Journal of Machine Learning Research 25 (2024) 1-100 Submitted 6/22; Revised 1/24; Published 5/24

Risk Measures and Upper Probabilities: Coherence and Stratification

Christian Fröhlich CHRISTIAN.FROEHLICH@UNI-TUEBINGEN.DE
Robert C. Williamson BOB.WILLIAMSON@UNI-TUEBINGEN.DE
University of Tübingen
and Tübingen AI Center
Tübingen, Germany

Editor: Bharath Sriperumbudur

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.

Keywords: coherent risk measures, imprecise probability, coherent upper previsions, rearrangement invariant function norms, Choquet integrals, spectral risk measures, ambiguity.

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,

©2024 Christian Fröhlich and Robert C. Williamson.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/22-0641.html.

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; 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.

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

1. This 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).

> 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 outcomes[2]. Urn II is an ambiguous urn, as the subject must entertain a 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).

Ellsberg's 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 Ellsberg's 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

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

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

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

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,

2. This holds when adopting the "principal principle" of Lewis (1980), which asserts that knowledge of chances requires that these be taken as subjective probabilities.

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., 2021), 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 and Prashanth (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.

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) and coherent risk measures (Section 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 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 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 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. 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.

2.1 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 $X : \Omega \to \mathbb{R}$, 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:

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

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:

$$\overline{P}(X) := \inf\{\alpha \in \mathbb{R} : X - \alpha \in \mathcal{D}\}. \tag{1}$$

We interpret $\overline{P}(X)$ 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*:

$$\begin{aligned}\underline{P}(X) &:= -\overline{P}(-X)\\ &= -\inf\{\alpha \in \mathbb{R} : -X - \alpha \in \mathcal{D}\}\\ &= \sup\{\alpha \in \mathbb{R} : \alpha - X \in \mathcal{D}\},\end{aligned}$$

which specifies the 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 (1), it can be shown to satisfy the properties (Walley, 1991, p. 65):

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

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

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

We call a functional satisfying P1-P3 a *coherent upper prevision.* The corresponding coherent set of desirable gambles can be defined as $\mathcal{D} := \{X : \overline{P}(X) \leqslant 0\}$, a definition which interacts well with (1). P2 and P3 together imply

P4. $\overline{P}(\alpha X + (1-\alpha)Y) \leqslant \alpha\overline{P}(X) + (1-\alpha)\overline{P}(Y)\ \forall \alpha \in [0,1]$ (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. $X(\omega) \leqslant Y(\omega)\ \forall \omega \in \Omega \Rightarrow \overline{P}(X) \leqslant \overline{P}(Y)$ (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 $\overline{P}(X) = -\overline{P}(-X)$. For a coherent prevision, it holds that $\underline{P}(X) \leqslant \overline{P}(X)$ and we may call the width of the interval $\left[\underline{P}(X), \overline{P}(X)\right]$ 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 1. *A functional $\overline{P}$ defined on $\mathcal{L}$ avoids sure loss if*

$$\forall n \in \mathbb{N} : \forall X_1, .., X_n \in \mathcal{L} : \sup_{\omega \in \Omega} \left[\sum_{j=1}^{n} \overline{P}(X_j) - X_j(\omega) \right] \geqslant 0. \tag{2}$$

Consider what happens if (2) fails. Then $\forall \omega \in \Omega : \sum_{j=1}^{n} \overline{P}(X_j) < \sum_{j=1}^{n} X_j(\omega)$. This means that our risk assessments $\overline{P}(X_j)$ 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 $\mathcal{Q} = \{Q : Q(X) \leqslant \overline{P}(X)\ \forall X \in \mathcal{L}, Q \text{ is linear prevision}\}$ is non-empty. We call such a set $\mathcal{Q}$ an *envelope*.[3] Then we can construct a canonical coherent extension of $\overline{P}$ by forming the supremum over this set

$$\overline{E}(X) = \sup_{Q \in \mathcal{Q}} Q(X). \tag{3}$$

3. In the literature on imprecise probabilities, the functional $\overline{E}$ in (3) 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).

This process is the *natural extension* of $\overline{P}$ and yields a coherent upper prevision if and only if $\overline{P}$ avoids sure loss. If $\overline{P}$ was already coherent, then $\overline{E} = \overline{P}$. On the other hand, if $\overline{P}$ 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 $\overline{P'}$ which is dominated by $\overline{P}$, meaning $\overline{P'}(X) \leqslant \overline{P}(X)\ \forall X \in \mathcal{L}$, the natural extension lies in-between: $\overline{P'}(X) \leqslant \overline{E}(X) \leqslant \overline{P}(X)\ \forall X \in \mathcal{L}$. In this sense, the natural extension is the most pessimistic one, as it reduces the risk assessment by $\overline{P}$ just as little as necessary to achieve coherence.

Conversely, every coherent upper prevision can be written in the form of (3) 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 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

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

so that $\overline{P}(A) := \overline{P}(\chi_A)$ is an *upper probability* and $\underline{P}(A) := 1 - \overline{P}(\chi_{A^C})$ a *lower probability*, where A^C is the complement of A, i.e. $\Omega \backslash A$. 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 $\overline{P}$ to be defined on a field[4] of events. Some consequences of coherence are then (Walley, 1991, p. 84):

Pa) $0 \leqslant \underline{P}(A) \leqslant \overline{P}(A) \leqslant 1$

Pb) $\underline{P}(\varnothing) = \overline{P}(\varnothing) = 0; \quad \underline{P}(\Omega) = \overline{P}(\Omega) = 1$

Pc) $A \subseteq B \Rightarrow \big(\underline{P}(A) \leqslant \underline{P}(B) \text{ and } \overline{P}(A) \leqslant \overline{P}(B)\big)$.

Like for previsions, the width of the interval $[\underline{P}(A), \overline{P}(A)]$ is a natural measure for the degree of imprecision. An interesting interpretation for this comes from a comparison to modal

4. A 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.

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:

K1) $P(A) \geqslant 0$

K2) $P(\Omega) = 1$

K3) $P(A \cup B) = P(A) + P(B)$, if $A \cap B = \varnothing$.

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. (3)) 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 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.

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,

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

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 properties[6]:

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

C2. $R(X + Y) \leqslant R(X) + R(Y)$ (subadditivity)

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

C4. $X(\omega) \leqslant Y(\omega)\ \forall \omega \Rightarrow R(X) \leqslant R(Y)$ (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 Vicig, 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.

Theorem 2. *(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 $R(X) \leqslant \sup(X)$. Since $X \leqslant \sup(X)$, we have by monotonicity and translation equivariance that $R(X) \leqslant R(\sup(X)) = R(0 + \sup(X)) = R(0) + \sup(X) = \sup(X)$. 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 3.7.

6. For 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.

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

$$\overline{P}(X) = \sup_{Q \in \mathcal{Q}} Q(X),$$

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 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 3.2)

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

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 3.2). Note that now the definition of monotonicity is adapted to the measure[7]:

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

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 $\overline{P}(X) = \overline{P}(X')$ 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 4.

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

3.2 Envelope Representations

In virtue of positive homogeneity and subadditivity, a risk measure is a *sublinear* functional and hence, assuming closedness[8] 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

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

In the case of $1 \leqslant p < \infty$, $\mathcal{L}^q$ 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

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

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 $\mathcal{L}^q_+ = \{Q \in \mathcal{L}^q : Q \geqslant 0 \ P\text{-a.s.}\}$

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. $R(X) \geqslant \mathbb{E}(X) \ \forall X \in \mathcal{L}^p$ 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

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

Due to $Q \in \mathcal{L}^q_+$, $\mu_Q(A) \geqslant 0$ 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

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

If X is bounded, then from this representation it is clear that $R(X) \leqslant \sup(X)$ 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

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

$$\begin{array}{ccc} \text{Risk } R & \longleftrightarrow & \text{Deviation } D \\ \uparrow & & \uparrow \\ \text{Regret } V & \longleftrightarrow & \text{Error } E \end{array}$$

Figure 1: The fundamental risk quadrangle (Rockafellar and Uryasev, 2013).

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).

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 Fig. 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*

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

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. $R(X + Y) \leqslant R(X) + R(Y)$ (subadditivity)

A3. $R(X + c) = R(X) + c \quad \forall c \in \mathbb{R}$ (translation equivariance)

A4. $X \leqslant Y$ P-a.s. $\Rightarrow R(X) \leqslant R(Y)$ (monotonicity)

A5. $R(c) = c$ $\forall c \in \mathbb{R}$, but $R(X) > \mathbb{E}[X]$ for nonconstant X

A6. R is closed, i.e. it has closed sublevel sets $\{X \in \mathrm{dom}(R) : R(X) \leqslant c\}$, $c \in \mathbb{R}$. Here, $\mathrm{dom}(R) := \{X : R(X) < \infty\}$.

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

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

$$D(X) \leqslant \text{ess sup}(X) - \mathbb{E}[X], \quad \text{ess sup}(X) := \inf\{\lambda \in \mathbb{R} : P(X > \lambda) = 0\}$$
$$D(X + c) = D(X) \ \forall c \in \mathbb{R} \quad \text{(translation invariance)}$$

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 subadditive. 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

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

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, subadditive, closed and averse in the sense that

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

Furthermore, we require

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

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

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

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 $\sigma_{\mathcal{Q}}(X) := \sup_{Q \in \mathcal{Q}} \langle X, Q \rangle$ be the support function of the set $\mathcal{Q} = \{Q \in \mathcal{L}^q : \langle X, Q \rangle \leqslant \sigma_{\mathcal{Q}}(X) \ \forall X \in \mathcal{L}^p\}$.

Theorem 3. *(Sun et al., 2020). Let $V = \sigma_{\mathcal{Q}}$ be a positively homogeneous, subadditive, monotonic and closed functional, i.e. a coherent regret measure, and $\mathcal{Q}$ be its supported set. Suppose $\mathcal{Q}' := \mathcal{Q} \cap \mathcal{E}_1 \neq \varnothing$. 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 2.1):

Theorem 4. *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

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

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

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

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:

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

for random variables $X_1,..,X_n$, 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

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

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 4 are also linked to coherent regret (and thus error) measures.

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 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:

$$R(X) = \mathrm{CVar}_\alpha(X), \qquad D(X) = \mathrm{CVar}_\alpha(X - \mathbb{E}(X))$$
$$V(X) := \frac{1}{1-\alpha}\mathbb{E}[X^+], \quad E(X) = \mathbb{E}\left[\frac{1}{1-\alpha}X^+ + X^-\right].$$

According to the projection from regret, $\mathrm{CVar}_\alpha(X) = \min_c\{\frac{1}{1-\alpha}\mathbb{E}((X-c)^+) + c\}$. We also define $\mathrm{CVar}_{\alpha=0} := \mathbb{E}$ in the same way, but this is only "weakly" averse in the degenerate sense that $\mathbb{E} \geqslant \mathbb{E}$. The random variable X has a right-continuous distribution function F_X with generalized inverse[9] $F_X^{-1}(q) = \sup\{\lambda \geqslant 0 : F_X(\lambda) < q\}$. Then CVar_α can be equivalently expressed as an integral over quantiles

$$\mathrm{CVar}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 F_X^{-1}(q)\ \mathrm{d}q.$$

If F_X is continuous, this can be further written as

$$\mathrm{CVar}_\alpha(X) = \mathbb{E}\left[X | X \geqslant F_X^{-1}(\alpha)\right],$$

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, $\mathrm{CVar}_{\alpha=0}$ corresponds to the expectation; in the other, $\mathrm{CVar}_{\alpha\to 1} := \lim_{\alpha\to 1}\mathrm{CVar}_\alpha$ gives the essential supremum (worst-case) of X.

The envelope of CVar_α is known to be $\mathcal{Q} = \{Q : 0 \leqslant Q \leqslant \frac{1}{1-\alpha}, \mathbb{E}[Q] = 1\}$, 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 $\alpha \to 1$, the reweighting may be arbitrarily large and hence the worst-case will receive all 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 $\{Q : 0 \leqslant Q \leqslant 1/(1-\alpha)\}$. The projection of V to CVar_α, i.e. the natural extension, entails intersecting this set with the constraint $\mathbb{E}[Q] = 1$.

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 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

$$R^\lambda(X) := \int_0^1 \mathrm{CVar}_\alpha(X)\ \mathrm{d}\lambda(\alpha), \tag{4}$$

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

$$R^\lambda(X) = R_{(w)}(X) := \int_0^1 F_X^{-1}(q)w(q)\ \mathrm{d}q, \tag{5}$$

9. For consistency with Section 4, we choose to work with the lower instead of the upper quantile.

with a spectral weighting function $w : [0,1] \to \mathbb{R}^+$. This generates a coherent risk measure if and only if w is nonnegative, monotonically increasing and $\int_0^1 w(q)\ \mathrm{d}q = 1$ (Acerbi, 2002). These properties are automatically satisfied when w is induced by a probability measure λ from (4). 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

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

hence all values below the α-th quantile are ignored, and values above it receive the constant weight $1/(1-\alpha) \geqslant 1$. Note that by demanding that λ does not have an atom at 1, we have excluded the supremum risk measure $\mathrm{CVar}_{\alpha\to 1}$, which is represented by the Dirac measure at 1. The corresponding weight function would be 0 everywhere, rendering a representation of the form (5) 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 4.

For an arbitrary spectral risk measure $R_{(w)}$ with spectrum w, the envelope representation is (Pflug, 2006):

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

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 4, which also supplies intuition.

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:

$$\begin{aligned} \mathbb{E}[X] &= -\int_{-\infty}^{0} F_X(x)\ \mathrm{d}x + \int_0^\infty (1 - F_X(x))\ \mathrm{d}x \\ &= \int_{-\infty}^{0} [S_X(x) - 1]\ \mathrm{d}x + \int_0^\infty S_X(x)\ \mathrm{d}x, \end{aligned}$$

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*:

$$R_\phi(X) := \int_{-\infty}^{0} [\phi(S_X(x)) - 1] \, \mathrm{d}x + \int_{0}^{\infty} \phi\left(S_X(x)\right) \, \mathrm{d}x, \tag{6}$$

where $\phi : [0,1] \to [0,1]$ 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 $R_\phi(X) \geqslant \mathbb{E}(X)$. 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 superadditive[10]. In Appendix B.3, we consider functionals of the form (6), Choquet integrals, in more depth. Here we observe that distortion risk measures are equivalent to spectral risk measures.

Theorem 5. *(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 A.1. For example, in the case of CVar_α we have $\phi(t) = \int_0^t w(1-u) \, \mathrm{d}u = \min\{t/(1-\alpha), 1\}$.

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 $\overline{\mu}(A) := \phi(P(A)) \; \forall A \in \mathcal{F}$ defines a *capacity* on events. A capacity on $(\Omega, \mathcal{F})$ is a set function $\overline{\mu} : \mathcal{F} \to \mathbb{R}$ with the normalization $\overline{\mu}(\varnothing) = 0$, $\overline{\mu}(\Omega) = 1$ and the monotonicity property $A \subseteq B \Rightarrow \overline{\mu}(A) \leqslant \overline{\mu}(B)$. 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

$$\overline{\mu}(A \cup B) + \overline{\mu}(A \cap B) \leqslant \overline{\mu}(A) + \overline{\mu}(B) \quad \forall A, B \in \mathcal{F}.$$

If and only if the function ϕ is concave and $\phi(0) = 0$, $\overline{\mu}$ 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 B.3 it has a rich interpretation in terms of systematic risk aversion. It is known that a monotone

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

submodular capacity is always coherent[11], due to its envelope being non-empty:

$$\begin{aligned}\operatorname{core}(\overline{\mu}) &= \{P : P(A) \leqslant \overline{\mu}(A)\ \forall A \in \mathcal{F},\ P \text{ probability measure}\}\\ \overline{\mu}(A) &= \sup_{P \in \operatorname{core}(\overline{\mu})} P(A)\end{aligned}$$

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 and therefore the core consists of countably additive measures. Employing the core, the Choquet integral (6) for a submodular capacity is equivalently given by

$$R_\phi(X) = \sup_{P \in \operatorname{core}(\overline{\mu})} \left\{ \int_{-\infty}^{\infty} X \ \mathrm{d}P \right\}. \tag{7}$$

Thus, for a submodular capacity, Walley's natural extension coincides with the Choquet integral. This can be understood by viewing (7) 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 $\overline{\mu}$, we can define a corresponding dual capacity $\underline{\mu}(A) := 1 - \overline{\mu}(A^C)\ \forall A \in \mathcal{F}$. This capacity is supermodular:

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

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

$$\begin{aligned}\overline{F}_X(x) &:= \overline{P}(X \leqslant x) = 1 - \underline{P}(X > x)\\ \underline{F}_X(x) &:= \underline{P}(X \leqslant x) = 1 - \overline{P}(X > x).\end{aligned}$$

Hence we can compute upper and lower densities as $\overline{f}_X = \overline{F}'_X$ 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 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*.

11. The 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.

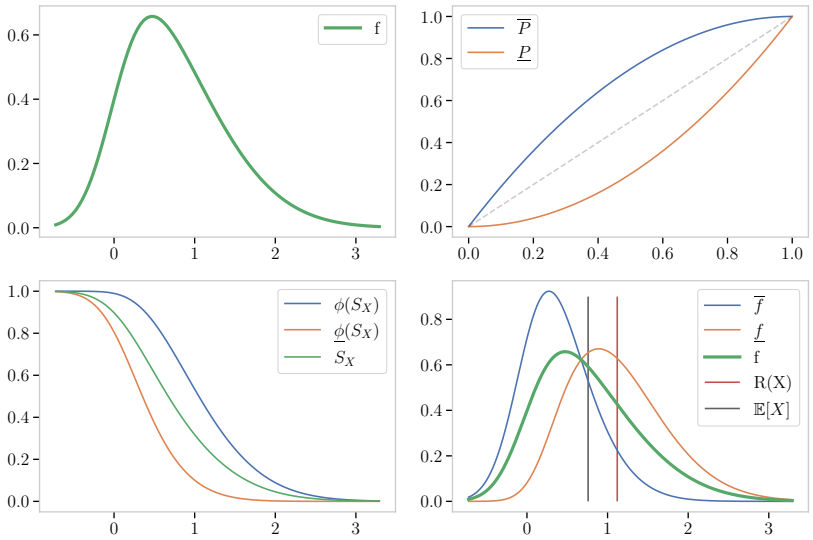

Figure 2: Top left: the density of an exemplary skew-normal distribution, belonging to some random variable X. Top right: lower and upper probabilities with distortion function $\phi(t) = 1 - (1-t)^2$. Bottom left: lower and upper distortion of the survival function, corresponding to the exemplary distribution. 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 $\mathbb{E}[X]$.

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, 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 2) is particularly interesting because the axioms are motivated 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 B, 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.

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

R1. $R(X) = 0 \Leftrightarrow X = 0$; $\quad R(\lambda X) = \lambda R(X)\ \forall \lambda \geqslant 0$; $\quad R(X + Y) \leqslant R(X) + R(Y)$

R2. $0 \leqslant X \leqslant Y \Rightarrow R(X) \leqslant R(Y)$

R3. $0 \leqslant X_n \uparrow X$ *μ-a.e.* $\Rightarrow R(X_n) \uparrow R(X)$

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

$$\|X\|_{\mathcal{R}} := R(|X|). \tag{8}$$

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 (8). 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 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* $\mu_X : \mathbb{R}^+ \to [0, 1]$ of $X \in \mathcal{M}$ as

$$\mu_X(\lambda) := \mu\left\{\omega \in \Omega : |X(\omega)| > \lambda\right\}. \tag{9}$$

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. $\mu_X(\lambda) = \mu_Y(\lambda)\ \forall \lambda \geqslant 0$.

Definition 7. *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 $X^* : [0,1] \to \mathbb{R}^+$ as the generalized inverse of its distribution function:

$$\begin{aligned} X^*(\omega) &:= \inf\{\lambda \geqslant 0 : \mu_X(\lambda) \leqslant \omega\}, \quad \omega \in [0,1) \\ X^*(1) &:= \lim_{\omega \uparrow 1} X^*(\omega). \end{aligned} \tag{10}$$

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 $|X|$:

$$\begin{aligned} X^*(\omega) &= \inf\{\lambda \geqslant 0 : \mu_X(\lambda) \leqslant \omega\} \\ &= \sup\{\lambda \geqslant 0 : \mu_X(\lambda) > \omega\} \\ &= \sup\{\lambda \geqslant 0 : 1 - \mu_X(\lambda) < 1 - \omega\} \\ &= \sup\{\lambda \geqslant 0 : F_{|X|}(\lambda) < 1 - \omega\} \\ &= F_{|X|}^{-1}(1 - \omega). \end{aligned}$$

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.

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 $u : \mathcal{R} \to \mathbb{R}$, and is equipped with the norm (Rubshtein et al., 2016, p. 83)

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

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

$$\begin{aligned} R'(X) &:= \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : R(Y) \leqslant 1, Y \in \mathcal{M}^+\right\} \\ \|X\|_{\mathcal{R}'} &:= \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : \|Y\|_{\mathcal{R}} \leqslant 1, Y \in \mathcal{R}\right\}. \end{aligned}$$

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

$$\int_\Omega |XY| \, \mathrm{d}\mu \leqslant \|X\|_{\mathcal{R}} \|Y\|_{\mathcal{R}'}.$$

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

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

where $\text{ess sup}(X) := \inf\{\lambda \geqslant 0 : \mu_X(\lambda) = 0\}$. 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).

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)

$$\|X\|_{\mathcal{R}} \leqslant c\|X\|_{\mathcal{S}} \quad \forall X \in \mathcal{S}.$$

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 $\mu(\Omega) = 1$):

$$\mathcal{L}^\infty \hookrightarrow \mathcal{R} \hookrightarrow \mathcal{L}^1,$$

and 1 is a feasible embedding constant:

$$\|X\|_{\mathcal{L}^1} \leqslant \|X\|_{\mathcal{R}} \quad \forall X \in \mathcal{R}, \quad \|X\|_{\mathcal{R}} \leqslant \|X\|_{\mathcal{L}^\infty} \quad \forall X \in \mathcal{L}^\infty.$$

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 3.2): the envelope of the expectation is the singleton $\{1\}$ (a singleton envelope c yields a constant multiple of the expectation[12]), 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 $\text{ess sup} = \text{CVar}_{\alpha\to 1}$, 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*.

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

4.3 Hardy-Littlewood's Inequality

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

$$\sum_{i=1}^{n} x_i y_i \leqslant \sum_{i=1}^{n} x_i^* y_i^*,$$

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):

$$\int_\Omega |XY| \, \mathrm{d}\mu \leqslant \int_0^1 X^*(\omega) Y^*(\omega) \, \mathrm{d}\omega. \tag{11}$$

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.

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

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

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

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

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:

$$\begin{aligned} \forall t \in (0,1] : X^{**}(t) &= \frac{1}{t} \int_0^t X^*(\omega) \, \mathrm{d}\omega \\ &= \frac{1}{t} \int_0^t F_{|X|}^{-1}(1-\omega) \, \mathrm{d}\omega \\ &= \frac{1}{1-t} \int_{1-t}^1 F_{|X|}^{-1}(\omega) \, \mathrm{d}\omega \\ &= \mathrm{CVar}_{1-t}(|X|). \end{aligned}$$

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)

$$(\forall \alpha \in [0,1) : \mathrm{CVar}_\alpha(|X|) \leqslant \mathrm{CVar}_\alpha(|Y|)) \implies \|X\|_{\mathcal{R}} \leqslant \|Y\|_{\mathcal{R}}$$

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 17). To this end, we need to introduce the *fundamental function* of an ri space.

4.4 The Fundamental Function

Definition 8. *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 $\phi_{\mathcal{R}} : [0,1] \to \mathbb{R}^+$ as*

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

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 then[13] $\phi(1) = 1$. 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):

$$\begin{aligned}
&\phi \text{ is non-decreasing and } \phi(0) = 0\\
&t \mapsto \phi(t)/t \text{ is decreasing}\\
&\phi \text{ is continuous except perhaps at the origin.}
\end{aligned}$$

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 $\phi : [0,1] \to [0,1]$ 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 4.8, 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*

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

(Van Heerwaarden and Kaas, 1992) (Du):

$$\begin{aligned} \mathrm{MaxV}(X) &:= \mathbb{E}\left[\max(X_1, X_2)\right], \; X_1, X_2 \overset{\text{ind}}{\sim} X \quad \forall X \in \mathcal{M}^+ \\ \mathrm{Du}(X) &:= \mathbb{E}\left[\max(X, \mathbb{E}[X])\right] \quad \forall X \in \mathcal{M}^+, \end{aligned} \tag{12}$$

where $X_1, X_2 \overset{\text{ind}}{\sim} X$ 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 $X \in \mathcal{M}^+$:

$$\begin{aligned} \mathrm{MaxV}(X) &= \int_0^\infty 1 - (1 - S_X(\omega))^2 \, \mathrm{d}\omega = \int_0^\infty 1 - (1 - (1 - F_X(\omega)))^2 \, \mathrm{d}\omega \\ &= \int_0^\infty 1 - F_X^2(\omega) \, \mathrm{d}\omega. \end{aligned}$$

This is just the expectation of a random variable with distribution function $F_Y = F_X^2$, i.e. $Y = \max(X_1, X_2)$, $X_1, X_2 \overset{\text{ind}}{\sim} X$. The Dutch risk measure, on the other hand, is also law invariant but not spectral. It is also easy to see that $\mathbb{E}[X] \leqslant \mathrm{Du}(X) \leqslant \mathrm{MaxV}(X)$ $\forall X \in \mathcal{M}^+$ by applying Jensen's inequality in (12). 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 20 that the Dutch risk measure is the most optimistic extension for its specific ϕ.

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

$$\begin{aligned} \|X\|_{\Lambda_\phi} &:= \int_0^1 X^*(\omega) \, \mathrm{d}\phi(\omega) \\ &= X^*(0)\phi(0+) + \int_0^1 X^*(\omega)\phi'(\omega) \, \mathrm{d}\omega \\ &= X^*(0)\phi(0+) + \int_0^1 F_{|X|}^{-1}(1-\omega)\phi'(\omega) \, \mathrm{d}\omega, \end{aligned}$$

where we immediately recognize the correspondence to the distortion (spectral) risk measure (19) 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 $\phi \in \Phi \backslash \Phi_{0+}$.

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

$$\begin{aligned}\|X\|_{M_\phi} &= \sup_{0<t\leqslant 1} \{\phi(t) X^{**}(t)\} \\ &= \sup_{0<t\leqslant 1} \left\{\phi(t) \frac{1}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega\right\} \\ &= \sup_{0<t\leqslant 1} \{\phi(t)\, \mathrm{CVar}_{1-t}(|X|)\}\end{aligned}$$

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 9. *(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*

$$\Lambda_\phi \hookrightarrow \mathcal{R} \hookrightarrow M_\phi$$

and 1 *is a feasible embedding constant:*

$$\|X\|_{M_\phi} \leqslant \|X\|_R \quad \forall X \in \mathcal{R}, \quad \|X\|_R \leqslant \|X\|_{\Lambda_\phi} \quad \forall X \in \Lambda_\phi.$$

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 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 $\mathrm{Du}(X) \leqslant \mathrm{MaxV}(X)\ \forall X \in \mathcal{M}^+$, 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 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 4.6,4.8) 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 space[14]. Then we can write the Lorentz norm as

$$\|X\|_{\Lambda_\phi} = \sup\left\{\int_0^1 X^*(\omega) Y^*(\omega)\ \mathrm{d}\omega : \|Y\|_{M_{\phi^*}} \leqslant 1, Y \in \mathcal{M}^+\right\},$$

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.

14. This 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.

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,

$$\begin{aligned}\|X\|_{M_\phi} &= \sup_{0<t\leqslant 1} \{\phi(t)X^{**}(t)\} = \sup_{0<t\leqslant 1} \{t \cdot X^{**}(t)\} \\ &= \sup_{0<t\leqslant 1} \left\{\int_0^t X^*(\omega)\ \mathrm{d}\omega\right\} = \int_0^1 X^*(\omega)\ \mathrm{d}\omega = \|X\|_{\mathcal{L}^1}.\end{aligned}$$

Similarly, one easily checks that the coincidence also holds for $\mathcal{L}^\infty$. Hence $\mathcal{L}^1$ and $\mathcal{L}^\infty$ are distuingished spaces as they allow only a single ri norm. We prove a novel result in Section 4.8 (Theorem 21): 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 $\alpha \to 1$, 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).

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 equivariance:

$$R(X) = R(X + c) - c,$$

for some constant c so that $\operatorname{ess\,inf}(X + c) \geqslant 0$. 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 10. *An ri function norm R is called* PTE *if*

$$\forall X \in \mathcal{M}^+, c \in \mathbb{R} \text{ s.t. } X + c \geqslant 0 : R(X + c) = R(X) + c.$$

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:

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

Theorem 11. *An ri function norm R can be represented in the following reduced form if and only if it is positive translation equivariant* (PTE)*:*

$$R(X) = \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+\right\}. \tag{13}$$

We call this the positive translation equivariant representation of R. The proof is in Appendix A.2.

Remark 12. 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 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 1. 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 $\{Y : R^\infty(Y) \leqslant 1\}$ 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 2. 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 $\alpha \to 1$, $\phi(t) = \chi_{(0,1]}$ (Theorem 21).

As an example of the above, consider the function norm R^1, which can be written as

$$R^1(X) = \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : R^\infty(Y) \leqslant 1, Y \in \mathcal{M}^+\right\} \quad \forall X \in \mathcal{M}^+.$$

For nonnegative functions $X \in \mathcal{M}^+$, this is the expectation $\mathbb{E}[X]$. However, if we we want to extend the above definition to work on general $X \in \mathcal{M}$, we must use its positive translation equivariant representation

$$R^1(X) = \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] = R^\infty(Y) = 1, Y \in \mathcal{M}^+\right\} \quad \forall X \in \mathcal{M}^+.$$

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 $\mu_X^- : \mathbb{R} \to [0,1]$ and the generalized decreasing rearrangement $X^{*-} : [0,1] \to \mathbb{R}$ as (cf. (9), (10))

$$\mu_X^-(\lambda) := \mu\{\omega \in \Omega : X(\omega) > \lambda\}$$
$$X^{*-}(\omega) := \inf\{\lambda \in \mathbb{R} : \mu_X^-(\lambda) \leqslant \omega\}.$$

Clearly, $X^{*-}(\omega) = F_X^{-1}(1-\omega)$.

Theorem 13. *Let R be an ri function norm which has a positive translation equivariant representation. Then the extended functional*

$$R^-(X) := \sup\left\{\int_0^1 X^{*-}(\omega)Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+\right\} \quad \forall X \in \mathcal{M} \tag{14}$$

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 20 for the construction of a PTE norm related to the Marcinkiewicz norm.

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 14. *Every ri function norm which is* PTE *admits a representation on the form*

$$R(X) = \sup_{\lambda \in \mathfrak{M}}\left\{\int_0^1 \mathrm{CVar}_\alpha(X)\ \mathrm{d}\lambda(\alpha)\right\} \quad \forall X \in \mathcal{M}^+ \tag{15}$$

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:

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

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

15. A 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.

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 11), that is $\int_0^1 Y^*(\omega)\ \mathrm{d}\omega = 1$. Then we can write this as ($Y \in \mathcal{M}^+$)

$$R(X) = R_{\mathcal{Y}_1}(X) = \sup\left\{\int_0^1 F_X^{-1}(1-\omega)Y^*(\omega)\ \mathrm{d}\omega : R'(Y) = \mathbb{E}[Y] = 1\right\} \quad \forall X \in \mathcal{M}^+. \tag{16}$$

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 3.5). Thus we can also write this as a supremum over distortion risk measures with concave distortions $\mathcal{Z} := \{t \mapsto \int_0^t Y^*(\omega)\ \mathrm{d}\omega : Y \in \mathcal{Y}_1\}$. For each $Z \in \mathcal{Z}$, $Z : [0,1] \to [0,1]$, we have $Z(0) = 0$, $Z(1) = 1$ (due to PTE), since $Z(1) = \int_0^1 Y^*(\omega)\ \mathrm{d}\omega$. Therefore we obtain a representation as the supremum over Choquet integrals:

$$R_{\mathcal{Y}_1}(X) = R_{\mathcal{Z}}(X) = \sup\left\{\int_0^\infty Z(\mu_X(\omega))\ \mathrm{d}\omega : Z \in \mathcal{Z}\right\} \quad \forall X \in \mathcal{M}^+.$$

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 $Z(1) = 1 \Leftrightarrow \int_0^1 Y^*(\omega)\ \mathrm{d}\omega = 1$. 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)

$$\lambda_w(E) := w(0)\delta_0(E) + \int_E 1-\alpha\ \mathrm{d}w(\alpha) \quad E \text{ measurable},$$

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 (15) is recovered. We remark that Kusuoka representations need not be unique in general. Conversely, we can specify a functional $R_{\mathcal{Y}}$ directly as

$$R_{\mathcal{Y}}(X) = \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : Y \in \mathcal{Y}\right\} \quad \forall X \in \mathcal{M}^+.$$

For any set of nonnegative decreasing functions $\mathcal{Y}$, we are guaranteed to obtain a valid ri norm[16], since the supremum over a set of Lorentz norms preserves the relevant properties (Lemma 52). 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 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 $\int_0^1 Y^*(\omega)\ \mathrm{d}\omega = 1$ holds, then (16) 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

16. The normalization $R_{\mathcal{Y}}(1) = 1$ may not hold in general when no constraints on $\mathcal{Y}$ are imposed.

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 equivalent 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

$$R(X) = \sup\left\{\int_0^1 X(\omega)Y(\omega)\ \mathrm{d}\omega : R(Y) \leqslant 1\right\} \quad \forall X \in \mathcal{M}^+.$$

Under the ri property, X and Y can be replaced by any distributionally equivalent choice. The Hardy-Littlewood inequality (11) 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)):

$$\max\{F(x) + G(y) - 1, 0\} \leqslant H(x,y) \leqslant \min\{F(x), G(y)\}, \quad \forall x, y \in \mathbb{R}.$$

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) $H(x,y) = \mu(X \leqslant x, Y \leqslant y) = \min(F(x), G(y))$

C2) $(X(\omega) - X(\omega'))(Y(\omega) - Y(\omega')) \geqslant 0 \quad \forall \omega, \omega' \in \Omega$

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 correlation[17]. It is known that

$$\mathbb{E}[XY] \leqslant \mathbb{E}[\tilde{X}\tilde{Y}],$$

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 obviously holds). Therefore we recover Hardy-Littlewoods inequality (11).

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:

$$R(X + Y) = R(X) + R(Y).$$

17. Antimonotonicity means that $(X(\omega) - X(\omega'))(Y(\omega) - Y(\omega')) \leqslant 0 \quad \forall \omega, \omega' \in \Omega$.

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 15. *(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.

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

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

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

$$\phi(t) = R(\chi_E) = \sup\left\{\int_0^t Y(\omega)\ \mathrm{d}\omega : Y \in \mathcal{Y}\right\} = \sup_{Z\in\mathcal{Z}} Z(t),$$

where $\mathcal{Z} = \{t \mapsto \int_0^t Y(\omega)\ \mathrm{d}\omega : Y \in \mathcal{Y}\}$. Depending on the context, we may call either $\mathcal{Y}$ or $\mathcal{Z}$ a Kusuoka set, as either fully describes the representation[18]. 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 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 16. *Hardy's lemma (Bennett and Sharpley, 1988). Let Y_1 and Y_2 be nonnegative measurable functions on* $(0,1]$ *and*

$$\int_0^t Y_1(\omega)\ \mathrm{d}\omega \leqslant \int_0^t Y_2(\omega)\ \mathrm{d}\omega \quad \forall t \in (0,1].$$

If η is any nonnegative decreasing function on $(0,1]$*, then*

$$\int_0^1 \eta(\omega)Y_1(\omega)\ \mathrm{d}\omega \leqslant \int_0^1 \eta(\omega)Y_2(\omega)\ \mathrm{d}\omega.$$

18. We 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.

As an example, in our context this implies that if a concave distortion $Z_1(t) = \int_0^t Y_1(\omega)\ \mathrm{d}\omega$ 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 $Z_1(t) = \int_0^t Y_1(\omega)\ \mathrm{d}\omega$.

Theorem 17. *(Bennett and Sharpley, 1988, p. 72). Given any ri norm R with fundamental function $\phi \in \Phi_{0+}$, we have $R(X) \leqslant \|X\|_{\Lambda_\phi}\ \forall X \in \mathcal{M}^+$.*

Proof Let $R(X) = \sup_{Y\in\mathcal{Y}}\{\int_0^1 X^*(\omega)Y(\omega)\ \mathrm{d}\omega\}$. We know that $\phi(t) = \sup_{Y\in\mathcal{Y}}\{\int_0^t Y(\omega)\ \mathrm{d}\omega\}$. Recall that $\|X\|_{\Lambda_\phi} = \int_0^1 X^*(\omega)\phi'(\omega)\ \mathrm{d}\omega$ if $\phi \in \Phi_{0+}$. But since ϕ majorizes all elements of $\{t \mapsto \int_0^t Y(\omega)\ \mathrm{d}\omega : Y \in \mathcal{Y}\}$ pointwise and $\phi \in \Phi_{0+}$, we immediately obtain from the qualification in Hardy's lemma

$$\int_0^t \phi'(\omega)\ \mathrm{d}\omega \geqslant \int_0^t Y(\omega)\ \mathrm{d}\omega \quad \forall 0 < t \leqslant 1 \quad \forall Y \in \mathcal{Y}$$

that it holds

$$\begin{aligned} R(X) &= \sup\left\{\int_0^1 X^*(\omega)Y(\omega)\ \mathrm{d}\omega : Y \in \mathcal{Y}\right\} \\ &\leqslant \sup\left\{\int_0^1 X^*(\omega)\phi'(\omega)\ \mathrm{d}\omega : Y \in \mathcal{Y}\right\} = \|X\|_{\Lambda_\phi}. \end{aligned}$$

The statement also holds true if $\phi \in \Phi\backslash\Phi_{0+}$ (Rubshtein et al., 2016). ■

Theorem 18. *(Bennett and Sharpley, 1988, p. 70). Given any ri norm R with fundamental function $\phi \in \Phi$, we have $\|X\|_{M_\phi} \leqslant R(X)\ \forall X \in \mathcal{R}$.*

Proof Write

$$R(X) = \sup_{Z_\gamma\in\mathcal{Z}}\left\{\int_0^1 X^*(\omega)Z'_\gamma(\omega)\ \mathrm{d}\omega\right\}, \quad \phi(t) = \sup_{Z_\gamma\in\mathcal{Z}} Z_\gamma(t)$$

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 functions[19]

$$\forall t \in (0,1] : Z_{\gamma,t}(x) := \begin{cases} Z_\gamma(t)\frac{x}{t} & ,\ x \leqslant t \\ Z_\gamma(t) & ,\ x > t. \end{cases}$$

The $Z_{\gamma,t}$ are constructed as the integrals of the functions[20]

$$\forall t \in (0,1] : Z'_{\gamma,t} := \begin{cases} \frac{Z_\gamma(t)}{t} & ,\ x \leqslant t \\ 0 & ,\ x > t. \end{cases}$$

19. Let $Z_\gamma : [0,1] \to \mathbb{R}^+$ be concave, i.e. $\forall\alpha \in [0,1] : Z_\gamma(\alpha t + (1-\alpha)x) \geqslant \alpha Z_\gamma(t) + (1-\alpha)Z_\gamma(x)$. Choosing $x = 0$ yields $Z_\gamma(\alpha t) \geqslant \alpha Z_\gamma(t)$. For any $x \leqslant t$ hence $x = \alpha t$ for some $\alpha \in [0,1]$, we obtain $Z_\gamma(x) \geqslant \frac{x}{t} Z_\gamma t$. For $x > t$ the statement is obvious, as the concave Z_i has nonnegative derivative, whereas $Z_{\gamma,t}$ has zero derivative.
20. The "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.

Applying Hardy's lemma yields

$$\begin{aligned}
R(X) &= \sup_{Z_\gamma \in \mathcal{Z}} \left\{ \int_0^1 X^*(\omega) Z'_\gamma(\omega)\ \mathrm{d}\omega \right\} \\
&\geqslant \sup_{Z_{\gamma,t}} \left\{ \int_0^1 X^*(\omega) Z'_{\gamma,t}(\omega)\ \mathrm{d}\omega \right\} \\
&= \sup_{Z_{\gamma,t}} \left\{ \frac{Z_\gamma(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega \right\} \\
&\geqslant \sup_{0<t\leqslant 1} \left\{ \sup \left\{ \frac{Z_\gamma(t)}{t} : Z_\gamma \in \mathcal{Z} \right\} \int_0^t X^*(\omega)\ \mathrm{d}\omega \right\} \\
&= \sup_{0<t\leqslant 1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega \right\} = \|X\|_{M_\phi},
\end{aligned}$$

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 19. Throughout the paper, we use shorthand notation to avoid explicitly writing 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 20. *Given any concave fundamental function $\phi \in \Phi$, we can construct the smallest positive translation equivariant ri norm as:*

$$\|X\|_{TM_\phi} = \sup_{0<t<1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega + \frac{\phi(t)-1}{t-1} \int_t^1 X^*(\omega)\ \mathrm{d}\omega \right\}.$$

For any other PTE *ri norm R with fundamental function ϕ we have $\|X\|_{M_\phi} \leqslant \|X\|_{TM_\phi} \leqslant R(X)\ \forall X \in \mathcal{R}$. We call TM_ϕ the positive translation equivariant Marcinkiewicz norm.*

Example 3. Recall that both the Dutch risk measure and the spectral MaxVar share the fundamental function $\phi(t) = 2t - t^2$. Then: $\|X\|_{TM_\phi} = \mathrm{Du}(|X|)$. 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 A.3.1. Next, we show when equality of all ri norms for a given fundamental function holds.

Theorem 21. *Given any concave fundamental function $\phi \in \Phi$, it holds that*

$$\|X\|_{M_\phi} = R(X) = \mathrm{CVar}_\alpha(X) = \|X\|_{\Lambda_\phi} \quad \forall X \in \mathcal{M}^+$$

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 $\alpha \to 1$, $\phi(t) = \phi_\infty(t) := \chi_{(0,1]}(t)$. For $\alpha \to 1$, $R(X) = \|X\|_{\mathcal{L}^\infty}$.

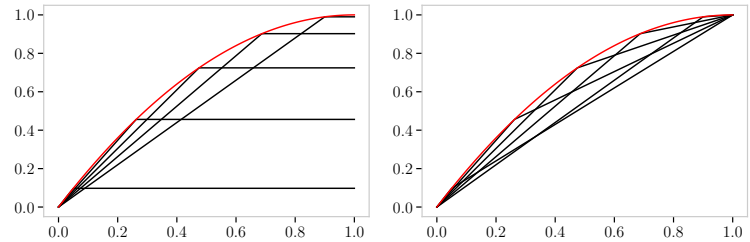

Figure 3: The red curve is the fundamental function $\phi(t) = 1-(1-t)^2$. Left: the black lines correspond to five selected ϕ_t in the Marcinkiewicz norm construction. Right: the black lines correspond to five selected ϕ_t in the positive translation equivariant Marcinkiewicz norm construction. In this 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 proof is in Appendix A.3.2. This result implies that for CVar_α-type fundamental functions (including $\mathcal{L}^1$ and $\mathcal{L}^\infty$ as special cases[21]), 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 22. *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*

$$\|X\|_{\Lambda_\phi} = \|X\|_{TM_\phi} = \beta\mathbb{E}[X] + (1-\beta)\,\mathrm{CVar}_\alpha(X) =: \mathrm{RIM}_{\alpha,\beta}(X) \quad \forall X \in \mathcal{M}^+.$$

Proof The Lorentz norm is easily computed

$$\begin{aligned}\|X\|_{\Lambda_\phi} &= \int_0^{1-\alpha} X^*(\omega)\left(\beta + (1-\beta)\frac{1}{1-\alpha}\right)\,\mathrm{d}\omega + \int_{1-\alpha}^1 X^*(\omega)\beta\,\mathrm{d}\omega \\ &= \beta\mathbb{E}[|X|] + (1-\beta)\,\mathrm{CVar}_\alpha(|X|).\end{aligned}$$

A Kusuoka set of concave functions for the PTE Marcinkiewicz norm is

$$\forall t \in (0,1) : \phi_{TM,t}(x) := \begin{cases} \phi(t)\frac{x}{t} & , \ x \leqslant t \\ \frac{1-\phi(t)}{1-t}x + \frac{\phi(t)-t}{1-t} & , \ x > t \end{cases}$$

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 $\|X\|_{TM_\phi} =$

21. More 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 $\alpha \to 1$ yields the $\mathcal{L}^\infty$ space. See Section 4.9.

$\int_0^1 X^*(\omega)\phi'_{TM,1-\alpha}(\omega)\ \mathrm{d}\omega = \|X\|_{\Lambda_\phi}$. ■

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

$$\mathrm{Du}(X) = \sup_{0<\beta<1} \{\beta\mathbb{E}[X] + (1-\beta)\cdot \mathrm{CVar}_\beta(X)\} \quad \forall X \in \mathcal{M}^+$$

we observe that the Dutch risk measure can be seen as an ambiguity set over $\mathrm{RIM}_{\alpha,\beta}$s, 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 $n \geqslant 2$. As n increases, risk aversion increases. Then:

$$\|X\|_{TM_\phi} = \sup_{0<\beta<1} \beta^{n-1}\mathbb{E}[|X|] + (1-\beta^{n-1})\cdot \mathrm{CVar}_\beta(|X|) \quad \forall X \in \mathcal{M}.$$

The corresponding Lorentz norm is (Cherny and Madan, 2009)

$$\|X\|_{\Lambda_\phi} = \mathbb{E}\left[\max(X_1, .., X_n)\right],\ X_1, .., X_n \overset{\text{ind}}{\sim} |X| \quad \forall X \in \mathcal{M}.$$

We can generalize this further to find that the PTE Marcinkiewicz norm has a family of RIMs as its basic building blocks.

Theorem 23. *Let $\phi \in \Phi_{0+}$ a fundamental function. Then*

$$\begin{aligned}\|X\|_{TM_\phi} &= \sup_{0<t<1} \frac{1-\phi(1-t)}{t}\mathbb{E}[|X|] + \left(1 - \frac{1-\phi(1-t)}{t}\right)\cdot \mathrm{CVar}_t(|X|)\\ &= \sup_{0<t<1} \mathrm{RIM}_{\alpha(t),\beta(t)}(|X|), \quad \textit{where } \alpha(t) = t, \beta(t) = \frac{1-\phi(1-t)}{t}.\end{aligned}$$

Proof Set $\phi_{TM,1-\alpha}(x) = \phi_{\mathrm{RIM}_{\alpha,\beta}}(x) = \beta x + (1-\beta)\min\{1, x/(1-\alpha)\}$. Both $\phi_{TM,1-\alpha}$ and $\phi_{\mathrm{RIM}_{\alpha,\beta}}$ 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 $\phi_{\mathrm{RIM}_{\alpha,\beta}}$, but these are just the $\mathrm{RIM}_{\alpha,\beta}$. ■

We remark that $\mathrm{RIM}_{\alpha,\beta}$ admits the following variational representation:

$$\mathrm{RIM}_{\alpha,\beta}(X) = \inf_{\mu\in\mathbb{R}} \mu + \mathbb{E}v(X-\mu) \quad \forall X \in \mathcal{M},$$

where the regret function v is given by the piecewise linear function

$$v(t) = \begin{cases} \beta t & t \leqslant 0 \\ \frac{\beta\alpha-1}{\alpha-1}t & t > 0. \end{cases}$$

For the proof, see Appendix A.3.3, 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 1).

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 $\exists c_1, c_2 > 0 : c_1\|X\|_{\mathcal{R}_1} \leqslant \|X\|_{\mathcal{R}_2} \leqslant c_2 \cdot \|X\|_{\mathcal{R}_1}\ \forall X \in \mathcal{R}_1 = \mathcal{R}_2$. 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 24. *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*

$$C := \sup_{\phi_{2\zeta}\in\mathcal{Z}_2} \inf_{\phi_{1\gamma}\in\mathcal{Z}_1} \sup_{0<\alpha\leqslant 1} \frac{\phi_{2\zeta}(\alpha)}{\phi_{1\gamma}(\alpha)} \tag{17}$$

is finite, we have the relationship

$$\|X\|_{\mathcal{R}_2} \leqslant C \cdot \|X\|_{\mathcal{R}_1} \quad \forall X \in \mathcal{R}_1$$

and $\mathcal{R}_1 \subseteq \mathcal{R}_2$, therefore $\mathcal{R}_1 \hookrightarrow \mathcal{R}_2$. If furthermore $\exists c > 0 : c \cdot \|X\|_{\mathcal{R}_1} \leqslant \|X\|_{\mathcal{R}_2} \forall Y \in \mathcal{R}_2$, 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 A.4.1, 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 $\mathcal{R}_1 \nsubseteq \mathcal{R}_2$). At first sight one might conjecture that $C = \infty$ implies nonequivalence, but we provide a counterexample in Theorem 31. However, we can state the following slightly refined result.

Theorem 25. *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 24, 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*

$$C' := \sup_{\alpha\to 0} \sup_{\phi_{2\zeta}\in\mathcal{Z}_2} \inf_{\phi_{1\gamma}\in\mathcal{Z}_1} \frac{\phi_{2\zeta}(\alpha)}{\phi_{1\gamma}(\alpha)}$$

is infinite, then the norms are not equivalent; we have $\mathcal{R}_1 \nsubseteq \mathcal{R}_2$ and

$$\nexists c : \|X\|_{\mathcal{R}_2} \leqslant c \cdot \|X\|_{\mathcal{R}_1} \quad \forall X \in \mathcal{R}_1.$$

Proof Suppose $C' = \infty$. The norm of the identity embedding $\mathcal{R}_1 \hookrightarrow \mathcal{R}_2$ is

$$\| \operatorname{id} \| = \sup \left\{ \frac{\|X\|_{\mathcal{R}_2}}{\|X\|_{\mathcal{R}_1}} : X \in \mathcal{R}_1 \right\}.$$

We restrict the supremum to measurable indicator functions and obtain:

$$\| \operatorname{id} \| \geqslant \sup_{\chi_A} \frac{\|\chi_A\|_{\mathcal{R}_2}}{\|\chi_A\|_{\mathcal{R}_1}} = \sup_{\alpha \to 0} \sup_{\phi_{2\zeta} \in \mathcal{Z}_2} \frac{\phi_{2\zeta}(\alpha)}{\sup_{\phi_{1\gamma} \in \mathcal{Z}_1} \phi_{1\gamma}(\alpha)} = \sup_{\alpha \to 0} \sup_{\phi_{2\zeta} \in \mathcal{Z}_2} \inf_{\phi_{1\gamma} \in \mathcal{Z}_1} \frac{\phi_{2\zeta}(\alpha)}{\phi_{1\gamma}(\alpha)} = C' = \infty.$$

Since $\| \operatorname{id} \|$ 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 $\mathcal{L}^\infty$ (Rubshtein 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 ($\phi \in \Phi_{0+}$):

$$\|X\|_{\Lambda_\phi} = \int_0^1 X^*(\omega)\phi'(\omega) \, \mathrm{d}\omega,$$

and recall that these are the basic building blocks of any ri norm (Section 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 $\alpha \to 1$ may grow arbitrarily large. In general, the most benign situation occurs when $\phi'(0)$ is finite.

Theorem 26. *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*

$$\|X\|_{\Lambda_\phi} \leqslant K \cdot \|X\|_{M_\phi} \quad \forall X \in M_\phi.$$

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 A.4.2. 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

$$\|X\|_{M_\phi} \leqslant \mathrm{Du}(|X|) \leqslant \mathrm{MaxV}(|X|) \leqslant \frac{4}{3}\|X\|_{M_\phi} \quad \forall X \in M_\phi.$$

The smallest K, however is achieved for CVar_α: $K = 1/\phi(\frac{1}{\phi'(0)}) = 1 \ \forall \alpha \in [0, 1)$, a sanity check for Theorem 21.

Remark 27. Assume R_1 and R_2 are coherent risk measures. Then $R_1(X) \leqslant K \cdot R_2(X)$ $\forall X \in \mathcal{M}$ 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 28. *Let $\phi \in \Phi_{0+}$ with $\phi'(0) = \infty$. Then the Marcinkiewicz and the Lorentz norm are not equivalent. We have*

$$\|X\|_{M_\phi} \leqslant \|X\|_{\Lambda_\phi} \forall X \in \Lambda_\phi \quad \text{but } \nexists K : \|X\|_{\Lambda_\phi} \leqslant K \cdot \|X\|_{M_\phi} \forall X \in M_\phi.$$

The proof is in Appendix A.4.3.

Theorem 29. *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 A.4.4.

Corollary 30. *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

$$\mathcal{R}_1 \subseteq M_{\phi_1} \subsetneq \mathcal{L}^1 = \mathcal{R}_2.$$

If the spaces do not coincide, the norms cannot be equivalent (Bennett and Sharpley, 1988, p. 7).

■

Theorem 31. *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 A.4.5.

Corollary 32. *Let $\phi \in \Phi_{0+}$ with $\phi'(0) = \infty$. In view of the embedding theorem, Theorem 28 and Theorem 31, we have the embeddings:*

$$\Lambda_\phi \subsetneq TM_\phi = M_\phi$$

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.

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 (14) can be constructed:

$$V^-(X) := \sup\left\{\int_0^1 X^{*-}(\omega)Y^*(\omega)\,\mathrm{d}\omega : V'(Y) \leqslant 1, Y \in \mathcal{M}^+\right\} \quad \forall X \in \mathcal{M}$$

It is easy to check that this fulfills the desiderata of a coherent regret measure except perhaps aversity[22]. The corresponding coherent risk measure can then be obtained by infimal convolution (Theorem 3). 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.

22. However, a "weak aversity" condition $V^-(X) \geqslant \mathbb{E}[X]$ 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 3.2. Positive homogeneity and subadditivity are easily checked.

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.

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 33. *Suppose $\phi\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ and $\phi(1) = 1$. If ϕ is quasiconcave then*

$$\forall t \geqslant 0, \qquad 1 \wedge t \leqslant \phi(t) \leqslant 1 \vee t.$$

The proof is in Appendix A.5.1.

Lemma 34. *(Rubshtein et al., 2016, p. 127). The function $\phi\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ is quasiconcave if and only if $t \mapsto t/\phi(t)$ is quasiconcave.*

Lemma 34 has a natural interpretation in terms of fundamental functions:

Lemma 35. *(Kreĭn et al., 1982, 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 36. *Suppose $\phi_i\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ 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 A.5.2. Lemma 36 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 37. *Suppose $\psi\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ is quasiconcave and $\psi(1) = 1$. The* perspective *of ψ is the function*

$$\breve{\psi}\colon \mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \ni (x, y) \mapsto y\psi(x/y).$$

Let $\mathscr{P}$ denote the set of functions $\mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ which are positively homogeneous, non-zero (except at $(0, 0)$) and nondecreasing (in both arguments), and let $\mathscr{Q}$ denote the set of quasiconcave functions on $\mathbb{R}_{\geqslant 0}$.

Lemma 38. *Suppose $\psi\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$. The perspective $\breve{\psi} \in \mathscr{P}$ if and only if $\psi \in \mathscr{Q}$.*

The proof is in Appendix A.5.3. The following lemma shows that combining two quasiconcave functions using $\breve{\psi}$ is guaranteed to result in a quasiconcave function, and that this is the only way to ensure such a preservation of quasiconcavity.

Lemma 39. *Suppose $\phi_0, \phi_1, \psi\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$, and let $f_{\phi_0,\phi_1}(t) := \breve{\psi}(\phi_0(t), \phi_1(t))$, $t \geqslant 0$. Then $[f_{\phi_0,\phi_1} \in \mathscr{Q},\ \forall \phi_1, \phi_2 \in \mathscr{Q}]$ if and only if $\psi \in \mathscr{Q}$.*

The proof is in Appendix A.5.4. Observe that if $\psi(1) = 1$, and $\phi_1 = \phi_0$, then

$$\breve{\psi}(\phi_0(t), \phi_1(t)) = \phi_0(t)\psi(\phi_0(t)/\phi_0(t)) = \phi_0(t).$$

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

$$\begin{aligned}
\|f\|_{\Delta(\mathcal{X}_0,\mathcal{X}_1)} &:= \|f\|_{\mathcal{X}_0} \vee \|f\|_{\mathcal{X}_1} \\
\|f\|_{\Sigma(\mathcal{X}_0,\mathcal{X}_1)} &:= \inf\{\|f_0\|_{\mathcal{X}_0} + \|f_1\|_{\mathcal{X}_1} : f = f_0 + f_1,\ f_0 \in \mathcal{X}_0,\ f_1 \in \mathcal{X}_1\}.
\end{aligned}$$

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 $\overset{1}{\hookrightarrow}$, and which can always be ensured by simple scaling, for any intermediate space $\mathcal{X}$, for all $f \in \mathcal{X}_0 + \mathcal{X}_1$,

$$\|f\|_{\Sigma(\mathcal{X}_0,\mathcal{X}_1)} \leqslant \|f\|_{\mathcal{X}} \leqslant \|f\|_{\Delta(\mathcal{X}_0,\mathcal{X}_1)}.$$

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 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 40. *(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 36 in terms of fundamental functions.

Lemma 41. *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 $t \geqslant 0$, let $E_t \subset \Omega$ be such that $\mu(E_t) = t$. Equation 5.2 implies that for all $t \geqslant 0$,

$$\phi_{\Delta(\mathcal{X}_0,\mathcal{X}_1)}(t) = \|\chi_{E_t}\|_{\Delta(\mathcal{X}_0,\mathcal{X}_1)} = \|\chi_{E_t}\|_{\mathcal{X}_0} \vee \|\chi_{E_t}\|_{\mathcal{X}_1} = \phi_{\mathcal{X}_0}(t) \vee \phi_{\mathcal{X}_1}(t).$$

Lemmas 35 and 40 together imply that for all $t \geqslant 0$,

$$\begin{aligned}\phi_{\Sigma(\mathcal{X}_0,\mathcal{X}_1)}(t) &= \frac{t}{\phi_{\Sigma(\mathcal{X}_0,\mathcal{X}_1)'}(t)} = \frac{t}{\phi_{\Delta(\mathcal{X}_0',\mathcal{X}_1')}(t)} = \frac{t}{\phi_{\mathcal{X}_0'}(t) \vee \phi_{\mathcal{X}_1'}(t)} \\ &= \frac{t}{\frac{t}{\phi_{\mathcal{X}_0}(t)} \vee \frac{t}{\phi_{\mathcal{X}_1}(t)}} = \phi_{\mathcal{X}_0}(t) \wedge \phi_{\mathcal{X}_1}(t).\end{aligned}$$

■

Since $\breve{\psi}(x, 1) = \psi(x)$, we see that the ψ functions from Lemma 39 corresponding to max and min are $\psi_{\max}(x) = \max(x, 1)$ and $\psi_{\min}(x) = \min(x, 1)$, which are indeed both quasiconcave.

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 $((\mathcal{Y}_0, Y_1); \mathcal{Y})$ 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 $T \in \mathscr{L}(\mathcal{X}, \mathcal{Y})$, $\|T\|_{\mathcal{X}\to\mathcal{Y}} \leqslant c\|T\|_{(\mathcal{X}_0,\mathcal{X}_1)\to(\mathcal{Y}_0,\mathcal{Y}_1)}$, where the operator norm $\|T\|_{\mathcal{X}\to\mathcal{Y}} = \sup\{\|TX\|_{\mathcal{Y}} \colon X \in \mathcal{X} \text{ and } \|X\|_{\mathcal{X}} \leqslant 1\}$. If $c \leqslant 1$ 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 $\bar{\mathfrak{B}}$ denote the category of Banach pairs (for what follows it suffices to just consider these as sets).

Definition 42. *An* interpolation functor *is a functor $\mathscr{F} \colon \bar{\mathfrak{B}} \to \mathfrak{B}$ which assigns to each Banach pair $\bar{\mathcal{X}} = (\mathcal{X}_0, \mathcal{X}_1)$ a Banach space $\mathscr{F}(\bar{\mathcal{X}})$ intermediate between $\mathcal{X}_0$ and $\mathcal{X}_1$, and to each operator $T \in \mathcal{L}(\bar{\mathcal{X}}, \bar{\mathcal{Y}})$ it assigns the restriction to the space $\mathscr{F}(\bar{\mathcal{X}})$.*

The triples $(\bar{\mathcal{X}}; \mathscr{F}(\bar{\mathcal{X}}))$ and $(\bar{\mathcal{Y}}; \mathscr{F}(\bar{\mathcal{Y}}))$ are interpolation triples relative to each other. If for any pairs $\bar{\mathcal{X}}$ and $\bar{\mathcal{Y}}$ the resulting triples are normalised then $\mathscr{F}$ is said to be a *normalised interpolation functor*. The functors Δ and Σ introduced in (5.2) and (5.2) 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 $\mathscr{F}$, if we apply it to the Banach pair $(\alpha\mathbb{R}, \beta\mathbb{R})$ we obtain $\mathscr{F}(\alpha\mathbb{R}, \beta\mathbb{R}) = \phi_{\mathscr{F}}(\alpha, \beta)\mathbb{R}$, where the constant $\phi_{\mathscr{F}}(\alpha, \beta)$ is known as the *fundamental function of the functor $\mathscr{F}$* (Brudnyĭ and Krugljak, 1991). (Sometimes $\phi_{\mathscr{F}}$ is called the *characteristic function of the functor $\mathscr{F}$* (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 $\mathscr{F}$, $(\alpha, \beta) \mapsto \phi_{\mathscr{F}}(\alpha, \beta)$ is positive, positively homogeneous, and nondecreasing in each argument. The *dual fundamental function of the*

functor $\mathscr{F}$ is given by $\phi^*_{\mathscr{F}}(\alpha,\beta) = \frac{1}{\phi(1/\alpha,1/\beta)}$. If $\mathscr{F}$ is normalised, $\phi_{\mathscr{F}}(1,1) = 1$. Normalised interpolation functors, when restricted to the Banach pair $(\alpha\mathbb{R},\beta\mathbb{R})$ are characterised by their fundamental function $\phi_{\mathscr{F}}$.

Given a Banach pair $\bar{\mathcal{X}} = (\mathcal{X}_0,\mathcal{X}_1)$, the K-*functional* is defined as

$$K(s_0,s_1,X,\bar{\mathcal{X}}) := \inf\{s_0\|X_0\|_{\mathcal{X}_0}+s_1\|X_1\|_{\mathcal{X}_1} : X_0\in\mathcal{X}_0, X_1\in\mathcal{X}_1 \text{ s.t. } X = X_0+X_1\}, \quad s_0,s_1\geqslant 0$$

Pick an arbitrary function $\phi\in\mathscr{P}$ (recall definition 37), and let $\mathscr{F}$ denote the interpolation functor on $(\alpha\mathbb{R},\beta\mathbb{R})$ with fundamental function $\phi_{\mathscr{F}} = \phi$. On one dimensional spaces, the interpolation functor is entirely determined by its fundamental function; taking ϕ as given, then $\mathscr{F}_\phi$ is given as $\mathscr{F}(\alpha\mathbb{R},\beta\mathbb{R}) = \phi(\alpha,\beta)$. 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 $\underline{\mathscr{F}}$ and upper extension $\overline{\mathscr{F}}$ such that for all pairs of Banach spaces $\overline{\mathcal{X}} = (\mathcal{X}_0,\mathcal{X}_1)$ and all interpolation functors $\mathscr{F}$,

$$\underline{\mathscr{F}}(\overline{\mathcal{X}}) \overset{1}{\hookrightarrow} \mathscr{F}(\overline{\mathcal{X}}) \overset{1}{\hookrightarrow} \overline{\mathscr{F}}(\overline{\mathcal{X}}).$$

The lower and upper extensions are characterised by the following (Brudnyi et al., 1986, Section 1.17):

Lemma 43. *Let $\bar{\mathcal{X}} = (\mathcal{X}_0,\mathcal{X}_1)$ be an arbitrary Banach pair. The lower and upper extensions $\underline{\mathscr{F}}(\bar{\mathcal{X}}) = \Lambda_\phi(\bar{\mathcal{X}})$ and $\overline{\mathscr{F}}(\bar{\mathcal{X}}) = M_\phi(\bar{\mathcal{X}}))$ correspond to the space of all elements of $\mathcal{X}_0+\mathcal{X}_1$ with (respectively) finite norms*

$$\|X\|_{\Lambda_\phi(\bar{\mathcal{X}})} := \inf\sum_k \phi(\|X_k\|_{\mathcal{X}_0},\|X_k\|_{\mathcal{X}_1}),$$

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

$$\|X\|_{M_\phi(\bar{\mathcal{X}})} := \sup_{s_0,s_1}\frac{K(s_0,s_1,X,\bar{\mathcal{X}})}{\phi^*(s_0,s_1)}.$$

The spaces $\Lambda_\phi(\bar{\mathcal{X}})$ and $M_\phi(\bar{\mathcal{X}})$ are called the *abstract Lorentz space* and *abstract Marcinkiewicz space* respectively[23], and the functors Λ_ϕ and M_ϕ are known as the *lower and upper*

23. That 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 $Z = X+Y$),

$$\begin{aligned}
&\|X\|_{\Lambda_\phi(\bar{\mathcal{X}})} + \|Y\|_{\Lambda_\phi(\bar{\mathcal{X}})} \geqslant \|Z\|_{\Lambda_\phi(\bar{\mathcal{X}})}\\
\Leftrightarrow\ & \inf_{\sum_{k_1} X_{k_1}=X}\sum_{k_1}\phi(\|X_{k_1}\|_{\mathcal{X}_0},\|X_{k_1}\|_{\mathcal{X}_1}) + \inf_{\sum_{k_2} Y_{k_2}=Y}\sum_{k_2}\phi(\|Y_{k_2}\|_{\mathcal{X}_0},\|Y_{k_2}\|_{\mathcal{X}_1}) \geqslant \inf_{\sum_k Z_k=Z}\sum_k\phi(\|Z_k\|_{\mathcal{X}_0},\|Z_k\|_{\mathcal{X}_1})\\
\Leftrightarrow\ & \inf_{\substack{\sum_{k_1} X_{k_1}=X\\ \sum_{k_2} Y_{k_2}=Y}}\left(\sum_{k_1}\phi(\|X_{k_1}\|_{\mathcal{X}_0},\|X_{k_1}\|_{\mathcal{X}_1}) + \sum_{k_2}\phi(\|Y_{k_2}\|_{\mathcal{X}_0},\|Y_{k_2}\|_{\mathcal{X}_1})\right) \geqslant \inf_{\sum_k Z_k=Z}\sum_k\phi(\|Z_k\|_{\mathcal{X}_0},\|Z_k\|_{\mathcal{X}_1}),
\end{aligned}$$

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.

extensions of the functor $\mathscr{F}_\phi$ defined on $(\alpha\mathbb{R}, \beta\mathbb{R})$ in terms of the fundamental function $\phi \in \mathscr{P}$ because they bound the behaviour of interpolation functors with a given fundamental function:

Lemma 44. *(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 $\mathscr{F}$ be some normalised interpolation functor for which $\mathscr{F}(\bar{\mathcal{X}}) = \mathcal{X}$ with fundamental function ϕ. Then*

$$\Lambda_\phi(\bar{\mathcal{X}}) \overset{1}{\hookrightarrow} \mathcal{X} \overset{1}{\hookrightarrow} M_\phi(\bar{\mathcal{X}}),$$

and thus

$$\|X\|_{M_\phi(\bar{\mathcal{X}})} \leqslant \|X\|_{\mathcal{X}} \leqslant \|X\|_{\Lambda_\phi(\bar{\mathcal{X}})}.$$

Recall that the Lorentz space Λ_ϕ is always positive translation equivariant (PTE) (Example 2). We might thus conjecture that the interpolation functor $\Lambda_\phi \colon \bar{\mathfrak{B}} \to \mathfrak{B}$ would preserve PTE; indeed that is the case as the lemma below shows.

Lemma 45. *Suppose $\mathcal{X}_0, \mathcal{X}_1$ are* PTE*, and $\phi \in \mathscr{Q}$, then $\Lambda_\phi(\mathcal{X}_0, \mathcal{X}_1)$ is* PTE*.*

The proof is in Appendix A.6.1. Lemma 41 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 $t \geqslant 0$,

$$\phi_0(t) \wedge \phi_1(t) \leqslant \phi(t) \leqslant \phi_0(t) \vee \phi_1(t).$$

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 $\phi_0 \leqslant \phi \leqslant \phi_0$ and hence $\phi = \phi_0$. We formalise this observation as follows.

Lemma 46. *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*: $\mathcal{X}_0 \overset{1}{\hookrightarrow} \mathcal{X} \overset{1}{\hookrightarrow} \mathcal{X}_1$. Then $\phi_{\mathcal{X}} = \phi$.*

Proof The embedding ensures that for all $X \in \mathcal{X}$, we have $\|X\|_{\mathcal{X}_1} \leqslant \|X\|_{\mathcal{X}} \leqslant \|X\|_{\mathcal{X}_0}$ and thus choosing $X = \chi_{[0,t]}$ for some arbitrary $t > 0$ we obtain

$$\phi(t) = \|\chi_{[0,t]}\|_{\mathcal{X}_1} \leqslant \|\chi_{[0,t]}\|_{\mathcal{X}} = \phi_{\mathcal{X}}(t) \leqslant \|\chi_{[0,t]}\|_{\mathcal{X}_0} = \phi(t).$$

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 41 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 47. *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 $\bar{\phi} \in \mathscr{P}$. Then the fundamental function of $M_{\bar{\phi}}(\bar{\mathcal{X}})$ satisfies*

$$\phi_{M_{\bar{\phi}}(\bar{\mathcal{X}})}(t) = \bar{\phi}(\phi_0(t), \phi_1(t)), \qquad t > 0.$$

The proof is in Appendix A.6.2. We have an analogous (one-sided) result for $\Lambda_\phi(\bar{\mathcal{X}})$:

Lemma 48. *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 $\bar{\phi} \in \mathscr{P}$. Then the fundamental function of $\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})$ satisfies*

$$\phi_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})}(t) \leqslant \bar{\phi}(\phi_0(t), \phi_1(t)), \qquad t > 0.$$

The proof is in Appendix A.6.3. Lemmas 43, 47, and 48 combined with (44) from Lemma 44 imply that for all $t > 0$,

$$\bar{\phi}(\phi_0(t), \phi_1(t)) = \phi_{M_{\bar{\phi}}(\bar{\mathcal{X}})} = \|\chi_{E_t}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} \leqslant \|\chi_{E_t}\|_{\mathcal{X}} = \phi_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})}(t) \leqslant \|\chi_{E_t}\|_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})} \leqslant \bar{\phi}(\phi_0(t), \phi_1(t)),$$

and thus by Lemma 46, we have for all $t > 0$,

$$\phi_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})}(t) = \bar{\phi}(\phi_0(t), \phi_1(t)).$$

We have thus proved:

Theorem 49. *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 $\mathscr{F}$ is a normalized interpolation functor with fundamental function $\phi_{\mathscr{F}}$. Then $\mathscr{F}(\mathcal{X}_0, \mathcal{X}_1)$ is an intermediate space for $(\mathcal{X}_0, \mathcal{X}_1)$ and its fundamental function satisfies*

$$\phi_{\mathscr{F}(\mathcal{X}_0, \mathcal{X}_1)} = \phi_{\mathscr{F}}(\phi_0, \phi_1).$$

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 $X \colon \Omega \to \mathbb{R}$ 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 norms[24].

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). Notwithstanding 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 49 are given below.

Example 4. When $\phi_{\mathcal{X}_0} = \phi_{\mathcal{X}_1}$, Theorem 49 implies $\phi_{\mathcal{X}}(t) = \phi(\phi_{\mathcal{X}_1}(t), \phi_{\mathcal{X}_1}(t)) = \phi_{\mathcal{X}_1}(t)$ using (5.1) and the fact that $\phi(1,1) = 1$ since $\mathscr{F}$ is a *normalised* interpolation functor. This agrees with the observation made earlier following (5.3).

Example 5. Consider two interpolation functors from Lemma 41: $\Delta(\mathcal{X}_0, \mathcal{X}_1) = \mathcal{X}_0 \cap \mathcal{X}_1$ 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 $\|X\|_{\Sigma(\alpha\mathbb{R},\beta\mathbb{R})} = \inf\{\alpha|X_0| + \beta|X_1| \colon X_0 + X_1 = X\}$. The infimum is achieved by choosing $X_0 = X$ and $X_1 = 0$ when $\alpha \leqslant \beta$ and $X_0 = 0$ and $X_1 = X$ when $\alpha \geqslant \beta$, 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 41.

Example 6. 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 $\|f\|_{\mathcal{X}_0 + \mathcal{X}_1} \leqslant \|f\|_{\mathcal{X}_0 \tilde{+} \mathcal{X}_1} \leqslant \|f\|_{\mathcal{X}_0 \cap \mathcal{X}_1}$, 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 $t \geqslant 0$, we have

$$\phi_{\mathcal{X}_0 \tilde{+} \mathcal{X}_1}(t) = \frac{1}{2}\left(\phi_{\mathcal{X}_0}(t) + \phi_{\mathcal{X}_1}(t)\right).$$

24. The 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).

Example 7. 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 $\alpha + \beta \leqslant \rho(\alpha, \beta) \leqslant \alpha \vee \beta$ and it immediately follows from the definition that for all $t \geqslant 0$,

$$\phi_{\rho(\mathcal{X}_0,\mathcal{X}_1)}(t) = \rho(\phi_0(t), \phi_1(t)).$$

Example 8. A special case of theorem 49 (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 $\mathscr{F}^\theta$ "of exponent θ", $\theta \in (0,1)$. They showed that for all $t \geqslant 0$,

$$\phi_{\mathscr{F}^\theta(\mathcal{X}_0,\mathcal{X}_1)}(t) = \phi^{1-\theta}_{\mathcal{X}_0}(t)\phi^{\theta}_{\mathcal{X}_1}(t),$$

which can be seen to be of the form of (49). 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 9. Consider two Lorentz norms Λ_{ϕ_0} and Λ_{ϕ_1} and an arbitrary interpolation functor $\mathscr{F}$. A natural question to ask is when (if ever) is $\mathscr{F}(\Lambda_{\phi_0}, \Lambda_{\phi_1})$ 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 $\mathscr{F}$ corresponds to the abstract Lorentz norm this is true.

We do know that if $\mathscr{F} = \Lambda_{\bar{\phi}}(\cdot,\cdot)$ then $\phi_{\mathscr{F}} = \bar{\phi}$, since $\Lambda_{\bar{\phi}}$ is the extension of the functor defined on one dimensional spaces with fundamental function $\bar{\phi}$. We also know by (5.3) 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 9, we thus have for all X,

$$\|X\|_{\Lambda_{\bar{\phi}}(\mathcal{X}_0,\mathcal{X}_1)} \leqslant \|X\|_{\Lambda_{\bar{\phi}(\phi_0,\phi_1)}}.$$

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 $\psi \in \mathscr{Q}$ and letting $\bar{\phi} = \breve{\psi}$ 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 $\breve{\psi}(x,y) = \breve{\psi}(y,x)$ for all $x, y > 0$, it is necessary and sufficient that $\psi^\diamond(t) = \psi(t)$ for all $t \in (0,1]$, where $\psi^\diamond(t) := t\psi(1/t)$ is the Csiszár conjugate of ψ. One can thus readily construct symmetric $\bar{\phi} = \breve{\psi}$ by choosing an arbitrary quasiconcave function ψ on $[0,1]$ and extending it to $[1,\infty)$ via $\psi(t) = t\psi(1/t)$ for $t \geqslant 1$, the resulting ψ is then guaranteed to satisfy the Csiszár conjugate condition and thus the induced perspective $\breve{\psi}$ is guaranteed symmetric.

Observe that in contrast to the method in Example 7, 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.

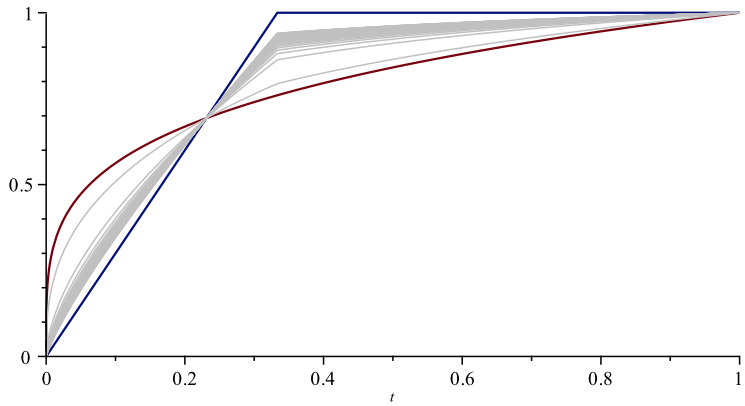

Figure 4: Illustration of the interpolation between two quasiconcave fundamental functions. The graph shows $\phi_{\text{red}}(t) = t^{1/4}$ (in red) and $\phi_{\text{blue}}(t) = 3t \wedge 1$ (in blue). The grey curves are obtained via $\phi(t) = \breve{\phi}_a(\phi_{\text{red}}(t), \phi_{\text{blue}}(t))$ where $\breve{\phi}_a$ 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.

Example 10. In order to provide some insight into (49), especially for its use in Example 9, in Figure 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.

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 and Prashanth (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

$$\operatorname*{argmin}_f R(\ell(f(X), Y))$$

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.

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:

$$(\forall \alpha \in [0,1) : \mathrm{CVar}_\alpha(X) \leqslant \mathrm{CVar}_\alpha(Y)) \Rightarrow R(X) \leqslant R(Y) \tag{18}$$

for *any* ri function norm R and $X, Y \in \mathcal{M}^+$. Interestingly, the condition that $\forall \alpha \in [0,1) : \mathrm{CVar}_\alpha(X) \leqslant \mathrm{CVar}_\alpha(Y)$ is equivalent to saying that X is dominated by Y in the *second stochastic order* (Ding, 2023; Bäuerle and Müller, 2006). Then (18) 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 $\mathrm{CVar}_\alpha(X)$ for some X. As an example, we draw 500 samples from a standard normal

25. Note 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).

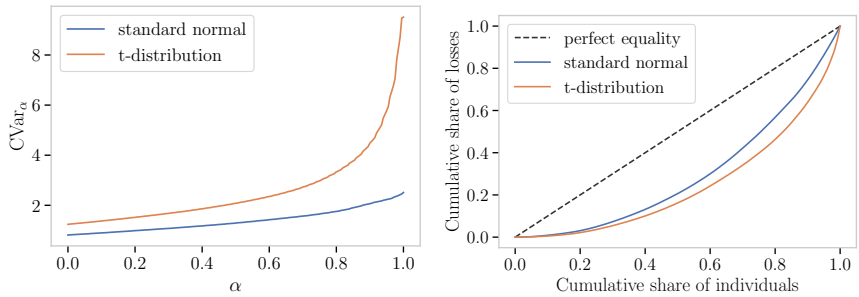

Figure 5: Left: CVar_α curves for 500 randomly drawn samples from a standard normal and a t-distribution with 2 degrees of freedom, respectively. Only nonnegative samples were kept. Both have approximately the same mean $CVar_{\alpha=0}$, but the t-distribution has substantially more weight in the tails. 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.

distribution and a t-distribution with 2 degrees of freedom. We keep only the nonnegative samples and interpret them as losses. The empirical CVar_α curves are shown in Figure 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_α curves from samples. Another possibility is to introduce a cutoff value, so only values until for example $\alpha = 0.99$ are plotted.

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):

$$L_X(q) := \frac{1}{\mathbb{E}[X]} \int_0^q F_X^{-1}(p)\ \mathrm{d}p, \quad 0 < q \leqslant 1; \quad L_X(0) = 0.$$

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 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:

$$(\forall \alpha \in [0,1) : \mathrm{CVar}_\alpha(X) \geqslant \mathrm{CVar}_\alpha(Y)) \Leftrightarrow (\forall q \in (0,1] : L_X(q) \leqslant L_Y(q)) \,.$$

For a visualization, see Figure 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.

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.

6.3.1 Data

We use the `MNIST`[26] and the `adult`[27] 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.

26. http://yann.lecun.com/exdb/mnist/
27. https://archive.ics.uci.edu/ml/datasets/Adult

6.3.2 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

$$\min_{V_k} \sum_{i=1}^{n} \|x_i - V_k^\intercal V_k x_i\|_{\mathcal{L}^2}^2 \quad \text{so that } V_k V_k^\intercal = I_k$$

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:

$$\min_{V_k} R\left(\|X - V_k^\intercal V_k X\|_{\mathcal{L}^2}^2\right)$$

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` $k = 5$. We train for 2000 epochs with a learning rate of 0.001. To avoid additional challenges from the stochasticity of mini-batches (see Section 6.5), we use the full data for each epoch. We repeat the experiments over 25 independent runs with random training and test splits.

6.3.3 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 $\mathrm{CVar}_{0.0} = \mathbb{E}$. As α increases, sensitivity to tail risk increases. Moreover, we compare variations of the risk measure for integrated risk management ($\mathrm{RIM}_{\alpha,\beta}$), where α is fixed at 0.7 and $\beta \in \{0.2, .., 0.8, 1.0\}$. Recall that

$$\mathrm{RIM}_{\alpha,\beta}(X) := \beta \mathbb{E}[X] + (1-\beta)\,\mathrm{CVar}_\alpha(X)$$

Hence $\beta = 1$ yields the expectation, whereas $\beta = 0$ would yield $\mathrm{CVar}_{\alpha=0.7}$. 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.

6.3.4 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 C. In Appendix C, 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 $\mathrm{RIM}_{\alpha,\beta}$ as a risk measure leads to lower tail risks as compared to the expectation. For

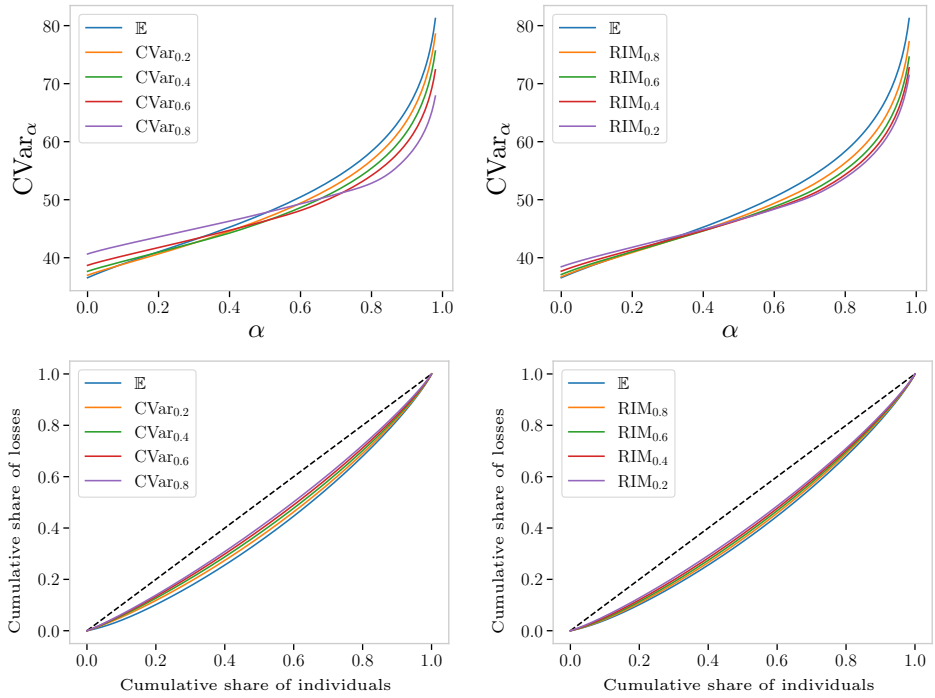

Figure 6: PCA* results on `MNIST`. Top row: CVar_α curves of test losses for CVar_α risk measures (left) with different α and RIMs risk measures (right) with $\alpha = 0.7$ and different β, indicated by subscript. For better visibility of the differences, we cut off α at 0.98. Bottom row: Lorenz curves of test losses for CVar_α (left) and RIMs (right) with $\alpha = 0.7$ and different β.

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 $\mathrm{RIM}_{\alpha,\beta}$ (with fixed α) gradually achieves a more equal distribution. This is most clearly visible in the Gini boxplots (Appendix C).

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 as S. The objective in standard expected risk minimization can be written in a two-stage manner as

$$\mathbb{E}[\ell(f(X), Y)] = \mathbb{E}_S \left[\mathbb{E}_{X,Y|S}[\ell(f(X), Y)]\right]$$

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.

6.4.1 Data

In this experiment we perform multi-class classification on `MNIST` (see Section 6.3.1) 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 $s \in [0, 9]$:

$$N(s) = 5000 \cdot \exp\left(\frac{-2s}{10}\right)$$

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 11 in Appendix C. 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 6.3.1.

6.4.2 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 6.5). We compare the same risk measures as in the first experiment (Section 6.3.3). 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.

6.4.3 Results

For `MNIST`, we report the average subgroup test accuracies and Gini coefficients of subgroup accuracies in Figure 7. Due to the data set shift setting, we find that the risk measures even

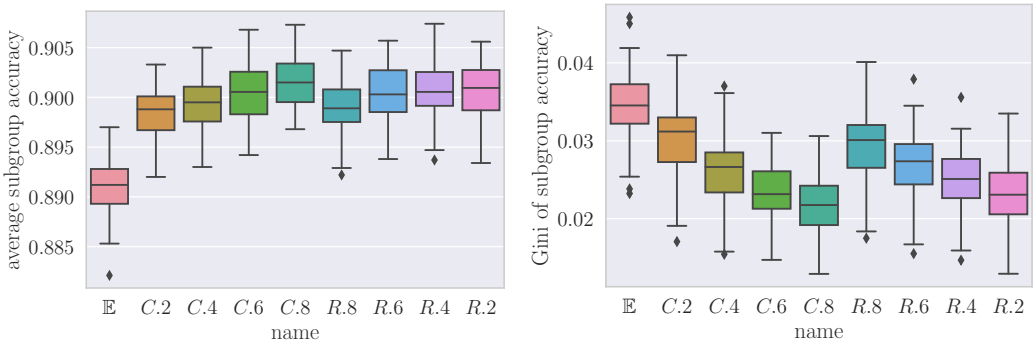

Figure 7: Results of logistic regression on `MNIST` across 50 independent runs. Left: average subgroup accuracies. Right: Gini coefficients of subgroup accuracies. We abbreviate $\mathrm{CVar}_{\alpha=0.f}$ as $C.f$ and $\mathrm{RIM}_{\alpha=0.7,\beta=0.f}$ as $R.f$.

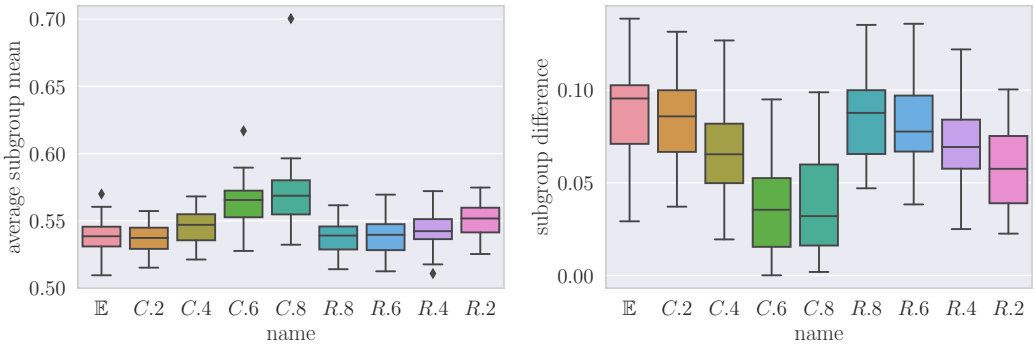

Figure 8: Results of linear regression on `winequality` across 50 independent runs. Left: averages of the two subgroup error means on test data. Right: subgroup absolute error differences on test data across 50 independent runs. We abbreviate $\mathrm{CVar}_{\alpha=0.f}$ as $C.f$ and $\mathrm{RIM}_{\alpha=0.7,\beta=0.f}$ as $R.f$.

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 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.

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 *estimatibility*. 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 $\mathrm{RIM}_{\alpha,\beta}$, 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. (2022) have obtained uniform convergence results that justify the optimization of a wide class of risk measures (including the spectrals) on the empirical distribution.

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 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 Rabanus Derr for many helpful discussions and comments.

Appendix A. Proofs

A.1 Proof of Theorem 5

Proof (Gzyl and Mayoral, 2008; Ridaoui and Grabisch, 2016):

$$
\begin{aligned}
R_\phi(X) &= \int_{-\infty}^{0} [\phi(S_X(x)) - 1] \, \mathrm{d}x + \int_0^\infty \phi\left(S_X(x)\right) \, \mathrm{d}x && (19)\\
&= \int_{-\infty}^{0} [\phi(1 - F_X(x)) - 1] \, \mathrm{d}x + \int_0^\infty \phi\left(1 - F_X(x)\right) \, \mathrm{d}x \\
&=^{t=F_X(x)} \int_0^{F_X(0)} [\phi(1-t) - 1] \, (F_X^{-1})'(t) \, \mathrm{d}t + \int_{F_X(0)}^{1} \phi(1-t)(F_X^{-1})'(t) \, \mathrm{d}t \\
&= \int_0^{F_X(0)} \phi'(1-t) F_X^{-1}(t) \, \mathrm{d}t + \int_{F_X(0)}^{1} \phi'(1-t) F_X^{-1}(t) \, \mathrm{d}t && (20)\\
&= \int_0^1 F_X^{-1}(1-t)\phi'(t) \, \mathrm{d}t = R_{(w)}(X),
\end{aligned}
$$

where the step (20) 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. ■

A.2 Proof of Theorem 11

Proof First observe that it is always true that $\mathbb{E}[Y] \leqslant R'(Y)$, 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 (13) as TErep, that is, the possibility of a representation of the form:

$$
R(X) = \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, \mathrm{d}\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+ \right\}.
$$

That TErep implies PTE is trivial:

$$
\begin{aligned}
R(X + c) &= \sup \left\{ \int_0^1 (X + c)^*(\omega) Y^*(\omega) \, \mathrm{d}\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+ \right\} \\
&= \sup \left\{ \int_0^1 X^*(\omega) Y^*(\omega) \, \mathrm{d}\omega + c \int_0^1 Y^*(\omega) \, \mathrm{d}\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+ \right\} \\
&= R(X) + c,
\end{aligned}
$$

where we used the fact that $(X + c)^* = X^* + c$ holds even if $c < 0$ as long as $X \geqslant 0$ and $X + c \geqslant 0$. The other direction is more involved. We show PTE $\Rightarrow$ TErep by showing $\neg$ TErep $\Rightarrow \neg$ 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:

$$
R(Z) = \sup \left\{ \int_0^1 Z^*(\omega) Y^*(\omega) \, \mathrm{d}\omega : \mathbb{E}[Y] < 1, R'(Y) \leqslant 1, Y \in \mathcal{M}^+ \right\}, \tag{21}
$$

that is, the supremum is not decreased when taking it only over the subset $\{\mathbb{E}[Y] < 1, R'(Y) \leqslant 1\}$. However, the supremum need not be actually attained.

Lemma 50. $\neg A \wedge \text{PTE} \Rightarrow \text{TErep}$*, 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

$$R(Z) = \sup\left\{\int_0^1 Z^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] = R'(Y) = 1, Y \in \mathcal{M}^+\right\}. \tag{22}$$

To see that this is indeed the negation of A, observe that $\neg A$ means $\forall Z \in \mathcal{M}^+, Z > \epsilon$, for some $\epsilon > 0$

$$R(Z) > \sup\left\{\int_0^1 Z^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] < 1, R'(Y) \leqslant 1, Y \in \mathcal{M}^+\right\}. \tag{23}$$

Negating the equality in (21) here must yield "strictly greater", since the set on the righthand side in (23) is a subset of the full envelope $\{Y : R'(Y) \leqslant 1\}$, which is implicit in the lefthand side of (23). Now we continue to analyze the statement (23). Formally, we can write it as $\sup_C f(Y) > \sup_{C\setminus B} f(Y)$, where $B \subseteq C$. Here, $C := \{Y : R'(Y) \leqslant 1\}$ and $B := \{Y : \mathbb{E}[Y] = R'(Y) = 1\}$. The function $f(Y) := \int_0^1 Z^*(\omega)Y^*(\omega)\ \mathrm{d}\omega$ 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 (22) 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 $Y \in \mathcal{M}^+$):

$$R(X) = \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : R'(Y) \leqslant 1\right\} \quad \text{(general representation for any } R\text{)}$$

$$\Rightarrow \forall\epsilon > 0 : R(X) = \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega + \epsilon\int_0^1 Y^*(\omega)\ \mathrm{d}\omega - \epsilon\int_0^1 Y^*(\omega)\ \mathrm{d}\omega : R'(Y) \leqslant 1\right\},$$

from which it follows due to PTE that:

$$\forall\epsilon > 0 : R(X) = \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega + \epsilon\int_0^1 Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] = R'(Y) = 1\right\} - \epsilon.$$

Since due to PTE: $R(X+\epsilon-\epsilon) = R(X+\epsilon)-\epsilon$ since $(X+\epsilon)-\epsilon \geqslant 0$. Taking the supremum over $\{Y : \mathbb{E}[Y] = R'(Y) = 1\}$ suffices due to $\neg A$. Then:

$$\forall \epsilon > 0 : R(X) = \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega + \epsilon\int_0^1 Y^*(\omega)\ \mathrm{d}\omega - \epsilon\int_0^1 Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] = R'(Y) = 1\right\}$$

$$\Rightarrow R(X) = \sup\left\{\int_0^1 X^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] = R'(Y) = 1\right\}.$$

Hence R has a positive translation equivariant representation, i.e. TErep holds. We have thus shown that $\neg A \wedge \mathrm{PTE} \Rightarrow \mathrm{TErep}$, which is equivalent to $\neg\,\mathrm{TErep} \Rightarrow A \vee \neg\,\mathrm{PTE}$. ■

Recall that our goal is to show that $\neg\,\mathrm{TErep} \Rightarrow \neg\,\mathrm{PTE}$. Using the lemma 50, that $\neg\,\mathrm{TErep} \Rightarrow A \vee \neg\,\mathrm{PTE}$, it only remains to show that $A \Rightarrow \neg\,\mathrm{PTE}$. So now assume A. This means that $\exists Z \in \mathcal{M}^+, Z > \epsilon$, for some $\epsilon > 0$, so that

$$R(Z) = \sup\left\{\int_0^1 Z^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] < 1, R'(Y) \leqslant 1, Y \in \mathcal{M}^+\right\}.$$

Write now $X + \epsilon = Z$, which is possible by assumption. Then $X \in \mathcal{M}^+$ and:

$$\begin{aligned} R(X+\epsilon) &= \sup\left\{\int_0^1 (X+\epsilon)^*(\omega)Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] < 1, R'(Y) \leqslant 1, Y \in \mathcal{M}^+\right\} \\ &\leqslant R(X) + \epsilon\sup\left\{\int_0^1 Y^*(\omega)\ \mathrm{d}\omega : \mathbb{E}[Y] < 1, R'(Y) \leqslant 1, Y \in \mathcal{M}^+\right\} \qquad \text{(subadditivity)} \\ &< R(X) + \epsilon. \end{aligned}$$

Therefore R is not PTE and the proof is complete. ■

A.3 Families of Fundamental Functions

A.3.1 Proof of Theorem 20.

Proof Suppose R is an ri norm which is PTE and has fundamental function ϕ. Write

$$R(X) = \sup_{Z\in\mathcal{Z}}\left\{\int_0^1 X^*(\omega)Z'(\omega)\ \mathrm{d}\omega\right\}, \quad \phi(t) = \sup_{Z\in\mathcal{Z}} Z(t).$$

for some Kusuoka set $\mathcal{Z}$ of R. And define

$$\forall t \in (0,1] : Z_t(x) := \begin{cases} \phi(t)\frac{x}{t} & , x \leqslant t \\ \frac{1-\phi(t)}{1-t}x + \frac{\phi(t)-t}{1-t} & , x > t \end{cases}$$

Then the TM norm can be written as:

$$\|X\|_{TM_\phi} = \sup\left\{\int_0^1 X^*(\omega)Z_t'(\omega)\ \mathrm{d}\omega : t \in (0,1]\right\}. \tag{24}$$

Lemma 51. *Using the above definitions, it holds that:*

$$\forall t \in (0,1] : R(X) \geqslant \int_0^1 X^*(\omega) Z_t'(\omega)\ \mathrm{d}\omega.$$

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 $\varepsilon > 0$, find some $Z_\varepsilon \in \mathcal{Z}$ so that $0 \leqslant \phi(t) - Z_\varepsilon(t) < \varepsilon$. Let $\varepsilon_n \downarrow 0$ 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

$$h_{\varepsilon_n}(x) = \begin{cases} Z_{\varepsilon_n}(t)\frac{x}{t} & , x \leqslant t \\ \frac{1-Z_{\varepsilon_n}(t)}{1-t}x + \frac{Z_{\varepsilon_n}(t)-t}{1-t} & , x > t \end{cases}, \quad h'_{\varepsilon_n}(x) := \begin{cases} \frac{Z_{\varepsilon_n}(t)}{t} & , x \leqslant t \\ \frac{Z_{\varepsilon_n}(t)-1}{t-1} & , x > t \end{cases}$$

where the derivative is defined first so that $h_{\varepsilon_n}(x) := \int_0^x h'_{\varepsilon_n}(\omega)\ \mathrm{d}\omega$. ■

We observe that by construction $Z_{\varepsilon_n} \geqslant h_{\varepsilon_n}$; the condition for Hardy's lemma is fulfilled, i.e. $\int_0^x Z'_{\varepsilon_n}(\omega)\ \mathrm{d}\omega \geqslant \int_0^x h'_{\varepsilon_n}(\omega)\ \mathrm{d}\omega$. Therefore

$$\mathrm{I}(\varepsilon_n) := \int_0^1 Z'_{\varepsilon_n}(\omega) X^*(\omega)\ \mathrm{d}\omega \geqslant \int_0^1 h'_{\varepsilon_n}(\omega) X^*(\omega)\ \mathrm{d}\omega =: \mathrm{II}(\varepsilon_n).$$

By definition of h'_{ε_n},

$$\mathrm{II}(\varepsilon_n) = \frac{Z_{\varepsilon_n}(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega + \frac{Z_{\varepsilon_n}(t)-1}{t-1} \int_t^1 X^*(\omega)\ \mathrm{d}\omega.$$

Now consider $\mathrm{I}(\varepsilon_n)$. We know that $R(X) \geqslant \mathrm{I}(\varepsilon_n)$ by our choice of the Z_{ε_n}. Hence

$$R(X) \geqslant \sup_{n \in \mathbb{N}} \left\{ \int_0^1 Z'_{\varepsilon_n}(\omega) X^*(\omega)\ \mathrm{d}\omega \right\} = \limsup_{\varepsilon_n \downarrow 0} \mathrm{I}(\varepsilon_n) \geqslant \limsup_{\varepsilon_n \downarrow 0} \mathrm{II}(\varepsilon_n),$$

since $0 \leqslant a_n \leqslant b_n$ implies $\limsup a_n \leqslant \limsup b_n$.

We find that

$$\lim_{n \to \infty} \frac{Z_{\varepsilon_n}(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega = \frac{\phi(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega,$$

since the integral term is constant, and as $n \to \infty$, $Z_{\varepsilon_n}(t) \to \phi(t)$ by construction (note that t is fixed). Similarly

$$\lim_{n \to \infty} \frac{Z_{\varepsilon_n}(t)-1}{t-1} \int_t^1 X^*(\omega)\ \mathrm{d}\omega = \frac{\phi(t)-1}{t-1} \int_t^1 X^*(\omega)\ \mathrm{d}\omega.$$

Since both limits exist, the limit of their sum exists:

$$\lim_{n \to \infty} \mathrm{II}(\varepsilon_n) = \frac{\phi(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega + \frac{\phi(t)-1}{t-1} \int_t^1 X^*(\omega)\ \mathrm{d}\omega = \int_0^1 Z'_t(\omega) X^*(\omega)\ \mathrm{d}\omega.$$

Therefore $R(X) \geqslant \int_0^1 Z'_t(\omega) X^*(\omega)\ \mathrm{d}\omega$ for fixed t. Since this holds for all $t \in (0,1]$, we also get $R(X) \geqslant \sup_{t \in (0,1]} \left\{ \int_0^1 X^*(\omega) Z'_t(\omega)\ \mathrm{d}\omega \right\} = \|X\|_{TM_\phi}$.

Thus, taking the embedding theorem into account, for any PTE ri norm ('coherent risk measure') R:

$$\|X\|_{M_\phi} \leqslant \|X\|_{TM_\phi} \leqslant R(X) \leqslant \|X\|_{\Lambda_\phi} \quad \forall X \in \Lambda_\phi.$$

We show explicitly that $\|\cdot\|_{TM_\phi}$ is positive translation equivariant. Let $X \in \mathcal{M}^+$ and $c \in \mathbb{R}$ so that $X + c \geqslant 0$:

$$\begin{aligned}
&\|X+c\|_{TM_\phi} \\
&= \sup_{0<t<1} \left\{ \frac{\phi(t)}{t} \int_0^t (X+c)^*(\omega)\ \mathrm{d}\omega + \frac{\phi(t)-1}{t-1} \int_t^1 (X+c)^*(\omega)\ \mathrm{d}\omega \right\} \\
&= \sup_{0<t<1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega + \frac{\phi(t)-1}{t-1} \int_t^1 X^*(\omega)\ \mathrm{d}\omega + \frac{\phi(t)}{t} \int_0^t c\ \mathrm{d}\omega + \frac{\phi(t)-1}{t-1} \int_t^1 c\ \mathrm{d}\omega \right\} \\
&= \sup_{0<t<1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega + \frac{\phi(t)-1}{t-1} \int_t^1 X^*(\omega)\ \mathrm{d}\omega + \frac{\phi(t)}{t} ct + \frac{\phi(t)-1}{t-1}(c-ct) \right\} \\
&= \sup_{0<t<1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega + \frac{\phi(t)-1}{t-1} \int_t^1 X^*(\omega)\ \mathrm{d}\omega + c \right\} \\
&= \|X\|_{TM_\phi} + c
\end{aligned}$$

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

$$\begin{aligned}
\|X\|_{TM_\phi} &= \sup_{0<t<1} \left\{ (2-t) \int_0^t X^*(\omega)\ \mathrm{d}\omega + (1-t) \int_t^1 X^*(\omega)\ \mathrm{d}\omega \right\} \\
&= \sup_{0<t<1} \left\{ 2 \int_0^t X^*(\omega)\ \mathrm{d}\omega - t \int_0^1 X^*(\omega)\ \mathrm{d}\omega + \int_t^1 X^*(\omega)\ \mathrm{d}\omega - t \int_t^1 X^*(\omega)\ \mathrm{d}\omega \right\} \\
&= \sup_{0<t<1} \left\{ (1-t) \int_0^1 X^*(\omega)\ \mathrm{d}\omega + \int_0^t X^*(\omega)\ \mathrm{d}\omega \right\} \\
&= \sup_{0<t<1} \{(1-t)\mathbb{E}[|X|] + t \cdot \mathrm{CVar}_{1-t}(|X|)\} = \mathrm{Du}(|X|).
\end{aligned}$$

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 (24), 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 52). Recall that we assume $R(\chi_\Omega) = 1$ throughout the paper.

Lemma 52. *Let $\{R_i : i \in \mathcal{I}\}$ be a non-empty family of ri function norms (Definitions 6,7). 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

$$(\forall i \in \mathcal{I} : 0 \leqslant X_n \uparrow X\ \mu\text{-a.e.} \Rightarrow R_i(X_n) \uparrow R_i(X)) \implies (0 \leqslant X_n \uparrow X\ \mu\text{-a.e.} \Rightarrow R(X_n) \uparrow R(X)).$$

So let us assume $\forall i \in \mathcal{I} : 0 \leqslant X_n \uparrow X$ μ-a.e. $\Rightarrow R_i(X_n) \uparrow R_i(X)$. Then $R(X) = \sup_{i\in\mathcal{I}} R_i(X) = \sup_{i\in\mathcal{I}} \lim_{n\to\infty} R_i(X_n)$. 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

$$\forall \delta > 0 : \forall i \in \mathcal{I} : \exists n_i \in \mathbb{N} : \forall n \geqslant n_i : R_i(X_n) > R(X) - \delta/2.$$

Choosing $\varepsilon := \delta$ and taking the corresponding $i(\delta)$ yields:

$$\begin{aligned}
&\forall \delta > 0 : \exists i(\delta) : \exists n_i \in \mathbb{N} : \forall n \geqslant n_i : R_i(X_n) > R_i(X) - \delta/2 > R(X) - \delta/2 \\
&\Leftrightarrow \forall \delta > 0 : \exists n_i \in \mathbb{N} : \forall n \geqslant n_i : \sup_{i\in\mathcal{I}} R_i(X_n) > R(X) - \delta \\
&\Leftrightarrow \lim_{n\to\infty} \sup_{i\in\mathcal{I}} R_i(X_n) \geqslant R(X). \qquad (25)
\end{aligned}$$

The statement $\lim_{n\to\infty} R(X_n) = R(X)$ can be written as

$$\lim_{n\to\infty} \sup_{i\in\mathcal{I}} R_i(X_n) = \sup_{i\in\mathcal{I}} \lim_{n\to\infty} R_i(X_n).$$

Since $\lim_{n\to\infty} \sup_{i\in\mathcal{I}} R_i(X_n) \leqslant \sup_{i\in\mathcal{I}} \lim_{n\to\infty} R_i(X_n)$ is obvious, we have, taking this together with (25), shown both inequalities and thus equality, i.e. $\lim_{n\to\infty} R(X_n) = R(X)$. 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 $R(\chi_E) \leqslant 1$ for measurable E. Also, $\int_E X \, \mathrm{d}\mu \leqslant \int_\Omega X \, \mathrm{d}\mu = \mathbb{E}[X] \leqslant R_i(X) \forall i \in \mathcal{I}$ by assumption that $R_i(\chi_\Omega) = 1$ and the embedding theorem. Hence choosing e.g. $c = 2$ gives $\int_E X \, \mathrm{d}\mu < cR(X)$ for any measurable E. Finally, the ri property is obvious. ■

The theorem is thus proved. ■

A.3.2 Proof of Theorem 21

Proof Let $\phi(t) = \min\{t/(1-\alpha), 1\}$ for some $\alpha \in [0, 1)$. Hence $\phi \in \Phi_{0+}$ We show that $\|X\|_{M_\phi} = \mathrm{CVar}_\alpha(|X|) = \|X\|_{\Lambda_\phi}$. Clearly, the Lorentz norm for such ϕ is CVar_α, as

$$\|X\|_{\Lambda_\phi} = \frac{1}{1-\alpha} \int_0^{1-\alpha} X^*(\omega) \, \mathrm{d}\omega = \mathrm{CVar}_\alpha(|X|).$$

The Marcinkiewicz norm is

$$\|X\|_{M_\phi} = \sup_{0<t\leqslant 1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega) \, \mathrm{d}\omega \right\}$$

We claim that the supremum is reached at $t = 1 - \alpha$. Then:

$$\|X\|_{M_\phi} = \frac{1}{1-\alpha} \int_0^{1-\alpha} X^*(\omega) \, \mathrm{d}\omega = \|X\|_{\Lambda_\phi} = \mathrm{CVar}_\alpha(|X|).$$

Hence it remains to show that the supremum is indeed reached at $t = 1-\alpha$. Let $t = 1-\alpha+\epsilon$, $\epsilon > 0$. Then

$$\frac{\phi(t)}{t} \int_0^t X^*(\omega) \, \mathrm{d}\omega = \phi(1-\alpha+\epsilon) X^{**}(1-\alpha+\epsilon) = 1 \cdot \mathrm{CVar}_{\alpha-\epsilon}(|X|) < \mathrm{CVar}_\alpha(|X|).$$

Let $t < 1-\alpha$. Then

$$\frac{\phi(t)}{t}\int_0^t X^*(\omega)\ \mathrm{d}\omega = \frac{1}{1-\alpha}\int_0^t X^*(\omega)\ \mathrm{d}\omega < \frac{1}{1-\alpha}\int_0^{(1-\alpha)} X^*(\omega)\ \mathrm{d}\omega = \mathrm{CVar}_\alpha(|X|),$$

since X^* is nonnegative. For $\alpha \to 1$, the Marcinkiewicz and the Lorentz norm both coincide with the $\mathcal{L}^\infty$ norm. If $\alpha \to 1$, then $\phi(t) = \chi_{(0,1]} \notin \Phi_{0+}$.

$$\|X\|_{\Lambda_\phi} = \int_0^1 X^*(\omega)\phi'(\omega)\ \mathrm{d}\omega + X^*(0)\phi(0+) = X^*(0) = \|X\|_{\mathcal{L}^\infty}$$
$$\|X\|_{M_\phi} = \sup_{0<t\leqslant 1}\left\{1\cdot\frac{1}{t}\int_0^t X^*(\omega)\ \mathrm{d}\omega\right\} = X^*(0) = \|X\|_{\mathcal{L}^\infty}.$$

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 $\phi(t) = \phi_\infty(t) = \lim_{\alpha\to 1}\min\{t/(1-\alpha), 1\}$. 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.

$$\forall t \in (0,1] : \phi_t(x) := \begin{cases}\phi(t)\frac{x}{t} & ,\ x \leqslant t\\ \phi(t) & ,\ x > t.\end{cases}$$

If the norm is PTE, we can reduce it (Theorem 11) 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 $\{t \mapsto \int_0^t Y^*(\omega)\ \mathrm{d}\omega : \|Y\|_{M'_\phi \leqslant 1}\}$, the proof of Theorem 11 is agnostic to the used Kusuoka representation; uniqueness is not assumed. But then, these $\phi_{t'}$ by their definition are

$$\phi_{t'}(x) := \begin{cases}\frac{x}{t'} & ,\ x \leqslant t'\\ 1 & ,\ x > t'\end{cases} = \min\{x/(1-\alpha), 1\}, \quad \alpha = 1-t'.$$

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 $\alpha \to 1$, for which an uncountable infinity of t' is needed, i.e. $\phi_{t'}(x) = \sup_{t'\to 0}\min\{\frac{x}{t'}, 1\}$, thus $\phi = \chi_{(0,1]}$. ■

A.3.3 Variational representation of $\mathrm{RIM}_{\alpha,\beta}$

Recall that (Theorem 22):

$$\mathrm{RIM}_{\alpha,\beta}(X) := \beta\mathbb{E}[X] + (1-\beta)\,\mathrm{CVar}_\alpha(X).$$

We show that $\mathrm{RIM}_{\alpha,\beta}$ admits the following variational representation:

$$\mathrm{RIM}_{\alpha,\beta}(X) = \inf_{\mu\in\mathbb{R}}\mu + \mathbb{E}v(X-\mu) \quad \forall X \in \mathcal{M},$$

where the regret function v is given by the piecewise linear

$$v(t) = \begin{cases} \beta t & t \leqslant 0 \\ \frac{\beta\alpha-1}{\alpha-1} t & t > 0. \end{cases}$$

Proof We here translate a result from (Pflug and Ruszczynski, 2001) to a loss-based formulation. Let $0 < \lambda_1 < 1 < \lambda_2$. Consider the function:

$$\begin{aligned} f_\mu(x) &= \mu + \lambda_2(x-\mu)^+ - \lambda_1(x-\mu)^- \\ &= \mu + (\lambda_2 - \lambda_1)(x-\mu)^+ + \lambda_1(x-\mu) \\ &= \mu + (\lambda_2 - \lambda_1)(x-\mu)^+ + \lambda_1 x - \lambda_1\mu \\ &= (\lambda_2 - \lambda_1)(x-\mu)^+ + \lambda_1 x + (1-\lambda_1)\mu \\ &= \lambda_1 x + (1-\lambda_1)\left(\mu + \frac{\lambda_2-\lambda_1}{1-\lambda_1}(x-\mu)^+\right). \end{aligned}$$

Now,

$$\begin{aligned} \mathrm{RIM}_{\alpha,\beta}(X) &= \inf_{\mu\in\mathbb{R}} \mathbb{E}[f_\mu(X)] \\ &= \lambda_1\mathbb{E}[X] + (1-\lambda_1)\inf_{\mu\in\mathbb{R}}\left(\mu + \frac{\lambda_2-\lambda_1}{1-\lambda_1}\mathbb{E}(X-\mu)^+\right) \\ &= \beta\mathbb{E}[X] + (1-\beta)\,\mathrm{CVar}_\alpha(X), \end{aligned}$$

where $\lambda_1 = \beta$ and $\frac{1}{1-\alpha} = \frac{\lambda_2-\lambda_1}{1-\lambda_1}$, hence $\lambda_2 = \frac{\beta\alpha-1}{\alpha-1}$. ■

A.4 Norm Equivalences and Tail Risk

A.4.1 Proof of Theorem 24

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 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

$$K := \sup_{0<\alpha\leqslant 1} \frac{\phi_2(\alpha)}{\phi_1(\alpha)}.$$

Thus $\forall\alpha \in (0,1]$: $\phi_2(\alpha) \leqslant K \cdot \phi_1(\alpha)$. Then it directly follows from Hardy's lemma that

$$\|X\|_{\mathcal{R}_2} = \int_0^1 X^*(\omega)\phi_2'(\omega)\,\mathrm{d}\omega \leqslant K \cdot \int_0^1 X^*(\omega)\phi_1'(\omega)\,\mathrm{d}\omega = K \cdot \|X\|_{\mathcal{R}_1}.$$

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 (17) is equal to (Pichler, 2013)

$$C = \inf\{c > 0 : \forall \phi_{2\zeta} \in \mathcal{Z}_2 : \exists \phi_{1\gamma} \in \mathcal{Z}_1 : \forall \alpha \in (0,1] : \phi_{2\zeta}(\alpha) \leqslant C \cdot \phi_{1\gamma}(\alpha)\},$$

which ensures that $\forall \epsilon > 0 : \forall \phi_{2\zeta} \in \mathcal{Z}_2 : \exists \phi_{1\gamma} \in \mathcal{Z}_1 : \forall \alpha \in (0,1]$: $\phi_{2\zeta}(\alpha) \leqslant (C+\epsilon) \cdot \phi_{1\gamma}(\alpha)$. But then,

$$\begin{aligned}\|X\|_{\mathcal{R}_2} &= \sup\left\{\int_0^1 X^*(\omega)\phi'_{2\zeta}(\omega)\ \mathrm{d}\omega : \phi_{2\zeta} \in \mathcal{Z}_2\right\} \\ &\leqslant (C+\epsilon) \cdot \sup\left\{\int_0^1 X^*(\omega)\phi'_{1\gamma}(\omega)\ \mathrm{d}\omega : \phi_{1\gamma} \in \mathcal{Z}_1\right\} = (C+\epsilon) \cdot \|X\|_{\mathcal{R}_1},\end{aligned}$$

where we applied Hardy's lemma to each of the replacements of $\phi_{2\zeta} \to \phi_{1\gamma}$ such that the above inequality holds. Let $\epsilon \downarrow 0$ to obtain $\|X\|_{\mathcal{R}_2} \leqslant C \cdot \|X\|_{\mathcal{R}_1}$. ■

The converse direction: Pichler (2013) stated that $C = \infty$ (17) implies non-equivalence of the norms. However, no proof was provided for the statement. In Theorem 31 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 $\forall \phi_1 : \phi_1(\alpha) \geqslant \alpha$, 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 $\lim_{\alpha \to 0}$ 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.

A.4.2 Proof of Theorem 26

Proof Denote by ϕ_t the family of functions

$$\forall t \in (0,1] : \phi_t(x) := \begin{cases} \phi(t)\frac{x}{t} & , x \leqslant t \\ \phi(t) & , x > t \end{cases}$$

which generate a Kusuoka set of the Marcinkiewicz norm with fundamental function ϕ. Then the problem of finding a C as in Theorem 24 reduces to

$$C = \inf_{\phi_t} \sup_{0<\alpha\leqslant 1} \frac{\phi(\alpha)}{\phi_t(\alpha)} = \inf_{\phi_t} \sup_{0<\alpha\leqslant 1} \begin{cases} \frac{\phi(\alpha)}{\alpha}\frac{t}{\phi(t)} & , \alpha \leqslant t \\ \frac{\phi(\alpha)}{\phi(t)} & , \alpha > t. \end{cases}$$

Fix any t. Then certainly

$$C \leqslant \sup_{0<\alpha\leqslant 1} \frac{\phi(\alpha)}{\phi_t(\alpha)} = \sup_{0<\alpha\leqslant 1} \begin{cases} \frac{\phi(\alpha)}{\alpha}\frac{t}{\phi(t)} & , \alpha \leqslant t \\ \frac{\phi(\alpha)}{\phi(t)} & , \alpha > t. \end{cases}$$

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 ($\alpha \leqslant t$), since $\frac{\phi(\alpha)}{\alpha}$ is decreasing in α (due to the quasiconcavity

of ϕ), the supremum must occur[28] for $\alpha \to 0$, or otherwise it occurs in the second term. Observe that $\phi'(0) = \lim_{\alpha\to 0} \frac{\phi(\alpha)}{\alpha}$ by definition of the difference quotient since $\phi(0) = 0$. Therefore:

$$\lim_{\alpha\to 0} \frac{\phi(\alpha)}{\alpha} \frac{t}{\phi(t)} = \frac{\phi'(0)}{1} \frac{t}{\phi(t)} < \infty.$$

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

$$\phi_t(\alpha) = \begin{cases} \frac{\phi(1/\phi'(0))}{1/\phi'(0)}\alpha & , \ \alpha \leqslant t \\ \phi(1/\phi'(0)) & , \ \alpha > t. \end{cases}$$

Then

$$\phi(\alpha) \leqslant K \cdot \phi_{1/\phi'(0)} = \begin{cases} \phi'(0) \cdot \alpha & , \ \alpha \leqslant 1/\phi'(0) \\ 1 & , \ \alpha > 1/\phi'(0). \end{cases}$$

To see that this holds, observe that $\forall \alpha \in (0,1]$ $\frac{\phi(\alpha)}{\alpha} \leqslant \phi'(0) = \lim_{t\to 0} \frac{\phi(t)}{t}$ due to quasiconcavity and also $\phi(\alpha) \leqslant 1$. Indeed K is the smallest constant such that the statement $\exists t' \in (0,1] : \forall \alpha \in (0,1]$ $\phi(\alpha) \leqslant K' \cdot \phi_{t'}(\alpha)$ holds. Assume $t' < t = 1/\phi'(0)$. For $\alpha > t'$ we require $K' \cdot \phi(t') \geqslant 1$, 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 $(K' \cdot \phi_{t'})'(0) \geqslant \phi'(0)$ for the majorization to hold in the limit as $\alpha \to 0$ (this comes from letting $\alpha \to 0$ in the condition $K' \cdot \frac{\phi(t')}{t'}\alpha \geqslant \phi(\alpha)$). 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 $(K' \cdot \phi_{t'})'(0) \geqslant \phi'(0)$ implies $K' > K$.

However, we note that K need not be the smallest constant such that $\|X\|_{\Lambda_\phi} \leqslant K \cdot \|X\|_{M_\phi}$ holds, but it is the smallest constant so that $\exists t' \in (0,1] : \forall \alpha \in (0,1]$ $\phi(\alpha) \leqslant K' \cdot \phi_{t'}(\alpha)$, which guarantees the previous statement to hold. ■

A.4.3 Proof of Theorem 28

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.

28. When writing *occur*, we do not mean to imply that a supremum is actually attained.
29. Note 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.

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:

$$C = \inf \{c > 0 : \exists \phi_t : \forall \alpha \in (0,1] : \phi(\alpha) \leqslant c \cdot \phi_t(\alpha)\} < \infty.$$

Then $\forall \epsilon > 0 : \exists t : \forall \alpha \in (0,1] : \phi(\alpha) \leqslant (C+\epsilon) \cdot \phi_t(\alpha)$. Fix some ϵ. With this choice of t,

$$1 \geqslant \sup_{0<\alpha\leqslant 1} \begin{cases} \frac{\phi(\alpha)}{\alpha} \frac{t}{(C+\epsilon)\cdot\phi(t)} & , \ \alpha \leqslant t \\ \frac{\phi(\alpha)}{(C+\epsilon)\cdot\phi(t)} & , \ \alpha > t. \end{cases}$$

If the supremum occurs in the first term, it occurs as $\alpha \to 0$ due to the quasiconcavity of ϕ. Recall that $\phi'(0) = \lim_{\alpha\to 0} \frac{\phi(\alpha)}{\alpha}$ by definition of the difference quotient. Therefore $1 \geqslant \sup_{0<\alpha\leqslant t} \frac{\phi(\alpha)}{\alpha} \frac{t}{(C+\epsilon)\cdot\phi(t)} = \lim_{\alpha\to 0} \frac{\phi(\alpha)}{\alpha} \frac{t}{(C+\epsilon)\cdot\phi(t)} = \phi'(0) \frac{t}{(C+\epsilon)\cdot\phi(t)} = \infty$, a contradiction. Thus no finite C exists. Finally, the inequality $\|X\|_{M_\phi} \leqslant \|X\|_{\Lambda_\phi}$ stems from the embedding theorem.

■

A.4.4 Proof of Theorem 29

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:

$$C = \sup_{\phi_{2\zeta}} \sup_{0<\alpha\leqslant 1} \frac{\phi_{2\zeta}(\alpha)}{\alpha}.$$

Due to the concavity of $\phi_{2\zeta}$, the fraction is decreasing in α and hence

$$C = \sup_{\phi_{2\zeta}} \lim_{\alpha\to 0} \frac{\phi_{2\zeta}(\alpha)}{\alpha} = \sup_{\phi_{2\zeta}} \phi'_{2\zeta}(0).$$

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

$$\phi_2(\alpha) = \sup_{\mathcal{Z}_2} \{\phi_{2\zeta}(\alpha)\}.$$

We know that

$$\frac{\phi_2(t)}{t} \geqslant \frac{\phi_{2\zeta}(t)}{t} \quad \forall t \quad \forall \phi_{2\zeta},$$

since we know the limit of the left hand side exists, we have

$$\infty > \phi'_2(0) = \lim_{h\to 0} \frac{\phi_2(t)}{t} \geqslant \lim_{h\to 0} \frac{\phi_{2\zeta}(t)}{t} = \phi'_{2\zeta}(0) \quad \forall \phi_{2\zeta}.$$

Therefore C is finite and we obtain $\|X\|_{\mathcal{L}^1} \leqslant \|X\|_{\mathcal{R}_1} \leqslant C \cdot \|X\|_{\mathcal{L}^1} \quad \forall X \in \mathcal{L}^1$. As a consequence, any two ri norms with finite $\phi'_1(0), \phi'_2(0)$ are equivalent. ■

A.4.5 Proof of Theorem 31

Proof First we show that $C = \infty$. Denote by $\phi_{TM,t}, \phi_{M,t}$ the respective Kusuoka sets, constructed as:

$$\forall t \in (0,1] : \phi_{TM,t}(x) := \begin{cases} \phi(t)\frac{x}{t} & , x \leqslant t \\ \frac{1-\phi(t)}{1-t}x + \frac{\phi(t)-t}{1-t} & , x > t \end{cases}, \quad \phi_{M,t}(x) := \begin{cases} \phi(t)\frac{x}{t} & , x \leqslant t \\ \phi(t) & , x > t. \end{cases}$$

The desired constant is

$$\begin{aligned} C &= \sup_{\phi_{TM,t}} \inf_{\phi_{M,t'}} \sup_{0<\alpha\leqslant 1} \frac{\phi_{TM,t}(\alpha)}{\phi_{M,t'}(\alpha)} \\ &= \inf \left\{ c > 0 : \forall \phi_{TM,t} \exists \phi_{M,t'} : \forall \alpha \in (0,1] : \phi_{TM,t}(\alpha) \leqslant c \cdot \phi_{M,t'}(\alpha) \right\}. \end{aligned}$$

We show that no such finite constant can exist. Let some t be given. We wish to find t' such that $\forall \alpha \in (0,1] : \phi_{TM,t}(\alpha) \leqslant K \cdot \phi_{M,t'}(\alpha)$. The argument in Theorem 26 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 $\phi_{TM,t}$

$$\frac{1}{K} = \frac{1-\phi(t)}{1-t}\frac{t}{\phi(t)} + \frac{\phi(t)-t}{1-t}.$$

For each fixed t, this is the best feasible constant (Theorem 26) and it is finite. However, as $t \to 0$, we find that $\lim_{t\to 0} K = \frac{\phi(t)}{t} = \phi'(0) = \infty$. As $t \to 0$, 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

$$\begin{aligned} \|X\|_{TM_\phi} &= \sup_{0<t<1} \left\{ \frac{\phi(t)}{t} \int_0^t X^*(\omega)\ \mathrm{d}\omega + \frac{\phi(t)-1}{t-1} \int_t^1 X^*(\omega)\ \mathrm{d}\omega \right\} \\ &\leqslant \|X\|_{M_\phi} + \sup_{0<t<1} \left\{ \frac{\phi(t)-1}{t-1} \int_t^1 X^*(\omega)\ \mathrm{d}\omega \right\}. \end{aligned} \tag{26}$$

Knowing that $\|X\|_{M_\phi} \leqslant \|X\|_{TM_\phi}$, 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 (26) behaves nicely: $\int_t^1 X^*(\omega)\ \mathrm{d}\omega \leqslant \int_0^1 X^*(\omega)\ \mathrm{d}\omega < \infty$ 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 $t \to 0$ and $t \to 1$:

$$\lim_{t\to 0} \frac{\phi(t)-1}{t-1} = 1, \quad \lim_{t\to 1} \frac{\phi(t)-1}{t-1} = \phi'(1) < \infty.$$

As $t \to 0$, we can apply the quotient rule, since both limits exist. As $t \to 1$, 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). ■

A.5 Properties of Quasiconcave Functions

A.5.1 Proof of Lemma 33

Proof By quasiconcavity, and assumption, we have $\phi(0) = 0$, $\phi(1) = 1$ and

$$0 \leqslant t_0 \leqslant t_1 \quad \Rightarrow \quad \phi(t_0) \leqslant \phi(t_1) \text{ and } \phi(t_0)/t_0 \geqslant \phi(t_1)/t_1.$$

Taking $t_1 = 1$ we have $\phi(t_0) \leqslant \phi(t_1) = 1$. Taking $1 = t_0 \leqslant t_1$ implies $1 = \phi(1) \leqslant \phi(t_1)$. Furthermore, $t_0 \leqslant t_1 = 1$ implies $\phi(t_0)/t_0 \geqslant 1$ which implies $\phi(t_0) \geqslant t_0$. Additionally, $1 = t_0 \leqslant t_1$ implies $1 \geqslant \phi(t_1)/t_1$ and thus $\phi(t_1) \leqslant t_1$. Combining all these facts we have shown

$$\begin{aligned} t \leqslant 1 \quad &\Rightarrow \quad t \leqslant \phi(t) \leqslant 1 \\ t \geqslant 1 \quad &\Rightarrow \quad 1 \leqslant \phi(t) \leqslant t. \end{aligned}$$

For $0 \leqslant t \leqslant 1$, $t = 1 \wedge t$ and $1 = 1 \vee t$. For $t \geqslant 1$, $1 = 1 \wedge t$ and $t = 1 \vee t$. Hence Lemma 33 holds. ■

A.5.2 Proof of Lemma 36

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 $t_0 \leqslant t_1$. Let $i^* = \arg\max_i \phi_i(t_1)/t_1$. Then

$$\frac{\bigvee_{i\in[n]} \phi_i(t_0)}{t_0} \geqslant \frac{\phi_{i^*}(t_0)}{t_0} \geqslant \frac{\phi_{i^*}(t_1)}{t_1} = \frac{\bigvee_{i\in[n]} \phi_i(t_1)}{t_1},$$

and thus $t \mapsto \left(\bigvee_{i\in[n]} \phi_i(t)\right)/t$ is decreasing. A similar argument holds for $\bigwedge_i \phi_i$. ■

A.5.3 Proof of Lemma 38

Proof (If): For $\alpha > 0$, $\breve{\psi}(\alpha x, \alpha y) = \alpha y \psi(\alpha x/(\alpha y)) = \alpha \breve{\psi}(x, y)$, and thus $\breve{\psi}$ is positively homogeneous. Furthermore, $\breve{\psi}(x, y)$ is nondecreasing in each argument as we now show. Let $y \in \mathbb{R}_{\geqslant 0}$ be arbitrary but fixed and consider $x \mapsto \breve{\psi}(x, y) = y\psi(x/y)$. This is nondecreasing since ψ is nondecreasing. Now let $x \in \mathbb{R}_{\geqslant 0}$ 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 $\breve{\psi}$ is nondecreasing in both of its arguments concluding the demonstration that $\breve{\psi} \in \mathscr{P}$.

(Only if): If $x \leqslant y$ then $\psi(x) = \breve{\psi}(x, 1) \leqslant \breve{\psi}(y, 1) = \psi(y)$ and thus ψ is nondecreasing. Since $\breve{\psi}$ is positively homogeneous and nonzero, $\psi(s) \neq 0$ unless $s = 0$. Finally, for $x \leqslant y$,

$$\frac{\psi(x)}{x} = \frac{\breve{\psi}(x, 1)}{x} = \breve{\psi}\left(1, \frac{1}{x}\right) \geqslant \breve{\psi}\left(1, \frac{1}{y}\right) = \frac{\breve{\psi}(y, 1)}{y} = \frac{\psi(y)}{y},$$

which shows that $x \mapsto \psi(x)/x$ is nonincreasing, demonstrating that ψ is quasiconcave. ■

A.5.4 Proof of Lemma 39

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 $\breve{\psi}$ is nondecreasing in each argument, then $f_{\phi_0,\phi_1}(t)$ is nondecreasing in t. Finally, we need to show that

$$t_0 \leqslant t_1 \;\Rightarrow\; \frac{\breve{\psi}(\phi_0(t_0), \phi_1(t_0))}{t_0} \geqslant \frac{\breve{\psi}(\phi_0(t_1), \phi_1(t_1))}{t_1}. \tag{27}$$

Since $\breve{\psi}$ is positively homogeneous, (27) is equivalent to

$$t_0 \leqslant t_1 \;\Rightarrow\; \breve{\psi}\left(\frac{\phi_0(t_0)}{t_0}, \frac{\phi_1(t_0)}{t_0}\right) \;\geqslant\; \breve{\psi}\left(\frac{\phi_0(t_1)}{t_1}, \frac{\phi_1(t_1)}{t_1}\right).$$

But by assumption on ϕ_0 and ϕ_1 we have

$$t_0 \leqslant t_1 \;\Rightarrow\; \frac{\phi_0(t_0)}{t_0} \geqslant \frac{\phi_0(t_1)}{t_1} \quad \text{and} \quad \frac{\phi_1(t_0)}{t_0} \geqslant \frac{\phi_1(t_1)}{t_1}.$$

Since $\breve{\psi}$ is nondecreasing in each argument, we can thus conclude that (27) holds, thus demonstrating the final property needed to show quasiconcavity of f_{ϕ_0,ϕ_1}.

(Only if): Using the definition of $\breve{\psi}$ 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 $[\forall \phi_0, \phi_1 \in \mathscr{Q},\ f_{\phi_0,\phi_1} \in \mathscr{Q}] \Rightarrow \psi \in \mathscr{Q}$. Now $f_{\phi_0,\phi_1} \in \mathscr{Q}$ means that

1. $t_0 \leqslant t_1 \Rightarrow \phi_1(t_0)\psi\left(\frac{\phi_0(t_0)}{\phi_1(t_0)}\right) \leqslant \phi_1(t_1)\psi\left(\frac{\phi_0(t_1)}{\phi_1(t_1)}\right)$.
2. $t_0 \leqslant t_1 \Rightarrow \frac{\phi_1(t_0)}{t_0}\psi\left(\frac{\phi_0(t)0)}{\phi_1(t_0)}\right) \geqslant \frac{\phi_1(t_1)}{t_1}\psi\left(\frac{\phi_0(t_1)}{\phi_1(t_1)}\right)$.
3. $f_{\phi_0,\phi_1}(t) = 0 \Leftrightarrow t = 0$.

Choose $\phi_1(t) = t$ and $\phi_0 \in \mathscr{Q}$. Then condition 2 above requires that

$$t_0 \leqslant t_1 \Rightarrow \psi\left(\frac{\phi_0(t_0)}{t_0}\right) \geqslant \psi\left(\frac{\phi_0(t_1)}{t_1}\right).$$

But since $\phi_0 \in \mathscr{Q}$, $t_0 \leqslant t_1 \Rightarrow \frac{\phi_0(t_0)}{t_0} \geqslant \frac{\phi_0(t_1)}{t_1}$ and thus ψ must be nondecreasing.

Now choose $\phi_1(t) = 1$ and $\phi_0(t) = t$. Then the second condition implies

$$t_0 \leqslant t_1 \Rightarrow \frac{\psi(t_0)}{t_0} \geqslant \frac{\psi(t_1)}{t_1}.$$

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 quasiconcavity for ψ. ■

A.6 Interpolation Functors and their Fundamental Functions

A.6.1 Proof of Lemma 45

Proof Recalling the definition of PTE (10), we show that for any $X \in \mathcal{M}^+$ and $c \in \mathbb{R}$ such that $X + c \geqslant 0$ 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 $c \geqslant 0$. 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

$$\begin{aligned}
\|X\|_\Lambda + c &= \inf\left\{\sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) + c \colon X_k \in \mathcal{X}_0 + \mathcal{X}_1,\ K \in \mathbb{N},\ X = \sum_k X_k\right\} \\
&= \inf\left\{\sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) + \phi(c, c) \colon X_k \in \mathcal{X}_0 + \mathcal{X}_1,\ K \in \mathbb{N},\ X = \sum_k X_k\right\}
\end{aligned}$$

since $\phi(1,1) = 1$ and ϕ is positively homogeneous,

$$\begin{aligned}
&= \inf\left\{\sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) + \sum_k \phi\left(\tfrac{c}{K}, \tfrac{c}{K}\right) : X_k \in \mathcal{X}_0 + \mathcal{X}_1,\ K \in \mathbb{N},\ X = \sum_k X_k\right\} \\
&= \inf\left\{\sum_k^K \left[\phi(\|X_k\|_0, \|X_k\|_1) + \phi\left(\tfrac{c}{K}, \tfrac{c}{K}\right)\right] : X_k \in \mathcal{X}_0 + \mathcal{X}_1,\ K \in \mathbb{N},\ X = \sum_k X_k\right\} \\
&\leqslant \inf\left\{\sum_k^K \phi(\|X_k\|_0 + \tfrac{c}{K}, \|X_k\|_1 + \tfrac{c}{K}) \colon X_k \in \mathcal{X}_0 + \mathcal{X}_1,\ K \in \mathbb{N},\ X = \sum_k X_k\right\}
\end{aligned}$$

since $\phi(z_1) + \phi(z_2) \leqslant \phi(z_1 + z_2)$ for arbitrary $z_1, z_2 \in \mathbb{R}_+^2$ because ϕ is a concave gauge function (Barbara and Crouzeix, 1994, Proposition 2.1)

$$= \inf\left\{\sum_k^K \phi(\|X_k + \tfrac{c}{K}\|_0, \|X_k + \tfrac{c}{K}\|_1) \colon X_k \in \mathcal{X}_0 + \mathcal{X}_1,\ K \in \mathbb{N},\ X = \sum_k X_k\right\}$$

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

$$\begin{aligned}
&= \inf\left\{\sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) \colon X'_k \in \mathcal{X}_0 + \mathcal{X}_1,\ K \in \mathbb{N},\ X = \sum_k (X'_k - \tfrac{c}{K})\right\} \\
&= \inf\left\{\sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) \colon X'_k \in \mathcal{X}_0 + \mathcal{X}_1,\ K \in \mathbb{N},\ X = \left(\sum_k X'_k\right) - c\right\} \\
&= \inf\left\{\sum_k^K \phi(\|X_k\|_0, \|X_k\|_1) \colon X'_k \in \mathcal{X}_0 + \mathcal{X}_1,\ K \in \mathbb{N},\ X + c = \left(\sum_k X'_k\right)\right\} \\
&= \|X + c\|_\Lambda.
\end{aligned}$$

Since $\|\cdot\|_\Lambda$ is a norm it satisfies the triangle inequality $\|X + c\|_\Lambda \leqslant \|X\|_\Lambda + \|c\|_\Lambda = \|X\|_\Lambda + c$. Thus combining with the above we have

$$\|X\|_\Lambda + c \leqslant \|X + c\|_\Lambda \leqslant \|X\|_\Lambda + c$$

and thus $\|X+c\|_\Lambda = \|X\|_\Lambda + c$ as required.

If instead we have $c<0$ but $X+c \geqslant 0$, then there exists some $c_0 \geqslant 0$ such that $c_0 \geqslant -c$ and $X = X_0 + c_0$, with $X_0 \in \mathcal{M}_+$. Consequently,

$$\|X+c\|_\Lambda = \|X_0+c_0+c\|_\Lambda = \|X_0+(c_0+c)\|_\Lambda \overset{(*)}{=} \|X_0\|_\Lambda + (c_0+c) \overset{(**)}{=} \|X_0+c_0\|_\Lambda + c = \|X\|_\Lambda + c,$$

where (*) holds from the case already shown, since $X_0 \geqslant 0$ and $(c_0+c) \geqslant 0$, and (**) holds similarly (since $c_0 \geqslant 0$). ■

A.6.2 Proof of Lemma 47

Proof For $t>0$, we need to compute

$$\phi_{M_{\bar\phi}(\bar{\mathcal{X}})}(t) = \|\chi_{[0,t]}\|_{M_{\bar\phi}(\bar{\mathcal{X}})} = \sup_{s_0,s_1\geqslant 0} \frac{K(s_0,s_1,\chi_{[0,t]},\bar{\mathcal{X}})}{\bar\phi^*(s_0,s_1)}.$$

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 9, the Marcinkiewicz norm minorises any ri norm with fundamental function $\phi_{\mathcal{X}}$: $\|X_i\|_{M_{\bar\phi}} \leqslant \|X\|_{\mathcal{X}}$, we have

$$K(s_0,s_1,\chi_{E_t},\bar{\mathcal{X}}) \geqslant \inf_{X_0+X_1=\chi_{E_t}} s_0\|X_0\|_{M_{\phi_0}} + s_1\|X_1\|_{M_{\phi_1}},$$

and since $\|X\|_{M_{\bar\phi}} = \sup_{0<r<\infty} X^{**}(r)\bar\phi(r)$, we have

$$\begin{aligned} K(s_0,s_1,\chi_{E_t},\bar{\mathcal{X}}) &\geqslant \inf_{X_0+X_1=\chi_{E_t}} s_0\left(\sup_{0<r<\infty} X_0^{**}(r)\phi_0(r)\right) + s_1\left(\sup_{0<r<\infty} X_1^{**}(r)\phi_1(r)\right) \\ &\geqslant \inf_{X_0+X_1=\chi_{E_t}} \sup_{0<r<\infty} (s_0\phi_0(r)X_0^{**}(r) + s_1\phi_1(r)X_1^{**}(r)) \end{aligned}$$

and by choosing $r=t$ we obtain

$$K(s_0,s_1,\chi_{E_t},\bar{\mathcal{X}}) \geqslant \inf_{X_0+X_1=\chi_{E_t}} (c_0X_0^{**}(t) + c_1X_1^{**}(t)),$$

where $c_0 = s_0\phi_0(t)$ and $c_1 = s_1\phi_1(t)$ are constants (since t is fixed),

$$= \inf_{X_0+X_1=\chi_{E_t}} \left((c_0X_0)^{**}(t) + (c_1X_1)^{**}(t)\right).$$

But $(f+g)^{**}(t) \leqslant f^{**}(t) + g^{**}(t)$ for all $t>0$ and any f,g, and so

$$K(s_0,s_1,\chi_{E_t},\bar{X}) \geqslant \inf_{X_0+X_1=\chi_{E_t}} (c_0X_0 + c_1X_1)^{**}(t)).$$

Since for any f we have $f^*(t) \leqslant f^{**}(t)$, for all t, we have

$$\begin{aligned} K(s_0, s_1, \chi_{E_t}, \bar{\mathcal{X}}) &\geqslant \inf_{X_0+X_1=\chi_{E_t}} (c_0X_1 + c_1X_1)^*(t) \\ &= \inf_{X_0+X_1=\chi_{E_t}} \inf\{\lambda\colon \mu_{c_0X_0+c_1X_1}(\lambda) \leqslant t\} \\ &= \inf_{X_0+X_1=\chi_{E_t}} \inf\{\lambda\colon \mu\{s \in \mathbb{R}\colon (c_0X_0 + c_1X_1)(s) > \lambda\} \leqslant t\}. \end{aligned}$$

Now let $A_\lambda := \mu\{s \in \mathbb{R}\colon (c_0X_0 + c_1X_1)(s) > \lambda\}$ and $B_\lambda := \mu\{s \in \mathbb{R}\colon (c_0 \wedge c_1)(X_0 + X_1)(s) > \lambda\}$. Since $(c_0 \wedge c_1)(X_0 + X_1) = (c_0 \wedge c_1)X_0 + (c_0 \wedge c_1)X_1 \leqslant c_0X_0 + c_1X_1$ we have that $B_\lambda \leqslant A_\lambda$ for all λ. Furthermore, $\lambda \mapsto A_\lambda$ and $\lambda \mapsto B_\lambda$ are nonincreasing and thus $\inf\{\lambda\colon A_\lambda \leqslant t\} \geqslant \inf\{\lambda\colon B_\lambda \leqslant t\}$, and hence

$$\begin{aligned} K(s_0, s_1, \chi_{E_t}, \bar{\mathcal{X}}) &\geqslant \inf_{X_0+X_1=\chi_{E_t}} \inf\{\lambda\colon \mu\{s \in \mathbb{R}\colon (c_0 \wedge c_1)(X_0 + X_1)(s) > \lambda\} \leqslant t\} \\ &= \inf_{X_0+X_1=\chi_{E_t}} (c_0 \wedge c_1)(X_0 + X_1)^*(t) \\ &= (c_0 \wedge c_1)\chi_{[0,t]}(t) \\ &= c_0 \wedge c_1 \\ &= s_0\phi_0(t) \wedge s_1\phi_1(t). \end{aligned}$$

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

$$\|\chi_{[0,t]}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} = \sup_{s_0,s_1\geqslant 0} \frac{s_0\phi_0(t) \wedge s_1\phi_1(t)}{\bar{\phi}^*(s_0, s_1)}.$$

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

$$\begin{aligned} s\phi_0(t) &\quad \text{if } s\phi_0(t) \leqslant (1-s)\phi_1(t) \\ (1-s)\phi_1(t) &\quad \text{if } s\phi_0(t) \geqslant (1-s)\phi_1(t). \end{aligned}$$

Setting $\gamma := \phi_1(t)/\phi_0(t)$, we have $s\phi_0(t) \leqslant (1-s)\phi_1(t) \Leftrightarrow s/(1-s) \leqslant \gamma \Leftrightarrow s \leqslant \gamma/(1+\gamma)$. Hence

$$s\phi_0(t) \wedge (1-s)\phi_1(t) = \begin{cases} s\phi_0(t), & s \leqslant \frac{\gamma}{1+\gamma} \\ (1-s)\phi_1(t), & s \geqslant \frac{\gamma}{1+\gamma} \end{cases}.$$

Hence

$$\phi_{M_{\bar{\phi}}(\bar{\mathcal{X}})}(t) = \|\chi_{[0,t]}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} = \min\left(\sup_{s \leqslant \frac{\gamma}{1+\gamma}} \frac{s\phi_0(t)}{(1-s)\bar{\psi}\left(\frac{s}{1-s}\right)}, \sup_{s \geqslant \frac{\gamma}{1+\gamma}} \frac{(1-s)\phi_1(t)}{(1-s)\bar{\psi}\left(\frac{s}{1-s}\right)} \right).$$

Since $\phi_0(t), \phi_1(t) \geqslant 0$, we only need to determine

$$a := \sup_{0 \leqslant s \leqslant \frac{\gamma}{1+\gamma}} f(s) \quad \text{and} \quad b := \sup_{\frac{\gamma}{1+\gamma} \leqslant s < \infty} g(s),$$

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 $a = \gamma/\psi(\gamma)$. Similarly for g, we have

$$b = \sup_{t \geqslant \gamma} \frac{1}{\bar{\psi}(t)}.$$

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

$$\begin{aligned} \|\chi_{[0,t]}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} &= \min\left(\frac{\gamma\phi_0(t)}{\bar{\psi}(\gamma)}, \frac{\phi_1(t)}{\bar{\psi}(\gamma)} \right) = \min\left(\frac{\phi_1(t)}{\bar{\psi}(\gamma)}, \frac{\phi_1(t)}{\bar{\psi}(\gamma)} \right) \\ &= \frac{\phi_1(t)}{\bar{\psi}(\phi_1(t)/\phi_0(t))} = \frac{\phi_1(t)\phi_0(t)}{\bar{\phi}^*(\phi_1(t), \phi_0(t))}. \end{aligned}$$

But $\bar{\phi}^*$ is positively homogeneous, and thus

$$\begin{aligned} \|\chi_{[0,t]}\|_{M_{\bar{\phi}}(\bar{\mathcal{X}})} &= \frac{\phi_1(t)\phi_0(t)}{\bar{\phi}^*(\phi_1(t), \phi_0(t))} = \phi_1(t)\phi_0(t)\bar{\phi}\left(\frac{1}{\phi_1(t)}, \frac{1}{\phi_0(t)} \right) \\ &= \bar{\phi}\left(\frac{\phi_1(t)\phi_0(t)}{\phi_1(t)}, \frac{\phi_1(t)\phi_0(t)}{\phi_0(t)} \right) = \bar{\phi}(\phi_0(t), \phi_1(t)). \end{aligned}$$

■

A.6.3 Proof of Lemma 48

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 $\mu(E_t) = t$,

$$= \inf_{(X_k)_k \colon \chi_{E_t} = \sum_k X_k} \sum_k \bar{\phi}(\|X_k\|_{\mathcal{X}_0}, \|X_k\|_{\mathcal{X}_1}).$$

Taking the particular choice $X_1 = \chi_{[0,t]}$ and $X_k = 0$ for $k > 1$, gives an upper bound on the infimum:

$$\begin{aligned}\phi_{\Lambda_{\bar{\phi}}(\bar{\mathcal{X}})}(t) &\leqslant \bar{\phi}(\|\chi_{E_t}\|_{\mathcal{X}_0}, \|\chi_{E_t}\|_{\mathcal{X}_1}) \\ &= \bar{\phi}(\phi_0(t), \phi_1(t)).\end{aligned}$$

■

Appendix B. (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 view[30].

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.

B.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.

$$\mathcal{P} = \{P : \mathcal{F} \to [0,1] : P(\{c\}) \neq 0 \text{ only for finitely many } c \in C, P(C) = 1\}.$$

An element $P \in \mathcal{P}$ is called a *lottery*. We characterize a decision maker by her preference relation $\succcurlyeq$ on $\mathcal{P}$. The meaning of $X \succcurlyeq Y$ 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 $\succcurlyeq$ is represented by a functional $R : \mathcal{P} \to \mathbb{R}$ if

$$P \succcurlyeq Q \iff R(P) \leqslant R(Q),$$

Assume that $\succcurlyeq$ satisfies the following axioms:

N1. $\forall P, Q \in \mathcal{P} : P \succcurlyeq Q$ or $Q \succcurlyeq P$ or both. (completeness)

N2. $\forall P, Q, S \in \mathcal{P} : P \succcurlyeq Q,\ Q \succcurlyeq S \Rightarrow P \succcurlyeq S$ (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)

30. For 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.

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 53. *(von Neumann and Morgenstern, 1947). If and only if N1-N4 are satisfied, the preference relation $\succcurlyeq$ allows an affine representation with a loss function $\ell : C \to \mathbb{R}$, unique up to an affine transformation:*

$$P \succcurlyeq Q \iff \int_C \ell(\omega)\ \mathrm{d}P(\omega) \leqslant \int_C \ell(\omega)\ \mathrm{d}Q(\omega) \iff \mathbb{E}_P[\ell] \leqslant \mathbb{E}_Q[\ell].$$

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)). 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

$$\begin{aligned} &\forall P \in \mathcal{P} : \mathbb{E}_P[id] \succcurlyeq P, \\ \Leftrightarrow\ &\ell\left(\int_C \omega\ \mathrm{d}P(\omega)\right) \leqslant \int_C \ell(\omega)\ \mathrm{d}P(\omega), \end{aligned}$$

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

31. This 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.

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 1.1 and Section 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 one's degree of confidence in one's 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.

B.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:

$$\mathcal{P} = \{P : \Omega \to [0,1] : P(\omega) \neq 0 \text{ only for finitely many } \omega, \sum_{\omega \in \Omega} P(\omega) = 1\}.$$

An *act* is a function $X : S \to \mathcal{P}$ 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. $\forall X, Y \in \mathcal{L} : X \succcurlyeq Y$ or $X \succcurlyeq Y$ or both. (completeness)

M2. $\forall X, Y, Z \in \mathcal{L} : X \succcurlyeq Y,\ Y \succcurlyeq Z \Rightarrow X \succcurlyeq Z$ (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 $\forall \omega \in \Omega : \omega' \mapsto X(\omega) \succcurlyeq \omega' \mapsto Y(\omega)$ then $X \succcurlyeq Y$ (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. $\forall X, Y \in \mathcal{L}, \alpha \in (0,1) : X \sim Y \Rightarrow \alpha X + (1-\alpha)Y \succcurlyeq X$ (ambiguity aversion)

M7. not for all $X, Y \in \mathcal{L} : X \succcurlyeq Y$ (non-degeneracy)

where $X \sim Y$ denotes the indifference relation, i.e. $X \succcurlyeq Y$ and $Y \succcurlyeq X$, 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 54. *(Gilboa and Schmeidler, 1989). If and only if the preference relation $\succcurlyeq$ satisfies M1-M7 then it allows a representation of the form*

$$X \succcurlyeq Y \Longleftrightarrow \max_{Q \in \mathcal{Q}} \int \ell \circ X \, \mathrm{d}Q \leqslant \max_{Q \in \mathcal{Q}} \int \ell \circ Y \, \mathrm{d}Q$$

for a loss function $\ell : \mathcal{P} \to \mathbb{R}$ 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 $R(X) = \max_{Q \in \mathcal{Q}} \int \ell \circ X \, \mathrm{d}Q$ 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 outcomes[32]. 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.

B.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 $\succcurlyeq$ 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 $\forall \omega \in \Omega : Y(\omega) \geqslant X(\omega) \Rightarrow X \succcurlyeq Y$ (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

$$(X(\omega) - X(\omega'))(Y(\omega) - Y(\omega')) \geqslant 0 \quad \forall \omega, \omega' \in \Omega$$

32. We 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 $\overline{\mu}(\emptyset) = 0$ and $\overline{\mu}(\Omega) = 1$,[33] which is monotone.

Theorem 55. *(Chateauneuf, 1994). If and only if $\succcurlyeq$ satisfies the above axioms, there exists a capacity $\overline{\mu}$ on $\mathcal{F}$ such that*

$$X \succcurlyeq Y \Longleftrightarrow \int X \, \mathrm{d}\overline{\mu} \leqslant \int Y \, \mathrm{d}\overline{\mu},$$

where the Choquet integral with respect to the capacity $\overline{\mu}$ is defined as

$$\int X \, \mathrm{d}\overline{\mu} := \int_{-\infty}^{0} [\overline{\mu}(\{X \geqslant x\}) - 1] \, \mathrm{d}x + \int_{0}^{\infty} \overline{\mu}(\{X \geqslant x\}) \, \mathrm{d}x.$$

Remark 56. In this single-stage formulation, no loss/utility function has entered the picture. Typically, Choquet expected utility refers to representations of the form $\int u \circ X \, \mathrm{d}\overline{\mu}$ 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 (6), 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 uncertainy aversion axiom

CH5. $\forall X, Y, Z \in \mathcal{L}$: If $X \sim Y$ and Y and Z are comonotone, then $X + Z \succcurlyeq Y + Z$

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 57. *(Chateauneuf, 1994). If and only if $\succcurlyeq$ satisfies CH1-CH3 and CH5, then the representation of Theorem 55 holds and the capacity $\overline{\mu}$ is furthermore submodular, that is:*

$$\overline{\mu}(A \cup B) + \overline{\mu}(A \cap B) \leqslant \overline{\mu}(A) + \overline{\mu}(B).$$

To get an intuition for the Choquet integral, let us consider a finite space $\Omega = \{\omega_1, .., \omega_n\}$. Assume X is a step function which takes on the values $x_1 \leqslant x_2 \leqslant .. \leqslant x_n$. Let $x_0 = 0$. Then the Choquet integral can be written as

$$\int X \, \mathrm{d}\overline{\mu} = \sum_{i}^{n} (x_i - x_{i-1}) \overline{\mu}(\{X \geqslant x_i\}).$$

33. The normalization $\overline{\mu}(\emptyset) = 0$ is required for any capacity. Capacities with $\overline{\mu}(\Omega) = 1$ are also called *normalized capacities*. We impose $\overline{\mu}(\Omega) = 1$ throughout, however, and simply call it a capacity.

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*:

$$\forall A \subseteq B \subset \Omega, c \notin B : \overline{\mu}(A \cup \{c\}) - \overline{\mu}(A) \geqslant \overline{\mu}(B \cup \{c\}) - \overline{\mu}(B).$$

This expresses that adding an element to a smaller set results in a greater increase in decision weight. Consequently, large losses (where $\overline{\mu}(\{X \geqslant x\})$ 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 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

$$\operatorname{core}(\overline{\mu}) = \{P : P(A) \leqslant \overline{\mu}(A)\ \forall A \in \mathcal{F},\ P \text{ finitely additive probability measure}\}$$

$$\int X \ \mathrm{d}\overline{\mu} = \sup_{P \in \operatorname{core}(\overline{\mu})} \left\{ \int_{-\infty}^{\infty} X \ \mathrm{d}P \right\}.$$

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 makers, 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 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 classsical 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 $\overline{\mu} = \phi \circ P$ 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, consisting of all[34] 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.

B.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:

$$X \succcurlyeq Y \iff \int \ell \circ X \ \mathrm{d}(\phi \circ P) \leqslant \int \ell \circ Y \ \mathrm{d}(\phi \circ P),$$

34. Probability measures which are absolutely continuous with respect to a base measure.

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 3.6). If the loss function is the identity, as we take it in machine learning[35], 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 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.

B.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 $X : \Omega \to \mathbb{R}$ 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.

35. Note again that in our machine learning setup $X(\omega)$ already corresponds to a loss value.

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.

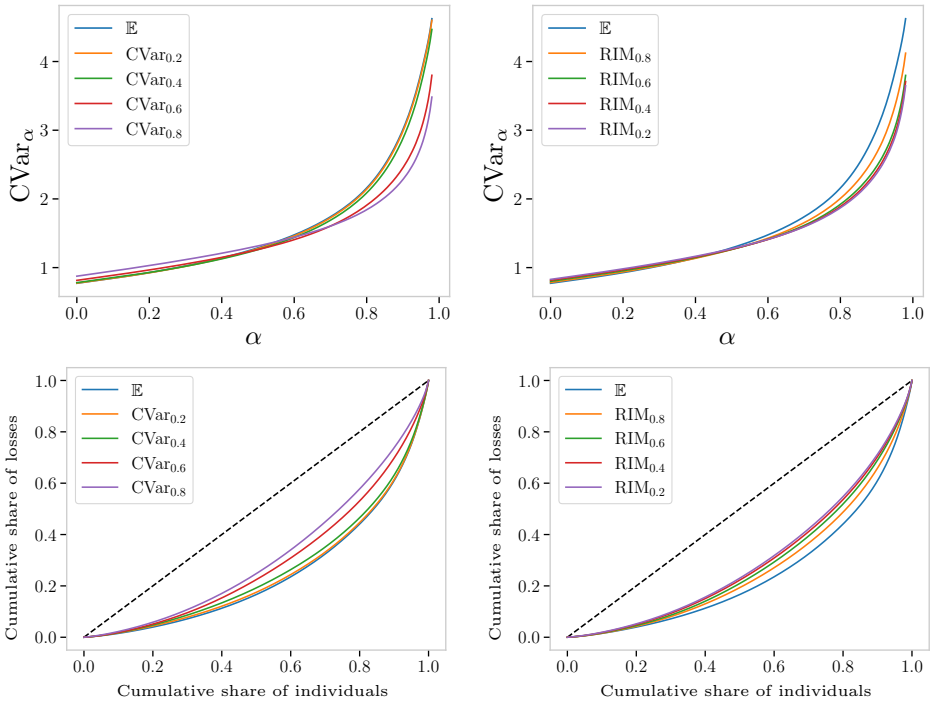

Figure 9: PCA* results on `adult`. Top row: CVar_α curves of test losses for CVar_α risk measures (left) with different α and RIMs risk measures (right) with different β, indicated by subscript, where $\alpha = 0.7$. For better visibility of the differences, we cut off α at 0.98. Bottom row: Lorenz curves of test losses for CVar_α (left) and RIMs (right) with $\alpha = 0.7$ and different β.

Spectral risk measures, provide a partial resolution to this conflict between aggregate and individual, between average utilitarianism and subgroup fairness. The extreme points of 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.

Appendix C. Experiments

Here we report additional results, not shown in the main paper. See Figures 9, 10, 11.

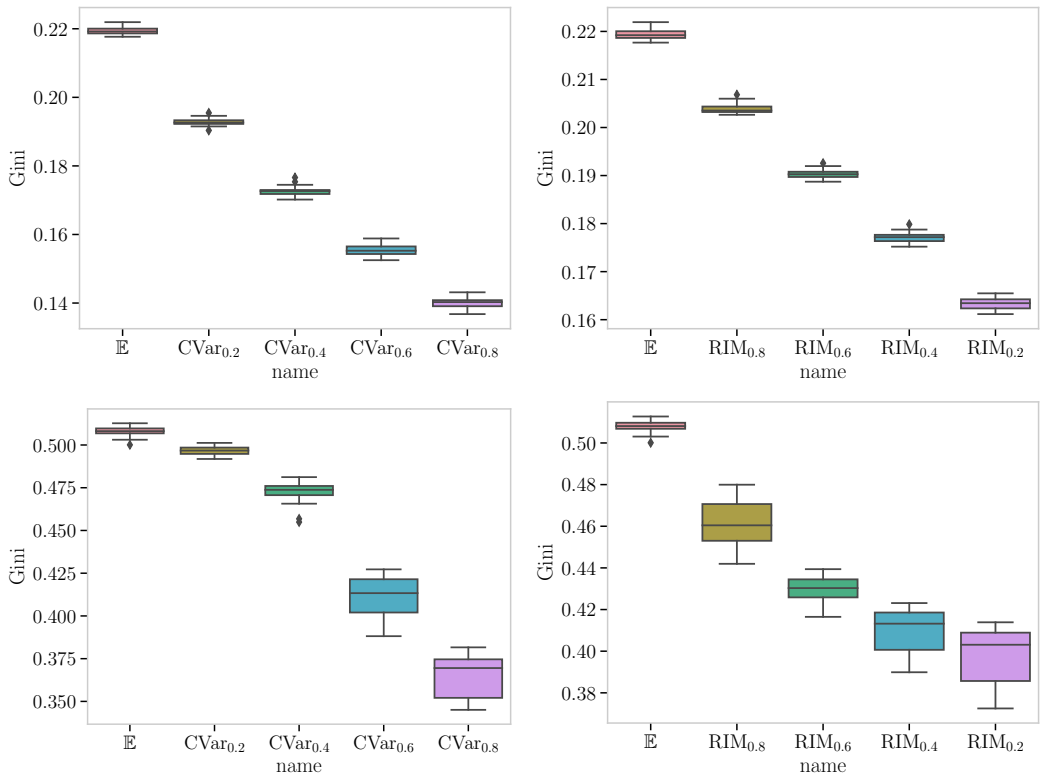

Figure 10: PCA* Gini coefficients for `MNIST` (top) and `adult` (bottom) over 25 runs.

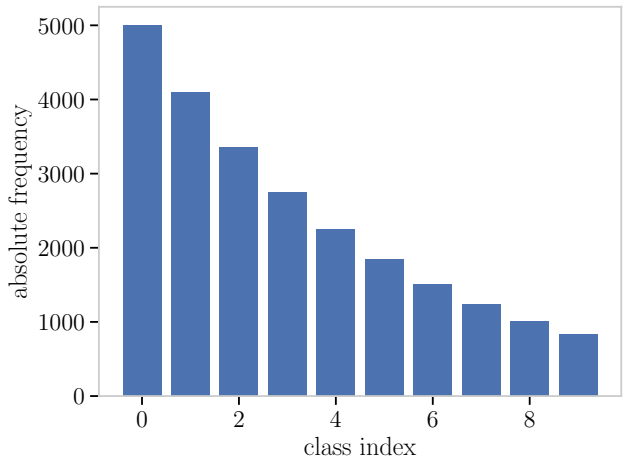

Figure 11: Class frequencies of imbalanced `MNIST`. Before each iteration, the assignments of digit class to frequency are randomly shuffled.

References

Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

Nabil I. Al-Najjar and Luciano De Castro. Subjective probability. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Ltd, 2010.

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*, 21:503–546, 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.

Francis J. Anscombe and Robert J. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1):199–205, 1963.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Thomas Augustin, Frank P.A. Coolen, Gert De Cooman, and Matthias C.M. Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

Abdessamad Barbara and Jean-Pierre Crouzeix. Concave gauge functions and applications. *Zeitschrift für Operations Research*, 40(1):43–74, 1994.

Nicole Bäuerle and Alfred Müller. Stochastic orders and risk measures: consistency and bounds. *Insurance: Mathematics and Economics*, 38(1):132–148, 2006.

Tadeusz Bednarski. On solutions of minimax test problems for special capacities. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 58(3):397–405, 1981.

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.

Seamus Bradley. Imprecise Probabilities. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2019 edition, 2019.

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. Brudnyĭ 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.

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.

Alain Chateauneuf. Modeling attitudes towards uncertainty and risk through the use of Choquet integral. *Annals of Operations Research*, 52(1):1–20, 1994.

Alexander Cherny and Dilip Madan. New measures for performance evaluation. *The Review of Financial Studies*, 22(7):2571–2606, 2009.

Fernado 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.

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.

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.

Philip Dawid. On individual risk. *Synthese*, 194(9):3445–3474, 2017.

Bruno de Finetti. *Theory of probability: A critical introductory treatment*. John Wiley & Sons, 1974/2017.

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.

Enrico Diecidue and Peter P. Wakker. On the intuition of rank-dependent utility. *Journal of Risk and Uncertainty*, 23(3):281–298, 2001.

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.

Daniel Ellsberg. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961.

Larry G. Epstein. A definition of uncertainty aversion. *The Review of Economic Studies*, 66 (3):579–608, 1999.

Johanna Etner, Meglena Jeleva, and Jean-Marc Tallon. Decision theory under ambiguity. *Journal of Economic Surveys*, 26(2):234–270, 2012.

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.

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.

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.

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.

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, 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.

Walter Bryce Gallie. Essentially contested concepts. *Proceedings of the Aristotelian society*, 56:167–198, 1955.

Joseph L. Gastwirth. A general definition of the Lorenz curve. *Econometrica*, 39(6): 1037–1039, 1971.

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.

Igor I. Gorban. *The statistical stability phenomenon*. Springer, 2017.

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.

Henryk Gzyl and Silvia Mayoral. On a relationship between distorted and spectral risk measures. *Revista de Economia Financera*, 15:8–21, 2008.

Ian Hacking. *The Taming of Chance*. Cambridge University Press, 1990.

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.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

Yoaav Isaacs, Alan Hájek, and John Hawthorne. Non-measurability, imprecise credences, and imprecise chances. *Mind*, 131(523):894–918, 2021.

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.

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.

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.

Peter Klibanoff. Characterizing uncertainty aversion through preference for mixtures. *Social Choice and Welfare*, 18(2):289–301, 2001.

Frank Hyneman Knight. *Risk, uncertainty and profit*. Houghton Mifflin, 1921.

Andrei N. Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Company, New York, 1950.

Jason Konek. Epistemic conservativity and imprecise credence, 2015. URL `https://philpapers.org/rec/KONECA`. Accessed on January 23, 2024.

Selim Grigor'evich Kreĭn, Jurii Ivanovvich Petunin, and Evgenii Mikhailovich Semenov. *Interpolation of Linear Operators*. American Mathematical Society, 1982.

Shigeo Kusuoka. On law invariant coherent risk measures. In *Advances in mathematical economics*, volume 3, pages 83–95. Springer, 2001.

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.

Liu Leqi, Audrey Huang, Zachary Lipton, and Kamyar Azizzadenesheli. Supervised learning with general risk functionals. In *International Conference on Machine Learning*, pages 12570–12592, 2022.

David K. Lewis. A subjectivist's 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.

Yuxi Liu. Beyond expectations, but within limits — the theory of coherent risk measures. Bachelor Thesis, Department of Mathematics, Australian National University, 2019.

Lola L. Lopes. Risk and distributional inequality. *Journal of Experimental Psychology: Human Perception and Performance*, 10(4):465, 1984.

Mark J. Machina and David Schmeidler. A more robust definition of subjective probability. *Econometrica*, 60(4):745–780, 1992.

Alexander Mafusalov and Stan Uryasev. CVaR (superquantile) norm: Stochastic case. *European Journal of Operational Research*, 249(1):200–208, 2016.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Ronak Mehta, Vincent Roulet, Krishna Pillutla, and Zaid Harchaoui. Distributionally robust optimization with bias and variance reduction. *arXiv preprint arXiv:2310.13863*, 2023.

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.

Ignacio Montes, Enrique Miranda, and Paolo Vicig. 2-monotone outer approximations of coherent lower probabilities. *International Journal of Approximate Reasoning*, 101: 181–205, 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.

Pietro Muliere and Marco Scarsini. A note on stochastic dominance and inequality measures. *Journal of Economic Theory*, 49(2):314–323, 1989.

Renato Pelessoni and Paolo Vicig. Imprecise previsions for risk measurement. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 11(04):393–412, 2003.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, volume 119, pages 7599–7609. PMLR, 2020.

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 Ruszczynski. Risk measures for income streams. Technical report, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Mathematik, 2001. preprint, DOI: 10.18452/8262.

Alois Pichler. The natural Banach space for version independent risk measures. *Insurance: Mathematics and Economics*, 53(2):405–415, 2013.

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.

Elad Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253*, 2018.

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.

John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

Mustapha Ridaoui and Michel Grabisch. Choquet integral calculus on a continuous support and its applications. *Operations Research and Decisions*, 26(1):73–93, 2016.

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.

Ben-Zion A. Rubshtein, Genady Ya. Grabarnik, Mustafa A. Muratov, and Yulia S. Pashkova. *Foundations of symmetric spaces of measurable functions*. Springer, 2016.

Leonard J. Savage. *The foundations of statistics*. John Wiley & Sons, 1954.

David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571–587, 1989.

Miriam Schoenfield. Chilling out on epistemic rationality. *Philosophical Studies*, 158(2): 197–219, 2012.

Moritz Schönherr and Friedemann Schuricht. Pure measures, density measures and the dual of L-infinity. *arXiv preprint arXiv:1710.02197*, 2017.

Stephen Semmes. Interpolation of Banach spaces, differential geometry and differential equations. *Revista Matemática Iberoamericana*, 4(1):155–176, 1988.

Alexander Shapiro. On Kusuoka representation of law invariant risk measures. *Mathematics of Operations Research*, 38(1):142–152, 2013.

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.

Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London, 1776.

H. Orri Stefánsson and Richard Bradley. What is risk aversion? *The British Journal for the Philosophy of Science*, 70(1):77–102, 2019.

Jie Sun, Xinmin Yang, Qiang Yao, and Min Zhang. Risk minimization, regret minimization and progressive hedging algorithms. *Mathematical Programming*, 181:509–530, 2020.

Akiko Takeda and Masashi Sugiyama. ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th international conference on Machine learning*, pages 1056–1063, 2008.

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.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. *arXiv preprint arXiv:2102.05371*, 2021.

Angela E. Van Heerwaarden and Rob Kaas. The Dutch premium principle. *Insurance: Mathematics and Economics*, 11(2):129–133, 1992.

Grant Van Horn, Oisin 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.

Nithia Vijayan and L.A. Prashanth. Policy gradient methods for distortion risk measures. *arXiv preprint arXiv:2107.04422*, 2021.

John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior, 2nd rev. ed.* Princeton university press, 1947.

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.

Shaun Wang. Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics*, 17(1):43–54, 1995.

Shaun S. Wang. A class of distortion operators for pricing financial and insurance risks. *Journal of Risk and Insurance*, pages 15–36, 2000.

Shaun S. Wang, Virginia R. Young, and Harry H. Panjer. Axiomatic characterization of insurance prices. *Insurance: Mathematics and Economics*, 21(2):173–183, 1997.

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 Aditya Menon. Fairness risk measures. In *International Conference on Machine Learning*, volume 97, pages 6786–6797. PMLR, 2019.

Menahem E. Yaari. The dual theory of choice under risk. *Econometrica*, 55(1):95–115, 1987.

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*, 2021.