Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning

Daniel Kuhn
Risk Analytics and Optimization Chair, EPFL, Lausanne 1015, Switzerland, daniel.kuhn@epfl.ch

Peyman Mohajerin Esfahani
Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands,
P.MohajerinEsfahani@tudelft.nl

Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh
Risk Analytics and Optimization Chair, EPFL, Lausanne 1015, Switzerland,
{viet-anh.nguyen@epfl.ch, soroosh.shafiee@epfl.ch}

Abstract Many decision problems in science, engineering and economics are affected by uncertain parameters whose distribution is only indirectly observable through samples. The goal of data-driven decision-making is to learn a decision from finitely many training samples that will perform well on unseen test samples. This learning task is difficult even if all training and test samples are drawn from the same distribution—especially if the dimension of the uncertainty is large relative to the training sample size. Wasserstein distributionally robust optimization seeks data-driven decisions that perform well under the most adverse distribution within a certain Wasserstein distance from a nominal distribution constructed from the training samples. In this tutorial we will argue that this approach has many conceptual and computational benefits. Most prominently, the optimal decisions can often be computed by solving tractable convex optimization problems, and they enjoy rigorous out-of-sample and asymptotic consistency guarantees. We will also show that Wasserstein distributionally robust optimization has interesting ramifications for statistical learning and motivates new approaches for fundamental learning tasks such as classification, regression, maximum likelihood estimation or minimum mean square error estimation, among others.

Keywords distributionally robust optimization; data-driven optimization; Wasserstein distance; optimizer's curse; machine learning; regularization

1. Introduction

We consider a decision problem under uncertainty, where each admissible decision results in an uncertain loss that is modeled by a measurable extended real-valued *loss function* $\ell(\xi)$. We assume that the random vector $\xi \in \mathbb{R}^m$ captures all decision-relevant risk factors and is governed by a probability distribution $\mathbb{P}$. The feasible set of all available loss functions is denoted by $\mathcal{L}$. The *risk* of a decision $\ell \in \mathcal{L}$ is defined as the expected loss under $\mathbb{P}$, that is,

$$\mathcal{R}(\mathbb{P}, \ell) = \mathbb{E}^{\mathbb{P}}[\ell(\xi)], \tag{1}$$

and the *optimal risk* is defined as the risk of the least risky admissible loss function, that is,

$$\mathcal{R}(\mathbb{P}, \mathcal{L}) = \inf_{\ell \in \mathcal{L}} \mathcal{R}(\mathbb{P}, \ell). \tag{2}$$

To ensure that the expectations in (1) and (2) are defined for all measurable loss functions, we set $\mathbb{E}^{\mathbb{P}}[\ell(\xi)] = \infty$ whenever the expectations of the positive and negative parts of $\ell(\xi)$ are both infinite. This convention means that infeasibility trumps unboundedness.

In most real decision-making situations, the distribution $\mathbb{P}$ is fundamentally unknown. However, $\mathbb{P}$ may be indirectly observable through *training samples* $\widehat{\xi}_i$, $i \in \{1, \ldots, N\}$, drawn independently from $\mathbb{P}$. In addition, some structural properties of $\mathbb{P}$ may be known. For example, if ξ represents a vector of uncertain prices, then $\mathbb{P}$ must be supported on the nonnegative orthant $\mathbb{R}^m_+$. Alternatively, $\mathbb{P}$ may be known to display certain symmetry or unimodality properties, or it may even be known to belong to some parametric distribution family.

If the distribution $\mathbb{P}$ is unknown, we lack an important input parameter for the risk evaluation problem (1) and the decision problem (2). In this case, the unknown true distribution $\mathbb{P}$ could be replaced with a *nominal distribution* $\widehat{\mathbb{P}}_N$ estimated from the N training samples. Note that unlike $\mathbb{P}$, the nominal distribution $\widehat{\mathbb{P}}_N$ is accessible as it is constructed from observable quantities. Therefore, the nominal risk evaluation and decision problems (that is, problems (1) and (2) with $\widehat{\mathbb{P}}_N$ instead of $\mathbb{P}$) are at least in principle solvable. The following example showcases common methods for constructing the nominal distribution $\widehat{\mathbb{P}}_N$.

Example 1 (Nominal distribution). In the remainder we will primarily work with the following *non-parametric* and *parametric* models for the nominal distribution.

(1) In the absence of any structural information, it is convenient to set $\widehat{\mathbb{P}}_N$ to the discrete *empirical distribution*, that is, the uniform distribution on the N training samples,

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

where $\delta_{\widehat{\xi}_i}$ denotes the Dirac point mass at the i^{th} training sample $\widehat{\xi}_i$.

(2) We say that $\mathbb{Q} = \mathcal{E}_g(\mu, \Sigma)$ is an *elliptical* probability distribution if it has a density function of the form $f(\xi) = C \det(\Sigma)^{-1} g((\xi - \mu)\Sigma^{-1}(\xi - \mu))$ with density generator $g(u) \geq 0$ for all $u \geq 0$, normalization constant $C > 0$, mean vector $\mu \in \mathbb{R}^m$ and covariance matrix $\Sigma \in \mathbb{S}^m_{++}$. Examples of elliptical distributions are reported in Table 3 of Appendix A. In the presence of structural information, it is often convenient to set $\widehat{\mathbb{P}}_N$ to an elliptical distribution with a structure-dependent density generator g, that is,

$$\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}, \widehat{\Sigma}), \tag{4}$$

where only the mean vector $\widehat{\mu}$ and the covariance matrix $\widehat{\Sigma}$ depend on the training samples and are constructed via maximum likelihood estimation.

As a function of the training data, the nominal distribution $\widehat{\mathbb{P}}_N$ constitutes itself a random object, which is governed by the distribution $\mathbb{P}^N$ of the N independent training samples. □

Even if the most sophisticated statistical tools are deployed, the nominal distribution $\widehat{\mathbb{P}}_N$ will invariably differ from the unknown true distribution $\mathbb{P}$ that generated the training samples. Moreover, if $\widehat{\mathbb{P}}_N$ is used instead of $\mathbb{P}$, the solutions of the risk evaluation problem (1) and the decision problem (2) are likely to inherit any estimation errors in $\widehat{\mathbb{P}}_N$. In the context of financial portfolio theory it has even been observed that estimation errors in the input parameters of an optimization problem are often amplified by the optimization [23, 59]. To make things worse, one can generally show that even if the distributional input parameters of a decision problem are unbiased, the optimization results tend to be optimistically biased. Thus, implementing the optimal decisions leads to disappointment in out-of-sample tests. In decision analysis this phenomenon is sometimes termed the *optimizer's curse* [101], and in stochastic optimization it is referred to as the *optimization bias* [27, 99].

Example 2 (Optimizer's curse). Let $\widehat{\mathbb{P}}_N$ be an unbiased estimator for $\mathbb{P}$. Thus, we have $\mathbb{E}^{\mathbb{P}^N}[\widehat{\mathbb{P}}_N] = \mathbb{P}$, where the expectation is taken with respect to the distribution $\mathbb{P}^N$ of the N independent training samples. Then, the risk of a fixed loss function $\ell \in \mathcal{L}$ satisfies

$$\mathbb{E}^{\mathbb{P}^N}\left[\mathcal{R}(\widehat{\mathbb{P}}_N, \ell)\right] = \mathbb{E}^{\mathbb{P}^N}\left[\mathbb{E}^{\widehat{\mathbb{P}}_N}[\ell(\xi)]\right] = \mathbb{E}^{\mathbb{P}}[\ell(\xi)] = \mathcal{R}(\mathbb{P}, \ell),$$

where the second equality holds because the inner expectation is linear in $\widehat{\mathbb{P}}_N$. This implies that $\mathcal{R}(\widehat{\mathbb{P}}_N, \ell)$ constitutes an unbiased estimator for the true risk $\mathcal{R}(\mathbb{P}, \ell)$. Moreover, we have

$$\mathbb{E}^{\mathbb{P}^N}\left[\mathcal{R}(\widehat{\mathbb{P}}_N, \mathcal{L})\right] = \mathbb{E}^{\mathbb{P}^N}\left[\inf_{\ell\in\mathcal{L}} \mathcal{R}(\widehat{\mathbb{P}}_N, \ell)\right] \leq \inf_{\ell\in\mathcal{L}} \mathbb{E}^{\mathbb{P}^N}\left[\mathcal{R}(\widehat{\mathbb{P}}_N, \ell)\right] = \inf_{\ell\in\mathcal{L}} \mathcal{R}(\mathbb{P}, \ell) = \mathcal{R}(\mathbb{P}, \mathcal{L}),$$

where the inequality holds because the infimum inside the expectation can adapt to $\widehat{\mathbb{P}}_N$. Hence, $\mathcal{R}(\widehat{\mathbb{P}}_N, \mathcal{L})$ constitutes an optimistically biased estimator for $\mathcal{R}(\mathbb{P}, \mathcal{L})$, *i.e.*, it underestimates the true risk. On the other hand, any optimizer $\ell^\star \in \arg\min_{\ell\in\mathcal{L}} \mathcal{R}(\widehat{\mathbb{P}}_N, \ell)$ satisfies

$$\mathcal{R}(\mathbb{P}, \ell^\star) \geq \inf_{\ell\in\mathcal{L}} \mathcal{R}(\mathbb{P}, \ell) = \mathcal{R}(\mathbb{P}, \mathcal{L}).$$

The above observations can be interpreted as follows. Someone solving the nominal decision problem *thinks* that the risk of $\ell^\star$ amounts to $\mathcal{R}(\widehat{\mathbb{P}}_N, \ell^\star) = \mathcal{R}(\widehat{\mathbb{P}}_N, \mathcal{L})$ (the in-sample risk), which is typically *smaller* than the optimal risk $\mathcal{R}(\mathbb{P}, \mathcal{L})$ attainable under full knowledge of $\mathbb{P}$. However, the *actual* risk $\mathcal{R}(\mathbb{P}, \ell^\star)$ of the optimizer $\ell^\star$ under the true distribution (the out-of-sample risk) is always *larger* than $\mathcal{R}(\mathbb{P}, \mathcal{L})$. The difference between the out-of-sample risk and the in-sample risk is termed the *post-decision disappointment*. The *optimizer's curse* refers to the observation that the post-decision disappointment is positive on average. □

In order to quantify the sensitivity of $\mathcal{R}(\mathbb{P}, \ell)$ and $\mathcal{R}(\mathbb{P}, \mathcal{L})$ with respect to the unknown true distribution $\mathbb{P}$, we must introduce a distance measure between probability distributions. As we will argue below, the Wasserstein distance is a particularly convenient choice.

Definition 1 (Wasserstein distance). For any $p \in [1, \infty)$, the type-p Wasserstein distance between two probability distributions $\mathbb{Q}$ and $\mathbb{Q}'$ on $\mathbb{R}^m$ is defined as

$$W_p(\mathbb{Q}, \mathbb{Q}') = \left(\inf_{\pi\in\Pi(\mathbb{Q},\mathbb{Q}')} \int_{\mathbb{R}^m\times\mathbb{R}^m} \|\xi - \xi'\|^p \, \pi(\mathrm{d}\xi, \mathrm{d}\xi')\right)^{\frac{1}{p}}, \tag{5}$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^m$, while $\Pi(\mathbb{Q}, \mathbb{Q}')$ denotes the set of all joint probability distributions of $\xi \in \mathbb{R}^m$ and $\xi' \in \mathbb{R}^m$ with marginals $\mathbb{Q}$ and $\mathbb{Q}'$, respectively.

One can show that the Wasserstein distance is a metric, that is, it is nonnegative, symmetric and subadditive, and it vanishes only if $\mathbb{Q} = \mathbb{Q}'$ [107, p. 94]. One can further show that $W_p(\mathbb{Q}, \mathbb{Q}')$ is finite whenever both $\mathbb{Q}$ and $\mathbb{Q}'$ have finite p^{th}-order moments [107, p. 95].

The optimal value of the optimization problem in (5) can be interpreted as the minimum cost of turning one pile of dirt represented by $\mathbb{Q}$ into another pile of dirt represented by $\mathbb{Q}'$, where the cost of moving a unit mass from ξ to ξ' amounts to $\|\xi - \xi'\|^p$. The decision variable π thus encodes a transportation plan, that is, for any measurable sets $A, B \subseteq \mathbb{R}^m$ the probability $\pi(A \times B)$ reflects the amount of mass that is moved from the source region A to the target region B. Because of this interpretation, the Wasserstein distance is often referred to as the earth mover's distance in statistics and computer science [90]. The theory of optimal transport was pioneered by Monge in 1781 [62] and formalized by Kantorovich in 1942 [50]. Accordingly, the Wasserstein distance is often referred to as the Monge-Kantorovich distance [107]. The Wasserstein distance is used in many areas of science. In the wider context of machine learning, for instance, the Wasserstein distance is used for the analysis of mixture models [56, 70] as well as for image processing [1, 33, 55, 75, 105], computer vision and graphics [78, 79, 90, 102, 103], data-driven bioengineering [34, 57, 108], clustering [49], dimensionality reduction [18, 35, 87, 93, 95], deep learning with generative adversarial networks [2, 41, 45], domain adaptation [24, 63], signal processing [106], etc. For a comprehensive survey of different applications of the optimal transport problem see [54, 82].

The optimization problem in (5) constitutes an infinite-dimensional linear program over the transportation plan π. This linear program admits a strong dual, which in turn provides an alternative characterization of the Wasserstein distance.

Theorem 1 (Dual Kantorovich problem). *For any $p \in [1,\infty)$, the p^{th} power of the type-p Wasserstein distance between $\mathbb{Q}$ and $\mathbb{Q}'$ admits the dual representation*

$$\begin{array}{lll} W_p^p(\mathbb{Q},\mathbb{Q}') = & \sup & \displaystyle\int_{\mathbb{R}^m} \psi(\xi')\,\mathbb{Q}'(\mathrm{d}\,\xi') - \int_{\mathbb{R}^m} \phi(\xi)\,\mathbb{Q}(\mathrm{d}\,\xi) \\ & \text{s.t.} & \phi \text{ and } \psi \text{ are bounded continuous functions on } \mathbb{R}^m \text{ with} \\ & & \psi(\xi) - \phi(\xi') \le \|\xi - \xi'\|^p \quad \forall \xi, \xi' \in \mathbb{R}^m. \end{array}$$

For a proof of Theorem 1 see [107, § 5]. The dual problem can be interpreted as the profit maximization problem of a third party that reallocates the dirt from $\mathbb{Q}$ to $\mathbb{Q}'$ on behalf of the problem owner by buying dirt at the origin ξ at unit price $\phi(\xi)$ and selling dirt at the destination ξ' at unit price $\psi(\xi')$. The constraints ensure that the problem owner prefers to use the services of the third party for every origin-destination pair (ξ, ξ') instead of reallocating the dirt independently at her own transportation cost $\|\xi - \xi'\|^p$. The optimal price functions $\phi^\star$ and $\psi^\star$—if they exist—are called Kantorovich potentials [107, p. 99].

If $p = 1$, the dual problem can be further simplified. To see this, we define the Lipschitz modulus of an extended real-valued function ϕ on $\mathbb{R}^m$ with respect to the norm $\|\cdot\|$ as

$$\mathrm{Lip}(\phi) = \sup_{\xi \neq \xi'} \frac{|\phi(\xi) - \phi(\xi')|}{\|\xi - \xi'\|}.$$

The Lipschitz modulus can be viewed as the slope of the steepest line segment connecting any two points on the graph of ϕ. The following result simplifies Theorem 1 for $p = 1$.

Theorem 2 (Kantorovich-Rubinstein theorem). *The type-1 Wasserstein distance between $\mathbb{Q}$ and $\mathbb{Q}'$ admits the dual representation*

$$W_1(\mathbb{Q},\mathbb{Q}') = \sup_{\mathrm{Lip}(\phi) \le 1} \int_{\mathbb{R}^m} \phi(\xi)\,\mathbb{Q}(\mathrm{d}\,\xi) - \int_{\mathbb{R}^m} \phi(\xi')\,\mathbb{Q}'(\mathrm{d}\,\xi').$$

Kantorovich and Rubinstein [51] originally established this result for compactly supported distributions. A modern proof for arbitrary distributions can be found in [107, Remark 6.5]. Theorem 2 asserts that the type-1 Wasserstein distance between $\mathbb{Q}$ and $\mathbb{Q}'$ equals the difference between the expected values of a test function ϕ under $\mathbb{Q}$ and $\mathbb{Q}'$, respectively, maximized across all Lipschitz-continuous test functions with Lipschitz modulus of at most 1.

The Kantorovich-Rubinstein theorem enables us to estimate the sensitivity of $\mathcal{R}(\mathbb{P},\ell)$ and $\mathcal{R}(\mathbb{P},\mathcal{L})$ with respect to the unknown true distribution $\mathbb{P}$. To see this, assume that the type-1 Wasserstein distance between $\mathbb{P}$ and its noisy estimate $\widehat{\mathbb{P}}_N$ is known to be at most ε. Thus, ε can be viewed as a measure of the estimation error. Assume further that a fixed loss function $\ell(\xi)$ is Lipschitz-continuous with Lipschitz constant L. The risk of ℓ then satisfies

$$\left|\mathcal{R}(\widehat{\mathbb{P}}_N,\ell) - \mathcal{R}(\mathbb{P},\ell)\right| = L \cdot \left|\mathbb{E}^{\widehat{\mathbb{P}}_N}[\ell(\xi)/L] - \mathbb{E}^{\mathbb{P}}[\ell(\xi)/L]\right| \le L \cdot W_1(\widehat{\mathbb{P}}_N,\mathbb{P}) \le L \cdot \varepsilon,$$

where the equality holds due to the definition of the risk, while the first inequality follows from the Kantorovich-Rubinstein theorem, which applies because $\mathrm{Lip}(\ell/L) \le 1$. Moreover, if all loss functions $\ell \in \mathcal{L}$ are Lipschitz continuous with the same Lipschitz constant L, then a similar reasoning implies that the optimal risk satisfies $|\mathcal{R}(\widehat{\mathbb{P}}_N,\mathcal{L}) - \mathcal{R}(\mathbb{P},\mathcal{L})| \le L \cdot \varepsilon$. This analysis offers a rough understanding of how estimation errors in the input distribution are propagated to the (optimal) risk: they are amplified at most by the Lipschitz constant of the involved loss functions. Arguments of this type are central to the stability theory of stochastic programming. For example, it is known that under standard regularity conditions, the optimal values of two-stage stochastic programs with random right hand sides are Lipschitz continuous in the distribution of the uncertainty with respect to the Wasserstein distance [89]. Classical stability results in stochastic programming are surveyed in [31, 88].

The above reasoning suggests that in order to approximate the (optimal) risk well, one should construct an estimator $\widehat{\mathbb{P}}_N$ that has a small Wasserstein distance to the unknown true distribution $\mathbb{P}$ with high confidence. Unfortunately, however, estimators are subject to fundamental performance limitations and cannot be improved beyond a certain level.

Example 3 (Limitations of estimator performance). Depending on the available structural information on $\mathbb{P}$, the nominal distributions portrayed in Example 1, which will be used throughout this tutorial, are essentially optimal within certain estimator families.

(1) **Discrete distributions:** Assume that $\mathbb{P}$ is only known to be supported on a compact set $\Xi \subseteq \mathbb{R}^m$, and let $\mathcal{P}_N$ be the family of all discrete distributions on Ξ with N atoms. The theory of optimal quantization shows that there exist $\underline{N} \in \mathbb{N}$ and $\underline{c} > 0$ such that $\inf_{\mathbb{Q}\in\mathcal{P}_N} W_1(\mathbb{Q},\mathbb{P}) \geq \underline{c} N^{-1/m}$ for all $N \geq \underline{N}$ [17, Theorem 3.3]. Thus, the type-1 Wasserstein distance between $\mathbb{P}$ and its closest N-point distribution cannot decay faster than $N^{-1/m}$. Maybe surprisingly, the empirical distribution $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_i}$ attains this optimal decay rate in a probabilistic sense even though it is constructed from N random samples but without knowledge of $\mathbb{P}$. Indeed, [36, Theorem 2] implies that for every $\eta \in (0,1)$ there exist $\overline{N} \in \mathbb{N}$ and $\overline{c} > 0$ such that $W_1(\widehat{\mathbb{P}}_N,\mathbb{P}) \leq \overline{c} N^{-1/m}$ with confidence $1-\eta$ for every $N \geq \overline{N}$. Thus, if we aim to approximate $\mathbb{P}$ with a sequence of *discrete* distributions, the empirical distribution $\widehat{\mathbb{P}}_N$ is essentially optimal in the sense that it attains the best possible convergence rate at any desired confidence level.

(2) **Elliptical distributions:** Assume that $\mathbb{P}$ is known to be an elliptical distribution with a known density generator g but unknown mean vector μ and covariance matrix Σ. In this case, the problem of finding an estimator $\widehat{\mathbb{P}}_N$ for the distribution $\mathbb{P}$ reduces to finding an estimator $\widehat{\theta}_N$ for the vector $\theta = (\mu, \Sigma)$ of unknown distribution parameters. Under mild regularity conditions, the Cramér-Rao inequality guarantees that the covariance matrix of $\sqrt{N}\cdot\widehat{\theta}_N$ exceeds the inverse Fisher information matrix in a positive semidefinite sense for *any* unbiased estimator $\widehat{\theta}_N$. As the maximum likelihood estimator $\widehat{\theta}_N^{\mathrm{ML}}$ is asymptotically unbiased and efficient, *i.e.*, the mean of $\widehat{\theta}_N^{\mathrm{ML}}$ converges to θ and the variance of $\sqrt{N}\cdot\widehat{\theta}_N^{\mathrm{ML}}$ converges to the inverse Fisher information matrix as N grows, it is asymptotically optimal among all conceivable unbiased estimators.

We emphasize that, by mobilising more powerful results from statistics, the above optimality guarantees could be extended to even larger families of estimators. □

Example 3 suggests that the accuracy of the nominal distribution cannot be increased beyond some fundamental limit by tuning the estimator. The only remaining option to reduce the estimation error is to increase the sample size N, which may be expensive or impossible. Indeed, additional training samples may only become available in the future. Thus, the optimizer's curse illustrated in Example 2 is fundamental and cannot be eliminated. However, once the potential to improve the estimator $\widehat{\mathbb{P}}_N$ is exhausted, it may still be possible to mitigate the optimizer's curse by altering the risk evaluation and decision problems (1) and (2) directly. Specifically, we propose here to robustify these problems against the uncertainty about the true distribution $\mathbb{P}$. Distributional uncertainty is often referred to as *ambiguity* or *Knightian uncertainty* and is conveniently captured by an *ambiguity set*, that is, an uncertainty set in the space of probability distributions. To formalize this idea, we let $\Xi \subseteq \mathbb{R}^m$ be a closed set that is known to contain the support of $\mathbb{P}$. In the absence of any structural information, we may simply set $\Xi = \mathbb{R}^m$. Moreover, we denote by $\mathcal{P}(\Xi)$ the family of all probability distributions supported on Ξ, and we define the ambiguity set

$$\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N) = \Big\{ \mathbb{Q} \in \mathcal{P}(\Xi) : W_p(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon \Big\}$$

as the ball of radius $\varepsilon \geq 0$ in $\mathcal{P}(\Xi)$ centered at the nominal distribution $\widehat{\mathbb{P}}_N$ with respect to the type-p Wasserstein distance. By construction, this ambiguity set contains all distributions supported on Ξ that can be obtained by reshaping the nominal distribution $\widehat{\mathbb{P}}_N$ at a

transportation cost of at most ε. We can think of $\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)$ as the set of all distributions for which the estimation error—as measured by the type-p Wasserstein distance—is at most ε, and we can interpret ε as the maximum estimation error against which we seek protection.

Using the proposed ambiguity set, we define the *worst-case risk* as

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) = \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \mathcal{R}(\mathbb{Q}, \ell) \tag{6}$$

and the *worst-case optimal risk* as

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \mathcal{L}) = \inf_{\ell\in\mathcal{L}} \mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell). \tag{7}$$

Problem (7) constitutes a distributionally robust optimization problem. It seeks decisions that have minimum risk under the most adverse distributions in the ambiguity set. Intuitively, problem (7) can thus be viewed as a zero-sum game, where the decision-maker first selects an admissible loss function with the goal to minimize the risk, in response to which some fictitious adversary or 'nature' selects a distribution from within the ambiguity set with the goal to maximize the risk. The hope is that by minimizing the worst-case risk, we actually push down the risk under *all* distributions in the ambiguity set—in particular under the unknown true distribution $\mathbb{P}$, which is contained in the ambiguity set if ε is large enough. Thus, there is reason to hope that the solutions of distributionally robust optimization problems with carefully calibrated ambiguity sets display low out-of-sample risk.

The distributionally robust risk evaluation and decision problems (6) and (7) are attractive for a multitude of diverse reasons.

- **Fidelity:** Distributionally robust models are more 'honest' than their nominal counterparts as they acknowledge the presence of distributional uncertainty. They also benefit from information about the type and magnitude of the estimation errors, which is conveniently encoded in the geometry and size of the ambiguity set.
- **Managing expectations:** Due to the optimizer's curse, the solutions of nominal decision problems equipped with noisy estimators display an optimistic in-sample risk, which cannot be realized out of sample; see Example 2. In contrast, the solutions of distributionally robust decision problems are guaranteed to display an out-of-sample risk that falls below the worst-case optimal risk whenever the ambiguity set contains the unknown true distribution. Thus, nominal decision problems over-promise and under-deliver, while distributionally robust decision problems under-promise and over-deliver.
- **Computational tractability:** The distributionally robust problems (6) and (7) can often be reformulated as (or tightly approximated by) finite convex programs that are solvable in polynomial time. Section 2 will showcase some key tractability results.
- **Performance guarantees:** For judiciously calibrated ambiguity sets, one can prove that the worst-case optimal risk for any fixed sample size N provides an upper confidence bound on the out-of-sample risk attained by the optimizers of (7) (finite sample guarantee) and that the optimizers of (7) converge almost surely to an optimizer of (2) as N tends to infinity (asymptotic guarantee); see Section 3.
- **Regularization by robustification:** The optimizer's curse is reminiscent of overfitting phenomena that plague most statistical learning models. One can show that distributionally robust learning models equipped with a Wasserstein ambiguity set are often equivalent to regularized learning models that minimize the sum of a nominal objective and a norm term that penalizes hypothesis complexity. Similarly, one can show that some distributionally robust maximum likelihood estimation models produce shrinkage estimators. Thus, Wasserstein distributional robustness offers new probabilistic interpretations for popular regularization techniques. The empirical success of regularization methods in statistics fuels hope that Wasserstein distributionally robust models can effectively combat the optimizer's curse across many application areas. Connections between robustification and regularization will be explored in Section 4.

- **Anticipating black swans:** If uncertainty is modeled by the empirical distribution, then the nominal decision problem evaluates the admissible loss functions only at the training samples. However, possible future uncertainty realizations that differ from all training samples but could have devastating consequences ('black swans') are ignored. If the empirical distribution may be perturbed within a Wasserstein ball with a positive radius, on the other hand, then (possibly small amounts of) probability mass can be moved anywhere in the support set Ξ. Thus, the Wasserstein distributionally robust decision problem faithfully anticipates the possibility of black swans. We emphasize that all distributions in a Kullback-Leibler divergence ball must be absolutely continuous with respect to the nominal distribution, which implies that the corresponding distributionally robust decision problems ignore the possibility of black swans.
- **Axiomatic justification:** If the random vector ξ may follow any distribution in some ambiguity set $\mathcal{Q}$ (*e.g.*, a Wasserstein ball), then the scalar random variable $\ell(\xi)$ corresponding to a fixed loss function $\ell \in \mathcal{L}$ may follow any distribution in the induced ambiguity set $\ell_*(\mathcal{Q}) = \{\ell_*(\mathbb{Q}) : \mathbb{Q} \in \mathcal{Q}\}$, where $\ell_*(\mathbb{Q})$ is the pushforward measure of $\mathbb{Q}$ under ℓ. We call a loss function $\ell \in \mathcal{L}$ unambiguous if $\ell_*(\mathcal{Q})$ is a singleton. Assume now that ℓ is preferred to ℓ' under any of the following natural conditions: (i) ℓ and ℓ' are unambiguous, and $\mathcal{R}(\mathbb{Q}, \ell) \leq \mathcal{R}(\mathbb{Q}, \ell')$ for some $\mathbb{Q} \in \mathcal{Q}$; (ii) $\ell_*(\mathcal{Q}) \subseteq \ell'_*(\mathcal{Q})$; (iii) $\mathcal{R}(\mathbb{Q}, \ell) \leq \mathcal{R}(\mathbb{Q}, \ell')$ for every $\mathbb{Q} \in \mathcal{Q}$. Under a mild technical condition, the loss functions must then be ranked by the worst-case risk $\sup_{\mathbb{Q}\in\mathcal{Q}} \mathcal{R}(\mathbb{Q}, \ell)$ [28, Theorem 12]. This result provides an axiomatic justification for adopting a distributionally robust approach.
- **Optimality principle:** Data-driven optimization aims to use the training data directly to construct an estimator for the objective of problem (2) (a predictor) and a decision that minimizes this predictor (a prescriptor) without the detour of constructing an estimator for $\mathbb{P}$. It has been shown that optimal predictors and the corresponding prescriptors can be constructed by solving a meta-optimization model that minimizes the in-sample risk of the predictor-prescriptor pairs subject to constraints guaranteeing that the in-sample risk is actually attainable out of sample. It has been shown that this meta-optimization problem admits a unique solution: the best predictor-prescriptor pair is obtained by solving a distributionally robust optimization problem over all distributions in some neighborhood of the empirical distribution [77, Theorem 7]. Thus, if one aims to transform training data to decisions, it is in some precise sense optimal to do this by solving a data-driven distributionally robust optimization problem.

Distributionally robust optimization models with Wasserstein ambiguity sets were introduced in [84]. Reformulations of these models as nonconvex optimization problems as well as initial attempts to solve these problems via algorithms from global optimization are reported in [111] and [83, § 7.1]. In the next section we will review convex reformulations and approximations that were discovered in [60, 113] and significantly generalized in [12, 38].

Notation. The conjugate of a function $\ell(\xi)$ on $\mathbb{R}^m$ is defined as $\ell^*(z) = \sup_\xi z^\top \xi - \ell(\xi)$. The indicator function of a set $\Xi \subseteq \mathbb{R}^m$ is defined as $\delta_\Xi(\xi) = 0$ if $\xi \in \Xi$ and $\delta_\Xi(\xi) = \infty$ if $\xi \notin \Xi$. The conjugate $\sigma_\Xi(z) = \sup_{\xi\in\Xi} z^\top \xi$ of the indicator function is termed the support function. If $\|\xi\|$ represents the norm of $\xi \in \mathbb{R}^m$, then $\|z\|_* = \sup_{\|\xi\|\leq 1} z^\top \xi$ denotes the corresponding dual norm. The set of all symmetric (positive semidefinite) matrices $A \in \mathbb{R}^{m\times m}$ is denoted by $\mathbb{S}^m$ ($\mathbb{S}^m_+$). For $A, B \in \mathbb{S}^m$, the relation $A \succeq B$ ($A \succ B$) means that $A - B$ is positive semidefinite (positive definite). The trace of $A \in \mathbb{R}^{m\times m}$ is denoted by $\mathrm{Tr}\,[A]$, the smallest and largest eigenvalues of $A \in \mathbb{S}^m$ are denoted by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively, and the Moore-Penrose pseudoinverse of $A \in \mathbb{S}^m_+$ is denoted by $A^\dagger$. For $N \in \mathbb{N}$ we set $[N] = \{1, \ldots, N\}$.

2. Computation

The aim of this section is to show that the worst-case risk evaluation problem (6) and the distributionally robust decision problem (7) are computationally tractable in many situations of practical interest. Note first that checking whether a fixed distribution $\mathbb{Q}$ is feasible

in (6) requires computing the Wasserstein distance $W_p(\mathbb{Q}, \widehat{\mathbb{P}}_N)$. It is therefore instructive to study the complexity of evaluating Wasserstein distances between arbitrary distributions.

Computing the Wasserstein distance between two discrete distributions amounts to solving a tractable linear program that is susceptible to the network simplex algorithm [7] as well as dual ascent methods [9] or specialized auction algorithms [5, 6], etc. The set of feasible transportation plans is termed the transportation polytope and displays many useful theoretical properties, which are surveyed in [15]. The need to evaluate Wasserstein distances between increasingly fine-grained histograms has recently motivated efficient approximation schemes. When augmented with an entropic regularization term, for instance, the finite-dimensional transportation problem can be solved quickly by using Sinkhorn's algorithm [22, 26, 52, 80, 81, 94, 102]. Variants of this approach use Tikhonov regularizers [32], Bregman divergences [4] or Tsallis entropies [65] instead of the entropic regularization term. A survey of algorithms for the finite-dimensional transportation problem is provided in [82].

As soon as at least one of the two involved distributions ceases to be discrete, the Wasserstein distance can no longer be evaluated in polynomial time. Even in the simplest imaginable scenario where one distribution is uniform on a hypercube and the other distribution is discrete with two atoms, computing the Wasserstein distance becomes intractable [100].

Theorem 3 (Hardness of computing Wasserstein distances). *Computing the type-p Wasserstein distance between two distributions $\mathbb{Q}$ and $\mathbb{Q}'$ is #P-hard even if $\|\cdot\|$ is the Euclidean norm, $\mathbb{Q}$ is the uniform distribution on the standard hypercube $[0,1]^m$, and $\mathbb{Q}'$ is a discrete distribution supported on only two points.*

If $p = 2$ and $\|\cdot\|$ is the Euclidean norm, then the Wasserstein distance admits an analytical lower bound that depends only on the distributions' first- and second-order moments. This bound is available for *any* pair of distributions even if their exact Wasserstein distance cannot be computed efficiently. Moreover, the bound is exact for elliptical distributions.

Theorem 4 (Gelbrich bound). *If $\|\cdot\|$ is the Euclidean norm, and the distributions $\mathbb{Q}$ and $\mathbb{Q}'$ have mean vectors $\mu, \mu' \in \mathbb{R}^m$ and covariance matrices $\Sigma, \Sigma' \in \mathbb{S}_+^m$, respectively, then*

$$W_2(\mathbb{Q}, \mathbb{Q}') \geq \sqrt{\|\mu - \mu'\|_2^2 + \mathrm{Tr}\left[\Sigma + \Sigma' - 2\left(\Sigma^{\frac{1}{2}}\Sigma'\Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]}. \tag{8}$$

The bound is exact if $\mathbb{Q}$ and $\mathbb{Q}'$ are elliptical distributions with the same density generator.

The inequality (8) may be loose if $\mathbb{Q}$ and $\mathbb{Q}'$ are elliptical distributions with different density generators. Maybe unexpectedly, however, the Wasserstein distance between two elliptical distributions with the same density generator g is actually independent of g. In its general form, Theorem 4 is due to Gelbrich [40]. The exact formula for the type-2 Wasserstein distance between normal distributions has been discovered earlier in [30, 43, 72].

As any Wasserstein ball with a strictly positive radius contains non-discrete distributions (the nominal distribution can be smeared out even if the transportation budget is small), it is perhaps surprising that the worst-case risk evaluation problem (6) may be tractable at all. Indeed, Theorem 3 indicates that checking feasibility is already hard in general. We will see below, however, that the extremal distributions determining the worst-case risk are often structurally equivalent to the nominal distribution. Thus, there is hope that problems (6) and (7) become tractable if we choose a nominal distribution with a particularly simple structure (*e.g.*, a discrete or an elliptical distribution).

In order to ensure that any admissible loss function $\ell \in \mathcal{L}$ has a finite expected value under the nominal distribution, we impose the following technical regularity condition borrowed from [12], which will tacitly be assumed to hold throughout the rest of the paper.

Assumption 1 (Regularity). *Any loss function $\ell \in \mathcal{L}$ is upper semicontinuous and integrable with respect to the nominal distribution $\widehat{\mathbb{P}}_N$, that is, $\int_{\mathbb{R}^m} |\ell(\xi)|\, \widehat{\mathbb{P}}_N(\mathrm{d}\xi) < \infty$.*

In the remainder of this section, we will first review tractable bounds on the worst-case risk and present a strong duality result that paves the way towards exact tractable reformulations (Section 2.1). Next, we will delineate efficient methods to compute the worst-case risk as well as the underlying worst-case distributions in situations when the nominal distribution is discrete (Section 2.2) or elliptical (Section 2.3).

2.1. General Analysis of the Worst-Case Risk

Before attempting to derive exact tractable reformulations for the worst-case risk (6), we focus on the simpler task of establishing efficiently computable upper and lower bounds. To derive a pessimistic upper bound, we note that the transportation cost $\|\xi - \xi'\|^p$ is a convex function of the random variable $\|\xi - \xi'\|$ for any $p \geq 1$. Jensen's inequality thus implies

$$W_p(\mathbb{Q}, \widehat{\mathbb{P}}_N) \geq W_1(\mathbb{Q}, \widehat{\mathbb{P}}_N) \quad \forall \mathbb{Q} \in \mathcal{P}(\Xi) \quad \Longrightarrow \quad \mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N) \subseteq \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_N).$$

Hence, the worst-case risk of a loss function $\ell \in \mathcal{L}$ over the type-p Wasserstein ball satisfies

$$\begin{aligned}\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \leq \mathcal{R}_{\varepsilon,1}(\widehat{\mathbb{P}}_N, \ell) &= \mathbb{E}^{\widehat{\mathbb{P}}_N}[\ell(\xi)] + \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\ell(\xi)] - \mathbb{E}^{\widehat{\mathbb{P}}_N}[\ell(\xi)] \\ &\leq \mathcal{R}(\widehat{\mathbb{P}}_N, \ell) + \varepsilon \cdot \mathrm{Lip}(\ell),\end{aligned}$$

where the equality follows from the definition of the worst-case risk, while the second inequality is a direct consequence of the Kantorovich-Rubinstein theorem (see Theorem 2). We summarize the above reasoning in the following theorem.

Theorem 5 (Lipschitz regularization). *The worst-case risk (6) of any fixed loss function $\ell \in \mathcal{L}$ is bounded above by the Lipschitz-regularized nominal risk, that is,*

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \leq \mathcal{R}(\widehat{\mathbb{P}}_N, \ell) + \varepsilon \cdot \mathrm{Lip}(\ell).$$

If the loss function ℓ fails to be Lipschitz continuous (*i.e.*, $\mathrm{Lip}(\ell) = \infty$), then Theorem 5 is trivially satisfied. Note that $\mathcal{R}(\widehat{\mathbb{P}}_N, \ell)$ is linear in ℓ for any choice of the nominal distribution and that $\mathrm{Lip}(\ell)$ is a convex function of ℓ. Thus, minimizing the upper bound of Theorem 5 amounts to solving a convex optimization problem whenever $\mathcal{L}$ is a convex set.

An optimistic lower bound on the worst-case risk can be obtained by replacing the Wasserstein ball in (6) with a smaller ambiguity set. If the distributions in the restricted Wasserstein ball admit a finite parameterization, then the lower bounding problem coincides with a finite optimization problem. Depending on the parameterization, this problem may even be convex. If $\widehat{\mathbb{P}}_N$ is the empirical distribution, for example, one may restrict the original Wasserstein ball to a subset that contains only *perturbed* empirical distributions of the form

$$\mathbb{Q}(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\widehat{\xi}_i + \theta_i},$$

where $\theta_i \in \mathbb{R}^m$ is the displacement of the i^{th} training sample. Thus, all distributions in the restricted Wasserstein ball are encoded by a perturbation matrix $\Theta = (\theta_1, \ldots, \theta_N) \in \mathbb{R}^{m \times N}$. The requirement that $\mathbb{Q}(\Theta) \in \mathcal{P}(\Xi)$ translates to $\widehat{\xi}_i + \theta_i \in \Xi$ for all $i \in [N]$, while the Wasserstein constraint $W_p(\mathbb{Q}(\Theta), \widehat{\mathbb{P}}_N) \leq \varepsilon$ is equivalent to the inequality $\frac{1}{N} \sum_{i=1}^{N} \|\theta_i\|^p \leq \varepsilon^p$.

Theorem 6 (Robust lower bound). *If $\widehat{\mathbb{P}}_N$ is the empirical distribution, then the worst-case risk (6) of any fixed loss function $\ell \in \mathcal{L}$ is bounded below by the worst-case empirical loss, where the worst case is taken over all perturbation matrices $\Theta = (\theta_1, \ldots, \theta_N) \in \mathbb{R}^{m \times N}$ of the training samples in an $L_{p,1}$-norm uncertainty set, that is, we have*

$$\begin{array}{rcll} \mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \;\geq\; \sup & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \ell(\widehat{\xi}_i + \theta_i) & & \\ \text{s.t.} & \theta_i \in \mathbb{R}^m & \forall i \in [N] & \\ & \widehat{\xi}_i + \theta_i \in \Xi & \forall i \in [N] & \\ & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \|\theta_i\|^p \leq \varepsilon^p. & & \end{array} \tag{9}$$

Note that if the loss function ℓ is concave, then the robust lower bounding problem of Theorem 6 constitutes a finite convex optimization problem. In the remainder we will argue that both the upper bound of Theorem 5 as well as the lower bound of Theorem 6 can become exact in situations of practical interest. To see this, we first derive the Lagrangian dual of the worst-case risk evaluation problem (6).

Theorem 7 (Strong duality). *The worst-case risk (6) of any fixed $\ell \in \mathcal{L}$ satisfies*

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) = \inf_{\gamma \geq 0} \; \mathbb{E}^{\widehat{\mathbb{P}}_N}\left[\ell_\gamma(\xi)\right] + \gamma\varepsilon^p, \tag{10}$$

where $\ell_\gamma(\xi) = \sup_{z \in \Xi} \; \ell(z) - \gamma\|z - \xi\|^p$ is a Moreau-Yosida regularization [76] of $\ell(\xi)$.

The minimization problem on the right hand side of (10) can indeed be identified with the strong Lagrangian dual of problem (6), where $\gamma \geq 0$ is the Lagrange multiplier of the Wasserstein constraint $W_p(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon$. We emphasize that the Moreau-Yosida regularization $\ell_\gamma(\xi)$ is jointly convex in γ and ℓ for every fixed uncertainty realization ξ. Thus the dual problem (10) represents a convex minimization problem whose optimal value is convex in ℓ. One can further show that (10) is solvable for any $\varepsilon > 0$ under the mild assumption that there exists $C > 0$ such that $|\ell(\xi)| \leq C(1 + \|\xi\|^p)$ for all $\xi \in \Xi$. For type-1 Wasserstein balls centered at the empirical distribution, Theorem 7 is a corollary of [60, Theorem 4.2] and [113, Proposition 2]. An extension of Theorem 7 to situations where ξ ranges over a Polish space is discussed in [12, 38]. It has been shown that Theorem 7 remains even valid if the transportation cost $\|\xi - \xi'\|^p$ in the definition of the Wasserstein distance is replaced with a general nonnegative and lower semicontinuous function $c(\xi, \xi')$ that vanishes if and only if $\xi = \xi'$ [12]. Note that the Wasserstein distance may cease to be a metric in this case.

In the next sections we will describe specific settings in which (6) and (10) are tractable.

2.2. Tractability Results for Empirical Nominal Distributions

Assume now that the Wasserstein ambiguity set is centered at the empirical distribution defined in (3). In this case, under a mild convexity assumption, the worst-case risk (6) can be exactly expressed as the optimal value of a finite convex optimization problem.

Theorem 8 (Piecewise concave loss I). *Assume that Ξ is convex and closed and that $\ell(\xi) = \max_{j \in [J]} \ell_j(\xi)$, where $-\ell_j$ is proper, convex and lower semicontinuous for all $j \in [J]$. If $\widehat{\mathbb{P}}_N$ is the empirical distribution and $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, then the worst-case risk (6) coincides with the optimal value of a finite convex minimization problem, that is,*

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \;=$$

$$\begin{array}{ll} \inf & \gamma\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{s.t.} & \gamma \in \mathbb{R}_+,\ s_i \in \mathbb{R},\ u_{ij} \in \mathbb{R}^m,\ v_{ij} \in \mathbb{R}^m \qquad\qquad \forall i \in [N],\ j \in [J] \\ & [-\ell_j]^*(u_{ij} - v_{ij}) + \sigma_\Xi(v_{ij}) - u_{ij}^\top \widehat{\xi}_i + \varphi(q)\,\gamma \left\| \dfrac{u_{ij}}{\gamma} \right\|_*^q \le s_i \quad \forall i \in [N],\ j \in [J], \end{array} \tag{11}$$

where $[-\ell_j]^*(z)$ *is the conjugate of* $-\ell_j(\xi)$, $\sigma_\Xi(z)$ *is the support function of* Ξ, *and* $\|\cdot\|_*$ *is the dual of the norm* $\|\cdot\|$ *on* $\mathbb{R}^m$, *while* $\varphi(q) = (q-1)^{q-1}/q^q$ *for* $q > 1$ *and* $\varphi(1) = 1$. *For* $\gamma = 0$, *the expression* $0\|u_{ij}/0\|_*^q$ *is interpreted as* $\lim_{\gamma \downarrow 0} \gamma \|u_{ij}/\gamma\|_*^q$.

The assumptions of Theorem 8 are unrestrictive because any continuous function $\ell(\xi)$ on a compact set Ξ can be uniformly approximated as closely as desired by a pointwise maximum of finitely many concave functions. Note that the loss function $\ell(\xi)$ and the support set Ξ enter problem (11) through the conjugates of the negative constituent functions $-\ell_j(\xi)$, $j \in [J]$, and the support function $\sigma_\Xi(z)$, all of which are convex. Moreover, the norm that determines the transportation cost in the definition of the Wasserstein distance enters (11) via the dual norm $\|\cdot\|_*$. The term $\gamma\|u_{ij}/\gamma\|_*^q$ can be identified with the perspective function of $\|u_{ij}\|_*^q$ and is thus jointly convex in γ and u_{ij} [14, § 3.2.6]. Therefore, problem (11) is manifestly convex. Tables 4–6 in Appendix B list common conjugates, support functions and dual norms. By substituting (11) into (7), one can reformulate the distributionally robust decision problem (7) as a single explicit minimization problem, which is convex whenever $\mathcal{L}$ is a convex set. To prove Theorem 8, one re-expresses the empirical expectation in the objective function of the dual problem (7) as a finite sum and dualizes the maximization problems in the Moreau-Yosida regularization terms. For further details see [60, Theorem 4.2] and [114].

Remark 1 (Limiting cases I). In the limit when p tends to 1 and q to ∞, the function $\varphi(q)$ decays as $1/q$, while $\|u_{ij}/\gamma\|_*^q$ grows exponentially whenever $\|u_{ij}\|_* > \gamma$. Thus, we have

$$\lim_{q \uparrow \infty}\ \varphi(q)\,\gamma \left\| \frac{u_{ij}}{\gamma} \right\|_*^q = \begin{cases} 0 & \text{if } \|u_{ij}\|_* \le \gamma, \\ \infty & \text{if } \|u_{ij}\|_* > \gamma. \end{cases}$$

For $p = 1$, the constraints of the finite convex program (11) are thus equivalent to

$$[-\ell_j]^*(u_{ij} - v_{ij}) + \sigma_\Xi(v_{ij}) - u_{ij}^\top \widehat{\xi}_i \le s_i, \quad \|u_{ij}\|_* \le \gamma \quad \forall i \in [N],\ j \in [J].$$

In the opposite limit when p tends to ∞ and q to 1, the function $\varphi(q)$ converges to 1, and therefore it is easy to see that the constraints of problem (11) simplify to

$$[-\ell_j]^*(u_{ij} - v_{ij}) + \sigma_\Xi(v_{ij}) - u_{ij}^\top \widehat{\xi}_i + \|u_{ij}\|_* \le s_i \quad \forall i \in [N],\ j \in [J].$$

Thus, the variable γ disappears from the constraints. As the objective function coefficient of γ is nonnegative, this implies that $\gamma = 0$ at optimality. □

As it expresses the worst-case risk $\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell)$ as the optimal value of a *minimization* problem, Theorem 8 primarily serves as a vehicle to solve the distributionally robust decision problem (7). In order to construct an extremal distribution that solves problem (6), one may dualize the finite convex program (11) to convert it back to a *maximization* problem.

Theorem 9 (Piecewise concave loss II). *Under the conditions of Theorem 8 we have*

$$\begin{array}{lll} \mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N,\ell) \;=\; \max & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{J}\alpha_{ij}\ell_j\Big(\widehat{\xi}_i+\dfrac{\theta_{ij}}{\alpha_{ij}}\Big) & \\ \text{s.t.} & \alpha_{ij}\in\mathbb{R}_+,\ \theta_{ij}\in\mathbb{R}^m & \forall i\in[N],\ \forall j\in[J] \\ & \widehat{\xi}_i+\dfrac{\theta_{ij}}{\alpha_{ij}}\in\Xi & \forall i\in[N],\ \forall j\in[J] \\ & \displaystyle\sum_{j=1}^{J}\alpha_{ij}=1 & \forall i\in[N] \\ & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{J}\alpha_{ij}\left\|\dfrac{\theta_{ij}}{\alpha_{ij}}\right\|^p\le\varepsilon^p, & \end{array} \tag{12}$$

where $0\,\ell_j(\widehat{\xi}_i+\theta_{ij}/0)$ *is defined as the value that makes the function* $\alpha_{ij}\ell_j(\widehat{\xi}_i+\theta_{ij}/\alpha_{ij})$ *upper semicontinuous at* $(\theta_{ij},\alpha_{ij})=(\theta_{ij},0)$. *Similarly, the constraint* $\widehat{\xi}_i+\theta_{ij}/0\in\Xi$ *means that* θ_{ij} *belongs to the recession cone of* Ξ, *and* $0\|\theta_{ij}/0\|^p$ *is interpreted as* $\lim_{\alpha_{ij}\downarrow 0}\alpha_{ij}\|\theta_{ij}/\alpha_{ij}\|^p$.

Problem (12) is the Lagrangian dual of (11) and thus convex by construction. Convexity can also be verified directly. As the constituent functions $\ell_j(\xi)$, $j\in[J]$, are concave by assumption, the objective of (12) represents a sum of concave perspective functions and is thus concave [14, § 3.2.6]. Also, the support constraints $\widehat{\xi}_i+\theta_{ij}/\alpha_{ij}\in\Xi$ require that $(\theta_{ij},\alpha_{ij})$ belongs to the preimage of the convex set $\{\widehat{\xi}_i+\xi:\xi\in\Xi\}$ under the perspective transformation, which is known to be convex [14, § 2.3.3]. The term $\alpha_{ij}\|\theta_{ij}/\alpha_{ij}\|^p$ can be identified with the perspective function of $\|\theta_{ij}\|^p$ and is thus jointly convex in α_{ij} and θ_{ij} [14, § 3.2.6].

For a proof of Theorem 9 we refer to [60, Theorem 4.4] and [114]. We emphasize that problem (12) is always solvable because it has a compact feasible set and an upper semicontinuous objective function, and thus the use of the maximization operator is justified.

Note that if $J=1$, then the loss function $\ell(\xi)=\ell_1(\xi)$ is globally concave, and the penultimate constraint group in (12) simplifies to the requirement that $\alpha_{i1}=1$ for every $i\in[N]$. In this case, the convex program (12), which is equivalent to the worst-case risk evaluation problem (6), reduces to the robust optimization problem (9), which maximizes only over perturbed empirical distributions in the Wasserstein ball. Thus, the robust lower bound portrayed in Theorem 6 is exact if the loss function $\ell(\xi)$ is concave.

Remark 2 (Limiting cases II). For $p=1$, the last constraint of (12) simplifies to

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\|\theta_{ij}\|\le\varepsilon.$$

To analyze the limit when p tends to ∞, we divide the last constraint of (12) by ε^p and observe that $\|\theta_{ij}/(\varepsilon\,\alpha_{ij})\|^p$ grows exponentially with p if $\|\theta_{ij}/\alpha_{ij}\|>\varepsilon$. Otherwise, $\|\theta_{ij}/(\varepsilon\,\alpha_{ij})\|^p$ remains bounded by 1 for all p. For $p=\infty$, the last constraint of (12) is therefore equivalent to the requirement that $\|\theta_{ij}\|\le\varepsilon\,\alpha_{ij}$ for all $i\in[N]$ and $j\in[J]$.

An intimate connection between distributionally robust optimization with type-∞ Wasserstein balls and classical robust optimization has first been discovered in [8]. □

Even though problem (12) is guaranteed to have an optimal solution, the worst-case risk (6) may not be attained by any distribution if $p=1$. An instance of problem (6) that fails to be solvable is constructed in Example 4 below, which replicates [60, Example 2].

Example 4 (Non-existence of extremal distributions). Assume that $p=1$, $\Xi=\mathbb{R}$, $N=1$ and $\widehat{\xi}_1=0$ implying that the nominal distribution $\widehat{\mathbb{P}}_1$ reduces to the Dirac distribution

at 0. Set the norm on $\mathbb{R}$ to the absolute value $|\cdot|$, and set $\ell(\xi) = \max\{0, \xi - 1\}$. As $\mathrm{Lip}(\ell) = 1$, Theorem 5 implies that $\mathcal{R}_{\varepsilon,1}(\widehat{\mathbb{P}}_1, \ell) \leq \varepsilon$. Next, define $\mathbb{Q}_n = (1 - 1/n)\,\delta_0 + (1/n)\,\delta_{\varepsilon n}$ for $n \in \mathbb{N}$, and note that the type-1 Wasserstein distance between $\mathbb{Q}_n$ and $\widehat{\mathbb{P}}_1$ amounts to ε, which is the cost of moving mass $1/n$ from εn to 0. Thus, $\mathbb{Q}_n \in \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_1)$. Moreover, we have $\mathbb{E}^{\mathbb{Q}_n}[\ell(\xi)] = \max\{0, \varepsilon - 1/n\}$, which implies that $\mathbb{Q}_n$ attains the upper bound ε on the worst-case risk asymptotically as n tends to infinity. Thus, the sequence $\mathbb{Q}_n$, $n \in \mathbb{N}$, is asymptotically optimal in (6). Next, we argue that the worst-case risk ε is not attained. Suppose to the contrary that there exists $\mathbb{Q}^\star \in \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_1)$ with $\mathbb{E}^{\mathbb{Q}^\star}[\ell(\xi)] = \varepsilon$. Thus, $\varepsilon = \mathbb{E}^{\mathbb{Q}^\star}[\ell(\xi)] < \mathbb{E}^{\mathbb{Q}^\star}[|\xi|] \leq \varepsilon$ where the strict inequality follows from the observation that $\ell(\xi) < |\xi|$ for any $\xi \neq 0$ and that $\mathbb{Q}^\star \neq \delta_0$, and the second inequality follows from Theorem 2 and the assumption that $\mathbb{Q}^\star \in \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_1)$. The contradiction implies that $\mathbb{Q}^\star$ cannot exist, and thus (6) is not solvable. □

Fix now any maximizer $\{\alpha^\star_{ij}, \theta^\star_{ij}\}_{i,j}$ of problem (12). This maximizer can be used to construct an extremal distribution $\mathbb{Q}^\star$ that solves problem (6) (if such a $\mathbb{Q}^\star$ exists) or a sequence of asymptotically optimal distributions $\{\mathbb{Q}_n\}_{n\in\mathbb{N}}$ (if such a $\mathbb{Q}^\star$ does not exist). Before describing this construction, we remark that $\theta^\star_{ij}$ is a recession direction of the support set Ξ whenever $\alpha^\star_{ij} = 0$ (*i.e.*, $\widehat{\xi}_i + t\,\theta^\star_{ij} \in \Xi$ for every $t \geq 0$). If Ξ is bounded, this implies that $\theta^\star_{ij} = 0$ whenever $\alpha^\star_{ij} = 0$. Next, define ν_+ as the set of pairs (i,j) with $\alpha^\star_{ij} > 0$, ν_0 as the set of pairs (i,j) with $\alpha^\star_{ij} = 0$ and $\theta^\star_{ij} = 0$, and ν_∞ as the set of pairs (i,j) with $\alpha^\star_{ij} = 0$ and $\theta^\star_{ij} \neq 0$. By construction, ν_+, ν_0 and ν_∞ form a partition of $[N] \times [J]$.

If $\nu_\infty = \emptyset$, one can show that

$$\mathbb{Q}^\star = \sum_{(i,j)\in\nu_+} \frac{\alpha^\star_{ij}}{N}\, \delta_{\widehat{\xi}_i + \theta^\star_{ij}/\alpha^\star_{ij}}$$

is an extremal distribution that solves (6). For $p > 1$, the last constraint in (12) ensures that $\theta_{ij} = 0$ whenever $\alpha_{ij} = 0$ because otherwise $\alpha_{ij}\|\theta_{ij}/\alpha_{ij}\|^p$ evaluates to ∞. This implies that the set ν_∞ is empty. Thus, for $p > 1$, the worst-case risk (6) of a piecewise concave loss function is always attained by the discrete distribution $\mathbb{Q}^\star$ constructed above.

If $\nu_\infty \neq \emptyset$, which is only possible in the special case $p = 1$, the distributions

$$\mathbb{Q}_n = \sum_{(i,j)\in\nu_+\cup\nu_\infty} \frac{\alpha_{ij}(n)}{N}\, \delta_{\widehat{\xi}_i + \theta^\star_{ij}/\alpha_{ij}(n)} \quad \text{with} \quad \alpha_{ij}(n) = \begin{cases} \alpha^\star_{ij}\left(1 - \frac{|\nu_\infty|}{n}\right) & \text{if } (i,j) \in \nu_+, \\ \frac{1}{n} & \text{if } (i,j) \in \nu_\infty, \end{cases}$$

are feasible and asymptotically optimal in (6) as $n \geq |\nu_\infty|$ tends to infinity. Intuitively, these distributions send some atoms with decaying probabilities to infinity along specific recession directions $\theta^\star_{ij}$, $(i,j) \in \nu_\infty$, of the support set. Note that moving an atom to infinity is possible even when only a finite (type-1) transportation budget is available provided that the probability mass transported is inversely proportional to the transportation distance.

For $p > 1$, atoms can also migrate to infinity at a finite transportation cost provided that their probabilities are inversely proportional to the p^{th} power of the transportation distance. As piecewise concave loss functions grow at most linearly, however, the decay in probability always outweighs the increase in loss. This reasoning provides an intuitive explanation for our insight that $\nu_\infty = \emptyset$ and that the supremum in (6) is always attained for $p > 1$.

Given the promising results for piecewise concave loss functions, it is natural to ask whether the convex reformulations of Theorems 8 and 9 can be generalized. Indeed, it has been discovered that similar results are available for convex (but not piecewise convex) loss functions under the additional condition that there are no support constraints ($\Xi = \mathbb{R}^m$).

Theorem 10 (Convex loss and $p = 1$). *Assume that $\Xi = \mathbb{R}^m$ and that the loss function $\ell(\xi)$ is convex. If $p = 1$ and $\widehat{\mathbb{P}}_N$ is the empirical distribution, then the worst-case risk (6) coincides with the Lipschitz-regularized empirical loss, that is,*

$$\mathcal{R}_{\varepsilon,1}(\widehat{\mathbb{P}}_N, \ell) = \mathcal{R}(\widehat{\mathbb{P}}_N, \ell) + \varepsilon\,\mathrm{Lip}(\ell).$$

For a proof of this result we refer to [60, Theorem 6.3]. Theorem 10 shows that the simple upper bound of Theorem 5 is exact if $p=1$, $\Xi=\mathbb{R}^m$ and the loss function ℓ is convex.

Remark 3 (Computing the Lipschitz modulus). By Theorem 10, computing the worst-case risk of a convex loss function $\ell(\xi)$ requires computing the Lipschitz modulus of $\ell(\xi)$ with respect to the prescribed norm $\|\cdot\|$ on $\mathbb{R}^m$. One can show that

$$\mathrm{Lip}(\ell)=\sup\left\{\|z\|_*:\ell^*(z)<\infty\right\}, \tag{13}$$

that is, the Lipschitz modulus of $\ell(\xi)$ coincides with the radius of the smallest dual norm ball around 0 that encloses the domain of the conjugate loss function $\ell^*(z)$ [60, § 6.2]. Unfortunately, problem (13) maximizes a convex function over a convex set and is therefore hard. More formally, assume that $\ell(\xi)=\mu^\top\xi+\|\Sigma^{\frac{1}{2}}\xi\|_2$ for an arbitrary $\mu\in\mathbb{R}^m$ and $\Sigma\in\mathbb{S}_+^m$. An elementary calculation shows that $\ell^*(z)=0$ if $z\in\mathcal{E}$ and $\ell^*(z)=\infty$ if $z\notin\mathcal{E}$, where $\mathcal{E}=\{\mu+\Sigma^{\frac{1}{2}}u:\|u\|_2\leq 1\}$ stands for the ellipsoid with center μ and shape matrix Σ. Hence, $\mathcal{E}$ is the domain of $\ell^*(z)$. In order to compute the Lipschitz modulus of $\ell(\xi)$ with respect to the ∞-norm, for example, we thus need to solve an instance of problem (13) that maximizes the 1-norm over $\mathcal{E}$. As maximizing the 1-norm over an arbitrary ellipsoid is NP-hard [47, Lemma 4.1], we conclude that the worst-case risk evaluation problem (6) is intractable even for polyhedral norms and for simple classes of (convex) conic quadratic loss functions. □

One can show that the supremum of the worst-case risk evaluation problem (6) is *never* attained under the conditions of Theorem 10, that is, any asymptotically optimal sequence of distributions must push some (decreasing amount of) probability mass to infinity. As in the case of a piecewise concave loss function, such a sequence can be constructed explicitly. To do so, choose a maximizer $z^\star$ of problem (13), which is generally intractable as pointed out in Remark 3. Moreover, select $i_0\in[N]$ and $\xi^\star\in\arg\max_{\|\xi\|\leq 1}\xi^\top z^\star$. Then, the distributions

$$\mathbb{Q}_n=\frac{1}{N}\sum_{i\neq i_0}^{N}\delta_{\widehat{\xi}_i}+\frac{n-1}{nN}\,\delta_{\widehat{\xi}_{i_0}}+\frac{1}{nN}\,\delta_{\widehat{\xi}_{i_0}+\varepsilon nN\xi^\star}$$

can be shown to be feasible and asymptotically optimal in (6) as $n\geq 1$ tends to infinity.

Assume next that $p=2$, the loss function $\ell(\xi)$ is quadratic and the transportation cost in the definition of the Wasserstein distance is induced by the Euclidean norm. Then, the worst-case risk (6) coincides with the optimal value of a tractable semidefinite program (SDP).

Theorem 11 (Indefinite quadratic loss and $p=2$ I). *Assume that $\Xi=\mathbb{R}^m$ and that $\ell(\xi)=\xi^\top Q\xi+2q^\top\xi$ with $Q\in\mathbb{S}^m$ and $q\in\mathbb{R}^m$ is a (possibly indefinite) quadratic loss function. If $p=2$, $\|\cdot\|=\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^m$ and $\widehat{\mathbb{P}}_N$ is the empirical distribution, then the worst-case risk (6) coincides with the optimal value of a tractable SDP, that is,*

$$\begin{array}{rcll}\mathcal{R}_{\varepsilon,2}(\widehat{\mathbb{P}}_N,\ell) & = & \inf & \gamma\varepsilon^2+\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}s_i & \\ & & \text{s.t.} & \gamma\in\mathbb{R}_+,\ s_i\in\mathbb{R}_+ & \forall i\in[N] \\ & & & \begin{bmatrix}\gamma I-Q & q+\gamma\widehat{\xi}_i\\ q^\top+\gamma\widehat{\xi}_i^\top & s_i+\gamma\|\widehat{\xi}_i\|_2^2\end{bmatrix}\succeq 0 & \forall i\in[N]\,.\end{array} \tag{14}$$

Note that substituting the SDP (14) into the distributionally robust decision problem (7) yields a tractable SDP if the set $\mathcal{L}$ of admissible loss functions is defined through SDP constraints in Q and q. In order to construct an extremal distribution that solves problem (6) for a *fixed* convex quadratic loss function, it is useful to derive the dual of the SDP (14).

Theorem 12 (Indefinite quadratic loss and $p=2$ II). *Suppose that all conditions of Theorem 11 hold. If $\lambda_{\max}(Q)$ denotes the largest eigenvalue of Q, then*

$$\begin{array}{rcll} \mathcal{R}_{\varepsilon,2}(\widehat{\mathbb{P}}_N,\ell) & = & \max & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(\widehat{\xi}_i+\theta_i)^\top Q(\widehat{\xi}_i+\theta_i)+2q^\top(\widehat{\xi}_i+\theta_i)+\alpha\lambda_{\max}(Q) \\ & & \text{s.t.} & \alpha\in\mathbb{R}_+,\ \theta_i\in\mathbb{R}^m \quad \forall i\in[N] \\ & & & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\|\theta_i\|_2^2+\alpha\leq\varepsilon^2. \end{array} \tag{15}$$

Problem (15) represents a quadratically constrained quadratic program (QCQP) with a compact feasible set and is therefore solvable. As Q is not necessarily negative semidefinite, problem (15) is generally *nonconvex*. This is perhaps puzzling because (15) is obtained by 'massaging' the dual of (14) and because dual optimization problems are convex by construction. The apparent contradiction is resolved by noting that nonconvex QCQPs of the form (15) with a *single* constraint are equivalent to convex SDPs by virtue of the celebrated $\mathcal{S}$-procedure [14, Appendix B.1].

Intuitively, problem (15) can be interpreted as a *finite reduction* of the worst-case risk evaluation problem (6), which maximizes only over discrete distributions in the Wasserstein ball. Denoting by $v_{\max}(Q)$ an eigenvector corresponding to $\lambda_{\max}(Q)$, any such discrete distribution assigns probability $1/N$ to the perturbed training samples $\widehat{\xi}_i+\theta_i$, $i\in[N]$, and a 'vanishing' probability to an atom located 'infinitely' far away in the direction of $v_{\max}(Q)$. More precisely, the product of the squared transportation distance and the probability of this last atom must converge to a finite value $\alpha\in\mathbb{R}_+$ (hence, the probability of this atom must be asymptotically proportional to the inverse of the squared transportation distance). In the same spirit one can show that if $\alpha^\star$ and $\{\theta_i^\star\}_i$ are optimal in (15) and $i_0\in[N]$, then the discrete distributions

$$\mathbb{Q}_n=\frac{1}{N}\sum_{i\neq i_0}^{N}\delta_{\widehat{\xi}_i+\theta_i^\star}+\frac{n-1}{nN}\,\delta_{\widehat{\xi}_{i_0}+\theta_{i_0}^\star}+\frac{1}{nN}\,\delta_{\widehat{\xi}_{i_0}+\sqrt{nN\alpha^\star}\,v_{\max}(Q)}$$

are feasible and asymptotically optimal in (6) as $n\geq 1$ tends to infinity.

The structure of the extremal distributions for the worst-case risk evaluation problem (6) with general loss functions and nominal distributions as well as necessary and sufficient conditions for their existence have been studied in [38, 74, 111]. The special case of a Wasserstein ball centered at a discrete distribution with N atoms has undergone particular scrutiny. Considerable effort was spent on proving the existence of discrete extremal distributions with as few atoms as possible. A first breakthrough was marked by the insight that the worst-case risk of any continuous bounded loss function is attained by a discrete distribution with at most $N+3$ atoms [111, Theorem 2.3]. As any $(N+3)$-point distribution on $\mathbb{R}^m$ can be encoded by $(N+3)\cdot(m+1)-1$ parameters (*i.e.*, the coordinates and probabilities of the $N+3$ atoms), this result motivates a *finite reduction*: when searching for an extremal distribution, one may restrict attention to discrete distributions supported on $N+3$ points, which amounts to searching a finite-dimensional parameter space. It was later shown that one may actually focus on discrete distributions with $N+2$ atoms [74, Theorem 2.3] or even only $N+1$ atoms [38, Corollary 1] without sacrificing optimality. These sharper results facilitate more parsimonious finite reductions that may be fruitfully used in algorithm design. Exact finite reductions involving fewer atoms are available only in special cases. For example, the discussion after Theorem 9 shows that the worst-case risk of a *concave* loss function is always attained by an N-point distribution. For more general loss functions, however, every N-point distribution may be suboptimal even if the worst-case risk is attained.

Example 5 (Non-existence of extremal distributions with N atoms). Suppose that $\Xi = (-\infty, 2]$, $N = 1$ and $\widehat{\xi}_1 = 0$, which implies that $\widehat{\mathbb{P}}_1$ is the Dirac distribution at 0. Set the norm on $\mathbb{R}$ to the absolute value $|\cdot|$, select $\varepsilon \in (0,2)$ and set $\ell(\xi) = \max\{0, \xi - 1\}$. Next, define $\mathbb{Q}^\star = (1 - \varepsilon/2)\,\delta_0 + (\varepsilon/2)\,\delta_2$, and note that the type-1 Wasserstein distance between $\widehat{\mathbb{P}}_1$ and $\mathbb{Q}^\star$ amounts to ε, which is the cost of moving mass $\varepsilon/2$ from 2 to 0. Thus, $\mathbb{Q}^\star \in \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_1)$. Moreover, we have $\mathbb{E}^{\mathbb{Q}^\star}[\ell(\xi)] = \varepsilon/2$, which provides a lower bound on the worst-case risk. In fact, by solving problem (12) one can show that $\mathbb{Q}^\star$ is optimal in (6). Any one-point distribution δ_z resides in the Wasserstein ball of radius ε only if $|z| \leq \varepsilon$, and therefore the maximum risk that any one-point distribution can attain is $\max\{0, \varepsilon - 1\}$, which is strictly smaller than $\varepsilon/2$ for any $\varepsilon \in (0,2)$. Thus, no one-point distribution can be extremal. □

If the worst-case risk over a Wasserstein ball centered at the empirical distribution is attained, then there always exists an extremal distribution with $N + 1$ atoms that can be characterized in quasi-closed form [38, Corollary 2]. In practice, however, it is often convenient to ignore this minimal representability and to search over candidate distributions with more than $N + 1$ atoms, *e.g.*, by solving a finite convex optimization problem such as (9). For generic nominal distributions, necessary and sufficient conditions for the existence of an extremal distribution are detailed in [38, Corollary 1].

2.3. Tractability Results for Elliptical Nominal Distributions

We will now demonstrate that the worst-case risk evaluation problem (6) and the distributionally robust decision problem (7) sometimes admit exact tractable reformulations or conservative tractable approximations even if the nominal distribution $\widehat{\mathbb{P}}$ is continuous. To show this, we assume throughout this section that $\widehat{\mathbb{P}}_N$ has mean vector $\widehat{\mu} \in \mathbb{R}^m$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^m$. Thus, we implicitly assume that $\widehat{\mathbb{P}}_N$ has finite second-order moments.

We first define an uncertainty set in the space of mean vectors and covariance matrices.

$$\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = \left\{ (\mu, \Sigma) \in \mathbb{R}^m \times \mathbb{S}_+^m : \|\widehat{\mu} - \mu\|_2^2 + \mathrm{Tr}\left[\widehat{\Sigma} + \Sigma - 2\left(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}} \right)^{\frac{1}{2}} \right] \leq \varepsilon^2 \right\}$$

This uncertainty set is of interest because it covers the projection of the type-2 Wasserstein ball $\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)$ onto the space of mean vectors and covariance matrices. Moreover, if the nominal distribution is elliptical, $\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ is actually equal to the projection of $\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)$.

Proposition 1 (Projection of $\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)$ onto the mean-covariance space). *If $\widehat{\mathbb{P}}_N$ has mean vector $\widehat{\mu} \in \mathbb{R}^m$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^m$, then*

$$\left\{ \left(\mathbb{E}^{\mathbb{Q}}[\xi], \mathbb{E}^{\mathbb{Q}}[(\xi - \mathbb{E}^{\mathbb{Q}}[\xi])(\xi - \mathbb{E}^{\mathbb{Q}}[\xi])^\top] \right) : \mathbb{Q} \in \mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N) \right\} \subseteq \mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}).$$

The inclusion becomes an equality if $\Xi = \mathbb{R}^m$ and $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}, \widehat{\Sigma})$ is an elliptical distribution.

Proposition 1 follows immediately from Theorem 4. The condition $\Xi = \mathbb{R}^m$ ensures that any elliptical distribution $\mathbb{Q} = \mathcal{E}_g(\mu, \Sigma)$ with the same density generator as the nominal distribution and with $W_2(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon$ belongs to $\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)$. One can show that $\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ is convex and compact [98, Lemma A.6], which is expected as it is a projection of a (Wasserstein) ball.

The uncertainty set $\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ can conveniently be used in classical robust optimization. Indeed, a robust constraint that requires a concave function $h(\mu, \Sigma)$ to be nonpositive for all $(\mu, \Sigma) \in \mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ can be reformulated as a convex constraint that involves the conjugate of $-h(\mu, \Sigma)$ and the support function of the uncertainty set $\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ [3, Theorem 2], that is,

$$h(\mu, \Sigma) \leq 0 \quad \forall (\mu, \Sigma) \in \mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) \quad \Longleftrightarrow \quad \begin{cases} \exists q \in \mathbb{R}^m,\, Q \in \mathbb{S}^m : \\ (-h)^*(-q, -Q) + \sigma_{\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})}(q, Q) \leq 0. \end{cases}$$

This constraint is computationally tractable for many commonly used constraint functions because the support function of $\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ is SDP-representable [67].

Lemma 1 (Support function of $\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$). *The support function of $\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ coincides with the optimal value of a tractable SDP, that is, for any $q\in\mathbb{R}^m$ and $Q\in\mathbb{S}^m$ we have*

$$\begin{array}{rcll}\sigma_{\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})}(q,Q) & = & \inf & q^\top\widehat{\mu}+\tau\,\|q\|_2^2+\gamma\left(\varepsilon^2-\mathrm{Tr}\,[\widehat{\Sigma}]\right)+\mathrm{Tr}\,[Z] \\ & & \text{s.t.} & \gamma\in\mathbb{R}_+,\ \tau\in\mathbb{R}_+,\ Z\in\mathbb{S}_+^m \\ & & & \begin{bmatrix}\gamma I-Q & \gamma\widehat{\Sigma}^{\frac{1}{2}}\\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z\end{bmatrix}\succeq 0,\quad \left\|\begin{pmatrix}1\\ \tau-\gamma\end{pmatrix}\right\|_2\leq\tau+\gamma.\end{array}$$

Unlike the mean vector $\mu=\mathbb{E}^{\mathbb{Q}}[\xi]$ and the second-order moment matrix $M=\mathbb{E}^{\mathbb{Q}}[\xi\xi^\top]$, both of which constitute linear functions of the underlying distribution $\mathbb{Q}$, the covariance matrix $\Sigma=M-\mu\mu^\top$ is nonlinear in $\mathbb{Q}$. The condition $(\mu,\Sigma)\in\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ thus appears to be nonconvex in $\mathbb{Q}$. To gain a clearer understanding, it is instructive to introduce the uncertainty set $\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ for (μ,M) induced by the uncertainty set $\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ for (μ,Σ), that is,

$$\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})=\left\{(\mu,M)\in\mathbb{R}^m\times\mathbb{S}_+^m:(\mu,M-\mu\mu^\top)\in\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})\right\}.$$

Maybe surprisingly, even though it is defined as the pre-image of a convex set under a *non*linear transformation, one can prove that $\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ is convex. This implies, counterintuitively, that the condition $(\mu,\Sigma)\in\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ is actually convex in $\mathbb{Q}$ because it is equivalent to the requirement $(\mu,M)\in\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ and because the moments (μ,M) are linear in $\mathbb{Q}$.

Thanks to its convexity, the uncertainty set $\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ can again conveniently be used in classical robust optimization. Indeed, a robust constraint that requires a concave function $h(\mu,M)$ to be nonpositive for all $(\mu,M)\in\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ can be reformulated as a simple convex constraint involving the conjugate of $-h(\mu,M)$ and the support function of $\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$. This constraint is computationally tractable for many commonly used constraint functions because the support function of $\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ is SDP-representable [67].

Lemma 2 (Support function of $\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$). *The support function of $\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ coincides with the optimal value of a tractable SDP, that is, for any $q\in\mathbb{R}^m$ and $Q\in\mathbb{S}^m$, we have*

$$\begin{array}{rcll}\sigma_{\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})}(q,Q) & = & \inf & \gamma(\varepsilon^2-\|\widehat{\mu}\|_2^2-\mathrm{Tr}[\widehat{\Sigma}])+z+\mathrm{Tr}\,[Z] \\ & & \text{s.t.} & \gamma\in\mathbb{R}_+,\ z\in\mathbb{R}_+,\ Z\in\mathbb{S}_+^m \\ & & & \begin{bmatrix}\gamma I-Q & \gamma\widehat{\Sigma}^{\frac{1}{2}}\\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z\end{bmatrix}\succeq 0,\quad \begin{bmatrix}\gamma I-Q & \gamma\widehat{\mu}+\frac{q}{2}\\ (\gamma\widehat{\mu}+\frac{q}{2})^\top & z\end{bmatrix}\succeq 0.\end{array}\tag{16}$$

A useful ambiguity set in the space of probability distributions is the *Gelbrich hull*, which is constructed as the pre-image of $\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ under the mean-covariance projection.

Definition 2 (Gelbrich hull). The Gelbrich hull is given by

$$\mathbb{G}_\varepsilon(\widehat{\mu},\widehat{\Sigma})=\left\{\mathbb{Q}\in\mathcal{P}(\Xi):\left(\mathbb{E}^{\mathbb{Q}}[\xi],\mathbb{E}^{\mathbb{Q}}[(\xi-\mathbb{E}^{\mathbb{Q}}[\xi])(\xi-\mathbb{E}^{\mathbb{Q}}[\xi])^\top]\right)\in\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})\right\}.$$

By definition, $\mathbb{G}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ contains all distributions supported on Ξ whose mean vectors and covariance matrices fall into the uncertainty set $\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$. Equivalently, by the definition of the induced uncertainty set $\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$, the Gelbrich hull can also be represented as

$$\mathbb{G}_\varepsilon(\widehat{\mu},\widehat{\Sigma})=\left\{\mathbb{Q}\in\mathcal{P}(\Xi):\left(\mathbb{E}^{\mathbb{Q}}[\xi],\mathbb{E}^{\mathbb{Q}}[\xi\xi^\top]\right)\in\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})\right\}.$$

Thus, the Gelbrich hull can be expressed as the pre-image of the convex set $\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ under a linear transformation, which shows that is is actually convex. We emphasize that convexity is not apparent from Definition 2, which introduces the Gelbrich hull as the pre-image of a convex set under a *non*linear transformation.

If we define $\mathcal{P}(\Xi, \mu, \Sigma)$ as the *Chebyshev ambiguity set* that contains all distributions on Ξ with mean vector μ and covariance matrix Σ, then the Gelbrich hull can also be expressed as

$$\mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = \bigcup_{(\mu,\Sigma)\in\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})} \mathcal{P}(\Xi, \mu, \Sigma). \tag{17}$$

From this representation it is evident that if $\mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ contains a distribution $\mathbb{Q}$, then it contains *all* distributions on Ξ that have the same mean vector and covariance matrix as $\mathbb{Q}$. It is easy to verify that the Gelbrich hull provides an outer approximation for any Wasserstein ball $\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)$ with $p \geq 2$. Indeed, if $\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)$ contains a distribution $\mathbb{Q}$ with mean vector μ and covariance matrix Σ, then $(\mu, \Sigma) \in \mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ by virtue of Proposition 1, which implies via (17) that $\mathbb{Q} \in \mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$. These insights culminate in the following theorem.

Theorem 13 (Gelbrich hull). *If the nominal distribution $\widehat{\mathbb{P}}_N$ has mean vector $\widehat{\mu} \in \mathbb{R}^m$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^m$, then we have $\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N) \subseteq \mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ for every $p \geq 2$.*

Theorem 13 shows that the Gelbrich hull provides an outer approximation for all Wasserstein balls $\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)$ with $p \geq 2$ solely on the basis of mean and covariance information. Discarding all information about $\widehat{\mathbb{P}}_N$ beyond its first- and second-order moments can be seen as a compression of the training dataset. This amounts to sacrificing higher-order moment information and may improve the tractability of the risk evaluation problem (6) and the distributionally robust decision problem (7). To show this, we define the *Gelbrich risk* as

$$\overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) = \sup_{\mathbb{Q}\in\mathbb{G}_\varepsilon(\widehat{\mu},\widehat{\Sigma})} \mathcal{R}(\mathbb{Q}, \ell) \tag{18}$$

and the *optimal Gelbrich risk* as

$$\overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \mathcal{L}) = \inf_{\ell\in\mathcal{L}} \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell). \tag{19}$$

Theorem 13 immediately implies that the (optimal) Gelbrich risk provides an upper bound on the (optimal) worst-case risk whenever $p \geq 2$.

Corollary 1 (Gelbrich risk). *If the nominal distribution $\widehat{\mathbb{P}}_N$ has mean vector $\widehat{\mu} \in \mathbb{R}^m$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^m$ and if $p \geq 2$, then*

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \leq \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) \quad \forall \ell \in \mathcal{L} \qquad \text{and} \qquad \mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \mathcal{L}) \leq \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \mathcal{L}).$$

The representation (17) of the Gelbrich hull as a union of Chebyshev ambiguity sets suggests that the Gelbrich risk of any fixed loss function $\ell(\xi)$ can be expressed as the optimal value of the following two-layer optimization problem [67].

$$\overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) = \sup_{(\mu,\Sigma)\in\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathcal{P}(\Xi,\mu,\Sigma)} \mathcal{R}(\mathbb{Q}, \ell) \tag{20a}$$

$$= \sup_{(\mu,M)\in\mathcal{V}_\varepsilon(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathcal{P}(\Xi,\mu,M-\mu\mu^\top)} \mathcal{R}(\mathbb{Q}, \ell) \tag{20b}$$

Note that (20b) follows immediately from the definition of the uncertainty set $\mathcal{V}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ and the formula for the covariance matrix in terms of the mean vector and the second-order moment matrix. The inner problems in (20a) and (20b) both represent the same distributionally robust optimization problem over a Chebyshev ambiguity set but with different parameterizations. This problem can be viewed as an infinite-dimensional linear program over all probability distributions $\mathbb{Q}$ that satisfy the linear equality constraints $\mathbb{E}^{\mathbb{Q}}[\xi] = \mu$ and $\mathbb{E}^{\mathbb{Q}}[\xi\xi^\top] = M$. Therefore, the optimal value of the inner maximization problem is concave in the right hand side parameters μ and M but generally nonconcave in the alternative

parameters μ and Σ. The outer problem in (20a) hedges against ambiguity in the mean vector and the covariance matrix, while the one in (20b) hedges against ambiguity in the first- and second-order moments. The formulation (20a) is conceptually appealing because of its connection to the Wasserstein distance and because it is more natural to characterize a distribution in terms of its mean vector and covariance matrix. The formulation (20b), on the other hand, is computationally attractive because it expresses the outer problem as a convex program that maximizes a manifestly concave function over the convex set $\mathcal{V}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$.

Remark 4 (Second layer of robustness). Distributionally robust optimization problems akin to (20a) and (20b) that accommodate a second layer of robustness to account for moment ambiguity have been investigated in [29, 42, 47, 66, 91, 116], among others. As the optimal value of the inner maximization problem is always concave in (μ, M) but typically nonconcave in (μ, Σ), moment ambiguity has mostly been modeled through convex uncertainty sets for (μ, M), thereby ensuring convexity of the outer maximization problem. For example, uncertainty sets that force μ to lie in an ellipsoid and M in the intersection of two positive semi-definite cones were studied in [29], while box-type uncertainty sets for (μ, M) were proposed in [66] and refined in [47, 116]. Convex uncertainty sets for (μ, Σ) were shown to render the outer maximization problems convex only in special cases, *e.g.*, when evaluating a worst-case value-at-risk of a linear or quadratic loss function [42, 91]. The convex uncertainty set $\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ for (μ, Σ) is remarkable because it leads to a second-layer maximization problem in (20a) that admits a convex reformulation for *all* loss functions $\ell(\xi)$. □

The decomposition (20b) of the Gelbrich risk evaluation problem into two consecutive maximization problems offers a systematic approach to derive convex reformulations for (18). A tractable SDP reformulation is available, for example, when the loss function $\ell(\xi)$ is a pointwise maximum of finitely many (possibly indefinite) quadratic functions.

Theorem 14 (Piecewise quadratic loss I). *Assume that $\Xi = \mathbb{R}^m$ and $\varepsilon > 0$ and that $\ell(\xi) = \max_{j\in[J]}\{\xi^\top Q_j \xi + 2q_j^\top \xi + q_j^0\}$ with $Q_j \in \mathbb{S}^m$, $q_j \in \mathbb{R}^m$, and $q_j^0 \in \mathbb{R}$ for any $j \in [J]$ is a piecewise quadratic loss function. If $\widehat{\mu} \in \mathbb{R}^m$ and $\widehat{\Sigma} \in \mathbb{S}_+^m$, then the Gelbrich risk (18) is equal to the optimal value of a tractable SDP, that is,*

$$\begin{array}{rcl} \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) & = & \inf \;\; y_0 + \gamma\left(\varepsilon^2 - \|\widehat{\mu}\|_2^2 - \mathrm{Tr}[\widehat{\Sigma}]\right) + z + \mathrm{Tr}\,[Z] \\ & & \text{s.t. } \gamma \in \mathbb{R}_+,\, y_0 \in \mathbb{R},\, y \in \mathbb{R}^m,\, Y \in \mathbb{S}^m,\, z \in \mathbb{R}_+,\, Z \in \mathbb{S}_+^m \\ & & \quad \begin{bmatrix} \gamma I - Y & y + \gamma\widehat{\mu} \\ y^\top + \gamma\widehat{\mu}^\top & z \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \gamma I - Y & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0 \\ & & \quad \begin{bmatrix} Y - Q_j & y - q_j \\ y^\top - q_j^\top & y_0 - q_j^0 \end{bmatrix} \succeq 0 \quad \forall j \in [J]. \end{array} \tag{21}$$

In order to construct an extremal distribution for the Gelbrich risk evaluation problem (18), it is again expedient to derive the dual of the SDP (21).

Theorem 15 (Piecewise quadratic loss II). *If all conditions of Theorem 14 hold, then we have*

$$\begin{array}{rcl} \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) = \max & \displaystyle\sum_{j=1}^J \mathrm{Tr}\,[Q_j\Theta_j] + 2q_j^\top\theta_j + q_j^0\alpha_j & \\ \text{s.t.} & \mu \in \mathbb{R}^m,\, \Sigma \in \mathbb{S}_+^m,\, \alpha_j \in \mathbb{R}_+,\, \theta_j \in \mathbb{R}^m,\, \Theta_j \in \mathbb{S}_+^m \quad \forall j \in [J] & \\ & \begin{bmatrix} \Theta_j & \theta_j \\ \theta_j^\top & \alpha_j \end{bmatrix} \succeq 0 \quad \forall j \in [J] & \\ & \displaystyle\sum_{j=1}^J \alpha_j = 1,\;\; \sum_{j=1}^J \theta_j = \mu,\;\; \sum_{j=1}^J \Theta_j = \Sigma + \mu\mu^\top,\;\; (\mu, \Sigma) \in \mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}). & \end{array} \tag{22}$$

Note that problem (22) has a continuous objective function as well as a compact feasible set and is therefore solvable. Any optimal solution $(\mu^\star, \Sigma^\star, \{\alpha_j^\star, \theta_j^\star, \Theta_j^\star\}_j)$ can in principle be used to construct an extremal distribution $\mathbb{Q}^\star$ that attains the supremum in the Gelbrich risk evaluation problem (18). Specifically, for any $j \in [J]$ let $\mathbb{Q}_j^\star$ be any distribution supported on

$$\Xi_j = \left\{\xi \in \mathbb{R}^m : \xi^\top Q_j \xi + 2q_j^\top \xi + q_j^0 \geq \xi^\top Q_{j'} \xi + 2q_{j'}^\top \xi + q_{j'}^0 \;\; \forall j' \neq j \right\}.$$

If $\alpha_j^\star > 0$, we impose the additional requirement that $\mathbb{Q}_j^\star$ has mean value $\theta_j^\star/\alpha_j^\star$ and second-order moment matrix $\Theta_j^\star/\alpha_j^\star$. Such a distribution is indeed guaranteed to exist. One can then show that the mixture distribution $\mathbb{Q}^\star = \sum_{j\in[J]} \alpha_j^\star \cdot \mathbb{Q}_j^\star$ is optimal in (18). By construction, this distribution $\mathbb{Q}^\star$ has mean vector $\mu^\star$ and covariance matrix $\Sigma^\star$. We emphasize that problem (22) can be reformulated as a tractable SDP by applying the variable substitution $M \leftarrow \Sigma + \mu\mu^\top$, replacing the constraint $(\mu, \Sigma) \in \mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ with $(\mu, M) \in \mathcal{V}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ and recalling from Lemma 2 that the uncertainty set $\mathcal{V}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ is SDP-representable. Thus, problem (22) can be solved in polynomial time. Even though the mixture components $\mathbb{Q}_j^\star$, $j \in [J]$, are guaranteed to exist, however, one can prove that it is NP-hard to construct them. In other words, even though it is easy to solve (22) and even though any solution of (22) gives rise to a solution $\mathbb{Q}^\star$ of the Gelbrich risk evaluation problem (18), constructing $\mathbb{Q}^\star$ remains hard.

While exactly computable in polynomial time, the Gelbrich risk of a piecewise quadratic loss function may only provide a loose upper bound on the worst-case risk under the Wasserstein ambiguity set, which is often the actual quantity of interest. One can prove, however, that the Gelbrich risk (18) coincides with the worst-case risk (6) with respect to a type-2 Wasserstein ball if the loss function is quadratic and the nominal distribution is elliptical.

Theorem 16 (Indefinite quadratic loss I). *Assume that $\Xi = \mathbb{R}^m$ and that $\ell(\xi) = \xi^\top Q \xi + 2q^\top \xi$ with $Q \in \mathbb{S}^m$ and $q \in \mathbb{R}^m$ is a quadratic loss function. If $\widehat{\mu} \in \mathbb{R}^m$ and $\widehat{\Sigma} \in \mathbb{S}_+^m$, then the Gelbrich risk (18) is equal to the optimal value of a tractable SDP, that is,*

$$\begin{array}{rcl} \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) & = & \inf \;\; \gamma \left(\varepsilon^2 - \|\widehat{\mu}\|_2^2 - \mathrm{Tr}\,[\widehat{\Sigma}]\right) + z + \mathrm{Tr}\,[Z] \\ & & \text{s.t. } \gamma \in \mathbb{R}_+,\; z \in \mathbb{R}_+,\; Z \in \mathbb{S}_+^m \\ & & \begin{bmatrix} \gamma I - Q & q + \gamma\widehat{\mu} \\ q^\top + \gamma\widehat{\mu}^\top & z \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \gamma I - Q & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0. \end{array} \tag{23}$$

Moreover, if $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}, \widehat{\Sigma})$ is an elliptical distribution with mean vector $\widehat{\mu}$ and covariance matrix $\widehat{\Sigma}$, $p = 2$ and $\|\cdot\| = \|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^m$, then the worst-case risk (6), the Gelbrich risk (18) and the optimal value of the SDP (23) are all equal.

The SDP (23) is easily obtained from (21) by setting $J = 1$ and noting that $Y = Q_1$, $y = q_1$ and $y_0 = 0$ at optimality. As usual, a discrete extremal distribution $\mathbb{Q}^\star$ for the Gelbrich risk evaluation problem (18) can be derived from the dual of the SDP (23). In the following we denote the mean vector and the covariance matrix of $\mathbb{Q}^\star$ by $\mu^\star$ and $\Sigma^\star$, respectively. As $\mathbb{Q}^\star \in \mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$, and as the Gelbrich hull is constructed solely on the basis of first- and second-order moment information, *any* distribution with mean vector $\mu^\star$ and covariance matrix $\Sigma^\star$ belongs to $\mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$, too. Moreover, as $\ell(\xi)$ is quadratic, the risk $\mathcal{R}(\mathbb{Q}^\star, \ell)$ depends on $\mathbb{Q}^\star$ only through its first- and second-order moments. This implies that *any* distribution with mean vector $\mu^\star$ and covariance matrix $\Sigma^\star$ is optimal in (18).

Consider now the problem of evaluating the worst-case risk (6) of the quadratic loss function $\ell(\xi)$ over a type-2 Wasserstein ball centered at an elliptical nominal distribution $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}, \widehat{\Sigma})$. Theorem 4 ensures that all elliptical distributions in the Gelbrich hull with the same density generator as $\widehat{\mathbb{P}}_N$ belong to the Wasserstein ball $\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)$. This implies that the special elliptical distribution $\mathbb{Q}^\star = \mathcal{E}_g(\mu^\star, \Sigma^\star)$ is feasible in (6). Moreover, we have

$$\overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) = \mathcal{R}(\mathbb{Q}^\star, \ell) \leq \mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \leq \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell),$$

where the equality holds because $\mathbb{Q}^\star$ is optimal in the Gelbrich risk evaluation problem (18), while the two inequalities follow from the feasibility of $\mathbb{Q}^\star$ in the worst-case risk evaluation problem (6) and Corollary 1, respectively. Thus, all inequalities in the above expression are exact, which implies that $\mathbb{Q}^\star$ is actually optimal in (6).

Next, we show how $\mathbb{Q}^\star$ can be constructed from the optimality conditions of the SDP (23).

Theorem 17 (Indefinite quadratic loss II). *If all conditions of Theorem 16 hold, $\widehat{\Sigma} \succ 0$ and there exists $\gamma^\star \geq 0$ with $\gamma^\star I \succ Q$ that solves the nonlinear algebraic equation*

$$\|\widehat{\mu} - (\gamma I - Q)^{-1}(q + \gamma\widehat{\mu})\|_2^2 + \mathrm{Tr}\left[\widehat{\Sigma}\left(I - \gamma(\gamma I - Q)^{-1}\right)^2\right] = \varepsilon^2, \tag{24}$$

then the Gelbrich risk (18) is attained by any distribution with mean vector

$$\mu^\star = (\gamma^\star I - Q)^{-1}(\gamma^\star\widehat{\mu} + q) \tag{25a}$$

and covariance matrix

$$\Sigma^\star = (\gamma^\star)^2(\gamma^\star I - Q)^{-1}\widehat{\Sigma}(\gamma^\star I - Q)^{-1}. \tag{25b}$$

Moreover, if $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}, \widehat{\Sigma})$ is elliptical, $p = 2$ and $\|\cdot\| = \|\cdot\|_2$ is the Euclidean norm, then the elliptical distribution $\mathbb{Q}^\star = \mathcal{E}_g(\mu^\star, \Sigma^\star)$ attains the worst-case risk in (6).

One can show that if $Q \succeq 0$, then $\gamma^\star$ exists and $\Sigma^\star \succeq \lambda_{\min}(\widehat{\Sigma})I$. To give an intuition for Theorem 17, note that the SDP (23) can be converted to an equivalent nonlinear program (NLP) in the single decision variable γ by using Schur complements to show that

$$z = (q + \gamma\widehat{\mu})^\top(\gamma I - Q)^{-1}(q + \gamma\widehat{\mu}) \quad \text{and} \quad Z = \gamma^2\,\widehat{\Sigma}^{\frac{1}{2}}(\gamma I - Q)^{-1}\widehat{\Sigma}^{\frac{1}{2}}$$

at optimality. The resulting NLP minimizes a strictly convex objective function that explodes as γ drops to $\lambda_{\max}(Q)$ or as γ tends to infinity. Equation (24) represents its first-order optimality condition, whose unique solution $\gamma^\star$ can be computed efficiently to any precision via bisection or the Newton-Raphson method. Using (24), one can then show that any distribution with mean vector $\mu^\star$ and covariance matrix $\Sigma^\star$ as defined in (25a) and (25b), respectively, is indeed feasible and optimal in (18).

It is instructive to contrast Theorem 16 with Theorem 11, both of which provide exact tractable SDP reformulations for the problem of evaluating the worst-case risk of a quadratic loss function with respect to a type-2 Wasserstein ball. We highlight that the SDP (23) derived in Theorem 16 for *elliptical* nominal distributions accommodates only two linear matrix inequalities, while the SDP (14) derived in Theorem 11 for *empirical* nominal distributions involves N linear matrix inequalities and may thus be considerably harder to solve.

3. Performance Guarantees

We now argue that for judiciously calibrated Wasserstein ambiguity sets, the worst-case risk (6) associated with a finite sample size N provides an upper confidence bound on the true risk (1) for all admissible loss functions (finite sample guarantee) and that the worst-case optimal risk (7) converges almost surely to the true optimal risk (2) as N tends to infinity (asymptotic guarantee). Intuitively, the finite sample guarantee ensures that the out-of-sample risk will fall short of the worst-case risk with high confidence when we implement an optimizer of the distributionally robust decision probelm (7), while the asymptotic guarantee formalizes the simple intuition that more data enables us to make better decisions.

Concentration inequalities for the nominal distribution $\widehat{\mathbb{P}}_N$ and its moments can be used to derive finite sample and asymptotic guarantees. If $\widehat{\mathbb{P}}_N$ is the empirical distribution, for instance, one can prove that $\widehat{\mathbb{P}}_N$ converges exponentially fast to the data-generating distribution $\mathbb{P}$, in probability with respect to the Wasserstein distance, as N tends to infinity.

Theorem 18 (Concentration inequalities I). *Suppose that $\widehat{\mathbb{P}}_N$ is the empirical distribution, while $p \neq m/2$, and the unknown true distribution $\mathbb{P}$ is light-tailed in the sense that there exist $\alpha > p$ and $A > 0$ such that $\mathbb{E}^{\mathbb{P}}[\exp(\|\xi\|^\alpha)] \leq A$. Then, there are constants $c_1, c_2 > 0$ that depend on $\mathbb{P}$ only through α, A, and m such that for any $\eta \in (0,1]$ the concentration inequality $\mathbb{P}^N[\mathbb{P} \in \mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)] \geq 1-\eta$ holds whenever ε exceeds*

$$\varepsilon_{p,N}(\eta) = \begin{cases} \left(\dfrac{\log(c_1/\eta)}{c_2 N}\right)^{\min\{p/m,1/2\}} & \text{if } N \geq \dfrac{\log(c_1/\eta)}{c_2}, \\ \left(\dfrac{\log(c_1/\eta)}{c_2 N}\right)^{p/\alpha} & \text{if } N < \dfrac{\log(c_1/\eta)}{c_2}. \end{cases} \tag{26}$$

Theorem 18 generalizes [60, Theorem 3.5] to arbitrary $p \geq 1$ and is a direct consequence of [36, Theorem 2]. The result remains valid for $p = m/2$ but with a more complicated formula for $\varepsilon_{p,n}(\eta)$ [36, Theorem 2]. Intuitively, Theorem 18 asserts that any Wasserstein ball $\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)$ with radius $\varepsilon \geq \varepsilon_{p,N}(\eta)$ represents a $(1-\eta)$-confidence set for the unknown data-generating distribution $\mathbb{P}$. For large uncertainty dimensions $m > 2p$, the critical radius $\varepsilon_{p,N}(\eta)$ of this confidence set decays as $\mathcal{O}(N^{-\frac{p}{m}})$. To reduce the critical radius by 50%, the sample size must increase by $2^{\frac{m}{p}}$. Unfortunately, this curse of dimensionality is fundamental, and the decay rate of $\varepsilon_{p,N}(\eta)$ is essentially optimal; see [36, § 1.3] or [109].

The concentration inequality portrayed in Theorem 18 gives rise to the following finite sample guarantees [60, Theorem 3.5].

Theorem 19 (Finite sample guarantees I). *Assume that all conditions of Theorem 18 hold and $\varepsilon_{p,N}(\eta)$ is defined as in (26). Then, for all $\eta \in (0,1)$ and $\varepsilon \geq \varepsilon_N(\eta)$ we have*

$$\mathbb{P}^N\Big\{\mathcal{R}(\mathbb{P},\ell) \leq \mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N,\ell) \quad \forall \ell \in \mathcal{L}\Big\} \geq 1-\eta. \tag{27a}$$

Moreover, if $\ell^\star$ is an optimizer of the distributionally robust decision problem (7), which is a function of the training samples, then for all $\eta \in (0,1)$ and $\varepsilon \geq \varepsilon_N(\eta)$ we have

$$\mathbb{P}^N\Big\{\mathcal{R}(\mathbb{P},\ell^\star) \leq \mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N,\ell^\star)\Big\} \geq 1-\eta. \tag{27b}$$

Theorem (19) asserts that the worst-case risk (6) provides an upper confidence bound on the true risk (1) under the unknown data-generating distribution uniformly across all loss functions $\ell \in \mathcal{L}$. Moreover, it also asserts that the optimal value of the distributionally robust decision problem (7) (*i.e.*, the worst-case optimal risk) provides an upper confidence bound on the out-of-sample performance of its optimizers. Note that the probabilities in (27a) and (27b) are evaluated under the distribution $\mathbb{P}^N$ of the N independent training samples.

Remark 5 (Improved finite sample guarantees). Requiring the Wasserstein ball to cover $\mathbb{P}$ with high confidence is only a sufficient but not a necessary condition for the finite sample guarantees (27a) and (27b). Indeed, these guarantees can be sustained even if the Wasserstein radius is reduced below $\varepsilon_{p,N}(\eta)$, which is essentially the smallest radius for which the Wasserstein ball represents a $(1-\eta)$-confidence set for $\mathbb{P}$. The minimal Wasserstein radius that preserves the finite sample guarantees (27a) and (27b) often decays significantly faster than $\mathcal{O}(N^{-\frac{p}{m}})$ without suffering from a curse of dimensionality. If $p = 1$, the data-generating distribution is absolutely continuous with respect to the Lebesgue measure and the set $\mathcal{L}$ of admissible loss functions admits a smooth parameterization, for example, one can show that a Wasserstein radius of the order $\mathcal{O}(\sqrt{\log m/N})$ maintains finite sample guarantees akin to (27a) and (27b), which is consistent with recent findings in the compressed sensing and high-dimensional statistics literature [10, Theorem 1]. □

As the number N of training samples grows, one can simultaneously reduce the Wasserstein radius ε and the significance level η without sacrificing the finite sample guarantees (27a) and (27b), which allows us to prove asymptotic consistency [60, Theorem 3.6].

Theorem 20 (Asymptotic consistency I). *Assume that all conditions of Theorem 18 hold. Select* $\eta_N \in (0,1]$ *and set* $\varepsilon_N = \varepsilon_{p,N}(\eta_N)$ *as in (26),* $N \in \mathbb{N}$*, such that* $\sum_{N=1}^{\infty} \eta_N < \infty$ *and* $\lim_{N\to\infty} \varepsilon_N = 0$.[1] *If there exists* $C > 0$ *with* $|\ell(\xi)| \leq C(1 + \|\xi\|^p)$ *for all* $\ell \in \mathcal{L}$ *and* $\xi \in \Xi$*, then we have* $\mathbb{P}^\infty$*-almost surely that* $\mathcal{R}_{\varepsilon_N,p}(\widehat{\mathbb{P}}_N, \mathcal{L}) \to \mathcal{R}(\mathbb{P}, \mathcal{L})$ *as* N *tends to infinity.*

Next, we describe a concentration inequality for the sample mean and the sample covariance matrix that has ramifications for the Gelbrich risk minimization problem (19).

Theorem 21 (Concentration inequalities II). *Suppose that the unknown true distribution* $\mathbb{P}$ *has mean vector* μ *and covariance matrix* Σ *and that there are* $\alpha > 2$ *and* $A > 0$ *such that* $\mathbb{E}^{\mathbb{P}}[\exp(\|\xi\|_2^\alpha)] \leq A$*. Then, there is* $c > 1$ *that depends on* $\mathbb{P}$ *only through* μ*,* Σ*,* α*,* A*, and* m *such that for any* $\eta \in (0,1]$ *the sample mean* $\widehat{\mu}$ *and the sample covariance matrix* $\widehat{\Sigma}$ *satisfy the concentration inequality* $\mathbb{P}^N[(\mu,\Sigma) \in \mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})] \geq 1-\eta$ *whenever* ε *exceeds*

$$\varepsilon_N(\eta) = \frac{\log(c/\eta)}{\sqrt{N}}. \tag{28}$$

Theorem 21 asserts that the uncertainty set $\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})$ with radius $\varepsilon \geq \varepsilon_N(\eta)$ represents a $(1-\eta)$-confidence set for the mean vector and covariance matrix of the unknown data-generating distribution $\mathbb{P}$. The critical radius $\varepsilon_N(\eta)$ of this confidence set decays as $\mathcal{O}(N^{-\frac{1}{2}})$ and is therefore—unlike the critical radius (26)—*not* subject to a curse of dimensionality.

Theorem 21 strengthens [85, Theorem 2.3], which leverages a generalized central limit theorem to show that the type-2 Wasserstein distance between two normal distributions with true and empirical moments, respectively, decays *asymptotically* as $\mathcal{O}(N^{-\frac{1}{2}})$. A generalization of this result to elliptical distributions is discussed in [85, Remark 2.4].

Theorem 22 (Finite sample guarantees II). *Assume that all conditions of Theorem 21 hold and* $\varepsilon_N(\eta)$ *is defined as in (28). Then, for all* $\eta \in (0,1)$ *and* $\varepsilon \geq \varepsilon_N(\eta)$ *we have*

$$\mathbb{P}^N\Big\{\mathcal{R}(\mathbb{P},\ell) \leq \overline{\mathcal{R}}_\varepsilon(\widehat{\mu},\widehat{\Sigma},\ell) \quad \forall \ell \in \mathcal{L}\Big\} \geq 1-\eta. \tag{29a}$$

Moreover, if $\ell^\star$ *is an optimizer of the Gelbrich risk optimization problem (19), which is a function of the training samples, then for all* $\eta \in (0,1)$ *and* $\varepsilon \geq \varepsilon_N(\eta)$ *we have*

$$\mathbb{P}^N\Big\{\mathcal{R}(\mathbb{P},\ell^\star) \leq \overline{\mathcal{R}}_\varepsilon(\widehat{\mu},\widehat{\Sigma},\ell^\star)\Big\} \geq 1-\eta. \tag{29b}$$

Theorem 22 asserts that the Gelbrich risk (18) offers an upper confidence bound on the true risk (1) under the unknown data-generating distribution uniformly across all loss functions. It also asserts that the optimal value of the Gelbrich risk optimization problem (19) provides an upper confidence bound on the out-of-sample performance of its optimizers.

Asymptotic consistency of the Gelbrich risk optimization problem can only be established if the unknown true distribution is elliptical, $\Xi = \mathbb{R}^m$ and all admissible loss functions are quadratic. By Theorem 16, these conditions imply that the Gelbrich risk optimization problem (19) is equivalent to the Wasserstein distributionally robust optimization problem (7) equipped with a type-2 Wasserstein ball centered at an elliptical nominal distribution, where the Wasserstein distance is induced by the Euclidean norm.

Theorem 23 (Asymptotic consistency II). *Assume that all conditions of Theorem 21 hold. Select* $\eta_N \in (0,1]$ *and set* $\varepsilon_N = \varepsilon_N(\eta_N)$ *as in (28),* $N \in \mathbb{N}$*, such that* $\sum_{N=1}^{\infty} \eta_N < \infty$ *and* $\lim_{N\to\infty} \varepsilon_N = 0$*. If* $\mathbb{P}$ *is an elliptical distribution,* $\Xi = \mathbb{R}^m$ *and* $\ell(\xi)$ *is quadratic for all* $\ell \in \mathcal{L}$*, then we have* $\mathbb{P}^\infty$*-almost surely that* $\overline{\mathcal{R}}_{\varepsilon_N,p}(\widehat{\mu},\widehat{\Sigma},\mathcal{L}) \to \mathcal{R}(\mathbb{P},\mathcal{L})$ *as* N *tends to infinity.*

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

4. Distributionally Robust Optimization in Machine Learning

We now demonstrate that the theory of data-driven distributionally robust optimization with Wasserstein ambiguity sets has interesting ramifications for statistical learning and motivates new approaches for addressing fundamental learning tasks such as classification (Section 4.1), regression (Section 4.2), maximum likelihood estimation (Section 4.3) or minimum mean square error estimation (Section 4.4). We conclude with an overview of other applications of distributionally robust optimization in machine learning (Section 4.5).

4.1. Distributionally Robust Classification

In binary classification problems the central object of study is a random vector $\xi = (x, y)$, where $x \in \mathbb{R}^n$ is termed the *input*, and $y \in \{-1, +1\}$ is referred to as the *output*. The distribution $\mathbb{P}$ of ξ is unknown but indirectly observable through finitely many training samples $\widehat{\xi}_i = (\widehat{x}_i, \widehat{y}_i)$, $i \in [N]$. The goal of binary classification is to predict the output y corresponding to a given input x. The classifier with the lowest possible misclassification probability is the one that predicts $y = 1$ if $\mathbb{P}[y = 1|x] \geq 0.5$ and $y = -1$ otherwise. Unfortunately, this classifier is not implementable when $\mathbb{P}$ is unknown.

Statistical learning aims to construct classifiers solely on the basis of the training data. One of the most popular approaches in practice is to construct a linear *scoring function* $w^\top x$, encoded by a weight vector $w \in \mathbb{R}^n$, and to predict y as the sign of $w^\top x$. In hindsight, the prediction was correct if the actual output y coincides with the predicted output $\text{sign}(w^\top x)$ or, equivalently, if the product $y \cdot w^\top x$ is positive. The *realized* prediction error can thus be quantified by $L(y \cdot w^\top x)$, where $L(z)$ is some nonnegative and non-increasing univariate loss function that is large for negative and small for positive values of z. Examples of popular loss functions are listed in Table 1. The best scoring function for a given choice of $L(z)$ is the one whose weight vector w minimizes the *expected* prediction error $\mathbb{E}^{\mathbb{P}}[L(y \cdot w^\top x)]$. Unfortunately, the expectation is evaluated under the unknown distribution $\mathbb{P}$, and thus the optimal scoring function cannot be computed. Promising near-optimal scoring functions can be found, however, by solving a distributionally robust classification model that minimizes the worst-case expected prediction error with respect to a type-1 Wasserstein ball, that is,

$$\inf_{w \in \mathbb{R}^n} \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}} \left[L(y \cdot w^\top x) \right], \tag{30}$$

where $\widehat{\mathbb{P}}_N$ is the empirical distribution on the training samples. We assume here that all distributions in the Wasserstein ball are supported on $\Xi = \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X} \subseteq \mathbb{R}^n$ is convex and closed, while $\mathbb{Y} = \{-1, +1\}$. We also assume that the norm on the input-output space used in the definition of the Wasserstein distance is additively separable, that is, $\|\xi\| = \|x\| + \frac{\kappa}{2}|y|$, where—by slight abuse of notation—$\|x\|$ stands for an arbitrary norm on the input space, while $\kappa > 0$ quantifies the relative importance of outputs versus inputs.

Table 1. Commonly used loss functions for binary classification.

name	$L(z)$	learning model
hinge loss	$\max\{0, 1 - z\}$	support vector machine
smooth hinge loss	$\begin{cases} \frac{1}{2} - z & \text{if } z \leq 0 \\ \frac{1}{2}(1 - z)^2 & \text{if } 0 < z < 1 \\ 0 & \text{if } z \geq 1 \end{cases}$	smooth support vector machine
logloss	$\log(1 + \exp(-z))$	logistic regression

The classification model (30) is easily recognized as an instance of the distributionally robust decision problem (7) that optimizes over all (multivariate) loss functions of the form

$\ell(\xi) = L(y \cdot w^\top x)$ parameterized by $w \in \mathbb{R}^n$. By leveraging Theorem 8, problem (30) can be recast as a finite convex program if $L(z)$ is convex and piecewise linear, while $\mathbb{X}$ is convex and closed. An alternative convex reformulation can be obtained from Theorem 10 if $L(z)$ is convex (but not necessarily piecewise linear), while $\mathbb{X} = \mathbb{R}^n$. For all univariate loss functions listed in Table 1, the convex reformulations of problem (30) are equivalent to tractable conic programs. Explicit formulations of these conic programs are reported in [96, § 3.2].

The distributionally robust classification problem (30) encapsulates two interesting special cases. First, if the Wasserstein radius is set to $\varepsilon = 0$, then (30) collapses to the standard empirical risk minimization problem that minimizes the average prediction error across the training samples. Moreover, if the parameter κ appearing in the definition of the norm tends to infinity, then (30) reduces to a classical *regularized* empirical risk minimization problem.

Proposition 2 (Regularization by robustification). *If $L(z)$ is any of the loss functions of Table 1, $\mathbb{X} = \mathbb{R}^n$ and $\kappa = \infty$, then problem (30) is equivalent to*

$$\inf_{w \in \mathbb{R}^n} \; \frac{1}{N} \sum_{i=1}^{N} L(\widehat{y}_i \cdot w^\top \widehat{x}_i) + \varepsilon \cdot \|w\|_*.$$

Recall that the norm $\|\xi\| = \|x\| + \frac{\kappa}{2}|y|$ on the input-output space encodes the transportation cost in the definition of the Wasserstein distance. Thus, κ can be viewed as the cost of switching an output from $+1$ to -1 or vice versa. If $\kappa = \infty$, then all distributions in the Wasserstein ball are obtained by perturbing the empirical distribution along the input space because perturbations along the output space would be infinitely expensive. By setting $\kappa = \infty$, one thus postulates that there is only input uncertainty but no output uncertainty.

Proposition 2 gives commonly used regularization techniques a robustness interpretation, which applies under the premise that there is no output uncertainty. It identifies the regularization weight with the Wasserstein radius ε and the regularization function with the *dual* of the norm that determines the transportation cost along the input space.

Proposition 2 can be deduced from Theorem 8 by observing that the Lipschitz modulus of the multivariate loss function $\ell(\xi) = L(y \cdot w^\top x)$ is given by $\mathrm{Lip}(L) \cdot \|w\|_*$ and that $\mathrm{Lip}(L) = 1$ for all univariate loss functions of Table 1. Distributionally robust classification models with Wasserstein ambiguity sets were first studied in the context of logistic regression [97]. Extensions to other classification models are discussed in [10, 37, 96].

4.2. Distributionally Robust Regression

The goal of regression is a to predict a *real* (as opposed to a categorical) output $y \in \mathbb{R}$ corresponding to a given input $x \in \mathbb{R}^n$. The regressor that attains the lowest possible mean squared error is the one that predicts the output as $\mathbb{E}^{\mathbb{P}}[y|x]$. Unfortunately, this regressor is not implementable when the distribution $\mathbb{P}$ of the random vector $\xi = (x, y)$ is unknown.

In practice it is often convenient to construct a linear regressor that predicts the output by a linear function $w^\top x$ encoded by a weight vector $w \in \mathbb{R}^n$. The *realized* prediction error can thus be quantified by $L(w^\top x - y)$, where $L(z)$ is some nonnegative univariate loss function that is large when z deviates from 0. Examples of popular loss functions for regression are listed in Table 2. The best linear regressor that minimizes the *expected* prediction error $\mathbb{E}^{\mathbb{P}}[L(w^\top x - y)]$ cannot be computed when $\mathbb{P}$ is unknown, but promising near-optimal linear regressors can be found by solving the distributionally robust regression model

$$\inf_{w \in \mathbb{R}^n} \; \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \; \mathbb{E}^{\mathbb{Q}} \left[L(w^\top x - y) \right], \tag{31}$$

which minimizes the worst-case expected prediction error in view of all distributions on a convex closed set $\Xi = \mathbb{X} \times \mathbb{Y} \subseteq \mathbb{R}^n \times \mathbb{R}$ within a type-p Wasserstein ball around the empirical distribution on N training samples. By Theorem 8, problem (31) can be reformulated as

a finite convex program if $p = 1$, $L(z)$ is convex and piecewise linear, and $\mathbb{X}$ and $\mathbb{Y}$ are convex and closed. A convex reformulation can also be obtained from Theorem 8 if $p = 1$, $L(z)$ is convex (but not necessarily piecewise linear), while $\mathbb{X} = \mathbb{R}^n$ and $\mathbb{Y} = \mathbb{R}$. Moreover, problem (31) can be reformulated as a finite convex program by using Theorem 11 if $p = 2$, $L(z)$ is convex quadratic, $\mathbb{X} = \mathbb{R}^n$ and $\mathbb{Y} = \mathbb{R}$. For details see [96, § 3.1] and [10, § 3].

TABLE 2. Commonly used loss functions for regression

name	$L(z)$	parameter	learning model
squared error	z^2	n/a	ordinary least squares
Huber loss	$\begin{cases} \frac{1}{2}\xi^2 & \text{if } \lvert z\rvert \le \delta \\ \delta(\lvert z\rvert - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$	$\delta \in \mathbb{R}_+$	Huber regression
δ-insensitive loss	$\max\{0, \lvert z\rvert - \delta\}$	$\delta \in \mathbb{R}_+$	support vector regression
pinball loss	$\max\{-\delta z, (1-\delta)z\}$	$\delta \in [0,1]$	quantile regression

Assume now that the norm on the input-output space satisfies $\|\xi\| = \|x\| + \frac{\kappa}{2}|y|$, where $\|x\|$ is an arbitrary norm on the input space, while $\kappa > 0$ quantifies the relative importance of outputs versus inputs. In the absence of output uncertainty (that is, for $\kappa = \infty$), there is again an intimate relation between robustification and regularization.

Proposition 3 (Regularization by robustification). *Assume that $\mathbb{X} = \mathbb{R}^n$ and $\kappa = \infty$. If $L(z)$ is convex and Lipschitz continuous and $p = 1$, then problem (31) is equivalent to*

$$\inf_{w \in \mathbb{R}^n} \; \frac{1}{N} \sum_{i=1}^{N} L(w^\top \widehat{x}_i - \widehat{y}_i) + \varepsilon \cdot \mathrm{Lip}(L) \cdot \|w\|_*.$$

Moreover, if $L(z)$ is the square error and $p = 2$, then problem (31) reduces to

$$\inf_{w \in \mathbb{R}^n} \; \Big[\Big(\frac{1}{N} \sum_{i=1}^{N} L(w^\top \widehat{x}_i - \widehat{y}_i) \Big)^{\frac{1}{2}} + \varepsilon \cdot \|w\|_* \Big]^2.$$

Proposition 3 asserts that if there is no output uncertainty, then the distributionally robust regression problem (31) reduces to a regularized empirical risk minimization problem, where the regularization function is given by the dual of the norm on the input space. For Lipschitz continuous univariate loss functions $L(z)$ and for $p = 1$, one simply minimizes the sum of the empirical risk and the regularization term weighted by the product of Wasserstein radius and the Lipschitz modulus of $L(z)$. Note that the Huber loss, the δ-insensitive loss and the pinball loss are all Lipschitz continuous with Lipschitz moduli δ, 1 and $\max\{\delta, 1-\delta\}$, respectively. For the squared loss we need to set $p = 2$ because the type-2 Wasserstein ball is the largest Wasserstein ball for which the worst-case expected loss is finite. In this case, one minimizes a combination of the square root of the empirical loss and the regularization term. If one measures distances in the input space using the ∞-norm, then this convex program reduces to the so-called generalized LASSO (Least Absolute Shrinkage and Selection Operator) estimation problem. For further details on distributionally robust regression see [10, 37, 96].

4.3. Distributionally Robust Maximum Likelihood Estimation

Consider now the problem of estimating the mean vector $\mu \in \mathbb{R}^m$ and the covariance matrix $\Sigma \in \mathbb{S}_+^m$ of a random vector $\xi \in \mathbb{R}^m$ from independent training samples $\widehat{\xi}_i$, $i \in [N]$. The

simplest estimators for μ and Σ are the *sample mean* $\widehat{\mu}$ and the *sample covariance matrix* $\widehat{\Sigma}$, which we define as the actual mean and covariance matrix of the empirical distribution, *i.e.*,

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\xi}_i \qquad \text{and} \qquad \widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (\widehat{\xi}_i - \widehat{\mu})(\widehat{\xi}_i - \widehat{\mu})^\top.$$

While Σ serves as an input for many problems in engineering, science or economics, it is often the precision matrix Σ^{-1} that appears in their solutions. For example, in mean-variance portfolio analysis the portfolio variance to be minimized depends on the covariance matrix of the asset returns, while the optimal portfolio weights depend on the precision matrix. Similarly, linear discriminant analysis uses the covariance matrix of the features as an input and outputs a maximum likelihood classifier that depends on the precision matrix. Moreover, the optimal fingerprint method for climate change detection requires the covariance matrix of the internal climate variability as an input and outputs a climate change signal depending on the precision matrix. Thus, it is often more important to know the precision matrix than the covariance matrix. To ensure that the precision matrix is well defined, we will henceforth assume that $\Sigma \succ 0$. Unfortunately, the sample covariance matrix is rank-deficient in the big-data regime when the dimension of ξ exceeds the sample size ($m > N$) even if Σ has full rank. In this case, one cannot invert $\widehat{\Sigma}$ to obtain a meaningful precision matrix estimator.

From now on we will assume that the unknown true distribution $\mathbb{P}$ of ξ is normal. Thus, the problem of maximizing the log-likelihood of the training samples reduces to the following convex program over all candidate mean vectors μ and precision matrices X [14, § 7.1].

$$\inf_{\mu\in\mathbb{R}^m,\, X\in\mathbb{S}_+^m} \left\{ -\log\det X + \frac{1}{N}\sum_{i=1}^{N} (\widehat{\xi}_i - \mu)^\top X (\widehat{\xi}_i - \mu) \right\} \tag{32}$$

Unfortunately, this maximum likelihood estimation (MLE) problem is unbounded for $N \leq m$ and (almost surely) solved by $\mu^\star = \widehat{\mu}$ and $X^\star = \widehat{\Sigma}^{-1}$ for $N > m$. Thus, we fail again to find an estimator in the big-data regime and simply recover the sample mean and the sample covariance matrix in the small-data regime. To overcome this deficiency, we robustify the MLE problem against all distributions within a type-2 Wasserstein ball centered at the *normal* nominal distribution $\widehat{\mathbb{P}}_N = \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$, that is, we solve the robust MLE problem

$$\inf_{\mu\in\mathbb{R}^m,\, X\in\mathbb{S}_+^m} \left\{ -\log\det X + \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[(\xi-\mu)^\top X (\xi-\mu)\right] \right\}. \tag{33}$$

If $\varepsilon = 0$, then the robust MLE problem (33) reduces to the nominal MLE problem (32) because—by the definition of the sample mean and the sample covariance matrix—the (normal) nominal distribution has the same first- and second-order moments as the (discrete) empirical distribution and because the loss function in the expectation is quadratic in ξ. One can show via Theorem 16 that (33) is equivalent to a convex SDP with a determinant term in the objective function. Provided that the Wasserstein radius ε is strictly positive, this SDP is solvable even in the big-data regime when $m > N$. Thus, it yields a valid precision matrix estimator even if the sample covariance matrix is rank-deficient. Moreover, as SDPs are tractable, the optimal estimator can be computed in polynomial time. In fact, the SDP at hand is highly symmetric and can therefore even be solved in closed form [68, Theorem 3.1].

Theorem 24 (Wasserstein shrinkage estimator). *If $\varepsilon > 0$ and $\widehat{\Sigma} \in \mathbb{S}_+^m$ admits the spectral decomposition $\widehat{\Sigma} = \sum_{i=1}^m \lambda_i \cdot v_i v_i^\top$ with eigenvalues $\lambda_i \geq 0$ and corresponding orthonormal eigenvectors v_i, $i \in [m]$, then the unique minimizer of the robust MLE problem (33) is given by $\mu^\star = \widehat{\mu}$ and $X^\star = \sum_{i=1}^m x_i^\star \cdot v_i v_i^\top$, where*

$$x_i^\star = \gamma^\star \left[1 - \frac{1}{2}\left(\sqrt{\lambda_i^2 (\gamma^\star)^2 + 4\lambda_i \gamma^\star} - \lambda_i \gamma^\star \right) \right] \qquad \forall i \in [m], \tag{34a}$$

and $\gamma^\star > 0$ is the unique positive solution of the algebraic equation

$$\left(\varepsilon^2 - \frac{1}{2}\sum_{i=1}^m \lambda_i\right)\gamma - m + \frac{1}{2}\sum_{i=1}^m \sqrt{\lambda_i^2\gamma^2 + 4\lambda_i\gamma} = 0. \tag{34b}$$

Theorem 24 asserts that the robust MLE estimator $\mu^\star$ for the mean vector coincides with the sample mean $\widehat{\mu}$. More interestingly, it further asserts that the robust MLE estimator $X^\star$ for the precision matrix has the same eigenvectors v_i as the sample covariance matrix $\widehat{\Sigma}$, while its eigenvalues $x_i^\star$ are obtained by applying the nonlinear transformation (34a) to the corresponding eigenvalues λ_i of $\widehat{\Sigma}$. This transformation involves a single unknown parameter $\gamma^\star$, which is the unique positive solution of the algebraic equation (34b). As $X^\star$ is obtained by transforming the eigenvalues of the sample covariance matrix, it can be interpreted as a nonlinear shrinkage estimator. We thus refer to it as the *Wasserstein shrinkage estimator*.

As $X^\star$ and $\widehat{\Sigma}$ share the same eigenvectors, $X^\star$ is *rotation-equivariant*, that is, the estimator applied to the rotated data $R\widehat{\xi}_i$, $i \in [N]$, coincides with the rotated estimator $RX^\star R$ of the original data for every possible rotation matrix R. Moreover, as all eigenvalues of $X^\star$ are strictly positive, the estimator is always invertible. Finally, Theorem 24 indicates that $X^\star$ can be computed highly efficiently by computing the spectral decomposition of $\widehat{\Sigma}$ and by solving the scalar algebraic equation (34b), which can be accomplished by bisection.

One can show that the Wasserstein shrinkage estimator displays numerous desirable properties [68, Proposition 3.5]. First, its eigenvalues $x_i^\star$ decrease with ε and eventually converge to 0. This makes intuitive sense as for large values of ε nothing is known about ξ, and thus the safest bet is that all of its components have high variance and low precision. Moreover, one can show that the order of the eigenvalues $x_i^\star$ matches the order of the inverse sample eigenvalues $1/\lambda_i$ irrespective of $\varepsilon > 0$, which is expected in the absence of any structural information. Finally, one can show that the condition number of $X^\star$ decreases monotonically to 1 as ε grows. Thus, the condition number of $X^\star$ improves with the level of ambiguity.

A statistical theory that shows how to optimally choose ε is developed in [13]. Surprisingly, the Wasserstein radius that attains the lowest possible out-of-sample loss scales as $\varepsilon \propto 1/N$ instead of the canonical inverse square-root scaling, which may be expected for this problem.

So far we have assumed that there is no structural information about the distribution of ξ besides normality. In some practical situation, however, the precision matrix X may have a known sparsity pattern. Indeed, one can show that an element X_{ij} of the precision matrix vanishes if and only if the random variables ξ_i and ξ_j are conditionally independent given all other components of ξ. Conditional independencies of this type naturally arise, for example, in the analysis of spatio-temporal data. In the presence of sparsity information, the robust MLE problem is still equivalent to a tractable SDP. Even though it loses its analytical solvability, one can devise a tailored sequential quadratic approximation algorithm with rigorous convergence guarantees to solve the problem numerically, see [68, § 4].

4.4. Distributionally Robust Minimum Mean Square Error Estimation

Consider next the problem of estimating a signal $x \in \mathbb{R}^{m_x}$ from a noisy observation $y \in \mathbb{R}^{m_y}$ under the premise that the distribution of the random vector $\xi = [x^\top, y^\top]^\top \in \mathbb{R}^m$, $m = m_x + m_y$, is ambiguous and only known to belong to a type-2 Wasserstein ball centered at an elliptical nominal distribution $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}, \widehat{\Sigma})$ with nominal mean vector $\widehat{\mu} \in \mathbb{R}^m$, nominal covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^m$ and density generator g. This elementary problem is fundamental for numerous applications in engineering (*e.g.*, linear systems theory [44, 71]), econometrics (*e.g.*, linear regression [104, 110], time series analysis [19, 46]), machine learning and signal processing (*e.g.*, Kalman filtering [53, 64, 73]) or information theory (*e.g.*, multiple-input multiple-output systems [25, 58]), etc. To formalize the estimation problem, we define an estimator as a measurable function $\psi(y)$ that maps the observation y to a prediction of the

signal x, and we denote by Ψ the family of all possible estimators. Moreover, we define the distributionally robust *minimum mean square error* (MMSE) estimator as an optimizer of

$$\inf_{\psi\in\Psi} \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[\|x-\psi(y)\|_2^2\right]. \tag{35}$$

Note that (35) constitutes an infinite-dimensional functional optimization problem and thus appears to be hard. However, by establishing a minimax theorem for (35) and exploiting the properties of elliptical distributions, one can show that the outer infimum in (35) is attained by an *affine* estimator. Combining this structural insight with Theorem 16 allows us to prove that the estimation problem (35) is in fact equivalent to a convex program [98].

Theorem 25 (Distributionally robust MMSE estimator). *If $\widehat{\Sigma}\succ 0$, then the estimation problem (35) is equivalent to the nonlinear convex SDP*

$$\begin{array}{ll}\max & f(S)=\operatorname{Tr}\left[S_{xx}-S_{xy}S_{yy}^{-1}S_{yx}\right]\\ \text{s.t.} & S=\begin{bmatrix}S_{xx} & S_{xy}\\ S_{yx} & S_{yy}\end{bmatrix}\in\mathbb{S}_+^m,\quad S_{xx}\in\mathbb{S}_+^{m_x},\quad S_{yy}\in\mathbb{S}_+^{m_y},\quad S_{xy}=S_{yx}^\top\in\mathbb{R}^{m_x\times m_y}\\ & \operatorname{Tr}\left[S+\widehat{\Sigma}-2\left(\widehat{\Sigma}^{\frac12}S\widehat{\Sigma}^{\frac12}\right)^{\frac12}\right]\le\varepsilon^2,\quad S\succeq\lambda_{\min}(\widehat{\Sigma})I,\end{array} \tag{36}$$

which is always solvable. If $S^\star$, $S_{xx}^\star$, $S_{yy}^\star$ and $S_{xy}^\star$ are optimal in (36), while $\widehat{\mu}_x\in\mathbb{R}^{m_x}$ and $\widehat{\mu}_y\in\mathbb{R}^{m_y}$ are the (known) mean vectors of x and y under $\widehat{\mathbb{P}}_N$, respectively, then the affine function $\psi^\star(y)=S_{xy}^\star(S_{yy}^\star)^{-1}(y-\widehat{\mu}_y)+\widehat{\mu}_x$ is a distributionally robust MMSE estimator.

It is possible to eliminate all nonlinearities in (36) by using Schur complements and to reformulate the nonlinear convex SDP as a standard linear SDP, which is formally tractable. However, larger problem instances quickly exceed the capabilities of general-purpose solvers. Instead, there is merit in addressing the nonlinear SDP (36) directly with a customized first-order Frank-Wolfe algorithm, which starts at $S^{(0)}=\widehat{\Sigma}$ and constructs iterates

$$S^{(k+1)}=\alpha_k D^{(k)}+(1-\alpha_k)S^{(k)}\qquad\forall k=0,1,2,\ldots$$

with stepsize α_k, where $D^{(k)}\in\mathbb{S}^m$ is the unique solution of the direction-finding subproblem

$$\begin{array}{ll}\max\limits_{D\in\mathbb{S}^m} & \operatorname{Tr}\left[D\,\nabla f(S^{(k)})\right]\\ \text{s.t.} & \operatorname{Tr}\left[D+\widehat{\Sigma}-2\left(\widehat{\Sigma}^{\frac12}D\widehat{\Sigma}^{\frac12}\right)^{\frac12}\right]\le\varepsilon^2,\quad D\succeq\lambda_{\min}(\widehat{\Sigma})I.\end{array} \tag{37}$$

The Frank-Wolfe algorithm is highly efficient because the direction-finding subproblem (37), which linearizes the objective function $f(S)$ of (36) around the current iterate $S^{(k)}$, can be solved in closed form. Indeed, using Theorem 17 one can show that (37) is solved by

$$D^{(k)}=(\gamma^\star)^2\left(\gamma^\star I-\nabla f(S^{(k)})\right)^{-1}\widehat{\Sigma}\left(\gamma^\star I-\nabla f(S^{(k)})\right)^{-1},$$

where $\gamma^\star$ is the unique solution with $\gamma^\star I\succ\nabla f(S^{(k)})$ of the algebraic equation

$$\operatorname{Tr}\left[\widehat{\Sigma}\left(I-\gamma(\gamma I-\nabla f(S^{(k)}))^{-1}\right)^2\right]=\varepsilon^2,$$

which can be solved via bisection [98, Theorem 3.2]. For a judiciously chosen step-size rule, the Frank-Wolfe algorithm also offers rigorous convergence guarantees [98, Theorem 3.3].

Theorem 26 (Convergence analysis). *If $\widehat{\Sigma}\succ 0$, $\varepsilon>0$ and $\alpha_k=\frac{2}{k+2}$ for every $k\in\mathbb{N}$, then the k^{th} iterate $S^{(k)}$ of the Frank-Wolfe algorithm is feasible in (36) and satisfies $f(S^\star)-f(S^{(k)})\le\frac{C}{k+2}$, where C depends only on $\widehat{\Sigma}$ and ε, and $S^\star$ is a maximizer of (36).*

In some applications one has additional structural information about the relation between the signal x and the observation y (*e.g.*, the measurement noise may be known to be independent of the signal, or the observation may be governed by a linear measurement model, etc.). Such structural information can be used to restrict the Wasserstein ambiguity set in (35), thereby reducing the conservativeness of the distributionally robust MMSE estimator [69].

4.5. Other Applications in Machine Learning

Ideas from distributionally robust optimization also permeate several other areas of statistics and machine learning. For example, a distributionally robust optimization model involving two Wasserstein balls centered at two distinct empirical distributions can be used to develop a computationally tractable convex approximation for the *minimax robust hypothesis testing problem* that aims to minimize the maximum of the worst-case type-I and type-II errors of a prescribed hypothesis test [39]. Another example is *data-driven inverse optimization*, where one observes random signals as well as optimal solutions of an optimization problem parameterized by these signals. The aim is to predict the solution corresponding to a new unseen signal from N independent historical observations without any knowledge of the optimization problem's objective function. This problem can be framed as a *structural regression problem* that minimizes the worst-case expected prediction loss with respect to a Wasserstein ambiguity set over a space of candidate objective functions [61]. Data-driven inverse optimization lends itself, for example, to learning the purchasing behavior of consumers, the production costs of electricity generators, the route choice preferences of passengers in a multimodal transportation system or the hidden optimality principles governing a biological system. As a third example, distributionally robust optimization models with Wasserstein ambiguity sets can be used to efficiently compute the *worst-case misclassification probability* of a given classifier, which amounts to evaluating the worst-case expectation of the (nonconvex) zero-one loss [96, 97]. Using similar techniques, one can also efficiently compute the worst-case probability of an undesirable event described by the conjunction or disjunction of several linear inequalities for the random vector ξ [48, 60]. If the undesirable event can be influenced so as drive its worst-case probability below a prescribed tolerance, we face a *distributionally robust chance constraint*. Even though distributionally robust chance constrained programs with Wasserstein ambiguity sets around the empirical distribution are intractable in general, they are sometimes equivalent to mixed-integer linear programs that can be solved with off-the-shelf software [20, 112]. In contrast, distributionally robust chance constrained programs with moment ambiguity sets can often be reformulated as (or tightly approximated by) tractable conic programs [16, 21, 48, 115].

To conclude, we highlight two opportunities for tailoring a distributionally robust decision problem with a Wasserstein ambiguity set around the empirical distribution to a given training dataset. Recall first that finite sample guarantees hold whenever ε is large enough for the Wasserstein ball to contain the unknown data-generating distribution with high confidence $1-\beta$. Recall also that the distributionally robust decision problem can often be reformulated as a tractable convex program whose size scales with the sample size N. If the computational burden is unmanageable for the given sample size, we can select $K \ll N$, approximate $\widehat{\mathbb{P}}_N$ with the closest K-point distribution $\mathbb{Q}_K^\star$ in Wasserstein distance and replace the original Wasserstein ball of radius ε around $\widehat{\mathbb{P}}_N$ with a new inflated Wasserstein ball of radius $\varepsilon + W_p(\widehat{\mathbb{P}}_N, \mathbb{Q}_K^\star)$ around $\mathbb{Q}_K^\star$. By construction, the inflated Wasserstein ball contains the data-generating distribution with the same confidence $1-\beta$. But the size of the corresponding decision problem is only proportional to K. This approach provides a systematic method for reducing the computational burden without sacrificing robustness guarantees (but at the expense of increasing the model's level of conservatism). The approximation of a rich N-point distribution with a sparse K-point distribution is referred to as *scenario reduction* in the stochastic programming literature. While the exact computation of $\mathbb{Q}_K^\star$ is hard, there exist efficient approximation algorithms for scenario reduction [92].

An important input for any distributionally robust optimization model with a Wasserstein ambiguity set is the norm that determines the transportation cost in the definition of the Wasserstein distance. The flexibility to choose this norm could be exploited to improve the out-of-sample performance of the model's optimizers. A method for learning the best Mahalanobis norm from the training data is described in [11]. It is shown that this metric learning framework encompasses *adaptive regularization* as a special case.

Acknowledgments. This research was funded by the SNSF grant BSCGI0_157733.

Appendix

A. Elliptical Distributions

We say that $\mathbb{Q} = \mathcal{E}_g(\mu, \Sigma)$ is an *elliptical* probability distribution if it has a density function of the form $f(\xi) = C \det(\Sigma)^{-1} g((\xi-\mu)\Sigma^{-1}(\xi-\mu))$ with density generator $g(u) \geq 0$ for all $u \geq 0$, normalization constant $C > 0$, mean vector $\mu \in \mathbb{R}^m$ and covariance matrix $\Sigma \in \mathbb{S}_{++}^m$.

TABLE 3. Examples of elliptical distributions.

distribution family	density generator $g(u)$	normalization constant C
Gaussian distribution	$\exp(-u/2)$	$(2\pi)^{-m/2}$
Logistic distribution	$\dfrac{\exp(-u)}{(1+\exp(-u))^2}$	$\dfrac{\pi^{m/2}}{\Gamma(m/2)} \displaystyle\int_0^\infty \dfrac{y^{m/2-1}\exp(-y)}{(1+\exp(-y))^2}\,\mathrm{d}y$
t-distribution	$(1+u/(\nu-2))^{-\frac{m+\nu}{2}}$	$\dfrac{\nu^{m/2}\Gamma((m+\nu)/2)}{\pi^{m/2}\Gamma(\nu/2)(\nu-2)^m}$

Note. $\nu > 2$ denotes the degrees of freedom of the t-distribution, and Γ is the gamma function.

B. Conjugates, Support Functions and Dual Norms

The conjugate of an extended real-valued function $\ell(\xi)$ on $\mathbb{R}^m$ is a function $\ell^*(z)$ on $\mathbb{R}^m$ defined through $\ell^*(z) = \sup_\xi z^\top \xi - \ell(\xi)$. If $\ell(\xi)$ is proper, convex and lower semicontinuous, then the conjugate of the conjugate coincides with the initial function, that is, $\ell^{**}(\xi) = \ell(\xi)$ [86, § 12].

TABLE 4. Examples of conjugates.

$\ell(\xi)$	$\mathrm{dom}(\ell)$	$\ell^*(z)$	$\mathrm{dom}(\ell^*)$
$a^\top\xi + b$	$\mathbb{R}^m$	$\delta_{\{a\}}(z) - b$	$\{a\}$
$\frac{1}{2}\xi^\top A\xi + a^\top\xi + b$	$\mathbb{R}^m$	$\frac{1}{2}(z-a)^\top A^\dagger(z-a) - b$	$\{a\} + \mathrm{range(A)}$
$\log(1+\exp(-\xi))$	$\mathbb{R}$	$z\log z + (1-z)\log(1-z)$	$[0,1]$
$\exp(\xi)$	$\mathbb{R}$	$z\log z - z$	$\mathbb{R}_+$
$\frac{1}{p}\|\xi\|^p$	$\mathbb{R}^m$	$\frac{1}{q}\|\xi\|_*^q$	$\mathbb{R}^m$
$\|\xi\|$	$\mathbb{R}^m$	$\delta_{\mathcal{B}_1^*(0)}(z)$	$\mathcal{B}_1^*(0)$

Note. Assume that $A \in \mathbb{S}_+^m$, $a \in \mathbb{R}^m$, $b \in \mathbb{R}$ and $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. Moreover, denote by $\mathcal{B}_1^*(0) = \{z \in \mathbb{R}^m : \|z\|_* \leq 1\}$ the standard ball with respect to the dual norm on $\mathbb{R}^m$.

The indicator function of a set $\Xi \subseteq \mathbb{R}^m$ is a function $\delta_\Xi(\xi)$ on $\mathbb{R}^m$ defined through $\delta_\Xi(\xi) = 0$ if $\xi \in \Xi$ and $\delta_\Xi(\xi) = \infty$ if $\xi \notin \Xi$. The support function of $\Xi \subseteq \mathbb{R}^m$ is a function $\sigma_\Xi(z)$ on $\mathbb{R}^m$ defined through $\sigma_\Xi(z) = \sup_{\xi\in\Xi} z^\top\xi$. The support function of Ξ coincides with the conjugate of the indicator function of Ξ, that is, $\delta_\Xi^*(z) = \sigma_\Xi(z)$. If Ξ is convex and closed, then the conjugate of the support function of Ξ coincides with the indicator function of Ξ, that is, $\sigma_\Xi^*(\xi) = \delta_\Xi(\xi)$ [86, § 13].

TABLE 5. Examples of support functions.

Ξ	$\sigma_\Xi(z)$	$\mathrm{dom}(\sigma_\Xi)$
$\{\xi : \|\xi\| \leq b\}$	$b\|z\|_*$	$\mathbb{R}^m$
$\{\xi : C\xi \leq d\}$	$\inf\{\lambda^\top d : \lambda \in \mathbb{R}^l_+, C^\top\lambda = z\}$	$\{C^\top\lambda : \lambda \in \mathbb{R}^l_+\}$
$\{\xi : f(\xi) \leq 0\}$	$\inf\{\lambda f^*(z/\lambda) : \lambda \in \mathbb{R}^l_+\}$	$-\,\mathrm{recc}(f)^*$
$\{\xi : \xi \in \Xi_k\ \forall k \in [K]\}$	$\inf\{\sum_{k=1}^K \sigma_{\Xi_k}(z_k) : \sum_{k=1}^K z_k = z\}$	$-\bigcap_{k\in[K]} \mathrm{recc}(\Xi_k)^*$

Note. Assume that $b \in \mathbb{R}_+$, $C \in \mathbb{R}^{l\times m}$ and $d \in \mathbb{R}^l$. Let $f(\xi)$ be a closed, proper and convex function, and let Ξ_k, $k \in [K]$, be convex closed sets with nonempty intersection. Denote by $\mathrm{recc}(f)^*$ and $\mathrm{recc}(\Xi_k)^*$ the cones dual to the recession cones of the function $f(\xi)$ and the set Ξ_k, respectively.

If $\|\cdot\|$ is a norm on $\mathbb{R}^m$, then its dual norm $\|\cdot\|_*$ on $\mathbb{R}^m$ is defined through $\|z\|_* = \sup_{\|\xi\|\leq 1} z^\top\xi$. The dual norm of the dual norm coincides with the original norm, that is, $\|\cdot\|_{**} = \|\cdot\|$ [86, § 15].

TABLE 6. Examples of dual norms.

$\|\xi\|$	$\|z\|_*$	comment
$\|\xi\|_p$	$\|z\|_q$	standard p-norms
$\|\xi\|_1$	$\|z\|_\infty$	limiting case when $p\downarrow 1$ and $q\uparrow\infty$
$\alpha\|\xi\|_p$	$\frac{1}{\alpha}\|z\|_q$	scaled p-norms
$\|A\xi\|_p$	$\|A^{-1}z\|_q$	scaled p-norms
$\sum_{k\in[K]} \|\xi_k\|_{p_k}$	$\max_{k\in[K]} \|z_k\|_{q_k}$	additively separable norms

Note. Assume that $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, $\alpha > 0$, $A \in \mathbb{S}^m_{++}$, and $p_k, q_k \geq 1$ with $\frac{1}{p_k} + \frac{1}{q_k} = 1$ for all $k \in [K]$. Moreover, $\xi = (\xi_1, \ldots, \xi_K)$ and $z = (z_1, \ldots, z_K)$, where $\xi_k, z_k \in \mathbb{R}^{m_k}$ and $\sum_{k=1}^K m_k = m$.

References

[1] D. Alvarez-Melis, T. Jaakkola, and S. Jegelka. Structured optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1771–1780, 2018.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[3] A. Ben-Tal, D. Den Hertog, and J. Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, 149(1-2):265–299, 2015.

[4] J. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[5] D.P. Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21(1):152–171, 1981.

[6] D.P. Bertsekas. Auction algorithms for network flow problems: A tutorial introduction. *Computational Optimization and Applications*, 1(1):7–66, 1992.

[7] D.P. Bertsekas. *Network Optimization: Continuous and Discrete Models*. Athena Scientific, 1998.

[8] D. Bertsimas, S. Shtern, and B. Sturt. A data-driven approach for multi-stage linear optimization. *Available from Optimization Online*, 2018.

[9] D. Bertsimas and J.N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[10] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.

[11] J. Blanchet, Y. Kang, F. Zhang, and K. Murthy. Data-driven optimal transport cost selection for distributionally robust optimization. *arXiv preprint arXiv:1705.07152*, 2017.

[12] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *To appear in Mathematics of Operations Research*, 2019.

[13] J. Blanchet and N. Si. Optimal uncertainty size in distributionally robust inverse covariance estimation. *arXiv preprint arXiv:1901.07693*, 2019.

[14] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[15] R.A. Brualdi. *Combinatorial Matrix Classes*. Cambridge University Press, 2006.

[16] G.C. Calafiore and L. El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.

[17] G. Canas and L. Rosasco. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems 25*, pages 2492–2500, 2012.

[18] E. Cazelles, V. Seguy, J. Bigot, M. Cuturi, and N. Papadakis. Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing*, 40(2):B429–B456, 2018.

[19] C. Chatfield. *The Analysis of Time Series: An Introduction*. CRC Press, 2016.

[20] Z. Chen, D. Kuhn, and W. Wiesemann. Data-driven chance constrained programs over Wasserstein balls. *arXiv preprint arXiv:1809.00210*, 2018.

[21] S. Cheung, A. Man-Cho So, and K. Wang. Linear matrix inequalities with stochastically dependent perturbations and applications to chance-constrained semidefinite optimization. *SIAM Journal on Optimization*, 22(4):1394–1430, 2012.

[22] L. Chizat, G. Peyré, B. Schmitzer, and F. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87:2563–2609, 2018.

[23] V. Chopra and W. Ziemba. The effect of errors in means, variances, and covariances on optimal portfolio choice. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, chapter 18, pages 249–257. World Scientific Publishing, 2011.

[24] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[25] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[26] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

[27] T. Homem de Mello and G. Bayraksan. Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.

[28] E. Delage, D. Kuhn, and W. Wiesemann. "Dice"-sion making under uncertainty: When can a random decision reduce risk? *To appear in Management Science*, 2019.

[29] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[30] D.C. Dowson and B.V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[31] J. Dupačová. Stability and sensitivity-analysis for stochastic programming. *Annals of Operations Research*, 27(1):115–142, 1990.

[32] M. Essid and J. Solomon. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4):A1961–A1986, 2018.

[33] S. Ferradans, N. Papadakis, G. Peyré, and J. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[34] J. Feydy, B. Charlier, F. Vialard, and G. Peyré. Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 291–299, 2017.

[35] R. Flamary, M. Cuturi, N. Courty, and A. Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.

[36] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[37] R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

[38] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[39] R. Gao, L. Xie, Y. Xie, and H. Xu. Robust hypothesis testing using Wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems 31*, pages 7913–7923, 2018.

[40] M. Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[41] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

[42] L. El Ghaoui, M. Oks, and F. Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, 51(4):543–556, 2003.

[43] C.R. Givens and R.M. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[44] F. Golnaraghi and B.C. Kuo. *Automatic Control Systems*. McGraw-Hill Education, 2017.

[45] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30*, pages 5767–5777, 2017.

[46] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[47] G.A. Hanasusanto, D. Kuhn, S. W. Wallace, and S. Zymler. Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Mathematical Programming*, 152(1-2):1–32, 2015.

[48] G.A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann. A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming*, 151(1):35–62, 2015.

[49] N. Ho, X. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel clustering via Wasserstein means. In *International Conference on Machine Learning*, pages 1501–1509, 2017.

[50] L.V. Kantorovich. On the translocation of masses. *Doklady Akademii Nauk USSR*, 37:199–201, 1942.

[51] L.V. Kantorovich and G. Rubinshtein. On a space of totally additive functions. *Vestnik Leningrad University*, 13(7):52–59, 1958.

[52] J. Karlsson and A. Ringh. Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport. *SIAM Journal on Imaging Sciences*, 10(4):1935–1962, 2017.

[53] S.M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

[54] S. Kolouri, S.R. Park, M. Thorpe, D. Slepcev, and G.K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[55] S. Kolouri and G.K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2015.

[56] S. Kolouri, G.K. Rohde, and H. Hoffmann. Sliced Wasserstein distance for learning Gaussian mixture models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3427–3436, 2018.

[57] S. Kundu, S. Kolouri, K. I. Erickson, A. F. Kramer, E. McAuley, and G.K. Rohde. Discovery and visualization of structural biomarkers from MRI using transport-based morphometry. *NeuroImage*, 167:256–275, 2018.

[58] D.J.C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[59] R. Michaud. The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42, 1989.

[60] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[61] P. Mohajerin Esfahani, S. Shafieezadeh-Abadeh, G.A. Hanasusanto, and D. Kuhn. Data-driven inverse optimization with imperfect information. *Mathematical Programming*, 167(1):191–234, 2018.

[62] G. Monge. Mémoire sur la théorie des déblais et des remblais. In *Histoire de l'Académie Royale des Sciences de Paris*, pages 666–704, 1781.

[63] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim. Image to image translation for domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.

[64] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[65] B. Muzellec, R. Nock, G. Patrini, and F. Nielsen. Tsallis regularized optimal transport and ecological inference. In *Association for the Advancement of Artificial Intelligence*, pages 2387–2393, 2017.

[66] K. Natarajan, M. Sim, and J. Uichanco. Tractable robust expected utility and risk models for portfolio optimisation. *Mathematical Finance*, 20(4):695–731, 2010.

[67] V.A. Nguyen, D. Filipovic, and D. Kuhn. Distributionally robust risk measures with structured Wasserstein ambiguity sets. *Working paper*, 2019.

[68] V.A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *arXiv preprint arXiv:1805.07194*, 2018.

[69] V.A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization. *Working paper*, 2019.

[70] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

[71] K. Ogata. *Modern Control Engineering*. Pearson, 2009.

[72] I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[73] A.V. Oppenheim and G.C. Verghese. *Signals, Systems and Inference*. Pearson, 2015.

[74] H. Owhadi and C. Scovel. Extreme points of a ball about a measure with finite support. *Communications in Mathematical Sciences*, 15(1):77–96, 2017.

[75] N. Papadakis and J. Rabin. Convex histogram-based joint image segmentation with regularized optimal transport cost. *Journal of Mathematical Imaging and Vision*, 59(2):161–186, 2017.

[76] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[77] B. Van Parys, P. Mohajerin Esfahani, and D. Kuhn. From data to decisions: Distributionally robust optimization is optimal. *arXiv preprint arXiv:1704.04118*, 2017.

[78] O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *European Conference on Computer Vision*, pages 495–508, 2008.

[79] O. Pele and M. Werman. Fast and robust earth mover's distances. In *International Conference on Computer Vision*, volume 9, pages 460–467, 2009.

[80] G. Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[81] G. Peyré, L. Chizat, F. Vialard, and J. Solomon. Quantum entropic regularization of matrix-valued optimal transport. *European Journal of Applied Mathematics*, pages 1–24, 2017.

[82] G. Peyré and M. Cuturi. Computational optimal transport. *arXiv preprint arXiv:1803.00567*, 2018.

[83] G. Pflug and A. Pichler. *Multistage Stochastic Optimization*. Springer, 2014.

[84] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[85] T. Rippl, A. Munk, and A. Sturm. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90–109, 2016.

[86] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

[87] A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *International Conference on Artificial Intelligence and Statistics*, pages 630–638, 2016.

[88] W. Römisch. Stability of stochastic programming problems. In *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 483–554. Elsevier, 2003.

[89] W. Römisch and R. Schultz. Stability analysis for stochastic programs. *Annals of Operations Research*, 30(1):241–266, 1991.

[90] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[91] N. Rujeerapaiboon, D. Kuhn, and W. Wiesemann. Robust growth-optimal portfolios. *Management Science*, 62(7):2090–2109, 2015.

[92] N. Rujeerapaiboon, K. Schindler, D. Kuhn, and W. Wiesemann. Scenario reduction revisited: Fundamental limits and guarantees. *To appear in Mathematical Programming*, 2019.

[93] M.A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, and J. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[94] B. Schmitzer. A sparse multiscale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259, 2016.

[95] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems 28*, pages 3312–3320, 2015.

[96] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Regularization via mass transportation. *arXiv preprint arXiv:1710.10016*, 2017.

[97] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584, 2015.

[98] S. Shafieezadeh-Abadeh, V.A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. Wasserstein distributionally robust Kalman filtering. In *Advances in Neural Information Processing Systems 31*, pages 8483–8492, 2018.

[99] A. Shapiro. Monte Carlo sampling methods. In *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 353–425. Elsevier, 2003.

[100] B. Taşkesen, S. Shafieezadeh-Abadeh, and D. Kuhn. On the complexity of computing Wasserstein distances. *Working paper*, 2019.

[101] J. Smith and R. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

[102] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66, 2015.

[103] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics*, 33(4):67, 2014.

[104] J.H. Stock and M.W. Watson. *Introduction to Econometrics*. Prentice Hall, 2015.

[105] G. Tartavel, G. Peyré, and Y. Gousseau. Wasserstein loss for image synthesis and restoration. *SIAM Journal on Imaging Sciences*, 9(4):1726–1755, 2016.

[106] M. Thorpe, S. Park, S. Kolouri, G.K. Rohde, and D. Slepčev. A transportation L^p distance for signal analysis. *Journal of Mathematical Imaging and Vision*, 59(2):187–210, 2017.

[107] C. Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

[108] W. Wang, J. A. Ozolek, D. Slepcev, A. B. Lee, C. Chen, and G.K. Rohde. An optimal transportation approach for nuclear structure-based pathology. *IEEE Transactions on Medical Imaging*, 30(3):621–631, 2011.

[109] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *To appear in Bernoulli*, 2019.

[110] J.M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2010.

[111] D. Wozabal. A framework for optimization under ambiguity. *Annals of Operations Research*, 193(1):21–47, 2012.

[112] W. Xie. On distributionally robust chance constrained program with Wasserstein distance. *arXiv preprint arXiv:1806.07418*, 2018.

[113] C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262 – 267, 2018.

[114] J. Zhen, D. Kuhn, and W. Wiesemann. Distributionally robust nonlinear optimization. *Working paper*, 2019.

[115] S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, 2013.

[116] S. Zymler, D. Kuhn, and B. Rustem. Worst-case value at risk of nonlinear portfolios. *Management Science*, 59(1):172–188, 2013.