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 optimization procedures which have the same sample and iteration complexity as a natural non-DRO benchmark algorithm such as stochastic gradient descent; and sometimes even better complexity. Our analysis provides insights into the fine structure and convexity properties of the DRO value function, the optimal policy and the worst-case optimal transport adversarial model.

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 $\delta > 0$

(*) 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.

parameterizes the size of the distributional uncertainty. Precisely, we assume that P_0 is an arbitrary distribution with finite second moments, that is $E_{P_0}\|X\|_2^2 < \infty$.

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 of a decision, β_0, which significantly under-performs in actual practice, relative to the 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 precise description of $\mathcal{U}_\delta(P_0)$. 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, [27, 19, 36, 6, 14, 4, 35, 31, 15, 32, 7] and references therein). In particular, 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, [27, 4, 28, 13]. 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.

In general, however, the inner maximization (2) is not easy to perform and its properties, parametrically as a function of β, are non-trivial to analyze. Nevertheless, to have algorithms that are scalable and problem formulations that are powerful, it is desirable to consider a flexible model for distributional uncertainty sets, $\mathcal{U}_\delta(P_0)$, which enables, precisely, 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). This is precisely our goal in this paper.

A description of the distributional uncertainty region $\mathcal{U}_\delta(P_0)$. In this paper we focus on DRO formulations based on extensions of the Wasserstein distance, called optimal transport discrepancies. 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 $c(x,x') = \|x - x'\|$ under any given norm. If $c(x,x')$ is not a distance, then $D_c(P_0,P)$ is not necessarily a distance.

Our interest in this paper is on the computational tractability of the DRO problem (2) assuming that,

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

for a flexible class of functions c. We concentrate on what we call local 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, but the methods here can be applied to cost functions of the form $c(x,x') = u(x-x')$ or $c(x,x') = u(x') - u(x) - \nabla u(x)^T(x'-x)$, for a strongly convex function $u(\cdot)$ with a Lipschitz gradient $\nabla u(\cdot)$.

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. It is not difficult to think of 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.

Naturally, one can always select $A(x)$ to be the identity matrix in order to recover a Wasserstein-*distance*-based DRO formulation. Even in this case, (2) is not entirely easy to analyze, due to the fact that $\ell(\cdot)$ has a flexible-enough structure that makes the inner optimization problem in (2) non-trivial to study.

In this paper we do not focus on the problem of fitting the cost function. Related questions have been explored, at least empirically, in classification settings, using manifold learning procedures ([33, 23, 5]). Our point 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.

However, to exploit these formulations one must develop algorithms which can be used to solve (2) efficiently. A natural benchmark to obtain a certificate of efficiency is to consider a canonical type of algorithm used in the (well-understood) setting of the non-DRO version (1). Such benchmark is given by stochastic gradient descent, whose complexity is well-understood based on natural convexity assumptions on the loss function $\ell(\cdot)$. So, to develop algorithms to solve (2) efficiently we need to study the structural properties (e.g., convexity, conditioning, etc.) of a formulation such as (2).

Main contributions: Structural study of (2) and algorithmic consequences. First, using a standard duality result, we write the inner maximization in (2) as,

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

for a dual objective function $\ell_{rob}(\cdot)$ and a dual variable $\lambda \geq 0$. We show that after a rescaling in λ, the right hand side of (6) is locally strongly convex in (β,λ) uniformly over a compact set containing the optimizer, with a strong convexity parameter of at least $\kappa_2\delta^{1/2}$ (for some $\kappa_2 > 0$ which we identify), under suitable convexity and growth assumptions on $\ell(\cdot)$; see Theorem 1 and Theorem 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. 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.

On the other hand, due to the strong convexity properties derived for $\ell_{rob}(\cdot)$, for $\delta > 0$, we are able to provide an iterative-scheme, based on stochastic gradient descent, which can be used to solve (2) in $O_p(\varepsilon^{-1}L)$ iteration complexity to reach $O(\varepsilon)$ error, where L is the iteration complexity of a one dimensional line search procedure. We also discuss in the Appendix how to execute this line search procedure efficiently, provided that suitable smoothness assumptions are imposed on $\ell(\cdot)$. In this sense, we obtain a provably efficient iterative procedure to solve (2) (which, given the complexity bounds, it might be even faster in many practical settings).

Another useful consequence of our results involve the application of standard Sample Average Approximation output analysis results to Optimal Transport based DRO. This enables the direct application of results in [30], to produce confidence regions for the solution of the DRO formulation.

Our structural results may also find use in the analysis of optimization algorithms based on deterministic convex programming formulations of (2) in which the size of the formulations grow with the cardinality of the support of P_0 (see, for example, [27, 19, 36, 28, 8, 17, 7, 34]). Finally, we mention [31], 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 [31] is similar to the ones that we discuss in Section 3, however, for a fixed λ, chosen suitably large. Our analysis suggests that a suitable rescaling may enhance performance, even in the case of the more general type of losses considered in [31].

Our second main contribution consists in studying the local structure of the worst-case optimal transport plan, including uniqueness and comparative statics results, see Proposition 2 and Theorem 3. 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.

Organization of the paper. The rest of the paper is organized as follows. Section 2, sets the stage for our analysis, by first obtaining the duality result (6). In Section 2, we also introduce our assumptions and our main structural results involving the convexity of the modified objective function (the right hand side of (6)), as well as the structure of the worst-case optimal transport plan. In Section 3 we discuss the iterative procedures which are naturally applicable given our structural results, together with the corresponding rate of convergence analysis. We provide several illustrative examples in Section 4, followed by a discussion on conclusions in Section 5.

The proofs of our main structural results are given in Section 6. 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 we consider a result of independent interest, 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 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 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 via (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 of (6).

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

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

which is an infinite-dimensional linear program that maximizes $E_\pi[\ell(\beta^T X')]$ over all joint distributions π of pair $(X, X') \in \mathbb{R}^d \times \mathbb{R}^d$ satisfying the linear marginal constraints that the law of X is P_0 and the cost constraint that $E_\pi[c(X,X')] \le \delta$ (see [6] for details). Theorem 1 below builds on a recent strong duality result applicable for this linear program when the chosen transport cost function $c(x,x')$ is not necessarily a metric. The local Mahalanobis costs we consider in this paper satisfy Assumption 1 below. As mentioned in the Introduction, such a cost function is not necessarily symmetric (hence need not be a metric).

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,

b) there exist positive constants $\underline{\rho}, \overline{\rho}$ satisfying $\sup_{\|v\|=1} v^T A(x) v \le \overline{\rho}$ and $\inf_{\|v\|=1} v^T A(x) v \ge \underline{\rho}$, for P_0-almost every $x \in \mathbb{R}^d$.

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

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

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

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

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

Proof. 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

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

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

$$\sup_{\Delta\in\mathbb{R}^d}\left\{\ell(\beta^T(X+\Delta)) - \lambda\left(\Delta^T A(X)\Delta - \delta\right)\right\} = \sup_{\gamma\in\mathbb{R}} F(\gamma,\beta,\lambda;X) =: \ell_{rob}(\beta,\lambda;X), \tag{8}$$

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. □

2.2. **Convexity properties of (6).** The convexity properties of the dual objective function $f(\beta,\lambda) := E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$ we derive here will be crucial towards establishing iteration complexities of the computational schemes developed in Section 3. Specifically, we establish convexity of $f(\cdot)$, restricted strong convexity of $f(\cdot,\lambda)$ for fixed λ, and restricted joint strong convexity of $f(\cdot)$ under increasingly stronger sets of assumptions listed below.

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 $E_{P_0}\|X\|^2 < \infty$.

Assumption 3. The set $B \subseteq \mathbb{R}^d$ is convex and compact. Specifically, $\sup_{\beta\in B}\|\beta\| =: R_\beta < \infty$.

Assumption 4. 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 $\ell''(\cdot) \leq M$.

Assumption 5. 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 $\beta \in B$.

Assumption 6. The loss function $\ell : \mathbb{R} \to \mathbb{R}$ is uniformly strongly convex in compact sets. In addition, the baseline distribution P_0 is such that there exist $C > 0$, $p \in (0,1)$ satisfying

$$P_0\left(|\beta^T X \ell'(\beta^T X)| > C\|\beta\|\right) \geq p,$$

for every $\beta \in B$.

2.2.1. *Some useful constants.* Recall the definition of the dual objective $f(\beta,\lambda)$ in the statement of Theorem 1 along with the fact that the function $f(\beta,\lambda)$ attains its infimum at $\lambda = \lambda_*(\beta)$ for every fixed $\beta \in B$. Define positive constants $\delta_0 := \rho_{\min}^2 \underline{L} R_\beta^{-2} M^{-2}\rho_{\max}^{-1}$, $K_1 := 2^{-1}(\underline{L}\rho_{\max}^{-1})^{1/2}$, $K_2 := 2^{-1}\delta_0^{1/2} M R_\beta \rho_{\min}^{-1} + (\overline{L}\rho_{\min}^{-1})^{1/2}$, $K_3 := R_\beta K_2$, $\delta_1 := \min\{\delta_0/4, C^2 p^2 \rho_{\min}^2 \rho_{\max}^{-1}\underline{L}\overline{L}^{-2}/256\}$, and the set

$$\mathbb{V} := \{(\beta,\lambda) \in B \times [0,K_3] : \ K_1\|\beta\| \leq \lambda \leq K_2\|\beta\|\}. \tag{9}$$

2.2.2. *Local strong convexity.* Recall from Theorem 1 that $\arg\min_{\lambda\geq 0} f(\beta,\lambda)$ is not empty for every $\beta \in B$.

Proposition 1. *Suppose that Assumptions 1 - 5 hold and $\delta < \delta_0$. Then for any $\beta \in B$ and dual optimizer $\lambda^*(\beta) \in \arg\min_{\lambda\geq 0} f(\beta,\lambda)$, we have $(\beta, \lambda_*(\beta)) \in \mathbb{V}$, for every $\beta \in B$.*

Theorem 2. *The function $f : B \times \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$ is*

a) *convex when Assumptions 1a and 2 hold;*
b) *such that $\partial^2 f/\partial\beta^2 \succeq \sqrt{\delta}\kappa_1\lambda^{-1}\mathbb{I}_d$, for $(\beta,\lambda) \in \mathbb{V}$ and a positive constant $\kappa_1 > 0$, when $\delta < \delta_0$ and Assumptions 1 - 5 hold;*
c) *such that the Hessian $\nabla^2 f \succeq \sqrt{\delta}\kappa_2\mathbb{I}_{d+1}$, for $(\beta,\lambda) \in \mathbb{V}$ and a positive constant $\kappa_2 > 0$, when $\delta < \delta_1$ and Assumptions 1 - 6 hold.*

Theorem 2 above identifies conditions under which $f(\cdot)$ is convex and strongly convex when restricted to the set $\mathbb{V}$. Indeed it turns out that it is sufficient to restrict attention to $\mathbb{V}$ to arrive at local strong convexity around $\arg\min_{\beta,\lambda} f(\beta,\lambda)$ because of Proposition 1. The proofs of Proposition 1 and Theorem 2 are presented in Section 6.1. As far as we know, Theorem 2 is the first result that characterizes strong convexity of the objective in Wasserstein distance based DRO in a suitable sense. Evidently, strong convexity is a property that crucially determines the iteration complexity of gradient based descent methods. We utilize this in Section 3.

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 let P_n be the empirical distribution corresponding to the n data samples (X_i, Y_i), the stochastic optimization objective to be minimized, $E_{P_n}[(Y-\beta^T X)^2] = n^{-1}\sum_{i=1}^n (Y_i - \beta^T X_i)^2$ is not strongly convex. Theorem 2 asserts that the respective DRO objective $f(\beta,\lambda)$ is, nevertheless, strongly convex in a region containing the minimizer (refer an example in Section 4.3 for a discussion on how a DRO formulation of the least squares linear regression problem results in the dual objective of the form $f(\beta,\lambda)$). Thus, due to Theorem 2, for a considerable class of useful loss functions $\ell(\cdot)$, the DRO dual objective to be minimized, $f(\beta,\lambda)$, is strongly convex in a suitable sense, even if the nonrobust counterpart $E_{P_0}[\ell(\beta^T X)]$ is not.

2.2.3. *Comments on Assumptions 1 - 6.* 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 assumptions are satisfied by a wide variety of loss functions $\ell(\cdot)$ and a flexible class of local Mahalanobis cost functions $c(\cdot)$ that includes commonly used Euclidean metric, Mahalanobis distances as special cases.

As we shall see in the proof of Theorem 2, the twice differentiability imposed in Assumption 4 is necessary to characterize the local strong convexity of f by means of the positive definiteness of Hessian of f. The boundedness of $\ell''(\cdot)$ and finiteness of $E\|X\|^2$ automatically imply the existence of $\overline{L} \in (0,\infty)$ in Assumption 5. Further, if $E_{P_0}[\ell'(\beta^T X)^2] = 0$ for some $\beta_0 \in B$, then $\ell(\beta_0^T X)$ is degenerate (P_0-almost surely a constant) and $E_{P_0}[\ell(\beta^T X)]$ is minimized at $\beta = \beta_0$ because of the convexity of $\ell(\cdot)$. With such degeneracy being not common in stochastic optimization models, it is not restrictive to assume $E_{P_0}[\ell'(\beta^T X)^2] > \underline{L} > 0$ in the light of compactness of the set B.

Moving to Assumption 6, the positive probability requirement in Assumption 6 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. As we shall see in Remark 2, this is necessary because the strong convexity coefficient of

$\ell_{rob}(\beta, \lambda; x)$ is directly proportional to $(\beta^T x \ell'(\beta^T x))^2 + \kappa_3 \delta^{3/2}$, for some nonnegative constant κ_3. Since $f(\beta, \lambda) = E_{P_0}[\ell_{rob}(\beta, \lambda; X)]$ (see Theorem 1), it could be argued as in Remark 2 that the strong convexity coefficient of $f(\cdot)$ is $O(\delta^{3/2})$ if the only non-negative value of C for which the probability requirement in Assumption 6 holds is $C = 0$. Therefore the probability requirement in Assumption 6 is necessary to ensure that the strong convexity coefficient κ_2 is bounded away from zero independent of the ambiguity radius δ.

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 $\inf_{\lambda \geq 0} f(\beta, \lambda)$ (see Theorem 1). Recall the notation that $\lambda_*(\beta)$ attains the infimum in $\inf_{\lambda \geq 0} f(\beta, \lambda)$ for fixed $\beta \in B$. For each $\beta \in B, \lambda \geq 0$ and $x \in \mathbb{R}^d$, define the set of optimal solutions to (7) as

$$\Gamma^*(\beta, \lambda; x) = \left\{ \gamma : \ F(\gamma, \beta, \lambda; x) = \sup_{c \in \mathbb{R}} F(c, \beta, \lambda; x) \right\}. \tag{10}$$

Finally, for a fixed $\beta \in B$, define $\lambda_{thr}(\beta)$ to be the P_0-essential supremum of $\sqrt{\delta} \kappa \beta^T A(x)^{-1} \beta$. Similarly, when Assumption 4 holds, define $\lambda'_{thr}(\beta)$ to be the P_0-essential supremum of $\sqrt{\delta} M \beta^T A(x)^{-1} \beta / 2$. Since $\kappa \leq M/2$, we have $\lambda'_{thr}(\beta) \geq \lambda_{thr}(\beta)$ for every $\beta \in B$.

Theorem 3. *Suppose that Assumptions 1,2 hold and $\beta \neq \mathbf{0}$. Take any dual optimizer $\lambda^*(\beta) \in \arg\min_{\lambda \geq 0} f(\beta, \lambda)$. Then*

a) *the dual optimizer $\lambda_*(\beta)$ is positive unless $\ell(\cdot)$ is a constant function. If $\ell(\cdot)$ is indeed a constant function, then any distribution in $\{P : \ D_c(P_0, P) \leq \delta\}$ attains the supremum in (6);*
b) *the dual optimizer $\lambda_*(\beta) \geq \lambda_{thr}(\beta)$ whenever $\ell(\cdot)$ is not a constant;*
c) *if $\lambda_*(\beta) > \lambda_{thr}(\beta)$, the law of*

$$X^* := X + \sqrt{\delta} G A(X)^{-1} \beta \tag{11}$$

attains the supremum in (6) and satisfies $E[c(X, X^*)] = \delta$; here the random variable G can be written as $G := ZG_- + (1 - Z)G_+$, with $G_- = \inf \Gamma(\beta, \lambda_*(\beta); X)$, $G_+ = \sup \Gamma(\beta, \lambda_*(\beta); X)$, P_0-almost surely, and Z is an independent Bernoulli random variable satisfying $P(Z = 1) = (\bar{c} - 1)/(\bar{c} - \underline{c})$, where $\bar{c} := E_{P_0}[G_+^2 \beta^T A(X)^{-1} \beta]$ and $\underline{c} := E_{P_0}[G_-^2 \beta^T A(X)^{-1} \beta]$;
d) *if $\lambda_*(\beta) = \lambda_{thr}(\beta)$, then a worst-case distribution attaining the supremum in (6) may not exist;*
e) *under additional Assumption 4, if $\lambda^*(\beta) > \lambda'_{thr}(\beta)$, the set $\Gamma^*(\beta, \lambda_*(\beta); x)$ is 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 3 is presented in Section 6.3.

Remark 1. Consider the case $\beta = \mathbf{0}$. Then $\lambda = 0$ attains the minimum in $\min_{\lambda \geq 0} f(\mathbf{0}, \lambda)$, $\sup_{D_c(P_0, P) \leq \delta} E_{P_0}[\ell(\beta^T X)] = \ell(0)$, and any distribution in $\{P : \ D_c(P_0, P) \leq \delta\}$ attains the supremum.

2.4. **Comparative statics analysis.** In this section we explain how the worst-case distribution structure explained in Section 2.3 changes for every realization of X when the radius of ambiguity δ is changed. Such a sample-wise description is facilitated by examining the derivative of the random variable G described in Part e) of Theorem 3, P_0-almost surely. First, we describe a sufficient condition for $\lambda_*(\beta)$ to exceed $\lambda'_{thr}(\beta)$, which is required for Part e) of Theorem 3 to hold. Recall the definition that $\delta_0 := \rho^2_{\min}\underline{L}R_\beta^{-2}M^{-2}\rho^{-1}_{\max}$.

Lemma 1. *Suppose that Assumptions 1 - 5 are satisfied. Consider any fixed $\beta \in B$ and let $\lambda_*(\beta)$ be such that it attains the minimum in $\min_{\lambda\geq 0} f(\beta,\lambda)$. Then if $\delta < \delta_0$, we have $\lambda_*(\beta) > \lambda'_{thr}(\beta)$.*

The proof of Lemma 1 is available in Section 6.1.

Proposition 2. *Suppose that Assumptions 1 - 5 are satisfied. For $\delta \in (0,\delta_1)$ and any fixed $\beta \in B\setminus\{\mathbf{0}\}$, suppose that we denote the unique worst-case distribution attaining the supremum in $\sup_{P:D_c(P_0,P)\leq\delta} E_P[\ell(\beta^T X)]$ by P^*_δ. Then 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 Proposition 2 is presented in Section 6.3. Interestingly, Proposition 2 asserts that the trajectory $\{X^*_\delta : \delta \in [0,\delta_0)\}$ is a straight-line, P_0-almost surely, with probability mass being transported to farther distances as δ increases in $[0,\delta_1)$.

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 algorithm enjoys optimal rates of convergence among the class of iterative algorithms that (i) utilize first-order oracle information and (ii) have 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 optimal transport 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, [27, 19]). 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 $f(\beta,\lambda)$ derived in Theorem 2. Guided by the strong convexity structure of $f(\beta,\lambda)$, 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 (10). With the DRO objective (6) being related to the dual objective $f(\beta,\lambda) := E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$ as in Theorem 1, the minimization can be restricted to the effective domain,

$$\mathbb{U} := \{(\beta,\lambda)\in B\times\mathbb{R}_+ :\ E_{P_0}\left[\ell_{rob}(\beta,\lambda;X)\right] < \infty\}\,. \tag{12}$$

Lemma 2 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

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

Lemma 2. *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 3. *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 $x\in\mathbb{R}^d$.*

Lemma 4 below utilizes envelope theorem (see [18]) to characterize the gradients of $\ell_{rob}(\cdot)$.

Lemma 4. *Suppose that Assumptions 1 - 2 hold and that $\ell(\cdot)$ is continuously differentiable. 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,*

$$\frac{\partial_-\ell_{rob}}{\partial\beta_j}(\beta,\lambda;x) = \min_{\gamma\in\Gamma^*(\beta,\lambda;x)} \ell'\left(\beta^T(x+\sqrt{\delta}\gamma A(x)^{-1}\beta)\right)(x+\sqrt{\delta}\gamma A(x)^{-1}\beta)_j, \tag{13a}$$

$$\frac{\partial_+\ell_{rob}}{\partial\beta_j}(\beta,\lambda;x) = \max_{\gamma\in\Gamma^*(\beta,\lambda;x)} \ell'\left(\beta^T(x+\sqrt{\delta}\gamma A(x)^{-1}\beta)\right)(x+\sqrt{\delta}\gamma A(x)^{-1}\beta)_j, \tag{13b}$$

$$\frac{\partial_-\ell_{rob}}{\partial\lambda}(\beta,\lambda;x) = \min_{\gamma\in\Gamma^*(\beta,\lambda;x)} -\sqrt{\delta}\left(\gamma^2\beta^T A(x)^{-1}\beta - 1\right), \tag{13c}$$

$$\frac{\partial_+\ell_{rob}}{\partial\lambda}(\beta,\lambda;x) = \max_{\gamma\in\Gamma^*(\beta,\lambda;x)} -\sqrt{\delta}\left(\gamma^2\beta^T A(x)^{-1}\beta - 1\right). \tag{13d}$$

Furthermore, $\lambda\mapsto\ell_{rob}(\beta,\lambda;x)$ and $\beta_j\mapsto\ell_{rob}(\beta,\lambda;x)$ is differentiable if and only if $\{\frac{\partial F}{\partial\gamma}(\gamma,\beta,\lambda;x) : \gamma\in\Gamma^*(\beta,\lambda;x)\}$ is a singleton; in that case, 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,

$$\frac{\partial \ell_{rob}}{\partial \beta}(\beta, \lambda; x) = \ell'\left(\beta^T \tilde{x}\right)\tilde{x} \quad \text{and} \quad \frac{\partial \ell_{rob}}{\partial \lambda}(\beta, \lambda; x) = -\sqrt{\delta}\left(g^2 \beta^T A(x)^{-1}\beta - 1\right). \tag{14}$$

The proof of Lemma 4 can be found in Appendix A. For a given $(\beta, \lambda) \in \mathbb{U}_1$, any univariate optimization procedure such as bisection (or) Newton-Raphson methods can be used to solve (7). Assuming that it is feasible to exchange the derivative and expectation operators in $\nabla_{(\beta,\lambda)} E_{P_0}[\ell_{rob}(\beta, \lambda; X)]$ (see Proposition 3 in Section 3.2), the derivatives of $\ell_{rob}(\beta, \lambda; X)$ yield noisy gradients of $f(\beta, \lambda)$.

A simple gradient descent (or) stochastic gradient 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 $\ell'(\cdot)$. Likewise, due to the characterization in Lemma 4, all the function evaluation information required to implement a stochastic gradient descent type iterative scheme for minimizing its robust counterpart $f(\beta, \lambda)$ are evaluations of $\ell(\cdot)$ and $\ell'(\cdot)$.

3.2. **A stochastic gradient descent scheme for the case $\delta < \delta_0$.** 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 radius of ambiguity, δ, satisfies $\delta < \delta_0$. As we shall see imminently, $f(\cdot)$ is differentiable when $\delta < \delta_0$.

Recall the definition of positive constants K_1, K_2 and K_3 in Section 2.2.1. Define the set,

$$\mathbb{W} := \{(\beta, \lambda) \in B \times [0, K_3] : K_1 \|\beta\| \leq \lambda\}. \tag{15}$$

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

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

Lemma 5. *Suppose that Assumptions 1 - 5 hold and $\delta < \delta_0$. Then*

a) *the set $\mathbb{W} \subseteq \{(\beta, \lambda) : \beta \in B, \lambda > \lambda'_{thr}(\beta)\}$; and*
b) *the map $\gamma \mapsto F(\gamma, \beta, \lambda; x)$ is proper and strongly concave for every $(\beta, \lambda) \in \mathbb{W}$, P_0-almost every x.*

The proof of Lemma 5 is presented in Section 6.1. Due to Lemma 5, we have that the set of maximizers $\Gamma^*(\theta; x) := \{\gamma \in \mathbb{R} : F(\gamma, \theta; x) = \sup_{c \in \mathbb{R}} F(c, \theta; x)\}$ is singleton for every $\theta \in \mathbb{W}$. Then it is immediate from Lemma 4b that $\nabla_\theta \ell_{rob}(\theta; x) = (\partial f/\partial\beta(\theta; x), \partial f/\partial\lambda(\theta; x))$ exists and can be computed from (14) for $\theta \in \mathbb{W}$, P_0-almost every x.

Proposition 3. *Suppose that Assumptions 1 - 5 hold and $\delta < \delta_0$. Then $E_{P_0}[\nabla_\theta \ell_{rob}(\theta; X)]$ is well-defined and $\nabla_\theta f(\theta) = E_{P_0}[\nabla_\theta \ell_{rob}(\theta; X)]$ for $(\beta, \lambda) \in \{(\beta, \lambda) : \beta \in B, \lambda > \lambda'_{thr}(\beta)\} \supset \mathbb{W}$.*

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 $\nabla_\theta f(\theta)$ and are called 'stochastic gradients' of $f(\theta)$. Utilising these noisy gradients, we generate averaged iterates $\{\bar{\theta}_k : k \geq 1\}$ according to the following scheme:
Fix $\xi \geq 0$ and initialize $\bar{\theta}_0 = \theta_0 \in \mathbb{V}$. 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 (14) 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:

$$\theta_k := \Pi_{\mathbb{W}}\big(\theta_{k-1} - \alpha_k \nabla_\theta \ell_{rob}(\theta_{k-1}; X_k)\big) \quad \text{and} \quad \bar{\theta}_k = \left(1 - \frac{\xi+1}{k+\xi}\right)\bar{\theta}_{k-1} + \frac{\xi+1}{k+\xi}\theta_k, \tag{16}$$

where $\Pi_{\mathbb{W}}(\cdot)$ denotes the projection operation on to the closed convex set $\mathbb{W}$ and $(\alpha_k)_{k\geq 1}$ is referred to as the step-size sequence (or) learning rate of the iterative scheme.

Assumption 7. *The step-size sequence $(\alpha_k)_{k\geq 1}$ 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 [26]). If $\xi = 0$ in the definition of $\bar{\theta}_k$ in (16), 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 [24] and references therein). On the other hand, the averaging scheme with $\xi > 0$ is referred as polynomial-decay averaging (see [29]).

3.2.2. *Rates of convergence.* Our objective here is to characterize the convergence of $(f(\bar{\theta}_k))_{k\geq 1}$ for the iteration scheme (16). Recall from the definition of the positive constant δ_1 in Section 2.2.1 that $\delta_1 < \delta_0$.

Let $f_* := \inf_{\theta\in B\times\mathbb{R}_+} f(\theta)$ be the optimal value. It is well-known that stochastic gradient descent schemes for smooth objective functions enjoy $f(\bar{\theta}_k) - f_* = O_p(k^{-1})$ rate of convergence if f is strongly convex and $f(\theta_k) - f_* = O_p(k^{-1/2})$ if f is simply convex, for suitable choices of step sizes (see, for example, [29] and references therein). While $f(\cdot)$ is convex for all $\delta \geq 0$, it follows from Theorem 2c that $f(\cdot)$ 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 $f(\bar{\theta}_k) - f_*$ when $\delta < \delta_1$. The proof of Proposition 4 is presented in Section 6.4.

Proposition 4. *Suppose that Assumptions 1 - 5 hold and $E\|X\|^4 < \infty$. Then we have,*

a) *$f(\bar{\theta}_k) - f_* = O_p(k^{-1/2})$ if $\delta < \delta_0$, $\xi \geq 1$ in (16) and $\tau = 1/2$ in Assumption 7;*

b) *$f(\bar{\theta}_k) - f_* = O_p(\sqrt{\delta} k^{-1})$ if $\delta < \delta_1$, $\xi = 0$, $\tau \in (1/2, 1)$ in Assumption 7, and Assumption 6 is satisfied.*

For the strongly convex case, the averaged procedure endows the sequence $(f(\bar{\theta}_k))_{k\geq 1}$ 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 7. 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 at θ minimizing $f(\theta)$, in order to have $f(\theta_k) - f_* = O_p(k^{-1})$ (see, for example, [22, 21] for discussions on the effect of step sizes on error $f(\theta_k) - f_*$).

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. 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 $O_p(\delta^{-1/2}\varepsilon^{-1})$ iterations in order to obtain an iterate $\bar{\theta}_k$ that satisfies $f(\bar{\theta}_k) - f_* \leq \varepsilon$. On the other hand, if $\delta > \delta_1$, we require the usual $O_p(\varepsilon^{-2})$ iteration complexity to obtain $f(\theta_k) - f_* \leq \varepsilon$, 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, [29]). Here, recall from the discussion following

Theorem 2 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 $f(\beta, \lambda)$ is jointly strongly convex in (β, λ) more generally under the conditions identified in Theorem 2.

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 $O_p(\delta^{-1/2}\varepsilon^{-1}L)$ 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.

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 the interval $[0, K_3]$ 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 Lemma 14a 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 [14, 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 strong convexity characterizations in Theorem 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:

$$\tilde{\beta}_k = \beta_{k-1} - \alpha_k \frac{\partial f}{\partial \beta}(\beta_{k-1}, \lambda_{k-1}; X_k) \tag{17a}$$

$$\tilde{\lambda}_k = \lambda_{k-1} - \gamma_k \frac{\partial f}{\partial \lambda}(\beta_{k-1}, \lambda_{k-1}; X_k), \text{ and} \tag{17b}$$

$$\theta_k = \Pi_{\mathbb{W}}\left((\tilde{\beta}_k, \tilde{\lambda}_k)\right). \tag{17c}$$

where the step-sizes $(\alpha_k)_{k\geq 1}, (\gamma_k)_{k\geq 1}$ satisfy the step-size requirement in Assumption 7 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 (17b) 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,

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

Once again, we consider the averaged iterates $\bar{\theta}_k$, defined as in (16) with $\xi = 0$. Similar to Section 3.2, if we let $f_* := \inf_{\theta \in B\times\mathbb{R}_+} f(\theta)$, it can be argued that $f(\bar{\theta}_k) - f_* = O_p(\sqrt{\delta}k^{-1})$ in the presence of strong convexity (see [20, Theorem 2]) that holds in the $\delta < \delta_1$ case. As a result, if $\delta < \delta_1$, it suffices to terminate after $O_p(\delta^{-1/2}\varepsilon^{-1})$ iterations in order to obtain an iterate $\bar{\theta}_k$ that satisfies $f(\bar{\theta}_k) - f_* \leq \varepsilon$. 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 2b asserts that $f(\beta, \lambda)$ satisfies strong convexity in the variable β for every fixed λ. This strong convexity in variable β holds even if $f(\beta, \lambda)$ 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 $h(\lambda) := \inf_{\beta\in B} f(\beta, \lambda)$ for any given λ and b) utilizes univariate line search for minimizing $h(\cdot)$ in a suitable interval.

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

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

where $(X_k)_{k\geq 1}$ are iid samples of P_0 and the step-sizes $(\alpha_k)_{k\geq 1}$ satisfy the requirement in Assumption 7 with $\tau \in (1/2, 1)$ and $\xi = 0$. Then it follows from the strong convexity characterization in Theorem 2b that $f(\bar{\beta}_k, \lambda) - h(\lambda) = O_p(\sqrt{\delta}\lambda^{-1}k^{-1})$ if $\delta < \delta_0$. With the ability to evaluate the function $h(\lambda) = \inf_{\beta\in B} f(\beta, \lambda)$ within desired precision, any standard line search method, such as triangle section method (see Algorithm 3 in [12]), 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 $[0, K_3]$. This is because, due to Proposition 1, we have that the interval $[0, K_3]$ contains optimal $\lambda_*(\beta)$ for every $\beta \in B$. It can be argued that the described approach results in iteration complexity of $O(\delta^{-1/2}\varepsilon^{-1}\text{poly}(\log \varepsilon^{-1}))$ to solve $\min f(\beta, \lambda)$ 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 Theorem 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, [16, 9]), or momentum based acceleration (see [1]). The strong convexity results in Theorem 2 could be used to establish improved rates of convergence for such extensions as well.

3.4. **SGD for nondifferentiable f.** The function $f(\cdot)$ need not be differentiable when the radius of ambiguity δ exceeds δ_0. 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 $\delta > \delta_0$ as it may not be the case that $\arg\min_\theta f(\theta)$ is contained in $\mathbb{W}$. In that case, with the characterization of the effective domain of f as in Lemma 2, define the family of closed convex sets, $(\mathbb{U}_\eta : \eta \geq 0)$ as,

$$\mathbb{U}_\eta := \{(\beta, \lambda) \in B \times \mathbb{R}_+ : \lambda \geq \lambda_{thr}(\beta) + \eta\}. \tag{18}$$

Let $\partial f(\beta, \lambda)$ and $\partial \ell_{rob}(\beta, \lambda; x)$, respectively, be the set of subgradients of $f(\cdot)$ and $\ell_{rob}(\,\cdot\,; x)$ at (β, λ). Then it follows from Lemma 4b that the set,

$$D(\beta, \lambda; x) := \operatorname{conv}\left\{ \begin{pmatrix} \ell'(\beta^T\tilde{x})\tilde{x} \\ \sqrt{\delta}\left(1 - g^2\beta^T A(x)^{-1}\beta\right) \end{pmatrix} : \begin{array}{c} \tilde{x} = x + \sqrt{\delta} g A(x)^{-1}\beta, \\ g \in \Gamma^*(\beta, \lambda; x) \end{array} \right\} = \partial \ell_{rob}(\beta, \lambda; x). \tag{19}$$

Similar to Proposition 3, Proposition 5 below helps in characterizing noisy subgradients of $f(\cdot)$.

Proposition 5. *Suppose that Assumptions 1, 2 hold and $\ell(\cdot)$ is continuously differentiable. 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 $E[h(\beta, \lambda; X)] \in \partial f(\beta, \lambda)$.*

The proof of Proposition 5 is available in Appendix A. Following Proposition 5, 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:

$$\theta_k := \Pi_{\mathbb{U}_\eta}\left(\theta_{k-1} - \alpha_k H_k\right) \quad \text{and} \quad \bar{\theta}_k = \left(1 - \frac{\xi + 1}{k + \xi}\right)\bar{\theta}_{k-1} + \frac{\xi + 1}{k + \xi}\theta_k, \tag{20}$$

where the step-size sequence $(\alpha_k)_{k \geq 1}$ satisfies Assumption 7 with $\tau = 1/2$ and H_k is computed as follows:

a) Generate a sample X_k independently from the distribution P_0;
b) Pick any $g \in \Gamma^*(\beta, \lambda; X_k)$ by solving the univariate search $\sup_{\gamma \in \mathbb{R}} F(\gamma, \beta_{k-1}, \lambda_{k-1}; X_k)$;
c) Let $\tilde{X}_k := X_k + \sqrt{\delta} g A(X_k)^{-1}\beta$, and take H_k as,

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

It is immediate from (19) that $H_k \in D(\beta_{k-1}, \lambda_{k-1}; X_k)$. Then due to Proposition 5, we have that $EH_k \in \partial f(\beta_{k-1}, \lambda_{k-1})$. Due to the convexity of $f(\cdot)$ characterized in Theorem 2a, we have the following rates of convergence for $f(\bar{\theta}_k) - f_*$, as $k \to \infty$. The proof of Proposition 6 is presented in Section 6.4.

Proposition 6. *Suppose that Assumptions 1 - 4 hold, $E\|X\|^4 < \infty$, $\xi \geq 1$ in (20) and $\tau = 1/2$ in Assumption 7. Then we have $f(\bar{\theta}_k) - f_* \leq \eta\sqrt{\delta} + O_p(k^{-1/2})$.*

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. Illustrative examples from supervised learning

In this section, we discuss how the DRO formulation (2) and the described algorithms can be utilized in the context of supervised learning, in which a response variable Y is typically present in addition to the predictor variables X in (2). We also report results of stylized numerical experiments that,

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 2) is relaxed.

The advantages of distributionally robust optimization formulations that utilize Wasserstein distances and optimal transport costs with Mahalanobis distances have been explored in [4, 14, 5, 31]. Therefore we restrict the focus our numerical experiments to studying the iteration complexity of the stochastic gradient descent scheme proposed in Section 3.2.

4.1. 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.

4.2. 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

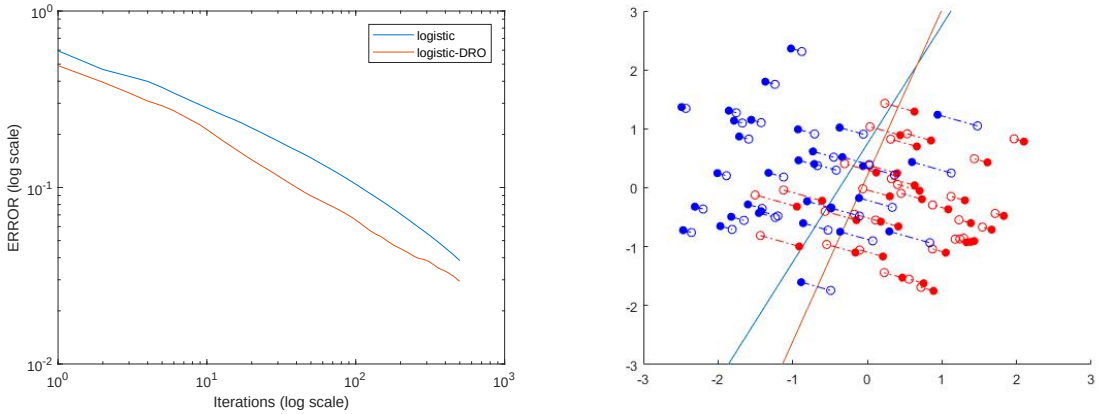
$$\ell(u; y) = \log\left(1 + \exp(-yu)\right).$$

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

In previous sections, several assumptions are imposed on the loss function $\ell(\cdot; y)$ and base line distribution P_0. Now we demonstrate they are natural assumptions on logistic regression.



(a) Convergence of loss function **(b)** Decision boundary and worst case distribution

FIGURE 1. Logistic Regression

Assumption 2. The logistic loss function is convex and asymptotically linear, so the growth condition is satisfied with $\kappa = 0$. In addition, as P_0 has finite support, it is immediate that $E_{P_0}\|X\|^2 < \infty$.

Assumption 4. For the first part, note that $\ell''(u;\pm 1) = (1/4)\cosh(u/2)^{-2} \leq 1/4$. For the second part, using the fact that P_0 are finitely supported and $0 < |\ell'(u;\pm 1)| < 1$, so the lower bound and upper bound on $E_{P_0}[\ell'(\beta^T X;Y)^2]$ always exist.

Assumption 5. The strong convexity in can be verified using the closed form expression $\ell''(u;\pm 1) = (1/4)\cosh(u/2)^{-2}$.

Assumption 6. This is satisfied 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 the predictors has a positive density with respect to the Lebesgue measure.

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 $n = 2048$, and the dimension of data is $d = 128$.

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 ($\delta > 0$). 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 7 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.

Then, we visualize the worst case distribution in Figure 1b, the number of data points is not too large: 32 data points on 2 dimensional space are generated for each class. In Figure 1b, we also visualize the decision boundary and worst case distribution corresponding to (nearly) optimal parameters. The solid points in different color denote the data in different classes. The blue line is the decision boundary implied by ordinary logistic regression. The red line is

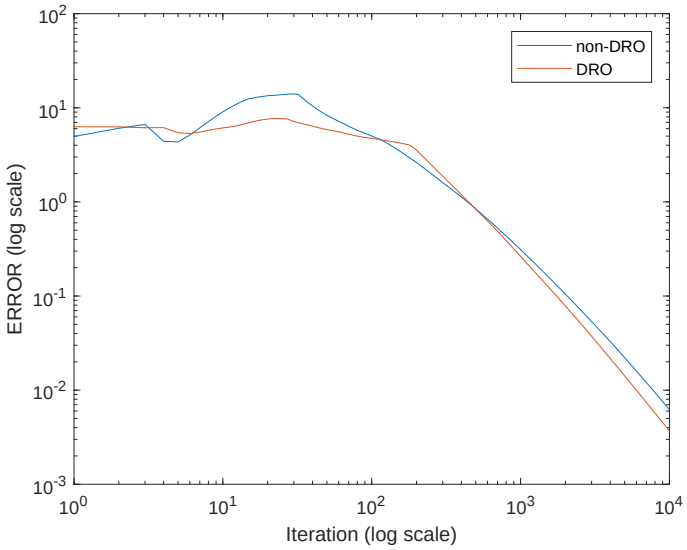


FIGURE 2. Convergence of loss function for Linear Regression.

the decision boundary implied by distributionally robust logistic regression. The dashed lines represent the optimal transport, and the small circles are the destinations of optimal transport (which also represents the worst case distribution).

4.3. **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 $P_n(dx) := \frac{1}{n}\sum_{i=1}^n \delta_{\{(X_i,Y_i)\}}(dx)$. Then, the distributionally robust linear regression problem is defined as

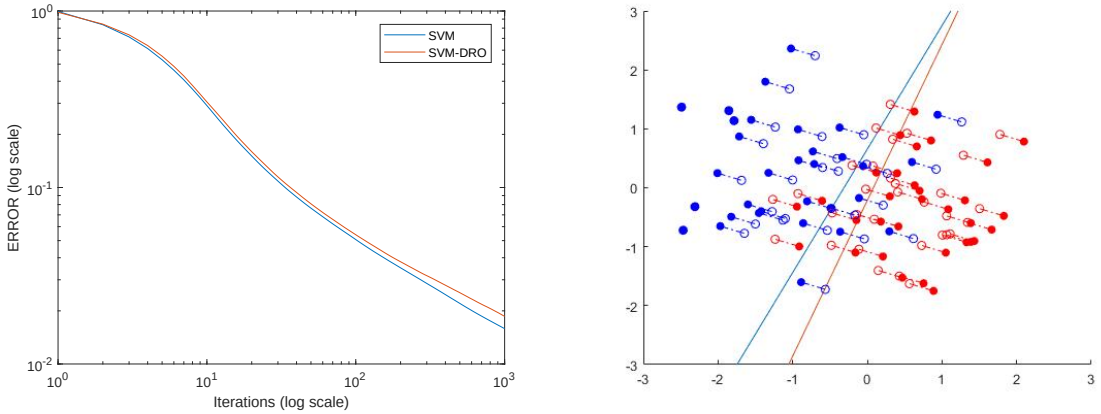
$$\inf_{\beta\in B} \sup_{P:D_c(P_n,P)\leq\delta} E_P\left[\ell(\beta^T X; Y)\right].$$

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.

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

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

Now we explain the setting of our numerical experiment in this example. The dimension of data is $d = 200$, and we randomly generate $n = 50$ training data points. 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 7. The plot of optimality gaps (ERROR) versus iterations for DRO-model and ordinary linear regression model is given in Figure 2.



(a) Convergence of loss function **(b)** Decision boundary and worst case distribution

FIGURE 3. Support Vector Machine

4.4. **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 problem,

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

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

Note that the hinge loss function is not continuously differentiable, so the theoretical result provided in previous sections can not be directly applied. However, it is still interesting to observe the performance of our algorithm when applied to the distributionally robust SVM. Interestingly, in the numerical experiment, the rate of convergence is as fast as the non-DRO benchmark algorithm.

In the numerical experiment, we use the same data as the example of logistic regression. Again, we set the learning rate of β to be same for DRO and non-DRO algorithms. Figure 3a 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, then we plot the optimality gap (ERROR) versus number of iterations in Figure 3a. Figure 3b visualizes the decision boundary and worst case distribution corresponding to (nearly) optimal parameters. The solid points in different color denotes the data in different classes. The blue line is the decision boundary implied by ordinary SVM. The red line is the decision boundary implied by distributionally robust SVM. The dashed lines represent the optimal transport, and the small circles are the destinations of optimal transport (which also represents the worst case distribution). Note that for hinge loss function, when a data is far from the decision boundary, then no transportation would happen. For the remaining data points, the size of optimal transport is same.

5. Conclusions

Our main objective in this paper has been to set the stage for algorithms and analysis of a flexible class of DRO problems. We are motivated by our belief that the choice of the distributional uncertainty region in DRO is crucial to fully exploit the advantages of robust decision rules which are informed by data-driven methods. We show that in the case of affine decision rules and convex loss functions, robustification with a flexible family of optimal transport costs introduces basically no additional computational complexity relative to the non-DRO counterpart (in terms of standard benchmark algorithms used to solve the non-DRO problem). We believe that some of our technical assumptions (such as twice differentiability) can be relaxed, as suggested by the performance observed in numerical experiments. This will be explored in future work.

Our philosophy is that by providing a general analysis of a flexible class of cost functions, modelers will be able to be free to choose a cost function that enhances out-of-sample performance in a convenient way relative to the needs of the modeler. Of course, an important problem that we leave untouched in this paper is, precisely, how to select a suitable cost function in a reasonable way, we expect to address this problem systematically in future work. A discussion in the context of classification tasks is given in [5].

6. Proofs of main results

6.1. **Bounds for dual optimizer $\lambda^*(\beta)$.** The main objective of this section is to derive bounds for any dual optimizer $\lambda^*(\beta)$ in $\arg\min_{\lambda\geq 0} f(\beta,\lambda)$. As we shall see later in this section, these bounds derived in Lemma 8 are crucial towards providing proofs for Lemma 1, 5 and Proposition 1. We begin with a proof of Lemma 3, which establishes the convexity of $\ell_{rob}(\beta,\lambda;x)$.

Proof of Lemma 3. Take any $\theta_1 := (\beta_1,\lambda_1)$ and $\theta_2 := (\beta_1,\lambda_1)$ in $B\times\mathbb{R}_+$. Given $\alpha\in[0,1]$, it follows from (8) that $\ell_{rob}(\alpha\theta_1+(1-\alpha)\theta_2;x)$ equals

$$\begin{aligned}&\sup_{\Delta\in\mathbb{R}^d}\left\{\ell\left((\alpha\beta_1+(1-\alpha)\beta_2)^T(x+\Delta)\right)-(\alpha\lambda_1+(1-\alpha)\lambda_2)\left(\Delta^TA(x)\Delta-\delta\right)\right\}\\&\quad=(\alpha\lambda_1+(1-\alpha)\lambda_2)\,\delta\\&\qquad+\sup_{\Delta\in\mathbb{R}^d}\left\{\ell\left(\alpha\beta_1^T(x+\Delta)+(1-\alpha)(\beta_2^T(x+\Delta))\right)-(\alpha\lambda_1+(1-\alpha)\lambda_2)\,\Delta^TA(x)\Delta\right\}.\end{aligned}\tag{21}$$

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 (21) is bounded from above by,

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

This observation, in conjunction with (21), 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 $\ell_{rob}(\,\cdot\,;x)$. □

Lemma 6 and 7 below, whose proofs are provided in Appendix A, are useful towards establishing the lower bound for $\lambda^*(\beta)$ in Lemma 8.

Lemma 6. *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 $g\in\Gamma^*(\beta,\lambda;x)$ we have,*

$$|g|\geq\frac{|\ell'(\beta^Tx)|}{2\lambda}. \tag{22}$$

Lemma 7. *Suppose that Assumptions 1, 2 are satisfied and $\beta\in B$. Then for any $\lambda_*(\beta)\in arg\min_{\lambda\geq 0}f(\beta,\lambda)$, we have $\Gamma^*(\beta,\lambda_*(\beta);x)\neq\varnothing$, for P_0-almost every $x\in\mathbb{R}^d$. Moreover,*

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

Lemma 8. *Suppose that Assumptions 1 - 4 hold. Then any minimizer $\lambda_*(\beta)\in arg\min_{\lambda\geq 0}f(\beta,\lambda)$ satisfies $\lambda_{\min}(\beta)\leq\lambda_*(\beta)\leq\lambda_{\max}(\beta)$, where*

$$\lambda_{\min}(\beta):=\frac{1}{2}\rho_{\max}^{-1/2}\|\beta\|\sqrt{E_{P_0}\left[\ell'(\beta^TX)^2\right]}\ \text{and}$$
$$\lambda_{\max}(\beta):=\rho_{\min}^{-1/2}\|\beta\|\sqrt{E_{P_0}\left[\ell'(\beta^TX)^2\right]}+\frac{1}{2}\sqrt{\delta}M\rho_{\min}^{-1}\|\beta\|^2.$$

Proof of Lemma 8. **Lower bound.** Combining the observations in Lemma 6 - 7 and the first order optimality condition that $\partial_+f(\beta,\lambda_*(\beta))/\partial\lambda\geq 0$, we obtain,

$$0\leq\frac{\partial_+f}{\partial\lambda}(\beta,\lambda_*(\beta))\leq\sqrt{\delta}\left(1-E_{P_0}\left[\beta^TA(X)^{-1}\beta\frac{\ell'(\beta^TX)^2}{4\lambda_*(\beta)^2}\right]\right).$$

Because of Assumption 1b, the above inequality results in,

$$\lambda_*(\beta)\geq\frac{1}{2}E_{P_0}^{1/2}\left[\ell'(\beta^TX)^2\beta^TA(X)^{-1}\beta\right]\geq\frac{1}{2}\rho_{\max}^{-1/2}\|\beta\|\sqrt{E_{P_0}\left[\ell'(\beta^TX)^2\right]}=:\lambda_{\min}(\beta).$$

Upper bound. As $\ell''(\cdot)\leq M$ due to Assumption 4, we have that $\ell_{rob}(\beta,\lambda;X)-\ell(\beta^TX)$ is bounded from above by,

$$\begin{aligned}&\sup_{\gamma\in\mathbb{R}}\left\{\ell\left(\beta^TX+\gamma\sqrt{\delta}\beta^TA(X)^{-1}\beta\right)-\ell\left(\beta^TX\right)-\lambda\sqrt{\delta}\beta^TA(X)^{-1}\beta\gamma^2\right\}\\&\quad\leq\sup_{\gamma\in\mathbb{R}}\left\{\ell'\left(\beta^TX\right)\sqrt{\delta}\beta^TA(X)^{-1}\beta\gamma+\frac{1}{2}M\left(\gamma\sqrt{\delta}\beta^TA(X)^{-1}\beta\right)^2-\lambda\sqrt{\delta}\beta^TA(X)^{-1}\beta\gamma^2\right\}\\&\quad=\frac{\sqrt{\delta}\beta^TA(X)^{-1}\beta[\ell'(\beta^TX)]^2}{(4\lambda-2M\sqrt{\delta}\beta^TA(X)^{-1}\beta)^+}.\end{aligned}$$

Next, since $\lambda_*(\beta)\sqrt{\delta}+E_{P_0}\left[\ell(\beta^TX)\right]\leq\inf_{\lambda\geq 0}E_{P_0}[\ell_{rob}(\beta,\lambda;X)]$, we use the above result and the bounds in Assumption 1b to write,

$$\begin{aligned}\lambda_*(\beta)&\leq\inf_{\lambda\geq 0}\left\{\lambda+\delta^{-1/2}E_{P_0}\left[\ell_{rob}(\beta,\lambda;X)-\ell(\beta^TX)\right]\right\}\\&\leq\inf_{\lambda>\frac{1}{2}\sqrt{\delta}M\rho_{\min}^{-1}\|\beta\|_2^2}\left\{\lambda+E_{P_0}\left[\frac{\beta^TA(X)^{-1}\beta[\ell'(\beta^TX)]^2}{4\lambda-2M\sqrt{\delta}\beta^TA(X)^{-1}\beta}\right]\right\}\\&\leq\inf_{\lambda>\frac{1}{2}\sqrt{\delta}M\rho_{\min}^{-1}\|\beta\|_2^2}\left\{\lambda+\frac{\rho_{\min}^{-1}\|\beta\|^2}{4\lambda-2M\sqrt{\delta}\rho_{\min}^{-1}\|\beta\|^2}E_{P_0}\left[\ell'(\beta^TX)^2\right]\right\}\end{aligned}$$

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 8. □

Recall that $\lambda'_{thr}(\beta)$ is the P_0-essential supremum of $M\sqrt{\delta}\beta^T A(x)^{-1}\beta/2$. Lemma 9 and Lemma 10 below are useful towards utlizing the above bounds on $\lambda_*(\beta)$ to provide proofs of Lemma 1, 5, Proposition 1 and Theorem 2.

Lemma 9. *Suppose that Assumptions 1 - 4 hold. 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 $x \in \mathbb{R}^d$.*

Proof of Lemma 9. 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 4), it follows from the definition of $F(\cdot)$ in (7) that

$$\begin{aligned}\frac{\partial^2 F}{\partial \gamma} &= \sqrt{\delta}\beta^T A(x)^{-1}\beta\left(\ell''(\beta^T x + \sqrt{\delta}\gamma\beta^T A(x)^{-1}\beta)\sqrt{\delta}\beta^T A(x)^{-1}\beta - 2\lambda\right) \\ &\leq \sqrt{\delta}\beta^T A(x)^{-1}\beta\left(M\sqrt{\delta}\beta^T A(x)^{-1}\beta - 2\lambda\right) \leq -2\varepsilon\delta\beta^T A(x)^{-1}\beta,\end{aligned} \tag{23}$$

for P_0-almost every x. This proves the desired strong concavity. □

Proof of Lemma 1. It follows from Lemma 8 and Assumption 5 that $\lambda^*(\beta) \geq 2^{-1}(\underline{L}\rho_{\max}^{-1})^{1/2}\|\beta\|$. Due to Assumptions 1b and 3, we also have that,

$$\lambda'_{thr}(\beta) \leq 2^{-1}\sqrt{\delta}M\rho_{\min}^{-1}\|\beta\|^2 < 2^{-1}\sqrt{\delta_0}M\rho_{\min}^{-1}R_\beta\|\beta\| \leq 2^{-1}(\underline{L}\rho_{\max}^{-1})^{1/2}\|\beta\|, \tag{24}$$

where the last inequality is immediate from the definition of δ_0 in Section 2.2.1. Therefore $\lambda'_{thr}(\beta) < \lambda^*(\beta)$ when $\delta < \delta_0$. This completes the proof of Lemma 1. □

Proof of Proposition 1. For a given $\beta \in B$, it follows from Lemma 8 that any optimal $\lambda_*(\beta)$ lies in the interval $[\lambda_{\min}(\beta), \lambda_{\max}(\beta)]$. Recalling the definitions of $R_\beta, \overline{L}$ and $\underline{L}$ from Assumptions 3, 5, we have from Lemma 8 that $\lambda_{\min}(\beta) \geq K_1\|\beta\|$ and $\lambda_{\max}(\beta) \leq K_2\|\beta\|$, where

$$K_1 := \frac{1}{2}\sqrt{\underline{L}\rho_{\max}^{-1}} \quad \text{and} \quad K_2 := \frac{1}{2}\sqrt{\delta_0}MR_\beta\rho_{\min}^{-1} + \sqrt{\rho_{\min}^{-1}\overline{L}}.$$

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

Proof of Lemma 5. Since $\ell(\cdot)$ is proper, we have that $F(\cdot, \beta, \lambda; x)$ is proper. For any $(\beta, \lambda) \in \mathbb{W}$, we have $\lambda \geq K_1\|\beta\| = 2^{-1}(\underline{L}\rho_{\max}^{-1})^{1/2}\|\beta\|$. Then it follows from (24) that $\lambda > \lambda'_{thr}(\beta)$. This verifies part a). The desired strong concavity in part b) is now simply a consequence of Lemma 9. This completes the proof of Lemma 5. □

6.2. **Proof of Theorem 2.** As $f(\beta, \lambda) := E_{P_0}[\ell_{rob}(\beta, \lambda; X)]$, the convexity of $f(\cdot)$ follows as a consequence of Lemma 3 and linearity of expectations. This verifies Theorem 2a. In order to prove the subsequent strong convexity claims, we first identify the Hessian of $f(\cdot)$.

Due to Lemma 9, there exists a set $D \subseteq \mathbb{R}^d$ with $P_0(X \in D) = 1$ such that the map $\gamma \mapsto F(\gamma, \beta, \lambda; x)$ is strongly concave for every for every $\beta \in B, \lambda > \lambda'_{thr}(\beta)$ and $x \in D$. Therefore $\Gamma^*(\beta, \lambda; x)$ is singleton for every $x \in D$. If we use $g(\beta, \lambda; x)$ to denote the unique maximizer in $\Gamma^*(\beta, \lambda; x)$ for $x \in D$, then the map g is measurable (see [3, Proposition 7.50b]).

While the gradient of $\ell_{rob}(\beta,\lambda;x) = F(g(\beta,\lambda;x),\beta,\lambda;x)$ for $x \in D$ can be computed as in (14), computing the Hessian will require the knowledge of $\partial g/\partial\beta$ and $\partial g/\partial\lambda$. We compute this information using implicit function theorem. For this purpose, define

$$\tilde{x} := x + \sqrt{\delta} g(\beta,\lambda;x) A(x)^{-1}\beta, \qquad \bar{x} := x + 2\sqrt{\delta} g(\beta,\lambda;x) A(x)^{-1}\beta, \tag{25}$$

$$\varphi(\beta,\lambda;x) := 2\lambda - \sqrt{\delta}\beta^T A(x)^{-1}\beta \ell''(\beta^T\tilde{x}) \quad \text{and} \quad \varphi_{\min} := \sqrt{L}\rho_{\max}^{-1/2} - \sqrt{\delta} R_\beta M \rho_{\min}^{-1}. \tag{26}$$

for any $\beta \in B, \lambda > \lambda'_{thr}(\beta)$ and $x \in D$. It follows from the definition of $\varphi_{\min} > 0$ in (26) and δ_0 in Section 2.2.1 that $\varphi_{\min} > 0$ when $\delta < \delta_0$.

Hereafter, we often suppress the arguments $(\beta,\lambda;x)$ while writing the functions such as $g(\beta,\lambda;x)$ and $\varphi(\beta,\lambda;x)$, in order to reduce clutter in the resulting expressions; for example, we simply write φ and g, respectively, for $\varphi(\beta,\lambda;x)$ and $g(\beta,\lambda;x)$.

Lemma 10. *Suppose that Assumptions 1 - 4 are satisfied. Consider any (β,λ) such that $\lambda > \lambda'_{thr}(\beta)$ and $x \in D$. Then,*

a) $\varphi(\beta,\lambda;x) > 0$;

b) *with $g(\beta,\lambda;x) \in \Gamma^*(\beta,\lambda;x)$ and $\tilde{x},\bar{x}$ defined as in (25), we have*

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

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

Proof of Lemma 10. Observe that $\varphi > 0$ because of the strong concavity of $F(\cdot,\beta,\lambda;x)$:

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

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

$$\ell'(\beta^T x + \sqrt{\delta} g(\beta,\lambda;x)\beta^T A(x)^{-1}\beta) - 2\lambda g(\beta,\lambda;x) = 0. \tag{27}$$

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

$$\frac{\partial g}{\partial\beta}(\beta,\lambda;x) = -\frac{\partial^2 F/\partial\beta\partial\gamma(g(\beta,\lambda;x),\beta,\lambda;x)}{\partial^2 F/\partial\gamma^2(g(\beta,\lambda;x),\beta,\lambda;x)} = \frac{\ell''(\beta^T\tilde{x})}{\varphi(\beta,\lambda;x)}\bar{x} \tag{28}$$

$$\frac{\partial g}{\partial\lambda}(\beta,\lambda;x) = -\frac{\partial^2 F/\partial\lambda\partial\gamma(g(\beta,\lambda;x),\beta,\lambda;x)}{\partial^2 F/\partial\gamma^2(g(\beta,\lambda;x),\beta,\lambda;x)} = -\frac{2g(\beta,\lambda;x)}{\varphi(\beta,\lambda;x)}. \tag{29}$$

Following these expressions for gradient of g, the hessian of $\ell_{rob}(\cdot;x)$ in the statement of Lemma 10 follow from elementary rules of differentiation. □

When $\delta < \delta_0$, we have from Lemma 5a that any (β,λ) in $\mathbb{W}$ satisfies $\lambda > \lambda'_{thr}(\beta) > \lambda_{thr}(\beta)$. Therefore it follows from the bounds for g and $\ell_{rob}(\beta,\lambda;x)$ in Lemma 14 that $E_{P_0}[g^2(\beta,\lambda;x)], E\|\bar{X}\|^2$ are all bounded from above for $(\beta,\lambda) \in \mathbb{W}$. Then due to the boundedness of $\ell''(\cdot)$, it follows as an application of dominated convergence theorem that $\nabla^2_\theta f(\theta) = E_{P_0}[\nabla^2_\theta \ell_{rob}(\theta;X)]$ for $\theta \in \mathbb{W}$.

Proof of Theorem 2b. Recall from Lemma 6 that $|g(\beta,\lambda;x)| \geq |\ell'(\beta^T x)|/(2\lambda)$. Therefore it follows from the expression of $\partial^2\ell_{rob}/\partial\beta^2$ from Lemma 10 that

$$\frac{\partial^2 f}{\partial\beta^2}(\beta,\lambda) = E_{P_0}\left[\frac{\partial^2\ell_{rob}}{\partial\beta^2}(\beta,\lambda;X)\right] \geq \sqrt{\delta}\frac{E_{P_0}[\ell'(\beta^T X)^2]}{2\lambda}\rho_{\max}^{-1} \geq \sqrt{\delta}\frac{\kappa_1}{\lambda},$$

where $\kappa_1 := 2^{-1}\underline{L}\rho_{\max}^{-1}$, thus proving Theorem 2b.

Proof of Theorem 2c. Next, let $\nabla^2 f(\beta,\lambda;x)$ denote the Hessian matrix of $f(\beta,\lambda;x)$ with respect to the variables β and λ. Let $B(x)$ be a $(d+1)\times(d+1)$-matrix defined as follows:

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

To estimate the strong convexity coefficient of $\ell_{rob}(\beta,\lambda;x)$, we aim to find a $\Lambda(x)\in\mathbb{R}$ as large as possible such that $\nabla^2\ell_{rob}(\beta,\lambda;x) - \Lambda(x)B(x) \succeq 0$. A possible (but not necessarily tight) choice of $\Lambda(x)$ is given in the following lemma, whose proof is presented in Appendix A.

Lemma 11. *For $(\beta,\lambda)\in\mathbb{W}$ and P_0 almost every x, we have $\nabla^2\ell_{rob}(\beta,\lambda;x) - \Lambda(x)B(x)\succeq 0$, where $\Lambda(x)$ defined as*

$$\Lambda(x) := \frac{4\left(\beta^T\tilde{x}\right)^2\ell''(\beta^T\tilde{x})}{1+\bar{x}^TA(x)\bar{x}\ell''(\beta^T\tilde{x})/(\sqrt{\delta}g^2\varphi)}\frac{1}{2\lambda\varphi + 4\beta^TA(x)^{-1}\beta}.$$

Lemma 12. *For $(\beta,\lambda)\in\mathbb{V}$, we have $\varphi\geq\varphi_{\min}\|\beta\|$ and*

$$\frac{|\ell'(\beta^Tx)|}{2K_2\|\beta\|} \leq |g(\beta,\lambda;x)| \leq \frac{|\ell'(\beta^Tx)|}{\varphi_{\min}\|\beta\|}.$$

Proof of Lemma 12. Observe that, as a consequence of the mean value theorem, the first order optimality condition (27) means that $g = \ell'(\beta^Tx)/(2\lambda - \sqrt{\delta}\beta^TA(x)^{-1}\beta\ell''(\eta))$, for some η between the real numbers β^Tx and $\beta^T\tilde{x}$. Since $\ell''(\cdot)\leq M$ and $\delta\leq\delta_0$, we have that $2\lambda - \sqrt{\delta}\beta^TA(x)^{-1}\beta\ell''(\eta) \geq 2\lambda_{\min}(\beta) - \sqrt{\delta}\beta^TA(x)^{-1}\beta\ell''(\eta) \geq \varphi_{\min}\|\beta\|$. Lemma 12 follows because $\varphi_{\min}\|\beta\| \leq 2\lambda - \sqrt{\delta}\beta^TA(x)^{-1}\beta\ell''(\eta) \leq 2\lambda \leq K_2\|\beta\|$. □

Using the bounds of $|g|$ and φ from Lemma 12 along with other immediate bounds such as $\varphi\leq 2\lambda$, $\lambda\in(K_1\|\beta\|, K_2\|\beta\|)$ and $\beta^TA(x)^{-1}\beta\leq\rho_{\min}^{-1}\|\beta\|^2$, the expression for $\Lambda(x)$ from Lemma 11 simplifies to,

$$\begin{aligned}\Lambda(x) &= \frac{4\sqrt{\delta}(g\beta^T\tilde{x})^2}{2\lambda\sqrt{\delta}g^2/\ell''(\beta^Tx) + \bar{x}^TA(x)\bar{x}}\cdot\frac{1}{2\lambda + 4\beta^TA(x)^{-1}\beta/\varphi} \qquad (30)\\ &\geq \frac{4\sqrt{\delta}(\beta^T\tilde{x}\ell'(\beta^Tx)/(2K_2\|\beta\|))^2}{2K_2\sqrt{\delta}\ell'(\beta^Tx)^2/(\varphi_{\min}^2\|\beta\|\ell''(\beta^Tx)) + \bar{x}^TA(x)\bar{x}}\cdot\frac{1}{2K_2\|\beta\| + 4\rho_{\min}^{-1}\|\beta\|^2/(\varphi_{\min}\|\beta\|)}\\ &\geq \sqrt{\delta}C_0\frac{\|\beta\|^{-2}\left(\beta^T\tilde{x}\ell'(\beta^Tx)\right)^2}{2K_2\sqrt{\delta}\varphi_{\min}^{-2}\ell'(\beta^Tx)^2/\ell''(\beta^T\tilde{x}) + \bar{x}^TA(x)\bar{x}\|\beta\|}, \qquad (31)\end{aligned}$$

where $C_0 := (2K_2 + 4\varphi_{\min}^{-1}\rho_{\min}^{-1})^{-1}$. Next, since $\beta^T\tilde{x} = \beta^Tx + \sqrt{\delta}g\beta^TA(x)^{-1}\beta$, we obtain from the bounds in Lemma 12 that

$$\begin{aligned}|\beta^T\tilde{x}\ell'(\beta^Tx)| &\geq |\beta^Tx\ell'(\beta^Tx)| - \sqrt{\delta}|g\ell'(\beta^Tx)|\beta^TA(x)^{-1}\beta\\ &\geq |\beta^Tx\ell'(\beta^Tx)| - \sqrt{\delta}\frac{\ell'(\beta^Tx)^2}{\varphi_{\min}\|\beta\|}\|\beta\|^2\rho_{\min}^{-1} \geq \left(C - \frac{4\sqrt{\delta\bar{L}}}{p\varphi_{\min}\rho_{\min}}\right)\|\beta\|, \qquad (32)\end{aligned}$$

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

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

where the constants C and p are given by Assumption 6. Since $E_{P_0}[\ell'(\beta^T X)^2] \leq \overline{L}$ for any $\beta \in \Xi$, we have from Markov's inequality that $\inf_{\beta\in\Xi} P_0(X \in A_2) \geq 1 - p/4$. Consequently, it follows from Assumption 6 and union bound that $\inf_{\beta\in\Xi} P_0(X \in A_1 \cap A_2) \geq 3p/4$.

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 $\delta < \delta_1 \leq C^2 p^2 \rho_{\min}^2 \rho_{\max}^{-1} \underline{L}\overline{L}^{-2}/256$, we have

$$C - 4\sqrt{\delta}\overline{L} p^{-1} \varphi_{\min}^{-1} \rho_{\min}^{-1} \geq C/2. \tag{33}$$

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 (31) 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 $\tilde{x} := x + \sqrt{\delta} g(\beta, \lambda; x) A(x)^{-1}\beta$ and $\bar{x} := x + 2\sqrt{\delta} g(\beta, \lambda; x) A(x)^{-1}\beta$, it follows from the bounds of $|g|$ in Lemma 12 that

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

and similarly, $\|\tilde{x}\| \leq C_2$ for $x \in A_2 \cap A_3$. Since $\|\beta^T\tilde{x}\| \leq R_\beta C_2 < \infty$ 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 (31) 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 (31),(32) and (33), we obtain that

$$\Lambda(x) \geq \sqrt{\delta} C_5 \mathbf{1}_{\{x\in A_1\cap A_2\cap A_3\}}$$

for $C_5 := (1/2)C_0 C C_4^{-1}$.

Finally, since $P_0(A_1 \cap A_2 \cap A_3) \geq p/2$, we have $E_{P_0}[\Lambda(X)B(X)] \succeq \sqrt{\delta}\kappa\mathbb{I}_{d+1}$ where $\kappa := pC_5\rho_{\max}^{-1}/2$. This verifies Theorem 2c, thus completing the proof of Theorem 2. □

Remark 2. Suppose that $C = 0$ is the only nonnegative number for which the probability requirement in Assumption 6 is satisfied. In this case, we have from the upper bound for g in Lemma 12 that $g\beta^T X = 0$, P_0 almost surely. As a result, the numerator of $\Lambda(x)$ in the right hand side of (30) 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 nonnegative constant κ_3. Since $\delta^{3/2} = o(\sqrt{\delta})$ as $\delta \to 0$, it is not possible to derive a positive constant κ_2 that is not dependent on δ as in the statement of Theorem 2c.

6.3. **Proofs of results on structure of the worst case distribution.** In this section we provide proofs of Theorem 3 and Proposition 2 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 3. Recall from Assumption 2 that $\ell(u)$ is convex and grows quadratically or sub-quadratically as $|u| \to \infty$. Therefore there exists $\lambda \geq 0$ such that $f(\beta, \lambda) < \infty$, and subsequently, $\inf_\lambda f(\beta, \lambda) < \infty$. According to Theorem 1, there exist a dual optimizer, $\lambda_*(\beta)$ in $\arg\min_{\lambda\geq 0} f(\beta, \lambda)$ for any $\beta \in B$.

a) When $\lambda_*(\beta) = 0$: We have $\inf_{\beta,\lambda} f(\beta, \lambda) = f(\beta, 0) = \sup_{u\in\mathbb{R}} \ell(u)$. Due to the convexity of $\ell(\cdot)$, the finiteness of the optimal value $f(\beta, 0) = \sup_u \ell(u)$ 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 $f(\cdot)$ in Lemma 2 that $f(\beta,\lambda)=\infty$ when $\lambda<\lambda_{thr}(\beta)$. Therefore, $\lambda^*(\beta)\geq\lambda_{thr}(\beta)$.

c) When $\lambda_*(\beta)>\lambda_{thr}(\beta)$: Recall from Lemma 4b the expressions for $\partial_+\ell_{rob}/\partial\lambda$ and $\partial_-\ell_{rob}/\partial\lambda$. Further we have $f(\beta,\lambda)<\infty$ 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 $\partial_+f/\partial\lambda$ and $\partial_-f/\partial\lambda$ satisfy,

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

for $(\beta,\lambda)\in\mathbb{U}_1$. Since $\lambda_*(\beta)>\lambda_{thr}(\beta)$, we have from Lemma 14a 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,

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

Since $\lambda_*(\beta)\in\arg\min_{\lambda\geq0}f(\beta,\lambda)$, we have from the first order optimality condition that $\partial_+f/\partial\lambda(\beta,\lambda_*(\beta))\geq0$ and $\partial_-f/\partial\lambda(\beta,\lambda_*(\beta))\leq0$. Thus $\underline{c}=E_{P_0}[G_-^2\beta^TA(X)^{-1}\beta]\leq1$ and $\overline{c}=E_{P_0}[G_+^2\beta^TA(X)^{-1}\beta]\geq1$. With $G:=ZG_-+(1-Z)G_+$ and Z being an independent Bernoulli random variable with $P(Z=1)=p$, we have that $E_{P_0}[G^2\beta^TA(X)^{-1}\beta]=1$. In addition, since $G\in\Gamma^*(\beta,\lambda;X)$ P_0-a.s., we have that

$$X^*\in\arg\max_{x'\in\mathbb{R}^d}\left\{\ell(\beta^Tx')-\lambda_*(\beta)c(X,x')\right\}\quad\text{and}\quad E\left[c(X,X^*)\right]=E[(\sqrt{\delta}G)^2\beta^TA(X)^{-1}\beta]=\delta.$$

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

d) When $\lambda_*(\beta)=\lambda_{thr}(\beta)$: The worst case distribution $P_*(\beta)$ attaining the supremum in $\sup_{P:D_c(P,P_0)\leq\delta}E_P[\ell(\beta^TX)]$ 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 $f(\beta,\lambda)=\lambda\sqrt{\delta}$ when $\lambda\geq\lambda_{thr}(\beta)$. Therefore $\lambda_*(\beta)=\lambda_{thr}(\beta)=\sqrt{\delta}$ and the dual optimal value $f(\beta,\lambda_*(\beta))=\delta$. However, this value is not attainable by $E_P[\ell(\beta^TX)]$ 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 $E_P[\ell(\beta^TX)]<\delta$ as in the following series of inequalities:

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

e) When $\lambda_*(\beta)>\lambda'_{thr}(\beta)$: In this case, it follows from Lemma 9 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 $E[G^2\beta^TA(X)^{-1}\beta]=1$. Therefore $E[c(X,X^*)]=\delta$. Moreover, the above described uniqueness in optimizer means that $X^*=X+\sqrt{\delta}GA(X)^{-1}\beta$ is

the unique element in $\arg\max_{x'\in\mathbb{R}^d}\{\ell(\beta^T x') - \lambda_*(\beta)c(X,x')\}$, P_0-almost surely. 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 Proposition 2. 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 $\min_{\lambda\geq 0} f(\beta,\lambda)$ 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 2b. For every $\delta<\delta_1$, we have from Lemma 1 that that $\lambda_*(\delta)>\lambda'_{thr}(\beta)$. As a result, it follows from Theorem 3e that,

a) 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 $\lambda>\lambda'_{thr}(\beta)$; and
b) since $E[c(X,X^*_\delta)]=\delta$, $g(\delta,\lambda_*(\delta);X)$ satisfies that $E_{P_0}[g^2(\delta,\lambda_*(\delta);X)\beta^T A(X)^{-1}\beta]=1$.

Following the implicit function theorem application in Lemma 10, we obtain that

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

where g and φ in the right hand side denote, respectively, $g(\delta,\lambda_*;x)$ and $\varphi(\beta,\lambda_*;x) := 2\lambda_*(\delta) - \sqrt{\delta}\beta^T A(X)^{-1}\beta\ell''(\beta^T X^*_\delta) > 0$ (see Lemma 10a).

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 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 Lemma 10, 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 Lemma 12). Moreover, $\beta^T A(X)^{-1}\beta\leq R_\beta\rho_{\min}^{-1}\|\beta\|$ and $\ell''(\cdot)\in(0,M]$ (see Assumptions 1b and 4). As a result, we obtain that

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

where the last equality follows from the definitions of δ_0 and $\varphi_{\min}$ in (26). 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 (27) 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 2b - 2d are verified. This completes the proof of Proposition 2. □

6.4. **Proofs of results on rates of convergence.** Lemma 13 below, establishing finite second moments for the gradients (or) subgradients utilized in SGD schemes, is useful towards proving Propositions 4 and 6. Recall the definition of $D(\beta,\lambda;X)$ in (19).

Lemma 13. *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 $E_{P_0}\|h(\theta;X)\|^2 \le G_\eta$ for any $\theta \in \mathbb{U}_\eta$.*

The proof of Lemma 13 is presented in Appendix A.

Proof of Proposition 4. a) When $\delta < \delta_0$, it follows from Lemma 4b 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 $\lambda > \lambda'_{thr}(\beta) \ge \lambda_{thr}(\beta)$ for every $(\beta,\lambda) \in \mathbb{W}$ (see Lemma 5a), it follows from Lemma 13 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 [29] that $E[f(\theta_k)] - f_* = O(k^{-1/2}\log k)$ and $E[f(\bar\theta_k)] - f_* = O(k^{-1/2})$, 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 $\theta_* := \arg\min f(\theta)$ (see Theorem 2c) that there exists $\varepsilon > 0$ satisfying $(\theta-\theta_*)^T\nabla_\theta f(\theta) \ge \kappa_2\sqrt{\delta}\|\theta-\theta_*\|^2$ for all $\theta \in \mathbb{V}$ and $\|\theta-\theta_*\| \le \varepsilon$. Further, due to the uniqueness of the minimizer, we also have $(\theta-\theta_*)^T\nabla_\theta f(\theta) > 0$. Similar to Part a), as $\lambda > \lambda'_{thr}(\beta) \ge \lambda_{thr}(\beta)$ for every $(\beta,\lambda) \in \mathbb{W}$ (see Lemma 5a), we have due to Lemma 13 that $\sup_{\theta\in\mathbb{W}} E\|\nabla_\theta \ell_{rob}(\theta;X)\|^2 < \infty$. Taylor's expansion of $\nabla_\theta f(\theta)$ results in,

$$\|\nabla_\theta f(\theta) - \nabla^2_\theta f(\theta_*)^T(\theta-\theta_*)\| = o\left(\|\theta-\theta_*\|\right), \tag{34}$$

for $\theta \in \mathbb{W}$. With these conditions being satisfied, it follows from [25, Theorem 2] that

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

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

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

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

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

Now it follows from the local joint strong convexity of $f(\cdot)$ in Theorem 2c and (34) that

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

This completes the proof of Proposition 4. □

Proof of Proposition 6. Due to the characterization of subgradients $\partial f(\beta,\lambda)$ in Proposition 5, we have that $\partial_+ f/\partial\lambda(\beta,\lambda) \leq \sqrt{\delta}$ for every $(\beta,\lambda) \in \mathbb{U}$. Recalling the definitions of $\mathbb{U}_\eta$ in (18) and $\mathbb{U}$ in (12), the above reasoning leads to concluding that, $f(\beta,\lambda+\varepsilon) - f(\beta,\lambda) \leq \varepsilon\sqrt{\delta}$ for any $\varepsilon > 0$. Let $(\beta_*,\lambda_*) \in \inf_{(\beta,\lambda)\in\mathbb{U}} f(\beta,\lambda)$. Then

$$\inf_{\theta\in\mathbb{U}_\eta} f(\theta) - f_* \leq f(\beta_*,\lambda_*+\eta) - f(\beta_*,\lambda_*) \leq \eta\sqrt{\delta} \tag{35}$$

It follows from Lemma 13 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 [29] that $E[f(\theta_k)] - \inf_{\theta\in\mathbb{U}_\eta} f(\theta) = O(k^{-1/2}\log k)$ and $E[f(\bar{\theta}_k)] - \inf_{\theta\in\mathbb{U}_\eta} f(\theta) = O(k^{-1/2})$. Combining this with the observation in (35), we obtain that $E[f(\bar{\theta}_k)] - f_* \leq \eta\sqrt{\delta} + O(k^{-1/2})$. As in the proof of Proposition 4a, the conclusion in Proposition 6 follows as a consequence of Markov's inequality. □

Appendix

Appendix A. Proofs of technical results

Lemma 14. *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) *$\ell_{rob}(\beta,\lambda;x) \leq \lambda\sqrt{\delta} + C_2(1+\varepsilon+\varepsilon^{-1})(1+|\beta^T x|)^2$.*

Proof of Lemma 14. a) Given $\varepsilon > 0$, it follows from the growth condition in Assumption 2 that there exist a positive constant C_ε satisfying $\ell(u) \leq (\kappa+\varepsilon/2)u^2 + C_\varepsilon$ 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 $F(g,\beta,\lambda;x) \geq F(0,\beta,\lambda;x)$. Recalling the definition of $F(\cdot,\beta,\lambda;x)$ from (7), the above inequality results in,

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

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

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

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

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

Since $\sqrt{x} \leq 1+x$ for $x \geq 0$, the above inequality verifies Part a) of Lemma 14.

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

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

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):

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

This verifies Part b) of Lemma 14. □

Proof of Lemma 2. 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 2.

To verify the inclusions in the final statement of Lemma 2 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 $\{(\beta,\lambda) : \beta \in B, \lambda \geq \lambda_{thr}(\beta)\}$. On the other hand, if $\lambda > \lambda_{thr}(\beta)$, we have $\lambda \geq (\kappa+\varepsilon)\sqrt{\delta}\beta^T A(x)^{-1}\beta$ for some $\varepsilon > 0$, P_0-almost every x. Since $\|\beta\| \leq R_\beta$ and $E\|X\|^2 < \infty$, it follows Lemma 14b that $f(\beta,\lambda) = E_{P_0}[\ell_{rob}(\beta,\lambda;X)] < \infty$. Therefore $\{(\beta,\lambda) : \beta \in B, \lambda > \lambda_{thr}(\beta)\}$ is contained in $\mathbb{U}$. This completes the proof of Lemma 2. □

Proof of Lemma 4. a) Since $\lambda > \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta$ for P_0-almost surely every x, Lemma 4a follows directly from Lemma 2.

b) Consider any fixed $x \in \mathbb{R}^d, C_3 < \infty$ and $\eta > 0$. Define the set $A := \{(\beta,\lambda) \in B \times \mathbb{R}_+ : \|\beta\| < C_3,\ \lambda \geq \lambda_{thr}(\beta)+\eta\}$. Then for $(\beta,\lambda) \in A$, we have the following two conditions satisfied: i) $\lambda \geq (\kappa+\varepsilon)\sqrt{\delta}\beta^T A(x)^{-1}\beta$ for some $\varepsilon > 0$, for P_0-almost every x; and ii) $\beta^T A(x)^{-1}\beta$ is bounded away from zero (due to Assumption 1). Therefore, for any (β,λ) in we have from Lemma 14a 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, $f(\beta,\lambda;x) = \sup_{\gamma\in[-C_x,C_x]} F(\gamma,\beta,\lambda;x)$.

Next, we observe that for $j \in \{1,\ldots,d\}$,

$$\frac{\partial}{\partial\beta_j}F(\beta,\lambda,\gamma;x) = \ell'(\beta^T x + \gamma\sqrt{\delta}\beta^T A(x)^{-1}\beta)(x_j + \gamma\sqrt{\delta}A(x)^{-1}\beta_j),$$
$$\frac{\partial}{\partial\lambda}F(\beta,\lambda,\gamma;x) = \sqrt{\delta}\big(1-\gamma^2\beta^T A(x)^{-1}\beta\big).$$

where x_j is the jth element of x, and β_j is the jth element of β. Therefore, according to Envelope theorem [18, Corollary 4], we arrive at the following two conclusions: i) when $(\beta,\lambda) \in A$, the functions $\lambda \mapsto f(\beta,\lambda;x)$, $\beta_j \mapsto f(\beta,\lambda;x)$ are absolutely continuous, and have left and right derivative given by (13a) - (13d); and ii) the partial derivatives exist as in (14) if and only if the respective sets $\{\partial F/\partial\beta_j(g,\beta,\lambda;x) : g \in \Gamma^*(\beta,\lambda;x)\}$, $\{\partial F/\partial\lambda(g,\beta,\lambda;x) : g \in \Gamma^*(\beta,\lambda;x)\}$ are singleton. Since these expressions hold for any $C_3,\eta \in (0,\infty)$, Lemma 4b stands verified. □

Proof of Proposition 5. Let $g(\beta,\lambda;X)$ be such that $g(\beta,\lambda;X) \in \Gamma^*(\beta,\lambda;X)$ and

$$h(\beta,\lambda;X) = \begin{pmatrix} \ell'(\beta^T\tilde{X})\tilde{X} \\ \sqrt{\delta}\left(1-g^2(\beta,\lambda;X)\beta^T A(X)^{-1}\beta\right)\end{pmatrix} \qquad P_0-\text{a.s.}, \tag{36}$$

where $\tilde{X} := X + \sqrt{\delta}g(\beta,\lambda;X)A(X)^{-1}\beta$. Since $\ell(\cdot)$ is convex, continuously differentiable with at most quadratic growth (see Assumption 2), there exist positive constants C_0, C_1 such that $|\ell'(u)| \leq C_0 + C_1|u|$. As $E\|X\|^2 < \infty$, it follows from Lemma 14a that $E[g^2(\beta,\lambda;X)], E\|\tilde{X}\|^2$ and

$E[\ell'(\beta^T\tilde{X})^2]$ are all finite. Then due to Cauchy-Schwarz inequality, we have that $Eh(\beta,\lambda;X)$ is well-defined. Here we have used that $\beta^T A(x)^{-1}\beta$ is bounded for P_0-almost every x (see Assumption 1b). Now, since $h(\beta,\lambda;X)\in\partial\ell_{rob}(\beta,\lambda)$ (see (19)), we have that

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

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

Proof of Lemma 13. 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 (36) 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 5, we have $|\ell'(u)|\leq C_0+C_1|u|$ as a consequence of convexity, continuous differentiability and at most quadratic growth of $\ell(\cdot)$. Since $E\|X\|^4<\infty$, $\beta^T A(x)^{-1}\beta$ is bounded for P_0-almost every x, $\lambda>\lambda_{thr}(\beta)+\eta$, and $\|\beta\|\leq R_\beta$, it follows from Lemma 14a 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-Schwarz inequality, we have that $\sup_{(\beta,\lambda)\in\mathbb{U}_\eta}E\|h(\beta,\lambda;X)\|^2<\infty$. □

Proof of Lemma 6. If $\ell'(\beta^T x)=0$, inequality (22) is trivial. Thus, in order to prove (22), it suffices to consider the case where $\ell'(\beta^T x)$ is strictly positive or strictly negative. Note that $\Gamma^*(\beta,\lambda;x)\neq\emptyset$ implies $\lambda\geq\kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta$ (see Lemma 2b). With $F(\cdot)$ being defined as in (7), any $g\in\Gamma^*(\beta,\lambda;x)$ must satisfy the first order optimality condition that,

$$2\lambda g=\ell'(\beta^T x+g\sqrt{\delta}\beta^T A(x)^{-1}\beta).\tag{37}$$

As $\ell'(\beta^T x)\neq 0$, we have $g\neq 0$. Therefore it is sufficient to establish (22) 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 nondecreasing, 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 (37); equivalently, $g\geq\ell'(\beta^T x)/(2\lambda)$. This verifies (22) 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,

$$\begin{aligned}F(g,\beta,\lambda;x)&\geq\sup_{\gamma\geq 0}\left\{\ell(\beta^T x)+\ell'(\beta^T x)\sqrt{\delta}\beta^T A(x)^{-1}\beta\gamma-\lambda\sqrt{\delta}\left(\gamma^2\beta^T A(x)^{-1}\beta-1\right)\right\}\\&=\ell(\beta^T x)+\lambda\sqrt{\delta}+\sqrt{\delta}\beta^T A(x)^{-1}\beta\frac{[\ell'(\beta^T x)]^2}{4\lambda}.\end{aligned}\tag{38}$$

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,

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

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 (37) that,

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

Combining this observation with that in (38), we obtain $|g|\geq\ell'(\beta^T x)/(2\lambda)$.

When $\ell'(\beta^T x) < 0$, (22) follows by an argument symmetric to that of the $\ell'(\beta^T x) > 0$ cases described above. This completes the proof of Lemma 6. □

Lemma 15. *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 15. Suppose that $\ell_{rob}(\beta, \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta; x) < \infty$. Then for any $\varepsilon > 0$, there exist $\gamma \in \mathbb{R}$ such that $F(\gamma, \beta, \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta; x) > \ell_{rob}(\beta, \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta; x) - \varepsilon$. Thanks to the continuity of $F(\gamma, \beta, \lambda; x)$ with respect to λ,

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

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

$$\liminf_{\lambda \to (\kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta)^+} \ell_{rob}(\beta, \lambda; x) \geq \ell_{rob}(\beta, \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta; x).$$

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

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

thus yielding, $\limsup_{\lambda \to (\kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta)^+} \ell_{rob}(\beta, \lambda; x) \leq \ell_{rob}(\beta, \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta; x)$, and $\lim_{\lambda \to (\kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta)^+} \ell_{rob}(\beta, \lambda; x) = \ell_{rob}(\beta, \kappa\sqrt{\delta}\beta^T A(x)^{-1}\beta; x)$. 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 3). 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 7. Fix any $\beta \in B$ and define the set $A := \{x \in \mathbb{R}^d : \Gamma^*(\beta, \lambda_*(\beta); x) = \varnothing\}$. It follows from the characterization of $\mathbb{U}$ in Lemma 2 that $f(\beta, \lambda) = +\infty$ if $\lambda < \lambda_{thr}(\beta)$ and $f(\beta, \lambda) < +\infty$ if $\lambda > \lambda_{thr}(\beta)$. Therefore, it is necessary that $f(\beta, \lambda_*(\beta))$ is finite and $\lambda_*(\beta) \geq \lambda_{thr}(\beta)$.

Case 1. Suppose that $\lambda_*(\beta) > \lambda_{thr}(\beta)$. In this case we have from Lemma 2 that $\Gamma^*(\beta, \lambda_*(\beta); x)$ is not empty and $\partial_+ \ell_{rob}/\partial\lambda(\beta, \lambda_*(\beta); x)$ is given as in (13d), for P_0-almost every x. Since $f(\cdot)$ is finite in the neighborhood of $\lambda = \lambda^*(\beta)$, it follows from [2, Proposition 2.1] that

$$\frac{\partial_+ f}{\partial\lambda}(\beta, \lambda_*(\beta)) = \sqrt{\delta} - \sqrt{\delta} E_{P_0}\left[\beta^T A(X)^{-1}\beta \min_{\gamma \in \Gamma^*(\beta, \lambda_*(\beta); X)} \gamma^2\right]. \tag{39}$$

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

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

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

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

for $h > 0$. Due to the convexity of f, we also have that $h^{-1}(f(\beta, \lambda_*(\beta) + h) - f(\beta, \lambda_*(\beta)))$ is nondecreasing 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 $f(\beta, \cdot)$ and finiteness of $f(\beta, \lambda)$ for any $\lambda > \lambda_*(\beta)$, we have from [2, Proposition 2.1] and Lemma 4b that

$$0 \leq \frac{\partial_+ f}{\partial\lambda}(\beta, \lambda_*(\beta)) \leq \lim_{\lambda \downarrow \lambda_*(\beta)} \frac{\partial_- f}{\partial\lambda}(\beta, \lambda) \leq \sqrt{\delta}\left(1 - \lim_{\lambda \downarrow \lambda_*(\beta)} E_{P_0}\left[\beta^T A(X)^{-1}\beta g_\lambda(X)^2\right]\right), \quad (40)$$

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].

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 $f(\beta, \lambda_{thr}(\beta)) < \infty$ (see Lemma 15), we have that

$$\ell_{rob}(\beta, \lambda_*(\beta); x) = \lim_{n\to\infty} \ell_{rob}(\beta, \lambda_n; x) = \lim_{n\to\infty} F(g_{\lambda_n}(x), \beta, \lambda_n; x) = F(g_0, \beta, \lambda_*(\beta); x), \quad (41)$$

where the last equality holds because $F(\gamma, \beta, \lambda; x)$ is a continuous function in (γ, β, λ). However, it follows from (41) 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 $\overline{\lim}_{\lambda \downarrow \lambda_*(\beta)} g_\lambda^2(x) < \infty$. 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 (40), we obtain from (40) that $E_{P_0}[\beta^T A(X)^{-1}\beta \underline{\lim}_{\lambda \downarrow \lambda_*(\beta)} g_\lambda^2(X)] \leq 1$. Since $\underline{\lim}_{\lambda \downarrow \lambda_*(\beta)} g_\lambda^2(x) = +\infty$ 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 [18, Corollary 4]) similar to that in Lemma 4b 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 3), we have $h^{-1}(\ell_{rob}(\beta, \lambda_*(\beta) + h; x) - \ell_{rob}(\beta, \lambda_*(\beta); x))$ is nondecreasing 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 (39). This completes the proof of Lemma 7. □

Proof of Proposition 3. When Assumptions 1 - 5 are satisfied, we have that $\Gamma^*(\theta; X)$ is singleton for every $\theta = (\beta, \lambda)$ such that $\beta \in B$, $\lambda > \lambda'_{thr}(\beta)$, and P_0-almost surely every x. This is because of the strong concavity of the map $\gamma \mapsto F(\gamma, \theta; x)$ (see Lemma 9). Since any $g(\theta; x)$ such that $g(\theta; x) \in \Gamma^*(\theta; X)$ is uniquely defined for P_0-almost every x, we have from Lemma 4 that $\nabla_\theta \ell_{rob}(\theta; X)$ exists P_0-almost surely. Then it follows from Proposition 5 that $E_{P_0}[\nabla_\theta \ell_{rob}(\theta; X)]$ is well-defined and $E_{P_0}[\nabla_\theta \ell_{rob}(\theta; X)] \in \partial f(\theta)$ is a subgradient of the convex function $f(\theta)$, for every θ in $\{(\beta, \lambda) : \beta \in B, \lambda > \lambda'_{thr}(\beta)\}$ and P_0-almost every x. Since $\ell_{rob}(\theta; x)$ is convex and $f(\theta)$ is finite-valued whenever $\theta \in \{(\beta, \lambda) : \beta \in B, \lambda > \lambda'_{thr}(\beta)\}$, we have that $\partial f(\beta, \lambda) = E[\partial \ell_{rob}(\beta, \lambda; X)]$ (see [2, Proposition 2.2], [22]). Therefore $\nabla_\theta f(\beta, \lambda) = E_{P_0}[\nabla_\theta \ell_{rob}(\beta, \lambda; X)]$ is the only subgradient of f. In the case that $\delta < \delta_0$, the inclusion that $\mathbb{W}$ is contained in $\{(\beta, \lambda) : \beta \in B, \lambda > \lambda_{thr}(\beta)\}$ follows from Lemma 5a. □

Proof of Lemma 11. First of all, we provide an equivalent characterization of $\nabla^2 f(\beta, \lambda; x) - \Lambda(x)B(x) \succeq 0$. For simplicity, we rescale $\Lambda(x)$ and pick a new parameter m such that $\Lambda(x) =$

$m\sqrt{\delta}g^2$. The matrix $\nabla^2 f(\beta,\lambda;x) - m\sqrt{\delta}g^2 B(x)$ can be written as a block matrix, namely,

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

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

$$\frac{4\sqrt{\delta}g^2\beta^T A(x)^{-1}\beta}{\varphi} - m\sqrt{\delta}g^2 > 4\delta g^4 z^T \left((2\lambda - m)\sqrt{\delta}g^2 A(x)^{-1} + \frac{(2\lambda - m)\ell''(\beta^T\tilde{x})}{\varphi}\bar{x}\bar{x}^T\right)^{-1} z. \tag{42}$$

Recalling from the assumptions that $m \in (0, 2\lambda)$ and $\ell(\cdot)$ is convex, the positive definiteness of $(2\lambda - m)\sqrt{\delta}g^2 A(x)^{-1} + \frac{(2\lambda-m)\ell''(\beta^T\tilde{x})}{\varphi}\bar{x}\bar{x}^T$ is automatically satisfied. Then, applying Sharman-Morrison formula, one can show that

$$\left((2\lambda - m)\sqrt{\delta}g^2 A(x)^{-1} + \frac{(2\lambda - m)\ell''(\beta^T\tilde{x})}{\varphi}\bar{x}\bar{x}^T\right)^{-1} = \frac{1}{(2\lambda - m)\sqrt{\delta}g^2} C, \tag{43}$$

where C is a matrix defined as

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

Thus, combining equation (42) and (43), if $(\beta,\lambda) \in \mathbb{W}$ and $m \in (0, 2\lambda)$, then the matrix $\nabla^2 f(\beta,\lambda;x) - m\sqrt{\delta}g^2 B(x)$ if and only if

$$(2\lambda - m)\left(\frac{4\beta^T A(x)^{-1}\beta}{\varphi} - m\right) > 4z^T C z. \tag{44}$$

Let a, b and θ be constant defined as

$$a = 2\lambda, \quad b = \frac{4\beta^T A(x)^{-1}\beta}{\varphi}, \quad \theta := 4z^T C z. \tag{45}$$

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

$$\nabla^2 f(\beta,\lambda;x) - m\sqrt{\delta}g^2 B(x) \succeq 0 \tag{46}$$

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

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,

$$ab - \theta \geq \frac{4(\beta^T\tilde{x})^2}{\bar{x}^T A(x)\bar{x} + \sqrt{\delta}g^2\varphi/\ell''(\beta^T\tilde{x})}. \tag{47}$$

To show (47), 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,

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

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 $\bar{x} = \tilde{x} + \sqrt{\delta}gA(x)^{-1}\beta$, 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 (47). Finally, combining (46) and (47), we have

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

which is obtained by plugging in the definitions of a, b from (45). □

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 $f(\beta, \lambda) = \infty$, so the line search algorithm will not be executed in this case.
(2) When $\lambda \geq \lambda'_{thr}(\beta)$, 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 gradient descent method or Newton-Raphson method.

It then remains 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* $\beta \neq \mathbf{0}$, $P_0(\cdot) = \delta_{\{\mathbf{0}\}}(\cdot)$ *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 16. *Recall the definition of* $\mathbb{U}_\eta$ *from* (18) *and suppose that Assumption 1-4 hold and* $\eta > 0$. *Then there exist a random variable* R *with* $E_{P_0}[R^2] < \infty$, *such that*

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

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 $\lambda \geq \lambda_{thr}(\beta) + \eta$. Then, according to the Assumptions we have $\beta^T A(X)^{-1}\beta \leq \rho_{\min}^{-1} R_\beta^2$. Thus, letting $\varepsilon = \eta\delta^{-1/2}\rho_{\min} R_\beta^{-2}$, we have $\lambda \geq (\kappa + \varepsilon)\sqrt{\delta}\beta^T A(x)^{-1}\beta$, and thus the result of Lemma 14a can be applied. As a result, there exist a constant C_1 such that

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

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

With the help of Lemma 16, 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 $\gamma \to F(\gamma\beta^T A(x)^{-1}\beta, \beta, \lambda; x)$ 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 17. *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 3. *Even if a function is C^∞, it may have infinitely many isolated local optima on a compact set. Consider the next example,*

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

Now we discuss the line search scheme and its complexity. If the loss function $\ell(\cdot)$ is piecewise real analytic, the function $F(\cdot, \beta, \lambda; x)$ is also piecewise real analytic. In addition, using the result of Lemma 16, the optimization problem $\max_{\gamma \in \mathbb{R}} F(\gamma, \beta, \lambda; x)$ is equivalent to the problem $\max_{\bar{\gamma} \in [-R,R]} F(\bar{\gamma}\beta^T A(x)^{-1}\beta, \beta, \lambda; x)$, 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, [10, 11]). This approach, which different from randomly re-started Newton's method repeatedly, is guaranteed to locate all of the stationary points. The use of Cauchy's theorem requires the evaluation of certain integrals in smooth trajectories. The evaluation of these trajectories 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 $\bar{\gamma} \to F(\bar{\gamma}\beta^T A(x)^{-1}\beta, \beta, \lambda; x)$ 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.

Acknowledgments.

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

References

[1] Z. 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] D. P. Bertsekas and S. E. Shreve. *Stochastic optimal control: the discrete time case*. Elsevier, 1978.

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

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

[6] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *arXiv preprint arXiv:1604.01446*, 2016.

[7] Z. Chen, D. Kuhn, and W. Wiesemann. Data-driven chance constrained programs over wasserstein balls. -, 2018. Available from Optimization Online.

[8] Z. Chen, M. Sim, and P. Xiong. Adaptive robust optimization with scenario-wise ambiguity sets. -, 2018. Available from Optimization Online.

[9] A. Defazio, F. Bach, and S. 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.

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

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

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

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

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

[15] R. Gao, L. Xie, Y. Xie, and H. Xu. Robust hypothesis testing using wasserstein uncertainty sets. *arXiv preprint arXiv:1805.10611*, 2018.

[16] R. Johnson and T. 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.

[17] F. Luo and S. Mehrotra. Decomposition algorithm for distributionally robust optimization using wasserstein metric. *arXiv preprint arXiv:1704.03920*, 2017.
[18] P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.
[19] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, Sep 2018.
[20] A. Mokkadem and M. Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Ann. Appl. Probab.*, 16(3):1671–1702, 08 2006.
[21] E. Moulines and F. 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.
[22] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.
[23] Y.-K. Noh, B.-T. Zhang, and D. D. Lee. Generative local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 1822–1830, 2010.
[24] B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
[25] B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
[26] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.
[27] S. Shafieezadeh-Abadeh, P. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584. 2015.
[28] S. Shafieezadeh Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Regularization via mass transportation. -, 2017. Available from Optimization Online.
[29] O. Shamir and T. 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.
[30] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.
[31] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.
[32] R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018.
[33] J. Wang, A. Kalousis, and A. Woznica. Parametric local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 1601–1609, 2012.
[34] W. Xie. On distributionally robust chance constrained program with wasserstein distance. *arXiv preprint arXiv:1806.07418*, 2018.
[35] 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.
[36] C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with wasserstein metric. *Operations Research Letters*, 46(2):262 – 267, 2018.