Ordered Risk Minimization: Learning More from Less Data

Peter Coppens and Panagiotis Patrinos[†]

Abstract— **We consider the worst-case expectation of a permutation invariant ambiguity set of discrete distributions as a proxy-cost for data-driven expected risk minimization. For this framework, we coin the term *ordered risk minimization* to highlight how results from order statistics inspired the proxy-cost. Specifically, we show how such costs serve as pointwise high-confidence upper bounds of the expected risk. The confidence level can be determined tightly for any sample size. Conversely we also illustrate how to calibrate the size of the ambiguity set such that the high-confidence upper bound has some user specified confidence. This calibration procedure notably supports ϕ-divergence based ambiguity sets. Numerical experiments then illustrate how the resulting scheme both generalizes better and is less sensitive to tuning parameters compared to the empirical risk minimization approach.**

I. INTRODUCTION

The problem of *expected risk minimization* is ubiquitous in machine learning and statistics [1], [2]. It is based on the idea that the quality of a model can be assessed by measuring its expected error, quantified by some loss function. The expectation should be evaluated with respect to the data-generating distribution. However, in practice, only samples are available. So the expectation needs to be replaced with a data-driven proxy, which aggregates the data. The common solution is *empirical risk minimization* or the *sample average approach (SAA)*, where one takes an average over the losses at the sampled data points.

Despite its advantages, the SAA often exhibits excessive sensitivity to the specific data realizations, particularly in high-dimensional settings [3, §8.H], leading to diminished generalization capabilities of the model. To address this, researchers have turned to *Distributionally Robust Optimization (DRO)*, aiming to robustify against disparities between the empirical and true data-generating distributions.

In DRO, a worst-case expectation with respect to distributions in an *ambiguity set* centered on the empirical distribution serves as a proxy for the true expected risk. This set can be based on the Wasserstein distance [4], ϕ-divergences [5], hypothesis tests [6], and others. See [7], [8] for recent surveys. However, DRO faces challenges when determining the ambiguity set's size. Current approaches rely on concentration inequalities [9] or asymptotic bounds [5], ensuring that true distribution is contained within the

ambiguity set with high probability. Unfortunately, this often results in conservatism, caused by either loose constants in the concentration inequalities or the shape of the ambiguity set. As an alternative, bootstrapping or cross validation techniques are often employed (cf. [4]). These can be computationally expensive and lack statistical guarantees for finite samples, similarly to the asymptotic bounds. Such guarantees are a requirement in safety-critical applications like control (e.g. constraint tightening in tube-based MPC [10], [11]).

To address conservativeness issues, [12], [13] focus on bounding the expectation directly as an alternative to creating a confidence bound for the entire distribution. This mimics the focus on the expectation in the statistical learning framework of [2, §1]. However, their bounds are asymptotic and therefore also lack strong statistical guarantees. In this paper, we take the first steps towards a finite-sample version of their scheme. To achieve this, we draw inspiration from results in *order statistics* [14] and *stochastic orders* [15] to motivate the use of permutation invariant ambiguity sets. Notably, the ϕ-divergences used in [12], [13], [16] produce a specific case. We coin the term *ordered risk minimization* to emphasize our statistical motivations. We demonstrate how to calibrate the ambiguity set's size to upper bound the true expectation with high probability, even when the true distribution does not fall within the ambiguity set. This probability serves as an intuitive tuning parameter.

The remainder of the paper continues as follows. We first present some notation, before moving on the the problem statement in §II. There the proxy costs we use are presented as well as the calibration problem. We present the statistical interpretation of these proxy costs as high confidence upper bounds in §III and solve the calibration problem in §IV. Numerical experiments are then presented in §V.

Notation: Let $\mathbb{R}$ denote the reals. For some convex function $\phi\colon \mathbb{R}^n \to \mathbb{R}$, let ϕ^* denote the convex conjugate. For integers a, b let $[a, b] = \{a, \dots, b\}$ and $[b] = \{1, \dots, b\}$. Let $[x]_+ = \max(0, x)$. For real vectors $x, y \in \mathbb{R}^n$ we use $\langle x, y \rangle$ to denote the Euclidean inner product. For a cone $\mathcal{K}$ let $\mathcal{K}^\circ := \{y\colon \langle x, y \rangle \leq 0, \forall x \in \mathcal{K}\}$ denote its polar cone. Let Π^n denote the permutations of $[n]$ (i.e., all bijections $[n] \to [n]$). We write $\pi x = (x_{\pi(1)}, \dots, x_{\pi(n)})$ for $\pi \in \Pi^n$, $x \in \mathbb{R}^n$ and similarly let $\Pi^n y = \{\pi y\colon \pi \in \Pi^n\}$ denote the orbit of y under Π^n. Let $\mathbb{R}^n_\uparrow := \{x\colon x_1 \leq x_2 \leq \dots \leq x_n\}$ denote the monotone cone and $\mathcal{M}^n$ its polar (cf. Lem. A.1 and (6)). Let $\Delta^n := \{\mu\colon \sum_{i=1}^n \mu_i = 1, \mu_i \geq 0, i \in [n]\}$ denote the probability simplex. For a vector $x \in \mathbb{R}^n$ let $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ be the increasing permutation of the elements of x with $x_\uparrow = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$. For sets A, B let $A + B := \{a + b\colon a \in A, b \in B\}$ denote

the Minkowski sum. For random variables X, Y we write $X \cong Y$ to say X is identically distributed to Y. Let $\mathrm{esssup}[X]$ denote the essential supremum. Let $X \cong \mathrm{U}[\ell, u]$ imply X is uniformly distributed over $[\ell, u]$. For a set A, $X \cong \mathrm{U}[A]$ is then uniformly distributed over A. Finally let $X \cong \mathcal{N}(\mu, \Sigma)$ denote X that is normally distributed with mean μ and covariance Σ.

II. PROBLEM STATEMENT

We consider expected risk minimization

$$\underset{\theta \in \Theta}{\mathrm{minimize}} \quad \mathbb{E}[\ell(\theta, \xi)]. \tag{1}$$

Here $\Theta \subseteq \mathbb{R}^{n_\theta}$ and $\xi \colon \Omega \to \Xi \subseteq \mathbb{R}^{n_\xi}$ is a random vector on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. As is common [1], [2], we then assume access to *independent and identically distributed (iid) samples* $\xi^{(1)}, \ldots, \xi^{(n-1)}$. Let $\ell_i(\theta) = \ell(\theta, \xi^{(i)})$ for $i = 1, \ldots, n-1$ and take $\ell_n(\theta)$ such that it upper bounds the cost almost surely. That is $\mathbb{P}[\ell(\theta, \xi) \leq \ell_n(\theta)] = 1$. It is assumed that $\ell_n(\theta)$ is finite for any θ. We do so for two reasons: *(i)* an assumption on the tail of the distribution of $\ell(\theta, \xi)$ is required to find a confidence interval for the mean [17]; *(ii)* the scheme is simplified considerably. An example of a valid bound is $\ell_n(\theta) = \sup_{\xi \in \Xi} \ell(\theta, \xi)$.

We will use a data-driven proxy for the expectation by introducing *permutation invariant ambiguity sets*. These are subsets $\mathscr{A}$ of the probability simplex Δ^n such that, for each $\mu \in \mathscr{A}$ any permutation of μ is also in $\mathscr{A}$. The *ordered risk minimization* problem is then:

$$\underset{\theta \in \Theta}{\mathrm{minimize}} \quad \sup_{\mu \in \mathscr{A}} \sum_{i=1}^{n} \mu_i \ell_i(\theta). \tag{2}$$

This proxy cost interpolates between the robust case for $\mathscr{A} = \Delta^n$ and the sample average (including a term associated with $\ell_n(\theta)$) when $\mathscr{A} = \{\mathbb{1}_n/n\}$. The interpolation interpretation is also common in DRO [5]. To find a good balance, our goal is to select $\mathscr{A}$ such that, for all $\theta \in \Theta$,

$$\mathbb{P}\left[\sup_{\mu \in \mathscr{A}} \sum_{i=1}^{n} \mu_i \ell_i(\theta) \geq \mathbb{E}[\ell(\theta, \xi)] \right] \geq 1 - \delta, \tag{3}$$

We refer to this problem as the *calibration problem*. It robustifies against disparities between the empirical and true distributions. The parameter δ then serves as an intuitive, user-determined parameter that controls the conservativeness of the method. However, as illustrated by experiments, our method is relatively insensitive to the value of δ.

To find an ambiguity set $\mathscr{A}$ satisfying (3) we need to somehow parametrize it. A well known class of permutation invariant ambiguity sets uses ϕ-divergences [5]. Let $\phi \colon \mathbb{R}_+ \to \mathbb{R}$ be lower semicontinuous, convex and $\phi(1) = 0$. Also, let[1] $I_\phi(\mu, \nu) := \sum_{i=1}^{n} \nu_i \phi(\mu_i/\nu_i)$ for all $\mu, \nu \in \Delta^n$. A *(centered) ϕ-divergence ambiguity set* is then

$$\mathscr{A}_\alpha := \left\{ \mu \in \Delta^n \colon I_\phi\left(\mu, \frac{\mathbb{1}_n}{n}\right) = \sum_{i=1}^{n} \frac{\phi(n\mu_i)}{n} \leq \alpha \right\}. \tag{4}$$

---

[1] We take the lower semicontinuous envelope of the terms inside the sum [18, Def. 6] to handle cases where ν_i equals zero.

In this work we consider two examples: *total variation (TV)* for which $\phi(t) = |t - 1|$ and *Kullback Leibler (KL) divergence* for which $\phi(t) = t \log t - t + 1$. However, our method works for any divergence. See [5] for more examples.

To calibrate $\mathscr{A}_\alpha$ the radius $\alpha \in \mathbb{R}$ should then be the smallest value such that (3) still holds. We also provide an alternative parametrization, related to a well known bound by Anderson [19] and the conditional value-at-risk.

It is important to note that the constraint in (3) is less stringent compared to DRO, which guarantees that the supremum in (3) acts as a high-confidence upper bound, uniformly over θ[2]. After all, we never require that the true distribution is contained within $\mathscr{A}$ (as is the case in [20]). The gap between the point-wise (3) and the uniform equivalent is examined for ϕ-divergences in [12], [13] in the asymptotic regime.

We numerically approximate the calibration problem without samples from ξ. So the parameters of the set $\mathscr{A}$ only need to be computed once and can be tabulated afterwards. This contrasts the complex derivations and the resulting conservative constants associated with analytical approaches used to compute the radius of an ambiguity set in DRO [4], [5], [9]. We show experimentally how our calibration of $\mathscr{A}$ according to (3) greatly improves generalization.

III. STATISTICAL FRAMEWORK

The analysis of this section investigates upper bounds for the mean of a scalar *random variable (rv)* $Z \colon \Omega \to \mathbb{R}$, defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. These findings remain applicable to the previous section, when considering $Z = \ell(\theta, \xi)$ for fixed θ. Let $F(z) = \mathbb{P}[Z \leq z]$ denote the *cumulative distribution function (cdf)* of Z. We assume access to iid samples $Z_1, \ldots, Z_{n-1}$ and an upper bound denoted as $Z_n = \mathrm{esssup}[Z]$ for notational convenience, which satisfies $F(Z_n) = 1$.

We then introduce the *coverages*

$$W_i = F(Z_{(i)}) - F(Z_{(i-1)}), \quad \forall i \in [n], \tag{5}$$

where $-\infty = Z_{(0)} \leq Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)}$ denotes an increasing permutation of $Z_1, \ldots, Z_n$ called the *order statistics* of Z and with $Z_{(0)}$ added for convenience.

The coverages will be used to bound the expectation. To do so, we require the following cone in $\mathbb{R}^n$:

$$\mathcal{M}^n = \left\{ x \colon \sum_{i=1}^{k} x_i \geq 0, \forall k \in [n-1], \sum_{i=1}^{n} x_i = 0 \right\}. \tag{6}$$

In Lem. A.1 we prove that it corresponds to the polar of the monotone cone $\mathbb{R}_\uparrow^n$. It has a history in isotonic regression [21], majorization [22] and also describes a stochastic order between discrete distributions [15, p. 4] as we illustrate later.

The expectation bound is then as follows:

---

[2] The mean bound will be uniform when

$$\mathbb{P}\left[\sup_{\mu \in \mathscr{A}} \sum_{i=1}^{n} \mu_i \ell_i(\theta) \geq \mathbb{E}[\ell(\theta, \xi)], \forall \theta \in \Theta \right] \geq 1 - \delta.$$

Proposition III.1. *Suppose that $\mathcal{A} \subseteq \Delta^n$ is permutation invariant[3]. Take W as in (5). Then, for any rv $Z \colon \Omega \to \mathbb{R}$ with iid samples $\{Z_i\}_{i=1}^{n-1}$ and $Z_n = \mathrm{esssup}[Z] < +\infty$,*

$$\mathbb{P}\left[\sup_{\mu \in \mathcal{A}} \sum_{i=1}^{n} \mu_i Z_i \geq \mathbb{E}[Z]\right] \geq \mathbb{P}[W \in \mathcal{A} + \mathcal{M}^n].$$

Proof. We split up the expectation[4] as follows:

$$\mathbb{E}[Z] = \int_{-\infty}^{Z_{(n)}} z\, \mathrm{d}F(z) = \sum_{i=1}^{n} \int_{Z_{(i-1)}}^{Z_{(i)}} z\, \mathrm{d}F(z), \quad (7)$$

with $Z_{(1)} \leq \cdots \leq Z_{(n-1)}$ the order statistics and $Z_{(n)} = \mathrm{esssup}[Z]$. The first equality follows from $\mathbb{P}[Z \leq Z_{(n)}] = 1$ and the second from [23, Thm. 16.9]. For each term $\int_{Z_{(i-1)}}^{Z_{(i)}} z\, \mathrm{d}F(z) \leq Z_{(i)} \int_{Z_{(i-1)}}^{Z_{(i)}} \mathrm{d}F(z) = Z_{(i)}(F(Z_{(i)}) - F(Z_{(i-1)}))$. Hence $\mathbb{E}[Z] \leq \sum_{i=1}^{n} Z_{(i)} W_i$.

Condition on $W \in \mathcal{A} + \mathcal{M}^n$. Then,

$$\mathbb{E}[Z] \leq \sum_{i=1}^{n} W_i Z_{(i)} \leq \sup_{\mu \in \mathcal{A} + \mathcal{M}^n} \sum_{i=1}^{n} \mu_i Z_{(i)}.$$

The expression on the right can be simplified by noting that

$$\textstyle\sup_{\mu \in \mathcal{A} + \mathcal{M}^n} \sum_{i=1}^{n} \mu_i Z_{(i)} = \sup_{\mu \in \mathcal{A}} \sum_{i=1}^{n} \mu_i Z_{(i)},$$

where we use $(Z_{(1)}, \ldots, Z_{(n)}) \in \mathbb{R}_\uparrow^n$, the definition of the polar cone and Lem. A.1, which implies $\sum_{i=1}^{n} s_i Z_{(i)} \leq 0$ for any $s \in \mathcal{M}^n$ with equality for $s = 0$.

Finally, let $\pi \in \Pi^n$ be the permutation such that $Z_{\pi(i)} = Z_{(i)}$ for $i \in [n]$ and let π^{-1} denote its inverse (which exists, since permutations are bijections). Then

$$\sup_{\mu \in \mathcal{A}} \sum_{i=1}^{n} \mu_i Z_i = \sup_{\mu \in \mathcal{A}} \sum_{i=1}^{n} \mu_{\pi^{-1}(i)} Z_{(i)}$$

$$= \sup_{\pi\mu \in \mathcal{A}} \sum_{i=1}^{n} \mu_i Z_{(i)} \overset{(i)}{=} \sup_{\mu \in \mathcal{A}} \sum_{i=1}^{n} \mu_i Z_{(i)},$$

where *(i)* uses permutation invariance. Hence, we showed $\mathbb{P}[\sup_{\mu \in \mathcal{A}} \sum_{i=1}^{n} \mu_i Z_{(i)} \geq \mathbb{E}[Z]] \geq \mathbb{P}[W \in \mathcal{A} + \mathcal{M}^n]$. □

From Prop. III.1, it is clear that the distribution of W is important. Interestingly, when F is continuous, then W is always uniformly distributed over Δ^n [24, Thm. 8.7.4]. For general distributions however, we can still establish a type of *stochastic order* between the two distributions using $\mathcal{M}^n$:

Lemma III.2. *Take $W = (W_1, \ldots, W_n) \in \Delta^n$ as in (5). Then, for any (Lebesgue measurable) $\mathcal{A} \subseteq \Delta^n$,*

$$\mathbb{P}[W \in \mathcal{A} + \mathcal{M}^n] \geq \mathbb{P}[\nu \in \mathcal{A} + \mathcal{M}^n],$$

with $\nu \cong \mathrm{U}[\Delta^n]$ uniformly distributed over Δ^n.

For continuous cdf we have $\mathbb{P}[W \in \mathcal{A}] = \mathbb{P}[\nu \in \mathcal{A}]$.

Proof. For continuous cdf we refer to [24, Thm. 8.7.4]. For discontinuous cdf we first introduce a construction of the joint distribution of random vectors W' and ν', such that $W' \cong W$ and $\nu' \cong \nu$ (i.e., the marginals are as specified in the lemma). For this construction, we show that $\nu' \in$
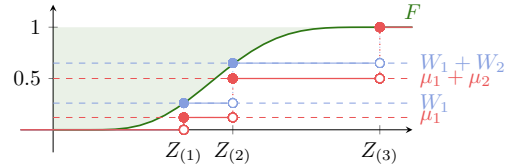
Fig. 1: Lower bounds of the cumulative distribution for non-negative Z with cdf F. The tightest lower bound supported on the samples is depicted in blue, while a feasible lower bound is depicted in red. The green area is the expectation.

$\mathcal{A} + \mathcal{M}^n$ implies $W' \in \mathcal{A} + \mathcal{M}^n$ almost surely. So $\mathbb{P}[W' \in \mathcal{A} + \mathcal{M}^n] \geq \mathbb{P}[\nu' \in \mathcal{A} + \mathcal{M}^n]$. From $W' \cong W$ and $\nu' \cong \nu$ we then get the required result.

The construction starts by taking $U_i \cong \mathrm{U}[0, 1]$ as uniform random variables, for $i \in [n - 1]$, with $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n-1)} \leq U_{(n)} = 1$ the uniform order statistics. Let $\sum_{i=1}^{k} \nu_i' = U_{(k)}$ for $k \in [n]$. Then, by [14, §6.4], $\nu' \cong \mathrm{U}[\Delta^n]$ or $\nu' \cong \nu$. Meanwhile, the *quantile transform* [25, Lem. 1.2.4(i)] states that $Z_i' = F^{-1}(U_i)$ has cdf F, where F^{-1} is the quantile function of Z. Since F (and therefore F^{-1}) is nondecreasing, we can apply [25, Lem. 1.2.1] to claim that $F^{-1}(U_{(i)})$ is distributed as the i-th order statistic $Z_{(i)}$. So with $Z_{(i)}' = F^{-1}(U_{(i)})$ we take W' analogously to (5). From $Z_{(i)}' \cong Z_{(i)}$ we then have $W' \cong W$.

Both marginals are related as, for $k \in [n]$,

$$\sum_{i=1}^{k} W_i' = F(Z_{(k)}')$$
$$= F(F^{-1}(U_{(k)})) = F(F^{-1}(\textstyle\sum_{i=1}^{k} \nu_i')), \quad (8)$$

where the first equality follows by summing (5) for $i \in [k]$ with W_i', $Z_{(i)}'$ in place of W_i, Z_i, the second by construction of $Z_{(k)}'$ and the third by construction of ν'. Note that, for general distributions, we have $F(F^{-1}(p)) \geq p$ for all $p \in [0, 1]$ [26, Ex. 3.2], with strict inequality iff $p \in (0, 1)$ is not in the range of F. Applying this to (8) gives

$$\sum_{i=1}^{k} \nu_i' \leq \sum_{i=1}^{k} W_i', \forall k \in [n-1] \text{ and } \sum_{i=1}^{n} \nu_i' = \sum_{i=1}^{n} W_i'. \quad (9)$$

Observe how (9) corresponds to a conic inequality under $\mathcal{M}^n$. The inequalities are strict iff $\sum_{i=1}^{k} \nu_i'$ is not in the range of F (i.e., it lies in a discontinuous jump of F). In that sense, (9) models the gap between $\mathrm{U}[\Delta^n]$ and the coverages W.

To complete the proof, assume that $\nu' \in \mathcal{A} + \mathcal{M}^n$. By definition of the Minkowski sum, this is equivalent to there being some $\mu \in \mathcal{A}$ such that $\nu - \mu \in \mathcal{M}^n$ or, $\sum_{i=1}^{k} \mu_i \leq \sum_{i=1}^{k} \nu_i'$ for $k \in [n-1]$ and $\sum_{i=1}^{n} \mu_i = \sum_{i=1}^{n} \nu_i'$ (cf. (6)). Thus, from (9), we have $\sum_{i=1}^{k} \mu_i \leq \sum_{i=1}^{k} W_i'$ for $k \in [n-1]$ and $\sum_{i=1}^{n} \mu_i = \sum_{i=1}^{n} W_i'$. By definition of $\mathcal{A} + \mathcal{M}^n$ and (6) this shows that $W' \in \mathcal{A} + \mathcal{M}^n$. So, by our arguments at the start of the proof, we showed the required result. □

The result in Lem. III.2 can be interpreted in terms of lower bounding the cdf of Z. To illustrate this, we introduce the weighted empirical cdf

$$F_\mu^n(x) = \textstyle\sum_{i=1}^{n} \mu_i \mathbb{1}_{[Z_{(i)}, +\infty)}, \quad (10)$$

where $\mu \in \Delta^n$ and $\mathbb{1}_{[Z_{(i)}, +\infty)}(z) = 1$ when $z \geq Z_{(i)}$ and zero otherwise. Note that $W \in \{\mu\} + \mathcal{M}^n$ holds iff

$$F_\mu^n(Z_{(i)}) = \textstyle\sum_{i=1}^k \mu_i \leq \sum_{i=1}^k W_i = F(Z_{(i)}), \forall k \in [n],$$

with $\sum_{i=1}^n \mu_i = \sum_{i=1}^n W_i = 1$. This relationship relates to (9) and (6) and implies that F_μ^n should lower bound the cdf F everywhere, as depicted in Fig. 1. This inequality between cdfs is the usual stochastic order [15, §1.A.1]. The event $W \in \mathcal{A} + \mathcal{M}^n$ is then equivalent to the existence of a cdf F_μ^n with weights $\mu \in \mathcal{A}$ that lower bounds the true cdf. In terms of this interpretation, Prop. III.1 follows from the fact that a lower bound on the cdf implies an upper bound on the expectation (cf. [19], [27, Eq. 1.A.5]).

By combining Prop. III.1 and Lem. III.2 we directly prove the main contribution of this paper:

Theorem III.3. *Assuming the setting of Prop. III.1 and taking $\nu \cong U[\Delta^n]$, then*

$$\mathbb{P}\left[\sup_{\mu \in \mathcal{A}} \sum_{i=1}^n \mu_i Z_i \geq \mathbb{E}[Z]\right] \geq \mathbb{P}[\nu \in \mathcal{A} + \mathcal{M}^n].$$

Note that, the important requirement of $\mathcal{A}$ in Thm. III.3 is that it is permutation invariant and a subset of the probability simplex. The resulting support functions are related to *law-invariant, coherent risk measures* in literature [28, §6.3.5], [6], where the *law* in our case is a permutation of the random vector. The ϕ-based ambiguity sets considered below are the most frequently studied case of such risk measures.

IV. CALIBRATION PROBLEM

This section considers calibrating ambiguity sets $\mathcal{A}$ such that (3) holds. To do so, we first consider the ϕ-divergence parametrization in (4) and try to upper bound the smallest α such that (3) still holds for $\mathcal{A} = \mathcal{A}_\alpha$, which we denote as $\alpha_\star$. Later we also provide an alternative parametrization similar to the conditional value-at-risk (cf. Cor. IV.3 and the discussion below).

We can use the previous result to simplify (3). Note that, for a fixed θ, $\ell(\theta, \xi)$ is simply a scalar random variable, which we will denote as Z. So we consider

$$\inf\left\{\alpha \colon \mathbb{P}\left[\sup_{\mu \in \mathcal{A}_\alpha} \sum_{i=1}^n \mu_i Z_i \geq \mathbb{E}[Z]\right] \geq 1 - \delta, \forall Z\right\}, \quad (11)$$

where we inherit the notation from the previous section. It's solution will upper bound $\alpha_\star$. Note that all previous results were distribution-free. They hold for all (bounded) random variables Z, invariant of their underlying distribution. As such, by using Thm. III.3, the constraint in (11) can be conservatively approximated by

$$\mathbb{P}[\nu \in \mathcal{A}_\alpha + \mathcal{M}^n] \geq 1 - \delta, \quad \text{with } \nu \cong U[\Delta^n]. \quad (12)$$

We use this to approximate the calibration problem:

Proposition IV.1. *Let I_ϕ denote the ϕ-divergence, with $\mathcal{A}_\alpha$ the associated ambiguity set as in (4). Let $\alpha_\star$ denote the smallest α such that (3) holds for $\mathcal{A}_\alpha$. Then*

$$\alpha_\star \leq \inf\left\{\alpha \in \mathbb{R} \colon \mathbb{P}\left[I_\phi^\diamond(\nu) \leq \alpha\right] \geq 1 - \delta\right\}, \quad (13)$$

with $\nu \cong U[\Delta^n]$ and, for $\phi^*(s) = \sup_{t \geq 0}\{ts - \phi(t)\}$,

$$I_\phi^\diamond(\nu) := \sup_{\lambda \in \mathbb{R}_\uparrow^n}\left\{\textstyle\sum_{i=1}^n \nu_i \lambda_i - \frac{1}{n}\phi^*(\lambda_i)\right\}. \quad (14)$$

Proof. Note that $\nu \in \mathcal{A}_\alpha + \mathcal{M}^n$ holds iff there is some $\mu \in \mathcal{A}_\alpha$ such that $\nu - \mu \in \mathcal{M}^n$. Equivalently, there should exist a $\mu \in \Delta^n$, which satisfies $I_\phi(\mu, \mathbb{1}_n/n) \leq \alpha$ and $\nu - \mu \in \mathcal{M}^n$. We already showed that the smallest α satisfying (12) upper bounds $\alpha_\star$. The left-hand side of (12) equals

$$\mathbb{P}\left[\inf_{\mu \in \Delta^n}\left\{\textstyle\sum_{i=1}^n \frac{1}{n}\phi(n\mu_i)\colon \nu - \mu \in \mathcal{M}^n\right\} \leq \alpha\right].$$

Since $\nu \cong U[\Delta^n]$, the constraint $\nu - \mu \in \mathcal{M}^n$, together with $\sum_{i=1}^n \nu_i = 1$ implies (cf. (6)) $\sum_{i=1}^n \mu_i = 1$. So the infimum can be taken over $\mathbb{R}_+^n$. We place the constraint inside of the cost by noting that the indicator function of a polar cone equals the support function of its dual [20, Ex. 2.26]. So we can rewrite (12) as follows:

$$\inf_{\mu \in \mathbb{R}_+^n} \sup_{\lambda \in \mathbb{R}_\uparrow^n}\left\{\textstyle\sum_{i=1}^n \frac{1}{n}\phi(n\mu_i) + \langle \lambda, \nu - \mu \rangle\right\},$$

$$\leq \sup_{\lambda \in \mathbb{R}_\uparrow^n} \textstyle\sum_{i=1}^n \inf_{t_i \geq 0}\left\{\phi(t_i) - t_i\lambda_i\right\}/n + \lambda_i\nu_i,$$

The inequality follows by weak duality, the substitution $t_i = n\mu_i$ and separability. The infima inside the sum are $-\phi^*(\lambda_i)$, completing the proof by the argument preceding (12). $\square$

Observe that the right-hand side of (13) is essentially the $1 - \delta$ quantile of the scalar random variable $I_\phi^\diamond(\nu)$. Unfortunately its distribution is unknown. However, we can sample from it. To do so note that $\nu \cong U[\Delta^n]$ is Dirichlet distributed with parameters $(1, \ldots, 1) \in \mathbb{R}^n$ [14, §6.4]. So we can sample it by either sampling from a Dirichlet distribution, or by sampling $n - 1$ uniform order statistics $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n-1)}$ and using $(U_{(1)}, U_{(2)} - U_{(1)}, \ldots, 1 - U_{(n-1)})$ as a sample from $U[\Delta^n]$ (cf. [14, §6.4]). We can then evaluate $I_\phi^\diamond(\nu)$ by solving (14), which is a special case of the optimization problem in [29]. That paper presents the *pool-adjacent violator (PAV)* algorithm. It solves (14) with complexity $\mathcal{O}(n)$.

The fact that we can sample $I_\phi^\diamond(\nu)$ efficiently implies that (11) can be estimated through data-driven means.

Theorem IV.2. *Let $\alpha_1, \ldots, \alpha_m$ denote m iid samples from $I_\phi^\diamond(\nu)$ with $\nu \cong U[\Delta^n]$ and $\alpha_{(1)} \leq \cdots \leq \alpha_{(m)}$ the associated order statistics. Then, for any $\delta \in [0, 1]$ and $k \in [m]$,*

$$\mathbb{P}\left[\alpha_\star \leq \alpha_{(k)}\right] \geq 1 - \beta \quad (15)$$

with $\alpha_\star$ as in Prop. IV.1, and $\beta = I_{1-\delta}(k, m - k + 1)$ the regularized incomplete beta function (i.e., the cdf of a beta distribution) at level $1 - \delta$.

Proof. The result follows directly from Prop. IV.1 and a one-sided data-driven bound of a quantile stated in [30, §G.2.2] applied to (13). Alternatively (13) can be interpreted as a *scenario program*, for which [31, Thm. 3.7] holds. $\square$

In practice, the user would select m (the number of samples from $I_\phi^\diamond(\nu)$ computed using PAV). A larger value

gives a tighter upper bound for $\alpha_\star$ at the cost of additional computation time. The confidence level is then determined by fixing some $\beta \in [0,1]$ and then finding the smallest k such that $I_{1-\delta}(k, m-k+1) \leq \beta$. Such a k is determined with a scalar root finder[5].

We can establish a connection with other results in literature, through the following corollary of Thm. III.3:

Corollary IV.3. *Let* $\mu \in \mathbb{R}_\uparrow^n \cap \Delta^n$. *Then, for uniform order statistics* $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n-1)} \leq U_{(n)} = 1$ *and* $Z_{(1)} \leq Z_{(2)} \leq \ldots Z_{(n-1)}$ *and* $Z_{(n)} = \mathrm{esssup}[Z]$ *the order statistics of rv* $Z \colon \Omega \to \mathbb{R}$,

$$\mathbb{P}\left[\sum_{i=1}^n \mu_i Z_{(i)} \geq \mathbb{E}[Z]\right] \geq \mathbb{P}\left[\sum_{i=1}^k \mu_i \leq U_{(k)}, \forall k \in [n]\right]. \tag{16}$$

Proof. Using [6] gives $\sum_{i=1}^n \mu_i Z_{(i)} = \sup_{\mu \in \mathscr{A}} \sum_{i=1}^n \mu_i Z_i$, with $\mathscr{A} = \mathrm{co}\,\Pi^n \mu$ the convex hull of all permutations of μ. Invoking Thm. III.3 then gives $\mathbb{P}\left[\sum_{i=1}^n \mu_i Z_{(i)} \geq \mathbb{E}[Z]\right] \geq \mathbb{P}[\nu \in \mathscr{A} + \mathcal{M}^n]$, with ν distributed according to $\mathrm{U}[\Delta^n]$. Assume that $\sum_{i=1}^k \mu_i \leq U_{(k)}$ for all $k \in [n-1]$. Noting that $\sum_{i=1}^k \nu_i \cong U_{(k)}$ by [14, §6.4] and using (6), this implies that $\nu - \mu \in \mathcal{M}^n$ and, since $\mu \in \mathscr{A}$, $\nu \in \mathscr{A} + \mathcal{M}^n$. So $\mathbb{P}[\nu \in \mathscr{A} + \mathcal{M}^n] \geq \mathbb{P}[\sum_{i=1}^k \mu_i \leq U_{(k)}, \forall k \in [n]]$. $\square$

Evaluating the right-hand side of (16) was studied in the context of lower bounding the cdf of a random variable (cf. [32] for an efficient and numerically stable algorithm).

The expression $\sum_{i=1}^n \mu_i Z_{(i)}$ is called a *distortion risk* [6], which is a convex function of $(Z_1, \ldots, Z_n)$ iff $\mu \in \mathbb{R}_\uparrow^n$. Since the uniform order statistics form a uniform distribution over a convex set (i.e., $\mathbb{R}_\uparrow^n \cap [0,1]^n$), their density is quasi-concave [28, Ex. 4.10]. Hence we can leverage [28, Ex. 4.17, Cor. 4.42] to claim that the calibration problem

$$\inf_{\mu \in \mathbb{R}_\uparrow^n \cap \Delta^n} \left\{\varphi(\mu) \colon \mathbb{P}\left[\sum_{i=1}^k \mu_i \leq U_{(k)}, \forall k \in [n]\right] \geq 1 - \delta\right\}$$

is a convex problem, whenever φ is convex. We will study the choice of φ, which should measure the size of $\mathscr{A}$, in future work as well as the tractability of the calibration problem. For now, we instead consider a specific parametrization:

$$\mu^{(\gamma)} := \left(0, \ldots, 0, \frac{\lceil(n-1)\gamma\rceil}{n-1} - \gamma, \frac{1}{n-1}, \ldots, \frac{1}{n-1}, \gamma\right), \tag{17}$$

with $\gamma \geq 1/(n-1)$, such that $\mu \in \Delta^n \cap \mathbb{R}_\uparrow^n$ holds. Then

$$\sum_{i=1}^n \mu_i^{(\gamma)} Z_{(i)} = \left(\frac{d}{n-1} - \gamma\right) Z_{(d)} + \sum_{i=d+1}^{n-1} \frac{Z_{(i)}}{n-1} + \gamma Z_{(n)}, \tag{18}$$

with $d = \lceil(n-1)\gamma\rceil$. The final expression is a well-known bound for the expectation of Z due to Anderson [19]. The re-interpretation in terms of a distortion risk is novel to the authors' knowledge. In the technical report [33, App. A-B] we also show that the final expression is an affine transformation of the well-known conditional value-at-risk. Hence we denote it as $\overline{\mathrm{CV@R}}$ in the experiments.

[5]We use `brentq` as implemented in `scipy 1.10.0`

The value of γ in the context of [19] is usually determined using an asymptotic bound (cf. [24, Thm. 11.6.2]). We find an accurate value by solving:

$$\inf_{\gamma \geq 1/(n-1)} \left\{\gamma \colon \mathbb{P}\left[\sum_{i=1}^k \mu_i^{(\gamma)} \leq U_{(k)}, \forall k \in [n]\right] \geq 1 - \delta\right\}, \tag{19}$$

for a user-specified $\delta \in [0,1]$ with a scalar root finder-foot:brentq. The probability is evaluated numerically using [32]. The calibration problem (19) in [19] is interpreted as moving the empirical cdf down as little as possible, while still guaranteeing that it lower bounds the true cdf with high probability. It is also comparable to producing the tightest mean bound as in the calibration problem (11).

V. CASE STUDIES

To illustrate the validity and potential of our method we provide several simple case studies. These are convex for maximum interpretability, as in the non-convex case a worse generalization performance might be caused by local optima. Nonetheless our method is also applicable in non-convex settings, where stochastic gradient descent methods can be used (cf. [34] for simple distortions and [18] for divergences). In the convex case we use duality to reformulate the proxy cost in (2). See [6], [35] for details.

We present problems of the form (1) and employ ordered risk minimization (2) or using (18), which we refer to as $\overline{\mathrm{CV@R}}$. For divergences we use either the *total variation (TV)* or *Kullback-Leibler (KL)* and the radius is calibrated using Thm. IV.2 (with $\beta = 0.005$ and $m = 10\,000$). The value of γ in (18) is calibrated using (19).

A. Newsvendor

We begin with a toy problem, illustrating the behavior of our method in low-sample settings. Let $\xi \colon \Omega \to \mathbb{R}$ be Beta distributed with $\alpha = 0.1$, $\beta = 0.2$, scaled by a factor $\overline{D} := 100$. Consider a newsvendor problem [28, §1.2.1]:

$$\underset{\theta \in \mathbb{R}}{\text{minimize}} \quad \mathbb{E}[\underbrace{c\theta + b[\xi - \theta]_+ + h[\theta - \xi]_+}_{\ell(\theta, \xi)}],$$

with $b = 14$, $h = 2$ amd $c = 1$. For samples $\{\xi_i\}_{i=1}^{n-1}$ with $n = 20$ let $\ell_i(\theta) = \ell(\theta, \xi_i)$ for $i \in [n-1]$, $\ell_n(\theta) = \max\{(c-b)\theta + b\overline{D}, (c+h)\theta\}$ a robust upper bound. We replace the expectation by a data-driven proxy as described at the start of the section. For the *sample average approach (SAA)* we take $\sum_{i=1}^{n-1} \ell_i(\theta)/(n-1)$.

The calibration problems are solved for $\delta = 0.2$. Their performance is compared over 200 sampled data sets in Fig. 2. The left plot shows the actual expected cost for the minimizers. The blue dashed line is the true optimum of §V-A. See [28, §1.2.1] for details on how to compute these values. Note how the SAA performs decently in the median, but has significantly more variance. The outliers above 240 were omitted, the largest of which was 428.2. Moreover, the right plot depicts the difference between the optimum value of the proxy cost, and the true cost. The SAA often underestimates its true cost, while our methods overestimate it. The dashed red line depicts the behavior when taking
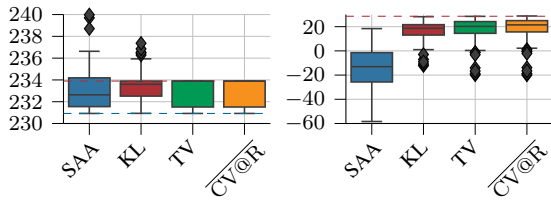
Fig. 2: Box plots showing newsvendor expected cost (left); and difference between the predicted cost and expected cost (right). The colored area is the *inter-quartile range (IQR)*, while the whiskers show the range of samples truncated to 1.5 times the IQR. Outliers outside of this range are depicted as diamonds. The red dashed lines depict the robust performance. The blue dashed line is the optimal cost.

$\mathscr{A} = \Delta^n$ in (2) (cf. [28, Eq. 1.9]). As we almost never perform worse than this robust method, this shows that our methods learn from data without over-fitting on the sample.

In large sample cases, we can use the largest sample as an approximation of $\ell_n(\theta)$. This heuristic is similar to the one used in the scenario approach, the consequences of which have been studied in detail (cf. [36]). In combination with some regularization, this significantly boosts the performance of our method, as shown in the next example.

B. Support Vector Machines

Consider a classification problem with $X \cong \mathcal{N}(0, I_2)$ normally distributed and $Y = 1$ if $X_1 X_2 \geq 0$ and $Y = -1$ otherwise. A *Support Vector Machine (SVM)* solves:

$$\underset{(f,b) \in \mathcal{H} \times \mathbb{R}}{\text{minimize}} \quad \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \lambda \mathbb{E}\left[1 - Y(f(X) - b)\right]_+$$

with $\lambda > 0$ and $\mathcal{H}$ some *reproducing kernel Hilbert Space (RKHS)* [37, Def. 2.9]. The resulting classifier is then given by $\text{sign}(f(X) - b)$. Henceforth $\mathcal{H}$ is the RKHS associated with the *radial basis function* kernel [37, §2.3] with some standard deviation σ. Solving the primal problem is difficult for two reasons: *(i)* the true expectation is often unknown; *(ii)* optimizing over the infinite dimensional $\mathcal{H}$ is intractable in general. We resolve *(i)* by replacing the expectation with a proxy-cost as described above and *(ii)* through the usual duality trick [37, §7.4]. Details are deferred to [33, App. B].

The proxy cost of three of the risks above – SAA, TV and $\overline{\text{CV@R}}$ – is a maximum of linear functions and the dual problem is a QP. The sample average – C-SVC in [37, §7.5] – is the usual choice. We illustrate the superior performance our calibrated risks.

In Fig. 3, the three classifiers produced by the three proxy costs above are depicted. Note how both TV and $\overline{\text{CV@R}}$ perform similarly and both visibly better than the usual SAA. Quantitative performance is compared through the fraction of incorrectly labeled samples in a test set of 10^5 samples, which we refer to as the misclassification rate. The performance is compared for several tunings in Tbl. I, where any parameters not mentioned are kept as specified above. It is of note that our methods are significantly less sensitive to tuning parameters compared to the SAA. In fact, even for
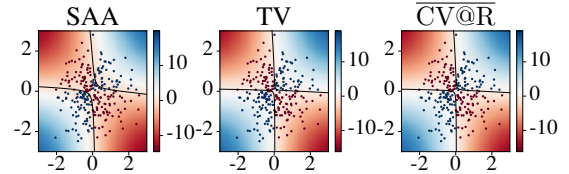

Fig. 3: SVM classifiers trained using $n = 250$ samples with $\sigma = 0.25$ and $\lambda = 10^4$ for different risks. The red and blue markers are samples for $Y = 1$ and -1 respectively. The line is the decision boundary and the color axis depicts $f(X) - b$.
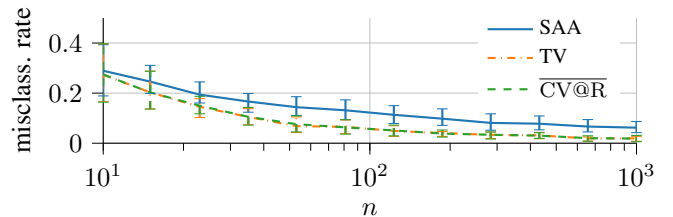

Fig. 4: SVM misclassification rates for varying sample counts n. The center line depicts the mean, while the intervals depicts the empirical 0.2-confidence interval.

the tunings where SAA performs best, our methods perform better for the same tuning, for reasonable choices of δ.

We can also examine the effect of varying the sample count n. For each such value we train the classifiers, again using the parameters used to produce Fig. 3, for 30 training sets. The resulting misclassification rates are depicted in Fig. 4. Again note that $\overline{\text{CV@R}}$ and TV both outperform SAA.

REFERENCES

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning*. Cambridge University Press, 2014.
[2] V. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.
[3] J. Royset and R. Wets, *An optimization primer*. Springer, 2022.
[4] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1-2, pp. 115–166, 2018.
[5] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science*, vol. 59, no. 2, pp. 341–357, 2013.
[6] D. Bertsimas and D. B. Brown, "Constructing Uncertainty Sets for Robust Linear Optimization," *Operations Research*, vol. 57, no. 6, pp. 1483–1495, 2009.
[7] H. Rahimian and S. Mehrotra, "Distributionally Robust Optimization: A Review," 2019, arXiv: 1908.05659.
[8] F. Lin, X. Fang, and Z. Gao, "Distributionally Robust Optimization: A review on theory and applications," *Numerical Algebra, Control & Optimization*, vol. 12, no. 1, p. 159, 2022.
[9] E. Delage and Y. Ye, "Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems," *Operations Research*, vol. 58, no. 3, pp. 595–612, 2010.
[10] M. Lorenzen, F. Dabbene, R. Tempo, and F. Allgower, "Constraint-Tightening and Stability in Stochastic Model Predictive Control," *IEEE TAC*, vol. 62, no. 7, pp. 3165–3177, 2017.
[11] L. Aolaritei, M. Fochesato, J. Lygeros, and F. Dörfler, "Wasserstein tube MPC with exact uncertainty propagation," 2023, arXiv: 2304.12093.
[12] J. C. Duchi, P. W. Glynn, and H. Namkoong, "Statistics of robust optimization: A generalized empirical likelihood approach," *Mathematics of Operations Research*, vol. 46, no. 3, pp. 946–969, 2021.
[13] H. Lam, "Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization," *Operations Research*, vol. 67, no. 4, pp. 1090–1105, 2019.

δ	σ	λ	SAA	TV	$\overline{\text{CV@R}}$
0.05	0.05	10^4	0.474 (0.052)	0.126 (0.100)	0.111 (0.054)
		10^6	0.170 (0.057)	0.052 (0.035)	0.052 (0.041)
		10^8	0.053 (0.020)	0.020 (0.008)	0.022 (0.011)
		10^{10}	0.039 (0.029)	0.027 (0.013)	0.022 (0.017)
	0.25	10^4	0.097 (0.047)	0.045 (0.038)	0.038 (0.035)
		10^6	0.042 (0.012)	0.019 (0.004)	0.016 (0.003)
		10^8	**0.022 (0.010)**	0.019 (0.004)	0.018 (0.003)
		10^{10}	0.067 (0.082)	0.061 (0.114)	0.043 (0.045)
	0.5	10^4	0.055 (0.021)	0.023 (0.007)	0.023 (0.006)
		10^6	0.030 (0.008)	0.020 (0.003)	0.020 (0.004)
		10^8	0.023 (0.008)	0.024 (0.007)	0.024 (0.007)
		10^{10}	0.202 (0.140)	0.145 (0.110)	0.114 (0.110)
0.1	0.05	10^4	0.474 (0.052)	0.154 (0.093)	0.145 (0.075)
		10^6	0.170 (0.057)	0.059 (0.048)	0.060 (0.046)
		10^8	0.053 (0.020)	0.022 (0.011)	0.021 (0.012)
		10^{10}	0.039 (0.029)	0.030 (0.021)	0.030 (0.020)
	0.25	10^4	0.097 (0.047)	0.046 (0.042)	0.041 (0.036)
		10^6	0.042 (0.012)	**0.015 (0.003)**	**0.015 (0.003)**
		10^8	**0.022 (0.010)**	0.020 (0.005)	0.018 (0.004)
		10^{10}	0.067 (0.082)	0.076 (0.119)	0.037 (0.039)
	0.5	10^4	0.055 (0.021)	0.025 (0.009)	0.024 (0.009)
		10^6	0.030 (0.008)	0.021 (0.004)	0.021 (0.005)
		10^8	0.023 (0.008)	0.024 (0.007)	0.024 (0.007)
		10^{10}	0.202 (0.140)	0.145 (0.097)	0.112 (0.110)
0.2	0.05	10^4	0.474 (0.052)	0.225 (0.142)	0.197 (0.100)
		10^6	0.170 (0.057)	0.074 (0.050)	0.075 (0.051)
		10^8	0.053 (0.020)	0.025 (0.015)	0.026 (0.015)
		10^{10}	0.039 (0.029)	0.033 (0.023)	0.036 (0.026)
	0.25	10^4	0.097 (0.047)	0.049 (0.044)	0.045 (0.040)
		10^6	0.042 (0.012)	0.019 (0.005)	0.016 (0.003)
		10^8	**0.022 (0.010)**	0.017 (0.004)	0.017 (0.004)
		10^{10}	0.067 (0.082)	0.070 (0.113)	0.039 (0.033)
	0.5	10^4	0.055 (0.021)	0.027 (0.014)	0.027 (0.015)
		10^6	0.030 (0.008)	0.021 (0.005)	0.020 (0.005)
		10^8	0.023 (0.008)	0.024 (0.007)	0.024 (0.007)
		10^{10}	0.202 (0.140)	0.186 (0.144)	0.112 (0.101)

TABLE I: SVM misclassification rates for various tuning parameters. Reported values are the *mean (standard deviation)* over 10 training sets. The same 10 sets were used for every parameter selection and method. Note that δ does not affect SAA. The lowest values in a column are bold. Observe that in the rows where SAA achieves its best performance, our methods still perform better.

[14] H. A. David and H. N. Nagaraja, *Order Statistics*. John Wiley & Sons, Inc., 2003.

[15] M. Shaked and J. G. Shanthikumar, *Stochastic Orders*. Springer Series in Statistics, Springer New York, 2007.

[16] B. P. Van Parys, P. M. Esfahani, and D. Kuhn, "From data to decisions: Distributionally robust optimization is optimal," *Management Science*, vol. 67, no. 6, pp. 3387–3402, 2021.

[17] R. R. Bahadur and L. J. Savage, "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *The Annals of Mathematical Statistics*, vol. 27, no. 4, pp. 1115 – 1122, 1956.

[18] É. Chouzenoux, H. Gérard, and J.-C. Pesquet, "General risk measures for robust machine learning," *Foundations of Data Science*, vol. 1, no. 3, pp. 249–269, 2019.

[19] T. W. Anderson, "Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function," tech. rep., Stanford University, 1969.

[20] A. Beck, *First-Order Methods in Optimization*. SIAM, 2017.

[21] R. E. Barlow and H. D. Brunk, "The Isotonic Regression Problem and its Dual," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 140–147, 1972.

[22] A. Steerneman, "G-Majorization, group-induced cone orderings, and reflection groups," *Linear Algebra and its Applications*, vol. 127, pp. 107–119, 1990.

[23] P. Billingsley, *Probability and Measure*. John Wiley & Sons, Inc., 1995.

[24] S. S. Wilks, *Mathematical Statistics*. John Wiley & Sons, Inc., 1964.

[25] R.-D. Reiss, *Approximate Distributions of Order Statistics*. Springer Series in Statistics, Springer New York, 1989.

[26] G. R. Shorack, *Probability for statisticians*. Springer, 2nd ed., 2017.

[27] G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*. SIAM, 2009.

[28] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory, Third Edition*. SIAM, 2021.

[29] M. J. Best, N. Chakravarti, and V. A. Ubhaya, "Minimizing Separable Convex Functions Subject to Simple Chain Constraints," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 658–672, 2000.

[30] L. A. E. William Q. Meeker, Gerald J. Hahn, *Statistical Intervals: A Guide for Practitioners and Researchers*. Wiley, 2ed. ed., 2017.

[31] M. C. Campi and S. Garatti, *Introduction to the Scenario Approach*. SIAM, 2018.

[32] A. Moscovich, "Fast calculation of p-values for one-sided Kolmogorov-Smirnov type statistics," 2020, arXiv: 2009.04954.

[33] P. Coppens and P. Patrinos, "Ordered risk minimization: Learning more from less data," 2023, arXiv: 2303.09196.

[34] R. Mehta, V. Roulet, K. Pillutla, L. Liu, and Z. Harchaoui, "Stochastic optimization for spectral risk measures," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, vol. 206, pp. 10112–10159, PMLR, 2023.

[35] M. Schuurmans and P. Patrinos, "A general framework for learning-based distributionally robust MPC of Markov jump systems," *IEEE Transactions on Automatic Control*, pp. 1–16, 2023.

[36] F. A. Ramponi, "Consistency of the scenario approach," *SIAM Journal on Optimization*, vol. 28, no. 1, pp. 135–162, 2018.

[37] B. Schölkopf, A. J. Smola, F. Bach, *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[38] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

APPENDIX A
PRELIMINARIES

A. Monotone Cone

Let $\mathbb{R}^n_\uparrow := \{x \in \mathbb{R}^n : x_1 \le x_2 \le \cdots \le x_n\}$ denote the monotone cone. This cone and its polar have a history in isotonic regression [21] and majorization [22].

We show that $\mathcal{M}^n$ and $\mathbb{R}^n_\uparrow$ are related.

Lemma A.1. *Let* $\mathbb{R}^n_\uparrow$ *be the monotone cone and* $\mathcal{M}^n$ *as in* (6). *Then* $\mathcal{M}^n$ *is the polar of* $\mathbb{R}^n_\uparrow$.

Proof. The monotone cone is polyhedral with $\mathbb{R}^n_\uparrow = \{x : Mx \le 0\}$ for $M \in \mathbb{R}^{n-1 \times n}$ with $Mx = (x_1 - x_2, x_2 - x_3, \ldots, x_{n-1} - x_n)$. The definition of the polar cone is thus

$$(\mathbb{R}^n_\uparrow)^\circ = \{y \in \mathbb{R}^n : \langle x, y \rangle \le 0, \forall x \text{ s.t. } Mx \le 0\}.$$

By Farkas' lemma [38, p. 263] we have either $Mx \le 0$ and $\langle x, y \rangle > 0$ or $M^\top \lambda = y$ and $\lambda \ge 0$. So

$$(\mathbb{R}^n_\uparrow)^\circ = \{y \in \mathbb{R}^n : y = M^\top \lambda, \lambda \ge 0\}.$$

Note that $\langle \lambda, Mx \rangle = \sum_{i=1}^{n-1} \lambda_i (x_i - x_{i+1}) = \sum_{j=1}^{n} x_j (\lambda_j - \lambda_{j-1}) = \langle M^\top \lambda, x \rangle$, where $\lambda_0 = \lambda_n = 0$. Thus $y \in (\mathbb{R}^n_\uparrow)^\circ$ iff $y = M^\top \lambda$ for $\lambda \ge 0$. Here $y = M^\top \lambda$ holds iff

$$y_j = \lambda_j - \lambda_{j-1}, \qquad \forall j \in [n].$$
$$\Leftrightarrow \quad \sum_{j=1}^{k} y_j = \sum_{j=1}^{k} \lambda_j - \lambda_{j-1} = \lambda_k, \quad \forall k \in [n].$$

Since $\lambda \ge 0$ we have $\sum_{j=1}^{k} y_j = \lambda_k \ge 0$ for $k \in [n-1]$ and $\sum_{j=1}^{n} y_j = \lambda_n = 0$. These are the constraints in (6). $\square$