Published version: Computational Mathematical Science
doi.org/10.1007/s10287-024-0516-z

DISTRIBUTIONAL ROBUSTNESS, STOCHASTIC DIVERGENCES, AND THE QUADRANGLE OF RISK

R. Tyrrell Rockafellar[1]

Abstract

In the distributional robustness approach to optimization under uncertainty, ambiguity about which probability distribution to use is addressed by turning to the worst that might occur with respect to a specified set of alternative probability distributions. Such sets are often taken to be neighborhoods of some nominal distribution with respect to a stochastic divergence like that of Kullback-Leibler or Wasserstein. Here that approach is coordinated with the fundamental quadrangle of risk with its quantifications not only of risk, but also regret, deviation and error, along with the functionals that dualize them.

Stochastic divergences are introduced axiomatically and shown to constitute the duals of risk measures in a special class. Rules are uncovered for how regret measures for those risk measures can be obtained by appropriate extensions of the divergence functional. This reveals clearly the pattern in which the robustness functionals coming from divergence neighborhoods can be provided with other formulas featuring minimization instead of maximization, which is beneficial for optimization schemes. To get everything to fit, however the aversity properties of risk and the rest that, until now, have been imposed in the quadrangle of relationships must be relaxed. A suitable substitute, called subaversity, is found that works while only differing from aversity for functionals that are not positively homogeneous.

Keywords: *distributionally robust optimization, coherent measures of risk, stochastic divergences, Kullbach-Leibler divergence, Wasserstein divergence, superquantile divergence, φ-divergences, stochastic ambiguity, ambiguity graduators, divergence neighborhoods, subaversity, risk quadrangle.*

Version of May 10, 2024

[1]University of Washington, Department of Mathematics, Box 354350, Seattle, WA 98195-4350;
E-mail: *rtr@uw.edu*, URL: http://sites.math.washington.edu/∼rtr/mypage.html

1 Uncertain expectations

The distinguishing feature of optimization problems in a host of applications is that the decision x to be optimized may have an uncertain outcome which depends on factors unknown until later, such as future weather or future demands, or the hidden weakness of some material. A simple way to think of this is that the outcome is a loss or cost expressed as a function $f(x, \omega)$ of x and an *uncertain state* ω as an element of a space Ω. It would be good to choose x to "minimize" the loss, but what can that mean when there is a different function of x for each different $\omega \in \Omega$?

The theory of risk, very generally, approaches this by organizing ways in which the different functions for different states ω can be consolidated into a single function of x on which minimization can be performed. Prominent in this are the *coherent* measures of risk that were introduced by Artzner, Delbaen, Eber and Heath [5] in 1999 with common-sense axioms that have since been refined or relaxed. In the fundamental quadrangle of risk proposed by Rockafellar and Uryasev [26] in 2013, measures of risk are combined with measures of deviation, error and regret in a framework that links issues in optimization with issues in statistics, often to a surprising extent. The four kinds of measures ("quantifiers" really, not measures in the sense of measure theory in mathematics) are convex functionals on a space of random variables, and they furthermore can be dualized through conjugacy, where even more of interest comes to light. Still, how does one first get to having a space of random variables in which the risk quadrangle can be articulated?

Looking at this on a more basic level, where a decision x doesn't need to enter, we can model uncertain losses as outcomes $X(\omega)$ of real-valued functions X on the state space Ω. That depiction of uncertainty doesn't make X be a *random variable*, though. To get that, we also need probabilities. They can be anchored by designating some probability distribution P_0 on Ω (technicalities about that being left aside, for now). With respect to P_0, X has a cumulative distribution function

$$F_{X,P_0}(\xi) = P_0\text{-probability of } \{\, \omega \in \Omega \,|\, X(\omega) \leq \xi \}. \tag{1.1}$$

In a problem with loss expression $f(x, \omega)$, there is likewise then for each x a P_0-based random variable $X_x = f(x, \cdot)$. We could consolidate the family of ω-dependent loss functions $x \mapsto f(x, \omega)$ by minimizing, say, the expected value $E[X_x]$ as a function of x, and that is just one of many good options.

This relies on having a *stochastic* model of uncertainty, which can be very effective when a natural designation of P_0 is at hand. There are many situations, however, in which it might not be at hand, through a lack of information, in particular. This has the motivated *robustness* approaches to optimization coming up next.

There are ways of avoiding probabilities entirely, and some have a long tradition. Faced with $f(x, \omega)$, we could try minimizing as a function of x the worst of losses as ω ranges over Ω (again with technicalities in interpretation left aside). More sophisticated in this direction, is the designation of a subset Ω_0 of Ω, as indicating the uncertain states that should really be of concern, and then minimizing the worst that could happen with ω restricted to Ω_0; see the 1998 paper of Ben-Tal and Nemirovski [8] and their 2009 book with El Ghaoui [6]. Such an approach has its virtues and successes, but can be overly conservative and thereby costly.

Distributional robustness offers a compromise of a sort. Instead of relying on a solitary distribution P_0 or avoiding probabilities altogether, we can turn to *sets* of distributions: $P \in \mathcal{P}$. Instead of working with $E_{P_0}[X]$, we can hedge by assigning to X the worst-case value

$$\sup_{P \in \mathcal{P}} E_P[X] \tag{1.2}$$

This keeps expectations in the foreground but regards those expectations themselves as uncertain, or as commonly said, *ambiguous*. Again there is a long history, going back to two-person games, but returning to prominence in ways explained in [16] and [30].

What might go into choosing a set $\mathcal{P}$ of alternative distributions P? The idea of designating a subset $\Omega_0 \subset \Omega$ and looking at the worst of outcomes for $\omega \in \Omega_0$ can be identified with choosing $\mathcal{P}$ to be the set of "all" probability distributions P having their support within Ω_0 (assigning zero probability to everything outside of Ω_0). Another idea, gaining in popularity, is taking $\mathcal{P}$ to be a "neighborhood" of some *nominal* (tentative, or best-guess) P_0, in the form

$$\{\, P \,|\, \mathcal{I}(P\|P_0) \leq \beta\} \;\text{ for some }\; \beta \in (0,\infty), \tag{1.3}$$

where $\mathcal{I}(P\|P_0)$ is a *stochastic divergence* expression giving a "distance" of P from P_0 that vanishes only when $P = P_0$. Important examples, which can serve as guidelines and will be discussed in detail later along with a variety of other examples, are the Kullbach-Leibler and Wasserstein divergences. Kullback-Leibler divergence has strong ties to information theory and Bolzano-Shannon entropy, but is essentially limited to probability distributions P expressible by a density with respect to the nominal P_0. Wasserstein divergence, which comes out of optimal transport, is valuable not so much for its interpretation as for its practical advantages in being able to compare P_0 with distributions P that might even have disjointly situated support.

Our aim here is to show how distributional robustness can be integrated into broader approaches in risk theory to the benefit of both. A key part of this is developing a very broad concept of stochastic divergence $\mathcal{I}(P\|P_0)$ and showing how the maximization formulas (1.2) for the associated neighborhoods (1.3), furnishing a nest of risk measures, can be partnered with minimization formulas involving *regret* measures. Up to now, such alternative minimization formulas have been developed in individual cases without the risk-regret pattern coming into view. But that pattern has a critical role in the quadrangle of risk in also designating a measure of *error* that can enter into regression and other tasks. Here we determine exactly the conditions on regret measures and error measures that tie them to a stochastic divergence. Futhermore, we bring out a direct duality between a stochastic divergence and a "parent" measure of risk which provides elementary alternative expressions for the risk measures (1.2) coming from (1.3). Fitting such parent risk measures into the fundamental quadrangle triggers a need for relaxing of the *aversity* property on which the quadrangle has so far relied. We find a good substitute and introduce it as *subaversity*.

We don't begin with quadrangle issues, because they reside in the random variable framework. Instead, we examine connections between distributional robustness and measures of risk that don't require designating a P_0 with all of Ω as its support. Coherency of a measure of risk in the sense in [5] is central to understanding the connections. But coherency in the slightly relaxed sense where the original axiom of positive homogeneity is suppressed turns out to be essential as well. It is needed in particular to cover the parent risk measure that dualizes a stochastic divergence.

The developments with the relaxed, or general, version of coherency lead to a consideration of *graduated robustness*. The robustness captured in (1.2) by designating an *ambiguity set* $\mathcal{P}$ of probability distributions P, instead of just a single distribution P_0, is "black-and-white" in allowing no gradation. Imagine more broadly the designation of not just $\mathcal{P}$ but also an expression $\mathcal{J}(P)$ with values in $[0,\infty]$, vanishing on $\mathcal{P}$ but nonzero for distributions not in $\mathcal{P}$. Make the interpretation that a distribution P with $\mathcal{J}(P) \in (0,\infty)$ is worth admitting in an evaluation of robustness, although at a lower level of influence than the ones in $\mathcal{P}$, as calibrated by the size of $\mathcal{J}(P)$. From that perspective, replace (1.2) by

$$\sup_P\{\, E_P[X] - \mathcal{J}(P)\}, \tag{1.4}$$

where of course only distributions P with $\mathcal{J}(P) < \infty$ really matter in the maximization. In this graduated version of robustness, $\mathcal{J}$ is an ambiguity *graduator* with $\mathcal{P}$ as its core. Stochastic divergences of P from a nominal P_0 will come out as certain cases of graduators $\mathcal{J}$ having $\{P_0\}$ as the core $\mathcal{P}$.

Here, in speaking of a *graduator* we are introducing a term that can be allied with "indicator" in referring to *any function with values in* $[0, \infty)$ *with* $\min = 0$ *and* $\operatorname{argmin} \neq \emptyset$. It "graduates" from that argmin set, its "core," where it has the value 0. In the extreme when the value jumps immediately from that set to ∞, it is the *indicator* of that set. In other words, a graduator is in concept a "fuzzy" indicator. When the graduator $\mathcal{J}$ in (1.4) is specialized to be the indicator of $\mathcal{P}$, with

$$\mathcal{J}(P) = 0 \text{ when } P \in \mathcal{P} \text{ but otherwise } \mathcal{J}(P) = \infty, \tag{1.5}$$

the supremum reduces to the earlier one in (1.2).

This brings us to questions of technical underpinnings. What structure should be provided for the state space Ω as a platform for comparing various different probability distributions on it? What class of functions $X : \Omega \to \mathbb{R}$ should be admitted as representing uncertain losses/costs? For many applications it would be good to target a probability space $(\Omega, \mathcal{A}, P_0)$ given by a field of sets $\mathcal{A}$ in the measure-theoretic sense and a probability measure P_0 on $\mathcal{A}$ and then restrict X to being in a particular linear function space $\mathcal{L}^p(\Omega, \mathcal{A}, P_0)$. But we hold back from that for two reasons. One, of course, is our reluctance to charge ahead by fixing a particular P_0 without investigating things from a wider perspective. Another is our desire to bypass a mire of technical complications, especially when no single best resolution of them is obvious. For example, it would be advantageous to compare distributions with finite support, arising empirically, with "continuous" distributions. We would run into that even in replacing $(\Omega, \mathcal{A}, P_0)$ by $(\Omega, \mathcal{A}, \mu)$ for a general measure μ, as in modeling a region of some $\mathbb{R}^d$ with μ as Lebesgue measure.

Our strategy here is therefore to leave those puzzles and complications aside and concentrate on the case of Ω being a *finite set*. This, after all, is the setting in which coherent risk was originally explored by Artzner et al. in [5]. It works well for explaining basic ideas and their relationships, and anyway is a case of major practical importance in its own right which can benefit from direct handling. Extensions beyond finite Ω are left to be carried out elsewhere, but Ruszczyński and Shapiro in [28] offer a particularly broad and sturdy foundation which could help with the technicalities in that. Also important in this picture is the paper of Shapiro [29] as a precedent for risk theory treatment of stochastic divergences in a setting of $\mathcal{L}^p$ spaces of random variables tied to a designated P_0. The divergences there, posed rather generally although not axiomatically, aren't themselves dualized to risk measures, nor are they given a quadrangle orientation, but much is developed about their law invariance, which is a topic we don't take up in this paper.

The first stage in our plan for placing distributional robustness and stochastic divergences in a larger matrix of interacting concepts is to explain connections between robustness and coherency. Section 2 is devoted to that. Divergences in example and generalizations are the next topic, in Section 3. Measures of regret and their role in producing alternative minimization formulas for the risk measures connected with divergences are taken up in Section 4. Finally, in Section 5, we pass to fully coordinating with the quadrangle of risk in its framework of random variables X backed up by a nominal probability distribution P_0.

2 Robustness in its relationship to coherency

In taking the state space Ω to be finite, we don't have to worry about which functions on it should be admitted. We can just work with the linear space

$$\mathbf{L}(\Omega) = \{ \text{ all functions } X : \Omega \to I\!R \}$$

as a finite-dimensional vector space. It could be identified with $I\!R^n$ under an indexing of the elements of Ω as $\omega_1, \ldots, \omega_n$, but that would get in our way. The would-be inner product of $I\!R^n$ comes out better for our purposes as

$$\langle X, Y \rangle = \sum\nolimits_{\omega \in \Omega} X(\omega) Y(\omega).$$

The convergence of a sequence $\{X^k\}$ to X is the convergence of $X^k(\omega)$ to $X(\omega)$ for every ω. The space of all probability distributions on Ω is simply

$$\mathbf{P}(\Omega) = \{ P \in \mathbf{L}(\Omega) \,|\, P \geq 0, \, \langle 1, P \rangle = 1 \},$$

where 1 in the inner product with P is the constant function "1" as an element of $\mathbf{L}(\Omega)$,[2] so that $\langle 1, P \rangle = \sum_{\omega \in \Omega} P(\omega)$. This is a compact convex subset of $\mathbf{L}(\Omega)$, its "canonical simplex." The expectation of an uncertain loss $X \in \mathbf{L}(\Omega)$ with respect to a distribution $P \in \mathbf{P}(\Omega)$ is

$$E_P(X) = \sum_{\omega \in \Omega} X(\omega) P(\omega) = \langle X, P \rangle.$$

In this framework, distributional robustness revolves around functionals $\mathcal{R}_{\mathcal{P}} : \mathbf{L}(\Omega) \to I\!R$ having the form

$$\mathcal{R}_{\mathcal{P}}(X) = \sup_{P \in \mathcal{P}} E_P(X) \;\text{ for nonempty }\; \mathcal{P} \subset \mathbf{P}(\Omega). \tag{2.1}$$

Note that the supremum doesn't change if $\mathcal{P}$ is replaced by its closure or by its convex hull, so that *only closed convex sets P matter* in the formula, and for them "sup" can be replaced by "max."

An immediate question is what distinguishes such functionals $\mathcal{R}_{\mathcal{P}}$ from other functionals on $\mathbf{L}(\Omega)$. What are their particular properties with respect to X? An answer to that was provided as the key contribution of the original paper of Artzner et al. [5] on risk. To explain it, we need to go over the axioms they introduced — as subsequently refined.

The idea behind a *measure of risk* $\mathcal{R}$ is that it consolidates an uncertain loss X into a single representative value $\mathcal{R}(X)$. Two fundamental examples are

$$\mathcal{R}(X) = E_P[X] \;\text{ for a } P \in \mathbf{P}(\Omega), \text{ or } \; \mathcal{R}(X) = \max X = \max_{\omega \in \Omega} X(\omega). \tag{2.2}$$

The first is *risk-neutral* in looking only at an average. That might be justified in circumstances where P is trusted, something occurs over and over, and the pain of a real loss, a positive $X(\omega)$, is perfectly balanced by later gaining back the same amount with equal probability. The second is concerned only with the worst that can happen and makes no allowance for balancing ups and downs.

Between these extremes there are many other possibilities, as will be seen. But it must be underscored that the "risk" in X to be evaluated by $\mathcal{R}$ is in the extent of loss, which is completely different from the degree of uncertainty in X. (Later, in the quadrangle framework, there will be measures of deviation $\mathcal{D}$ as introduced in [27], which evaluate how uncertain X might be.)

[2] In general we use the same symbol C for a number and for the corresponding constant function on Ω. It will always be clear from the context which interpretation is intended. But sometimes for emphasis we write $X \equiv C$ instead of just $X = C$, and on the other hand $X \not\equiv C$ as shorthand for X not being a constant function for any C.

Definition 2.1 (coherent measures of risk)**.** *A functional $\mathcal{R}$ on $\mathbf{L}(\Omega)$ is a coherent measure of risk in the general sense if it satisfies:*

(R1) *$\mathcal{R}$ is convex with closed level sets $\{\, X \,|\, \mathcal{R}(X) \le \xi \}$, $\xi < \infty$,*

(R2) *$\mathcal{R}(X) = C$ when $X \equiv C$,*

(R3) *$\mathcal{R}(X) \le \mathcal{R}(X')$ when $X \le X'$.*

It is a coherent measure of risk in the basic sense if, in addition, it satisfies

(R4) *$\mathcal{R}(\lambda X) = \lambda \mathcal{R}(X)$ for $\lambda > 0$.*

It needs to be pointed out right away that the part of (R1) about level sets being closed is redundant in our setting of finite Ω and has been included only as a bridge of clarity toward other choices of Ω. From the combination of (R2) and (R3) we have $\mathcal{R}(X) \le C$ when $X \le C$, so that $\mathcal{R}(X) < \infty$ when X is bounded from above. Here that's true for every $X \in \mathbf{L}(\Omega)$, so $\mathcal{R}$ must be *finite everywhere* and therefore, by the convexity in (R1), continuous everywhere.

The axioms (R1)–(R4), following [26], are equivalent to the ones in the original definition of coherency in [5], but in contrast they ask for convexity outright in (R1). Under (R4), $\mathcal{R}$ is convex if and only if

$$\mathcal{R}(X + X') \le \mathcal{R}(X) + \mathcal{R}(X'), \tag{2.3}$$

and in [5] this subadditivity property was the axiom combined with (R4) instead of direct convexity. However, that formulation is inconvenient when the omission of (R4) is contemplated as an extension of coherency. Another advantage of directly imposing (R1) is that

$$\text{(R1)+(R2)} \quad \Longrightarrow \quad \mathcal{R}(X + C) = \mathcal{R}(X) + C \ \text{ for constants } C. \tag{2.4}$$

In [5], the property in (2.4) is taken as an axiom instead of the simpler (R2), but that's redundant — as well as trickier to interpret and justify. The general principle of convex analysis which leads in particular to (2.4) will be useful here for more than just that, so we record it for reference as follows.

Proposition 2.2 (recession properties of convex functions [19, Theorem 8.5+])**.** *For a finite convex function $\mathcal{R}$, if there exist X_0, X' and ξ such that $\mathcal{R}(X_0 + \tau X') \le \mathcal{R}(X_0) + \tau\xi$ when $\tau \ge 0$, then $\mathcal{R}(X + \tau X') \le \mathcal{R}(X) + \tau\xi$ for all X when $\tau \ge 0$. If in fact $\mathcal{R}(X_0 + \tau X') \le \mathcal{R}(X_0) + \tau\xi$ for all τ, both positive and negative, then $\mathcal{R}(X + \tau X') = \mathcal{R}(X) + \tau\xi$ for all X and all τ.*

The risk measures in (2.2) are coherent in the basic sense, and more examples will soon be in hand. The desirability of insisting always on the positive homogeneity in (R4) came into question, however. The landmark book of Föllmer and Schied [14] in mathematical finance, in forgoing that property, speaks of "convex measures of risk." But that designation seems inadequate, because the functionals

$$\mathcal{R}(X) = E_P[X] + \gamma E_P[\,(X - E_P[X]\,)^2\,], \quad \gamma > 0, \tag{2.5}$$

satisfy (R1) and (R2) without satisfying (R3). Such convex mean-variance functionals have long been employed as measures of risk in finance, but their lack of the monotonicity in (R3) is disturbing, and signals to us an "incoherency" as crucial as a lack of convexity. That's behind our preference for speaking of convex functionals satisfying (R1), (R2) and (R3), but not (R4), as still being coherent, which began with [20]. It's a terminology that emphasizes what truly counts in making sense of risk and thereby leaves out measures $\mathcal{R}$ like those in (2.5).

Artzner et al. in [5] established the correspondence in the theorem we state next. But the result can easily be derived from basic rules of convex analysis, as we record in the proof given here.

Theorem 2.3 (robustness dualization of basic coherency)**.** *The measures of risk $\mathcal{R}$ on $\mathbf{L}(\Omega)$ that are coherent in the basic sense correspond one-to-one, through the formula*

$$\mathcal{R}(X) = \max_{P\in\mathcal{P}} E_P[X], \qquad (\mathcal{P} = \text{“risk envelope” for } \mathcal{R}) \tag{2.6}$$

with the nonempty closed, convex sets $\mathcal{P} \subset \mathbf{P}(\Omega)$. Thus, the robustness functionals $\mathcal{R}_\mathcal{P}$ in (2.1) are precisely the measures of risk on $\mathbf{L}(\Omega)$ that are coherent in the basic sense.

Proof. As known from [19, Section 13], the formula $\mathcal{S}_\mathcal{C}(X) = \sup_{Y\in\mathcal{C}}\langle X, Y\rangle$ yields a one-to-one correspondence between the nonempty closed convex sets $C \subset \mathbf{L}(\Omega)$ and the closed proper convex functions on $\mathbf{L}(\Omega)$ that are positively homogeneous — their support functions, with $\mathcal{S}_\mathcal{C}$ being finite if and only if $\mathcal{C}$ is bounded. Here we are specializing that to $\mathcal{C} = \mathcal{P} \subset \mathbf{P}(\Omega)$, and therefore are dealing with finite $\mathcal{R}$ satifying (R1) and (R4), but the question remains of what additional properties of $\mathcal{R}$ correspond to constraining $Y \in \mathcal{C}$ to be nonnegative with $\langle 1, Y\rangle = 1$. Those properties can be identified through the rule that

$$\mathcal{C} \subset \mathcal{C}' \iff \mathcal{S}_\mathcal{C} \le \mathcal{S}_{\mathcal{C}'}.$$

Taking $\mathcal{C}' = \{\, Y \,|\, Y \ge 0\}$ we get as $\mathcal{S}_{\mathcal{C}'}$ the indicator of $\{\, X \,|\, X \le 0\}$, hence the property that $\mathcal{R}(X) \le 0 = \mathcal{R}(0)$ when $X \le 0$. But by Proposition 2.2 this is equivalent to the seemingly stronger property that $\mathcal{R}(X + X') \le \mathcal{R}(X)$ when $X' \le 0$, which is (R3). Taking $\mathcal{C}' = \{\, Y \,|\, \langle 1, Y\rangle = 1\}$, on the other hand, we have $\mathcal{S}_{\mathcal{C}'}(X) = C$ when $X \equiv C$ and $\mathcal{S}_{\mathcal{C}'}(X) = \infty$ for nonconstant X. That tells us that $\mathcal{R}(C) \le C$ for all C, positive and negative, and then equality must hold by Proposition 2.2. □

Having characterized the *basic* robustness functionals $\mathcal{R}_\mathcal{P}$ in (2.1), we now take up the task of characterizing *graduated* robustness functionals of the form

$$\mathcal{R}_\mathcal{J}(X) = \sup_{P\in\mathbf{P}(\Omega)} \{\, E_P[X] - \mathcal{J}(P)\} \text{ for a graduator } \mathcal{J} \text{ on } \mathbf{P}(\Omega). \tag{2.7}$$

Recall from the introduction of the term “graduator” in Section 1, ahead of (1.4), that it refers here to a function from $\mathbf{P}(\Omega)$ to $[0, \infty]$ that is 0 on a nonempty set $\mathcal{P} \subset \mathbf{P}(\Omega)$, its argmin. As in (2.1), where we only had a set $\mathcal{P}$ by itself, and we noted that $\mathcal{P}$ might just as well be closed and convex, both closure and convexification of $\mathcal{J}$ yield the same $\mathcal{R}_\mathcal{J}$. We pin these natural properties down in the following axioms.

Definition 2.4 (coherent graduators)**.** *A functional $\mathcal{J}$ on $\mathbf{P}(\Omega)$ will be called a coherent graduator of ambiguity if it satisfies:*

(J1) *$\mathcal{J}$ is convex with values in $[0, \infty]$,*
(J2) *the level sets $\{\, P \in \mathbf{P}(\Omega) \,|\, \mathcal{J}(P) \le \beta\}$ are closed for $\beta \in [0, \infty)$,*
(J3) *$\min \mathcal{J} = 0$: $\exists\, P$ with $\mathcal{J}(P) = 0$.*

In line with the interpretation of (2.7) as the earlier (1.4), $\mathcal{J}$ graduates ambiguity from the set $\mathcal{P} = \{\, P \,|\, \mathcal{J}(P) = 0\}$, which the axioms ensure is nonempty, closed, and convex in $\mathbf{P}(\Omega)$. In the extreme case where $\mathcal{J}$ is simply the indicator of $\mathcal{P}$ as in (1.5), (2.7) reduces to (2.1). But we are interested now in situations beyond that, where some “graduation” is definitely offered and the correspondence in Theorem 2.3 is a specialization of something broader.

Theorem 2.5 (robustness dualization of general coherency)**.** *The measures of risk $\mathcal{R}$ on $\mathbf{L}(\Omega)$ that are coherent in the general sense correspond one-to-one, through the formula*

$$\mathcal{R}(X) = \sup_{P\in\mathbf{P}(\Omega)} \Big\{\, E_P[X] - \mathcal{J}(P)\Big\}, \qquad (\mathcal{J} = \text{“dualizing graduator” for } \mathcal{R}) \tag{2.8}$$

with the coherent graduators $\mathcal{J}$ in Definition 2.4. Thus, the graduated robustness functionals $\mathcal{R}_\mathcal{J}$ in (2.7) are precisely the measures of risk on $\mathbf{L}(\Omega)$ that are coherent in the general sense.

Proof. By taking $\mathcal{J}$ to be ∞ outside of $\mathbf{P}(\Omega)$, we get on the basis of (J1)–(J3) a closed proper convex function on all of $\mathbf{L}(\Omega)$. Then (2.8) can be restated equivalently as

$$\mathcal{R}(X) = \sup_{Y \in \mathbf{L}(\Omega)} \{ \langle X, Y \rangle - \mathcal{J}(Y) \},$$

in saying that $\mathcal{R}$ is the conjugate $\mathcal{J}^*$, and then $\mathcal{J}$ is the conjugate of $\mathcal{R}$. This is the fundamental one-to-one duality correspondence in convex analysis, extending the one in Theorem 2.3, for indicators $\mathcal{J}$ by dropping positive homogeneity (R4) from the properties required of $\mathcal{R}$. It obeys the rule that $\mathcal{J} \geq \mathcal{J}'$ if and only if $\mathcal{R} \leq \mathcal{R}'$, when $\mathcal{R}'$ and $\mathcal{J}'$ are likewise conjugate to each other. From that, the argument that (R2) and (R3) complete the characterization is the same as the one using Proposition 2.2 in the proof of Theorem 2.3, since $\mathcal{R}(0) = -\inf \mathcal{J}$ in the conjugacy. □

3 Bringing in stochastic divergences

In applications of distributional robustness in optimization, special attention is given to cases where the set of distributions P is a sort of neighborhood of a given distribution P_0 with respect to some version of *stochastic divergence* $\mathcal{I}(P\|P_0)$ indicating how far P is from P_0,

$$\mathcal{R}_\beta(X) = \max_{P \in \mathcal{P}_\beta} E_P[X] \;\text{ for }\; \mathcal{P}_\beta = \Big\{ P \in \mathbf{P}(\Omega) \,\Big|\, \mathcal{I}(P\|P_0) \leq \beta \Big\},\ \beta \in (0, \infty). \tag{3.1}$$

Two examples were mentioned in the introduction: Kullbach-Leibler divergence and Wasserstein divergence. We review them now, with other examples, and go on to paint a picture of how stochastic divergences in general conception fit into the framework of coherent risk.

Definition 3.1 (Wasserstein divergence)**.** *Consider on $\Omega \times \Omega$ any expression $W(\omega, \omega')$ such that*

$$W(\omega, \omega') \in [0, \infty), \quad \textit{with}\ \ W(\omega, \omega') = 0 \iff \omega = \omega'. \tag{3.2}$$

Let $\mathbf{P}(\Omega \times \Omega)$ denote the set of all probability distributions Π on $\Omega \times \Omega$. Then

$$\mathcal{I}(P\|P_0) = \begin{cases} \textit{minimum of}\ \ E_\Pi[W] = \sum_{\omega,\omega'} W(\omega, \omega')\Pi(\omega, \omega')\ \ \textit{over} \\ \textit{all}\ \ \Pi \in \mathbf{P}(\Omega \times \Omega)\ \textit{having}\ P\ \textit{and}\ P_0\ \textit{as its marginals,} \end{cases} \tag{3.3}$$

where the marginality constraint means that $\sum_{\omega'} \Pi(\omega, \omega') = P(\omega)$ and $\sum_{\omega} \Pi(\omega, \omega') = P_0(\omega')$.

This revolves around "probability transport" with $\Pi(\omega, \omega')$ being the amount of probability taken from $P_0(\omega')$ and transported to ω to be part of $P(\omega)$, the transportation cost per unit being $W(\omega, \omega')$. The minimization problem is in the category of linear programming, where an optimal solution is sure to exist and the minimum value as a function of P is convex and piecewise linear. Under (3.2), the minimum value is 0 only when $P = P_0$, so a "distance" of P from P_0 is expressed in general.

A reason for the popularity of Wasserstein divergence in many of the applications made of it is that no restriction is placed on the supports of the distributions P and P_0, the support of P being in our framework

$$\operatorname{supp} P = \{ \omega \in \Omega \,|\, P(\omega) > 0 \}. \tag{3.4}$$

The sets $\operatorname{supp} P$ and $\operatorname{supp} P_0$ might even be disjoint. There has been much written about this, and the Gao-Kleywegt paper [15] could be a good entry point.

Other notions of stochastic divergence, like the one we'll look at next, depend on P being *representable by a density with respect to* P_0:

$$P(\omega) = Q(\omega)P_0(\omega), \quad \text{requiring} \quad \operatorname{supp} P \subset \operatorname{supp} P_0. \tag{3.5}$$

The densities that fill this role are the functions $Q \geq 0$ on $\operatorname{supp} P_0$ having $E_{P_0}[Q] = 1$.

Definition 3.2 (Kullbach-Leibler divergence)**.** *For P density-representable as in (3.5), let*

$$\mathcal{I}(P\|P_0) = E_{P_0}\Big[Q \log Q\Big], \tag{3.6}$$

but otherwise take $\mathcal{I}(P\|P_0) = \infty$. (Here $0 \log 0 = 0$ in the usual convention from taking limits.)

The formula in (3.6), an expression of relative entropy coming from information theory, is well known in convex analysis as giving an example of a continuous convex function of P having positive values unless $P = P_0$, in which case $Q \equiv 1$ and the value is 0.

Robustness neighborhoods based on Kullbach-Leibler divergence are the centerpiece of the 2012 paper of Ahmadi-Javid [2]. For him, the neighborhood risk measures $\mathcal{R}_\beta$ in this case constituted *entropic value-at-risk*, $\mathrm{EVaR}_\alpha(X) = \mathcal{R}_\beta(X)$ with $\alpha = 1 - e^{-\beta}$, which switches the β-range $(0, \infty)$ to an α-range $(0, 1)$.[3]

But potentially useful neighborhoods beyond these, that have been touted for usefulness in robust optimization by Ben-Tal et al. in [7], can be obtained by putting different expressions $\varphi(Q)$ in place of $Q \log Q$ in (3.6) to get *φ-divergences*,

$$\mathcal{I}(P\|P_0) = \begin{cases} E_{P_0}[\,\varphi(Q)\,] & \text{if} \quad \operatorname{supp} P \subset \operatorname{supp} P_0, \\ \infty & \text{otherwise,} \end{cases} \tag{3.7}$$

which have various roles in statistics as explained by Liese and Vajda [17]. This divergence idea, going back at least to Csiszár [11] (1963) and independently to Ali and Silbey [4] (1966), was more recently taken up by Ahmadi-Javid in [1] (2011) and [2] (2012) as well as by Breuer and Csiszár [10] (2013), who went somewhat further with what they called *Bregman-divergences* — for which there needs to be a different φ for each ω. The Breuer-Csiszár results, couched in terms of moment constraints, are challenging to coordinate with numerical optimization methodology, however.

Dommel and Pichler, still more recently in [12] (2021), pursued the matter more like Ahmadi-Javid and with outlook and terminology more closely aligned with ours here. They focused on φ-divergences defined from finite convex functions φ on $[0, \infty)$ with $\varphi(1) = 0$ that are continous from the right at 0 and have $\varphi(q)/q \to \infty$ as $q \to \infty$. Breuer and Csiszár in [10] instead imposed *strict convexity* on φ without making any assumption at 1 or about a q limit at ∞, whereas Ahmed-Javid only asked for $\varphi(1) = 0$ (although from the context it's clear that right continuity at 0 was also intended). Observe, though, that by taking any $r \in \partial\varphi(1)$ (thus $r = \varphi'(1)$ if the left and right derivatives of φ agree at 1), and replacing φ in (3.7) by φ_1 with $\varphi_1(q) := \varphi(q) - \varphi(1) - r[q - 1]$, there would be *no change at all in the resulting $\mathcal{I}(P\|P_0)$*. Therefore, in dealing with (3.7), nothing is lost by "normalizing" to the common ground of

$$\varphi \text{ finite convex on } [0, \infty) \text{ with } \lim_{q \searrow 0} \varphi(q) = \varphi(0), \ \min \varphi = 0, \text{ and } 1 \in \operatorname{argmin} \varphi. \tag{3.8}$$

This has the advantage of making immediately obvious that $\mathcal{I}(P\|P_0)$ is a convex functional of P that *attains at P_0 its minimum value,* 0. If no point other than 1 belongs to the interval $\operatorname{argmin} \varphi$,

[3] The α here is $1 - \alpha$ in [2].

as would hold under the strict convexity demanded by Breuer and Csiszár, then surely $\mathcal{I}(P\|P_0) > 0$ when $P \neq P_0$, but that would fail if the interval has 1 in its interior, as allowed in (3.8) in reflection of the assumptions of Dommel and Pichler.

Note that the normalization to (3.8) amounts in the case of Kullback-Leibler divergence to harmlessly replacing the strictly convex expression $q \log q$ in (3.6) by $q \log q - q + 1$.

The conditions in (3.8), perhaps strengthed to $\operatorname{argmin} \varphi = \{1\}$ without requiring strict convexity, might be relaxed in other directions. We'll pick up again on φ-divergences at the end of Section 4.

Standards for what might, very generally, be called a stochastic divergence will be proposed below. The next example, like Wasserstein divergence, falls outside the φ-divergence pattern and offers another perspective on what needs to be included in such generality.

Definition 3.3 (CVaR divergence, or superquantile divergence)**.** *For P density-representable as in (3.5), let*

$$\mathcal{I}(P\|P_0) = P_0\text{-}\max Q - 1 = \max_{\omega \in \operatorname{supp} P_0} Q(\omega) - 1, \tag{3.9}$$

but otherwise take $\mathcal{I}(P\|P_0) = \infty$.

Once more we have from this formula a convex function of P that vanishes when $P = P_0$, the case when $Q \equiv 1$, but otherwise is positive, since in other cases there must be some ω with $Q(\omega) > 1$ or we wouldn't have P in $\mathbf{P}(\Omega)$ along with P_0.

What's behind the CVaR name of this divergence, not previously brought to anyone's attention? The worst-case functionals $\mathcal{R}_\beta$ associated with its neighborhoods $\mathcal{P}_\beta$ in (3.1) turn out to produce the CVaR family of risk measures. What are they? For the P_0-random variable obtained from X with its cumulative distribution function F_{X,P_0} in (1.1), the *upper α-tail* distribution for $\alpha \in (0,1)$, referring to the worst $100(1-\alpha)\%$ outcomes,[4] is the conditional probability distribution for X in that tail. The *P_0-conditional-value-at-risk* of X at level $\alpha \in (0,1)$ is

$$P_0\text{-CVaR}_\alpha(X) = \text{ expected value of } X \text{ in its upper } \alpha\text{-tail distribution}, \tag{3.10}$$

as rigorously pinned down in [25] to mean the expected value associated with the cumulative distribution function

$$F^\alpha_{X,P_0}(\xi) = \frac{1}{1-\alpha} \max\{F_{X,P_0}(\xi) - \alpha, 0\}. \tag{3.11}$$

In these terms, our assertion about the stochastic divergence introduced in Definition 3.1 as CVaR divergence is the following:

$$\begin{array}{l} \text{the neighborhoods } \mathcal{P}_\beta \text{ in (3.1) for (3.9) produce} \\ \mathcal{R}_\beta(X) = P_0\text{-CVaR}_\alpha(X), \ \text{ where } \ 1 - \alpha = (1+\beta)^{-1}. \end{array} \tag{3.12}$$

This rests on specializing Theorem 2.3 to $\mathcal{R}(X) = P_0\text{-CVaR}_\alpha(X)$, knowing that the risk envelope $\mathcal{P}$ consists then of all P given by densities Q in (3.5) such that $Q \leq (1-\alpha)^{-1}$ on $\operatorname{supp} P_0$.

The CVaR definition in (3.10) is delicate because there can be α such that, for $\xi = P_0\text{-VaR}_\alpha(X)$, the *$P_0$-value-at-risk* (the least ξ such that $F_{X,P_0}(\xi) \geq \alpha$), the probability P_0 assigns to $[\xi, \infty)$ is more than $1-\alpha$, yet the probability it assigns to (ξ, ∞) is less than $1-\alpha$, due to a jump in F_{X,P_0} at ξ. Then there seems to be no α-tail at all. But that's resolved through (3.11), in effect by splitting a probability atom at ξ.[5] Much the same concept of risk was floated earlier under names like tail

[4] For instance, for $\alpha = 0.9$ this means the worst 10% of outcomes.

[5] In our discrete-probability setting with Ω finite, atoms are of course unavoidable. Cumulatative distribution functions are always step functions.

risk and expected shortfall, but without this necessary tail refinement for having a measure of risk that's coherent in all situations. Fölmer and Schied in their very influencial book [14] in mathematical finance have promoted "average-value-at-risk" rather than "conditional-value-at-risk." On the other hand, because of numerous applications where financial terminology seems inappropriate, and "value-at-risk" is identified as "quantile," a standard notion in statistics, *superquantile* has been proposed as a suitable alternative name for "conditional value-at-risk" [23]. That explains the double name for the divergence in Definition 3.3.

Building on these examples of stochastic divergences and the properties they have that seem essential, we now offer a general description to work with.

Definition 3.4 (stochastic divergence, in general)**.** *A function $\mathcal{I}(P\|P_0)$ of $P \in \mathbf{P}(\Omega)$ will be said to give a stochastic divergence from $P_0 \in \mathbf{P}(\Omega)$ if it satisfies:*

(I1) *$\mathcal{I}(\cdot\|P_0)$ is convex with values in $[0,\infty]$,*
(I2) *the level sets $\{\, P \in \mathbf{P}(\Omega) \,|\, \mathcal{I}(P\|P_0) \le \beta\}$ are closed for $\beta \in [0,\infty)$,*
(I3) *$\min \mathcal{I}(\cdot\|P_0) = 0$ with $\operatorname{argmin} \mathcal{I}(\cdot\|P_0) = \{P_0\}$,*
(I4) *$\{\, P \in \mathbf{P}(\Omega) \mid \operatorname{supp} P \subset \operatorname{supp} P_0\} \subset \operatorname{cl}\{\, P \in \mathbf{P}(\Omega) \,|\, \mathcal{I}(P\|P_0) < \infty\}$.*

Clearly (I1), (I2) and (I3) mimic the axioms (J1), (J2) and (J3) for a coherent graduator in Definition 2.4, except that (I3) strengthens (J3) in making P_0 be the only distribution in the argmin. The purpose of this strengthening, and the further axiom (I4), is to guarantee that a worthy nest of ambiguity neighborhoods of P_0 is generated in considering distributions P with $\mathcal{I}(P\|P_0) \le \beta$ for a "distance" $\beta > 0$. The associated worst-case functionals are available then for deployment in distributionally robust optimization in a vast extension of the methodology already in widespread use for Wasserstein and Kullbach-Leibler divergences.

Much more about divergence neighborhoods will be examined in the next section of this paper. The remainder of this section is aimed at identifying where stochastic divergences fit in the coherency correspondence of Theorem 2.5.

Theorem 3.5 (risk dualization of stochastic divergences)**.** *A functional $\mathcal{J}$ on $\mathbf{P}(\Omega)$ is a stochastic divergence $\mathcal{I}(\cdot\|P_0)$ from $P_0 \in \mathbf{P}(\Omega)$ if and only if it is a coherent graduator for which the corresponding risk measure $\mathcal{R}$ in Theorem 2.5 satisfies for all X and the nondecreasing function $\tau \to \tau^{-1}\mathcal{R}(\tau X)$ on $(0,\infty)$,*

$$\begin{aligned} &\mathcal{R}(X) \ge E_{P_0}[X] \;\text{ and }\; \lim_{\tau \searrow 0} \mathcal{R}(\tau X)/\tau = E_{P_0}[X], \\ &\lim_{\tau \nearrow \infty} \mathcal{R}(\tau X)/\tau \ge P_0\text{-}\max X := \max\{\, X(\omega) \,|\, \omega \in \operatorname{supp} P_0\}, \end{aligned} \tag{3.13}$$

with the limit as $\tau \searrow 0$ meaning that $\mathcal{R}$ is differentiable at 0 and $\nabla\mathcal{R}(0) = P_0$. Here $\mathcal{R}$ is coherent in the general sense, but definitely not in the basic sense.

Proof. The issue is how the strenghening of graduator axioms from (J1)–(J3) to (I1)–(I4) affects the correspondence in Theorem 2.5. Because $\mathcal{R}(0) = 0$ by (R2), $\mathcal{R}(\tau X)/\tau$ is the difference quotient $\Delta_\tau\mathcal{R}(X) = [\mathcal{R}(0+\tau X) - \mathcal{R}(0)]/\tau$. The finite convexity of $\mathcal{R}$ is inherited by $\Delta_\tau\mathcal{R}$, and $\Delta_\tau\mathcal{R}(X)$ as a function of $\tau > 0$ is nondecreasing, thus having limits when $\tau \nearrow \infty$ and when $\tau \searrow 0$. The limits are

$$\lim_{\tau \nearrow \infty} \Delta_\tau\mathcal{R}(X) = \mathcal{R}0^+(X), \qquad \lim_{\tau \searrow 0} \Delta_\tau\mathcal{R}(X) = \mathcal{R}'(0;X),$$

the first being the recession function associated with $\mathcal{R}$ according to [19, Theorem 8.5] and the second being by definition the directional derivative function at the origin.

The recession function $\mathcal{R}0^+$ is known to be the support function of the closure of the effective domain of the conjugate, here the closed convex set $\operatorname{cl}[\operatorname{dom}\mathcal{J}]$ [19, Theorem 13.3]; that's therefore

the limit on the left side in the second line of (3.13). The right side of that relation gives the support function of the set of probability distributions P supported by P_0. The inequality between left and right corresponds exactly to the inclusion required by (I4).

On the other hand, the directional derivative, here a finite convex function of X, is the support function of the subgradient set $\partial\mathcal{R}(0)$ [19, Theorem 23.4], which in conjugacy is the set argmin $\mathcal{J}$ [19, Theorem 23.5]. Differentiability is the case where this subgradient set is a singleton consisting of the gradient [19, Theorem 25.1]. With the limit being $E_{P_0}[X] = \langle P_0, X\rangle$, the gradient is P_0.

Although $\mathcal{R}$ is coherent in the general sense by Theorem 2.5, it can't be coherent in the basic sense, because that would put it in the sway of Theorem 2.3, where $\mathcal{J}$ is an indicator function. Axioms (I3) and (I4) preclude $\mathcal{J}$ as a graduator from actually being an indicator. □

Example 3.6 (dualization of Wasserstein divergence)**.** *For $\mathcal{J}(P) = \mathcal{I}(P\|P_0)$ in the case of (3.2), the dualizing measure of risk that is coherent in the general sense is*

$$\mathcal{R}(X) = E_{P_0}[X^\circ] \;\; \textit{where} \;\; X^\circ(\omega') = \max_{\omega\in\Omega}\{\, X(\omega) - W(\omega,\omega')\}. \tag{3.14}$$

Detail. Because $\mathcal{R}$ is the convex function conjugate to $\mathcal{J}(P) = \mathcal{I}(P\|P_0)$ that's defined by the linear programming formula in (3.2), it is biven by

$$\max\Big\{ \sum_{\omega} X(\omega) \sum_{\omega'} \Pi(\omega,\omega') - \sum_{\omega,\omega'} W(\omega,\omega')\Pi(\omega,\omega') \;\Big|\; \Pi(\omega,\omega') \geq 0,\, \sum_{\omega} \Pi(\omega,\omega') = P_0(\omega')\Big\}.$$

Letting $\pi(\omega,\omega') = \Pi(\omega,\omega')/P_0(\omega')$ when $\omega' \in \operatorname{supp} P_0$ and noting that $\Pi(\omega,\omega')$ must be 0 when $\omega' \notin \operatorname{supp} P_0$, we can express this as

$$\begin{aligned}&\max\Big\{ \sum_{\omega,\omega'} P_0(\omega')\pi(\omega,\omega')\Big[X(\omega) - W(\omega,\omega')\Big] \;\Big|\; \pi(\omega,\omega') \geq 0,\, \textstyle\sum_\omega \pi(\omega,\omega') = 1\Big\}\\ &\quad = \textstyle\sum_{\omega'} P_0(\omega') \max_\omega\{\, X(\omega) - W(\omega,\omega')\} = E_{P_0}[X^\circ],\end{aligned}$$

as claimed in (3.14). □

Example 3.7 (dualization of Kullbach-Leibler divergence)**.** *For $\mathcal{J}(P) = \mathcal{I}(P\|P_0)$ in the case of (3.6), the dualizing measure of risk that is coherent in the general sense is*

$$\mathcal{R}(X) = \log E_{P_0}[\exp X]. \tag{3.15}$$

Detail. The duality between the expressions in (3.6) and (3.15) is a familiar example in the theory of conjugate convex functions, cf. [19, pp. 148–149]. □

Example 3.8 (dualization of CVaR divergence)**.** *For $\mathcal{J}(P) = \mathcal{I}(P\|P_0)$ in the case of (3.9), the dualizing measure of risk that is coherent in the general sense is*

$$\mathcal{R}(X) = 1 + C \textit{ for the unique } C \textit{ with } E_{P_0}[\max\{X - C, 0\}] = 1. \tag{3.16}$$

Detail. Here $\max\{X - C, 0\}$ denotes the function $\omega \mapsto \max\{X(\omega) - C, 0\}$ in $\mathbf{L}(\Omega)$. We want to determine the functional $\mathcal{R}$ conjugate to the functional $\mathcal{J}$ that's given on all of $\mathbf{L}(\Omega)$ by $\mathcal{J} = \mathcal{J}_1 + \mathcal{J}_2 + J_3 - 1$, where

$$\begin{aligned}\mathcal{J}_1(Y) &= \text{ indicator of } \{\, Y \,|\, Y(\omega) \geq 0 \text{ for } \omega \in \operatorname{supp} P_0, \text{ but } =0 \text{ otherwise}\,\}\\ \mathcal{J}_2(Y) &= \text{ indicator of } \{\, Y \mid \textstyle\sum_\omega Y(\omega) = 1\},\\ \mathcal{J}_3(Y) &= \text{ max of } Y(\omega)/P_0(\omega) \;\text{ for }\; \omega \in \operatorname{supp} P_0.\end{aligned}$$

The conjugate is given by infimal convolution of the conjugates [19, Theorem 16.4],

$$\mathcal{R}(X) = 1 + \min\{\, \mathcal{J}_1^*(X_1) + \mathcal{J}_2^*(X_2) + \mathcal{J}_3^*(X_3) \,|\, X_1 + X_2 + X_3 = X\}, \tag{3.17}$$

where the minimum is sure to be attained because $\mathcal{R}(X)$ is finite and the functions $\mathcal{J}_i^*$, being conjugate to polyhedral convex functions $\mathcal{J}_i$, are themselves polyhedral convex [19, Theorem 19.2]. The conjugates on $\mathbf{L}(\Omega)$ are

$$\begin{array}{ll} \mathcal{J}_1^*(X_1) = & \text{indicator of } \{\, X_1 \,|\, X_1(\omega) \le 0 \text{ for } \omega \in \operatorname{supp} P_0\}, \\ \mathcal{J}_2^*(X_2) = & C \text{ when } X_2 \equiv C \text{ but } = \infty \text{ otherwise}, \\ \mathcal{J}_3^*(X_3) = & \text{indicator of } S = \{\, X_3 \,|\, X_3(\omega) \ge 0 \text{ for } \omega \in \operatorname{supp} P_0,\ E_{P_0}[X_3] = 1\}, \end{array} \tag{3.18}$$

so (3.17) refers to minimizing $1 + C$ over all C and X_1 such that $X - X_1 - C \in S$ with $X_1 \le 0$ on $\operatorname{supp} P_0$. In terms of $X' = X - X_1$, we can characterize the feasible values of C as the ones for which there exists $X' \ge X$ on $\operatorname{supp} P_0$ such that also $X' \ge C$ on $\operatorname{supp} P_0$ and $E_{P_0}[X'] = 1 + C$. That existence obviously corresponds to $E_{P_0}[\max\{X, C\}] \le 1 + C$, which can be written as $E_{P_0}[\max\{X - C, 0\}] \le 1$. Thus in (3.17) we are minimizing $1 + C$ subject to that inequality. The left side of the inequality is a function of C that has the constant value 0 when $C \ge \max X$, but grows strictly from that toward ∞ as C decreases toward $-\infty$. Then there is a unique value of C for which the inequality is an equation, and that C furnishes the minimum in (3.17). □

4 Supporting formulas for divergence-based robustness

Stochastic divergences have been valuable in distributionally robust optimization for providing ambiguity sets as neighborhoods of a nominal distribution P_0. Working with the generalization of stochastic divergence in Definiton 3.4, we can investigate this on a much broader level than ever before (although here just for finite Ω). The neighborhoods

$$\mathcal{P}_\beta = \{\, P \in \mathbf{P}(\Omega) \,|\, \mathcal{I}(P\|P_0) \le \beta\} \text{ for } 0 < \beta < \infty, \tag{4.1}$$

are closed convex sets by (I1) and (I2) which have only P_0 in their intersection by (I3), yet through (I4) contain more than just P_0. They form a nest which strictly increases with respect to β up to some limit, namely

$$\bar\beta = \sup\{\, \beta \,|\, \exists\, P,\ \mathcal{I}(P\|P_0) = \beta\} \in (0, \infty], \tag{4.2}$$

after which they evermore coincide. Their union, by (I4), encompasses, among the distributions P that can be represented by densities with respect to P_0, all but perhaps some extreme cases. This nest gives rise to a "spectrum" of coherent measures of risk in the basic sense as the worst-case functionals

$$\mathcal{R}_\beta(X) = \max_{P \in \mathcal{P}_\beta} E_P[X] \text{ for } \mathcal{P}_\beta,\ \beta \le \bar\beta, \text{ as in (4.1)–(4.2)}, \tag{4.3}$$

which interestingly produces parallels to the familiar CVaR spectrum of risk measures that was observed in (3.12).

The task now is developing formulas of *minimization* type for the functionals $\mathcal{R}_\beta$ which offer alternatives to the maximization in (4.3) and can better be integrated into computational schemes in distributionally robust optimization. A fundamental rule comes first.

Theorem 4.1 (neighborhood risk from dualized risk)**.** *Relative to the risk measure $\mathcal{R}$ in Theorem 3.5 that dualizes a stochastic divergence $\mathcal{I}(P\|P_0)$ for a given P_0, the worst-case functionals $\mathcal{R}_\beta$ in (4.3) satisfy*

$$\mathcal{R}_\beta(X) = \inf_{0<\lambda<\infty}\Big\{\,\lambda\beta + \lambda\mathcal{R}(\lambda^{-1}X\,)\Big\}. \tag{4.4}$$

Here the infimum over $(0,\infty)$ can be sharpened to a minimum over $[0,\infty]$ by interpreting $\lambda\mathcal{R}(\lambda^{-1}X)$ for $\lambda = 0$ as the recession function $\mathcal{R}0^+(X)$, its limit as $\lambda\searrow 0$.

Proof. This just invokes the standard formula in [19, Theorem 13.5] for support functions of level sets of a convex function, here $\mathcal{J}(P) = \mathcal{I}(P\|P_0)$, in terms of the conjugate function, here $\mathcal{R}$. □

Applying this rule to Wasserstein divergence by way of its dualization in Example 3.6 we get

$$\mathcal{R}_\beta(X) = \min_{\lambda\ge 0}\Big[\lambda\beta + \textstyle\sum_{\omega'}\max_{\omega\in\Omega}\{\,X(\omega) - \lambda W(\omega,\omega')\}P_0(\omega')\Big], \tag{4.5}$$

as was observed in [13, (12b)] and noted again in [15, Theorem 1]. This can also be cast as a linear programming formula:

$$\mathcal{R}_\beta(X) = \min\{\,\lambda\beta + E_{P_0}[\Lambda]\,|\,\lambda\ge 0,\ \Lambda(\omega')\ge X(\omega) - \lambda W(\omega,\omega'),\,\forall\omega,\,\omega'\}. \tag{4.6}$$

Applying the rule instead to Kullbach-Leibler divergence by way of its dualization in Example 3.7, yields

$$\mathcal{R}_\beta(X) = \inf_{0<\lambda<\infty}\lambda\Big[\beta + \log E_{P_0}[\exp(\lambda^{-1}X)]\Big], \tag{4.7}$$

in harmony with the results of Ahmadi-Javid [2] about his EVaR, entropic value-at-risk.

For CVaR divergence, we already know from (3.12) that the rule in Theorem 4.1 must lead to $\mathcal{R}_\beta$ being the risk measure P_0-CVaR$_\alpha$ where $1-\alpha$ and $1+\beta$ are reciprocals to each other, but it's instructive to see how that comes out through (4.4) with the $\mathcal{R}$ in Example 3.8. With C' in place of C in the formula for $\mathcal{R}$ in (3.16), we are minimizing in (4.4) $\lambda(\beta+1+C')$ over $\lambda>0$ and $C'\in\mathbb{R}$ subject to the constraint $E_{P_0}[\max\{\lambda^{-1}X - C', 0\}] = 1$. Introducing C as $\lambda C'$, this transforms to minimizing $\lambda(1+\beta)+C$ over $\lambda>0$ and $C\in\mathbb{R}$ subject to $E_{P_0}[\max\{X-C,0\}] = \lambda$. That reduces to minimizing $E_{P_0}[\max\{X-C,0\}](1+\beta)+C$ over $C\in\mathbb{R}$, hence to

$$\min_C\Big\{\,C + \frac{1}{1-\alpha}E_{P_0}\Big[\,\max\{X-C,0\}\Big]\,\Big\}\ \text{ where }\ 1-\alpha = (1+\beta)^{-1}. \tag{4.8}$$

That's in fact the alternative CVaR$_\alpha$ formula that was brought to light in [24] and refined to allow for probability atoms in [25].

Risk formulas involving a trade-off between C and something about $X-C$, as in (4.8), are "regret" formulas in the quadrangle theory of [26]. They take the form of minimizing $C + \mathcal{V}(X-C)$ for a functional $\mathcal{V}$ acting as a "measure of regret." The interpretation is that, by accepting an immediate loss of cost or loss of the amount C, it's only the residual loss $X-C$ that remains uncertain, and $\mathcal{V}$ assesses the perceived present impact of that future prospect. Similar formulas were explored by Ben-Tal and Teboulle in 2007 [9] in connection with obtaining a "certainty equivalent" as the optimal C, but with a focus on expected utility. This connection is explained in [26] in terms of regret being a sort of anti-utility, but not limited to an expectation form, and that's what we develop here next. An extended philosophical discussion of regret and utility from such a general perspective is available in [21].

Definition 4.2 (coherent measures of regret)**.** *By a coherent measure of regret will be meant a functional $\mathcal{V}$ on $\mathbf{L}(\Omega)$ with values in $(-\infty,\infty]$ that satisfies:*

(V1) *$\mathcal{V}$ is convex with closed level sets $\{\,X\,|\,\mathcal{V}(X)\le\gamma\}$, $\gamma<\infty$,*

(V2) *$\mathcal{V}(C)\ge C$, and the interval $\{\,C\,|\,\mathcal{V}(C)=C\}$ is bounded, containing 0,*

(V3) *$\mathcal{V}(X)\le\mathcal{V}(X')$ when $X\le X'$.*

It is positively homogeneous if also

(V4) *$\mathcal{V}(\lambda X)=\lambda\mathcal{V}(X')$ for $\lambda>0$.*

Theorem 4.3 (deriving risk from regret)**.** *For any coherent measure of regret $\mathcal{V}$, the formula*

$$\mathcal{R}(X)=\min_C\{\,C+\mathcal{V}(X-C)\},\tag{4.9}$$

where "min" signals that the infimum is surely attained, gives a measure of risk $\mathcal{R}$ that is coherent in the general sense. It is sure to be coherent in the basic sense if $\mathcal{V}$ satisfies the additional axiom (V4).

If $\mathcal{V}$ is a measure of regret yielding in (4.9) the risk measure $\mathcal{R}$ that dualizes a stochastic divergence $\mathcal{I}(\cdot\|P_0)$ in the manner of Theorem 3.5, then the associated neighborhood risk measures (4.3) satisfy

$$\mathcal{R}_\beta(X)=\min_C\{\,C+\mathcal{V}_\beta(X-C)\}\tag{4.10}$$

for the coherent and positively homogeneous measures of regret derived from $\mathcal{V}$ by

$$\mathcal{V}_\beta(X)=\inf_{\lambda>0}\Big\{\,\lambda\Big[\beta+\mathcal{V}(\lambda^{-1}X)\Big]\,\Big\}.\tag{4.11}$$

Here the infimum over $(0,\infty)$ can be sharpened to a minimum over $[0,\infty]$ by interpreting $\lambda\mathcal{V}(\lambda^{-1}X)$ for $\lambda=0$ as the recession function $\mathcal{V}0^+(X)$, which is its limit as $\lambda\searrow 0$.

Proof. The formula in (4.9) expresses $\mathcal{R}$ as resulting from inf-convolution of $\mathcal{V}$ with the convex function $\mathcal{W}$ that has $\mathcal{W}(X)=C$ if $X\equiv C$ but $\mathcal{W}(X)=\infty$ otherwise. By (V1), this operation is covered by [19, Corollary 9.2.2]. It says that $\mathcal{R}$ will be closed proper convex with attainment of the minimum under the recession function condition that $\mathcal{V}0^+(X)+\mathcal{W}0^+(-X)>0$ for all $X\ne 0$. Here $\mathcal{W}0^+=\mathcal{W}$, because $\mathcal{W}$ is positively homogeneous, so $\mathcal{W}0^+(-X)=-C$ if $X\equiv C$, but otherwise $\mathcal{W}0^+(-X)=\infty$. The condition thus comes down to requiring $\mathcal{V}0^+(C)>C$ for $C\ne 0$. That's guaranteed by (V2). It's easily seen that (4.9) preserves the monotonicity in (V3). Next, application of Theorem 4.1 to $\mathcal{R}$ expressed by (4.9) leads to

$$\mathcal{R}_\beta(X)=\inf_{\lambda>0,C}\Big\{\,\lambda\beta+\lambda C+\mathcal{V}(\lambda^{-1}X-C)\Big\}.\tag{4.12}$$

In changing variables from C to $C'=\lambda C$, this transforms to minimizing $\lambda\beta+C'+\mathcal{V}(\lambda^{-1}[X-C'])$, where the minimization in λ can be carried out first. That's the meaning of the combined formulas (4.10) and (4.11). □

Example 4.4 (CVaR regret)**.** *A coherent measure of regret $\mathcal{V}$ that yields in (4.9) the risk measure $\mathcal{R}$ in Example 3.8 that dualizes CVaR divergence is given by*

$$\mathcal{V}(X)=E_{P_0}[\max\{X,0\}]\ \ \textit{if}\ \ E_{P_0}[\max\{X,0\}]\le 1,\ \ \textit{otherwise}\ \ \mathcal{V}(X)=\infty.\tag{4.13}$$

The corresponding regret measures in (4.11) for the risk measures $\mathcal{R}_\beta$, already identified as P_0-CVaR risk measures (3.12) are

$$\mathcal{V}_\beta(X)=(1-\alpha)^{-1}E_{P_0}[\max\{X,0\}],\ \ \textit{where}\ \ 1-\alpha=(1+\beta)^{-1},\tag{4.14}$$

so that (4.10) corresponds to the formula already seen in (4.8).

Detail. With (4.13) giving $\mathcal{V}$, (4.9) concerns the minimization of $C + E_{P_0}[\max\{X - C, 0\}]$ subject to $E_{P_0}[\max\{X - C, 0\}] \leq 1$, but that's attained when $E_{P_0}[\max\{X - C, 0\}] = 1$. Thus (4.9) says $\mathcal{R}(X) = 1+C$ for C making $E_{P_0}[\max\{X-C, 0\}] = 1$, and that agrees with the formula in Example 3.8. Turning to (4.11) in the case of the $\mathcal{V}$ in (4.13), we face the minimization of $\lambda[\beta + E_{P_0}[\max\{\lambda^{-1}X, 0\}]$ subject to $E_{P_0}[\max\{\lambda^{-1}X, 0\}] \leq 1$, which means $E_{P_0}[\max\{X, 0\}] \leq \lambda$. The minimum value is therefore $(1 + \beta)E_{P_0}[\max\{X, 0\}]$ as claimed in (4.14). □

Example 4.5 (Kullbach-Leibler regret)**.** *The risk measure $\mathcal{R}$ associated with Kullbach-Leibler divergence in (3.15) fits (4.9) with*

$$\mathcal{V}(X) = E_{P_0}[\exp X] - 1. \tag{4.15}$$

Correspondingly then in (4.11),

$$\mathcal{V}_\beta(X) = \inf_{\lambda>0} \{\, \lambda[\beta + E_{P_0}[\exp \lambda^{-1}X] - 1\}, \tag{4.16}$$

and therefore in (4.10)

$$\mathcal{R}_\beta(X) = \inf_{\lambda>0,\, C} \Big\{\, C + \lambda\Big[\beta + E_{P_0}[\exp \lambda^{-1}(X - C)] - 1\Big]\Big\}. \tag{4.17}$$

Detail. The regret measure in (4.16) was recorded already in [26, Example 8]. The rest just follows from the prescriptions in Theorem 4.3. □

For Wasserstein divergence, a suitable choice of a measure of regret hasn't been determined. More about that will be explained after Theorem 4.6.

Every coherent measure of risk $\mathcal{R}$ can be expressed as in (4.9) for some coherent measure of regret $\mathcal{V}$, but there's no uniquess to $\mathcal{V}$ and the real challenge is determining a $\mathcal{V}$ that's *natural* for $\mathcal{R}$ and somehow useful in adding information. The construction platform emerges through duality.

Theorem 4.6 (dualization of regret)**.** *Coherent measures of regret $\mathcal{V}$ correspond one-to-one, through the formula*

$$\mathcal{V}(X) = \sup_{Y \geq 0} \{\, \langle X, Y\rangle - \mathcal{K}(Y)\}, \tag{4.18}$$

with the functionals $\mathcal{K}$ on $\mathbf{L}_+(\Omega) = \{\, Y \in \mathbf{L}(\Omega) \,|\, Y \geq 0\}$ that satisfy:

(K1) *$\mathcal{K}$ is convex with values in $[0, \infty]$,*

(K2) *the level sets $\{\, Y \geq 0 \,|\, \mathcal{K}(Y) \leq \beta\}$ are closed for $\beta \in [0, \infty)$,*

(K3) *$\min \mathcal{K} = 0$ and $\operatorname{argmin} \mathcal{K}$ meets $\mathcal{P}(\Omega)$,*

(K4) *$\operatorname{dom} \mathcal{K}$ in $\mathbf{L}_+(\Omega)$ contains Y with $\sum_\omega Y(\omega) > 1$ and Y with $\sum_\omega Y(\omega) < 1$.*

In this correspondence, $\mathcal{V}$ produces the risk measure $\mathcal{R}$ if and only if the graduator $\mathcal{J}$ on $\mathbf{P}(\Omega)$ that dualizes $\mathcal{R}$ in Theorem 2.5 is the restriction of $\mathcal{K}$ to $\mathbf{P}(\Omega)$. The case where $\mathcal{J}(P) = \mathcal{I}(P\|P_0)$ for a stochastic divergence as in Definition 3.4, requiring also

(K5) *$\mathbf{P}(\Omega) \cap \operatorname{argmin} \mathcal{K} = \{P_0\}$ and $\operatorname{cl}[\operatorname{dom} \mathcal{K}] \supset \{\, P \in \mathbf{P}(\Omega) \,|\, \operatorname{supp} P \subset \operatorname{supp} P_0\}$,*

is the case where $\mathcal{V}$ further satisfies:

(V5) *$\inf_C\{\, C + \mathcal{V}'(0; X - C)\} = E_{P_0}[X]$ for the directional derivative function $\mathcal{V}'$,*

(V6) *$\lim_{\tau \nearrow \infty} \mathcal{V}(\tau X)/\tau \geq P_0\text{-}\max X := \max\{\, X(\omega) \,|\, \omega \in \operatorname{supp} P_0\}$.*

Proof. In taking $\mathcal{K}(Y)$ to be ∞ for Y outside of $\mathbf{L}_+(\Omega)$, we get $\mathcal{K}$ to be a closed proper convex function on $\mathbf{L}(\Omega)$ having $\mathcal{V}$ as its conjugate, and therefore reciprocally

$$\mathcal{K}(Y) = \sup_X \{\, \langle X, Y\rangle - \mathcal{V}(X)\}. \tag{4.19}$$

The recession function $\mathcal{V}0^+$ is then the support function of $\mathrm{cl}[\mathrm{dom}\,\mathcal{K}]$ [19, Theorem 13.3], so having $\mathrm{dom}\,\mathcal{K} \subset \mathbf{L}_+(\Omega)$ means having $\mathcal{V}0^+ \leq$ the support function of $\mathbf{L}_+(\Omega)$, which is the indicator of $\mathbf{L}_-(\Omega)$, or in other words, $\mathcal{V}0^+(X') \leq 0$ when $X' \leq 0$. That corresponds to the monotonicity of $\mathcal{V}$ in (V3), because $\mathcal{V}0^+(X') \leq \alpha$ if and only if $\mathcal{V}(X + X') \leq \mathcal{V}(X) + \alpha$ for all X [19, Theorem 8.5].

It was observed in the proof of Theorem 4.3 that (4.9) represents $\mathcal{R}$ as coming from inf-convolution of $\mathcal{V}$ with the convex function $\mathcal{W}$ having

$$\mathcal{W}(X) = C \text{ if } X \equiv C, \text{ but } W(X) = \infty \text{ otherwise.} \tag{4.20}$$

In duality, that means the function conjugate to $\mathcal{R}$, namely its dualizing graduator $\mathcal{J}$, is the sum of the conjugates $\mathcal{V}^*$ and $\mathcal{W}^*$, with $\mathcal{V}^* = \mathcal{K}$ and $\mathcal{W}^*$ being the indicator of the hyperplane

$$\mathcal{H} = \Big\{ Y \,\Big|\, \sum\nolimits_\omega Y(\omega) = 1 \Big\}. \tag{4.21}$$

This tells us that $\mathcal{V}$ gives us $\mathcal{R}$ if and only if $\mathcal{J}$ is obtained by adding the indicator to $\mathcal{H}$ to $\mathcal{K}$, which in terms of domains amounts to restricting $\mathcal{K}$ to $\mathbf{P}(\Omega)$.

From that, (K3) is confirmed as a consequence of the properties specified for $\mathcal{V}$. Conversely, since $\mathrm{argmin}\,\mathcal{K}$ is in conjugacy the subgradient set $\mathcal{V}$, while $\mathcal{V}(0) = \min \mathcal{K}$, (K3) provides some $P \in \mathbf{P}(\Omega)$ such that $\mathcal{V}(X) \geq \langle X, P\rangle = E_P[X]$ for all X. Then in particular, $\mathcal{V}(C) \geq C$ for all constants C.

What about (K4), which requires the existence of elements $Y \in \mathrm{dom}\,\mathcal{K}$ on both sides of the hyperplane $\mathcal{H}$? Because $\mathcal{V}0^+$ is the support function of $\mathrm{cl}[\mathrm{dom}\,\mathcal{K}]$, that's captured by requiring both $\mathcal{V}0^+(1) > 0$ and $\mathcal{V}0^+(-1) > 0$. But that's also equivalent to the condition in (V2) about the C interval there being bounded.

All that remains is confirming the extra conditions associated with stochastic divergence. The properties of $\mathcal{K}$ in (K5) obviously correspond, in restricting $\mathcal{K}$ to get $\mathcal{J}$, to the extra properties $\mathcal{J}$ must have beyond Definition 2.4 to have the form $\mathcal{I}(\cdot\|P_0)$ in Definition 3.4. The claim is that these properties dualize to (V5) and (V6).

The infimum in (V5) represents the inf-convolution of the $\mathcal{W}$ function in (4.20) and the directional derivative function $\mathcal{V}'(0,\cdot)$. The first is the support function of the hyperplane $\mathcal{H}$ in (4.15), while the second is the support function of the subgradient set $\partial\mathcal{V}(0)$ (inasmuch as $\mathcal{V}$ is finite and closures aren't needed) [19, Theorem 23.2]. Through conjugacy, that subgradient set is $\mathrm{argmin}\,\mathcal{K}$ [19, Theorem 23.5]. Inf convolution of the support functions of two convex sets produces the support function of their intersection, at least under finiteness, as here [19, Corollary 16.4.1]. Thus, (V5) says $\mathcal{H} \cap \mathrm{argmin}\,\mathcal{K} = \{P_0\}$. That's the same as the argmin condition in (K5), inasmuch as $\mathrm{dom}\,\mathcal{K} \subset \mathbf{L}_+(\Omega)$.

In (V6), the max gives the support function of the set $\{\, P \in \mathbf{P}(\Omega) \,|\, \mathrm{supp}\, P \subset \mathrm{supp}\, P_0\}$, while the limit gives $\mathcal{V}0^+(X)$ [19, Theorem 8.5]. Because $\mathcal{V}0^+$ is the support function of $\mathrm{cl}[\mathrm{dom}\,\mathcal{K}]$, the inequality is equivalent to the inclusion in (K5). That completes the proof. □

Theorem 4.6 provides a full and clear picture of how to get a coherent measure of regret $\mathcal{V}$ that can serve through formula (4.10) in giving an alternative minimization description of the neighborhood robustness functionals (4.3) associated with a general stochastic divergence $\mathcal{I}(\cdot\|P_0)$ by way of (4.10) and (4.11). First, extend $\mathcal{J} = \mathcal{I}(\cdot\|P_0)$ as a functional on $\mathbf{P}(\Omega)$ satisfying (I1)–(I4) to a functional $\mathcal{K}$ on $\mathbf{L}_+(\Omega)$ satisfying (K1)–K(5). The key points in that are making sure that the extended function still has minimum value 0 and that its effective domain contains elements on both sides of the hyperplane $\mathcal{H}$ in (4.21). Second, determine $\mathcal{V}$ from $\mathcal{K}$ by (4.18).

It's evident that the prescribed extension of $\mathcal{J} = \mathcal{I}(\cdot\|P_0)$ is always possible and by no means unique. As in finding a good regret measure for a given risk measure, the real goal is coming up with an extension $\mathcal{K}$ that is "natural," It should be convenient for carrying out the dualization to $\mathcal{V}$ and having that $\mathcal{V}$ furnish a practical expression for use in (4.10)–(4.11).

Can the prescription in Theorem 4.6 also be followed to get a regret $\mathcal{V}$ for Wasserstein divergence? The trouble there is that it's hard to see how the formula (3.2) for Wasserstein divergence can be extended "naturally" beyond $\mathbf{P}(\Omega)$ to $\mathbf{L}_+(\Omega)$, because having $Y(\omega) = \sum_{\omega'} \Pi(\omega, \omega')$ for a function $\Pi(\omega, \omega') \geq 0$ with $\sum_\omega \Pi(\omega, \omega') = P_0(\omega')$ makes Y have to be some $P \in \mathbf{P}(\Omega)$.

Here's a good-looking way of extending $\mathcal{J}$ to $\mathcal{K}$ that follows all the rules and can always be employed, but may offer less help in the end than might be wished. Take

$$\mathcal{K}(Y) = \lambda \mathcal{J}(\lambda^{-1} Y) \text{ for } \lambda = \sum\nolimits_\omega Y(\omega) \text{ if } Y \geq 0,\, Y \not\equiv 0, \text{ and } \mathcal{K}(0) = 0. \tag{4.22}$$

In this case $\mathcal{K}$ is positively homogeneous and vanishes along the ray $\{\, \tau P_0 \,|\, \tau \geq 0\}$, in consequence of $\mathcal{J}$ having minimum 0 attained at P_0. That makes $\mathcal{V}$ be the indicator of a closed convex set $\mathcal{C}$ having the origin on its boundary and P_0 as a normal vector there, moreover with $X \in \mathcal{C}$ implying $X' \in \mathcal{C}$ for all $X' \leq X$. The corresponding formula (4.9) for $\mathcal{R}$ boils down then to saying that $\mathcal{R}(X)$ is the smallest C such that $X - C \in \mathcal{C}$. But $\mathcal{R}(X) \leq C$ if and only if $\mathcal{R}(X - C) \leq 0$, so this means $\mathcal{C}$ is simply the level set $\{\, X \,|\, \mathcal{R}(X) \leq 0\}$.

In other situations, there is an easy and more interesting path to follow between stochastic divergence and regret because of underlying separability in the divergence formula.

Theorem 4.7 (expectational regret and its relationship to divergence)**.** *With respect to a pair of closed proper convex functions v and k on $\mathbb{R}$ that are conjugate to each other:*

$$v(x) = \sup_y \{\, xy - k(y)\}, \qquad k(y) = \sup_x \{\, xy - v(x)\}, \tag{4.23}$$

the functionals $\mathcal{V}$ on $\mathbf{L}(\Omega)$ having the form

$$\mathcal{V}(X) = E_{P_0}[v(X)] \tag{4.24}$$

with respect to a given distribution P_0 are conjugate to the functionals $\mathcal{K}$ on $\mathbf{L}(\Omega)$ having the form

$$\mathcal{K}(Y) = \begin{cases} \sum_\omega k\Big(Y(\omega)/P_0(\omega)\Big) P_0(\omega) & \text{if } Y(\omega) = 0 \text{ when } P_0(\omega) = 0, \\ \infty & \text{otherwise.} \end{cases} \tag{4.25}$$

In this relationship $\mathcal{K}$ satisfies (K1)–(K5), the properties under which its restriction to $\mathbf{P}(\Omega)$ gives a stochastic divergence $\mathcal{I}(\cdot \| P_0)$, when k satisfies

$$1 \in \operatorname{int}[\operatorname{dom} k] \subset (0, \infty),\ \min k = 0,\ \operatorname{argmin} k = \{1\}. \tag{4.26}$$

The corresponding dual properties to those, under which $\mathcal{V}$ satisfies (V1)–(V3) and (V5)–(V6), are

$$\begin{array}{l} \text{v is a nondecreasing function with } v(0) = 0 \text{ and } v'(0) = 1, \text{ hence} \\ v(x) \geq x \text{ for all } x, \text{ but having the interval } \{\, x \,|\, v(x) = x\} \text{ bounded.} \end{array} \tag{4.27}$$

Detail. In (4.24) we have $\mathcal{V}(X) = \sum_\omega V_\omega(X(\omega))$ for $V_\omega(x) = v(x) P_0(\omega)$. The conjugate functional $\mathcal{K} = \mathcal{V}^*$, obtained by maximizing $\sum_\omega X(\omega) Y(\omega) - \mathcal{V}(X)$ over all possibilities for $X(\omega)$, is given then by $\mathcal{K}(Y) = \sum_\omega V_\omega^*(Y(\omega))$. From (4.23) we have $V_\omega^*(y) = P_0(\omega) k\Big(y/P_0(\omega)\Big)$ for ω with $P_0(\omega) > 0$, but $V_\omega^* =$ indicator of 0 for ω with $P_0(\omega) = 0$, since then $V_\omega(x) \equiv 0$. That establishes (4.25) as the formula for the conjugate.

Properties (K1) and (K2) for $\mathcal{K}$ and (V1)–(V2) for $\mathcal{V}$ are, of course, immediate from k and v being closed proper convex, and (V2) reflects (4.27). Having $\min \mathcal{K} = 0$ with $\mathbf{P}(\Omega) \cap \operatorname{argmin} \mathcal{K} = \{P_0\}$

corresponds to having $\min k = 0$ with $\operatorname{argmin} k = \{1\}$, in which case actually $\operatorname{argmin} \mathcal{K} = \{P_0\}$. That takes care of (K3) and (K5), because the inclusion in (K5) is automatic for this form of $\mathcal{K}$. Finally, having (K4) corresponds to having the interval $\operatorname{dom} k$ lie in $[0, \infty)$ with 1 belonging to its interior. The k properties in (4.26) thus fill the role that was claimed for them.

We could next go through the properties (V3),(V5), (V6), and determine what they demand of v, but there's a shortcut. We already know that (V1)–(V3) and (V5)–(V6) dualize (K1)–(K5), so it suffices to observe that the properties of v in (4.27) dualize those of k in (4.26). In fact, by the elementary rules for the one-dimensional conjugacy in (4.23), v nondecreasing corresponds to $\operatorname{dom} k \subset [0, \infty)$, while

$$-v(0) = \inf k, \;\; \partial v(0) = \operatorname{argmin} k, \;\; -k(1) = \inf\{\, v(x) - x\}, \;\; \partial k(1) = \operatorname{argmin}\{\, v(x) - x\},$$

with $\partial v(0)$ reducing to $\{1\}$ if and only if v is differentiable at 0 with $v'(0) = 1$, and on the other hand, $\partial k(1)$ bounded if and only if $1 \in \operatorname{int}[\operatorname{dom} k]$. □

Example 4.8 (φ-divergences)**.** *These divergences, defined in (3.7) with the basic assumptions on φ taken to be the normalized ones in (3.8), fit into the framework of Theorem 4.7 to the specialization to $k = \varphi$ with (3.8) strengthened to insist on* $\operatorname{argmin} \varphi = \{1\}$.

The condition in (3.8) that $\operatorname{dom} \varphi$ *be all of* $[0, \infty)$ *corresponds to requiring $v(x)/x \to \infty$ as $x \to \infty$). Imposing on φ the additional assumption that $\varphi(q)/q \to \infty$ as $q \to \infty$ corresponds to requiring* $\operatorname{dom} v = (-\infty, \infty)$. *On the other hand imposing strict convexity on φ corresponds to requiring differentiability everywhere of v.*

Example 4.8 confirms that in the case of a φ-divergence $\mathcal{I}(P\|P_0) = E_{P_0}[\varphi(Q)]$ as in (3.7) and the neighborhood risk measures $\mathcal{R}_\beta$ derived from it via (3.1), we have the alternative formula

$$\mathcal{R}_\beta(X) = \inf_{\lambda > 0,\, C} \Big\{\, C + \lambda\Big[\beta + E_{P_0}\Big[v\Big(\lambda^{-1}[X - C]\Big)\Big]\Big]\Big\} \tag{4.28}$$

coming from Theorem 4.3 as the combination of (4.10) and (4.11). The explorations of φ-divergence by Ahmadi-Javid [1], [2], and by Dommel and Pichler [12], reached this same formula as an all-in-one result, whereas we have taken pains to separate the pattern of derivation into the (4.10) part as "immediate from convex analysis" and the trickier (4.11) part. It may be recalled from Section 3, though, that the assumptions of Dommel and Pichler in [12] also brought in the condition about $\varphi(q)/q$ while, in normalized form, they allowed more than just 1 to be in $\operatorname{argmin} \varphi$.

The work of Breuer and Csiszár [10] on φ-divergences relied on *strict* convexity of φ, but on the other hand allowed for $\varphi(\omega, q)$ instead of just $\varphi(q)$, producing divergences that in our notation here come out as

$$\mathcal{I}(P\|P_0) = \sum_{\omega \in \operatorname{supp} P_0} \varphi(\omega, Q(\omega))P_0(\omega) \;\text{ for }\; Q(\omega) = P(\omega)/P_0(\omega) \quad \text{when} \;\; \operatorname{supp} P \subset \operatorname{supp} P_0, \;\; \text{but } \infty \text{ otherwise.} \tag{4.29}$$

Theorem 4.7 can easily be extended this by taking conjugate functions $v(\omega, \cdot)$ and $k(\omega, \cdot)$ in (4.23).

5 Integrating stochastic divergences into the risk quadrangle

From now on, we operate with a fixed distribution P_0 having $\operatorname{supp} P_0 = \Omega$. This puts us in the framework where every $X \in \mathbf{L}(\Omega)$ can be interpreted as a random variable with its cumulative distribution function F_{X,P_0} in (1.1) being written as F_X, and with its expectation $E_{P_0}[X]$ and highest value

$\max_{P_0} X$ just as $E[X]$ and $\max X$. Despite it being fixed, we view P_0 as a nominal distribution subject perhaps to uncertainty. It's the focus of our continued treatment of ambiguity and distributional robustness.

Everything about stochastic divergences comes out simpler now, with $\mathcal{I}(P\|P_0)$ just $\mathcal{I}(P)$, say, but we can only compare P_0 to distributions P representable by a density Q with respect to P_0. In that, we forgo a capability of Wasserstein divergence, in particular, but open the way to other capabilities that stochastic divergences might provide. We examine the possible connections they have to the *fundamental quadrangle of risk* in the pattern

$$\begin{array}{ccccc} & & \text{risk } \mathcal{R} & \longleftrightarrow & \mathcal{D} \text{ deviation} \\ \text{optimization} & & \uparrow\downarrow & \mathcal{S} & \downarrow\uparrow & & \text{estimation} \\ & & \text{regret } \mathcal{V} & \longleftrightarrow & \mathcal{E} \text{ error} \end{array}$$

So far, we have effectively only been looking at the optimization side of this quadrangle, where a measure of risk $\mathcal{R}$ can be obtained from a measure of regret $\mathcal{V}$. On the other side a *measure of error* $\mathcal{E}(X)$ assesses the *nonzeroness* of X, and a *measure of deviation* $\mathcal{D}$ assesses the *uncertainty* of X as its nonconstancy. The two sides are related by

$$\mathcal{D}(X) = \mathcal{R}(X) - E[X], \qquad \mathcal{E}(X) = \mathcal{V}(X) - E[X], \tag{5.1}$$

with respect to which

$$\mathcal{D}(X) = \inf_C \mathcal{E}(X - C) \;\text{ corresponds to }\; \mathcal{R}(X) = \inf_C\{\, C + \mathcal{V}(X - C)\}. \tag{5.2}$$

This pairing of $\mathcal{R}$ and $\mathcal{V}$ with $\mathcal{D}$ and $\mathcal{E}$ may seem a tiny distinction to make but has profound consequences in affording a different perspective.

As a baseline for the relationships between $\mathcal{R}$, $\mathcal{V}$, $\mathcal{D}$ and $\mathcal{E}$ in our context of ambiguity concerns centered on P_0, we can take for granted now that

$$\begin{array}{l} \mathcal{R}(X) \geq E[X] \;\text{ for all } X, \\ \mathcal{V}(X) \geq E[X] \;\text{ for all } X, \\ \mathcal{D}(X) \geq 0 \;\text{ for all } X, \\ \mathcal{E}(X) \geq 0 \;\text{ for all } X. \end{array} \tag{5.3}$$

For positively homogeneous $\mathcal{R}$, the first inequality just corresponds to P_0 belonging to the risk envelope $\mathcal{P}$ associated with $\mathcal{R}$ in Theorem 2.3, and without homogeneity its membership in $\operatorname{argmin} \mathcal{J}$ for the dualizing graduator in Theorem 2.5. The second inequality is a natural counterpart in relation to the infimum in (5.2), and the rest then is automatic from (5.1). But a stricter version of these properties was deemed necessary in developing the quadrangle in [26].

Definition 5.1 (aversity)**.** *The functionals $\mathcal{R}$, $\mathcal{V}$, $\mathcal{D}$, $\mathcal{E}$, are averse when*

$$\begin{array}{l} \mathcal{R}(X) \geq E[X] \;\text{ for all } X, \text{ with} > \text{when } \; X \not\equiv C, \\ \mathcal{V}(X) \geq E[X] \;\text{ for all } X, \text{ with} > \text{when } \; X \not\equiv 0, \\ \mathcal{D}(X) \geq 0 \;\text{ for all } X, \text{ with} > \text{when } \; X \not\equiv C, \\ \mathcal{E}(X) \geq 0 \;\text{ for all } X, \text{ with} > \text{when } \; X \not\equiv 0. \end{array} \tag{5.4}$$

The terminology of aversity coordinates with risk aversity in stochastic optimization. In optimizing a decision by minimizing an expected cost, an allowance is being made for higher and lower cost outcomes to balance each other out. That makes sense in applications where the same situation is

faced over and over again, and consequences over the long run are the appropriate focus. But in many applications the prospect of an unusually high loss outweighs most other prospects. Thus, the risk-neutral risk measure $\mathcal{R}(X) = E[X]$ may need to be replaced by one satisfying the first of the aversity conditions in (5.4), according to which the difference between $\mathcal{R}(X)$ and $E[X]$ for an uncertain X should be a positive *risk premium*. It's evident from (5.1) and (5.2) that the $\mathcal{V}$ aversity in (5.4) implies the $\mathcal{R}$ aversity and is equivalent to the $\mathcal{E}$ aversity. That in turn implies the $\mathcal{D}$ diversity, which is equivalent to the $\mathcal{R}$ aversity.

The properties of $\mathcal{E}$ and $\mathcal{D}$ are crucial to understanding their side of the quadrangle. The minimization formula for $\mathcal{D}$ in (5.2) seeks the constant C that best represents the random variable X with respect to the error in the difference $X - C$, as measured by $\mathcal{E}$. The $\mathcal{S}$ in the quadrangle refers to $\mathcal{S}(X)$ being the argmin of $\mathcal{E}(X - C)$ in C, that being called the *statistic* associated with X by $\mathcal{E}$. Much about it, with lots of examples, is in the paper [26] and needn't be reviewed here.

The remarkable insight provided by the optimization-statistics connection in the quadrangle is that risk-averse optimization, centered on averse $\mathcal{R}(X)$ assisted by averse regret $\mathcal{V}(X)$ *inevitably ties into a pattern error of* $\mathcal{E}(X)$*, deviation* $\mathcal{D}(X)$ *and statistic* $\mathcal{S}(X)$ *that supports generalized regression and other potentially valuable tools of analysis*, such as in [22].

The task before us is figuring out how stochastic divergences fit into this quadrangle picture. It turns out that the aversity in Definition 5.1 demands too much and would exclude the risk measures $\mathcal{R}$ that dualize divergences in Theorem 3.5. Weaker properties are needed, but they must still allow for a reasonable interpretation of error and deviation in relinquishing the strictness in (5.3). We propose the following concept of *sub*aversity.

Definition 5.2 (subaversity)**.** *The functionals $\mathcal{R}$, $\mathcal{V}$, $\mathcal{D}$, $\mathcal{E}$, are subaverse when*

$$\begin{array}{ll} \text{(R0)} & \mathcal{R}(X) \geq E[X], \text{ and } \forall X \not\equiv C,\ \exists \lambda > 0 \text{ with } \mathcal{R}(\lambda X) > \lambda E[X], \\ \text{(V0)} & \mathcal{V}(X) \geq E[X], \text{ and } \forall X \not\equiv 0,\ \exists \lambda > 0 \text{ with } \mathcal{V}(\lambda X) > \lambda E[X], \\ \text{(D0)} & \mathcal{D}(X) \geq 0, \text{ and } \forall X \not\equiv C,\ \exists \lambda > 0 \text{ with } \mathcal{D}(\lambda X) > 0, \\ \text{(E0)} & \mathcal{E}(X) \geq 0,\ \forall X \not\equiv 0,\ \exists \lambda > 0 \text{ with } \mathcal{E}(\lambda X) > \lambda E[X]. \end{array} \tag{5.5}$$

Note in (R0) that, if $\mathcal{R}(\lambda X) = \lambda E[X]$ when $\lambda > 0$, the condition reduces to requiring $\mathcal{R}(X) > E[X]$ for nonconstant X. Similarly tor (V0), (D0) and (E0). Thus, subadversity only differs from aversity for functionals that are not positively homogeneous!

Theorem 5.3 (aversity and subaversity from stochastic divergences)**.** *For a stochastic divergence $\mathcal{I} = \mathcal{I}(\cdot \| P_0)$ under the axioms (I1)–(I4) in Definition 3.4, the coherent risk measure $\mathcal{R}$ that dualizes it as a graduator in Theorem 2.5 and Theorem 3.5 is generally subaverse rather than averse. However, the coherent risk measures $\mathcal{R}_\beta$ in (4.3), tied to the nested divergence neighborhoods, are indeed averse.*

The regret measures $\mathcal{V}$ coming from an extension of $\mathcal{I}$ to a functional $\mathcal{K}$ as described in Theorem 4.7 likewise are generally just subaverse. The same also then for the associated error measure $\mathcal{E}$ and deviation measure $\mathcal{D}$.

Proof. We know from Theorem 3.5 that $\mathcal{R}(X) \geq E[X]$ for all X. Any X such that $\mathcal{R}(\lambda X) = \lambda E[X]$ for all $\lambda > 0$ has $E[X] \geq \max X$ by (3.13), but that only holds for constant X.

Similarly, because $P_0 \in \mathcal{P}_\beta$ in (4.1) we know $\mathcal{R}_\beta(X) \geq E[X]$ for all X. For X such that $\mathcal{R}_\beta(X) = E[X]$, we have from Theorem 4.7 that

$$0 = \inf_{0<\lambda<\infty} \Big\{ \lambda\beta + \Big(\lambda \mathcal{R}(\lambda^{-1} X) - E[X]\Big)\Big\}, \tag{5.6}$$

where by Theorem 3.5 the term in parentheses is nonnegative and nondecreasing with limit as $\tau \nearrow \infty$ being $\geq \max X$. Those properties make (5.5) impossible unless X is constant. Thus $\mathcal{R}_\beta$ is averse. □

The quadrangle scheme in [26] required aversity, but Theorem 5.3 makes clear that subaversity must be accommodated in order to take full advantage of the ideas surrounding stochastic divergence. Where might this lead in quadrangle modification? Coherency deserves consideration in that as well, already on the baseline in (5.3).

They translate through (5.2) to the following properties of $\mathcal{D}$ and $\mathcal{E}$, which build on the baseline in (5.3), before any upgrade to subaversity.

Proposition 5.4 (translations of coherency)**.** *Through (5.2), the risk properties (R1)–(R4) in Definition 2.1 translate to the following properties of the functional $\mathcal{D} \geq 0$:*

(D1) *$\mathcal{D}$ is convex with closed level sets, $\{\, X \,|\, \mathcal{D}(X) \leq \xi\}$, $\xi < \infty$,*

(D2) *$\mathcal{D}(X) = 0$ when $X \equiv C$,*

(D3) *$\mathcal{D}(-X) \leq E[X]$ when $X \geq 0$,*

(D4) *$\mathcal{D}(\lambda X) = \lambda \mathcal{D}(X)$ for $\lambda > 0$.*

On the other hand, the regret properties (V1)–(V4) in Definition 4.2 translate through (5.2) to the following properties of the functional $\mathcal{E} \geq 0$:

(E1) *$\mathcal{E}$ is convex with closed level sets $\{\, X \,|\, \mathcal{E}(X) \leq \xi\}$, $\xi < \infty$,*

(E2) *$\{\, C \,|\, \mathcal{E}(C) = 0\}$ is bounded, containing 0,*

(E3) *$\mathcal{E}(-X) \leq E[X]$ when $X \geq 0$,*

(E4) *$\mathcal{E}(\lambda X) = \lambda \mathcal{E}(X)$ for $\lambda > 0$.*

Proof. All this is elementary, except that (D3) and (E3) may be mysterious. Observe first that the monotonicity axiom (R3) can equivalently be expressed, in the presence of the other axioms, simply as $\mathcal{R}(-X) \leq 0$ when $X \geq 0$; this is a consequence of the facts in Proposition 2.2. But $\mathcal{R}(-X) = \mathcal{D}(-X) - E[X]$ in (5.2), so this becomes (D3). The same path leads to (E3). □

As explained after Definition 2.1, the combination of (R2) and (R3) ensures in our case of finite Ω that $\mathcal{R}$ is finite on all of $\mathbf{L}(\Omega)$. That carries over in Proposition 5.4 to (D2) and (D3) ensuring such finiteness. Then, as through convexity, $\mathcal{D}$ is continuous and the level set requirement is axiomatically fulfilled.

Definition 5.5 (coherency of error and deviation)**.** *A functional $\mathcal{D} \geq 0$ on $\mathbf{L}(\Omega)$ will be called a coherent measure of deviation, in the general sense, if it satisfies (D1), (D2) and (D3), and in the basic sense if it also satisfies (D4). A functional $\mathcal{E} \geq 0$ on $\mathbf{L}(\Omega)$ will be called a coherent measure of error if it satisfies (E1), (E2) and (E3).*

From the perspective of statistics on the right side of the quadrangle, the coherency axioms (E3) and (D3) are clearly a sharp restriction. They exclude functionals like $\mathcal{E}(X) = E[X^2]$ and $\mathcal{D}(X) = E[(X - E[X])^2] = \sigma^2(X)$ at the heart of statistical theory. This returns us, though, to the reasons why "coherency" was introduced. Popular risk measures of the form $\mathcal{R}(X) = E[X] + \gamma\sigma^2(X)$ for $\gamma > 0$, fitting the quadrangle pattern in the case of $\mathcal{E}(X) = \gamma E[X^2]$, lack monotonicity. That's such a powerful consideration for risk, but still — the quadrangle could well have serious applications in a mode where $\mathcal{E}$ and $\mathcal{D}$ are allowed to violate (E3) and (D3) at the expense of getting nonmonotone $\mathcal{R}$ and $\mathcal{V}$. That was in fact the framework adopted in the original quadrangle formulation in [26]. But our purposes here, we'll keep to coherency and what it entails. In doing this, we explore a kind of asymmetric statistics in which errors on the good side, here outcomes ≤ 0, receive a distinctive treatment through (E3).

Theorem 5.6 (The risk quadrangle featuring coherency and subaversity)**.** *The quadrangle relationships in (5.1) and (5.2) are fully coordinated and sustained when*

(R) *the risk measure $\mathcal{R}$ satisfies (R0), (R1), (R2) and (R3).*

(V) *the regret measure $\mathcal{V}$ satisfies (V0), (V1), (V2) and (V3).*
(D) *the deviation measure $\mathcal{D}$ satisfies (D0), (D1), (D2) and (D3).*
(E) *the error measure $\mathcal{E}$ satisfies (E0), (E1), (E2) and (E3).*

These properties propogate through the specified relationships, moreover with the minimum in (5.2) being attained on a nonempty closed and bounded interval $\mathcal{S}(X)$. The same holds with the addition of the axioms (R4), (V4), (D4) and (E4).

Proof. Mostly this just summarizes facts we already know from Proposition 5.4 and Theorem 4.3. Details related to subaversity are the only thing demanding attention. The coordination of subaversity between $\mathcal{R}$ and $\mathcal{D}$ through (5.1), and between $\mathcal{V}$ and $\mathcal{E}$ is apparent. We just have to confirm that subaversity passes through the minimization formulas in (5.2) with the claims about minimizing C values being correct. The two formulas are equivalent, so we can concentrate on the one for $\mathcal{D}$ in terms of $\mathcal{E}$ while observing that, in terms of recession functions,

$$\begin{array}{l}\text{subaversity of } \mathcal{E} \text{ corresponds to } \mathcal{E}0^+(X') > 0 \text{ for } X' \not\equiv 0,\\ \text{subaversity of } \mathcal{D} \text{ corresponds to } \mathcal{D}0^+(X') > 0 \text{ for } X' \not\equiv C.\end{array} \tag{5.7}$$

Posing the formula as

$$\mathcal{D}(X) = \inf_C \mathcal{F}(X,C) \;\text{ for }\; \mathcal{F}(X,C) = \mathcal{E}(X-C) \;\text{ on } \mathbf{L}(\Omega)\times I\!R, \tag{5.8}$$

we can apply [19, Theorem 9.2] with respect to the linear mapping $A : (X, C) \mapsto X$ and the recession function

$$\mathcal{F}0^+(X',C') = \mathcal{E}0^+(X'-C'). \tag{5.9}$$

According to the cited rule, if

$$\mathcal{F}0^+(X',C') \le 0,\ \mathcal{F}0^+(-X',-C') > 0 \quad\Longrightarrow\quad X' \neq 0, \tag{5.10}$$

then the minimum is sure to be attained in (5.8) and will result in having

$$\mathcal{D}0^+(X') = \min_{C'} \mathcal{F}0^+(X',C'). \tag{5.11}$$

We're assuming the subaversity of $\mathcal{E}$ in (5.7) and therefore have in (5.9) that $\mathcal{F}0^+(X', C') > 0$ unless $X' \equiv C'$, so the implication in (5.10) is true by virtue of its assumption being impossible to be satisfied. Then in (5.11) from (5.9) we have $\mathcal{D}0^+(X') = \min_{C'} \mathcal{E}0^+(X'-C')$, and that comes out 0 when $X' = C$ for some constant C, but since $\mathcal{E}0^+(X'-C') > 0$ for all C' otherwise, the attained minimum is positive otherwise. □

Finally, we would like to mention that, since this paper was written, more about fitting risk measures from φ-divergence neighborhoods into the quadrangle of risk has been put together by Malandii, Gupta, Peng and Uryasev in [18],

References

[1] Ahmadi-Javid, A., "An information-theoretic approach to constructing coherent risk measures," *2011 IEEE Internationsl Symposium on Information Theory (SIT) Proceedings* (2011), 2125–2127.

[2] Ahmadi-Javid, A., "Entropic value-at-risk: a new coherent risk measure," *Journal of Optimization Theory and Applications* **155/3** (2012), 1105–1123.

[3] Ahmadi-Javid, A., Fallah-Tafti, M. "Portfolio optimization with entropic value-at-risk," *European Journal of Operational Research* **279/3** (2019), 225-241.

[4] Ali, S.M., Silvey, S.D., "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society, Series B (Methodoligical)* **28** (1966), 131–142.

[5] Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D., "Coherent measures of risk," *Mathematical Finance* **9** (1999), 203–227.

[6] Ben-Tal, A., El Ghaoui, L., and Nemirovski, A., *Robust Optimization*, Princeton University Press, 2009.

[7] Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, Gijs, "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science* **59/2** (2013), 341–357.

[8] Ben-Tal, A., and Nemirovski, A., "Robust convex optimization," *Mathematics of Operations Research* **23** (1998), 769–805.

[9] Ben-Tal, A., and Teboulle, M., "An old-new concept of convex risk measures: The optimized certainty equivalent," *Mathematical Finance* **17/3** (2007), 449–476.

[10] Breuer, T., and Csiszár, I., "Measuring distribution model risk," *Mathematical Finance* **26** (2016), 395–411.

[11] Csiszár, I., "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten," *Publ. Math. Inst. Hungarian Acad. Sci.* **8** (1963), 85–108.

[12] Dommel, P., and Pichler, A., "Convex risk measures based on divergence," Pure and Applied Functional Analysis **6** (2021), 1157–1181.

[13] Esfahani, P. M. and Kuhn, D., "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming]/* **171/1** (2018), 115–166.

[14] Föllmer, H., and Schied, A., *Stochastic Finance*, 2nd edition, de Gruyter, New York, 2004.

[15] Gao, R., and Kleywegt, A. J., Distributionally robust stochastic optimization with Wasserstein distance," *Mathematics of Operations Research* **48/2** (2023), 603–655.

[16] Goh, J., and Sim, M., "Distributionally robust optimization and its tractable approximations," *Operations Research* **58** (2010), 902–917.

[17] Liese, F., and Vajda, I., "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory* **52/10**, 4394–4412.

[18] Malandii, A., Gupte, S., Peng, Ch., and Uryasev, S., "Risk quadrangle based on φ-divergence," working paper, Stony Brook University, 2023.

[19] Rockafellar, R.T., *Convex Analysis*, Princeton University Press, 1970.

[20] Rockafellar, R.T., "Coherent approaches to risk in optimization under uncertainty," *Tutorials in Operations Research INFORMS 2007,* 38–61.

[21] Rockafellar, R.T., "Risk and utility in the duality framework of convex analysis," Chapter 3 in *From Analysis to Visualization: A celebration of the Life and Legacy of Jonathan M. Borwein*, Springer Proceedings in Mathematics and Statistics, 2020

[22] Rockafellar, R.T., and Royset, J.O., "Measures of residual risk with connections to regression, risk tracking, surrogate models and ambiguity," *SIAM J. Optimization* **28** (2015), 1179–1208.

[23] Rockafellar, R.T., and Royset, J.O., "Superquantile/CVaR risk measures: second-order theory," *Annals of Operations Research* **262** (2018), 3–29.

[24] Rockafellar, R. T. and Uryasev, S., "Optimization of conditional value-at-risk," *J. Risk* **2** (2000), 21–42.

[25] Rockafellar, R. T. and Uryasev, S., "Conditional value-at-risk for general loss distributions," *J. Banking and Finance* **26** (2002), 1443–1471.

[26] Rockafellar, R. T. and Uryasev, S. "The fundamental risk quadrangle in risk management, optimization and statistical estimation," *Surveys in Operations Research and Management Science* **18** (2013), 33–53.

[27] Rockafellar, R. T., Uryasev, S., and Zabarankin, M., "Generalized deviations in risk analysis," *Finance and Stochastics* **10** (2006), 51–74.

[28] Ruszczyński, A., and Shapiro, A., "Optimization of convex risk functions," *Mathematics of Operations Research* **31** (2006), 433–452.

[29] Shapiro, A., "Distributionally robust stochastic programming," *Mathematics of operations research* **39** (2014), 1222–1259.

[30] Wiesemann, W., Kuhn, D., and Sim, M., "Distributionally robust convex optimization," *Operations Research* **62** (2014), 1358–1376.