Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations

PEYMAN MOHAJERIN ESFAHANI AND DANIEL KUHN

Abstract. We consider stochastic programs where the distribution of the uncertain parameters is only observable through a finite training dataset. Using the Wasserstein metric, we construct a ball in the space of (multivariate and non-discrete) probability distributions centered at the uniform distribution on the training samples, and we seek decisions that perform best in view of the worst-case distribution within this Wasserstein ball. The state-of-the-art methods for solving the resulting distributionally robust optimization problems rely on global optimization techniques, which quickly become computationally excruciating. In this paper we demonstrate that, under mild assumptions, the distributionally robust optimization problems over Wasserstein balls can in fact be reformulated as finite convex programs—in many interesting cases even as tractable linear programs. Leveraging recent measure concentration results, we also show that their solutions enjoy powerful finite-sample performance guarantees. Our theoretical results are exemplified in mean-risk portfolio optimization as well as uncertainty quantification.

1. Introduction

Stochastic programming is a powerful modeling paradigm for optimization under uncertainty. The goal of a generic single-stage stochastic program is to find a decision $x \in \mathbb{R}^n$ that minimizes an expected cost $\mathbb{E}^{\mathbb{P}}[h(x,\xi)]$, where the expectation is taken with respect to the distribution $\mathbb{P}$ of the continuous random vector $\xi \in \mathbb{R}^m$. However, classical stochastic programming is challenged by the large-scale decision problems encountered in today's increasingly interconnected world. First, the distribution $\mathbb{P}$ is never observable but must be inferred from data. However, if we calibrate a stochastic program to a given dataset and evaluate its optimal decision on a different dataset, then the resulting out-of-sample performance is often disappointing—even if the two datasets are generated from the same distribution. This phenomenon is termed the *optimizer's curse* and is reminiscent of overfitting effects in statistics [48]. Second, in order to evaluate the objective function of a stochastic program for a fixed decision x, we need to compute a multivariate integral, which is #P-hard even if $h(x,\xi)$ constitutes the positive part of an affine function, while ξ is uniformly distributed on the unit hypercube [24, Corollary 1].

Distributionally robust optimization is an alternative modeling paradigm, where the objective is to find a decision x that minimizes the *worst-case* expected cost $\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}^{\mathbb{Q}}[h(x,\xi)]$. Here, the worst-case is taken over an ambiguity set $\mathcal{P}$, that is, a family of distributions characterized through certain known properties of the unknown data-generating distribution $\mathbb{P}$. Distributionally robust optimization problems have been studied since Scarf's seminal treatise on the ambiguity-averse newsvendor problem in 1958 [43], but the field has gained thrust only with the advent of modern robust optimization techniques in the last decade [3, 9].

Date: June 13, 2017.

The authors are with the Delft Center for Systems and Control, TU Delft, The Netherlands (P.MohajerinEsfahani@tudelft.nl), and the Risk Analytics and Optimization Chair, EPFL, Switzerland (daniel.kuhn@epfl.ch).

Distributionally robust optimization has the following striking benefits. First, adopting a worst-case approach regularizes the optimization problem and thereby mitigates the optimizer's curse characteristic for stochastic programming. Second, distributionally robust models are often tractable even though the corresponding stochastic model with the true data-generating distribution (which is generically continuous) are $\#P$-hard. So even if the data-generating distribution was known, the corresponding stochastic program could not be solved efficiently.

The ambiguity set $\mathcal{P}$ is a key ingredient of any distributionally robust optimization model. A good ambiguity set should be rich enough to contain the true data-generating distribution with high confidence. On the other hand, the ambiguity set should be small enough to exclude pathological distributions, which would incentivize overly conservative decisions. The ambiguity set should also be easy to parameterize from data, and—ideally—it should facilitate a tractable reformulation of the distributionally robust optimization problem as a structured mathematical program that can be solved with off-the-shelf optimization software.

Distributionally robust optimization models where ξ has finitely many realizations are reviewed in [2, 7, 39]. This paper focuses on situations where ξ can have a continuum of realizations. In this setting, the existing literature has studied three types of ambiguity sets. Moment ambiguity sets contain all distributions that satisfy certain moment constraints, see for example [18, 22, 51] or the references therein. An attractive alternative is to define the ambiguity set as a ball in the space of probability distributions by using a probability distance function such as the Prohorov metric [20], the Kullback-Leibler divergence [27, 25], or the Wasserstein metric [38, 52] etc. Such metric-based ambiguity sets contain all distributions that are close to a *nominal* or *most likely* distribution with respect to the prescribed probability metric. By adjusting the radius of the ambiguity set, the modeler can thus control the degree of conservatism of the underlying optimization problem. If the radius drops to zero, then the ambiguity set shrinks to a singleton that contains only the nominal distribution, in which case the distributionally robust problem reduces to an ambiguity-free stochastic program. In addition, ambiguity sets can also be defined as confidence regions of goodness-of-fit tests [7].

In this paper we study distributionally robust optimization problems with a *Wasserstein ambiguity set* centered at the uniform distribution $\widehat{\mathbb{P}}_N$ on N independent and identically distributed training samples. The Wasserstein distance of two distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ can be viewed as the minimum transportation cost for moving the probability mass from $\mathbb{Q}_1$ to $\mathbb{Q}_2$, and the Wasserstein ambiguity set contains all (continuous or discrete) distributions that are sufficiently close to the (discrete) empirical distribution $\widehat{\mathbb{P}}_N$ with respect to the Wasserstein metric. Modern measure concentration results from statistics guarantee that the unknown data-generating distribution $\mathbb{P}$ belongs to the Wasserstein ambiguity set around $\widehat{\mathbb{P}}_N$ with confidence $1-\beta$ if its radius is a sublinearly growing function of $\log(1/\beta)/N$ [11, 21]. The optimal value of the distributionally robust problem thus provides an upper confidence bound on the achievable out-of-sample cost.

While Wasserstein ambiguity sets offer powerful out-of-sample performance guarantees and enable the decision maker to control the model's conservativeness, moment-based ambiguity sets appear to display better tractability properties. Specifically, there is growing evidence that distributionally robust models with moment ambiguity sets are more tractable than the corresponding stochastic models because the intractable high-dimensional integrals in the objective function are replaced with tractable (generalized) moment problems [18, 22, 51]. In contrast, distributionally robust models with Wasserstein ambiguity sets are believed to be harder than their stochastic counterparts [36]. Indeed, the state-of-the-art method for computing the worst-case expectation over a Wasserstein ambiguity set $\mathcal{P}$ relies on global optimization techniques. Exploiting the fact that the extreme points of $\mathcal{P}$ are discrete distributions with a fixed number of atoms [52], one may reformulate the original worst-case expectation problem as a finite-dimensional non-convex program, which can be solved via "difference of convex programming" methods, see [52] or [36, Section 7.1]. However, the

computational effort is reported to be considerable, and there is no guarantee to find the global optimum. Nevertheless, tractability results are available for special cases. Specifically, the worst case of a convex law-invariant risk measure with respect to a Wasserstein ambiguity set $\mathcal{P}$ reduces to the sum of the nominal risk and a regularization term whenever $h(x, \xi)$ is affine in ξ and $\mathcal{P}$ does not include any support constraints [53]. Moreover, while this paper was under review we became aware of the PhD thesis [54], which reformulates a distributionally robust two-stage unit commitment problem over a Wasserstein ambiguity set as a semi-infinite linear program, which is subsequently solved using a Benders decomposition algorithm.

The main contribution of this paper is to demonstrate that the worst-case expectation over a Wasserstein ambiguity set can in fact be computed efficiently via convex optimization techniques for numerous loss functions of practical interest. Furthermore, we propose an efficient procedure for constructing an extremal distribution that attains the worst-case expectation—provided that such a distribution exists. Otherwise, we construct a sequence of distributions that attain the worst-case expectation asymptotically. As a by-product, our analysis shows that many interesting distributionally robust optimization problems with Wasserstein ambiguity sets can be solved in polynomial time. We also investigate the out-of-sample performance of the resulting optimal decisions—both theoretically and experimentally—and analyze its dependence on the number of training samples. We highlight the following main contributions of this paper.

- We prove that the worst-case expectation of an uncertain loss $\ell(\xi)$ over a Wasserstein ambiguity set coincides with the optimal value of a finite-dimensional convex program if $\ell(\xi)$ constitutes a pointwise maximum of finitely many concave functions. Generalizations to convex functions or to sums of maxima of concave functions are also discussed. We conclude that worst-case expectations can be computed efficiently to high precision via modern convex optimization algorithms.
- We describe a supplementary finite-dimensional convex program whose optimal (near-optimal) solutions can be used to construct exact (approximate) extremal distributions for the infinite-dimensional worst-case expectation problem.
- We show that the worst-case expectation reduces to the optimal value of an explicit linear program if the 1-norm or the ∞-norm is used in the definition of the Wasserstein metric and if $\ell(\xi)$ belongs to any of the following function classes: (1) a pointwise maximum or minimum of affine functions; (2) the indicator function of a closed polytope or the indicator function of the complement of an open polytope; (3) the optimal value of a parametric linear program whose cost or right-hand side coefficients depend linearly on ξ.
- Using recent measure concentration results from statistics, we demonstrate that the optimal value of a distributionally robust optimization problem over a Wasserstein ambiguity set provides an upper confidence bound on the out-of-sample cost of the worst-case optimal decision. We validate this theoretical performance guarantee in numerical tests.

If the uncertain parameter vector ξ is confined to a fixed finite subset of $\mathbb{R}^m$, then the worst-case expectation problems over Wasserstein ambiguity sets simplify substantially and can often be reformulated as tractable conic programs by leveraging ideas from robust optimization. An elegant second-order conic reformulation has been discovered, for instance, in the context of distributionally robust regression analysis [32], and a comprehensive list of tractable reformulations of distributionally robust risk constraints for various risk measures is provided in [39]. Our paper extends these tractability results to the practically relevant case where ξ has uncountably many possible realizations—without resorting to space tessellation or discretization techniques that are prone to the curse of dimensionality.

When $\ell(\xi)$ is linear and the distribution of ξ ranges over a Wasserstein ambiguity set without support constraints, one can derive a concise closed-form expression for the worst-case risk of $\ell(\xi)$ for various convex

risk measures [53]. However, these analytical solutions come at the expense of a loss of generality. We believe that the results of this paper may pave the way towards an efficient computational procedure for evaluating the worst-case risk of $\ell(\xi)$ in more general settings where the loss function may be non-linear and ξ may be subject to support constraints.

Among all metric-based ambiguity sets studied to date, the Kullback-Leibler ambiguity set has attracted most attention from the robust optimization community. It has first been used in financial portfolio optimization to capture the distributional uncertainty of asset returns with a Gaussian nominal distribution [19]. Subsequent work has focused on Kullback-Leibler ambiguity sets for discrete distributions with a fixed support, which offer additional modeling flexibility without sacrificing computational tractability [14, 2]. It is also known that distributionally robust chance constraints involving a generic Kullback-Leibler ambiguity set are equivalent to the respective classical chance constraints under the nominal distribution but with a rescaled violation probability [27, 26]. Moreover, closed-form counterparts of distributionally robust expectation constraints with Kullback-Leibler ambiguity sets have been derived in [25].

However, Kullback-Leibler ambiguity sets typically fail to represent confidence sets for the unknown distribution $\mathbb{P}$. To see this, assume that $\mathbb{P}$ is absolutely continuous with respect to the Lebesgue measure and that the ambiguity set is centered at the discrete empirical distribution $\widehat{\mathbb{P}}_N$. Then, any distribution in a Kullback-Leibler ambiguity set around $\widehat{\mathbb{P}}_N$ must assign positive probability mass to each training sample. As $\mathbb{P}$ has a density function, it must therefore reside outside of the Kullback-Leibler ambiguity set irrespective of the training samples. Thus, Kullback-Leibler ambiguity sets around $\widehat{\mathbb{P}}_N$ contain $\mathbb{P}$ with probability 0. In contrast, Wasserstein ambiguity sets centered at $\widehat{\mathbb{P}}_N$ contain discrete as well as continuous distributions and, if properly calibrated, represent meaningful confidence sets for $\mathbb{P}$. We will exploit this property in Section 3 to derive finite-sample guarantees. A comparison and critical assessment of various metric-based ambiguity sets is provided in [45]. Specifically, it is shown that worst-case expectations over Kullback-Leibler and other divergence-based ambiguity sets are law invariant. In contrast, worst-case expectations over Wasserstein ambiguity sets are not. The law invariance can be exploited to evaluate worst-case expectations via the sample average approximation.

The models proposed in this paper fall within the scope of data-driven distributionally robust optimization [20, 16, 7, 23]. Closest in spirit to our work is the robust sample average approximation [7], which seeks decisions that are robust with respect to the ambiguity set of all distributions that pass a prescribed statistical hypothesis test. Indeed, the distributions within the Wasserstein ambiguity set could be viewed as those that pass a multivariate goodness-of-fit test in light of the available training samples. This amounts to interpreting the Wasserstein distance between the empirical distribution $\widehat{\mathbb{P}}_N$ and a given hypothesis $\mathbb{Q}$ as a test statistic and the radius of the Wasserstein ambiguity set as a threshold that needs to be chosen in view of the test's desired significance level β. The Wasserstein distance has already been used in tests for normality [17] and to devise nonparametric homogeneity tests [40].

The rest of the paper proceeds as follows. Section 2 sketches a generic framework for data-driven distributionally robust optimization, while Section 3 introduces our specific approach based on Wasserstein ambiguity sets and establishes its out-of-sample performance guarantees. In Section 4 we demonstrate that many worst-case expectation problems over Wasserstein ambiguity sets can be reduced to finite-dimensional convex programs, and we develop a systematic procedure for constructing worst-case distributions. Explicit linear programming reformulations of distributionally robust single and two-stage stochastic programs as well as uncertainty quantification problems are derived in Section 5. Section 6 extends the scope of the basic approach to broader classes of objective functions, and Section 7 reports on numerical results.

Notation. We denote by $\mathbb{R}_+$ the non-negative and by $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{-\infty, \infty\}$ the extended reals. Throughout this paper, we adopt the conventions of extended arithmetics, whereby $\infty \cdot 0 = 0 \cdot \infty = 0/0 = 0$ and $\infty - \infty = -\infty + \infty = 1/0 = \infty$. The inner product of two vectors $a, b \in \mathbb{R}^m$ is denoted by $\langle a, b \rangle \coloneqq a^\intercal b$. Given a norm $\|\cdot\|$ on $\mathbb{R}^m$, the dual norm is defined through $\|z\|_* \coloneqq \sup_{\|\xi\| \leq 1} \langle z, \xi \rangle$. A function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is proper if $f(\xi) < +\infty$ for at least one ξ and $f(\xi) > -\infty$ for every ξ in $\mathbb{R}^m$. The conjugate of f is defined as $f^*(z) \coloneqq \sup_{\xi \in \mathbb{R}^m} \langle z, \xi \rangle - f(\xi)$. Note that conjugacy preserves properness. For a set $\Xi \subseteq \mathbb{R}^m$, the indicator function $\mathbb{1}_\Xi$ is defined through $\mathbb{1}_\Xi(\xi) = 1$ if $\xi \in \Xi$; $= 0$ otherwise. Similarly, the characteristic function χ_Ξ is defined via $\chi_\Xi(\xi) = 0$ if $\xi \in \Xi$; $= \infty$ otherwise. The support function of Ξ is defined as $\sigma_\Xi(z) \coloneqq \sup_{\xi \in \Xi} \langle z, \xi \rangle$. It coincides with the conjugate of χ_Ξ. We denote by δ_ξ the Dirac distribution concentrating unit mass at $\xi \in \mathbb{R}^m$. The product of two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on Ξ_1 and Ξ_2, respectively, is the distribution $\mathbb{P}_1 \otimes \mathbb{P}_2$ on $\Xi_1 \times \Xi_2$. The N-fold product of a distribution $\mathbb{P}$ on Ξ is denoted by $\mathbb{P}^N$, which represents a distribution on the Cartesian product space Ξ^N. Finally, we set the expectation of $\ell : \Xi \to \overline{\mathbb{R}}$ under $\mathbb{P}$ to $\mathbb{E}^{\mathbb{P}}[\ell(\xi)] = \mathbb{E}^{\mathbb{P}}\big[\max\{\ell(\xi), 0\}\big] + \mathbb{E}^{\mathbb{P}}\big[\min\{\ell(\xi), 0\}\big]$, which is well-defined by the conventions of extended arithmetics.

2. Data-Driven Stochastic Programming

Consider the stochastic program

$$J^\star \coloneqq \inf_{x \in \mathbb{X}} \left\{ \mathbb{E}^{\mathbb{P}}\big[h(x, \xi)\big] = \int_\Xi h(x, \xi)\, \mathbb{P}(\mathrm{d}\xi) \right\} \tag{1}$$

with feasible set $\mathbb{X} \subseteq \mathbb{R}^n$, uncertainty set $\Xi \subseteq \mathbb{R}^m$ and loss function $h : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$. The loss function depends both on the decision vector $x \in \mathbb{R}^n$ and the random vector $\xi \in \mathbb{R}^m$, whose distribution $\mathbb{P}$ is supported on Ξ. Problem (1) can be viewed as the first-stage problem of a two-stage stochastic program, where $h(x, \xi)$ represents the optimal value of a subordinate second-stage problem [46]. Alternatively, problem (1) may also be interpreted as a generic learning problem in the spirit of [49].

Unfortunately, in most situations of practical interest, the distribution $\mathbb{P}$ is not precisely known, and therefore we miss essential information to solve problem (1) *exactly*. However, $\mathbb{P}$ is often partially observable through a finite set of N independent samples, *e.g.*, past realizations of the random vector ξ. We denote the training dataset comprising these samples by $\widehat{\Xi}_N \coloneqq \{\widehat{\xi}_i\}_{i \leq N} \subseteq \Xi$. We emphasize that—before its revelation—the dataset $\widehat{\Xi}_N$ can be viewed as a random object governed by the distribution $\mathbb{P}^N$ supported on Ξ^N.

A *data-driven solution* for problem (1) is a feasible decision $\widehat{x}_N \in \mathbb{X}$ that is constructed from the training dataset $\widehat{\Xi}_N$. Throughout this paper, we notationally suppress the dependence of $\widehat{x}_N$ on the training samples in order to avoid clutter. Instead, we reserve the superscript '$\widehat{\ }$' for objects that depend on the training data and thus constitute random objects governed by the product distribution $\mathbb{P}^N$. The *out-of-sample performance* of $\widehat{x}_N$ is defined as $\mathbb{E}^{\mathbb{P}}\big[h(\widehat{x}_N, \xi)\big]$ and can thus be viewed as the expected cost of $\widehat{x}_N$ under a new sample ξ that is independent of the training dataset. As $\mathbb{P}$ is unknown, however, the exact out-of-sample performance cannot be evaluated in practice, and the best we can hope for is to establish *performance guarantees* in the form of tight bounds. The feasibility of $\widehat{x}_N$ in (1) implies $J^\star \leq \mathbb{E}^{\mathbb{P}}\big[h(\widehat{x}_N, \xi)\big]$, but this lower bound is again of limited use as $J^\star$ is unknown and as our primary concern is to bound the costs from above. Thus, we seek data-driven solutions $\widehat{x}_N$ with performance guarantees of the type

$$\mathbb{P}^N\Big\{\widehat{\Xi}_N \;:\; \mathbb{E}^{\mathbb{P}}\big[h(\widehat{x}_N, \xi)\big] \leq \widehat{J}_N\Big\} \geq 1 - \beta, \tag{2}$$

where $\widehat{J}_N$ constitutes an upper bound that may depend on the training dataset, and $\beta \in (0, 1)$ is a *significance parameter* with respect to the distribution $\mathbb{P}^N$, which governs both $\widehat{x}_N$ and $\widehat{J}_N$. Hereafter we refer to $\widehat{J}_N$ as a *certificate* for the out-of-sample performance of $\widehat{x}_N$ and to the probability on the left-hand side of (2) as its

reliability. Our ideal goal is to find a data-driven solution with the lowest possible out-of-sample performance. This is impossible, however, as $\mathbb{P}$ is unknown, and the out-of-sample performance cannot be computed. We thus pursue the more modest but achievable goal to find a data-driven solution with a low certificate and a high reliability.

A natural approach to generate data-driven solutions $\widehat{x}_N$ is to approximate $\mathbb{P}$ with the discrete empirical probability distribution

$$\widehat{\mathbb{P}}_N \coloneqq \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i}, \tag{3}$$

that is, the uniform distribution on $\widehat{\Xi}_N$. This amounts to approximating the original stochastic program (1) with the *sample-average approximation* (SAA) problem

$$\widehat{J}_{\mathrm{SAA}} \coloneqq \inf_{x\in\mathbb{X}} \left\{ \mathbb{E}^{\widehat{\mathbb{P}}_N}\big[h(x,\xi)\big] = \frac{1}{N}\sum_{i=1}^{N} h(x,\widehat{\xi}_i) \right\}. \tag{4}$$

If the feasible set $\mathbb{X}$ is compact and the loss function is uniformly continuous in x across all $\xi \in \Xi$, then the optimal value and optimal solutions of the SAA problem (4) converge almost surely to their counterparts of the true problem (1) as N tends to infinity [46, Theorem 5.3]. Even though finite sample performance guarantees of the type (2) can be obtained under additional assumptions such as Lipschitz continuity of the loss function (see *e.g.*, [47, Theorem 1]), the SAA problem has been conceived primarily for situations where the distribution $\mathbb{P}$ is known and additional samples can be acquired cheaply via random number generation. However, the optimal solutions of the SAA problem tend to display a poor out-of-sample performance in situations where N is small and where the acquisition of additional samples would be costly.

In this paper we address problem (1) with an alternative approach that explicitly accounts for our ignorance of the true data-generating distribution $\mathbb{P}$, and that offers attractive performance guarantees even when the acquisition of additional samples from $\mathbb{P}$ is impossible or expensive. Specifically, we use $\widehat{\Xi}_N$ to design an ambiguity set $\widehat{\mathcal{P}}_N$ containing all distributions that could have generated the training samples with high confidence. This ambiguity set enables us to define the certificate $\widehat{J}_N$ as the optimal value of a distributionally robust optimization problem that minimize the *worst-case* expected cost.

$$\widehat{J}_N \coloneqq \inf_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}\big[h(x,\xi)\big] \tag{5}$$

Following [38], we construct $\widehat{\mathcal{P}}_N$ as a ball around the empirical distribution (3) with respect to the Wasserstein metric. In the remainder of the paper we will demonstrate that the optimal value $\widehat{J}_N$ as well as any optimal solution $\widehat{x}_N$ (if it exists) of the distributionally robust problem (5) satisfy the following conditions.

(i) **Finite sample guarantee:** For a carefully chosen size of the ambiguity set, the certificate $\widehat{J}_N$ provides a $1-\beta$ confidence bound of the type (2) on the out-of-sample performance of $\widehat{x}_N$.

(ii) **Asymptotic consistency:** As N tends to infinity, the certificate $\widehat{J}_N$ and the data-driven solution $\widehat{x}_N$ converge—in a sense to be made precise below—to the optimal value $J^\star$ and an optimizer $x^\star$ of the stochastic program (1), respectively.

(iii) **Tractability:** For many loss functions $h(x,\xi)$ and sets $\mathbb{X}$, the distributionally robust problem (5) is computationally tractable and admits a reformulation reminiscent of the SAA problem (4).

Conditions (i)–(iii) have been identified in [7] as desirable properties of data-driven solutions for stochastic programs. Precise statements of these conditions will be provided in the remainder. In Section 3 we will use the Wasserstein metric to construct ambiguity sets of the type $\widehat{\mathcal{P}}_N$ satisfying the conditions (i) and

(ii). In Section 4, we will demonstrate that these ambiguity sets also fulfill the tractability condition (iii). We see this last result as the main contribution of this paper because the state-of-the-art method for solving distributionally robust problems over Wasserstein ambiguity sets relies on global optimization algorithms [36].

3. Wasserstein Metric and Measure Concentration

Probability metrics represent distance functions on the space of probability distributions. One of the most widely used examples is the Wasserstein metric, which is defined on the space $\mathcal{M}(\Xi)$ of all probability distributions $\mathbb{Q}$ supported on Ξ with $\mathbb{E}^{\mathbb{Q}}\big[\|\xi\|\big] = \int_{\Xi} \|\xi\|\, \mathbb{Q}(\mathrm{d}\xi) < \infty$.

Definition 3.1 (Wasserstein metric [29]). *The Wasserstein metric $d_{\mathrm{W}} : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \to \mathbb{R}$ is defined via*

$$d_{\mathrm{W}}\big(\mathbb{Q}_1, \mathbb{Q}_2\big) \coloneqq \inf \left\{ \int_{\Xi^2} \|\xi_1 - \xi_2\|\, \Pi(\mathrm{d}\xi_1, \mathrm{d}\xi_2) \;:\; \begin{array}{l} \Pi \text{ is a joint distribution of } \xi_1 \text{ and } \xi_2 \\ \text{with marginals } \mathbb{Q}_1 \text{ and } \mathbb{Q}_2, \text{ respectively} \end{array} \right\}$$

for all distributions $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{M}(\Xi)$, where $\|\cdot\|$ represents an arbitrary norm on $\mathbb{R}^m$.

The decision variable Π can be viewed as a *transportation plan* for moving a mass distribution described by $\mathbb{Q}_1$ to another one described by $\mathbb{Q}_2$. Thus, the Wasserstein distance between $\mathbb{Q}_1$ and $\mathbb{Q}_2$ represents the cost of an optimal mass transportation plan, where the norm $\|\cdot\|$ encodes the transportation costs. We remark that a generalized p-Wasserstein metric for $p \geq 1$ is obtained by setting the transportation cost between ξ_1 and ξ_2 to $\|\xi_1 - \xi_2\|^p$. In this paper, however, we focus exclusively on the 1-Wasserstein metric of Definition 3.1, which is sometimes also referred to as the Kantorovich metric.

We will sometimes also need the following dual representation of the Wasserstein metric.

Theorem 3.2 (Kantorovich-Rubinstein [29]). *For any distributions $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{M}(\Xi)$ we have*

$$d_{\mathrm{W}}\big(\mathbb{Q}_1, \mathbb{Q}_2\big) = \sup_{f \in \mathcal{L}} \left\{ \int_{\Xi} f(\xi)\, \mathbb{Q}_1(\mathrm{d}\xi) - \int_{\Xi} f(\xi)\, \mathbb{Q}_2(\mathrm{d}\xi) \right\},$$

where $\mathcal{L}$ denotes the space of all Lipschitz functions with $|f(\xi) - f(\xi')| \leq \|\xi - \xi'\|$ for all $\xi, \xi' \in \Xi$.

Kantorovich and Rubinstein [29] originally established this result for distributions with bounded support. A modern proof for unbounded distributions is due to Villani [50, Remark 6.5, p. 107]. The optimization problems in Definition 3.1 and Theorem 3.2, which provide two equivalent characterizations of the Wasserstein metric, constitute a primal-dual pair of infinite-dimensional linear programs. The dual representation implies that two distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ are close to each other with respect to the Wasserstein metric if and only if all functions with uniformly bounded slopes have similar integrals under $\mathbb{Q}_1$ and $\mathbb{Q}_2$. Theorem 3.2 also demonstrates that the Wasserstein metric is a special instance of an integral probability metric (see *e.g.* [33]) and that its generating function class coincides with a family of Lipschitz continuous functions.

In the remainder we will examine the ambiguity set

$$\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N) \coloneqq \Big\{ \mathbb{Q} \in \mathcal{M}(\Xi) \;:\; d_{\mathrm{W}}\big(\widehat{\mathbb{P}}_N, \mathbb{Q}\big) \leq \varepsilon \Big\}, \tag{6}$$

which can be viewed as the Wasserstein ball of radius ε centered at the empirical distribution $\widehat{\mathbb{P}}_N$. Under a common light tail assumption on the unknown data-generating distribution $\mathbb{P}$, this ambiguity set offers attractive performance guarantees in the spirit of Section 2.

Assumption 3.3 (Light-tailed distribution). *There exists an exponent $a > 1$ such that*

$$A \coloneqq \mathbb{E}^{\mathbb{P}}\big[\exp(\|\xi\|^a)\big] = \int_{\Xi} \exp(\|\xi\|^a)\, \mathbb{P}(\mathrm{d}\xi) < \infty.$$

Assumption 3.3 essentially requires the tail of the distribution $\mathbb{P}$ to decay at an exponential rate. Note that this assumption trivially holds if Ξ is compact. Heavy-tailed distributions that fail to meet Assumption 3.3 are difficult to handle even in the context of the classical sample average approximation. Indeed, under a heavy-tailed distribution the sample average of the loss corresponding to any fixed decision $x \in \mathbb{X}$ may not even converge to the expected loss; see *e.g.* [13, 15]. The following modern measure concentration result provides the basis for establishing powerful finite sample guarantees.

Theorem 3.4 (Measure concentration [21, Theorem 2]). *If Assumption 3.3 holds, we have*

$$\mathbb{P}^N\Big\{d_{\mathrm{W}}\big(\mathbb{P},\widehat{\mathbb{P}}_N\big)\geq\varepsilon\Big\}\leq\begin{cases} c_1\exp\big(-c_2N\varepsilon^{\max\{m,2\}}\big) & \text{if } \varepsilon\leq 1,\\ c_1\exp\big(-c_2N\varepsilon^{a}\big) & \text{if } \varepsilon>1,\end{cases} \tag{7}$$

for all $N\geq 1$, $m\neq 2$, and $\varepsilon>0$, where c_1,c_2 are positive constants that only depend on a, A, and m.[1]

Theorem 3.4 provides an a priori estimate of the probability that the unknown data-generating distribution $\mathbb{P}$ resides outside of the Wasserstein ball $\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$. Thus, we can use Theorem 3.4 to estimate the radius of the smallest Wasserstein ball that contains $\mathbb{P}$ with confidence $1-\beta$ for some prescribed $\beta\in(0,1)$. Indeed, equating the right-hand side of (7) to β and solving for ε yields

$$\varepsilon_N(\beta):=\begin{cases}\left(\frac{\log(c_1\beta^{-1})}{c_2N}\right)^{1/\max\{m,2\}} & \text{if } N\geq\frac{\log(c_1\beta^{-1})}{c_2},\\ \left(\frac{\log(c_1\beta^{-1})}{c_2N}\right)^{1/a} & \text{if } N<\frac{\log(c_1\beta^{-1})}{c_2}.\end{cases} \tag{8}$$

Note that the Wasserstein ball with radius $\varepsilon_N(\beta)$ can thus be viewed as a confidence set for the unknown true distribution as in statistical testing; see also [7].

Theorem 3.5 (Finite sample guarantee). *Suppose that Assumption 3.3 holds and that $\beta\in(0,1)$. Assume also that $\widehat{J}_N$ and $\widehat{x}_N$ represent the optimal value and an optimizer of the distributionally robust program (5) with ambiguity set $\widehat{\mathcal{P}}_N=\mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)$. Then, the finite sample guarantee (2) holds.*

Proof. The claim follows immediately from Theorem 3.4, which ensures via the definition of $\varepsilon_N(\beta)$ in (8) that $\mathbb{P}^N\{\mathbb{P}\in\mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)\}\geq 1-\beta$. Thus, $\mathbb{E}^{\mathbb{P}}[h(\widehat{x}_N,\xi)]\leq\sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N}\mathbb{E}^{\mathbb{Q}}[h(\widehat{x}_N,\xi)]=\widehat{J}_N$ with probability $1-\beta$. □

It is clear from (8) that for any fixed $\beta>0$, the radius $\varepsilon_N(\beta)$ tends to 0 as N increases. Moreover, one can show that if β_N converges to zero at a carefully chosen rate, then the solution of the distributionally robust optimization problem (5) with ambiguity set $\widehat{\mathcal{P}}_N=\mathbb{B}_{\varepsilon_N(\beta_N)}(\widehat{\mathbb{P}}_N)$ converges to the solution of the original stochastic program (1) as N tends to infinity. The following theorem formalizes this statement.

Theorem 3.6 (Asymptotic consistency). *Suppose that Assumption 3.3 holds and that $\beta_N\in(0,1)$, $N\in\mathbb{N}$, satisfies $\sum_{N=1}^\infty\beta_N<\infty$ and $\lim_{N\to\infty}\varepsilon_N(\beta_N)=0$.*[2] *Assume also that $\widehat{J}_N$ and $\widehat{x}_N$ represent the optimal value and an optimizer of the distributionally robust program (5) with ambiguity set $\widehat{\mathcal{P}}_N=\mathbb{B}_{\varepsilon_N(\beta_N)}(\widehat{\mathbb{P}}_N)$, $N\in\mathbb{N}$.*

(i) If $h(x,\xi)$ is upper semicontinuous in ξ and there exists $L\geq 0$ with $|h(x,\xi)|\leq L(1+\|\xi\|)$ for all $x\in\mathbb{X}$ and $\xi\in\Xi$, then $\mathbb{P}^\infty$-almost surely we have $\widehat{J}_N\downarrow J^\star$ as $N\to\infty$ where $J^\star$ is the optimal value of (1).

(ii) If the assumptions of assertion (i) hold, $\mathbb{X}$ is closed, and $h(x,\xi)$ is lower semicontinuous in x for every $\xi\in\Xi$, then any accumulation point of $\{\widehat{x}_N\}_{N\in\mathbb{N}}$ is $\mathbb{P}^\infty$-almost surely an optimal solution for (1).

The proof of Theorem 3.6 will rely on the following technical lemma.

[1] A similar but slightly more complicated inequality also holds for the special case $m=2$; see [21, Theorem 2] for details.

[2] A possible choice is $\beta_N=\exp(-\sqrt{N})$.

Lemma 3.7 (Convergence of distributions)**.** *If Assumption 3.3 holds and $\beta_N \in (0,1)$, $N \in \mathbb{N}$, satisfies $\sum_{N=1}^{\infty} \beta_N < \infty$ and $\lim_{N\to\infty} \varepsilon_N(\beta_N) = 0$, then, any sequence $\widehat{\mathbb{Q}}_N \in \mathbb{B}_{\varepsilon_N(\beta_N)}(\widehat{\mathbb{P}}_N)$, $N \in \mathbb{N}$, where $\widehat{\mathbb{Q}}_N$ may depend on the training data, converges under the Wasserstein metric (and thus weakly) to $\mathbb{P}$ almost surely with respect to $\mathbb{P}^\infty$, that is,*

$$\mathbb{P}^\infty \left\{ \lim_{N\to\infty} d_{\mathrm{W}}\big(\mathbb{P}, \widehat{\mathbb{Q}}_N\big) = 0 \right\} = 1.$$

Proof. As $\widehat{\mathbb{Q}}_N \in \mathbb{B}_{\delta_N}(\widehat{\mathbb{P}}_N)$, the triangle inequality for the Wasserstein metric ensures that

$$d_{\mathrm{W}}\big(\mathbb{P}, \widehat{\mathbb{Q}}_N\big) \le d_{\mathrm{W}}\big(\mathbb{P}, \widehat{\mathbb{P}}_N\big) + d_{\mathrm{W}}\big(\widehat{\mathbb{P}}_N, \widehat{\mathbb{Q}}_N\big) \le d_{\mathrm{W}}\big(\mathbb{P}, \widehat{\mathbb{P}}_N\big) + \varepsilon_N(\beta_N).$$

Moreover, Theorem 3.4 implies that $\mathbb{P}^N\{d_{\mathrm{W}}\big(\mathbb{P}, \widehat{\mathbb{P}}_N\big) \le \varepsilon_N(\beta_N)\} \ge 1-\beta_N$, and thus we have $\mathbb{P}^N\{d_{\mathrm{W}}\big(\mathbb{P}, \widehat{\mathbb{Q}}_N\big) \le 2\varepsilon_N(\beta_N)\} \ge 1-\beta_N$. As $\sum_{N=1}^{\infty} \beta_N < \infty$, the Borel-Cantelli Lemma [28, Theorem 2.18] further implies that

$$\mathbb{P}^\infty \left\{ d_{\mathrm{W}}\big(\mathbb{P}, \widehat{\mathbb{Q}}_N\big) \le \varepsilon_N(\beta_N) \text{ for all sufficiently large } N \right\} = 1.$$

Finally, as $\lim_{N\uparrow\infty} \varepsilon_N(\beta_N) = 0$, we conclude that $\lim_{N\uparrow\infty} d_{\mathrm{W}}\big(\mathbb{P}, \widehat{\mathbb{Q}}_N\big) = 0$ almost surely. Note that convergence with respect to the Wasserstein metric implies weak convergence [10]. □

Proof of Theorem 3.6. As $\widehat{x}_N \in \mathbb{X}$, we have $J^\star \le \mathbb{E}^{\mathbb{P}}[h(\widehat{x}_N, \xi)]$. Moreover, Theorem 3.5 implies that

$$\mathbb{P}^N \left\{ J^\star \le \mathbb{E}^{\mathbb{P}}[h(\widehat{x}_N, \xi)] \le \widehat{J}_N \right\} \ge \mathbb{P}^N \left\{ \mathbb{P} \in \mathbb{B}_{\varepsilon_N(\beta_N)}(\widehat{\mathbb{P}}_N) \right\} \ge 1 - \beta_N,$$

for all $N \in \mathbb{N}$. As $\sum_{N=1}^{\infty} \beta_N < \infty$, the Borel-Cantelli Lemma further implies that

$$\mathbb{P}^\infty \left\{ J^\star \le \mathbb{E}^{\mathbb{P}}[h(\widehat{x}_N, \xi)] \le \widehat{J}_N \text{ for all sufficiently large } N \right\} = 1.$$

To prove assertion (i), it thus remains to be shown that $\limsup_{N\to\infty} \widehat{J}_N \le J^\star$ with probability 1. As $h(x,\xi)$ is upper semicontinuous and grows at most linearly in ξ, there exists a non-increasing sequence of functions $h_k(x,\xi)$, $k \in \mathbb{N}$, such that $h(x,\xi) = \lim_{k\to\infty} h_k(x,\xi)$, and $h_k(x,\xi)$ is Lipschitz continuous in ξ for any fixed $x \in \mathbb{X}$ and $k \in \mathbb{N}$ with Lipschitz constant $L_k \ge 0$; see Lemma A.1 in the appendix. Next, choose any $\delta > 0$, fix a δ-optimal decision $x_\delta \in \mathbb{X}$ for (1) with $\mathbb{E}^{\mathbb{P}}[h(x_\delta, \xi)] \le J^\star + \delta$, and for every $N \in \mathbb{N}$ let $\widehat{\mathbb{Q}}_N \in \widehat{\mathcal{P}}_N$ be a δ-optimal distribution corresponding to x_δ with

$$\sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}[h(x_\delta, \xi)] \le \mathbb{E}^{\mathbb{Q}_N}[h(x_\delta, \xi)] + \delta.$$

Then, we have

$$\begin{aligned}
\limsup_{N\to\infty} \widehat{J}_N \le \limsup_{N\to\infty} \sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}[h(x_\delta, \xi)] &\le \limsup_{N\to\infty} \mathbb{E}^{\widehat{\mathbb{Q}}_N}[h(x_\delta, \xi)] + \delta \\
&\le \lim_{k\to\infty} \limsup_{N\to\infty} \mathbb{E}^{\widehat{\mathbb{Q}}_N}[h_k(x_\delta, \xi)] + \delta \\
&\le \lim_{k\to\infty} \limsup_{N\to\infty} \Big(\mathbb{E}^{\mathbb{P}}[h_k(x_\delta, \xi)] + L_k\, d_{\mathrm{W}}\big(\mathbb{P}, \widehat{\mathbb{Q}}_N\big) \Big) + \delta \\
&= \lim_{k\to\infty} \mathbb{E}^{\mathbb{P}}[h_k(x_\delta, \xi)] + \delta, \quad \mathbb{P}^\infty\text{-almost surely} \\
&= \mathbb{E}^{\mathbb{P}}[h(x_\delta, \xi)] + \delta \le J^\star + 2\delta,
\end{aligned}$$

where the second inequality holds because $h_k(x,\xi)$ converges from above to $h(x,\xi)$, and the third inequality follows from Theorem 3.2. Moreover, the almost sure equality holds due to Lemma 3.7, and the last equality follows from the Monotone Convergence Theorem [30, Theorem 5.5], which applies because $|\mathbb{E}^{\mathbb{P}}[h_k(x_\delta, \xi)]| < \infty$. Indeed, recall that $\mathbb{P}$ has an exponentially decaying tail due to Assumption 3.3 and that $h_k(x_\delta, \xi)$ is Lipschitz continuous in ξ. As $\delta > 0$ was chosen arbitrarily, we thus conclude that $\limsup_{N\to\infty} \widehat{J}_N \le J^\star$.

To prove assertion (ii), fix an arbitrary realization of the stochastic process $\{\widehat{\xi}_N\}_{N\in\mathbb{N}}$ such that $J^\star = \lim_{N\to\infty} \widehat{J}_N$ and $J^\star \leq \mathbb{E}^{\mathbb{P}}[h(\widehat{x}_N,\xi)] \leq \widehat{J}_N$ for all sufficiently large N. From the proof of assertion (i) we know that these two conditions are satisfied $\mathbb{P}^\infty$-almost surely. Using these assumptions, one easily verifies that

$$\liminf_{N\to\infty} \mathbb{E}^{\mathbb{P}}[h(\widehat{x}_N,\xi)] \leq \lim_{N\to\infty} \widehat{J}_N = J^\star. \tag{9}$$

Next, let $x^\star$ be an accumulation point of the sequence $\{\widehat{x}_N\}_{N\in\mathbb{N}}$, and note that $x^\star \in \mathbb{X}$ as $\mathbb{X}$ is closed. By passing to a subsequence, if necessary, we may assume without loss of generality that $x^\star = \lim_{N\to\infty} \widehat{x}_N$. Thus,

$$J^\star \leq \mathbb{E}^{\mathbb{P}}[h(x^\star,\xi)] \leq \mathbb{E}^{\mathbb{P}}[\liminf_{N\to\infty} h(\widehat{x}_N,\xi)] \leq \liminf_{N\to\infty} \mathbb{E}^{\mathbb{P}}[h(\widehat{x}_N,\xi)] \leq J^\star,$$

where the first inequality exploits that $x^\star \in \mathbb{X}$, the second inequality follows from the lower semicontinuity of $h(x,\xi)$ in x, the third inequality holds due to Fatou's lemma (which applies because $h(x,\xi)$ grows at most linearly in ξ), and the last inequality follows from (9). Therefore, we have $\mathbb{E}^{\mathbb{P}}[h(x^\star,\xi)] = J^\star$. □

In the following we show that all assumptions of Theorem 3.6 are necessary for asymptotic convergence, that is, relaxing any of these conditions can invalidate the convergence result.

Example 1 (Necessity of regularity conditions).

(1) *Upper semicontinuity of $\xi \mapsto h(x,\xi)$ in Theorem 3.6 (i):*
Set $\Xi = [0,1]$, $\mathbb{P} = \delta_0$ and $h(x,\xi) = \mathbb{1}_{(0,1]}(\xi)$, whereby $J^\star = 0$. As $\mathbb{P}$ concentrates unit mass at 0, we have $\widehat{\mathbb{P}}_N = \delta_0 = \mathbb{P}$ irrespective of $N \in \mathbb{N}$. For any $\varepsilon > 0$, the Dirac distribution δ_ε thus resides within the Wasserstein ball $\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$. Hence, $\widehat{J}_N$ fails to converge to $J^\star$ for $\varepsilon \to 0$ because

$$\widehat{J}_N \geq \mathbb{E}^{\delta_\varepsilon}[h(x,\xi)] = h(x,\varepsilon) = 1, \quad \forall \varepsilon > 0.$$

(2) *Linear growth of $\xi \mapsto h(x,\xi)$ in Theorem 3.6 (i):*
Set $\Xi = \mathbb{R}$, $\mathbb{P} = \delta_0$ and $h(x,\xi) = \xi^2$, which implies that $J^\star = 0$. Note that for any $\rho > \varepsilon$, the two-point distribution $\mathbb{Q}_\rho = (1 - \frac{\varepsilon}{\rho})\delta_0 + \frac{\varepsilon}{\rho}\delta_\rho$ is contained in the Wasserstein ball $\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ of radius $\varepsilon > 0$. Hence, $\widehat{J}_N$ fails to converge to $J^\star$ for $\varepsilon \to 0$ because

$$\widehat{J}_N \geq \sup_{\rho>\varepsilon} \mathbb{E}^{\mathbb{Q}_\rho}[h(x,\xi)] = \sup_{\rho>\varepsilon} \varepsilon\rho = \infty, \quad \forall \varepsilon > 0.$$

(3) *Lower semicontinuity of $x \mapsto h(x,\xi)$ in Theorem 3.6 (ii):*
Set $\mathbb{X} = [0,1]$ and $h(x,\xi) = \mathbb{1}_{[0.5,1]}(x)$, whereby $J^\star = 0$ irrespective of $\mathbb{P}$. As the objective is independent of ξ, the distributionally robust optimization problem (5) is equivalent to (1). Then, $\widehat{x}_N = \frac{N-1}{2N}$ is a sequence of minimizers for (5) whose accumulation point $x^\star = \frac{1}{2}$ fails to be optimal in (1).

A convergence result akin to Theorem 3.6 for goodness-of-fit-based ambiguity sets is discussed in [7, Section 4]. This result is complementary to Theorem 3.6. Indeed, Theorem 3.6(i) requires $h(x,\xi)$ to be upper semicontinuous in ξ, which is a necessary condition in our setting (see Example 1) that is absent in [7]. Moreover, Theorem 3.6(ii) only requires $h(x,\xi)$ to be lower semicontinuous in x, while [7] asks for equicontinuity of this mapping. This stronger requirement provides a stronger result, that is, the almost sure convergence of $\sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}[h(x,\xi)]$ to $\mathbb{E}^{\mathbb{P}}[h(x,\xi)]$ uniformly in x on any compact subset of $\mathbb{X}$.

Theorems 3.5 and 3.6 indicate that a careful a priori design of the Wasserstein ball results in attractive finite sample and asymptotic guarantees for the distributionally robust solutions. In practice, however, setting the Wasserstein radius to $\varepsilon_N(\beta)$ yields over-conservative solutions for the following reasons:

- Even though the constants c_1 and c_2 in (8) can be computed based on the proof of [21, Theorem 2], the resulting Wasserstein ball is larger than necessary, *i.e.*, $\mathbb{P} \notin \mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)$ with probability $\ll \beta$.
- Even if $\mathbb{P} \notin \mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)$, the optimal value $\widehat{J}_N$ of (5) may still provide an upper bound on $J^\star$.
- The formula for $\varepsilon_N(\beta)$ in (8) is independent of the training data. Allowing for random Wasserstein radii, however, results in a more efficient use of the available training data.

While Theorems 3.5 and 3.6 provide strong theoretical justification for using Wasserstein ambiguity sets, in practice, it is prudent to calibrate the Wasserstein radius via bootstrapping or cross-validation instead of using the conservative a priori bound $\varepsilon_N(\beta)$; see Section 7.2 for further details. A similar approach has been advocated in [7] to determine the sizes of ambiguity sets that are constructed via goodness-of-fit tests.

So far we have seen that the Wasserstein metric allows us to construct ambiguity sets with favorable asymptotic and finite sample guarantees. In the remainder of the paper we will further demonstrate that the distributionally robust optimization problem (5) with a Wasserstein ambiguity set (6) is not significantly harder to solve than the corresponding SAA problem (4).

4. Solving Worst-Case Expectation Problems

We now demonstrate that the inner worst-case expectation problem in (5) over the Wasserstein ambiguity set (6) can be reformulated as a finite convex program for many loss functions $h(x, \xi)$ of practical interest. For ease of notation, throughout this section we suppress the dependence on the decision variable x. Thus, we examine a generic worst-case expectation problem

$$\sup_{\mathbb{Q} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\big[\ell(\xi)\big] \tag{10}$$

involving a decision-*in*dependent loss function $\ell(\xi) \coloneqq \max_{k \leq K} \ell_k(\xi)$, which is defined as the pointwise maximum of more elementary measurable functions $\ell_k : \mathbb{R}^m \to \overline{\mathbb{R}}$, $k \leq K$. The focus on loss functions representable as pointwise maxima is non-restrictive unless we impose some structure on the functions ℓ_k. Many tractability results in the remainder of this paper are predicated on the following convexity assumption.

Assumption 4.1 (Convexity)**.** *The uncertainty set $\Xi \subseteq \mathbb{R}^m$ is convex and closed, and the negative constituent functions $-\ell_k$ are proper, convex, and lower semicontinuous for all $k \leq K$. Moreover, we assume that ℓ_k is not identically $-\infty$ on Ξ for all $\leq K$.*

Assumption 4.1 essentially stipulates that $\ell(\xi)$ can be written as a maximum of concave functions. As we will showcase in Section 5, this mild restriction does not sacrifice much modeling power. Moreover, generalizations of this setting will be discussed in Section 6. We proceed as follows. Subsection 4.1 addresses the reduction of (10) to a finite convex program, while Subsection 4.2 describes a technique for constructing worst-case distributions.

4.1. Reduction to a Finite Convex Program

The worst-case expectation problem (10) constitutes an infinite-dimensional optimization problem over probability distributions and thus appears to be intractable. However, we will now demonstrate that (10) can be re-expressed as a finite-dimensional convex program by leveraging tools from robust optimization.

Theorem 4.2 (Convex reduction)**.** *If the convexity Assumption 4.1 holds, then for any $\varepsilon \geq 0$ the worst-case expectation (10) equals the optimal value of the finite convex program*

$$\left\{\begin{array}{lll} \inf\limits_{\lambda, s_i, z_{ik}, \nu_{ik}} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N} s_i & \\ \text{s.t.} & [-\ell_k]^*(z_{ik} - \nu_{ik}) + \sigma_\Xi(\nu_{ik}) - \left\langle z_{ik}, \widehat{\xi}_i \right\rangle \leq s_i & \forall i \leq N, \quad \forall k \leq K \\ & \|z_{ik}\|_* \leq \lambda & \forall i \leq N, \quad \forall k \leq K. \end{array}\right. \tag{11}$$

Recall that $[-\ell_k]^*(z_{ik} - \nu_{ik})$ denotes the conjugate of $-\ell_k$ evaluated at $z_{ik} - \nu_{ik}$ and $\|z_{ik}\|_*$ the dual norm of z_{ik}. Moreover, χ_Ξ represents the characteristic function of Ξ and σ_Ξ its conjugate, that is, the support function of Ξ.

Proof of Theorem 4.2. By using Definition 3.1 we can re-express the worst-case expectation (10) as

$$\sup_{\mathbb{Q} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\big[\ell(\xi)\big] = \left\{\begin{array}{ll} \sup\limits_{\Pi, \mathbb{Q}} & \int_\Xi \ell(\xi)\, \mathbb{Q}(\mathrm{d}\xi) \\ \text{s.t.} & \int_{\Xi^2} \|\xi - \xi'\|\, \Pi(\mathrm{d}\xi, \mathrm{d}\xi') \leq \varepsilon \\ & \left\{\begin{array}{l} \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \\ \text{with marginals } \mathbb{Q} \text{ and } \widehat{\mathbb{P}}_N, \text{ respectively} \end{array}\right. \end{array}\right.$$

$$= \left\{\begin{array}{ll} \sup\limits_{\mathbb{Q}_i \in \mathcal{M}(\Xi)} & \frac{1}{N}\sum\limits_{i=1}^{N} \int_\Xi \ell(\xi)\, \mathbb{Q}_i(\mathrm{d}\xi) \\ \text{s.t.} & \frac{1}{N}\sum\limits_{i=1}^{N} \int_\Xi \|\xi - \widehat{\xi}_i\|\, \mathbb{Q}_i(\mathrm{d}\xi) \leq \varepsilon. \end{array}\right.$$

The second equality follows from the law of total probability, which asserts that any joint probability distribution Π of ξ and ξ' can be constructed from the marginal distribution $\widehat{\mathbb{P}}_N$ of ξ' and the conditional distributions $\mathbb{Q}_i$ of ξ given $\xi' = \widehat{\xi}_i$, $i \leq N$, that is, we may write $\Pi = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i} \otimes \mathbb{Q}_i$. The resulting optimization problem represents a generalized moment problem in the distributions $\mathbb{Q}_i$, $i \leq N$. Using a standard duality argument, we obtain

$$\begin{aligned} \sup_{\mathbb{Q} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\big[\ell(\xi)\big] &= \sup_{\mathbb{Q}_i \in \mathcal{M}(\Xi)} \inf_{\lambda \geq 0} \frac{1}{N}\sum_{i=1}^{N} \int_\Xi \ell(\xi)\, \mathbb{Q}_i(\mathrm{d}\xi) + \lambda\Big(\varepsilon - \frac{1}{N}\sum_{i=1}^{N} \int_\Xi \|\xi - \widehat{\xi}_i\|\, \mathbb{Q}_i(\mathrm{d}\xi)\Big) \\ &\leq \inf_{\lambda \geq 0} \sup_{\mathbb{Q}_i \in \mathcal{M}(\Xi)} \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \int_\Xi \Big(\ell(\xi) - \lambda\|\xi - \widehat{\xi}_i\|\Big)\, \mathbb{Q}_i(\mathrm{d}\xi) \qquad (12a) \\ &= \inf_{\lambda \geq 0} \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \sup_{\xi \in \Xi} \Big(\ell(\xi) - \lambda\|\xi - \widehat{\xi}_i\|\Big), \qquad (12b) \end{aligned}$$

where (12a) follows from the max-min inequality, and (12b) follows from the fact that $\mathcal{M}(\Xi)$ contains all the Dirac distributions supported on Ξ. Introducing epigraphical auxiliary variables s_i, $i \leq N$, allows us to reformulate (12b) as

$$\left\{\begin{array}{ll} \inf\limits_{\lambda, s_i} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N} s_i \\ \text{s.t.} & \sup\limits_{\xi \in \Xi} \Big(\ell(\xi) - \lambda\|\xi - \widehat{\xi}_i\|\Big) \leq s_i \quad \forall i \leq N \\ & \lambda \geq 0 \end{array}\right. \tag{12c}$$

$$= \left\{\begin{array}{ll} \inf\limits_{\lambda, s_i} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N} s_i \\ \text{s.t.} & \sup\limits_{\xi \in \Xi} \Big(\ell_k(\xi) - \max\limits_{\|z_{ik}\|_* \leq \lambda} \left\langle z_{ik}, \xi - \widehat{\xi}_i \right\rangle\Big) \leq s_i \quad \forall i \leq N, \quad \forall k \leq K \\ & \lambda \geq 0 \end{array}\right. \tag{12d}$$

$$\leq \left\{ \begin{array}{ll} \inf\limits_{\lambda, s_i} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N} s_i \\ \text{s.t.} & \min\limits_{\|z_{ik}\|_* \leq \lambda} \sup\limits_{\xi\in\Xi} \Big(\ell_k(\xi) - \big\langle z_{ik}, \xi - \widehat{\xi}_i \big\rangle\Big) \leq s_i \quad \forall i \leq N, \quad \forall k \leq K \\ & \lambda \geq 0. \end{array} \right. \tag{12e}$$

Equality (12d) exploits the definition of the dual norm and the decomposability of $\ell(\xi)$ into its constituents $\ell_k(\xi)$, $k \leq K$. Interchanging the maximization over z_{ik} with the minus sign (thereby converting the maximization to a minimization) and then with the maximization over ξ leads to a restriction of the feasible set of (12d). The resulting upper bound (12e) can be re-expressed as

$$\begin{array}{l} \left\{ \begin{array}{lll} \inf\limits_{\lambda, s_i, z_{ik}} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N} s_i & \\ \text{s.t.} & \sup\limits_{\xi\in\Xi} \Big(\ell_k(\xi) - \big\langle z_{ik}, \xi \big\rangle\Big) + \big\langle z_{ik}, \widehat{\xi}_i \big\rangle \leq s_i \quad \forall i \leq N, \quad \forall k \leq K \\ & \|z_{ik}\|_* \leq \lambda & \forall i \leq N, \quad \forall k \leq K \end{array} \right. \\ = \left\{ \begin{array}{lll} \inf\limits_{\lambda, s_i, z_{ik}} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N} s_i & \\ \text{s.t.} & [-\ell_k + \chi_\Xi]^*(z_{ik}) - \big\langle z_{ik}, \widehat{\xi}_i \big\rangle \leq s_i \quad \forall i \leq N, \quad \forall k \leq K \\ & \|z_{ik}\|_* \leq \lambda & \forall i \leq N, \quad \forall k \leq K, \end{array} \right. \end{array} \tag{12f}$$

where (12f) follows from the definition of conjugacy, our conventions of extended arithmetic, and the substitution of z_{ik} with $-z_{ik}$. Note that (12f) is already a finite convex program.

Next, we show that Assumption 4.1 reduces the inequalities (12a) and (12e) to equalities. Under Assumption 4.1, the inequality (12a) is in fact an equality for any $\varepsilon > 0$ by virtue of an extended version of a well-known strong duality result for moment problems [44, Proposition 3.4]. One can show that (12a) continues to hold as an equality even for $\varepsilon = 0$, in which case the Wasserstein ambiguity set (6) reduces to the singleton $\{\widehat{\mathbb{P}}_N\}$, while (10) reduces to the sample average $\frac{1}{N}\sum_{i=1}^{N} \ell(\widehat{\xi}_i)$. Indeed, for $\varepsilon = 0$ the variable λ in (12b) can be increased indefinitely at no penalty. As $\ell(\xi)$ constitutes a pointwise maximum of upper semicontinuous concave functions, an elementary but tedious argument shows that (12b) converges to the sample average $\frac{1}{N}\sum_{i=1}^{N} \ell(\widehat{\xi}_i)$ as λ tends to infinity.

The inequality (12e) also reduces to an equality under Assumption 4.1 thanks to the classical minimax theorem [4, Proposition 5.5.4], which applies because the set $\{z_{ik} \in \mathbb{R}^m : \|z_{ik}\|_* \leq \lambda\}$ is compact for any finite $\lambda \geq 0$. Thus, the optimal values of (10) and (12f) coincide.

Assumption 4.1 further implies that the function $-\ell_k + \chi_\Xi$ is proper, convex and lower semicontinuous. Properness holds because ℓ_k is not identically $-\infty$ on Ξ. By [42, Theorem 11.23(a), p. 493], its conjugate essentially coincides with the *epi-addition* (also known as *inf-convolution*) of the conjugates of the functions $-\ell_k$ and σ_Ξ. Thus,

$$\begin{aligned} [-\ell_k + \chi_\Xi]^*(z_{ik}) &= \inf_{\nu_{ik}} \Big([-\ell_k]^*(z_{ik} - \nu_{ik}) + [\chi_\Xi]^*(\nu_{ik}) \Big) \\ &= \mathrm{cl}\Big[\inf_{\nu_{ik}} \Big([-\ell_k]^*(z_{ik} - \nu_{ik}) + \sigma_\Xi(\nu_{ik}) \Big) \Big], \end{aligned}$$

where $\mathrm{cl}[\,\cdot\,]$ denotes the closure operator that maps any function to its largest lower semicontinuous minorant. As $\mathrm{cl}[f(\xi)] \leq 0$ if and only if $f(\xi) \leq 0$ for any function f, we may conclude that (12f) is indeed equivalent to (11) under Assumption 4.1. □

Note that the semi-infinite inequality in (12c) generalizes the nonlinear uncertain constraints studied in [1] because it involves an additional norm term and as the loss function $\ell(\xi)$ is not necessarily concave under

Assumption 4.1. As in [1], however, the semi-infinite constraint admits a robust counterpart that involves the conjugate of the loss function and the support function of the uncertainty set.

From the proof of Theorem 4.2 it is immediately clear that the worst-case expectation (10) is conservatively approximated by the optimal value of the finite convex program (12f) even if Assumption 4.1 fails to hold. In this case the sum $-\ell_k + \chi_\Xi$ in (12f) must be evaluated under our conventions of extended arithmetics, whereby $\infty - \infty = \infty$. These observations are formalized in the following corollary.

Corollary 4.3 (Approximate convex reduction)**.** *For any $\varepsilon \geq 0$, the worst-case expectation (10) is smaller or equal to the optimal value of the finite convex program (12f).*

4.2. Extremal Distributions

Stress test experiments are instrumental to assess the quality of candidate decisions in stochastic optimization. Meaningful stress tests require a good understanding of the extremal distributions from within the Wasserstein ball that achieve the worst-case expectation (10) for various loss functions. We now show that such extremal distributions can be constructed systematically from the solution of a convex program akin to (11).

Theorem 4.4 (Worst-case distributions)**.** *If Assumption 4.1 holds, then the worst-case expectation (10) coincides with the optimal value of the finite convex program*

$$\left\{\begin{array}{lll} \sup\limits_{\alpha_{ik}, q_{ik}} & \frac{1}{N}\sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K}\alpha_{ik}\ell_k\big(\widehat{\xi}_i - \frac{q_{ik}}{\alpha_{ik}}\big) & \\ \text{s.t.} & \frac{1}{N}\sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K}\|q_{ik}\| \leq \varepsilon & \\ & \sum\limits_{k=1}^{K}\alpha_{ik} = 1 & \forall i \leq N \\ & \alpha_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K \\ & \widehat{\xi}_i - \frac{q_{ik}}{\alpha_{ik}} \in \Xi & \forall i \leq N, \quad \forall k \leq K \end{array}\right. \tag{13}$$

irrespective of $\varepsilon \geq 0$. Let $\big\{\alpha_{ik}(r), q_{ik}(r)\big\}_{r\in\mathbb{N}}$ be a sequence of feasible decisions whose objective values converge to the supremum of (13). Then, the discrete probability distributions

$$\mathbb{Q}_r \coloneqq \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_{ik}(r)\delta_{\xi_{ik}(r)} \qquad \text{with} \qquad \xi_{ik}(r) \coloneqq \widehat{\xi}_i - \frac{q_{ik}(r)}{\alpha_{ik}(r)}$$

belong to the Wasserstein ball $\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ and attain the supremum of (10) asymptotically, i.e.,

$$\sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\big[\ell(\xi)\big] = \lim_{r\to\infty}\mathbb{E}^{\mathbb{Q}_r}\big[\ell(\xi)\big] = \lim_{k\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_{ik}(r)\ell\big(\xi_{ik}(r)\big).$$

We highlight that all fractions in (13) must again be evaluated under our conventions of extended arithmetics. Specifically, if $\alpha_{ik} = 0$ and $q_{ik} \neq 0$, then q_{ik}/α_{ik} has at least one component equal to $+\infty$ or $-\infty$, which implies that $\widehat{\xi}_i - q_{ik}/\alpha_{ik} \notin \Xi$. In contrast, if $\alpha_{ik} = 0$ and $q_{ik} = 0$, then $\widehat{\xi}_i - q_{ik}/\alpha_{ik} = \widehat{\xi}_i \in \Xi$. Moreover, the ik-th component in the objective function of (13) evaluates to 0 whenever $\alpha_{ik} = 0$ regardless of q_{ik}.

The proof of Theorem 4.4 is based on the following technical lemma.

Lemma 4.5. *Define $F : \mathbb{R}^m \times \mathbb{R}_+ \to \overline{\mathbb{R}}$ through $F(q,\alpha) = \inf_{z\in\mathbb{R}^m}\big\langle z, q - \alpha\widehat{\xi}\big\rangle + \alpha f^*(z)$ for some proper, convex, and lower semicontinuous function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ and reference point $\widehat{\xi} \in \mathbb{R}^m$. Then, F coincides with*

the (extended) perspective function of the mapping $q \mapsto -f(\widehat{\xi} - q)$, that is,

$$F(q,\alpha) = \begin{cases} -\alpha f\big(\widehat{\xi} - q/\alpha\big) & \text{if } \alpha > 0, \\ -\chi_{\{0\}}(q) & \text{if } \alpha = 0. \end{cases}$$

Proof. By construction, we have $F(q,0) = \inf_{z\in\mathbb{R}^m} \langle z, q\rangle = -\chi_{\{0\}}(q)$. For $\alpha > 0$, on the other hand, the definition of conjugacy implies that

$$F(q,\alpha) = -[\alpha f^*]^*(\alpha\widehat{\xi} - q) = -\alpha[f^*]^*\big(\widehat{\xi} - q/\alpha\big).$$

The claim then follows because $[f^*]^* = f$ for any proper, convex, and lower semicontinuous function f [4, Proposition 1.6.1(c)]. Additional information on perspective functions can be found in [12, Section 2.2.3, p. 39]. □

Proof of Theorem 4.4. By Theorem 4.2, which applies under Assumption 4.1, the worst-case expectation (10) coincides with the optimal value of the convex program (11). From the proof of Theorem 4.2 we know that (11) is equivalent to (12f). The Lagrangian dual of (12f) is given by

$$\left\{\begin{array}{lll} \sup\limits_{\beta_{ik},\alpha_{ik}} & \inf\limits_{\lambda,s_i,z_{ik}} \lambda\varepsilon + \sum\limits_{i=1}^N \Big[\frac{s_i}{N} + \sum\limits_{k=1}^K \big[\beta_{ik}\big(\|z_{ik}\|_* - \lambda\big) + \alpha_{ik}\big([-\ell_k + \chi_\Xi]^*(z_{ik}) - \big\langle z_{ik}, \widehat{\xi}_i\big\rangle - s_i\big)\big]\Big] & \\ \text{s.t.} & \alpha_{ik} \geq 0 \qquad\qquad \forall i \leq N, \quad \forall k \leq K & \\ & \beta_{ik} \geq 0 \qquad\qquad \forall i \leq N, \quad \forall k \leq K, & \end{array}\right.$$

where the products of dual variables and constraint functions in the objective are evaluated under the standard convention $0 \cdot \infty = 0$. Strong duality holds since the function $[-\ell_k + \chi_\Xi]^*$ is proper, convex, and lower semicontinuous under Assumption 4.1 and because this function appears in a constraint of (12f) whose right-hand side is a free decision variable. By explicitly carrying out the minimization over λ and s_i, one can show that the above dual problem is equivalent to

$$\left\{\begin{array}{lll} \sup\limits_{\beta_{ik},\alpha_{ik}} & \inf\limits_{z_{ik}} \sum\limits_{i=1}^N \sum\limits_{k=1}^K \beta_{ik}\|z_{ik}\|_* + \alpha_{ik}[-\ell_k + \chi_\Xi]^*(z_{ik}) - \alpha_{ik}\big\langle z_{ik}, \widehat{\xi}_i\big\rangle & \\ \text{s.t.} & \sum\limits_{i=1}^N \sum\limits_{k=1}^K \beta_{ik} = \varepsilon & \\ & \sum\limits_{k=1}^K \alpha_{ik} = \frac{1}{N} \qquad\qquad \forall i \leq N & \\ & \alpha_{ik} \geq 0 \qquad\qquad \forall i \leq N, \quad \forall k \leq K & \\ & \beta_{ik} \geq 0 \qquad\qquad \forall i \leq N, \quad \forall k \leq K. & \end{array}\right. \tag{14a}$$

By using the definition of the dual norm, (14a) can be re-expressed as

$$\left\{\begin{array}{lll} \sup\limits_{\beta_{ik},\alpha_{ik}} & \inf\limits_{z_{ik}} \sum\limits_{i=1}^N \sum\limits_{k=1}^K \max\limits_{\|q_{ik}\|\leq\beta_{ik}} \big\langle z_{ik}, q_{ik}\big\rangle + \alpha_{ik}[-\ell_k + \chi_\Xi]^*(z_{ik}) - \alpha_{ik}\big\langle z_{ik}, \widehat{\xi}_i\big\rangle\Big] & \\ \text{s.t.} & \sum\limits_{i=1}^N \sum\limits_{k=1}^K \beta_{ik} = \varepsilon & \\ & \sum\limits_{k=1}^K \alpha_{ik} = \frac{1}{N} \qquad\qquad \forall i \leq N & \\ & \alpha_{ik} \geq 0 \qquad\qquad \forall i \leq N, \quad \forall k \leq K & \\ & \beta_{ik} \geq 0 \qquad\qquad \forall i \leq N, \quad \forall k \leq K & \end{array}\right. \tag{14b}$$

$$= \left\{ \begin{array}{lll} \sup\limits_{\beta_{ik},\alpha_{ik}} & \max\limits_{\|q_{ik}\|\leq\beta_{ik}} \inf\limits_{z_{ik}} \sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K} \big\langle z_{ik}, q_{ik}\big\rangle + \alpha_{ik}[-\ell_k + \chi_\Xi]^*(z_{ik}) - \alpha_{ik}\big\langle z_{ik}, \widehat{\xi}_i\big\rangle & \\ \text{s.t.} & \sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K} \beta_{ik} = \varepsilon & \\ & \sum\limits_{k=1}^{K} \alpha_{ik} = \frac{1}{N} & \forall i \leq N \\ & \alpha_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K \\ & \beta_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K, \end{array} \right. \tag{14c}$$

where (14c) follows from the classical minimax theorem and the fact that the q_{ik} variables range over a non-empty and compact feasible set for any finite ε; see [4, Proposition 5.5.4]. Eliminating the β_{ik} variables and using Lemma 4.5 allows us to reformulate (14c) as

$$\left\{ \begin{array}{lll} \sup\limits_{\alpha_{ik},q_{ik}} & \inf\limits_{z_{ik}} \sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K} \big\langle z_{ik}, q_{ik} - \alpha_{ik}\widehat{\xi}_i\big\rangle + \alpha_{ik}[-\ell_k + \chi_\Xi]^*(z_{ik}) & \\ \text{s.t.} & \sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K} \|q_{ik}\| \leq \varepsilon & \\ & \sum\limits_{k=1}^{K} \alpha_{ik} = \frac{1}{N} & \forall i \leq N \\ & \alpha_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K \end{array} \right. \tag{14d}$$

$$= \left\{ \begin{array}{lll} \sup\limits_{\alpha_{ik},q_{ik}} & \sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K} -\alpha_{ik}\Big(- \ell_k\big(\widehat{\xi}_i - \frac{q_{ik}}{\alpha_{ik}}\big) + \chi_\Xi\big(\widehat{\xi}_i - \frac{q_{ik}}{\alpha_{ik}}\big)\Big)\mathbb{1}_{\{\alpha_{ik}>0\}} - \chi_{\{0\}}(q_{ik})\mathbb{1}_{\{\alpha_{ik}=0\}} & \\ \text{s.t.} & \sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K} \|q_{ik}\| \leq \varepsilon & \\ & \sum\limits_{k=1}^{K} \alpha_{ik} = \frac{1}{N} & \forall i \leq N \\ & \alpha_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K. \end{array} \right. \tag{14e}$$

Our conventions of extended arithmetics imply that the ik-th term in the objective function of problem (14e) simplifies to

$$\alpha_{ik}\ell_k\big(\widehat{\xi}_i - \frac{q_{ik}}{\alpha_{ik}}\big) - \chi_\Xi\big(\widehat{\xi}_i - \frac{q_{ik}}{\alpha_{ik}}\big). \tag{14f}$$

Indeed, for $\alpha_{ik} > 0$, this identity trivially holds. For $\alpha_{ik} = 0$, on the other hand, the ik-th objective term in (14e) reduces to $-\chi_{\{0\}}(q_{ik})$. Moreover, the first term in (14f) vanishes whenever $\alpha_{ik} = 0$ regardless of q_{ik}, and the second term in (14f) evaluates to 0 if $q_{ik} = 0$ (as $0/0 = 0$ and $\widehat{\xi}_i \in \Xi$) and to $-\infty$ if $q_{ik} \neq 0$ (as $q_{ik}/0$ has at least one infinite component, implying that $\widehat{\xi}_i + q_{ik}/0 \notin \Xi$). Therefore, (14f) also reduces to $-\chi_{\{0\}}(q_{ik})$ when $\alpha_{ik} = 0$. This proves that the ik-th objective term in (14e) coincides with (14f). Substituting (14f) into (14e) and re-expressing $-\chi_\Xi\big(\widehat{\xi}_i - \frac{q_{ik}}{\alpha_{ik}}\big)$ in terms of an explicit hard constraint yields

$$\left\{ \begin{array}{lll} \sup\limits_{\alpha_{ik},q_{ik}} & \sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K} \alpha_{ik}\ell_k\big(\widehat{\xi}_i - \frac{q_{ik}}{\alpha_{ik}}\big) & \\ \text{s.t.} & \sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K} \|q_{ik}\| \leq \varepsilon & \\ & \sum\limits_{k=1}^{K} \alpha_{ik} = \frac{1}{N} & \forall i \leq N \\ & \alpha_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K \\ & \widehat{\xi}_i - \frac{q_{ik}}{\alpha_{ik}} \in \Xi & \forall i \leq N, \quad \forall k \leq K. \end{array} \right. \tag{14g}$$

Finally, replacing $\big\{\alpha_{ik}, q_{ik}\big\}$ with $\frac{1}{N}\big\{\alpha_{ik}, q_{ik}\big\}$ shows that (14g) is equivalent to (13). This completes the first part of the proof.

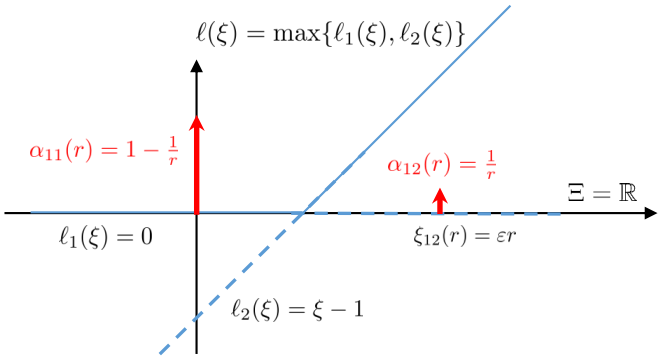


FIGURE 1. Example of a worst-case expectation problem without a worst-case distribution

As for the second claim, let $\{\alpha_{ik}(r), q_{ik}(r)\}_{r\in\mathbb{N}}$ be a sequence of feasible solutions that attains the supremum in (13), and set $\xi_{ik}(r) \coloneqq \widehat{\xi}_i - \frac{q_{ik}(r)}{\alpha_{ik}(r)} \in \Xi$. Then, the discrete distribution

$$\Pi_r \coloneqq \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_{ik}(r)\delta_{\left(\xi_{ik}(r),\widehat{\xi}_i\right)}$$

has the distribution $\mathbb{Q}_r$ defined in the theorem statement and the empirical distribution $\widehat{\mathbb{P}}_N$ as marginals. By the definition of the Wasserstein metric, Π_r represents a feasible mass transportation plan that provides an upper bound on the distance between $\widehat{\mathbb{P}}_N$ and $\mathbb{Q}_r$; see Definition 3.1. Thus, we have

$$d_{\mathrm{W}}\big(\mathbb{Q}_r, \widehat{\mathbb{P}}_N\big) \leq \int_{\Xi^2} \|\xi - \xi'\|\, \Pi_r(\mathrm{d}\xi, \mathrm{d}\xi') = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_{ik}(r)\big\|\xi_{ik}(r) - \widehat{\xi}_i\big\| = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \big\|q_{ik}(r)\big\| \leq \varepsilon,$$

where the last inequality follows readily from the feasibility of $q_{ik}(r)$ in (13). We conclude that

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\big[\ell(\xi)\big] &\geq \limsup_{k\to\infty} \mathbb{E}^{\mathbb{Q}_r}\big[\ell(\xi)\big] = \limsup_{k\to\infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_{ik}(r)\ell\big(\xi_{ik}(r)\big) \\ &\geq \limsup_{k\to\infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_{ik}(r)\ell_k\big(\xi_{ik}(r)\big) = \sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\big[\ell(\xi)\big],\end{aligned}$$

where the first inequality holds as $\mathbb{Q}_r \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ for all $k \in \mathbb{N}$, and the second inequality uses the trivial estimate $\ell \geq \ell_k$ for all $k \leq K$. The last equality follows from the construction of $\alpha_{ik}(r)$ and $\xi_{ik}(r)$ and the fact that (13) coincides with the worst-case expectation (10). □

In the rest of this section we discuss some notable properties of the convex program (13).

In the *ambiguity-free* limit, that is, when the radius of the Wasserstein ball is set to zero, then the optimal value of the convex program (13) reduces to the expected loss under the empirical distribution. Indeed, for $\varepsilon = 0$ all q_{ik} variables are forced to zero, and α_{ik} enters the objective only through $\sum_{k=1}^{K} \alpha_{ik} = \frac{1}{N}$. Thus, the objective function of (13) simplifies to $\mathbb{E}^{\widehat{\mathbb{P}}_N}[\ell(\xi)]$.

We further emphasize that it is not possible to guarantee the existence of a worst-case distribution that attains the supremum in (10). In general, as shown in Theorem 4.4, we can only construct a sequence of distributions that attains the supremum asymptotically. The following example discusses an instance of (10) that admits no worst-case distribution.

Example 2 (Non-existence of a worst-case distribution). Assume that $\Xi = \mathbb{R}$, $N = 1$, $\widehat{\xi}_1 = 0$, $K = 2$, $\ell_1(\xi) = 0$ and $\ell_2(\xi) = \xi - 1$. In this case we have $\widehat{\mathbb{P}}_N = \delta_{\{0\}}$, and problem (13) reduces to

$$\sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\delta_0)} \mathbb{E}^{\mathbb{Q}}\big[\ell(\xi)\big] = \left\{ \begin{array}{ll} \sup\limits_{\alpha_{1j},q_{1j}} & -q_{12} - \alpha_{12} \\ \text{s.t.} & |q_{11}| + |q_{12}| \le \varepsilon \\ & \alpha_{11} + \alpha_{12} = 1 \\ & \alpha_{11} \ge 0, \quad \alpha_{12} \ge 0. \end{array} \right.$$

The supremum on the right-hand side amounts to ε and is attained, for instance, by the sequence $\alpha_{11}(r) = 1 - \frac{1}{k}$, $\alpha_{12}(r) = \frac{1}{k}$, $q_{11}(r) = 0$, $q_{12}(r) = -\varepsilon$ for $k \in \mathbb{N}$. Define

$$\mathbb{Q}_r = \alpha_{11}(r)\, \delta_{\xi_{11}(r)} + \alpha_{12}(r)\, \delta_{\xi_{12}(r)},$$

with $\xi_{11}(r) = \widehat{\xi}_1 - \frac{q_{11}(r)}{\alpha_{11}(r)} = 0$, and $\xi_{12}(r) = \widehat{\xi}_1 - \frac{q_{12}(r)}{\alpha_{12}(r)} = \varepsilon k$. By Theorem 4.4, the two-point distributions $\mathbb{Q}_r$ reside within the Wasserstein ball of radius ε around δ_0 and asymptotically attain the supremum in the worst-case expectation problem. However, this sequence has no weak limit as $\xi_{12}(r) = \varepsilon k$ tends to infinity, see Figure 1. In fact, no single distribution can attain the worst-case expectation. Assume for the sake of contradiction that there exists $\mathbb{Q}^\star \in \mathbb{B}_\varepsilon(\delta_0)$ with $\mathbb{E}^{\mathbb{Q}^\star}[\ell(\xi)] = \varepsilon$. Then, we find $\varepsilon = \mathbb{E}^{\mathbb{Q}^\star}[\ell(\xi)] < \mathbb{E}^{\mathbb{Q}^\star}[|\xi|] \le \varepsilon$, where the strict inequality follows from the relation $\ell(\xi) < |\xi|$ for all $\xi \neq 0$ and the observation that $\mathbb{Q}^\star \neq \delta_0$, while the second inequality follows from Theorem 3.2. Thus, $\mathbb{Q}^\star$ does not exist.

The existence of a worst-case distribution can, however, be guaranteed in some special cases.

Corollary 4.6 (Existence of a worst-case distribution)**.** *Suppose that Assumption 4.1 holds. If the uncertainty set Ξ is compact or the loss function is concave (i.e., $K = 1$), then the sequence $\{\alpha_{ik}(r), \xi_{ik}(r)\}_{r\in\mathbb{N}}$ constructed in Theorem 4.4 has an accumulation point $\{\alpha^\star_{ik}, \xi^\star_{ik}\}$, and*

$$\mathbb{Q}^\star := \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \alpha^\star_{ik} \delta_{\xi^\star_{ik}}$$

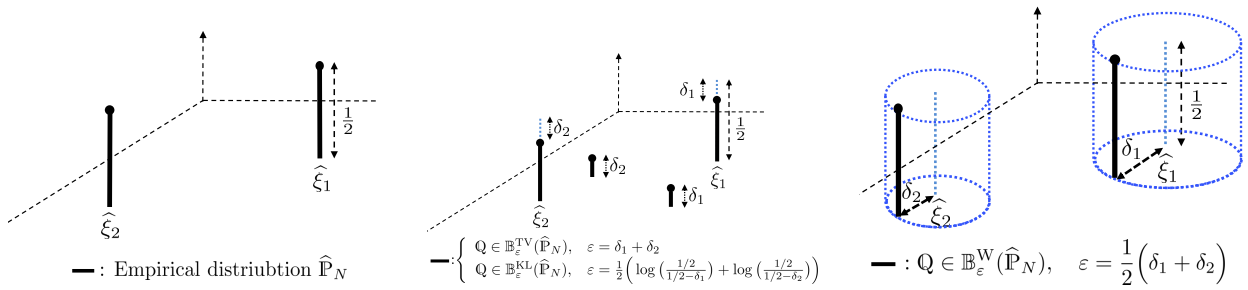
is a worst-case distribution achieving the supremum in (10).

Proof. If Ξ is compact, then the sequence $\{\alpha_{ik}(r), \xi_{ik}(r)\}_{r\in\mathbb{N}}$ has a converging subsequence with limit $\{\alpha^\star_{ik}, \xi^\star_{ik}\}$. Similarly, if $K = 1$, then $\alpha_{i1} = 1$ for all $i \le N$, in which case (13) reduces to a convex optimization problem with an upper semicontinuous objective function over a compact feasible set. Hence, its supremum is attained at a point $\{\alpha^\star_{ik}, \xi^\star_{ik}\}$. In both cases, Theorem 4.4 guarantees that the distribution $\mathbb{Q}^\star$ implied by $\{\alpha^\star_{ik}, \xi^\star_{ik}\}$ achieves the supremum in (10). □

The worst-case distribution of Corollary 4.6 is discrete, and its atoms $\xi^\star_{ik}$ reside in the neighborhood of the given data points $\widehat{\xi}_i$. By the constraints of problem (13), the probability-weighted cumulative distance between the atoms and the respective data points amounts to

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_{ik} \|\xi^\star_{ik} - \widehat{\xi}_i\| = \sum_{i=1}^{N} \sum_{k=1}^{K} \|q_{ik}\| \le \varepsilon,$$

which is bounded above by the radius of the Wasserstein ball. The fact that the worst-case distribution $\mathbb{Q}^\star$ (if it exists) is supported outside of $\widehat{\Xi}_N$ is a key feature distinguishing the Wasserstein ball from the ambiguity sets induced by other probability metrics such as the total variation distance or the Kullback-Leibler divergence; see Figure 2. Thus, the worst-case expectation criterion based on Wasserstein balls advocated in this paper should appeal to decision makers who wish to immunize their optimization problems against perturbations of the data points.



(a) Empirical distribution on a training dataset with $N = 2$ samples

(b) A representative discrete distribution in the total variation or the Kullback-Leiber ball

(c) A representative discrete distribution in the Wasserstein ball

FIGURE 2. Representative distributions in balls centered at $\widehat{\mathbb{P}}_N$ induced by different metrics

Remark 4.7 (Weak coupling)**.** *We highlight that the convex program* (13) *is amenable to decomposition and parallelization techniques as the decision variables associated with different sample points are only coupled through the norm constraint. We expect the resulting scenario decomposition to offer a substantial speedup of the solution times for problems involving large datasets. Efficient decomposition algorithms that could be used for solving the convex program* (13) *are described, for example, in* [35] *and* [5, Chapter 4].

5. Special Loss Functions

We now demonstrate that the convex optimization problems (11) and (13) reduce to computationally tractable conic programs for several loss functions of practical interest.

5.1. Piecewise Affine Loss Functions

We first investigate the worst-case expectations of convex and concave piecewise affine loss functions, which arise, for example, in option pricing [8], risk management [34] and in generic two-stage stochastic programming [6]. Moreover, piecewise affine functions frequently serve as approximations of *smooth* convex or concave loss functions.

Corollary 5.1 (Piecewise affine loss functions)**.** *Suppose that the uncertainty set is a polytope, that is,* $\Xi = \{\xi \in \mathbb{R}^m : C\xi \leq d\}$ *where* C *is a matrix and* d *a vector of appropriate dimensions. Moreover, consider the affine functions* $a_k(\xi) \coloneqq \langle a_k, \xi \rangle + b_k$ *for all* $k \leq K$*.*

(i) If $\ell(\xi) = \max_{k \leq K} a_k(\xi)$*, then the worst-case expectation* (10) *evaluates to*

$$\left\{\begin{array}{lll} \inf\limits_{\lambda, s_i, \gamma_{ik}} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N} s_i & \\ \text{s.t.} & b_k + \langle a_k, \widehat{\xi}_i \rangle + \langle \gamma_{ik}, d - C\widehat{\xi}_i \rangle \leq s_i & \forall i \leq N, \quad \forall k \leq K \\ & \|C^\intercal \gamma_{ik} - a_k\|_* \leq \lambda & \forall i \leq N, \quad \forall k \leq K \\ & \gamma_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K. \end{array}\right. \tag{15a}$$

(ii) If $\ell(\xi) = \min_{k\leq K} a_k(\xi)$, then the worst-case expectation (10) evaluates to

$$\left\{\begin{array}{lll} \inf\limits_{\lambda,s_i,\gamma_i,\theta_i} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N} s_i & \\ \text{s.t.} & \big\langle \theta_i, b + A\widehat{\xi}_i\big\rangle + \big\langle \gamma_i, d - C\widehat{\xi}_i\big\rangle \leq s_i & \forall i \leq N \\ & \|C^\intercal\gamma_i - A^\intercal\theta_i\|_* \leq \lambda & \forall i \leq N \\ & \big\langle \theta_i, e\big\rangle = 1 & \forall i \leq N \\ & \gamma_i \geq 0 & \forall i \leq N \\ & \theta_i \geq 0 & \forall i \leq N, \end{array}\right. \tag{15b}$$

where A is the matrix with rows $a_k^\intercal$, $k \leq K$, b is the column vector with entries b_k, $k \leq K$, and e is the vector of all ones.

Proof. Assertion (i) is an immediate consequence of Theorem 4.2, which applies because $\ell(x)$ is the pointwise maximum of the affine functions $\ell_k(\xi) = a_k(\xi)$, $k \leq K$, and thus Assumption 4.1 holds for $J = K$. By definition of the conjugacy operator, we have

$$[-\ell_k]^*(z) = [-a_k]^*(z) = \sup_{\xi} \big\langle z, \xi\big\rangle + \big\langle a_k, \xi\big\rangle + b_k = \left\{\begin{array}{ll} b_k & \text{if } z = -a_k, \\ \infty & \text{else,} \end{array}\right.$$

and

$$\sigma_\Xi(\nu) = \left\{\begin{array}{ll} \sup\limits_{\xi} & \big\langle \nu, \xi\big\rangle \\ \text{s.t.} & C\xi \leq d \end{array}\right. = \left\{\begin{array}{ll} \inf\limits_{\gamma\geq 0} & \big\langle \gamma, d\big\rangle \\ \text{s.t.} & C^\intercal\gamma = \nu, \end{array}\right.$$

where the last equality follows from strong duality, which holds as the uncertainty set is non-empty. Assertion (i) then follows by substituting the above expressions into (11).

Assertion (ii) also follows directly from Theorem 4.2 because $\ell(\xi) = \ell_1(\xi) = \min_{k\leq K} a_j(\xi)$ is concave and thus satisfies Assumption 4.1 for $J = 1$. In this setting, we find

$$[-\ell]^*(z) = \sup_{\xi} \big\langle z, \xi\big\rangle + \min_{k\leq K}\Big\{\big\langle a_k, \xi\big\rangle + b_k\Big\} = \left\{\begin{array}{ll} \sup\limits_{\xi,\tau} & \big\langle z, \xi\big\rangle + \tau \\ \text{s.t.} & A\xi + b \geq \tau e \end{array}\right. = \left\{\begin{array}{ll} \inf\limits_{\theta\geq 0} & \big\langle \theta, b\big\rangle \\ \text{s.t.} & A^\intercal\theta = -z \\ & \big\langle \theta, e\big\rangle = 1 \end{array}\right.$$

where the last equality follows again from strong linear programming duality, which holds since the primal maximization problem is feasible. Assertion (ii) then follows by substituting $[-\ell]^*$ as well as the formula for σ_Ξ from the proof of assertion (i) into (11). □

As a consistency check, we ascertain that in the *ambiguity-free limit*, the optimal value of (15a) reduces to the expectation of $\max_{k\leq K} a_k(\xi)$ under the empirical distribution. Indeed, for $\varepsilon = 0$, the variable λ can be set to any positive value at no penalty. For this reason and because all training samples must belong to the uncertainty set (*i.e.*, $d - C\widehat{\xi}_i \geq 0$ for all $i \leq N$), it is optimal to set $\gamma_{ik} = 0$. This in turn implies that $s_i = \max_{k\leq K} a_k(\widehat{\xi}_i)$ at optimality, in which case $\frac{1}{N}\sum_{i=1}^N s_i$ represents the sample average of the convex loss function at hand.

An analogous argument shows that, for $\varepsilon = 0$, the optimal value of (15b) reduces to the expectation of $\min_{k\leq K} a_k(\xi)$ under the empirical distribution. As before, λ can be increased at no penalty. Thus, we conclude that $\gamma_i = 0$ and

$$s_i = \min_{\theta_i\geq 0}\Big\{\big\langle \theta_i, b + A\widehat{\xi}_i\big\rangle : \big\langle \theta_i, e\big\rangle = 1\Big\} = \min_{k\leq K} a_k(\widehat{\xi}_i)$$

at optimality, in which case $\frac{1}{N}\sum_{i=1}^N s_i$ is the sample average of the given concave loss function.

5.2. Uncertainty Quantification

A problem of great practical interest is to ascertain whether a physical, economic or engineering system with an uncertain state ξ satisfies a number of safety constraints with high probability. In the following we denote by $\mathbb{A}$ the set of states in which the system is safe. Our goal is to quantify the probability of the event $\xi \in \mathbb{A}$ ($\xi \notin \mathbb{A}$) under an ambiguous state distribution that is only indirectly observable through a finite training dataset. More precisely, we aim to calculate the *worst-case* probability of the system being *unsafe*, *i.e.*,

$$\sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{Q}\left[\xi \notin \mathbb{A}\right], \tag{16a}$$

as well as the *best-case* probability of the system being *safe*, that is,

$$\sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{Q}\left[\xi \in \mathbb{A}\right]. \tag{16b}$$

Remark 5.2 (Data-dependent sets)**.** *The set $\mathbb{A}$ may even depend on the samples $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$, in which case $\mathbb{A}$ is renamed as $\widehat{\mathbb{A}}$. If the Wasserstein radius ε is set to $\varepsilon_N(\beta)$, then we have $\mathbb{P} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ with probability $1-\beta$, implying that (16a) and (16b) still provide $1-\beta$ confidence bounds on $\mathbb{P}[\xi \notin \widehat{\mathbb{A}}]$ and $\mathbb{P}[\xi \in \widehat{\mathbb{A}}]$, respectively.*

Corollary 5.3 (Uncertainty quantification)**.** *Suppose that the uncertainty set is a polytope of the form $\Xi = \{\xi \in \mathbb{R}^m : C\xi \leq d\}$ as in Corollary 5.1.*

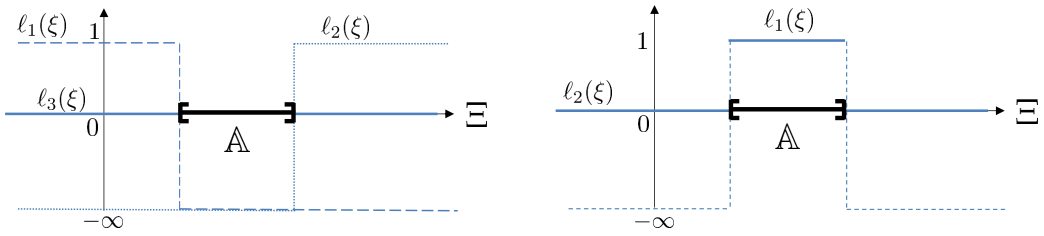
(i) If $\mathbb{A} = \{\xi \in \mathbb{R}^m : A\xi < b\}$ is an open polytope and the halfspace $\{\xi : \langle a_k, \xi\rangle \geq b_k\}$ has a nonempty intersection with Ξ for any $k \leq K$, where a_k is the k-th row of the matrix A and b_k is the k-th entry of the vector b, then the worst-case probability (16a) is given by

$$\left\{\begin{array}{lll} \inf\limits_{\lambda, s_i, \gamma_{ik}, \theta_{ik}} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^N s_i & \\ \text{s.t.} & 1 - \theta_{ik}\big(b_k - \langle a_k, \widehat{\xi}_i\rangle\big) + \langle \gamma_{ik}, d - C\widehat{\xi}_i\rangle \leq s_i & \forall i \leq N, \quad \forall k \leq K \\ & \|a_k\theta_{ik} - C^\intercal\gamma_{ik}\|_* \leq \lambda & \forall i \leq N, \quad \forall k \leq K \\ & \gamma_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K \\ & \theta_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K \\ & s_i \geq 0 & \forall i \leq N. \end{array}\right. \tag{17a}$$

(ii) If $\mathbb{A} = \{\xi \in \mathbb{R}^m : A\xi \leq b\}$ is a closed polytope that has a nonempty intersection with Ξ, then the best-case probability (16b) is given by

$$\left\{\begin{array}{lll} \inf\limits_{\lambda, s_i, \gamma_i, \theta_i} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^N s_i & \\ \text{s.t.} & 1 + \langle \theta_i, b - A\widehat{\xi}_i\rangle + \langle \gamma_i, d - C\widehat{\xi}_i\rangle \leq s_i & \forall i \leq N \\ & \|A^\intercal\theta_i + C^\intercal\gamma_i\|_* \leq \lambda & \forall i \leq N \\ & \gamma_i \geq 0 & \forall i \leq N \\ & \theta_i \geq 0 & \forall i \leq N \\ & s_i \geq 0 & \forall i \leq N. \end{array}\right. \tag{17b}$$

Proof. The uncertainty quantification problems (16a) and (16b) can be interpreted as instances of (10) with loss functions $\ell = 1 - \mathbb{1}_\mathbb{A}$ and $\ell = \mathbb{1}_\mathbb{A}$, respectively. In order to be able to apply Theorem 4.2, we should represent these loss functions as finite maxima of concave functions as shown in Figure 3.



(a) Indicator function of the unsafe set (b) Indicator function of the safe set

FIGURE 3. Representing the indicator function of a convex set and its complement as a pointwise maximum of concave functions

Formally, assertion (i) follows from Theorem 4.2 for a loss function with $K+1$ pieces if we use the following definitions. For every $k \leq K$ we define

$$\ell_k(\xi) = \begin{cases} 1 & \text{if } \langle a_k, \xi \rangle \geq b_k, \\ -\infty & \text{otherwise.} \end{cases}$$

Moreover, we define $\ell_{K+1}(\xi) = 0$. As illustrated in Figure 3(a), we thus have $\ell(\xi) = \max_{k \leq K+1} \ell_k(\xi) = 1 - \mathbb{1}_{\mathbb{A}}(\xi)$ and

$$\sup_{\mathbb{Q} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{Q}[\xi \notin \mathbb{A}] = \sup_{\mathbb{Q} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\ell(\xi)].$$

Assumption 4.1 holds due to the postulated properties of $\mathbb{A}$ and Ξ. In order to apply Theorem 4.2, we must determine the support function σ_Ξ, which is already known from Corollary 5.1, as well as the conjugate functions of $-\ell_k$, $k \leq K+1$. A standard duality argument yields

$$[-\ell_k]^*(z) = \begin{cases} \sup\limits_{\xi} & \langle z, \xi \rangle + 1 \\ \text{s.t.} & \langle a_k, \xi \rangle \geq b_k \end{cases} = \begin{cases} \inf\limits_{\theta \geq 0} & 1 - b_k \theta \\ \text{s.t.} & a_k \theta = -z, \end{cases}$$

for all $k \leq K$. Moreover, we have $[-\ell_{K+1}]^* = 0$ if $\xi = 0$; $= \infty$ otherwise. Assertion (ii) then follows by substituting the formulas for $[-\ell_k]^*$, $k \leq K+1$, and σ_Ξ into (11).

Assertion (ii) follows from Theorem 4.2 by setting $K = 2$, $\ell_1(\xi) = 1 - \chi_{\mathbb{A}}(\xi)$ and $\ell_2(\xi) = 0$. As illustrated in Figure 3(b), this implies that $\ell(\xi) = \max\{\ell_1(\xi), \ell_2(\xi)\} = \mathbb{1}_{\mathbb{A}}(\xi)$ and

$$\sup_{\mathbb{Q} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{Q}[\xi \in \mathbb{A}] = \sup_{\mathbb{Q} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\ell(\xi)].$$

Assumption 4.1 holds by our assumptions on $\mathbb{A}$ and Ξ. In order to apply Theorem 4.2, we thus have to determine the support function σ_Ξ, which was already calculated in Corollary 5.1, and the conjugate functions of $-\ell_1$ and $-\ell_2$. By the definition of the conjugacy operator, we find

$$[-\ell_1]^*(z) = \sup_{\xi \in \mathbb{A}} \langle z, \xi \rangle + 1 = \begin{cases} \sup\limits_{\xi} & \langle z, \xi \rangle + 1 \\ \text{s.t.} & A\xi \leq b \end{cases} = \begin{cases} \inf\limits_{\theta_k \geq 0} & \langle \theta, b \rangle + 1 \\ \text{s.t.} & A^\intercal \theta = z \end{cases}$$

where the last equality follows from strong linear programming duality, which holds as the safe set is non-empty. Similarly, we find $[-\ell_2]^* = 0$ if $\xi = 0$; $= \infty$ otherwise. Assertion (ii) then follows by substituting the above expressions into (11). □

In the *ambiguity-free limit* (*i.e.*, for $\varepsilon = 0$) the optimal value of (17a) reduces to the fraction of training samples residing outside of the open polytope $\mathbb{A} = \{\xi : A\xi < b\}$. Indeed, in this case the variable λ can be set to any positive value at no penalty. For this reason and because all training samples belong to the

uncertainty set (*i.e.*, $d - C\widehat{\xi}_i \geq 0$ for all $i \leq N$), it is optimal to set $\gamma_{ik} = 0$. If the i-th training sample belongs to $\mathbb{A}$ (*i.e.*, $b_k - \langle a_k, \widehat{\xi}_i \rangle > 0$ for all $k \leq K$), then $\theta_{ik} \geq 1/(b_k - \langle a_k, \widehat{\xi}_i \rangle)$ for all $k \leq K$ and $s_i = 0$ at optimality. Conversely, if the i-th training sample belongs to the complement of $\mathbb{A}$, (*i.e.*, $b_k - \langle a_k, \widehat{\xi}_i \rangle \leq 0$ for some $k \leq K$), then $\theta_{ik} = 0$ for some $k \leq K$ and $s_i = 1$ at optimality. Thus, $\sum_{i=1}^N s_i$ coincides with the number of training samples outside of $\mathbb{A}$ at optimality. An analogous argument shows that, for $\varepsilon = 0$, the optimal value of (17b) reduces to the fraction of training samples residing inside of the closed polytope $\mathbb{A} = \{\xi : A\xi \leq b\}$.

5.3. **Two-Stage Stochastic Programming**

A major challenge in linear two-stage stochastic programming is to evaluate the expected recourse costs, which are only implicitly defined as the optimal value of a linear program whose coefficients depend linearly on the uncertain problem parameters [46, Section 2.1]. The following corollary shows how we can evaluate the worst-case expectation of the recourse costs with respect to an ambiguous parameter distribution that is only observable through a finite training dataset. For ease of notation and without loss of generality, we suppress here any dependence on the first-stage decisions.

Corollary 5.4 (Two-stage stochastic programming)**.** *Suppose that the uncertainty set is a polytope of the form $\Xi = \{\xi \in \mathbb{R}^m : C\xi \leq d\}$ as in Corollaries 5.1 and 5.3.*

(i) If $\ell(\xi) = \inf_y \{\langle y, Q\xi \rangle : Wy \geq h\}$ is the optimal value of a parametric linear program with objective uncertainty, and if the feasible set $\{y : Wy \geq h\}$ is non-empty and compact, then the worst-case expectation (10) is given by

$$\left\{ \begin{array}{lll} \inf\limits_{\lambda, s_i, \gamma_i, y_i} & \lambda\varepsilon + \frac{1}{N} \sum\limits_{i=1}^N s_i & \\ \text{s.t.} & \langle y_i, Q\widehat{\xi}_i \rangle + \langle \gamma_i, d - C\widehat{\xi}_i \rangle \leq s_i & \forall i \leq N \\ & Wy_i \geq h & \forall i \leq N \\ & \|Q^\intercal y_i - C^\intercal \gamma_i\|_* \leq \lambda & \forall i \leq N \\ & \gamma_i \geq 0 & \forall i \leq N. \end{array} \right. \tag{18a}$$

(ii) If $\ell(\xi) = \inf_y \{\langle q, y \rangle : Wy \geq H\xi + h\}$ is the optimal value of a parametric linear program with right-hand side uncertainty, and if the dual feasible set $\{\theta \geq 0 : W^\intercal \theta = q\}$ is non-empty and compact with vertices v_k, $k \leq K$, then the worst-case expectation (10) is given by

$$\left\{ \begin{array}{lll} \inf\limits_{\lambda, s_i, \gamma_{ik}} & \lambda\varepsilon + \frac{1}{N} \sum\limits_{i=1}^N s_i & \\ \text{s.t.} & \langle v_k, h \rangle + \langle H^\intercal v_k, \widehat{\xi}_i \rangle + \langle \gamma_{ik}, d - C\widehat{\xi}_i \rangle \leq s_i & \forall i \leq N, \quad \forall k \leq K \\ & \|C^\intercal \gamma_{ik} - H^\intercal v_k\|_* \leq \lambda & \forall i \leq N, \quad \forall k \leq K \\ & \gamma_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K. \end{array} \right. \tag{18b}$$

Proof. Assertion (i) follows directly from Theorem 4.2 because $\ell(\xi)$ is concave as an infimum of linear functions in ξ. Indeed, the compactness of the feasible set $\{y : Wy \geq h\}$ ensures that Assumption 4.1 holds for $K = 1$. In this setting, we find

$$\begin{aligned} [-\ell]^*(z) &= \sup_\xi \left\{ \langle z, \xi \rangle + \inf_y \{\langle y, Q\xi \rangle : Wy \geq h\} \right\} \\ &= \inf_y \left\{ \sup_\xi \{\langle z + Q^\intercal y, \xi \rangle\} : Wy \geq h \right\} \end{aligned}$$

$$= \begin{cases} 0 & \text{if there exists } y \text{ with } Q^\intercal y = -z \text{ and } Wy \geq h, \\ \infty & \text{otherwise,} \end{cases}$$

where the second equality follows from the classical minimax theorem [4, Proposition 5.5.4], which applies because $\{y : Wy \geq h\}$ is compact. Assertion (i) then follows by substituting $[-\ell]^*$ as well as the formula for σ_Ξ from Corollary 5.1 into (11).

Assertion (ii) relies on the following reformulation of the loss function,

$$\ell(\xi) = \left\{ \begin{array}{ll} \inf\limits_{y} & \langle q, y \rangle \\ \text{s.t.} & Wy \geq H\xi + h \end{array} \right. = \left\{ \begin{array}{ll} \sup\limits_{\theta \geq 0} & \langle \theta, H\xi + h \rangle \\ \text{s.t.} & W^\intercal \theta = q \end{array} \right. = \max_{k \leq K} \langle v_k, H\xi + h \rangle$$
$$= \max_{k \leq K} \langle H^\intercal v_k, \xi \rangle + \langle v_k, h \rangle,$$

where the first equality holds due to strong linear programming duality, which applies as the dual feasible set is non-empty. The second equality exploits the elementary observation that the optimal value of a linear program with non-empty, compact feasible set is always adopted at a vertex. As we managed to express $\ell(\xi)$ as a pointwise maximum of linear functions, assertion (ii) follows immediately from Corllary 5.1 (i). □

As expected, in the *ambiguity-free limit*, problem (18a) reduces to a standard SAA problem. Indeed, for $\varepsilon = 0$, the variable λ can be made large at no penalty, and thus $\gamma_i = 0$ and $s_i = \langle y_i, Q\widehat{\xi}_i \rangle$ at optimality. In this case, problem (18a) is equivalent to

$$\inf_{y_i} \left\{ \frac{1}{N} \sum_{i=1}^{N} \langle y_i, Q\widehat{\xi}_i \rangle : Wy_i \geq h \quad \forall i \leq N \right\}.$$

Similarly, one can verify that for $\varepsilon = 0$, (18b) reduces to the SAA problem

$$\inf_{y_i} \left\{ \frac{1}{N} \sum_{i=1}^{N} \langle y_i, q \rangle : Wy_i \geq H\widehat{\xi}_i \quad \forall i \leq N \right\}.$$

We close this section with a remark on the computational complexity of all the convex optimization problems derived in this section.

Remark 5.5 (Computational tractability)**.**

- *If the Wasserstein metric is defined in terms of the* 1*-norm (i.e.,* $\|\xi\| = \sum_{k=1}^{m} |\xi_k|$*) or the* ∞*-norm (i.e.,* $\|\xi\| = \max_{k \leq m} |\xi_k|$*), then the optimization problems* (15a)*,* (15b)*,* (17a)*,* (17b)*,* (18a) *and* (18b) *all reduce to linear programs whose sizes scale with the number* N *of data points and the number* J *of affine pieces of the underlying loss functions.*
- *Except for the two-stage stochastic program with right-hand side uncertainty in* (18b)*, the resulting linear programs scale polynomially in the problem description and are therefore computationally tractable. As the number of vertices* v_k*,* $k \leq K$*, of the polytope* $\{\theta \geq 0 : W^\intercal \theta = q\}$ *may be exponential in the number of its facets, however, the linear program* (18b) *has generically exponential size.*
- *Inspecting* (15a)*, one easily verifies that the distributionally robust optimization problem* (5) *reduces to a finite convex program if* $\mathbb{X}$ *is convex and* $h(x, \xi) = \max_{k \leq K} \langle a_k(x), \xi \rangle + b_k(x)$*, while the gradients* $a_k(x)$ *and the intercepts* $b_k(x)$ *depend linearly on* x*. Similarly,* (5) *can be reformulated as a finite convex program if* $\mathbb{X}$ *is convex and* $h(x, \xi) = \inf_y \{\langle y, Q\xi \rangle : Wy \geq h(x)\}$ *or* $h(x, \xi) = \inf_y \{\langle q, y \rangle : Wy \geq H(x)\xi + h(x)\}$*, while the right hand side coefficients* $h(x)$ *and* $H(x)$ *depend linearly on* x*; see* (18a) *and* (18b)*, respectively. In contrast, problems* (15b)*,* (17a) *and* (17b) *result in non-convex optimization problems when their data depends on* x*.*

- *We emphasize that the computational complexity of all convex programs examined in this section is independent of the radius ε of the Wasserstein ball.*

6. Tractable Extensions

We now demonstrate that through minor modifications of the proofs, Theorems 4.2 and 4.4 extend to worst-case expectation problems involving even richer classes of loss functions. First, we investigate problems where the uncertainty can be viewed as a stochastic process and where the loss function is additively separable. Next, we study problems whose loss functions are convex in the uncertain variables and are therefore not necessarily representable as finite maxima of concave functions as postulated by Assumption 4.1.

6.1. Stochastic Processes with a Separable Cost

Consider a variant of the worst-case expectation problem (10), where the uncertain parameters can be interpreted as a stochastic process $\xi = (\xi_1, \ldots, \xi_T)$, and assume that $\xi_t \in \Xi_t$, where $\Xi_t \subseteq \mathbb{R}^m$ is non-empty and closed for any $t \leq T$. Moreover, assume that the loss function is additively separable with respect to the temporal structure of ξ, that is,

$$\ell(\xi) \coloneqq \sum_{t=1}^{T} \max_{k \leq K} \ell_{tk}(\xi_t), \tag{19}$$

where $\ell_{tk} : \mathbb{R}^m \to \overline{\mathbb{R}}$ is a measurable function for any $k \leq K$ and $t \leq T$. Such loss functions appear, for instance, in open-loop stochastic optimal control or in multi-item newsvendor problems. Consider a process norm $\|\xi\|_{\mathrm{T}} = \sum_{t=1}^{T} \|\xi_t\|$ associated with the base norm $\|\cdot\|$ on $\mathbb{R}^m$, and assume that its induced metric is the one used in the definition of the Wasserstein distance. Note that if $\|\cdot\|$ is the 1-norm on $\mathbb{R}^m$, then $\|\cdot\|_{\mathrm{T}}$ reduces to the 1-norm on $\mathbb{R}^{mT}$.

By interchanging summation and maximization, the loss function (19) can be re-expressed as

$$\ell(\xi) = \max_{k_t \leq K} \sum_{t=1}^{T} \ell_{tk_t}(\xi_t),$$

where the maximum runs over all K^T combinations of $k_1, \ldots, k_T \leq K$. Under this representation, Theorem 4.2 remains applicable. However, the resulting convex optimization problem would involve $\mathcal{O}(K^T)$ decision variables and constraints, indicating that an efficient solution may not be available. Fortunately, this deficiency can be overcome by modifying Theorem 4.2.

Theorem 6.1 (Convex reduction for separable loss functions)**.** *Assume that the loss function ℓ is of the form (19), and the Wasserstein ball is defined through the process norm $\|\cdot\|_{\mathrm{T}}$. Then, for any $\varepsilon \geq 0$, the worst-case expectation (10) is smaller or equal to the optimal value of the finite convex program*

$$\left\{\begin{array}{lll} \inf\limits_{\lambda, s_{ti}, z_{tik}, \nu_{tik}} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N}\sum\limits_{t=1}^{T} s_{ti} & \\ \text{s.t.} & [-\ell_{tk}]^*(z_{tik} - \nu_{tik}) + \sigma_{\Xi_t}(\nu_{tik}) - \langle z_{tik}, \widehat{\xi}_{ti}\rangle \leq s_{ti} \quad \forall i \leq N, \quad \forall k \leq K, \quad \forall t \leq T, \\ & \|z_{tik}\|_* \leq \lambda \qquad\qquad\qquad\qquad\qquad\qquad\quad \forall i \leq N, \quad \forall k \leq K, \quad \forall t \leq T. \end{array}\right. \tag{20}$$

If Ξ_t and $\{\ell_{tk}\}_{k \leq K}$ satisfy the convexity Assumption 4.1 for every $t \leq T$, then the worst-case expectation (10) coincides exactly with the optimal value of problem (20).

Proof. Up until equation (12d), the proof of Theorem 6.1 parallels that of Theorem 4.2. Starting from (12d), we then have

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\big[\ell(\xi)\big] &= \inf_{\lambda\geq 0}\ \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\sup_{\xi}\Big(\ell(\xi) - \lambda\Big\|\xi-\widehat{\xi}_i\Big\|_{\mathrm{T}}\Big)\\ &= \inf_{\lambda\geq 0}\ \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sup_{\xi_t\in\Xi_t}\Big(\max_{k\leq K}\ell_{tk}(\xi_t) - \lambda\big\|\xi_t-\widehat{\xi}_{ti}\big\|\Big),\end{aligned}$$

where the interchange of the summation and the maximization is facilitated by the separability of the overall loss function. Introducing epigraphical auxiliary variables yields

$$\left\{\begin{array}{ll}\inf\limits_{\lambda,s_{ti}} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N}\sum\limits_{t=1}^{T} s_{ti}\\ \text{s.t.} & \sup\limits_{\xi_t\in\Xi_t}\Big(\ell_{tk}(\xi_t) - \lambda\big\|\xi_t-\widehat{\xi}_{ti}\big\|\Big)\leq s_{ti}\quad \forall i\leq N,\ \forall k\leq K,\ \forall t\leq T\\ & \lambda\geq 0\end{array}\right.$$

$$\leq\left\{\begin{array}{lll}\inf\limits_{\lambda,s_{ti},z_{tik}} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N}\sum\limits_{t=1}^{T} s_{ti} & \\ \text{s.t.} & \sup\limits_{\xi_t\in\Xi_t}\Big(\ell_{tk}(\xi_t) - \big\langle z_{tik},\xi_t\big\rangle\Big) + \big\langle z_{tik},\widehat{\xi}_{ti}\big\rangle\leq s_{ti} & \forall i\leq N,\ \forall k\leq K,\ \forall t\leq T\\ & \|z_{tik}\|_*\leq\lambda & \forall i\leq N,\ \forall k\leq K,\ \forall t\leq T\end{array}\right.$$

$$=\left\{\begin{array}{lll}\inf\limits_{\lambda,s_{ti},z_{tik}} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^{N}\sum\limits_{t=1}^{T} s_{ti} & \\ \text{s.t.} & [-\ell_{tk}+\chi_{\Xi_t}]^*\big(-z_{tik}\big) + \big\langle z_{tik},\widehat{\xi}_{ti}\big\rangle\leq s_{ti} & \forall i\leq N,\ \forall k\leq K,\ \forall t\leq T\\ & \|z_{tik}\|_*\leq\lambda & \forall i\leq N,\ \forall k\leq K,\ \forall t\leq T,\end{array}\right.$$

where the inequality is justified in a similar manner as the one in (12e), and it holds as an equality provided that Ξ_t and $\{\ell_{tk}\}_{k\leq K}$ satisfy Assumption 4.1 for all $t\leq T$. Finally, by [42, Theorem 11.23(a), p. 493], the conjugate of $-\ell_{tk}+\chi_{\Xi_t}$ can be replaced by the inf-convolution of the conjugates of $-\ell_{tk}$ and χ_{Ξ_t}. This completes the proof. □

Note that the convex program (20) involves only $\mathcal{O}(NKT)$ decision variables and constraints. Moreover, if ℓ_{tk} is affine for every $t\leq T$ and $k\leq K$, while $\|\cdot\|$ represents the 1-norm or the ∞-norm on $\mathbb{R}^m$, then (20) reduces to a tractable linear program (see also Remark 5.5). A natural generalization of Theorem 4.4 further allows us to characterize the extremal distributions of the worst-case expectation problem (10) with a separable loss function of the form (19).

Theorem 6.2 (Worst-case distributions for separable loss functions)**.** *Assume that the loss function ℓ is of the form (19), and the Wasserstein ball is defined through the process norm $\|\cdot\|_{\mathrm{T}}$. If Ξ_t and $\{\ell_{tk}\}_{k\leq K}$ satisfy Assumption 4.1 for all $t\leq T$, then the worst-case expectation (10) coincides with the optimal value of the finite convex program*

$$\left\{\begin{array}{lll}\sup\limits_{\alpha_{tik},q_{tik}} & \frac{1}{N}\sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K}\sum\limits_{t=1}^{T}\alpha_{tik}\ell_{tk}\Big(\widehat{\xi}_{ti}-\frac{q_{tik}}{\alpha_{tik}}\Big) & \\ \text{s.t.} & \frac{1}{N}\sum\limits_{i=1}^{N}\sum\limits_{k=1}^{K}\sum\limits_{t=1}^{T}\|q_{tik}\|\leq\varepsilon & \\ & \sum\limits_{k=1}^{K}\alpha_{tik}=1 & \forall i\leq N,\quad\forall t\leq T\\ & \alpha_{tik}\geq 0 & \forall i\leq N,\quad\forall t\leq T,\quad\forall k\leq K\\ & \widehat{\xi}_{ti}-\frac{q_{tik}}{\alpha_{tik}}\in\Xi_t & \forall i\leq N,\quad\forall t\leq T,\quad\forall k\leq K\end{array}\right.\tag{21}$$

irrespective of $\varepsilon \geq 0$. *Let* $\big\{\alpha_{tik}(r), q_{tik}(r)\big\}_{r\in\mathbb{N}}$ *be a sequence of feasible decisions whose objective values converge to the supremum of* (21). *Then, the discrete (product) probability distributions*

$$\mathbb{Q}_r \coloneqq \frac{1}{N}\sum_{i=1}^{N}\bigotimes_{t=1}^{T}\Big(\sum_{k=1}^{K}\alpha_{tik}(r)\delta_{\xi_{tik}(r)}\Big) \quad \text{with} \quad \xi_{tik}(r) \coloneqq \widehat{\xi}_{ti} - \frac{q_{tik}(r)}{\alpha_{tik}(r)}$$

belong to the Wasserstein ball $\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ *and attain the supremum of* (10) *asymptotically, i.e.,*

$$\sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\big[\ell(\xi)\big] = \lim_{r\to\infty}\mathbb{E}^{\mathbb{Q}_r}\big[\ell(\xi)\big] = \lim_{r\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{t=1}^{T}\alpha_{tik}(r)\ell_{tk}\big(\xi_{tik}(r)\big).$$

Proof. As in the proof of Theorem 4.4, the claim follows by dualizing the convex program (20). Details are omitted for brevity of exposition. □

We emphasize that the distributions $\mathbb{Q}_r$ from Theorem 6.2 can be constructed efficiently by solving a convex program of polynomial size even though they have NK^T discretization points.

6.2. Convex Loss Functions

Consider now another variant of the worst-case expectation problem (10), where the loss function ℓ is proper, convex and lower semicontinuous. Unless ℓ is piecewise affine, we cannot represent such a loss function as a pointwise maximum of finitely many concave functions, and thus Theorem 4.2 may only provide a loose upper bound on the worst-case expectation (10). The following theorem provides an alternative upper bound that admits new insights into distributionally robust optimization with Wasserstein balls and becomes exact for $\Xi = \mathbb{R}^m$.

Theorem 6.3 (Convex reduction for convex loss functions)**.** *Assume that the loss function* ℓ *is proper, convex, and lower semicontinuous, and define* $\kappa \coloneqq \sup\big\{\|\theta\|_* : \ell^*(\theta) < \infty\big\}$. *Then, for any* $\varepsilon \geq 0$, *the worst-case expectation* (10) *is smaller or equal to*

$$\kappa\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\ell(\widehat{\xi}_i). \tag{22}$$

If $\Xi = \mathbb{R}^m$, *then the worst-case expectation* (10) *coincides exactly with* (22).

Remark 6.4 (Radius of effective domain)**.** *The parameter* κ *can be viewed as the radius of the smallest ball containing the effective domain of the conjugate function* ℓ^* *in terms of the dual norm. By the standard conventions of extended arithmetic, the term* $\kappa\varepsilon$ *in* (22) *is interpreted as* 0 *if* $\kappa = \infty$ *and* $\varepsilon = 0$.

Proof. Equation (12b) in the proof of Theorem 4.2 implies that

$$\sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\big[\ell(\xi)\big] = \inf_{\lambda\geq 0}\ \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\sup_{\xi\in\Xi}\Big(\ell(\xi) - \lambda\|\xi - \widehat{\xi}_i\|\Big) \tag{23}$$

for every $\varepsilon > 0$. As ℓ is proper, convex, and lower semicontinuous, it coincides with its bi-conjugate function ℓ^{**}, see *e.g.* [4, Proposition 1.6.1(c)]. Thus, we may write

$$\ell(\xi) = \sup_{\theta\in\Theta}\big\langle\theta,\xi\big\rangle - \ell^*(\theta),$$

where $\Theta \coloneqq \{\theta\in\mathbb{R}^m : \ell^*(\theta) < \infty\}$ denotes the effective domain of the conjugate function ℓ^*. Using this dual representation of ℓ in conjunction with the definition of the dual norm, we find

$$\sup_{\xi\in\Xi}\Big(\ell(\xi) - \lambda\|\xi - \widehat{\xi}_i\|\Big) = \sup_{\xi\in\Xi}\ \sup_{\theta\in\Theta}\Big(\big\langle\theta,\xi\big\rangle - \ell^*(\theta) - \lambda\|\xi - \widehat{\xi}_i\|\Big)$$

$$= \sup_{\xi\in\Xi} \sup_{\theta\in\Theta} \inf_{\|z\|_*\leq\lambda} \Big(\langle \theta, \xi \rangle - \ell^*(\theta) + \langle z, \xi \rangle - \langle z, \widehat{\xi}_i \rangle \Big).$$

The classical minimax theorem [4, Proposition 5.5.4] then allows us to interchange the maximization over ξ with the maximization over θ and the minimization over z to obtain

$$\begin{aligned} \sup_{\xi\in\Xi} \Big(\ell(\xi) - \lambda\|\xi - \widehat{\xi}_i\| \Big) &= \sup_{\theta\in\Theta} \inf_{\|z\|_*\leq\lambda} \sup_{\xi\in\Xi} \Big(\langle \theta + z, \xi \rangle - \ell^*(\theta) - \langle z, \widehat{\xi}_i \rangle \Big) \\ &= \sup_{\theta\in\Theta} \inf_{\|z\|_*\leq\lambda} \sigma_\Xi(\theta + z) - \ell^*(\theta) - \langle z, \widehat{\xi}_i \rangle. \end{aligned} \tag{24}$$

Recall that σ_Ξ denotes the support function of Ξ. It seems that there is no simple exact reformulation of (24) for arbitrary convex uncertainty sets Ξ. Interchanging the maximization over θ with the minimization over z in (24) would lead to the conservative upper bound of Corollary 4.3. Here, however, we employ an alternative approximation. By definition of the support function, we have $\sigma_\Xi \leq \sigma_{\mathbb{R}^m} = \chi_{\{0\}}$. Replacing σ_Ξ with $\chi_{\{0\}}$ in (24) thus results in the conservative approximation

$$\sup_{\xi\in\mathbb{R}^m} \Big(\ell(\xi) - \lambda\|\xi - \widehat{\xi}_i\| \Big) \leq \begin{cases} \ell(\widehat{\xi}_i) & \text{if } \sup\big\{\|\theta\|_* : \theta \in \Theta\big\} \leq \lambda, \\ \infty & \text{otherwise.} \end{cases} \tag{25}$$

The inequality (22) then follows readily by substituting (25) into (23) and using the definition of κ in the theorem statement. For $\Xi = \mathbb{R}^m$ we have $\sigma_\Xi = \chi_{\{0\}}$, and thus the upper bound (22) becomes exact. Finally, if $\varepsilon = 0$, then (10) trivially coincides with (22) under our conventions of extended arithmetic. Thus, the claim follows. □

Theorem 6.3 asserts that for $\Xi = \mathbb{R}^m$, the worst-case expectation (10) of a convex loss function reduces the sample average of the loss adjusted by the simple correction term $\kappa\varepsilon$. The following proposition highlights that κ can be interpreted as a measure of maximum steepness of the loss function. This interpretation has intuitive appeal in view of Definition 3.1.

Proposition 6.5 (Steepness of the loss function)**.** *Let κ be defined as in Theorem 6.3.*

(i) If ℓ is $\overline{L}$-Lipschitz continuous, i.e., if there exists $\xi' \in \mathbb{R}^m$ such that $\ell(\xi) - \ell(\xi') \leq \overline{L}\|\xi - \xi'\|$ for all $\xi \in \mathbb{R}^m$, then $\kappa \leq \overline{L}$.

(ii) If ℓ majorizes an affine function, i.e., if there exists $\theta \in \mathbb{R}^m$ with $\|\theta\|_* =: \underline{L}$ and $\xi' \in \mathbb{R}^m$ such that $\ell(\xi) - \ell(\xi') \geq \langle \theta, \xi - \xi' \rangle$ for all $\xi \in \mathbb{R}^m$, then $\kappa \geq \underline{L}$.

Proof. The proof follows directly from the definition of conjugacy. As for (i), we have

$$\begin{aligned} \ell^*(\theta) = \sup_{\xi\in\mathbb{R}^m} \langle \theta, \xi \rangle - \ell(\xi) &\geq \sup_{\xi\in\mathbb{R}^m} \langle \theta, \xi \rangle - \overline{L}\|\xi - \xi'\| - \ell(\xi') \\ &= \sup_{\xi\in\mathbb{R}^m} \inf_{\|z\|_*\leq\overline{L}} \langle \theta, \xi \rangle - \langle z, \xi - \xi' \rangle - \ell(\xi'), \end{aligned}$$

where the last equality follows from the definition of the dual norm. Applying the minimax theorem [4, Proposition 5.5.4] and explicitly carrying out the maximization over ξ yields

$$\ell^*(\theta) \geq \begin{cases} \langle \theta, \xi' \rangle - \ell(\xi') & \text{if } \|\theta\|_* \leq \overline{L}, \\ \infty & \text{otherwise.} \end{cases}$$

Consequently, $\ell^*(\theta)$ is infinite for all θ with $\|\theta\|_* > \overline{L}$, which readily implies that the $\|\cdot\|_*$-ball of radius $\overline{L}$ contains the effective domain of ℓ^*. Thus, $\kappa \leq \overline{L}$.

As for (ii), we have

$$\ell^*(\theta) = \sup_{\xi\in\mathbb{R}^m} \langle \theta, \xi \rangle - \ell(\xi) \leq \sup_{\xi\in\mathbb{R}^m} \langle \theta, \xi \rangle - \langle z, \xi - \xi' \rangle - \ell(\xi')$$

$$= \sigma_{\mathbb{R}^m}(\theta - z) + \langle z, \xi' \rangle - \ell(\xi'),$$

which implies that $\ell^*(\theta) \leq \langle \theta, \xi' \rangle - \ell(\xi') < \infty$. Thus, θ belongs to the effective domain of ℓ^*. We then conclude that $\kappa \geq \|\theta\|_* = \underline{L}$. □

Remark 6.6 (Consistent formulations)**.** *If $\Xi = \mathbb{R}^m$ and the loss function is given by $\ell(\xi) = \max_{k \leq K} \{\langle a_k, \xi \rangle + b_k\}$, then both Corollary 5.1 and Theorem 6.3 offer an exact reformulation of the worst-case expectation (10) in terms of a finite-dimensional convex program. On the one hand, Corollary 5.1 implies that (10) is equivalent to*

$$\left\{ \begin{array}{ll} \min\limits_{\lambda} & \lambda\varepsilon + \frac{1}{N} \sum\limits_{i=1}^{N} \ell(\widehat{\xi}_i) \\ \text{s.t.} & \|a_k\|_* \leq \lambda \qquad \forall k \leq K, \end{array} \right.$$

which is obtained by setting $C = 0$ and $d = 0$ in (15a). At optimality we have $\lambda^\star = \max_{k \leq K} \|a_k\|_*$, which corresponds to the (best) Lipschitz constant of $\ell(\xi)$ with respect to the norm $\|\cdot\|$. On the other hand, Theorem 6.3 implies that (10) is equivalent to (22) with $\kappa = \lambda^\star$. Thus, Corollary 5.1 and Theorem 6.3 are consistent.

Remark 6.7 (ε-insensitive optimizers[3])**.** *Consider a loss function $h(x, \xi)$ that is convex in ξ, and assume that $\Xi = \mathbb{R}^m$. In this case Theorem 6.3 remains valid, but the steepness parameter $\kappa(x)$ may depend on x. For loss functions whose Lipschitz modulus with respect to ξ is independent of x (e.g., the newsvendor loss), however, $\kappa(x)$ is constant. In this case the distributionally robust optimization problem (5) and the SAA problem (4) share the same minimizers irrespective of the Wasserstein radius ε. This phenomenon could explain why the SAA solutions tend to display a surprisingly strong out-of-sample performance in these problems.*

7. Numerical Results

We validate the theoretical results of this paper in the context of a stylized portfolio selection problem. The subsequent simulation experiments are designed to provide additional insights into the performance guarantees of the proposed distributionally robust optimization scheme.

7.1. Mean-Risk Portfolio Optimization

Consider a capital market consisting of m assets whose yearly returns are captured by the random vector $\xi = [\xi_1, \ldots, \xi_m]^\intercal$. If short-selling is forbidden, a portfolio is encoded by a vector of percentage weights $x = [x_1, \ldots, x_m]^\intercal$ ranging over the probability simplex $\mathbb{X} = \{x \in \mathbb{R}^m_+ : \sum_{i=1}^m x_i = 1\}$. As portfolio x invests a percentage x_i of the available capital in asset i for each $i = 1, \ldots, m$, its return amounts to $\langle x, \xi \rangle$. In the remainder we aim to solve the single-stage stochastic program

$$J^\star = \inf_{x \in \mathbb{X}} \left\{ \mathbb{E}^{\mathbb{P}} \big[-\langle x, \xi \rangle \big] + \rho\, \mathbb{P}\text{-CVaR}_\alpha \big(-\langle x, \xi \rangle \big) \right\}, \tag{26}$$

which minimizes a weighted sum of the mean and the conditional value-at-risk (CVaR) of the portfolio loss $-\langle x, \xi \rangle$, where $\alpha \in (0, 1]$ is referred to as the confidence level of the CVaR, and $\rho \in \mathbb{R}_+$ quantifies the investor's risk-aversion. Intuitively, the CVaR at level α represents the average of the $\alpha \times 100\%$ worst (highest) portfolio losses under the distribution $\mathbb{P}$. Replacing the CVaR in the above expression with its formal definition [41], we obtain

$$\begin{aligned} J^\star &= \inf_{x \in \mathbb{X}} \left\{ \mathbb{E}^{\mathbb{P}} \big[-\langle x, \xi \rangle \big] + \rho \inf_{\tau \in \mathbb{R}} \mathbb{E}^{\mathbb{P}} \Big[\tau + \frac{1}{\alpha} \max \big\{ -\langle x, \xi \rangle - \tau, 0 \big\} \Big] \right\} \\ &= \inf_{x \in \mathbb{X}, \tau \in \mathbb{R}} \mathbb{E}^{\mathbb{P}} \Big[\max_{k \leq K} a_k \langle x, \xi \rangle + b_k \tau \Big], \end{aligned}$$

[3]We are indepted to Vishal Gupta who has brought this interesting observation to our attention.

where $K = 2$, $a_1 = -1$, $a_2 = -1 - \frac{\rho}{\alpha}$, $b_1 = \rho$ and $b_2 = \rho(1 - \frac{1}{\alpha})$. An investor who is unaware of the distribution $\mathbb{P}$ but has observed a dataset $\widehat{\Xi}_N$ of N historical samples from $\mathbb{P}$ and knows that the support of $\mathbb{P}$ is contained in $\Xi = \{\xi \in \mathbb{R}^m : C\xi \leq d\}$ might solve the distributionally robust counterpart of (26) with respect to the Wasserstein ambiguity set $\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, that is,

$$\widehat{J}_N(\varepsilon) \coloneqq \inf_{x\in\mathbb{X},\tau\in\mathbb{R}} \sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\Big[\max_{k\leq K} a_k\langle x,\xi\rangle + b_k\tau\Big],$$

where we make the dependence on the Wasserstein radius ε explicit. By Corollary 5.1 we know that

$$\widehat{J}_N(\varepsilon) = \left\{ \begin{array}{lll} \inf\limits_{x,\tau,\lambda,s_i,\gamma_{ik}} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^N s_i & \\ \text{s.t.} & x \in \mathbb{X} & \\ & b_k\tau + a_k\langle x,\widehat{\xi}_i\rangle + \langle \gamma_{ik}, d - C\widehat{\xi}_i\rangle \leq s_i & \forall i \leq N, \quad \forall k \leq K \\ & \|C^\intercal\gamma_{ik} - a_k x\|_* \leq \lambda & \forall i \leq N, \quad \forall k \leq K \\ & \gamma_{ik} \geq 0 & \forall i \leq N, \quad \forall k \leq K. \end{array} \right. \tag{27}$$

Before proceeding with the numerical analysis of this problem, we provide some analytical insights into its optimal solutions when there is significant ambiguity. In what follows we keep the training data set fixed and let $\widehat{x}_N(\varepsilon)$ be an optimal distributionally robust portfolio corresponding to the Wasserstein ambiguity set of radius ε. We will now show that, for natural choices of the ambiguity set, $\widehat{x}_N(\varepsilon)$ converges to the equally weighted portfolio $\frac{1}{m}e$ as ε tends to infinity, where $e \coloneqq (1,\ldots,1)^\intercal$. The optimality of the equally weighted portfolio under high ambiguity has first been demonstrated in [37] using analytical methods. We identify this result here as an immediate consequence of Theorem 4.2, which is primarily a computational result.

For any non-empty set $S \subseteq \mathbb{R}^m$ we denote by $\operatorname{recc}(S) \coloneqq \{y \in \mathbb{R}^m : x + \lambda y \in S \ \forall x \in S,\ \forall \lambda \geq 0\}$ the recession cone and by $S^\circ \coloneqq \{y \in \mathbb{R}^m : \langle y, x\rangle \leq 0 \ \forall x \in S\}$ the polar cone of S.

Lemma 7.1. *If $\{\varepsilon_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+$ tends to infinity, then any accumulation point $x^\star$ of $\big\{\widehat{x}_N(\varepsilon_k)\big\}_{k\in\mathbb{N}}$ is a portfolio that has minimum distance to $(\operatorname{recc}(\Xi))^\circ$ with respect to $\|\cdot\|_*$.*

Proof. Note first that $\widehat{x}_N(\varepsilon_k)$, $k \in \mathbb{N}$, and $x^\star$ exist because $\mathbb{X}$ is compact. For large Wasserstein radii ε, the term $\lambda\varepsilon$ dominates the objective function of problem (27). Using standard epi-convergence results [42, Section 7.E], one can thus show that

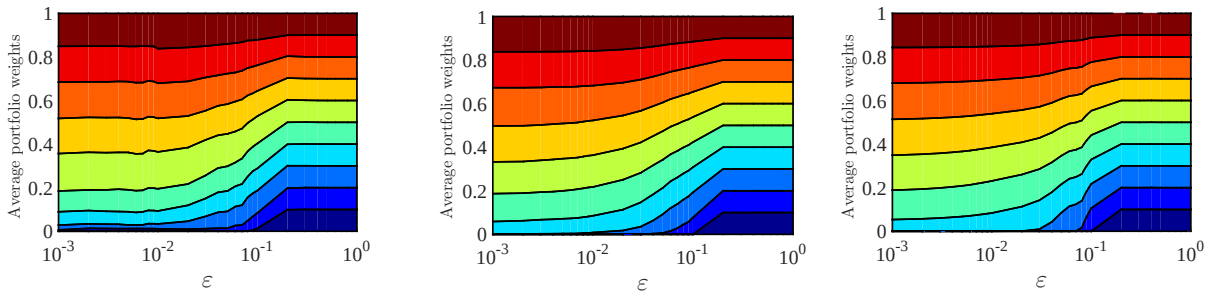
$$\begin{aligned} x^\star &\in \arg\min_{x\in\mathbb{X}} \min_{\gamma_{ik}\geq 0} \max_{i\leq N,\, k\leq K} \|C^\intercal\gamma_{ik} - a_k x\|_* \\ &= \arg\min_{x\in\mathbb{X}} \max_{i\leq N,\, k\leq K} \min_{\gamma\geq 0} \|C^\intercal\gamma + |a_k|\, x\|_* \\ &= \arg\min_{x\in\mathbb{X}} \min_{\gamma\geq 0} \|C^\intercal\gamma + x\|_* \max_{k\leq K} |a_k| \\ &= \arg\min_{x\in\mathbb{X}} \min_{\gamma\geq 0} \|C^\intercal\gamma + x\|_*, \end{aligned}$$

where the first equality follows from the fact that $a_k < 0$ for all $k \leq K$, the second equality uses the substitution $\gamma \to \gamma|a_k|$, and the last equality holds because the set of minimizers of an optimization problem is not affected by a positive scaling of the objective function. Thus, $x^\star$ is the portfolio nearest to the cone $\mathcal{C} = \{C^\intercal\gamma : \gamma \geq 0\}$. The claim now follows as the polar cone

$$\mathcal{C}^\circ \coloneqq \{y \in \mathbb{R}^m : y^\intercal x \leq 0 \ \forall x \in \mathcal{C}\} = \{y \in \mathbb{R}^m : y^\intercal C^\intercal\gamma \leq 0 \ \forall \gamma \geq 0\} = \{y \in \mathbb{R}^m : Cy \geq 0\}$$

is readily recognized as the recession cone of Ξ and as $\mathcal{C} = (\mathcal{C}^\circ)^\circ$. □

Proposition 7.2 (Equally weighted portfolio)**.** *Assume that the Wasserstein metric is defined in terms of the p-norm in the uncertainty space for some $p \in [1,\infty)$. If $\{\varepsilon_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+$ tends to infinity, then $\big\{\widehat{x}_N(\varepsilon_k)\big\}_{k\in\mathbb{N}}$ converges to the equally weighted portfolio $x^\star = \frac{1}{m}e$ provided that the uncertainty set is given by*



(a) $N = 30$ training samples (b) $N = 300$ training samples (c) $N = 3000$ training samples

FIGURE 4. Optimal portfolio composition as a function of the Wasserstein radius ε averaged over 200 simulations; the portfolio weights are depicted in ascending order, *i.e.*, the weight of asset 1 at the bottom (dark blue area) and that of asset 10 at the top (dark red area)

(i) the entire space, i.e., $\Xi = \mathbb{R}^m$*, or*

(ii) the nonnegative orthant shifted by $-e$*, i.e.,* $\Xi = \{\xi \in \mathbb{R}^m : \xi \geq -e\}$*, which captures the idea that no asset can lose more than* 100% *of its value.*

Proof. (i) One easily verifies from the definitions that $(\mathrm{recc}(\Xi))^\circ = \{0\}$. Moreover, we have $\|\cdot\|_* = \|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$. As $p \in [1, \infty)$, we conclude that $q \in (1, \infty]$, and thus the unique nearest portfolio to $(\mathrm{recc}(\Xi))^\circ$ with respect to $\|\cdot\|_*$ is $x^\star = \frac{1}{m}e$. The claim then follows from Lemma 7.1. Assertion (ii) follows in a similar manner from the observation that $(\mathrm{recc}(\Xi))^\circ$ is now the non-positive orthant. □

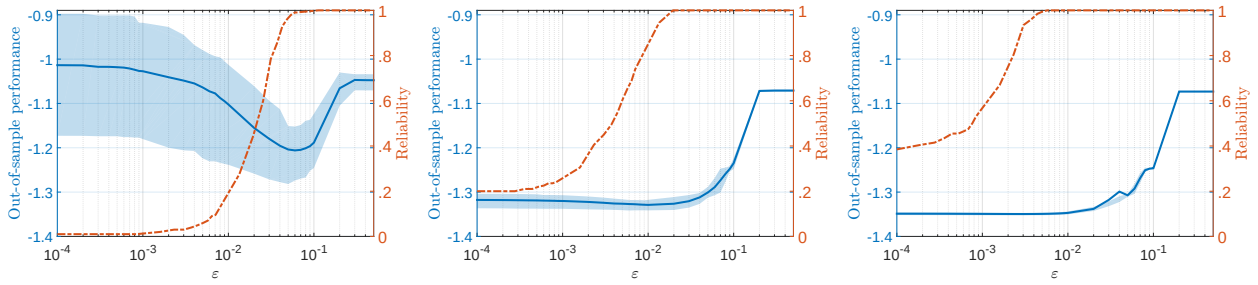
With some extra effort one can show that for every $p \in [1, \infty)$ there is a threshold $\bar{\varepsilon} > 0$ with $\widehat{x}_N(\varepsilon) = x^\star$ for all $\varepsilon \geq \bar{\varepsilon}$, see [37, Proposition 3]. Moreover, for $p \in \{1, 2\}$ the threshold $\bar{\varepsilon}$ is known analytically.

7.2. **Simulation Results: Portfolio Optimization**

Our experiments are based on a market with $m = 10$ assets considered in [7, Section 7.5]. In view of the capital asset pricing model we may assume that the return ξ_i is decomposable into a systematic risk factor $\psi \sim \mathcal{N}(0, 2\%)$ common to all assets and an unsystematic or idiosyncratic risk factor $\zeta_i \sim \mathcal{N}(i \times 3\%, i \times 2.5\%)$ specific to asset i. Thus, we set $\xi_i = \psi + \zeta_i$, where ψ and the idiosyncratic risk factors ζ_i, $i = 1, \ldots, m$, constitute independent normal random variables. By construction, assets with higher indices promise higher mean returns at a higher risk. Note that the given moments of the risk factors completely determine the distribution $\mathbb{P}$ of ξ. This distribution has support $\Xi = \mathbb{R}^m$ and satisfies Assumption 3.3 for the tail exponent $a = 1$, say. We also set $\alpha = 20\%$ and $\rho = 10$ in all numerical experiments, and we use the 1-norm to measure distances in the uncertainty space. Thus, $\|\cdot\|_*$ is the ∞-norm, whereby (27) reduces to a linear program.

7.2.A. **Impact of the Wasserstein Radius**

In the first experiment we investigate the impact of the Wasserstein radius ε on the optimal distributionally robust portfolios and their out-of-sample performance. We solve problem (27) using training datasets of cardinality $N \in \{30, 300, 3000\}$. Figure 4 visualizes the corresponding optimal portfolio weights $\widehat{x}_N(\varepsilon)$ as a function of ε, averaged over 200 independent simulation runs. Our numerical results confirm the theoretical insight of Proposition 7.2 that the optimal distributionally robust portfolios converge to the equally weighted portfolio as the Wasserstein radius ε increases; see also [37].



(a) $N = 30$ training samples (b) $N = 300$ training samples (c) $N = 3000$ training samples

FIGURE 5. Out-of-sample performance $J(\widehat{x}_N(\varepsilon))$ (left axis, solid line and shaded area) and reliability $\mathbb{P}^N[J(\widehat{x}_N(\varepsilon)) \leq \widehat{J}_N(\varepsilon)]$ (right axis, dashed line) as a function of the Wasserstein radius ε and estimated on the basis of 200 simulations

The out-of-sample performance

$$J\big(\widehat{x}_N(\varepsilon)\big) := \mathbb{E}^{\mathbb{P}}\big[-\big\langle \widehat{x}_N(\varepsilon), \xi\big\rangle\big] + \rho\, \mathbb{P}\text{-CVaR}_\alpha\big(-\big\langle \widehat{x}_N(\varepsilon), \xi\big\rangle\big)$$

of any fixed distributionally robust portfolio $\widehat{x}_N(\varepsilon)$ can be computed analytically as $\mathbb{P}$ constitutes a normal distribution by design, see, *e.g.*, [41, p. 29]. Figure 5 shows the tubes between the 20% and 80% quantiles (shaded areas) and the means (solid lines) of the out-of-sample performance $J\big(\widehat{x}_N(\varepsilon)\big)$ as a function of ε—estimated using 200 independent simulation runs. We observe that the out-of-sample performance improves (decreases) up to a critical Wasserstein radius $\varepsilon_{\text{crit}}$ and then deteriorates (increases). This stylized fact was observed consistently across all of simulations and provides an empirical justification for adopting a distributionally robust approach.

Figure 5 also visualizes the reliability of the performance guarantees offered by our distributionally robust portfolio model. Specifically, the dashed lines represent the empirical probability of the event $J\big(\widehat{x}_N(\varepsilon)\big) \leq \widehat{J}_N(\varepsilon)$ with respect to 200 independent training datasets. We find that the reliability is nondecreasing in ε. This observation has intuitive appeal because $\widehat{J}_N(\varepsilon) \geq J(\widehat{x}_N(\varepsilon))$ whenever $\mathbb{P} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, and the latter event becomes increasingly likely as ε grows. Figure 5 also indicates that the certificate guarantee sharply rises towards 1 near the critical Wasserstein radius $\varepsilon_{\text{crit}}$. Hence, the out-of-sample performance of the distributionally robust portfolios improves as long as the reliability of the performance guarantee is noticeably smaller than 1 and deteriorates when it saturates at 1. Even though this observation was made consistently across all simulations, we were unable to validate it theoretically.

7.2.B. **Portfolios Driven by Out-of-Sample Performance**

Different Wasserstein radii ε may result in robust portfolios $\widehat{x}_N(\varepsilon)$ with vastly different out-of-sample performance $J(\widehat{x}_N(\varepsilon))$. Ideally, one should select the radius $\widehat{\varepsilon}_N^{\text{opt}}$ that minimizes $J(\widehat{x}_N(\varepsilon))$ over all $\varepsilon \geq 0$; note that $\widehat{\varepsilon}_N^{\text{opt}}$ inherits the dependence on the training data from $J(\widehat{x}_N(\varepsilon))$. As the true distribution $\mathbb{P}$ is unknown, however, it is impossible to evaluate and minimize $J(\widehat{x}_N(\varepsilon))$. In practice, the best we can hope for is to approximate $\widehat{\varepsilon}_N^{\text{opt}}$ using the training data. Statistics offers several methods to accomplish this goal:

- *Holdout method:* Partition $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ into a training dataset of size N_T and a validation dataset of size $N_V = N - N_T$. Using only the training dataset, solve (27) for a large but finite number of candidate radii ε to obtain $\widehat{x}_{N_T}(\varepsilon)$. Use the validation dataset to estimate the out-of-sample performance of $\widehat{x}_{N_T}(\varepsilon)$ via the sample average approximation. Set $\widehat{\varepsilon}_N^{\text{hm}}$ to any ε that minimizes this quantity. Report $\widehat{x}_N^{\text{hm}} = \widehat{x}_{N_T}(\widehat{\varepsilon}_N^{\text{hm}})$ as the data-driven solution and $\widehat{J}_N^{\text{hm}} = \widehat{J}_{N_T}(\widehat{\varepsilon}_N^{\text{hm}})$ as the corresponding certificate.

- *k-fold cross validation:* Partition $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ into k subsets, and run the holdout method k times. In each run, use exactly one subset as the validation dataset and merge the remaining $k-1$ subsets to a training dataset. Set $\widehat{\varepsilon}_N^{\mathrm{cv}}$ to the average of the Wasserstein radii obtained from the k holdout runs. Resolve (27) with $\varepsilon = \widehat{\varepsilon}_N^{\mathrm{cv}}$ using all N samples, and report $\widehat{x}_N^{\mathrm{cv}} = \widehat{x}_N(\widehat{\varepsilon}_N^{\mathrm{cv}})$ as the data-driven solution and $\widehat{J}_N^{\mathrm{cv}} = \widehat{J}_N(\widehat{\varepsilon}_N^{\mathrm{cv}})$ as the corresponding certificate.

The holdout method is computationally cheaper, but cross validation has superior statistical properties. There are several other methods to estimate the best Wassertein radius $\widehat{\varepsilon}_N^{\mathrm{opt}}$. By construction, however, no method can provide a radius $\widehat{\varepsilon}_N$ such that $\widehat{x}_N(\widehat{\varepsilon}_N)$ has a better out-of-sample performance than $\widehat{x}_N(\widehat{\varepsilon}_N^{\mathrm{opt}})$.

In all experiments we compare the distributionally robust approach based on the Wasserstein ambiguity set with the classical sample average approximation (SAA) and with a state-of-the-art data-driven distributionally robust approach, where the ambiguity set is defined via a linear-convex ordering (LCX)-based goodness-of-fit test [7, Section 3.3.2]. The size of the LCX ambiguity set is determined by a single parameter, which should be tuned to optimize the out-of-sample performance. While the best parameter value is unavailable, it can again be estimated using the holdout method or via cross validation. To our best knowledge, the LCX approach represents the only existing data-driven distributionally robust approach for *continuous* uncertainty spaces that enjoys strong finite-sample guarantees, asymptotic consistency as well as computational tractability.[4]

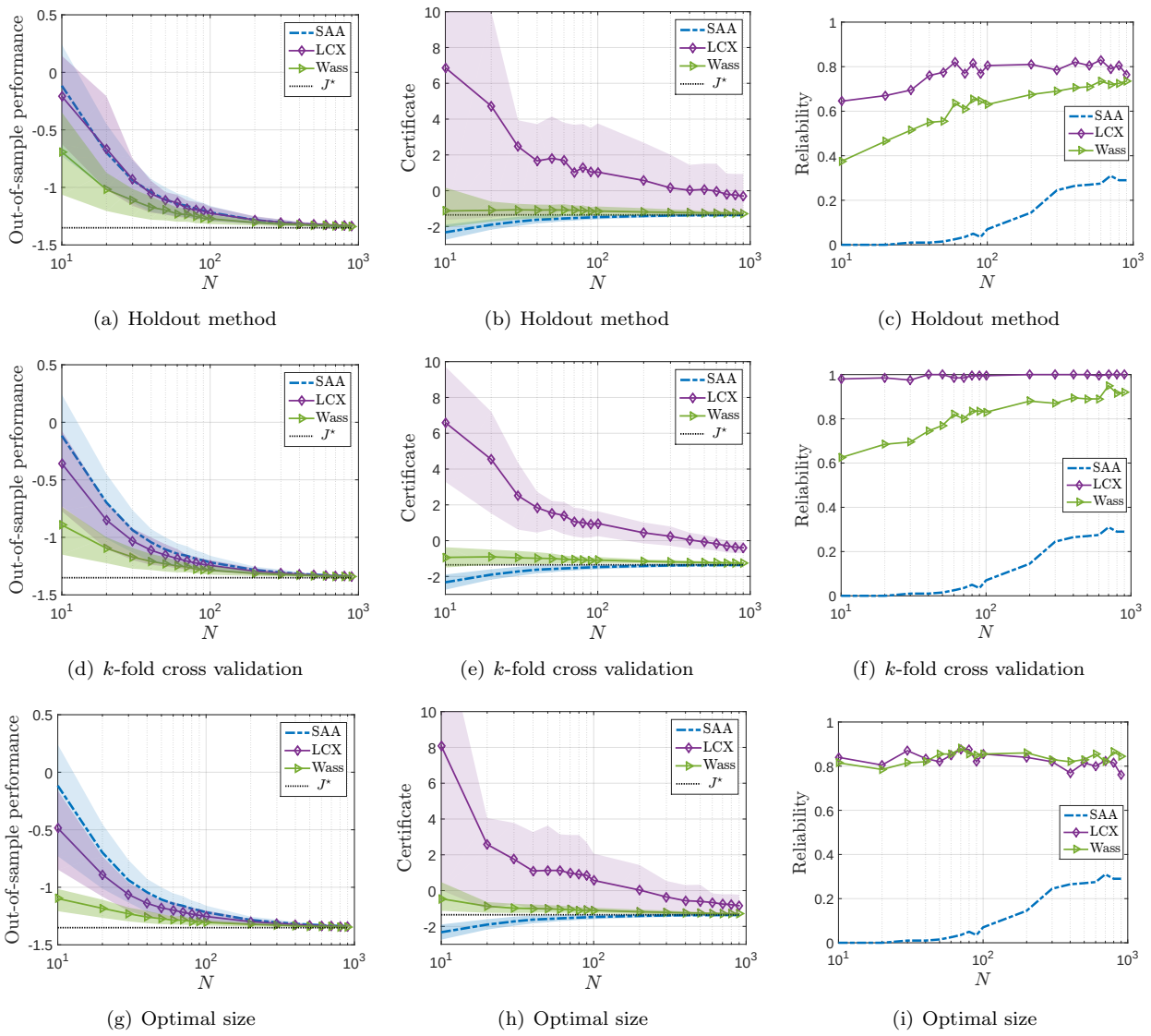
To keep the computational burden manageable, in all experiments we select the Wasserstein radius as well as the LCX size parameter from within the discrete set $\mathcal{E} = \{\varepsilon = b \cdot 10^c : b \in \{0, \ldots, 9\},\ c \in \{-3, -2, -1\}\}$ instead of $\mathbb{R}_+$. We have verified that refining or extending $\mathcal{E}$ has only a marginal impact on our results, which indicates that $\mathcal{E}$ provides a sufficiently rich approximation of $\mathbb{R}_+$.

In Figures 6(a)–6(c) the sizes of the (LCX and Wasserstein) ambiguity sets are determined via the holdout method, where 80% of the data are used for training and 20% for validation. Figure 6(a) visualizes the tube between the 20% and 80% quantiles (shaded areas) as well as the mean value (solid lines) of the out-of-sample performance $J(\widehat{x}_N)$ as a function of the sample size N and based on 200 independent simulation runs, where $\widehat{x}_N$ is set to the minimizer of the SAA (blue), LCX (purple) and Wasserstein (green) problems, respectively. The constant dashed line represents the optimal value $J^\star$ of the original stochastic program (1), which is computed through an SAA problem with $N = 10^6$ samples. We observe that the Wasserstein solutions tend to be superior to the SAA and LCX solutions in terms of out-of-sample performance.

Figure 6(b) shows the optimal values $\widehat{J}_N$ of the SAA, LCX and Wasserstein problems, where the sizes of the ambiguity sets are chosen via the holdout method. Unlike Figure 6(a), Figure 6(b) thus reports *in-sample* estimates of the achievable portfolio performance. As expected, the SAA approach is over-optimistic due to the optimizer's curse, while the LCX and Wasserstein approaches err on the side of caution. All three methods are known to enjoy asymptotic consistency, which is in agreement with all in-sample and out-of-sample results.

Figure 6(c) visualizes the reliability of the different performance certificates, that is, the empirical probability of the event $J(\widehat{x}_N) \leq \widehat{J}_N$ evaluated over 200 independent simulation runs. Here, $\widehat{x}_N$ represents either an optimal portfolio of the SAA, LCX or Wasserstein problems, while $\widehat{J}_N$ denotes the corresponding optimal value. The optimal SAA portfolios display a disappointing out-of-sample performance relative to the optimistically biased mimimum of the SAA problem—particularly when the training data is scarce. In contrast, the out-of-sample performance of the optimal LCX and Wasserstein portfolios often undershoots $\widehat{J}_N$.

[4]Much like worst-case expectations over Wasserstein balls, worst-case expectations over LCX ambiguity sets can be reformulated as finite convex programs whenever the underlying loss function represents a pointwise maximum of K concave component functions. Unlike problem (11) in Theorem 4.2, however, the resulting convex program scales exponentially with K.



(a) Holdout method (b) Holdout method (c) Holdout method

(d) k-fold cross validation (e) k-fold cross validation (f) k-fold cross validation

(g) Optimal size (h) Optimal size (i) Optimal size

FIGURE 6. Out-of-sample performance $J(\widehat{x}_N)$, certificate $\widehat{J}_N$, and certificate reliability $\mathbb{P}^N\big[J(\widehat{x}_N) \leq \widehat{J}_N\big]$ for the performance-driven SAA, LCX and Wasserstein solutions as a function of N

Figures 6(d)–6(f) show the same graphs as Figures 6(a)–6(c), but now the sizes of the ambiguity sets are determined via k-fold cross validation with $k = 5$. In this case, the out-of-sample performance of both distributionally robust methods improves slightly, while the corresponding certificates and their reliabilities increase significantly with respect to the naïve holdout method. However, these improvements come at the expense of a k-fold increase in the computational cost.

One could think of numerous other statistical methods to select the size of the Wasserstein ambiguity set. As discussed above, however, if the ultimate goal is to minimize the out-of-sample performance of $\widehat{x}_N(\varepsilon)$, then the best possible choice is $\varepsilon = \widehat{\varepsilon}_N^{\,\mathrm{opt}}$. Similarly, one can construct a size parameter for the LCX ambiguity set that leads to the best possible out-of-sample performance of any LCX solution. We emphasize that these optimal Wasserstein radii and LCX size parameters are not available in practice because computing $J(\widehat{x}_N(\varepsilon))$ requires knowledge of the data-generating distribution. In our experiments we evaluate $J(\widehat{x}_N(\varepsilon))$ to high

accuracy for every fixed $\varepsilon \in \mathcal{E}$ using $2 \cdot 10^5$ validation samples, which are independent from the (much fewer) training samples used to compute $\widehat{x}_N(\varepsilon)$. Figures 6(g)–6(i) show the same graphs as Figures 6(a)–6(c) for optimally sized ambiguity sets. By construction, no method for sizing the Wasserstein or LCX ambiguity sets can result in a better out-of-sample performance, respectively. In this sense, the graphs in Figure 6(g) capture the fundamental limitations of the different distributionally robust schemes.

7.2.C. **Portfolios Driven by Reliability**

In Section 7.2.B the Wasserstein radii and LCX size parameters were calibrated with the goal to achieve the best out-of-sample performance. Figures 6(c), 6(f) and 6(i) reveal, however, that by optimizing the out-of-sample performance one may sacrifice reliability. An alternative objective more in line with the general philosophy of Section 2 would be to choose Wasserstein radii that guarantee a prescribed reliability level. Thus, for a given $\beta \in [0,1]$ we should find the smallest Wasserstein radius $\varepsilon \geq 0$ for which the optimal value $\widehat{J}_N(\varepsilon)$ of (27) provides an upper $1-\beta$ confidence bound on the out-of-sample performance $J(\widehat{x}_N(\varepsilon))$ of its optimal solution. As the true distribution $\mathbb{P}$ is unknown, however, the optimal Wasserstein radius corresponding to a given β cannot be computed exactly. Instead, we must derive an estimator $\widehat{\varepsilon}_N^\beta$ that depends on the training data. We construct $\widehat{\varepsilon}_N^\beta$ and the corresponding reliability-driven portfolio via bootstrapping as follows:

(1) Construct k resamples of size N (with replacement) from the original training dataset. It is well known that, as N grows, the probability that any fixed training data point appears in a particular resample converges to $\frac{e-1}{e} \approx \frac{2}{3}$. Thus, about $\frac{N}{3}$ training samples are absent from any resample. We collect all unused samples in a validation dataset.

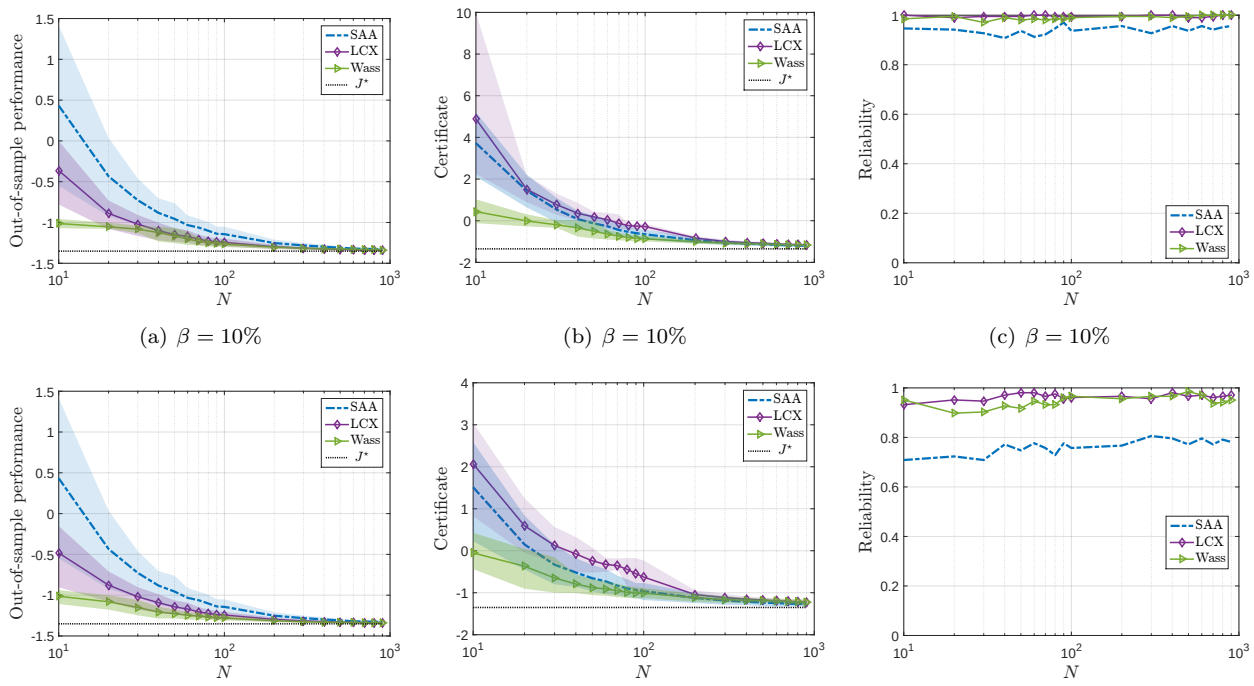
(2) For each resample $\kappa = 1, \ldots, k$ and $\varepsilon \geq 0$, solve problem (27) using the Wasserstein ball of radius ε around the empirical distribution $\widehat{\mathbb{P}}_N^\kappa$ on the κ-th resample. The resulting optimal decision and optimal value are denoted as $\widehat{x}_N^\kappa(\varepsilon)$ and $\widehat{J}_N^\kappa(\varepsilon)$, respectively. Next, estimate the out-of-sample performance $J(\widehat{x}_N^\kappa(\varepsilon))$ of $\widehat{x}_N^\kappa(\varepsilon)$ using the sample average over the κ-th validation dataset.

(3) Set $\widehat{\varepsilon}_N^\beta$ to the smallest $\varepsilon \geq 0$ so that the certificate $\widehat{J}_N^\kappa(\varepsilon)$ exceeds the estimate of $J(\widehat{x}_N^\kappa(\varepsilon))$ in at least $(1-\beta) \times k$ different resamples.

(4) Compute the data-driven portfolio $\widehat{x}_N = \widehat{x}_N(\widehat{\varepsilon}_N^\beta)$ and the corresponding certificate $\widehat{J}_N = \widehat{J}_N(\widehat{\varepsilon}_N^\beta)$ using the original training dataset.

As in Section 7.2.B, we compare the Wasserstein approach with the LCX and SAA approaches. Specifically, by using bootstrapping, we calibrate the size of the LCX ambiguity set so as to guarantee a desired reliability level $1-\beta$. The SAA problem, on the other hand, has no free parameter that can be tuned to meet a prescribed reliability target. Nevertheless, we can construct a meaningful certificate of the form $\widehat{J}_N(\Delta) := \widehat{J}_{\mathrm{SAA}} + \Delta$ for the SAA portfolio by adding a non-negative constant to the optimal value of the SAA problem. Our aim is to find the smallest offset $\Delta \geq 0$ with the property that $\widehat{J}_N(\Delta)$ provides an upper $1-\beta$ confidence bound on the out-of-sample performance $J(\widehat{x}_{\mathrm{SAA}})$ of the optimal SAA portfolio $\widehat{x}_{\mathrm{SAA}}$. The optimal offset corresponding to a given β cannot be computed exactly. Instead, we must derive an estimator $\widehat{\Delta}_N^\beta$ that depends on the training data. Such an estimator can be found through a simple variant of the above bootstrapping procedure.

In all experiments we set the number of resamples to $k = 50$. Figures 7(a)–7(c) visualize the out-of-sample performance, the certificate and the empirical reliability of the reliability-driven portfolios obtained with the SAA, LCX and Wasserstein approaches, respectively, for the reliability target $1-\beta = 90\%$ and based on 200 independent simulation runs. Figures 7(d)–7(f) show the same graphs as Figures 7(a)–7(c) but for the reliability target $1-\beta = 75\%$. We observe that the new SAA certificate now overestimates the true optimal value of the portfolio problem. Moreover, while the empirical reliability of the SAA solution now closely matches the desired reliability target, the empirical reliabilities of the LCX and Wasserstein



(a) $\beta = 10\%$ (b) $\beta = 10\%$ (c) $\beta = 10\%$

(d) $\beta = 25\%$ (e) $\beta = 25\%$ (f) $\beta = 25\%$

FIGURE 7. Out-of-sample performance $J(\widehat{x}_N)$, certificate $\widehat{J}_N$, and certificate reliability $\mathbb{P}^N\left[J(\widehat{x}_N) \leq \widehat{J}_N\right]$ for the reliability-driven SAA, LCX and Wasserstein portfolios as a function of N

solutions are similar but noticeably exceed the prescribed reliability threshold. A possible explanation for this phenomenon is that the k resamples generated by the bootstrapping algorithm are not independent, which may give rise to a systematic bias in estimating the Wasserstein radii required for the desired reliability levels.

7.2.D. **Impact of the Sample Size on the Wasserstein Radius**

It is instructive to analyze the dependence of the Wasserstein radii on the sample size N for different data-driven schemes. As for the performance-driven portfolios from Section 7.2.B, Figure 8 depicts the best possible Wasserstein radius $\widehat{\varepsilon}_N^{\mathrm{opt}}$ as well as the Wasserstein radii $\widehat{\varepsilon}_N^{\mathrm{hm}}$ and $\widehat{\varepsilon}_N^{\mathrm{cv}}$ obtained by the holdout method and via k-fold cross validation, respectively. As for the reliability-driven portfolios from Section 7.2.C, Figure 8 further depicts the Wasserstein radii $\widehat{\varepsilon}_N^{\beta}$, for $\beta \in \{10\%, 25\%\}$, obtained by bootstrapping. All results are averaged across 200 independent simulation runs. As expected from Theorem 3.6, all Wasserstein radii tend to zero as N increases. Moreover, the convergence rate is approximately equal to $N^{-\frac{1}{2}}$. This rate is likely to be optimal. Indeed, if $\mathbb{X}$ is a singleton, then every quantile of the sample average estimator $\widehat{J}_{\mathrm{SAA}}$ converges to $J^\star$ at rate $N^{-\frac{1}{2}}$ due to the central limit theorem. Thus, if $\widehat{\varepsilon}_N = o(N^{-\frac{1}{2}})$, then $\widehat{J}_N$ also converges to $J^\star$ at leading order $N^{-\frac{1}{2}}$ by Theorem 6.3, which applies as the loss function is convex. This indicates that the a priori rate $N^{-\frac{1}{m}}$ suggested by Theorem 3.4 is too pessimistic in practice.

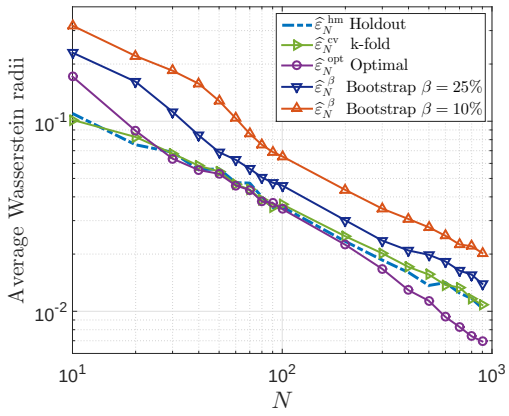


FIGURE 8. Optimal performance-driven Wasserstein radius $\widehat{\varepsilon}_N^{\mathrm{opt}}$ and its estimates $\widehat{\varepsilon}_N^{\mathrm{hm}}$ and $\widehat{\varepsilon}_N^{\mathrm{cv}}$ obtained via the holdout method and k-fold cross validation, respectively, as well as the reliability-driven Wasserstein radius $\widehat{\varepsilon}_N^{\beta}$ for $\beta \in \{10\%, 25\%\}$ obtained via bootstrapping

7.3. **Simulation Results: Uncertainty Quantification**

Investors often wish to determine the probability that a given portfolio will outperform various benchmark indices or assets. Our results on uncertainty quantification developed in Section 5.2 enable us to compute this probability in a meaningful way—solely on the basis of the training dataset.

Assume for example that we wish to quantify the probability that any data-driven portfolio $\widehat{x}_N$ outperforms the three most risky assets in the market *jointly*. Thus, we should compute the probability of the closed polytope

$$\widehat{\mathbb{A}} = \Big\{ \xi \in \mathbb{R}^m \;:\; \langle \widehat{x}_N, \xi \rangle \geq \xi_i \;\forall i = 8, 9, 10 \Big\}.$$

As the true distribution $\mathbb{P}$ is unknown, the probability $\mathbb{P}[\xi \in \widehat{\mathbb{A}}]$ cannot be evaluated exactly. Note that $\widehat{\mathbb{A}}$ as well as $\mathbb{P}[\xi \in \widehat{\mathbb{A}}]$ constitute random objects that depend on $\widehat{x}_N$ and thus on the training data. Using the *same* training dataset that was used to compute $\widehat{x}_N$, however, we may estimate $\mathbb{P}[\xi \in \widehat{\mathbb{A}}]$ from above and below by

$$\sup_{\mathbb{Q} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{Q}\left[\xi \in \widehat{\mathbb{A}}\right] \qquad \text{and} \qquad \inf_{\mathbb{Q} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{Q}\left[\xi \in \widehat{\mathbb{A}}\right] = 1 - \sup_{\mathbb{Q} \in \mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{Q}\left[\xi \notin \widehat{\mathbb{A}}\right],$$

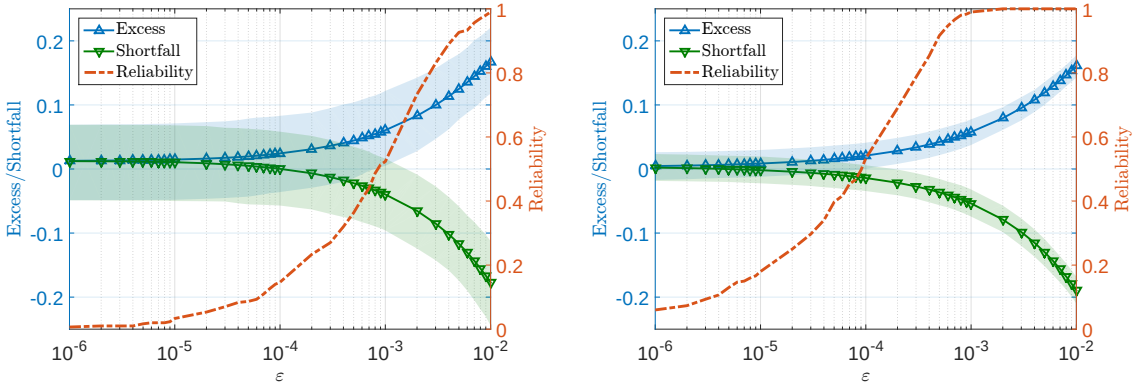
respectively. Indeed, recall that the true data-generating probability distribution resides in the Wasserstein ball of radius $\varepsilon_N(\beta)$ defined in (8) with probability $1 - \beta$. Therefore, we have

$$\begin{aligned} 1 - \beta \leq \mathbb{P}^N\Big[\widehat{\Xi}_N : \mathbb{P} \in \mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)\Big] &\leq \mathbb{P}^N\Big[\widehat{\Xi}_N : \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)} \mathbb{Q}[\mathbb{A}] \geq \mathbb{P}[\mathbb{A}] \quad \forall \mathbb{A} \in \mathfrak{B}(\Xi)\Big] \\ &= \mathbb{P}^N\Big[\widehat{\Xi}_N : \inf_{\mathbb{A} \in \mathfrak{B}(\Xi)} \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)} \mathbb{Q}[\mathbb{A}] - \mathbb{P}[\mathbb{A}] \geq 0\Big], \end{aligned}$$

where $\mathfrak{B}(\Xi)$ denotes the set of all Borel subsets of Ξ. The data-dependent set $\widehat{\mathbb{A}}_N$ can now be viewed as a (measurable) mapping from $\widehat{\Xi}_N$ to the subsets in $\mathfrak{B}(\Xi)$. The above inequality then implies

$$\mathbb{P}^N\Big[\widehat{\Xi}_N : \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)} \mathbb{Q}[\widehat{\mathbb{A}}_N] - \mathbb{P}[\widehat{\mathbb{A}}_N] \geq 0\Big] \geq 1 - \beta.$$

Thus, $\sup\{\mathbb{Q}[\widehat{\mathbb{A}}_N] : \mathbb{Q} \in \mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)\}$ provides indeed an upper bound on $\mathbb{P}[\widehat{\mathbb{A}}_N]$ with confidence $1 - \beta$. Similarly, one can show that $\inf\{\mathbb{Q}[\widehat{\mathbb{A}}_N] : \mathbb{Q} \in \mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)\}$ provides a lower confidence bound on $\mathbb{P}[\widehat{\mathbb{A}}_N]$.



(a) $N = 30$ (b) $N = 300$

FIGURE 9. Excess $\widehat{J}_N^+(\varepsilon) - \mathbb{P}[\widehat{\mathbb{A}}]$ and shortfall $\widehat{J}_N^-(\varepsilon) - \mathbb{P}[\widehat{\mathbb{A}}]$ (solid lines, left axis) as well as reliability $\mathbb{P}^N[\widehat{J}_N^-(\varepsilon) \leq \mathbb{P}[\widehat{\mathbb{A}}] \leq \widehat{J}_N^+(\varepsilon)]$ (dashed lines, right axis) as a function of ε

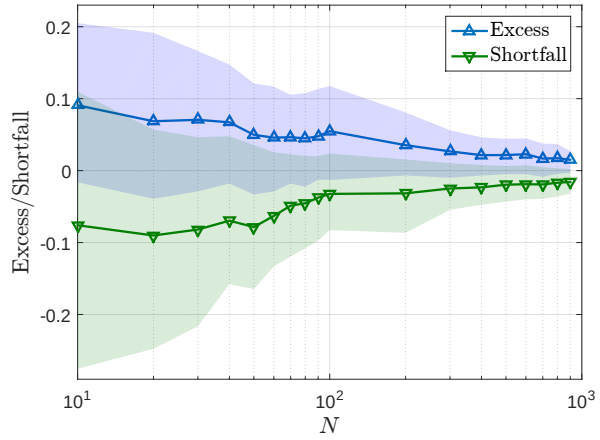
The upper confidence bound can be computed by solving the linear program (17a). Replacing $\widehat{\mathbb{A}}$ with its interior in the lower confidence bound leads to another (potentially weaker) lower bound that can be computed by solving the linear program (17b). We denote these computable bounds by $\widehat{J}_N^+(\varepsilon)$ and $\widehat{J}_N^-(\varepsilon)$, respectively. In all subsequent experiments $\widehat{x}_N$ is set to a solution of the distributionally robust program (27) calibrated via k-fold cross validation as described in Section 7.2.B.

7.3.A. **Impact of the Wasserstein Radius**

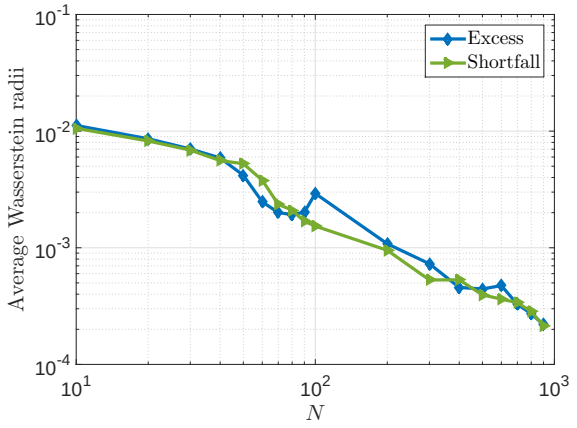
As $\widehat{J}_N^+(\varepsilon)$ and $\widehat{J}_N^-(\varepsilon)$ estimate a random target $\mathbb{P}[\widehat{\mathbb{A}}]$, it makes sense to filter out the randomness of the target and to study only the differences $\widehat{J}_N^+(\varepsilon) - \mathbb{P}[\widehat{\mathbb{A}}]$ and $\widehat{J}_N^-(\varepsilon) - \mathbb{P}[\widehat{\mathbb{A}}]$. Figures 9(a) and 9(b) visualize the empirical mean (solid lines) as well as the tube between the empirical 20% and 80% quantiles (shaded areas) of these differences as a function of the Wasserstein radius ε, based on 200 training datasets of cardinality $N = 30$ and $N = 300$, respectively. Figure 9 also shows the empirical reliability of the bounds (dashed lines), that is, the empirical probability of the event $\widehat{J}_N^-(\varepsilon) \leq \mathbb{P}[\widehat{\mathbb{A}}] \leq \widehat{J}_N^+(\varepsilon)$. Note that the reliability drops to 0 for $\varepsilon = 0$, in which case both $\widehat{J}_N^+(0)$ and $\widehat{J}_N^-(0)$ coincide with the SAA estimator for $\mathbb{P}[\widehat{\mathbb{A}}]$. Moreover, at $\varepsilon = 0$ the set $\widehat{\mathbb{A}}$ is constructed from the SAA portfolio $\widehat{x}_N$, whose performance is overestimated on the training dataset. Thus, the SAA estimator for $\mathbb{P}[\widehat{\mathbb{A}}]$, which is evaluated using the same training dataset, is positively biased. For $\varepsilon > 0$, finally, the reliability increases as the shaded confidence intervals move away from 0.

7.3.B. **Impact of the Sample Size**

We propose a variant of the k-fold cross validation procedure for selecting ε in uncertainty quantification. Partition $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ into k subsets and repeat the following holdout method k times. Select one of the subsets as the validation set of size N_V and merge the remaining $k-1$ subsets to a training dataset of size $N_T = N - N_V$. Use the validation set to compute the SAA estimator of $\mathbb{P}[\widehat{\mathbb{A}}]$, and use the training dataset to compute $\widehat{J}_{N_T}^+(\varepsilon)$ for a large but finite number of candidate radii ε. Set $\widehat{\varepsilon}_N^{\mathrm{hm}}$ to the smallest candidate radius for which the SAA estimator of $\mathbb{P}[\widehat{\mathbb{A}}]$ is not larger than $\widehat{J}_{N_T}^+(\varepsilon)$. Next, set $\widehat{\varepsilon}_N^{\mathrm{cv}}$ to the average of the Wasserstein radii obtained from the k holdout runs, and report $\widehat{J}_N^+ = \widehat{J}_N^+(\widehat{\varepsilon}_N^{\mathrm{cv}})$ as the data-driven upper bound on $\mathbb{P}[\widehat{\mathbb{A}}]$. The data-driven lower bound $\widehat{J}_N^-$ is constructed analogously in the obvious way.



(a) Excess $\widehat{J}_N^+ - \mathbb{P}[\widehat{\mathbb{A}}]$ and shortfall $\widehat{J}_N^- - \mathbb{P}[\widehat{\mathbb{A}}]$ of the data-driven confidence bounds for $\mathbb{P}[\widehat{\mathbb{A}}]$



(b) Data-driven Wasserstein radius $\widehat{\varepsilon}_N^{\mathrm{cv}}$ obtained via k-fold cross validation

FIGURE 10. Dependence of the confidence bounds and the Wasserstein radius on N

Figure 10(a) visualizes the empirical means (solid lines) as well as the tubes between the empirical 20% and 80% quantiles (shaded areas) of $\widehat{J}_N^+ - \mathbb{P}[\widehat{\mathbb{A}}]$ and $\widehat{J}_N^- - \mathbb{P}[\widehat{\mathbb{A}}]$ as a function of the sample size N, based on 300 independent training datasets. As expected, the confidence intervals shrink and converge to 0 as N increases. We emphasize that $\widehat{J}_N^+$ and $\widehat{J}_N^-$ are computed solely on the basis of N training samples, whereas the computation of $\mathbb{P}[\widehat{\mathbb{A}}]$ necessitates a much larger dataset, particularly if $\widehat{\mathbb{A}}$ constitutes a rare event.

Figure 10(b) shows the Wasserstein radius $\widehat{\varepsilon}_N^{\mathrm{cv}}$ obtained via k-fold cross validation (both for $\widehat{J}_N^+$ and $\widehat{J}_N^-$). As usual, all results are averaged across 300 independent simulation runs. A comparison with Figure 8 reveals that the data-driven Wasserstein radii in uncertainty quantification display a similar but faster polynomial decay than in portfolio optimization. We conjecture that this is due to the absence of decisions, which implies that uncertainty quantification is less susceptible to the optimizer's curse. Thus, nature (*i.e.*, the fictitious adversary choosing the distribution in the ambiguity set) only has to compensate for noise but not for bias. A smaller Wasserstein radius seems to be sufficient for this purpose.

Acknowledgments. We thank Soroosh Shafieezadeh Abadeh for helping us with the numerical experiments. The authors are grateful to Vishal Gupta, Ruiwei Jiang and Nathan Kallus for their valuable comments. This research was supported by the Swiss National Science Foundation under Grant BSCGI0_157733.

APPENDIX A.

The following technical lemma on the pointwise approximation of an upper semicontinuous function by a non-increasing sequence of Lipschitz continuous majorants strengthens [31, Theorem 4.2], which focuses on bounded domains and continuous (but not necessarily Lipschitz continuous) majorants.

Lemma A.1. *If $h : \Xi \to \mathbb{R}$ is upper semicontinuous and satisfies $h(\xi) \leq L(1 + \|\xi\|)$ for some $L \geq 0$, then there exists a non-increasing sequence of Lipschitz continuous functions that converge pointwise to h on Ξ.*

Proof. The proof is constructive. Define the functions

$$h_k(\xi) = \sup_{\xi' \in \Xi} h(\xi') - kL\|\xi - \xi'\|, \quad k \in \mathbb{N},$$

where L is the linear growth rate of h. Note that by construction $h_k(\xi) \leq L(1 + \|\xi\|)$. As $\xi' = \xi$ is feasible in the maximization problem defining $h_k(\xi)$, we have $h_k(\xi) \geq h(\xi)$ for all $\xi \in \Xi$ and $k \in \mathbb{N}$. Moreover, $h_k(\xi)$ is Lipschitz continuous with Lipschitz constant kL (as $h_k(\xi)$ constitutes a supremum of norm functions with this property). Given any $\xi \in \Xi$, it remains to be shown that $\lim_{k\to\infty} h_k(\xi) = h(\xi)$. Thus, choose $\xi'_k \in \Xi$ with

$$h_k(\xi) = \sup_{\xi' \in \Xi} h(\xi') - kL\|\xi - \xi'\| \leq h(\xi'_k) - kL\|\xi - \xi'_k\| + \frac{1}{k}.$$

We first show that ξ_k converges to ξ as k tends to ∞. Indeed, we have

$$\begin{aligned} h(\xi) \leq h_k(\xi) &\leq h(\xi'_k) - kL\|\xi - \xi'_k\| + \frac{1}{k} \leq L(1 + \|\xi'_k\|) - kL\|\xi - \xi'_k\| + \frac{1}{k} \\ &\leq L(1 + \|\xi - \xi'_k\| + \|\xi\|) - kL\|\xi - \xi'_k\| + \frac{1}{k} = L(1 + \|\xi\|) + \frac{1}{k} - (k-1)L\|\xi - \xi'_k\|, \end{aligned}$$

which implies

$$\|\xi - \xi'_k\| \leq \frac{1}{L(k-1)} \left(h(\xi) - L(1 + \|\xi\|) - \frac{1}{k} \right),$$

that is, $\|\xi - \xi'_k\| \to 0$ as $k \to \infty$. Therefore, we find

$$h(\xi) \leq \lim_{k\to\infty} h_k(\xi) \leq \limsup_{k\to\infty} h(\xi'_k) - kL\|\xi - \xi'_k\| + \frac{1}{k} \leq \limsup_{k\to\infty} h(\xi'_k) \leq h(\xi),$$

where the last inequality is due to the upper semicontinuity of h. This concludes the proof. □

References

[1] A. Ben-Tal, D. den Hertog, and J.-P. Vial, *Deriving robust counterparts of nonlinear uncertain inequalities*, Mathematical Programming, 149 (2015), pp. 265–299.

[2] A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen, *Robust solutions of optimization problems affected by uncertain probabilities*, Management Science, 59 (2013), pp. 341–357.

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

[4] D. P. Bertsekas, *Convex Optimization Theory*, Athena Scientific, 2009.

[5] ———, *Convex Optimization Algorithms*, Athena Scientific, 2015.

[6] D. Bertsimas, X. V. Doan, K. Natarajan, and C.-P. Teo, *Models for minimax stochastic linear optimization problems with risk aversion*, Mathematics of Operations Research, 35 (2010), pp. 580–602.

[7] D. Bertsimas, V. Gupta, and N. Kallus, *Robust SAA*. Available at arXiv:1408.4445, 2014.

[8] D. Bertsimas and I. Popescu, *On the relation between option and stock prices: A convex optimization approach*, Operations Research, 50 (2002), pp. 358–374.

[9] D. Bertsimas and M. Sim, *The price of robustness*, Operations Research, 52 (2004), pp. 35–53.

[10] E. Boissard, *Simple bounds for convergence of empirical and occupation measures in 1-Wasserstein distance*, Electronic Journal of Probability, 16 (2011), pp. 2296–2333.

[11] F. Bolley, A. Guillin, and C. Villani, *Quantitative concentration inequalities for empirical measures on non-compact spaces*, Probability Theory and Related Fields, 137 (2007), pp. 541–593.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2009.

[13] C. Brownlees, E. Joly, and G. Lugosi, *Empirical risk minimization for heavy-tailed losses*, The Annals of Statistics, 43 (2015), pp. 2507–2536.

[14] G. C. Calafiore, *Ambiguous risk measures and optimal robust portfolios*, SIAM Journal on Optimization, 18 (2007), pp. 853–877.

[15] O. Catoni, *Challenging the empirical mean and empirical variance: A deviation study*, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 48 (2012), pp. 1148–1185.

[16] N. Chehrazi and T. A. Weber, *Monotone approximation of decision problems*, Operations Research, 58 (2010), pp. 1158–1177.

[17] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, et al., *Tests of goodness of fit based on the l_2-Wasserstein distance*, The Annals of Statistics, 27 (1999), pp. 1230–1239.

[18] E. Delage and Y. Ye, *Distributionally robust optimization under moment uncertainty with application to data-driven problems*, Operations Research, 58 (2010), pp. 595–612.

[19] L. El Ghaoui, M. Oks, and F. Oustry, *Worst-case Value-at-Risk and robust portfolio optimization: A conic programming approach*, Operations Research, 51 (2003), pp. 543–556.
[20] E. Erdoğan and G. Iyengar, *Ambiguous chance constrained problems and robust optimization*, Mathematical Programming, 107 (2006), pp. 37–61.
[21] N. Fournier and A. Guillin, *On the rate of convergence in Wasserstein distance of the empirical measure*, Probability Theory and Related Fields, (2014), pp. 1–32.
[22] J. Goh and M. Sim, *Distributionally robust optimization and its tractable approximations*, Operations Research, 58 (2010), pp. 902–917.
[23] G. A. Hanasusanto and D. Kuhn, *Robust data-driven dynamic programming*, in Advances in Neural Information Processing Systems, 2013, pp. 827–835.
[24] G. A. Hanasusanto, D. Kuhn, and W. Wiesemann, *A comment on "Computational complexity of stochastic programming problems"*, Mathematical Programming, (2016), pp. 557–569.
[25] Z. Hu and L. J. Hong, *Kullback-Leibler divergence constrained distributionally robust optimization*. Available at Optimization Online, 2013.
[26] Z. Hu, L. J. Hong, and A. M.-C. So, *Ambiguous probabilistic programs*. Available at Optimization Online, 2013.
[27] R. Jiang and Y. Guan, *Data-driven chance constrained stochastic program*, Mathematical Programming, 158 (2016), pp. 291–327.
[28] O. Kallenberg, *Foundations of Modern Probability*, Probability and its Applications (New York), Springer-Verlag, New York, 1997.
[29] L. V. Kantorovich and G. S. Rubinshtein, *On a space of totally additive functions*, Vestnik Leningradskogo Universiteta, 13 (1958), pp. 52–59.
[30] S. Lang, *Real and Functional Analysis*, Springer-Verlag, third ed., 1993.
[31] J. Mashreghi, *Representation Theorems in Hardy Spaces*, Cambridge University Press, 2009.
[32] S. Mehrotra and H. Zhang, *Models and algorithms for distributionally robust least squares problems*, Mathematical Programming, 146 (2014), pp. 123–141.
[33] A. Müller, *Integral probability metrics and their generating classes of functions*, Advances in Applied Probability, (1997), pp. 429–443.
[34] K. Natarajan, M. Sim, and J. Uichanco, *Tractable robust expected utility and risk models for portfolio optimization*, Mathematical Finance, 20 (2010), pp. 695–731.
[35] N. Parikh and S. Boyd, *Block splitting for distributed optimization*, Mathematical Programming Computation, 6 (2014), pp. 77–102.
[36] G. C. Pflug and A. Pichler, *Multistage Sochastic Optimization*, Springer, 2014.
[37] G. C. Pflug, A. Pichler, and D. Wozabal, *The 1/N investment strategy is optimal under high model ambiguity*, Journal of Banking & Finance, 36 (2012), pp. 410 – 417.
[38] G. C. Pflug and D. Wozabal, *Ambiguity in portfolio selection*, Quantitative Finance, 7 (2007), pp. 435–442.
[39] K. Postek, D. den Hertog, and B. Melenberg, *Tractable counterparts of distributionally robust constraints on risk measures*, forthcoming in SIAM Review, (2016).
[40] A. Ramdas, N. Garcia, and M. Cuturi, *On Wasserstein two sample testing and related families of nonparametric tests*. Available at arXiv:1509.02237, 2015.
[41] R. T. Rockafellar and S. Uryasev, *Optimization of conditional Value-at-Risk*, Journal of Risk, 2 (2000), pp. 21–42.
[42] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, 2010.
[43] H. E. Scarf, *A min-max solution of an inventory problem*, in Studies in the Mathematical Theory of Inventory and Production, K. J. Arrow, S. Karlin, and H. E. Scarf, eds., Stanford University Press, 1958, pp. 201–209.
[44] A. Shapiro, *On duality theory of conic linear problems*, in Semi-Infinite Programming, M. A. Goberna and M. A. López, eds., Kluwer Academic Publishers, 2001, pp. 135–165.
[45] ———, *Distributionally robust stochastic programming*. Available at Optimization Online, 2015.
[46] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming*, SIAM, second ed., 2014.
[47] A. Shapiro and A. Nemirovski, *On complexity of stochastic programming problems*, in Continuous Optimization, Springer, New York, 2005, pp. 111–146.
[48] J. E. Smith and R. L. Winkler, *The optimizers curse: Skepticism and postdecision surprise in decision analysis*, Management Science, 52 (2006), pp. 311–322.
[49] V. N. Vapnik, *Statistical Learning Theory*, Wiley, 1998.
[50] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2003.
[51] W. Wiesemann, D. Kuhn, and M. Sim, *Distributionally robust convex optimization*, Operations Research, 62 (2014), pp. 1358–1376.

[52] D. Wozabal, *A framework for optimization under ambiguity*, Annals of Operations Research, 193 (2012), pp. 21–47.

[53] ———, *Robustifying convex risk measures for linear portfolios: A nonparametric approach*, Operations Research, 62 (2014), pp. 1302–1315.

[54] C. Zhao, *Data-Driven Risk-Averse Stochastic Program and Renewable Energy Integration*, PhD thesis, University of Florida, 2014.