presented above. For brevity, we refer to a data model parameterized by some imprecise probability $\mathcal{P}$ as a “ $\mathcal{P}$ data model”.

6.2 IP SCORING RULES

In this section we present an account of proper scoring rules for imprecise probabilities. Our primary goal is to enable decision making for an agent who is associated with a loss function. This means that our forecaster (the machine learning model) should recommend “favorable” actions to the agent, where a “favorable” action is one that leads to low loss values. That is, for us the evaluation of a forecast is inextricably intertwined with action recommendation, which shapes our perspective throughout the paper. The aggregation of these losses, which determines the overall performance, is grounded in the data model. We foreshadow that in Section 6.3 we will introduce another, orthogonal goal for the decision maker: intuitively, forecasts should also be trustworthy in the sense that their associated uncertainty should be “appropriate” (calibrated).

Our setup is as follows. Assume a possibility set $\Omega = \mathcal{X} \times \mathcal{Y}$, with features $X : \Omega \rightarrow \mathcal{X} \subset \mathbb{R}^9$ and labels $Y : \Omega \rightarrow \mathcal{Y} \subset \mathbb{R}$. To reduce the burden of technicalities, we assume $2 \leq |\Omega| = k < \infty$ throughout the paper. The set of bounded functions is $L^\infty := \{Z : \Omega \rightarrow \mathbb{R}\}$ in this case, and we call such functions $Z \in L^\infty$ *gambles*. We assume that an agent needs to make a decision in the form of choosing an action $a^* \in \mathcal{A}$, where the action space $\mathcal{A}$ is assumed finite for simplicity and $|\mathcal{A}| \geq 2$. The consequence is then evaluated by a loss function $\ell : \mathcal{A} \times \Omega \rightarrow \mathbb{R}$, where $\ell_a = \ell(a, \cdot) \in L^\infty \forall a \in \mathcal{A}$. We denote the set of loss functions by $\mathcal{L}$. The quantity $\ell(a, \omega)$ represents the loss (disutility) that the agent incurs when following action $a \in \mathcal{A}$ and the state of the world $\omega \in \Omega$ actually materializes. Note that we impose no further constraints on ℓ . Throughout the paper, we often need to access the elementary event directly, and thus write expressions of the form $\mathbb{E}_P[\ell(a, \omega)]$, where P is a probability measure on Ω , as shorthand for $\mathbb{E}_P[\omega \mapsto \ell(a, \omega)]$.¹⁰

Example 284. As a standard example, consider predicting whether it will rain ($Y = 1$) or not ($Y = 0$) on a given day, which is represented by some features X (temperature

⁹We could also have $\mathbb{R}^d$ -valued features, since $\mathbb{R}$ and $\mathbb{R}^d$ have the same cardinality.

¹⁰In other words, in expressions such as $\mathbb{E}_P[\ell(a, \omega)]$ we implicitly treat ω as the identity random variable.

etc.). The action space $\mathcal{A}$ could consist of the actions “bringing an umbrella” ($a = 1$) and “not bringing an umbrella” ($a = 0$). A choice of a loss function ℓ is then equivalent to a real-valued 2×2 matrix, assessing the possible consequences. Negative values would correspond to utility (gain), whereas positive values correspond to disutility (loss). We are then interested in evaluating an imprecise forecaster in light of an actual outcome sequence, in this case a binary sequence indicating “rain” or “no rain”. We extend this example in Appendix 6.7.6.

The main goal is to recommend a supposedly “optimal” action to the agent in terms of the consequences. There are two main open questions:

- Q1.** Given an imprecise forecast, how does the agent choose an action?
- Q2.** How do we aggregate losses, that is, how do we summarize the performance in the long run over many ω s?

To answer Q1, we first define the notion of an imprecise forecast. Since $|\Omega| < \infty$, a probability measure P on Ω can be identified with a point in the simplex $\Delta^k := \{(p_1, p_2, \dots, p_k) : \sum_{j=1}^k p_j = 1\}$.¹¹ We write $\mathcal{IP}$ for the set of nonempty subsets of Δ^k , the imprecise probabilities. A precise probability $P \in \Delta^k$ can be seen as an imprecise probability $\{P\} \in \mathcal{IP}$, a viewpoint which we adopt implicitly throughout the paper. An imprecise forecast is then a function $\mathcal{Q} : \mathcal{X} \rightarrow \mathcal{IP}$, which takes as input the features $X = x$ and outputs a set of probabilities $\mathcal{Q}(x) := \mathcal{Q}(X(\omega)) \in \mathcal{IP}$ on Ω . Clearly, a reasonable forecast $\mathcal{Q}(x)$ would assign zero mass to those ω where $X(\omega) \neq x$, at least if the forecaster considers the $X = x$ information as perfectly trustworthy. However, we do not impose this restriction in this paper for additional flexibility, although this introduces some new translation effort. We slightly abuse notation and also write $\mathcal{Q} \in \mathcal{IP}$ for a constant forecast (be it precise or actually imprecise).

Having access to a forecast $\mathcal{Q}(x)$, the agent needs to choose an action. There are multiple, non-equivalent decision-making principles for imprecise probabilities [Troffaes, 2007]. In this paper we consider a *worst-case* (pessimistic) attitude for multiple reasons, primarily because it can interplay nicely with aggregation in the data model. This principle is also known as (Γ) -MinMax [Gilboa and Schmeidler, 1989].¹²

Assumption 285. *The agent chooses the action by the MinMax principle:*

$$a_{\mathcal{Q}(x)}^* \leftarrow \operatorname{argmin}_{a \in \mathcal{A}} \sup_{Q \in \mathcal{Q}(x)} \mathbb{E}_Q[\ell(a, \omega)], \quad (6.1)$$

where $\mathcal{Q}(x) \in \mathcal{IP}$ is the imprecise forecast. The “ $\leftarrow$ ” operator picks a single element from the potentially set-valued argmin in a consistent way, see Appendix 6.7.1.

¹¹On Δ^k we can assume the Euclidean topology without loss of generality [Fröhlich and Williamson, 2024b].

¹²Or “ Γ -MaxMin” if one works with a utility-based orientation, equivalent up to a sign flip.

For brevity, we usually write $\mathcal{Q}(X)$ or $\mathcal{Q}(\omega)$ instead of $\mathcal{Q}(X(\omega))$, but sometimes choose the latter for emphasis. After having committed to an action $a_{\mathcal{Q}(X(\omega))}^*$, the agent incurs the loss $\ell(a_{\mathcal{Q}(X(\omega))}^*, \omega)$. This key quantity of interest defines our IP scoring rule.

Definition 286. For a loss function $\ell \in \mathcal{L}$, we define the IP scoring rule as:

$$\tilde{S}_\ell(\omega) := S_\ell(\mathcal{Q}(X(\omega)), \omega) := \ell(a_{\mathcal{Q}(X(\omega))}^*, \omega). \quad (6.2)$$

Note that the definition of this function might implicitly depend on how tie-breaking in the argmin (Equation 6.1) is resolved. The IP scoring rule quantifies the adverse consequences resulting from following the forecast $\mathcal{Q}$ when ω materializes. This way of obtaining a scoring rule essentially generalizes the concept of a *tailored scoring rule* [Dawid and Sebastiani, 1999, Johnstone et al., 2011] by incorporating a MinMax assumption. Here, *tailored* means that the scoring rule S_ℓ is constructed in such a way so as to enable decision making for an agent with underlying loss function ℓ .

There is still one crucial ingredient missing here, which is the data model. Introducing a data model then allows answering Q2, that is, how to aggregate the scores over repeated outcomes $\omega_1, \omega_2, \dots$ and so on. For the experiments in this paper, we will focus on the NSLP data model presented in Section 6.1.1, which is parameterized by a set of probabilities $\mathcal{P}$ on Ω ; the theoretical developments however can also be instantiated with other models, as long as they are parameterized by an imprecise probability. For such decision problems, we propose a MinMax assumption, which will establish the link to DRO (Section 6.2.2):

D1. MinMax-Assumption 2: we aim for worst-case (“robust”) performance over the underlying set of probabilities.

Combining this with our IP scoring rule, we express D1 as a MinMax objective:

Assumption 287. When the data model is parameterized by $\mathcal{P} \in \mathcal{IP}$, the overall problem (minimizing IP-Score) from the perspective of the forecaster is:

$$\text{IP-Score}^*(\ell, \mathcal{P}) := \min_{\mathcal{Q}: \mathcal{X} \rightarrow \mathcal{IP}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[S_\ell(\mathcal{Q}(X(\omega)), \omega)], \quad (6.3)$$

where the minimization is over all imprecise forecasts.¹³

Using the upper expectation, $\text{IP-Score}^*(\ell, \mathcal{P}) = \min_{\mathcal{Q}: \mathcal{X} \rightarrow \mathcal{IP}} \bar{R}_{\mathcal{P}}(S_\ell(\mathcal{Q}(X(\omega)), \omega))$. We refer to the inner quantity $\bar{R}_{\mathcal{P}}(S_\ell(\mathcal{Q}(X(\omega)), \omega))$ as the IP score of $\mathcal{Q}$ under a $\mathcal{P}$ data model. Even though in practice, we of course work with constrained hypothesis classes, we consider the unconstrained objective in theory. Observe that the MinMax assumptions 285

¹³The minimum is indeed attained since with a finite action space, $S_\ell(\mathcal{Q}(X(\omega)), \omega)$ can only take on finitely many distinct values.

and 287 appear similar, but are conceptually distinct. Whereas 285 refers to how the agent subjectively picks an action, 287 is about aggregation of losses in the data model. However, it is the shared MinMax form that enables their constructive interplay. What we believe is unique about the MinMax criterion, in contrast to other criteria [Troffaes, 2007], is that it appears sensible both from an agent’s as well as a data model’s perspective.

Remark/Def. 288. We note that if $\bar{co}(\mathcal{P}) = \bar{co}(\mathcal{P}')$, then $\bar{R}_{\mathcal{P}} = \bar{R}_{\mathcal{P}'}$, where $\bar{co}(\cdot)$ denotes the closed convex hull, that is, upper expectations are unique only up to closed convex hull [Walley, 1991, Section 3.6]; see also [Hiriart-Urruty and Lemaréchal, 2004, Proposition 2.2.1]. If $\bar{co}(\mathcal{P}) = \bar{co}(\mathcal{P}')$, we write $\mathcal{P} \simeq \mathcal{P}'$, and otherwise $\mathcal{P} \not\simeq \mathcal{P}'$.

6.2.1 WEAK AND STRONG PROPRIETY

A first question to ask, and also a sanity check, is whether our IP scoring rules satisfy a notion of *propriety*. From a data model perspective, we suggest, a scoring rule may be called (strictly) proper if the optimal forecast coincides with the parameterization of the data model. This is the key for estimation procedures to work. In the precise case of an *i.i.d.* model parameterized by P , this means that P is an optimal constant forecast under a proper scoring rule.

To state the definition of propriety, we need to ignore the features for a moment and focus on constant forecasts. In the precise case, a scoring rule $s : \Delta^k \times \Omega \rightarrow \mathbb{R}$ is called weakly proper if $\mathbb{E}_P[s(P, \omega)] \leq \mathbb{E}_P[s(Q, \omega)] \quad \forall P, Q \in \Delta^k$. That is, no other forecast Q can do better than the underlying P . This can straightforwardly be generalized to our IP scoring rules.

Proposition 289 (Weak propriety). *Let $\ell \in \mathcal{L}$ and $\mathcal{P}, \mathcal{Q} \in \mathcal{IP}$. Then it holds that*

$$\bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{P}, \omega)) \leq \bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{Q}, \omega)). \quad (6.4)$$

The proof is in Appendix 6.7.2. Note that the existence of weakly proper IP scoring rules is not in contradiction to previous impossibility theorems in the literature [Seidenfeld et al., 2012, Mayo-Wilson and Wheeler, 2016, Schoenfeld, 2017], which preclude *strictly* proper IP scoring rules. Strict propriety would mean that equality in Equation 6.4 only holds if $\mathcal{P} = \mathcal{Q}$. Clearly, if $\mathcal{P} \simeq \mathcal{Q}$, then $S_{\ell}(\mathcal{P}, \omega) = S_{\ell}(\mathcal{Q}, \omega)$, meaning the set of extreme points is an equivalent forecast since $\text{ext}(\mathcal{P}) \simeq \mathcal{P}$. This could be seen as a failure of strict propriety, but we only really differentiate between sets of probabilities up to closed convex hull. Another question is whether this can happen if $\mathcal{P} \not\simeq \mathcal{Q}$. Thus a further question is whether strict propriety can fail with an actually different forecast. First, we note that there are $\ell \in \mathcal{L}$ which lead to complete failure of strict propriety (consider for example the constant zero loss). Thus the question of interest is whether there exist any ℓ such that the induced IP scoring rule is strictly proper (when accounting for the “ $\simeq$ ” equivalence).

Indeed our IP scoring rules have the following defect when looking at precise forecasts for comparison.

Proposition 290 (Partial failure of strict propriety). *The following statements hold:*

(290.1) $\forall \ell \in \mathcal{L} : \exists \mathcal{P} \in \mathcal{IP} : \exists P \in \bar{\text{co}}(\mathcal{P})$ so that $\bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{P}, \omega)) = \bar{R}_{\mathcal{P}}(S_{\ell}(P, \omega))$.

(290.2) Let $|\mathcal{A}| \geq 3$. Then $\exists \ell : \exists \mathcal{P}$ so that $\forall P \in \Delta^k : \bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{P}, \omega)) < \bar{R}_{\mathcal{P}}(S_{\ell}(P, \omega))$.

The proof is in Appendix 6.7.3. In summary, if we mean by strict propriety that the inequality in Definition 6.4, for a fixed $\ell \in \mathcal{L}$, holds *strictly* for all pairs of $\mathcal{P}, \mathcal{Q}$, even if restricting us to $\mathcal{P} \not\subseteq \mathcal{Q}$, then our IP scoring rules are not strictly proper. Statement (290.1) asserts that for any ℓ , strict propriety must fail in combination with some $\mathcal{P}$ due to the existence of an equivalent precise forecast $P \in \bar{\text{co}}(\mathcal{P})$; although statement (290.2) assures us that for some ℓ this need not happen *for all* $\mathcal{P}$, meaning there exist pairs $(\ell, \mathcal{P})$ so that no precise forecast can be optimal. The take-away is that the situation is intricate: whether we can reduce the problem of finding an optimal constant IP forecast to finding an constant optimal precise forecast may depend on both ℓ and $\mathcal{P}$. This will generalize to the non-constant case, see Section 6.2.2.

Moreover, observe that in statement (290.1) the precise forecast $P \in \bar{\text{co}}(\mathcal{P})$ which is equivalent to $\mathcal{P}$ may depend on ℓ . In practice, the loss function relevant at test time may not always be known at training time, where we typically optimize a surrogate loss instead of the loss we actually care about. Recently, the paradigm of *omniprediction* emerged in the machine learning community [Gopalan et al., 2021], where the goal is to perform well with respect to many loss functions simultaneously. Related to this framework, we make the following observation about our IP scoring rules: for any other candidate forecast $\mathcal{Q} \in \mathcal{IP}$, there exists a loss function ℓ so that the IP score of the forecast $\mathcal{P}$ under a $\mathcal{P}$ data model will be lower than that of $\mathcal{Q}$.

Proposition 291 (“Strong” propriety). *Let $\mathcal{P}, \mathcal{Q} \in \mathcal{IP}$ be any constant forecast so that $\mathcal{Q} \not\subseteq \mathcal{P}$. Then there must exist some loss function $\ell \in \mathcal{L}$ such that*

$$\bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{P}, \omega)) < \bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{Q}, \omega)).$$

The proof is in Appendix 6.7.4. Note the order of the quantifiers, contrasting with that of Proposition 290. In words, if our goal is “omniprediction”, then the underlying $\mathcal{P}$ will win over all constant competitor forecasts in the above sense, whether these competitors are precise or imprecise.

6.2.2 REDUCTION TO PRECISE FORECASTS VIA MAXIMUM ENTROPY

We have so far investigated propriety for the case of constant forecasts, which is also how propriety is formulated in the precise case. Yet our interest in practice lies with forecasts that take the features X into account. In the precise case, the unconditional and the conditional viewpoints interact nicely due to additivity of the expectation, and it is possible to derive that for a weakly proper scoring rule (with propriety defined in the constant case) an optimal forecast is $f^*(x) := P(\cdot|X = x)$ when $P(X = x) > 0$, under a P data model. In a sense, this forecast is exactly tuned to the underlying data model. In the case of an imprecise probability $\mathcal{P}$, it is not immediately clear what the corresponding “conditional” forecast is. Thus we now focus on the IP score objective in Equation 6.3; we recall:

$$\text{IP-Score}^*(\ell, \mathcal{P}) := \min_{Q: \mathcal{X} \rightarrow \mathcal{IP}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[S_\ell(Q(X(\omega)), \omega)].$$

This is our notion of a “Bayes risk” for imprecise forecasts, as it quantifies the inherent difficulty of the task. We now show that under a suitable assumption on ℓ and $\mathcal{P}$ jointly, we can restrict ourselves to optimizing over only precise forecasts and indeed then characterize this optimum. For trivial (constant) X , this yields a reduction of imprecise to precise forecasts for the unconditional case of Section 6.2.1.

The key will be to leverage the *maximum entropy* bridge of Grünwald and Dawid [2004], which we now work towards.

Definition 292. For an arbitrary loss $\ell \in \mathcal{L}$, we can define a generalized conditional entropy of a probability $P \in \Delta^k$ as

$$H_\ell(P|X) := \mathbb{E}_P[S_\ell(P(\cdot|X), \omega)], \quad P \in \Delta^k.$$

(Compare also [Xu, 2020]).

Intuitively, this quantity measures the variability in P (as measured by the difficulty of forecasting when loss is measured by ℓ) when allowing the forecaster to see the value of X . When ℓ is the log-loss, this quantity becomes the familiar conditional Shannon entropy.¹⁴ In the decision-theoretic view, the entropy concept expresses the inherent difficulty of a task: *entropy is Bayes risk* [Osterreicher and Vajda, 1993, Williamson and Cranko, 2024]. In this way, $H_\ell(P|X)$ is a conditional version of the decision-theoretic entropy of Dawid and Sebastiani [1999]. It turns out that, under a subtle condition, an optimal forecast in the imprecise case is simply an optimal precise forecast for the “most difficult” $P \in \text{c}\bar{\text{co}}(\mathcal{P})$.

¹⁴ Due to our finiteness assumption on $\mathcal{A}$ we cannot exactly recover the log-loss in our framework, but approximately. Note that usually conditional (Shannon) entropy is defined for random variables, i.e. $H(Y|X)$, and not with a probability measure as $H(P|X)$, but the relation is straightforward.

Proposition 293. Let $\ell \in \mathcal{L}$ and $\mathcal{P} \in \mathcal{IP}$. Compute a maximum entropy probability over $\bar{\text{co}}(\mathcal{P})$:¹⁵

$$P^* \in \operatorname{argmax}_{P \in \bar{\text{co}}(\mathcal{P})} H_\ell(P|X). \quad (6.5)$$

If $P^*(X = x) > 0 \forall x \in \mathcal{X}$ and $\operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}_{P^*(\cdot|X=x)}[\ell(a, \omega)]$ is a singleton for each $X = x$ (the action recommendation is unique), then the forecast $f^*(x) := P^*(\cdot|X = x)$ is optimal up to weak inequality:

$$\text{IP-Score}^*(\ell, \mathcal{P}) = \bar{R}_\mathcal{P}(S_\ell(f^*(X), \omega)) = \mathbb{E}_{X, \omega \sim P^*}[S_\ell(f^*(X), \omega)].$$

The proof of this result, which draws on the maximum entropy principle of Grünwald and Dawid [2004], is in Appendix 6.7.5. When $\underline{P}_\mathcal{P}(X = x) > 0 \forall x \in \mathcal{X}$, then also $P^*(X = x) > 0 \forall x \in \mathcal{X}$. We emphasize that when computing the maximum entropy probability, it is crucial to consider the closed convex hull of $\mathcal{P}$; indeed P^* can lie in the interior of $\bar{\text{co}}(\mathcal{P})$. If the maximum entropy probability does not provide unique action recommendations, it can be that

$$\mathbb{E}_{P^*}[S_\ell(f^*(X), \omega)] < \text{IP-Score}^*(\ell, \mathcal{P}) < \bar{R}_\mathcal{P}(S_\ell(f^*(X), \omega)),$$

(for an example see the proof of (290.2)).

To compare to the precise case: under an *i.i.d.* data model with probability P , an optimal forecast is $f^*(x) = P(\cdot|X = x)$ under any ℓ (due to the “automatic” weak propriety of S_ℓ), which would amount to the action recommendation $a^*(x) = f^*(x)$ if we had $\mathcal{A} = \Delta^k$ and ℓ were already a strictly proper scoring rule.¹⁶ Similarly, in the imprecise case, the optimal forecast of the maximum entropy principle corresponds to one of the probabilities $P \in \mathcal{P}$ in the same way, but ℓ determines which $P \in \mathcal{P}$ we pick. The condition that the action recommendations are unique is subtle, however. The proof of (290.2) already includes such an example. We provide another example to illustrate this failure of uniqueness in Appendix 6.7.6, meaning we cannot characterize the optimal forecast using Proposition 293.

We remark that when taking X to be trivial, $X(\omega) = c \forall \omega \in \Omega$, the findings here apply also to the unconditional case of Section 6.2.1. In this case, the generalized conditional entropy of Definition 292 becomes a generalized unconditional entropy $H_\ell(P) := \min_{a \in \mathcal{A}} \mathbb{E}_P[\ell(a, \omega)]$.

Remark 294 (Connection to standard setup and intuition for uniqueness condition). Usually a loss function would depend on ω only through $Y(\omega)$, and a forecast would correspond to a probability on $\mathcal{Y}$, not on $\Omega = \mathcal{X} \times \mathcal{Y}$ as in our case. For example, for binary

¹⁵Any P^* in the arg max will do, but there is no general uniqueness guarantee.

¹⁶In this paper, we cannot have ℓ be a proper precise scoring rule due to $|\mathcal{A}| < \infty$. See 14 and Remark 294.

$\mathcal{Y} = \{0, 1\}$ the log-loss is

$$\ell_{\log} : \Delta^2 \times \mathcal{Y} \rightarrow \mathbb{R}, \quad \ell_{\log}(q, y) := -y \log(q) - (1 - y) \log(1 - q).$$

The log-loss is strictly proper in the sense of a precise scoring rule $s : \Delta^{|\mathcal{Y}|} \times \mathcal{Y} \rightarrow \mathbb{R}$, meaning

$$\mathbb{E}_P[s(P, Y)] \leq \mathbb{E}_P[s(Q, Y)],$$

for $P, Q \in \Delta^{|\mathcal{Y}|}$, with equality only if $P = Q$. Note here the possibly confusing terminology: a tailored *scoring rule* S_ℓ depends on an underlying *loss function* ℓ , but when the action space is the simplex itself and ℓ formally fulfills strict propriety, then $S_\ell = \ell$ because $a_Q^* = Q$ [Dawid and Musio, 2014]. In this paper, we stick to finite action spaces since it simplifies some aspects of the mathematics and we believe that from a conceptual viewpoint there is no loss of generality. But if ℓ were a strictly proper scoring rule like the log-loss, then the uniqueness condition “ $\operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}_{P^*(\cdot|X=x)}[\ell(a, \omega)]$ is a singleton for each $X = x$ ” in Proposition 293 would be satisfied, since then only $P^*(\cdot|X = x)$ would be in the argmin. Since strict propriety of ℓ implies that $a \mapsto \ell(a, y)$ is continuous on the interior of $\Delta^{|\mathcal{Y}|}$ for any fixed $y \in \mathcal{Y}$ [Mhammedi and Williamson, 2018, Lemma 26], we can then safely approximate $\Delta^{|\mathcal{Y}|}$ by a finite discretization in practice (e.g. the set of floats in a computer) and apply Proposition 293.

THE LINK TO DISTRIBUTIONAL ROBUSTNESS

Observe that under the condition in Proposition 293, the problem of finding an optimal forecast is reduced to optimizing over only precise forecasts:

$$\min_{Q: \mathcal{X} \rightarrow \mathcal{IP}} \bar{R}_{\mathcal{P}}(S_\ell(Q(X), \omega)) = \min_{Q: \mathcal{X} \rightarrow \Delta^k} \bar{R}_{\mathcal{P}}(S_\ell(Q(X), \omega)).$$

This means that, under the uniqueness condition, our IP scoring rule objective turns out to be approximately equivalent to the objective considered in distributionally robust optimization (DRO), well-known in the machine learning community [Sagawa et al., 2019, Duchi and Namkoong, 2021, Fröhlich and Williamson, 2024a]. To our knowledge, the connection between DRO and the maximum entropy principle has not yet been laid out in the machine learning literature.

What still distinguishes the objective here from standard DRO is that we use the tailored scoring rule S_ℓ . We might call this *tailored DRO*, and our results in Section 6.4 highlight the importance of the choice of loss function in DRO. Remark 294 tells us that if we use a strictly proper ℓ (in the precise sense) and a finite discretization of the simplex as our action space, then we can effectively apply Proposition 293 to obtain the optimal forecast. *Prima facie* it seems unsettling that for optimal performance we do not need imprecision in the

forecasts — what is more fundamental is the imprecision in the data model, expressed by $\bar{R}_{\mathcal{P}}$. Observe, however, that the above statement holds for a fixed ℓ . We do not always know at training time the loss function of the agent, or we may want to enable decision making for multiple agents with different loss functions. Yet the minimizer in the DRO objective is tailored to the exact loss function, which at training time is usually a convenient surrogate loss function. In summary, we highlight that the choice of loss function matters in DRO, since this determines the maximum entropy probability.

6.2.3 A VARIETY OF ENTROPIES

Recall the idea that entropy is Bayes risk (e.g. [Williamson and Cranko, 2024]). The IP score, our generalized Bayes risk, can be considered as a generalized conditional entropy notion for imprecise probabilities, which we might write as $\bar{H}_{\ell}(\mathcal{P}|X) := \text{IP-Score}^*(\ell, \mathcal{P})$, where X is implicit on the right side. We may call it the *imprecise conditional ℓ -entropy* of $\mathcal{P}$. In contrast, the quantity $\text{MaxEnt} := \max_{P \in \text{co}(\mathcal{P})} H_{\ell}(P|X)$, featuring in Equation 6.5, could be called the *maximum conditional ℓ -entropy*. In the unconditional case, this has been indeed been proposed as a generalized entropy for imprecise probabilities [Abellán and Moral, 2003, 2005, Shaker and Hüllermeier, 2021]. Importantly, we have seen that possibly $\text{MaxEnt} < \bar{H}_{\ell}(\mathcal{P}|X)$, although they sometimes coincide. Our interpretation of the situation is that the IP score $\bar{H}_{\ell}(\mathcal{P}|X)$ is more suited as a generalization of the entropy concept, as it quantifies the inherent difficulty of the prediction task. For clarification of these concepts, we specialize the insight of Proposition 293 to the unconditional case of constant forecasts.

Def./Observation 295. Let $\mathcal{P} \in \mathcal{IP}$. In analogy to $\bar{H}_{\ell}(\mathcal{P}|X)$, we call $\bar{H}_{\ell}(\mathcal{P}) := \bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{P}, \omega))$ the *imprecise ℓ -entropy* of $\mathcal{P}$. Observe that it corresponds to the IP score of the optimal constant forecast under a $\mathcal{P}$ data model due to weak propriety (Proposition 6.4). Let $P^* \in \arg \max_{P \in \text{co}(\mathcal{P})} \mathbb{E}_P[\ell(a_P^*, \omega)]$ be a maximum entropy probability. If it recommends a unique action, meaning $\{a^*\} = \arg \min_{a \in \mathcal{A}} \mathbb{E}_{P^*}[\ell(a, \omega)]$ is a singleton, then $\bar{H}_{\ell}(\mathcal{P}) = \bar{R}_{\mathcal{P}}(S_{\ell}(P^*, \omega)) = \mathbb{E}_{P^*}[\ell(a^*, \omega)]$.

We view the unconditional quantity as a generalization of the decision-theoretic entropy of Dawid and Sebastiani [1999] to the imprecise case.

6.2.4 THE GENERALIZED BAYES RULE

The optimal forecast f^* of Proposition 293, if the condition holds, picks exactly one $P \in \mathcal{P}$ and conditions it. A more conservative, intuitive approach would be to simply forecast the whole set of conditionals for a given $X = x$. This is effectively a generalization of Bayes rule to a set of probabilities, and can be seen as conditioning the data model. As a prerequisite for Section 6.3, we first define it for the general case of an arbitrary partition $\mathcal{B}$

of Ω and then specialize. For a set $B \subseteq \Omega$, we denote its indicator function by $\chi_B(\omega) := 1$ if $\omega \in B$, and 0 otherwise.

Definition 296 ([Walley, 1991, Chapter 6]). Let $\mathcal{P} \in \mathcal{IP}$. For a partition $\mathcal{B}$ of Ω , where $\underline{P}_{\mathcal{P}}(B) > 0 \forall B \in \mathcal{B}$, define

$$\mathcal{P}_{|B} := \{P(\cdot|B) : P \in \mathcal{P}\}, \quad \text{GBR}_{\mathcal{P}}(Z|B) := \bar{R}_{\mathcal{P}_{|B}}(Z), \quad B \in \mathcal{B}, Z \in L^{\infty}.$$

It follows that $\text{GBR}_{\mathcal{M}}(Z|B) = \min\{\alpha \in \mathbb{R} : \bar{R}_{\mathcal{P}}(\chi_B(Z - \alpha)) = 0\}$. Due to the 1:1 relation, we refer to either $\mathcal{P}_{|B}$ or $\text{GBR}_{\mathcal{P}}(\cdot|B)$ as the *generalized Bayes rule (GBR)*.

We will be mostly interested in the special case where the partition is induced by the features X . In this case we obtain:

Definition 297. Let $\mathcal{P} \in \mathcal{IP}$. Define, if $\underline{P}(X = x) > 0 \forall x \in \mathcal{X}$:

$$\mathcal{P}_{|x} := \{P(\cdot|X = x) : P \in \mathcal{P}\}, \quad \text{GBR}_{\mathcal{P}}(Z|x) = \bar{R}_{\mathcal{P}_{|x}}(Z), \quad Z \in L^{\infty}.$$

The associated forecast is

$$\mathcal{P}_{|X}(\omega) := \mathcal{P}_{|x} \text{ if } X(\omega) = x.$$

We refer to $\mathcal{P}_{|X}$ as the *generalized Bayes rule (GBR) forecast*.

While this forecast has an intuitive appeal, one can observe by example that it need not be optimal in terms of IP score (see Appendix 6.7.6). Since $f^*(x) \in \mathcal{P}_{|x}$ (f^* of Proposition 293), the GBR is a more conservative (pessimistic) forecast than f^* . However, as we now shift the focus to a distinct goal in decision making, we will see what the strength of the GBR is, making it attractive again. In the next section, we present a general account of *calibration* in terms of entropy, which applies to both the precise and the imprecise case, and then relate it to familiar notions of calibration.

6.3 IP CALIBRATION

In a recent paper, Fröhlich and Williamson [2024c] put forward the thesis that there are strong conceptual parallels between *insurance* and fair machine learning. While the authors focused on fairness and its interplay with uncertainty, their arguments suggest that the relevance of insurance as an analogy extends beyond fairness to issues of uncertainty in general. This heuristic motivates us to consider decision making with imprecise forecasts in an abstract insurance context, in order to arrive at a notion of calibration for imprecise forecasts. More generally, we are motivated here by the game-theoretic account of probability (see Section 6.5).

6.3.1 CALIBRATION AS ACTUARIALLY FAIR INSURANCE

What was lacking in the previous section was a justification for why the agent might *trust* the forecasts. So far, we have assumed that the agent must choose an action based on the forecasts without any guarantees for the quality of the forecasts. As such, the burden of uncertainty lies fully on the agent. Insurance manages uncertainty by pooling losses, which in our case means shifting this burden to the forecaster (the insurer), whereas the agent who repeatedly makes decisions (or a group of distinct agents) corresponds to the insurance pool of policyholders. To instantiate the analogy, when the agent has chosen the action a_Q^* as recommended by the forecaster, the loss to be incurred is $\ell(a_Q^*, \omega)$, and this is the uncertain claim that the forecaster insures. Putting aside normative questions of solidarity (risk subsidizing), a standard desideratum for insurance is *actuarial fairness* [Arrow, 1963].

Desideratum 298. *The forecaster should offer actuarially fair insurance for the uncertain loss $\tilde{S}_\ell(\omega) = \ell(a_{Q(X(\omega))}^*, \omega)$. Actuarial fairness here means that in the long run, the forecaster neither loses nor profits from offering insurance under the data model.*

In ML practice, the burden of uncertainty will still usually lie on the agent; but if we know that forecasts are actuarially fair, there is a certain adequateness to their associated uncertainty. Within the ML research community, this philosophy aligns with Yoo and Kweon [2019] and Kirchhof et al. [2024], who ground the quality of uncertainty estimates in accurate loss prediction. For fixed $\omega \in \Omega$, consider the decomposition

$$\tilde{S}_\ell(\omega) = \left(\tilde{S}_\ell(\omega) - \bar{R}_{Q(\omega)}(\omega' \mapsto \tilde{S}_\ell(\omega')) \right) + \bar{R}_{Q(\omega)}(\omega' \mapsto \tilde{S}_\ell(\omega')),$$

where we highlight the dependence on ω vs. ω' . A gamble of the form $Z - \bar{R}(Z)$, $Z \in L^\infty$, is known as a *marginally desirable gamble* in the IP literature [Augustin et al., 2014], since a forecaster with subjective upper expectation $\bar{R}$ should be indifferent to it. Such gambles can be used to evaluate forecasts, an idea discussed in depth by Derr and Williamson [2024]. In our case, the forecaster will shoulder the uncertain gamble $\tilde{S}_\ell(\omega) - \bar{R}_{Q(\omega)}(\tilde{S}_\ell)$ and the agent will shoulder the constant¹⁷ loss $\bar{R}_{Q(\omega)}(\tilde{S}_\ell)$, which is the price for the insurance as specified by the forecaster. The quantity $\bar{R}_{Q(\omega)}(\tilde{S}_\ell)$ represents a generalization of the entropy concept to the imprecise case (recall Section 6.2.3), quantifying the uncertainty inherent in the task, with evaluation by ℓ , from the subjective viewpoint of the forecaster. If the forecaster believes their own forecast in the sense that if $X(\omega) = x$, then $\underline{P}_{Q(\omega)}(X = x) = 1$, then indeed $\bar{R}_{Q(\omega)}(\tilde{S}_\ell) = \bar{H}_\ell(Q(\omega))$ is the imprecise ℓ -entropy of $Q(\omega)$.¹⁸

¹⁷By *constant* loss, we here mean that the price of insurance can be specified after learning $X = x$, which the agent then shoulders, but before all information in ω is revealed.

¹⁸The reason is that under this condition, the forecast is constant according to $\bar{R}_{Q(\omega)}$, and then the quantity $\bar{R}_{Q(\omega)}(\tilde{S}_\ell)$ is of the form $\bar{R}_Q(S_\ell(Q, \omega))$ for $Q \in \mathcal{IP}$.

In the transaction above, the agent desires to have as little constant loss as possible (cheap insurance), whereas the forecaster wants to avoid bankruptcy. The tension is resolved by the principle of actuarial fairness. Fröhlich and Williamson [2024c] have demonstrated a tight relationship between actuarial fairness and calibration. We now use the term (IP) *calibration*, but it might be equivalently called (IP) *actuarial fairness*.

6.3.2 DEFINING IP CALIBRATION

Definition 299 (IP calibration without groups). *Let $\ell \in \mathcal{L}, \mathcal{P} \in \mathcal{IP}$. The forecast $\mathcal{Q}$ is ℓ -calibrated (without groups) under a $\mathcal{P}$ data model if, with $\tilde{S}_\ell(\omega) = S(\mathcal{Q}(\omega), \omega)$:*

$$\bar{R}_{\mathcal{P}} \left(\tilde{S}_\ell(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\tilde{S}_\ell) \right) = 0. \quad (6.6)$$

We say that $\mathcal{Q}$ is sub-calibrated if Equation 6.6 holds with a “ $\leq$ ” inequality.

Throughout this section, we continue to assume that the data model is parameterized by $\mathcal{P}$. Intuitively, calibration as a criterion is about a matching of the amount of loss to be expected *under the data model* and the entropy *with respect to the forecaster*. A calibrated forecaster makes adequate entropy assessments, hence accurately predicts the loss to be incurred. This criterion in itself is rather weak. For instance, the constant forecast $\mathcal{P}$ clearly achieves this. Central in insurance, particularly in debates revolving around fairness, is the question of forming groups. An individual is then treated as equivalent to the “average human” of their group [Krippner and Hirschman, 2022]. In the above definition, there is only a single group. The condition can be strengthened by introducing a partition into groups.

We first illustrate this logic with a highly relevant example, the partition that is induced by the action recommendations. Assume the forecaster announces $\mathcal{Q}(x)$, based on which the agent chooses the action $a_{\mathcal{Q}(x)}^*$. The agent would now like to know which loss they get in terms of worst-case expected loss under the data model, conditional on this action recommendation; then forecasts are trustworthy when their uncertainty estimates are actuarially fair. Formally, the action-induced partition is

$$\mathcal{B}_{\mathcal{Q}, \mathcal{A}} := \bigcup_{a \in \mathcal{A}} B_{\mathcal{Q}, a}; \quad B_{\mathcal{Q}, a} := \{\omega \in \Omega : a_{\mathcal{Q}(\omega)}^* = a\}.$$

This leads us to the following definition of IP calibration with respect to a partition.

Definition 300 (IP calibration with groups). *Let $\ell \in \mathcal{L}, \mathcal{P} \in \mathcal{IP}$ and $\mathcal{B}$ be a partition of Ω . We say that a forecast $\mathcal{Q}$ is $(\ell, \mathcal{B})$ -calibrated under a $\mathcal{P}$ data model if, with $\tilde{S}_\ell(\omega) = S_\ell(\mathcal{Q}(\omega), \omega)$:*

$$\forall B \in \mathcal{B} : \bar{R}_{\mathcal{P}} \left(\chi_B(\omega) \left(\tilde{S}_\ell(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\tilde{S}_\ell) \right) \right) = 0.$$

We say that $\mathcal{Q}$ is $(\ell, \mathcal{B})$ -sub-calibrated if the above holds with a “ $\leq$ ” inequality.

In comparison to the formulation without groups, we now demand adequate entropy assessments *conditionally*. IP calibration has a useful structural property. We say that a partition $\mathcal{B}_1$ is coarser than $\mathcal{B}_2$ if every set in $\mathcal{B}_2$ is a subset of some set in $\mathcal{B}_1$ (e.g. [Höltgen and Williamson, 2023]). We then have the following implication.

Proposition 301. *Let $\ell \in \mathcal{L}$ and $\mathcal{B}_1, \mathcal{B}_2$ be partitions of Ω , with $\mathcal{B}_1$ coarser than $\mathcal{B}_2$. Then, if $\mathcal{Q}$ is $(\ell, \mathcal{B}_2)$ -sub-calibrated, it is $(\ell, \mathcal{B}_1)$ -sub-calibrated. In particular, $(\ell, \mathcal{B}_2)$ -calibration implies $(\ell, \mathcal{B}_1)$ -sub-calibration.*

The proof is in Appendix 6.7.7. Why care about sub-calibration? Due to the loss orientation of the scoring rule S_ℓ , our situation is an asymmetric one; we choose to prioritize *not underestimating* the loss, which aligns with the goal of deploying robust and safe machine learning systems. In insurantal terms this would mean ensuring that the price for insurance is at least sufficient to cover the losses, but potentially more pessimistic.

Since calibration embodies trustworthiness of uncertainty estimates, it appears as an important goal. Yet we observe that it can diverge from the goal of Section 6.2, i.e. recommending actions with favorable consequences.

Proposition 302. *There exist $\ell \in \mathcal{L}$ and $\mathcal{P} \in \mathcal{IP}$ so that the condition of Proposition 293 applies, and the optimal forecast f^* of Proposition 293 fails to be ℓ -sub-calibrated and therefore also fails to be $(\ell, \mathcal{B}_{f^*, A})$ -sub-calibrated.*

As a proof, see the example in Appendix 6.7.6, where this failure is illustrated both in an unconditional (without features) fashion, as well as in a conditional (with features) fashion. Note that a failure of sub-calibration implies a failure of calibration. We remark that it can be that $\mathcal{Q} \neq \mathcal{Q}'$ are both optimal in Equation 6.3 (e.g. if they recommend the same actions), but one of them is calibrated and the other is not (see Appendix 6.7.6). This negative result means that the optimal forecast can be overly optimistic, as it underestimates the loss to be incurred on some $P \in \mathcal{IP}$, which can happen on both the unconditional level (without groups) and the conditional (group) level. In other words, the price for insurance is too low, leading to long-run bankruptcy.

6.3.3 CALIBRATION AND THE GBR

On the other hand, the GBR has an attractive calibration guarantee, as it is more conservative. First, we write the calibration condition in terms of the GBR itself.

Remark 303. *With $\tilde{S}_\ell(\omega) = S_\ell(\mathcal{Q}(\omega), \omega)$, the IP calibration condition in Definition 300 can be equivalently rewritten as*

$$\forall B \in \mathcal{B} : \text{GBR}_{\mathcal{P}}(\tilde{S}_\ell(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\tilde{S}_\ell) | B) = 0.$$

Proof From $\text{GBR}_{\mathcal{P}}(Z|B) = \min\{\alpha \in \mathbb{R} : \bar{R}_{\mathcal{P}}(\chi_B(Z - \alpha)) = 0\}$, it follows that, with $Y(\omega) := \tilde{S}_{\ell}(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\tilde{S}_{\ell})$, $\text{GBR}(Y|B) = 0 \Leftrightarrow \bar{R}_{\mathcal{P}}(\chi_B(\omega)Y(\omega)) = 0$, which is the definition of calibration with groups. ■

Rewriting the calibration criterion in terms of the GBR suggests a tight conceptual connection. Comparing two forecasts is particularly of interest when they “get to see the same information”, intuitively; in our context, this means usually the features X but we generalize.

Definition 304. Let $\mathcal{B}$ be a partition of Ω and $\mathcal{Q}_{\mathcal{B}} : \mathcal{X} \rightarrow \mathcal{IP}$ be a forecast. We say that $\mathcal{Q}_{\mathcal{B}}$ is $\mathcal{B}$ -measurable if it is constant on each $B \in \mathcal{B}$ (following terminology of Walley [1991, Section 6.2.5]). We then write $\mathcal{Q}_B \in \mathcal{IP}$ for the value it takes on $B \in \mathcal{B}$.

Using this definition, we find that, for any loss ℓ , the GBR is ℓ -calibrated and its entropy provides a lower bound to that of other calibrated forecasts. We write $\mathfrak{B}$ for the set of partitions of Ω with the property that each element of the partition has positive lower probability under the data model, meaning $\mathcal{B} \in \mathfrak{B}$ if $\underline{P}_{\mathcal{P}}(B) > 0 \forall B \in \mathcal{B}$.

Proposition 305. Let $\ell \in \mathcal{L}, \mathcal{P} \in \mathcal{IP}$ and $\mathcal{B} \in \mathfrak{B}$ a partition of Ω . Then the GBR forecast is $(\ell, \mathcal{B})$ -calibrated:

$$\forall B \in \mathcal{B} : \text{GBR}_{\mathcal{P}}(S_{\ell}(\mathcal{P}_{|B}, \omega) - \text{GBR}_{\mathcal{P}}(S_{\ell}(\mathcal{P}_{|B}, \omega)|B) | B) = 0.$$

Moreover, if $\mathcal{Q}_{\mathcal{B}}$ is any other $\mathcal{B}$ -measurable forecast which is $(\ell, \mathcal{B})$ -calibrated, then

$$\forall B \in \mathcal{B} : \bar{H}_{\ell}(\mathcal{P}_{|B}) = \text{GBR}_{\mathcal{P}}(S_{\ell}(\mathcal{P}_{|B}, \omega)|B) \leq \bar{R}_{\mathcal{Q}_{\mathcal{B}}}(S(\mathcal{Q}_{\mathcal{B}}, \omega)) = \bar{H}_{\ell}(\mathcal{Q}_{\mathcal{B}}).$$

In words, for each $B \in \mathcal{B}$ the imprecise ℓ -entropy of the “conditional” data model parameterized by $\mathcal{P}_{|B}$ is a lower bound for the imprecise ℓ -entropy of any calibrated competitor forecast $\mathcal{Q}_{\mathcal{B}}$. In addition, there exist $(\ell, \mathcal{P}, \mathcal{B}, \mathcal{Q}_{\mathcal{B}})$ satisfying the assumptions so that the inequality in Equation 305 is strict.¹⁹

The proof is in Appendix 6.7.8. Elaborating on this result, if a forecast is calibrated, we can use its entropy to get a bound on the IP score.

Proposition 306. Let $\ell \in \mathcal{L}, \mathcal{P} \in \mathcal{IP}$ and $\mathcal{B} \in \mathfrak{B}$ a partition of Ω . Let $\mathcal{Q}_{\mathcal{B}}$ be a $\mathcal{B}$ -measurable and $\mathcal{B}$ -calibrated forecast. Then

$$\bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{Q}_{\mathcal{B}}(X), \omega)) \leq \sum_{B \in \mathcal{B}} H_{\ell}(\mathcal{Q}_B) \cdot \bar{R}_{\mathcal{P}}(\chi_B).$$

¹⁹Meaning that there exists a pair $(\ell, \mathcal{P}, \mathcal{B}, \mathcal{Q}_{\mathcal{B}})$, so that the forecast $\mathcal{Q}_{\mathcal{B}}$ is also $(\ell, \mathcal{B})$ -calibrated but $\bar{H}_{\ell}(\mathcal{P}_{|B}) < \bar{H}_{\ell}(\mathcal{Q}_{\mathcal{B}}) \forall B \in \mathcal{B}$.

The proof is in Appendix 6.7.9.

For a fixed loss ℓ , a distinct ($\neq$) forecast could yield the same entropy assessments as the GBR. When considering all losses, however, the GBR has a nice uniqueness guarantee.

Proposition 307. *Let $\ell \in \mathcal{L}$, $\mathcal{P} \in \mathcal{IP}$ and $\mathcal{B} \in \mathfrak{B}$ a partition of Ω . Let $\mathcal{Q}_{\mathcal{B}}$ be any $\mathcal{B}$ -measurable forecast. Assume without loss of generality that $\mathcal{Q}_{\mathcal{B}}$ and $\mathcal{P}_{|\mathcal{B}}$ are closed convex $\forall B \in \mathcal{B}$. Then*

1. *(More optimistic than GBR): If $\mathcal{Q}_{\mathcal{B}} \subseteq \mathcal{P}_{|\mathcal{B}} \forall B$ and $\mathcal{Q}_{\mathcal{B}} \subsetneq \mathcal{P}_{|\mathcal{B}}$ for at least one B , then $\mathcal{Q}_{\mathcal{B}}$ cannot be $(\ell, \mathcal{B})$ -sub-calibrated for all loss functions.*
2. *(More pessimistic than GBR): If $\mathcal{P}_{|\mathcal{B}} \subseteq \mathcal{Q}_{\mathcal{B}} \forall B$, then $\mathcal{Q}_{\mathcal{B}}$ is $(\ell, \mathcal{B})$ -sub-calibrated.*
3. *(Different to GBR): If $\exists B \in \mathcal{B} : \mathcal{P}_{|\mathcal{B}} \neq \mathcal{Q}_{\mathcal{B}}$, then $\mathcal{Q}$ cannot be $(\ell, \mathcal{B})$ -calibrated for all loss functions.*

The proof, based on the general monotonicity property that if $\mathcal{Q} \subseteq \mathcal{Q}'$, then $\bar{R}_{\mathcal{Q}}(Z) \leq \bar{R}_{\mathcal{Q}'}(Z)$ for any $Z \in L^{\infty}$, is in Appendix 6.7.10.

The take-away of this section is that calibration, which expresses trustworthiness of uncertainty estimates in terms of entropy, is a goal which is distinct to the goal of recommending actions with favorable consequences. In the classical, precise case, this distinction has been observed on a formal level in the decomposition of proper scoring rules [DeGroot and Fienberg, 1983, Bröcker, 2009], but the key difference is that these goals are aligned in the precise case (the Bayes classifier is the optimal forecast and it is calibrated), but they can diverge in the imprecise case. In particular, the optimal forecast need not be calibrated.

6.4 EXPERIMENTS

In the present paper we have proposed a general conceptual framework for the evaluation of imprecise probabilistic forecasts, which opens up ample room for practical investigations and algorithmic developments in the future. Many questions arise (see the discussion in Section 6.6) which require more detailed attention than we can give here. Our focus lies therefore on illustrative experiments, with deliberately simple model architecture (i.e. logistic regression) to avoid confounding due to contingencies of the training process. Our focus lies on demonstrating theoretical insights in practice. Others works have already shown the usefulness of group DRO in deep learning practice ([Sagawa et al., 2019]). We here also stick to the group DRO setting as our imprecise data model due to its high relevance. Thus, to connect back to the presented hierarchy in Section 6.1.1, in this setting the imprecision is assumed to be fundamental to the data model, and thus does not express our epistemic uncertainty or risk aversion. Recall (Section 6.1.1) that the evaluation of imprecise forecasts may also be studied under other (possibly even a precise) data

model. Our aim here is to showcase subtle potential pitfalls with (im)precise forecasts under group DRO. In particular, we illustrate how the choice of the loss function *can* matter in group DRO, and how the two key qualities of good forecasts, embodied by scoring rules and calibration, *can* diverge in the group DRO setting; our experiments should be interpreted with an existential quantifier instead of a universal quantifier in mind. Code for reproducing our experiments can be found at https://github.com/froec/ip_scoring_rules_experiments.

6.4.1 DATA AND DECISION PROBLEMS

Inspired by Ding et al. [2021], the first dataset that we use, ACS PUMS, stems from the american community survey. In our variant, the binary classification task is to predict whether a person has a job ($Y = 1$), or is unemployed/not on the labor market ($Y = 0$), based on demographic information about the person (their age, marital status, gender etc.). We form the groups based on membership to the 50 U.S. states. The sample size across groups is highly uneven, with the largest group California having 302,640 examples²⁰, contrasting with the smallest, Wyoming, having only 4,552 examples. For each group, we have a training as well as a test set. In the group DRO setting, we keep the data from these groups (states) separate; in the empirical risk minimization (ERM) setting, we stack the data of the groups.

The other datasets share the same structure, each paired with a binary classification task ($Y = 0$ or $Y = 1$). The next dataset we use is FRAMINGHAM, where the goal is to predict 10-year heart disease risk based on information such as age, blood pressure, having diabetes etc.; We form two groups based on thresholding age at 60. Finally, loosely inspired by [Sagawa et al., 2019], on CELEBA, we use the embeddings of a pre-trained (on ImageNet) Resnet50 to detect whether a person wears an earring or not. We form two groups based on gender (“male” or “not male”, where the latter includes non-binary identities). For detailed information about the datasets see Appendix 6.7.12. In the main paper, we report results for ACS PUMS and shift results for the other datasets to Appendix 6.7.16.

Besides the data model, the second key ingredient in our framework is the emphasis on the decision-making context. In machine learning practice, it is common to choose a proper scoring rule as a *surrogate loss*, such as the log-loss (also known as cross-entropy loss) or Brier loss (squared loss) at training time. Crucially, these are continuous and differentiable which enables training, and they are moreover symmetric. Test performance is often summarized using accuracy on the test set. Accuracy is, up to a constant factor, a special case ($c = 0.5$) of the following family of loss functions:

²⁰We remark that a datum of the ACS PUMS dataset does not correspond to a single person, but rather to a sort of *statistical individual*.

Definition 308 (e.g. [Buja et al., 2005]²¹). The cost-sensitive loss ℓ_c with parameter $c \in (0, 1)$ and action space $\mathcal{A} := \{0, 1\}$ is defined by $\ell_c(0, Y(\omega) = 1) := 1 - c$, $\ell_c(1, Y(\omega) = 0) := c$, and 0 otherwise.

Importantly, these cost-sensitive losses form a “basis” of all binary proper scoring rules [Schervish, 1989]. The parameter c controls the asymmetry in the decision problem: often, a false positive is much more consequential than a false negative, or conversely. For example, in heart disease prediction, accuracy might be a poor measure of performance, and we might be inclined to choose a small c , which means we get high loss when falsely predicting “has no risk of heart disease”. While cost-sensitive losses are flexible, they are unfortunately hardly usable for training due to discontinuity: with only the binary action space $\mathcal{A} = \{0, 1\}$, a small change in the forecast can imply a big change in the resulting loss. We therefore use a similarly parameterized family of continuous surrogate losses at training time due to Winkler [1994], which we here specialize to use the Brier loss as the base.²²

Definition 309 ([Winkler, 1994]). With $\mathcal{A} := [0, 1]$ and $c \in (0, 1)$, an asymmetric, strictly proper scoring rule²³ can be defined as:

$$\ell_{W;c}(a, Y(\omega) = y) = 1 - \frac{\ell^2(a, y) - \ell^2(c, y)}{T(c, a)}; \quad T(c, a) := \begin{cases} -\ell^2(c, 1) & a \geq c \\ -\ell^2(c, 0) & a < c \end{cases}; \quad \ell^2(a, y) := (a - y)^2.$$

Note that for any $c \in (0, 1)$, $a \mapsto \ell_{W;c}(a, y)$ is continuous for fixed $y \in \{0, 1\}$, even at $a = c$. To improve optimizer behaviour near the point c of non-differentiability, in practice we replace the step function T by a very close sigmoid approximation $\tilde{T}$ (see Appendix 6.7.13).

Remark 310. In the theoretical framework we have presented, the focus was on the idealized case where we minimize over *all* forecasts $Q : \mathcal{X} \rightarrow \Delta^k$. This is of course infeasible in practice, and we thus use a constrained model class throughout. Concretely, in our experiments, we train simple models consisting of a linear layer followed by a sigmoid (the output of which is considered as a precise probabilistic forecast) and then a loss function. We denote the respective constrained entropy quantities as $\hat{H}_\ell$, and the (precise) model class as $\text{Lin} := \{Q : \mathcal{X} \rightarrow \Delta^2 : Q \text{ is linear followed by sigmoid}\}$. We restrict ourselves to

²¹What Buja et al. [2005] calls “cost-weighted misclassification losses” is the tailored scoring rule S_{ℓ_c} induced by ℓ_c .

²²We also remark that Buja et al. [2005] have proposed an asymmetric family of proper scoring rules based on the Beta distribution, with analytical form for some values of the asymmetry parameters.

²³Recall from Section 6.2.2 [Dawid and Musio, 2014] that if ℓ is a precise, strictly proper scoring rule with $\mathcal{A} = \Delta^{|\mathcal{Y}|}$ (here $\mathcal{A} = [0, 1]$ due to the binary setting), then $S_\ell = \ell$. That is, we could call ℓ either (proper) “loss” or (proper) “scoring rule” [Williamson and Cranko, 2023]. In our view, it is beneficial to generally distinguish the roles of ℓ and S_ℓ though.

Δ^2 since our loss functions will depend on ω only through Y and we now assume that the forecaster trusts the $X = x$ information.

To compare the associated decision problems, we plot the generalized unconditional entropies $\hat{H}_{\ell_c}$ and $\hat{H}_{\ell_{W;c}}$ (obtained by taking constant X in Definition 292). Both the entropies of ℓ_c and $\ell_{W;c}$ have their maximum at c ; in this way, $\ell_{W;c}$ can be viewed as a surrogate loss for ℓ_c (see Figure 6.3). We will therefore use $\ell_{W;c}$ at training time when we are actually interested in the test loss ℓ_c . We caution that $a \mapsto \ell_{W;c}(a, y)$ need not be convex. Since $\ell_{W;c}$ is strictly proper, Proposition 293 can be applied approximately (see Remark 294), and we can therefore follow the maximum entropy principle.

6.4.2 MODELS AND TRAINING

We now train different models on the ACS PUMS as well as the other datasets (Appendix 6.7.16). Again and throughout the paper, as the base of all our models we use a linear regression followed by a sigmoid which outputs a precise probabilistic forecast, followed by a loss function. This means that for the log-loss, we recover the familiar logistic regression, and otherwise we get asymmetric variants of it. First, we train a standard logistic regression model (**ERM(log)**), using the log-loss, on the full data, meaning the stacked data of all states. We also train group DRO models according to Proposition 293 for different loss functions. Our training follows the maximum entropy principle:

$$\max_{\lambda \in \Delta^G} \min_{Q \in \text{Lin}} \mathbb{E}_{\hat{P}_\lambda} [S_\ell(Q(X), Y)], \quad \hat{P}_\lambda := \sum_{i=1}^G \lambda_i \hat{P}_i.$$

where $\hat{P}_i$, $i = 1..G$, is the empirical distribution (training data) for the i -th group. Here, we express the closed convex hull as a weighting (by λ) of the extreme points. Our loss functions depend on ω only through Y , hence a forecast can be identified with a mapping $Q : \mathcal{X} \rightarrow \Delta^2$ for binary Y .

Our approach to training is naïve: at iteration $t = 1..N_{\text{outer}}$, our forecast Q_{θ_t} is parameterized by θ_t and the weighting is λ_t . We perform an inner minimization loop for N_{inner} number of iterations which updates θ_t . At the end of this, the gradient for λ_t is simply the vector of mean group losses, i.e. $\nabla \lambda_t = [\mathbb{E}_{\hat{P}_1} [S_\ell(Q_{\theta_t}(X), Y)], \dots]^\top$. To update λ_t so as to enforce $\lambda_t \in \Delta^k$ (unit L1 norm and positivity), we use exponentiated gradient descent [Kivinen and Warmuth, 1997, Wei et al., 2023]:

$$\lambda_{t+1} := \lambda_t \cdot \exp(-\eta \cdot \nabla \lambda_t), \tag{6.7}$$

$$\lambda_{t+1} := \lambda_{t+1} / \sum \lambda_{t+1}, \tag{6.8}$$

with some learning rate $\eta > 0$. In a DRO context, exponentiated gradient descent to op-

Predictor	$\ell_{0.1}$	$\ell_{0.3}$	$\ell_{0.7}$	$\ell_{0.9}$
GBR	0.0492	0.1259	0.1475	0.0614
DRO(0.1)	0.0447	0.1044	0.1343	0.0769
DRO(0.3)	0.0456	0.0998	0.1350	0.0748
DRO(0.7)	0.0473	0.1144	0.1242	0.0589
DRO(0.9)	0.0474	0.1139	0.1315	0.0574
ERM(log)	0.0466	0.1114	0.1329	0.0715
DRO(log)	0.0463	0.1071	0.1299	0.0637

Predictor	$\ell_{0.1}$	$\ell_{0.3}$	$\ell_{0.7}$	$\ell_{0.9}$
GBR	0.0504	0.1267	0.1440	0.0609
DRO(0.1)	0.0463	0.1054	0.1364	0.0779
DRO(0.3)	0.0474	0.1050	0.1349	0.0743
DRO(0.7)	0.0493	0.1146	0.1294	0.0593
DRO(0.9)	0.0493	0.1146	0.1310	0.0579
ERM(log)	0.0487	0.1120	0.1337	0.0709
DRO(log)	0.0476	0.1104	0.1331	0.0632

Table 6.1: Train (left) and test (right) IP Scores of different forecasts (rows), evaluated under different loss functions (columns) on ACS PUMS data.

timize the group weights has already been employed by [Wei et al. \[2023\]](#). Additional details on our procedure using batches can be found in Appendix 6.7.15. We train a DRO model for each of the losses $\ell \in \{\ell_{\log}, \ell_{W;0.1}, \ell_{W;0.3}, \ell_{W;0.7}, \ell_{W;0.9}\}$. We refer to the models as **DRO(log)** and **DRO(c)**, respectively.

For the **GBR**, we choose a naïve estimation method: we fit a logistic regressor, using the log-loss, for each of the groups, and then take the set of these precise forecasts as our imprecise forecast. Setting aside the constraint of the linear model class, this would recover in the limit of infinite data the group-wise conditional probabilities due to propriety of the log-loss.

6.4.3 EVALUATION VIA IP SCORE

We now subject these trained models to an evaluation by the losses $\{\ell_{0.1}, \ell_{0.3}, \ell_{0.7}, \ell_{0.9}\}$, where the action space is now $\mathcal{A} = \{0, 1\}$. Note the difference between *training* a model using some $\ell_{W;c}$ and *evaluating* by some $\ell_{c'}$, with possibly $c \neq c'$. Evaluation, in any case, means selecting the optimal action $a_{Q(X)}^*$ based on the evaluation loss function, which may not coincide with the loss function used at training time. However, Table 6.1 shows that indeed, $\ell_{W;c}$ is a reasonable surrogate loss to use in place of ℓ_c . Recall that the IP score of the forecast Q under a $\mathcal{P}$ data model is $\bar{R}_{\mathcal{P}}(S_{\ell}(Q(X(\omega)), \omega))$. The *train data model* is $\hat{\mathcal{P}} := \{\hat{P}_i : i = 1..G\}$, and similarly, the *test data model* $\hat{P}_{\text{test}}$ consists of the test sets (their empirical distributions) of the groups. Typically (this happens in most of our cases), the model with the best IP score under the training data model with ℓ_c is the **DRO(c)** model. Due to generalization gaps, under the test data model and $\ell_{c'}$, it sometimes happens that the best performing model is not the **DRO(c)** model (see Appendix 6.7.16). In some cases, the **GBR** in fact performs best in terms of IP score under the test data model.

We reiterate the key insight: for DRO, the choice of loss function can indeed matter in practice, in particular when we face highly asymmetric decision problems. The reason is that the shape of the entropy may depend on the loss function. As a consequence, we

cannot hope to achieve optimal downstream performance by simply training using e.g. log-loss and then later using these predictions in an asymmetric downstream task.

6.4.4 EVALUATION VIA IP CALIBRATION

In Section 6.3 we have discussed IP calibration as a quality of forecasts in the sense of trustworthiness, conceptually separate from achieving optimal IP scores. While the numerical differences in Table 6.1 may appear rather small (albeit potentially significant in high-stakes decision making), we obtain comparatively drastic differences when evaluating forecasts in terms of IP calibration. We evaluate IP calibration under both the training as well as the test data model. Recall the definition of ℓ -calibration without groups (Definition 299):

$$\bar{R}_{\mathcal{P}} \left(\tilde{\mathcal{S}}_{\ell}(\omega) - \bar{R}_{Q(\omega)}(\tilde{\mathcal{S}}_{\ell}) \right) = 0.$$

Moreover, since we are interested in the trustworthiness of the action recommendations ($a = 0$ or $a = 1$), we evaluate via $(\ell, \mathcal{B}_{Q, \mathcal{A}})$ -calibration:

$$\forall a \in \{0, 1\} : I_a := \bar{R}_{\mathcal{P}} \left(\chi_{A=a}(\omega) \left(\ell(a, \omega) - \bar{R}_{Q(\omega)}(\ell_a) \right) \right) = 0.$$

Note that if we look at the term I_a for some $a \in \mathcal{A}$, the magnitude of the term depends on how often a is recommended. But we might also care about trustworthiness with respect to actions that are not recommended often. So we suggest instead plotting as a diagnostic, for each $a \in \{0, 1\}$:

$$II_a := \text{GBR}_{\mathcal{P}} \left(\ell(a, \omega) - \bar{R}_{Q(\omega)}(\ell_a) \mid a_{Q(\omega)}^* = a \right). \quad (6.9)$$

Clearly, $I_a = 0 \Leftrightarrow II_a = 0$. A perfectly trustworthy forecast would satisfy the equality; otherwise, recall that in Section 6.3 we have argued for preferring sub-calibration (" ≤ 0 ") over the converse direction (" ≥ 0 "). This choice simply amounts to preferring pessimism over optimism, which in the context of high-stakes decision making may be warranted.

Since we do not have access to $\mathcal{P}$, we evaluate under the train data model $\hat{P}$ as well as the test data model $\hat{P}_{\text{test}}$. Figure 6.1 depicts the diagnostic results under both the train and test data model on ACS PUMS for both calibration without groups as well as the diagnostic in Equation 6.9 for the action-induced partition. For the other datasets, see Appendix 6.7.16.

We evaluate IP calibration under $\{\ell_{0.1}, \ell_{0.3}, \ell_{0.7}, \ell_{0.9}\}$. Some general observations: The **GBR** is sub-calibrated in almost all cases, whereas the other forecasts can be wildly overoptimistic (see in particular the recommendation " $a = 1$ " on CELEBA in Appendix 6.7.16). A general, approximate tendency is that the **DRO(c)** models for small c perform better (in terms of achieving sub-calibration) when evaluated under small c' (referring to evalua-

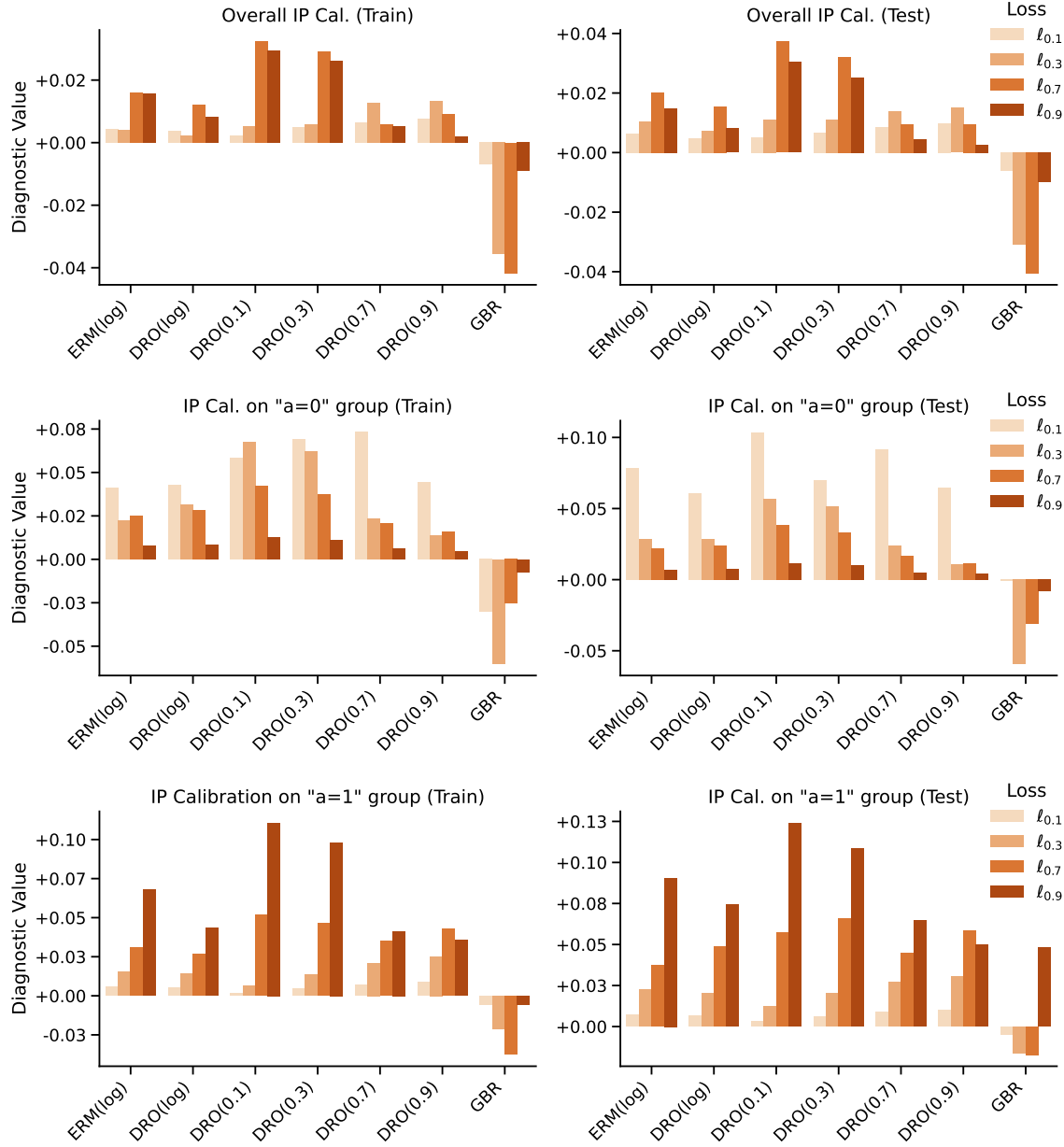


Figure 6.1: Evaluation of IP calibration under the train (left column) and test (right column) data models on ACS PUMS. In each subplot, the forecasts Q vary along the x -axis and the loss function used for evaluation corresponds to a hue. The top row shows diagnostic values for IP calibration without groups, that is, the y -axis shows $\bar{R}_{\hat{P}} \left(\tilde{S}_{\ell}(\omega) - \bar{R}_{Q(\omega)}(\tilde{S}_{\ell}) \right)$ (top left) and $\bar{R}_{\hat{P}_{\text{test}}} \left(\tilde{S}_{\ell}(\omega) - \bar{R}_{Q(\omega)}(\tilde{S}_{\ell}) \right)$ (top right). In the other subplots, the y -axis shows the diagnostics for IP calibration with respect to the action-induced partition (of Equation 6.9) under $\hat{P}$ (left) and $\hat{P}_{\text{test}}$ (right), respectively. We observe that the **GBR** forecast, in contrast to the other forecasts, is sub-calibrated in almost all cases (recall that we consider negative values on the y -axis as more desirable than positive values).

tion by $\ell_{c'}$) and worst for high c' , and conversely for **DRO(c)** models for large c . This observation highlights that rather than attempting to generalize the familiar concept of distribution calibration (see Appendix 6.7.11), which is agnostic of the decision problem, it was indeed meaningful to emphasize the primacy of the decision problem in the evaluation of uncertainty. For the **ERM(log)** and **DRO(log)** forecasts, it depends on the concrete problem whether they perform better when evaluated under small or large c' .

6.5 RELATED WORK

A highly relevant work to mention is that of [Zhao and Ermon \[2021\]](#), who are interested in mechanisms to provide trustworthy individual predictions. Like us, they are motivated by a betting interpretation of probability, citing [de Finetti \[1931\]](#), and also take an insurantal perspective. Although they make use of specific imprecise probabilities in the binary case, parameterized as $[\mu - c, \mu + c] \subseteq [0, 1]$, and corresponding marginally desirable gambles (see their Lemma 1) [[Augustin et al., 2014](#)],²⁴ they do not refer to literature on imprecise probabilities, or in particular the highly related recent developments on game-theoretic imprecise probabilities (see e.g. [[Cooman and De Bock, 2017](#), [De Cooman and De Bock, 2021, 2022](#), [Persiau et al., 2022](#)]). In the latter stream, which has been a major motivation to us, calibration appears in the form of *supermartingales*, built from marginally desirable gambles. In the case of the cost-sensitive loss, we can establish a close relation between our IP calibration and the framework of [Zhao and Ermon \[2021\]](#), see Appendix 6.7.17. Another question of interest is: what are the semantics of the imprecision in [Zhao and Ermon \[2021\]](#)? The paper uses a non-stationary, but precise probabilistic setup, where the probability may vary over time; this appears to fit with the group DRO setting. However, the authors go on to suggest that the semantics of the imprecision refers to the lack of knowledge (from the perspective of the forecaster) about the true, underlying probability. To add to the complexity, their definition of *asymptotic soundness* (essentially asymptotic calibration) fits with yet other frequentist semantics, studied by [Walley and Fine \[1982\]](#) and [Fröhlich et al. \[2024\]](#).²⁵ Our take-away is that we appeal to clarify which semantics one has in mind when working with imprecise probabilities.

Also in relation to this, [Zhao et al. \[2021a\]](#) have proposed the notion of decision calibration. We realized that our IP calibration can be seen as a generalization of decision calibration

²⁴The gambles of the form $b(Y - \mu) - |b|c$ are exactly the marginally desirable gambles for a binary imprecise probability parameterized as $[\mu - c, \mu + c]$; this is effectively what Lemma 1 of [Zhao and Ermon \[2021\]](#) states.

²⁵The conceptually puzzling situation can be cleared by consulting [Fröhlich and Williamson \[2024b\]](#) on two frequentist manifestations of imprecision; one is about hidden heterogeneity, and fits with the group distributional robustness perspective (where probability varies over time); the second is about instability of an aggregate, and fits with the definition of asymptotic soundness (the $\limsup$ is indicative of this), which the algorithm of [Zhao and Ermon \[2021\]](#) achieves. However, we note that neither semantics is, in our view, of epistemic character.

to the imprecise case. In our view (and our motivation stems from there), these concepts can be seen as grounded in the notion of marginal desirability and in game-theoretic probability [Shafer and Vovk, 2001, 2019] (see also [Augustin et al., 2014, Chapter 6]). In Appendix 6.7.11, we discuss the relations between IP calibration, decision calibration of Zhao et al. [2021a] and the familiar notion of calibration. For the relation between scoring rules and calibration, see also the recent proposal of *U-calibration* [Kleinberg et al., 2023].

In the literature there are various proposals for learning predictors which output imprecise probabilities,²⁶ e.g. [Corani and Zaffalon, 2008, Caprio et al., 2023, Nguyen et al., 2023, Wang et al., 2024, Hofman et al., 2024], with epistemic semantics for the imprecision in mind. For example, the imprecise forecast can be obtained by an ensemble (e.g. [Hofman et al., 2024]), or, in the subjectivist approach, by applying the generalized Bayes rule to an imprecise prior (e.g. [Corani and Zaffalon, 2008]). Following different routes and motivated by domain generalization issues, the recent papers by Caprio et al. [2023] and Singh et al. [2024] work with imprecise probabilities in a setting where data comes from a set of distributions. Yet in these works, the model class is a set of functions from $\mathcal{X}$ to $\mathcal{Y}$, so the imprecision is not located in the forecasts.²⁷ Also, these works employ evaluative criteria (specific generalization bounds and “C-pareto improvements”, respectively) which are conceptually distinct from ours. In light of these previous works, our hope is that the framework we have proposed here can aid the unification in the field.

6.6 DISCUSSION

In this work, we have proposed a framework for the evaluation of (im)precise forecasts under (im)precise data models based on a generalization of proper scoring rules and calibration.

A key insight, already appearing in the decomposition of DeGroot and Fienberg [1983], was the clear separation between two conceptually distinct qualities of good forecasts: first, recommending actions with favorable consequences (low loss values); second, accurately estimating the consequences (loss) to be expected when following the forecast. These two qualities, related to proper scoring rules and calibration, respectively, appear conflated in the precise case. We do not believe that their divergence in the imprecise case is pathological, but rather nicely reveals their distinct conceptual character. Even further, we observed that it is possible that two distinct forecasts are both optimal in the sense of the first goal, but only one of them is calibrated, meaning only this forecast is appropriately aware of

²⁶The final output is not necessarily an imprecise probability. For example, the classifier proposed by Corani and Zaffalon [2008] works with imprecise probabilities yet finally outputs a set-valued classification, i.e. a set of labels.

²⁷Of course a function $h : \mathcal{X} \rightarrow \mathcal{Y}$ can be seen as a degenerate imprecise forecast in our sense, outputting a precise probability with all the mass on a single $Y \in \mathcal{Y}$, which in turn can be seen as an imprecise probability.

“how well it is doing”.

In the opposite direction, we have investigated a phenomenon which can be disentangled under a precise data model, but not anymore under an imprecise data model. In the precise case, *any* proper scoring rule can be used to learn the “true” probabilities in the limit of infinite data. These precise forecasts can then be used for optimal decision making under *any* loss function (not necessarily proper) in the sense of minimizing expected loss. Whether data comes from a single, stable distribution in the real world so that the limit exists is another question [Fröhlich et al., 2024] which we put aside here, but at least in theory this shows that *belief* and *decision* (action) can be disentangled in the precise case. In contrast, we have seen in Proposition 293 that this breaks down in the imprecise case *even in theory* (putting aside the problem of finite data, which can only make matters worse). Again, we do not view this as pathological about the imprecise case, but rather take it as a fundamental lesson about the entanglement of belief and decision. To give one example, this observation is crucial when group DRO is applied in a machine learning context with fairness desiderata. Here, it is not generally possible to choose a “neutral” proper scoring rule at training time which is later optimal for downstream decision making in the sense of minimizing worst-case (over the groups) expected loss. This contrasts with proposals in the literature that urge to cleanly separate belief from decision [Beigang, 2022, Kuppler et al., 2022].

To sum it up: we have observed that two dualities (the qualities of forecasts and the belief-decision duality) behave surprisingly different in the imprecise case. In both cases, although in opposite directions, we have argued that the more general imprecise case actually better reflects underlying conceptual subtleties.

We believe that many opportunities for future research arise from the framework presented here. Some of them being very practical, for example how to estimate the generalized Bayes rule appropriately, or to address training/test generalization gaps. Another interesting, potentially challenging avenue would be to go beyond the MinMax assumptions. We highlight that due to the MinMax assumption in Equation 6.1, imprecision in our framework refers more to pessimism (caution) than to abstaining or indecision, which is a common interpretation in the literature, and which is accompanied by other criteria for action selection [Troffaes, 2007]. Hence some advocates of imprecise probabilities might be dissatisfied with our proposal. We are tentatively skeptical whether a similar mathematical development could be conducted with other decision criteria; perhaps the Hurwicz criterion (e.g. [Konek, 2019]) provides a starting point.

A fundamental difference to much of the literature is that here we have not focused on epistemic semantics of imprecision in combination with a precise data model, as in “the data is *i.i.d.* from P but we are ignorant about P ”. A proper scoring rule, according to our definition, would then suggest $P(\cdot|X)$ as the optimal forecast, a consequence of our

viewpoint that propriety refers to the recovery of the parameterization of an underlying data model. We conjecture that this is essentially why previous attempts to develop proper scoring rules for imprecise probabilities with epistemic semantics have not reached their goal and instead yielded impossibility theorems. It would be premature to conclude that our framework offers nothing to those interested in epistemic semantics for imprecision, however, but the problem would be approached in a different way. For example, under a precise P data model, the optimal forecast $P(\cdot|X)$ is precise and it is calibrated for any loss. If we do not have access to P but to some estimate $\hat{P}$ (e.g. the empirical distribution), we can construct an ambiguity set $\mathcal{P}(\hat{P})$ around it,²⁸ as in the DRO and coherent risk measures literature, and then consider two kinds of statements: first, it would be of interest to obtain bounds that relate, under P , the maximum entropy forecast with respect to $\mathcal{P}(\hat{P})$ or the generalized Bayes rule forecast $\mathcal{P}(\hat{P})_X$ — note that these are random quantities under P — to the optimal but unknown $P(\cdot|X)$; second, statements which control the size of the ambiguity set so that sub-calibration of the forecast $\mathcal{P}(\hat{P})_X$ holds with high probability under P . Formally, with $\tilde{S}_\ell(\omega) = S_\ell(\mathcal{P}(\hat{P})_{X(\omega)}, \omega)$,

$$P\left(\mathbb{E}_P[\tilde{S}_\ell] - \mathbb{E}_P\left[\text{GBR}_{\mathcal{P}(\hat{P})}(\tilde{S}_\ell|X)\right] \geq \alpha\right) \leq \beta,$$

for some $\alpha, \beta > 0$, and similar for the corresponding maximum entropy forecast. In this way, we can obtain imprecise forecasts which satisfy desired “safety” guarantees under precise data models.

We conclude with an appeal to further investigate which benefits imprecise forecasts can bring to the practice of machine learning and statistics.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy EXC number 2064/1 Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christian Fröhlich. Many thanks to Rabanus Derr for the insightful and helpful discussions.

²⁸For example, by using a law invariant coherent risk measure. See e.g. [Duchi and Namkoong, 2021] for such risk measures based on f -divergences.

6.7 APPENDIX

6.7.1 A REMARK ON TIE-BREAKING

Throughout this paper, for picking an action we often face expressions of the form

$$a_{\mathcal{P}}^* \leftarrow \operatorname{argmin}_{a \in \mathcal{A}} \max_{P \in \bar{\operatorname{co}}(\mathcal{P})} \mathbb{E}_P[\ell(a, \omega)]. \quad (6.10)$$

We resolve the potentially set-valued argmin to a single action $a_{\mathcal{P}}^* \in \mathcal{A}$ in a consistent way (for example by installing a total order on $\mathcal{A}$ and then using this order for tie-breaking); we denote this using the “ $\leftarrow$ ” operator. While the exact nature of the tie-breaking does not matter for our purposes, it must be so that we can view Equation 6.10 as a well-defined function from $\mathcal{IP}$ to $\mathcal{A}$.

When the action $a_{\mathcal{P}}^*$ is fixed, we can extract a corresponding P^* in the argmax , which also may not be unique, so that the value of the MinMax objective is

$$\min_{a \in \mathcal{A}} \max_{P \in \bar{\operatorname{co}}(\mathcal{P})} \mathbb{E}_P[\ell(a, \omega)] = \mathbb{E}_{P^*}[\ell(a_{\mathcal{P}}^*, \omega)].$$

The potential non-uniqueness of P^* will not matter in this paper, although it could be resolved by installing a total order on Δ^k (e.g. by restricting the lexicographic order on $\mathbb{R}^k$).

6.7.2 PROOF OF PROPOSITION 289

Proof We want to show that

$$\bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{P}, \omega)) \leq \bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{Q}, \omega)) \quad (6.11)$$

$$\Leftrightarrow \max_{P \in \bar{\operatorname{co}}(\mathcal{P})} \mathbb{E}_P[S_{\ell}(\mathcal{P}, \omega)] \leq \max_{P \in \bar{\operatorname{co}}(\mathcal{P})} \mathbb{E}_P[S_{\ell}(\mathcal{Q}, \omega)] \quad \forall \mathcal{Q} \in \mathcal{IP}. \quad (6.12)$$

Note that when taking the closed convex hull, we can use $\max$ instead of $\sup$. To this end, consider

$$S_{\ell}(\omega, \mathcal{P}) = \ell(a_{\mathcal{P}}^*, \omega), \quad a_{\mathcal{P}}^* \leftarrow \operatorname{argmin}_{a \in \mathcal{A}} \max_{P \in \bar{\operatorname{co}}(\mathcal{P})} \mathbb{E}_P[\ell(a, \omega)],$$

where we pick the $a_{\mathcal{P}}^*$ from the argmin in accordance with Remark 6.7.1. But then, by definition of argmin we have that

$$\max_{P \in \bar{\operatorname{co}}(\mathcal{P})} \mathbb{E}_P[\ell(a_{\mathcal{P}}^*, \omega)] \leq \max_{P \in \bar{\operatorname{co}}(\mathcal{P})} \mathbb{E}_P[\ell(a', \omega)] \quad \forall a' \in \mathcal{A}, \quad (6.13)$$

since otherwise a' and some $P \in \bar{\operatorname{co}}(\mathcal{P})$ would achieve a strictly lower MinMax objective value. Hence it holds in particular for $a' = a_{\mathcal{Q}}^*$, and thus $\bar{R}_{\mathcal{P}}(\ell(a_{\mathcal{P}}^*, \omega)) \leq \bar{R}_{\mathcal{P}}(\ell(a_{\mathcal{Q}}^*, \omega))$.

We remark that Equation 6.13 makes it clear that it does not matter which action a^* is picked from the potentially set-valued argmin when the forecast is $\mathcal{P}$ itself, as the resulting $\max_{P \in \text{co}(\mathcal{P})} \mathbb{E}_P[\ell(a^*, \omega)]$ is guaranteed to be the same. By contrast, consider $\bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{Q}, \omega)) = \bar{R}_{\mathcal{P}}(S_{\ell}(a_{\mathcal{Q}}^*, \omega))$, where

$$a_{\mathcal{Q}}^* \leftarrow \operatorname{argmin}_{a \in \mathcal{A}} \max_{Q \in \text{co}(\mathcal{Q})} \mathbb{E}_Q[\ell(a, \omega)].$$

Here, it is clear that we require a consistent way of picking the action from the argmin, as there is no guarantee that the resulting $\bar{R}_{\mathcal{P}}(S_{\ell}(a_{\mathcal{Q}}^*, \omega))$ is the same for all choices. ■

6.7.3 PROOF OF PROPOSITION 290

Proof

We begin by showing (290.1): $\forall \ell \in \mathcal{L} : \exists \mathcal{P} \in \mathcal{IP} : \exists P \in \text{co}(\mathcal{P})$ so that $\bar{R}_{\mathcal{P}}(S_{\ell}(\mathcal{P}, \omega)) = \bar{R}_{\mathcal{P}}(S_{\ell}(P, \omega))$. To this end, we make use of a Lemma.

Lemma 311. *Define $H : \Delta^k \rightarrow \mathbb{R}$ by $H(P) := \min_{a \in \mathcal{A}} \mathbb{E}_P[\ell(a, \omega)]$. Pick any Q_0, Q_1 in the interior of Δ^k . Then H is differentiable on all points of the segment $[Q_0, Q_1]$ except perhaps on a finite set of points.*

Proof It is well-known that $H(P)$ is concave as a minimum over a set of linear functions, and any concave function is differentiable except perhaps on a countable set of points. We now exploit our assumption that the action space is finite, $|\mathcal{A}| < \infty$, to show that indeed the set must be finite. We restrict ourselves to looking at the segment $[Q_0, Q_1]$.

For any $a_i \in \mathcal{A}$, define $f_i : \mathbb{R}^k \rightarrow \mathbb{R}$ given by $f_i(P) := \mathbb{E}_P[\ell(a_i, \omega)]$, so that $H(P) = \min_{a \in \mathcal{A}} f_i(P)$. Note that f_i is linear, and it is differentiable. We acknowledge that when $P \in \mathbb{R}^k$ but $P \notin \Delta^k$, we should prefer an inner product notation over $\mathbb{E}$, but for simplicity we stick to using $\mathbb{E}$. Thus H is given as the minimum over a finite set of differentiable functions. The function H can only fail to be differentiable at a point P if $f_i(P) = f_j(P)$ for some pair (i, j) (note that this is not *sufficient* for a failure of differentiability, for example when $f_i(P) = f_j(P) \forall P \in \Delta^k$, meaning a_i and a_j always have the same consequences). Thus, to show that H is differentiable except perhaps at a finite number of points, it suffices to bound the number of problematic points where $f_i(P) = f_j(P)$ for each pair (i, j) , noting that there are only finitely many such pairs.

Consider the line $y(\alpha) := \alpha Q_0 + (1 - \alpha)Q_1$, $y : \mathbb{R} \rightarrow \mathbb{R}^k$. For $\alpha \in [0, 1]$, we get the segment $[Q_0, Q_1]$. Due to linearity of f_i , $f_i(y(\alpha)) = \alpha f_i(Q_0) + (1 - \alpha)f_i(Q_1)$. The function $f_i \circ y : \mathbb{R} \rightarrow \mathbb{R}$ is a line. Therefore $f_i \circ y$ and $f_j \circ y$ either fully coincide, do not intersect at all, or intersect in a single point. But now since $f_i(P) = f_j(P) \Leftrightarrow f_i(y(\alpha^*)) = f_j(y(\alpha^*))$, if $P = \alpha^* Q_0 + (1 - \alpha^*)Q_1$ for some $\alpha^* \in \mathbb{R}$, we either have $f_i(P) = f_j(P) \forall P \in [Q_0, Q_1]$,

or $f_i(P) = f_j(P)$ for at most a single $P \in \mathbb{R}^k$. If f_i and f_j fully coincide, then $P \mapsto \min\{f_i(P), f_j(P)\}$ is differentiable, and thus this pair (i, j) poses no challenge to differentiability; also note that if $\ell(a_i, \omega) = \ell(a_j, \omega)$, we could have removed one of them from $\mathcal{A}$ in the setup of our scoring rule without any significant consequence. ■

We now exploit Theorem 6.2 from Grünwald and Dawid [2004]. We slightly restate it using our notation and specialize to our needs.²⁹

Proposition 312 (Grünwald and Dawid [2004, Theorem 6.2, adapted]). *Let $\mathcal{P} \subseteq \Delta^k$ be convex and P^* so that $H(P^*) = \max_{P \in \text{co}(\mathcal{P})} H(P)$. Suppose that $\forall P \in \mathcal{P}$ there exists $P_0 \in \Delta^k$ so that, with $Q_\lambda := (1 - \lambda)P_0 + \lambda P$ it holds:*

$$(312.1) \quad P^* = Q_{\lambda^*} \text{ for some } \lambda^* \in (0, 1),$$

$$(312.2) \quad \text{and } \lambda \mapsto H(Q_\lambda) \text{ is differentiable at } \lambda = \lambda^*. \text{ Then } a_{P^*}^* \leftarrow \operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}_{P^*}[\ell(a, \omega)] \text{ satisfies}$$

$$\min_{a \in \mathcal{A}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(a, \omega)] = \sup_{P \in \mathcal{P}} \min_{a \in \mathcal{A}} \mathbb{E}_P[\ell(a, \omega)] = \mathbb{E}_{P^*}[\ell(a_{P^*}^*, \omega)] = \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(a_{P^*}^*, \omega)].$$

If the conditions of the theorem are satisfied, then $\bar{R}_{\mathcal{P}}(S_\ell(\mathcal{P}, \omega)) = \bar{R}_{\mathcal{P}}(S_\ell(P^*, \omega))$, because $\min_{a \in \mathcal{A}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(a, \omega)] = \bar{R}_{\mathcal{P}}(S_\ell(\mathcal{P}, \omega))$. This would mean that we obtain (290.1).

We now show that the conditions of the statement are satisfied. We take any segment $[Q_0, Q_1]$ in the interior of Δ^k , and then take some $\mathcal{P}$ so that $\mathcal{P} \subsetneq \text{int}([Q_0, Q_1])$, and so that H is differentiable on all of $\mathcal{P}$. In particular, H is then continuous on $\mathcal{P}$ (a concave function is continuous on the interior of its domain), and thus attains its maximum on $\mathcal{P}$ at some P^* . The previous Lemma assures us that such $\mathcal{P}$ exists, since we can have non-differentiability at only finitely many points of $[Q_0, Q_1]$. The choice that $\mathcal{P} \subsetneq \text{int}([Q_0, Q_1])$ guarantees that for every $P \in \mathcal{P}$, there exists some $P_0 \in [Q_0, Q_1]$ so that with $Q_\lambda = (1 - \lambda)P_0 + \lambda P$ we get that the desired $P^* = Q_{\lambda^*}$ for some $\lambda^* \in (0, 1)$.

It remains to show that then $\lambda \mapsto H(Q_\lambda)$ is differentiable at $\lambda = \lambda^*$. This mapping can be seen as a composition $H(g(\lambda))$, where $g : (0, 1) \rightarrow \Delta^k$, $g(\lambda) = (1 - \lambda)P_0 + \lambda P$, hence $H \circ g : (0, 1) \rightarrow \mathbb{R}$. Differentiability follows from the multivariate chain rule, considering that H is differentiable on $[Q_0, Q_1]$ and g is differentiable on $(0, 1)$. Therefore Theorem 6.2 of Grünwald and Dawid [2004] applies.

Finally, we show by example that (290.2): $\exists \ell : \exists \mathcal{P}$ so that $\forall P \in \Delta^k : \bar{R}_{\mathcal{P}}(S_\ell(\mathcal{P}, \omega)) < \bar{R}_{\mathcal{P}}(S_\ell(P, \omega))$.

²⁹We remark that the original statement of Theorem 6.2 by Grünwald and Dawid [2004] works with *any* potentially randomized Bayes actions (“acts”) ζ^* (“Bayes act against P^* ”). In our setup, we do not include randomized actions, and there always exists a non-randomized Bayes act $a_{P^*}^* \leftarrow \operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}_{P^*}[\ell(a, \omega)]$. In our statement of (312.2), we have inlined the definition of a “saddle point” in this setting; see Theorem 4.3 of Grünwald and Dawid [2004].

Let $\Omega = \{\omega_1, \omega_2\}$, $\mathcal{A} = \{a_1, a_2, a_3\}$ and $\mathcal{P} = \{P_1, P_2\}$. For the tie-breaking (see Appendix 6.7.1) we fix the ordering $a_1 > a_2 > a_3$. Let $P_1(\{\omega_1\}) = 1$, $P_2(\{\omega_2\}) = 1$. Define the loss function by $\ell(a_1, \omega_1) = 2$, $\ell(a_1, \omega_2) = 7$; $\ell(a_2, \omega_1) = 6$, $\ell(a_2, \omega_2) = 3$; and $\ell(a_3, \omega) = 0.5\ell(a_1, \omega) + 0.5\ell(a_2, \omega)$, $\omega \in \Omega$. Hence $\mathbb{E}_{P_1}[\ell(a_3, \omega)] = 4$ and $\mathbb{E}_{P_2}[\ell(a_3, \omega)] = 5$. Observe that a_3 is the unique optimal action under $\mathcal{P}$:

$$\{a_3\} = \operatorname{argmin}_{a \in \mathcal{A}} \max_{P \in \{P_1, P_2\}} \mathbb{E}_P[\ell(a, \omega)] < \min_{a \in \{a_1, a_2\}} \max_{P \in \{P_1, P_2\}} \mathbb{E}_P[\ell(a, \omega)].$$

Now assume that there exists some $P^* \in \Delta^k$ which *uniquely* recommends a_3 :

$$\{a_3\} = \operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}_{P^*}[\ell(a, \omega)],$$

where *uniquely* means that the argmin is the singleton $\{a_3\}$. If the argmin was not a singleton, then the action recommendation would depend on the contingent tie-breaking procedure, which in this case would favor a_1 and a_2 over a_3 , leading to a sub-optimal recommendation.

Assume $P^*(\{\omega_1\}) = p$ and $P^*(\{\omega_2\}) = 1 - p$. Then

$$\mathbb{E}_{P^*}[\ell(a_1, \omega)] = p \cdot 2 + (1 - p) \cdot 7, \quad (6.14)$$

$$\mathbb{E}_{P^*}[\ell(a_2, \omega)] = p \cdot 6 + (1 - p) \cdot 3, \quad (6.15)$$

$$\mathbb{E}_{P^*}[\ell(a_3, \omega)] = p \cdot 4 + (1 - p) \cdot 5. \quad (6.16)$$

If we want P^* to uniquely recommend a_3 , we need

$$p \cdot 4 + (1 - p) \cdot 5 < \min(p \cdot 2 + (1 - p) \cdot 7, p \cdot 6 + (1 - p) \cdot 3),$$

but this has no solution. Observe that for $P^* = 0.5P_1 + 0.5P_2$, which uniquely maximizes H in this case, all actions get the same expected score, but the balance cannot be tilted towards a_3 . This situation relates to an insight of Grünwald and Dawid [2004]: if a P^* maximizing H on $\operatorname{co}(\mathcal{P})$ exists but it has no unique Bayes action, then it is unclear which of these Bayes actions (where we would have to allow for randomized actions, as well) is “robust Bayes”, which means solving the $\min_{a \in \mathcal{A}} \max_{P \in \operatorname{co}(\mathcal{P})} \mathbb{E}_P[\ell(a, \omega)]$ objective. In the above example, a_3 can be essentially interpreted as a randomized action, where a coin flip decides between a_1 and a_2 . Note that in this example,

$$H(P^*) = \mathbb{E}_{P^*}[S_\ell(P^*, \omega)] = \max_{P \in \operatorname{co}(\mathcal{P})} \min_{a \in \mathcal{A}} \mathbb{E}_P[\ell(a, \omega)] = 4.5 \quad (6.17)$$

$$< \overline{R}_{\mathcal{P}}(S_\ell(\mathcal{P}, \omega)) = \overline{R}_{\mathcal{P}}(\ell(a_3, \omega)) = 5 \quad (6.18)$$

$$< \overline{R}_{\mathcal{P}}(S_\ell(P^*, \omega)) = \overline{R}_{\mathcal{P}}(\ell(a_1, \omega)) = 7. \quad (6.19)$$

meaning the maximum entropy is strictly smaller than the IP score of the optimal constant forecast, which is in turn smaller than the IP score of the forecast corresponding to the maximum entropy probability.

Finally, we note that if $|\Omega| > 2$, the above proof can be straightforwardly adapted, where P_1 and P_2 only have the support on two elementary events. The logic regarding P^* also works since a rescaling of P^* does not affect the order of the relevant expected losses. ■

6.7.4 PROOF OF PROPOSITION 291

To obtain the statement, we prove the following proposition.

Proposition 313. *Let $\mathcal{P}, \mathcal{Q} \in \mathcal{IP}$ and $\mathcal{P} \not\preceq \mathcal{Q}$. Then there exist $\ell \in \mathcal{L}$ such that*

$$\left(a_{\mathcal{P}}^* \leftarrow \operatorname{argmin}_{a \in \mathcal{A}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(a, \omega)] \right) \neq \left(a_{\mathcal{Q}}^* \leftarrow \operatorname{argmin}_{a \in \mathcal{A}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[\ell(a, \omega)] \right),$$

and

$$\exists \omega \in \Omega : S(\mathcal{P}, \omega) \neq S(\mathcal{Q}, \omega),$$

and indeed that

$$\overline{R}_{\mathcal{P}}(S_{\ell}(\mathcal{P}, \omega)) < \overline{R}_{\mathcal{P}}(S_{\ell}(\mathcal{Q}, \omega)) \tag{6.20}$$

$$\Leftrightarrow \overline{R}_{\mathcal{P}}(\ell_{a_{\mathcal{P}}^*}) < \overline{R}_{\mathcal{P}}(\ell_{a_{\mathcal{Q}}^*}). \tag{6.21}$$

Proof Since $\mathcal{P} \not\preceq \mathcal{Q}$, we can find some $X : \Omega \rightarrow \mathbb{R}$ such that either $\overline{R}_{\mathcal{P}}(X) < b < \overline{R}_{\mathcal{Q}}(X)$ or $\overline{R}_{\mathcal{Q}}(X) < b < \overline{R}_{\mathcal{P}}(X)$ ³⁰ — but we cannot choose which way (consider $\mathcal{P} = \Delta^k$, this never gives a smaller risk assessment than any other $\mathcal{Q}$). Then construct the loss function as $\ell(a_1) := X(\omega)$, and $\ell(a_2) = b$, and $\ell(a) = c > \max(X(\omega), b)$ otherwise.

First, assume that the direction is such that $\overline{R}_{\mathcal{P}}(X) < b < \overline{R}_{\mathcal{Q}}(X)$. Clearly, the credal set $\mathcal{P}$ will recommend a_1 whereas $\mathcal{Q}$ will recommend a_2 . That is, they recommend different actions $a_{\mathcal{P}}^* \neq a_{\mathcal{Q}}^*$. Now we also show that:

$$\overline{R}_{\mathcal{P}}(S_{\ell}(\mathcal{P}, \omega)) < \overline{R}_{\mathcal{P}}(S_{\ell}(\mathcal{Q}, \omega))$$

We know that $\mathcal{P}$ recommends a_1 hence the left term is $\overline{R}_{\mathcal{P}}(\ell(a_1, \omega)) = \overline{R}_{\mathcal{P}}(X)$, whereas

³⁰This holds since there is a one-to-one correspondence of closed convex sets of probabilities and upper expectations [Walley, 1991, Section 3.6.1]. We remark that Walley [1991] does not work with probabilities in Δ^k , but with so-called “linear prevision”. For the equivalence of these viewpoints when $|\Omega| < \infty$, see [Fröhlich et al., 2024].

$\mathcal{Q}$ recommends a_2 , meaning the right term is $\bar{R}_{\mathcal{P}}(\ell(a_2, \omega)) = b$. Since $\bar{R}_{\mathcal{P}}(X) < b$ by assumption, this proves the statement for this direction.

Now assume conversely that $\bar{R}_{\mathcal{P}}(X) > b > \bar{R}_{\mathcal{Q}}(X)$ and use the same ℓ . Then $\mathcal{P}$ will recommend a_2 , whereas $\mathcal{Q}$ will recommend a_1 . But $\bar{R}_{\mathcal{P}}(\ell(a_2, \omega)) = b < \bar{R}_{\mathcal{P}}(X)$. ■

6.7.5 PROOF OF PROPOSITION 293

We first lift the action space to combine actions for the different $X = x$ in a single higher-order action; If $\mathcal{X} = \{x_1, \dots, x_J\}$, we define $\tilde{\mathcal{A}} := \{(a_1, \dots, a_J) : a_j \in \mathcal{A}, j = 1..J\}$, and the lifted loss function is $\ell' : \tilde{\mathcal{A}} \times \Omega \rightarrow \mathbb{R}$, where

$$\ell'(\tilde{a}, \omega) := \ell(a_j, \omega) \text{ if } X(\omega) = x_j.$$

We then consider

$$\min_{\tilde{a} \in \tilde{\mathcal{A}}} \max_{P \in \text{co}(\mathcal{P})} \mathbb{E}_P[\ell'(\tilde{a}, \omega)] \leq \min_{\mathcal{Q} : \mathcal{X} \rightarrow \mathcal{IP}} \max_{P \in \text{co}P} \mathbb{E}_P[S_{\ell}(\mathcal{Q}(X), \omega)] = \text{IP-Score}^*(\ell, \mathcal{P}).$$

That the inequality holds is clear since for any $S_{\ell}(\mathcal{Q}(X), \omega)$, we have the correspond lifted action $\tilde{a}_{\mathcal{Q}} := (a_{\mathcal{Q}(x_1)}^*, \dots, a_{\mathcal{Q}(x_J)}^*)$ on the left side, and $S_{\ell}(\mathcal{Q}(X), \omega) = \ell'(\tilde{a}_{\mathcal{Q}}, \omega)$. Thus we aim to characterize the optimal $\tilde{a}^* = (a_1^*, \dots, a_J^*)$ of the MinMax objective on the left side first. If we can then further exhibit some (in this case precise) forecast $f^* : \mathcal{X} \rightarrow \Delta^k$, so that we have $a_j^* \leftarrow \text{argmin}_{a \in \mathcal{A}} \mathbb{E}_{f^*(x_j)}[\ell(a, \omega)] \forall x_j \in \mathcal{X}$, then f^* is optimal on the right side.

We now focus on $\min_{\tilde{a} \in \tilde{\mathcal{A}}} \max_{P \in \text{co}(\mathcal{P})} \mathbb{E}_P[\ell'(\tilde{a}, \omega)]$, and apply to it the maximum entropy principle of Grünwald and Dawid [2004]. Adapted to our setup, Theorem 5.2 of Grünwald and Dawid [2004] provides the following, here stated informally.

Proposition 314 (adapted from Theorem 5.2 of Grünwald and Dawid [2004]). *Compute some maximum entropy probability*

$$P^* \in \arg \max_{P \in \text{co}(\mathcal{P})} \min_{\tilde{a} \in \tilde{\mathcal{A}}} \mathbb{E}_P[\ell'(\tilde{a}, \omega)],$$

noting that any P^* in the arg max suffices. Then P^* admits a potentially randomized Bayes action ζ^* , which has the desired property of being a “saddle point”.

From Proposition 3.1.iii) of Grünwald and Dawid [2004], we know that if a probability P has a randomized Bayes act, then it also has some non-randomized Bayes action. In combination, we therefore know that if P^* has a *unique* Bayes action, meaning $\arg \min_{\tilde{a} \in \tilde{\mathcal{A}}} \mathbb{E}_{P^*}[\ell'(\tilde{a}, \omega)]$ is a singleton, then this must be a non-randomized one, so it fits our setup. Thus we obtain, by inlining the definition of saddle point:

Proposition 315. *Compute some maximum entropy probability*

$$P^* \in \arg \max_{P \in \text{co}(\mathcal{P})} \min_{\tilde{a} \in \tilde{\mathcal{A}}} \mathbb{E}_P[\ell'(\tilde{a}, \omega)].$$

If

$$\{\tilde{a}_{P^*}^*\} := \operatorname{argmin}_{\tilde{a} \in \tilde{\mathcal{A}}} \mathbb{E}_{P^*}[\ell'(\tilde{a}, \omega)],$$

i.e. the action recommendation on the lifted space is unique, then

$$\min_{\tilde{a} \in \tilde{\mathcal{A}}} \max_{P \in \text{co}(\mathcal{P})} \mathbb{E}_P[\ell'(\tilde{a}, \omega)] = \max_{P \in \text{co}(\mathcal{P})} \min_{\tilde{a} \in \tilde{\mathcal{A}}} \mathbb{E}_P[\ell'(\tilde{a}, \omega)] = \mathbb{E}_{P^*}[\ell'(\tilde{a}_{P^*}^*, \omega)] = \max_{P \in \text{co}(\mathcal{P})} \mathbb{E}_P[\ell'(\tilde{a}_{P^*}^*, \omega)]$$

We want to show that indeed the forecast P^* , if this uniqueness condition holds, will recommend actions which when viewed as a single lifted action coincide exactly this lifted action $\tilde{a}_{P^*}^*$.

Lemma 316. *Assume $P^*(X = x_j) > 0 \ \forall x_j \in \mathcal{X}$. Let $I := \operatorname{argmin}_{\tilde{a} \in \tilde{\mathcal{A}}} \mathbb{E}_{P^*}[\ell'(\tilde{a}, \omega)]$ and consider $II_j := \operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}_{P^*(\cdot|X=x_j)}[\ell(a, \omega)]$.*

Then I is a singleton (unique) if and only if II_j is a singleton (unique) for each $X = x_j, j = 1..J$. In the case of uniqueness, we get that $I = (II_1, \dots, II_J)$.

Proof Consider

$$\min_{\tilde{a} \in \tilde{\mathcal{A}}} \mathbb{E}_{P^*}[\ell'(a, \omega)] \tag{6.22}$$

$$= \min_{\tilde{a}=(a_1, \dots, a_J) \in \tilde{\mathcal{A}}} \sum_{j=1}^J \mathbb{E}_{P^*(\cdot|X=x_j)}[\ell(a_j, \omega)] P^*(X = x_j) \tag{6.23}$$

The minimization with respect to the j -th component of $\tilde{a}$ depends only on the j -th summand. Note that if we had $P^*(X = x_j) = 0$, then the j -th component would remain fully undetermined. Otherwise, it is clear that with $\tilde{a}^* = (\tilde{a}_1^*, \dots, \tilde{a}_J^*)$, the set of actions which we can obtain in the j -th component of an optimal solution $\tilde{a}^*$ coincides exactly with the potentially set-valued $\operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}_{P^*(\cdot|X=x_j)}[\ell(a, \omega)]$, since the scaling factor $P^*(X = x_j)$ is irrelevant for the solution, as long as it is strictly positive. Thus uniqueness is preserved in both directions. ■

The desired statement now follows from the Lemma: the uniqueness of $a_{P^*}^*$ guarantees the uniqueness of the $II_j := \operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}_{P^*(\cdot|X=x_j)}[\ell(a, \omega)]$, and moreover that $a_{P^*}^* = (II_1, \dots, II_J)$. Therefore $\text{IP-Score}^*(\ell, \mathcal{P}) = \overline{R}_{\mathcal{P}}(S_{\ell}(f^*(X), \omega))$ upon defining $f^*(x_j) := P^*(\cdot|X = x_j)$, $x_j \in \mathcal{X}$.

Finally, to obtain the form of Proposition 293, it remains to show:

Lemma 317. *We can rewrite the generalized conditional entropy as follows:*

$$H_\ell(P|X) := \mathbb{E}_{X, \omega \sim P} [S_\ell(P(\cdot|X), \omega)] = \min_{\tilde{a} \in \mathcal{A}} \mathbb{E}_P[\ell'(\tilde{a}, \omega)].$$

Proof Assume without loss of generality that $P(X = x_j) > 0 \forall x_j \in \mathcal{X}$ (convince yourself that this does not matter). We begin by rewriting the left side, giving

$$\mathbb{E}_{X, \omega \sim P} [S_\ell(P(\cdot|X), \omega)] = \sum_{j=1}^J \mathbb{E}_{\omega \sim P(\cdot|X=x_j)} [S_\ell(P(\cdot|X=x_j), \omega)] P(X = x_j) \quad (6.24)$$

$$= \sum_{j=1}^J \mathbb{E}_{\omega \sim P(\cdot|X=x_j)} [\ell(a_{P(\cdot|X=x_j)}^*, \omega)] P(X = x_j). \quad (6.25)$$

Now the right side gives

$$\min_{\tilde{a} \in \mathcal{A}} \mathbb{E}_P[\ell'(\tilde{a}, \omega)] \quad (6.26)$$

$$= \min_{\tilde{a} = (a_1, \dots, a_J) \in \tilde{\mathcal{A}}} \sum_{j=1}^J \mathbb{E}_{P(\cdot|X=x_j)} [\ell(a_j, \omega)] P(X = x_j). \quad (6.27)$$

Write $\tilde{a}^* = (\tilde{a}_1^*, \dots, \tilde{a}_J^*)$ for any of the solutions (without uniqueness guarantee). Observe that

$$\mathbb{E}_{\omega \sim P(\cdot|X=x_j)} [\ell(a_j^*, \omega)] = \mathbb{E}_{\omega \sim P(\cdot|X=x_j)} [\ell(a_{P(\cdot|X=x_j)}^*, \omega)],$$

although possibly $a_j^* \neq a_{P(\cdot|X=x_j)}^*$. We can thus conclude that $\mathbb{E}_{X, \omega \sim P} [S_\ell(P(\cdot|X), \omega)] = \min_{\tilde{a} \in \mathcal{A}} \mathbb{E}_P[\ell'(\tilde{a}, \omega)]$. ■

6.7.6 AN EXTENDED EXAMPLE

We first begin with an unconditional example without features, and then expand on it to include features. Let $\Omega := \{\text{rain}=0, \text{rain}=1\}$ and $\mathcal{A} := \{\text{umbrella}=0, \text{umbrella}=1\}$. Consider the cost-sensitive loss function ℓ with parameter $c = 0.1$, given by the table:

	rain=0	rain=1
umbrella=0	0	0.9
umbrella=1	0.1	0

To simplify notation, we implicitly identify a probability measure P for which $P(\text{rain}=1) = p$ with the point $p \in [0, 1]$. We abbreviate “umbrella=0” as “u=0”. Consider $\mathcal{P} := [0.85, 0.95]$.

Then the IP scores of the actions can be computed as:

$$\bar{R}_{\mathcal{P}}(\ell(u=0, \omega)) = \max_{P \in [0.85, 0.95]} P \cdot 0.9 = 0.95 \cdot 0.9 = 0.855 \quad (6.28)$$

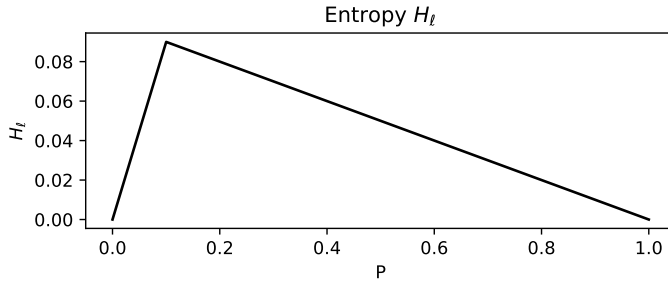
$$\bar{R}_{\mathcal{P}}(\ell(u=1, \omega)) = \max_{P \in [0.85, 0.95]} (1 - P) \cdot 0.1 = (1 - 0.85) \cdot 0.1 = 0.015. \quad (6.29)$$

So without question, the optimal action is to bring an umbrella. Note that a precise forecast P uniquely recommends bringing an umbrella if $P > 0.1$, and for $P = 0.1$ both actions $u = 0$ and $u = 1$ have the same expected loss. Hence, in this case all $P \in \mathcal{P}$ agree on the action recommendation.

Without features, the conditional entropy of Definition 292 (take a trivial, constant X) simply becomes the unconditional

$$H_{\ell}(P) := \min_{a \in \mathcal{A}} \mathbb{E}_P[\ell(a, \omega)].$$

Then $P^* = 0.1$, $H(P^*) = 0.09$, is indeed the maximum entropy probability over $[0, 1]$, as the following graphic shows:



And for this P^* , the condition of Proposition 293 is not satisfied, since the action recommendation is not unique. To observe this failure, we can therefore consider an interval $\mathcal{P}' := [0.05, 0.15]$, which includes P^* . The IP scores of the actions are:

$$\bar{R}_{\mathcal{P}'}(\ell(u=0, \omega)) = \max_{P \in [0.05, 0.15]} P \cdot 0.9 = 0.15 \cdot 0.9 = 0.135 \quad (6.30)$$

$$\bar{R}_{\mathcal{P}'}(\ell(u=1, \omega)) = \max_{P \in [0.05, 0.15]} (1 - P) \cdot 0.1 = (1 - 0.05) \cdot 0.1 = 0.095. \quad (6.31)$$

Thus, it is again uniquely optimal to bring an umbrella under $\mathcal{P}'$, but now there is disagreement among the precise probabilities in $\mathcal{P}'$, and the maximum entropy principle cannot be applied since P^* does not give a unique recommendation. Clearly, any precise forecast $P > 0.1$ is optimal in this example, as well as any imprecise forecast $[P, Q]$ for which $P > 0.1$ (we always assume for an interval $[P, Q]$ that $P \leq Q$).

We now lift this example to include features $\mathcal{X} := \{\text{cloudy}, \text{sunny}\}$. Then $\Omega = \mathcal{X} \times \mathcal{Y}$,

$\mathcal{Y} := \{\text{rain}=0, \text{rain}=1\}$. Assume that for some country A, the weather can be described by

$$P^1(X = \text{cloudy}) = 0.4, \quad (6.32)$$

$$P^1(Y = \text{rain}=1|X = \text{cloudy}) = 0.95, \quad (6.33)$$

$$P^1(Y = \text{rain}=1|X = \text{sunny}) = 0.05. \quad (6.34)$$

In contrast, for country B:

$$P^2(X = \text{cloudy}) = 0.9, \quad (6.35)$$

$$P^2(Y = \text{rain}=1|X = \text{cloudy}) = 0.85, \quad (6.36)$$

$$P^2(Y = \text{rain}=1|X = \text{sunny}) = 0.15. \quad (6.37)$$

Let us consider the lifted action space

$$\tilde{A} := \{(u = 0, u = 0), (u = 0, u = 1), (u = 1, u = 0), (u = 1, u = 1)\},$$

as in the proof of Proposition 293. Correspondingly, the lifted loss function is

$$\ell'(\tilde{a} = (a_1, a_2), (x, y)) := \ell(a_1, y) \text{ if } x = \text{cloudy}; \ell(a_2, y) \text{ otherwise.}$$

The GBR forecast (see Section 6.2.4) is $[0.85, 0.95]$ for $X = \text{cloudy}$, and $[0.05, 0.15]$ for $X = \text{sunny}$. Therefore, in light of what we have seen before, it recommends the lifted action $(u = 1, u = 1)$.

We now compute the IP scores of the lifted actions under $\tilde{\mathcal{P}} = \{P^1, P^2\}$ (to compute upper expectations it suffices to consider the extreme points):

$$\overline{R}_{\tilde{\mathcal{P}}}(\ell'((u = 0, u = 0), \omega)) = \max_{P \in \{P^1, P^2\}} P(X = \text{cloudy})P(Y = \text{rain}=1|X = \text{cloudy}) \cdot 0.9 \quad (6.38)$$

$$+ (1 - P(X = \text{cloudy}))P(Y = \text{rain}=1|X = \text{sunny}) \cdot 0.9 \quad (6.39)$$

$$= 0.702, \quad (6.40)$$

$$\bar{R}_{\tilde{P}}(\ell'((u = 0, u = 1), \omega)) = \max_{P \in \{P^1, P^2\}} P(X = \text{cloudy})P(Y = \text{rain}=1|X = \text{cloudy}) \cdot 0.9 \quad (6.41)$$

$$+ (1 - P(X = \text{cloudy}))P(Y = \text{rain}=0|X = \text{sunny}) \cdot 0.1 \quad (6.42)$$

$$\approx 0.697, \quad (6.43)$$

$$\bar{R}_{\tilde{P}}(\ell'((u = 1, u = 0), \omega)) = \max_{P \in \{P^1, P^2\}} P(X = \text{cloudy})P(Y = \text{rain}=0|X = \text{cloudy}) \cdot 0.1 \quad (6.44)$$

$$+ (1 - P(X = \text{cloudy}))P(Y = \text{rain}=1|X = \text{sunny}) \cdot 0.9 \quad (6.45)$$

$$= \mathbf{0.029}, \quad (6.46)$$

$$\bar{R}_{\tilde{P}}(\ell'((u = 1, u = 1), \omega)) = \max_{P \in \{P^1, P^2\}} P(X = \text{cloudy})P(Y = \text{rain}=0|X = \text{cloudy}) \cdot 0.1 \quad (6.47)$$

$$+ (1 - P(X = \text{cloudy}))P(Y = \text{rain}=0|X = \text{sunny}) \cdot 0.1 \quad (6.48)$$

$$= 0.059. \quad (6.49)$$

Thus, we see that the lifted action $(u = 1, u = 0)$ is uniquely optimal, that is, bringing an umbrella on cloudy days but not on sunny days. Recall that country A recommends an umbrella only on cloudy days, hence the optimal forecast with respect to $\tilde{P}$ coincides with the optimal forecast for P^1 , whereas P^2 recommends an umbrella for both cloudy and sunny days. It can be checked that indeed P^1 is the unique maximum entropy probability in this example, where we now have to consider the conditional entropy of Definition 292, i.e. the full Bayes risk. Here, the marginals on X are crucial, since they affect the full Bayes risk. Since P^1 uniquely recommends $(u = 1, u = 0)$, the maximum entropy principle of Proposition 293 applies (at least approximately).

To observe that P^1 is the unique maximum entropy probability, first note that

$$H_\ell(P) = \min((1 - 0.1)P, 0.1(1 - P)),$$

and then compute

$$\dagger := \mathbb{E}_{P^\lambda} \left[H_\ell(P^\lambda(\cdot|X)) \right], \quad P^\lambda := \lambda P^1 + (1 - \lambda)P^2,$$

which gives

$$\dagger = (0.6 - 0.2 \cdot \lambda) \cdot \min(0.9 \cdot y, 0.1 \cdot (1 - y)) + (1 - (0.6 - 0.2 \cdot \lambda)) \cdot \min(0.9 \cdot z, 0.1 \cdot (1 - z)), \quad (6.50)$$

$$y := (\lambda \cdot 0.95 \cdot 0.4 + (1 - \lambda) \cdot 0.85 \cdot 0.9) / (\lambda \cdot 0.4 + (1 - \lambda) \cdot 0.6), \quad (6.51)$$

$$z := (\lambda \cdot 0.05 \cdot 0.6 + (1 - \lambda) \cdot 0.15 \cdot 0.1) / (1 - (\lambda \cdot 0.4 + (1 - \lambda) \cdot 0.6)). \quad (6.52)$$

This term has a unique maximum at $\lambda = 1$ (a computer algebra system comes to the rescue).

Failure of sub-calibration. First, we illustrate that it is possibly that two distinct ($\neq$) forecasts are optimal, but only of them is sub-calibrated. Consider again $\mathcal{P}' = [0.05, 0.15]$. An optimal forecast is $Q = 0.15$ with action recommendation $u = 1$, but it is not sub-calibrated without groups (in this case, the action-induced partition is the trivial one, since the action recommendation is constant). Sub-calibration would mean that:

$$\bar{R}_{\mathcal{P}'}(\ell(u = 1, \omega) - \mathbb{E}_{Q=0.15}[\ell(u = 1, \omega)]) \leq 0 \quad (6.53)$$

$$\Leftrightarrow \forall P' \in \mathcal{P}' : \mathbb{E}_{P'}[\ell(u = 1, \omega)] \leq \mathbb{E}_{Q=0.15}[\ell(u = 1, \omega)], \quad (6.54)$$

but for $P' = 0.05$, we have $\mathbb{E}_{P'=0.05}[\ell(u = 1, \omega)] = 0.095 > \mathbb{E}_{Q=0.15}[\ell(u = 1, \omega)] = 0.085$, and therefore sub-calibration does not hold. In contrast, the forecast $\mathcal{P}'$ is also optimal (it also recommends $u = 1$) and is trivially (sub-)calibrated; but the crucial difference is in the *risk assessment* of the consequences.

We can also show that in the extended example above, which includes features “cloudy” and “sunny”, where the condition of Proposition 293 applies, the optimal forecast f^* of Proposition 293 is neither sub-calibrated without groups nor $\mathcal{B}_{f^*, \mathcal{A}}$ -sub-calibrated. We know that P^1 which recommends $(u = 1, u = 0)$ is the unique maximizer of the generalized conditional entropy and thus corresponds to the optimal forecast in Proposition 293. To check sub-calibration without groups, we need to see if

$$T := \bar{R}_{\mathcal{P}} \left(\ell'((u = 1, u = 0), \omega) - \mathbb{E}_{P^1(\cdot|X=.)}[\omega' \mapsto \ell'((u = 1, u = 0), \omega')] \right) \leq 0. \quad (6.55)$$

Write $T = \max(T_1, T_2)$, noting that we only need to look at the extreme points of $\mathcal{P}$, with

$$T_i := \underbrace{\mathbb{E}_{P^i}[\ell'((u = 1, u = 0), \omega)]}_{\text{actual expected loss of } P^1 \text{ forecast under } P^i} - \underbrace{\mathbb{E}_{X \sim P^i}}_{\text{sampling the forecast}} \left[\underbrace{\mathbb{E}_{\omega \sim P^1(\cdot|X)}[\ell'((u = 1, u = 0), \omega)]}_{\text{entropy of the } P^1 \text{ forecast for fixed } X} \right] \quad (6.56)$$

which is equal to

$$T_i := \left(P^i(X = \text{cloudy})P^i(Y = \text{rain}=0|X = \text{cloudy}) \cdot 0.1 \right. \quad (6.57)$$

$$\left. + P^i(X = \text{sunny})P^i(Y = \text{rain}=1|X = \text{sunny}) \cdot 0.9 \right) \quad (6.58)$$

$$- \left(P^i(X = \text{cloudy})P^1(Y = \text{rain}=0|X = \text{cloudy}) \cdot 0.1 \right. \quad (6.59)$$

$$\left. + P^i(X = \text{sunny})P^1(Y = \text{rain}=1|X = \text{sunny}) \cdot 0.9 \right) \quad (6.60)$$

By plugging in the respective numbers we obtain that $T = \max(0, 0.011) > 0$, however. Hence the P^1 forecast is not sub-calibrated without groups, and therefore also cannot be (sub-)calibrated on any finer partition, in particular the action-induced partition.

6.7.7 PROOF OF PROPOSITION 301

Recall that a partition $\mathcal{B}_1$ is coarser than $\mathcal{B}_2$ if every set in $\mathcal{B}_2$ is a subset of some set in $\mathcal{B}_1$. This can usefully be reformulated as follows.

Lemma 318 (known). *If $\mathcal{B}_1$ is coarser than $\mathcal{B}_2$, then every set in $\mathcal{B}_1$ can be written as a disjoint union of some sets from $\mathcal{B}_2$.*

Proof Fix $B_1 \in \mathcal{B}_1$ which we want to write as a disjoint union of sets from $\mathcal{B}_2$. Consider all those $B_{2i} \in \mathcal{B}_2$ for which $B_{2i} \subseteq B_1$. We claim that the union of those disjoint sets is exactly B_1 . First, it is clear that the union cannot be larger than B_1 . To show that the union is exactly B_1 we must show that every $b \in B_1$ is in some B_{2i} . Since $\mathcal{B}_2$ is a partition, $b \in \tilde{B}_2$ for some $\tilde{B}_2 \in \mathcal{B}_2$. From the refinement property, we know that there exists some $\tilde{B}_1 \in \mathcal{B}_1$ for which $\tilde{B}_2 \subseteq \tilde{B}_1$, so $b \in \tilde{B}_1$. But $\mathcal{B}_1$ is a partition, so either $\tilde{B}_1 = B_1$ (then we are done, since $\tilde{B}_2$ is one of the B_{2i}), or $\tilde{B}_1 \cap B_1 = \emptyset$, but this contradicts the assumption that $b \in B_1$. ■

We can now prove the Proposition.

Proof Consider some $B_1 \in \mathcal{B}_1$. Write $B_1 = \bigcup_{i=1}^I B_{2i}$ for the disjoint union of sets from $\mathcal{B}_2$. Then $\chi_{B_1} = \sum_{i=1}^I \chi_{B_{2i}}$. Since $\bar{R}_{\mathcal{P}}$ is subadditive, we get

$$\bar{R}_{\mathcal{P}} \left(\chi_{B_1}(\omega) \left(\tilde{S}(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\tilde{S}) \right) \right) \quad (6.61)$$

$$= \bar{R}_{\mathcal{P}} \left(\sum_{i=1}^I \chi_{B_{2i}}(\omega) \left(\tilde{S}(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\tilde{S}) \right) \right) \quad (6.62)$$

$$\leq \sum_{i=1}^I \bar{R}_{\mathcal{P}} \left(\chi_{B_{2i}}(\omega) \left(\tilde{S}(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\tilde{S}) \right) \right) \leq 0, \quad (6.63)$$

by the assumption of $(\ell, \mathcal{B}_2)$ -sub-calibration, and therefore we have shown $(\ell, \mathcal{B}_1)$ -sub-calibration. ■

6.7.8 PROOF OF PROPOSITION 305

Proof First we clarify notation. When writing expressions of the form

$$\overline{R}_{\mathcal{P}} \left(f(\omega) - \overline{R}_{Q(\omega)}(\omega' \mapsto f(\omega')) \right),$$

it can be necessary to distinguish in the notation the ω corresponding to the gamble on which we here evaluate the outer $\overline{R}_{\mathcal{P}}$, and an additional inner variable. For each $\omega \in \Omega$, the term $\overline{R}_{Q(\omega)}(\omega' \mapsto f(\omega'))$ here is simply a constant. When the distinction is not necessary, we fall back to using ω as the variable.

To understand that the GBR is $(\ell, \mathcal{B})$ -calibrated, we rewrite the definition of calibration using Remark 303 and plug in the definition of the GBR forecast:

$$\forall B \in \mathcal{B} : \text{GBR}_{\mathcal{P}} \left(S_{\ell}(\mathcal{P}_{|B}, \omega) - \text{GBR}_{\mathcal{P}} \left(\omega' \mapsto S_{\ell}(\mathcal{P}_{|B}, \omega') \right) | B \right) = 0.$$

We highlight again to the reader the subtle distinction of ω vs. ω' here. Note that for each $B \in \mathcal{B}$, the quantity $\text{GBR}_{\mathcal{P}} \left(\omega' \mapsto S_{\ell}(\mathcal{P}_{|B}, \omega') \right) | B$ is a constant, and can therefore be pulled outside³¹ to obtain

$$\forall B \in \mathcal{B} : \text{GBR}_{\mathcal{P}}(S_{\ell}(\mathcal{P}_{|B}, \omega) | B) = \text{GBR}_{\mathcal{P}} \left(S_{\ell}(\mathcal{P}_{|B}, \omega) \right) | B,$$

which is true. For the second statement, let Q_B be another $\mathcal{B}$ -measurable and $(\ell, \mathcal{B})$ -calibrated forecast, so

$$\forall B \in \mathcal{B} : \text{GBR}_{\mathcal{P}} \left(S_{\ell}(Q_B, \omega) - \overline{R}_{Q_B}(\omega' \mapsto S_{\ell}(Q_B, \omega')) \right) | B = 0.$$

using Remark 303. Since the forecast is $\mathcal{B}$ -measurable, we can pull the constant outside to obtain

$$\forall B \in \mathcal{B} : \text{GBR}_{\mathcal{P}}(S_{\ell}(Q_B, \omega) | B) = \overline{R}_{Q_B}(S_{\ell}(Q_B, \omega)).$$

We know from weak propriety (Proposition 289) that

$$\text{GBR}_{\mathcal{P}}(S_{\ell}(\mathcal{P}_{|B}, \omega) | B) \leq \text{GBR}_{\mathcal{P}}(S_{\ell}(Q_B, \omega) | B),$$

and therefore $\text{GBR}_{\mathcal{P}}(S_{\ell}(\mathcal{P}_{|B}, \omega) | B) \leq \overline{R}_{Q_B}(S_{\ell}(Q_B, \omega))$. Equivalently, $\overline{H}_{\ell}(\mathcal{P}_{|B}) \leq \overline{H}_{\ell}(Q_B)$.

³¹We have $\overline{R}(X - c) = 0 \Leftrightarrow \overline{R}(X) = c$ when c is a constant.

Finally, we show by example that it is possible to find $(\ell, \mathcal{P}, \mathcal{B}, \mathcal{Q}_B)$ satisfying the respective assumptions so that $\overline{H}_\ell(\mathcal{P}_{|B}) < \overline{H}_\ell(\mathcal{Q}_B)$. Take the trivial partition $\mathcal{B} = \{\Omega\}$ and $\mathcal{P} = \{p\}$. Consider the constant forecast p , which equals the GBR in this case, and a constant (i.e. $\mathcal{B}$ -measurable) competitor forecast q which we assume to be also $(\ell, \mathcal{B})$ -calibrated. Now construct p, q and ℓ in such a way that $\mathbb{E}_p[\ell(a_p^*, \omega)] < \mathbb{E}_p[\ell(a_q^*, \omega)] = \mathbb{E}_q[\ell(a_q^*, \omega)] < \mathbb{E}_q[\ell(a_p^*, \omega)]$, which is clearly possible when $|\mathcal{A}| \geq 2$. That is, q will recommend a sub-optimal action a_q^* (under the p data model), but will agree with p on the risk assessment for this action. Then both p and q will be $(\ell, \mathcal{B})$ -calibrated under a p data model; for p this is obvious; since $\mathbb{E}_p[\ell(a_q^*, \omega)] = \mathbb{E}_q[\ell(a_q^*, \omega)]$, the q forecast is also $(\ell, \mathcal{B})$ -calibrated. To sum up, in this case both p and q are $\mathcal{B}$ -calibrated, but $\overline{H}_\ell(p) < \overline{H}_\ell(q)$. ■

6.7.9 PROOF OF PROPOSITION 306

Proof The total IP score of $\mathcal{Q}_B$ can be decomposed as

$$\overline{R}_\mathcal{P}(S_\ell(\mathcal{Q}_B(X), \omega)) = \overline{R}_\mathcal{P} \left(\sum_B \chi_B(\omega) S_\ell(\mathcal{Q}_B, \omega) \right) \quad (6.64)$$

$$\leq \sum_B \overline{R}_\mathcal{P}(\chi_B(\omega) S_\ell(\mathcal{Q}_B, \omega)) \quad (6.65)$$

$$= \sum_B \sup_{P \in \mathcal{P}} \mathbb{E}_{P(\cdot|B)}[S_\ell(\mathcal{Q}_B, \omega)] P(B) \quad (6.66)$$

$$\leq \sum_B \text{GBR}_\mathcal{P}(S_\ell(\mathcal{Q}_B, \omega)|B) \cdot \overline{R}_\mathcal{P}(\chi_B) \quad (6.67)$$

$$= \sum_B \overline{H}_\ell(\mathcal{Q}_B) \cdot \overline{R}_\mathcal{P}(\chi_B). \quad (6.68)$$

■

6.7.10 PROOF OF PROPOSITION 307

Proof of 1. We want to show: If $\mathcal{Q}_B \subseteq \mathcal{P}_{|B} \forall B$ and $\mathcal{Q}_B \subsetneq \mathcal{P}_{|B}$ for at least one B , then $\mathcal{Q}_B$ cannot be $\mathcal{B}$ -sub-calibrated for all loss functions, therefore in particular not $\mathcal{B}$ -calibrated for all loss functions. Recall that $(\ell, \mathcal{B})$ -sub-calibration requires that

$$\forall B \in \mathcal{B} : \text{GBR}_\mathcal{P}(S_\ell(\mathcal{Q}_B, \omega)) \leq \overline{R}_{\mathcal{Q}_B}(S_\ell(\mathcal{Q}_B, \omega))$$

(see the proof of Proposition 305). To show failure of $(\ell, \mathcal{B})$ -sub-calibration, we therefore exhibit some $B \in \mathcal{B}$ and $\ell \in \mathcal{L}$ so that

$$\text{GBR}_{\mathcal{P}}(S_{\ell}(\mathcal{Q}_B, \omega)) > \bar{R}_{\mathcal{Q}_B}(S_{\ell}(\mathcal{Q}_B, \omega)).$$

If we can find some $Z \in L^{\infty}$ so that $\text{GBR}_{\mathcal{P}}(Z|B) > \bar{R}_{\mathcal{Q}_B}(Z)$, then when defining the trivial $\ell(a, \omega) := Z(\omega)$, ignoring the action entirely, we are done (then $S_{\ell}(\mathcal{Q}_B, \omega) = Z(\omega)$); the optimal action is determined only by tie-breaking (Appendix 6.7.1).

In general, if $\mathcal{Q}, \mathcal{Q}' \in \mathcal{IP}$ are closed convex and $\mathcal{Q} \subsetneq \mathcal{Q}'$, then we have the monotonicity

$$\forall Z \in L^{\infty} : \bar{R}_{\mathcal{Q}}(Z) \leq \bar{R}_{\mathcal{Q}'}(Z),$$

and the inequality has to be strict for at least one $Z^* \in L^{\infty}$, otherwise $\bar{R}_{\mathcal{Q}}$ and $\bar{R}_{\mathcal{Q}'}$ would always coincide, but there is a one-to-one relation between closed convex sets of probabilities and upper expectations (see the footnote in Appendix 6.7.4). But for this Z^* , we obtain in the above that $\text{GBR}_{\mathcal{P}}(S_{\ell}(\mathcal{Q}_B, \omega)) > \bar{R}_{\mathcal{Q}_B}(S_{\ell}(\mathcal{Q}_B, \omega))$, and therefore $\mathcal{Q}_B$ does not satisfy the $(\ell, \mathcal{B})$ -sub-calibration condition on $B \in \mathcal{B}$.

The proofs of 2 and 3 are analogous.

Proof of 2. We want to show: If $\mathcal{P}|_B \subseteq \mathcal{Q}_B \forall B$, then $\mathcal{Q}_B$ is $(\ell, \mathcal{B})$ -sub-calibrated. Upper expectations are monotone in the sense that if $\mathcal{P}|_B \subseteq \mathcal{Q}_B$, then $\bar{R}_{\mathcal{P}|_B}(Z) \leq \bar{R}_{\mathcal{Q}_B}(Z) \forall Z \in L^{\infty}$; recall $\text{GBR}(Z|B) = \bar{R}_{\mathcal{P}|_B}(Z)$. Now take $Z(\omega) := S(\mathcal{Q}_B, \omega)$. For $\mathcal{Q}_B$ to be $(\ell, \mathcal{B})$ -sub-calibrated, we need

$$\forall B \in \mathcal{B} : \text{GBR}_{\mathcal{P}}\left(S_{\ell}(\mathcal{Q}_B, \omega) - \bar{R}_{\mathcal{Q}_B}(\omega' \mapsto S_{\ell}(\mathcal{Q}_B, \omega'))|B\right) \leq 0 \quad (6.69)$$

$$\Leftrightarrow \forall B \in \mathcal{B} : \text{GBR}_{\mathcal{P}}(S_{\ell}(\mathcal{Q}_B, \omega)|B) \leq \bar{R}_{\mathcal{Q}_B}(S_{\ell}(\mathcal{Q}_B, \omega)). \quad (6.70)$$

which holds due to the monotonicity.

Proof of 3. We want to show: if $\exists B \in \mathcal{B} : \mathcal{P}|_B \neq \mathcal{Q}_B$, then $\mathcal{Q}$ cannot be $\mathcal{B}$ -calibrated for all loss functions. Since $\mathcal{P}|_B \not\subseteq \mathcal{Q}_B$ (they are closed convex by assumption, then “ $\neq$ ” implies “ $\not\subseteq$ ”), we can find some $Z \in L^{\infty}$ so that

$$\bar{R}_{\mathcal{P}|_B}(Z) < \bar{R}_{\mathcal{Q}_B}(Z) \text{ or } \bar{R}_{\mathcal{P}|_B}(Z) > \bar{R}_{\mathcal{Q}_B}(Z),$$

(see the proof of Proposition 291 in Appendix 6.7.4). Now set $\ell(a, \omega) = Z(\omega)$. Then

$$\bar{R}_{\mathcal{P}|_B}(S_{\ell}(\mathcal{Q}_B, \omega)) \neq \bar{R}_{\mathcal{Q}_B}(S_{\ell}(\mathcal{Q}_B, \omega))$$

and therefore calibration fails.

6.7.11 RELATION TO CLASSICAL CALIBRATION

While we were differently motivated by the centrality of the decision-theoretic entropy concept, the notion of *marginally desirable gambles* in the IP literature [Walley, 1991, Augustin et al., 2014] as well as an insurantal perspective, it turns out that our notion of IP calibration (approximately) specializes, in the precise case, to *decision calibration* of Zhao et al. [2021a]. To illustrate the connection, we specialize both the data model as well as the forecast to be precise — note that it may also be of interest to consider IP sub-calibration of imprecise forecasts under a precise data model to model epistemic uncertainty.

Remark 319 (compare Definition 2 of Zhao et al. [2021a]). *Let $\ell \in \mathcal{L}$. When $\mathcal{P} = \{P\}$ and the forecast $Q : \mathcal{X} \rightarrow \Delta^k$ is precise in Definition 299 (calibration without groups), we obtain, with $\tilde{S}(\omega) = S_\ell(Q(\omega), \omega)$:*

$$\mathbb{E}_P [\tilde{S}(\omega) - \mathbb{E}_Q[\tilde{S}]] = 0. \quad (6.71)$$

A slight difference in the setup is that for Zhao et al. [2021a], a forecast returns a probability on the set of labels, whereas in our setup a forecast returns a probability on the whole Ω , allowing the forecast $f(X = x)$ even to disbelieve the $X = x$ information. Also, in their setup a loss function has the type signature $\ell : \mathcal{Y} \times \mathcal{A} \rightarrow \mathbb{R}$, and can therefore not directly depend on the elementary event.

Another difference is that Zhao et al. [2021a] allow that some loss function ℓ is used to pick the optimal action (in this case, the Bayes optimal action under ℓ), but that *another* loss function is used to evaluate the consequences of the action.³² We are not aware of a reason to choose different loss functions for selecting the action and evaluating its consequences.

Interestingly, Zhao et al. [2021a] prove the following (here informally stated) proposition:³³ decision calibration holds for all bounded loss functions if and only if distribution calibration holds. When the loss function depends on ω only through Y , distribution calibration demands that $\forall q : \mathbb{E}_P[Y|Q(X) = q] = q$; this is the familiar notion of calibration.

6.7.12 DETAILED INFORMATION ABOUT DATASETS

In this appendix, we provide detailed information about how we obtained the datasets.

³²For example, the proof of part 3 of Theorem 1 (available in Appendix C.1, part 3, in the arxiv version [Zhao et al., 2021b]), uses the log-loss to choose the optimal action (with $\mathcal{A}$ as the probability simplex, the optimal action a_P^* is then the probability P^* itself due to strict propriety of the log-loss), but another loss (in this case any bounded loss) to evaluate the consequences.

³³We are unable to comprehend some steps of the proof.

ACS PUMS

Inspired by Ding et al. [2021], we use census data from the ACS PUMS survey (year: 2018). From the raw data (see the github repository for scripts which download and preprocess the data, we keep the following subset of columns:

AGEP, SCHL, MAR, SEX, ESP, MIG, GCL, HICOV, SCIENGP, DIS, DEAR, DEYE,
DREM, CIT, NATIVITY, ANC, RAC1P, LANX

For the meaning of these features, we refer to the official ACS PUMS documentation, available at https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2018.pdf (Accessed: 2024-10-07). A row in the dataset corresponds not exactly to an individual person, but rather to a *statistical individual*, carrying a statistical weight (see the official ACS PUMS documentation for how the weights are computed; we neglect the weights in our experiments, and treat a row essentially as a person). For example, “AGEP” refers to age, “SCHL” represents level of education. We add two features to the dataset: a binary feature “isHisp”, indicating Hispanic origin in a binary way (based on the raw, more fine-grained “HISP” feature); and the binary “isWhiteOnly” feature, indicating whether a person is “white and not Hispanic”.

As the binary target label, we construct the target based on the raw “ESR” feature. We consider a person as “having a job” ($Y = 1$), if they are either “Civilian employed, at work”, “Civilian employed, with a job but not at work”, “Armed forces, at work” or “Armed forces, with a job but not at work”; contrariwise, we consider a person as “not having a job” ($Y = 0$) if they are either “Unemployed” (unemployed proper) or “Not in labor force”. In this way, we follow Ding et al. [2021], who introduced an employment prediction task with the ACS PUMS data. We remark that, in the economic sense, a person who is “Not in labor force” is not considered to be unemployed.

As Ding et al. [2021], we filter for $16 < \text{age} < 90$. We form 50 groups based on dividing the data into the 50 U.S. states (e.g. we do not include data for Puerto Rico and D.C.). The sizes of the group-wise datasets are highly uneven; we do not perform any balancing, since in practice, we may wish to use all data which is available, without throwing data away to achieve balancing.

In addition, we then used `PolynomialFeatures(degree=2, interaction_only=True, include_bias=False)` from the `sklearn.preprocessing` library to obtain polynomial features, which means introducing interaction terms in the model.

To the reader who wonders about why we have chosen certain features and the task; in the present paper we use all datasets rather in a toy-like fashion, to have realistic data but not in an intentionally meaningful real-world prediction/decision context, but we have previously used the ACS PUMS dataset described here for fairness-related investigations in

a different context.

As a final step in our preprocessing pipeline, for each model which we train, we apply a standard scaler which we fit on the training data.

FRAMINGHAM

We downloaded the dataset from

<https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset>

(Accessed: 2024-08-16). The dataset is of tabular character, with 16 columns (such as gender, age, education, being a smoker, having diabetes etc.), and the target label is the binary “10-year risk of future coronary heart disease” ($Y = 1$ means “at risk” and $Y = 0$ means “not at risk”).

For preprocessing, we imputed missing values following <https://www.kaggle.com/code/kapiluniyara/framingham-dataset>. In contrast to them, we did not remove outliers (an exploratory analysis suggested a very limited effect of this). As in the ACS PUMS dataset, we used `PolynomialFeatures(degree=2, interaction_only=True, include_bias=False)` to obtain polynomial features.

We formed two groups, “young” and “old”, by thresholding the “age” feature at 60 (≤ 60 vs. > 60). For each group, we then use `RandomOverSampler` of the `imblearn.over_sampling` library to balance the frequency of the labels (so approximately 50% have a positive label in each group); otherwise, there are relatively few examples of $Y = 0$ in the dataset. By doing this we intend to illustrate that balancing is not sufficient to remove the influence of asymmetric loss functions (in DRO and in ERM). Note that this way of balancing is different to balancing the group sizes (see Appendix 6.7.12).

We use a random subset of relative size 0.2 as the test set, and the rest as the training data. As a result, the “young” group has 5,091 training examples and 2,567 positive training labels, and the old group has 6,62 training examples and 330 positive training labels. As to the test data, the “young” group has 1,237 examples and 615 positive labels, and the old group has 166 examples and 84 positive labels.

As a final step in our preprocessing pipeline, for each model which we train, we apply a standard scaler which we fit on the training data.

CELEBA

We downloaded the dataset from <https://www.kaggle.com/datasets/jessicali9530/celeba-dataset> (Accessed: 2024-09-03), where the train and test split is also defined (we do not use the validation data). We use the “Wearing_Earrings” attribute as our binary target label, where $Y = 1$ means “wearing earrings” and $Y = 0$ means “not wearing

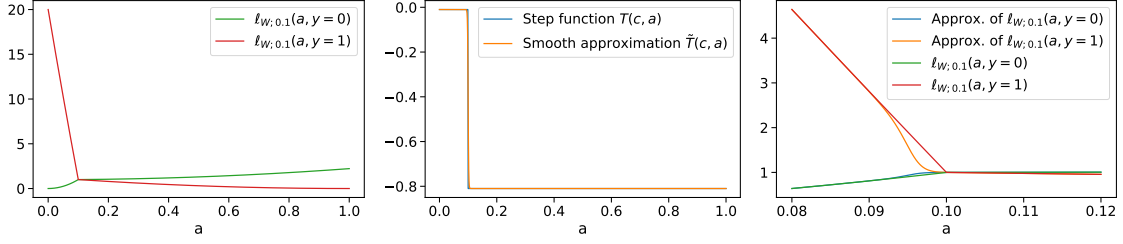


Figure 6.2: Left: a visualization of the asymmetric loss function (the two curves are the partial loss functions, meaning each curve has a fixed outcome $y \in \{0, 1\}$). Middle: the discrete step function and the sigmoid approximation. The difference is hardly discernible. Right: by zooming in near $c = 0.1$, the difference between the original loss function and its approximation can be detected.

earrings”. We divide the data based on the “male” attribute in two groups, “male” and “non-male”, where the latter included non-binary identities.

We use a pretrained Resnet50 from the pytorch library (with the default weights) to obtain embeddings of train and test data, whereby “embedding” we mean computing a forward pass of the model and saving the output of the average pooling layer (just before the final linear layer), which is simply achieved by replacing the final linear layer with the identity mapping. The embeddings are then of dimension 2,048. We then perform a PCA, which we fit on the combined training data, with 100 components to reduce the dimensionality.

From the train data, we used only the first 501 batches of size 64 for each group, giving 32,064 training examples for each group. We use all the available test data (7,715 examples for the “male” group and 12,247 for the “non-male” group). The “male” group then has 480 positive train examples and 182 positive test examples; the “non-male” group has 12,247 positive train examples and 3,943 positive test examples. In this case, we do not balance the label frequency to have this as a contrasting example to framingham.

As a final step in our preprocessing pipeline, for each model which we train, we apply a standard scaler which we fit on the training data.

6.7.13 SIGMOID APPROXIMATION OF THE ASYMMETRIC SCORING RULE

When specializing to the Brier (squared) loss, the asymmetric scoring rule proposed by [Winkler \[1994\]](#) is given by:

$$\ell_{W;c}(a, Y(\omega) = y) = 1 - \frac{\ell^2(a, y) - \ell^2(c, y)}{T(c, a)}; \quad T(c, a) := \begin{cases} -\ell^2(c, 1) & a \geq c \\ -\ell^2(c, 0) & a < c \end{cases}; \quad \ell^2(a, y) := (a - y)^2.$$

This loss function is depicted in Figure 6.2 (left) for $c = 0.1$. At the point c , this loss is not differentiable due to the discrete step function $a \mapsto T(c, a)$. We therefore use the

following variant in practice:

$$\tilde{T}(c, a) := -\alpha \ell^2(c, 1) - (1 - \alpha) \ell^2(c, 0), \quad \alpha := \sigma(f \cdot (a - c)),$$

where $\sigma(z) = 1/(1 + \exp(-z))$ is the sigmoid function, and we choose $f = 1000$ for a close approximation. With such a high value of f , the difference between the step function and its approximation is hardly discernible visually (see Figure 6.2 (middle)). By zooming in around c , the small difference in the loss function can be discerned (Figure 6.2 (right)). Since the difference between the original loss function and its smooth approximation is tiny and appears negligible in practice, we do not introduce new notation for the approximation (which we use in the experiments).

6.7.14 ENTROPY COMPARISON

Figure 6.3 compares the unconditional entropy for the loss functions $\ell_{0.1}$ and $\ell_{W,0.1}$ (recall Section 6.4.1).

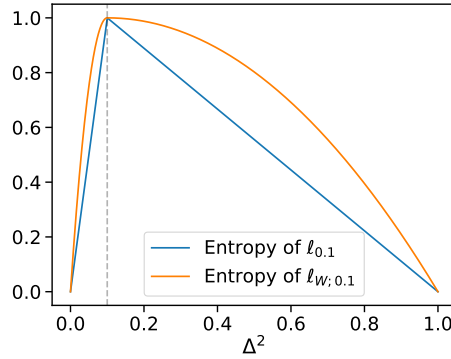


Figure 6.3: Comparison of unconditional entropies $\hat{H}_{\ell_c}$ and $\hat{H}_{\ell_{W,c}}$ (rescaled to obtain maximum value of 1) of ℓ_c and $\ell_{W,c}$ with $c = 0.1$. The x-axis corresponds to the set of probabilities Δ^2 on the binary Y .

6.7.15 DETAILED INFORMATION ABOUT TRAINING PROCEDURES

Our models have the following structure: a linear layer followed by a sigmoid, which when the input data is $X \in \mathbb{R}^{N \times D}$ outputs a precise probabilistic forecast $Q \in \mathbb{R}^{N \times 2}$ (where the forecast for a single datum is in Δ^2 , since $Y \in \{0, 1\}$). This forecast is then fed into a loss function of signature $\ell : \Delta^2 \times \{0, 1\} \rightarrow \mathbb{R}$.

ERB(log): For the ERM model, the training objective is:

$$\min_{Q \in \text{Lin}} \mathbb{E}_{\hat{P}} [S_{\ell}(Q(X), Y)],$$

where $\hat{P}$ is the empirical distribution of the stacked data, i.e. the combined data of all groups. We train using the Adam optimizer (with the pytorch default options) for 5,000 iterations and a learning rate of 0.001. If we use batches (see below), then we draw a random batch at each iteration.

GBR: To obtain the GBR, we perform a similar procedure, where for each group we train an individual model, meaning

$$\min_{Q_i \in \text{Lin}} \mathbb{E}_{\hat{P}_i} [S_\ell(Q_i(X), Y)],$$

where $\hat{P}_i$ is the empirical train distribution (the training data) of group i . We then use

$$\mathcal{Q}(X) := \{Q_i(X) : i = 1..G\}$$

as an estimate of the GBR. We also train with the Adam optimizer (with the pytorch default options) for 5,000 iterations and a learning rate of 0.001. As above, if we use batches (see below), then we draw a random batch at each iteration.

DRO(log) and **DRO(c)**: We apply the maximum entropy principle, so the minimization of the IP score can be expressed as maximizing entropy:

$$\max_{\lambda \in \Delta^G} \min_{Q \in \text{Lin}} \mathbb{E}_{\hat{P}_\lambda} [S_\ell(Q(X), Y)], \quad \hat{P}_\lambda := \sum_{i=1}^G \lambda_i \hat{P}_i,$$

where $\lambda = (\lambda_1, \dots, \lambda_G)^\top \in \Delta^G$. The outer (maximizing) routine is described in Section 6.4.2; more concretely, we perform 2,000 such outer iterations with a learning rate of $\eta = 0.1$. In our examples, we have observed that this choice worked well as a compromise between stability and updating λ sufficiently.

For each outer (maximizing) iteration, we approximate the inner minimization for fixed λ . We do this using the Adam optimizer (with the pytorch default options) for 500 iterations with a learning rate of 0.001. At each iteration, we draw a random batch (if we use batches) for each group, or use the full data for each group (if we don't use batches), and then compute the mean loss on the training distribution for each group. Weighting by the currently fixed λ then gives an estimate of $\mathbb{E}_{\hat{P}_\lambda} [S_\ell(Q(X), Y)]$.

At the end of the inner (minimizing) routine, we need to estimate the gradient $\nabla \lambda$, which simply corresponds to the current mean group losses. When using batches, these can be highly variable, so we instead compute the group losses for multiple batches simultaneously to reduce the variance in our estimate for $\nabla \lambda$. In our experiments, we fixed the number of these batches to 10.

Batch Training: For ACS PUMS and CELEBA, we use a relatively large batch size of 512 to

Predictor	$\ell_{0.1}$	$\ell_{0.3}$	$\ell_{0.7}$	$\ell_{0.9}$	Predictor	$\ell_{0.1}$	$\ell_{0.3}$	$\ell_{0.7}$	$\ell_{0.9}$
GBR	0.0502	0.1477	0.1412	0.0497	GBR	0.0516	0.1535	0.1376	0.0482
DRO(0.1)	0.0662	0.1522	0.1438	0.0666	DRO(0.1)	0.0737	0.1599	0.1831	0.0916
DRO(0.3)	0.0487	0.1333	0.1370	0.0586	DRO(0.3)	0.0500	0.1430	0.1542	0.0771
DRO(0.7)	0.0509	0.1465	0.1316	0.0491	DRO(0.7)	0.0545	0.1500	0.1524	0.0536
DRO(0.9)	0.0497	0.1461	0.1666	0.0530	DRO(0.9)	0.0518	0.1488	0.1922	0.0886
ERM(log)	0.0500	0.1459	0.1576	0.0492	ERM(log)	0.0507	0.1452	0.1759	0.0705
DRO(log)	0.0492	0.1429	0.1347	0.0507	DRO(log)	0.0508	0.1434	0.1548	0.0633

Table 6.2: Train (left) and test (right) IP Scores of different forecasts (rows), evaluated under different loss functions (columns) on FRAMINGHAM data.

Predictor	$\ell_{0.1}$	$\ell_{0.3}$	$\ell_{0.7}$	$\ell_{0.9}$	Predictor	$\ell_{0.1}$	$\ell_{0.3}$	$\ell_{0.7}$	$\ell_{0.9}$
GBR	0.0878	0.1323	0.0919	0.0306	GBR	0.0880	0.1375	0.0966	0.0322
DRO(0.1)	0.0660	0.1362	0.0896	0.0310	DRO(0.1)	0.0672	0.1396	0.0957	0.0326
DRO(0.3)	0.0861	0.1282	0.0874	0.0306	DRO(0.3)	0.0863	0.1280	0.0913	0.0324
DRO(0.7)	0.0874	0.1446	0.0851	0.0306	DRO(0.7)	0.0880	0.1449	0.0893	0.0323
DRO(0.9)	0.0869	0.1435	0.0855	0.0302	DRO(0.9)	0.0869	0.1442	0.0902	0.0324
ERM(log)	0.0769	0.1487	0.0899	0.0308	ERM(log)	0.0774	0.1559	0.0945	0.0326
DRO(log)	0.0882	0.1450	0.0855	0.0306	DRO(log)	0.0886	0.1452	0.0909	0.0324

Table 6.3: Train (left) and test (right) IP Scores of different forecasts (rows), evaluated under different loss functions (columns) on CELEBA data.

improve stability. For FRAMINGHAM, we use the full data instead of batches due to its manageable size.

6.7.16 ADDITIONAL RESULTS

In this appendix, we provide additional results for the datasets, which are described in Appendix 6.7.12.

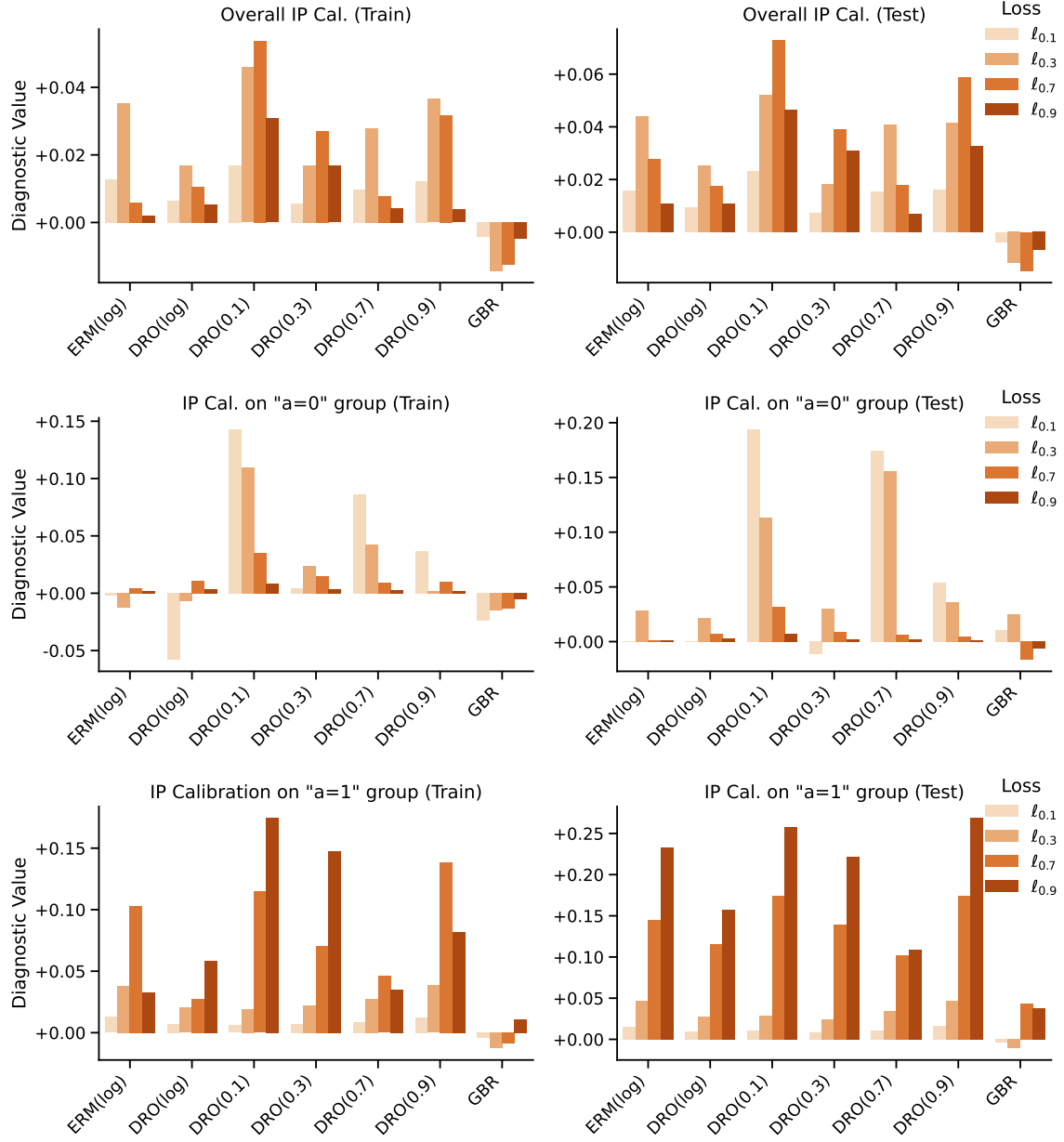


Figure 6.4: Evaluation of IP calibration under the train (left column) and test (right column) data models on FRAMINGHAM. In each subplot, the forecasts $\mathcal{Q}$ vary along the x -axis and the loss function used for evaluation corresponds to a hue. The top row shows diagnostic values for IP calibration without groups, that is, the y -axis shows $\bar{R}_{\hat{P}}(\hat{\xi}_{\ell}(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\hat{\xi}_{\ell}))$ (top left) and $\bar{R}_{\hat{P}_{\text{test}}}(\hat{\xi}_{\ell}(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\hat{\xi}_{\ell}))$ (top right). In the other subplots, the y -axis shows the diagnostics for IP calibration with respect to the action-induced partition (of Equation 6.9) under $\hat{P}$ (left) and $\hat{P}_{\text{test}}$ (right), respectively.

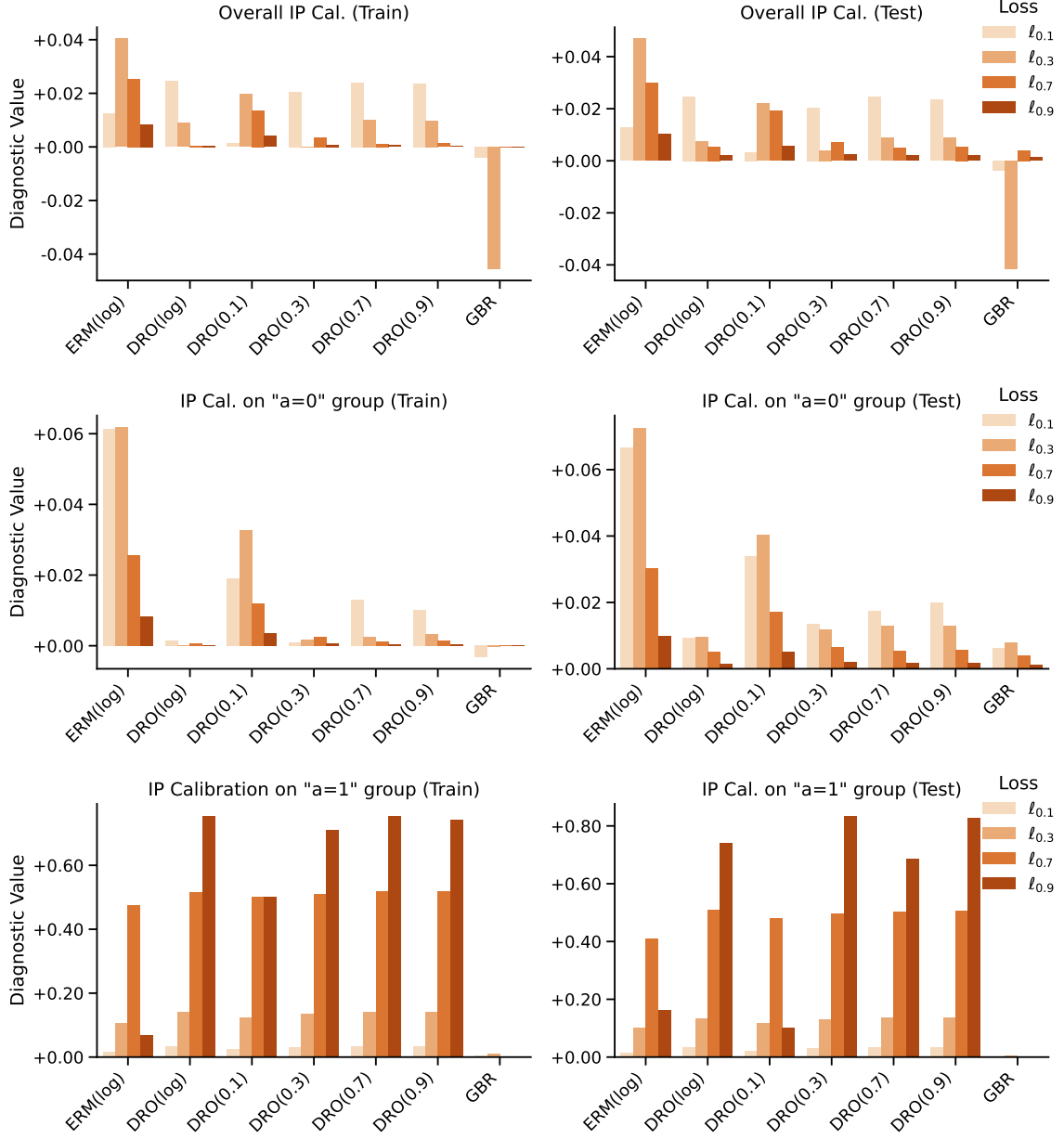


Figure 6.5: Evaluation of IP calibration under the train (left column) and test (right column) data models on CELEBA. In each subplot, the forecasts $\mathcal{Q}$ vary along the x-axis and the loss function used for evaluation corresponds to a hue. The top row shows diagnostic values for IP calibration without groups, that is, the y-axis shows $\bar{R}_{\hat{P}}(\hat{\xi}_{\ell}(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\hat{\xi}_{\ell}))$ (top left) and $\bar{R}_{\hat{P}_{\text{test}}}(\hat{\xi}_{\ell}(\omega) - \bar{R}_{\mathcal{Q}(\omega)}(\hat{\xi}_{\ell}))$ (top right). In the other subplots, the y-axis shows the diagnostics for IP calibration with respect to the action-induced partition (of Equation 6.9) under $\hat{P}$ (left) and $\hat{P}_{\text{test}}$ (right), respectively.

6.7.17 THE LINK TO ZHAO AND ERMON [2021]

To formally relate our IP calibration to the framework of Zhao and Ermon [2021], consider the case of a binary event ω and an imprecise forecast $\mathcal{Q} = [\underline{q}, \bar{q}] \subseteq [0, 1]$ and the cost-sensitive loss ℓ_c . Then, $\bar{R}_{\mathcal{Q}}(\ell(a_{\mathcal{Q}}^*, \omega)) = (1 - c)\bar{q}$ if $a_{\mathcal{Q}}^* = 0$, and $\bar{R}_{\mathcal{Q}}(\ell(a_{\mathcal{Q}}^*, \omega)) = c(1 - \underline{q})$ if $a_{\mathcal{Q}}^* = 1$. Zhao and Ermon [2021] suggest using the stakes $b = \ell(a, 1) - \ell(a, 0)$ (their Proposition 2) in the marginally desirable gamble $\text{MG} := b(Y - \mu) - |b|c$; the gambles of this form are marginally desirable if $[\underline{q}, \bar{q}] = [\mu - c, \mu + c]$ which we now assume. Now consider setting the stakes b based on the action recommendation $a_{\mathcal{Q}}^*$. The agent will then shoulder the loss:

$$L := \ell(a_{\mathcal{Q}}^*, \omega) - \text{MG}.$$

First, note that MG can be rewritten as:

$$\begin{cases} +b(\omega - \bar{q}) & b \geq 0, \\ -b(\underline{q} - \omega) & b < 0 \end{cases}$$

We now compute L (plugging in the respective stakes b) for all combinations of $a_{\mathcal{Q}}^*$ and ω .

If $a_{\mathcal{Q}}^* = 1$ and $\omega = 0$:

$$L = \ell(0, 0) - (1 - c)(0 - \bar{q}) = (1 - c)\bar{q} = \bar{R}_{\mathcal{Q}}(\ell_{a_{\mathcal{Q}}^*}).$$

If $a_{\mathcal{Q}}^* = 1$ and $\omega = 1$:

$$L = \ell(0, 1) - (1 - c)(1 - \bar{q}) = (1 - c)\bar{q} = \bar{R}_{\mathcal{Q}}(\ell_{a_{\mathcal{Q}}^*}).$$

If $a_{\mathcal{Q}}^* = 0$ and $\omega = 0$:

$$L = \ell(1, 0) - c(\underline{q} - 0) = c(1 - \underline{q}) = \bar{R}_{\mathcal{Q}}(\ell_{a_{\mathcal{Q}}^*}).$$

If $a_{\mathcal{Q}}^* = 0$ and $\omega = 1$:

$$L = \ell(1, 1) - c(\underline{q} - 1) = c(1 - \underline{q}) = \bar{R}_{\mathcal{Q}}(\ell_{a_{\mathcal{Q}}^*}).$$

Hence the agent will, in any case, shoulder the constant loss $\bar{R}_{\mathcal{Q}}(\ell_{a_{\mathcal{Q}}^*})$.

Part III

An Actuarial Perspective

7

Insights From Insurance for Fair Machine Learning

AUTHOR CONTRIBUTIONS

Christian Fröhlich (CF) first learned about the idea that insurance could be an interesting analogon to machine learning with respect to fairness issues in a lecture given by Robert Williamson (RW) in autumn 2021, along with a helpful literature pointer. RW also made CF aware of the fundamental tension between aggregate and individual. RW contributed through supervision of the project, e.g. through discussions and proof-reading. The remaining work is due to CF.

ABSTRACT

We argue that insurance can act as an analogon for the social situatedness of machine learning systems, hence allowing machine learning scholars to take insights from the rich and interdisciplinary insurance literature. Tracing the interaction of uncertainty, fairness and responsibility in insurance provides a fresh perspective on fairness in machine learning. We link insurance fairness conceptions to their machine learning relatives, and use this bridge to problematize fairness as calibration. In this process, we bring to the forefront two themes that have been largely overlooked in the machine learning literature: responsibility and aggregate-individual tensions.

7.1 INTRODUCTION

Insurance is “interestingly uninteresting”.¹ In this work, we argue that in fact insurance is far from uninteresting and indeed a rich source of inspiration and insight to scholarship interested in social issues surrounding machine learning, specifically the field now known as fair machine learning. Our proposal is that insurance can be viewed as an analogon to machine learning with respect to these issues arising from the social situatedness. While machine learning is a relatively recent technology, debates regarding social issues in the context of insurance have been ongoing for a long time. Thus, we argue that taking inspiration from studies of insurance can contribute to a more integrative view of machine learning systems as *socio-technical* systems [Selbst et al., 2019].

Both machine learning and insurance are firmly based on a statistical, probabilistic mode of reasoning — an *actuarial* mode. Indeed, insurance can be viewed as the first commercial test of probability theory [Gigerenzer et al., 1989, McFall, 2011]. Insurance, a technology for *doing risk*, transforms uncertainty into calculable risk [Lehtonen and Van Hoyweghen, 2014]. The key idea is to share the risk of a loss in a collective, organized through an abstract mutuality; due to the ‘law’ of large numbers, uncertainty thus becomes manageable and the effect of chance can be offset [Ewald, 1991]. In this way, insurance creates a “community of fate” in the face of uncertainty [Heimer, 1985]. To enter into this community (the insurance *pool*), the insurer demands a certain fee, called *premium*, from the policyholder.

In insurance, questions of fairness inevitably arise, and have been the subject of much debate. The central point of debate is the tension between risk assessment and distribution [Abraham, 1985]. In other words, who is to be mutualized in the pool. Some form of segmentation is found in many insurantal arrangements: the pool of policyholders can be stratified by separating high and low risk individuals. But the specific nature that such segmentation takes typically depends not only on risk assessment, but on further considerations such as assignment of responsibility, modulated by social context; in this way, insurance is not a neutral technology [Baker and Simon, 2002, Glenn, 2003a].

Our non-comprehensive outline of the history of insurance illustrates how uncertainty, fairness and responsibility interact, and can be entangled and disentangled. From this background, we can extract conceptual insights which also apply to machine learning. The tension between risk assessment and distribution is mirrored in formal fairness principles: *solidarity*, which can be linked to independence in fair machine learning, contrasts with *actuarial fairness*, linked to calibration. Briefly, actuarial fairness demands that each policyholder should pay only for their own risk, that is, mutualization should occur only between individuals with the same ‘true’ risk. In contrast, solidarity calls for equal contri-

¹McFall et al. [2020] call insurance “interestingly uninteresting”, referring to how insurance is “hugely underresearched” given its societal importance, which is typically not recognized [Ewald, 1989].

bution to the pool. On one level of this text, we problematize actuarial fairness (by extension, calibration) as a notion of fairness in the normative sense by taking inspiration from insurance. This perspective is aligned with recent proposals that stress the discrepancy of formal algorithmic fairness and “substantive” fairness [Green, 2022], which some prefer to call *justice* [Vredenburg, 2022]. Parallel to this runs a distinct textual level, where we emphasize two intricately interacting themes: *responsibility* and *tensions between aggregate and individual*. Both entail criticism of actuarial fairness, but we suggest that they additionally provide much broader, fruitful lessons for machine learning from insurance.

At the highest level of abstraction, our goal is to establish a general conceptual bridge between insurance and machine learning. Traversing this bridge, machine learning scholars can obtain new perspectives on the social situatedness of a probabilistic, statistical technology — we attempt to offer a new ‘cognitive toolkit’ for thinking about the social situatedness of machine learning. Our point of view is that fairness cannot be reduced to a formal, mathematical issue, but that it requires taking broader social context into account, reasoning for instance about responsibility. And for this, we suggest, insurance is an insightful analogon. Therefore, our objective is to furnish the reader with a guide that charts the landscape of insurance with respect to social issues and to establish links to machine learning.

On a formal level, we use the following analogy. In a machine learning task, we are given some features X and associated outcomes Y , which we attempt to approximate by predictions $\hat{Y}$. The structural relation to insurance is established by conceiving of X as the features of policyholders (e.g. age, gender) with outcomes Y (e.g. having an accident or not), and the task is to set a corresponding premium $\hat{Y}$.

7.2 A BRIEF HISTORY OF INSURANCE RATIONALITIES

Insurance is not a monolithic technology, but rather a general principle of risk management, which is instantiated in multiple distinct forms. Insurance and the conceptual resources it deploys are not immutable and stable over time, as exemplified by Baker’s [1996] study on *moral hazard*, attesting to their evolving nature. In this section, we provide a succinct, necessarily non-comprehensive description of three historical modes of insurantal operation: the welfare state, neoclassical economics and personalized insurance. Throughout, we focus on the role that uncertainty, fairness and responsibility play in each of the three modes. To each mode we ascribe a set of attitudes towards these three aspects and in what manner they are entangled or disentangled. Fairness conceptions in insurance are contingent upon prevailing societal norms, particularly regarding responsibility, but concurrently insurance shapes the moral fabric of the society in which it is embedded [Glenn, 2003b, Van Hoyweghen et al., 2006, Lehtonen and Liukko, 2015]. Furthermore, fairness

conceptions in insurance are historically intertwined with their accompanying (statistical) epistemologies, ways of ‘knowing’ the risk in the face of uncertainty. A common thread is also that distinct forms of insurance correspond to distinct ways of governing society. Importantly, we do not want to suggest a linear historic progression here² — different forms of insurance co-exist at any given time. For instance, contemporary health care systems tend to operate with the logic of the welfare state, while actuarial fairness undergirds the private insurance sector. What follows is our synthesis of the literature, particularly drawing on the works of Ewald [1986] and Frezal and Barry [2020], with a discerning focus on elucidating the intricate interplay between uncertainty, fairness, and responsibility. With this background, we are then able to extract conceptual lessons that apply also to machine learning.

7.2.1 BROAD SOLIDARITY IN THE WELFARE STATE

In his seminal work *L’État providence*, Ewald [1986] gave an influential account of the rise of the welfare state and explicates how insurance became its prime way of government. The point of departure is the predominance of liberal reasoning, which operates with the categories of *fault* and *foresight* in risk management. Here, it is presumed that the occurrence of an accident (a damage, loss or injury) must be due to fault, a lack of foresight. Liberal thought held individuals responsible for their own fate and inequalities were naturalized as just consequences of individual responsibility [Landes, 2013]. An accident implies a trial under the regime of juridical responsibility [Ewald, 1986], where the goal is trying to establish the fault of one party. Responsibility is then borne by the person who is assigned fault, who is seen as having caused the accident. In turn, voluntary charity is the preferred means of supporting the poor and unlucky. With the rise of industrialization in different countries, an “accident crisis” unfolded roughly between 1870 and 1910 [Krippner, 2023]: the number of workplace accidents dramatically increased and there appeared a new regularity at the aggregate level, suggesting a kind of determinism in the phenomenon. This led to the *objectification* of the accident and its management became a question of the collective, the *social*. By objectification, we understand an act of aggregation combined with the law of large numbers. As a consequence, we find a conception of insurance as broad solidarity and the rise of the welfare states. While at the individual level it was impossible to predict *who* will suffer an accident, the new regularity observed at the aggregate level could provide an effective pattern of risk management. Since equal ignorance in the fate of uncertainty was emphasized [Ewald, 1986], solidarity, implying equal contribution to the pool, was considered fair. Indeed, this conception of insurance was so firmly based on the aggregate that it led Ewald [1991] to assert that

²See Baker [2000, p. 571] on Ewald’s work for this point.

Strictly speaking there is no such thing as an individual risk; otherwise insurance would be no more than a wager. Risk only becomes something calculable when it is spread over a population. The work of the insurer is, precisely, to constitute that population by selecting and dividing risks. [...] It makes each person a part of the whole.

The business of insurance, then, was construed as the constitution of abstract mutualities³ [Lehtonen and Liukko, 2011] and the sharing of responsibility to counteract the effect of fate.

The epistemology of the welfare state is one of the aggregate, the collective, the social. In this respect, it is interesting and instructive how insurance became intertwined with probability and statistics. Early forms of insurance were more like *gambling*, and insurance was often accused of being immoral and faced prohibitions.⁴ Insurance gained more legitimacy when it became based on ‘objective’ probability and statistics [Daston, 2023, p. 162ff]; by the end of the nineteenth century, the morality of insurance was established [Baker, 1996]. The crucial conceptual move was the marriage of probability theory and statistics. While early probabilists were more concerned with reasonable subjective judgment, the nineteenth century shows an increasing shift towards “objective calculation” [McFall, 2011]. The idea was that in a large insurance pool losses occur randomly, so the law of large numbers applies and the total loss can be predicted at the level of the aggregate. Frequentist (‘objective’) probability thus combines aggregate regularity with individual irregularity, as explained by Venn [1876, p. 4].

A key player in this development was Adolphe Quetelet, a Belgian astronomer, who initiated the study of “social physics” by attempting to discover natural laws about human behaviour [McFall, 2011]. Quetelet’s innovation was the transposition of one sense of the concept *average* to a different one. Consider first the familiar aggregating sense of average, where a set of commensurate objects is summarized in a single number. A *prima facie* different sense of average is as the single true value of some measurement problem, from which one can obtain a set of noisy measurements. The radical conceptual move was then to transpose the second to the first sense, thereby viewing the individuals of a population as many realizations of some abstract *average human*⁵ [Ewald, 1990, McFall, 2011]. In

³Typically policyholders are unaware with whom they are mutualized, however, so this mutualization does not require a shared sense of groupness [Krippner and Hirschman, 2022].

⁴Cooper and Grinder [2009] provide some examples. In 18th century London life insurance could be bought on the life of celebrities, without an insurable interest to the policyholder. In fact, the notion of an insurable interest was put forward by the insurers to counter the allegation of gambling [Baker, 1996, McFall and Moor, 2018]. Particularly interesting is also insurance in Islamic law, which prohibits gambling and contracts based on usury: the morality of insurance is justified then by emphasizing the solidaristic nature of the arrangement [Baker, 2002], in contrast to the view of insurance as a bilateral contract that is more prevalent in Western societies.

⁵The original term was “average man”.

this way, the rates of birth, death and other social phenomena could be attributed to this fictional average human. In the context of the workplace accident and insurance, in line with the imaginary of the average human, the focus shifted from the unique, individual experience (the object of a juridical trial) to an objectification based on the average occurrence [Krippner, 2023]. With regard to how this influences the notion of personhood, Dean [1998] writes

Insurance practices displace the abstract, invariant norm of a responsible juridical subject with an individuality relative to other members of an insured population, an ‘average sociological individuality’.

Quetelet’s fiction of the average human has made an impact on the insurance sector [McFall, 2011]: identifying the individual with an average is at the core of actuarial practice. Indeed, the question of how individuals relate to the aggregates they make up is, as we suggest in line with Krippner [2023], runs through the history of insurance. Moreover, we argue in Section 7.5 that it is a major concern for fair machine learning, too.

Although insurance increasingly relied on probability and statistics, the available quantification methods at the level of the collective severely limited the possibility of actually ‘knowing’ the risk of an individual [Barry, 2019]. With improved actuarial methods, new possibilities for segmentation of the pool have opened up, which have led also to the rise of a new fairness notion that is successively contributing to the erosion of solidarity.

7.2.2 NEOCLASSICAL ECONOMICS AND ACTUARIAL FAIRNESS

A distinct mode of insurance, which undergirds contemporary (private) insurance, is based on neo-classical economics, which construes individuals as rational expected utility maximizers — the human is viewed as a *homo oeconomicus*. The assumption in this paradigm is that insurance is purchased due to risk aversity from the perspective of the policyholder, while the insurer is risk neutral. This configuration of the individual is tied to a different notion of fairness which contrasts starkly with the solidarity of the welfare state: *actuarial fairness*. The idea is that the pure premium (i.e. what the policyholder pays before adding additional expenses such as for administration) should equal the expected risk for each policyholder. While the idea of “equality in risk” was around for a long time [Heras et al., 2020], its modern formulation is due to Arrow [1963]. On the one hand, actuarial fairness can be understood as a purely descriptive, technical notion; but it is also advanced in the literature and by insurers as a notion of *fairness* in a normative sense, as legitimate practice, see for instance [Walters, 1981, Clifford and Iuculano, 1987, Daniels, 1990, Stone, 2001, Thiery and Van Schoubroeck, 2006]; Interestingly, actuarial fairness can be traced back to the Aristotelian consistency principle “fairness is to treat equal people equally and

unequal people unequally” [Landes, 2015], in a context where ‘equal’ means ‘equal risk’ [Heras et al., 2020].

The definition and especially the Aristotelian motivation betrays that actuarial fairness is in practice always a *group-based* notion [Miller, 2009], since insurers needed (traditionally, at least) to make use of large segments for calculating expected losses. For this, actuaries choose a set of relevant variables, while ignoring others. This seems fundamentally in conflict with the idea of adjustment to the ‘individual risk’ of the policyholder — and indeed this was not how actuarial fairness was construed until roughly the 1970s, it remained firmly group-based [Barry, 2020]. In line with Quetelet, Thiery and Van Schoubroeck [2006] describe the logic succinctly as follows:

[I]nsurance classification schemes rely on the assumption that individuals answer to the average (stereotypical) characteristics of a group to which they belong.

Hence the justification for group-based actuarial fairness relies on assuming an “average sociological individuality” [Dean, 1998] — each individual is assigned to a segment and is then identified with the corresponding average [Krippner and Hirschman, 2022]. In this sense, we find again the logic of the welfare state but now only *segmentwise*, with the aspiration to reduce solidarity between groups as much as possible, given practical constraints. In the terminology of Lehtonen and Liukko [2011], this means that ideally only *chance solidarity* is left, which compensates for the effects of aleatoric uncertainty;⁶ in contrast, *risk subsidizing solidarity* refers to a solidarity between individuals of different expected loss.⁷

The role of responsibility in actuarial fairness is subtle. Actuarial fairness, when understood as a normative principle, rests on the assumption that people can be held responsible for their individual risk to some extent [Lehtonen and Liukko, 2015]. However, we should distinguish conceptually between *responsibility* and *responsibilization*, that is, *holding* someone responsible for something.⁸ While normative philosophical literature may be careful about this distinction, in practice it can appear blurry. In fact, there are two principles that provide *non-responsibility based reasons for responsibilization* in insurance [Andersen and Nielsen, 2015], therefore in favor of actuarial fairness: *moral hazard* and *adverse selection* (see Appendix 7.8.1). Thus, responsibility is, in the neoclassical framework, not

⁶We put aside for now the question of whether the conceptual distinction between aleatoric and epistemic uncertainty is well-defined.

⁷Lehtonen and Liukko [2011] further mention *subsidizing income solidarity*, occurring when premia are adjusted based on income; this is more like a tax than ‘genuine’ insurance solidarity.

⁸We use the term ‘responsibilization’ in line with Andersen and Nielsen [2015]. The term originally appeared in the governmentality literature and refers to a neoliberal mode of governing that frames individuals as autonomous and responsible, see for example [Shamir, 2008, Pyysiäinen et al., 2017].

central to actuarial fairness [Landes, 2015]. However, the role of responsibility (more precisely, responsibility-based reasons for responsabilization) in insurance is currently being emphasized more and more. To this we turn now.

7.2.3 THE CLIMAX: PERSONALIZED INSURANCE

From the 1980s on, the insurance industry was increasingly challenged by anti-discrimination legislation. Social movements attacked the average human (woman) logic that insurance based its actuarially ‘fair’ premia on. A prominent example is the campaign initiated by the National Organization for Women (NOW) [Krippner and Hirschman, 2022, Krippner, 2023], aimed at ending gender-based risk segmentation. In line with the civil rights movement, feminists considered such underwriting (that is, risk classification) practices to be unfairly discriminatory, as they rely on group-based generalizations. Instead, they asked for a finer adjustment to *individual risk*. What was under attack here is the fundamental group-based logic of actuarial fairness. The US supreme court asserted in the context of insurance that “[e]ven a true generalization about [a] class cannot justify class-based treatment” (Norris, 1983, as cited in Avraham [2018]). A more recent case is the *ECJ Test-Achats* ruling in the EU, which highly restricts gender-based underwriting [Rebert and Van Hoyweghen, 2015, Cevolini and Esposito, 2022]. In response to such anti-discrimination legislation and in anticipation of a continuation of this trend, the practical meaning of actuarial fairness gradually began to shift. Consequently we find two highly entangled trends in the contemporary insurance industry: the *individualization of risk* and *behaviour-based personalization* [Cevolini and Esposito, 2020, McFall et al., 2020].

The individualization of risk can be understood as taking group-based actuarial fairness to the limit and (hypothetically) forming ‘groups of one’; instead of spreading risk over a pool, each policyholder would pay exactly for her own *individual risk* [Cevolini and Esposito, 2020]. In terms of solidarity, this implies a complete erosion of subsidizing risk solidarity, so that ideally only chance solidarity remains. Rephrasing this, what is now being emphasized is the non-homogeneity within previous groups of policyholders [Barry, 2020]. To this end, insurers begin to shift the focus from attributes which are considered uncontrollable (e.g. gender, race)⁹ to controllable, dynamic data about the individual, and adjust premia accordingly, yielding behaviour-based personalization. Of course, the hope is that behaviour is closely linked to the individual risk, otherwise personalization would hardly be reasonable; in this way individualization and personalization are correlated.¹⁰

Personalization is linked to *InsurTech*, that is, technology-driven innovation in insurance

⁹While the case of gender demonstrates that what is considered controllable can change, for insurance purposes, gender is arguably still uncontrollable.

¹⁰One would (in most contexts) not try to personalize premia based on the binary feature ‘having attached earlobes’.

[McFall et al., 2020]. A prominent example is the use of wearable devices such as fitness trackers in health insurance [Lupton, 2016, McFall, 2019], where discounts and rewards are supposed to incentivize “healthy behaviour”. In car insurance, the use of telematics is gaining popularity [Verbelen et al., 2018, Meyers and Van Hoyweghen, 2018, Cevolini and Esposito, 2022], where a small device installed in the car dynamically provides information about driving behaviour from proxy variables such as speed to the insurer, as well as feedback to the policyholder. Here, the premium is continuously adjusted to behaviour, which contrasts with previous static underwriting. In both examples, the premium is supposed to act *on* the behaviour with the aim of loss prevention [McFall and Moor, 2018]; this opens up the possibility for feedback loops. In other words, the premium is *performative* — see Appendix 7.8.4.

The accompanying epistemology associated with the shift towards individualization and personalization is the aim to tailor the premium to the ‘individual risk’ and it is assumed that a combination of big data (often behavioural, with a fine temporal resolution) and machine learning enables ‘knowing’ this risk [Cevolini and Esposito, 2020]. Hence, in this currently unfolding chapter, machine learning enters into a dynamic interaction with insurance; we expect conceptual lessons to flow in both directions in the future (compare also [Williamson, 2004]); in this paper, however, we specifically focus on lessons from insurance for machine learning. Barry [2019] describes the new epistemology as follows:

Hence what was once considered as ‘noise,’ the individual specificities that had to be averaged out by statistics, is now the core of the analysis and the focus of the new knowledge.

The upshot, according to Barry [2019], is the “deconstruction of the aggregate viewpoint that produced collectives”. Commentators speak of emerging “segments of one” [Prainsack and Van Hoyweghen, 2020].

The individualization and personalization of risk is associated with a shift in fairness: actuarial fairness is taken to the limit and now clearly carries a normative flavour based on a linkage to *responsibility*. We call this utopia of individual risk adjustment *perfect actuarial fairness*, to demarcate it from practical, group-based actuarial fairness. Reviving pre-Welfare liberal thought, individual responsibility is stressed [Dean, 1998]. For example, Ericson et al. [2000] document a shift in the concept of accident; the new rhetoric, speaking of a “crash” in the case of a car accident, underscores that *someone must be at fault* and thus responsible. Without taking a philosophical stance on actual responsibility, the trend is one of the *responsibilization* of the individual. Even in the context of health insurance, individuals face such responsibilization [Van Hoyweghen et al., 2006, Prainsack and Van Hoyweghen, 2020]. For example, self-tracking favors a view of individuals as “managers” of their health [Lupton, 2016, Sharon, 2017]. Overall, responsibilization is linked to

neoliberal modes of government [Dean, 1998, Ericson et al., 2000, Meyers and Van Hoyweghen, 2018].

As a consequence, many commentators argue that personalization undermines solidarity [Rosanvallon, 2000, Prainsack and Van Hoyweghen, 2020, Barry, 2020, Cevolini and Esposito, 2020]; For instance, Swedloff [2014] claims that big data is in contradiction to the risk-spreading mechanism of insurance and Heras et al. [2020] observe that, when taken to the extreme, actuarial fairness contradicts the very logic of insurance.¹¹ When risk spreading disappears, insurance becomes more like personal saving. With respect to distributive consequences, the individualization of risk has been “profoundly inequalitarian” [Armstrong, 2005]. The highest-at-risk individuals can even face exclusion from the pool [Lehtonen and Liukko, 2015, Cevolini and Esposito, 2020].

In summary, we have described three broad modes of insurance and their associated attitudes towards uncertainty, fairness and responsibility. We now investigate multiple dimensions of responsibility and then establish a link to fair machine learning, where we argue that reflections on responsibility should be foregrounded.

7.3 RESPONSIBILITY

Insurance actively constructs and distributes responsibility [Baker, 2002]; the boundaries of individual responsibility are drawn by accounting for some factors but not for others in the premium. Adding to the terminology of Landes and Holtug [2015], we distinguish four dimensions of responsibility: causal (*who is causally responsible for the accident?*), control-based (*who could control the happening?*), moral (*who is normatively responsible?*) and material (*who bears the consequences?*). By responsabilization we understand holding individuals responsible, with an emphasis on the material dimension, based on narratives about causal, control-based and moral dimensions of some phenomenon. What is distinctive about insurance as a technology of risk management is the tendency to separate these dimensions [Landes and Holtug, 2015]. However, we have observed a recent trend towards a renewed entanglement when compared to the mode of the welfare state. In particular, the notion of actuarial fairness has been increasingly linked to responsibility in contrast to mere non-responsibility based responsabilization.

What is intriguing is also how the entanglement of uncertainty and responsibility has changed historically. A thick veil of ignorance, when the individual level remains out of reach, appears to favor a collective responsibility for risk management; when this veil is gradually lifted, it seems easier to assign responsibility to individuals [Frezal and Barry,

¹¹Some argue that there is no place for risk subsidizing solidarity in insurance, that insurance is concerned only with chance solidarity. However, the conceptual distinction between these forms of solidarity is unclear and rather heuristic [Frezal and Barry, 2020]; cf. also the discussion in Section 7.5 on individual risk.

2020, Barry, 2020]. However, this is not a necessity: ‘knowing’ individual risk (to some extent) does not necessarily imply that individuals are morally responsible or that we should hold them materially responsible (see Section 7.3.2).

7.3.1 CAUSALITY AND CONTROL

Causality has received major attention in recent machine learning research and is also widely discussed in the literature on insurance. A highly related, but subtly distinct notion is that of control — indeed, the recent trend of responsabilization can to some extent be explained as a response to legislation that prohibits differentiating premia based on variables beyond individual control [Meyers and Van Hoyweghen, 2018]. While actuarial practice is correlation-based, causality and control are emphasized by legal commentators on insurance discrimination (see e.g. [Abraham, 1985, Gaulding, 1994, Avraham et al., 2013, Avraham, 2018]). For example, Avraham [2018] demands that a variable used for calculating premia must be both causally linked to the outcome of interest *and* within individual control. Why demand a causal relationship? If a variable is merely non-causally correlated with the outcome, then it is a proxy for another variable, which should be used in its place in order to avoid differential inaccuracy [Avraham et al., 2013].

Another argument is that the importance of causality is derived from control, and that it is control which is at the heart of many controversies. To fix a rough notion of causality, we understand “*X* causes *Y*” as “intervening on *X* changes the probability for *Y*”, where an intervention means changing the value. In this way, we can view causality as hypothetical control. Further, if an individual can *actually* intervene on *X*, then we may say simply that *X* is under control of the individual.¹² Observe that “*X* causes *Y*” is thus a necessary but not sufficient condition for an individual’s ability to control *Y* by controlling *X*. The importance of control in turn derives from responsibility: how could an individual be responsible (in the moral and perhaps in the material sense) for a variable which is beyond control [Abraham, 1985, Avraham, 2018]? Mere causality seems insufficient for this. Hence, we view causality as the conceptual entry point to get at control. Moreover, control is key for downstream effects, as we discuss in Appendix 7.8.4. From a normative perspective on responsibility, the importance of control (or more precisely, choice) is emphasized in theories of *luck egalitarianism*, often discussed as a justification of risk classification [Knight, 2013, Lippert-Rasmussen, 2015, Huseby, 2016, Björk et al., 2020].

Conversely, control is arguably not *sufficient* for responsibility. For instance, control without causality can yield ‘discrimination by proxy’: in Swedloff’s [2014] hypothetical example, liking Vampire novels (arguably within control) is correlated with risky behaviour, but

¹²As an example, genes might be causally related to some outcome *Y* of interest (hypothetically controlling genes would change the probability for *Y*), but are not actually under control of the individual.

in a non-causal way. This suggests that here a controllable variable works as a proxy for a potentially non-controllable one such as gender. To further complicate matters, the argument by [Hu and Kohler-Hausmann \[2020\]](#) demonstrates that a conflict may arise when a seemingly controllable stands to a non-controllable variable such as gender in a *constitutive* relation;¹³ responsabilizing for the controllable variable then effectively leads to responsabilization for the non-controllable one, too.

In the context of (fair) machine learning (see Section 7.4), causality has received much attention, but due to the previous considerations we suggest putting reflections on control-based responsibility at the center. This also implies shifting the focus to downstream (‘performative’) effects of deploying a machine learning system, since consequences of responsabilization are linked to control (see Appendix 7.8.4). Problematically, however, we must answer the question of what is under control. We now argue, in line with other social studies of insurance scholars [[Abraham, 1985](#), [Gaulding, 1994](#)], that the variant of this question which is relevant for insurance and machine learning purposes is in fact fundamentally *normative* in character.

7.3.2 SOCIAL CONTINGENCIES IN RESPONSIBILIZATION: WHAT IS UNDER CONTROL?

The distinction of *control vs. no-control* plays a major role in justifying the responsabilization of individuals, and thus actuarial fairness in the mode of personalization.¹⁴ Here we illustrate with examples from insurance that this dichotomy is a slippery one, however, and thus provides only a shaky basis for actuarial fairness. The question of whether a variable is within individual control is often unclear and in fact insurance scholars have argued that it is a normative question [[Abraham, 1985](#), [Gaulding, 1994](#)], not a descriptive one. Our sketch of an argument is as follows.

An instructive example is risk classification based on smoking in life insurance. It was only in the 1980s that life insurance companies widely started to differentiate premia with respect to smoking, even though it was already known in the 1960s that the associated difference in lifespan is substantial — the decision *not* to responsabilize depended on the social acceptability of smoking [[Wilkie, 1997](#), [Glenn, 2003b](#)]. Today, ‘lifestyle’-based responsabilization increasingly plays a role in insurance and smoking is considered a prime example for a variable within individual control [[Van Hoyweghen et al., 2006, 2007](#)]. We should thus pay close attention to shifting societal narratives about control and responsibility. A contrasting example is the case of HIV in underwriting practices, which [Daniels \[1990\]](#) explicitly contrasts with the (previous) neglect of smoking as a rating variable. [Daniels \[1990\]](#) discusses a widespread denial of health insurance coverage for individuals with HIV with

¹³For constitutive relations, see also the discussion on performativity in Appendix 7.8.4.

¹⁴Recall that in the neoclassical mode, the question of control is disregarded, but it is emphasized in the mode of personalization.

a justification based on actuarial fairness. However, [Daniels \[1990\]](#) suspects that homophobia and social antipathy to drug users may play a role in this, favoring responsabilization for risky ‘lifestyle’, conceived as controllable.

Underwriting based on ‘lifestyle’ risks can be most starkly contrasted with the use of genetic information in health and life insurance, which is now tightly regulated in some countries [[Van Hoyweghen, 2018](#)] albeit welcomed by the industry [[Rechfeld, 2016](#)]. The notion that “we are all carriers of genes” has successfully invoked a solidaristic imaginary in this context [[Van Hoyweghen et al., 2007](#)]: many argue that nobody should be penalized for their genes, since they are clearly beyond individual control. Here, a ‘genetic veil of ignorance’ is mobilized in the debate. This has even given rise to the paradigm of *genetic exceptionalism*, holding that genetic information is normatively distinct from other medical information (for a critique see [[Lippert-Rasmussen, 2015](#)]). In the context of genetic information, solidarity is thus emphasized [[Liukko, 2010](#)], but this has on the other hand contributed to a responsabilization of ‘lifestyle’ risk which continues to justify actuarial fairness [[Van Hoyweghen, 2018](#)].

Do the previous cases really show that the question of control is a normative one? We are not opposed to the idea that there exists a prior, descriptive question of control: considering *all* possible actions¹⁵ that an individual can embark on, does one of them intervene on X? In this way, it seems reasonable to say that for instance driving behaviour, but not genes, are controllable. This, however, misses the point as it is not the relevant question in the context of responsibility. Control-relevant actions will bring about different consequences for the individual, and what is more, those consequences will differ among individuals. To give up smoking might give more negative utility to an individual with genetic dispositions that favor addictive behaviour. To move to a less earthquake-prone area, in order to lower one’s insurance premium, might require investing a large amount of one’s resources. Reasoning about the control dimension of responsibility then amounts to setting the boundaries of which actions we may justifiably demand from an individual, and hence becomes normative in character, in effect a matter of distributive justice. Acknowledging the normative element in questions of control offers a new lens on the *fairness* of actuarial fairness, demonstrating that actuarial calculations are not as ‘objective and neutral’ as they are promoted, a point which we further develop in Section 7.8.4. While our examples are from insurance, we want to transport conceptual insights to machine learning.

7.4 THE LINK TO FAIR MACHINE LEARNING

In recent years, as machine learning is increasingly being deployed in sensitive domains, the field of *fair machine learning* has flourished [[Barocas et al., 2019](#), [Mitchell et al., 2021](#),

¹⁵Even granting that such an expression is sensible.

Mehrabi et al., 2021, Castelnovo et al., 2022]. Different mathematical formalizations of fairness have been proposed in the literature (see e.g. [Barocas et al., 2019]); we here focus on statistical, group-based definitions. As is common in the literature we fix a probability space, for simplicity assumed finite, where we define the following random variables: $X \in \mathcal{X}$ represents the features of the individuals under consideration, Y represents the true outcomes associated with those individuals, $\hat{Y}$ represents the predictions generated by our model, and $S \in \mathcal{S}$ represents a ‘sensitive feature’¹⁶ related to the individuals. We assume that S can be perfectly predicted from X , e.g. $X = (\tilde{X}, S)$ for some $\tilde{X}$. For simplicity, we assume binary $Y \in \{0, 1\}$ and probabilistic scores $\hat{Y} \in [0, 1]$. For example, in a credit lending scenario, X contains features such as age, income etc., S could represent “having migrant background” in a binary way, Y indicates whether an individual defaulted or not and $\hat{Y} \in [0, 1]$ represents the probabilistic prediction of the model. Group-fairness definitions are often based on binary decisions, but for the analogy with insurance we use probabilistic scores: we intuitively think of Y as representing the true outcome (an accident, damage or loss) and $\hat{Y}$ as representing the premium that the insurer (by analogy, the ML engineer) demands for shouldering the risk of the uncertain outcome Y . By $\perp\!\!\!\perp$ we denote statistical independence of random variables.

Definition 320. *A model satisfies independence if $\hat{Y} \perp\!\!\!\perp S$.*

If, for instance, S represents migrant background, independence demands that the distribution of scores is the same for people with and without migrant background. Independence starkly contrasts with calibration.

Definition 321. *A model satisfies calibration by groups with respect to S if*

$$\mathbb{E}[Y \mid S = s, \hat{Y} = \hat{y}] = \hat{y}, \quad \forall s \in \mathcal{S} \ \forall \hat{y} \in [0, 1].$$

As a fairness criterion, calibration embodies the aim of matching our probabilistic predictions well to the true outcomes. If our model predicts a probability of $p\%$ for defaulting, then indeed $p\%$ should default if our model is adequate. Recently, Höltingen and Williamson [2023] have demonstrated that in fact calibration is a richer notion than what is captured by the traditional definition. While they focus on the case of finite data, we present a corresponding theoretical definition. Assume again some set of groups $\mathcal{G} \subseteq 2^{\mathcal{X}}$. Calibration can then be defined based on this choice of groups.

Definition 322. *A model satisfies (theoretical) calibration with respect to a set of groups $\mathcal{G}$ if:*

$$\forall G \in \mathcal{G} : \mathbb{E}[(\hat{Y}(X) - Y) \mid X \in G] = 0.$$

¹⁶In the fair machine learning literature, a sensitive feature relates to membership in a socially salient group, for instance based on gender, race or religion.

While this looks deceptively similar to the recently popularized *multi-calibration*, it abstracts away from the prediction-based binning, which is partly due to historical reasons [Höltgen and Williamson, 2023]. To gain intuition, it is instructive to consider what this means in the case of finite data. Assume a finite dataset $(X_1, Y_1), \dots, (X_n, Y_n)$ with associated predictions $\hat{Y}(X_1), \dots, \hat{Y}(X_n)$. A set of groups is then equivalent to choosing a subset of the data, i.e. a set $\mathcal{G} \subseteq 2^{\{1, \dots, n\}}$. If we use the empirical distribution associated with this dataset in Definition 322, we obtain the following empirical variant.

Definition 323. *A model satisfies (empirical) calibration with respect to a set of groups $\mathcal{G}$ if:*

$$\forall G \in \mathcal{G} : \frac{1}{|G|} \sum_{i \in G} (\hat{Y}(X_i) - Y_i) = 0.$$

This offers an insurantal interpretation: $\hat{Y}(X_i)$ is the premium we demand for shouldering the risk of the uncertain Y_i . Indeed, calibration is formally reminiscent of the subjective betting interpretation for probability theory proposed by de Finetti [1974/2017]. There, the expectation $\mathbb{E}[Y]$ is viewed as the fair betting price for a gamble Y .¹⁷ Calibration is however not tied to a subjective or objective interpretation of probability, but a criterion for model evaluation.

Actuarial fairness and calibration are in close correspondence. For a given set of groups $\mathcal{G} \subset 2^{\mathcal{X}}$, which we assume forms a partition of $\mathcal{X}$, we define the actuarially fair predictor $\hat{Y}_{\text{af}}(x) := \mathbb{E}[Y | X \in G_x]$, where G_x is the unique group that contains x . The goal of actuarial fairness is to make this partition as fine as possible (cf. Section 7.2.2, 7.2.3).

Consider first the extreme choice of a single group as the whole population in Definition 323. Then calibration in the theoretical (respectively, empirical) case demands that

$$\mathbb{E}[\hat{Y}(X) - Y] = 0, \quad \frac{1}{n} \sum_{i=1}^n (\hat{Y}(X_i) - Y_i) = 0,$$

which is called *global balance* in insurance [Denuit et al., 2021]; intuitively, we need to collect sufficient premia $\hat{Y}(X_i)$ to cover all claims Y_i . In this case of a single group, calling $\hat{Y}_{\text{af}}$ ‘actuarially fair’ is some abuse of naming, since in this case it corresponds to full solidarity.

Proposition 324. *Given a partition $\mathcal{G} \subset 2^{\mathcal{X}}$ of $\mathcal{X}$, the actuarially fair predictor $\hat{Y}_{\text{af}}$ satisfies (theoretical) calibration with respect to $\mathcal{G}$, and is furthermore the coarsest calibrated predictor in the sense that any other predictor which is calibrated with respect to $\mathcal{G}$ either coincides with it or is not group-wise constant.*

The trivial proof is in Appendix 7.8.2. If we had access to the true outcomes Y , we could use them for the finest, calibrated predictions. In this way, calibration is still a coarser

¹⁷Accordingly, probability refers to indicator gambles, that is, indicator functions of events.

criterion as it is based on the expectation, the theoretical average, and thus aligned with actuarial fairness; perfect accuracy is in general not demanded. In the extreme, making the partition finer and finer we reach segments of one. Calibration then demands that $\hat{Y}(x) = \mathbb{E}[Y|X = x]$, which we called perfect actuarial fairness. In this way, we obtain a ‘spectrum of calibration’, where refining the choice of groups interpolates between two extremes. Group-based actuarial fairness attempts to approximate the extreme of segments of one given practical constraints. Hence, actuarial fairness is well-aligned with fairness-unaware machine learning, where the goal is to approximate the conditional expectation $\mathbb{E}[Y|X]$ as closely as possible. In the limit, the distinction between group-based approaches to fairness and *individual fairness* (in the sense of the machine learning literature [Dwork et al., 2012]) then becomes blurry: perfect actuarial fairness corresponds to the notion of *individual merit* [Joseph et al., 2016], but in line with Binns [2020] we argue in Section 7.5 that this does not yield actually ‘individual’ fairness.

Actuarial fairness is closely related to calibration due to Proposition 324; however finer predictors (e.g. the perfect predictor) satisfy calibration, as well. For a precise conceptual correspondence with perfect actuarial fairness, we could consider the following class of fairness measures, inspired and slightly generalized from R  z [2021].¹⁸

Definition 325. *A fairness measure is probabilistically conservative if it is necessarily satisfied by the perfect actuarially fair predictor $\hat{Y}_{\text{af}}(x) = \mathbb{E}[Y|X = x]$.*

For simplicity and due to the insurantal interpretation, we focus on calibration.

Not only for calibration, but similarly for independence (Definition 320) the choice of grouping is crucial. At the one extreme, choosing S to be the indicator of the whole population, independence becomes vacuous as it is trivially satisfied. In contrast, we can add more and more groups (sensitive features) for which we demand independence, so that less and less variations in $\hat{Y}$ are allowed. In the limit, then, we reach *full solidarity* with a constant $\hat{Y}$ (see Appendix 7.8.3). Contrasting independence and calibration, the question is whether we allow the prediction $\hat{Y}$ (the premium) to be sensitive to a certain group or not — where the within-group variation is however neglected.

Calibration and independence can be mapped onto *responsibilization vs. non-responsibilization*. The fairness notion that is embodied by calibration is that of accurately reflecting the ‘true’ probabilities, that is, holding individuals responsible for their risk. In contrast, independence works to decouple predictions from ‘true’ probabilities to some extent and thus can be viewed as non-responsibilizing — independence is similar, albeit not equivalent, to affirmative action [R  z, 2021]. As a consequence, we expect different *performative effects* (see Appendix 7.8.4), i.e. downstream effects, when applying calibration vs. independence.

¹⁸R  z [2021] defines a fairness measure as conservative if it is necessarily satisfied by the perfect predictor $\hat{Y} = Y$.

For practice, a simple suggestion is as follows. Since calibration comes practically for free for loss-minimizing predictive models [Barocas et al., 2019, p. 62f], we may focus on demanding certain independence relationships. Assume that we have designated a subset of features X_R which we aim to responsibilize for, and a set of features X_{NR} which we aim *not* to responsibilize for. Hence $X = (X_R, X_{NR}, X_{\text{other}})$. Conditional independence [Castelnovo et al., 2022] then demands that

$$\hat{Y} \perp\!\!\!\perp X_{NR} \mid X_R$$

The features X_{other} , on which we withhold judgement, can then be used by the model in a way restricted by the conditional independence. In the spirit of Section 7.3.2, however, the choice of features for (non)responsibilization should be the outcome of a reflexive process of inquiry.

Beyond calibration and independence, other proposals have been put forward in the machine learning literature, which may also be linked to (non)responsibilization; hence the lessons from insurance can be applied, too. For instance, within an equality of opportunity framework, Heidari et al. [2019] suggest splitting the whole set of features into a set of “accountability” features and “irrelevant” features. From our perspective, this maps onto responsibilization and non-responsibilization. As another prominent example, in a causal fairness framework, Kilbertus et al. [2017] assume that a set of “resolving variables” is given, which are influenced by a sensitive feature in a way that it considered “non-discriminatory”; but the authors do not provide guidance on how to choose them. We believe that the lessons from insurance about causality and control (Section 7.3.1), and more broadly on responsibility in general, can on the one hand guide the selection of such features. On the other hand, we have seen that causal and control-based dimensions of responsibility are highly sensitive to social context (Section 7.3.2). This highlights the normative element in making such a distinction (responsibilizing or not), which can be problematic and must be recognized as such. The choice can be side stepped by favoring solidarity over actuarial fairness.

7.5 TENSIONS BETWEEN AGGREGATE AND INDIVIDUALS

Throughout the history of insurance, and also highly relevant to machine learning, we find tensions between *aggregate and individual*. The mode of the welfare states operates with the imaginary of a collective, in which the individual is mutualized in solidarity. This aggregate viewpoint, where an individual is always identified with the average of some group, finds continuity in group-based actuarial fairness [Thiery and Van Schoubroeck, 2006] — consistent with Quetelet’s *average human*. Social movements, however, argued that the group-based actuarially fair price is not fair from the viewpoint of the individual — the desire was to “navigate the social world *unmarked* by the social stereotypes (fashioned by

actuarial science as ‘objective’ statistical classifications) [...]” [Krippner, 2023, emphasis in original]. This critique has prompted a shift in the enactment of actuarial fairness [Meyers and Van Hoyweghen, 2018], giving rise to the individualization and personalization of risk and thereby threatening (risk subsidizing) solidarity by dissolving the aggregate. At the heart of this aggregate vs. individual tension is the multifaceted concept of responsibility: what links the individual to the aggregate is mutualization based on establishing shared responsibility. *Personhood*, being an individual, is deeply intertwined with assigning responsibility, as insurance scholars have pointed out [McFall and Moor, 2018, Moor and Lury, 2018]. It is never the whole of a person that a premium is attached to in insurance, but specific, contextually relevant aspects [McFall and Moor, 2018], thereby transforming a person into an insurance risk [Van Hoyweghen, 2014, McFall and Moor, 2018]. It would be intriguing to explore how personhood is negotiated within machine learning systems, drawing parallels with similar studies in insurance (e.g. [Tanninen, 2020]).

The mode of machine learning is a paradoxical one: on the one hand, it fits with the mode of individualization and personalization. The goal is to provide highly tailored predictions for the individual. On the other hand, aggregates are central to the workings of machine learning: they appear in the input data due to categorization processes; second, the fairness of machine learning systems is typically evaluated based on groups (with the exception of *individual fairness*, see below); third, machine learning in general, whether fairness-unaware or not, rests on aggregate criteria such as average training error. We suggest that a large share of the social worries and issues surrounding machine learning can be understood by framing them in the context of the aggregate vs. individual tension.

Group-based actuarial fairness, which relies on historical data aggregated by groups, is prone to reproduce past injustice [Daniels, 1990, Lehtonen and Liukko, 2015], see also Appendix 7.8.4. In contrast, the allure of perfect actuarial fairness associated with the personalization of risk, driven by big data and machine learning, is that it is supposedly *individually fair* — the goal being ‘segments of one’ and setting the premium as $\mathbb{E}[Y|X = x]$. However, we contend that this elusive goal cannot be reached. The core issue lies with the ‘hidden collective’. The working of a neural network is similarity-based computation, arguably interpolation [Hasson et al., 2020]. Predictions are invariably grounded in data from individuals *similar to you*, where the similarity is with respect to the opaque nonlinear character of the network. This argument has been made in the context of insurance: ‘individualized’ risk is still relative to the other members of the collective [Tanninen, 2020, Prainsack and Van Hoyweghen, 2020]. Yet individual justice in an Aristotelian tradition requires treating people *as individuals* [Thiery and Van Schoubroeck, 2006, Jorgensen, 2022], not based on the data of others.¹⁹ For the same reason, what is called *individual fairness* in machine learning fails to be genuinely individual, as pointed out by Binns [2020]. Problematically the

¹⁹We leave open the question to what extent this can be realized by human decision makers.

collective is implicit, hidden, in the mode of personalization; without transparency and explainability, individuals cannot recognize their own context. Insurance scholars have also argued that this diminishes opportunity for collective action [Moor and Lury, 2018, McFall and Moor, 2018, Krippner and Hirschman, 2022] — the study of collective action in machine learning has just begun [Hardt et al., 2023].

Another way of framing this consists in problematizing the conceptual foundation of probability and statistics itself. Besides the subjective variant of probability, which we consider unfit for decision making affecting people,²⁰ probability and statistics are fundamentally based on aggregates [Desrosières, 1998]. Currently, no viable concept of individual probability is available [Dawid, 2017]; instead, probability relies on a reference class [Reichenbach, 1949, Hájek, 2007]. While multi-calibration aims at finer aggregates, it is still not individual [Dawid, 2017]. Thus, even speaking of an ‘individual probability’, which perfect actuarial fairness aims at, has no sound conceptual basis. In fact, Friedman [2020, Chapter 4] provides an insightful account for the close link of frequentist probability and insurance in the solidaristic mode of the welfare state. This account invites us to consider a reference class as a class of solidarity, which implies disregarding the quest for the single ‘right’ reference class and instead recognizing the normative element in this choice. Furthermore, randomness can then also be viewed as a normative assumption in the face of uncertainty, establishing shared responsibility: “anyone of us could have had the accident”. Thus, the kind of data that Venn [1876] has in mind, combining aggregate regularity with local irregularity (randomness), corresponds to the prerequisites for insurance. As a consequence, we find normative character in frequentist probability itself and a link to aggregate-based solidarity. Attempting to individualize frequentist probability then raises a paradox. In the context of insurance, Frezal and Barry [2020] have argued that the actuarially fair expected value is only adequate from the economic, aggregate viewpoint of the insurer, but conceptually inadequate (and hence in particular not necessarily ‘fair’) for the individual (see also [Frezal, 2016]). Or, in the words of Abraham [1985]: “No one has a true expected loss”. When taken to the limit, actuarial fairness thus undermines the logic of insurance itself [Heras et al., 2020]. In the context of machine learning, we argue that it is problematic to evoke the idea of individual probability (risk), particularly when stakes are high such as in the COMPAS case [Angwin et al., 2016], even when individual probability is beyond reach and has no conceptual foundation to rest on. We thus encourage more modesty about the epistemic potential of machine learning.

²⁰Setting insurance premia based on subjective probabilities seems objectionable when it affects the welfare of people; similarly for machine learning.

7.6 RELATED WORK

Conceptual literature at the intersection of fair machine learning and insurance is sparse. [Donahue and Barocas \[2021\]](#) study the problem of *externalities of size* by taking inspiration from insurance and challenge the clear distinction between actuarial fairness and solidarity as a consequence. [Loi and Christen \[2021\]](#) study the interplay of machine learning and (non)discrimination in an insurance setting from a normative, philosophical perspective. [Frees and Huang \[2023\]](#) and [Charpentier \[2022\]](#) provide broad overviews on discrimination in insurance and consider implications of using machine learning. [Xin and Huang \[2023\]](#) link formal fairness definitions to insurance on a technical level. The closest work to ours that we are aware of is by [Barry and Charpentier \[2022\]](#), who investigate how the use of machine learning in insurance is related to classical fairness debates, but they set other foci. A general difference from ours to related work is that we do not study the use of machine learning *in* insurance, but are interested in a more abstract conceptual linkage. We also note that [Frezal and Barry’s \[2020\]](#) critique of actuarial fairness was a major source of inspiration to us.

7.7 DISCUSSION

The main claim of our work is that insurance is an insightful analogon for the social situatedness and impact of machine learning systems. By traversing this conceptual bridge, machine learning scholars can make use of the rich and interdisciplinary literature on insurance. In particular, we suggest that the multifaceted concept of responsibility, tightly linked to causality and control, deserves more attention. We have illustrated problems with actuarial fairness as a notion of fairness in the normative sense. In this way, our suggestions are in line with others who demand moving beyond formal fairness to substantive fairness [[Green, 2022](#)] and argue that accurate predictive models need not be ‘fair’ [[Eidelson, 2021](#)].

While in this text we have focused on social issues, there are also technical lessons that machine learning could take from insurance, a technology for handling uncertainty. For instance, the problems of *dataset shift* and *model ambiguity* have been recognized in insurance as well as machine learning; for contributions from insurance see e.g. [[Milevsky et al., 2006](#), [Cabantous, 2007](#), [Pichler, 2014](#), [Dietz and Niehörster, 2021](#)]. On the other hand, the use of machine learning in insurance is increasing. We thus believe that a research agenda linking machine and learning and insurance may lead to a fruitful, two-way interaction of these fields.

In summary, we offer the following insights. Recent impossibility theorems in the fair machine learning literature [[Kleinberg et al., 2017](#)] are not as surprising when considering

them in the light of the old, fundamental tension in insurance between solidarity and actuarial fairness. In essence, this tension is grounded in how individuals are related to the aggregates they form. This relation rests on responsabilization. Responsibility and responsabilization should be conceptually distinguished, even if in the recent mode of personalized insurance (tightly linked to machine learning) the two are increasingly intertwined. For insurance and machine learning purposes, reasoning about responsibility crucially requires reasoning about causality and control. We have emphasized that the relevant control question has a normative flavour, and thus cannot be left to engineers alone. What is under individuals' control is often hotly contested. As a general research heuristic, many case studies by social scholars of insurance can inspire analogous studies in the context of machine learning, covering a diverse set of topics; mining the literature is thus a rich source of inspiration.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC number 2064/1 Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christian Fröhlich. Thanks to Benedikt Höltingen, Sebastian Zezulka, Renate Baumgartner and Maiju Tanninen for helpful discussions and comments.

7.8 APPENDIX

7.8.1 NON-RESPONSIBILITY BASED REASONS FOR RESPONSIBILIZATION

The difference between responsibility and responsabilization can be subtle. In insurance, two classical principles supply reasons for justifying responsabilization *without* being based on responsibility, however. We might also call them “efficiency-based” reasons for responsabilization [Andersen and Nielsen, 2015].

Adverse selection [Akerlof, 1970, Thiery and Van Schoubroeck, 2006, Avraham et al., 2013] refers to the informational asymmetry between insurer and insured (policyholder) at the time of underwriting. Typically, the insured is better informed about their risk; higher-risk individuals are more likely to seek insurance for protection. The reasoning is then that without segmentation, insurers are more likely to attract high-risk individuals who profit from the subsidy of the pool; ultimately, leading to bankruptcy of the insurer. Thus competition drives increasing segmentation. Public insurance has the benefit of compulsory participation, so that adverse selection cannot occur and solidarity can be implemented.

In contrast, adverse selection is advanced by the industry as a justification for actuarial fairness [Miller, 2009].

Moral hazard [Heimer, 1985, Baker, 1996, 2000] (in the context of insurance) on the other hand refers to a *performative*, behaviour-shaping aspect of insurance premia. Simply put, the idea is that policyholders are more inclined to behave in a risky way due to the protection offered by insurance coverage. For instance, there is less incentive to purchase precautionary measures such as alarm systems. Like adverse selection, moral hazard is invoked as an argument in favor of actuarial fairness. The welfare state is seen as the ultimate source of moral hazard [Ericson et al., 2000]. For instance, public health insurance might lead to more visits to the doctor.²¹ Indeed, moral hazard is *the* responsabilization force in the neoliberal era [Ericson et al., 2000]. In particular, moral hazard favors behaviour-based personalization [Verbelen et al., 2018]. Insurance companies thus increasingly emphasize loss prevention, that is, acting on behaviour [Baker and Simon, 2002, Cevolini and Esposito, 2022]. Note that the applicability of moral hazard presupposes control for a feedback loop to exist.

Both adverse selection and moral hazard concern the performative dimensions of “calculative devices” [Van Hoyweghen, 2014]. Hence we suggest that these concepts might be usefully transposed onto machine learning.

7.8.2 PROOF OF PROPOSITION 324

Proof

Plugging in the definition of the actuarially fair predictor, we find that $\forall G \in \mathcal{G}$:

$$\mathbb{E}[\hat{Y}(X)|X \in G] = \mathbb{E}[Y|X \in G] \quad (7.1)$$

$$\Leftrightarrow \mathbb{E}[(x \mapsto \mathbb{E}[Y|X \in G_x])(X)|X \in G] = \mathbb{E}[Y|X \in G] \quad (7.2)$$

$$\Leftrightarrow \mathbb{E}[Y|X \in G] = \mathbb{E}[Y|X \in G], \quad (7.3)$$

which is true. For the second statement, observe that Equation 7.1 implies that a predictor $\hat{Y}$ which is constant on each group G must equal the conditional expectation for that group.

■

²¹Noting the negative connotation in the term *moral hazard*, Baker [1996] asks: “What, after all, is wrong with enabling people to go to the doctor when they feel the need, and why should we be concerned when they do so?”

7.8.3 INDEPENDENCE WITH RESPECT TO ALL GROUPS IMPLIES FULL SOLIDARITY

Proposition 326. *Assume independence holds with respect to all groups, that is*

$$\hat{Y} \perp\!\!\!\perp \chi_A, \quad \forall A \subseteq \Omega, A \neq \emptyset.$$

Then $\forall \omega \in \Omega : P(\{\omega\}) > 0 : \hat{Y}(x) = c$ for some $c \in \mathbb{R}$.

Proof Recall that we assume a finite Ω . Then we can assume without loss of generality that $P(\{\omega\}) > 0 \forall \omega \in \Omega$, otherwise we could work on an altered probability space by discarding sets of measure zero. From the independence assumption it follows that $P(\{\omega : \hat{Y}(\omega) = y | A\}) = P(\{\omega : \hat{Y}(\omega) = y\})$ for any $y \in \mathbb{R}$ and $A \subseteq \Omega, A \neq \emptyset$. Pick any ω_1 so that $\hat{Y}(\omega_1) = y_1$. Then it must hold $P(\{\omega : \hat{Y}(\omega) = y_1 | \{\omega_1\}\}) = 1 = P(\{\omega : \hat{Y}(\omega) = y_1\})$, from which we conclude that $\hat{Y}$ must be constant on Ω . ■

7.8.4 PERFORMATIVITY

A central and unifying theme that emerges when considering social issues surrounding insurance, and as we contend, also machine learning, is that of *performativity*. This concept allows us to comprehend many of the previously raised points through a new lense. The idea of performativity originates from John L. Austin’s seminal work “How to do things with words” [Austin, 1962], where Austin notes that the function of language is often not only descriptive, but also has constitutive and causal effects. For example, uttering “I promise” itself *constitutes* a promise and has the causal effect of establishing certain expectations. Austin [1962] refers to sentences with such *performative* force as *speech acts*. Since Austin, the term ‘performativity’ has travelled far into multiple disciplines and acquired lives of its own, so that it can be hard to pin down precisely a common core (see [Mäki, 2013] for a critique). We will use the term in the broadest possible sense to emphasize similarity of perspectives instead of differences. An influential account has been put forward in economics [Callon, 1998, MacKenzie et al., 2008, MacKenzie, 2008], which has inspired also a recent formalization in machine learning [Perdomo et al., 2020, Hardt et al., 2022]. The central claim of this line of work is that economics is not simply in the business of describing or representing an independent, passive reality, but also actively shapes it, for instance by encouraging people to act in accordance with its models [Boldyrev and Svetlova, 2016]. Another general framework and a source of inspiration to us, can be found in the work of Mol [2002]. To avoid the dualist connotation of the term performance, implying a ‘backstage reality’,²² Mol [2002] instead coins the term *enactment*, referring to the

²²However, for a critical examination of whether this dualism is inherently associated with ‘performance’, see [Hafermalz et al., 2016].

multiple and ongoing work that sustains a reality:

It is possible to say that in practices objects are *enacted*. This suggests that activities take place — but leaves the actors vague. It also suggests that in the act, and only then and there, something *is* — being enacted. [emphasis in original]

Perhaps the simplest way to understand what is at the heart of performativity, we suggest, is to assert that *representation and intervention are entangled* [Vosselman, 2014]. Performativity hence contrasts with the commonsense view that perception and action can be neatly separated; in the latter, the task of machine learning is simply to extract patterns from a passive reality ‘out there’ in an objective way. For instance, Mitchell et al. [2021] distinguish between “world as it is” and “world as it should and could be”, mapping onto prediction and decision task. Similarly, Kuppler et al. [2021] urge to cleanly separate prediction and decision. A word that frequently occurs in this context is ‘bias’, which we like to avoid, since we associate it with the notion of an objective ‘backstage’ reality, whose representation is then distorted.

What then is the relevance of performativity for insurance, and by analogy, machine learning? Actuarial fairness (calibration), or more broadly the fairness of ‘accurate’ statistical methods, carries with it an aura of objectivity and neutrality. *If* we choose responsibilization, *then* our predictive models better be ‘objective and neutral’. Recall that actuarial fairness aims to set premia in accordance with the expected risk for each policyholder, and it is assumed that insurers can know this risk through statistical methods. Glenn [2003a] succinctly captures this as follows:

[T]here is a general belief that insurance practices are predicated on objective statistics, what has elsewhere been called the myth of the actuary. The myth of the actuary is the idea that there is a reality in the world that can be captured by rational choice models and statistical analysis and that insurance companies do this ethically, objectively, and correctly.

Such objectivity then is supposed to be a source of authority and fairness. In the words of Van Hoyweghen [2014]:

The dominant view is that insurance technologies of risk assessment are somehow measuring, observing or describing peoples insurance risks. This paper calls for a different approach, namely a pragmatist analysis of the *performativity of insurance calculative devices*. Contrary to the financial realism of the everyday categories of insurance numbers, I argue that insurance calculative devices not only represent but generate, intervene and rearrange the worlds in which they are deployed. [emphasis added]

The performativity of insurance and machine learning becomes especially relevant due to ethical implications. Many scholars have argued, providing insightful examples, that insurance is fundamentally a *normative technology* [Baker and Simon, 2002, Glenn, 2003a, Van Hoyweghen et al., 2006, 2007, Lehtonen and Liukko, 2015, Tanninen, 2020, Prainsack and Van Hoyweghen, 2020], depending on causality, control and responsibility. *Doing* insurance or machine learning involves *enacting* certain realities and suppressing others, as we have sketched in Section 7.3.2. For instance, in the process of collecting data, only some features are considered, and others neglected. Expanding on this, a performativity perspective would emphasize that there is no objective data ‘collection’ process, that quantification and categorization require significant and ongoing work; such work may be influenced by implicit normative judgements, which becomes ingrained and hidden in the ‘representation’. There is now a vibrant, if still nascent, research field on the *sociology of quantification* (including categorization), owing much to the seminal work of Desrosières [1998]; for overviews of this field see [Espeland and Stevens, 2008, Diaz-Bone and Didier, 2016, Mennicken and Espeland, 2019], where the reader finds plenty of evidence for such work. Central in this research field is again performativity, or what has been called the *constitutive potential* of quantification [Mennicken and Espeland, 2019]. As a noteworthy example, it has been demonstrated that the census, through the introduction of statistical categories, can contribute to the establishment of a collective identity among the individuals it aims to describe [Lee, 1993, Bowker and Star, 2000, Mora, 2014]. Thus, a category that was initially intended to merely represent acquires performativity by actively shaping the formation of this particular group [Grommé and Scheel, 2020].²³

We propose that insurance can act as a model for the performativity of statistical, “calculative devices” [Van Hoyweghen, 2014] that arise from their social situatedness. In machine learning, performativity shows up in at least two ways: on the one hand, it requires training data, and this training data has been shaped by performative forces in a broader social context — referring to the quantification and categorization processes. Using the data thus imports this performativity into the model. On the other hand, by deploying a machine learning model its predictions may acquire performative force in both a constitutive and causal sense. The predictions may act as interventions and through responsabilization shape the behaviour of people, who for instance may strategically adapt to the predictions — this sense of performativity is also referred to as *reactivity* [Espeland and Sauder, 2007, Espeland and Stevens, 2008], and is closer to the formalization proposed by Perdomo et al. [2020], which emphasizes the causal dimension, but neglects the constitutive one. The extent of this phenomenon, i.e. how much a company can steer a population using a model,

²³To anticipate a criticism, this does of course not imply that arbitrary sets of people can become a mutually recognizing group: the hard conceptual work is to investigate how performativity of groups functions and what its limits are. In Austinian terms, this means delineating the “felicity” conditions, which make performative utterances successful [Brisset, 2016].

has been termed *performative power* [Hardt et al., 2022].

As explicated in Section 7.4, depending on the choice of fairness metric (or none) and to which features it is applied (i.e. the choice of groups), machine learning can exert responsabilizing and non-responsibilizing force. We suggest that two classes of performative effects can then be broadly distinguished, which is however not a clear dichotomy in light of Section 7.3.2. When machine learning responsabilizes for a controllable feature, individuals may adapt to the prediction so as to change it — if they receive feedback; this is the hope of personalized insurance, and this setting also motivates the concept of moral hazard (Appendix 7.8.1). In contrast, blindly applying the principle of actuarial fairness can lead to responsabilizing for non-controllable features, which then runs the risk of reproducing past injustice implicit in the training data. Yet against a background of such past injustice, it is not clear why actuarial *fairness* should be considered as a principle of *justice* — this argument has been made both in the insurance [Daniels, 1990, Lehtonen and Liukko, 2015, Barry, 2020] as well as the machine learning literature [Mitchell et al., 2021, Vredenburg, 2022, Green, 2022, Kasirzadeh, 2022]; see also [Eidelson, 2021] — however, the insurance literature provides illuminating examples. In this way, machine learning (resp. insurance) can implicitly responsabilize for sensitive features such as gender or race; the situation is particularly intricate when a feature is considered as controllable which stands in a constitutive relation to a sensitive feature [Hu and Kohler-Hausmann, 2020], for instance due to performativity.

In response to the performativity of machine learning, we advocate for explicit reflection about how performative forces have shaped the present input data, and furthermore how a model in conjunction with a choice of fairness metric might exert performative force by acting on people. Focusing on (non)responsibilization and performativity implies taking a dynamic perspective. Thus, it becomes imperative to foreground and explicitly model the effects of deploying machine learning systems [Hu and Chen, 2018, Liu et al., 2018, D’Amour et al., 2020, Schwöbel and Remmers, 2022], contrasting with rather static vocabulary such as *bias* or *discrimination*. In this process of reflexive inquiry, we suggest to pay more attention to enactments of causality, control and responsibility — framing them in this way rather than as immutable *facts* implies making them contestable, that is, putting them on the stage for scrutiny [Glenn, 2003a].

Part IV

Conclusion

8

An Outlook

Very briefly, I want to reflect on key outcomes of this thesis. We have contributed to the theory of imprecise probabilities, mostly from a generalized frequentist perspective. Throughout this research, we have developed an approach to uncertainty that emphasizes two crucial elements: data model and decision problem. We have demonstrated that when these elements are properly considered, open questions that arise from underdetermination (such as how to evaluate forecasts) can often be resolved. Additionally, we have proposed that insurance serves as an illuminating paradigm for understanding uncertainty more broadly. While Part III explicitly explored this connection, this thread runs throughout the entire thesis.

I hope that by now the reader is convinced that imprecise probabilities are interesting and relevant, and can be an appropriate response to some of the headaches arising from precision. In any case, many insights that we have generated in the process also illuminate the familiar precise case, and I think *for that reason alone* it was already worth it.

As to future research, we have opened many avenues for investigation. The papers contain detailed suggestions, so here I would like to focus only on two avenues which I think are particularly worthwhile. First, the generalized frequentist theory we have developed calls for a correspond finite case version which helps to handle aggregate instability in practice. And as a highly related research project, I believe there are now many ingredients on the table for the development of a comprehensive picture that lays out the connections between imprecise probabilities, heavy tails, (non)stationary, (non)equilibrium and

further concepts. There appears to be a fundamental story with highly relevant applications emerging here, and we have already taken steps in this direction, but many aspects of this picture are yet to be worked out.

References

- Joaquín Abellán and Serafín Moral. Maximum of entropy for credal sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(05):587–597, 2003.
- Joaquín Abellán and Serafín Moral. Upper entropy of credal sets. Applications to credal classification. *International Journal of Approximate Reasoning*, 39(2-3):235–255, 2005.
- Kenneth S. Abraham. Efficiency and fairness in insurance risk classification. *Virginia Law Review*, 71(3):403–451, 1985.
- Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.
- George A. Akerlof. The market for lemons: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.
- Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.
- Amir Ahmadi-Javid and Alois Pichler. An analytical study of norms and Banach spaces induced by the entropic value-at-risk. *Mathematics and Financial Economics*, 11(4):527–550, 2017.
- Nabil I. Al-Najjar and Luciano De Castro. Subjective probability. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Ltd, 2010.
- Sergio Albeverio, Mykola Pratsiovytyi, and Grygoriy Torbin. Topological and fractal properties of real numbers which are not normal. *Bulletin des Sciences Mathématiques*, 129(8):615–630, 2005.
- Sergio Albeverio, Iryna Garko, Muslem Ibragim, and Grygoriy Torbin. Non-normal numbers: Full Hausdorff dimensionality vs zero dimensionality. *Bulletin des Sciences Mathématiques*, 141(2):1–19, 2017.
- Aurélien Alfonsi. A simple proof for the convexity of the Choquet integral. *Statistics & Probability Letters*, 104:22–25, 2015.
- Maurice Allais. Le comportement de l’homme rationnel devant le risque: critique des postulats et axiomes de l’école Américaine. *Econometrica*, 21(4):503–546, 1953.
- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

- Paul Anand, Prasanta Pattanaik, and Clemens Puppe, editors. *The handbook of rational and social choice*. Oxford University Press, 2009.
- Martin Marchman Andersen and Morten Ebbe Juul Nielsen. Luck egalitarianism, universal health care, and non-responsibility-based reasons for responsabilization. *Res Publica*, 21:201–216, 2015.
- Alexandr Andoni, Chengyu Lin, Ying Sheng, Peilin Zhong, and Ruiqi Zhong. Subspace embedding and linear regression with Orlicz norm. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 224–233. PMLR, 10–15 Jul 2018.
- Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There’s software used across the country to predict future criminals, and its biased against blacks. *ProPublica*, 2016. URL <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. Accessed on June 2, 2023.
- Francis J. Anscombe and Robert J. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1):199–205, 1963.
- Takuji Arai. Convex risk measures on Orlicz spaces: inf-convolution and shortfall. *Mathematics and Financial Economics*, 3(2):73–88, 2010.
- Takuji Arai. Good deal bounds induced by shortfall risk. *SIAM Journal on Financial Mathematics*, 2(1):1–21, 2011.
- Chris Armstrong. Equality, risk and responsibility: Dworkin on the insurance market. *Economy and Society*, 34(3):451–473, 2005.
- Kenneth J. Arrow. Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5):941–973, 1963.
- Philippe Artzner. Application of coherent risk measures to capital requirements in insurance. *North American Actuarial Journal*, 3(2):11–25, 1999.
- Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.
- Miroslav D. Ašić and Dušan D. Adamović. Limit points of sequences in metric spaces. *The American Mathematical Monthly*, 77(6):613–616, 1970.
- Thomas Augustin. Statistics with imprecise probabilities a short survey. In *Uncertainty in Engineering: Introduction to Methods and Applications*, pages 67–80. 2022.
- Thomas Augustin, Frank P.A. Coolen, Gert De Cooman, and Matthias C.M. Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.
- John L. Austin. *How to Do Things With Words*. Harvard University Press, Cambridge, 1962.
- Andrea Aveni and Paolo Leonetti. Most numbers are not normal. In *Mathematical Proceedings of the Cambridge Philosophical Society*, pages 1–11. Cambridge University Press, 2022.

- Ronen Avraham. Discrimination and insurance. In *The Routledge Handbook of the Ethics of Discrimination*. Routledge, 2018.
- Ronen Avraham, Kyle D. Logue, and Daniel Schwarcz. Understanding insurance antidiscrimination law. *Southern California Law Review*, 87:195–274, 2013. URL https://scholarship.law.umn.edu/faculty_articles/576. Accessed on June 2, 2023.
- Ralph Alger Bagnold. The nature and correlation of random distributions. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 388(1795):273–291, 1983.
- Drumi D. Bainov and Pavel S. Simeonov. *Integral inequalities and applications*. Springer Science & Business Media, 2013.
- Tom Baker. On the genealogy of moral hazard. *Texas Law Review*, 75:237, 1996.
- Tom Baker. Insuring morality. *Economy and Society*, 29(4):559–577, 2000.
- Tom Baker. *Risk, insurance, and the social construction of responsibility*. University of Chicago Press, 2002.
- Tom Baker and Anja Shortland. The government behind insurance governance: Lessons for ransomware. *Regulation & Governance*, 17(4):1000–1020, 2023.
- Tom Baker and Jonathan Simon. *Embracing risk: The changing culture of insurance and responsibility*. University of Chicago Press, 2002.
- Abdessamad Barbara and Jean-Pierre Crouzeix. Concave gauge functions and applications. *Zeitschrift für Operations Research*, 40(1):43–74, 1994.
- Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. 2019. URL <http://www.fairmlbook.org>. Accessed on January 5, 2024.
- Laurence Barry. The rationality of the digital governmentality. *Journal for Cultural Research*, 23(4):365–380, 2019.
- Laurence Barry. Insurance, big data and changing conceptions of fairness. *European Journal of Sociology/Archives Européennes de Sociologie*, 61(2):159–184, 2020.
- Laurence Barry and Arthur Charpentier. The fairness of machine learning in insurance: New rags for an old man? *arXiv preprint arXiv:2205.08112*, 2022.
- František Bartoš, Alexandra Sarafoglou, Henrik R. Godmann, Amir Sahrani, David Klein Leunk, Pierre Y. Gui, David Voss, Kaleem Ullah, Malte J. Zoubek, Franziska Nippold, et al. Fair coins tend to land on the same side they started: Evidence from 350,757 flips. *arXiv preprint arXiv:2310.04153*, 2023.
- Gilbert W. Bassett Jr., Roger Koenker, and Gregory Kordas. Pessimistic portfolio allocation and Choquet expected utility. *Journal of Financial Econometrics*, 2(4):477–492, 2004.
- Nicole Bäuerle and Alexander Glauner. Minimizing spectral risk measures applied to Markov decision processes. *Mathematical Methods of Operations Research*, 94(1):35–69, 2021.

- Nicole Bäuerle and Alfred Müller. Stochastic orders and risk measures: consistency and bounds. *Insurance: Mathematics and Economics*, 38(1):132–148, 2006.
- Heinz H. Bauschke, Minh N. Dao, and Walaa M. Moursi. On Fejér monotone sequences and nonexpansive mappings. *Linear and Nonlinear Analysis*, 1(2):287–295, 2015.
- Güzin Bayraksan and David K. Love. Data-driven stochastic programming using Phi-divergences. In Dionne M. Aleman and Aurélie C. Thiele, editors, *The Operations Research Revolution*, chapter 1, pages 1–19. 2015.
- Tadeusz Bednarski. On solutions of minimax test problems for special capacities. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 58(3):397–405, 1981.
- Fabian Beigang. On the advantages of distinguishing between predictive and allocative fairness in algorithmic decision-making. *Minds and Machines*, 32(4):655–682, 2022.
- Fabio Bellini and Valeria Bignozzi. On elicitable risk measures. *Quantitative Finance*, 15(5):725–733, 2015.
- Fabio Bellini and Emanuela Rosazza Gianin. On Haezendonck risk measures. *Journal of Banking & Finance*, 32(6):986–994, 2008.
- Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.
- Colin Bennett and Robert Sharpley. *Interpolation of Operators*. Academic Press, 1988.
- James O. Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.
- Jöran Bergh and Jörgen Löfström. *Interpolation Spaces: An Introduction*. Springer, 1976.
- Daniel Bernoulli. Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):23–36, 1738/1954.
- Sara Biagini and Marco Frittelli. On the extension of the Namioka-Klee theorem and on the Fatou property for risk measures. In Freddy Delbaen, Miklós Rásonyi, and Christophe Stricker, editors, *Optimality and Risk – Modern Trends in Mathematical Finance*, pages 1–28. Springer, 2009.
- Laurent Bienvenu, Glenn Shafer, and Alexander Shen. On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics*, 5(1):1–40, 2009.
- Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524, 2020.
- Christopher J. Bishop and Yuval Peres. *Fractals in probability and analysis*. Cambridge University Press, 2017.
- Joar Björk, Gert Helgesson, and Niklas Juth. Better in theory than in practise? challenges when applying the luck egalitarian ethos in health care policy. *Medicine, Health Care and Philosophy*, 23:735–742, 2020.

- Neeltje E. Blankenstein, Jiska S. Peper, Eveline A. Crone, and Anna C.K. van Duijvenvoorde. Neural mechanisms underlying risk and ambiguity attitudes. *Journal of Cognitive Neuroscience*, 29(11):1845–1859, 2017.
- Avrim Blum, Nika Haghtalab, Ariel D. Procaccia, and Mingda Qiao. Collaborative PAC learning. In *Advances in Neural Information Processing Systems*, volume 30. 2017.
- Ivan Boldyrev and Ekaterina Svetlova. *Enacting dismal science: New perspectives on the performativity of economics*. Springer, 2016.
- Émile Borel. *Probability and Certainty*. Walker and Company, 1963. Translation of *Probabilité et Certitude*, Presses Universitaires de France, 1950.
- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. The MIT Press, 2000.
- Seamus Bradley. Imprecise Probabilities. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2019 edition, 2019.
- Seamus Bradley and Katie Steele. Should subjective probabilities be sharp? *Episteme*, 11(3):277–289, 2014.
- Thomas Breuer and Imre Csiszár. Measuring distribution model risk. *Mathematical Finance*, 26(2):395–411, 2016.
- Nicolas Brisset. Economics is not always performative: some limits for performativity. *Journal of Economic Methodology*, 23(2):160–184, 2016.
- Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009.
- Efim M. Bronshteyn and L.D. Ivanov. The approximation of convex sets by polyhedra. *Siberian Mathematical Journal*, 16(5):852–853, 1975.
- Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.
- Yurii Abramovich Brudnyi, Selim Grigor’evich Krein, and Evgenii Mikhailovich Semenov. Interpolation of linear operators. *Itogi Nauki i Tekhniki, Seriya Matematicheskii Analiz*, 24: 3–163, 1986. English translation in *Journal of Soviet Mathematics*, 42(6), 2009–2113, September 1988.
- Yuri A. Brudnyi and Natan Ya. Krugljak. *Interpolation Functors and Interpolation Spaces*, volume 1. North-Holland, 1991.
- Lara Buchak. *Risk and rationality*. Oxford University Press, 2013.
- Lara Buchak. Taking risks behind the veil of ignorance. *Ethics*, 127(3):610–644, 2017.

- Christoph Bühren, Fabian Meier, and Marco Pleßner. Ambiguity aversion: bibliometric analysis and literature review of the last 60 years. *Management Review Quarterly*, 73(2): 495–525, 2023.
- Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *preprint*, 2005. URL <http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf>. Accessed: 2024-10-14.
- Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (ACM FAT*)*, volume 81, pages 77–91. PMLR, 2018.
- Laure Cabantous. Ambiguity aversion in the field of insurance: Insurers attitude to imprecise and conflicting probability estimates. *Theory and Decision*, 62(3):219–240, 2007.
- Michel Callon, editor. *The laws of the markets*. Oxford: Blackwell, 1998.
- Cristian Calude and Tudor Zamfirescu. Most numbers obey no probability laws. *Publicationes Mathematicae Debrecen*, 54(Supplement):619–623, 1999.
- Cristian S. Calude. *Information and randomness: an algorithmic perspective*. Springer Science & Business Media, 2002.
- Cristian S. Calude, Solomon Marcus, and Ludwig Staiger. A topological characterization of random sequences. *Information Processing Letters*, 88(5):245–250, 2003.
- Colin Camerer and Martin Weber. Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of risk and uncertainty*, 5:325–370, 1992.
- Georges Canguilhem. *On the Normal and the Pathological*. D. Riedel, 1978.
- Alexander W. Cappelen, James Konow, Erik Ø. Sørensen, and Bertil Tungodden. Just luck: An experimental study of risk-taking and fairness. *American Economic Review*, 103(4): 1398–1413, 2013.
- Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Imprecise Bayesian neural networks. *arXiv preprint arXiv:2302.09656*, 2023.
- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*, pages 39–57. IEEE, 2017.
- Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12, 2022. Article number: 4209.
- Marco E. G. V. Cattaneo. Empirical interpretation of imprecise probabilities. In Alessandro Antonucci, Giorgio Corani, Inés Couso, and Sébastien Destercke, editors, *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, volume 62 of *Proceedings of Machine Learning Research*, pages 61–72. PMLR, 2017.

- Nicholas R. Cavanaugh, Alexander Gershunov, Anna K. Panorska, and Tomasz J. Kozubowski. The probability distribution of intense daily precipitation. *Geophysical Research Letters*, 42(5):1560–1567, 2015.
- Alberto Cevolini and Elena Esposito. From pool to profile: Social consequences of algorithmic prediction in insurance. *Big Data & Society*, 7(2):1–11, 2020.
- Alberto Cevolini and Elena Esposito. From actuarial to behavioural valuation. the impact of telematics on motor insurance. *Valuation Studies*, 9(1):109–139, 2022.
- Linda Chamakh, Emmanuel Gobet, and Zoltán Szabó. Orlicz random Fourier features. *Journal of Machine Learning Research*, 21(145):1–37, 2020.
- Linda Chamakh, Emmanuel Gobet, and Wenjun Liu. Orlicz norms and concentration inequalities for β -heavy tailed random variables. *HAL preprint*, hal-03175697v3, 2021.
- Jake Chandler. Subjective probabilities need not be sharp. *Erkenntnis*, 79:1273–1286, 2014.
- Arthur Charpentier. Insurance: Discrimination, biases & fairness. *Opinions & Debates*, 2022. URL <https://www.institutlouisbachelier.org/en/insurance-discrimination-biases-fairness/>. Accessed on June 2, 2023.
- Alain Chateauneuf. Modeling attitudes towards uncertainty and risk through the use of Choquet integral. *Annals of Operations Research*, 52(1):1–20, 1994.
- Ping Chen. Nonequilibrium and nonlinearity — a bridge between the two cultures. In George P. Scott, editor, *Time, rhythms, and chaos in the new dialogue with nature*, pages 67–85. Iowa State University Press, 1991.
- Zengjing Chen, Panyu Wu, and Baoming Li. A strong law of large numbers for non-additive probabilities. *International Journal of Approximate Reasoning*, 54(3):365–377, 2013.
- Patrick Cheridito and Tianhui Li. Risk measures on Orlicz hearts. *Mathematical Finance*, 19(2):189–214, 2009.
- Alexander Cherny and Dilip Madan. New measures for performance evaluation. *The Review of Financial Studies*, 22(7):2571–2606, 2009.
- Emilie Chouzenoux, Henri Gérard, and Jean-Christophe Pesquet. General risk measures for robust machine learning. *arXiv preprint arXiv:1904.11707*, 2019.
- Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. *Advances in Neural Information Processing Systems*, 27, 2014.
- Alonzo Church. On the concept of a random sequence. *Bulletin of the American Mathematical Society*, 46(2):130–135, 1940.
- Jessi Cisewski, Joseph B. Kadane, Mark J. Schervish, Teddy Seidenfeld, and Rafael Stern. Standards for modest Bayesian credences. *Philosophy of Science*, 85(1):53–78, 2018.
- Karen A. Clifford and Russel P. Iuculano. AIDS and insurance: the rationale for AIDS-related testing. *Harvard Law Review*, 100(7):1806–1825, 1987.

- Fernando Cobos and Joaquim Martín. On interpolation of function spaces by methods defined by means of polygons. *Journal of Approximation Theory*, 132(2):182–203, 2005.
- Fernando Cobos and Luz M. Fernández-Cabrera. The fundamental function of certain interpolation spaces generated by n -tuples of rearrangement-invariant spaces. In Pankaj Jain and Hans-Jürgen Schmeisser, editors, *Function Spaces and Inequalities*, pages 1–14. Springer, 2017.
- Gert Cooman and Jasper De Bock. Computable randomness is inherently imprecise. In *International Symposium on Imprecise Probability: Theories and Applications*, volume 62, pages 133–144. PMLR, 2017.
- Dan Cooper and Brian Grinder. Probability, gambling and the origins of risk management. *Financial History Magazine*, 93:10–11, 2009.
- Giorgio Corani and Marco Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9(20):581–621, 2008.
- Inés Couso and Serafín Moral. Sets of desirable gambles: conditioning, representation, and precise probabilities. *International Journal of Approximate Reasoning*, 52(7):1034–1055, 2011.
- Fabio Cozman. Concentration inequalities and laws of large numbers under epistemic and regular irrelevance. *International Journal of Approximate Reasoning*, 51(9):1069–1084, 2010.
- Fabio Cozman and Lonnie Chrisman. Learning convex sets of probability from data. Technical report, Carnegie Mellon University, 1997. CMU-RI-TR 97-25.
- Yunan Cui, Lifen Duan, Henryk Hudzik, and Marek Wisła. Basic theory of p -Amemiya norm in Orlicz spaces ($1 \leq p \leq \infty$): extreme points and rotundity in Orlicz spaces endowed with these norms. *Nonlinear Analysis: Theory, Methods & Applications*, 69(5-6):1796–1816, 2008.
- Sebastian Curi, Kfir Y. Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 1036–1047, 2020.
- Shawn P. Curley, J. Frank Yates, and Richard A. Abrams. Psychological sources of ambiguity avoidance. *Organizational behavior and human decision processes*, 38(2):230–256, 1986.
- Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, volume 80, pages 1096–1105. PMLR, 2018.
- Alexander D’Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.
- Norman Daniels. Insurability and the HIV epidemic: ethical issues in underwriting. *The Milbank Quarterly*, 68(4):497–525, 1990.

- Lorraine Daston. *Classical Probability in the Enlightenment*. Princeton University Press, 2023.
- Philip Dawid. On individual risk. *Synthese*, 194(9):3445–3474, 2017.
- Philip A. Dawid and Monica Musio. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014.
- Philip A. Dawid and Paola Sebastiani. Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, 27(1):65–81, 1999.
- Gert De Cooman and Jasper De Bock. Randomness and imprecision: a discussion of recent results. In *International Symposium on Imprecise Probability: Theories and Applications*, pages 110–121. PMLR, 2021.
- Gert De Cooman and Jasper De Bock. Randomness is inherently imprecise. *International Journal of Approximate Reasoning*, 141:28–68, 2022.
- Gert De Cooman and Enrique Miranda. Weak and strong laws of large numbers for coherent lower previsions. *Journal of Statistical Planning and Inference*, 138(8):2409–2432, 2008.
- Gert De Cooman, Filip Hermans, and Erik Quaeghebeur. Imprecise markov chains and their limit behavior. *Probability in the Engineering and Informational Sciences*, 23(4):597–635, 2009.
- Bruno de Finetti. Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, 17: 298–329, 1931.
- Bruno de Finetti. *Theory of probability: A critical introductory treatment*. John Wiley & Sons, 1974/2017.
- Abraham de Moivre. *The Doctrine of Chances: A Method of Calculating the Probabilities of Events in Play*. Frank Cass and Company Limited, 2nd edition, 1738/1967.
- Mitchell Dean. Risk, calculable and incalculable. *Soziale Welt*, pages 25–42, 1998.
- Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1):12–22, 1983.
- Anton Deitmar. *Analysis*. Springer, 2016.
- Freddy Delbaen. Coherent risk measures on general probability spaces. In Klaus Sandmann and Philipp J. Schönbucher, editors, *Advances in Finance and Stochastics: Essays in Honour of Dieter Sondermann*, pages 1–37. Springer, 2002.
- Michel Denuit, Jan Dhaene, Marc Goovaerts, Rob Kaas, and Roger Laeven. Risk measurement with equivalent utility principles. *Statistics & Risk Modeling*, 24(1):1–25, 2006.
- Michel Denuit, Arthur Charpentier, and Julien Trufin. Autocalibration and Tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics*, 101:485–497, 2021.
- Esther Derman and Shie Mannor. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

- Rabanus Derr and Robert C. Williamson. Fairness and randomness in machine learning: Statistical independence and relativization. *arXiv preprint arXiv:2207.13596*, 2022.
- Rabanus Derr and Robert C. Williamson. Systems of precision: Coherent probabilities on pre-Dynkin systems and coherent previsions on linear subspaces. *Entropy*, 25(9), 2023.
- Rabanus Derr and Robert C. Williamson. Four facets of forecast felicity: Calibration, predictiveness, randomness and regret. *arXiv preprint arXiv:2401.14483*, 2024.
- Alain Desrosières. *The politics of large numbers: A history of statistical reasoning*. Harvard University Press, 1998.
- Jan Dhaene, Steven Vanduffel, Qihe Tang, Marc J Goovaerts, Rob Kaas, David Vyncke, et al. Capital requirements, risk measures and comonotonicity. *Belgian Actuarial Bulletin*, 4(1): 53–61, 2004.
- Rainer Diaz-Bone and Emmanuel Didier. Introduction: The sociology of quantification — perspectives on an emerging field in the social sciences. *Historical Social Research*, 41(2): 7–26, 2016.
- Enrico Diecidue and Peter P. Wakker. On the intuition of rank-dependent utility. *Journal of Risk and Uncertainty*, 23(3):281–298, 2001.
- Simon Dietz and Falk Nienhörster. Pricing ambiguity in catastrophe risk insurance. *The Geneva Risk and Insurance Review*, 46(2):112–132, 2021.
- Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in neural information processing systems*, volume 34, pages 6478–6490. JMLR, 2021.
- Rui Ding. Stochastic dominance, risk, and weak sub-majorization with applications to portfolio optimization. In *Proceedings of the 2023 6th International Conference on Mathematics and Statistics, ICoMS '23*, pages 63–71, 2023.
- Paul Dommel and Alois Pichler. Convex risk measures based on divergence. *arXiv preprint arXiv:2003.07648*, 2020.
- Kate Donahue and Solon Barocas. Better together? how externalities of size complicate notions of solidarity and actuarial fairness. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, pages 185–195, 2021.
- Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. Adversarial distributional training for robust deep learning. *Advances in Neural Information Processing Systems*, 33:8270–8283, 2020.
- James Dow and Sérgio Ribeiro da Costa Werlang. Laws of large numbers for non-additive probabilities. 1993. URL <https://repositorio.fgv.br/server/api/core/bitstreams/5e83654c-e8a5-4ad7-b6e5-6ba93ee58a6f/content>. Accessed: 2024-04-03.
- John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378 – 1406, 2021.

- John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.
- Nelson Dunford and Jacob T. Schwartz. *Linear operators, part 1: general theory*. John Wiley & Sons, 1988.
- Théodora Dupont-Courtade. Insurance demand under ambiguity and conflict for extreme risks: Evidence from a large representative survey. *Assurances et gestion des risques*, 80(2): 291–322, 2012.
- Detlef Dürr. Bohmian mechanics. In Jean Bricmont, Detlef Dürr, Maria Carla Galavotti, Giancarlo Ghirardi, Francesco Petruccione, and Nino Zanghi, editors, *Chance in Physics: Foundations and Perspectives*, pages 115–131. Springer, 2001.
- Detlef Dürr and Ward Struyve. Typicality in the foundations of statistical physics and Borns rule. *Do Wave Functions Jump? Perspectives of the Work of GianCarlo Ghirardi*, pages 35–43, 2021.
- Detlef Dürr and Stefan Teufel. *Bohmian Mechanics: The Physics and Mathematics of Quantum Theory*. Springer, 2009.
- Detlef Dürr, Sheldon Goldstein, and Nino Zanghi. Quantum equilibrium and the origin of absolute uncertainty. *Journal of Statistical Physics*, 67:843–907, 1992.
- Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019.
- Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.
- Antony Eagle. Chance versus Randomness. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.
- Gerald A. Edgar and Louis Sucheston. *Stopping times and directed processes*. Cambridge University Press, 1992.
- Harold Gordon Eggleston. The fractional dimension of a set defined by decimal properties. *The Quarterly Journal of Mathematics*, 20:31–36, 1949.
- Benjamin Eidelson. Patterned inequality, compounding injustice, and algorithmic prediction. *American Journal of Law and Equality*, 1:252–276, 2021.
- Adam Elga. Subjective probabilities should be sharp. *Philosophers’ Imprint*, 10(5):1–11, 2010.
- Daniel Ellsberg. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961.
- Daniel Ellsberg. *Risk, ambiguity and decision*. Routledge, 2015.

- Larry G. Epstein. A definition of uncertainty aversion. *The Review of Economic Studies*, 66 (3):579–608, 1999.
- Larry G. Epstein and Martin Schneider. IID: independently and indistinguishably distributed. *Journal of Economic Theory*, 113(1):32–50, 2003.
- Richard Ericson, Dean Barry, and Aaron Doyle. The moral hazards of neo-liberalism: lessons from the private insurance industry. *Economy and Society*, 29(4):532–558, 2000.
- Wendy Nelson Espeland and Michael Sauder. Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, 113(1):1–40, 2007.
- Wendy Nelson Espeland and Mitchell L. Stevens. A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie*, 49(3):401–436, 2008.
- Johanna Etner, Meglena Jeleva, and Jean-Marc Tallon. Decision theory under ambiguity. *Journal of Economic Surveys*, 26(2):234–270, 2012.
- European Union. Directive 138/2009/ec of the european parliament and of the council of 25 november 2009 on the taking-up and pursuit of the business of insurance and reinsurance (solvency ii). Official Journal of the European Union, L 335, 2009. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32009L0138>.
- François Ewald. Die Versicherungs-Gesellschaft. *Kritische Justiz*, 22(4):385–393, 1989.
- François Ewald. Norms, discipline, and the law. *Representations*, 30:138–161, 1990.
- François Ewald. Insurance and risk. In *The Foucault effect: Studies in governmentality*, pages 197–210. The University of Chicago Press, 1991.
- François Ewald. *L’État providence*. Grasset, 1986.
- Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with average top-k loss. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Alberto Feduzi, Jochen Runde, and Carlo Zappia. De Finetti on the insurance of risks and uncertainties. *The British journal for the philosophy of science*, 2012.
- Oriel FeldmanHall, Paul Glimcher, Augustus L Baker, and Elizabeth A Phelps. Emotion and decision-making under uncertainty: Physiological arousal predicts increased gambling during ambiguity but not risk. *Journal of Experimental Psychology: General*, 145(10):1255, 2016.
- William Feller. *An introduction to probability theory and its applications, Volume 2, 2nd edition*. John Wiley & Sons, 1991.
- Luz M. Fernández-Cabrera. The fundamental function of spaces generated by interpolation methods associated to polygons. *Mediterranean Journal of Mathematics*, 14(17):1–15, 2017.
- Pablo I. Fierens. *Towards a chaotic probability model for frequentist probability*. PhD thesis, Cornell University, 2003.

- Pablo I. Fierens. An extension of chaotic probability models to real-valued variables. *International Journal of Approximate Reasoning*, 50(4):627–641, 2009.
- Pablo I. Fierens and Terrence L. Fine. Towards a frequentist interpretation of sets of measures. In *International Symposium on Imprecise Probabilities: Theories and Applications*, pages 179–187, 2001.
- Pablo I. Fierens and Terrence L. Fine. Towards a chaotic probability model for frequentist probability: The univariate case. In *International Symposium on Imprecise Probabilities: Theories and Applications*, pages 245–259, 2003.
- Pablo I. Fierens, Leonardo C. Rêgo, and Terrence L. Fine. A frequentist understanding of sets of measures. *Journal of Statistical Planning and Inference*, 139(6):1879–1892, 2009.
- Financial Services Authority. The Turner review - a regulatory response to the global banking crisis. Available at http://www.actuaries.org/CTTEES_TFRISKCRISIS/Documents/turner_review.pdf, 2009.
- Terrence L. Fine. On the apparent convergence of relative frequency and its implications. *IEEE Transactions on Information Theory*, 16(3):251–257, 1970.
- Terrence L. Fine. *Theories of probability: An examination of foundations*. Academic Press, 1973.
- Terrence L. Fine. A computational complexity viewpoint on the stability of relative frequency and on stochastic independence. In William L. Harper and Clifford Alan Hooker, editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Volume 1*, pages 29–40. Springer, 1976.
- Terrence L. Fine. Lower probability models for uncertainty and nondeterministic processes. *Journal of Statistical Planning and Inference*, 20(3):389–411, 1988.
- Terrence L. Fine. Mathematical alternatives to standard probability that provide selectable degrees of precision. In Alan Hájek and Christopher Hitchcock, editors, *The Oxford Handbook of Probability and Philosophy*. Oxford University Press, 2016.
- Hans Föllmer and Thomas Knispel. Entropic risk measures: Coherence vs. convexity, model ambiguity and robust large deviations. *Stochastics and Dynamics*, 11(02n03):333–351, 2011.
- Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, 2002.
- Hans Föllmer and Alexander Schied. *Stochastic Finance*. de Gruyter, 2016.
- Hans Föllmer and Stefan Weber. The axiomatic approach to risk measures for capital determination. *Annual Review of Financial Economics*, 7:301–337, 2015.
- Craig R. Fox, Michael Goedde-Menke, and David Tannenbaum. Ambiguity aversion and epistemic uncertainty. preprint SSRN 3922716, 2021. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3922716. Accessed: 2024-10-14.

- Christian L.E. Franzke, Susana Barbosa, Richard Blender, Hege-Beate Fredriksen, Thomas Laepple, Fabrice Lambert, Tine Nilsen, Kristoffer Rypdal, Martin Rypdal, Manuel G. Scotto, et al. The structure of climate variability across scales. *Reviews of Geophysics*, 58 (2):1–44, 2020. Paper number e2019RG000657.
- Edward W. Frees and Fei Huang. The discriminating (pricing) actuary. *North American Actuarial Journal*, 27(1):2–24, 2023.
- Sylvestre Frezal. Alea and heterogeneity: the tyrannous conflation, 2016. URL https://www.chaire-pari.fr/wp-content/uploads/2016/09/Alea-and-Heterogeneity_the-Tyrannous-Conflation_eng_MEP.pdf. Accessed on June 14, 2023.
- Sylvestre Frezal and Laurence Barry. Fairness in uncertainty: Some limits and misinterpretations of actuarial fairness. *Journal of Business Ethics*, 167:127–136, 2020.
- Rachel Z. Friedman. *Probable Justice: Risk, Insurance, and the Welfare State*. University of Chicago Press, 2020.
- Christian Fröhlich and Robert C. Williamson. Tailoring to the tails: Risk measures for fine-grained tail sensitivity. *arXiv preprint arXiv:2208.03066*, 2022.
- Christian Fröhlich and Robert C. Williamson. Tailoring to the tails: Risk measures for fine-grained tail sensitivity. *Transactions on Machine Learning Research*, 2023. URL <https://openreview.net/forum?id=UntUoeLwwu>.
- Christian Fröhlich and Robert C. Williamson. Risk measures and upper probabilities: Coherence and stratification. *Journal of Machine Learning Research*, 25, 2024a.
- Christian Fröhlich and Robert C. Williamson. Data models with two manifestations of imprecision. *arXiv preprint arXiv:2404.09741 (v1)*, 2024b.
- Christian Fröhlich and Robert C. Williamson. Insights from insurance for fair machine learning. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 407–421, 2024c.
- Christian Fröhlich and Robert C. Williamson. Insights from insurance for fair machine learning. *arXiv preprint arXiv:2306.14624*, 2024d.
- Christian Fröhlich and Robert C. Williamson. Scoring rules and calibration for imprecise probabilities. *arXiv preprint arXiv:2410.23001*, 2024e.
- Christian Fröhlich and Robert C. Williamson. Data models with two manifestations of imprecision. *arXiv preprint arXiv:2404.09741 (v2)*, 2025.
- Christian Fröhlich, Rabanus Derr, and Robert C. Williamson. Towards a strictly frequentist theory of imprecise probability. In *International Symposium on Imprecise Probability: Theories and Applications*, pages 230–240, 2023.
- Christian Fröhlich, Rabanus Derr, and Robert C. Williamson. Strictly frequentist imprecise probability. *International Journal of Approximate Reasoning*, 168, 2024.

- Walter Bryce Gallie. Essentially contested concepts. *Proceedings of the Aristotelian society*, 56:167–198, 1955.
- Francis Galton. *Natural Inheritance*. Macmillan and Company, 1889.
- Bruno Galvan. Bohmian mechanics and typicality without probability. *arXiv preprint quant-ph/0605162*, 2006.
- Bruno Galvan. Typicality vs. probability in trajectory-based formulations of quantum mechanics. *Foundations of Physics*, 37(11):1540–1562, 2007.
- Bruno Galvan. Non-probabilistic typicality, with application to quantum mechanics. *arXiv preprint arXiv:2209.14985*, 2022.
- William A. Gardner. *Statistical Spectral Analysis: A Nonprobabilistic Theory*. Prentice-Hall, Inc., USA, 1986.
- William A. Gardner. Transitioning away from stochastic process models, 2022. URL <https://cyclostationarity.com/wp-content/uploads/2022/07/Transitioning-Away-from-Stochastic-Process-Models.pdf>. Accessed: 2023-01-27.
- Joseph L. Gastwirth. A general definition of the Lorenz curve. *Econometrica*, 39(6):1037–1039, 1971.
- Jill Gaubling. Race, sex and genetic discrimination in insurance: What’s fair? *Cornell Law Review*, 80:1646, 1994.
- Paolo Ghirardato and Massimo Marinacci. Ambiguity made precise: A comparative foundation. *Journal of Economic Theory*, 102(2):251–289, 2002.
- Ronald N. Giere. Objective single-case probabilities and the foundations of statistics. In *Studies in Logic and the Foundations of Mathematics*, volume 74, pages 467–483. Elsevier, 1973.
- Kay Giesecke, Thorsten Schmidt, and Stefan Weber. Measuring the risk of large losses. *Journal of Investment Management (JOIM)*, Fourth Quarter, 2008. URL <https://ssrn.com/abstract=810886>.
- Gerd Gigerenzer, Zeno Swijtink, Theodore Porter, Lorraine Daston, and Lorenz Kruger. *The empire of chance: How probability changed science and everyday life*. Cambridge University Press, 1989.
- Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.
- Itzhak Gilboa, Andrew Postlewaite, and David Schmeidler. Is it always rational to satisfy Savage’s axioms? *Economics & Philosophy*, 25(3):285–296, 2009.
- Brian J. Glenn. Postmodernism: the basis of insurance. *Risk Management and Insurance Review*, 6(2):131–143, 2003a.
- Brian J. Glenn. Risk, insurance, and the changing nature of mutual obligation. *Law & Social Inquiry*, 28(1):295–314, 2003b.

- Tilman Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
- Sheldon Goldstein. Typicality and notions of probability in physics. In *Probability in physics*, pages 59–71. Springer, 2012.
- Ruobin Gong and Xiao-Li Meng. Judicious judgment meets unsettling updating: Dilation, sure loss and Simpson’s paradox. 2021.
- Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. *arXiv preprint arXiv:2109.05389*, 2021.
- Igor I. Gorban. Statistical instability of physical processes. *Radioelectronics and Communications Systems*, 54(9):499, 2011.
- Igor I. Gorban. *The Statistical Stability Phenomenon*. Springer, 2017.
- Igor I. Gorban. *Randomness and Hyper-randomness*. Springer, 2018.
- Jun-ya Gotoh and Stan Uryasev. Two pairs of families of polyhedral norms versus ℓ_p -norms: proximity and applications in optimization. *Mathematical Programming*, 156(1-2): 391–431, 2016.
- Ben Green. Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology*, 35, 2022. Article number: 90.
- Ido Greenberg, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Efficient risk-averse reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 32639–32652, 2022.
- Yves L. Grize and Terrence L. Fine. Continuous lower probability-based models for stationary processes with bounded and divergent time averages. *The Annals of Probability*, 15(2):783–803, 1987.
- Francisca Grommé and Stephan Scheel. Doing statistics, enacting the nation: The performative powers of categories. *Nations and nationalism*, 26(3):576–593, 2020.
- Peter D. Grünwald and Philip A. Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367 – 1433, 2004.
- Xiaoyang Gu and Jack H. Lutz. Effective dimensions and relative frequencies. *Theoretical computer science*, 412(48):6696–6711, 2011.
- Stanley P. Gudder. *Stochastic methods in quantum mechanics*. North Holland, 1979.
- Massimo Guidolin and Francesca Rinaldi. Ambiguity in asset pricing and portfolio choice: A review of the literature. *Theory and Decision*, 74:183–217, 2013.
- Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3964–3975. PMLR, 18–24 Jul 2021.

- Yair M. Guttman. *The concept of probability in statistical physics*. Cambridge University Press, 1999.
- Henryk Gzyl and Silvia Mayoral. On a relationship between distorted and spectral risk measures. *Revista de Economia Financiera*, 15:8–21, 2008.
- Ian Hacking. *The Taming of Chance*. Cambridge University Press, 1990.
- Ian Hacking. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press, 2nd edition, 2006.
- Ella Hafermalz, Kai Riemer, and Sebastian Boell. Enactment or performance? a non-dualist reading of Goffman. In *Beyond Interpretivism? New Encounters with Technology and Organization: IFIP WG 8.2 Working Conference on Information Systems and Organizations, IS&O 2016*, pages 167–181. Springer, Cham, 2016.
- Nika Haghtalab, Michael Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. In *Advances in Neural Information Processing Systems*, volume 35, pages 406–419. 2022.
- Alan Hájek. The reference class problem is your problem too. *Synthese*, 156:563–585, 2007.
- Alan Hájek. Fifteen arguments against hypothetical frequentism. *Erkenntnis*, 70(2):211–235, 2009.
- Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. In *Advances in Neural Information Processing Systems*, volume 35, pages 22969–22981, 2022.
- Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. 2023.
- Dorothee D. Haroske. *Envelopes and sharp embeddings of function spaces*. Chapman and Hall/CRC, 2006.
- Dorothee D. Haroske. Envelope functions in real interpolation spaces. A first approach. *Contemporary Mathematics*, 445:93–102, 2007.
- Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.
- Uri Hasson, Samuel A. Nastase, and Ariel Goldstein. Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, 2020.
- Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. A moral framework for understanding fair ML through economic models of equality of opportunity. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 181–190, 2019.
- Carol Anne Heimer. *Reactive risk and rational action: Managing moral hazard in insurance contracts*. University of California Press, 1985.

- Antonio J. Heras, Pierre-Charles Pradier, and David Teira. What was fair in actuarial fairness? *History of the Human Sciences*, 33(2):91–114, 2020.
- Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.
- Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty: A credal approach. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024.
- Robin M. Hogarth and Howard Kunreuther. Risk, ambiguity, and insurance. *Journal of risk and uncertainty*, 2:5–35, 1989.
- Matthew Holland and El Mehdi Haress. Learning with risk-averse feedback under potentially heavy tails. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 892–900. PMLR, 13–15 Apr 2021.
- Matthew Holland and El Mehdi Haress. Spectral risk-based learning using unbounded losses. In *International Conference on Artificial Intelligence and Statistics*, pages 1871–1886. PMLR, 2022.
- Richard Holmes. *Geometric functional analysis and its applications*. Springer, 1975.
- Benedikt Höltingen and Robert C. Williamson. On the richness of calibration. In *The 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1124–1138, 2023.
- Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.
- Feng Hu, Zengjing Chen, and Panyu Wu. A general strong law of large numbers for non-additive probabilities and its applications. *Statistics*, 50(4):733–749, 2016.
- Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.
- Lily Hu and Issa Kohler-Hausmann. What’s sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.
- Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, volume 80, pages 2029–2037. PMLR, 2018.
- Peter J. Huber. *Robust Statistics*. John Wiley & Sons, Inc., 1981.
- Henryk Hudzik and Lech Maligranda. Amemiya norm equals Orlicz norm in general. *Indagationes Mathematicae*, 11(4):573–585, 2000.
- Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

- Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Uncertainty in Artificial Intelligence*, pages 548–557. PMLR, 2022.
- Robert Huseby. Can luck egalitarianism justify the fact that some are worse off than others? *Journal of Applied Philosophy*, 33(3):259–269, 2016.
- Laura Iacovissi, Nan Lu, and Robert C. Williamson. A general framework for learning under corruption: Label noise, attribute noise, and beyond. *arXiv preprint arXiv:2307.08643*, 2023.
- Marat Ibragimov, Rustam Ibragimov, and Johan Walden. *Heavy-tailed distributions and robustness in economics and finance*. Springer, 2015.
- Yoaav Isaacs, Alan Hájek, and John Hawthorne. Non-measurability, imprecise credences, and imprecise chances. *Mind*, 2021.
- Victor I. Ivanenko. *Decision Systems and Nonstochastic Randomness*. Springer, 2010.
- Victor I. Ivanenko and Valery A. Labkovskii. On the functional dependence between the available information and the chosen optimality principle. In Vadim I. Arkin, A. Shiraev, and R. Wets, editors, *Stochastic Optimization*, pages 388–392. Springer, 1986a.
- Victor I. Ivanenko and Valery A. Labkovskii. A class of criterion-choosing rules. *Doklady Akademii Nauk SSSR*, 31(3):204–205, 1986b.
- Victor I. Ivanenko and Valery A. Labkovskii. A model of nonstochastic randomness. *Doklady Akademii Nauk SSSR*, 310(5):10591062, 1990.
- Victor I. Ivanenko and Valery A. Labkovskii. On the studying of mass phenomena. In *Proceedings of the 12th IFAC Triennial World Congress, Sydney, Australia*, pages 693–696, 1993.
- Victor I. Ivanenko and Valery A. Labkovskii. On regularities of mass phenomena. *Sankhya A*, 77:237–248, 2015.
- Victor I. Ivanenko and Bertrand Munier. Decision making in ‘random in a broad sense’ environments. *Theory and Decision*, 49:127–150, 2000.
- Victor I. Ivanenko and Illia Pasichnichenko. Expected utility for nonstochastic risk. *Mathematical Social Sciences*, 86:18–22, 2017.
- Alireza Javanmardi, David Stutz, and Eyke Hüllermeier. Conformalized credal set predictors. *arXiv preprint arXiv:2402.10723*, 2024.
- Edwin T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620–630, 1957.
- David J. Johnstone, Victor Richmond R. Jose, and Robert L. Winkler. Tailored scoring rules for probabilities. *Decision Analysis*, 8(4):256–268, 2011.
- Renée Jorgensen. Algorithms and the individual in criminal law. *Canadian Journal of Philosophy*, 52(1):61–77, 2022.

- Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, volume 29. 2016.
- Kapil D. Joshi. *Introduction to general topology*. New Age International, 1983.
- James M. Joyce. How probabilities reflect evidence. *Philosophical Perspectives*, 19:153–178, 2005.
- James M. Joyce. A defense of imprecise credences in inference and decision making. *Philosophical perspectives*, 24:281–323, 2010.
- Atoosa Kasirzadeh. Algorithmic fairness and structural injustice: Insights from feminist political philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’22, page 349356, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471.
- Konstantinos V. Katsikopoulos, Özgür Şimşek, Marcus Buckmann, and Gerd Gigerenzer. *Classification in the wild: The science and art of transparent decision making*. MIT Press, 2021.
- John Maynard Keynes. *A treatise on probability*. Macmillan and Company, limited, 1921.
- John Maynard Keynes. The general theory of employment. *The Quarterly Journal of Economics*, 51(2):209–223, 1937.
- Andrei Y. Khrennikov. *Interpretations of Probability – Second revised and extended edition*. Walter de Gruyter, 2009.
- Andrei Y. Khrennikov. *p-Adic valued distributions in mathematical physics*. Springer, 2013.
- Andrei Y. Khrennikov. *Probability and randomness: quantum versus classical*. Imperial College Press, 2016.
- Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in neural information processing systems*, volume 30, 2017.
- Mervyn King and John Kay. *Radical uncertainty: Decision-making for an unknowable future*. Hachette UK, 2020.
- Michael Kirchhof, Mark Collier, Seong Joon Oh, and Enkelejda Kasneci. Pretrained visual uncertainties. *arXiv preprint arXiv:2402.16569*, 2024.
- Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. Distributionally robust Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2174–2184. PMLR, 2020.
- Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.
- Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *The Thirty Sixth Annual Conference on Learning Theory*, volume 195, pages 5143–5145. PMLR, 2023.

- Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, pages 43:1–43:23, 2017.
- Achim Klenke. *Probability theory: a comprehensive course*. Springer, 2013.
- Peter Klibanoff. Characterizing uncertainty aversion through preference for mixtures. *Social Choice and Welfare*, 18(2):289–301, 2001.
- Carl Knight. Luck egalitarianism. *Philosophy Compass*, 8(10):924–934, 2013.
- Frank Knight. *Risk, uncertainty and profit*. Houghton Mifflin, 1921.
- Martin G. Kocher, Amrei Marie Lahno, and Stefan T. Trautmann. Ambiguity aversion is not universal. *European Economic Review*, 101:268–283, 2018.
- Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.
- Andreï Nikolaevich Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933.
- Andreï Nikolaevich Kolmogorov. Letter to Maurice Fréchet. 3 August 1939; reproduced in [Bienvenu et al., 2009], 1939.
- Andreï Nikolaevich Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Company, New York, 1950.
- Andreï Nikolaevich Kolmogorov. *Foundations of the theory of probability: Second English Edition*. Chelsea Publishing Company, 1956.
- Andreï Nikolaevich Kolmogorov. On logical foundations of probability theory. In K. Ito and J.V. Prokhorov, editors, *Probability theory and mathematical statistics: Proceedings of the Fourth USSR – Japan Symposium, held at Tbilisi, USSR, August 23-29, 1982*, pages 1–5. Springer, 1983.
- Jason Konek. Epistemic conservativity and imprecise credence. *preprint*, 2015. URL <https://philpapers.org/rec/KONECA>. Accessed: 2024-10-14.
- Jason Konek. IP scoring rules: foundations and applications. In *International Symposium on Imprecise Probabilities: Theories and Applications*, volume 103, pages 256–264. PMLR, 2019.
- Jason Konek. Evaluating imprecise forecasts. In *International Symposium on Imprecise Probability: Theories and Applications*, volume 215, pages 270–279. PMLR, 2023.
- Peter Kosmol and Dieter Müller-Wichards. *Optimization in function spaces*. De Gruyter, 2011.
- M. A. Krasnoselskii and YA. B. Rutickii. *Convex functions and Orlicz spaces*. P. Noordhoff, Groningen, 1961.

- Selim Grigor'evich Kreĭn, Jurii Ivanovvich Petunin, and Evgenii Mikhailovich Semenov. *Interpolation of Linear Operators*. American Mathematical Society, 1982a.
- Selim Grigor'evich Kreĭn, Jurii Ivanovvich Petunin, and Evgenii Mikhailovich Semenov. *Interpolation of Linear Operators*. American Mathematical Society, 1982b.
- Greta R. Krippner. Unmasked: A history of the individualization of risk. *Sociological Theory*, pages 83–104, 2023.
- Greta R. Krippner and Daniel Hirschman. The person of the category: the pricing of risk and the politics of classification in insurance and credit. *Theory and Society*, 51(5):685–727, 2022.
- Jiddu Krishnamurti. *Think on These Things*. HarperPerennial, 1989.
- Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pages 5436–5446. PMLR, 2020.
- Anton Kuhberger. The rationality of risky decisions: A changing message. *Theory & Psychology*, 12(4):427–452, 2002.
- Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. Informa, 2019.
- Anurag Kumar and Terrence L. Fine. Stationary lower probabilities and unstable averages. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 69(1):1–17, 1985.
- Matthias Kuppler, Christoph Kern, Ruben L Bach, and Frauke Kreuter. Distributive justice and fairness metrics in automated decision-making: How much overlap is there? *arXiv preprint arXiv:2105.01441*, 2021.
- Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. From fair predictions to just decisions? conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology*, 7, 2022.
- Shigeo Kusuoka. On law invariant coherent risk measures. In *Advances in mathematical economics*, volume 3, pages 83–95. Springer, 2001.
- Adam La Caze. Frequentism. In Alan Hájek and Christopher Hitchcock, editors, *The Oxford Handbook of Probability and Philosophy*. Oxford University Press, 2016.
- Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued and Variational Analysis*, 29(4):967–996, 2021.
- Xavier Landes. The normative foundations of (social) insurance: Technology, social practices and political philosophy. 2013. URL https://www.centroeinaudi.it/images/abook_file/WP-LPF_6_2013_Landes.pdf. Accessed on June 14, 2023.

- Xavier Landes. How fair is actuarial fairness? *Journal of Business Ethics*, 128:519–533, 2015.
- Xavier Landes and Nils Holtug. Insurance, equality and the welfare state: Political philosophy and (of) public insurance. *Res Publica*, 21:111–118, 2015.
- J. Dennis Lawrence. *A Catalog of Special Plane Curves*. Dover, 1972.
- Jaeho Lee, Sejun Park, and Jinwoo Shin. Learning bounds for risk-sensitive learning. *Advances in Neural Information Processing Systems*, 33:13867–13879, 2020.
- Sharon M. Lee. Racial classifications in the US census: 1890–1990. *Ethnic and Racial Studies*, 16(1):75–94, 1993.
- Turo-Kimmo Lehtonen and Jyri Liukko. The forms and limits of insurance solidarity. *Journal of Business Ethics*, 103:33–44, 2011.
- Turo-Kimmo Lehtonen and Jyri Liukko. Producing solidarity, inequality and exclusion through insurance. *Res publica*, 21(2):155–169, 2015.
- Turo-Kimmo Lehtonen and Ine Van Hoyweghen. Editorial: Insurance and the economization of uncertainty journal. *Journal of Cultural Economy*, 7(4):532–540, 2014.
- Johannes Lenhard and Wolfgang Krohn. The law of large numbers. edgar zilsels attempt at the foundation of physical and socio-historical laws. In *Edgar Zilsel: Philosopher, Historian, Sociologist*, pages 121–142. Springer, 2022.
- Liu Leqi, Audrey Huang, Zachary Lipton, and Kamyar Azizzadenesheli. Supervised learning with general risk functionals. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12570–12592. PMLR, 17–23 Jul 2022a.
- Liu Leqi, Audrey Huang, Zachary Lipton, and Kamyar Azizzadenesheli. Supervised learning with general risk functionals. In *International Conference on Machine Learning*, pages 12570–12592, 2022b.
- Jacek Leśkow and Antonio Napolitano. Foundations of the functional approach for signal analysis. *Signal Processing*, 86(12):3796–3825, 2006.
- Isaac Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press, 1980.
- Leonid A. Levin. A concept of independence with applications in various fields of mathematics. Technical Report MIT/LCS/TR-235, MIT, Laboratory for Computer Science, 1980.
- Ifat Levy, Jason Snell, Amy J. Nelson, Aldo Rustichini, and Paul W. Glimcher. Neural representation of subjective value under risk and ambiguity. *Journal of Neurophysiology*, 103(2):1036–1047, 2010.
- David K. Lewis. A subjectivists guide to objective chance. In Richard C. Jeffrey, editor, *Studies in Inductive Logic and Probability, Volume II*, pages 263–293. Berkeley: University of California Press, 1980.

- Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.
- Kasper Lippert-Rasmussen. Genetic discrimination and health insurance. *Res Publica*, 21(2):185–199, 2015.
- Hsin-Hsien Liu and Andrew M. Colman. Ambiguity aversion in the long run: Repeated decisions under risk and uncertainty. *Journal of Economic Psychology*, 30(3):277–284, 2009.
- Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158, 2018.
- Yuxi Liu. Beyond expectations, but within limits — the theory of coherent risk measures. Bachelor Thesis, Department of Mathematics, Australian National University, 2019.
- Jyri Liukko. Genetic discrimination, insurance, and solidarity: an analysis of the argumentation for fair risk classification. *New Genetics and Society*, 29(4):457–475, 2010.
- Edward H. Lockwood. *A Book of Curves*. Cambridge University Press, 1961.
- Michele Loi and Markus Christen. Choosing how to discriminate: Navigating ethical trade-offs in fair algorithmic design for the insurance sector. *Philosophy & Technology*, 34:967–992, 2021.
- Lola L. Lopes. Risk and distributional inequality. *Journal of Experimental Psychology: Human Perception and Performance*, 10(4):465, 1984.
- S. Lovejoy. A voyage through scales, a missing quadrillion and why the climate is not what you expect. *Climate Dynamics*, 44:3187–3210, 2015.
- Deborah Lupton. The diverse domains of quantified selves: self-tracking modes and dataveillance. *Economy and Society*, 45(1):101–122, 2016.
- Wilhelmus A. J. Luxemburg. *Banach function spaces*. PhD thesis, Technische Hogeschool te Delft, Netherlands, 1955.
- Fabio Maccheroni and Massimo Marinacci. A strong law of large numbers for capacities. *The Annals of Probability*, 33(3):1171–1178, 2005.
- Mark J. Machina and David Schmeidler. A more robust definition of subjective probability. *Econometrica*, 60(4):745–780, 1992.
- Donald MacKenzie. *An engine, not a camera: How financial models shape markets*. MIT Press, 2008.
- Donald MacKenzie, Fabian Muniesa, and Leung-Sea Siu, editors. *Do economists make markets?: on the performativity of economics*. Princeton University Press, 2008.
- Alexander Mafusalov and Stan Uryasev. CVaR (superquantile) norm: Stochastic case. *European Journal of Operational Research*, 249(1):200–208, 2016.

- Uskali Mäki. Performativity: Saving austin from MacKenzie. In *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, pages 443–453, 2013.
- Massimo Marinacci. Limit laws for non-additive probabilities and their frequentist interpretation. *Journal of Economic Theory*, 84(2):145–195, 1999.
- Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.
- Conor Mayo-Wilson and Gregory Wheeler. Scoring imprecise credences. *Philosophy and Phenomenological Research*, 93(1):55–78, 2016.
- Liz McFall. A good, average man: Calculation and the limits of statistics in enrolling insurance customers. *The Sociological Review*, 59(4):661–684, 2011.
- Liz McFall. Personalizing solidarity? the role of self-tracking in health insurance pricing. *Economy and Society*, 48(1):52–76, 2019.
- Liz McFall and Liz Moor. Who, or what, is insurtech personalizing?: persons, prices and the historical classifications of risk. *Distinktion: journal of social theory*, 19(2):193–213, 2018.
- Liz McFall, Gert Meyers, and Ine Van Hoyweghen. Editorial: The personalisation of insurance: Data, behaviour and innovation. *Big Data & Society*, 7(2):1–11, 2020.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1273–1282. PMLR, 2017.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.
- Ronak Mehta, Vincent Roulet, Krishna Pillutla, Lang Liu, and Zaid Harchaoui. Stochastic optimization for spectral risk measures. *arXiv preprint arXiv:2212.05149*, 2022.
- Ronak Mehta, Vincent Roulet, Krishna Pillutla, and Zaid Harchaoui. Distributionally robust optimization with bias and variance reduction. *arXiv preprint arXiv:2310.13863*, 2023.
- Celestino G. Méndez. On the law of large numbers, infinite games, and category. *The American Mathematical Monthly*, 88(1):40–42, 1981.
- Andrea Mennicken and Wendy Nelson Espeland. What’s new with numbers? sociological approaches to the study of quantification. *Annual Review of Sociology*, 45:223–245, 2019.
- Bruno Merz, Stefano Basso, Svenja Fischer, David Lun, Günter Blöschl, Ralf Merz, Björn Guse, Alberto Viglione, Sergiy Vorogushyn, Elena Macdonald, et al. Understanding heavy tails of flood peak distributions. *Water Resources Research*, 58(6):e2021WR030506, 2022.

- Gert Meyers and Ine Van Hoyweghen. Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. *Science as Culture*, 27(4):413–438, 2018.
- Zakaria Mhammedi and Robert C. Williamson. Constant regret, generalized mixability, and mirror descent. *arXiv preprint arXiv:1802.06965*, 2018.
- Mary Midgley. *Science as salvation: A modern myth and its meaning*. Routledge, 2013.
- M. A. Milevsky, S. D. Promislow, and V. R. Young. Killing the law of large numbers: Mortality risk premiums and the sharpe ratio. *Journal of Risk and Insurance*, 73(4):673–686, 2006.
- Michael J. Miller. Disparate impact and unfairly discriminatory insurance rates. In *Casualty Actuarial Society E-Forum*, Winter 2009, 2009.
- John R. Milton. The origin and development of the concept of the ‘laws of nature’. *European Journal of Sociology/Archives Européennes de Sociologie*, 22(2):173–195, 1981.
- Enrique Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658, 2008.
- Enrique Miranda and Gert de Cooman. Lower previsions. In *Introduction to Imprecise Probabilities*, pages 28–55. John Wiley & Sons, Ltd, 2014.
- Enrique Miranda, Inés Couso, and Pedro Gil. Extreme points of credal sets generated by 2-alternating capacities. *International Journal of Approximate Reasoning*, 33(1):95–115, 2003.
- Shira Mitchell, Eric Potash, Solon Barocas, Alexander D’Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.
- Annemarie Mol. *The body multiple: Ontology in medical practice*. Duke University Press, 2002.
- Ignacio Montes, Enrique Miranda, and Paolo Vicig. 2-monotone outer approximations of coherent lower probabilities. *International Journal of Approximate Reasoning*, 101:181–205, 2018.
- Liz Moor and Celia Lury. Price and the person: Markets, discrimination, and personhood. *Journal of Cultural Economy*, 11(6):501–513, 2018.
- Don A. Moore, Elizabeth R. Tenney, and Uriel Haran. Overprecision in judgment. In *The Wiley Blackwell Handbook of Judgment and Decision Making*, chapter 6, pages 182–209. John Wiley & Sons, Ltd, 2015.
- G. Cristina Mora. *Making Hispanics: How Activists, Bureaucrats, and Media Constructed a New American*. University of Chicago Press, 2014.
- Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning*, pages 799–806, 2010.

- Emanuel D. Moss. *The objective function: Science and society in the age of machine intelligence*. PhD thesis, City University of New York, 2021.
- Pietro Muliere and Marco Scarsini. A note on stochastic dominance and inequality measures. *Journal of Economic Theory*, 49(2):314–323, 1989.
- Gabriel Nagy. Topological vector spaces III: Finite dimensional spaces – Notes from the Functional Analysis Course (Fall 07 - Spring 08), 2007. URL <https://www.math.ksu.edu/~nagy/func-an-2007-2008/top-vs-3.pdf>. Accessed: 2023-01-13.
- Jayakrishnan Nair, Adam Wierman, and Bert Zwart. *The fundamentals of heavy tails: Properties, emergence, and estimation*. Cambridge University Press, 2022.
- Hongseok Namkoong. *Reliable Machine Learning Via Distributional Robustness*. PhD thesis, 2019.
- Hongseok Namkoong and John C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
- Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- Antonio Napolitano and William A. Gardner. Fraction-of-time probability: Advancing beyond the need for stationarity and ergodicity assumptions. *IEEE Access*, 10:34591–34612, 2022.
- Vu-Linh Nguyen, Haifei Zhang, and Sébastien Destercke. Learning sets of probabilities through ensemble methods. In *The 17th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2023)*, pages 270–283. Springer, 2023.
- Grégoire Nicolis and Ilya Prigogine. *Self-organization in nonequilibrium systems*. John Wiley & Sons, 1977.
- William D. Nordhaus. Economic policy in the face of severe tail events. *Journal of Public Economic Theory*, 14(2):197–219, 2012.
- Lars Olsen. Extremely non-normal numbers. *Mathematical Proceedings of the Cambridge Philosophical Society*, 137(1):43–53, 2004.
- Ferdinand Osterreicher and Igor Vajda. Statistical information and discrimination. *IEEE Transactions on Information Theory*, 39(3):1036–1039, 1993.
- John C. Oxtoby. *Measure and category: A survey of the analogies between topological and measure spaces (2nd edition)*. Springer, 1980.

- Ajay Kumar Pandey, L.A. Prashanth, and Sanjay P. Bhat. Estimation of spectral risk measures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12166–12173, 2021.
- Adrian Papamarcou and Terrence L. Fine. Stationarity and almost sure divergence of time averages in interval-valued probability. *Journal of Theoretical Probability*, 4(2):239–260, 1991a.
- Adrian Papamarcou and Terrence L. Fine. Unstable collectives and envelopes of probability measures. *The Annals of Probability*, 19(2):893–906, 1991b.
- Renato Pelessoni and Paolo Vicig. Imprecise previsions for risk measurement. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 11(04):393–412, 2003.
- Liang Peng and Yongcheng Qi. *Inference for heavy-tailed data: applications in insurance and finance*. Academic press, 2018.
- Shige Peng. Law of large numbers and central limit theorem under nonlinear expectations. *arXiv preprint math/0702358*, 2007.
- Shige Peng. *Nonlinear expectations and stochastic calculus under uncertainty: with robust CLT and G-Brownian motion*. Springer Nature, 2019.
- Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609, 2020.
- Floris Persiau, Jasper De Bock, and Gert De Cooman. A remarkable equivalence between non-stationary precise and stationary imprecise uncertainty models in computable randomness. In *International Symposium on Imprecise Probability: Theories and Applications*, volume 147, pages 244–253. PMLR, 2021.
- Floris Persiau, Jasper De Bock, and Gert de Cooman. On the (dis)similarities between stationary imprecise and non-stationary precise uncertainty models in algorithmic randomness. *International Journal of Approximate Reasoning*, 151:272–291, 2022.
- Richard Pettigrew. *Accuracy and the Laws of Credence*. Oxford University Press, 2016.
- Georg Ch. Pflug. Subdifferential representations of risk measures. *Mathematical programming*, 108(2):339–354, 2006.
- Georg Ch. Pflug and Werner Romisch. *Modeling, Measuring and Managing Risk*. World Scientific, 2007.
- Georg Ch. Pflug and Andrzej Ruszczyński. Risk measures for income streams. Technical report, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Mathematik, 2001. preprint, DOI: 10.18452/8262.
- Graeme M. Philip and D.F. Watson. Some speculations on the randomness of nature. *Mathematical Geology*, 19(6):571–573, 1987.
- Alois Pichler. The natural Banach space for version independent risk measures. *Insurance: Mathematics and Economics*, 53(2):405–415, 2013.

- Alois Pichler. Insurance pricing under ambiguity. *European Actuarial Journal*, 4(2):335–364, 2014.
- Alois Pichler. Premiums and reserves, adjusted by distortions. *Scandinavian Actuarial Journal*, 2015(4):332–351, 2015.
- Alois Pichler. A quantitative comparison of risk measures. *Annals of Operations Research*, 254(1):251–275, 2017.
- Alois Pichler and Alexander Shapiro. Uniqueness of Kusuoka representations. *arXiv preprint arXiv:1210.7257*, 2012.
- Luboř Pick, Alois Kufner, Oldřich John, and Svatopluk Fucík. *Function spaces*, 1. de Gruyter, 2013.
- Itamar Pitowsky. Typicality and the role of the Lebesgue measure in statistical mechanics. In *Probability in physics*, pages 41–58. Springer, 2012.
- Elad Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253*, 2018.
- Barbara Prainsack and Ine Van Hoyweghen. Shifting solidarities: Personalisation in insurance and medicine. *Shifting solidarities: Trends and developments in European societies*, pages 127–151, 2020.
- Ilya Prigogine. Time, structure, and fluctuations (Nobel prize lecture). *Science*, 201(4358):777–785, 1978.
- Ilya Prigogine. *From Being to Becoming: Time and Complexity in the Physical Sciences*. W.H. Freeman and Company, 1980.
- Ilya Prigogine and Isabelle Stengers. *Order out of chaos: Man’s new dialogue with nature*. Bantam Books, 1985.
- Antonio Punzo, Luca Bagnato, and Antonello Maruotti. Compound unimodal distributions for insurance losses. *Insurance: Mathematics and Economics*, 81:95–107, 2018.
- Helen Pushkarskaya, Michael Smithson, Jane E Joseph, Christine Corbly, and Ifat Levy. Neural correlates of decision-making under ambiguity and conflict. *Frontiers in behavioral neuroscience*, 9:325, 2015.
- Jarkko Pyysiäinen, Darren Halpin, and Andrew Guilfoyle. Neoliberal governance and responsabilization of agents: reassessing the mechanisms of responsibility-shift in neoliberal discursive environments. *Distinktion: Journal of Social Theory*, 18(2):215–235, 2017.
- John Quiggin. A theory of anticipated utility. *Journal of economic behavior & organization*, 3(4):323–343, 1982.
- John Quiggin. *Generalized expected utility theory: The rank-dependent model*. Springer Science & Business Media, 2012.
- Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset shift in machine learning*. MIT Press, 2008.

- Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.
- K.P.S. Bhaskara Rao and M. Bhaskara Rao. *Theory of charges: a study of finitely additive measures*. Academic Press, 1983.
- John Rawls. *A Theory of Justice*. Harvard University Press, 1971.
- Tim R  z. Group fairness: Independence revisited. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, pages 129–137, 2021.
- Lisa Rebert and Ine Van Hoyweghen. The right to underwrite gender: The goods & services directive and the politics of insurance pricing. *Tijdschrift Voor Genderstudies*, 18(4): 413–431, 2015.
- Yann R  bill  . Law of large numbers for non-additive measures. *Journal of Mathematical Analysis and Applications*, 352(2):872–879, 2009.
- Florian Rechfeld. Personalised genetic testing and its impact to insurance. *Swiss Re*, 2016. URL https://www.swissre.com/dam/jcr:24995a5d-5b66-42ea-a2b9-660458bc6e26/Personalised_genetic_testing_and_its_impact_to_insurance.pdf. Accessed on June 14, 2023.
- Leandro C. R  go and Terrence L. Fine. Estimation of chaotic probabilities. In *International Symposium on Imprecise Probabilities: Theories and Applications*, volume 5, pages 297–305, 2005.
- Hans Reichenbach. *The Theory of Probability: An Inquiry Into the Logical and Mathematical Foundations of the Calculus of Probability*. University of California Press, 1949.
- George E. Rejda. *Principles of risk management and insurance*. Addison Wesley Longman, Inc, 7th edition, 2001.
- Sidney I. Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.
- Mustapha Ridaoui and Michel Grabisch. Choquet integral calculus on a continuous support and its applications. *Operations Research and Decisions*, 26(1):73–93, 2016.
- Marc Rigter, Bruno Lacerda, and Nick Hawes. Risk-averse Bayes-adaptive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1142–1154, 2021.
-   ngel Rivas. On the role of joint probability distributions of incompatible observables in Bell and Kochen–Specker theorems. *Annals of Physics*, 411:167939, 2019.
- James A. Roberts, Tjeerd W. Boonstra, and Michael Breakspear. The heavy tail of the human brain. *Current Opinion in Neurobiology*, 31:164–172, 2015.
- R. Tyrrell Rockafellar. Integrals which are convex functionals. *Pacific Journal of Mathematics*, 24(3):525–539, 1968.

- R. Tyrrell Rockafellar and Johannes O Royset. Measures of residual risk with connections to regression, risk tracking, surrogate models, and ambiguity. *SIAM Journal on Optimization*, 25(2):1179–1208, 2015.
- R. Tyrrell Rockafellar and Stan Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1-2):33–53, 2013.
- R. Tyrrell Rockafellar, Stan Uryasev, and Michael Zabarankin. Risk tuning with generalized linear regression. *Mathematics of Operations Research*, 33(3):712–729, 2008.
- Pierre Rosanvallon. *The New Social Question: Rethinking the Welfare State*. Princeton University Press, 2000.
- Todd Rose. *The end of average: How to succeed in a world that values sameness*. Penguin UK, 2016.
- Ben-Zion A. Rubshtein, Genady Ya. Grabarnik, Mustafa A. Muratov, and Yulia S. Pashkova. *Foundations of symmetric spaces of measurable functions*. Springer, 2016.
- Jane E. Ruby. The origins of scientific “law”. *Journal of the History of Ideas*, 47(3):341–359, 1986.
- Amir Sadrolhefazi and Terrence L. Fine. Finite-dimensional distributions and tail behavior in stationary interval-valued probability models. *The Annals of Statistics*, 22(4):1840–1870, 1994.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. Published as a conference paper at ICLR 2020.
- Leonard J. Savage. *The foundations of statistics*. John Wiley & Sons, 1954.
- Eric Schechter. *Handbook of Analysis and its Foundations*. Academic Press, 1997.
- Mark J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989.
- David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571–587, 1989.
- Claus P. Schnorr. *Zufälligkeit und Wahrscheinlichkeit: eine algorithmische Begründung der Wahrscheinlichkeitstheorie*. Springer, 2007.
- Miriam Schoenfield. Chilling out on epistemic rationality. *Philosophical Studies*, 158(2): 197–219, 2012.
- Miriam Schoenfield. The accuracy and rationality of imprecise credences. *Noûs*, 51(4): 667–685, 2017.
- Miriam Schoenfield. Can imprecise probabilities be practically motivated? a challenge to the desirability of ambiguity aversion. *Philosophers’ Imprint*, 20(30):1–21, 2020.

- Moritz Schönherr and Friedemann Schüricht. Pure measures, density measures and the dual of L^∞ . *arXiv preprint arXiv:1710.02197*, 2017.
- Pola Schwöbel and Peter Remmers. The long arc of fairness: Formalisations and ethical discourse. In *Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency*, pages 2179–2188, 2022.
- Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, 53(8):1248–1261, 2012.
- Teddy Seidenfeld, Jessica Cisewski, Jay Kadane, Mark Schervish, and Rafael Stern. When large also is small conflicts between measure theoretic and topological senses of a negligible set. In *Presented at Pitt Workshop 3/17*, 2017.
- Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, page 5968. Association for Computing Machinery, 2019.
- Stephen Semmes. Interpolation of Banach spaces, differential geometry and differential equations. *Revista Matemática Iberoamericana*, 4(1):155–176, 1988.
- Glenn Shafer and Vladimir Vovk. *Probability and finance: it's only a game!* John Wiley & Sons, 2001.
- Glenn Shafer and Vladimir Vovk. *Game-theoretic foundations for probability and finance*. John Wiley & Sons, 2019.
- Mohammad Hossein Shaker and Eyke Hüllermeier. Ensemble-based uncertainty quantification: Bayesian versus credal inference. *arXiv preprint arXiv:2107.10384*, 2021.
- Ronen Shamir. The age of responsabilization: On market-embedded morality. *Economy and Society*, 37(1):1–19, 2008.
- Alexander Shapiro. On Kusuoka representation of law invariant risk measures. *Mathematics of Operations Research*, 38(1):142–152, 2013.
- Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.
- Tamar Sharon. Self-tracking for health and the quantified self: Re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare. *Philosophy & Technology*, 30(1):93–121, 2017.
- Alexander Shen. Algorithmic information theory and foundations of probability. *arXiv preprint arXiv:0906.4411v1*, 2009.
- Yun Shen, Michael J. Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328, 2014.
- Anurag Singh, Siu Lun Chau, Shahine Bouabid, and Krikamol Muandet. Domain generalisation via imprecise learning. *arXiv preprint arXiv:2404.04669*, 2024.

- Rahul Singh, Qinsheng Zhang, and Yongxin Chen. Improving robustness via risk averse distributional reinforcement learning. In *Learning for Dynamics and Control*, volume 120, pages 958–968. PMLR, 2020.
- Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.
- Lawrence Sklar. *Physics and chance: Philosophical issues in the foundations of statistical mechanics*. Cambridge University Press, 1995.
- Lawrence Sklar. Topology versus measure in statistical mechanics. *The Monist*, 83(2):258–273, 2000.
- Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London, 1776.
- Michael Smithson. Conflict aversion: preference for ambiguity vs conflict in sources and evidence. *Organizational behavior and human decision processes*, 79(3):179–198, 1999.
- Michael Smithson. Probability judgments under ambiguity and conflict. *Frontiers in psychology*, 6:674, 2015.
- Tasuku Soma and Yuichi Yoshida. Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*, 2020.
- Zhao Song, Ruosong Wang, Lin Yang, Hongyang Zhang, and Peilin Zhong. Efficient symmetric norm regression via linear sketching. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- George Soros. General theory of reflexivity. *The Financial Times*, 27 October 2009.
- Wolfgang Spohn. Knightian uncertainty meets ranking theory. *Homo Oeconomicus*, 34:293–311, 2017.
- Peter J.C. Spreij. Measure theoretic probability. *Lecture Notes*.
- H. Orri Stefánsson and Richard Bradley. What is risk aversion? *The British Journal for the Philosophy of Science*, 70(1):77–102, 2019.
- Deborah A. Stone. Ad missions. *The American Prospect*, 2001. URL <https://prospect.org/health/ad-missions/>. Accessed on May 17, 2023.
- Anastasios Stylianou. A typical number is extremely non-normal. *arXiv preprint arXiv:2006.02202*, 2020.
- Jie Sun, Xinmin Yang, Qiang Yao, and Min Zhang. Risk minimization, regret minimization and progressive hedging algorithms. *Mathematical Programming*, 181:509–530, 2020.
- Rick Swedloff. Risk classification’s big data (r) evolution. *Connecticut Insurance Law Journal*, 143:339–373, 2014.

- Akiko Takeda and Masashi Sugiyama. v -support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th international conference on Machine learning*, pages 1056–1063, 2008.
- Nassim Nicholas Taleb. *The black swan: The impact of the highly improbable*. Random House, 2007.
- Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In *Advances in neural information processing systems*, volume 28, 2015.
- Maiju Tanninen. Contested technology: Social scientific perspectives of behaviour-based insurance. *Big Data & Society*, 7(2):1–14, 2020.
- Terence Tao. 245B, notes 11: The strong and weak topologies. <https://terrytao.wordpress.com/2009/02/21/245b-notes-11-the-strong-and-weak-topologies/>, 2009. Accessed: 2023-01-13.
- Pedro Terán. Laws of large numbers without additivity. *Transactions of the American Mathematical Society*, 366(10):5431–5451, 2014.
- Yves Thiery and Caroline Van Schoubroeck. Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 31(2):190–211, 2006.
- Stefan T. Trautmann and Gijs Van De Kuilen. Ambiguity attitudes. In *The Wiley Blackwell Handbook of Judgment and Decision Making*, pages 89–116. John Wiley & Sons, Ltd., 2015.
- Matthias C.M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.
- Matthias C.M. Troffaes and Gert De Cooman. *Lower previsions*. John Wiley & Sons, 2014.
- Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.
- Natan TJoens, Thomas Krak, Jasper De Bock, and Gert de Cooman. A recursive algorithm for computing inferences in imprecise markov chains. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 15th European Conference, ECSQARU 2019*, pages 455–465. Springer, 2019.
- Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. *arXiv preprint arXiv:2102.05371*, 2021.
- Sara van de Geer and Johannes Lederer. The Bernstein–Orlicz norm and deviation inequalities. *Probability theory and related fields*, 157(1):225–250, 2013.
- Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.
- A.E. Van Heerwaarden and R. Kaas. The Dutch premium principle. *Insurance: Mathematics and Economics*, 11(2):129–133, 1992.

- Grant Van Horn, Oisín Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.
- Ine Van Hoyweghen. On the politics of calculative devices: performing life insurance markets. *Journal of Cultural Economy*, 7(3):334–352, 2014.
- Ine Van Hoyweghen. Genomics and insurance: The lock-in effects of a politics of genetic solidarity. In *Handbook of Genomics, Health and Society*, pages 203–211. Routledge, 2018.
- Ine Van Hoyweghen, Klasien Horstman, and Rita Schepers. Making the normal deviant: The introduction of predictive medicine in life insurance. *Social Science & Medicine*, 63(5):1225–1235, 2006.
- Ine Van Hoyweghen, Klasien Horstman, and Rita Schepers. Genetic risk carriers and lifestyle risk takers. which risks deserve our legal protection in insurance? *Health Care Analysis*, 15:179–193, 2007.
- Michiel van Lambalgen. Von Mises’ definition of random sequences reconsidered. *The Journal of Symbolic Logic*, 52(3):725–755, 1987.
- John Venn. *The Logic of Chance*. MacMillan, 1876.
- Roel Verbelen, Katrien Antonio, and Gerda Claeskens. Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1275–1304, 2018.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. 2020. URL <https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.html>. Version of 2020-06-09.
- Paolo Vicig. Financial risk measurement with imprecise probabilities. *International Journal of Approximate Reasoning*, 49(1):159–174, 2008.
- Nithia Vijayan et al. Policy gradient methods for distortion risk measures. *arXiv preprint arXiv:2107.04422*, 2021.
- Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and Julyan Arbel. Understanding priors in Bayesian neural networks at the unit level. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6458–6467. PMLR, 09–15 Jun 2019.
- Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.
- Elart Von Collani. A note on the concept of independence. *Economic Quality Control*, 21(1): 155–164, 2006.

- Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5(1):52–99, 1919.
- Richard von Mises. *Wahrscheinlichkeitsrechnung und ihre Anwendungen in der Statistik und Theoretischen Physik*. Deuticke, 1931.
- Richard von Mises. *Probability, Statistics, and Truth*. Dover, 1981.
- Richard von Mises and Hilda Geiringer. *Mathematical theory of probability and statistics*. Academic press, 1964.
- John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*, 2nd rev. Princeton University Press, 1947.
- Ed Vosselman. The performativity thesis and its critics: Towards a relational ontology of management accounting. *Accounting and Business Research*, 44(2):181–203, 2014.
- Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting. In *International Workshop on Artificial Intelligence and Statistics*, pages 365–372. PMLR, 2005.
- Kate Vredenburg. Fairness. In *The Oxford Handbook of AI Governance*. Oxford University Press, 2022.
- Peter Walley. *Statistical reasoning with imprecise probabilities*. Chapman-Hall, 1991.
- Peter Walley and Terrence L. Fine. Towards a frequentist theory of upper and lower probability. *The Annals of Statistics*, 10(3):741–761, 1982.
- Daniel J. Walters, Gülden Ülkümen, David Tannenbaum, Carsten Erner, and Craig R. Fox. Investor behavior under epistemic vs. aleatory uncertainty. *Management Science*, 69(5): 2761–2777, 2023.
- Michael A. Walters. Risk classification standards. In *Proceedings of the Casualty Actuarial Society*, volume 68, pages 1–18, 1981.
- Kaizheng Wang, Keivan Shariatmadar, Shireen Kudukil Manchingal, Fabio Cuzzolin, David Moens, and Hans Hallez. CreINNs: Credal-set interval neural networks for uncertainty estimation in classification tasks. *arXiv preprint arXiv:2401.05043*, 2024.
- Shaun Wang. Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics*, 17(1):43–54, 1995.
- Shaun Wang. A class of distortion operators for pricing financial and insurance risks. *Journal of Risk and Insurance*, pages 15–36, 2000.
- Shaun Wang and Jan Dhaene. Comonotonicity, correlation order and premium principles. *Insurance: Mathematics and Economics*, 22(3):235–242, 1998.
- Shaun Wang, Virginia R. Young, and Harry H. Panjer. Axiomatic characterization of insurance prices. *Insurance: Mathematics and Economics*, 21(2):173–183, 1997.

- Jiaheng Wei, Harikrishna Narasimhan, Ehsan Amid, Wen-Sheng Chu, Yang Liu, and Abhishek Kumar. Distributionally robust post-hoc classifiers under prior shifts. *arXiv preprint arXiv:2309.08825*, 2023. Published at ICLR 2023.
- Friedel Weinert, editor. *Laws of nature: Essays on the philosophical, scientific and historical dimensions*, volume 8. Walter de Gruyter, 1995.
- Jon A. Wellner. The Bennett-Orlicz norm. *Sankhya A*, 79(2):355–383, 2017.
- Gregory Wheeler. A gentle approach to imprecise probability. In Thomas Augustin, Fabio Cozman, and Gregory Wheeler, editors, *Reflections on the Foundations of Statistics: Essays in Honor of Teddy Seidenfeld*. Springer, 2021.
- David Wilkie. Mutuality and solidarity: assessing risks and sharing losses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1357):1039–1044, 1997.
- David Williams. *Probability with martingales*. Cambridge University Press, 1991.
- Jon Williamson. A dynamic interaction between machine learning and the philosophy of science. *Minds and Machines*, 14(4):539–549, 2004.
- Robert C. Williamson and Zac Cranko. The geometry and calculus of losses. *Journal of Machine Learning Research*, 24(342):1–72, 2023.
- Robert C. Williamson and Zac Cranko. Information processing equalities and the information–risk bridge. *Journal of Machine Learning Research*, 25(103):1–53, 2024.
- Robert C. Williamson and Aditya Menon. Fairness risk measures. In *International Conference on Machine Learning*, volume 97, pages 6786–6797. PMLR, 2019.
- Robert L. Winkler. Evaluating probabilities: Asymmetric scoring rules. *Management Science*, 40(11):1395–1405, 1994.
- Shuyi Wu, Sai Sun, Julia A Camilleri, Simon B Eickhoff, and Rongjun Yu. Better the devil you know than the devil you don’t: Neural processing of risk and ambiguity. *NeuroImage*, 236:118109, 2021.
- Xi Xin and Fei Huang. Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal*, pages 1–35, 2023.
- Aolin Xu. Continuity of generalized entropy and statistical learning. *arXiv preprint arXiv:2012.15829*, 2020.
- Menahem E. Yaari. The dual theory of choice under risk. *Econometrica*, 55(1):95–115, 1987.
- Donggeun Yoo and In So Kweon. Learning loss for active learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102, 2019.
- Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

- Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216, 2021a. Paper number 106775.
- Jiaqi Zhang, Yanyan Tang, and Jie Xiong. Conditional strong law of large numbers under g-expectations. *Symmetry*, 16(3):272, 2024.
- Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. In *International Conference on Learning Representations*, 2021b.
- Ning Zhang and Yuting Lan. A strong law of large numbers for independent random variables under non-additive probabilities. *Communications in Statistics — Theory and Methods*, 49(21):5252–5272, 2020.
- Shengjia Zhao and Stefano Ermon. Right decisions from wrong predictions: A mechanism design alternative to individual calibration. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pages 2683–2691. PMLR, 2021.
- Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324, 2021a.
- Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *arXiv preprint arXiv:2107.05719*, 2021b.
- Edgar Zilsel. *Das Anwendungsproblem. Ein philosophischer Versuch über das Gesetz der groSSen Zahlen und die Induktion. Leipzig*. Barth, 1916.
- Edgar Zilsel. The genesis of the concept of physical law. *The Philosophical Review*, 51(3): 245–279, 1942.