Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimal Transport-Based Distributionally Robust Optimization: Structural Properties and Iterative Schemes

Jose Blanchet, Karthyek Murthy, Fan Zhang

To cite this article:
Jose Blanchet, Karthyek Murthy, Fan Zhang (2022) Optimal Transport-Based Distributionally Robust Optimization: Structural Properties and Iterative Schemes. Mathematics of Operations Research 47(2):1500-1529. https://doi.org/10.1287/moor.2021.1178

http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 47, No. 2, May 2022, pp. 1500–1529
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Optimal Transport-Based Distributionally Robust Optimization: Structural Properties and Iterative Schemes

Jose Blanchet,[a] Karthyek Murthy,[b] Fan Zhang[a]

[a] Management Science and Engineering, Stanford University, Stanford, California 94305; [b] Engineering Systems and Design, Singapore University of Technology & Design, Singapore 487372, Singapore

Contact: jose.blanchet@stanford.edu, https://orcid.org/0000-0001-5895-0912 (JB); karthyek_murthy@sutd.edu.sg, https://orcid.org/0000-0003-1071-0981 (KM); fzh@stanford.edu, https://orcid.org/0000-0001-6735-4901 (FZ)

Received: October 8, 2018
Revised: May 26, 2020; February 1, 2021
Accepted: April 20, 2021
Published Online in Articles in Advance: November 10, 2021

MSC2020 Subject Classification: Primary: 90C15; Secondary: 65K05, 90C47

https://doi.org/10.1287/moor.2021.1178

Abstract. We consider optimal transport-based distributionally robust optimization (DRO) problems with locally strongly convex transport cost functions and affine decision rules. Under conventional convexity assumptions on the underlying loss function, we obtain structural results about the value function, the optimal policy, and the worst-case optimal transport adversarial model. These results expose a rich structure embedded in the DRO problem (e.g., strong convexity even if the non-DRO problem is not strongly convex, a suitable scaling of the Lagrangian for the DRO constraint, etc., which are crucial for the design of efficient algorithms). As a consequence of these results, one can develop efficient optimization procedures that have the same sample and iteration complexity as a natural non-DRO benchmark algorithm, such as stochastic gradient descent.

Funding: Material in this paper is based upon work supported by the Air Force Office of Scientific Research [Grant FA9550-20-1-0397]. Additional support is gratefully acknowledged from the Singapore University of Technology and Design [Grant MOE SRG ESD 2018 134], China Merchants Bank, the Defense Advanced Research Projects Agency [Grant N660011824028], and the National Science Foundation [Grants 1915967, 1820942, and 1838576].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/moor.2021.1178.

Keywords: distributionally robust optimization • stochastic gradient descent • optimal transport • Wasserstein distances • adversarial • strong convexity • comparative statics • rate of convergence

1. Introduction

In this paper, we study the distributionally robust optimization (DRO) version of stochastic optimization models with linear decision rules of the form

$$\inf_{\beta \in B} E_{P^*}[\ell(\beta^T X)], \tag{1}$$

where $E_{P^*}[\cdot]$ represents the expectation operator associated to the probability model P^*, which describes the random element $X \in \mathbb{R}^d$. The decision (or optimization) variable β is assumed to take values on a convex set $B \subseteq \mathbb{R}^d$, and the loss function $\ell : \mathbb{R} \to \mathbb{R}$ is assumed to satisfy certain convexity and regularity assumptions discussed in the sequel. The formulation also includes affine decision rules by simply redefining X by $(X, \mathbf{1})$.

Stochastic optimization problems such as (1) include standard formulations in important operations research and machine learning applications, including newsvendor models, portfolio optimization via utility maximization, and a large portion of the most conventional generalized linear models in the setting of statistical learning problems.

The corresponding DRO version of (1) takes the form

$$\inf_{\beta \in B} \sup_{P \in \mathcal{U}_\delta(P_0)} E_P[\ell(\beta^T X)], \tag{2}$$

where $\mathcal{U}_\delta(P_0)$ is a so-called distributional uncertainty region "centered" around some benchmark model, P_0, which may be data-driven (for example, an empirical distribution), and $\delta > 0$ parameterizes the size of the distributional uncertainty. Precisely, we assume that P_0 is an arbitrary distribution with suitably bounded moments.

The DRO counterpart of (1) is motivated by the fact that the underlying model P^* generally is unknown although the benchmark model, P_0, is typically chosen to be a tractable model that, in principle, should retain as much model fidelity as possible (i.e., P_0 should at least capture the most relevant features present in P^*).

However, simply replacing P^* by P_0 in Formulation (1) may result in the selection of a decision, β_0, which significantly underperforms in actual practice relative to the optimal decision for the actual problem (based on P^*).

The DRO formulation (2) introduces an adversary (represented by the inner sup) that explores the implications of any decision β as the benchmark model P_0 varies within $\mathcal{U}_\delta(P_0)$. The adversary should be seen as a powerful modeling tool whose goal is to explore the impact of potential decisions in the phase of distributional uncertainty. The DRO formulation then prescribes a choice that minimizes the worst-case expected cost induced by the models in the distributional uncertainty region.

An important ingredient in the DRO formulation is the description of the distributional uncertainty region $\mathcal{U}_\delta(P_0)$. In recent years, there has been significant interest in distributional uncertainty regions satisfying

$$\mathcal{U}_\delta(P_0) = \{P : \mathcal{W}(P_0, P) \le \delta\},$$

where $\mathcal{W}(P_0, P)$ is a Wasserstein distance (see, for example, Blanchet and Murthy [4], Blanchet et al. [5], Chen et al. [8], Gao and Kleywegt [12], Gao et al. [14], Mohajerin Esfahani and Kuhn [19], Shafieezadeh-Abadeh et al. [29], Sinha et al. [32], Volpi et al. [33], Yang [36], Zhao and Guan [37], and references therein).

The Wasserstein distance is a particular case of optimal transport discrepancies, which we review momentarily. A general optimal transport discrepancy computes the cheapest cost of transporting the mass of P_0 to the mass of P so that a unit of mass transported from position x to position y is measured according to a transportation cost function, $c(\cdot)$. The definition of $\mathcal{W}(P_0, P)$ requires that $c(\cdot)$ be a norm or a distance, but this is not necessary, and endowing modelers with increased flexibility in choosing $c(\cdot)$ is an important part of our motivation.

The use of the Wasserstein distance is closely related to norm-regularization, and DRO formulations have been shown to recover approximately and exactly a wide range of machine learning estimators; see, for example, Blanchet et al. [5], Gao et al. [13], and Shafieezadeh-Abadeh et al. [28, 29]. These and some other applications of the DRO formulation (2) based on Wasserstein distance lead to a reduction from (2) back to a problem of the form (1), in which the objective loss function is modified by adding a regularization penalty expressed in terms of the norm of β and a regularization penalty parameter as an explicit function of δ.

We stress that, in many of these settings, particularly the cases in which $\ell(\cdot)$ is Lipschitz and convex, the worst-case distribution is degenerate (i.e., it is realized by moving infinitesimally small mass toward infinity or moving no mass at all).

We enable efficient algorithms that can be applied to more flexible cost functions $c(\cdot)$ and losses $\ell(\cdot)$ in order to induce adversarial distributions that can be both informed by side information and endowed with meaningful interpretations.

For other special cases that are amenable to either analytical solutions or software implementations, P_0 is either assumed to have a special structure (e.g., Gaussian distribution as in Nguyen et al. [24]) or ultimately requires robust optimization formulations that require P_0 to have finite support; see Chen et al. [8, 9], Luo and Mehrotra [17], Mohajerin Esfahani and Kuhn [19], Shafieezadeh-Abadeh et al. [28, 29], Xie [35], and Zhao and Guan [37].

As we shall see, our analysis enables the application of stochastic gradient descent (SGD) algorithms to approximate the solution to (2) and that are applicable to cases in which P_0 has unbounded support (under suitable moment constraints). Moreover, by enabling the use of stochastic gradient descent algorithms, we open the door to further research on accelerated stochastic gradient methods. In this paper, we focus on providing stochastic gradient descent implementation to demonstrate the direct application of our structural results.

We mention Sinha et al. [32], in which relaxed Wasserstein DRO formulations are explored in the context of certifying robustness in deep neural networks. The stochastic gradient descent type employed in Sinha et al. [32] is similar to the ones we discuss in Section 3. Nevertheless, these algorithms are designed for a fixed value of the dual parameter (which we call λ), chosen to be large. Our analysis suggests that rescaling λ so that $\left(\lambda = O\left(\delta^{-1/2}\right)\right)$ may enhance performance even in the case of the more general type of losses considered in Sinha et al. [32]. The impact of this type of rescaling in terms of performance guarantees for computational algorithms has not been studied in the literature, and we believe that our analysis could prove useful in future studies. Additional discussion on the rescaling is given at the end of Section 3.2.2.

The challenge in our study lies in the inner maximization (2), which is not easy to perform, and its properties, parametrically as a function of both β and δ, are nontrivial to analyze. So much of our effort goes into understanding these properties. But, before we describe our results, we first describe a flexible class of models for distributional uncertainty sets, $\mathcal{U}_\delta(P_0)$.

1.1. A Description of the Distributional Uncertainty Region $\mathcal{U}_\delta(P_0)$

We focus on DRO formulations based on extensions of the Wasserstein distance, called optimal transport discrepancies. Formally, an optimal transport discrepancy between distributions P and P_0 with respect to the (lower semicontinuous) cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$ is defined as follows.

First, let $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ be the set of Borel probability measures on $\mathbb{R}^d \times \mathbb{R}^d$. So, for any $X \in \mathbb{R}^d$ and $X' \in \mathbb{R}^d$ random elements living on the same probability space, there exists $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, which governs the joint distribution of (X, X').

If we use π_X to denote the marginal distribution of X under π and $\pi_{X'}$ to denote the marginal distribution of X' under π, then the optimal transport cost between P and P_0 can be written as

$$D_c(P_0, P) = \inf\{E_\pi[c(X, X')] : \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d),\ \pi_X = P_0,\ \pi_{X'} = P\}. \tag{3}$$

The Wasserstein distance is recovered if $c(x, x') = \|x - x'\|$ under any given norm. If $c(x, x')$ is not a distance, then $D_c(P_0, P)$ is not necessarily a distance.

Ultimately, we are interested in the computational tractability of the DRO problem (2) assuming

$$\mathcal{U}_\delta(P_0) = \{P : D_c(P_0, P) \le \delta\}, \tag{4}$$

for a flexible class of functions c. We concentrate on what we call local Mahalanobis or state-dependent Mahalanobis cost functions of the form

$$c(x, x') = (x - x')^T A(x)(x - x'), \tag{5}$$

where $A(x)$ is a positive definite matrix for each x. For this choice of $c(\cdot)$, the distributions in $\mathcal{U}_\delta(P_0)$ are unrestricted in support. We explain in the Conclusions section how our results can be applied to other cost functions.

The family of cost functions that we consider is motivated by the perspective that the adversary introduced in the DRO formulation (2) (represented by the inner sup) is a modeling tool that explores the impact of potential decisions.

Let us consider, for example, a situation in which we are interested in choosing an optimal portfolio strategy. In this setting, historical returns can naturally be used to fit a statistical model. However, there is also current market information that is not of a statistical nature but of an economic nature in the form of, for instance, implied volatilities (i.e., the volatility that is implied by the current supply and demand reflected by the prices of derivative securities). The implied volatility differs from the historical volatility, and it is more sensible to capturing current market perceptions. An enhanced DRO formulation that uses a cost function such as (5) could incorporate market information as follows. Returns with higher implied volatility, maybe even depending on the current stock values, could be assigned a lower cost of transportation, and returns with lower implied volatility may be given higher transportation costs. The intuition is that high implied volatilities correspond to potentially higher future fluctuations (as perceived by the market), so the adversary should be given higher ability relative (and, thus, lower costs) to explore the potential implications of such future out-of-sample fluctuations on portfolio choices.

In general, just as we discuss in the previous paragraph, it is not difficult to imagine more situations in which the optimizer may be more concerned about the impact of distributional uncertainty on certain regions of the outcome space relative to other regions. Such situations may arise as a consequence of different amounts of information available in different regions of the outcome space or perhaps due to data contamination or measurement errors, which may be more prone to occur for certain values of x.

In this paper, we do not focus on the problem of fitting the cost function, but we do consider the portfolio optimization discussed earlier and the use of implied volatilities in an empirical study in Section 4. We point out, however, that related questions have been explored, at least empirically, in classification settings, using manifold learning procedures (Blanchet et al. [7], Noh et al. [25], Wang et al. [34]). Our motivation is that flexible formulations based on cost functions such as (5) are useful if one wishes to fully exploit the role of the artificial adversary in (2) as a modeling tool.

Now, leaving aside the modeling advantages of choosing a cost function such as (5) and coming back to the computational challenges, even if one selects $A(x)$ to be the identity (thus, recovering a more traditional Wasserstein DRO formulation), solving (2) is not entirely easy because the inner optimization problem in (2) is nontrivial to study. An exact convex optimization reformulation has been demonstrated only for losses taking a specific form. For example, Hanasusanto and Kuhn [15] provide a conic reformulation for the data-driven DRO problem with piecewise linear convex losses. With the number of conic constraints being proportional to the data set size, it is, however, computationally less suited for handling large data sets.

To exploit these DRO formulations, one must develop scalable algorithms with guaranteed good performance for solving (2). By good performance, we mean that we can easily develop algorithms for solving (2) with complexity that is comparable to that of natural benchmark algorithms for solving (1). Enabling these good-performing algorithms is precisely one of the goals of this paper. To this end, several properties, such as duality

representations, convexity, and the structure of worst-case adversaries, are studied. These results have far-ranging implications as we discuss next.

1.2. A More In-Depth Discussion of Our Technical Contributions

First, using a standard duality result, we write the inner maximization in (2) as

$$\sup_{P:D_c(P_0,P)\le\delta} E_P\big[\ell(\beta^T X)\big] = \inf_{\lambda\ge 0} E_{P_0}[\ell_{rob}(\beta,\lambda;X)], \tag{6}$$

for a dual objective function $\ell_{rob}(\cdot)$ and a dual variable $\lambda \ge 0$.

Then, we show that, after a rescaling in λ, that the objective function, $E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$, is locally strongly convex in (β,λ) uniformly over a compact set containing the optimizer, the strong convexity parameter of at least $\kappa_1\delta^{1/2}$ (for some $\kappa_1 > 0$ that we identify), under suitable convexity and growth assumptions on $\ell(\cdot)$; see Theorems 1–4.

It turns out that the function $\ell_{rob}(\cdot)$ can be computed by solving a one-dimensional search problem on a compact interval. This can be solved quite efficiently (exponentially fast rate of convergence) under the setting of Theorems 1–4.

We then study a natural stochastic gradient descent algorithm for solving (2) that, because of the strong convexity properties derived for $\ell_{rob}(\cdot)$, achieves an iteration complexity of order $O_p(\varepsilon^{-1}L)$ to reach $O(\varepsilon)$ error, where L is the cost of solving the one-dimensional search problem. We also discuss in the online appendix how to execute this line search procedure efficiently, provided that suitable smoothness assumptions are imposed on $\ell(\cdot)$ (leading to an extra factor of order $L = O(\log(1/\varepsilon))$ in total cost. In this sense, we obtain a provably efficient iterative procedure to solve (2).

It is important to note that the non-DRO version of the problem, namely (1), corresponding to the case $\delta = 0$ may not be strongly convex even if $\ell(\cdot)$ is strongly convex; see Remark 1 following Theorem 4. So, in principle, (1) may require $O(1/\varepsilon^2)$ stochastic gradient descent iterations to reach $O(\varepsilon)$ error of the optimal value. Indeed, if $\ell(\cdot)$ is convex, the problem is always convex in β (for $\delta \ge 0$) because the supremum of convex functions is convex.

Of course, $\delta > 0$ may be seen as a form of "regularization" in some cases, as discussed earlier, and this is a feature that could explain, at least intuitively, the convexity properties of the objective function. But the goal of Formulation (2) is not to regularize for the sake of making the problem better posed from an optimization standpoint. Rather, the point of Formulation (2) is enabling the flexibility in choosing effective DRO formulations (via (5)) in order to improve out-of-sample properties. This flexibility could come at a price in terms of computational tractability. The point of keeping the case $\delta = 0$ in mind as a benchmark is that such a price is not incurred, and therefore, our results enable modelers to use formulations such as (2) to improve out-of-sample performance based on side information as in the portfolio-optimization example mentioned earlier.

Another useful consequence of our results involves the application of standard sample average approximation statistical analysis results to optimal transport based DRO. This enables the direct application of results in conjunction with, for example, Shapiro et al. [31], to produce confidence regions for the solution of the DRO formulation.

Another interesting contribution of our analysis consists in studying the local structure of the worst-case optimal transport plan, including uniqueness and comparative statics results; see Theorems 6 and 7.

The structure of the optimal transport plan, we believe, could prove helpful in the development of statistical results to certify robustness and in providing insights for robustification in nonconvex objective functions. Some of the statistical implications are studied in Blanchet et al. [6].

1.3. Organization of the Paper

We now describe how to navigate the results in the paper. Throughout the rest of the paper, we introduce assumptions as we need them. Often these assumptions and the corresponding results that are obtained involved constants, which are surveyed in a table presented in Online Appendix D.

Section 2 sets the stage for our analysis by first obtaining the duality result (6). The duality result in (6) is given only under the assumption that $\ell(\cdot)$ is upper semicontinuous and $c(\cdot)$ is as in (5), assuming $A(x)$ is uniformly well conditioned in x.

In Section 2.2.1, under the assumption that $\ell(\cdot)$ is convex with at most quadratic growth and fourth order moments of P_0, we establish convexity and finiteness in the right-hand side of (6).

In Section 2.2.2, we add the assumptions that $\ell(\cdot)$ is twice differentiable with a natural nondegeneracy condition on P_0 and that the feasible set, B, is convex and compact. We characterize a useful region (compact and with

convenient analytical properties), called $\mathbb{V}$, which contains the dual optimizer $\lambda_*(\beta)$ parametrically as a function of each decision β. Then, we show smoothness and strong convexity in β of the right-hand side of (6) on $\mathbb{V}$.

Also in Section 2.2.2, now under a local strong convexity condition on $\ell(\cdot)$ and a strengthening of the nondegeneracy condition on P_0 mentioned earlier, we extend the smoothness and strong convexity of the right-hand side of (6) in both β and the dual variable $\lambda > 0$ provided that δ is chosen suitably small throughout $\mathbb{V}$.

The assumption that B is compact is imposed to simplify the strong convexity analysis and comparative statics (i.e., the structure of the worst-case distribution and comparative statics). We show in Section 2.2.3 that the compactness of B can be relaxed at the expense of additional technical burden.

The structure of the worst case is studied in Section 2.3 in Theorem 6. The result includes the amount of displacement (parametrically in δ) of the optimal transport plan and the existence of a Monge map (i.e., a direct "matching" between outcomes of P_0 and those of the worst-case distribution). We also discuss situations in which the optimal transport plan may not exist (even if an optimal solution to (6) exists) among other results.

Comparative statics results, including the uniqueness of the worst-case distribution as a Monge map as well as monotonicity in the amount of the displacement as a function of δ for every single outcome of P_0 are also discussed in Section 2.3. Also, the geometry of the worst-case transportation parametrically in δ is shown to follow straight lines.

In Section 3, we examine the wide range of algorithmic implications that follow from the results in earlier sections. Section 3.1 studies how to evaluate subgradients of the function $\ell_{rob}(\cdot)$ inside the expectation in (6). This is discussed under mild assumptions that do not require the loss $\ell(\cdot)$ to be differentiable. So the result can be applied to developing stochastic subgradient descent algorithms for nondifferentiable losses if derivatives and expectations can be swapped.

This swapping is explored in Section 3.2. We evaluate gradients for the expectation in the right-hand side of (6) under the assumptions imposed in Section 2.2.1, and a formal stochastic gradient descent scheme is given in Section 3.2.1, together with the corresponding iteration complexity analysis discussed in Section 3.2.2.

In Section 3.3, we discuss potential enhancements of the basic stochastic gradient descent strategy introduced in Section 3.2.1. These include a two-scale stochastic approximation scheme for dealing with the evaluation of the gradients of $\ell_{rob}(\cdot)$ and the case in which δ may not be small enough to apply the smoothness results from Section 2.2.2, and we need to deal with nondifferentiable losses as well.

We provide several specific examples in Section 4. These are designed to derive the expressions of the structural results that we present and explore the structure of the worst-case probability model and its behavior parametrically in δ. The various constants summarized required in the assumptions for application of our structural results are summarized in Online Appendix D. With the complexity of the SGD approach not scaling with the data size, the numerical study in Section 4.1.5 demonstrates the distinct computational advantage enjoyed by the proposed SGD scheme over second order cone formulations derived from piecewise linear approximation to the loss $\ell(\cdot)$.

In Section 4.2, we provide a discussion related to the portfolio optimization discussed earlier in the Introduction. The set of matrices, $A(x)$, is calibrated based on an implied volatility index, and P_0 is constructed based on several years of historical data for the S&P 500 index.

The proofs of our main structural results are given in Section 5. Additional discussion involving technical lemmas and propositions, which are auxiliary to our main structural results, are given in Online Appendix A. The discussion on the complexity of the line search, which underlies the gradient evaluation of $\ell_{rob}(\cdot)$, is given in Online Appendix B.

1.4. Notations

In the sequel, the symbol $\mathcal{P}(S)$ is used to denote the set of all probability measures defined on a complete separable metric space S. A collection of random variables $\{X_n : n \geq 1\}$ is said to satisfy the relationship $X_n = O_p(1)$ if it is tight; in other words, for any $\varepsilon > 0$, there exists a constant C_ε such that $\sup_n P(|X_n| > C_\varepsilon) < \varepsilon$. Following this notation, we write $X_n = O_p(g(n))$ to denote that the family $\{X_n/g(n) : n \geq 1\}$ is tight. The notation $X \sim P$ is to write that the law of X is P. For any measurable function $f : S \to \mathbb{R}$, we denote the essential supremum of f under measure $P \in \mathcal{P}(S)$ as $P\text{–ess-sup}_x f(x) := \inf\{a \in \mathbb{R} : P(f^{-1}(a, \infty)) = 0\}$. For any real-symmetric matrix A, we write $A \succeq 0$ to denote that A is a positive semidefinite matrix. The set of d-dimensional positive definite matrices with real entries is denoted by $\mathbb{S}_d^{++}$. The d-dimensional identity matrix is denoted by $\mathbb{I}_d$. The norm $\|\cdot\|$ is written to denote the ℓ_2–Euclidean norm unless specified otherwise. For any real vector x and $r > 0$, $\mathcal{N}_r(x)$ denotes the neighborhood $\mathcal{N}_r(x) := \{y : \|y - x\| < r\}$. We say that a collection of random variables $\{X_c : c \in \mathcal{C}\}$ is L_2–bounded (or bounded in the L_2–norm) if $\sup_{c\in\mathcal{C}} E\|X_c\|^2 < \infty$. For any function $f : \mathbb{R}^d \to \mathbb{R}$, the notation ∇f and $\nabla^2 f$ are written to denote, respectively, the gradient and Hessian of f. In instances in which it is helpful to clarify the variable with which

partial derivatives are taken, we resort to writing, for example, $\nabla_x f(x,y)$, $\nabla_x^2 f(x,y)$, or equivalently, $\partial f/\partial x$, $\partial^2 f/\partial x^2$ to denote that the partial derivative is taken with respect to the variable x. We write $\partial_+ f$, $\partial_- f$ to denote the right and left derivatives.

2. Dual Reformulation and Convexity Properties

In this section, we first reexpress the robust (worst-case) objective as in (6). Such reformulation, entirely in terms of the baseline probability distribution P_0, is useful in deriving the convexity and other structural properties to be examined in Sections 2.2–2.4. In turn, the reformulation (6) is helpful in developing the stochastic gradient–based iterative descent schemes described in Section 3.

2.1. Dual Reformulation

It follows from the definition of the optimal transport costs $D_c(P_0,P)$ (see (3)) that the worst-case objective in (6) equals

$$\sup\left\{\int \ell(\beta^T x')d\pi(x,x') : \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d),\ \pi(\cdot \times \mathbb{R}^d) = P_0(\cdot),\ \int c(x,x')d\pi(x,x') \le \delta\right\},$$

which is an infinite-dimensional linear program that maximizes $E_\pi[\ell(\beta^T X')]$ over all joint distributions π of pair $(X,X') \in \mathbb{R}^d \times \mathbb{R}^d$ satisfying the linear marginal constraints that the law of X is P_0 and the cost constraint that $E_\pi[c(X,X')] \le \delta$ (see Blanchet and Murthy [4, section 2.2] for details). A precise description of the state-dependent Mahalanobis transport costs $c(\cdot,\cdot)$ we consider in this paper is given in Assumption 1.

Assumption 1. *The transport cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is of the form*

$$c(x,x') = (x-x')^T A(x)(x-x'),$$

where $A : \mathbb{R}^d \to \mathbb{S}_d^{++}$ is such that (a) $c(\cdot)$ is lower-semicontinuous, and (b) there exist positive constants $\rho_{\min}$, $\rho_{\max}$ satisfying $\sup_{\|v\|=1} v^T A(x) v \le \rho_{\max}$ and $\inf_{\|v\|=1} v^T A(x) v \ge \rho_{\min}$, for P_0–almost every $x \in \mathbb{R}^d$.

As mentioned in the Introduction, a transport cost function satisfying Assumption 1 is not necessarily symmetric (and, hence, need not be a metric). The special case of $A(x)$ being the identity matrix (for all x) corresponds to $D_c^{1/2}(\cdot)$ being the well-known Wasserstein distance (in this case, the constants $\rho_{\max} = \rho_{\min} = 1$). Theorem 1 builds on a general strong duality result applicable for this linear program when the chosen transport cost function $c(x,x')$ is not necessarily a metric.

Theorem 1. *Suppose that $\ell : \mathbb{R} \to \mathbb{R}$ is upper semicontinuous. Then, under Assumption 1, the worst-case objective,*

$$\sup_{P: D_c(P_0,P) \le \delta} E_P\left[\ell(\beta^T X)\right] = \inf_{\lambda \ge 0} f_\delta(\beta,\lambda),$$

where $f_\delta(\beta,\lambda) := E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$, $\ell_{rob}(\beta,\lambda;x) := \sup_{\gamma \in \mathbb{R}} F(\gamma,\beta,\lambda;x)$, and

$$F(\gamma,\beta,\lambda;x) := \ell\left(\beta^T x + \gamma\sqrt{\delta}\beta^T A(x)^{-1}\beta\right) - \lambda\sqrt{\delta}(\gamma^2 \beta^T A(x)^{-1}\beta - 1). \tag{7}$$

For any $\beta \in B$, there exists a dual optimizer $\lambda_*(\beta) \ge 0$ such that $f_\delta(\beta,\lambda_*(\beta)) = \inf_{\lambda \ge 0} f_\delta(\beta,\lambda)$.

The proof of Theorem 1 is provided in Section 5.1.

2.2. Convexity and Smoothness Properties of the Dual DRO Objective

Here, we study the convexity and smoothness properties of the dual objective function $f_\delta(\beta,\lambda)$.

2.2.1. Convexity. We first identify conditions under which the function $f_\delta(\cdot)$ is proper and convex.

Assumption 2. *The loss function $\ell : \mathbb{R} \to \mathbb{R}$ is convex, and it satisfies the growth condition that $\kappa := \inf\{s \ge 0 : \sup_{u \in \mathbb{R}}(\ell(u) - su^2) < \infty\}$ is finite. In addition, the baseline distribution P_0 is such that $E_{P_0}\|X\|^4 < \infty$.*

Theorem 2. *The function $f_\delta : B \times \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$ is proper and convex when Assumptions 1 and 2 hold.*

The proof of Theorem 2 can be found in Section 5.1.

2.2.2. Smoothness and Strong Convexity. Next, we establish smoothness, strong convexity of $f_\delta(\cdot,\lambda)$ for fixed λ, and joint strong convexity of $f_\delta(\cdot)$, when restricted to the domain $\mathbb{V}$, under increasingly stronger sets of

assumptions. Although these assumptions are helpful in understanding smoothness and strong convexity properties, the development of iterative schemes in Online Appendix E does not require these stronger assumptions.

Assumption 3. *The loss function $\ell : \mathbb{R} \to \mathbb{R}$ is twice differentiable with bounded second derivatives. Specifically, we have a positive constant M such that $\ell''(\cdot) \le M$. Moreover, the baseline distribution P_0 is such that $\ell'(\beta^T X)$ is not identically 0 for any $\beta \in B$.*

Assumption 4. *The set $B \subseteq \mathbb{R}^d$ is convex and compact. Specifically, $\sup_{\beta \in B} \|\beta\| =: R_\beta < \infty$.*

Recall from Theorem 1 that $\arg\min_{\lambda \ge 0} f_\delta(\beta, \lambda)$ is not empty for every $\beta \in B$.

Proposition 1. *Suppose that Assumptions 1–4 hold. Then, for any $\beta \in B$ and dual optimizer $\lambda_*(\beta) \in \arg\min_{\lambda \ge 0} f_\delta(\beta, \lambda)$, we have $(\beta, \lambda_*(\beta)) \in \mathbb{V}$, where*

$$\mathbb{V} := \{(\beta, \lambda) \in B \times \mathbb{R}_+ : K_1 \|\beta\| \le \lambda \le K_2 \|\beta\|\}, \tag{8}$$

for some positive constants K_1, K_2 that can be explicitly determined in terms of parameters $\delta, M, R_\beta, \rho_{\max}, \rho_{\min}$.

To avoid clutter, we provide explicit characterizations for the constants K_1, K_2 in the proof of Proposition 1 (see Section 5.2) and as well in Online Table 2 (see Online Appendix D).

Theorem 3. *Suppose that Assumptions 1–4 are satisfied. Then, there exist positive constants δ_0, κ_0 such that the following hold: whenever $\delta < \delta_0$, the function $f_\delta : B \times \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$ satisfies the following properties:*

a. *The function $f_\delta(\cdot)$ is twice differentiable throughout the domain $\mathbb{V}$ with a uniformly bounded Hessian.*

b. *The second derivative of $f_\delta(\cdot)$ satisfies*

$$\frac{\partial^2 f_\delta}{\partial \beta^2}(\beta, \lambda) \succeq \sqrt{\delta} \kappa_0 \lambda^{-1} \mathbb{I}_d, \qquad \text{for } (\beta, \lambda) \in \mathbb{V}.$$

Theorem 3 identifies conditions under which the dual DRO objective $f_\delta(\cdot)$ has Lipschitz continuous gradients (smoothness) and also points toward strong convexity in terms of the parameter β (for any fixed λ). Similar to Proposition 1, we provide explicit characterizations for the constants δ_0, κ_0 in the proof of Theorem 3 in Section 5.3 (see also Online Appendix D for tables summarizing useful constants). We next focus on characterizing strong convexity jointly in the parameters (β, λ).

Assumption 5. *The loss function $\ell : \mathbb{R} \to \mathbb{R}$ is locally strongly convex. In addition, for every $\beta \in B$, the baseline distribution P_0 is such that there exist $c_1, c_2 \in (0, \infty)$, $p \in (0, 1)$ satisfying $P_0\Big(|\ell'(\beta^T X)| > c_1, |\beta^T X| > c_2 \|\beta\|\Big) \ge p$.*

Theorem 4. *Suppose that Assumptions 1–5 hold. Then, there exist constants $\delta_1 \in (0, \delta_0)$ and $\kappa_1 \in (0, \infty)$ such that, whenever $\delta < \delta_1$, the Hessian of the function $f_\delta : B \times \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$ satisfies*

$$\nabla^2 f_\delta(\theta) \succeq \sqrt{\delta} \kappa_1 \mathbb{I}_{d+1},$$

for $\theta \in \mathbb{V}$.

The proof of Theorem 4, along with an explicit characterization of the constant δ_1, is presented in Section 5.3. Theorem 4 identifies conditions under which $f_\delta(\cdot)$ is strongly convex (jointly over (β, λ)) when restricted to the set $\mathbb{V}$. Indeed, because of Proposition 1, it is sufficient to restrict attention to $\mathbb{V}$ to arrive at local strong convexity around $\arg\min_{\beta, \lambda} f_\delta(\beta, \lambda)$. To the best of our knowledge, Theorem 4 is the first result that presents strong convexity of the objective in Wasserstein distance–based DRO in a suitable sense. As is well known, strong convexity is a property that determines the iteration complexity of gradient-based descent methods. We utilize this in Section 3 to derive convergence properties of the proposed iterative schemes.

Remark 1. It is instructive to recall that $\ell(\cdot)$ being strongly convex does not mean $E_{P_0}[\ell(\beta^T X)]$ is necessarily strongly convex. For example, consider the underdetermined case of least-squares linear regression in which $\ell(u) = (y - u)^2$ and the number of samples $n < d$. If we take P_0 to be the empirical distribution corresponding to the n data samples (X_i, Y_i), the stochastic optimization objective to be minimized, $E_{P_0}[(Y - \beta^T X)^2] = n^{-1} \sum_{i=1}^n (Y_i - \beta^T X_i)^2$ is not strongly convex. Theorem 4 asserts that the respective dual DRO objective $f_\delta(\beta, \lambda)$ is, nevertheless, strongly convex in a region containing the minimizer (refer to an example in Section 4.1.3 for a discussion on how a DRO formulation of the least squares linear regression problem results in the dual objective of the form $f_\delta(\beta, \lambda)$). Thus, because of Theorem 4, for a considerable class of useful loss functions $\ell(\cdot)$, the DRO dual

objective to be minimized, $f_\delta(\beta,\lambda)$, is strongly convex in a suitable sense, even if the nonrobust counterpart $E_{P_0}[\ell(\beta^T X)]$ is not.

***2.2.2.1. Comments on Assumptions* 1–5.** Assumptions 1 and 2 ensure that the DRO objective (6) is convex and proper and that the strong duality utilized in Theorem 1 is indeed applicable. These nonrestrictive assumptions serve the purpose of clearly stating the framework considered. Indeed, Assumptions 1 and 2 are satisfied by a wide variety of loss functions $\ell(\cdot)$ and a flexible class of state-dependent Mahalanobis cost functions $c(\cdot)$, which include commonly used Euclidean metric, Mahalanobis distances as special cases. As we see in the proof of Theorem 4, the twice differentiability imposed in Assumption 3 is necessary to characterize the local strong convexity of f_δ by means of the positive definiteness of the Hessian of f_δ. The assumption of boundedness of the set B, though not necessary for strong convexity (see Section 2.2.3), is essential for guaranteeing the differentiability of $f_\delta(\cdot)$. Moving to Assumption 5, the positive probability requirement in Assumption 5 rules out the degeneracy that P_0 is not concentrated entirely in the regions in which either $|\ell'(\beta^T x)|$ or $|\beta^T x|$ is small. See Remark 4 (following the proof of Theorem 4 in Section 5.3) for an explanation of why the positivity of c_1, c_2 is necessary to identify the coefficient κ_1, which is independent of the ambiguity radius δ. We reiterate that the development of iterative schemes in Online Appendix E does not require Assumptions 3–5.

2.2.3. Strong Convexity Property for Noncompact *B*. As we see in Theorem 5, compactness of the set B (as in Assumption 4) is not crucial for strong convexity of the DRO objective around the minimizer. Assumption 4 is merely a simplifying assumption that allows the study of additional structural properties, such as differentiability, smoothness (see Theorem 3), and comparative statics (see Section 2.4). A proof of Theorem 5 is presented in Online Appendix A.

Theorem 5. *Suppose that Assumptions 1–3 and 5 are satisfied. In addition, suppose we have positive constants* k_1, k_2 *such that* $|u|\ell''(u) \le k_1 + k_2|\ell'(u)|$, *for* $u \in \mathbb{R}$. *Then, there exists* $\delta_2 > 0$ *such that, for every* $\delta < \delta_2$, *the following property holds: for any* $\beta \in B$, *we have positive constants* κ, r *such that*

$$f_\delta(\alpha\theta_1 + (1-\alpha)\theta_2) \le \alpha f_\delta(\theta_1) + (1-\alpha) f_\delta(\theta_2) - \frac{1}{2}\kappa\alpha(1-\alpha)\|\theta_1 - \theta_2\|^2,$$

for every $\theta_1, \theta_2 \in \mathcal{N}_r((\beta, \lambda_*(\beta)))$.

2.3. Structure of the Worst-Case Distribution

Fixing $\beta \in B$, we explain the structure of the worst-case distribution(s) that attains the supremum in (6) by utilizing the solution of the respective dual problem $\inf_{\lambda\ge 0} f_\delta(\beta,\lambda)$ (see Theorem 1). Recall the notation that $\lambda_*(\beta)$ attains the infimum in $\inf_{\lambda\ge 0} f_\delta(\beta,\lambda)$ for fixed $\beta \in B$. For each $\beta \in B$, $\lambda \ge 0$ and $x \in \mathbb{R}^d$, define the set of optimal solutions to (7) as

$$\Gamma^*(\beta,\lambda;x) = \left\{\gamma : F(\gamma,\beta,\lambda;x) = \sup_{c\in\mathbb{R}} F(c,\beta,\lambda;x)\right\}. \tag{9}$$

Finally, for a fixed $\beta \in B$, define

$$\lambda_{thr}(\beta) = \kappa\sqrt{\delta}(P_0 - \text{ess-sup}_x \beta^T A(x)^{-1}\beta).$$

Similarly, when Assumption 3 holds, define

$$\lambda'_{thr}(\beta) = \frac{1}{2}M\sqrt{\delta}(P_0 - \text{ess-sup}_x \beta^T A(x)^{-1}\beta).$$

Because $\kappa \le M/2$, we have $\lambda'_{thr}(\beta) \ge \lambda_{thr}(\beta)$ for every $\beta \in B$.

Theorem 6. *Suppose that Assumptions 1 and 2 hold and* $\beta \neq \mathbf{0}$. Take any dual optimizer $\lambda_*(\beta) \in \arg\min_{\lambda\ge 0} f_\delta(\beta,\lambda)$. Then,

a. *The dual optimizer* $\lambda_*(\beta)$ *is strictly positive unless* $\ell(\cdot)$ *is a constant function. If* $\ell(\cdot)$ *is indeed a constant function, then any distribution in* $\mathcal{U}_\delta(P_0) = \{P : D_c(P_0, P) \le \delta\}$ *attains the supremum in* (6).

b. *The dual optimizer* $\lambda_*(\beta) \ge \lambda_{thr}(\beta)$ *whenever* $\ell(\cdot)$ *is not a constant.*

c. *If* $\lambda_*(\beta) > \lambda_{thr}(\beta)$, *the law of*

$$X^* := X + \sqrt{\delta} G A(X)^{-1}\beta \tag{10}$$

attains the supremum in (6) *and satisfies* $E[c(X, X^*)] = \delta$; *here, the random variable* G *can be written as* $G := ZG_- + (1-Z)G_+$, *with* $G_- = \inf\Gamma(\beta,\lambda_*(\beta);X)$, $G_+ = \sup\Gamma(\beta,\lambda_*(\beta);X)$, P_0*–almost surely, and* Z *is an independent Bernoulli random variable satisfying* $P(Z=1) = (\bar{c}-1)/(\bar{c}-\underline{c})$, *where* $\bar{c} := E_{P_0}[G_+^2\beta^T A(X)^{-1}\beta]$ *and* $\underline{c} := E_{P_0}[G_-^2\beta^T A(X)^{-1}\beta]$;

d. *If* $\lambda_*(\beta) = \lambda_{thr}(\beta)$, *then a worst-case distribution attaining the supremum in* (6) *may not exist.*

e. *Under additional Assumption* 3, *if* $\lambda_*(\beta) > \lambda'_{thr}(\beta)$, *the set* $\Gamma^*(\beta, \lambda_*(\beta); x)$ *is a singleton for every* $x \in \mathbb{R}^d$. *Then, for the random variable* G *being the unique element in* $\Gamma^*(\beta, \lambda_*(\beta); X)$, P_0*–almost surely, we have that the law of* $X^* := X + \sqrt{\delta} G A(X)^{-1}\beta$ *is the only distribution that attains the supremum in* (6). *In addition,* $E[c(X, X^*)] = \delta$.

The proof of Theorem 6 is presented in Section 5.4.

Remark 2. Consider the case $\beta = \mathbf{0}$. Then, $\lambda = 0$ attains the minimum in $\min_{\lambda \geq 0} f_\delta(\mathbf{0}, \lambda)$, $\sup_{D_c(P_0, P) \leq \delta} E_{P_0}[\ell(\beta^T X)] = \ell(0)$, and any distribution in $\{P : D_c(P_0, P) \leq \delta\}$ attains the supremum.

2.4. Comparative Statics Analysis

In this section, we explain how the worst-case distribution structure explained in Section 2.3 changes for every realization of X when the radius of ambiguity δ is changed. Such a sample-wise description is facilitated by examining the derivative of the random variable G described in part (e) of Theorem 6, P_0- almost surely.

Theorem 7. *Suppose that the assumptions in Theorem* 3 *are satisfied. For any* $\delta \in (0, \delta_1)$ *and fixed* $\beta \in B \backslash \{\mathbf{0}\}$, *there exists a unique worst-case distribution* P^*_δ *that attains the supremum in* $\sup_{P: D_c(P_0, P) \leq \delta} E_P[\ell(\beta^T X)]$. *In particular, there exist random variables* $\{G_\delta : \delta \in (0, \delta_1)\}$ *such that*

a. *The law of* $X^*_\delta := X + \sqrt{\delta} G_\delta A(X)^{-1}\beta$ *is* P^*_δ;

b. $0 < \sqrt{\delta} G_\delta < \sqrt{\delta'} G_{\delta'}$ *whenever* $0 < \delta < \delta' < \delta_1$ *and* $\ell'(\beta^T X) > 0$;

c. $\sqrt{\delta'} G_{\delta'} < \sqrt{\delta} G_\delta < 0$ *whenever* $0 < \delta < \delta' < \delta_1$ *and* $\ell'(\beta^T X) < 0$; *and*

d. $G_\delta = 0$ *whenever* $\delta \in (0, \delta_1)$ *and* $\ell'(\beta^T X) = 0$.

Therefore, $\|X^*_\delta - X\| \leq \|X^*_{\delta'} - X\|$, P_0*–almost surely whenever* $0 < \delta < \delta' < \delta_1$.

The proof of Theorem 7 is presented in Section 5.4. Interestingly, Theorem 7 asserts that the trajectory $\{X^*_\delta : \delta \in [0, \delta_1)\}$ is a straight line P_0- almost surely with probability mass being transported to farther distances as δ increases in $[0, \delta_1)$. A pictorial description of this phenomenon is provided in Section 4.1.2 for numerical demonstration.

3. Algorithmic Implications of the Strong Convexity Properties

A key component of this section is a stochastic gradient–based iterative scheme that exhibits the following desirable convergence properties:

a. The proposed scheme enjoys optimal rates of convergence among the class of iterative algorithms that utilize first-order oracle information and possesses per-iteration effort not dependent on the size of the support of P_0.

b. Compared with the "nonrobust" counterpart $\inf_{\beta \in B} E_{P_0}[\ell(\beta^T X)]$, the proposed first-order method yields similar (or) superior rates of convergence for the DRO formulation (2).

In the case of data-driven problems in which P_0 is taken to be the empirical distribution, the size of the support of P_0 is simply the size of the data set. In such cases, property (a) is a particularly pleasant property as it allows Wasserstein distance–based DRO formulations to be amenable for big data problems that have become common in machine learning and operations research. Alternative approaches that directly solve the resulting convex program reformulations without resorting to stochastic gradients suffer from a large problem size when employed for large data sets (see, for example, Mohajerin Esfahani and Kuhn [19], Shafieezadeh-Abadeh et al. [29]). Further, the proposed stochastic gradient–based approaches are also immediately applicable to problems in which P_0 has uncountably infinite support.

Property (b) makes sure that computational intractability is not a reason that should deter the use of the DRO approach toward optimization under uncertainty. In fact property (b) describes that it may be computationally more advantageous, in addition to the desired robustness, to work with the DRO formulation (2) compared with its stochastic optimization counterpart $\inf_{\beta \in B} E_{P_0}[\ell(\beta^T X)]$. As we see in Section 3.2, this computational benefit for the proposed stochastic gradient descent scheme is endowed by the strong convexity properties of the dual objective $f_\delta(\beta, \lambda)$ derived in Theorem 4. Guided by the strong convexity structure of $f_\delta(\beta, \lambda)$, we also discuss enhancements to the vanilla SGD scheme in Sections 3.3.1 and 3.3.2.

3.1. Extracting First-Order Information

Recall the univariate maximization (7) that defines $\ell_{rob}(\beta, \lambda; x)$ for $\beta \in B, \lambda \geq 0, x \in \mathbb{R}^d$ and the set of maximizers $\Gamma^*(\beta, \lambda; x)$ in (9). With the DRO objective (6) being related to the dual objective $f_\delta(\beta, \lambda) := E_{P_0}[\ell_{rob}(\beta, \lambda; X)]$ as in Theorem 1, the minimization can be restricted to the effective domain

$$\mathbb{U} := \{(\beta, \lambda) \in B \times \mathbb{R}_+ : E_{P_0}[\ell_{rob}(\beta, \lambda; X)] < \infty\}. \tag{11}$$

Lemma 1, whose proof is presented in Online Appendix A, provides a characterization of the effective domain $\mathbb{U}$. Here, recall the earlier definition that $\lambda_{thr}(\beta)$ is the P_0–essential supremum of $\sqrt{\delta}\kappa\beta^T A(x)^{-1}\beta$. Define

$$\mathbb{U}_1 := \{(\beta,\lambda) \in B \times \mathbb{R}_+ : \lambda > \lambda_{thr}(\beta)\} \quad \text{and} \quad \mathbb{U}_2 : = \{(\beta,\lambda) \in \mathrm{B} \times \mathbb{R}_+ : \lambda \geq \lambda_{thr}(\beta)\}.$$

Lemma 1. *Suppose that Assumptions 1 and 2 hold. Then, for any $\beta \in B, \lambda \geq 0$ and $x \in \mathbb{R}^d$,*
a. *$\Gamma^*(\beta,\lambda;x)$ is nonempty and $\ell_{rob}(\beta,\lambda;x)$ is finite if $\lambda > \kappa\sqrt{\delta}\beta A(x)^{-1}\beta$;*
b. *$\Gamma^*(\beta,\lambda;x)$ is empty and $\ell_{rob}(\beta,\lambda;x) = \infty$ if $\lambda < \kappa\sqrt{\delta}\beta A(x)^{-1}\beta$.*
Consequently, $\mathbb{U}_1 \subseteq \mathbb{U} \subseteq \mathbb{U}_2$.

Lemma 2. *Suppose that Assumptions 1(a) and 2 hold. Then, the function $\ell_{rob}(\beta,\lambda;x)$ is convex in $(\beta,\lambda) \in B \times \mathbb{R}_+$ for any $x \in \mathbb{R}^d$.*

Proposition 2 utilizes the envelope theorem (see Milgrom and Segal [18]) to characterize the gradients of $\ell_{rob}(\cdot)$. Recall that we use $\partial_-\ell(u)$, $\partial_+\ell(u)$ to denote the left and right derivatives of $\ell(\cdot)$ when evaluated at $u \in \mathbb{R}$.

Proposition 2. *Suppose that $\ell : \mathbb{R} \to \mathbb{R}$ satisfies Assumption 2 and is of the form $\ell(u) = \max_{i=1,\ldots,K}\ell_i(u)$ for continuously differentiable $\ell_i : \mathbb{R} \to \mathbb{R}$ and a positive integer K. The following statements hold for P_0–almost every x:*
a. *The set of maximizers, $\Gamma^*(\beta,\lambda;x) \neq \emptyset$, for any $(\beta,\lambda) \in \mathbb{U}_1$.*
b. *The maps $\lambda \longmapsto \ell_{rob}(\beta,\lambda;x)$, $\beta_j \longmapsto \ell_{rob}(\beta,\lambda;x)$ are absolutely continuous for $(\beta,\lambda) \in \mathbb{U}_1$, and their directional derivatives are given by*

$$\frac{\partial_-\ell_{rob}}{\partial\beta_j}(\beta,\lambda;x) = \min_{\gamma\in\Gamma^*(\beta,\lambda;x)} \partial_-\ell\Big(\beta^T(x+\sqrt{\delta}\gamma A(x)^{-1}\beta)\Big)(x+\sqrt{\delta}\gamma A(x)^{-1}\beta)_j, \tag{12a}$$

$$\frac{\partial_+\ell_{rob}}{\partial\beta_j}(\beta,\lambda;x) = \max_{\gamma\in\Gamma^*(\beta,\lambda;x)} \partial_+\ell\Big(\beta^T(x+\sqrt{\delta}\gamma A(x)^{-1}\beta)\Big)(x+\sqrt{\delta}\gamma A(x)^{-1}\beta)_j, \tag{12b}$$

$$\frac{\partial_-\ell_{rob}}{\partial\lambda}(\beta,\lambda;x) = \min_{\gamma\in\Gamma^*(\beta,\lambda;x)} -\sqrt{\delta}\Big(\gamma^2\beta^T A(x)^{-1}\beta - 1\Big), \tag{12c}$$

$$\frac{\partial_+\ell_{rob}}{\partial\lambda}(\beta,\lambda;x) = \max_{\gamma\in\Gamma^*(\beta,\lambda;x)} -\sqrt{\delta}\Big(\gamma^2\beta^T A(x)^{-1}\beta - 1\Big). \tag{12d}$$

Furthermore, $\lambda \longmapsto \ell_{rob}(\beta,\lambda;x)$ is differentiable if and only if $\left\{\frac{\partial_+F}{\partial\lambda}(\gamma,\beta,\lambda;x), \frac{\partial_-F}{\partial\lambda}(\gamma,\beta,\lambda;x) : \gamma \in \Gamma^*(\beta,\lambda;x)\right\}$ is a singleton. Likewise, for j in $\{1,\ldots,d\}$ $\beta_j \longmapsto \ell_{rob}(\beta,\lambda;x)$ is differentiable if and only if the respective set $\left\{\frac{\partial_+F}{\partial\beta_j}(\gamma,\beta,\lambda;x), \frac{\partial_-F}{\partial\beta_j}(\gamma,\beta,\lambda;x) : \gamma \in \Gamma^*(\beta,\lambda;x)\right\}$ is a singleton. When all these sets are singleton, if we let $\tilde{x} := x + \sqrt{\delta}gA(x)^{-1}\beta$ for any $g \in \Gamma^*(\beta,\lambda;x)$, and then the derivative is given by

$$\frac{\partial\ell_{rob}}{\partial\beta}(\beta,\lambda;x) = \ell'\Big(\beta^T\tilde{x}\Big)\tilde{x} \quad \text{and} \quad \frac{\partial\ell_{rob}}{\partial\lambda}(\beta,\lambda;x) = -\sqrt{\delta}\Big(g^2\beta^T A(x)^{-1}\beta - 1\Big). \tag{13}$$

A proof of Proposition 2 can be found in Online Appendix A. Recall that a simple subgradient descent (or) stochastic subgradient descent for solving the nonrobust problem $\inf_{\beta\in B} E_{P_0}[\ell(\beta^T X)]$ assumes access to first-order oracle evaluations $\ell(\cdot)$ and $\partial_+\ell(\cdot)$, $\partial_-\ell(\cdot)$. Likewise, because of the characterization in Proposition 2, all the function evaluation information required to implement a stochastic subgradient descent type iterative scheme for minimizing its robust counterpart $f_\delta(\beta,\lambda)$ are evaluations of $\ell(\cdot)$ and $\partial_+\ell(\cdot)$, $\partial_-\ell(\cdot)$. Indeed, when it is feasible to exchange the gradient (or subgradient) and the expectation operators in $\nabla_{(\beta,\lambda)}E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$ (as in Proposition 3 in Section 3.2), the subgradients of $\ell_{rob}(\beta,\lambda;X)$ yield noisy subgradients of $f_\delta(\beta,\lambda)$. For a given $(\beta,\lambda) \in \mathbb{U}_1$, a univariate optimization procedure, such as bisection (or) Newton–Raphson methods. is used to solve (7).

3.2. A Stochastic Gradient Descent Scheme for Differentiable $f_\delta(\cdot)$

For ease of notation, we write θ in place of $(\beta,\lambda) \in B \times \mathbb{R}_+$. We describe the algorithm initially assuming that the conditions in Theorem 3 are satisfied. Then, as a consequence of Theorem 3, we have that $f_\delta(\cdot)$ is differentiable over the set $\mathbb{V}$. Here, recall the characterization of the set $\mathbb{V}$ in Proposition 1 and the constants K_1, K_2 therein and the constant R_β in Assumption 4. Define the set

$$\mathbb{W} := \{(\beta,\lambda) \in B \times \mathbb{R} : K_1\|\beta\| \leq \lambda \leq K_2 R_\beta\}. \tag{14}$$

See that $\mathbb{W}$ is a closed convex set containing $\mathbb{V}$. Therefore, when $\delta < \delta_0$, as a consequence of Theorem 1 and Proposition 1, we have that

$$\inf_{\beta\in B} \sup_{P:D_c(P,P_0)\le\delta} E_P\big[\ell(\beta^T X)\big] = \inf_{\theta\in\mathbb{W}} f_\delta(\theta).$$

Proposition 3. *Suppose that Assumptions 1–4 hold and $\delta < \delta_0$. Then, $E_{P_0}[\nabla_\theta \ell_{rob}(\theta; X)]$ is well defined and*

$$\nabla_\theta f_\delta(\theta) = E_{P_0}[\nabla_\theta \ell_{rob}(\theta; X)],$$

for any $\theta \in \{(\beta,\lambda) : \beta \in B, \lambda > \lambda'_{thr}(\beta)\} \supset \mathbb{W}$.

The proof of Proposition 3 is available in Online Appendix A.

3.2.1. The Iterative Scheme. Because of Proposition 3, samples of the random vector $\nabla_\theta \ell_{rob}(\theta; X)$, where $X \sim P_0$, are unbiased estimators of the desired gradient $\nabla_\theta f_\delta(\theta)$ and are called "stochastic gradients" of $f_\delta(\theta)$. Utilizing these noisy gradients, we generate averaged iterates $\{\bar\theta_k : k \ge 1\}$ according to the following scheme:

Fix $\xi \ge 0$ and initialize $\bar\theta_0 = \theta_0 \in \mathbb{W}$. For $k > 0$, given the iterate θ_{k-1} from the $(k-1)$th step,

a. Generate an independent sample X_k from the distribution P_0,

b. Compute $\nabla_\theta \ell_{rob}(\theta_k; X_k)$ characterized in (13) by solving $\sup_{\gamma\in\mathbb{R}} F(\gamma, \theta; X_k)$, and

c. Compute the kth iterate θ_k and its weighted running average $\bar\theta_k$ as follows:

$$\theta_k := \Pi_{\mathbb{W}}(\theta_{k-1} - \alpha_k \nabla_\theta \ell_{rob}(\theta_{k-1}; X_k)) \quad \text{and} \quad \bar\theta_k = \left(1 - \frac{\xi+1}{k+\xi}\right)\bar\theta_{k-1} + \frac{\xi+1}{k+\xi}\theta_k, \tag{15}$$

where $\Pi_{\mathbb{W}}(\cdot)$ denotes the projection operation on to the closed convex set $\mathbb{W}$ and $(\alpha_k)_{k\ge1}$ is referred to as the step-size sequence (or) learning rate of the iterative scheme. A closed-form expression for the projection $\Pi_{\mathbb{W}}$ is given in Online Appendix C, and a detailed algorithmic description of these steps is described in Online Appendix E.

Assumption 6. *The step-size sequence $(\alpha_k)_{k\ge1}$ is taken to satisfy $\alpha_k = \alpha k^{-\tau}$, for some constants $\alpha > 0$ and $\tau \in [1/2, 1]$.*

The iterates $(\theta_k)_{k\ge1}$ are the classical Robbins–Monro iterates with slower step sizes (see Robbins and Monro [27]). If $\xi = 0$ in the definition of $\bar\theta_k$ in (15), the iterate $\bar\theta_k$ is simply the running average of $\theta_1, \ldots \theta_{k-1}$, and the averaging scheme is the well-known Polyak–Ruppert averaging for stochastic gradient descent (see Polyak and Juditsky [26] and references therein). On the other hand, the averaging scheme with $\xi > 0$ is referred to as polynomial-decay averaging (see Shamir and Zhang [30]).

3.2.2. Rates of Convergence. Our objective here is to characterize the convergence of $(f_\delta(\bar\theta_k))_{k\ge1}$ for the iteration scheme (15). Let $f_* := \inf_{\theta\in B\times\mathbb{R}_+} f_\delta(\theta)$ be the optimal value. It is well known that stochastic gradient descent schemes for smooth objective functions enjoy $f_\delta(\bar\theta_k) - f_* = O_p(k^{-1})$ rate of convergence if f_δ is strongly convex and $f_\delta(\theta_k) - f_* = O_p(k^{-1/2})$ if f_δ is simply convex for suitable choices of step sizes (see, for example Shamir and Zhang [30] and references therein). Although $f_\delta(\cdot)$ is convex for all $\delta \ge 0$, it follows from Theorem 4 that $f_\delta(\cdot)$ is locally strongly convex in the region containing the optimizer when $\delta < \delta_1$. As a result, we have the following better rate of convergence for $f_\delta(\bar\theta_k) - f_*$ when $\delta < \delta_1$. The proof of Proposition 4 is presented in Section 5.5.

Proposition 4. *Suppose that Assumptions 1–4 hold. Then, we have*

a. *$f_\delta(\bar\theta_k) - f_* = O_p(k^{-1/2})$ if $\delta < \delta_0$, $\xi \ge 1$ in (15) and $\tau = 1/2$ in Assumption 6.*

b. *$f_\delta(\bar\theta_k) - f_* = O_p(k^{-1})$ if $\delta < \delta_1$, $\xi = 0$, $\tau \in (1/2, 1)$ in Assumption 6, and Assumption 5 is satisfied.*

For the strongly convex case, the averaged procedure endows the sequence $(f_\delta(\bar\theta_k))_{k\ge1}$ with the robustness property that the precise choice of step size $(\alpha_k)_{k\ge1}$ does not affect the convergence behavior as long as the step size choice satisfies Assumption 6. Contrast this with the vanilla stochastic approximation iterates $(\theta_k)_{k\ge1}$ with step size $\alpha_k = \alpha k^{-1}$, in which case the constant α has to be chosen larger than a threshold that depends on the Hessian of f_δ at θ minimizing $f_\delta(\theta)$, in order to have $f_\delta(\theta_k) - f_* = O_p(k^{-1})$ (see, for example, Moulines and Bach [22], Nemirovski et al. [23] for discussions on the effect of step sizes on error $f_\delta(\theta_k) - f_*$).

Recall that δ_0, δ_1 are positive constants that do not depend on the size of the support of P_0. For data-driven optimization problems, the radius of ambiguity, δ, is typically chosen to decrease to zero with the number of data samples n (see, for example, Blanchet et al. [5], Shafieezadeh-Abadeh et al. [29]). Therefore the requirement that $\delta < \delta_1$ is typically satisfied in practice in data-driven applications.

Indeed, if $\delta < \delta_1$, because of Proposition 4(b)), it suffices to terminate after $O_p(\varepsilon^{-1})$ iterations in order to obtain an iterate $\bar\theta_k$ that satisfies $f_\delta(\bar\theta_k) - f_* \le \varepsilon$. On the other hand, if $\delta > \delta_1$, we require the usual $O_p(\varepsilon^{-2})$ iteration

complexity to obtain $f_\delta(\theta_k) - f_* \le \varepsilon$, which is identical to the sample complexity of stochastic gradient descent for the nonrobust problem $\inf_\beta E_{P_0}[\ell(\beta^T X)]$ in the presence of convexity (see, for example, Shamir and Zhang [30]). Here, recall from the discussion following Theorem 4 that the nonrobust stochastic optimization objective $\inf_\beta E_{P_0}[\ell(\beta^T X)]$ need not be strongly convex even if $\ell(\cdot)$ is strongly convex, whereas the corresponding worst-case objective $f_\delta(\beta, \lambda)$ is jointly strongly convex in (β, λ) more generally under the conditions identified in Theorem 4.

As a result, if we let L denote the complexity of the univariate line search that solves $\sup_{\gamma\in\mathbb{R}} F(\gamma, \theta; x)$ for any $(\beta, \lambda) \in \mathbb{W}$, then the computational effort involved in solving (2) scales as $O_p(\varepsilon^{-1}L)$ when $\delta < \delta_1$ and $O_p(\varepsilon^{-2}L)$ when $\delta \in [\delta_1, \delta_0)$. As mentioned earlier, this complexity does not scale with the size of the support of P_0 for a given δ. See Online Appendix B for a brief discussion on L, the complexity introduced by line search schemes.

The analysis of stochastic gradient descent with small bias can be done without significant complications under regularity conditions. The following result summarizes the overall rate of convergence analysis for the classical Robbins–Monro iterates $(\theta_k : k \ge 1)$, including bias induced by the line search in the strongly convex case. The proof of Proposition 5 is presented in Online Appendix A.

Proposition 5. *Suppose that Assumptions 1–6 hold and $\delta < \delta_1$. At the kth iteration, the bisection method is employed with at least $\tau\log_2(k) - \log_2(\alpha) + 2\log_2(1 + \|X_k\|)$ cuts to compute $\nabla_\theta \ell_{rob}(\theta_{k-1}; X_k)$. Then, we have*

a. *$f_\delta(\theta_k) - f_* = O_p(k^{-\tau})$ if $\tau \in (1/2, 1)$ in Assumption 6.*

b. *$f_\delta(\theta_k) - f_* = O_p(k^{-1})$ if α is larger than the smallest eigenvalue of $\nabla^2_\theta f_\delta(\theta_*)$ and $\tau = 1$ in Assumption 6.*

Remark 3. Proposition 5 indicates that, if the bisection method is applied with $O(\log_2(k))$ cuts at the kth iterates, then the classical Robbins–Monro algorithm still achieves the optimal $O_p(1/k)$ rate even if the bias of line search is taken into consideration. The assumption in part (b) on requiring a lower bound on α is standard. Typically, avoiding an estimate of such a lower bound can be done by Polyak–Ruppert–Juditsky averaging and choosing $\tau \in (1/2, 1)$. This is most often studied in the case of unbiased gradients. An adaptation is required for the case of biased gradients. Although we believe that such an adaptation should be quite doable, we do not pursue it in this paper as it would be a significant distraction from our objective. Our goal here is to showcase the applicability of the structural results in Section 2.2 toward designing efficient algorithms for DRO based on flexible cost functions.

To complete this discussion, recall that the dual formulation,

$$\inf_{\lambda\ge 0} E_{P_0}\left[\sup_{\gamma\in\mathbb{R}} F(\gamma, \beta, \lambda; X)\right],$$

that we are working with is a result of the change of variables $c = \sqrt{\delta}\gamma\beta^T A(X)^{-1}\beta$ and $\lambda\sqrt{\delta}$ to λ in the proof of Theorem 1. Evidently, this change of variables involves scaling by a factor $\sqrt{\delta}$. It is a consequence of this scaling by $\sqrt{\delta}$ that an optimal $\lambda_*(\beta)$ is bounded, thus allowing the optimization to be restricted to values of λ over a compact interval $[0, K_2 R_\beta]$ regardless of how small the radius of ambiguity δ is. Moreover, if we let $g_\delta(x)$ denote a maximizer for the inner maximization $\sup_{\gamma\ge 0} F(\gamma, \beta, \lambda_*(\beta); x)$ for any δ, x and a fixed $\beta \in B$, we also witness in Proposition 9(b) that $g_\delta(X) = O_p(1)$, as $\delta \to 0$. These two properties ensure that the inner and outer optimization problems $\inf_{\lambda\ge 0} E_{P_0}\left[\sup_{\gamma\in\mathbb{R}} F(\gamma, \beta, \lambda; X)\right]$ are well conditioned and their solutions remain scale-free (with respect to δ).

For algorithms that directly proceed with the dual reformulation in Blanchet and Murthy [4, theorem 1] or Gao and Kleywegt [12, theorem 1] without employing the described scaling of variables by factor $\sqrt{\delta}$, the resulting dual formulation has the property that the solutions to the inner and outer optimization problems are $O_p(\sqrt{\delta})$ and $O(\delta^{-1/2})$, respectively. Consequently, the local strong convexity coefficient of the dual reformulation obtained without scaling can be shown to be $O(\delta)$, which is inferior when compared with the $O(\sqrt{\delta})$ strong convexity coefficient that we have identified in Theorem 1. Indeed, the focus on strong convexity and its effect of computational performance in this paper has helped bring out this nuanced and important effect of the scaling that appears to be absent in the existing algorithmic approaches for Wasserstein DRO.

3.3. Enhancements to the SGD Scheme in Section 3.2

Our focus in this section is to describe natural enhancements to the vanilla SGD scheme described in Section 3.2 by utilizing the convexity characterizations in Section 2.2.

3.3.1. A Two Time Scale Stochastic Approximation Scheme. Because λ is an auxiliary variable introduced by the duality formulation, it is rather natural to update the variables β and λ at different learning rates (step sizes) as follows: given iterate $(\beta_{k-1}, \lambda_{k-1})$, generate a sample X_k independently from P_0 in order to update as follows:

$$\tilde{\beta}_k = \beta_{k-1} - \alpha_k \frac{\partial f_\delta}{\partial \beta}(\beta_{k-1}, \lambda_{k-1}; X_k), \tag{16a}$$

$$\tilde{\lambda}_k = \lambda_{k-1} - \gamma_k \frac{\partial f_\delta}{\partial \lambda}(\beta_{k-1}, \lambda_{k-1}; X_k), \text{ and} \tag{16b}$$

$$\theta_k = \Pi_{\mathbb{W}}\left((\tilde{\beta}_k, \tilde{\lambda}_k)\right). \tag{16c}$$

Here, the step-sizes $(\alpha_k)_{k\geq 1}, (\gamma_k)_{k\geq 1}$ satisfy the step-size requirement in Assumption 6 with $\tau \in (1/2, 1)$ and $\alpha_k/\gamma_k \to 0$. Because α_k is very small relative to γ_k, the iterates β_k remain relatively static compared with λ_k, thus having an effect of fixing β_k and running (16b) for a long time. As a result, the iterates λ_k appear "most of the time" as $\lambda_*(\beta_k)$ in the view of β_k, thus resulting in effective updates of the form

$$\beta_k = \beta_{k-1} - \alpha_k \frac{\partial f_\delta}{\partial \beta}(\beta_{k-1}, \lambda_*(\beta_{k-1}); X_k).$$

Once again, we consider the averaged iterates $\bar{\theta}_k$, defined as in (15) with $\xi = 0$. Similar to Section 3.2, if we let $f_* := \inf_{\theta \in B \times \mathbb{R}_+} f_\delta(\theta)$, it can be argued that $f_\delta(\bar{\theta}_k) - f_* = O_p(k^{-1})$ in the presence of strong convexity (see Mokkadem and Pelletier [20, theorem 2]) that holds in the $\delta < \delta_1$ case. As a result, if $\delta < \delta_1$, it suffices to terminate after $O_p(\varepsilon^{-1})$ iterations in order to obtain an iterate $\bar{\theta}_k$ that satisfies $f_\delta(\bar{\theta}_k) - f_* \leq \varepsilon$. We leave it as a question for future research to develop a precise understanding of the effect of two time scales in affecting the convergence behavior.

3.3.2. Line Search–Based SGD Scheme. When $\delta < \delta_0$, Theorem 3 asserts that $f_\delta(\beta, \lambda)$ satisfies strong convexity in the variable β for every fixed λ. This strong convexity in variable β holds even if $f_\delta(\beta, \lambda)$ may not be jointly strongly convex in (β, λ) (for example, when $\delta \in [\delta_1, \delta_0)$). We make use of this observation in this section to describe an SGD scheme that (a) quickly evaluates $h(\lambda) := \inf_{\beta \in B} f_\delta(\beta, \lambda)$ for any given λ and (b) utilizes univariate line search for minimizing $h(\cdot)$ in a suitable interval.

Because $f_\delta(\cdot)$ is a convex function, the partial minimization $h(\lambda) := \inf_{\beta \in B} f_\delta(\beta, \lambda)$ defines a univariate convex function in λ. For any fixed $\lambda > 0$, consider stochastic gradient descent iterates of the form

$$\beta_k := \beta_{k-1} - \alpha_k \frac{\partial f_\delta}{\partial \beta}(\beta_{k-1}, \lambda; X_k), \quad \text{and} \quad \bar{\beta}_k := \frac{1}{k}\sum_{i=1}^{k} \beta_i,$$

where $(X_k)_{k\geq 1}$ are independent and identically distributed (i.i.d.) samples of P_0, and the step sizes $(\alpha_k)_{k\geq 1}$ satisfy the requirement in Assumption 6 with $\tau \in (1/2, 1)$ and $\xi = 0$. Then, it follows from the strong convexity characterization in Theorem 3 that $f_\delta(\bar{\beta}_k, \lambda) - h(\lambda) = O_p(k^{-1})$ if $\delta < \delta_0$. With the ability to evaluate the function $h(\lambda) = \inf_{\beta \in B} f_\delta(\beta, \lambda)$ within a desired precision, any standard line search method, such as the triangle section method (see den Boef and den Hertog [11, algorithm 3]), that exploits the convexity of $h(\cdot)$ to achieve linear convergence for line search can be employed to evaluate $\min_\lambda h(\lambda)$ to any desired precision.

With line searches requiring identification of an interval (in which the minimum is attained) to begin with, we restrict the line search over λ to the interval $[0, K_2 R_\beta]$. This is because, as a result of Proposition 1 and that $\|\beta\| \leq R_\beta$, we have that the interval $[0, K_2 R_\beta]$ contains optimal $\lambda_*(\beta)$ for every $\beta \in B$. It can be argued that the described approach results in iteration complexity of $O_p(\varepsilon^{-1}\text{poly}(\log \varepsilon^{-1}))$ to solve $\min f_\delta(\beta, \lambda)$ within $\varepsilon-$precision when $\delta < \delta_0$. We do not pursue this derivation here as our objective is to simply demonstrate the versatility of applications of the structural insights given by the results in Section 2.2.

Likewise, one could consider a variety of algorithms that accelerate SGD at a greater computational cost per iteration; such algorithms utilize either variance reduction (see, for example, Defazio et al. [10], Johnson and Zhang [16]) or momentum-based acceleration (see Allen-Zhu [1]). The strong convexity results in Section 2.2 could be used to establish improved rates of convergence for such extensions as well.

3.4. SGD for Nondifferentiable f_δ

The function $f_\delta(\cdot)$ need not be differentiable when the radius of ambiguity δ exceeds δ_0 (or) when the set B is not bounded. The iterative algorithms described in Sections 3.2 and 3.3 rely on restricting the iterates θ_k to the set $\mathbb{W}$. Such an approach is not feasible when $\delta > \delta_0$. In that case, with the characterization of the effective domain of f_δ

as in Lemma 1, define the family of closed convex sets, $(\mathbb{U}_\eta : \eta \geq 0)$ as

$$\mathbb{U}_\eta := \{(\beta, \lambda) \in B \times \mathbb{R}_+ : \lambda \geq \lambda_{thr}(\beta) + \eta\}. \tag{17}$$

Let $\partial f_\delta(\beta, \lambda)$ and $\partial \ell_{rob}(\beta, \lambda; x)$, respectively, be the set of subgradients of $f_\delta(\cdot)$ and $\ell_{rob}(\cdot\,; x)$ at (β, λ). Likewise, let $\partial \ell(u) := \text{conv}\{\partial_- \ell(u)/\partial u,\ \partial_+ \ell(u)/\partial u\}$ denote the subgradient set of the univariate function $\ell(\cdot)$ evaluated at u. Then, it follows from Proposition 2(b) that the set

$$D(\beta, \lambda; x) := \text{conv}\left\{ \begin{pmatrix} \partial \ell(\beta^T \tilde{x})\tilde{x} \\ \sqrt{\delta}\left(1 - g^2 \beta^T A(x)^{-1} \beta\right) \end{pmatrix} : \begin{matrix} \tilde{x} = x + \sqrt{\delta} g A(x)^{-1} \beta, \\ g \in \Gamma^*(\beta, \lambda; x) \end{matrix} \right\} \tag{18}$$

comprises the subgradient set $\partial \ell_{rob}(\beta, \lambda; x)$. Similar to Proposition 3, Proposition 6 helps in characterizing noisy subgradients of $f_\delta(\cdot)$.

Proposition 6. *Suppose that Assumptions 1 and 2 are satisfied and the loss $\ell(\cdot)$ is of the form $\ell(u) = \max_{i=1,\ldots,K} \ell_i(u)$ for continuously differentiable $\ell_i : \mathbb{R} \to \mathbb{R}$ and a positive integer K. For any $\eta > 0$ and fixed $(\beta, \lambda) \in \mathbb{U}_\eta$, let $(X, h(\beta, \lambda; X))$ be such that $X \sim P_0$ and $h(\beta, \lambda, X) \in D(\beta, \lambda; X)$, P_0−almost surely. Then, $E[h(\beta, \lambda; X)]$ is well defined, and $E[h(\beta, \lambda; X)] \in \partial f_\delta(\beta, \lambda)$.*

The proof of Proposition 6 is available in Online Appendix A. Following Proposition 6, consider an iterative scheme utilizing noisy subgradients as follows. Given fixed $\eta > 0, \xi \geq 1$ and iterate $\theta_{k-1} = (\beta_{k-1}, \lambda_{k-1})$ from the $(k-1)$st iteration, the kth iterate is computed as follows:

$$\theta_k := \Pi_{\mathbb{U}_\eta}(\theta_{k-1} - \alpha_k H_k) \quad \text{and} \quad \bar{\theta}_k = \left(1 - \frac{\xi + 1}{k + \xi}\right)\bar{\theta}_{k-1} + \frac{\xi + 1}{k + \xi}\theta_k, \tag{19}$$

where the step-size sequence $(\alpha_k)_{k \geq 1}$ satisfies Assumption 6 with $\tau = 1/2$ and H_k is computed as follows:

a. Generate a sample X_k independently from the distribution P_0;

b. Pick any $g \in \Gamma^*(\beta, \lambda; X_k)$ by solving the univariate search $\sup_{\gamma \in \mathbb{R}} F(\gamma, \beta_{k-1}, \lambda_{k-1}; X_k)$;

c. Let $\tilde{X}_k := X_k + \sqrt{\delta} g A(X_k)^{-1} \beta$, and take $H_k \in D(\beta_{k-1}, \lambda_{k-1}; X_k)$ as

$$H_k := \begin{pmatrix} L' \tilde{X}_k \\ \sqrt{\delta}\left(1 - g^2 \beta_{k-1}^T A(X_k)^{-1} \beta_{k-1}\right) \end{pmatrix},$$

where L' is selected uniformly at random from the interval $[\partial_- \ell(\beta_{k-1}^T \tilde{X}_k)/\partial u,\ \partial_+ \ell(\beta_{k-1}^T \tilde{X}_k)/\partial u] =: \partial \ell(\beta_{k-1}^T \tilde{X}_k)$.

It is immediate from (18) that $H_k \in D(\beta_{k-1}, \lambda_{k-1}; X_k)$. Then, because of Proposition 6, we have that $EH_k \in \partial f_\delta(\beta_{k-1}, \lambda_{k-1})$. Because of the convexity of $f_\delta(\cdot)$ characterized in Theorem 4, we have the following rates of convergence for $f_\delta(\bar{\theta}_k) - f_*$, as $k \to \infty$. The proof of Proposition 7 is presented in Online Appendix A.

Proposition 7. *Suppose that Assumptions 1 and 2 are satisfied and the loss $\ell(\cdot)$ is of the form $\ell(u) = \max_{i=1,\ldots,K} \ell_i(u)$ for continuously differentiable $\ell_i : \mathbb{R} \to \mathbb{R}$ and a positive integer K. In addition, suppose that the constants ξ in (19) and τ in Assumption 6 are such that $\xi \geq 1$ and $\tau = 1/2$. Then, we have $f_\delta(\bar{\theta}_k) - f_* \leq \eta\sqrt{\delta} + O_p(k^{-1/2})$.*

Consequently, if we choose η small enough and use L to denote the computational effort needed to solve the line search $\sup_\gamma F(\gamma, \beta, \lambda; X)$ for any $(\beta, \lambda) \in \mathbb{U}_\eta$, then the total computational effort needed to obtain estimates of f_* within ε−precision is $O_p(L\varepsilon^{-2})$. A brief description of the complexity L introduced by the line search can be found in Online Appendix B.

4. Numerical Experiments

In this section, we provide some illustrative examples in the contexts of supervised learning and portfolio optimization. All the numerical examples were carried out on a laptop computer with a 2.2 GHz Intel Core i7 CPU and 16 GB memory. We keep in mind that our goal in this section is to demonstrate empirically the structural properties that we derived and their implications for algorithmic performance. We are not concerned with a specific choice of δ, which is typically done via cross-validation in a typical data-driven setting.

4.1. Illustrative Examples from Supervised Learning

The out-of-sample performance advantages of utilizing optimal transport costs with Mahalanobis distances have been demonstrated comprehensively with real data classification examples in Blanchet et al. [7]. Therefore, in the interest of space and to avoid repetition, we restrict the focus in this section to reporting the results of stylized numerical experiments that accomplish the following enumerated goals: (1) compare the iteration complexity of

the iterative scheme proposed in Section 3.2 for the DRO formulation (2) with that of the benchmark stochastic gradient descent for its nonrobust counterpart (1), (2) provide a visualization of the worst case distribution, and (3) study the iteration complexity when the twice differentiability assumption (made in order to prove Theorem 4) is relaxed.

4.1.1. Modifications of Notations for Supervised Learning. As supervised learning problems typically involve a response variable in addition to the predictor variables X, we first discuss how the DRO formulation in (2) can be utilized in the presence of the additional response variable. Let us use Y to denote the response variable in the rest of this section. We begin by treating the response Y as a random parameter of the loss function $\ell(\cdot)$, so the assumptions applied to $\ell(\cdot)$ should be replaced by that of $\ell(\cdot;Y)$ when considering problems with response variable Y. In addition, the reference measure $P_0 \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R})$ is modified to characterize the joint distribution of (X, Y). Further, as we assume the ambiguity only appears on the predictors X, we defined the optimal transport between $P \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R})$ and $P_0 \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R})$ can be modified as

$$D_c(P_0, P) = \inf\left\{ E_\pi\left[c(X, X')\right] : \begin{array}{l} \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}), \pi(Y = Y') = 1, \\ \pi_{(X,Y)} = P_0, \pi_{(X',Y')} = P. \end{array} \right\},$$

where π is the joint distribution of (X, Y, X', Y').

Using the modified model, if $\ell(\cdot;y)$ satisfies the assumptions of $\ell(\cdot)$ for P_0-almost every y, then all the results and algorithms developed in the previous sections are still valid. The proof of the generalized result is essentially same as before as we just need to replace $\ell(\cdot)$ by $\ell(\cdot;Y)$ in the proof as well.

4.1.2. Logistic Regression. We consider the case of binary classification, in which the data are given by $\{(X_i, Y_i)\}_{i=1}^n$ with predictor $X_i \in \mathbb{R}^d$ and label $Y_i \in \{-1, 1\}$. In this case, the logistic loss function is

$$\ell(u; y) = \log(1 + \exp(-yu)).$$

We are interested in solving the distributionally robust logistic regression problem

$$\inf_{\beta \in B} \sup_{P: D_c(P_n, P) \le \delta} E_P\left[\ell(\beta^T X; Y)\right]$$

where $P_0 = P_n(dx, dy) := \frac{1}{n}\sum_{i=1}^n \delta_{\{(X_i, Y_i)\}}(dx, dy)$ is the empirical measure of data.

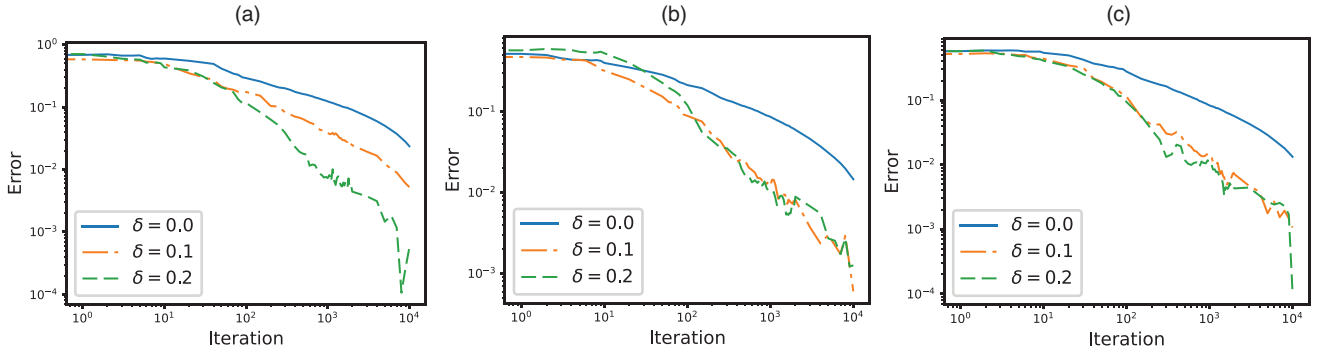
In Online Appendix D, we demonstrate that the assumptions in Section 2.2 are naturally satisfied by the logistic loss $\ell(\cdot;y)$, and therein, we also include computation of related constants. Consequently, all of the algorithms and theoretical results developed in this paper are applicable to the logistic regression example.

We design a numerical experiment to test the performance of our algorithm on a distributionally robust logistic regression. The data are generated from normal distribution with a different mean for each class and the same variance. The total number of data points ranges among $n \in \{64, 256, 1024\}$, and the dimension of data is $d = 32$.

We implement the iterative scheme provided in Section 3.2.1 to solve the ordinary logistic regression (with $\delta = 0$) and its distributionally robust counterpart ($\delta > 0$). In the numerical experiment, we choose $A(x) = \mathbb{I}_d$. To compare the rates of convergence of these two models, the same learning rate (or step size) on β is adapted. The parameter τ in Assumption 6 is chosen to be 0.55. We use the value of loss function at 10^5 iterations as the approximate optimal loss, and then, we plot the optimality gap (error) versus number of iterations for the DRO model and the ordinary logistic model in Figure 1.

Next, in the sequence of subplots in Figure 2, we attempt to visualize how the worst-case distributions $\{X_\delta^* : \delta > 0\}$ change as the radius δ is increased. In the first subplot corresponding to $\delta = 0$, we have 64 independent samples of $X \in \mathbb{R}^2$ and the decision boundary obtained from the ordinary logistic regression. The different markers denote the data in different classes: on the lower left side the data are classified to be circles, and on the upper right side the data are classified to be triangles. Naturally, when $\delta = 0$, most of the data points are correctly classified. Then, fixing the decision boundary to be the same as that obtained from the ordinary logistic regression, we increase the transportation budget δ and display the respective worst-case distribution computed with β fixed to that obtained from the ordinary logistic regression estimator. The worst-case distributions X_δ^* for different δ are visualized in the subsequent plots. We can observe that more and more points are misclassified when δ is increasing, and in the last plot, the misclassification rate is larger than 50%. In addition, the trajectory of X_δ^* forms a straight line moving toward the wrong side of the decision boundary, which is aligned with our observations pertaining to comparative statics in Theorem 6 (see Section 2.4).

Figure 1. (Color online) Convergence of loss function for logistic regression. (a) $n = 64$. (b) $n = 256$. (c) $n = 1{,}024$.



4.1.3. Linear Regression. Now, we turn to consider the example of linear regression with a squared loss function. In this, data are given by $\{(X_i, Y_i)\}_{i=1}^n$ with predictor $X_i \in \mathbb{R}^d$ and label $Y_i \in \mathbb{R}$. We consider the squared loss function $\ell(u; y) = (y - u)^2$ in this example, and the reference measure is defined as the empirical measure $P_0 = P_n(dx, dy) := \frac{1}{n}\sum_{i=1}^n \delta_{\{(X_i,Y_i)\}}(dx, dy)$. Then, the distributionally robust linear regression problem is defined as

$$\inf_{\beta \in B} \sup_{P: D_c(P_n, P) \le \delta} E_P\big[\ell(\beta^T X; Y)\big].$$

Following a similar argument as in the example of logistic regression, it is not hard to verify that the squared loss function satisfies all the assumptions regarding the loss function. We refer the interested readers to Online Appendix D for verification of assumptions and computation of related constants.

Actually, in this example, the dual objective function can be computed in closed form. The distributionally robust linear regression problem is equivalent to

$$\inf_{\beta \in B} \inf_{\lambda \ge 0} \left\{ \lambda \sqrt{\delta} + \frac{1}{n} \sum_{i=1}^n \frac{\lambda (\beta^T X_i - Y_i)^2}{\lambda - \sqrt{\delta} \beta^T A(X_i)^{-1} \beta} \right\}.$$

Now, we explain the setting of our numerical experiment in this example. The dimension of data is $d = 16$, and we randomly generate three different training data sets of size $n \in \{64, 256, 1024\}$. The matrix that appears in the cost function is chosen as $A(x) = \mathbb{I}_d$. We apply the iterative scheme in Section 3.2.1 to solve the ordinary linear regression model (with $\delta = 0$) and its distributionally robust counterpart ($\delta > 0$). Again, we adapt the same learning rate for both model and chosen parameter $\tau = 0.55$ in Assumption 6. The plot of optimality gaps (error) versus iterations for the DRO model and the ordinary linear regression model is given in Figure 3.

4.1.4. Support Vector Machines. We consider the case of binary classification, in which the data are given by $\{(X_i, Y_i)\}_{i=1}^n$, the same as the data in the example of logistic regression. The hinge loss function is $\ell(u; y) = \max(0, 1 - yu)$. We are interested in solving the distributionally robust hinge loss minimization problem

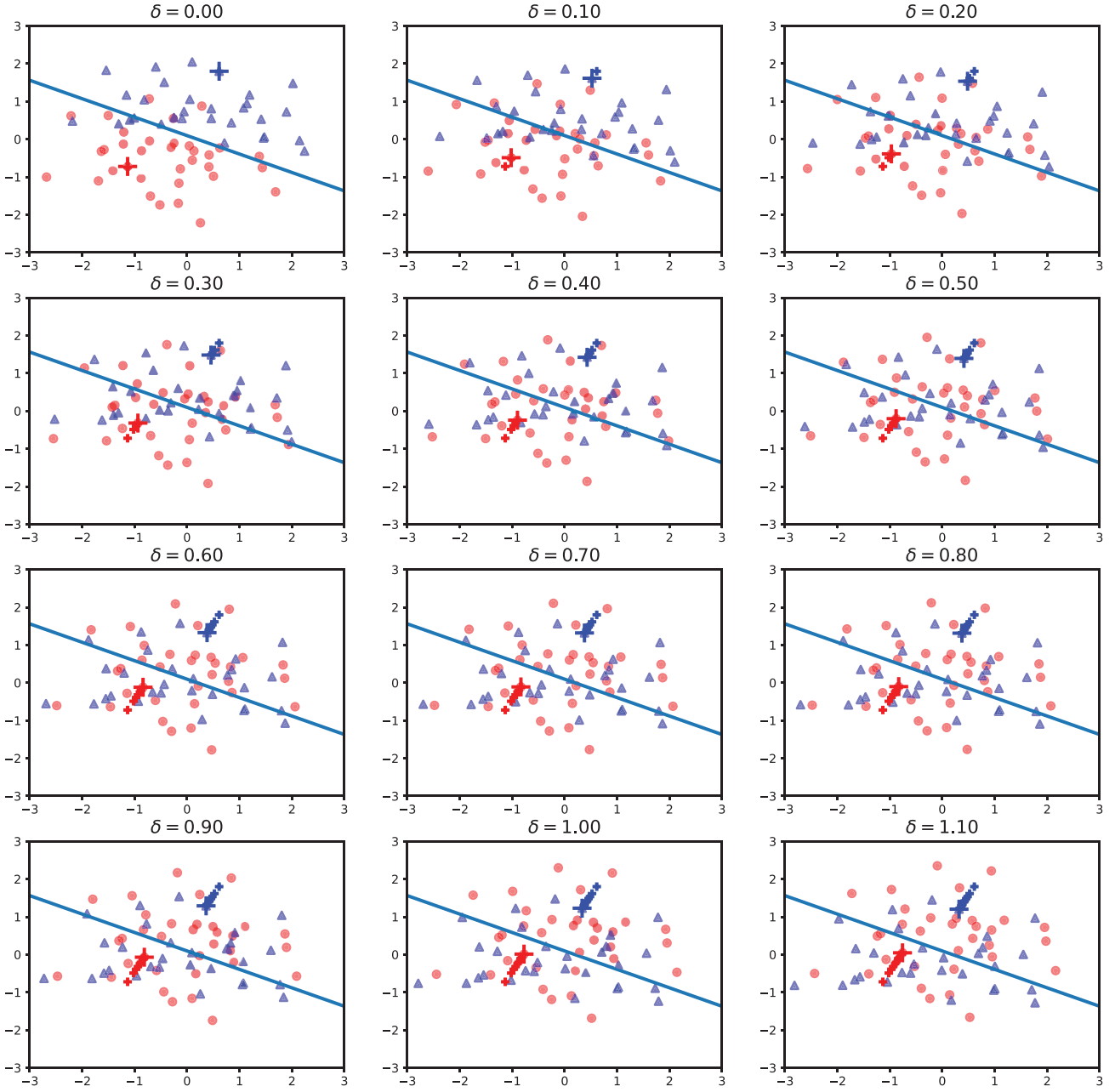
$$\inf_{\beta \in B} \sup_{P: D_c(P_n, P) \le \delta} E_P\big[\ell(\beta^T X; Y)\big],$$

where $P_0 = P_n(dx, dy) := \frac{1}{n}\sum_{i=1}^n \delta_{\{(X_i,Y_i)\}}(dx, dy)$ is the empirical measure of data.

The algorithm to solve the DRO with the piecewise continuously differentiable function is discussed in Section 3.4. We present the procedure of verification of related assumptions and computation of constants in Online Appendix D.

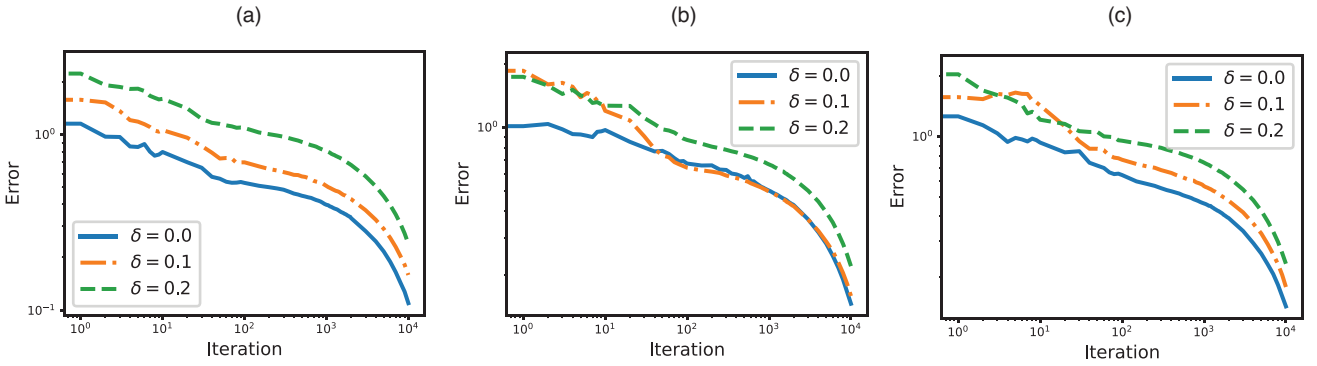
In the numerical experiment, we use the same data as in the example of logistic regression. Again, we set the learning rate to be same for DRO and non-DRO algorithms. Figure 4 shows the path of optimality gaps of loss functions during iterations. We use the value of loss function at 10^5 iterations as the approximate optimal loss given the training samples and plot the optimality gap (error) versus the number of iterations in Figure 4.

Figure 2. (Color online) Decision boundary and worst-case distribution. To facilitate tracking the change of X_δ^* when δ is increasing, we select one point from each class and use a big + to mark its position. We also employ a small + to mark its previous position when δ is smaller so that the trajectory of the point is visible. We can observe, as predicted by our theoretical results, that X_δ^* moves parametrically in a linear direction as δ changes. Moreover, the speed of displacement is decreasing, which is consistent with the $\sqrt{\delta}$ scaling size discussed in Theorem 7. It is worth noting the dynamics of the worse-case distribution, which transports the different classes in opposite directions in order to maximize the loss for misclassification.



4.1.5. Comparison Against Conic Programming Reformulation. Here, we provide a comparative numerical example against a direct convex optimization approach (Hanasusanto and Kuhn [15, proposition 4]. For data-driven DRO with a piecewise linear convex loss function of the form $\ell(u;y) = \max_{k=1,\dots,K}\{a_k \cdot (u-y) + b_k\}$, and a matrix appears in the cost function chosen as $A(x) = \mathbb{I}_d$, the second order cone program (SOCP) reformulation in Hanasusanto and Kuhn [15, proposition 4] obtained by letting $P_0 = P_n(dx,dy) := \frac{1}{n}\sum_{i=1}^n \delta_{\{(X_i,Y_i)\}}(dx,dy)$ is given in (20).

Figure 3. (Color online) Convergence of loss function for linear regression. (a) $n = 64$. (b) $n = 256$. (c) $n = 1{,}024$.



$$
\begin{aligned}
\inf \quad & \lambda\delta + \frac{1}{n}\sum_{i=1}^{n} s_i \\
\text{s.t.} \quad & \beta \in B,\ \lambda \in \mathbb{R}_+,\ s \in \mathbb{R}_+^n, \\
& s_i + a_k Y_i - b_k - a_k \beta^T X_i \geq 0 \qquad \forall k = 1,\ldots,K, \quad \forall i = 1,\ldots,n, \\
& \left\| \begin{matrix} 2c_k\beta \\ s_i + a_k Y_i - b_k - a_k\beta^T X_i - \lambda \end{matrix} \right\|_2 \leq s_i + a_k Y_i - b_k - a_k\beta^T X_i + \lambda \qquad \begin{matrix} \forall k = 1,\ldots,K, \\ \forall i = 1,\ldots,n. \end{matrix}
\end{aligned} \tag{20}
$$

If the loss function is not piecewise linear, such as square or logistic loss, one may solve the SOCP reformulation corresponding to a piecewise linear approximation. We invoke the linear regression model in Section 4.1.3 as an example for comparing the numerical performances of the direct convex optimization approach and the proposed SGD approach. We approximate the square loss function $\ell(u;y) = (u-y)^2$ with piecewise linear functions comprising $K = 9$ and $K = 19$ linear functions in separate instances. The linear functions are chosen to be the tangent line of the loss function $\ell(u;y) = (u-y)^2$ with distinct integer supporting points u satisfying $|u| \leq (K-1)/2$. We reformulated the resulting DRO with approximated loss as SOCP (20), which thereafter is solved using MOSEK ApS [21]. For the SGD approach, we terminate the algorithm if its optimality gap is smaller than the optimality gap of the SOCP solution (the SOCP solution is suboptimal because of the linear approximation error). The data-generating process of $\{(X_i, Y_i)\}_{i=1}^n$ is same as Section 4.1.3 with varying sample size n to test the scalability of the algorithms.

We compare the required time to solve SGD and SOCP in Figure 5. One can quickly remark that the SGD approach outperforms the SOCP approach for medium and large sample sizes. Though SOCP is a more efficient method for minimal sample sizes, its computational complexity rapidly deteriorates when n is increasing because of the $n \times K$ of cone constraints involved in the problem. In contrast, the computational time required by SGD is independent of the sample size.

Figure 4. (Color online) Convergence of loss function for support vector machines. (a) $n = 64$. (b) $n = 256$. (c) $n = 1{,}024$.

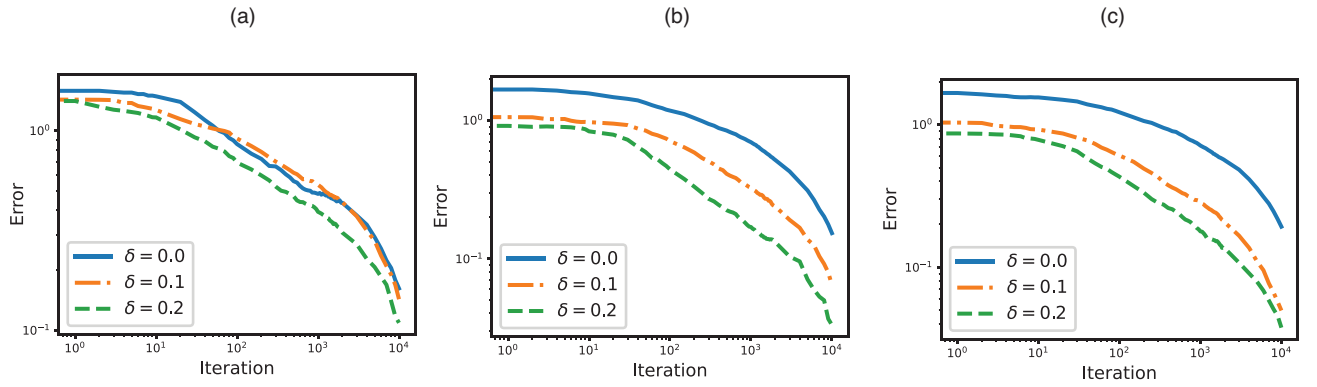
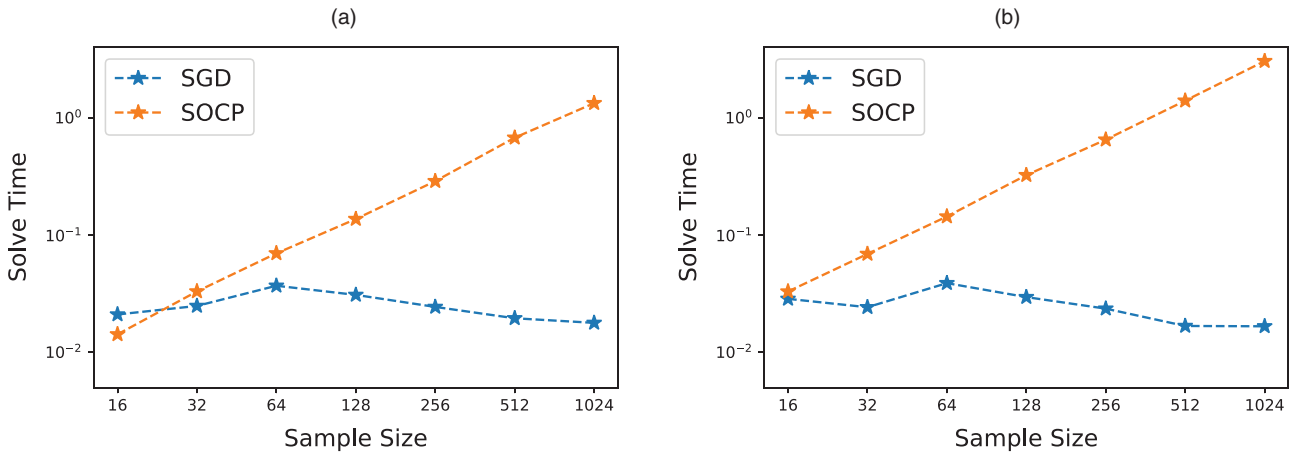


Figure 5. (Color online) Comparison of computational efforts for SGD and SOCP approaches for sample sizes $n \in \{16, 32, 64, 128, 256, 512, 1024\}$. The number of linear functions are (a) $K = 9$ and (b) $K = 19$.



4.2. Portfolio Optimization

In this section, we demonstrate an example application of the proposed DRO framework in the context of mean-variance portfolio optimization. Suppose that X is an $\mathbb{R}^d$−valued random vector representing the relative monthly returns of d securities. Let us use P^* to denote the probability distribution of X. The classical Markowitz mean-variance model suggests that the portfolio choices lying on the efficient frontier can be determined by solving an optimization problem of the form

$$\min_{\beta:\beta^T\mathbf{1}=1} \mathrm{Var}_{P^*}[\beta^T X] - \zeta \cdot E_{P^*}[\beta^T X], \tag{21}$$

where β is a d−dimensional weight vector and $\zeta \in [0, \infty)$ is a suitable parameter choice determining the extent of risk aversion. By adding an additional variable $\mu \in \mathbb{R}$ representing the mean return of the portfolio, Formulation (21) can be rewritten as the following stochastic optimization problem with affine decision rules:

$$\inf_{\mu} \inf_{\beta:\beta^T\mathbf{1}=1} E_{P^*}[(\beta^T X - \mu)^2 - \zeta \cdot \beta^T X]. \tag{22}$$

In practice, the probability distribution P^* is not known, and it is common to work with historical returns data to arrive at a suitable portfolio choice. Suppose that we use $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{\{X_i\}}$ to denote the empirical distribution corresponding to n historical return samples $\{X_1, \ldots, X_n\}$. Because of the discrepancy between the ground-truth measure P^* and the reference measure $P_0 = P_n$, we consider the following distributionally robust variant of (22):

$$\inf_{\mu} \inf_{\beta:\beta^T\mathbf{1}=1} \sup_{P:D_c(P_0,P)\le\delta} E_{P_0}[(\beta^T X - \mu)^2 - \zeta \cdot \beta^T X]. \tag{23}$$

As with most data-driven DRO formulations, the insertion of the inner supremum allows quantifying the impact of the model mismatch between the empirical distribution and plausible model variations that are a result of future market interactions. Additional information about such future variations can typically be inferred from current market data in the form of, for example, the implied volatility that can be elicited from the derivative prices. In such instances, a suitable choice of state-dependent Mahalanobis cost function $c(\cdot)$ in the proposed framework allows us to include this additional market information in the ambiguity set $\{P : D_c(P_0, P) \le \delta\}$, which corresponds to the set of plausible model variations. To demonstrate this idea in the portfolio example, suppose that we observe the implied volatility time series $\{V_i : i = 1, \ldots, n\}$ in addition to the returns data $\{X_i : i = 1, \ldots, n\}$; here, V_i is a positive scalar that represents the implied volatilities corresponding to the ith observation X_i. Let $\bar{V} = n^{-1}\sum_{i=1}^n V_i$ denote the average implied volatility. Corresponding to every point X_i in the support of P_n, we take the state-dependent Mahalanobis cost to be $c(X_i, x) = (X_i - x)^T A_i (X_i - x)$, where

$$A_i = \frac{\bar{V}}{V_i}\mathbb{I}_d, \quad i = 1, \ldots, n. \tag{24}$$

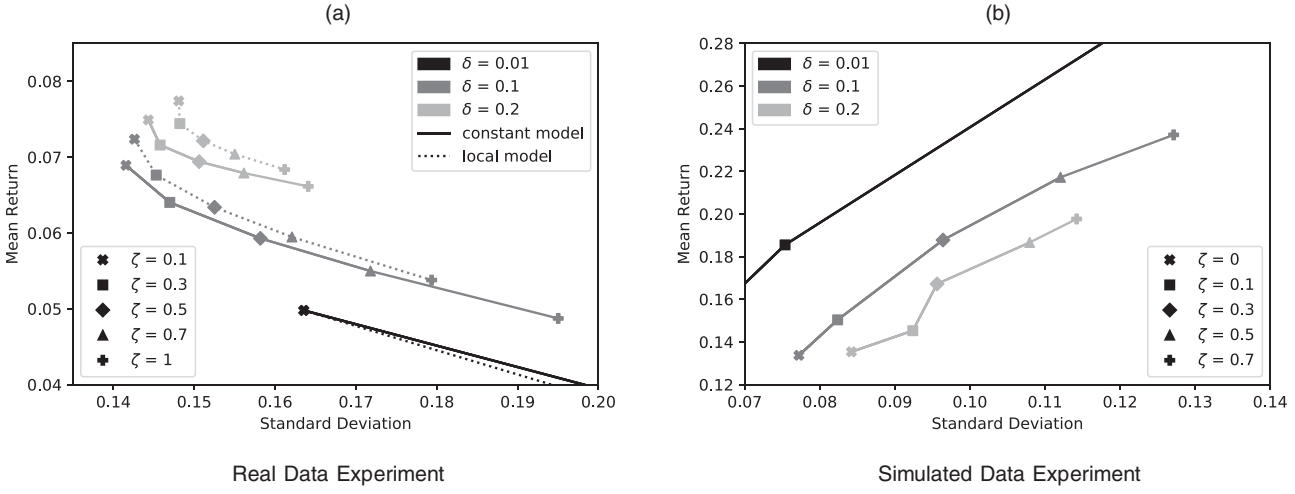
The rationale behind this choice is the hypothesis that a large implied volatility is suggestive of the anticipation of larger price uncertainty in future returns by the collective market. As a result, the inverse proportionality relationship $A_i \propto V_i^{-1}\mathbb{I}_d$ in (24) is such that it is cheaper to perturb returns (or transport mass) for observations with higher implied volatility. The normalization by $\bar{V}$ is introduced to allow comparisons with the choice of standard Euclidean squared norm (corresponding to the choice $A(x) := \mathbb{I}_d$) as the transportation cost function.

To test the effectiveness of the DRO formulation (23) with real data, we randomly pick 20 stocks from the constituents of the S&P 500 as the stock pool. The weights of the portfolio are adjusted on a monthly basis during the test period constituting the years 2000–2017. For every month in this test period, the portfolio weights are obtained by training Formulation (23) with the respective stock pool data from the previous 10 years. For example, at the beginning of January 2000, the training data $\{X_1, \ldots, X_n\}$ for Model (23) is the monthly historical returns of the selected 20 stocks observed during the period January 1990–January 2000 (thus, $n = 119$ and $d = 20$). The Chicago Board Options Exchange's volatility index, which is a popular gauge of the stock market's forward looking volatility implied by S&P 500 index options, is used to inform the market implied volatility. The parameter δ is treated as a hyperparameter, and the out-of-sample efficient frontier is generated by considering different values of the parameter ζ. In Figure 6(a), we report the mean and the standard deviation of the portfolio returns (during the test period 2000–2017) obtained from 100 random stock pool choices.

The data used for computing an optimal portfolio is different from the data used for evaluating the portfolio, which is the reason we address the efficient frontiers in Figure 6(a) as "out-of-sample." These out-of-sample efficient frontiers reveal that the DRO formulation (23) with state-dependent Mahalanobis cost choice (as in (24)) performs uniformly better than that obtained with the Euclidean distance choice (corresponding to constant $A(x) = \mathbb{I}_d$, addressed as the constant model in Figure 6(a)). We also observe that a larger value of distributional uncertainty δ results in a larger mean annualized return. Unlike the case of an efficient frontier generated and tested with samples from the same probability distribution, the negative slopes in the out-of-sample efficient frontiers in Figure 6(a) suggest that the out-of-sample effects (such as nonstationarity in data) are significant.

As a sanity check to verify our implementation, we also report the results of the same experiment with simulated data constituting i.i.d. training and test samples (see Figure 6(b)) for the choice $A(x) = \mathbb{I}_d$. In this simulation experiment, the DRO model is observed to produce less efficient portfolios relative to nonrobust formulations, which is not surprising given that the experiment has been designed with simulated data and there is little model error. The efficient frontiers of the DRO model, as expected for relatively small values of δ, have positive slopes in out-of-sample simulated frontiers, and this is consistent with the observations of the classical Markowitz theory. These experiment results can be viewed as underscoring the need for DRO model formulations such as the one we study in this paper. In addition to historical returns data, these model formulations incorporate the

Figure 6. Out-of-sample efficient frontier. The mean and the standard deviation are annualized. We use solid lines to represent models with a constant optimal transport cost function and use dashed lines to represent the models with a state-dependent Mahalanobis optimal transport cost function. The different choices of δ are denoted by different grayscales. The values of ζ are represented by different shapes of the markers. (a) Real data experiment. (b) Simulated data experiment.



flexibility to use additional information, such as implied volatilities, to elicit collective market expectations about future uncertainty.

5. Proofs of Main Results

We provide proofs of all the main results in Sections 2 and 3 in this section. The proofs of auxiliary results, which are technical in nature, are provided in the subsequent online technical appendix Section A for ease of reading.

5.1. Proofs of the Results on Dual Reformulation and Convexity

In this section, we see the proofs of Theorems 1 and 2 and Lemma 2.

Proof of Theorem 1. Because $c(\cdot)$ is lower semicontinuous and $\ell(\cdot)$ is upper semicontinuous, it follows from the strong duality result in Blanchet and Murthy [4, theorem 1] that

$$\sup_{P:D_c(P_0,P)\le\delta} E_P[\ell(\beta^T X)] = \inf_{\lambda\ge 0} E_{P_0}\left[\sup_{\Delta\in\mathbb{R}^d}\left\{\ell(\beta^T(X+\Delta)) - \lambda\left(\Delta^T A(X)\Delta - \delta\right)\right\}\right]$$
$$= \inf_{\lambda\ge 0} E_{P_0}\left[\sup_{c\in\mathbb{R}}\left\{\ell(\beta^T X + c) - \lambda\left(\inf_{\Delta:\beta^T\Delta=c}\Delta^T A(X)\Delta - \delta\right)\right\}\right],$$

and that the infimum on the right-hand side is attained for every $\beta\in B$. Because

$$\inf\{\Delta^T A(X)\Delta : \beta^T\Delta = c\} = c^2/(\beta^T A(X)^{-1}\beta)$$

for $\beta\neq\mathbf{0}$, changing variables as in $c=\sqrt{\delta}\gamma\beta^T A(X)^{-1}\beta$ and from $\lambda\sqrt{\delta}$ to λ lets us conclude that

$$\sup_{\Delta\in\mathbb{R}^d}\left\{\ell(\beta^T(X+\Delta)) - \lambda\left(\Delta^T A(X)\Delta-\delta\right)\right\} = \sup_{\gamma\in\mathbb{R}} F(\gamma,\beta,\lambda;X) =: \ell_{rob}(\beta,\lambda;X), \tag{25}$$

thus resulting in $\sup_{P:D_c(P_0,P)\le\delta} E_P[\ell(\beta^T X)] = \inf_{\lambda\ge 0} E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$. This completes the proof of Theorem 1. □

The proof of Theorem 2 follows immediately as a consequence of Lemma 2 (stated in Section 3.1) and Lemma 3, whose proof is furnished in the technical Online Appendix A.

Lemma 3. *Suppose that Assumptions 1 and 2 hold. Consider any $\varepsilon>0$, $x\in\mathbb{R}^d$, and $\beta\in B$. If $\lambda\ge(\kappa+\varepsilon)\sqrt{\delta}\beta^T A(x)^{-1}\beta$, then there exist positive constants C_1, C_2 such that*

a. *Any $g\in\Gamma^*(\beta,\lambda;x)$ satisfies $\sqrt{\delta}|g|\beta^T A(x)^{-1}\beta\le 1+C_1\varepsilon^{-1}(1+|\beta^T x|)$;*
b. $\ell_{rob}(\beta,\lambda;x)\le\lambda\sqrt{\delta}+C_2(1+\varepsilon+\varepsilon^{-1})(1+|\beta^T x|)^2$.

We first see the proof of Lemma 2 before proceeding to the proof of Theorem 2.

Proof of Lemma 2. Take any $\theta_1:=(\beta_1,\lambda_1)$ and $\theta_2:=(\beta_2,\lambda_2)$ in $B\times\mathbb{R}_+$. Given $\alpha\in[0,1]$, it follows from (25) that $\ell_{rob}(\alpha\theta_1+(1-\alpha)\theta_2;x)$ equals

$$\sup_{\Delta\in\mathbb{R}^d}\{\ell((\alpha\beta_1+(1-\alpha)\beta_2)^T(x+\Delta)) - (\alpha\lambda_1+(1-\alpha)\lambda_2)(\Delta^T A(x)\Delta-\delta)\}$$
$$= (\alpha\lambda_1+(1-\alpha)\lambda_2)\delta + \sup_{\Delta\in\mathbb{R}^d}\left\{\ell\left(\alpha\beta_1^T(x+\Delta)+(1-\alpha)(\beta_2^T(x+\Delta))\right) - (\alpha\lambda_1+(1-\alpha)\lambda_2)\Delta^T A(x)\Delta\right\}. \tag{26}$$

Because $\ell(\cdot)$ is convex, we have $\ell(\alpha u_1+(1-\alpha)u_2)\le\alpha\ell(u_1)+(1-\alpha)\ell(u_2)$ for $u_1,u_2\in\mathbb{R}$. Combining this with the fact that $\sup_\Delta(\alpha f_1(\Delta)+(1-\alpha)f_2(\Delta))\le\alpha\sup_\Delta f_1(\Delta)+(1-\alpha)\sup_\Delta f_2(\Delta)$ for any two functions f_1,f_2, we have that the term involving supremum in (26) is bounded from above by

$$\alpha\sup_{\Delta\in\mathbb{R}^d}\left\{\ell(\beta_1^T(x+\Delta)) - \lambda_1\Delta^T A(x)\Delta\right\} + (1-\alpha)\sup_{\Delta\in\mathbb{R}^d}\left\{\ell(\beta_2^T(x+\Delta)) - \lambda_2\Delta^T A(x)\Delta\right\}.$$

This observation, in conjunction with (26), establishes that $\ell_{rob}(\alpha\theta_1+(1-\alpha)\theta_2;x)\le\alpha\ell_{rob}(\theta_1;x)+(1-\alpha)\ell_{rob}(\theta_2;x)$, thus verifying the desired convexity of $\ell_{rob}(\cdot;x)$. □

Proof of Theorem 2. Because $f_\delta(\beta,\lambda) := E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$, the convexity of $f_\delta(\cdot)$ follows as a consequence of Lemma 2 and the linearity of expectations. The fact that $f_\delta(\cdot)$ is proper follows from the observation that $\ell_{rob}(\beta,\lambda;X)$ is almost surely finite for all λ sufficiently large (see Lemma 3(b)) and the assumption that $E_{P_0}\|X\|^2 < \infty$ (see Assumption 2). □

5.2. Bounds for Dual Optimizer $\lambda_*(\beta)$ and a Proof of Proposition 1

It follows from Theorem 1 that $\arg\min_{\lambda\geq 0} f_\delta(\beta,\lambda)$ is nonempty for any $\beta \in B$. Lemmas 4–6, whose proofs are provided in Online Appendix A, are useful toward establishing bounds for any $\lambda_*(\beta)$ in $\arg\min_{\lambda\geq 0} f_\delta(\beta,\lambda)$ (see Lemma 7). In turn, these bounds are useful toward identifying the region $\mathbb{V}$ in the main results Proposition 1 and Theorem 3.

Lemma 4. *Suppose that Assumptions 1 and 2 are satisfied and $\beta \in B$. Then, for any $\lambda_*(\beta) \in \arg\min_{\lambda\geq 0} f_\delta(\beta,\lambda)$, we have $\Gamma^*(\beta,\lambda_*(\beta);x) \neq \varnothing$, for P_0–almost every $x \in \mathbb{R}^d$. Moreover,*

$$\frac{\partial_+ f_\delta}{\partial\lambda}(\beta,\lambda_*(\beta)) = \sqrt{\delta}\left(1 - E_{P_0}\left[\beta^T A(X)^{-1}\beta \min_{\gamma\in\Gamma^*(\beta,\lambda_*(\beta);X)} \gamma^2\right]\right).$$

Lemma 5. *Suppose that Assumptions 1 and 2 are satisfied and $\Gamma^*(\beta,\lambda,x)$ is not empty for a given $\beta \in B$, $x \in \mathbb{R}^d$, and $\lambda \geq 0$. Then, for any $\gamma \in \Gamma^*(\beta,\lambda;x)$, we have $\gamma = \ell'(\beta^T x + \sqrt{\delta}\gamma\beta^T A(x)^{-1}\beta)/(2\lambda)$, and consequently,*

$$|\gamma| \geq \frac{|\ell'(\beta^T x)|}{2\lambda}. \tag{27}$$

Lemma 6. *Suppose that Assumptions 2–4 are satisfied. Then, there exist positive constants $\underline{L}, \bar{L}$ such that $\underline{L} \leq E_{P_0}[\ell'(\beta^T X)^2] \leq \bar{L}$ for every $\beta \in B$.*

Lemma 7. *Suppose that Assumptions 1–3 are satisfied. Then, any minimizer $\lambda_*(\beta) \in \arg\min_{\lambda\geq 0} f_\delta(\beta,\lambda)$ satisfies $\lambda_{\min}(\beta) \leq \lambda_*(\beta) \leq \lambda_{\max}(\beta)$, where*

$$\lambda_{\min}(\beta) := \frac{1}{2}\rho_{\max}^{-1/2}\|\beta\|\sqrt{E_{P_0}\left[\ell'(\beta^T X)^2\right]} \text{ and}$$

$$\lambda_{\max}(\beta) := \rho_{\max}^{-1/2}\|\beta\|\sqrt{E_{P_0}\left[\ell'(\beta^T X)^2\right]} + \frac{1}{2}\sqrt{\delta}M\rho_{\max}^{-1}\|\beta\|^2.$$

Proof of Lemma 7. Lower bound: Combining the observations in Lemmas 4 and 5 and the first order optimality condition that $\partial_+ f_\delta(\beta,\lambda_*(\beta))/\partial\lambda \geq 0$, we obtain

$$0 \leq \frac{\partial_+ f_\delta}{\partial\lambda}(\beta,\lambda_*(\beta)) \leq \sqrt{\delta}\left(1 - E_{P_0}\left[\beta^T A(X)^{-1}\beta \frac{\ell'(\beta^T X)^2}{4\lambda_*(\beta)^2}\right]\right).$$

Because of Assumption 1(b), the preceding inequality results in

$$\lambda_*(\beta) \geq \frac{1}{2}E_{P_0}^{1/2}\left[\ell'(\beta^T X)^2\beta^T A(X)^{-1}\beta\right] \geq \frac{1}{2}\rho_{\max}^{-1/2}\|\beta\|\sqrt{E_{P_0}\left[\ell'(\beta^T X)^2\right]} =: \lambda_{\min}(\beta).$$

Upper bound: As $\ell''(\cdot) \leq M$ because of Assumption 3, we have that $\ell_{rob}(\beta,\lambda;X) - \ell(\beta^T X)$ is bounded from above by

$$\begin{aligned}
&\sup_{\gamma\in\mathbb{R}}\left\{\ell\left(\beta^T X + \gamma\sqrt{\delta}\beta^T A(X)^{-1}\beta\right) - \ell\left(\beta^T X\right) - \lambda\sqrt{\delta}\beta^T A(X)^{-1}\beta\gamma^2\right\} \\
&\quad\leq \sup_{\gamma\in\mathbb{R}}\left\{\ell'\left(\beta^T X\right)\sqrt{\delta}\beta^T A(X)^{-1}\beta\gamma + \frac{1}{2}M\left(\gamma\sqrt{\delta}\beta^T A(X)^{-1}\beta\right)^2 - \lambda\sqrt{\delta}\beta^T A(X)^{-1}\beta\gamma^2\right\} \\
&\quad= \frac{\sqrt{\delta}\beta^T A(X)^{-1}\beta[\ell'(\beta^T X)]^2}{(4\lambda - 2M\sqrt{\delta}\beta^T A(X)^{-1}\beta)^+}.
\end{aligned}$$

Next, because $\lambda_*(\beta)\sqrt{\delta} + E_{P_0}[\ell(\beta^T X)] \le f_\delta(\beta, \lambda_*(\beta)) = \inf_{\lambda\ge 0} E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$, we use the preceding result and the bounds in Assumption 1(b) to write

$$\begin{aligned}\lambda_*(\beta) &\le \inf_{\lambda\ge 0}\left\{\lambda + \delta^{-1/2}E_{P_0}[\ell_{rob}(\beta,\lambda;X) - \ell(\beta^T X)]\right\}\\ &\le \inf_{\lambda > \frac{1}{2}\sqrt{\delta M\rho_{\min}^{-1}}\|\beta\|_2^2}\left\{\lambda + E_{P_0}\left[\frac{\beta^T A(X)^{-1}\beta[\ell'(\beta^T X)]^2}{4\lambda - 2M\sqrt{\delta}\beta^T A(X)^{-1}\beta}\right]\right\}\\ &\le \inf_{\lambda > \frac{1}{2}\sqrt{\delta M\rho_{\min}^{-1}}\|\beta\|_2^2}\left\{\lambda + \frac{\rho_{\min}^{-1}\|\beta\|^2}{4\lambda - 2M\sqrt{\delta}\rho_{\min}^{-1}\|\beta\|^2}E_{P_0}\left[\ell'(\beta^T X)^2\right]\right\}.\end{aligned}$$

The expression in the right-hand side is a one-dimensional convex optimization problem that can be solved in closed form to obtain

$$\lambda_*(\beta) \le \frac{1}{2}\sqrt{\delta M\rho_{\min}^{-1}}\|\beta\|^2 + \rho_{\max}^{-1/2}\|\beta\|\sqrt{E_{P_0}\left[\ell'(\beta^T X)^2\right]} =: \lambda_{\max}(\beta).$$

This completes the proof of Lemma 7. □

Proof of Proposition 1. For a given $\beta \in B$, it follows from Lemma 7 that any optimal $\lambda_*(\beta)$ lies in the interval $[\lambda_{\min}(\beta), \lambda_{\max}(\beta)]$. Recalling the definitions of R_β from Assumption 4 and the characterization of $\bar{L}$ and $\underline{L}$ in Lemma 6, we have from Lemma 7 that $\lambda_{\min}(\beta) \ge K_1\|\beta\|$ and $\lambda_{\max}(\beta) \le K_2\|\beta\|$, where

$$K_1 := \frac{1}{2}\sqrt{\underline{L}\rho_{\min}^{-1}} \quad \text{and} \quad K_2 := \frac{1}{2}\sqrt{\delta M R_\beta \rho_{\min}^{-1}} + \sqrt{\rho_{\min}^{-1}\bar{L}}. \tag{28}$$

Thus, we obtain that $(\beta, \lambda_*(\beta)) \in \mathbb{V}$ for all $\beta \in B$. □

5.3. Verifying Smoothness and Strong Convexity of the Dual DRO Objective

In this section, we provide proofs of Theorems 3 and 4. We accomplish this primarily by identifying the Hessian matrix of the dual DRO objective $f_\delta(\beta,\lambda) = E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$.

Recall the definition of the functions $\ell_{rob}(\cdot)$ and $F(\cdot)$ in Theorem 1. Let S_X be the support of the distribution P_0. For a given (β,λ) and $x \in S_X$, we use the set $\Gamma^*(\beta,\lambda;x)$ to denote the respective set of maximizers $\arg\max_\gamma F(\gamma,\beta,\lambda;x)$ (see (9)). A characterization of the gradient of the function $\ell_{rob}(\beta,\lambda;x)$ is derived in Proposition 2 with the help of the envelope theorem. Likewise, if the loss $\ell(\cdot)$ is twice differentiable, the implicit function theorem allows us to characterize the Hessian of $\ell_{rob}(\beta,\lambda;x)$. To accomplish this, define

$$\mathcal{U} := \{(\beta,\lambda,x) \in B \times \mathbb{R}_+ \times S_X : \Gamma^*(\beta,\lambda;x) \ne \emptyset,\ \varphi(\gamma,\beta,\lambda;x) > 0 \text{ for some } \gamma \in \Gamma^*(\beta,\lambda;x)\},$$

where

$$\varphi(\gamma,\beta,\lambda;x) := 2\lambda - \sqrt{\delta}\beta^T A(x)^{-1}\beta\ell''\left(\beta^T x + \sqrt{\delta}\gamma\beta^T A(x)^{-1}\beta\right).$$

Further consider the set valued map $x \mapsto \mathcal{U}(x)$ to be the projection

$$\mathcal{U}(x) := \{(\beta,\lambda) : (\beta,\lambda,x) \in \mathcal{U}\}.$$

Then, as a consequence of the implicit function theorem, the function $\ell_{rob}(\beta,\lambda;x)$ is twice differentiable for every (β,λ) in the interior of $\mathcal{U}(x)$. Indeed, this follows from the observation that $\partial^2 F/\partial\gamma^2(\cdot) = -2\sqrt{\delta}\beta^T A(x)^{-1}\beta\varphi(\cdot)$ is negative when $(\beta,\lambda,x) \in \mathcal{U}$. Next, consider any measurable selection $g : \mathcal{U} \to \mathbb{R}$ such that

$$g(\beta,\lambda;x) \in \Gamma^*(\beta,\lambda;x) \quad \text{and} \quad \varphi(g(\beta,\lambda;x),\beta,\lambda,x) > 0, \tag{29}$$

for P_0−almost every x and almost every $(\beta,\lambda) \in \mathcal{U}(x)$. The existence of such a measurable selection follows from the Jankov–Von Neumann theorem (see, for example, Bertsekas and Shreve [3, proposition 7.50]). To proceed further, define

$$T_g(x) := x + \sqrt{\delta}g(\beta,\lambda;x)A(x)^{-1}\beta, \quad \bar{T}_g(x) := x + 2\sqrt{\delta}g(\beta,\lambda;x)A(x)^{-1}\beta, \quad \text{and} \quad \varphi_g(\beta,\lambda;x) := \varphi(g(\beta,\lambda;x),\beta,\lambda;x), \tag{30}$$

for any $(\beta,\lambda,x) \in \mathcal{U}$, where the dependence on (β,λ) is hidden in the notation of the transport maps $T_g(x), \bar{T}_g(x)$ and has to be understood implicitly. Likewise, once the choice of measurable selection $g(\cdot)$ is fixed, we often suppress the arguments $(\beta,\lambda;x)$ when writing functions such as $g(\beta,\lambda;x)$ and $\varphi_g(\beta,\lambda;x)$ in order to reduce clutter in the resulting expressions; for example, we simply write φ_g and g, respectively, for $\varphi_g(\beta,\lambda;x)$ and $g(\beta,\lambda;x)$.

Proposition 8. *Suppose that Assumptions 1–3 are satisfied, $\mathcal{U}$ is not empty, and $g:\mathcal{U}\to\mathbb{R}$ is a measurable selection satisfying (29). Then, for almost every $x\in S_X$, $(\beta,\lambda)\in \mathrm{int}(\mathcal{U}(x))$, we have*

$$\frac{\partial^2 \ell_{rob}}{\partial \beta^2}(\beta,\lambda;x)=2\sqrt{\delta}\lambda g^2 A(x)^{-1}+\frac{2\lambda \ell''\left(\beta^T T_g(x)\right)}{\varphi_g}\bar{T}_g(x)\bar{T}_g(x)^T,\quad \frac{\partial^2 \ell_{rob}}{\partial \lambda^2}(\beta,\lambda;x)=\frac{4\sqrt{\delta}g^2\beta^T A(x)^{-1}\beta}{\varphi_g},$$

$$\frac{\partial^2 \ell_{rob}}{\partial \lambda\partial \beta}(\beta,\lambda;x)=-2\sqrt{\delta}g^2\left(A(x)^{-1}\beta+\frac{\beta^T A(x)^{-1}\beta \ell''(\beta^T T_g(x))}{g\varphi_g}\bar{T}_g(x)\right),$$

where $T_g(\cdot),\bar{T}_g(\cdot),\varphi_g$ are defined as in (30). Moreover, we have

$$\nabla_\theta^2 \ell_{rob}(\theta;x)-\Lambda(\theta;x)B(x)\geq 0, \tag{31}$$

where

$$\Lambda(\beta,\lambda;x):=\frac{4\left(\beta^T T_g(x)\right)^2\ell''(\beta^T T_g(x))}{1+\bar{T}_g(x)^T A(x)\bar{T}_g(x)\ell''(\beta^T T_g(x))/(\sqrt{\delta}g^2\varphi)}\frac{1}{2\lambda\varphi_g+4\beta^T A(x)^{-1}\beta}, \tag{32}$$

and

$$B(x)=\begin{bmatrix} A(x)^{-1}+\dfrac{\ell''(\beta^T T_g(x))}{\sqrt{\delta}g^2\varphi}\bar{T}_g(x)\bar{T}_g(x)^T & \mathbf{0}\\ \mathbf{0}^T & 1\end{bmatrix}.$$

The proofs of Proposition 8 and Lemma 8 are provided in the technical Online Appendix A. For every $\beta\in B$, recall that we have defined $\lambda'_{thr}(\beta)$ to be the P_0-essential supremum of $\sqrt{\delta}M\beta^T A(x)^{-1}\beta/2$.

Lemma 8. *Suppose Assumptions 1–3 are satisfied. Then, the map $\gamma\longmapsto F(\gamma,\beta,\lambda;x)$ is strongly concave for every $\beta\in B,\lambda>\lambda'_{thr}(\beta)$ and P_0-almost every x. Consequently, $\Gamma^*(\beta,\lambda;x)$ is singleton for every $\beta\in B$, $\lambda>\lambda'_{thr}(\beta)$, and*

$$\{(\beta,\lambda):\beta\in B,\lambda>\lambda'_{thr}(\beta)\}\subseteq \mathcal{U}(x)$$

for P_0-almost every x.

The proof of Lemma 8 is available in the technical Online Appendix A.

Proposition 9 allows us to characterize the Hessian matrix of the dual DRO objective $f_\delta(\cdot)$. To state Proposition 9, define

$$\delta_0:=\rho_{\min}^2\underline{L}R_\beta^{-2}M^{-2}\rho_{\max}^{-1},\quad \text{and}\quad \varphi_{\min}:=\sqrt{\underline{L}}\rho_{\max}^{-1/2}-\sqrt{\delta}R_\beta M\rho_{\min}^{-1},$$

where the constants $\rho_{\min},\rho_{\max}$ are as in Assumption 1(b), $\underline{L},\bar{L}$ in Lemma 6, R_β in Assumption 4, and M in Assumption 3. Recall the definition of the constants K_1, K_2 in (28) and that of the previously defined sets

$$\mathbb{W}:=\{(\beta,\lambda)\in B\times\mathbb{R}_+:K_1\|\beta\|\leq\lambda\leq K_2R_\beta\}\quad\text{and}\quad \mathbb{V}:=\{(\beta,\lambda)\in B\times\mathbb{R}_+:K_1\|\beta\|\leq\lambda\leq K_2\|\beta\|\},$$

which contain the partial minimizers $\{(\beta,\lambda_*(\beta)):\beta\in B\}$ when Assumptions 1–4 are satisfied (see Proposition 1). The proof of Proposition 9 is provided in the technical Online Appendix A.

Proposition 9. *Suppose Assumptions 1–4 are satisfied, and $\delta<\delta_0$. Then,*

a. *$\mathbb{V}\subseteq\mathbb{W}\subset\{(\beta,\lambda):\beta\in B,\lambda>\lambda'_{thr}(\beta)\}\subseteq\mathcal{U}(x)$ for P_0-almost every x;*

b. *Any map $g:\mathcal{U}\to\mathbb{R}$ satisfying (29) is uniquely specified for almost every (β,λ,x) in the subset $\mathbb{W}\times S_X\subseteq\mathcal{U}$, and it satisfies the following relationships: for P_0-almost every x, we have $\varphi_g(\beta,\lambda;x)>\varphi_{\min}\|\beta\|$ if $(\beta,\lambda)\in\mathbb{W}$, and*

$$|g(\beta,\lambda;x)|\geq\frac{|\ell'(\beta^T x)|}{2K_2\|\beta\|}\ \text{if}\ (\beta,\lambda)\in\mathbb{V},\qquad |g(\beta,\lambda;x)|\leq\frac{|\ell'(\beta^T x)|}{\varphi_{\min}\|\beta\|}\ \text{if}\ (\beta,\lambda)\in\mathbb{W}. \tag{33}$$

c. *With $X\sim P_0$, the collection $\{g^2(\beta,\lambda;X),(T_g(X))^2,(\bar{T}_g(X))^2,\ell(\beta^T T_g(X)),\ell'(\beta^T T_g(X))^2:(\beta,\lambda)\in\mathbb{V}\}$ is L_2-bounded.*

d. *The Hessian matrix $\nabla_\theta^2 f_\delta(\theta)=E_{P_0}\left[\nabla_\theta^2\ell_{rob}(\theta;X)\right]$ for every $\theta\in\mathbb{V}$, where the Hessian $\nabla_\theta^2\ell_{rob}(\theta;x)$ can be taken to be specified in terms of the second order partial derivative expressions in Proposition 8.*

The proofs of Theorems 3 and 4 are reliant on the observations made in Proposition 9.

Proof of Theorem 3.

a. It follows from part (c) of Proposition 9 and the expressions of partial derivatives in Proposition 8 that the norms of the respective entries (Frobenius norm $\|\cdot\|_F$ in the case of a matrix or ℓ_2-norm in the case of a vector), $\|\partial^2\ell_{rob}/\partial\beta^2(\beta,\lambda;X)\|_F$, $\|\partial^2\ell_{rob}/\partial\beta\partial\lambda(\beta,\lambda;X)\|$, $\partial^2\ell_{rob}/\partial\lambda^2(\beta,\lambda;X)$, are all bounded in L_2-norm over the set $(\beta,\lambda)\in\mathbb{V}$. Consequently, we have from part (d) of Proposition 9 that $\partial^2 f_\delta/\partial\beta^2$, $\partial^2 f_\delta/\partial\beta\partial\lambda$, and $\partial^2 f_\delta/\partial\lambda^2$ are all bounded over $(\beta,\lambda)\in\mathbb{V}$. As a result, the Frobenius norm of the Hessian matrix $\nabla^2_\theta f_\delta(\theta)$ is bounded over $\theta=(\beta,\lambda)\in\mathbb{V}$, and hence, the function $f_\delta(\cdot)$ is smooth over the interior of $\mathbb{V}$.

b. To argue that $\partial^2 f_\delta/\partial\beta^2$ is positive definite, we proceed as follows: First observe that $A(x)^{-1}\geq\rho_{\max}^{-1}\mathbb{I}_d$ for P_0-almost every x (that is, $A(x)^{-1}\rho_{\max}^{-1}\mathbb{I}_d$ is positive semidefinite). Next, recall from (27) in Lemmas 5 and 6 that $|g(\beta,\lambda;x)|\geq|\ell'(\beta^Tx)|/(2\lambda)$ and $\underline{L}>0$. Then, it follows from part (d) of Proposition 9 and the expression of $\partial^2\ell_{rob}/\partial\beta^2$ from Proposition 8 that, for any $(\beta,\lambda)\in\mathbb{V}$,

$$\frac{\partial^2 f_\delta}{\partial\beta^2}(\beta,\lambda)=E_{P_0}\left[\frac{\partial^2\ell_{rob}}{\partial\beta^2}(\beta,\lambda;X)\right]\geq\sqrt{\delta}\frac{E_{P_0}[\ell'(\beta^TX)^2]}{2\lambda}\rho_{\max}^{-1}\mathbb{I}_d\geq\sqrt{\delta}\frac{\kappa_0}{\lambda}\mathbb{I}_d,$$

where $\kappa_0:=2^{-1}\underline{L}\rho_{\max}^{-1}>0$, thus proving Theorem 3. □

In order to proceed with the proof of Theorem 4, define

$$\delta_1:=\min\{\delta_0/4,c_1^2c_2^2p^2\rho_{\max}^2\rho_{\max}^{-1}\underline{L}\bar{L}^{-2}/256\}.$$

Proof of Theorem 4. Using the bounds of $|g(\cdot)|$ and $\varphi_g(\cdot)$ from Proposition 9(b) along with other immediate bounds, such as $\varphi_g\leq 2\lambda,\lambda\in[K_1\|\beta\|,K_2\|\beta\|]$, and $\beta^TA(x)^{-1}\beta\leq\rho_{\max}^{-1}\|\beta\|^2$, the expression for $\Lambda(\beta,\lambda;x)$ from (32) simplifies to

$$\begin{aligned}\Lambda(\beta,\lambda;x)&=\frac{4\sqrt{\delta}(g\beta^TT_g(x))^2}{2\lambda\sqrt{\delta}g^2/\ell''(\beta^TT_g(x))+\bar{T}_g(x)^TA(x)\bar{T}_g(x)}\cdot\frac{1}{2\lambda+4\beta^TA(x)^{-1}\beta/\varphi_g}\qquad(34)\\&\geq\frac{4\sqrt{\delta}(\beta^TT_g(x)\ell'(\beta^Tx)/(2K_2\|\beta\|))^2}{2K_2\sqrt{\delta}\ell'(\beta^Tx)^2/(\varphi_{\max}^2\|\beta\|\ell''(\beta^TT_g(x)))+\bar{T}_g(x)^TA(x)\bar{T}_g(x)}\cdot\frac{1}{2K_2\|\beta\|+4\rho_{\max}^{-1}\|\beta\|^2/(\varphi_{\min}\|\beta\|)}\\&\geq\sqrt{\delta}C_0\frac{\|\beta\|^{-2}\left(\beta^TT_g(x)\ell'(\beta^Tx)\right)^2}{2K_2\sqrt{\delta}\varphi_{\max}^{-2}\ell'(\beta^Tx)^2/\ell''(\beta^TT_g(x))+\bar{T}_g(x)^TA(x)\bar{T}_g(x)\|\beta\|},\qquad(35)\end{aligned}$$

where $C_0:=(2K_2+4\varphi_{\max}^{-1}\rho_{\max}^{-1})^{-1}$. Next, because $\beta^TT_g(x)=\beta^Tx+\sqrt{\delta}g\beta^TA(x)^{-1}\beta$, we obtain from the bounds in (33) that

$$\begin{aligned}|\beta^TT_g(x)\ell'(\beta^Tx)|&\geq|\beta^Tx\ell'(\beta^Tx)|-\sqrt{\delta}|g\ell'(\beta^Tx)|\beta^TA(x)^{-1}\beta\\&\geq|\beta^Tx\ell'(\beta^Tx)|-\sqrt{\delta}\frac{\ell'(\beta^Tx)^2}{\varphi_{\min}\|\beta\|}\|\beta\|^2\rho_{\max}^{-1}\geq\left(\frac{c_1c_2-4\sqrt{\delta}\bar{L}}{p\varphi_{\min}\rho_{\min}}\right)\|\beta\|,\qquad(36)\end{aligned}$$

whenever $X\in A_1\cap A_2$; here, the sets A_1 and A_2 are defined as follows:

$$A_1:=\{x:|\beta^Tx\ell'(\beta^Tx)|>c_1c_2\|\beta\|\}\quad\text{and}\quad A_2:=\{x:\ell'(\beta^Tx)^2\leq 4\bar{L}/p\},$$

where the constants c_1,c_2,p are given by Assumption 5. Because $E_{P_0}[\ell'(\beta^TX)^2]\leq\bar{L}$ for any $\beta\in B$, we have from Markov's inequality that $\inf_{\beta\in B}P_0(X\in A_2)\geq 1-p/4$. Consequently, it follows from Assumption 5 and the union bound that $\inf_{\beta\in B}P_0(X\in A_1\cap A_2)\geq 3p/4$.

Recall that $\delta_0:=\rho_{\min}^2\underline{L}R_\beta^{-2}M^{-2}\rho_{\max}^{-1}$. In addition, note that, when $\delta\leq\delta_0/4$, we have $\varphi_{\min}=\sqrt{\underline{L}}\rho_{\max}^{-1/2}-\sqrt{\delta}R_\beta M\rho_{\max}^{-1}\geq\frac{1}{2}\sqrt{\underline{L}}\rho_{\max}^{-1/2}$. Further, because $\delta<\delta_1\leq c_1^2c_2^2p^2\rho_{\min}^2\rho_{\max}^{-1}\underline{L}\bar{L}^{-2}/256$, we have

$$c_1c_2-4\sqrt{\delta}\bar{L}p^{-1}\varphi_{\max}^{-1}\rho_{\max}^{-1}\geq c_1c_2/2.\qquad(37)$$

Next, if we choose $C_1>0$ large enough such that the set $A_3:=\{x:\|x\|\leq C_1\}$ satisfies $P_0(X\in A_3)\geq 1-p/4$, then we have $\inf_{\beta\in\Xi}P_0(X\in A_1\cap A_2\cap A_3)\geq p/2$. The denominator in (35) is bounded from above as follows whenever $x\in A_1\cap A_2\cap A_3$ and $\lambda\in[K_1\|\beta\|,K_2\|\beta\|]$: recalling that $T_g(x):=x+\sqrt{\delta}g(\beta,\lambda;x)A(x)^{-1}\beta$ and

$\bar{T}_g(x) := x + 2\sqrt{\delta} g(\beta,\lambda;x)A(x)^{-1}\beta$, it follows from the bounds of $|g|$ in (33) that

$$\|\bar{T}_g(x)\| \le \|x\| + 2\sqrt{\delta}|g|\rho_{\max}^{-1}\|\beta\| \le C_1 + 4\sqrt{\delta\bar{L}p^{-1}}\left(\frac{1}{2}\sqrt{\underline{L}}\rho_{\max}^{-1/2}\right)^{-1}\rho_{\max}^{-1} =: C_2,$$

and similarly, $\|T_g(x)\| \le C_2$ for $x \in A_2 \cap A_3$. Because $\|\beta^T T_g(x)\| \le R_\beta C_2 < \infty$ when $x \in A_2 \cap A_3$, if we let $C_3 := \inf_{|u|\le R_\beta C_2} \ell''(u) > 0$, we obtain that the denominator in (35) is bounded from above by $C_4 := 8K_2\delta^{1/2}\bar{L}p^{-1}C_3^{-1}(\frac{1}{2}\sqrt{\underline{L}}\rho_{\max}^{-1/2})^{-2} + \rho_{\max}C_2R_\beta$ whenever $x \in A_2 \cap A_3$. Combining this observation with that of (35)–(37), we obtain that $\Lambda(x) \ge \sqrt{\delta}C\mathbf{1}_{\{x\in A_1\cap A_2\cap A_3\}}$ for $C := (1/2)C_0c_1c_2C_4^{-1}$.

Finally, because $P_0(A_1 \cap A_2 \cap A_3) \ge p/2$, we have $E_{P_0}[\Lambda(\beta,\lambda;X)B(X)] \ge \sqrt{\delta}\kappa_1\mathbb{I}_{d+1}$, where $\kappa_1 := pC\rho_{\max}^{-1}/2$. As a consequence, we have that $\nabla_\theta^2 f_\delta(\theta) \ge \sqrt{\delta}\kappa_1\mathbb{I}_{d+1}$ in Theorem 4. □

Remark 4. Suppose that $c_1c_2 = 0$ is the only nonnegative number for which the probability requirement in Assumption 5 is satisfied. In this case, we have from the upper bound for g in Proposition 9(b) that $g\beta^T X = 0$, P_0-almost surely. As a result, the numerator of $\Lambda(x)$ in the right-hand side of (34) is bounded from above by $4\sqrt{\delta}(0 + \sqrt{\delta}g^2\beta^T A(x)^{-1}\beta)^2 \le 4\delta^{3/2}\ell'(\beta^T x)^2\varphi_{\max}^{-2}\rho_{\max}^{-2}$, P_0-almost surely. Because the denominator of $\Lambda(x)$ is bounded away from zero by a constant not dependent on δ, it follows that $E_{P_0}[\Lambda(X)] = \kappa_3\delta^{3/2}$, for some nonnegative constant κ_3. Because $\delta^{3/2} = o(\sqrt{\delta})$ as $\delta \to 0$, it is not possible to derive a positive constant κ_1 that is not dependent on δ as in the statement of Theorem 4.

5.4. Proofs of the Results Pertaining to the Structure of the Worst-Case Distribution

In this section, we provide proofs of Theorems 6 and 7, which shed light on the structure of the adversarial distribution(s) attaining the supremum in $\sup_{P:D_c(P_0,P)\le\delta} E_P[\ell(\beta^T X)]$.

Proof of Theorem 6. Recall from Assumption 2 that $\ell(u)$ is convex and grows quadratically or subquadratically as $|u| \to \infty$. Therefore, there exists $\lambda \ge 0$ such that $f_\delta(\beta,\lambda) < \infty$, and subsequently, $\inf_\lambda f_\delta(\beta,\lambda) < \infty$. According to Theorem 1, there exist a dual optimizer, $\lambda_*(\beta)$ in $\arg\min_{\lambda\ge 0} f_\delta(\beta,\lambda)$ for any $\beta \in B$.

a. When $\lambda_*(\beta) = 0$: we have $\inf_{\beta,\lambda} f_\delta(\beta,\lambda) = f_\delta(\beta,0) = \sup_{u\in\mathbb{R}} \ell(u)$. Because of the convexity of $\ell(\cdot)$, the finiteness of the optimal value $f_\delta(\beta,0) = \sup_u \ell(u)$ implies that $\ell(\cdot)$ is a constant function. In this case, any distribution P satisfying $D_c(P,P_0) \le \delta$ is a worst-case distribution attaining the supremum in $\sup_{P:D_c(P,P_0)\le\delta} E_{P_0}[\ell(\beta^T X)]$.

b. It follows from the characterization of the effective domain of $f_\delta(\cdot)$ in Lemma 1 that $f_\delta(\beta,\lambda) = \infty$ when $\lambda < \lambda_{thr}(\beta)$. Therefore, $\lambda_*(\beta) \ge \lambda_{thr}(\beta)$.

c. When $\lambda_*(\beta) > \lambda_{thr}(\beta)$: recall from Proposition 2 the expressions for $\partial_+\ell_{rob}/\partial\lambda$ and $\partial_-\ell_{rob}/\partial\lambda$. Further, we have $f_\delta(\beta,\lambda) < \infty$ for $(\beta,\lambda) \in \mathbb{U}_1 := \{(\beta,\lambda) : \beta \in B, \lambda > \lambda_{thr}(\beta)\}$. Then, it follows from Bertsekas [2, proposition 2.1] that the left and right derivatives $\partial_+ f_\delta/\partial\lambda$ and $\partial_- f_\delta/\partial\lambda$ satisfy

$$\frac{\partial_+ f_\delta}{\partial\lambda}(\beta,\lambda) = \sqrt{\delta}\left(1 - E_{P_0}\left[\beta^T A(X)^{-1}\beta \inf_{g\in\Gamma^*(\beta,\lambda;X)} g^2\right]\right) \text{ and}$$
$$\frac{\partial_- f_\delta}{\partial\lambda}(\beta,\lambda) = \sqrt{\delta}\left(1 - E_{P_0}\left[\beta^T A(X)^{-1}\beta \sup_{g\in\Gamma^*(\beta,\lambda;X)} g^2\right]\right),$$

for $(\beta,\lambda) \in \mathbb{U}_1$. Because $\lambda_*(\beta) > \lambda_{thr}(\beta)$, we have from Lemma 3(a) and the continuous differentiability of $\ell(\cdot)$ that $\Gamma^*(\beta,\lambda_*(\beta);x)$ is compact for P_0-almost every x. Consequently, there exist measurable selections $g_+(\beta,\lambda_*(\beta);x)$ and $g_-(\beta,\lambda_*(\beta);x)$ such that $g_+^2(\beta,\lambda_*(\beta);x) = \sup_{g\in\Gamma^*(\beta,\lambda_*(\beta);X)} g^2$ and $g_-(\beta,\lambda_*(\beta);x) = \inf_{g\in\Gamma^*(\beta,\lambda_*(\beta);X)} g^2$ (see Bertsekas and Shreve [3, proposition 7.50b]). Letting $g_+(\beta,\lambda_*(\beta);X) = G_+$ and $g_-(\beta,\lambda_*(\beta);X) = G_-$, we obtain that

$$\frac{\partial_+ f_\delta}{\partial\lambda}(\beta,\lambda_*(\beta)) = \sqrt{\delta}\left(1 - E_{P_0}\left[G_-^2\beta^T A(X)^{-1}\beta\right]\right) \quad \text{and}$$
$$\frac{\partial_- f_\delta}{\partial\lambda}(\beta,\lambda_*(\beta)) = \sqrt{\delta}\left(1 - E_{P_0}\left[G_+^2\beta^T A(X)^{-1}\beta\right]\right).$$

Because $\lambda_*(\beta) \in \arg\min_{\lambda\ge 0} f_\delta(\beta,\lambda)$, we have from the first order optimality condition that $\partial_+ f_\delta/\partial\lambda(\beta,\lambda_*(\beta)) \ge 0$ and $\partial_- f_\delta/\partial\lambda(\beta,\lambda_*(\beta)) \le 0$. Thus, $\underline{c} = E_{P_0}[G_-^2\beta^T A(X)^{-1}\beta] \le 1$ and $\bar{c} = E_{P_0}[G_+^2\beta^T A(X)^{-1}\beta] \ge 1$. With $G := ZG_- + (1-Z)G_+$ and Z being an independent Bernoulli random variable with $P(Z=1) = (\bar{c}-1)/(\bar{c}-\underline{c})$, we have that

$E_{P_0}[G^2\beta^T A(X)^{-1}\beta] = 1$. In addition, because $G \in \Gamma^*(\beta,\lambda;X)$ P_0-almost surely, we have that

$$X^* \in \arg\max_{x'\in\mathbb{R}^d}\{\ell(\beta^T x') - \lambda_*(\beta)c(X,x')\} \quad \text{and} \quad E[c(X,X^*)] = E[(\sqrt{\delta}G)^2\beta^T A(X)^{-1}\beta] = \delta.$$

As the complementary slackness conditions in Blanchet and Murthy [4, theorem 1] are satisfied, we have that the distribution of X^* attains the supremum in $\sup_{P:D_c(P,P_0)\le\delta} E_P[\ell(\beta^T X)]$.

d. When $\lambda_*(\beta) = \lambda_{thr}(\beta)$: the worst-case distribution $P^*(\beta)$ attaining the supremum in $\sup_{P:D_c(P,P_0)\le\delta} E_P[\ell(\beta^T X)]$ may not exist as demonstrated in the following example. Suppose that $\ell(u) := u^2 - |u|(1-e^{-|u|})$, $\|\beta\| = 1$, $P_0(dx) = \delta_{\{0\}}(dx)$, $\delta > 0$, and $A(x) = \mathbb{I}_d$. For this example, $\ell(\cdot)$ satisfies Assumption 2 with $\kappa = 1$, and $c(\cdot)$ satisfies Assumption 1 with $\rho_{\max} = \rho_{\min} = 1$. For any $\lambda \ge \lambda_{thr}(\beta) = \sqrt{\delta}$, we have $\Gamma^*(\beta,\lambda;\mathbf{0}) = \{\mathbf{0}\}$, and it follows that $f_\delta(\beta,\lambda) = \lambda\sqrt{\delta}$ when $\lambda \ge \lambda_{thr}(\beta)$. Therefore, $\lambda_*(\beta) = \lambda_{thr}(\beta) = \sqrt{\delta}$ and the dual optimal value $f_\delta(\beta,\lambda_*(\beta)) = \delta$. However, this value is not attainable by $E_P[\ell(\beta^T X)]$ for any P satisfying $D_c(P,P_0) \le \delta$. This is because we have $E\|X\|^2 \le \delta$ for any P such that $D_c(P,P_0) \le \delta$, and as a result, we have $E_P[\ell(\beta^T X)] < \delta$ as in the following series of inequalities:

$$E_P[\ell(\beta^T X)] = E_P\left[(\beta^T X)^2 - |\beta^T X|(1-\exp(-|\beta^T X|))\right] < E_P(\beta^T X)^2 \le E_P\|X\|^2 \le \delta.$$

e. When $\lambda_*(\beta) > \lambda'_{thr}(\beta)$: in this case, it follows from Lemma 8 that the map $\gamma \longmapsto F(\gamma,\beta,\lambda_*(\beta);x)$ is strongly concave for P_0-almost every x. As a result, $\Gamma^*(\beta,\lambda_*(\beta);X)$ is singleton, P_0-almost surely. As a result, the random variables, G, G_+, G_-, identified in part (c) satisfy that $P_0(G = G_+ = G_-) = 1$ and $E[G^2\beta^T A(X)^{-1}\beta] = 1$. Therefore, $E[c(X,X^*)] = \delta$. Moreover, the described uniqueness in the optimizer means that $X^* = X + \sqrt{\delta}GA(X)^{-1}\beta$ is the unique element in $\arg\max_{x'\in\mathbb{R}^d}\{\ell(\beta^T x') - \lambda_*(\beta)c(X,x')\}$, P_0-almost surely. Because any distribution $\bar{P}$ attaining the supremum in $\sup_{P:D_c(P,P_0)\le\delta} E_P[\ell(\beta^T X)]$ must satisfy that, if $\bar{X} \sim \bar{P}$, then $\bar{X} \in \arg\max_{x'\in\mathbb{R}^d}\{\ell(\beta^T x') - \lambda_*(\beta)c(X,x')\}$. As a result, we must have that $\bar{X} = X^*$, P_0-almost surely. This verifies that the distribution of X^* is the unique choice that attains the supremum in $\sup_{P:D_c(P,P_0)\le\delta} E_P[\ell(\beta^T X)]$. □

Proof of Theorem 7. Because $\beta \in B$ is fixed throughout the proof, we hide the dependence on β from the parameters $\lambda_*(\beta)$ and $g(\beta,\lambda;x)$ in the notation. Instead, to capture the dependence on δ, we let $\lambda_*(\delta)$ be the choice of λ that solves $\min_{\lambda\ge0} f_\delta(\beta,\lambda)$ for a given choice of $\delta \in (0,\delta_1)$; here, the minimizing $\lambda_*(\delta)$ is unique because of the strong convexity characterization in Theorem 4. For every $\delta < \delta_1$, we have from part (a) of Proposition 9 that $\lambda_*(\delta) > \lambda'_{thr}(\beta)$. Then, we obtain the following reasoning from part (e) of Theorem 6:

i. For every $\delta < \delta_1$, the distribution of $X^*_\delta = X + \sqrt{\delta}G_\delta A(x)^{-1}\beta$ is the unique choice that attains the supremum in $\sup_{P:D_c(P,P_0)\le\delta_i} E_P[\ell(\beta^T X)]$, with $G_\delta := g(\delta,\lambda_*(\delta);X)$, where $g(\delta,\lambda;x)$ is the unique real number that maximizes $F(\gamma,\beta,\lambda;x)$ for P_0-almost every x and $\lambda > \lambda'_{thr}(\beta)$;

ii. Moreover, we have that $E[c(X,X^*_\delta)] = \delta$, and consequently, $g(\delta,\lambda_*(\delta);X)$ satisfies $E_{P_0}[g^2(\delta,\lambda_*(\delta);X)\beta^T A(X)^{-1}\beta] = 1$.

Following the implicit function theorem application in the proof of Proposition 8 (see Online Appendix A), we obtain that

$$\frac{\partial g}{\partial\delta}(\delta,\lambda_*(\delta);x) = -\frac{\partial^2 F/\partial\delta}{\partial^2 F/\partial\gamma}(g(\delta,\lambda_*(\delta);x),\beta,\lambda_*(\delta);x) = \frac{\ell''(\beta^T X^*_\delta)g\beta^T A(X)^{-1}\beta}{2\sqrt{\delta}\varphi_g},$$

where g and φ in the right-hand side denote, respectively, $g(\delta,\lambda_*;x)$ and $\varphi_g(\beta,\lambda_*;x) := 2\lambda_*(\delta) - \sqrt{\delta}\beta^T A(X)^{-1}\beta\ell''(\beta^T X^*_\delta) > \varphi_{\min}\|\beta\| > 0$ (see Proposition 9(b)).

Next, define $H(\delta,\lambda) := E_{P_0}[g(\delta,\lambda;X)^2\beta^T A(X)^{-1}\beta] - 1$. Because $\lambda_*(\delta)$ satisfies $H(\delta,\lambda_*(\delta)) = 0$, a similar application of the implicit function theorem results in

$$\frac{\partial\lambda_*(\delta)}{\partial\delta} = -\frac{\partial H/\partial\delta}{\partial H/\partial\lambda}(\delta,\lambda_*(\delta)) = \frac{E_{P_0}[\ell''(\beta^T X^*_\delta)(g\beta^T A(X)^{-1}\beta)^2/\varphi]}{4\sqrt{\delta}E_{P_0}[g^2\beta^T A(X)^{-1}\beta/\varphi]}.$$

If we let $L(\delta) := \sqrt{\delta}g(\delta,\lambda_*(\delta);x)$, then with an application of chain rule and use of the preceding expressions for $\partial g/\partial\delta, \partial\lambda_*(\delta)/\partial\delta$ and that of $\partial g/\partial\lambda$ in the proof of Proposition 8 (see (46)), we obtain that

$$\frac{\partial L}{\partial\delta}(\delta) = \frac{g}{2\sqrt{\delta}} + \frac{g\beta^T A(X)^{-1}\beta\ell''(\beta^T X^*_\delta)}{2\varphi} - \frac{g}{2\varphi}\frac{E_{P_0}[\ell''(\beta^T X^*_\delta)(g\beta^T A(X)^{-1}\beta)^2/\varphi]}{E_{P_0}[g^2\beta^T A(X)^{-1}\beta/\varphi]},$$

if $g \neq 0$. When $\delta < \delta_1$, we have $\varphi > \varphi_{\min}\|\beta\| > 0$ (see Proposition 9(b)). Moreover, $\beta^T A(X)^{-1}\beta \leq R_\beta \rho_{\max}^{-1}\|\beta\|$ and $\ell''(\cdot) \in (0, M]$ (see Assumptions 1–3). As a result, we obtain that

$$\frac{2}{g}\frac{\partial L}{\partial \delta}(\delta) > \frac{1}{\sqrt{\delta}} - \frac{\rho_{\max}^{-1} M R_\beta \|\beta\|}{\varphi_{\min}\|\beta\|} = \frac{1}{\sqrt{\delta}} - \frac{1}{\sqrt{\delta_0} - \sqrt{\delta}},$$

where the last equality follows from the definitions of δ_0 and $\varphi_{\min}$ in Section 5.3. Because $\delta < \delta_1 \leq \delta_0/4$, we have that $2g^{-1}\partial L(\delta)/\partial\delta > 0$ if $g \neq 0$ and $\partial L(\delta)/\partial\delta = 0$ if $g = 0$. Further, observe that, as a consequence of the mean value theorem, the first order optimality condition (44) means that $g(\delta, \lambda_*(\delta); X) = \ell'(\beta^T X)/(2\lambda_*(\delta) - \sqrt{\delta}\beta^T A(X)^{-1}\beta\ell''(\eta))$, for some η between the real numbers $\beta^T X$ and $\beta^T X_\delta^*$. Because $2\lambda_*(\delta) - \sqrt{\delta}\beta^T A(X)^{-1}\beta\ell''(\eta) \geq \varphi_{\min}\|\beta\| > 0$, we have that the sign of $G_\delta := g(\delta, \lambda_*(\delta); X)$ matches with that of $\ell'(\beta^T X)$. As a result, with $L(\delta) := \sqrt{\delta} g(\delta, \lambda; X) = \sqrt{\delta} G_\delta$, the claims made in Proposition 7, (b)–(d), are verified. This completes the proof of Theorem 7. □

5.5. Proofs of the Results on Rates of Convergence

Lemma 9, establishing finite second moments for the gradients (or) subgradients utilized in SGD schemes, is useful toward proving Propositions 4 and 7. Recall the definitions of $\mathbb{U}_\eta$ in (17) and $D(\beta, \lambda; X)$ in (18).

Lemma 9. *Suppose that Assumptions 1 and 2 are satisfied, $\ell(\cdot)$ is continuously differentiable, $\eta > 0$, and $E_{P_0}\|X\|^4 < \infty$. For any $\theta \in \mathbb{U}_\eta$, let $h(\theta; X)$ be such that $h(\theta; X) \in D(\theta; X)$, P_0−almost surely. Then, there exists a positive constant G_η such that $E_{P_0}\|h(\theta; X)\|^2 \leq G_\eta$ for any $\theta \in \mathbb{U}_\eta$.*

The proof of Lemma 9 is presented in Online Appendix A.

Proof of Proposition 4.

a. When $\delta < \delta_0$, it follows from Propositions 2 and 3 that the subgradient set $\partial\ell_{rob}(\beta, \lambda; X) = \{\nabla_\theta \ell_{rob}(\beta, \lambda; X)\}$, P_0−almost surely. Because $\lambda > \lambda'_{thr}(\beta) \geq \lambda_{thr}(\beta)$ for every $(\beta, \lambda) \in \mathbb{W}$ (see Proposition 9(a)), it follows from Lemma 9 that $\sup_{\theta\in\mathbb{W}} E\|\nabla_\theta \ell_{rob}(\theta; X)\|^2 < \infty$, when $\delta < \delta_0$. As a consequence, we have from Theorem 2 and the remark in Shamir and Zhang [30, theorem 4] that $E[f_\delta(\theta_k)] - f_* = O(k^{-1/2}\log k)$ and $E[f_\delta(\bar{\theta}_k)] - f_* = O(k^{-1/2})$, as $k \to \infty$. Proposition 4(a) now follows as a consequence of Markov's inequality.

b. When $\delta < \delta_1$, it follows from the positive definiteness of the Hessian around the unique minimizer $\theta_* := \arg\min f_\delta(\theta)$ (see Theorem 4) that there exists $\varepsilon > 0$ satisfying $(\theta - \theta_*)^T \nabla_\theta f_\delta(\theta) \geq \kappa_1\sqrt{\delta}\|\theta - \theta_*\|^2$ for all $\theta \in \mathbb{V}$ and $\|\theta - \theta_*\| \leq \varepsilon$. Further, because of the uniqueness of the minimizer, we also have $(\theta - \theta_*)^T \nabla_\theta f_\delta(\theta) > 0$. Similar to part (a), as $\lambda > \lambda'_{thr}(\beta) \geq \lambda_{thr}(\beta)$ for every $(\beta, \lambda) \in \mathbb{W}$, we have because of Lemma 9 that $\sup_{\theta\in\mathbb{W}} E\|\nabla_\theta \ell_{rob}(\theta; X)\|^2 < \infty$. Taylor's expansion of $\nabla_\theta f_\delta(\theta)$ results in

$$\|\nabla_\theta f_\delta(\theta) - \nabla_\theta^2 f_\delta(\theta_*)^T(\theta - \theta_*)\| = o(\|\theta - \theta_*\|), \tag{38}$$

for $\theta \in \mathbb{W}$. With these conditions being satisfied, it follows from Polyak and Juditsky [26, theorem 2] that $\sqrt{k}(\bar{\theta}_k - \theta_*) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \Sigma)$, as $k \to \infty$, where $\Sigma := (\nabla_\theta^2 f_\delta(\theta_*))^{-1}\mathrm{Cov}[\nabla_\theta \ell_{rob}(\theta_*; X)]((\nabla_\theta^2 f_\delta(\theta_*))^{-1})^T$. If we let $Z \sim \mathcal{N}(0, \mathbb{I}_{d+1})$, then because of the continuous mapping theorem, we have that the distribution of $k(\bar{\theta}_k - \theta_*)^T \nabla_\theta^2 f_\delta(\theta_*)(\bar{\theta}_k - \theta_*)$ is convergent to that of

$$Z^T \Sigma^{1/2} \nabla_\theta^2 f_\delta(\theta_*) \Sigma^{1/2} Z = Z^T \nabla_\theta^2 f_\delta(\theta_*)^{-1/2} \mathrm{Cov}[\nabla_\theta \ell_{rob}(\theta_*; X)] \nabla_\theta^2 f_\delta(\theta_*)^{-1/2} Z.$$

The local strong convexity characterization in Theorem 4 yields that that the maximum eigenvalue of $\nabla_\theta^2 f_\delta(\theta_*)^{-1/2}$ is bounded from above by a constant times $\delta^{-1/4}$. As a result of the described convergence in distribution, we have that

$$(\bar{\theta}_k - \theta_*)^T \nabla_\theta^2 f_\delta(\theta_*)(\bar{\theta}_k - \theta_*) = O_p(k^{-1}).$$

Now, it follows from the local joint strong convexity of $f_\delta(\cdot)$ in Theorem 4 and (38) that

$$\begin{aligned} f_\delta(\bar{\theta}_k) - f_* &\leq \nabla_\theta f_\delta(\bar{\theta}_k)^T(\bar{\theta}_k - \theta_*) - \frac{\kappa\sqrt{\delta}}{2}\|\theta_k - \theta_*\|^2 \\ &= (\bar{\theta}_k - \theta_*)^T \nabla_\theta^2 f_\delta(\theta_*)(\bar{\theta}_k - \theta_*) - \left(\frac{\kappa\sqrt{\delta}}{2} + o(1)\right)\|\bar{\theta}_k - \theta_*\|^2 = O_p(k^{-1}). \end{aligned}$$

This completes the proof of Proposition 4. □

6. Conclusions

Our main objective in this paper has been to set the stage for algorithms and analysis of a flexible class of DRO problems. Our motivation stems from the observations that (i) a flexible choice of the distributional uncertainty region is useful toward fully exploiting the advantages of DRO in data-driven contexts and (ii) the existing computational methods largely pertain to Lipschitz losses and do not scale well with data size. We show that, in the case of affine decision rules and convex loss functions, robustification with a more flexible state-dependent Mahalanobis cost function does not introduce significantly additional computational complexity relative to the non-DRO counterpart (in terms of standard benchmark iterative algorithms used to solve the non-DRO problem). In some cases, interestingly, DRO introduces strong-convexity, which results in lower iteration complexity.

Naturally, the algorithmic approach and structural analysis presented in this paper can be considered in DRO formulations with further general cost functions of the form $c(x,x') = u(x-x')$ or $c(x,x') = u(x') - u(x) - \nabla u(x)(x'-x)^T$ for a strongly convex function $u(\cdot)$ with Lipschitz-continuous gradients. Although such extensions may render the inner maximization in (6) as a multidimensional optimization problem (as opposed to the line search in the state-dependent Mahalanobis case), a number of observations and structural properties are expected to continue to hold; for example, observations relating to convexity properties, magnitude of mass transportation in the worst-case distribution being of size $O_p(\sqrt{\delta})$, computation of stochastic gradients by means of envelope theorem, etc., are expected to generalize to the families of strongly convex, smooth transportation cost functions. We leave this exploration as a question for future research.

Our philosophy is that, by providing a general analysis for a flexible class of cost functions, a modeler will be able to choose a cost function that enhances out-of-sample performance in a way that is convenient and meaningful for the needs of the modeling situation. Although examples of how one may choose the transportation cost function in a data-driven way are available in existing literature (see, for example Blanchet et al. [7]), systemic treatment of the contextual choice of transportation cost is an essential question for future research.

Acknowledgments

The authors thank the editors and anonymous reviewers for their careful review and insightful comments and suggestions.

References

[1] Allen-Zhu Z (2017) Katyusha: The first direct acceleration of stochastic gradient methods. Hatami H, McKenzie P, King V, eds. *Proc. 49th Annual ACM SIGACT Sympos. Theory Comput.* (ACM, New York), 1200–1205.
[2] Bertsekas DP (1973) Stochastic optimization problems with nondifferentiable cost functionals. *J. Optim. Theory Appl.* 12:218–231.
[3] Bertsekas DP, Shreve SE (1978) *Stochastic Optimal Control: The Discrete Time Case* (Elsevier, Amsterdam).
[4] Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* 44(2):565–600.
[5] Blanchet J, Kang Y, Murthy K (2019) Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.* 56(3): 830–857.
[6] Blanchet J, Murthy K, Si N (2019) Confidence regions in Wasserstein distributionally robust estimation. Preprint, submitted June 4, https://arxiv.org/abs/1906.01614.
[7] Blanchet J, Kang Y, Murthy K, Zhang F (2019) Data-driven optimal transport cost selection for distributionally robust optimization. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 3740–3751.
[8] Chen Z, Kuhn D, Wiesemann W (2018) Data-driven chance constrained programs over Wasserstein balls. Working paper, City University of Hong Kong.
[9] Chen Z, Sim M, Xiong P (2018) Adaptive robust optimization with scenario-wise ambiguity sets. Working paper, City University of Hong Kong.
[10] Defazio A, Bach F, Lacoste-Julien S (2014) SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Red Hook, NY), 1646–1654.
[11] den Boef E, den Hertog D (2007) Efficient line search methods for convex functions. *SIAM J. Optim.* 18(1):338–363.
[12] Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with Wasserstein distance. Working paper, Georgia Institute of Technology, Atlanta.
[13] Gao R, Chen X, Kleywegt AJ (2017) Wasserstein distributional robustness and regularization in statistical learning. Working paper, Georgia Institute of Technology, Atlanta.
[14] Gao R, Xie L, Xie Y, Xu H (2018) Robust hypothesis testing using Wasserstein uncertainty sets. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 31 (Curran Associates, Red Hook, NY), 7902–7912.
[15] Hanasusanto GA, Kuhn D (2018) Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. *Oper. Res.* 66(3):849–869.
[16] Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 26 (Curran Associates, Red Hook, NY), 315–323.

[17] Luo F, Mehrotra S (2019) Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models. *Eur. J. Oper. Res.* 278(1):20–35.
[18] Milgrom P, Segal I (2002) Envelope theorems for arbitrary choice sets. *Econometrica* 70(2):583–601.
[19] Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming.* 171(1):115–166.
[20] Mokkadem A, Pelletier M (2006) Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Ann. Appl. Probab.* 16(3):1671–1702.
[21] MOSEK ApS (2019) MOSEK Optimizer API for Python 9.2.10. Accessed April 1, 2021, https://docs.mosek.com/9.2/pythonapi/index.html.
[22] Moulines E, Bach F (2011) Non-asymptotic analysis of stochastic approximation algorithms for machine learning. Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 24 (Curran Associates, Red Hook, NY), 451–459.
[23] Nemirovski A, Juditsky A, Lan G, Shapiro A (2008) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* 19(4):1574–1609.
[24] Nguyen VA, Kuhn D, Mohajerin Esfahani P (2018) Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. Preprint, submitted May 18, https://arxiv.org/abs/1805.07194.
[25] Noh YK, Zhang BT, Lee D (2010) Generative local metric learning for nearest neighbor classification. Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A, eds. *Advances in Neural Information Processing Systems*, vol. 23 (Curran Associates, Red Hook, NY), 1822–1830.
[26] Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* 30(4):838–855.
[27] Robbins H, Monro S (1951) A stochastic approximation method. *Ann. Math. Statist.* 22(3):400–407.
[28] Shafieezadeh-Abadeh S, Kuhn D, Mohajerin Esfahani P (2019) Regularization via mass transportation. *J. Machine Learn. Res.* 20:1–68.
[29] Shafieezadeh-Abadeh S, Mohajerin Esfahani P, Kuhn D (2015) Distributionally robust logistic regression. Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 28 (Curran Associates, Red Hook, NY), 1576–1584.
[30] Shamir O, Zhang T (2013) Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. Dasgupta S, McAllester D, eds. *Proc. 30th Internat. Conf. Machine Learn.*, vol. 28 (PMLR, Atlanta), 71–79.
[31] Shapiro A, Dentcheva D, Ruszczyński A (2014) *Lectures on Stochastic Programming*, MOS-SIAM Series on Optimization, vol. 9 (SIAM, Philadelphia).
[32] Sinha A, Namkoong H, Duchi J (2018) Certifiable distributional robustness with principled adversarial training. *Internat. Conf. Learn. Representations*.
[33] Volpi R, Namkoong H, Sener O, Duchi JC, Murino V, Savarese S (2018) Generalizing to unseen domains via adversarial data augmentation. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 31 (Curran Associates, Red Hook, NY), 5339–5349.
[34] Wang J, Kalousis A, Woznica A (2012) Parametric local metric learning for nearest neighbor classification. Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 25 (Curran Associates, Red Hook, NY), 1–9.
[35] Xie W (2021) On distributionally robust chance constrained programs with Wasserstein distance. *Math. Programming* 186(1–2):115–155.
[36] Yang I (2017) A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance. *IEEE Control Systems Lett.* 1(1):164–169.
[37] Zhao C, Guan Y (2018) Data-driven risk-averse stochastic optimization with Wasserstein metric. *Oper. Res. Lett.* 46(2):262–267.