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

JOSE BLANCHET*, KARTHYEK MURTHY†, AND FAN ZHANG*

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 was 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 which have the same sample and iteration complexity as a natural non-DRO benchmark algorithm such as stochastic gradient descent.

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 (OR) and Machine Learning (ML) 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, while the benchmark model, P_0, is typically chosen to be a tractable model which 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 the formulation (1) may result in the selection

(∗) Management Science and Engineering, Stanford University

(†) Engineering Systems & Design, Singapore University of Technology & Design

E-mail addresses: jose.blanchet@stanford.edu, karthyek_murthy@sutd.edu.sg, fzh@stanford.edu.

Key words and phrases. Distributionally Robust Optimization, Stochastic Gradient Descent, Optimal Transport, Wasserstein distances, Adversarial, Strong convexity, Comparative statics, Rate of convergence.

of a decision, β_0, which significantly under-performs 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) which 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 which 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

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

where $\mathcal{W}(P_0, P)$ is a Wasserstein distance (see, for example, [4, 6, 8, 15, 16, 21, 29, 33, 34, 37, 38] and references therein).

The Wasserstein distance is a particular case of optimal transport discrepancies, which we will 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, [4, 14, 29, 30]. 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 towards infinity or moving no mass at all).

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

For other special cases which 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 [25]) or ultimately requires robust optimization formulations which require P_0 to have finite support; see [8, 9, 19, 21, 29, 30, 36, 38]

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

We mention [33], 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 [33] is similar to the ones that 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 ($\lambda = O\left(\delta^{-1/2}\right)$) may enhance performance, even in the case of the more general type of losses considered in [33]. 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 non-trivial to analyze. So, much of our effort will go into understanding these properties. But before we describe our results, we first describe a flexible class of models for distributional uncertainty sets, $\mathcal{U}_\delta(P_0)$.

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,

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

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

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

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

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. We explain in the Conclusions section, however, 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 which 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 which is not of statistical nature but of 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 which 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 lower cost of transportation; while 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 discussed 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 ([5, 26, 35]). 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 non-trivial to study.

However, to exploit these formulations one must develop algorithms which enable 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 which 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 also have far ranging implications, as we discuss next.

A more in-depth discussion of our technical contributions.

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

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

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

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 $\kappa_1 > 0$ which 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 under (exponentially fast rate of convergence) under the setting of Theorems 1 - 4.

We then study a natural stochastic gradient descent algorithm for solving (2) which, due to 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 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 $O\left(1/\varepsilon^2\right)$ 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 $\delta \geq 0$), 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 possed 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, [32], 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 Proposition 7 and Theorem 6.

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 non-convex objective functions. Some of the statistical implications are studied in (see [7]).

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 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 semi-continuous 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 of (6).

In Section 2.2.2, we add the assumptions that $\ell(\cdot)$ is twice differentiable, with a natural non-degeneracy 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 non-degeneracy condition on P_0 mentioned earlier, we extend the smoothness and strong convexity of the right hand side of (6) both in 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 which 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 which do not require the loss $\ell(\cdot)$ to be differentiable. So, the result can be applied to developing stochastic subgradient descent algorithms for non-differentiable 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 corresponding iteration complexity analysis discussed in Section 3.2.2.

In Section 3.3 we discussed 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 non-differentiable 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, 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 Appendix D.

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&P500 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 the appendix, in Section A. The discussion on the complexity of the line search, which underlying the gradient evaluation of $\ell_{rob}(\cdot)$ and is given in Section B.

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 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 donate, respectively, the gradient and Hessian of f. In instances where 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 re-express 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 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

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

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 $E_\pi[c(X, X')] \leq \delta$ (see [6, 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 below.

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 $\sup_{\|v\|=1} v^T A(x)v \leq \rho_{\max}$ and $\inf_{\|v\|=1} v^T A(x)v \geq \rho_{\min}$, 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 (hence need not be a metric). The special case of $A(x)$ being the identity matrix (for all x) corresponds to the $D_c^{1/2}(\cdot)$ being the well-known Wasserstein distance (in this case, the constants $\rho_{\max} = \rho_{\min} = 1$). Theorem 1 below 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,*

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

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

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

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

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 $\kappa := \inf\{s \geq 0 : \sup_{u\in\mathbb{R}}(\ell(u) - su^2) < \infty\}$ 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. While these assumptions are helpful in understanding smoothness and strong convexity properties, the development of iterative schemes in Section 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 $\ell''(\cdot) \leq M$. 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 $\arg\min_{\lambda\geq 0} f_\delta(\beta, \lambda)$ 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 $\lambda_*(\beta) \in \arg\min_{\lambda \geq 0} f_\delta(\beta, \lambda)$, we have $(\beta, \lambda_*(\beta)) \in \mathbb{V}$, where*

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

for some positive constants K_1, K_2 which 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 Table 2 (see 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) *$f_\delta(\cdot)$ 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 towards 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 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 $P_0\left(|\ell'(\beta^T X)| > c_1, \ |\beta^T X| > c_2 \|\beta\|\right) \geq p$.*

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 above 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 where $\ell(u) = (y - u)^2$ 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 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, due to 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 non-robust counterpart $E_{P_0}[\ell(\beta^T X)]$ is not.*

Comments on Assumptions 1 - 5. Assumptions 1 - 2 above ensure that the DRO objective (6) is convex, proper and that the strong duality utilized in Theorem 1 is indeed applicable. These non-restrictive assumptions serve the purpose of clearly stating the framework considered. Indeed, Assumptions 1 - 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 shall 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 Hessian of f_δ. The assumption of boundedness of the set B, though not necessary for strong convexity (see following Section 2.2.3), is essential for guaranteeing 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 where 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 would like to reiterate that the development of iterative schemes in Section E does not require Assumptions 3 - 5.

2.2.3. *Strong convexity property for non-compact B..* As we shall see in Theorem 5 below, 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 which allows to study 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 Section 5.4.

Theorem 5. *Suppose that Assumptions 1 - 3 and Assumption 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 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

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

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

Theorem 6. *Suppose that Assumptions 1,2 hold and $\beta \ne \mathbf{0}$. 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* $P(Z = 1) = (\overline{c} - 1)/(\overline{c} - \underline{c})$, *where* $\overline{c} := E_{P_0}[G_+^2 \beta^T A(X)^{-1}\beta]$ *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 5.5.

Remark 2. Consider the case $\beta = \mathbf{0}$. Then $\lambda = 0$ attains the minimum in $\min_{\lambda \geq 0} f_\delta(\mathbf{0}, \lambda)$, and $\sup_{D_c(P_0,P) \leq \delta} E_{P_0}[\ell(\beta^T X)] = \ell(0)$, 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* $\beta \in B \setminus \{\mathbf{0}\}$, *there exists a unique worst-case distribution* P_δ^* *which 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 5.5. 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 can be inferred from Figure 2 in Section 4 devoted to numerical demonstrations.

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 'non-robust' 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 where 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) above 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, [29, 21]). Further, the proposed stochastic gradients based approaches are also immediately applicable to problems where P_0 has uncountably infinite support.

Property b) above makes sure that computational intractability is not a reason that should deter the use of DRO approach towards 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 to its stochastic optimization counterpart $\inf_{\beta\in B} E_{P_0}[\ell(\beta^T X)]$. As we shall 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,

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

Lemma 1 below, whose proof is presented in 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,

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

Lemma 1. *Suppose that Assumptions 1 - 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 $\lambda > \kappa\sqrt{\delta}\beta A(x)^{-1}\beta$; and*
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 1a 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 below utilizes envelope theorem (see [20]) 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, $\Gamma^*(\beta,\lambda;x)\neq\varnothing$, for any $(\beta,\lambda)\in\mathbb{U}_1$.*
b) *The maps $\lambda\mapsto\ell_{rob}(\beta,\lambda;x)$, $\beta_j\mapsto\ell_{rob}(\beta,\lambda;x)$ are absolutely continuous for $(\beta,\lambda)\in\mathbb{U}_1$, and their directional derivatives are given by,*

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

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

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

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

Furthermore, $\lambda\mapsto\ell_{rob}(\beta,\lambda;x)$ is differentiable if and only if $\{\frac{\partial_+F}{\partial\lambda}(\gamma,\beta,\lambda;x), \frac{\partial_-F}{\partial\lambda}(\gamma,\beta,\lambda;x) : \gamma\in\Gamma^*(\beta,\lambda;x)\}$ is a singleton. Likewise, for j in $\{1,\ldots,d\}$ $\beta_j\mapsto\ell_{rob}(\beta,\lambda;x)$ is differentiable

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

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

A proof of Proposition 2 can be found in Appendix A. Recall that a simple subgradient descent (or) stochastic subgradient descent for solving the 'non-robust' problem $\inf_{\beta\in B} E_{P_0}\left[\ell(\beta^T X)\right]$ assumes access to first-order oracle evaluations $\ell(\cdot)$ and $\partial_+\ell(\cdot), \partial_-\ell(\cdot)$. Likewise, due to 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 below), 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,

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

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

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

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 Appendix A.

3.2.1. *The iterative scheme.* Due to 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 $f_\delta(\theta)$. Utilising 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 k-th iterate θ_k and its weighted running average $\bar{\theta}_k$ as follows:

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

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 Appendix C and a detailed algorithmic description of the above steps is described in 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 $(\theta_k)_{k\geq 1}$ are the classical Robbins-Monro iterates with slower step-sizes (see [28]). If $\xi = 0$ in the definition of $\bar{\theta}_k$ in (15), the iterate $\bar{\theta}_k$ is simply the running average of $\theta_1, ...\theta_{k-1}$ and the averaging scheme is the well-known Polyak-Ruppert averaging for stochastic gradient descent (see [27] and references therein). On the other hand, the averaging scheme with $\xi > 0$ is referred as polynomial-decay averaging (see [31]).

3.2.2. *Rates of convergence.* Our objective here is to characterize the convergence of $(f_\delta(\bar{\theta}_k))_{k\geq 1}$ 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, [31] and references therein). While $f_\delta(\cdot)$ is convex for all $\delta \geq 0$, 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 5.6.

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

a) $f_\delta(\bar{\theta}_k) - f_* = O_p(k^{-1/2})$ *if* $\delta < \delta_0$, $\xi \geq 1$ *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 $(f_\delta(\bar{\theta}_k))_{k\geq 1}$ with the robustness property that the precise choice of step-size $(\alpha_k)_{k\geq 1}$ does not affect the convergence behaviour as long as the step size choice satisfies Assumption 6. Contrast this with the vanilla stochastic approximation iterates $(\theta_k)_{k\geq 1}$ 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, [23, 24] 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, [4, 29]). Therefore the requirement that $\delta < \delta_1$ is typically satisfied in practice in data-driven applications.

Indeed if $\delta < \delta_1$, due to Proposition 4b), it suffices to terminate after $O_p(\varepsilon^{-1})$ iterations in order to obtain an iterate $\bar{\theta}_k$ that satisfies $f_\delta(\bar{\theta}_k) - f_* \leq \varepsilon$. On the other hand, if $\delta > \delta_1$, we require the usual $O_p(\varepsilon^{-2})$ iteration complexity to obtain $f_\delta(\theta_k) - f_* \leq \varepsilon$, which is identical to the sample complexity of stochastic gradient descent for the non-robust problem $\inf_\beta E_{P_0}[\ell(\beta^T X)]$ in the presence of convexity (see, for example, [31]). Here, recall from the discussion following Theorem 4 that the non-robust 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 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 $(\theta_k : k \geq 1)$, including bias induced by the line search, in the strongly convex case.

Proposition 5. *Suppose that Assumptions 1 - 6 hold and* $\delta < \delta_1$. *At the k-th 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* k*-th 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. 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. While 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 towards designing efficient algorithms for DRO based on flexible cost functions.*

To complete this discussion, recall that the dual formulation,

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

that we are working with is 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, these change of variables involve 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 $\sup_{\gamma \geq 0} F(\gamma, \beta, \lambda_*(\beta); x)$ for any δ, x and a fixed $\beta \in B$, we shall also witness in Proposition 9b that $g_\delta(X) = O_p(1)$, as $\delta \to 0$. These two properties ensure that the inner and outer optimization problems $\inf_{\lambda \geq 0} E_{P_0}\left[\sup_{\gamma \in \mathbb{R}} F(\gamma, \beta, \lambda; X)\right]$ are well-conditioned and their solutions remain scale-free (with respect to δ).

For algorithms that directly proceed with the dual reformulation in [6, Theorem 1] or [15, Theorem 1] without employing the above described scaling of variables by factor $\sqrt{\delta}$, the resulting dual formulation will have 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 to 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.* Since 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:

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

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

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

where 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 $\alpha_k/\gamma_k \to 0$. Since α_k is very small relative to γ_k, the iterates β_k remain relatively static compared

to 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 [22, 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 behaviour.

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.

Since $f_\delta(\cdot)$ 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 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 desired precision, any standard line search method, such as triangle section method (see Algorithm 3 in [13]), that exploits 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 (where 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, due to 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, [17, 10]), or momentum based acceleration (see [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 2b that the set,

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

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

Proposition 6. *Suppose that Assumptions 1-2 are satisfied and the loss $\ell(\cdot)$ is of the form $\ell(u) = \max_{i=1,\dots,K}\ell_i(u)$ 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 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 $(k-1)$-st iteration, the k-th iterate is computed as follows:

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

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 $\tilde{X}_k := X_k+\sqrt{\delta}gA(X_k)^{-1}\beta$, 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 due to Proposition 6, we have that $EH_k\in\partial f_\delta(\beta_{k-1},\lambda_{k-1})$. Due to 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 Section 5.6.

Proposition 7. *Suppose that Assumptions 1-2 are satisfied and the loss $\ell(\cdot)$ is of the form $\ell(u) = \max_{i=1,\dots,K}\ell_i(u)$ 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 $\varepsilon-$precision is $O_p(L\varepsilon^{-2})$. A brief description of the complexity L introduced by the line search can be found in 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 in a laptop computer with a 2.2GHz Inter Core i7 GPU and 16GB 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 [5]. Therefore, in the interest of space and to avoid repetition, we restrict the focus in this subsection to reporting the results of stylized numerical experiments which 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 non-robust 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,

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

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, where the data is 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

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

We are interested in solving the distributionally robust logistic regression problem,

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

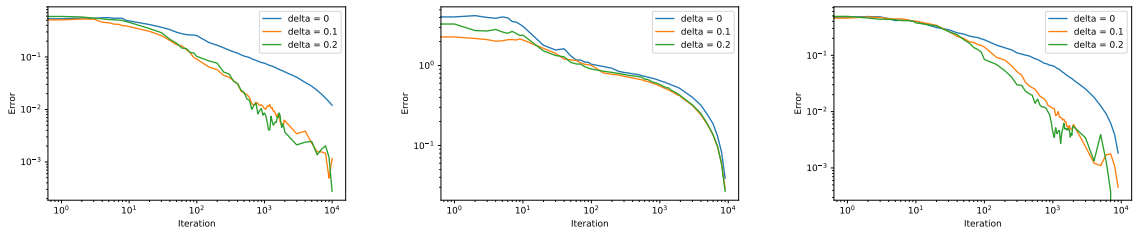
where $P_0 = P_n(dx, dy) := \frac{1}{n}\sum_{i=1}^n \delta_{\{(X_i, Y_i)\}}(dx, dy)$ is the empirical measure of data.

In 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 distributionally robust logistic regression. The data is generated from normal distribution, with different mean for each class and same variance. The total number of data points is $n = 1024$, 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, 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, then we plot the optimality gap (Error) versus number of iterations for DRO-model and ordinary logistic model in Figure 1a.

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.



(a) Logistic Regression **(b)** Linear Regression. **(c)** Support Vector Machine

FIGURE 1. Convergence of loss function

The dots in different color denote the data in different classes: on the lower left side the data is classified to be red and on the upper right side the data is classified to be blue. 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 towards the wrong side of the decision boundary, which are aligned with our observations pertaining to comparative statics in Theorem 6 (see Section 2.4).

4.1.3. *Linear regression.* Now we turn to consider the example of linear regression with squared loss function. In this data is 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

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

Following a similar argument as in the example of logistic regression, it is not hard to verify the squared loss function satisfies all the assumptions regarding the loss function. We refer the interested readers to 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

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

Now we explain the setting of our numerical experiment in this example. The dimension of data is $d = 16$, and we randomly generate $n = 256$ training data points. The matrix 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 DRO-model and ordinary linear regression model is given in Figure 1b.

4.1.4. *Support vector machines.* We consider the case of binary classification, where the data is given by $\{(X_i, Y_i)\}_{i=1}^n$, 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

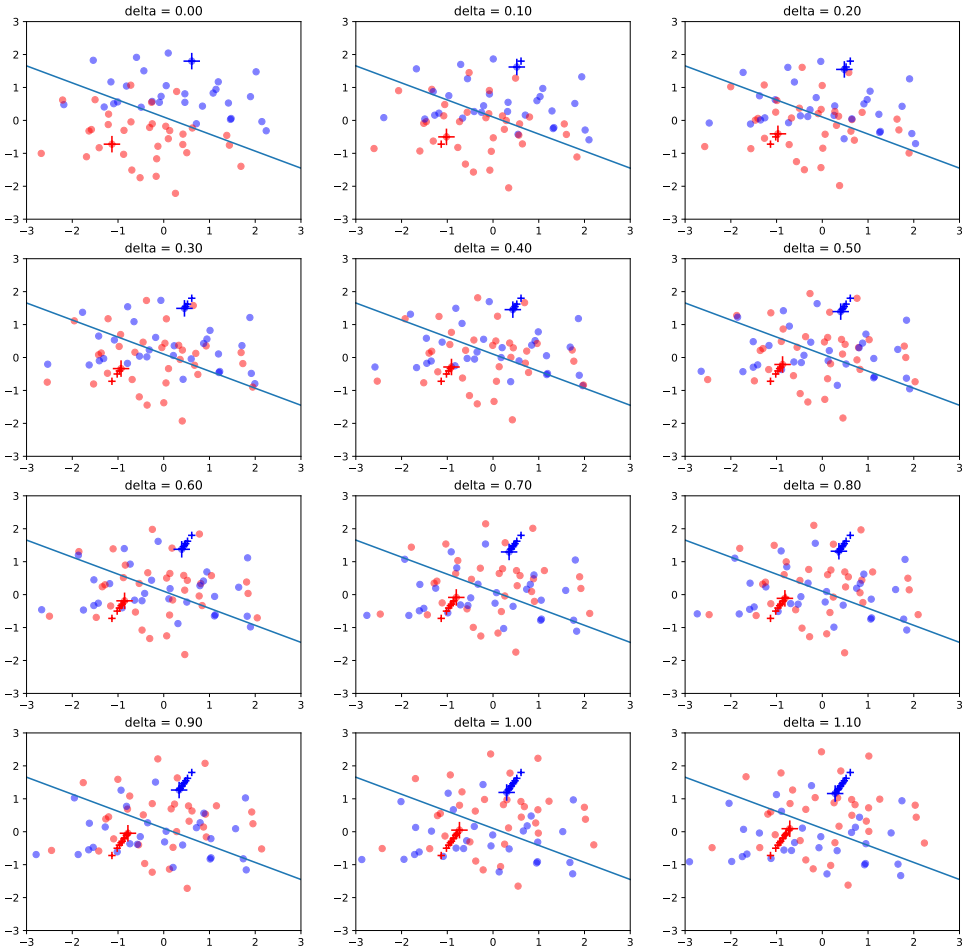


FIGURE 2. Decision boundary and worst case distribution. To facilitate tracking the change of X_δ^* when δ is increasing, we select one point from each class and use a big $+$ to mark its position. We also employ a small $+$ to mark its previous position when δ is smaller so that the trajectory of the 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.

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.

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

In the numerical experiment, we use the same data as the example of logistic regression. Again, we set the learning rate to be same for DRO and non-DRO algorithms. Figure 1c shows the path of optimality gaps of loss functions during iterations. We use the value of loss function at 10^5 iteration as the approximate optimal loss given training samples, and plot the optimality gap (Error) versus number of iterations in Figure 1c.

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,

$$\min_{\beta:\beta^T\mathbf{1}=1} \mathrm{Var}_{P^*}[\beta^T X] - \zeta\cdot E_{P^*}[\beta^T X], \tag{20}$$

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 (20) can be rewritten as the following stochastic optimization problem with affine decision rules:

$$\inf_{\mu}\inf_{\beta:\beta^T\mathbf{1}=1} E_{P^*}[(\beta^T X-\mu)^2 - \zeta\cdot\beta^T X]. \tag{21}$$

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 $\{X_1,\dots,X_n\}$. Due to the discrepancy between the ground-truth measure P^* and the reference measure $P_0 = P_n$, we consider the following distributionally robust variant of (21):

$$\inf_{\mu}\inf_{\beta:\beta^T\mathbf{1}=1}\sup_{P:D_c(P_0,P)\le\delta} E_{P_0}[(\beta^T X-\mu)^2 - \zeta\cdot\beta^T X]. \tag{22}$$

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 which 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 which 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 $\{V_i : i=1,\dots,n\}$ in addition to the returns data $\{X_i : i=1,\dots,n\}$; here, V_i is a positive scalar that represents the implied volatilities corresponding to the i-th observation X_i. Let $\overline{V} = n^{-1}\sum_{i=1}^n V_i$ 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

$$A_i = \frac{\overline{V}}{V_i}\mathbb{I}_d, \quad i=1,\dots,n. \tag{23}$$

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 (23) is such that it is cheaper to perturb returns (or

transport mass) for observations with higher implied volatility. The normalization by $\overline{V}$ 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 (22) with real data, we randomly pick 20 stocks from the constituents of 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 the formulation (22) 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 the model (22) 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 CBOE volatility index (VIX), 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 hyper-parameter and the out-of-sample efficient frontier is generated by considering different values of the parameter ζ. In Figure 3a, 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 3a as "out-of-sample". These out-of-sample efficient frontiers reveal that the DRO formulation (22) with state-dependent Mahalanobis cost choice (as in (23)) performs uniformly better than that obtained with the Euclidean distance choice (corresponding to constant $A(x) = \mathbb{I}_d$, addressed as constant model in Figure 3a). We also observe that a larger value of distributional uncertainty δ results in 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 3a suggest that the out-of-sample effects (such as non-stationarity 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 3b) for the choice $A(x) = \mathbb{I}_d$. In this simulation experiment, the DRO model is observed to produce less efficient portfolios relative to non-robust 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 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 flexibility to use additional information such as implied volatilities to elicit collective market expectations about future uncertainty.

5. Proofs of main results.

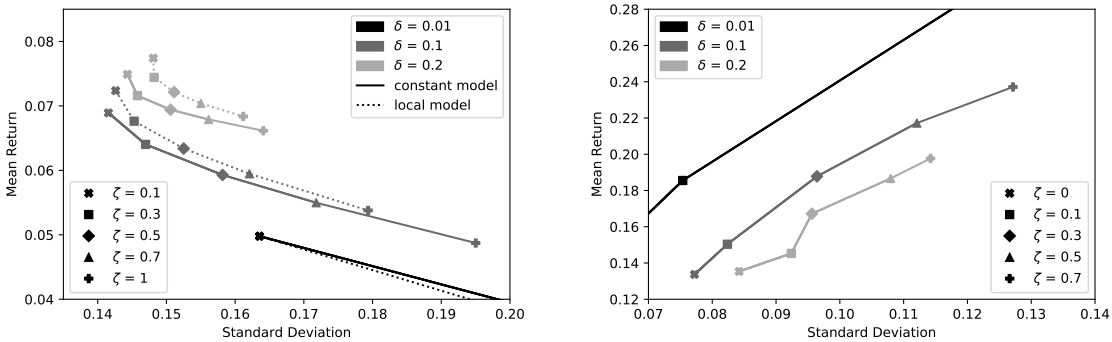
We shall provide proofs of all the main results in Sections 2 - 3 in this section. The proofs of auxiliary results, which are technical in nature, are provided in the subsequent technical appendix Section A for ease of reading.

5.1. Proofs of the results on Dual reformulation and convexity.

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

Proof of Theorem 1. Since $c(\cdot)$ is lower semicontinuous and $\ell(\cdot)$ is upper semicontinuous, it follows from the the strong duality result in Theorem 1 of [6] that

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



(a) Real Data Experiment **(b)** Simulated Data Experiment

FIGURE 3. Out of sample efficient frontier. The mean and the standard deviation are annualized. We use solid lines to represent models with constant optimal transport cost function, and use dashed lines to represent the models with state-dependent Mahalanobis optimal transport cost function. The different choices of δ are denoted by different colors. The values of ζ are represented by different shapes of the markers.

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

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

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

$$\sup_{\Delta\in\mathbb{R}^d}\left\{\ell(\beta^T(X+\Delta)) - \lambda\left(\Delta^T A(X)\Delta - \delta\right)\right\} = \sup_{\gamma\in\mathbb{R}} F(\gamma,\beta,\lambda;X) =: \ell_{rob}(\beta,\lambda;X), \tag{24}$$

thus resulting in $\sup_{P:D_c(P_0,P)\leq\delta} E_P[\ell(\beta^T X)] = \inf_{\lambda\geq 0} E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$. 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 below, whose proof is furnished in the technical Appendix A.

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

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

We shall 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 (24) that $\ell_{rob}(\alpha\theta_1 + (1-\alpha)\theta_2; x)$ equals

$$\begin{aligned}
&\sup_{\Delta\in\mathbb{R}^d}\left\{\ell\left((\alpha\beta_1+(1-\alpha)\beta_2)^T(x+\Delta)\right) - (\alpha\lambda_1+(1-\alpha)\lambda_2)\left(\Delta^T A(x)\Delta-\delta\right)\right\}\\
&\quad = (\alpha\lambda_1+(1-\alpha)\lambda_2)\,\delta\\
&\qquad + \sup_{\Delta\in\mathbb{R}^d}\left\{\ell\left(\alpha\beta_1^T(x+\Delta)+(1-\alpha)(\beta_2^T(x+\Delta))\right) - (\alpha\lambda_1+(1-\alpha)\lambda_2)\,\Delta^T A(x)\Delta\right\}.
\end{aligned} \tag{25}$$

Since $\ell(\cdot)$ is convex, we have $\ell(\alpha u_1 + (1-\alpha)u_2) \leq \alpha\ell(u_1) + (1-\alpha)\ell(u_2)$ for $u_1, u_2 \in \mathbb{R}$. Combining this with the fact that $\sup_\Delta(\alpha f_1(\Delta) + (1-\alpha)f_2(\Delta)) \leq \alpha\sup_\Delta f_1(\Delta) + (1-\alpha)\sup_\Delta f_2(\Delta)$ for any two

functions f_1, f_2, we have that the term involving supremum in (25) 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 (25), establishes that $\ell_{rob}(\alpha\theta_1 + (1-\alpha)\theta_2; x) \leq \alpha\ell_{rob}(\theta_1; x) + (1-\alpha)\ell_{rob}(\theta_2; x)$, thus verifying the desired convexity of $\ell_{rob}(\,\cdot\,; x)$. □

Proof of Theorem 2. Since $f_\delta(\beta,\lambda) := E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$, the convexity of $f_\delta(\cdot)$ follows as a consequence of Lemma 2 and 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 3b) 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 $\beta \in B$. Lemma 4 - 6 below, whose proofs are provided in Appendix A, are useful towards 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 towards identifying the region $\mathbb{V}$ in the main results Proposition 1 and Theorem 3.

Lemma 4. *Suppose that Assumptions 1 - 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 - 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,*

$$|\gamma| \geq \frac{|\ell'(\beta^T x)|}{2\lambda}. \tag{26}$$

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

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

Proof of Lemma 7. **Lower bound.** Combining the observations in Lemma 4 - 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 1b, the above 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 $\ell''(\cdot) \leq M$ due to 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, since $\lambda_*(\beta)\sqrt{\delta} + E_{P_0}\left[\ell(\beta^T X)\right] \leq f_\delta(\beta, \lambda_*(\beta)) = \inf_{\lambda \geq 0} E_{P_0}[\ell_{rob}(\beta, \lambda; X)]$, we use the above result and the bounds in Assumption 1b to write,

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

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

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

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 $\overline{L}$ and $\underline{L}$ in Lemma 6, we have from Lemma 7 above that $\lambda_{\min}(\beta) \geq K_1\|\beta\|$ and $\lambda_{\max}(\beta) \leq K_2\|\beta\|$, where

$$K_1 := \frac{1}{2}\sqrt{\underline{L}\rho_{\max}^{-1}} \quad \text{and} \quad K_2 := \frac{1}{2}\sqrt{\delta}MR_\beta\rho_{\min}^{-1} + \sqrt{\rho_{\min}^{-1}\overline{L}}. \tag{27}$$

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 - 4. We accomplish this primarily by identifying the Hessian matrix of the dual DRO objective $f_\delta(\beta, \lambda) = E_{P_0}\left[\ell_{rob}(\beta, \lambda; X)\right]$.

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 envelope theorem. Likewise, if the loss $\ell(\cdot)$ is twice differentiable, implicit function theorem allows us to characterize the Hessian of $\ell_{rob}(\beta, \lambda; x)$. To accomplish this, define

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

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,

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

Then, as a consequence of 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

$$g(\beta,\lambda;x)\in\Gamma^*(\beta,\lambda;x) \quad \text{and} \quad \varphi(g(\beta,\lambda;x),\beta,\lambda,x)>0, \tag{28}$$

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

$$T_g(x) := x+\sqrt{\delta}g(\beta,\lambda;x)A(x)^{-1}\beta, \quad \bar{T}_g(x) := x+2\sqrt{\delta}g(\beta,\lambda;x)A(x)^{-1}\beta, \text{ and}$$
$$\varphi_g(\beta,\lambda;x) := \varphi\left(g(\beta,\lambda;x),\beta,\lambda;x\right), \tag{29}$$

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 $(\beta,\lambda;x)$ while writing the 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 (28). Then for almost every $x\in S_X$, $(\beta,\lambda)\in\operatorname{int}(\mathcal{U}(x))$, 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 (29). Moreover, we have

$$\nabla^2_\theta\ell_{rob}(\theta;x) - \Lambda(\theta;x)B(x) \succeq 0, \tag{30}$$

where

$$\Lambda(\beta,\lambda;x) := \frac{4\left(\beta^T T_g(x)\right)^2\ell''(\beta^T T_g(x))}{1+\bar{T}_g(x)^T A(x)\bar{T}_g(x)\ell''(\beta^T T_g(x))/(\sqrt{\delta}g^2\varphi)}\frac{1}{2\lambda\varphi_g+4\beta^T A(x)^{-1}\beta} \tag{31}$$

and

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

The proofs of Proposition 8 and Lemma 8 below are provided in the technical 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 $\gamma\mapsto F(\gamma,\beta,\lambda;x)$ 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 is available in the technical Appendix A.

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

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

where the constants $\rho_{\min}, \rho_{\max}$ are as in Assumption 1b, $\underline{L}, \overline{L}$ in Lemma 6, R_β in Assumption 4 and M in Assumption 3. Recall the definition of the constants K_1, K_2 in (27) and that of the previously defined sets,

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

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 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 (28) 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*

$$|g(\beta, \lambda; x)| \ge \frac{|\ell'(\beta^T x)|}{2K_2\|\beta\|} \text{ if } (\beta, \lambda) \in \mathbb{V}, \qquad |g(\beta, \lambda; x)| \le \frac{|\ell'(\beta^T x)|}{\varphi_{\min}\|\beta\|} \text{ if } (\beta, \lambda) \in \mathbb{W}. \tag{32}$$

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

The proofs of Theorem 3 - 4 provided next are reliant on the observations made in proposition 9 above.

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 case of matrix, or ℓ_2-norm in case of 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 $\partial^2 f_\delta/\partial\beta^2$, $\partial^2 f_\delta/\partial\beta\partial\lambda$, $\partial^2 f_\delta/\partial\lambda^2$ 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 $A(x)^{-1} \succeq \rho^{-1}_{\max}\mathbb{I}_d$ for P_0-almost every x (that is, $A(x)^{-1} - \rho^{-1}_{\max}\mathbb{I}_d$ is positive semidefinite). Next, recall from (26) in Lemma 5 and Lemma 6 that $|g(\beta, \lambda; x)| \ge |\ell'(\beta^T x)|/(2\lambda)$ 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 $(\beta, \lambda) \in \mathbb{V}$,

$$\frac{\partial^2 f_\delta}{\partial\beta^2}(\beta, \lambda) = E_{P_0}\left[\frac{\partial^2\ell_{rob}}{\partial\beta^2}(\beta, \lambda; X)\right] \succeq \sqrt{\delta}\frac{E_{P_0}[\ell'(\beta^T X)^2]}{2\lambda}\rho^{-1}_{\max}\mathbb{I}_d \succeq \sqrt{\delta}\frac{\kappa_0}{\lambda}\mathbb{I}_d,$$

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

In order the proceed with the proof of 4, define

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

Proof of Theorem 4. Using the bounds of $|g(\cdot)|$ and $\varphi_g(\cdot)$ from Proposition 9b along with other immediate bounds such as $\varphi_g \le 2\lambda, \lambda \in [K_1\|\beta\|, K_2\|\beta\|]$, and $\beta^T A(x)^{-1}\beta \le \rho^{-1}_{\min}\|\beta\|^2$, the expression for

$\Lambda(\beta,\lambda;x)$ from (31) simplifies to,

$$\Lambda(\beta,\lambda;x) = \frac{4\sqrt{\delta}(g\beta^T T_g(x))^2}{2\lambda\sqrt{\delta}g^2/\ell''(\beta^T T_g(x)) + \bar{T}_g(x)^T A(x)\bar{T}_g(x)} \cdot \frac{1}{2\lambda + 4\beta^T A(x)^{-1}\beta/\varphi_g} \tag{33}$$

$$\geq \frac{4\sqrt{\delta}(\beta^T T_g(x)\ell'(\beta^T x)/(2K_2\|\beta\|))^2}{2K_2\sqrt{\delta}\ell'(\beta^T x)^2/(\varphi_{\min}^2\|\beta\|\ell''(\beta^T T_g(x))) + \overline{T}_g(x)^T A(x)\overline{T}_g(x)} \cdot \frac{1}{2K_2\|\beta\| + 4\rho_{\min}^{-1}\|\beta\|^2/(\varphi_{\min}\|\beta\|)}$$

$$\geq \sqrt{\delta}C_0 \frac{\|\beta\|^{-2}\left(\beta^T T_g(x)\ell'(\beta^T x)\right)^2}{2K_2\sqrt{\delta}\varphi_{\min}^{-2}\ell'(\beta^T x)^2/\ell''(\beta^T T_g(x)) + \overline{T}_g(x)^T A(x)\overline{T}_g(x)\|\beta\|}, \tag{34}$$

where $C_0 := (2K_2 + 4\varphi_{\min}^{-1}\rho_{\min}^{-1})^{-1}$. Next, since $\beta^T T_g(x) = \beta^T x + \sqrt{\delta}g\beta^T A(x)^{-1}\beta$, we obtain from the bounds in (32) that

$$|\beta^T T_g(x)\ell'(\beta^T x)| \geq |\beta^T x\ell'(\beta^T x)| - \sqrt{\delta}|g\ell'(\beta^T x)|\beta^T A(x)^{-1}\beta$$

$$\geq |\beta^T x\ell'(\beta^T x)| - \sqrt{\delta}\frac{\ell'(\beta^T x)^2}{\varphi_{\min}\|\beta\|}\|\beta\|^2\rho_{\min}^{-1} \geq \left(c_1c_2 - \frac{4\sqrt{\delta}\overline{L}}{p\varphi_{\min}\rho_{\min}}\right)\|\beta\|, \tag{35}$$

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

$$A_1 := \{x : |\beta^T x\ell'(\beta^T x)| > c_1c_2\|\beta\|\} \quad \text{and} \quad A_2 := \{x : \ell'(\beta^T x)^2 \leq 4\overline{L}/p\},$$

where the constants c_1, c_2, p are given by Assumption 5. Since $E_{P_0}[\ell'(\beta^T X)^2] \leq \overline{L}$ 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 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 $\varphi_{\min} = \sqrt{\underline{L}}\rho_{\max}^{-1/2} - \sqrt{\delta}R_\beta M\rho_{\min}^{-1} \geq \frac{1}{2}\sqrt{\underline{L}}\rho_{\max}^{-1/2}$. Further, since $\delta < \delta_1 \leq c_1^2c_2^2p^2\rho_{\min}^2\rho_{\max}^{-1}\underline{L}\overline{L}^{-2}/256$, we have

$$c_1c_2 - 4\sqrt{\delta}\overline{L}p^{-1}\varphi_{\min}^{-1}\rho_{\min}^{-1} \geq c_1c_2/2. \tag{36}$$

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 (34) 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 (32) that

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

and similarly, $\|T_g(x)\| \leq C_2$ for $x \in A_2 \cap A_3$. Since $\|\beta^T T_g(x)\| \leq R_\beta C_2 < \infty$ when $x \in A_2 \cap A_3$, if we let $C_3 := \inf_{|u|\leq R_\beta C_2} \ell''(u) > 0$, we obtain that the denominator in (34) 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 (34),(35) and (36), we obtain that $\Lambda(x) \geq \sqrt{\delta}C\mathbf{1}_{\{x\in A_1\cap A_2\cap A_3\}}$ for $C := (1/2)C_0c_1c_2C_4^{-1}$.

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

Remark 4. Suppose that $c_1c_2 = 0$ is the only non-negative number for which the probability requirement in Assumption 5 is satisfied. In this case, we have from the upper bound for g in Proposition 9b that $g\beta^T X = 0$, P_0 almost surely. As a result, the numerator of $\Lambda(x)$ in the right hand side of (33) is bounded from above by $4\sqrt{\delta}(0 + \sqrt{\delta}g^2\beta^T A(x)^{-1}\beta)^2 \leq 4\delta^{3/2}\ell'(\beta^T x)^2\varphi_{\min}^{-2}\rho_{\min}^{-2}$, P_0-almost surely. Since 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 non-negative constant κ_3. Since $\delta^{3/2} = o(\sqrt{\delta})$ 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. **Verifying strong convexity in the absence of boundedness of set B.** In this section, we provide a proof of Theorem 5. For any map $g : \mathcal{U} \to \mathbb{R}$ satisfying (28), consider $\varphi_g(\cdot)$ in (29) and define the functions,

$$I_0(\beta,\lambda) := E\left[g^2(\beta,\lambda;x)\beta^T A(X)^{-1}\beta\right], \quad I_1(\beta,\lambda;x) := |g\beta^T T_g(x)|\frac{\varphi_g(\beta,\lambda;x)}{2\lambda},$$
$$I_2(\beta,\lambda;x) := \frac{\sqrt{\beta^T A(x)^{-1}\beta}}{2\lambda}, \qquad \text{and} \qquad I_3(\beta,\lambda;x) := \sqrt{\delta}2\lambda g^2 + \bar{T}_g^T A(x)\bar{T}_g(x)\ell''(\beta^T T_g(x)).$$

It follows from the definition of $\varphi_g(\cdot)$ that $\varphi_g/2\lambda \leq 1$. Then, for any $(\beta,\lambda,x) \in \mathcal{U}$ for which the Hessian $\nabla^2\ell_{rob}(\beta,\lambda;x)$ (computed with respect to variables β,λ) exists, we have from Proposition 8 that $\nabla^2\ell_{rob}(\beta,\lambda;x) - \Lambda(\beta,\lambda;x)B(x) \succeq 0$; here, $\Lambda(\beta,\lambda;x)$, defined as in (31), satisfies,

$$\frac{2\lambda}{\sqrt{\delta}}\Lambda(\beta,\lambda;x) \geq \frac{I_1^2(\beta,\lambda;x)\ell''(\beta^T T_g(x))}{I_3(\beta,\lambda;x)(1+4I_2^2(\beta,\lambda;x))}. \tag{37}$$

Before beginning the proof of Theorem 5, define,

$$\delta_2 := c_1^2 c_2^2 p\rho_{\min}^2\left(2k_1\rho_{\max}^{1/2} + 4c_1\rho_{\min}^{1/2}(1+k_2)\right)^{-2}.$$

Proof of Theorem 5. Fix any $\beta \in B$ and $\lambda_*(\beta) \in \arg\min_{\lambda\geq 0} f_\delta(\beta,\lambda)$. Then it follows from the first-order optimality condition that,

$$0 \leq \frac{\partial_+ f_\delta}{\partial\lambda}(\beta,\lambda_*(\beta)) \leq \sqrt{\delta}\left(1 - E_{P_0}\left[g^2(\beta,\lambda_*(\beta))\beta^T A(X)^{-1}\beta\right]\right)$$

(see the proof of Lemma 7 in the earlier Subsection 5.2 for a similar application of the first order optimality condition). Consequently, for a given $\rho > 0$, there exists $r_1 > 0$ such that

$$\left|E_{P_0}\left[g^2(\tilde{\beta},\lambda;x)\tilde{\beta}^T A(X)^{-1}\tilde{\beta}\right] - 1\right| \leq \rho \tag{38}$$

for all $(\tilde{\beta},\lambda) \in \mathcal{N}_{r_1}((\beta,\lambda_*(\beta)))$. This follows from the continuity properties of $\ell'(\cdot)$. Since $|g(\tilde{\beta},\lambda;x)| \geq \ell'(\tilde{\beta}^T x)/(2\lambda)$, we have from (38) and Assumption 5 that

$$2\lambda \geq \frac{E_{P_0}[\ell'(\tilde{\beta}^T X)^2]^{1/2}\rho_{\max}^{-1/2}\|\beta\|}{(1+\rho)^{1/2}} \geq c_1(p\rho_{\max}^{-1}/2)^{1/2}\|\beta\|,$$

for any choice of $\rho < 1$ and $(\tilde{\beta},\lambda) \in \mathcal{N}_{r_1}((\beta,\lambda_*(\beta)))$. Consequently, we have that

$$I_2(\tilde{\beta},\lambda;x) := \frac{\sqrt{\beta^T A(x)^{-1}\beta}}{2\lambda} \leq \frac{\rho_{\min}^{-1/2}\|\beta\|}{c_1(p\rho_{\max}^{-1}/2)^{1/2}\|\beta\|} = \frac{1}{c_1}\frac{(2\rho_{\max})^{1/2}}{(p\rho_{\min})^{1/2}}, \tag{39}$$

for any $(\tilde{\beta},\lambda) \in \mathcal{N}_{r_1}((\beta,\lambda_*(\beta)))$. Moreover, we have from the definition of $\varphi_g(\cdot)$ that,

$$I_1(\tilde{\beta},\lambda;x) \geq |g\beta^T T_g(x)| - \frac{\sqrt{\delta}}{2\lambda}|g|\beta^T A(x)^{-1}\beta|\beta^T T_g(x)|\ell''(\beta^T T_g(x)).$$

Since $|u|\ell''(u) \leq k_1 + k_2|\ell'(u)|$, for any $u \in \mathbb{R}$ (see the assumption in the statement of Theorem 5), we have that

$$\begin{aligned}
I_1(\tilde{\beta},\lambda;x) &\geq |g\tilde{\beta}^T T_g(x)| - \frac{\sqrt{\delta}}{2\lambda}|g|\tilde{\beta}^T A(x)^{-1}\tilde{\beta}\left(k_1 + k_2|\ell'(\tilde{\beta}^T T_g(x))|\right) \\
&= |g\tilde{\beta}^T T_g(x)| - \frac{\sqrt{\delta}}{2\lambda}|g|\tilde{\beta}^T A(x)^{-1}\tilde{\beta}\left(k_1 + 2\lambda k_2|g|\right) \\
&\geq |g\tilde{\beta}^T x| - \sqrt{\delta}g^2\tilde{\beta}^T A(x)^{-1}\tilde{\beta} - \frac{\sqrt{\delta}}{2\lambda}|g|\tilde{\beta}^T A(x)^{-1}\tilde{\beta}\left(k_1 + 2\lambda k_2|g|\right),
\end{aligned}$$

where the equality follows from the first-order optimality condition satisfied by $g(\cdot)$, and the last inequality is a simple consequence of triangle inequality applied to $\tilde{\beta}^T T_g(x) = \tilde{\beta}^T x + \sqrt{\delta} g \tilde{\beta}^T A(x)^{-1}\tilde{\beta}$. As a result,

$$\frac{I_1(\tilde{\beta},\lambda;x)}{\sqrt{g^2\tilde{\beta}^T A(x)^{-1}\tilde{\beta}}} \geq \frac{\beta^T x}{\sqrt{\beta^T A(x)^{-1}\beta}} - \sqrt{\delta}\left(k_1 I_2(\tilde{\beta},\lambda;x) + k_2\sqrt{g^2\beta^T A(x)^{-1}\beta}\right).$$

By applying the upper bound for $I_2(\cdot)$ derived in (39), we arrive at,

$$\frac{I_1(\tilde{\beta},\lambda;x)}{\sqrt{g^2\tilde{\beta}^T A(x)^{-1}\tilde{\beta}}} \geq \frac{\beta^T x}{\sqrt{\beta^T A(x)^{-1}\beta}} - \sqrt{\delta}\left(\frac{k_1}{c_1}\frac{(2\rho_{\max})^{1/2}}{(p\rho_{\min})^{1/2}} + k_2\sqrt{g^2\beta^T A(x)^{-1}\beta}\right) \tag{40}$$

Next, as in the proof of Theorem 4, we define the following subsets of $\mathbb{R}^d$: $A_1 := \{x : |\ell'(\tilde{\beta}^T x)| \geq c_1/2, |\tilde{\beta}^T x| \geq c_2\|\tilde{\beta}\|/2 \text{ for all } \tilde{\beta} \in \mathcal{N}_{r_2}(\beta)\}$, $A_2 := \{x : g^2(\tilde{\beta},\lambda;x)\beta^T A(x)^{-1}\beta \leq C_0 \text{ for all } (\tilde{\beta},\lambda) \in \mathcal{N}_{r_1}(\beta,\lambda_*(\beta))\}$, and $A_3 := \{x : \|x\| \leq C_1\}$, where the constants C_0, C_1, r_2, ρ are to be chosen imminently. Define $\underline{r} = r_1 \wedge r_2$ and take the constant C_1 large enough such that $P_0(A_3) \geq p/4$, where p is specified as in Assumption 5. Due to Markov's inequality and (38), we also have that $P_0(A_2) \geq 1 - (1+\rho)/C_0$. For any choice of $\rho < 1$, if we take $C_0 = 8/p$, then $P_0(A_2) \geq 3p/4$. Likewise, due to Assumption 5, we have $P(A_1) \geq p$ for a suitably small r_2. If we let $A := A_1 \cap A_2 \cap A_3$, then it follows from union bound that $P(A) \geq p/2$.

Moreover, we have from (40) that

$$\frac{I_1(\tilde{\beta},\lambda;x)}{\sqrt{g^2\tilde{\beta}^T A(x)^{-1}\tilde{\beta}}} \geq \left(c_2\rho_{\min}^{1/2} - \sqrt{\delta}\left(\frac{k_1}{c_1}\frac{(2\rho_{\max})^{1/2}}{(p\rho_{\min})^{1/2}} + (1+k_2)(8/p)^{1/2}\right)\right), \tag{41}$$

whenever $x \in A$ and $(\beta,\lambda) \in \mathcal{N}_{\underline{r}}(\beta,\lambda_*(\beta))$. For any fixed $\delta < \delta_2$, it follows from the definition of δ_2 that $I_1(\tilde{\beta},\lambda;x) > 0$ for all $x \in A$, and $(\tilde{\beta},\lambda) \in \mathcal{N}_{\underline{r}}(\beta,\lambda_*(\beta))$. Since $\varphi_g(\tilde{\beta},\lambda;x)$ is positive whenever $I_1(\tilde{\beta},\lambda;x)$ is positive, we have (from Proposition 8) that the Hessian $\nabla^2\ell_{rob}(\tilde{\beta},\lambda;x)$ exists and it satisfies,

$$\nabla^2\ell_{rob}(\tilde{\beta},\lambda;x) - \Lambda(\tilde{\beta},\lambda;x)B(x) \succeq 0,$$

for $x \in A$ and $(\tilde{\beta},\lambda) \in \mathcal{N}_{\underline{r}}(\beta,\lambda_*(\beta))$. Following the same reasoning as in the proof of Theorem 4, one can obtain upper bound C_2 for $\|\bar{T}_g(x)\|$. Moreover, due to (38) and the property that $2\lambda|g(\tilde{\beta},\lambda;x)| \geq |\ell'(\tilde{\beta}^T x)| \geq c_1$ for $x \in A$, we have that $g^2(\tilde{\beta},\lambda;x)\tilde{\beta}^T A(x)^{-1}\tilde{\beta}$ is bounded away from zero, for every $x \in A$. Utilizing these observations and the bounds for I_1, I_2 (see (41) and (39)) in the expression for $\Lambda(\cdot)$ in (37), we arrive at the following conclusion: For any $x \in A$, there exists $\kappa(x) > 0$ such that

$$\ell_{rob}(\alpha\theta_1 + (1-\alpha)\theta_2;x) \leq \ell_{rob}(\theta_1;x) + (1-\alpha)\ell_{rob}(\theta_2;x) - \frac{\kappa(x)}{2}\alpha(1-\alpha)\|\theta_1-\theta_2\|^2,$$

for $\theta_1,\theta_2 \in \mathcal{N}_{\underline{r}}(\beta,\lambda_*(\beta))$. Since $f_\delta(\theta) := E_{P_0}\left[\ell_{rob}(\beta,\theta;X)\right]$, taking expectations on both sides, we arrive at the conclusion that

$$f_\delta(\alpha\theta_1 + (1-\alpha)\theta_2;x) \leq f_\delta(\theta_1;x) + (1-\alpha)f_\delta(\theta_2;x) - E\left[\frac{\kappa(X)}{2}\mathbb{I}(X \in A)\right]\alpha(1-\alpha)\|\theta_1-\theta_2\|^2,$$

for all $\theta_1,\theta_2 \in \mathcal{N}_{\underline{r}}(\beta,\lambda_*(\beta))$. With $P_0(A) \geq p/2$ being positive, we have that the constant $\kappa := E[\kappa(X)\mathbb{I}(X \in A)]$ is positive as well. This concludes the proof of Theorem 5. □

5.5. **Proofs of the results pertaining to the structure of the worst case distribution.** In this section we provide proofs of Theorem 6 and Proposition 7 which shed light on the structure of the adversarial distribution(s) attaining the supremum in $\sup_{P:D_c(P_0,P)\leq\delta} E_P[\ell(\beta^T X)]$.

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 $\lambda \geq 0$ 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 $\arg\min_{\lambda\geq 0} f_\delta(\beta,\lambda)$ for any $\beta \in B$.

a) When $\lambda_*(\beta) = 0$: We have $\inf_{\beta,\lambda} f_\delta(\beta,\lambda) = f_\delta(\beta,0) = \sup_{u\in\mathbb{R}} \ell(u)$. Due to 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 $D_c(P,P_0) \leq \delta$ is a worst case distribution attaining the supremum in $\sup_{P:D_c(P,P_0)\leq\delta} E_{P_0}[\ell(\beta^T X)]$.

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, $\lambda_*(\beta) \geq \lambda_{thr}(\beta)$.

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 [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 $(\beta,\lambda) \in \mathbb{U}_1$. Since $\lambda_*(\beta) > \lambda_{thr}(\beta)$, we have from Lemma 3a 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 [3, Proposition 7.50b]). Letting $g_+(\beta,\lambda_*(\beta);X) = G_+$ and $g_-(\beta,\lambda_*(\beta);X) = G_-$, we obtain that,

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

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

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

As the complementary slackness conditions in Theorem 1 of [6] are satisfied, we have that the distribution of X^* attains the supremum in $\sup_{P:D_c(P,P_0)\leq\delta} E_P[\ell(\beta^T X)]$.

d) When $\lambda_*(\beta) = \lambda_{thr}(\beta)$: The worst case distribution $P^*(\beta)$ attaining the supremum in $\sup_{P:D_c(P,P_0)\leq\delta} E_P[\ell(\beta^T X)]$ may not exist as demonstrated in the following example. Suppose that $\ell(u) := u^2 - |u|(1-e^{-|u|})$, $\|\beta\| = 1$, $P_0(dx) = \delta_{\{\mathbf{0}\}}(dx)$, $\delta > 0$ 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 $\lambda \geq \lambda_{thr}(\beta) = \sqrt{\delta}$, we have $\Gamma^*(\beta,\lambda;\mathbf{0}) = \{\mathbf{0}\}$, and it follows that $f_\delta(\beta,\lambda) = \lambda\sqrt{\delta}$ when $\lambda \geq \lambda_{thr}(\beta)$. 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 for any P satisfying $D_c(P,P_0) \leq \delta$. This is because, we have $E\|X\|^2 \leq \delta$ for any P such that $D_c(P,P_0) \leq \delta$, and as a result we have $E_P[\ell(\beta^T X)] < \delta$ as in the following series of inequalities:

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

e) When $\lambda_*(\beta) > \lambda'_{thr}(\beta)$: In this case, it follows from Lemma 8 that the map $\gamma \mapsto F(\gamma,\beta,\lambda_*(\beta);x)$ 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 above described uniqueness

in optimizer means that $X^* = X + \sqrt{\delta} G A(X)^{-1}\beta$ 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. Since any distribution $\bar{P}$ attaining the supremum in $\sup_{P:D_c(P,P_0)\leq\delta} E_P[\ell(\beta^T X)]$ 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 $\sup_{P:D_c(P,P_0)\leq\delta} E_P[\ell(\beta^T X)]$. □

Proof of Theorem 7. Since $\beta \in B$ 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 $\min_{\lambda\geq 0} f_\delta(\beta, \lambda)$ 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 $\sup_{P:D_c(P,P_0)\leq\delta_i} E_P[\ell(\beta^T X)]$, 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 appendix Section 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 9b).

Next, define $H(\delta, \lambda) := E_{P_0}[g(\delta, \lambda; X)^2\beta^T A(X)^{-1}\beta] - 1$. Since $\lambda_*(\delta)$ 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 above 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 (52)), 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 9b). Moreover, $\beta^T A(X)^{-1}\beta \leq R_\beta \rho^{-1}_{\min}\|\beta\|$ and $\ell''(\cdot) \in (0, M]$ (see Assumptions 1 - 3). As a result, we obtain that

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

where the last equality follows from the definitions of δ_0 and $\varphi_{\min}$ in the earlier Subsection 5.3. Since $\delta < \delta_1 \leq \delta_0/4$, 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 (50) 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 $\beta^T X^*_\delta$. Since $2\lambda_*(\delta) - \sqrt{\delta}\beta^T A(X)^{-1}\beta\ell''(\eta) \geq \varphi_{\min}\|\beta\| > 0$, 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 7b - 7d are verified. This completes the proof of Theorem 7. □

5.6. **Proofs of the results on rates of convergence.** Lemma 9 below, establishing finite second moments for the gradients (or) subgradients utilized in SGD schemes, is useful towards 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, 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 Appendix A.

Proof of Proposition 4. a) When $\delta < \delta_0$, it follows from Proposition 2 and Proposition 3 that the subgradient set $\partial\ell_{rob}(\beta, \lambda; X) = \{\nabla_\theta \ell_{rob}(\beta, \lambda; X)\}$, P_0-almost surely. Since $\lambda > \lambda'_{thr}(\beta) \geq \lambda_{thr}(\beta)$ for every $(\beta, \lambda) \in \mathbb{W}$ (see Proposition 9a), 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 following Theorem 4 in [31] 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 4a now follows as a consequence of Markov's inequality.

b) When $\delta < \delta_1$, it follows from the positive definiteness of 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, due to 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 due to 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,

$$\|\nabla_\theta f_\delta(\theta) - \nabla^2_\theta f_\delta(\theta_*)^T(\theta - \theta_*)\| = o\left(\|\theta - \theta_*\|\right), \tag{42}$$

for $\theta \in \mathbb{W}$. With these conditions being satisfied, it follows from [27, Theorem 2] that

$$\sqrt{k}(\bar{\theta}_k - \theta_*) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \Sigma),$$

as $k \to \infty$, where $\Sigma := (\nabla^2_\theta f_\delta(\theta_*))^{-1}\mathrm{Cov}[\nabla_\theta \ell_{rob}(\theta_*; X)]((\nabla^2_\theta f_\delta(\theta_*))^{-1})^T$. If we let $Z \sim \mathcal{N}(0, \mathbb{I}_{d+1})$, then due to continuous mapping theorem, we have that the distribution of $k(\bar{\theta}_k - \theta_*)^T \nabla^2_\theta f_\delta(\theta_*)(\bar{\theta}_k - \theta_*)$ is convergent to that of

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

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

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

Now it follows from the local joint strong convexity of $f_\delta(\cdot)$ in Theorem 4 and (42) that

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

This completes the proof of Proposition 4. □

Proof of Proposition 5. We use ε_k to denote the estimation error of $\nabla_\theta \ell_{rob}(\theta_{k-1}; X_k)$ induced by line search. Due to Lemma 11, there exists a constant $C > 0$ such that the estimation error can be controlled as $\|\varepsilon_k\| \leq C\alpha_k$, if we apply bisection method for at least $\log_2(\alpha_k^{-1}(1 + \|X_k\|)^2)$ steps.

We first show that $\|\theta_k - \theta_*\| \to 0$ with probability one. The update step is

$$\theta_k = \Pi_{\mathbb{W}}\big(\theta_{k-1} - \alpha_k \nabla_\theta \ell_{rob}(\theta_{k-1}; X_k) - \alpha_k\varepsilon_k\big). \tag{43}$$

Since the following conditions are satisfied:

i) $\sup_{\theta\in\mathbb{W}} E\|\nabla_\theta \ell_{rob}(\theta; X_k) + \varepsilon_k\|^2 < \infty$, due to Lemma 9 and the boundedness of ε_k;

ii) $\theta \mapsto \nabla_\theta f_\delta(\theta) = E[\nabla_\theta \ell_{rob}(\theta; X)]$ is continuous ;
iii) $\sum \alpha_k^2 < \infty$ and $\sum \alpha_k \|\varepsilon_k\| < \infty$

Applying [18, Theorem 5.2.1], we conclude that $\|\theta_k - \theta_*\| \to 0$ with probability one.

Now we show that $f_\delta(\theta_k) - f_* = O_p(\alpha_k)$. In view of the second order differentiability of f_δ at θ_*, it is sufficient to show that $\|\theta_k - \theta_*\| = O_p(\sqrt{\alpha_k})$, i.e., the sequence $\{\|\theta_k - \theta_*\|/\sqrt{\alpha_k}\}$ is tight. To this end, we shall slightly modify the proof of [18, Theorem 10.4.1] by allowing a small bias term ε_k appears in each update step, as shown in (43). As discussed in the proof of [18, Theorem 10.4.1], since $\|\theta_k - \theta_*\| \to 0$ with probability one, given any small $\nu > 0$, there is an $N_{\nu,\rho}$ such that $\|\theta_k - \theta_*\| \le \rho$ for $k \ge N_{\nu,\rho}$ with probability larger than $1 - \nu$. By shifting the time origin by $N_{\nu,\rho}$, we can suppose that $\|\theta_k - \theta_*\| \le \rho$. If for every $\nu > 0$ the time-shifted sequence is shown to be tight, then the original sequence is tight. Thus for the purposes of the tightness proof, it can be supposed without loss of generality that $\|\theta_k - \theta_*\| \le \rho$ for all k for the original process, where $\rho > 0$ is arbitrarily small.

To analyze the convergence rate of $\{\theta_k\}$, we define the Lyapunov function $V(\theta) = \|\theta - \theta_*\|^2$. Since the region $\mathbb{W}$ is convex, we have

$$V(\theta_k) \le \|\theta_{k-1} - \theta_* - \alpha_k \nabla_\theta \ell_{rob}(\theta_{k-1}; X_k) - \alpha_k \varepsilon_k\|^2$$

Notice that $\{\theta_k\}$ is a Markov Chain, by taking conditional expectation on both side we have the following inequalities hold with probability one,

$$\begin{aligned} E[V(\theta_k)|\theta_{k-1}] &\le V(\theta_{k-1}) - 2\alpha_k(\theta_{k-1} - \theta_*)^T \nabla_\theta f_\delta(\theta_{k-1}) \\ &\quad - \alpha_k^2 E\|\nabla_\theta \ell_{rob}(\theta_{k-1}; X_k) + \varepsilon_k\|^2 - 2\alpha_k(\theta_{k-1} - \theta_*)^T \varepsilon_k \\ &\le V(\theta_{k-1}) - 2\alpha_k(\theta_{k-1} - \theta_*)^T \nabla_\theta f_\delta(\theta_{k-1}) + O(\alpha_k^2), \end{aligned}$$

where the second inequality is due to $\|\theta_k - \theta_*\| \le \rho$, $\|\varepsilon_k\| \le C\alpha_k$, and $\sup_{\theta \in \mathbb{W}} E\|\nabla_\theta \ell_{rob}(\theta; X_k) + \varepsilon_k\|^2 < \infty$. Since $\rho > 0$ can be arbitrarily small, applying Taylor expansion we have

$$(\theta_{k-1} - \theta_*)^T \nabla_\theta f_\delta(\theta_{k-1}) = (\theta_{k-1} - \theta_*)^T \nabla_\theta^2 f_\delta(\theta_*)(\theta_{k-1} - \theta_*) + o(1)V(\theta_{k-1}).$$

Thus for any $\rho > 0$ that is smaller than the smallest eigenvalue of $\nabla_\theta^2 f_\delta(\theta_*)$, we have

$$E[V(\theta_k)|\theta_{k-1}] - V(\theta_{k-1}) \le -\rho\alpha_k V(\theta_{k-1}) + O(\alpha_k^2).$$

Recall that $\alpha_k = \alpha k^{-\tau}$ as required by Assumption 6, if either of the following is true (i) $\tau = 1$ and $\rho\alpha > 1$; (ii) $\tau \in [1/2, 1)$; then we have $E\|\theta_k - \theta_*\|^2 = E\|V(\theta_k)\| = O(\alpha_k)$ [18, Proof of Theorem 10.4.1], which implies that $\|\theta_k - \theta_*\| = O_p(\sqrt{\alpha_k})$ and $f_\delta(\theta_k) - f_* = O_p(\alpha_k)$. □

Proof of Proposition 7. Due to the characterization of subgradients $\partial f_\delta(\beta, \lambda)$ in Proposition 6, we have that $\partial_+ f_\delta / \partial\lambda(\beta, \lambda) \le \sqrt{\delta}$ for every $(\beta, \lambda) \in \mathbb{U}$. Recalling the definitions of $\mathbb{U}_\eta$ in (17) and $\mathbb{U}$ in (11), the above reasoning leads to concluding that, $f_\delta(\beta, \lambda + \varepsilon) - f_\delta(\beta, \lambda) \le \varepsilon\sqrt{\delta}$ for any $\varepsilon > 0$. Let $(\beta_*, \lambda_*) \in \inf_{(\beta,\lambda)\in\mathbb{U}} f_\delta(\beta, \lambda)$. Then

$$\inf_{\theta \in \mathbb{U}_\eta} f_\delta(\theta) - f_* \le f_\delta(\beta_*, \lambda_* + \eta) - f_\delta(\beta_*, \lambda_*) \le \eta\sqrt{\delta} \tag{44}$$

It follows from Lemma 9 that $\sup_k E_{P_0}\|H_k\|^2 < G_\eta$. As a consequence, we have from Theorem 2 and the remark following Theorem 4 in [31] that $E[f_\delta(\theta_k)] - \inf_{\theta\in\mathbb{U}_\eta} f_\delta(\theta) = O_p(k^{-1/2} \log k)$ and $E[f_\delta(\bar{\theta}_k)] - \inf_{\theta\in\mathbb{U}_\eta} f_\delta(\theta) = O_p(k^{-1/2})$. Combining this with the observation in (44), we obtain that $E[f_\delta(\bar{\theta}_k)] - f_* \le \eta\sqrt{\delta} + O_p(k^{-1/2})$. As in the proof of Proposition 4a, the conclusion in Proposition 7 follows as a consequence of Markov's inequality. □

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 stem from the observations that i) a flexible choice of the

distributional uncertainty region is useful towards to fully exploiting the advantages of DRO in data-driven contexts, and that 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 counter-part (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. While such extensions may render the inner maximization in (6) as a multi-dimensional 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 above 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. While examples of how one may choose the transportation cost function in a data-driven way are available in existing literature (see, for example, [5]), systemic treatment of the contextual choice of transportation cost is an essential question for future research.

Appendix A. Proofs of technical results.

The proofs of technical results in this section are presented in a logical order determined by their dependence on earlier proved results (rather than being based on the order in which they appear in the paper).

Proof of Lemma 3. a) Given $\varepsilon > 0$, it follows from the growth condition in Assumption 2 that there exist a positive constant C_ε satisfying $\ell(u) \le (\kappa+\varepsilon/2)u^2 + C_\varepsilon$ for all $u \in \mathbb{R}$. Since any $g \in \Gamma^*(\beta,\lambda;x)$ is a maximizer of $F(\cdot,\beta,\lambda;x)$, it follows immediately that $F(g,\beta,\lambda;x) \ge F(0,\beta,\lambda;x)$. Recalling the definition of $F(\cdot,\beta,\lambda;x)$ from (7), the above inequality results in,

$$(\kappa+\varepsilon/2)\left(\beta^T x + g\sqrt{\delta}\beta^T A(x)^{-1}\beta\right)^2 + C_\varepsilon - (\kappa+\varepsilon)\delta\left(\beta^T A(X)^{-1}\beta g\right)^2 \ge \ell(\beta^T x),$$

once we utilize that $\lambda \ge (\kappa+\varepsilon)\sqrt{\delta}\beta^T A(x)^{-1}\beta$ and $\ell(u) \le (\kappa+\varepsilon/2)u^2 + C_\varepsilon$. The above inequality can be equivalently written after a few basic algebraic steps as,

$$\left(g\sqrt{\delta}\beta^T A(x)^{-1}\beta - \frac{2\kappa+\varepsilon}{\varepsilon}\beta^T x\right)^2 \le \frac{2}{\varepsilon}C_\varepsilon - \frac{2}{\varepsilon}\ell(\beta^T x) + \frac{(\beta^T x)^2}{\varepsilon^2}(2\varepsilon^2 + 6\kappa\varepsilon + 4\kappa^2).$$

We first upper bound the right hand side by using $\ell(\beta^T x) \ge \ell(0) + \beta^T x \ell'(0)$, which holds due to the convexity of $\ell(\cdot)$. Next, utilizing the inequality $|a-b| \ge ||a|-|b||$ in the left hand side, we arrive at,

$$\sqrt{\delta}|g|\beta^T A(x)^{-1}\beta \le \sqrt{\frac{2}{\varepsilon}\left(C_\varepsilon + |\ell(0)| + |\ell'(0)||\beta^T x|\right)} + 4\frac{\kappa+\varepsilon}{\varepsilon}|\beta^T x|.$$

Since $\sqrt{x} \le 1 + x$ for $x \ge 0$, the above inequality verifies Part a) of Lemma 3.

b) Utilizing the bounds $\lambda \ge (\kappa+\varepsilon)\sqrt{\delta}\beta^T A(x)^{-1}\beta$ and $\ell(u) \le (\kappa+\varepsilon/2)u^2 + C_\varepsilon$ in the expression for $F(\cdot)$ in (7), we obtain that

$$F(g,\beta,\lambda;x) \le \lambda\sqrt{\delta} + C_\varepsilon + (\kappa+\varepsilon/2)(\beta^T x)^2 + 2(\kappa+\varepsilon/2)|\beta^T x|\sqrt{\delta}|g|\beta^T A(x)^{-1}\beta.$$

Since $\ell_{rob}(\beta,\lambda;x) = F(g,\beta,\lambda;x)$ for $g \in \Gamma(\beta,\lambda;x)$, we obtain the following bound for $\ell_{rob}(\beta,\lambda;x)$ once we substitute the bound for $\sqrt{\delta}|g|\beta^T A(x)^{-1}\beta$ from Part a):

$$\ell_{rob}(\beta,\lambda;x) \le \lambda\sqrt{\delta} + C_\varepsilon + (2\kappa+\varepsilon)|\beta^T x|\left(1+|\beta^T x|\right)\left(1+C_1\varepsilon^{-1}\right).$$

This verifies Part b) of Lemma 3. □

Proof of Lemma 1. For any fixed β,λ and x, it follows from the growth condition in Assumption 2 that i) $\lim_{\gamma\to\pm\infty} F(\gamma,\beta,\lambda;x) = -\infty$ if $\lambda > \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta$ and ii) $\lim_{\gamma\to\pm\infty} F(\gamma,\beta,\lambda;x) = +\infty$ if $\lambda < \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta$. Further, $F(\gamma,\beta,\lambda,x)$ is continuous in γ because of the continuity of $\ell(\cdot)$. Therefore we obtain that $\Gamma^*(\beta,\lambda;x) \neq \varnothing$ when $\lambda > \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta$. Likewise, $\Gamma^*(\beta,\lambda;x) = \varnothing$ when $\lambda < \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta$. This completes the proof of Parts a) and b) of Lemma 1.

To verify the inclusions in the final statement of Lemma 1 we proceed as follows: Whenever $\lambda < \lambda_{thr}(\beta)$ we have $\ell_{rob}(\beta,\lambda;x) = +\infty$ with positive probability. Therefore, $\mathbb{U}$ is contained in $\{(\beta,\lambda) : \beta \in B, \lambda \ge \lambda_{thr}(\beta)\}$. On the other hand, if $\lambda > \lambda_{thr}(\beta)$, we have $\lambda \ge (\kappa+\varepsilon)\sqrt{\delta}\beta^T A(x)^{-1}\beta$ for some $\varepsilon > 0$, P_0-almost every x. Since $\|\beta\| \le R_\beta$ and $E\|X\|^2 < \infty$, it follows Lemma 3b that $f_\delta(\beta,\lambda) = E_{P_0}[\ell_{rob}(\beta,\lambda;X)] < \infty$. Therefore $\{(\beta,\lambda) : \beta \in B, \lambda > \lambda_{thr}(\beta)\}$ is contained in $\mathbb{U}$. This completes the proof of Lemma 1. □

Proof of Proposition 2. a) Since $\lambda > \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta$ for P_0-almost surely every x, Proposition 2 follows directly from Lemma 1.
b) Consider any fixed $x \in \mathbb{R}^d, C_3 < \infty$ and $\eta > 0$. Define the set $A := \{(\beta,\lambda) \in B \times \mathbb{R}_+ : \|\beta\| < C_3,\ \lambda \ge \lambda_{thr}(\beta) + \eta\}$. Then for $(\beta,\lambda) \in A$, we have the following two conditions satisfied: i) $\lambda \ge (\kappa+\varepsilon)\sqrt{\delta}\beta^T A(x)^{-1}\beta$ for some $\varepsilon > 0$, for P_0-almost every x; and ii) $\beta^T A(x)^{-1}\beta$ is bounded away from zero if $\beta \neq \mathbf{0}$ (due to Assumption 1). Therefore, for any (β,λ) in A, we have from Lemma 3a that there exists a positive constant C_x such that $\Gamma^*(\beta,\lambda;x) \subseteq [-C_x, C_x]$. Thus for $(\beta,\lambda) \in A$, it suffices to restrict the univariate optimization problem (7) within the compact set $[-C_x, C_x]$, as in, $f_\delta(\beta,\lambda;x) = \sup_{\gamma\in[-C_x,C_x]} F(\gamma,\beta,\lambda;x)$.

Next, for $i = 1,\dots,K$, define

$$G_x(i,\gamma,\beta,\lambda) := \ell_i\left(\beta^T x + \gamma\sqrt{\delta}\beta^T A(x)^{-1}\beta\right) - \lambda\sqrt{\delta}\big(\gamma^2\beta^T A(x)^{-1}\beta - 1\big).$$

Then see that $F(\gamma,\beta,\lambda;x) = \max_{i=1,\dots,K} G_x(i,\gamma,\beta,\lambda)$ and

$$f_\delta(\beta,\lambda;x) = \sup_{\substack{i\in\{1,\dots,k\},\\ \gamma\in[-C_x,C_x]}} G_x(i,\gamma,\beta,\lambda).$$

Considering discrete topology for the variable i, see that $G_x(i,\gamma,\beta,\lambda)$ is upper semicontinuous in (γ,i) and continuously differentiable over variables (β,λ). Specifically, for $i \in \{1,\dots,K\}$, $j \in \{1,\dots,d\}$ and $\gamma^* \in \arg\max_x G_x(i,\gamma^*,\beta,\lambda)$, we have from the first-order optimality condition $(\ell_i'(\beta^T x+\gamma^*\sqrt{\delta}\beta^T A(x)^{-1}\beta) - 2\lambda\gamma^* = 0)$ that

$$\frac{\partial G_x}{\partial \beta_j}(i,\gamma^*,\beta,\lambda) = \ell_i'(\beta^T x + \gamma^*\sqrt{\delta}\beta^T A(x)^{-1}\beta)(x + \gamma^*\sqrt{\delta}A(x)^{-1}\beta)_j, \text{ and}$$
$$\frac{\partial G_x}{\partial \lambda}(i,\gamma^*,\beta,\lambda) = \sqrt{\delta}\big(1 - \gamma^{*2}\beta^T A(x)^{-1}\beta\big),$$

where $(x + \gamma^*\sqrt{\delta}A(x)^{-1}\beta)_j$ is the jth element of the vector $x + \gamma^*\sqrt{\delta}A(x)^{-1}\beta$. Moreover, since $\ell(u) := \max_{i=1,\dots,K} \ell_i(u)$, we have that

$$\frac{\partial_+\ell}{\partial u}(u) = \max_{i\,\in\,\mathrm{argmax}_j \ell_j(u)} \ell_i'(u) \quad \text{and} \quad \frac{\partial_-\ell}{\partial u}(u) = \min_{i\,\in\,\mathrm{argmax}_j \ell_j(u)} \ell_i'(u),$$

for any $u \in \mathbb{R}$. Equipped with these observations and the fact that $\ell_{rob}(\beta,\lambda;x) = \sup_{\gamma,i} G_x(\gamma,i,\beta,\lambda)$, we arrive at the following conclusions (i) and (ii) below as a consequence of Envelope theorem [20, Corollary 4]: i) When $(\beta,\lambda) \in A$, the functions $\lambda \mapsto \ell_{rob}(\beta,\lambda;x)$, $\beta_j \mapsto \ell_{rob}(\beta,\lambda;x)$ are absolutely continuous, and

have left and right derivative given by (12a) - (12d). Indeed, as an example for deriving $\partial_+\ell_{rob}/\partial\beta_j$, see that,

$$\begin{aligned}\frac{\partial_+\ell_{rob}}{\partial\beta_j}(\beta,\lambda;x) &= \max\left\{\frac{\partial G_x}{\partial\beta_j}(i^*,\gamma^*,\beta,\lambda) : (i^*,\gamma^*)\in\arg\max_{i,\gamma} G_x(i,\gamma,\beta,\lambda)\right\}\\ &= \max\left\{\frac{\partial G_x}{\partial\beta_j}(i^*,\gamma^*,\beta,\lambda) : \gamma^*\in\Gamma^*(\beta,\lambda;x),\ i^*\in\arg\max_{i=1,..,K} G_x(i,\gamma^*,\beta,\lambda)\right\}\\ &= \max_{\gamma\in\Gamma^*(\beta,\lambda;x)}\left\{\max_{i\in\arg\max_j \ell_j\left(\beta^Tx+\gamma\sqrt{\delta}\beta^TA(x)^{-1}\beta\right)} \ell_i'(\beta^Tx+\gamma\sqrt{\delta}\beta^TA(x)^{-1}\beta)(x+\gamma\sqrt{\delta}A(x)^{-1}\beta)_j\right\}\\ &= \max_{\gamma\in\Gamma^*(\beta,\lambda;x)}\frac{\partial_+\ell}{\partial u}\left(\beta^Tx+\gamma\sqrt{\delta}\beta^TA(x)^{-1}\beta\right)(x+\gamma\sqrt{\delta}A(x)^{-1}\beta)_j.\end{aligned}$$

Likewise, $\frac{\partial_-\ell_{rob}}{\partial\beta_j}(\beta,\lambda;x)$ can be seen equal to,

$$\begin{aligned}&\min_{\gamma\in\Gamma^*(\beta,\lambda;x)}\left\{\min_{i\in\arg\max_j \ell_j\left(\beta^Tx+\gamma\sqrt{\delta}\beta^TA(x)^{-1}\beta\right)} \ell_i'(\beta^Tx+\gamma\sqrt{\delta}\beta^TA(x)^{-1}\beta)(x+\gamma\sqrt{\delta}A(x)^{-1}\beta)_j\right\}\\ &\quad= \max_{\gamma\in\Gamma^*(\beta,\lambda;x)}\frac{\partial_-\ell}{\partial u}\left(\beta^Tx+\gamma\sqrt{\delta}\beta^TA(x)^{-1}\beta\right)(x+\gamma\sqrt{\delta}A(x)^{-1}\beta)_j.\end{aligned}$$

The directional derivatives with respect to the variable λ can be derived similarly. ii) Then we have that the partial derivatives exist as in (13) if and only if the respective sets,

$$\left\{\frac{\partial G_x}{\partial\beta_j}(i^*,\gamma^*,\beta,\lambda) : (i^*,\gamma^*)\in\arg\max_{i,\gamma} G_x(i,\gamma,\beta,\lambda)\right\},$$
$$\left\{\frac{\partial G_x}{\partial\lambda}(i^*,\gamma^*,\beta,\lambda) : (i^*,\gamma^*)\in\arg\max_{i,\gamma} G_x(i,\gamma,\beta,\lambda)\right\}$$

are singleton; this condition of being singleton is satisfied if and only if the respective sets,

$$\left\{\frac{\partial_+F}{\partial\beta_j}(\gamma,\beta,\lambda;x), \frac{\partial_-F}{\partial\beta_j}(\gamma,\beta,\lambda;x) : \gamma\in\Gamma^*(\beta,\lambda;x)\right\},$$
$$\left\{\frac{\partial_+F}{\partial\lambda}(\gamma,\beta,\lambda;x), \frac{\partial_-F}{\partial\lambda}(\gamma,\beta,\lambda;x) : \gamma\in\Gamma^*(\beta,\lambda;x)\right\}$$

are singleton. Since these expressions hold for any $C_3,\eta\in(0,\infty)$, Proposition 2 stands verified. □

The following technical result is useful towards proving Lemma 4.

Lemma 10. *Suppose that Assumptions 1,2 hold and $\ell(\cdot)$ is continuously differentiable. Then for fixed $\beta\in B$ and $x\in\mathbb{R}$, the map $\lambda\mapsto\ell_{rob}(\beta,\lambda;x)$ is right-continuous at $\lambda=\lambda_{thr}(\beta)$ if $\ell_{rob}(\beta,\lambda_{thr}(\beta);x)<\infty$.*

Proof of Lemma 10. Suppose that $\ell_{rob}(\beta,\kappa\sqrt{\delta}\beta^TA(x)^{-1}\beta;x)<\infty$. Then for any $\varepsilon>0$, there exist $\gamma\in\mathbb{R}$ such that $F(\gamma,\beta,\kappa\sqrt{\delta}\beta^TA(x)^{-1}\beta;x)>\ell_{rob}(\beta,\kappa\sqrt{\delta}\beta^TA(x)^{-1}\beta;x)-\varepsilon$. Thanks to the continuity of $F(\gamma,\beta,\lambda;x)$ with respect to λ,

$$\begin{aligned}\liminf_{\lambda\downarrow(\kappa\sqrt{\delta}\beta^TA(x)^{-1}\beta)}\ell_{rob}(\beta,\lambda;x) &\geq \lim_{\lambda\downarrow(\kappa\sqrt{\delta}\beta^TA(x)^{-1}\beta)}F(\gamma,\beta,\lambda;x)\\ &= F(\gamma,\beta,\kappa\sqrt{\delta}\beta^TA(x)^{-1}\beta;x)>\ell_{rob}(\beta,\kappa\sqrt{\delta}\beta^TA(x)^{-1}\beta;x)-\varepsilon.\end{aligned}$$

Since $\varepsilon>0$ is arbitrary, we have

$$\liminf_{\lambda\downarrow(\kappa\sqrt{\delta}\beta^TA(x)^{-1}\beta)}\ell_{rob}(\beta,\lambda;x)\geq\ell_{rob}(\beta,\kappa\sqrt{\delta}\beta^TA(x)^{-1}\beta;x).$$

Moreover, as $\ell_{rob}(\beta,\lambda;x)-\lambda\sqrt{\delta}$ is decreasing in λ, we also have that,

$$\limsup_{\lambda\downarrow(\kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta)} \ell_{rob}(\beta,\lambda;x)-\lambda\sqrt{\delta}\le \ell_{rob}(\beta,\kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta;x)-\kappa\delta\beta^T A(x)^{-1}\beta,$$

thus yielding, $\limsup_{\lambda\downarrow(\kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta)} \ell_{rob}(\beta,\lambda;x)\le \ell_{rob}(\beta,\kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta;x)$, and consequently,

$$\lim_{\lambda\downarrow(\kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta)} \ell_{rob}(\beta,\lambda;x)= \ell_{rob}(\beta,\kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta;x).$$

In addition, $\ell_{rob}(\beta,\lambda;x)$ is continuous at λ even if $\lambda>\kappa\sqrt{\delta}\beta A(x)^{-1}\beta$ (due to the convexity of $\ell_{rob}(\cdot;x)$ as in Lemma 2). Therefore, $\lambda\mapsto\ell_{rob}(\beta,\lambda;x)$ is right-continuous at $\lambda=\lambda_{thr}(\beta)$ if $\ell_{rob}(\beta,\lambda_{thr}(\beta);x)<\infty$. □

Proof of Lemma 4. Fix any $\beta\in B$. It follows from the characterization of $\mathbb{U}$ in Lemma 1 that $f_\delta(\beta,\lambda)=+\infty$ if $\lambda<\lambda_{thr}(\beta)$ and $f_\delta(\beta,\lambda)<+\infty$ if $\lambda>\lambda_{thr}(\beta)$. Therefore, it is necessary that $f_\delta(\beta,\lambda_*(\beta))$ is finite and $\lambda_*(\beta)\ge\lambda_{thr}(\beta)$.

Case 1. Suppose that $\lambda_*(\beta)>\lambda_{thr}(\beta)$. In this case we have from Lemma 1 that $\Gamma^*(\beta,\lambda_*(\beta);x)$ is not empty and $\partial_+\ell_{rob}/\partial\lambda(\beta,\lambda_*(\beta);x)$ is given as in (12d), for P_0-almost every x. Since $f_\delta(\cdot)$ is finite in the neighborhood of $\lambda=\lambda_*(\beta)$, it follows from [2, Proposition 2.1] that

$$\frac{\partial_+ f}{\partial\lambda}(\beta,\lambda_*(\beta))=\sqrt{\delta}-\sqrt{\delta}E_{P_0}\left[\beta^T A(X)^{-1}\beta\min_{\gamma\in\Gamma^*(\beta,\lambda_*(\beta);X)}\gamma^2\right]. \tag{45}$$

Case 2. Suppose that $\lambda_*(\beta)=\lambda_{thr}(\beta)$. We first argue that $\partial_+ f/\partial\lambda(\beta,\lambda_*(\beta);x)\in[0,\sqrt{\delta}]$. For this purpose, observe that

$$f_\delta(\beta,\lambda)=\lambda\sqrt{\delta}+E_{P_0}\left[\sup_{\gamma\in\mathbb{R}}\left\{\ell\left(\beta^T X+\gamma\sqrt{\delta}\beta^T A(X)^{-1}\beta\right)-\lambda\sqrt{\delta}\gamma^2\beta^T A(X)^{-1}\beta\right\}\right],$$

as a consequence of the duality representation in Theorem 1. Since the second term in the right hand side of the above equality is non-increasing in λ and $\lambda_*(\beta)$ is a minimizer, we have that

$$0\le f_\delta(\beta,\lambda_*(\beta)+h)-f_\delta(\beta,\lambda_*(\beta))\le\sqrt{\delta}h,$$

for $h>0$. Due to the convexity of f, we also have that $h^{-1}(f_\delta(\beta,\lambda_*(\beta)+h)-f_\delta(\beta,\lambda_*(\beta)))$ is non-decreasing in h. Therefore the right derivative $\partial_+ f/\partial\lambda(\beta,\lambda_*(\beta))\in[0,\sqrt{\delta}]$. As a result, due to the convexity of $f_\delta(\beta,\cdot)$ and finiteness of $f_\delta(\beta,\lambda)$ for any $\lambda>\lambda_*(\beta)$, we have from [2, Proposition 2.1] and Proposition 2b that

$$0\le\frac{\partial_+ f}{\partial\lambda}(\beta,\lambda_*(\beta))\le\lim_{\lambda\downarrow\lambda_*(\beta)}\frac{\partial_- f}{\partial\lambda}(\beta,\lambda)\le\sqrt{\delta}\left(1-\lim_{\lambda\downarrow\lambda_*(\beta)}E_{P_0}\left[\beta^T A(X)^{-1}\beta g_\lambda(X)^2\right]\right), \tag{46}$$

where $g_\lambda(x)$ is such that $g_\lambda(x)\in\Gamma^*(\beta,\lambda;x)$, P_0-almost every x. The existence of measurable maps $\{g_\lambda(\cdot):\lambda>\lambda_*(\beta)\}$ follow from Proposition 7.50(b) of [3].

For the chosen $\beta\in B$, define the set $A:=\{x\in\mathbb{R}^d:\Gamma^*(\beta,\lambda_*(\beta);x)=\varnothing\}$. Take any $x\in A$. For any sequence $\{g_\lambda(x):\lambda>\lambda_{thr}(\beta)\}$ such that $g_\lambda(x)\in\Gamma^*(\beta,\lambda;x)$, we next show that $\lim_{\lambda\downarrow\lambda_*(\beta)}g_\lambda^2(x)=+\infty$. If otherwise, there exist a real number g_0 and a decreasing sequence $\{\lambda_n:n\in\mathbb{N}\}$ satisfying $\lim_{n\to\infty}\lambda_n=\lambda_*(\beta)$ and $\lim_{n\to\infty}g_{\lambda_n}(x)=g_0$. Since $\ell_{rob}(\beta,\lambda;x)$ is right-continuous at $\lambda=\lambda_{thr}(\beta)$ when $f_\delta(\beta,\lambda_{thr}(\beta))<\infty$ (see Lemma 10), we have that

$$\ell_{rob}(\beta,\lambda_*(\beta);x)=\lim_{n\to\infty}\ell_{rob}(\beta,\lambda_n;x)=\lim_{n\to\infty}F(g_{\lambda_n}(x),\beta,\lambda_n;x)=F(g_0,\beta,\lambda_*(\beta);x), \tag{47}$$

where the last equality holds because $F(\gamma,\beta,\lambda;x)$ is a continuous function in (γ,β,λ). However, it follows from (47) that $g_0\in\Gamma(\beta,\lambda_*(\beta);x)$, which contradicts that $x\in A$ as $\Gamma(\beta,\lambda_*(\beta);x)$ is not an empty set if $\limsup_{\lambda\downarrow\lambda_*(\beta)}g_\lambda^2(x)<\infty$. Therefore $\lim_{\lambda\downarrow\lambda_*(\beta)}g_\lambda^2(x)=+\infty$ for $x\in A$.

Applying Fatou's lemma to the right hand side of (46), we obtain from (46) that

$$E_{P_0}[\beta^T A(X)^{-1}\beta\liminf_{\lambda\downarrow\lambda_*(\beta)}g_\lambda^2(X)]\le 1.$$

Since $\liminf_{\lambda\downarrow\lambda_*(\beta)} g_\lambda^2(x) = +\infty$ for $x \in A$, this inequality results in $\infty \times P_0(X \in A) \leq 1$. Therefore $P_0(X \in A) = 0$. In other words, the set of maximizers $\Gamma^*(\beta, \lambda_*(\beta); x)$ is not empty, for P_0-almost every x.

Consequently, an application of envelope theorem (see [20, Corollary 4]) similar to that in Proposition 2b results in $\partial_+\ell_{rob}/\partial\lambda(\beta, \lambda_*(\beta); x) = \sqrt{\delta}(1 - \beta^T A(x)^{-1}\beta \min_{\gamma\in\Gamma^*(\beta,\lambda_*(\beta);x)} \gamma^2)$, for $x \in A$. Since $\ell_{rob}(\beta, \lambda_*(\beta); x)$ is convex in $\lambda_*(\beta)$ for every x (see Lemma 2), we have $h^{-1}(\ell_{rob}(\beta, \lambda_*(\beta) + h; x) - \ell_{rob}(\beta, \lambda_*(\beta); x))$ is non-decreasing in h for $h \geq 0$, and the limit as $h \to 0$ is given by $\partial_+\ell_{rob}(\beta, \lambda_*(\beta); x)$ for $x \in A$. With $P_0(X \in A) = 1$, due to monotone convergence theorem, it follows that $\partial_+ f/\partial\lambda(\beta, \lambda_*(\beta)) = E_{P_0}[\partial_+\ell_{rob}/\partial\lambda(\beta, \lambda_*(\beta); X)]$, thus resulting in (45). This completes the proof of Lemma 4.

□

Proof of Lemma 5. Observe that $\Gamma^*(\beta, \lambda; x) \neq \varnothing$ implies $\lambda \geq \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta$ (see Lemma 1b). With $F(\cdot)$ being defined as in (7), any $g \in \Gamma^*(\beta, \lambda; x)$ must satisfy the first order optimality condition that,

$$2\lambda g = \ell'(\beta^T x + g\sqrt{\delta}\beta^T A(x)^{-1}\beta). \tag{48}$$

This verifies the first part of the statement of Lemma 5. To prove the inequality in (26), we proceed by considering the following cases depending on the signs of $\ell'(\beta^T x)$ and g. If $\ell'(\beta^T x) = 0$, inequality (26) is trivial. Thus, in order to prove (26), it suffices to consider the case where $\ell'(\beta^T x)$ is strictly positive or strictly negative. As $\ell'(\beta^T x) \neq 0$, we have $g \neq 0$. Therefore it is sufficient to establish (26) by considering cases where $\ell'(\beta^T x)$, g are strictly positive or negative.

Case 1 - Suppose that $\ell'(\beta^T x) > 0$ and $g > 0$. Since the convexity of $\ell(\cdot)$ in Assumption 2 ensures that $\ell'(\cdot)$ is non-decreasing, we have $2\lambda g = \ell'(\beta^T x + g\sqrt{\delta}\beta^T A(x)^{-1}\beta) \geq \ell'(\beta^T x)$, due to (48); equivalently, $g \geq \ell'(\beta^T x)/(2\lambda)$. This verifies (26) when both $\ell'(\beta^T x)$ and g are positive.

Case 2 - Suppose that $\ell'(\beta^T x) > 0$ and $g < 0$. Due to convexity of $\ell(\cdot)$, and optimality of g,

$$\begin{aligned} F(g, \beta, \lambda; x) &\geq \sup_{\gamma\geq 0} \Big\{ \ell(\beta^T x) + \ell'(\beta^T x)\sqrt{\delta}\beta^T A(x)^{-1}\beta\gamma - \lambda\sqrt{\delta}\big(\gamma^2\beta^T A(x)^{-1}\beta - 1\big) \Big\} \\ &= \ell(\beta^T x) + \lambda\sqrt{\delta} + \sqrt{\delta}\beta^T A(x)^{-1}\beta \frac{[\ell'(\beta^T x)]^2}{4\lambda}. \end{aligned} \tag{49}$$

An application of the fundamental theorem of calculus to the terms $\ell(\beta^T x + \sqrt{\delta}g\beta^T A(x)^{-1}\beta)$ and g^2 in the definition of $F(g, \beta, \lambda; x)$ (see (7)) allows us to rewrite the left hand side as,

$$F(g, \beta, \lambda; x) = \ell(\beta^T x) + \lambda\sqrt{\delta} + \sqrt{\delta}\beta^T A(x)^{-1}\beta \int_g^0 \Big(2\lambda\gamma - \ell'(\beta^T x + \gamma\sqrt{\delta}\beta^T A(x)^{-1}\beta)\Big) d\gamma.$$

For any γ in $(g, 0)$, we have from the monotonicity of $\ell'(\cdot)$ that $2\lambda\gamma - \ell'(\beta^T x + \gamma\sqrt{\delta}\beta^T A(x)^{-1}\beta)$ does not exceed the positive part of $2\lambda\gamma - \ell'(\beta^T x + g\sqrt{\delta}\beta^T A(x)^{-1}\beta)$. Consequently, it follows from the optimality condition in (48) that,

$$\begin{aligned} F(g, \beta, \lambda; x) &\leq \ell(\beta^T x) + \lambda\sqrt{\delta} + \sqrt{\delta}\beta^T A(x)^{-1}\beta \int_g^0 \Big(2\lambda\gamma - 2\lambda g\Big)_+ d\gamma \\ &= \ell(\beta^T x) + \lambda\sqrt{\delta} + \sqrt{\delta}\lambda\beta^T A(x)^{-1}\beta g^2. \end{aligned}$$

Combining this observation with that in (49), we obtain $|g| \geq \ell'(\beta^T x)/(2\lambda)$.

When $\ell'(\beta^T x) < 0$, (26) follows by an argument symmetric to that of the $\ell'(\beta^T x) > 0$ cases described above. This completes the proof of Lemma 5. □

Proof of Lemma 6. Define the function $L : B \to \mathbb{R}_+$ as $L(\beta) = E_{P_0}[\ell'(\beta^T X)^2]^{1/2}$. With $\ell'(\beta^T X) \neq 0$ almost surely, we have that $L(\beta) > 0$ for any $\beta \in B$. Moreover, we have that $L(\beta)$ is continuous in β due to the continuity of $\ell'(\cdot)$. Then the existence of finite $\bar{L}, \underline{L}$, as in the statement of Lemma 6, follows immediately from the fact that continuous functions attain their extrema over compact sets.

□

Proof of Proposition 8. With $\varphi(g,\beta,\lambda;x)=\varphi_g(\beta,\lambda;x)>0$, we have,

$$\frac{\partial^2 F}{\partial\gamma}(g,\beta,\lambda;x)=-\sqrt{\delta}\beta^T A(x)^{-1}\beta\varphi(g,\beta,\lambda;x)<0.$$

Moreover, we have that $g(\beta,\lambda;x)\in\Gamma^*(\beta,\lambda;x)$ satisfies the first order optimality condition that

$$\ell'(\beta^T x+\sqrt{\delta}g(\beta,\lambda;x)\beta^T A(x)^{-1}\beta)-2\lambda g(\beta,\lambda;x)=0. \tag{50}$$

Using implicit function theorem, the partial derivatives of $g(\beta,\lambda;x)$ are given as follows:

$$\frac{\partial g}{\partial\beta}(\beta,\lambda;x)=-\frac{\partial^2 F/\partial\beta\partial\gamma(g(\beta,\lambda;x),\beta,\lambda;x)}{\partial^2 F/\partial\gamma^2(g(\beta,\lambda;x),\beta,\lambda;x)}=\frac{\ell''(\beta^T T_g(x))}{\varphi_g(\beta,\lambda;x)}\bar{T}_g(x) \tag{51}$$

$$\frac{\partial g}{\partial\lambda}(\beta,\lambda;x)=-\frac{\partial^2 F/\partial\lambda\partial\gamma(g(\beta,\lambda;x),\beta,\lambda;x)}{\partial^2 F/\partial\gamma^2(g(\beta,\lambda;x),\beta,\lambda;x)}=-\frac{2g(\beta,\lambda;x)}{\varphi_g(\beta,\lambda;x)}. \tag{52}$$

Following these expressions for the gradient of g, the Hessian of $\ell_{rob}(\,\cdot\,;x)$ in the statement of Proposition 8 follows from the first order derivative information in Proposition 2 and elementary rules of differentiation. Next, to establish (30), we first provide an equivalent characterization of the relationship $\nabla^2 f_\delta(\beta,\lambda;x)-\Lambda(x)B(x)\succeq 0$. For simplicity, we re-scale $\Lambda(x)$ and pick a new parameter m such that $\Lambda(x)=m\sqrt{\delta}g^2$. To avoid clutter in expressions, we write $\tilde{x}:=T_g(x)$ and $\bar{x}:=\bar{T}_g(x)$ throughout this proof. The matrix $\nabla^2\ell_{rob}(\beta,\lambda;x)-m\sqrt{\delta}g^2B(x)$ can be written as a block matrix, namely,

$$\nabla^2\ell_{rob}(\beta,\lambda;x)-m\sqrt{\delta}g^2B(x)=\begin{bmatrix}(2\lambda-m)\sqrt{\delta}g^2A(x)^{-1}+\frac{(2\lambda-m)\ell''(\beta^T\tilde{x})}{\varphi}\bar{x}\bar{x}^T & -2\sqrt{\delta}g^2z\\ -2\sqrt{\delta}g^2z^T & \frac{4\sqrt{\delta}g^2\beta^TA(x)^{-1}\beta}{\varphi}-m\sqrt{\delta}g^2\end{bmatrix},$$

where $z:=A(x)^{-1}\beta+\frac{\beta^TA(x)^{-1}\beta}{\psi}\bar{x}$ and $\psi:=g\varphi/\ell''(\beta^T\tilde{x})$. According to Schur complement condition, the matrix $\nabla^2\ell_{rob}(\beta,\lambda;x)-m\sqrt{\delta}g^2B(x)$ is positive definite if and only if $(2\lambda-m)\sqrt{\delta}g^2A(x)^{-1}+\frac{(2\lambda-m)\ell''(\beta^T\tilde{x})}{\varphi}\bar{x}\bar{x}^T$ is positive definite and

$$\frac{4\sqrt{\delta}g^2\beta^TA(x)^{-1}\beta}{\varphi}-m\sqrt{\delta}g^2>4\delta g^4z^T\left((2\lambda-m)\sqrt{\delta}g^2A(x)^{-1}+\frac{(2\lambda-m)\ell''(\beta^T\tilde{x})}{\varphi}\bar{x}\bar{x}^T\right)^{-1}z. \tag{53}$$

Recalling from the assumptions that $m\in(0,2\lambda)$ and $\ell(\cdot)$ is convex, the positive definiteness of

$$(2\lambda-m)\sqrt{\delta}g^2A(x)^{-1}+\frac{(2\lambda-m)\ell''(\beta^T\tilde{x})}{\varphi}\bar{x}\bar{x}^T$$

is automatically satisfied. Then, applying Sharman-Morrison formula, one can show that

$$\left((2\lambda-m)\sqrt{\delta}g^2A(x)^{-1}+\frac{(2\lambda-m)\ell''(\beta^T\tilde{x})}{\varphi}\bar{x}\bar{x}^T\right)^{-1}=\frac{1}{(2\lambda-m)\sqrt{\delta}g^2}C, \tag{54}$$

where C is a matrix defined as

$$C:=A(x)-\frac{A(x)\bar{x}\bar{x}^TA(x)}{\bar{x}^TA(x)\bar{x}+\sqrt{\delta}g\psi}$$

Thus, combining equation (53) and (54), if $(\beta,\lambda)\in\mathbb{V}$ and $m\in(0,2\lambda)$, then the matrix $\nabla^2\ell_{rob}(\beta,\lambda;x)-m\sqrt{\delta}g^2B(x)$ if and only if

$$(2\lambda-m)\left(\frac{4\beta^TA(x)^{-1}\beta}{\varphi}-m\right)>4z^TCz.$$

Let a,b and θ be constant defined as

$$a:=2\lambda,\quad b:=\frac{4\beta^TA(x)^{-1}\beta}{\varphi},\quad \theta:=4z^TCz. \tag{55}$$

If $\theta \in (0, ab)$, then we have $m := (ab-\theta)/(a+b)$ satisfying $m \in (0, a \wedge b)$ and $(a-m)(b-m) > \theta$. So it follows that

$$\nabla^2 \ell_{rob}(\beta, \lambda; x) - m\sqrt{\delta} g^2 B(x) \succeq 0 \tag{56}$$

for any $(\beta, \lambda) \in \mathbb{V}$.

The rest of this proof is devoted to arguing that $\theta \in (0, ab)$, and to obtain a simplified lower bound for $m = (ab-\theta)/(a+b)$. We accomplish this by claiming that,

$$ab - \theta \geq \frac{4(\beta^T \tilde{x})^2}{\bar{x}^T A(x) \bar{x} + \sqrt{\delta} g^2 \varphi / \ell''(\beta^T \tilde{x})}. \tag{57}$$

To show (57), first we derive an alternative expression of θ. It follows from the definition of z and C that

$$\frac{\theta}{4} = \left(A(x)^{-1}\beta + \frac{\beta^T A(x)^{-1}\beta}{\psi} \bar{x} \right)^T \left(A(x) - \frac{A(x)\bar{x}\bar{x}^T A(x)}{\bar{x} A(x) \bar{x} + \sqrt{\delta} g \psi} \right) \left(A(x)^{-1}\beta + \frac{\beta^T A(x)^{-1}\beta}{\psi} \bar{x} \right)$$

On expanding the bracket,

$$\begin{aligned} \frac{\theta}{4} &= \beta^T A(x)^{-1}\beta + \left(\frac{\beta^T A(x)^{-1}\beta}{\psi} \right)^2 \bar{x}^T A(x) \bar{x} + 2 \frac{\beta^T A(x)^{-1}\beta}{\psi} \beta^T \bar{x} \\ &\quad - \frac{(\beta^T \bar{x})^2 + \left(\frac{\beta^T A(x)^{-1}\beta}{\psi} \right)^2 (\bar{x} A(x) \bar{x})^2 + 2 \frac{\beta^T A(x)\beta}{\psi} \bar{x}^T A(x) \bar{x} (\beta^T \bar{x})}{\bar{x} A(x) \bar{x} + \sqrt{\delta} g \psi}, \end{aligned}$$

which further implies

$$\begin{aligned} \frac{\theta}{4} &= \beta^T A(x)^{-1}\beta - \frac{(\beta^T \bar{x})^2}{\bar{x}^T A(x) \bar{x}} + \frac{(\beta^T \bar{x})^2}{\bar{x}^T A(x) \bar{x}} + \left(\frac{\beta^T A(x)^{-1}\beta}{\psi} \right)^2 \bar{x}^T A(x) \bar{x} + 2 \frac{\beta^T A(x)^{-1}\beta}{\psi} \beta^T \bar{x} \\ &\quad - \frac{(\beta^T \bar{x})^2 + \left(\frac{\beta^T A(x)^{-1}\beta}{\psi} \right)^2 (\bar{x} A(x) \bar{x})^2 + 2 \frac{\beta^T A(x)\beta}{\psi} \bar{x}^T A(x) \bar{x} (\beta^T \bar{x})}{\bar{x} A(x) \bar{x} + \sqrt{\delta} g \psi} \\ &= \beta^T A(x)^{-1}\beta - \frac{(\beta^T \bar{x})^2}{\bar{x}^T A(x) \bar{x}} + \frac{\left(\beta^T A(x)^{-1}\beta \frac{\sqrt{\bar{x}^T A(x) \bar{x}}}{\psi} + \frac{\beta^T \bar{x}}{\sqrt{\bar{x}^T A(x) \bar{x}}} \right)^2 (\sqrt{\delta} g \psi)}{\bar{x} A(x) \bar{x} + \sqrt{\delta} g \psi} \\ &= \beta^T A(x)^{-1}\beta - \frac{(\beta^T \bar{x})^2}{\bar{x}^T A(x) \bar{x}} + \frac{\left(\beta^T A(x)^{-1}\beta \frac{\sqrt{\bar{x}^T A(x) \bar{x}}}{\psi} + \frac{\beta^T \bar{x}}{\sqrt{\bar{x}^T A(x) \bar{x}}} \right)^2}{1 + \frac{\bar{x}^T A(x) \bar{x}}{\sqrt{\delta} g \psi}}. \end{aligned}$$

Then, using above upper bound for θ and the definition of a and b, we obtain,

$$\begin{aligned} \frac{ab-\theta}{4} \left(1 + \frac{\bar{x}^T A(x) \bar{x}}{\sqrt{\delta} g \psi} \right) &\geq \left[\left(\frac{2\lambda}{\varphi} - 1 \right) \beta^T A(x)^{-1}\beta + \frac{(\beta^T \bar{x})^2}{\bar{x}^T A(x) \bar{x}} \right] \left(1 + \frac{\bar{x}^T A(x) \bar{x}}{\sqrt{\delta} g \psi} \right) \\ &\quad - \left(\frac{\beta^T A(x)^{-1}\beta}{\psi} \sqrt{\bar{x}^T A(x) \bar{x}} + \frac{\beta^T \bar{x}}{\sqrt{\bar{x}^T A(x) \bar{x}}} \right)^2. \end{aligned}$$

Since $2\lambda - \varphi = \sqrt{\delta}(\beta^T A(x)^{-1}\beta) \ell''(\beta^T \tilde{x})$, $\psi = g\varphi / \ell''(\beta^T \tilde{x})$ and $\bar{x} = \tilde{x} + \sqrt{\delta} g A(x)^{-1}\beta$, on expanding the squares in the last term, the above inequality simplifies to,

$$\frac{ab-\theta}{4} \left(1 + \frac{\bar{x}^T A(x) \bar{x}}{\sqrt{\delta} g \psi} \right) \geq \frac{\ell''(\beta^T \tilde{x})}{\sqrt{\delta} g^2 \varphi} \left(\beta^T \bar{x} - \sqrt{\delta} g \beta^T A(x)^{-1}\beta \right)^2 = \frac{\ell''(\beta^T \tilde{x})}{\sqrt{\delta} g^2 \varphi} \left(\beta^T \tilde{x} \right)^2.$$

This establishes (57). Finally, combining (56) and (57), we have

$$\nabla^2 \ell_{rob}(\beta, \lambda; x) - \frac{4 \left(\beta^T \tilde{x} \right)^2 \ell''(\beta^T \tilde{x})}{1 + \bar{x}^T A(x) \bar{x} \ell''(\beta^T \tilde{x}) / (\sqrt{\delta} g^2 \varphi)} \frac{1}{2\lambda\varphi + 4\beta^T A(x)^{-1}\beta} B(x) \succeq 0,$$

which is obtained by plugging in the definitions of a, b from (55). □

Proof of Lemma 8. Since $\lambda > \lambda'_{thr}(\beta)$, there exist $\varepsilon > 0$ such that $\lambda \geq (M/2+\varepsilon)\sqrt{\delta}\beta^T A(x)^{-1}\beta$, for P_0-almost every x. Since $\ell(\cdot)$ is twice differentiable and $\ell''(\cdot) \leq M$ (see Assumption 3), it follows from the definition of $F(\cdot)$ in (7) that

$$\begin{aligned}\frac{\partial^2 F}{\partial \gamma^2}(\gamma,\beta,\lambda;x) &= \sqrt{\delta}\beta^T A(x)^{-1}\beta\left(\ell''(\beta^T x+\sqrt{\delta}\gamma\beta^T A(x)^{-1}\beta)\sqrt{\delta}\beta^T A(x)^{-1}\beta - 2\lambda\right) \\ &\leq \sqrt{\delta}\beta^T A(x)^{-1}\beta\left(M\sqrt{\delta}\beta^T A(x)^{-1}\beta - 2\lambda\right) \leq -2\varepsilon\delta\beta^T A(x)^{-1}\beta,\end{aligned} \tag{58}$$

for P_0-almost every x. Thus the map $\gamma \mapsto F(\gamma,\beta,\lambda;x)$ is strongly concave for every $\beta \in B, \lambda > \lambda'_{thr}(\beta)$ and P_0-almost every $x \in \mathbb{R}^d$, and attains maximum at a unique point $g(\beta,\lambda;x)$. In such case, the set $\Gamma^*(\beta,\lambda;x) = \{g(\beta,\lambda;x)\}$ is singleton. Moreover, for any $\beta \in B, \lambda > \lambda'_{thr}(\beta)$, we have,

$$\varphi(g(\beta,\lambda;x),\beta,\lambda;x) = -\frac{\partial^2 F}{\partial \gamma^2}(g(\beta,\lambda;x),\beta,\lambda;x) < 0,$$

for P_0-almost every x. Thus, every element of $\{(\beta,\lambda) : \beta \in B, \lambda > \lambda'_{thr}(\beta)\}$ lies in the set $\mathcal{U}(x)$. □

Proof of Proposition 9. a) The inclusion that $\mathbb{V} \subseteq \mathbb{W}$ is immediate from their respective definitions. To verify the second inclusion, see that for any $\beta \in B$,

$$\lambda'_{thr}(\beta) \leq 2^{-1}\sqrt{\delta}M\rho_{\min}^{-1}\|\beta\|^2 < 2^{-1}\sqrt{\delta_0}M\rho_{\min}^{-1}R_\beta\|\beta\| \leq 2^{-1}(\underline{L}\rho_{\max}^{-1})^{1/2}\|\beta\| = K_1\|\beta\|,$$

due to Assumption 1b. Since we have that $\lambda'_{thr}(\beta) < K_1\|\beta\|$, any $(\beta,\lambda) \in \mathbb{W}$ is also an element of the set $\{(\beta,\lambda) : \beta \in B, \lambda > \lambda'_{thr}(\beta)\}$. The final inclusion in the statement of Proposition 9a follows from Lemma 8. b) For every $(\beta,\lambda) \in \mathbb{W}$, we have $\lambda > \lambda'_{thr}(\beta)$. Then it is immediate from Lemma 8 that $\Gamma^*(\beta,\lambda;x)$ is singleton for $(\beta,\lambda) \in \mathbb{W}$ and P_0-almost every x. As a result, any measurable selection $g(\cdot)$ satisfying (28) is uniquely specified for almost every (β,λ,x) in the subset $\mathbb{W} \times S_X \subseteq \mathcal{U}$. Moreover, due to mean value theorem, the first order optimality condition (50) means that $g(\beta,\lambda;x) = \ell'(\beta^T x)/(2\lambda - \sqrt{\delta}\beta^T A(x)^{-1}\beta\ell''(\eta))$, for some η between the real numbers $\beta^T x$ and $\beta^T \tilde{x}$. Since $\ell''(\cdot) \leq M$ and $\delta \leq \delta_0$, for $(\beta,\lambda) \in \mathbb{W}$ we have that

$$2\lambda - \sqrt{\delta}\beta^T A(x)^{-1}\beta\ell''(\eta) \;\geq\; 2\lambda_{\min}(\beta) - \sqrt{\delta}\beta^T A(x)^{-1}\beta\ell''(\eta) \;\geq\; \varphi_{\min}\|\beta\|.$$

Also note that if $(\beta,\lambda) \in \mathbb{V}$ we have

$$2\lambda - \sqrt{\delta}\beta^T A(x)^{-1}\beta\ell''(\eta) \leq 2\lambda \leq 2K_2\|\beta\|$$

Then the conclusion in Proposition 9b follows. c) Since $|\ell'(\beta^T X) - \ell'(0)| \leq M\|\beta\|\|X\|$ (due to Assumption 3), $E_{P_0}\|X\|^4$ is finite and $\|\beta\| \leq R_\beta$ (see Assumptions 2 and 4), we have from the bounds in (32) that $\sup_{(\beta,\lambda)\in\mathbb{V}} E_{P_0}\left[g^4(\beta)\right] < \infty$. Therefore, the collection $\{g^2(\beta,\lambda;X) : (\beta,\lambda) \in \mathbb{V}\}$ is L_2-bounded. Then it is immediate from the definitions $T_g(x) := x + \sqrt{\delta}g(\beta,\lambda;x)A(x)^{-1}\beta$, $\bar{T}_g(\cdot) := x + 2\sqrt{\delta}g(\beta,\lambda;x)A(x)^{-1}\beta$ and Cauchy-Schwarz inequality that the collections $\{(T_g(X))^2, (\bar{T}_g(X))^2 : (\beta,\lambda) \in \mathbb{V}\}$ are L_2-bounded. Consequently, the collections $\{\ell(\beta^T T_g(X)), \ell'(\beta^T X_g)^2 : (\beta,\lambda) \in \mathbb{V}\}$ are L_2-bounded as well due to the at most quadratic growth property of $\ell(\cdot)$ (see Assumption 2). d) Recall from Part a) that $\mathbb{V}$ is strictly contained in $\mathcal{U}(x)$, for P_0-almost every x. With the DRO objective $\sup_{P:D_c(P,P_n)\leq\delta} E_P\left[\ell(\beta^T X)\right] = E_{P_0}[\ell_{rob}(\beta,\lambda)] =: f_\delta(\beta,\lambda)$ defined in terms of the convex loss $\ell(\cdot)$ specified over the entire real line, the second order partial derivative expressions of $\ell_{rob}(\cdot)$ in the statement of Proposition 8 hold throughout the set $\mathbb{V}$. It follows from the L_2-boundedness just established in Part c) and these partial derivative expressions that the norms of the individual entries of the Hessian matrix $\nabla^2_\theta \ell_{rob}(\theta;X)$ are all bounded in L_2-norm over the set $\theta \in \mathbb{V}$. With this L_2-boundedness of the collection $\{\nabla^2_\theta \ell_{rob}(\theta;X) : \theta \in \mathbb{V}\}$, the desired exchange of derivative and expectation in $\nabla^2_\theta f_\delta(\theta) = E_{P_0}[\nabla^2_\theta \ell_{rob}(\theta;X)]$, for $\theta \in \mathbb{V}$, follows as a consequence of dominated convergence theorem. □

Proof of Proposition 6. Let $g(\beta,\lambda;X)$ be such that $g(\beta,\lambda;X)\in\Gamma^*(\beta,\lambda;X)$ and

$$h(\beta,\lambda;X)=\begin{pmatrix} L'\tilde{X} \\ \sqrt{\delta}\left(1-g^2(\beta,\lambda;X)\beta^TA(X)^{-1}\beta\right)\end{pmatrix} \qquad P_0-\text{a.s.}, \tag{59}$$

where $\tilde{X}:=X+\sqrt{\delta}g(\beta,\lambda;X)A(X)^{-1}\beta$ and L' is any arbitrary (measurable) choice from the subgradient interval $\partial\ell(\beta^T\tilde{X})$. Since $\ell(\cdot)$ is convex, with at most quadratic growth (see Assumption 2), there exist positive constants C_0,C_1 such that $|L'|\le C_0+C_1|\beta^T\tilde{X}|$. As $E\|X\|^2<\infty$, it follows from Lemma 3a that $E[g^2(\beta,\lambda;X)],E\|\tilde{X}\|^2,E|\beta^T\tilde{X}|$ are finite. Then due to Cauchy-Schwartz inequality, we have that $Eh(\beta,\lambda;X)$ is well-defined. Here we have used that $\beta^TA(x)^{-1}\beta$ is bounded for P_0-almost every x (see Assumption 1b). Now, since $h(\beta,\lambda;X)\in D(\beta,\lambda;X)=\partial\ell_{rob}(\beta,\lambda)$, we have that

$$\ell_{rob}(\beta',\lambda';X)\ge\ell_{rob}(\beta,\lambda;X)+h(\beta,\lambda;X)^T\begin{pmatrix}\beta'-\beta\\ \lambda'-\lambda\end{pmatrix}, \qquad P_0-\text{a.s.}$$

Taking expectations on both sides of the above inequality, we obtain $Eh(\beta,\lambda;X)\in\partial f_\delta(\beta,\lambda)$. □

Proof of Lemma 9. Let $g(\beta,\lambda;X)$ be such that $g(\beta,\lambda;X)\in\Gamma^*(\beta,\lambda;X)$ and the given subgradient $h(\beta,\lambda;X)$ is defined as in (59) in terms of $g(\beta,\lambda;X)$ and $\tilde{X}:=X+\sqrt{\delta}g(\beta,\lambda;X)A(X)^{-1}\beta$. As in the proof of Proposition 6, we have $|\ell'(u)|\le C_0+C_1|u|$ as a consequence of convexity, continuous differentiability and at most quadratic growth of $\ell(\cdot)$. Since $E\|X\|^4<\infty$, $\beta^TA(x)^{-1}\beta$ is bounded for P_0-almost every x, $\lambda>\lambda_{thr}(\beta)+\eta$, and $\|\beta\|\le R_\beta$, it follows from Lemma 3a that $E[g^4(\beta,\lambda;X)],E\|\tilde{X}\|^4$ and $E[\ell'(\beta^T\tilde{X})^4]$ are all uniformly bounded for every $(\beta,\lambda)\in U_\eta$. Then due to Cauchy-Schwartz inequality, we have that $\sup_{(\beta,\lambda)\in\mathbb{U}_\eta}E\|h(\beta,\lambda;X)\|^2<\infty$. □

Appendix B. Line search scheme.

Our iterative procedure requires evaluating

$$\ell_{rob}(\beta,\lambda;x)=\sup_{\gamma\in\mathbb{R}}F(\gamma,\beta,\lambda;x)$$

and obtaining a maximizer $\gamma^*=g(\beta,\lambda;x)\in\Gamma^*(\beta,\lambda;x)$. This task involves a one dimensional optimization problem over γ. This problem, we claim, can be solved through a line search. This can be done efficiently on a case-by-case basis given $\ell(\cdot)$ (as we do in our numerical examples). However, our goal here is to provide reasonably general conditions which can be used to efficiently implement a line search procedure to compute $\ell_{rob}(\beta,\lambda;x)$. Unfortunately, however, the function $F(\cdot,\beta,\lambda;x)$ is not necessarily concave. So, to show that the line search can be implemented efficiently, we need to use study the definition of $F(\cdot)$ and introduce assumptions on $\ell(\cdot)$, which we believe are reasonable. The general line search scheme is easy to develop for λ small or large enough. Recall that

(1) When $\lambda<\lambda_{thr}(\beta)$, the dual objective $f_\delta(\beta,\lambda)=\infty$, so the line search algorithm will not be executed in this case.
(2) When $\lambda\ge\lambda'_{thr}(\beta)$, the function $F(\cdot,\beta,\lambda;x)$ is concave for P_0-almost every x. Consequently, finding $g(\beta,\lambda;x)$ is a convex optimization problem, and therefore can be solved by bisection method or Newton-Raphson method. We will always be in this case if $\delta<\delta_0$.
(3) The third case is the most challenging case, as we shall explain. It requires imposing smoothness assumptions on our loss function.

Lemma 11. *Suppose Assumptions 1 - 4 are satisfied then $\delta<\delta_0$. In turn, this implies that we are in case 2. above (i.e. $\lambda\ge\lambda'_{thr}(\beta)$) and a unique maximizer with $|g(\beta,\lambda,x)|\le|\ell'(\beta^Tx)|/(\varphi_{\min}\|\beta\|)$ exist for P_0-almost surely x due to Proposition 9 b). Then, there exist constants C_1 and C_2, such that for every $(\beta,\lambda)\in\mathbb{W}$ and $x\in\mathbb{R}^d$, it is required at most $\log_2(C_1\varepsilon^{-1}\|\beta\|^{-1}(1+\|x\|))$ steps for binary search method to solve for a solution γ such that $|\gamma-\gamma^*|\le\varepsilon$. In turn, the binary search procedure generates estimates for $\partial\ell_{rob}/\partial\beta$ and $\partial\ell_{rob}/\partial\lambda$ with ε accuracy and $\log_2(C_2\varepsilon^{-1}(1+\|x\|)^2)$ steps. (The explicit construction of the derivative estimates is summarized in Appendix E.)*

Proof. When solving the one dimensional optimization problem, $\max_\gamma F(\gamma,\beta,\lambda;x)$, we consider a scaled problem by setting $\tilde\gamma = \|\beta\|\gamma$ and it suffices to consider a scaled problem $\max_{\tilde\gamma} F(\tilde\gamma,\beta,\lambda;x)$, for $|\tilde\gamma| \le \varphi_{\min}^{-1}|\ell'(\beta^T x)|$. Note that there exist some constant C, independent of β,λ,x, such that the bound $\varphi_{\min}^{-1}|\ell'(\beta^T x)| \le \varphi_{\min}^{-1}(|\ell'(0)| + M\|\beta\|\|x\|) \le C(1+\|x\|)$. Notice that the function $F(\cdot,\beta,\lambda;x)$ is concave by Lemma 8, so using binary search method to solve for the optimal $\tilde\gamma$ up to an ε error requires at most $\log_2(C\varepsilon^{-1}(1+\|x\|))$ steps. Finally, consider the inverse scaling $\gamma = \|\beta\|^{-1}\tilde\gamma$, the error of optimal $\tilde\gamma$ need to be bounded by $\|\beta\|\varepsilon$ in order to get γ with error bounded by ε. Thus an ε-accuracy solution for γ requires at most $\log_2(C\varepsilon^{-1}\|\beta\|^{-1}(1+\|x\|))$ steps.

Now we consider the error of $\partial\ell_{rob}/\partial\beta$ and $\partial\ell_{rob}/\partial\lambda$ induced by the error of line search. Recall from Proposition (2) that $\partial\ell_{rob}/\partial\beta = \ell'(\beta^T x + \sqrt{\delta} g\beta^T A(x)^{-1}\beta)(x + \sqrt{\delta} gA(x)^{-1}\beta)$ and $\partial\ell_{rob}/\partial\lambda = -\sqrt{\delta}\left(g^2\beta^T A(x)^{-1}\beta - 1\right)$. For every $(\beta,\lambda)\in\mathbb{W}$ and $x\in\mathbb{R}^d$, there exist some constant L uniform in β,λ,x, such that $g\mapsto \partial\ell_{rob}/\partial\theta = (\partial\ell_{rob}/\partial\beta, \partial\ell_{rob}/\partial\lambda)$ for $g \le C\|\beta\|^{-1}(1+\|x\|)$ is Lipschitz continuous with Lipschitz constant $L\|\beta\|(1+\|x\|)$. Consequently, in order to get an ε-accuracy evaluation for $\partial\ell_{rob}/\partial\theta$, we need to solve for an $\varepsilon L^{-1}\|\beta\|^{-1}(1+\|x\|)^{-1}$-accuracy solution for g. Consequently the bisection method is required to run for at most $\log_2(CL\varepsilon^{-1}(1+\|x\|)^2)$ steps. □

It then remains to discuss case 3. Namely, to develop an algorithm to compute $g(\beta,\lambda;x)$ when $\lambda \in [\lambda_{thr}(\beta), \lambda'_{thr}(\beta))$, which, requires a more delicate analysis. The following example shows that the function $F(\cdot,\beta,\lambda;x)$ can have infinitely many local optima.

Example 1. *Suppose that $\beta \ne \mathbf{0}$, $P_0(\cdot) = \delta_{\{\mathbf{0}\}}(\cdot)$ and $\ell(u) = u^2 - \cos u$. It then follows that $\kappa = 1$ and $\lambda_{thr}(\beta) = \sqrt{\delta}\beta^T A(\mathbf{0})^{-1}\beta$. Thus, $F(\gamma,\beta,\lambda_{thr}(\beta);\mathbf{0}) = -\cos\left(\sqrt{\delta}\beta^T A(\mathbf{0})^{-1}\beta\gamma\right)$, which has infinitely many local optima.*

So, to solve the global nonconvex optimization problem, it is necessary to reduce the feasible region of optimization problem to a compact interval. To this end, we consider the scaled line search problem $\max_{\bar\gamma\in\mathbb{R}} F(\bar\gamma\beta^T A(x)^{-1}\beta,\beta,\lambda;x)$, instead of considering the original line search problem $\max_{\gamma\in\mathbb{R}} F(\gamma,\beta,\lambda;x)$. In the following Lemma, we show that when $(\beta,\lambda)\in\mathbb{U}_\eta$, it suffices to consider the scaled line search problem with a compact feasible region.

Lemma 12. *Recall the definition of $\mathbb{U}_\eta$ from (17) and suppose that Assumption 1-3 hold and $\eta > 0$. Then there exist a random variable R with $E_{P_0}[R^2] < \infty$, such that*

$$|g\beta^T A(X)^{-1}\beta| \le R$$

for any $(\beta,\lambda)\in\mathbb{U}_\eta$ and $g\in\Gamma^*(\beta,\lambda;X)$.

Proof. The fact that $(\beta,\lambda)\in\mathbb{U}_\eta$ implies that $\lambda \ge \lambda_{thr}(\beta)+\eta$. Then, according to the Assumptions we have $\beta^T A(X)^{-1}\beta \le \rho_{\min}^{-1}R_\beta^2$. Thus, letting $\varepsilon = \eta\delta^{-1/2}\rho_{\min}R_\beta^{-2}$, we have $\lambda \ge (\kappa+\varepsilon)\sqrt{\delta}\beta^T A(x)^{-1}\beta$, and thus the result of Lemma 3a can be applied. As a result, there exist a constant C_1 such that

$$|g\beta^T A(X)^{-1}\beta| \le 1 + C_1\varepsilon^{-1}(1+|\beta^T X|) =: R$$

and the squared integrability of R is easy to verified. □

With the help of Lemma 12, we know it suffices to consider the scaled line search problem

$$\max_{\bar\gamma\in[-R,R]} F(\bar\gamma\beta^T A(X)^{-1}\beta,\beta,\lambda;X)$$

with a bounded feasible region $[-R,R]$, where the length of interval $2R$ is squared integrable, controlling the average complexity of the line search. Next, we need to rule out the pathological case that the stationary points of $\gamma\mapsto F(\gamma\beta^T A(x)^{-1}\beta,\beta,\lambda;x)$ in $[-R,R]$ contain infinitely many connected components. To this ends, we further impose an assumption that $\ell(\cdot)$ is *piecewise real analytic* in any compact set K. A function f is *real analytic* on an open set D if for any $x_0\in D$ one can write $f(x) = \sum_{n=0}^\infty a_n(x-x_0)^n$, in which the coefficients a_n are real numbers and the series is convergent to $f(x)$ for x in a neighborhood

of x_0. A function f is *piecewise real analytic* in a compact set K if there exist $n \in \mathbb{N}$ and closed intervals $D_1, \ldots, D_n$, such that $K \subset \bigcup_{i=1}^n D_i$, and for each D_i, the restriction of f on D_i has a real analytic extension. In other words, for each set D_i, there exists an open set $D_i \subset \tilde{D}_i$ and a real analytic function g_i on $\tilde{D}_i$, such that $f(x) = g_i(x)$ for all $x \in D_i$.

Lemma 13. *Suppose that f is piecewise real analytic in compact set K, then the stationary points of f in K are contained in only finitely many connected components.*

Proof. If a connected component of stationary points is not a discrete point, then it must contains an open interval that disjoint with the remaining connected components. Thus, as the set K is compact, the total number of non-singleton connected components is finite.

It remains to prove the number of discrete stationary points of f is finite. To this end, it suffices to prove g_i has finite discrete stationary points in D_i. We claim that there does not exist an accumulation point of discrete stationary points of g_i in set D_i. Otherwise, we can find a sequence of discrete stationary points $\{x_n : n \geq 1\}$, and x_n converge to a point $x \in D_i$. Consider the Taylor series of the function g_i around x, if the Taylor series is zero except the constant term. Then by the real analytic property, the function is a constant in a neighborhood around x, violating the assumption that all the x_n are discrete stationary points. If the Taylor series has non-zero higher order terms, then there exist a neighbourhood of x such that x is the only stationary point in that neighborhood, violating the assumption that x_n converge to x. So the discrete stationary points of g_i does not have an accumulation point in D_i. As a result, we can find an open cover of D_i such that each open set in the open cover contains at most one discrete stationary point of g_i. Since D_i is also compact, g_i has finite discrete stationary points in D_i. The result follows. □

Note that it is important for the series to be absolutely and uniformly convergent; smoothness alone does not imply the existence of finitely many stationary points on a compact interval, as the next example shows.

Remark 5. *Even if a function is in $C^\infty(\mathbb{R})$, it may have infinitely many isolated local optima on a compact set. Consider*

$$f(x) := \begin{cases} \cos\left(-(1-x^2)^{-1}\right) \exp\left(-(1-x^2)^{-1}\right) & \text{if } -1 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Now we discuss the line search scheme and its complexity. If the loss function $\ell(\cdot)$ is piecewise real analytic, the function $\gamma \mapsto F(\gamma, \beta, \lambda; x)$ is also piecewise real analytic. In addition, using the result of Lemma 12, the optimization problem $\max_{\gamma \in \mathbb{R}} F(\gamma, \beta, \lambda; x)$ is equivalent to $\max_{\bar{\gamma} \in [-R,R]} F(\bar{\gamma}\beta^T A(x)^{-1}\beta, \beta, \lambda; x)$, a one dimensional optimization problem with compact feasible region. We denote the closed intervals partitioning $[-R, R]$ by $D_1, \ldots, D_n$. Thus, $F(\bar{\gamma}\beta^T A(x)^{-1}\beta, \beta, \lambda; x)$ has finite local optimal points in compact interval $[-R, R]$, which are either stationary points in the interior of a interval, or a hinge point connecting two adjacent intervals. One possible approach for computing stationary points of a real analytic function is to consider the holomorphic extension of the function and then apply Cauchy's theorem (see, for example, [11, 12]). This approach is guaranteed to locate all of the stationary points. However, the use of Cauchy's theorem requires the evaluation of certain integrals in smooth trajectories. The evaluation of these trajectories can be done with high precision integration rules which take advantage of the analytic properties of the integrands, evaluating $o\left(\varepsilon^{-\delta}\right)$ (for any $\delta > 0$) points in the integrand to achieve a ε relative error, for example, applying Newton integration rules. The complexity of finding all the stationary points of function $\bar{\gamma} \mapsto F(\bar{\gamma}\beta^T A(x)^{-1}\beta, \beta, \lambda; x)$ is proportional to $2R$, the length of the searching interval. Therefore, the total complexity of the line search scheme is $O_p(\varepsilon^{-\delta})$, for any $\delta > 0$, uniformly for all $(\beta, \lambda) \in \mathbb{U}_\eta$. This complexity includes the evaluation of the global maxima by comparing the value of $F(\gamma, \beta, \lambda; x)$ at all local optimal points.

Another approach, instead of using Cauchy's theorem, applying Newton's method repeatedly. Because we have established that there are finitely many roots if the loss is piecewise real analytic, by restarting

Newton's method from randomly chosen initial conditions we will be able to locate, with an exponential decaying error rate in the number of retrials, the global optimum. While this algorithms is easy to implement, its analysis is of independent interest and too long to include in this paper. So, we will discuss it in future work.

Appendix C. Analytical solution to projection $\Pi_{\mathbb{W}}$.

The algorithms described in Sections 3.2 and 3.3 requires to project the variables $\theta = (\beta, \lambda)$ to the set $\mathbb{W}$. In this section we provide an analytical solution to the projection. Recall that the definition of $\mathbb{W}$ as

$$\mathbb{W} = \{(\beta, \lambda) \in B \times \mathbb{R} \mid K_1\|\beta\| \le \lambda \le K_2 R_\beta\}.$$

Suppose that $\theta = (\beta, \lambda) \in B \times \mathbb{R}$. The projection $\Pi_{\mathbb{W}}(\theta) = (\beta', \lambda')$ is determined by the convex optimization problem

$$\begin{aligned} \min \quad & \|(\beta, \lambda) - (\beta', \lambda')\|^2 \\ \text{s.t.} \quad & K_1\|\beta'\| \le \lambda' \le K_2 R_\beta, \\ & (\beta', \lambda') \in B \times \mathbb{R}. \end{aligned}$$

Thus, analytical solution of the projection $\Pi_{\mathbb{W}}(\theta)$ is given by

$$\Pi_{\mathbb{W}}(\theta) = \begin{cases} (\beta, \lambda) & \text{if } (\beta, \lambda) \in \mathbb{W}, \\ (\beta, K_2 R_\beta) & \text{if } \|\beta\| \le K_2 R_\beta / K_1, \lambda > K_2 R_\beta, \\ (\mathbf{0}, 0) & \text{if } \lambda < -\|\beta\|/K_1, \\ \left(\frac{\|\beta\|+\lambda}{1+K_1^2}\frac{\beta}{\|\beta\|}, \frac{K_1\|\beta\|+K_1^2\lambda}{1+K_1^2}\right) & \text{if } -\|\beta\|/K_1 \le \lambda < \min\{K_1\|\beta\|, K_2 R_\beta(1+K_1^{-2}) - \|\beta\|/K_1\}, \\ \left(\frac{K_2 R_\beta}{K_1}\frac{\beta}{\|\beta\|}, K_2 R_\beta\right), & \text{otherwise.} \end{cases}$$

The optimality of above solution can be verified using the KKT condition.

Appendix D. Tables specifying useful constants and an illustration of computation of some of these constants in Examples

Tables 1 - 3 below present a compilation of useful constants used in the main text. A demonstration of how some of the relevant constants can be computed is presented in the subsequent sections D.1 and D.2.

Table 1. Constants which are specified as part of the framework

constant	description
δ	the radius of the Wasserstein ball specified in the DRO formulation (2)
$\rho_{\max}$	the largest possible eigenvalue of the matrices $\{A(x) : x \in \mathbb{R}^d\}$
$\rho_{\min}$	the smallest possible eigenvalue of the matrices $\{A(x) : x \in \mathbb{R}^d\}$
κ	quadatic growth rate of $\ell(\cdot)$ characterized by $\kappa := \inf\{s \ge 0 : \sup_{u\in\mathbb{R}} \left(\ell(u) - su^2\right) < \infty\}$
M	the maximum possible value for the second derivative $\ell''(\cdot)$, whenever it exists
R_β	equals $\sup_{\beta\in B} \|\beta\|$ if the set B is taken to be bounded
c_1, c_2, p	positive constants which ensure $P_0\left(\lvert\ell'(\beta^T X)\rvert > c_1, \lvert\beta^T X\rvert > c_2\|\beta\|\right) \ge p$
k_1, k_2	positive constants k_1, k_2 satisfying $\lvert u\rvert\ell''(u) \le k_1 + k_2\lvert\ell'(u)\rvert$ required by Theorem 5
K	the number of piecewise components in the loss $\ell(u) = \max_{i=1,\ldots,K} \ell_i(u)$

TABLE 2. Some useful constants which are specified in the analysis

constant	value specified in the analysis
δ_0	$\rho_{\min}^2 \underline{L} R_\beta^{-2} M^{-2} \rho_{\max}^{-1}$
δ_1	$\min\{\delta_0/4, c_1^2 c_2^2 p^2 \rho_{\min}^2 \rho_{\max}^{-1} \underline{L} \overline{L}^{-2}/256\}$
δ_2	$c_1^2 c_2^2 p \rho_{\min}^2 \left(2k_1 \rho_{\max}^{1/2} + 4c_1 \rho_{\min}^{1/2}(1+k_2)\right)^{-2}$
κ_0	$2^{-1} \underline{L} \rho_{\max}^{-1}$
κ_1	$pC\rho_{\max}^{-1}/2$
K_1	$\underline{L}^{1/2} \rho_{\max}^{-1/2}/2$

constant	value specified in the analysis
K_2	$\sqrt{\delta} M R_\beta \rho_{\min}^{-1} + \rho_{\min}^{-1/2} \overline{L}$
$\lambda_{thr}(\beta)$	$\sqrt{\delta}\kappa P_0 - \text{ess-sup}_x \beta^T A(x)^{-1}\beta$
$\lambda'_{thr}(\beta)$	$\sqrt{\delta} M P_0 - \text{ess-sup}_x \beta^T A(x)^{-1}\beta$
$\overline{L}$	$\max_{\beta\in B} E_{P_0}\left[\ell'(\beta^T X)^2\right]^{1/2}$
$\underline{L}$	$\min_{\beta\in B} E_{P_0}\left[\ell'(\beta^T X)^2\right]^{1/2}$
$\varphi_{\min}$	$\underline{L}^{-1/2}\rho_{\max}^{-1/2} - \sqrt{\delta} R_\beta M \rho_{\min}^{-1}$

TABLE 3. Constants involved in the SGD algorithm

constant	description
$\{\alpha_k : k \geq 1\}$	step-size sequence satisfying $\alpha_k = \alpha k^{-\tau}$ for some $\alpha > 0$, $\tau \in [1/2, 1)$
η	specifies the set $\mathbb{U}_\eta$ (see (17)) onto which iterates are projected
ξ	polynomial averaging constant in (15)

D.1. **Example 4.2: logistic regression.** The logistic loss function is given by

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

Theorem 4 and/or 5 are applied to analyze the locally strong convexity of f_δ; Proposition 4 and 5 guarantees the efficacy of the proposed algorithm. The constants appearing in the related assumptions can be chosen as:

- $\rho_{\min}, \rho_{\max}$: Determined by the selection of $A(x)$. If $A(x)$ is an identity matrix, then $\rho_{\min} = \rho_{\max} = 1$.
- κ: Since the loss function is asymptotically linear, $\kappa = 0$.
- k_1, k_2: $k_1 = k_2 = 1$, because $\sup_{u\in\mathbb{R}} |u|\ell''(u; \pm 1) \leq 1$
- M: $M = 1/4$, because $\ell''(u; \pm 1) = (1/4)\cosh(u/2)^{-2} \leq 1/4$,
- c_1, c_2, p: Depends on the distribution of X. The constants with desired properties exist if and only if rank $\{Y_i \cdot X_i\}_{i=1}^n = n$, which would happen almost surely if the data generating distribution of X has a density.

D.2. **Example 4.3: linear regression.** The squared loss function is given by

$$\ell(u; y) = (u - y)^2$$

Theorem 4 and/or 5 are applied to analyze the locally strong convexity of f_δ; Proposition 4 and 5 guarantees the efficacy of the proposed algorithm. The constants appearing in the related assumptions can be chosen as:

- $\rho_{\min}, \rho_{\max}$: Determined by the selection of $A(x)$. If $A(x)$ is an identity matrix, then $\rho_{\min} = \rho_{\max} = 1$.
- κ: Since the loss function is quadratic, $\kappa = 1$.
- k_1, k_2: $k_1 = \max_{i=1,\ldots,n} |Y_i|$ and $k_2 = 1$.
- M: $M = 1$ because of $\ell''(u; y) = 1$,
- c_1, c_2, p: Depends on the distribution of X. The constants with desired properties exist almost surely if the data generating distribution of X has a density.

D.3. **Example 4.4: support vector machines.** The hinge loss function is given by

$$\ell(u; y) = \max\left(0, 1 - yu\right)$$

Proposition 6 and 7 provide theoretical foundation for hinge loss function, in which Assumption 1 and 2 are imposed. The related constants can be chosen as:

- $\rho_{\min}, \rho_{\max}$: Determined by the selection of $A(x)$. If $A(x)$ is an identity matrix, then $\rho_{\min} = \rho_{\max} = 1$.
- κ: Since the loss function is asymptotically linear, $\kappa = 0$.

Appendix E. Algorithm

Algorithm 1 Stochastic Gradient Descent for the case $\delta < \delta_0$

input: Initial parameter $\bar{\theta}_0 = \theta_0 = (\beta_0, \lambda_0) \in \mathbb{W}$, step-size sequence $(\alpha_k)_{k\geq 1}$, total number of iterations N.

for $k = 1, 2, \ldots, N$ **do**

Generate an independent sample X_k from the distribution P_0.

Set $n_k \geq \tau \log_2(k) - \log_2(\alpha) + 2\log_2(1 + \|X_k\|)$ as total cuts for bisection method.

Set $I_k = \big[-|\ell'(\beta_{k-1}^T X_k)|/(\varphi_{\min}\|\beta_{k-1}\|), |\ell'(\beta_{k-1}^T X_k)|/(\varphi_{\min}\|\beta_{k-1}\|)\big]$ as the initial interval.

Solve $\gamma_k = \arg\max_{\gamma \in I_k} F(\gamma, \theta_{k-1}; X_k)$ using bisection method for n_k steps.

Compute $\tilde{X}_k = X_k + \sqrt{\delta} A(X_k)^{-1}\beta_{k-1}$.

Compute $\nabla_\theta \ell_{rob}(\theta_{k-1}; X_k)$ using the closed-form expression

$$\frac{\partial \ell_{rob}}{\partial \beta}(\theta_{k-1}; X_k) = \ell'\left(\beta_{k-1}^T \tilde{X}_k\right)\tilde{X}_k \quad \text{and} \quad \frac{\partial \ell_{rob}}{\partial \lambda}(\theta_{k-1}; X_k) = -\sqrt{\delta}\left(\gamma_k^2 \beta_{k-1}^T A(X_k)^{-1}\beta_{k-1} - 1\right).$$

Update the parameter by $\theta_k := \Pi_{\mathbb{W}}\big(\theta_{k-1} - \alpha_k \nabla_\theta \ell_{rob}(\theta_{k-1}; X_k)\big)$.

Update the trajectory average $\bar{\theta}_k = \left(\frac{k-1}{k}\right)\bar{\theta}_{k-1} + \frac{1}{k}\theta_k$.

end for

output: $\bar{\theta}_N$ and θ_N.

Acknowledgments.

Support from NSF grant 1820942, DARPA grant N660011824028, MOE SRG ESD 2018 134 and China Merchants Bank are gratefully acknowledged.

References

[1] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 1200–1205, New York, NY, USA, 2017. ACM.

[2] D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, Aug 1973.

[3] Dimitri P Bertsekas and Steven E Shreve. *Stochastic optimal control: the discrete time case*. Elsevier, 1978.

[4] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[5] Jose Blanchet, Yang Kang, Karthyek Murthy, and Fan Zhang. Data-driven optimal cost selection for distributionally robust optimization. *In the proceedings of Winter Simulation Conference*, 2019.

[6] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[7] Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in wasserstein distributionally robust estimation. *arXiv preprint arXiv:1906.01614*, 2019.

[8] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-driven chance constrained programs over Wasserstein balls. -, 2018. Available from Optimization Online.

[9] Zhi Chen, Melvyn Sim, and Peng Xiong. Adaptive robust optimization with scenario-wise ambiguity sets. -, 2018. Available from Optimization Online.

[10] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27*, pages 1646–1654. Curran Associates, Inc., 2014.

[11] Michael Dellnitz, Oliver Schütze, and Qinghua Zheng. Locating all the zeros of an analytic function in one complex variable. *Journal of Computational and Applied mathematics*, 138(2):325–333, 2002.

[12] LM Delves and JN Lyness. A numerical method for locating the zeros of an analytic function. *Mathematics of computation*, 21(100):543–560, 1967.

[13] E. den Boef and D. den Hertog. Efficient line search methods for convex functions. *SIAM Journal on Optimization*, 18(1):338–363, 2007.

[14] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

[15] Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[16] Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. Robust hypothesis testing using wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pages 7902–7912, 2018.

[17] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.

[18] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[19] Fengqiao Luo and Sanjay Mehrotra. Decomposition algorithm for distributionally robust optimization using wasserstein metric with an application to a class of regression models. *European Journal of Operational Research*, 278(1):20 – 35, 2019.

[20] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[21] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, Sep 2018.

[22] Abdelkader Mokkadem and Mariane Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Ann. Appl. Probab.*, 16(3):1671–1702, 08 2006.

[23] Eric Moulines and Francis R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 451–459. Curran Associates, Inc., 2011.

[24] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[25] Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The wasserstein shrinkage estimator. *arXiv preprint arXiv:1805.07194*, 2018.

[26] Yung-Kyun Noh, Byoung-Tak Zhang, and Daniel D Lee. Generative local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 1822–1830, 2010.

[27] B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[28] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, (3):400–407, 09 1951.

[29] Soroosh Shafieezadeh-Abadeh, Peyman Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584. 2015.

[30] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[31] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages I–71–I–79. JMLR.org, 2013.

[32] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[33] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[34] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5339–5349, Red Hook, NY, USA, 2018. Curran Associates Inc.

[35] Jun Wang, Alexandros Kalousis, and Adam Woznica. Parametric local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 1601–1609, 2012.

[36] Weijun Xie. On distributionally robust chance constrained program with Wasserstein distance. *Mathematical Programming*, 2019.

[37] I. Yang. A convex optimization approach to distributionally robust markov decision processes with Wasserstein distance. *IEEE Control Systems Letters*, 1(1):164–169, July 2017.

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