ROBUST WASSERSTEIN PROFILE INFERENCE AND APPLICATIONS TO MACHINE LEARNING

Jose Blanchet Yang Kang Karthyek Murthy

Columbia University

Abstract. We introduce RWPI (Robust Wasserstein Profile Inference), a novel class of statistical tools which exploits connections between Empirical Likelihood, Distributionally Robust Optimization and the Theory of Optimal Transport (via the use of Wasserstein distances). A key element of RWPI is the so-called Robust Wasserstein Profile function, whose asymptotic properties we study in this paper. We illustrate the use of RWPI in the context of machine learning algorithms, such as the generalized LASSO (Least Absolute Shrinkage and Selection) and regularized logistic regression, among others. For these algorithms, we show how to optimally select the regularization parameter without the use of cross validation. The use of RWPI for such optimal selection requires a suitable distributionally robust representation for these machine learning algorithms, which is also novel and of independent interest. Numerical experiments are also given to validate our theoretical findings.

1. Introduction

The goal of this paper is to introduce and investigate a novel inference methodology which we call RWPI (Robust Wasserstein-distance Profile-based Inference – pronounced similar to Rupee[1]). RWPI combines ideas from three different areas: Empirical Likelihood (EL), Distributionally Robust Optimization, and the Theory of Optimal Transport. While RWPI can be applied to a wide range of inference problems, in this paper we use several well known algorithms in machine learning to illustrate the use and implications of this methodology.

We will explain, by means of several examples of interest, how RWPI can be used to optimally choose the regularization parameter in machine learning applications without the need of cross validation. The examples of interest that we study in this paper include generalized LASSO (Least Absolute Shrinkage and Selection) and Regularized Logistic Regression, among others.

In order to explain RWPI let us walk through a simple application in a familiar context, namely, that of linear regression.

1.1. RWPI for optimal regularization of generalized LASSO.

Consider a given a set of training data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. The input $X_i \in \mathbb{R}^d$ is a vector of d predictor variables, and $Y_i \in \mathbb{R}$ is the response variable. (Throughout the paper any vector is understood to be a column vector and the transpose of x is denoted by x^T.) It is postulated that

$$Y_i = \beta_*^T X_i + e_i,$$

for some $\beta_* \in \mathbb{R}^d$ and errors $\{e_1, ..., e_n\}$. Under suitable statistical assumptions (such as independence of the samples in the training data) one may be interested in estimating β_*. Underlying

[1]The acronym RWPI is chosen to sound just as RUPI ("u" as in put and "i" as in bit). In turn, RUPI means beautiful in Sanskrit.

there is a general loss function, $l(x,y;\beta)$, which we shall take for simplicity in this discussion to be the quadratic loss, namely, $l(x,y;\beta)=\left(y-\beta^T x\right)^2$.

Now, let P_n be the empirical distribution function, namely,

$$P_n(dx,dy):=\frac{1}{n}\sum_{i=1}^{n}\delta_{\{(X_i,Y_i)\}}(dx,dy).$$

Throughout out the paper we use the notation $E_P[\cdot]$ to denote expectation with respect to a probability distribution P.

In the last two decades, regularized estimators have been introduced and studied. Many of them have gained substantial popularity because of their good empirical performance and insightful theoretical properties, (see, for example, [41] for an early reference and [16] for a discussion on regularized estimators). One such regularized estimator, implemented, for example in the "flare" package, see [19], is the so-called generalized LASSO estimator; which is obtained by solving the following convex optimization problem in β

$$\min_{\beta\in\mathbb{R}^d}\left\{\sqrt{E_{P_n}\left[l(X,Y;\beta)\right]}+\lambda\left\|\beta\right\|_1\right\}=\min_{\beta\in\mathbb{R}^d}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}l(X_i,Y_i;\beta)}+\lambda\left\|\beta\right\|_1\right\}, \tag{1}$$

where $\left\|\beta\right\|_p$ denotes the p-th norm in the Euclidean space. The parameter λ, commonly referred to as the regularization parameter, is crucial for the performance of the algorithm and it is often chosen using cross validation.

1.1.1. *Distributionally robust representation of generalized LASSO* . We shall illustrate how to choose λ, satisfying a natural optimality criterion, as the quantile of a certain object which we call the Robust Wasserstein Profile (RWP) function evaluated at β_*. This will motivate a systematic study of the RWP function as the sample size, n, increases. However, before we define the associated RWP function, we first introduce a class of representations which are of independent interest and which are necessary to motivate the definition of the RWP function for choosing λ.

One of our contributions in this paper (see Section 3) is a representation of (1) in terms of a Distributionally Robust Optimization formulation. In particular, we construct a discrepancy measure, $\mathcal{D}_c(P,Q)$, based on a suitable Wasserstein-type distance, between two probability measures P and Q satisfying that

$$\min_{\beta\in\mathbb{R}^d}\left\{\sqrt{E_{P_n}\left[l(X,Y;\beta)\right]}+\lambda\left\|\beta\right\|_1\right\}^2=\min_{\beta\in\mathbb{R}^d}\max_{P:\ \mathcal{D}_c(P,P_n)\leq\delta}E_P\left[l(X,Y;\beta)\right], \tag{2}$$

where $\delta=\lambda^{1/2}$. Observe that the regularization parameter is fully determined by the size of the uncertainty, δ, in the distributionally robust formulation on the right hand side of (2).

The set $\mathcal{U}_\delta(P_n)=\{P:\mathcal{D}_c(P,P_n)\leq\delta\}$ is called the uncertainty set in the language of distributionally robust optimization, and it represents the class of models that are, in some sense, plausible variations of P_n.

For every selection P in $\mathcal{U}_\delta(P_n)$, there is an optimal choice $\beta=\beta(P)$ which minimizes the risk $E_P[l(X,Y;\beta)]$. We shall define $\Lambda_n(\delta)=\{\beta(P):P\in\mathcal{U}_\delta(P_n)\}$ to be the set of plausible selections of the parameter β.

Now, for the definition of $\Lambda_n(\delta)$ to be sensible, we must have that the estimator obtained from the left hand side of (2) is plausible. This follows from the following result, which is

established with the aid of a min-max theorem in Section 4,

$$\min_{\beta\in\mathbb{R}^d}\ \max_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\left[l(X,Y;\beta)\right] = \min_{\beta\in\Lambda_n(\delta)}\ \max_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\left[l(X,Y;\beta)\right].$$

Then, we will say that β_* is **plausible** with $(1-\alpha)$ confidence, or simply, $(1-\alpha)$-plausible if δ is large enough so that $\beta_* \in \Lambda_n(\delta)$ with probability at least $1-\alpha$. This definition leads us to the optimality criterion that we shall consider.

Our optimal selection criterion for δ is formulated as follows: *Choose $\delta > 0$ as small as possible in order to guarantee that β_* is plausible with $(1-\alpha)$ confidence.*

As an additional desirable property, we shall verify that if β_* is $(1-\alpha)$-plausible, then $\Lambda_n(\delta)$ is a $(1-\alpha)$-confidence region for β_*. A computationally efficient procedure for evaluating $\Lambda_n(\delta)$ will be studied in future work. Our focus in this paper is on the optimal selection of δ.

1.1.2. *The associated Wasserstein Profile Function .* In order to formally setup an optimization problem for the choice of $\delta > 0$, note that for any given P, by convexity, any optimal selection β is characterized by the first order optimality condition, namely,

$$E_P\left[\left(Y-\beta^T X\right)X\right] = \mathbf{0}. \tag{3}$$

We then introduce the following object, which is the RWP function associated with the estimating equation (3),

$$R_n(\beta) = \inf\left\{\mathcal{D}_c(P,P_n) : E_P\left[\left(Y-\beta^T X\right)X\right] = \mathbf{0}\right\}. \tag{4}$$

Finally, we claim that the optimal choice of δ is precisely the $1-\alpha$ quantile, $\chi_{1-\alpha}$, of $R_n(\beta_*)$; that is

$$\chi_{1-\alpha} = \inf\left\{z : P\left(R_n(\beta_*)\leq z\right)\geq 1-\alpha\right\}.$$

To see this note that if $\tilde{\delta} > \chi_{1-\alpha}$ then indeed β^* is plausible with probability at least $1-\alpha$, but $\tilde{\delta}$ is not minimal. In turn, note that $R_n(\beta)$ allows to provide an explicit characterization of $\Lambda_n(\chi_{1-\alpha})$,

$$\Lambda_n(\chi_{1-\alpha}) = \{\beta : R_n(\beta)\leq\chi_{1-\alpha}\}.$$

Moreover, we clearly have

$$P\left(\beta_*\in\Lambda_n(\chi_{1-\alpha})\right) = P\left(R_n(\beta_*)\leq\chi_{1-\alpha}\right) = 1-\alpha,$$

so $\Lambda_n(\chi_{1-\alpha})$ is a $(1-\alpha)$-confidence region for β^*.

In order to further explain the role of $R_n(\beta_*)$, let us define $\mathcal{P}_{opt}$ to be the set of probability measures, P, supported on a subset of $\mathbb{R}^d\times\mathbb{R}$ for which (3) holds with $\beta=\beta_*$. Formally,

$$\mathcal{P}_{opt} := \left\{P : E_P\left[\left(Y-\beta_*^T X\right)X\right] = \mathbf{0}\right\}.$$

In simple words, $\mathcal{P}_{opt}$ is the set of probability measures for which β_* is an optimal risk minimization parameter. Observe that using this definition we can write

$$R_n(\beta_*) = \inf\{\mathcal{D}_c(P,P_n) : P\in\mathcal{P}_{opt}\}.$$

Consequently, the set

$$\{P : \mathcal{D}_c(P,P_n)\leq R_n(\beta_*)\}$$

denotes the smallest uncertainty region around P_n (in terms of $\mathcal{D}_c$) for which one can find a distribution P satisfying the optimality condition $E_P\left[(Y-\beta_*^T X)X\right] = \mathbf{0}$, see Figure 1 for a pictorial representation of $\mathcal{P}_{opt}$.

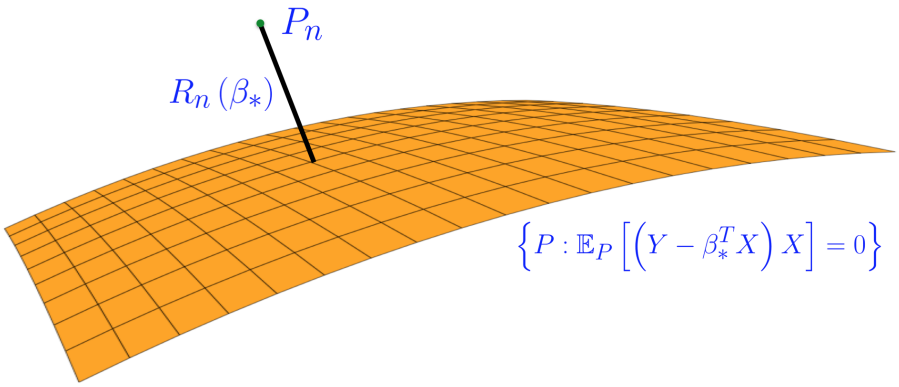

FIGURE 1. Illustration of RWP function evaluated at β_*

In summary, $R_n(\beta_*)$ denotes the smallest size of uncertainty that makes β_* *plausible*. If we were to choose a radius of uncertainty smaller than $R_n(\beta_*)$, then no probability measure in the neighborhood will satisfy the optimality condition $E_P\left[(Y - \beta_*^T X)X\right] = \mathbf{0}$. On the other hand, if $\delta > R_n(\beta_*)$, the set

$$\left\{P : E_P\left[(Y - \beta_*^T X)\right] = \mathbf{0}, \mathcal{D}_c\left(P, P_n\right) \leq \delta\right\}$$

is nonempty. Given the importance of $R_n(\beta_*)$ in the optimal selection of the regularization parameter λ, if is of interest to analyze its asymptotic properties as $n \to \infty$.

It is important to note, however, that the estimating equations given in (3) are just one of potentially many ways in which β_* can be characterized. In the case of Gaussian input there is a (well known) intimate connection between (3) an maximum likelihood estimation. In general it appears sensible, at least from the standpoint of philosophical consistency to connect the choice of estimating equation with the loss function $l(x, y; \beta)$ used in the Distributionally Robust Representation (2).

1.2. **A broad perspective of our contributions.** The previous discussion in the context of linear regression highlights two key ideas: a) the RWP function as a key object of analysis, and b) the role of distributionally robust representation of regularized estimators.

The RWP function can be applied much more broadly than in the context of regularized estimators. This paper is written with the goal of studying the RWP function for estimating equations generally and systematically. We showcase the study of the RWP function in a context of great importance, namely, the optimal selection of regularization parameters in several machine learning algorithms.

Broadly speaking, RWPI is a statistical tool which consists in building a suitable RWP function in order to estimate a parameter of interest. From a philosophical standpoint, RWPI borrows heavily from Empirical Likelihood (EL), introduced in the seminal work of [23, 24]. There are important methodological differences, however, as we shall discuss in the sequel. In the last three decades, there have been a great deal of successful applications of Empirical Likelihood for inference [25, 46, 7, 17, 29]. In principle all of those applications can be revisited using the RWP function and its ramifications. Therefore, we spend the first part of the paper, namely Section 2, discussing general properties of the RWP function.

The application of RWPI for the optimal selection of regularization parameters in various machine learning settings is given in Section 4. Once a suitable RWP function is obtained,

the results in Section 4 are obtained directly from applications of our results in Section 2. In order to obtain the correct RWP function formulation for each of the machine learning settings of interest, however, we will need to derive a suitable distributionally robust representations which, analogous to those discussed in the generalized LASSO setting. These representations are given in Section 3 of this paper.

We now provide a more precise description of our contributions:

A) We provide general limit theorems for the asymptotic distribution (as the sample size increases) of the RWP function defined for general estimating equations, not only those arising from linear regression problems. Hence, providing tools to apply RWPI in substantial generality (see the results in Section 2.4).

B) We will explain how, by judiciously choosing $\mathcal{D}_c(\cdot)$, we can define a family of regularized regression estimators (See Section 3). In particular, we will show how generalized LASSO (see and Theorem 2), and regularized logistic regression (see Theorem 3) arise as a particular case of a RWPI formulation.

C) The results in **B)** allow to obtain the appropriate RWP function to select an optimal regularization parameter. We then will illustrate how to analyze the distribution of $R_n(\beta_*)$ using our results form **A)** (see Section 4).

D) We analyze our regularization selection in the high-dimensional setting for generalized LASSO. Under standard regularity conditions, we show (see Theorem 6) that the regularization parameter λ might be chosen as,

$$\lambda = \frac{\pi}{\pi - 2} \frac{\Phi^{-1}(1 - \alpha/2d)}{\sqrt{n}},$$

where $\Phi(\cdot)$ is the cumulative distribution of the standard normal random variable and $1 - \alpha$ is a user-specified confidence level. The behavior of λ as a function of n and d is consistent with regularization selections studied in the literature motivated by different considerations.

E) We analyze the empirical performance of RWPI for the selection of the optimal regularization parameter in the context of generalized LASSO. This is done in Section 5 of the paper. We apply our analysis both to simulated and real data and compare against the performance of cross validation. We conclude that our approach is comparable (not worst) than cross validation.

We now provide a discussion on topics which are related to RWPI.

1.3. **Connections to related inference literature.** Le us first discuss the connections between RWPI and EL. In EL one builds a Profile Likelihood for an estimating equation. For instance, in the context of EL applied to estimating β satisfying (3), one would build a Profile Likelihood Function in which the optimization object is only defined as the likelihood (or the log-likelihood) between a given distribution P with respect to P_n. Therefore, the analogue of the uncertainty set $\{P : \mathcal{D}_c(P, P_n) \leq \delta\}$, in the context of EL, will typically contain distributions whose support coincides with that of P_n. In contrast, the definition of the RWP function does not require the likelihood between an alternative plausible model P, and the empirical

distribution, P_n, to exist. Owing to this flexibility, for example, we are able to establish the connection between regularization estimators and a suitable profile function.

There are other potential benefits of using a profile function which does not restrict the support of alternative plausible models. For example, it has been observed in the literature that in some settings EL might exhibit low coverage [26, 9, 43]. It is not the goal of this paper to examine the coverage properties of RWPI systematically, but it is conceivable that relaxing the support of alternative plausible models, as RWPI does, can translate into desirable coverage properties.

From a technical standpoint, the definition of the Profile Function in EL gives rise to a finite dimensional optimization problem. Moreover, there is a substantial amount of smoothness in the optimization problems defining the EL Profile Function. This degree of smoothness can be leveraged in order to obtain the asymptotic distribution of the Profile Function as the sample size increases. In contrast, the optimization problem underlying the definition of RWP function in RWPI is an infinite dimensional linear program. Therefore, the mathematical techniques required to analyze the associated RWP function are different (more involved) than the ones which are commonly used in the EL setting.

A significant advantage of EL, however, is that the limiting distribution of the associated Profile Function is typically chi-squared. Moreover, such distribution is self-normalized in the sense that no parameters need to be estimated from the data. Unfortunately, this is typically not the case in the case of RWPI. In many settings, however, the parameters of the distribution can be easily estimated from the data itself.

Another set of tools, strongly related to RWPI, have also been studied recently by the name of SOS (Sample-Out-of-Sample) inference [5]. In this setting, also a RWP function is built, but the support of alternative plausible models is assumed to be finite (but not necessarily equal to that of P_n). Instead, the support of alternative plausible models is assumed to be generated not only by the available data, but additional samples coming from independent distributions (defined by the user). The mathematical results obtained for the RWP function in the context of SOS are different from those obtained in this paper. For example, in the SOS setting, the rates of convergence are dimension-dependent, which is not the case in the RWPI case.

1.4. **Some connections to Distributionally Robust Optimization and Optimal Transport.** Connection between robust optimization and regularization procedures such as LASSO and Support Vector Machines have been studied in the literature, see [44, 45]. The methods proposed here differ subtly: While the papers [44, 45] add deterministic perturbations of a certain size to the predictor vectors X to quantify uncertainty, the Distributionally Robust Representations that we derive measure perturbations in terms of deviations from the empirical distribution. While this change may appear cosmetic, it brings a significant advantage: measuring deviations from empirical distribution, in turn, lets us derive suitable limit laws (or) probabilistic inequalities that can be used to choose the size of uncertainty, δ, in the uncertainty region $\mathcal{U}_\delta(P_n) = \{P : \mathcal{D}_c(P, P_n) \leq \delta\}$.

Now, it is intuitively clear that as the number of samples n increase, the deviation of the empirical distribution from the true distribution decays to zero, as a function of n, at a specific rate of convergence. To begin with, one can simply use, as a direct approach to choosing the size of δ, a concentration inequality that measures this rate of convergence. Such simple specification of the size of uncertainty, suitably as a function of n, does not arise naturally in the deterministic robust optimization approach. For a concentration inequality that measures

such deviations in terms of the Wasserstein distance, we refer to [13] and references there in. For an application of these concentration inequalities to choose the size of uncertainty set in the context of distributionally robust logistic regression, refer [35]. It is important to note that, despite imposing severe tail assumptions, these concentration inequalities dictate the size of uncertainty to decay at the rate $O(n^{-1/d})$; unfortunately, this prescription scales non-graciously as the number of dimensions increase. Since most of the modern learning problems have huge number of covariates, application of such concentration inequalities with poor rate of decay with dimensions may not be most suitable for applications.

In contrast to directly using concentration inequalities, the prescription that we advocate typically has a rate of convergence of order $O\left(n^{-1/2}\right)$ as $n \to \infty$ (for fixed d). Moreover, as we discuss in the case of LASSO, according to our results corresponding to contribution **E),** our prescription of the size of uncertainty actually can be shown (under suitable regularity conditions) to decay at rate $O(\sqrt{\log d/n})$ (uniformly over d and n), which is in agreement with the findings of compressed sensing and high-dimensional statistics literature (see [8, 20, 3] and references therein). Interestingly, the regularization parameter prescribed by RWPI methodology is automatically obtained without looking into the data (unlike cross-validation).

Although we have focused our discussion on the context of regularized estimators, our results are directly applicable to the area of data-driven Distributionally Robust Optimization whenever the uncertainty sets are defined in terms of a Wasserstein distance or, more generally, an optimal transport metric. In particular, consider a given distributionally robust formulation of the form

$$\min_{\theta: G(\theta) \leq 0} \quad \max_{P:\ \mathcal{D}_c(P,P_n) \leq \delta} \quad E_P\left[H(W,\theta)\right],$$

for a random element W and a convex function $H(W,\cdot)$ defined over a convex region $\{\theta : G(\theta) \leq 0\}$ (assuming $G : \mathbb{R}^d \to \mathbb{R}$ convex). Here P_n is the empirical measure of the sample $\{W_1, ..., W_n\}$. One can then follows a reasoning parallel to what we advocate throughout our LASSO discussion.

Argue, by applying the corresponding KKT (Karush-Kuhn-Tucker) conditions, if possible, that an optimal solution θ_* to the problem

$$\min_{\theta: G(\theta) \leq 0} E_{P_{true}}[H(W,\theta)]$$

satisfies a system of estimating equations of the form

$$E_{P_{true}}[h(W,\theta_*)] = 0, \tag{5}$$

for a suitable $h(\cdot)$ (where P_{true} is the weak limit of the empirical measure P_n as $n \to \infty$).

Then, given a confidence level $1-\alpha$, one should choose δ as the $(1-\alpha)$ quantile of the RWP function function

$$R_n(\theta_*) = \inf\{\mathcal{D}_c(P, P_n) : E_P[h(W,\theta_*)] = 0\}.$$

The results in Section 2 can then be used directly to approximate the $(1-\alpha)$ quantile of $R_n(\theta_*)$. Just as we explain in our discussion of the generalized LASSO example, the selection of δ is the smallest possible choice for which θ_* is plausible with $(1-\alpha)$ confidence.

1.5. **Organization of the paper.** The rest of the paper is organized as follows. Section 2 deals with contribution **A)**. We first introduce Wasserstein distances. Then, we discuss the Robust Wasserstein Profile function as an inference tool in a way which is parallel to the Profile Likelihood in EL. We derive the asymptotic distribution of the RWP function for general estimating equations. Section 3 corresponds to contribution **B**, namely, distributionally robust

representations of popular machine learning algorithms. Section 4 discusses contributions **C)**, namely the combination of the results from contributions **A)** and optimal regularization parameter selection. Our high-dimensional analysis of the RWP function in the case of generalized LASSO is also given in Section 4. The proofs for the main results are given in Section 5. Finally, our numerical experiments using both simulated and real data set are given in Section 6.

2. The Robust Wasserstein Profile Function

Given an estimating equation $E_{P_n}[h(W,\theta)] = \mathbf{0}$, the objective of this section is to study the asymptotic behavior of the associated RWP function $R_n(\theta)$. To do this, we first introduce some notation to define optimal transport costs and Wasserstein distances. Following this, we provide evidence, initially with a simple example, followed by results for general estimating equations, that the profile function defined using Wasserstein distances is tractable.

2.1. Optimal Transport Costs and Wasserstein Distances.

Let $c : \mathbb{R}^m \times \mathbb{R}^m \to [0,\infty]$ be any lower semi-continuous function such that $c(u,w) = 0$ if and only if $u = w$. Given two probability distributions $P(\cdot)$ and $Q(\cdot)$ supported on $\mathbb{R}^m$, one can define the optimal transport cost or discrepancy between P and Q, denoted by $\mathcal{D}_c(P,Q)$, as

$$\mathcal{D}_c(P,Q) = \inf\left\{E_\pi\left[c(U,W)\right] : \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m),\ \pi_U = P,\ \pi_W = Q\right\}. \tag{6}$$

Here, $\mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m)$ is the set of joint probability distributions π of (U,W) supported on $\mathbb{R}^m \times \mathbb{R}_m$, and π_U and π_W denote the marginals of U and W under π, respectively.

Intuitively, in our formulation, the quantity $c(u,w)$ specifies the cost of transporting unit mass from u in $\mathbb{R}^m$ to another element w in $\mathbb{R}^m$. As a result, the expectation $E_\pi[c(U,W)]$ denotes the expected transport cost associated with the joint distribution π. In the sequel, $\mathcal{D}_c(P,Q)$, defined in (6), represents the optimal transport cost associated with probability measures P and Q.

In addition to the stated assumptions on the cost function $c(\cdot)$, if $c(\cdot)$ is symmetric (that is, $c(u,w) = c(w,u)$), and there exists $\rho \geq 1$ such that $c^{1/\rho}(u,w) \leq c^{1/\rho}(u,z) + c^{1/\rho}(z,w)$ for all $u,w,z \in \mathbb{R}^m$ (that is, $c^{1/\rho}(\cdot)$ satisfies the triangle inequality), it can be easily verified that $\mathcal{D}_c^{1/\rho}(P,Q)$ is a metric (see [42] for a proof, and other properties of the metric D_c).

For example, if $c(u,w) = \|u-w\|_2^2$, where $\|\cdot\|_2$ is the Euclidean distance in $\mathbb{R}^m$, then $\rho = 2$ yields that $c(u,w)^{1/2} = \|u-w\|_2$ is symmetric, non-negative, lower semi-continuous and it satisfies the triangle inequality. In that case,

$$\mathcal{D}_c^{1/2}(P,Q) = \inf\left\{\sqrt{E_\pi\left[\|U-W\|_2^2\right]} :\ \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m),\ \pi_U = P,\ \pi_W = Q\right\}$$

coincides with the Wasserstein distance of order two.

More generally, if we choose $c^{1/\rho}(u,w) = d(u,w)$ for any metric d defined on $\mathbb{R}^m$, then $\mathcal{D}_c^{1/\rho}(\cdot)$ is the standard Wasserstein distance of order $\rho \geq 1$.

Wasserstein distances metrize weak convergence of probability measures under suitable moment assumptions, and have received immense attention in probability theory (see [30, 31, 42] for a collection of classical applications). In addition, earth-mover's distance, a particular example of Wasserstein distances, has been of interest in image processing (see [33, 38]). More recently, optimal transport metrics and Wasserstein distances are being actively investigated for

its use in various machine learning applications as well (see [34, 28, 32, 37, 14, 39] and references therein for a growing list of new applications).

Throughout this paper, we shall select $\mathcal{D}_c$ for a judiciously chosen cost function $c(\cdot)$ in formulations such as (2). It is useful to allow $c(\cdot)$ to be lower semi-continuous and potentially be infinite in some region to accommodate some of the applications, such as regularization in the context of logistic regression, as we shall see in Section 3. So, our setting requires discrepancy choices which are slightly more general than standard Wasserstein distances.

2.2. **The RWP Function for Estimating Equations and Its Use as an Inference Tool.** The Robust Wasserstein Profile function's definition is inspired by the notion of the Profile Likelihood function, introduced in the pioneering work of Art Owen in the context of EL (see [26]). We provide the definition of the RWP function for estimating $\theta_* \in \mathbb{R}^l$, which we assume satisfies

$$E\left[h\left(W,\theta_*\right)\right] = \mathbf{0}, \tag{7}$$

for a given random variable W taking values in $\mathbb{R}^m$ and an integrable function $h : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}^r$. The parameter θ_* will typically be unique to ensure consistency, but uniqueness is not necessary for the limit theorems that we shall state, unless we explicitly indicate so.

Given a set of samples $\{W_1, ..., W_n\}$, which are assumed to be i.i.d. copies of W, we define the Wasserstein Profile function for the estimating equation (7) as,

$$R_n\left(\theta\right) := \inf\left\{D_c(P, P_n) : \ E_P\left[h(W,\theta)\right] = \mathbf{0}\right\}. \tag{8}$$

Here, recall that P_n denotes the empirical distribution associated with the training samples $\{W_1, \ldots, W_n\}$ and $c(\cdot)$ is a chosen cost function. In this section, we are primarily concerned with cost functions of the form,

$$c\left(u, w\right) = \left\|w - u\right\|_q^\rho, \tag{9}$$

where $\rho \geq 1$ and $q \geq 1$. We remark, however, that the methods presented here can be easily adapted to more general cost functions. For simplicity, we assume that the samples $\{W_1, \ldots, W_n\}$ are distinct.

Since, as we shall see, that the asymptotic behavior of the RWP function $R_n(\theta)$ is dependent on the exponent ρ in (9), we shall sometimes write $R_n\left(\theta; \rho\right)$ to make this dependence explicit; but whenever the context is clear, we drop ρ to avoid notational burden. Also, observe that the profile function defined in (4) for the linear regression example is obtained as a particular case by selecting $W = (X, Y)$, $\beta = \theta$ and defining $h\left(x, y, \theta\right) = (y - \theta^T x)x$.

Our goal in this section is to develop an asymptotic analysis of the RWP function which parallels that of the theory of EL. In particular, we shall establish,

$$n^{\rho/2} R_n\left(\theta_*; \rho\right) \Rightarrow \bar{R}\left(\rho\right). \tag{10}$$

for a suitably defined random variable $\bar{R}\left(\rho\right)$ (throughout the rest of the paper, the symbol "$\Rightarrow$" denotes convergence in distribution).

As the empirical distribution weakly converges to the underlying probability distribution from which the samples are obtained from, it follows from the definition of RWP function in (10) that $R_n(\theta; \rho) \to 0$, as $n \to \infty$, if and only if θ satisfies $E[h(W, \theta)] = \mathbf{0}$; for every other θ, we have that $n^{\rho/2} R_n(\theta; \rho) \to \infty$. Therefore, the result in (10) can be used to provide confidence

regions (at least conceptually) around θ_*. In particular, given a confidence level $1-\alpha$ in (0,1), if we denote η_α as the $(1-\alpha)$ quantile of $\bar{R}(\rho)$, that is, $P\left(\bar{R}(\rho) \leq \eta_\alpha\right) = (1-\alpha)$, then

$$\bar{\Lambda}_n\left(\frac{\eta_\alpha}{n}\right) = \left\{\theta : R_n(\theta;\rho) \leq \frac{\eta_\alpha}{n}\right\}$$

yields an approximate $(1-\alpha)$ confidence region for θ_*. This is because, by definition of $\bar{\Lambda}_n(\eta_\alpha/n)$, we have

$$P\left(\theta_* \in \bar{\Lambda}_n(\eta_\alpha/n)\right) = P\left(n^{\rho/2} R_n(\theta_*;\rho) \leq \eta_\alpha\right) \approx P\left(\bar{R}(\rho) \leq \eta_\alpha\right) = 1-\alpha.$$

Throughout the development in this section, the dimension m of the underlying random vector W is kept fixed and the sample size n is sent to infinity; the function $h(\cdot)$ can be quite general. In Section 4.3, we extend the analysis of RWP function to the case where the ambient dimension could scale with the number of training samples n, in the specific context of generalized LASSO for linear regression.

2.3. **The dual formulation of RWP function.** The first step in the analysis of the RWP function $R_n(\theta)$ is to use the definition of the discrepancy measure D_c to rewrite $R_n(\theta)$ as,

$$R_n(\theta) = \inf\left\{E_\pi[c(U,W)] : \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m),\ E_\pi[h(U,\theta)] = \mathbf{0},\ \pi_{_W} = P_n\right\},$$

which is a *problem of moments* of the form,

$$R_n(\theta) = \inf_{\pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m)} \left\{E_\pi[c(U,W)] :\ E_\pi[h(U,\theta)] = \mathbf{0},\ E_\pi[\mathbb{I}(W = W_i)] = \frac{1}{n}, i = \overline{1,n}\right\}. \tag{11}$$

The problem of moments is a classical linear programming problem for which the respective dual formulation and strong duality have been well-studied (see, for example, [18, 36]). The linear program problem over the variable π in (11) admits a simple dual semi-infinite linear program of form,

$$\begin{aligned}
&\sup_{a_i \in \mathbb{R}, \lambda \in \mathbb{R}^r} \left\{a_0 + \frac{1}{n}\sum_{i=1}^n a_i :\ a_0 + \sum_{i=1}^n a_i \mathbf{1}_{\{w=W_i\}}(u,w) + \lambda^T h(u,\theta) \leq c(u,w) \text{ for } \forall u,w \in \mathbb{R}^m\right\} \\
&\quad = \sup_{\lambda \in \mathbb{R}^r}\left\{\frac{1}{n}\sum_{i=1}^n \inf_{u \in \mathbb{R}^m}\left\{c(u,W_i) - \lambda^T h(u,\theta)\right\}\right\} \\
&\quad = \sup_{\lambda \in \mathbb{R}^r}\left\{-\frac{1}{n}\sum_{i=1}^n \sup_{u \in \mathbb{R}^m}\left\{\lambda^T h(u,\theta) - c(u,W_i)\right\}\right\}.
\end{aligned}$$

Proposition 1 below states that strong duality holds under mild assumptions, and the dual formulation above indeed equals $R_n(\theta)$.

Proposition 1. *Let $h(\cdot,\theta)$ be Borel measurable, and $\Omega = \{(u,w) \in \mathbb{R}^m \times \mathbb{R}^m : c(u,w) < \infty\}$ be Borel measurable and non-empty. Further, suppose that $\mathbf{0}$ lies in the interior of the convex hull of $\{h(u,\theta) : u \in \mathbb{R}^m\}$. Then,*

$$R_n(\theta) = \sup_{\lambda \in \mathbb{R}^r}\left\{-\frac{1}{n}\sum_{i=1}^n \sup_{u \in \mathbb{R}^m}\left\{\lambda^T h(u,\theta) - c(u,W_i)\right\}\right\}.$$

A proof of Proposition 1, along with an introduction to the problem of moments, is provided in Appendix A.

2.4. **Asymptotic Distribution of the RWP Function.** In order to gain intuition behind (10), let us first consider the simple example of estimating the expectation $\theta_* = E[W]$ of a real-valued random variable W, using $h(w,\theta) = w - \theta$.

Example 1. Let $h(w,\theta) = w - \theta$ with $m = 1 = l = r$. First, suppose that the choice of cost function is $c(u,w) = |u-w|^\rho$ for some $\rho > 1$. As long as θ lies in the interior of convex hull of support of W, Proposition (1) implies,

$$\begin{aligned} R_n(\theta;\rho) &= \sup_{\lambda\in\mathbb{R}} \left\{ -\frac{1}{n}\sum_{i=1}^n \sup_{u\in\mathbb{R}} \left\{ \lambda(u-\theta) - |W_i - u|^\rho \right\} \right\} \\ &= \sup_{\lambda\in\mathbb{R}} \left\{ -\frac{\lambda}{n}\sum_{i=1}^n (W_i - \theta) - \frac{1}{n}\sum_{i=1}^n \sup_{u\in\mathbb{R}} \left\{ \lambda(u - W_i) - |W_i - u|^\rho \right\} \right\}. \end{aligned}$$

As $\max_\Delta\{\lambda\Delta - |\Delta|^\rho\} = (\rho-1)|\lambda/\rho|^{\rho/(\rho-1)}$, we obtain

$$\begin{aligned} R_n(\theta;\rho) &= \sup_{\lambda} \left\{ -\frac{\lambda}{n}\sum_{i=1}^n (W_i - \theta) - (\rho-1)\left|\frac{\lambda}{\rho}\right|^{\frac{\rho}{\rho-1}} \right\} \\ &= \left| \frac{1}{n}\sum_{i=1}^n (W_i - \theta) \right|^\rho. \end{aligned}$$

Then, under the hypothesis that $E[W] = \theta_*$, and assuming $\mathrm{Var}[W] = \sigma_W^2 < \infty$, we obtain,

$$n^{\rho/2} R_n(\theta_*;\rho) \Rightarrow \bar{R}(\rho) \sim \sigma_W^\rho |N(0,1)|^\rho,$$

where $N(0,1)$ denotes a standard Gaussian random variable. The limiting distribution for the case $\rho = 1$ can be formally obtained by setting $\rho = 1$ in the above expression for $\bar{R}(\rho)$, but the analysis is slightly different. When $\rho = 1$,

$$\begin{aligned} R_n(\theta) &= \sup_{\lambda\in\mathbb{R}} \left\{ -\frac{\lambda}{n}\sum_{i=1}^n (W_i - \theta) - \frac{1}{n}\sum_{i=1}^n \sup_{u\in\mathbb{R}} \left\{ \lambda(u - W_i) - |u - W_i| \right\} \right\} \\ &= \sup_{\lambda} \left\{ -\frac{\lambda}{n}\sum_{i=1}^n (W_i - \theta) - \sup_{\Delta\in\mathbb{R}} \left\{ \lambda\Delta - |\Delta| \right\} \right\}. \end{aligned}$$

Following the notion that $\infty \times 0 = 0$,

$$\begin{aligned} R_n(\theta) &= \sup_{\lambda} \left\{ \frac{\lambda}{n}\sum_{i=1}^n (W_i - \theta) - \infty I(|\lambda| > 1) \right\} \\ &= \max_{|\lambda|\le 1} \frac{\lambda}{n}\sum_{i=1}^n (W_i - \theta) = \left| \frac{1}{n}\sum_{i=1}^n (W_i - \theta) \right|. \end{aligned}$$

So, indeed if $E[W] = \theta_*$ and $Var[W] = \sigma_W^2 < \infty$, we obtain

$$n^{1/2} R_n(\theta_*) \Rightarrow \sigma_W |N(0,1)|.$$

We now discuss far reaching extensions to the developments in Example 1 by considering estimating equations that are more general. First, we state a general asymptotic stochastic upper bound, which we believe is the most important result from an applied standpoint as it captures the speed of convergence of $R_n(\theta_*)$ to zero. Following this, we obtain an asymptotic

stochastic lower bound that matches with the upper bound (and therefore the weak limit) under mild, additional regularity conditions. We discuss the nature of these additional regularity conditions, and also why the lower bound in the case $\rho = 1$ can be obtained basically without additional regularity.

For the asymptotic upper bound we shall impose the following assumptions.

Assumptions:

A1) Assume that $c(u,w) = \|u-w\|_q^\rho$ for $q \geq 1$ and $\rho \geq 1$. For a chosen $q \geq 1$, let p be such that $1/p + 1/q = 1$.

A2) Suppose that $\theta_* \in \mathbb{R}^l$ satisfies $\mathbb{E}[h(W,\theta_*)] = \mathbf{0}$ and $\mathbb{E}\|h(W,\theta_*)\|_2^2 < \infty$. (While we do not assume that θ_* is unique, the results are stated for a fixed θ_* satisfying $E[h(W,\theta_*)] = \mathbf{0}$.)

A3) Suppose that the function $h(\cdot,\theta_*)$ is continuously differentiable with derivative $D_w h(\cdot,\theta_*)$.

A4) Suppose that for each $\zeta \neq 0$,

$$P\left(\left\|\zeta^T D_w h(W,\theta_*)\right\|_p > 0\right) > 0. \tag{12}$$

In order to state the theorem, let us introduce the notation for asymptotic stochastic upper bound,

$$n^{\rho/2} R_n(\theta_*;\rho) \lesssim_D \bar{R}(\rho),$$

which expresses that for every continuous and bounded non-decreasing function $f(\cdot)$ we have that

$$\overline{\lim}_{n\to\infty} E\left[f\left(n^{\rho/2} R_n(\theta_*;\rho)\right)\right] \leq E\left[f\left(\bar{R}(\rho)\right)\right].$$

Similarly, we write $\gtrsim_D$ for an asymptotic stochastic lower bound, namely

$$\underline{\lim}_{n\to\infty} E\left[f\left(n^{\rho/2} R_n(\theta_*;\rho)\right)\right] \geq E\left[f\left(\bar{R}(\rho)\right)\right].$$

Therefore, if both stochastic upper and lower bounds hold, then $n^{\rho/2} R_n(\theta_*;\rho) \Rightarrow \bar{R}(\rho)$ as $n \to \infty$. (see, for example, [4]). Now we are ready to state our asymptotic upper bound.

Theorem 1. *Under Assumptions A1) to A4) we have, as $n \to \infty$,*

$$n^{\rho/2} R_n(\theta_*;\rho) \lesssim_D \bar{R}(\rho),$$

where, for $\rho > 1$,

$$\bar{R}(\rho) := \max_{\zeta\in\mathbb{R}^r} \left\{\rho\zeta^T H - (\rho-1) E\left\|\zeta^T D_w h(W,\theta_*)\right\|_p^{\rho/(\rho-1)}\right\},$$

and if $\rho = 1$,

$$\bar{R}(1) := \max_{\zeta: P\left(\|\zeta^T D_w h(W,\theta_*)\|_p > 1\right) = 0} \{\zeta^T H\}.$$

In both cases $H \sim \mathcal{N}(\mathbf{0}, Cov[h(W,\theta_*)])$, and $Cov[h(W,\theta_*)] = E\left[h(W,\theta_*)h(W,\theta_*)^T\right]$.

We remark that as $\rho \to 1$, one can verify that $\bar{R}(\rho) \Rightarrow \bar{R}(1)$, so formally one can simply keep in mind the expression $\bar{R}(\rho)$ with $\rho > 1$. In turn, it is intersting to note that $\bar{R}(\rho)$ is Fenchel transform as a function of H_n. We now study some sufficient conditions which guarantee that $\bar{R}(\rho)$ is also an asymptotic lower bound for $n^{\rho/2} R_n(\theta_*; \rho)$. We consider the case $\rho = 1$ first, which will be used in applications to logistic regression discussed later in the paper.

Proposition 2. *In addition to assuming A1) to A4), suppose that W has a positive density (almost everywhere) with respect to the Lebesgue measure. Then,*

$$n^{1/2} R_n(\theta_*; 1) \Rightarrow \bar{R}(1).$$

The following set of assumptions can be used to obtain tight asymptotic stochastic lower bounds when $\rho > 1$; the corresponding result will be applied to the context of generalized LASSO.

A5) (*Growth condition*) Assume that there exists $\kappa \in (0, \infty)$ such that for $\|w\|_q \geq 1$,

$$\|D_w h(w, \theta_*)\|_p \leq \kappa \|w\|_q^{\rho - 1}, \tag{13}$$

and that $E\|W_i\|^\rho < \infty$.

A6) (*Locally Lipschitz continuity*) Assume that there exists exists $\bar{\kappa} : \mathbb{R}^m \to [0, \infty)$ such that,

$$\|D_w h(w + \Delta, \theta_*) - D_w h(w, \theta_*)\|_p \leq \bar{\kappa}(W_i) \|\Delta\|_q,$$

for $\|\Delta\|_q \leq 1$, and

$$E[\bar{\kappa}(W_i)^c] < \infty,$$

for $c \leq \max\{2, \frac{\rho}{\rho - 1}\}$.

We now summarize our last weak convergence result of this section.

Proposition 3. *If Assumptions A1) to A6) are in force and $\rho > 1$, then*

$$n^{\rho/2} R_n(\theta_*; \rho) \Rightarrow \bar{R}(\rho).$$

Before we move on with the applications of the previous results, it is worth discussing the nature of the additional assumptions introduced to ensure that an asymptotic lower bound can be obtained which matches the upper bound in Theorem 1.

As we shall see in the technical development in Section 5.1 where the proofs of the above results are furnished, the dual formulation of RWP function in Proposition 1 can be re-expressed, assuming only A1) to A4), as,

$$n^{\rho/2} R_n(\theta_*; \rho) = \sup_{\zeta} \left\{ \zeta^T H_n - \frac{1}{n} \sum_{k=1}^{n} \sup_{\Delta} \left\{ \int_0^1 \zeta^T Dh\left(W_i + \Delta u / n^{1/2}, \theta_*\right) \Delta du - \|\Delta\|_q^\rho \right\} \right\}. \tag{14}$$

In order to make sure that the lower bound asymptotically matches the upper bound obtained in Theorem 1 we need to make sure that we rule out cases in which the inner supremum is infinite in (14) with positive probability in the prelimit.

In Proposition 2 we assume that W has a positive density with respect to the Lebesgue measure because in that case the condition

$$P\left(\left\|\zeta^T Dh\left(W,\theta_*\right)\right\|_p \leq 1\right) = 1,$$

(which appears in the upper bound obtained in Theorem 1) implies that $\left\|\zeta^T Dh\left(w,\theta_*\right)\right\|_p \leq 1$ almost everywhere with respect to the Lebesgue measure. Due to the appearance of the integral in the inner supremum in (14), an upper bound can be obtained for the inner supremum, which translates into a tight lower bound for $n^{\rho/2} R_n\left(\theta_*\right)$.

Moving to the case $\rho > 1$ studied in Proposition 3, condition (13) in A5) guarantees that (for fixed W_i and n)

$$\left\|Dh\left(W_i + \Delta u/n^{1/2}, \theta_*\right)\Delta\right\| = O\left(\|\Delta\|_q^\rho / n^{(\rho-1)/2}\right),$$

as $\|\Delta\|_q \to \infty$. Therefore, the cost term $\left(-\|\Delta\|_q^\rho\right)$ in (14) will ensure a finite optimum in the prelimit for large n. The condition that $E\|W\|_q^\rho < \infty$ is natural because we are using a optimal transport cost $c\left(u,w\right) = \|u-w\|_q^\rho$. If this condition is not satisfied, then the underlying nominal distribution is at infinite transport distance from the empirical distribution.

The local Lipschitz assumption A6) is just imposed to simplify the analysis and can be relaxed; we have opted to keep A6) because we consider it mild in view of the applications that we will study in the sequel.

3. Distributionally Robust Estimators for Machine Learning Algorithms

A common theme in machine learning problems is to find the best fitting parameter in a family of parameterized models that relate a vector of predictor variables $X \in \mathbb{R}^d$ to a response $Y \in \mathbb{R}$. In this section, we shall focus on a useful class of such models, namely, linear and logistic regression models. Associated with these models, we have a loss function $l(X_i, Y_i; \beta)$ which evaluates the fit of regression coefficient β for the given data points $\{(X_i, Y_i) : i = 1, \ldots, n.\}$ Then, just as we explained in the case of generalized LASSO in the Introduction, our first step will be to show that regularized linear and logistic regression estimators admit a Distributionally Robust Optimization (DRO) formulation of the form,

$$\min_{\beta\in\mathbb{R}^d} \sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta} E_P\left[l\left(X,Y;\beta\right)\right]. \tag{15}$$

Once we derive a representation such as (15) then we will proceed, in the next section to find the optimal choice of δ, which, as explained in the Introduction, will immediately characterize the optimal regularization parameter.

In contrast to the empirical risk minimization that performs well only on the training data, the DRO problem (15) finds an optimizer β that performs uniformly well over all probability measures in the neighborhood that can be perceived as perturbations to the empirical training data distribution. Hence the solution to (15) is said to be "distributionally robust", and can be expected to generalize better. See [44, 45] and [35] for works that relate robustness and generalization.

Recasting regularized regression as a DRO problem of form (15) lets us view these regularized estimators under the lens of distributional robustness. The regularized estimators that we consider in this section, in particular, include the following.

Example 2 (Generalized-LASSO)**.** We have already started discussing this example in the Introduction, namely given a set of training data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, with predictor $X_i \in \mathbb{R}^d$ and response $Y_i \in \mathbb{R}$, the postulated model is $Y_i = \beta_*^T X_i + e_i$ for some $\beta_* \in \mathbb{R}^d$ and errors $\{e_1, ..., e_n\}$. The underlying loss function is $l(x, y; \beta) = \left(y - \beta^T x\right)^2$ and the generalized LASSO estimator, is obtained by solving the problem,

$$\min_{\beta \in \mathbb{R}^d} \left\{ \sqrt{E_{P_n}\left[l\left(X, Y; \beta\right)\right]} + \lambda \left\|\beta\right\|_1 \right\},$$

see [3, 1, 27] for more on generalized LASSO. As P_n denotes the empirical distribution corresponding to training samples, $E_{P_n}\left[l\left(X, Y; \beta\right)\right]$ is just the mean square training loss. In addition to the Generalized LASSO estimator above with ℓ_1 penalty, we derive a DRO representation of the form (15) for ℓ_p-penalized estimators obtained by solving,

$$\min_{\beta \in \mathbb{R}^d} \left\{ \sqrt{E_{P_n}\left[l\left(X, Y; \beta\right)\right]} + \lambda \left\|\beta\right\|_p \right\}, \tag{16}$$

for any $p \in [1, \infty)$.

Example 3 (Regularized Logistic Regression)**.** We next consider the context of binary classification, in which case the data is of the form $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, with $X_i \in \mathbb{R}^d$, response $Y_i \in \{-1, 1\}$ and the model postulates that

$$\log\left(\frac{P\left(Y_i = 1 | X_i = x\right)}{1 - P\left(Y_i = 1 | X_i = x\right)}\right) = \beta_*^T x$$

for some $\beta_* \in \mathbb{R}^d$. In this case, the log-exponential loss function (or negative log-likelihood for binomial distribution) is

$$l\left(x, y; \beta\right) = \log\left(1 + \exp(-y \cdot \beta^T x)\right),$$

and one is interested in estimating β_* by solving

$$\min_{\beta \in \mathbb{R}^d} \left\{ E_{P_n}\left[l\left(X, Y; \beta\right)\right] + \lambda \left\|\beta\right\|_p \right\}, \tag{17}$$

for $p \in [1, \infty)$ (see [16] for a discussion on regularized logistic regressions).

The rest of this section is to show that generalized LASSO and Regularized Logistic Regression estimators are distributionally robust (in the sense, they admit a representation of the form (15)).

While these particular examples may be certainly interesting, we emphasize that the DRO formulation (15) should be viewed, in its entirety, as a framework for generating distributionally robust inference procedures for different models and loss functions, without having to prove equivalences with an existing or popular algorithm.

3.1. **Dual form of the DRO formulation (15).** Though the DRO formulation (15) involves optimizing over uncountably many probability measures, the following result ensures that the inner supremum in (15) over the neighborhood $\{P : \mathcal{D}_c(P, P_n) \leq \delta\}$ admits a reformulation which is a simple, univariate optimization problem. Before stating the result, we recall that the definition of discrepancy measure $\mathcal{D}_c$ (defined in (6)) requires the specification of cost function $c\left((x, y), (x', y')\right)$ between any two predictor-response pairs $(x, y), (x', y') \in \mathbb{R}^{d+1}$.

Proposition 4. *Let $c(\cdot)$ be a nonnegative, lower semi-continuous cost function such that the set $\{((x,y),(x',y')) : c((x,y),(x',y')) < \infty\}$ is Borel measurable and nonempty. For $\gamma \geq 0$ and loss functions $l(x,y;\beta)$ that are upper semi-continuous in (x,y) for each β, let*

$$\phi_\gamma(X_i,Y_i;\beta) = \sup_{u\in\mathbb{R}^d,\ v\in\ \mathbb{R}} \left\{ l(u,v;\beta) - \gamma c\big((u,v),(X_i,Y_i)\big) \right\}. \tag{18}$$

Then

$$\sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\big[l(X,Y;\beta)\big] = \min_{\gamma\geq 0}\left\{ \gamma\delta + \frac{1}{n}\sum_{i=1}^n \phi_\gamma(X_i,Y_i;\beta)\right\}.$$

Consequently, the DR regression problem (15) reduces to

$$\inf_{\beta\in\mathbb{R}^d}\sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\big[l(X,Y;\beta)\big] = \inf_{\beta\in\mathbb{R}^d}\min_{\gamma\geq 0}\left\{ \gamma\delta + \frac{1}{n}\sum_{i=1}^n \phi_\gamma(X_i,Y_i;\beta)\right\}. \tag{19}$$

Such reformulations have recently gained much attention in the literature of distributionally robust optimization (see [12, 6, 15]). For a proof of Proposition 4, see Appendix A.

3.2. **Distributionally Robust Representations.**

3.2.1. *Example 2 (continued): Recovering regularized estimators for linear regression.* We examine the right-hand side of (19) for the square loss function for the linear regression model $Y = \beta^T X + e$, and obtain the following result without any further distributional assumptions on X, Y and the error e. For brevity, let $\bar{\beta} = (-\beta, 1)$, and recall the definition of the discrepancy measure $\mathcal{D}_c$ in (6).

Proposition 5 (DR linear regression with square loss)**.** *Fix $q \in (1,\infty]$. Consider the square loss function and second order discrepancy measure $\mathcal{D}_c$ defined using ℓ_q-norm. In other words, take $l(x,y;\beta) = (y-\beta^T x)^2$ and $c\big((x,y),(u,v)\big) = \|(x,y)-(u,v)\|_q^2$. Then,*

$$\min_{\beta\in\mathbb{R}^d}\sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\big[l(X,Y;\beta)\big] = \min_{\beta\in\mathbb{R}^d}\left(\sqrt{MSE_n(\beta)} + \sqrt{\delta}\ \|\bar{\beta}\|_p\right)^2, \tag{20}$$

where $MSE_n(\beta) = E_{P_n}[(Y-\beta^T X)^2] = \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^T X_i)^2$ is the mean square error for the coefficient choice β, and p is such that $1/p + 1/q = 1$.

As an important special case, we consider $q=\infty$ and identify the following equivalence for DR regression applying discrepancy measure based on neighborhoods defined using ℓ_∞ norm:

$$\arg\min_{\beta\in\mathbb{R}^d}\sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\big[l(X,Y;\beta)\big] = \arg\min_{\beta\in\mathbb{R}^d}\left\{\sqrt{MSE_n(\beta)} + \sqrt{\delta}\ \|\beta\|_1\right\}.$$

Here the right hand side is same as the generalized LASSO estimator with $\lambda = \sqrt{\delta}$ in Example 2.

The right hand side of (20) resembles ℓ_p-norm regularized regression (except for the fact that we have $\|\bar{\beta}\|_p$ instead of $\|\beta\|_p$). In order to obtain a closer equivalence we must introduce a slight modification to the norm $\|\cdot\|_q$ to be used as the cost function, $c(\cdot)$, in defining $\mathcal{D}_c$. We define

$$N_q\big((x,y),(u,v)\big) = \begin{cases} \|x-u\|_q, & \text{if } y = v \\ \infty, & \text{otherwise.}\end{cases}, \tag{21}$$

to use $c(\cdot) = N_q(\cdot)$ as the cost instead of the standard ℓ_q norm $\|(x,y)-(u,v)\|_q$. Subsequently, one can consider modified cost functions of form $c((x,y),(u,v)) = (N_q((x,y),(u,v)))^a$. As this modified cost function assigns infinite cost when $y \neq v$, the infimum in (4) is effectively over joint distributions that do not alter the marginal distribution of Y. As a consequence, the resulting neighborhood set $\{P : \mathcal{D}_c(P,P_n) \leq \delta\}$ admits distributional ambiguities only with respect to the predictor variables X.

The following result is essentially the same as Proposition 5 except for the use of the modified cost N_q and the resulting norm regularization of form $\|\beta\|_p$ (instead of $\|\bar{\beta}\|_p$ as in Proposition 5), thus exactly recovering the regularized regression estimators in Example 2.

Theorem 2. *Consider the square loss and discrepancy measure $\mathcal{D}_c(P,P_n)$ defined as in (6) using the cost function $c((x,y),(u,v)) = (N_q((x,y),(u,v)))^2$ (the function N_q is defined in (21)). Then,*

$$\min_{\beta\in\mathbb{R}^d} \sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\big[l(X,Y;\beta)\big] = \min_{\beta\in\mathbb{R}^d} \Big(\sqrt{MSE_n(\beta)} + \sqrt{\delta}\ \|\beta\|_p\Big)^2,$$

where $MSE_n(\beta) = E_{P_n}[(Y-\beta^T X)^2] = n^{-1}\sum_{i=1}^n (Y_i - \beta^T X_i)^2$ is the mean square error for the coefficient choice β, and p is such that $1/p + 1/q = 1$.

3.2.2. *Example 3 (continued): Recovering regularized estimators for classification.* Apart from exactly recovering well-known norm regularized estimators for linear regression, the discrepancy measure $\mathcal{D}_c$ based on the modified norm N_q in (21) is natural when our interest is in learning problems where the responses Y_i take values in a finite set – as in the binary classification problem where the response variable Y takes values in $\{-1,+1\}$.

The following result allows us to recover the DRO formulation behind the regularized logistic regression estimators discussed in Example 3.

Theorem 3 (Regularized regression for Classification)**.** *Consider the discrepancy measure $\mathcal{D}_c(\cdot)$ defined using the cost function $c((x,y),(u,v)) = N_q((x,y),(u,v))$ in (21). Then, for logistic regression with log-exponential loss function and Support Vector Machine (SVM) with Hinge loss,*

$$\min_{\beta\in\mathbb{R}^d} \sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\big[\log(1+e^{-Y\beta^T X})\big] = \min_{\beta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \log\Big(1+e^{-Y_i\beta^T X_i}\Big) + \delta\,\|\beta\|_p,$$

and

$$\min_{\beta\in\mathbb{R}^d} \sup_{P:\ D_c(P,P_n)\leq\delta} E_P\big[(1-Y\beta^T X)^+\big] = \frac{1}{n}\sum_{i=1}^n (1-Y_i\beta^T X_i)^+ + \ \delta\,\|\beta\|_p,$$

where p is such that $1/p + 1/q = 1$.

The proof of all of the results in this subsection are provided in Section 5.2.

4. Using RWPI for optimal regularization

Our goal in this section is to use RWP function for optimal regularization in Examples 2 and 3. As explained in the Introduction, the key step is to propose a reasonable optimality criterion for the selection of δ in the DRO formulation (15). Then, owing to the DRO representations derived in Section 3.2, this would imply an automatic choice of regularization parameter $\lambda = \sqrt{\delta}$ in generalized LASSO example (following Theorem 2), or $\lambda = \delta$ in regularized logistic

regression (following Theorem 3). In the development below, we follow the logic described in the Introduction for the generalized LASSO setting.

We write $\mathcal{U}_\delta(P_n)$ to denote the uncertainty set, namely $\mathcal{U}_\delta(P_n) = \{P : \mathcal{D}_c\big(P, P_n\big) \leq \delta\}$, and β_* to denote the underlying linear or logistic regression model parameter from which the training samples $\{(X_i, Y_i) : i = 1, \ldots, n\}$ are obtained. Now, for each P, convexity considerations involving the loss functions $l\left(x, y; \beta\right)$, as a function of β, will allow us to conclude that the set

$$\mathcal{P}_{opt}(\beta) := \Big\{P \in \mathcal{P}\big(\mathbb{R}^d \times \mathbb{R}\big) : E_P\big[D_\beta l(X, Y; \beta_*)\big] = \mathbf{0}\Big\}$$

is the set of probability measures for which β is an optimal risk minimization parameter.

As indicated in the Introduction, we shall say that β_* is plausible for a given choice of δ if,

$$\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing.$$

If this intersection is empty, we say that β_* is implausible. Moreover, we remark that β_* is plausible with confidence at least $1 - \alpha$ if,

$$P\left(\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing\right) \geq 1 - \alpha.$$

We shall argue in Appendix B that the inf sup in the corresponding DRO formulation (15) of each of the machine learning algorithms that we consider can be exchanged as below:

Lemma 1. *In the settings of Theorems 2 and 3, if $E\|X\|_2^2 < \infty$, we have that*

$$\inf_{\beta \in \mathbb{R}^d} \sup_{P \in\, \mathcal{U}_\delta(P_n)} E_P\left[l\big(X, Y; \beta\big)\right] = \sup_{P \in\, \mathcal{U}_\delta(P_n)} \inf_{\beta \in \mathbb{R}^d} E_P\left[l\big(X, Y; \beta\big)\right]. \tag{22}$$

The representation in the right hand side of (22) implies that

$$\begin{aligned}
&\sup_{P \in\, \mathcal{U}_\delta(P_n)} \inf_{\beta \in \mathbb{R}^d} E_P\left[l\big(X, Y; \beta\big)\right] \\
&\quad = \sup_{P \in\, \mathcal{U}_\delta(P_n)} \Big\{E_P\left[l\big(X, Y; \beta\big)\right] : \beta \in \mathbb{R}^d \text{ such that } E_P\big[D_\beta l(X, Y; \beta)\big] = \mathbf{0}\Big\} \\
&\quad = \sup\Big\{E_P\left[l\big(X, Y; \beta\big)\right] : \beta \in \mathbb{R}^d \text{ such that } \mathcal{P}_{opt}(\beta) \cap \mathcal{U}_\delta(P_n) = \varnothing\Big\},
\end{aligned}$$

and this motivates our interest in finding a δ such that

$$\inf\big\{\delta : P\big(\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing\big) \geq 1 - \alpha\big\}. \tag{23}$$

asymptotically, as $n \to \infty$. In simple words, we wish to find the smallest value of δ for which β_* is plausible with at least $1 - \alpha$ confidence (see Figure 1).

Observe that as

$$R_n(\beta_*) = \inf\big\{\mathcal{D}_c(P, P_n) : P \in \mathcal{P}_{opt}(\beta_*)\big\},$$

we have,

$$P\big(\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing\big) = P\big(R_n(\beta_*) \leq \delta\big)$$

and therefore (23) is equivalent to

$$\inf\big\{\delta : P\left(R_n(\beta_*) \leq \delta\right) \geq 1 - \alpha\big\}, \tag{24}$$

thus obtaining that the optimal selection of δ as the $1 - \alpha$ quantile of $R_n(\beta_*)$.

Now, without knowing β_*, it is, of course, difficult to compute $R_n(\beta_*)$. However, assuming i.i.d. training data, we can obtain a limiting distribution for the quantity $nR_n(\beta_*)$ or $\sqrt{n}R_n(\beta_*)$, by applying results from Section 2.4.

Another consequence of Lemma 1 is that the set $\Lambda_n(\delta)$ of plausible values of β (i.e. β for which there exists $P \in \mathcal{U}_\delta(P_n)$ such that $E_P\big[D_\beta l(X,Y;\beta)\big] = \mathbf{0}$), contains the optimal solution obtained by solving the problem in the left hand side of (22). (If this was not the case, the left hand side in (22) would be strictly smaller than the right hand side of (22).) The fact that the estimator for β_* obtained by solving the left hand side in (22) is plausible, we believe, is a property which makes our selection of δ logically consistent with the ultimate goal of the overall estimation procedure, namely, choosing β_*.

4.1. **Linear regression models with squared loss function.** In this section, we derive the asymptotic limiting distribution of suitably scaled profile function corresponding to the estimating equation

$$E[(Y - \beta^T X)X] = \mathbf{0}.$$

The chosen estimating equation describes the optimality condition for square loss function $l(x,y;\beta) = (y - \beta^T x)^2$, and therefore, the corresponding $R_n(\beta_*)$ is a suitable for choosing δ as in (24), and the regularization parameter $\lambda = \sqrt{\delta}$ in Example 2.

Let H_0 denote the null hypothesis that the training samples $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ are obtained independently from the linear model $Y = \beta_*^T X + e$, where the error term e has zero mean, variance σ^2, and is independent of X. Let $\Sigma = \mathrm{Cov}[X]$.

Theorem 4. *Consider the discrepancy measure $\mathcal{D}_c(\cdot)$ defined as in (6) using the cost function $c((x,y),(u,v)) = (N_q((x,y),(u,v)))^2$ (the function N_q is defined in (21)). For $\beta \in \mathbb{R}^d$, let*

$$R_n(\beta) = \inf\big\{D_c(P,P_n) :\ E_P\big[(Y - \beta^T X)X\big] = \mathbf{0}\big\}.$$

Then, under the null hypothesis H_0,

$$nR_n(\beta_*) \Rightarrow L_1 := \max_{\xi \in \mathbb{R}^d}\left\{2\sigma\xi^T Z - E\left\|e\xi - (\xi^T X)\beta_*\right\|_p^2\right\},$$

as $n \to \infty$. In the above limiting relationship, $Z \sim \mathcal{N}(\mathbf{0},\Sigma)$. Further,

$$L_1 \overset{D}{\leq} L_2 := \frac{E[e^2]}{E[e^2] - (E|e|)^2}\|Z\|_q^2.$$

Specifically, if the additive error term e follows centered normally distribution, then

$$L_1 \overset{D}{\leq} L_2 := \frac{\pi}{\pi - 2}\|Z\|_q^2.$$

In the above theorem, the relationship $L_1 \overset{D}{\leq} L_2$ denotes that the limit law L_1 is stochastically dominated by L_2. We remark this notation $\overset{D}{\leq}$ for stochastic upper bound here is different from the notation $\lesssim_D$ introduced in Section 2.4 to denote asymptotic stochastic upper bound. A proof of Theorem 4 as an application of Theorem 1 and Proposition 3 is presented in Section 5.3.

Using Theorem 4 to obtain regularization parameter for (16). Let $\eta_{1-\alpha}$ denote the $(1-\alpha)$ quantile of the limiting random variable L_1 in Theorem 4, or its stochastic upper bound

L_2. If we choose $\delta = \eta_{1-\alpha}/n$, it follows from Theorem 4 that

$$P(R_n(\beta_*) \leq \delta) \geq 1 - \alpha,$$

asymptotically as $n \to \infty$, and consequently,

$$P\big(\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \emptyset\big) \geq 1 - \alpha.$$

In other words, the optimal regression coefficient β_* remains plausible (for the DRO formulation (15)) with probability exceeding $1 - \alpha$ with this choice of δ. Due to the distributionally robust representation derived in Theorem 2, a prescription for the uncertainty set size δ naturally provides the prescription, $\lambda = \sqrt{\delta}$, for the regularization parameter as well. The following steps summarize the guidelines for choosing the regularization parameter in the ℓ_p-penalized linear regression (16):

1) Draw samples Z from $\mathcal{N}(\mathbf{0}, \Sigma)$ to estimate the $1 - \alpha$ quantile of one of the random variables L_1 or L_2 in Theorem 4. Let us use $\hat{\eta}_{1-\alpha}$ to denote the estimated quantile. While L_2 is simply the norm of Z, obtaining realizations of limit law L_1 involves solving an optimization problem for each realization of Z. If $\Sigma = \mathrm{Cov}[X]$ is not known, one can use a simple plug-in estimator for $\mathrm{Cov}[X]$ in place of Σ.
2) Choose the regularization parameter λ to be

$$\lambda = \sqrt{\delta} = \sqrt{\hat{\eta}_{1-\alpha}/n}.$$

It is interesting to note that unlike the traditional LASSO algorithm, the prescription of regularization parameter in the above procedure is self-normalizing, in the sense that it does not depend on the variance of e.

4.2. **Logistic Regression with log-exponential loss function.** In this section, we apply results in Section 2.4 to prescribe regularization parameter for ℓ_p-penalized logistic regression in Example 3.

Let H_0 denote the null hypothesis that the training samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are obtained independently from a logistic regression model satisfying

$$\log\left(\frac{P(Y = 1|X = x)}{1 - P(Y = 1|X = x)}\right) = \beta_*^T x,$$

for predictors $X \in \mathbb{R}^d$ and corresponding responses $Y \in \{-1, 1\}$; further, under null hypothesis H_0, the predictor X has positive density almost everywhere with respect to the Lebesgue measure on $\mathbb{R}^d$. The log-exponential loss (or negative log-likelihood) that evaluates the fit of a logistic regression model with coefficient β is given by

$$l(x, y; \beta) = -\log p(y|x; \beta) = \log\big(1 + \exp(-y\beta^T x)\big).$$

If we let

$$h(x, y; \beta) = D_\beta l(x, y; \beta) = \frac{-yx}{1 + \exp(y\beta^T x)}, \tag{25}$$

then the optimality condition that the coefficient β^* satisfies is $E\left[h(x, y; \beta_*)\right] = \mathbf{0}$.

Theorem 5. *Consider the discrepancy measure $\mathcal{D}_c(\cdot)$ defined as in (6) using the cost function $c((x, y), (u, v)) = N_q((x, y), (u, v))$ (the function N_q is defined in (21)). For $\beta \in \mathbb{R}^d$, let*

$$R_n(\beta) = \inf\left\{D_c(P, P_n) : E_P\big[h(x, y; \beta)\big] = \mathbf{0}\right\},$$

where $h(\cdot)$ is defined in (25). Then, under the null hypothesis H_0,

$$\sqrt{n}R_n(\beta_*) \Rightarrow \; L_3 := \sup_{\xi \in A} \xi^T Z$$

as $n \to \infty$. In the above limiting relationship,

$$Z \sim \mathcal{N}\left(\mathbf{0}, E\left[\frac{XX^T}{(1+\exp(Y\beta_*^T X))^2}\right]\right) \text{ and } A = \left\{\xi \in \mathbb{R}^d : \; ess\ sup_{x,y}\left\|\xi^T D_x h(x,y;\beta)\right\|_p \; \leq 1\right\}.$$

Moreover, the limit law L_3 admits the following simpler stochastic bound:

$$L_3 \overset{D}{\leq} \; L_4 := \|\tilde{Z}\|_q,$$

where $\tilde{Z} \sim \mathcal{N}(\mathbf{0}, E[XX^T])$.

A proof of Theorem 4 as an application of Theorem 1 and Proposition 2 is presented in Section 5.3.

Using Theorem 5 to obtain regularization parameter for (17) . Similar to linear regression, the regularization parameter for Regularized Logistic Regression discussed in Example 3 can be chosen by the following procedure:

1) Estimate the $(1-\alpha)$ quantile of $L_4 := \|\tilde{Z}\|_q$, where $\tilde{Z} \sim \mathcal{N}(\mathbf{0}, E[XX^T])$. Let us use $\hat{\eta}_{1-\alpha}$ to denote the estimate of the quantile.
2) Choose the regularization parameter λ in the norm regularized logistic regression estimator (17) in Example 3 to be

$$\lambda = \delta = \hat{\eta}_{1-\alpha}/\sqrt{n}.$$

4.3. **Optimal regularization in high-dimensional generalized LASSO .** In this section, let us restrict our attention to the square-loss function $l(x,y;\beta) = (y - \beta^T x)^2$ for the linear regression model and the discrepancy measure D_c defined using the cost function $c = N_q$ with $q = \infty$ in (21). Then, due to Theorem 2, this corresponds to the interesting case of generalized LASSO or ℓ_2-LASSO that was rather a particular example in the class of ℓ_p penalized linear regression estimators considered in Section 4.1.

As an interesting byproduct of the RWP function analysis, the following theorem presents a prescription for regularization parameter even in high dimensional settings where the ambient dimension d is larger than the number of samples n. We introduce the growth parameter,

$$C(n,d) := \frac{E\,\|X\|_\infty}{\sqrt{n}} = \frac{E\left[\max_{i=1,\ldots,d}|X_i|\right]}{\sqrt{n}},$$

as a function of n and d, that will be useful in stating our results. In addition, we say that the predictors X have *sub-gaussian tails* if there exists a constant $a > 0$,

$$E\left[\exp(t^T X)\right] \leq \exp(a^2\|t\|_2^2/2)$$

for every $t \in \mathbb{R}^d$.

Theorem 6. *Let $E[X_i] = 0$ and $E[X_i^2] = 1$ for all $i = 1, \ldots, d$. Suppose the assumptions of Theorem 4 hold and assume the largest eigenvalue of $\Sigma = Cov[X]$ be $o(nC(n,d)^2)$. In addition, suppose that β_* satisfies a weak sparsity condition that $\|\beta_*\|_1 = o(1/C(n,d))$. Then*

$$nR_n(\beta_*) \lesssim_D \frac{\|Z_n\|_\infty^2}{Var|e|},$$

as $n, d \to \infty$. *Here,* $Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i X_i$. *In particular, if the predictors* X *have subgaussian tails, then we have*

$$nR_n(\beta_*) \lesssim_D \frac{Ee^2}{Ee^2 - \left(E|e|\right)^2} \|\tilde{Z}\|_\infty^2$$

where, $\tilde{Z}$ *follows the distribution* $\mathcal{N}(0, \Sigma)$. *Moreover, if the additive error* e *is normally distributed and* Σ *is the identity matrix, then the above stochastic bounds simplify to*

$$\sqrt{R_n(\beta_*)} \lesssim_D \frac{\pi}{\pi - 2} \frac{\Phi^{-1}(1 - \alpha/2d)}{\sqrt{n}},$$

with probability asymptotically larger than $1 - \alpha$. *Here,* $\Phi^{-1}(1-\alpha)$ *denotes the quantile* x *of the standard normal distribution* $\Phi(x) = 1 - \alpha$.

The prescription of regularization parameter as

$$\lambda = \sqrt{\delta} = \frac{\pi}{\pi - 2} \frac{\Phi^{-1}(1 - \alpha/2d)}{\sqrt{n}} = O\left(\sqrt{\frac{\log d}{n}}\right), \tag{26}$$

as in Theorem 6, is consistent with the findings in the literature of high-dimensional linear regression (see, for example, [3, 22, 46, 2]). This agreement strengthens the interpretation of regularization parameter in regularized regression as $\sqrt{R_n(\beta_*)}$, which, in turn, corresponds to the distance of the empirical distribution P_n from the set $\{P : E_P[(Y - \beta^T X)X] = \mathbf{0}\}$.

It is also interesting to note that unlike traditional LASSO algorithm, the prescription of regularization parameter as in (26) is self-normalizing, in the sense that it does not depend on the variance of e, even if the number of predictors d is larger than n.

5. Proofs of main results

This section, comprising the proofs of the main results, is organized as follows. Subsection 5.1 contains the proofs of stochastic upper and lower bounds (and hence weak limits) presented in Section 2.4. While Subsection 5.2 is devoted to derive the results on distributionally robust representations presented in Section 3.2, Subsection 5.3 contains the proofs of Theorems 4 and 5 as applications of the stochastic upper and lower bounds presented in Section 2.4. Some of the useful technical results that are not central to the argument are presented in the appendix.

5.1. Proofs of asymptotic stochastic upper and lower bounds of RWP function in Section 2.4.

We first use Proposition 1 to derive a dual formulation for $n^{\rho/2} R_n(\theta_*)$ which will be the starting point of our analysis. Due to Assumption A2) $E[h(W, \theta_*)] = \mathbf{0}$, and therefore, $\mathbf{0}$ lies in the interior of convex hull of $\{h(u, \theta_*) : u \in \mathbb{R}^m\}$. Therefore, due to Proposition 1,

$$R_n(\theta_*) = \sup_{\lambda \in \mathbb{R}^r} \left\{ -\frac{1}{n} \sum_{i=1}^n \sup_{u \in \mathbb{R}^m} \left\{ \lambda^T h(u, \theta_*) - \|u - W_i\|_q^\rho \right\} \right\}.$$

In order to simplify the notation, throughout the rest of the proof we will write $h\left(W_i\right)$ instead of $h\left(W_i, \theta_*\right)$ and $Dh\left(W_i\right)$ for $D_w h\left(W_i, \theta_*\right)$.

Letting $H_n = n^{-1/2} \sum_{i=1}^n h(W_i)$ and changing variables to $\Delta = u - W_i$, we obtain

$$R_n(\theta_*) = \sup_{\lambda} \left\{ -\lambda^T \frac{H_n}{n^{1/2}} - \frac{1}{n} \sum_{i=1}^n \sup_{\Delta} \left\{ \lambda^T \big(h(W_i + \Delta) - h(W_i)\big) - \|\Delta\|_q^\rho \right\} \right\}.$$

Due to Fundamental calculus (using Assumption A3)), we have that

$$h\left(W_i+\Delta\right)-h\left(W_i\right)=\int_0^1 Dh\left(W_i+u\Delta\right)\Delta du.$$

Now, redefining $\zeta=\lambda n^{(\rho-1)/2}$ and $\Delta=\Delta/n^{1/2}$ we arrive at following representation

$$n^{\rho/2}R_n(\theta_*)=\max_{\zeta}\left\{-\zeta^T H_n-M_n\left(\zeta\right)\right\}, \tag{27}$$

where

$$M_n\left(\zeta\right)=\frac{1}{n}\sum_{i=1}^{n}\max_{\Delta}\left\{\zeta^T\int_0^1 Dh\left(W_i+n^{-1/2}\Delta u\right)\Delta du-\|\Delta\|_q^{\rho}\right\}. \tag{28}$$

The reformulation in (27) is our starting point of the analysis.

To proceed further, we first state a result which will allow us to apply a localization argument in the representation of $n^{\rho/2}R_n\left(\theta_*\right)$ in (27). Recall the definition of M_n above in (28) and that $H_n=n^{-1/2}\sum_{i=1}^n h(W_i)$.

Lemma 2. *Suppose that the Assumptions A2) to A4) are in force. Then, for every $\varepsilon>0$, there exists $n_0>0$ and $b\in(0,\infty)$ such that*

$$P\left(\max_{\|\zeta\|_p\geq b}\left\{-\zeta^T H_n-M_n\left(\zeta\right)\right\}>0\right)\leq\varepsilon,$$

for all $n\geq n_0$.

Proof of Lemma 2. For $\zeta\neq 0$, we write $\bar{\zeta}=\zeta/\left\|\zeta\right\|_p$. Let us define the vector $V_i\left(\bar{\zeta}\right)=\left(Dh\left(W_i\right)^T\bar{\zeta}\right)$, and put

$$\Delta_i'=\Delta_i'\left(\bar{\zeta}\right)=\left|V_i\left(\bar{\zeta}\right)\right|^{p/q}sgn\left(V_i\left(\bar{\zeta}\right)\right). \tag{29}$$

Define the set $C_0=\{\|W_i\|_p\leq c_0\}$, where c_0 will be chosen large enough momentarily. Then, for any $c>0$, plugging in $\Delta=c\Delta_i'$, we have $\zeta^T Dh(W_i)\Delta=c\|\zeta^T Dh(W_i)\|_p\|\Delta'\|_q$, and therefore,

$$\begin{aligned}
&\max_{\Delta}\left\{\zeta^T\int_0^1 Dh\left(W_i+n^{-1/2}\Delta u\right)\Delta-\|\Delta\|_q^{\rho}\right\}\\
&\quad\geq\max\Bigg\{c\left\|\zeta^T Dh\left(W_i\right)\right\|_p\left\|\Delta_i'\right\|_q-c^{\rho}\left\|\Delta_i'\right\|_q^{\rho}\\
&\qquad+c\zeta^T\int_0^1\left[Dh\left(W_i+n^{-1/2}\Delta_i'u\right)-Dh\left(W_i\right)\right]\Delta_i',0\Bigg\}I\left(W_i\in C_0\right).
\end{aligned} \tag{30}$$

Due to Hölder's inequality,

$$\begin{aligned}
&I\left(W_i\in C_0\right)\left|\zeta^T\int_0^1\left[Dh\left(W_i+n^{-1/2}\Delta_i'u\right)-Dh\left(W_i\right)\right]\Delta_i'du\right|\\
&\quad\leq\|\zeta\|_p\int_0^1 I\left(W_i\in C_0\right)\left\|\left[Dh\left(W_i+n^{-1/2}\Delta_i'u\right)-Dh\left(W_i\right)\right]\Delta_i'\right\|_q du.
\end{aligned}$$

Because of continuity $Dh\left(\cdot\right)$ and the fact that $W_i\in C_0$ (so the integrand is bounded), we have that the previous expression converges to zero as $n\to\infty$. Therefore for any $\varepsilon'>0$ (note than

convergence is uniform on $W_i \in C_0$), there exists n_0 such that for all $n \geq n_0$

$$cI\left(W_i \in C_0\right)\left|\zeta^T \int_0^1 [Dh\left(W_i + n^{-1/2}\Delta_i' u\right) - Dh\left(W_i\right)]\Delta_i' du\right| \leq c\varepsilon' \left\|\zeta\right\|_p. \tag{31}$$

Next, as $\|Dh(W_i)^T\bar{\zeta}\|_p = \|\Delta_i'\|_q^{q/p}$ and $q/p + 1 = q$,

$$c\left\|\zeta^T Dh\left(W_i\right)\right\|_p \left\|\Delta_i'\right\|_q - c^\rho \left\|\Delta_i'\right\|_q^\rho = c\left\|\zeta\right\|_p \left\|\Delta_i'\right\|_q^q - c^\rho \left\|\Delta_i'\right\|_q^\rho.$$

Now, since $\bar{\zeta} \to \left\|\Delta_i'\left(\bar{\zeta}\right)\right\|_q^q$ is Lipschitz continuous on $\left\|\bar{\zeta}\right\|_p = 1$, we conclude that,

$$\frac{1}{n}\sum_{i=1}^n \left\|\Delta_i'\left(\bar{\zeta}\right)\right\|_q^q I\left(W_i \in C_0\right) \to E\left[\left\|D_w^T h\left(W\right)\bar{\zeta}\right\|_q^q I\left(W \in C_0\right)\right],$$

with probability one as $n \to \infty$. Moreover, due to Fatou's lemma we have that the map $\bar{\zeta} \hookrightarrow P\left(\left\|\bar{\zeta}^T D_w h\left(W\right)\right\|_q > 0\right)$ is lower semi-continuous. Therefore, by A4), we have that there exists $\delta > 0$ such that

$$\inf_{\bar{\zeta}} E\left(\left\|\bar{\zeta}^T D_w h\left(W\right)\right\|_q^2\right) > \delta, \tag{32}$$

we conclude, by selecting $c_0 > 0$ large enough, using the equivalence between norms in Euclidean space, we have that for $n \geq N'\left(\delta\right)$, then

$$\frac{1}{n}\sum_{i=1}^n \left\|\Delta_i'\left(\bar{\zeta}\right)\right\|_q^q I\left(W_i \in C_0\right) > \frac{\delta}{2}. \tag{33}$$

Similarly, the map $\bar{\zeta} \hookrightarrow \left\|\Delta_i'\left(\bar{\zeta}\right)\right\|_q^\rho$ is Lipschitz continuous on $\left\|\bar{\zeta}\right\|_q = 1$, therefore one might assume that $N'\left(\delta\right)$ is actually chosen so that for $n \geq N'\left(\delta\right)$,

$$\frac{1}{n}\sum_{i=1}^n \left\|\Delta_i'\left(\bar{\zeta}\right)\right\|_q^\rho I\left(W_i \in C_0\right) < 2E\left[\left\|\Delta_i'\left(\bar{\zeta}\right)\right\|_q^\rho I\left(W_i \in C_0\right)\right].$$

As $\|\Delta_i'(\bar{\zeta})\|_q = \|\bar{\zeta}^T Dh(W_i)\|_p^{p/q}$, and $c_1 := \sup_{w \in C_0} \|\bar{\zeta}^T Dh(w)\|_p < \infty$,

$$\frac{1}{n}\sum_{i=1}^n \left\|\Delta_i'\left(\bar{\zeta}\right)\right\|_q^\rho I\left(W_i \in C_0\right) < 2c_1^{\rho\frac{p}{q}},$$

for all $n > N'(\delta)$. As a consequence, if $n \geq N'\left(\delta\right)$, it follows from (30), (31) and (33) that

$$\begin{aligned}\sup_{\left\|\zeta\right\|_p > b}\left\{-\zeta^T H_n - M_n\left(\zeta\right)\right\} &\leq \sup_{\left\|\zeta\right\|_p > b}\left\{-\zeta^T H_n - c\left\{\frac{\delta}{2}\left\|\zeta\right\|_p - \frac{2c_1^{\rho p/q}}{b} - \varepsilon'\left\|\zeta\right\|_p\right\}\right\} \\ &\leq \sup_{\left\|\zeta\right\|_p > b}\left\{-\zeta^T H_n - c\left\|\zeta\right\|_p\left\{\frac{\delta}{2} - \frac{2c_1^{\rho p/q}}{b} - \varepsilon'\right\}\right\}.\end{aligned}$$

From the previous expression, on the set $\left\|H_n\right\|_q \leq b'$, we have that if $b \geq 1$,

$$\sup_{\left\|\zeta\right\|_p > b}\left\{-\zeta^T H_n - M_n\left(\zeta\right)\right\} \leq \sup_{\left\|\zeta\right\|_p > b}\left\|\zeta\right\|_p\left[b' - c\left\{\frac{\delta}{2} - \frac{2c_1^{\rho p/q}}{b} - \varepsilon'\right\}\right].$$

Now, fix $c = 4b'/\delta$, $b > 16c_1^{\rho p/q}/\delta$, and $\varepsilon' = \delta/8$ to conclude that on $n \geq N'(\delta)$ and $\|H_n\|_q \leq b'$

$$b' - c\left\{\frac{\delta}{2} - \frac{2c_1^{\rho p/q}}{b} - \varepsilon'\right\} < 0.$$

Therefore, if $n \geq n_0$ (see (31)), then

$$P\left(\max_{\|\zeta\|_p > b}\left\{-2\zeta^T H_n - M_n(\zeta)\right\} > 0\right) \leq P\left(\|H_n\|_q > b'\right) + P\left(N'(\delta) > n\right).$$

The result now follows immediately from the previous inequality by choosing b' large enough so that $P(\|H_n\|_q > b') \leq \varepsilon/2$ and later n_0 so that $P(N'(\delta) > n_0) \leq \varepsilon/2$. The selection of b' is feasible due to A2). This proves the statement of Lemma 2 when $\rho > 1$.

The case $\rho = 1$ is similar. Since (32) holds we must have that there exists a compact set C such that $P(W \in C) > 0$ and (by continuity of $Dh(\cdot)$, ensured due to A3)) with the property that for any $w \in C$ and v such that $\|v\|_p \leq \delta'$

$$\min_{\|\zeta\|_p = 1} \left\|\zeta^T Dh(w+v)\right\|_q \geq \delta', \tag{34}$$

for some $\delta' > 0$. Once again, we define $\Delta_i' = \Delta_i'(\bar{\zeta})$ as in (29) – recall that $\|\bar{\zeta}\|_p = 1$. Observe that for $W_i \in C$, because of (34), since C is compact and $Dh(\cdot)$ is continuous, we may assume that $\delta' > 0$ actually satisfies (uniformly over $\bar{\zeta}$),

$$1/\delta' \geq \left\|\bar{\Delta}_i\right\|_q \geq \delta' > 0.$$

Consequently, letting $\bar{\Delta}_i = c\Delta_i'$, for $c > 1$, we have that for $W_i \in C$ that if $n^{1/2} \geq c/(\delta'\delta)$ then, by (34),

$$\begin{aligned}
&\max_{\bar{\Delta}_i}\left\{\zeta^T \int_0^1 Dh\left(W_i + \Delta_i u/n^{1/2}\right)\Delta_i du - \|\Delta_i\|_q\right\} \\
\geq\ &\left(c\zeta^T \int_0^1 Dh\left(W_i + c\Delta_i' u/n^{1/2}\right)\Delta_i' du - c\left\|\Delta_i'\right\|_q\right)^+ \\
\geq\ &\left(c\|\zeta\|_p \left\|\Delta_i'\right\|_q (\delta')^2 - c\left\|\Delta_i'\right\|_q\right)^+ \geq c\delta'\left\|\Delta_i'\right\|_q \left(\|\zeta\|_p (\delta')^2 - 1\right)^+.
\end{aligned} \tag{35}$$

Now choose $c = n^{1/2-\varepsilon'}$, for some $\varepsilon' > 0$ then we have that $cn^{1/2} \geq 1/(\delta'\delta)$, for $n > 1/(\delta'\delta)^{1/\varepsilon'}$, and therefore

$$\begin{aligned}
&\max_{\|\zeta\|_p > 1/(\delta')^2}\left\{-\zeta^T H_n - M_n(\zeta)\right\} \\
\leq\ &\max_{\|\zeta\|_p > 1/(\delta')^2}\left\{\|\zeta\|_p \|H_n\|_q - \frac{c\delta'\left(\|\zeta\|_p (\delta')^2 - 1\right)}{n}\sum_{i=1}^n I(W_i \in C)\right\} \to -\infty
\end{aligned}$$

as $n \to \infty$ in probability (by the Law of Large Numbers since $P(W \in C) > 0$). So, the lemma follows by choosing $b = 1/(\delta')^2$ and n sufficiently large. □

Lemma 3. *For any $b > 0$ and $c_0 \in (0, \infty)$*

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} \left\|\zeta^T Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)} I\left(\left\|Dh\left(W_i\right)\right\|_q \leq c_0\right) \\ \to \quad & E\left[\left\|\zeta^T Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)} I\left(\left\|Dh\left(W_i\right)\right\|_q \leq c_0\right)\right] \end{aligned}$$

uniformly over $\|\zeta\|_p \leq b$ in probability as $n \to \infty$.

Proof of Lemma 3. We first argue a suitable Lipschitz property for the map $\zeta \hookrightarrow \left\|\zeta^T Dh\left(W_i\right)\right\|_p^{\rho/(\rho-1)}$. It follows elementary that for any $0 \leq a_0 < a_1$ and $\gamma > 1$

$$a_1^\gamma - a_0^\gamma = \gamma \int_{a_0}^{a_1} t^{\gamma-1} dt \leq \gamma a_1^{\gamma-1}\left(a_1 - a_0\right).$$

Applying this observation with

$$\begin{aligned} a_1 &= \max\left(\left\|\zeta_1^T Dh\left(W_i\right)\right\|_q, \left\|\zeta_0^T Dh\left(W_i\right)\right\|_q\right), \\ a_0 &= \min\left(\left\|\zeta_1^T Dh\left(W_i\right)\right\|_q, \left\|\zeta_0^T Dh\left(W_i\right)\right\|_q\right), \\ \gamma &= \rho/(1-\rho), \end{aligned}$$

and using that

$$\left\|\zeta^T Dh\left(W_i\right)\right\|_q \leq b\left\|Dh\left(W_i\right)^T\right\|_q,$$

for $\|\zeta\|_q \leq b$ then we obtain

$$\begin{aligned} & \left|\left\|\zeta_0^T Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)} - \left\|\zeta_1^T Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)}\right| \\ \leq \quad & b^{\rho/(1-\rho)}\left\|Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)}\left\|\zeta_0 - \zeta_1\right\|_q. \end{aligned}$$

Therefore,

$$\begin{aligned} & \left|\left\|\zeta_0^T Dh\left(W_i\right)\right\|_q^{\rho/(1-\rho)} - \left\|\zeta_1^T Dh\left(W_i\right)\right\|_q^{\rho/(1-\rho)}\right| \\ \leq \quad & b^{\rho/(1-\rho)}\left\|\zeta_0 - \zeta_1\right\|_q\left\|Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)}. \end{aligned}$$

From this inequality we have that there exists a constant $c_0\left(b, \rho\right) \in (0, \infty)$

$$\begin{aligned} & \left|\frac{1}{n}\sum_{i=1}^{n}\left\|\zeta_0^T Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)} - \frac{1}{n}\sum_{i=1}^{n}\left\|\zeta_1^T Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)}\right| \\ \leq \quad & \left\|\zeta_0 - \zeta_1\right\|_q \frac{b^{\rho/(1-\rho)}}{n}\sum_{i=1}^{n}\left\|Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)}. \end{aligned}$$

Since

$$E\left(\left\|Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)} I\left(\left\|Dh\left(W_i\right)\right\|_q \leq c_0\right)\right) < \infty,$$

then we conclude tightness under the uniform topology on compact sets. The Strong Law of Large Numbers guarantees that finite dimensional distributions converge, and, since the limit is deterministic, we obtain convergence in probability. □

Proof of Theorem 1. Due to Lemma 2, since $R_n(\theta_*) \geq 0$ (choosing $\zeta = 0$), we can conclude that, as long as $n \geq n_0$, then one can select b such that the event

$$\mathcal{A}_n = \{n^{\rho/2} R_n(\theta_*) = \max_{\|\zeta\|_q \leq b} \{-2\zeta^T H_n - M_n(\zeta)\} \tag{36}$$

occurs with probability at least $1 - \varepsilon$.

We first consider the case $\rho > 1$. For $\zeta \neq 0$, we write $\bar{\zeta} = \zeta / \|\zeta\|_p$ and define the vector $V(\bar{\zeta})$ via $V_j(\bar{\zeta}) = \left(Dh(W_i)^T \bar{\zeta}\right)_j$ (i.e. $V_j(\bar{\zeta})$ is the j-th entry of $Dh(W_i)^T \bar{\zeta}$), and put

$$\Delta_i' = \Delta_i'(\bar{\zeta}) = |V_i(\bar{\zeta})|^{p/q} sgn(V_i(\bar{\zeta})). \tag{37}$$

Then, let $\bar{\Delta}_i = c^* \Delta_i'$ with c^* chosen so that

$$\|\bar{\Delta}_i\|_q = \frac{1}{\rho} \|\bar{\zeta}^T Dh(W_i)\|_p^{1/(\rho-1)}.$$

In such case we have that

$$\begin{aligned} & \max_{\Delta} \left\{\zeta^T Dh(W_i)\Delta - \|\Delta\|_p^\rho\right\} \\ = & \max_{\|\Delta\|_p \geq 0} \left\{\|\zeta^T Dh(W_i)\|_q \|\Delta\|_p - \|\Delta\|_p^\rho\right\} \\ = & \zeta^T Dh(W_i)\bar{\Delta}_i - \|\bar{\Delta}_i\|_p^\rho \\ = & \|\zeta^T Dh(W_i)\|_q^{\rho/(\rho-1)} \left(\frac{1}{\rho}\right)^{1/(\rho-1)} \left(1 - \frac{1}{\rho}\right). \end{aligned} \tag{38}$$

Pick $c_0 \in (0, \infty)$ and define $C_0 = \{\|W_i\|_p \leq c_0\}$. Note that

$$M_n(\zeta) \geq M_n'(\zeta, c_0),$$

where

$$M_n'(\zeta, c_0) = \frac{1}{n} \sum_{i=1}^n I(W_i \in C_0) \left\{\zeta^T \int_0^1 Dh\left(W_i + n_i^{-1/2} \bar{\Delta}_i u\right) \bar{\Delta}_i du - \|\bar{\Delta}_i\|_q^\rho\right\}^+.$$

Therefore

$$\max_{\|\zeta\|_q \leq b} \{-2\zeta^T H_n - M_n(\zeta)\} \leq \max_{\|\zeta\|_q \leq b} \{-2\zeta^T H_n - M_n'(\zeta, c_0)\}. \tag{39}$$

Define

$$\begin{aligned} \widehat{M}_n(\zeta, c_0) &= \frac{1}{n} \sum_{i=1}^n I(W_i \in C_0) \left\{\zeta^T Dh(W_i) \bar{\Delta}_i du - \|\bar{\Delta}_i\|_q^\rho\right\}^+ \\ &= \frac{1}{n} \sum_{i=1}^n I(W_i \in C_0) \|\zeta^T Dh(W_i)\|_p^{\rho/(\rho-1)} \left(\frac{1}{\rho}\right)^{1/(\rho-1)} \left(1 - \frac{1}{\rho}\right), \end{aligned}$$

where the equality follows from (38). We then claim that

$$\sup_{\|\zeta\|_q \leq b} \left|\widehat{M}_n(\zeta, c_0) - M_n'(\zeta, c_0)\right| \to 0. \tag{40}$$

In order to verify (40) note, using the continuity of $Dh(\cdot)$ that for any $\varepsilon' > 0$ there exists n_0 such that if $n \geq n_0$ then (uniformly over $\|\zeta\|_p \leq b$),

$$\left| \int_0^1 I(W_i \in C_0) \left\| Dh\left(W_i + n^{-1/2}\bar{\Delta}_i u\right) - Dh(W_i) \right\|_q \|\bar{\Delta}_i\|_q \, du \right| \leq \varepsilon'.$$

Therefore, if $n \geq n_0$,

$$\frac{1}{n}\sum_{i=1}^n I(W_i \in C_0) \left| \zeta^T \int_0^1 \left[Dh\left(W_i + n_i^{-1/2}\bar{\Delta}_i u\right) - Dh(W_i) \right] \bar{\Delta}_i du \right| \leq \varepsilon'.$$

Since $\varepsilon' > 0$ is arbitrary we conclude (40).

Then, applying Lemma 3 we obtain

$$\widehat{M}_n(\zeta, c_0) \to E\left\{ \zeta^T Dh(W_i)\, \bar{\Delta}_i du - \|\bar{\Delta}_i\|_q^\rho \right\}^+$$

uniformly over $\|\zeta\|_p \leq b$ as $n \to \infty$, in probability. Therefore, applying the continuous mapping principle, we have that

$$\begin{aligned} & \max_{\|\zeta\|_q \leq b} \left\{ -2\zeta^T H_n - M_n'(\zeta, c_0) \right\} \\ \Rightarrow \quad & \max_{\|\zeta\|_q \leq b} \left\{ -2\zeta^T H - \kappa(\rho) E\left[\left\| \zeta^T Dh(W_i) \right\|_q^{\rho/(\rho-1)} I\left(\|Dh(W_i)\|_q \leq c_0 \right) \right] \right\}, \end{aligned} \tag{41}$$

as $n \to \infty$, where

$$\kappa(\rho) = \left(\frac{1}{\rho}\right)^{1/(\rho-1)} \left(1 - \frac{1}{\rho}\right),$$

and $H \sim N(0, Cov[h(W, \theta_*)])$. From (39) and the construction of (36), we can easily obtain that $n^{\rho/2} R_n(\theta_*)$ is stochastically bounded (asymptotically) by

$$\max_{\zeta} \left\{ -2\zeta^T H - \kappa(\rho) E\left[\left\| \zeta^T Dh(W_i) \right\|_q^{\rho/(\rho-1)} \right] \right\}.$$

So, the first part of the theorem, concerning $\rho > 1$ follows.

Now, for $\rho = 1$, we will follow very similar steps. Again, due to Lemma 2 we concentrate on the region $\|\zeta\|_p \leq b$ for some $b > 0$. For the upper bound, define Δ_i' as in (37). Using a localization technique similar to that described in the proof of Lemma 2 in which the set C_0 as introduced we might assume that $\|\Delta_i'\|_p \leq c_0$ for some $c_0 > 0$. Then, for a given constant $c > 0$, set $\bar{\Delta}_i = c\Delta_i'$. We obtain that

$$\begin{aligned} & \max_{\|\zeta\| \leq b} \left\{ -\zeta^T H_n - \frac{1}{n}\sum_{i=1}^n \sup_{\bar{\Delta}_i} \left\{ \zeta^T \int_0^1 Dh\left(W_i + \Delta_i u/n^{1/2}\right) \Delta_i du - \|\Delta_i\|_q \right\} \right\} \\ \leq \quad & \max_{\|\zeta\| \leq b} \left\{ -\zeta^T H_n - \frac{1}{n}\sum_{i=1}^n \left(c\zeta^T \int_0^1 Dh\left(W_i + c\Delta_i' u/n^{1/2}\right) \Delta_i' du - c\left\|\Delta_i'\right\|_q \right) \right\}. \end{aligned}$$

As in the case $\rho > 1$ we have that

$$\frac{1}{n}\sum_{i=1}^n \int_0^1 \zeta^T \left[Dh\left(W_i + c\Delta_i' u/n^{1/2}\right) - Dh(W_i) \right] \Delta_i' du \to 0$$

in probability uniformly on ζ-compact sets. Similarly, for any $c > 0$ and any $b > 0$

$$\begin{aligned} & \max_{\|\zeta\| \leq b} \left\{ -\zeta^T H_n - \frac{1}{n} \sum_{i=1}^{n} c \left(\left\| \zeta^T Dh\left(W\right) \right\|_p - 1 \right)^+ \right\} \\ \Rightarrow \quad & \max_{\|\zeta\| \leq b} \left\{ -\zeta^T H - cE \left(\left\| \zeta^T Dh\left(W\right) \right\|_p - 1 \right)^+ \right\}, \end{aligned}$$

the constant c is arbitrary, so we obtain a stochastic upper bound of the form

$$\max_{\|\zeta\| \leq b : P\left(\|\zeta^T Dh(W)\|_p \leq 1\right) = 1} \left\{ -\zeta^T H \right\} \leq \max_{\zeta : P\left(\|\zeta^T Dh(W)\|_p \leq 1\right) = 1} \left\{ -\zeta^T H \right\}.$$

The proof of the Theorem follows. □

Proof of Proposition 2. We follow the notation introduced in the proof of Theorem 1. Also, in the proof of Theorem 1, we obtained that

$$n^{1/2} R_n\left(\theta_*\right) = \sup_{\zeta} [\zeta^T H_n - \frac{1}{n} \sum_{k=1}^{n} \sup_{\Delta} \{ \int_0^1 \zeta^T Dh\left(W_i + \Delta u / n^{1/2}\right) \Delta du - \|\Delta\|_q \}].$$

Now let $A = \{\zeta : \operatorname{esssup} \left\|\zeta^T Dh\left(w\right)\right\|_p \leq 1\}$, where the essential supremum is taken with respect to the Lebesgue measure, then if $\zeta \in A$,

$$\begin{aligned} & \sup_{\Delta} \{ \int_0^1 \zeta^T Dh\left(W_i + \Delta u / n^{1/2}\right) \Delta du - \|\Delta\|_q \} \\ \leq \quad & \sup_{\Delta} \{ \int_0^1 \left\| \zeta^T Dh\left(W_i + \Delta u / n^{1/2}\right) \right\|_p \|\Delta\|_q \, du - \|\Delta\|_q \} \\ \leq \quad & \sup_{\Delta} \|\Delta\|_q \{ \int_0^1 \left(\left\| \zeta^T Dh\left(W_i + \Delta u / n^{1/2}\right) \right\|_p - 1 \right) du \} \leq 0. \end{aligned}$$

Consequently,

$$n^{1/2} R_n\left(\theta_*\right) \geq \sup_{\zeta \in A} \zeta^T H_n.$$

Letting $n \to \infty$ we conclude that

$$\sup_{\zeta \in A} \zeta^T H_n \Rightarrow \sup_{\zeta \in A} \zeta^T H.$$

Because W_i is assumed to have a density with respect to the Lebesgue measure it follows that $P\left(\left\| \zeta^T Dh\left(W_i\right) \right\|_p \leq 1 \right) = 1$ if and only if $\zeta \in A$ and the result follows. □

Finally, we provide the proof of Proposition 3.

Proof of Proposition 3. As in the proof of Theorem 1, due to Lemma 2, we might assume that $\|\zeta\|_p \leq b$ for some $b > 0$. Next, we adopt the notation introduced in the proof of Theorem 1, in which we also established the representation

$$n^{\rho/2} R_n\left(\theta_*\right) = \sup_{\zeta} [\zeta^T H_n - \frac{1}{n} \sum_{k=1}^{n} \sup_{\Delta} \{ \int_0^1 \zeta^T Dh\left(W_i + \Delta u / n^{1/2}\right) \Delta du - \|\Delta\|_q^\rho \}].$$

The strategy will be to split the inner supremum in values of $\|\Delta\|_q \leq \delta n^{1/2}$ and values $\|\Delta\|_q > \delta n^{1/2}$ for $\delta > 0$ small. We will show that the supremum is achieved with high probability in the interval former region. Then, we will analyze the region in which $\|\Delta\|_q \leq \delta n^{1/2}$ and argue that the integral can be replaced by $\zeta^T Dh(W_i)\Delta$. Once this substitution is performed we can solve the inner maximization problem explicitly and, finally, we will apply a weak convergence result on ζ-compact sets to conclude the result. We now proceed to execute this strategy.

Pick $\delta > 0$ small, to be chosen in the sequel, then note that A5) implies (by redefining κ if needed, due to the continuity of $Dh(\cdot)$) that

$$\|Dh(w)\|_p \leq \kappa\left(1 + \|w\|_q^{\rho-1}\right).$$

Therefore,

$$\begin{aligned}
&\sup_{\|\Delta\|_q \geq \delta n^{1/2}} \{\int_0^1 \left|\zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right)\Delta\right| du - \|\Delta\|_q^\rho\} \\
\leq &\sup_{\|\Delta\|_q \geq \delta n^{1/2}} \{b\kappa\left(1 + \int_0^1 \left\|W_i + \Delta u/n^{1/2}\right\|_q^{\rho-1} du\right)\|\Delta\|_q - \|\Delta\|_q^\rho\}.
\end{aligned}$$

Note that if $\rho \in (1,2)$, then $0 < \rho - 1 < 1$ and therefore by the triangle inequality and concavity

$$\left\|W_i + \Delta u/n^{1/2}\right\|_q^{\rho-1} \leq \left(\|W_i\|_q + \left\|\Delta/n^{1/2}\right\|_q\right)^{\rho-1} \leq \|W_i\|_q^{\rho-1} + \left\|\Delta/n^{1/2}\right\|_q^{\rho-1}.$$

On the other hand, if $\rho \geq 2$, then $\rho - 1 \geq 1$ and the triangle inequality combined with Jensen's inequality applied as follows:

$$\|a+c\|^{\rho-1} \leq 2^{\rho-1}\|a/2 + c/2\|^{\rho-1} \leq 2^{\rho-2}\left(\|a\|^{\rho-1} + \|c\|^{\rho-1}\right),$$

yields

$$\left\|W_i + \Delta u/n^{1/2}\right\|_q^{\rho-1} \leq 2^{\rho-2}\left(\|W_i\|_q^{\rho-1} + \left\|\Delta/n^{1/2}\right\|_q^{\rho-1}\right).$$

So, in both cases we can write

$$\begin{aligned}
&\sup_{\|\Delta\|_q \geq \delta n^{1/2}} \{\int_0^1 \left|\zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right)\Delta\right| du - \|\Delta\|_q^\rho\} \\
\leq &\sup_{\|\Delta\|_q \geq \delta n^{1/2}} \{b\kappa\left(1 + 2^{\rho-2}\left(\|W_i\|_q^{\rho-1} + \|\Delta\|_q^{\rho-1}/n^{1/2}\right)\right)\|\Delta\|_q - \|\Delta\|_q^\rho\} \\
\leq &\sup_{\|\Delta\|_q \geq \delta n^{1/2}} \{b\kappa\left(\|\Delta\|_q + 2^{\rho-2}\|W_i\|_q^{\rho-1}\|\Delta\|_q + 2^{\rho-2}\|\Delta\|_q^\rho/n^{1/2}\right) - \|\Delta\|_q^\rho\}.
\end{aligned}$$

Next, we have that for any $\varepsilon' > 0$,

$$P\left(\|W_n\|_q^\rho \geq \varepsilon' n \text{ i.o.}\right) = 0,$$

therefore we might assume that there exists N_0 such that for all $i \leq n$ and $n \geq n_0$, $\|W_i\|_q^{\rho-1} \leq (\varepsilon' n)^{(\rho-1)/\rho}$. Therefore, if $(\varepsilon')^{(\rho-1)/\rho} \leq \delta^{\rho-1}/\left(b\kappa 2^{\rho-1}\right)$, we conclude that if $\|\Delta\|_q \geq \delta n^{1/2}$,

$$\begin{aligned}
b\kappa 2^{\rho-1}\|W_i\|_q^{\rho-1}\|\Delta\|_q &\leq b\kappa 2^{\rho-1}\left(\varepsilon' n\right)^{(\rho-1)/\rho}\|\Delta\|_q \\
&\leq \frac{1}{2}\delta^{\rho-1} n^{(\rho-1)/2}\|\Delta\|_q \leq \frac{1}{2}\|\Delta\|_q^\rho.
\end{aligned}$$

Similarly, choosing n sufficiently large we can guarantee that

$$b\kappa\left(\|\Delta\|_q + 2^{\rho-2}\|\Delta\|_q^\rho/n^{1/2}\right) \leq \frac{1}{2}\|\Delta\|_q^\rho.$$

Therefore, we conclude that

$$\sup_{\|\Delta\|_q \geq \delta\sqrt{n}} \{\int_0^1 \left|\zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right)\Delta\right| du - \|\Delta\|_q^\rho\} \leq 0$$

provided n is large enough, for any $\delta > 0$.

Now, we let $\varepsilon'' > 0$, and note that

$$\begin{aligned}
& \sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \{\int_0^1 \zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right)\Delta du - \|\Delta\|_q^\rho\} \\
\leq & \sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \{\int_0^1 \zeta^T \left[Dh\left(W_i + \Delta u/n^{1/2}\right) - Dh\left(W_i\right)\right]\Delta du - \varepsilon''\|\Delta\|_q^\rho\} \\
& + \sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \{\zeta^T Dh\left(W_i\right)\Delta - (1-\varepsilon'')\|\Delta\|_q^\rho\}.
\end{aligned}$$

We now argue locally, using A6), for the first term in the right hand side of the previous display. To wit,

$$\begin{aligned}
& \sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \{\int_0^1 \zeta^T \left[Dh\left(W_i + \Delta u/n^{1/2}\right) - Dh\left(W_i\right)\right]\Delta du - \varepsilon''\|\Delta\|_q^\rho\} \\
\leq & \sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \{b\bar{\kappa}\left(W_i\right)\|\Delta\|_q^2/n^{1/2} - \varepsilon''\|\Delta\|_q^\rho\} \\
= & \sup_{\|\bar{\Delta}\|_q \leq 1} \{b\bar{\kappa}\left(W_i\right)\|\bar{\Delta}\|_q^2 \delta^2 n^{1/2} - \varepsilon''\|\bar{\Delta}\|_q^\rho \left(\delta n^{1/2}\right)^\rho\} \\
= & \left(\delta b\bar{\kappa}\left(W_i\right)\right)^{\rho/(\rho-1)} \left(\varepsilon''\right)^{-1/(\rho-1)} (\rho-1)/\rho. \qquad (42)
\end{aligned}$$

Using A6), we have that

$$E\left(\bar{\kappa}\left(W_i\right)^{\rho/(\rho-1)}\right) < \infty.$$

Therefore, by the Strong Law of Large Numbers,

$$\begin{aligned}
& \overline{\lim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \sup_{\|\Delta\|_q \leq \delta\sqrt{n}} \{\int_0^1 \zeta^T \left[Dh\left(W_i + \Delta u/n^{1/2}\right) - Dh\left(W_i\right)\right]\Delta du - \varepsilon''\|\Delta\|_q^\rho\} \\
\leq & \left(\delta b\right)^{\rho/(\rho-1)} \left(\varepsilon''\right)^{-1/(\rho-1)} \left(1 - \frac{1}{\rho}\right) E\left(\bar{\kappa}\left(W_i\right)^{\rho/(\rho-1)}\right) \leq \delta,
\end{aligned}$$

assuming that $\delta > 0$ is chosen (given ε'' and b)

$$\delta^{1/(\rho-1)} \leq b^{-\rho/(\rho-1)} \left(\varepsilon''\right)^{1/(\rho-1)} \left(1 - \frac{1}{\rho}\right)^{-1} E\left(\bar{\kappa}\left(W_i\right)^{\rho/(\rho-1)}\right)^{-1}.$$

We then have that for any $\varepsilon'', \delta > 0$, there exists N_0 so that if $n \geq N_0$,

$$\begin{aligned} & \sup_{\Delta}\{\int_0^1 \zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right)\Delta du - \|\Delta\|_q^\rho\} \\ \leq & \sup_{\Delta \leq \delta n}\{\zeta^T Dh\left(W_i\right)\Delta du - (1-\varepsilon'')\|\Delta\|_q^\rho\} + \delta \\ = & \min(\kappa\left(\rho, \varepsilon'\right)\left\|\zeta^T Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)}, c_n), \end{aligned}$$

where

$$\kappa\left(\rho, \varepsilon''\right) = \left(\frac{1}{\rho(1-\varepsilon'')}\right)^{1/(\rho-1)}\left(1 - \frac{1}{\rho}\right),$$

and $c_n \to \infty$ as $n \to \infty$ (the exact value of c_n is not important).

Then, note that A5) implies that

$$\|Dh\left(W_i\right)\|_p^{\rho/(1-\rho)} I\left(\|W_i\| \geq 1\right) \leq \kappa I\left(\|W_i\| \geq 1\right)\|W_i\|_q^\rho \leq \kappa\|W_i\|_q^\rho$$

and, therefore, since $Dh\left(\cdot\right)$ is continuous (therefore locally bounded) and $E\|W_i\|_q^\rho < \infty$ also by A5), we have (because $\|w\|_p \leq d^{1/p}\|w\|_q$) that

$$E\|Dh\left(W_i\right)\|_q^{\rho/(1-\rho)} < \infty.$$

An argument similar to Lemma 3 shows that

$$\begin{aligned} & \sup_{\|\zeta\| \leq b}[\zeta^T H_n - \frac{1}{n}\sum_{i=1}^n \min(\kappa\left(\rho, \varepsilon'\right)\left\|\zeta^T Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)}, c_n)] \\ \Rightarrow & \sup_{\|\zeta\| \leq b}[\zeta^T H_n - \kappa\left(\rho, \varepsilon''\right) E\left\|\zeta^T Dh\left(W_i\right)\right\|_q^{\rho/(\rho-1)}] \end{aligned}$$

as $n \to \infty$ (where $\Rightarrow$ denotes weak convergence). Finally, we can send $\varepsilon'', \delta \to 0$ and $b \to \infty$ to obtain the desired lower bound. □

5.2. **Proofs of the distributionally robust representations in Section 3.2.** Here, we provide proofs for results in Section 3.2 that recover various norm regularized regressions as a special case of distributionally robust regression (Theorems 5, 2 and 3).

Proof of Theorem 5. We utilize the duality result in Proposition 4 to prove Proposition 5. For brevity, let $\bar{X}_i = (X_i, Y_i)$ and $\bar{\beta} = (-\beta, 1)$. Then the loss function becomes $l(X_i, Y_i; \beta) = (\bar{\beta}^T \bar{X}_i)^2$. We first decipher the function $\phi_\gamma(X_i, Y_i; \beta)$ defined in Proposition 4:

$$\phi_\gamma(X_i, Y_i; \beta) = \sup_{\bar{u} \in \mathbb{R}^{d+1}}\left\{(\bar{\beta}^T \bar{u})^2 - \gamma\|\bar{X}_i - \bar{u}\|_q^2.\right\}$$

To proceed further, we change the variable to $\Delta = \bar{u} - \bar{X}_i$, and apply Hölder's inequality to see that $|\bar{\beta}^T \Delta| \leq \|\bar{\beta}\|_p \|\Delta\|_q$, where the equality holds for some $\Delta \in \mathbb{R}^{d+1}$. Therefore,

$$\begin{aligned} \phi_\gamma(\bar{X}_i; \beta) &= \sup_{\Delta \in \mathbb{R}^{d+1}}\left\{\left(\bar{\beta}^T \bar{X}_i + \bar{\beta}^T \Delta\right)^2 - \gamma\|\Delta\|_q^2\right\} \\ &= \sup_{\Delta \in \mathbb{R}^{d+1}}\left\{\left(\bar{\beta}^T \bar{X}_i + \operatorname{sign}\left(\bar{\beta}^T \bar{X}_i\right)\left|\bar{\beta}^T \Delta\right|\right)^2 - \gamma\|\Delta\|_q^2\right\} \\ &= \sup_{\Delta \in \mathbb{R}^{d+1}}\left\{\left(\bar{\beta}^T \bar{X}_i + \operatorname{sign}\left(\bar{\beta}^T \bar{X}_i\right)\|\Delta\|_q\left\|\bar{\beta}\right\|_p\right)^2 - \gamma\|\Delta\|_q^2\right\}. \end{aligned}$$

On expanding the squares, the above expression simplifies as below:

$$\phi_\gamma(\bar{X}_i;\beta) = \left(\bar{\beta}^T\bar{X}_i\right)^2 + \sup_{\Delta\in\mathbb{R}^{d+1}} \left\{ -\left(\gamma - \left\|\bar{\beta}\right\|_p^2\right)\|\Delta\|_q^2 + 2\left|\bar{\beta}^T\bar{X}_i\right|\left\|\bar{\beta}\right\|_p\|\Delta\|_q \right\}$$

$$= \begin{cases} \left(\bar{\beta}^T\bar{X}_i\right)^2\gamma/(\gamma - \left\|\bar{\beta}\right\|_p^2) & \text{if } \gamma > \left\|\bar{\beta}\right\|_p^2, \\ +\infty & \text{if } \gamma \le \left\|\bar{\beta}\right\|_p^2. \end{cases} \tag{43}$$

With this expression for $\phi_\gamma(X_i, Y_i;\beta)$, we next investigate the right hand side of the duality relation in Proposition 4. As $\phi_\gamma(x,y;\beta) = \infty$ when $\gamma \le \|\beta\|_p^2$, we obtain from the dual formulation in Proposition 4 that

$$\sup_{P:\mathcal{D}_c(P,P_n)\le\delta} E_P\left[l(X,Y;\beta)\right] = \inf_{\gamma\ge 0}\left\{\gamma\delta + \frac{1}{n}\sum_{i=1}^n \phi_\gamma(X_i,Y_i;\beta)\right\}$$

$$= \inf_{\gamma>\|\beta\|_p^2}\left\{\gamma\delta + \frac{\gamma}{\gamma - \|\bar{\beta}\|_p^2}\frac{1}{n}\sum_{i=1}^n(\bar{\beta}^T\bar{X}_i)^2\right\}. \tag{44}$$

Now, see that $\sum_{i=1}^n(\bar{\beta}^T\bar{X}_i)^2/n$ is nothing but the mean square error $MSE_n(\beta)$. Next, as the right hand side of (44) is a convex function growing to ∞ (when $\gamma\to\infty$ or $\gamma\to\|\bar{\beta}\|_p^2$), its global minimizer can be characterized uniquely via first order optimality condition. This, in turn, renders the right hand side of (44) as

$$\sup_{P:\mathcal{D}_c(P,P_n)\le\delta} E_P\left[l(X,Y;\beta)\right] = \left(\sqrt{MSE_n(\beta)} + \sqrt{\delta}\|\bar{\beta}\|_p\right)^2.$$

This completes the proof of Proposition 5. □

Outline of a proof of Theorem 2. The proof of Theorem 2 is essentially the same as the proof of Proposition 5, except for adjusting for ∞ in the definition of cost function $N_q((x,y),(u,v))$ when $y\neq v$ (as in the derivation leading to $\phi_\gamma(X_i,Y_i;\beta)$ defined in (18)). First, see that

$$\phi_\gamma(X_i,Y_i;\beta) = \sup_{x'\in\mathbb{R}^d, y'\in\mathbb{R}}\left\{(y'^Tx'^2 - \gamma N_q\big((x',y'),(X_i,Y_i)\big)\right\}.$$

As $N_q((x',y'),(X_i,Y_i)) = \infty$ when $y'\neq Y_i$, the supremum in the above expression is effectively over only (x',y') such that $y' = Y_i$. As a result, we obtain,

$$\phi_\gamma(X_i,Y_i;\beta) = \sup_{x'\in\mathbb{R}^d}\left\{(Y_i - \beta^Tx'^2 - \gamma N_q\big((x',Y_i),(X_i,Y_i)\big)\right\}.$$

$$= \sup_{x'\in\mathbb{R}^d}\left\{(Y_i - \beta^Tx'^2 - \gamma\|x'-X_i\|_q^2\big)\right\}.$$

Now, following same lines of reasoning as in the proof of Theorem 5 and the derivation leading to (43), we obtain

$$\phi_\gamma(x,y;\beta) = \begin{cases} \frac{\gamma}{\gamma-\|\beta\|_p^2} & (Y_i - \beta^TX_i)^2 \qquad \text{when } \lambda > \|\beta\|_p^2, \\ & +\infty \qquad \text{otherwise.} \end{cases}$$

The rest of the proof is same as in the proof of Proposition 5.

Proof of Theorem 3. As in the proof of Proposition 5, we apply the duality formulation in Proposition 4 to write the worst case expected log-exponential loss function as:

$$\sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\big[l(X,Y;\beta)\big] = \inf_{\lambda\geq 0}\left\{\delta\lambda + \frac{1}{n}\sum_{i=1}^{n}\sup_x\left\{\log\left(1+\exp(-Y_i\beta^T x)\right) - \lambda\left\|x-X_i\right\|_p\right\}\right\}.$$

For each (X_i, Y_i), following Lemma 1 in [35], we obtain

$$\sup_x\left\{\log\left(1+\exp(-Y_i\beta^T x)\right) - \lambda\left\|x-X_i\right\|_p\right\} = \begin{cases} \log\left(1+\exp(-Y_i\beta^T X_i)\right) & \text{if } \left\|\beta\right\|_q \leq \lambda, \\ +\infty & \text{if } \left\|\beta\right\|_q > \lambda. \end{cases}$$

Then we can write the worst case expected loss function as,

$$\begin{aligned}
&\inf_{\lambda\geq 0}\left\{\delta\lambda + \frac{1}{n}\sum_{i=1}^{n}\sup_x\left\{\log\left(1+\exp(-Y_i\beta^T x)\right) - \lambda\left\|x-X_i\right\|_p\right\}\right\} \\
&= \inf_{\lambda\geq 0}\left\{\delta\lambda + \frac{1}{n}\sum_{i=1}^{n}\left(\log\left(1+\exp(-Y_i\beta^T X_i)\right)1_{\{\lambda>\|\beta\|_q\}} + \infty 1_{\{\lambda\leq\|\beta\|_q\}}\right)\right\} \\
&= \inf_{\lambda>\|\beta\|_q}\left\{\delta\lambda + \frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp(-Y_i\beta^T X_i)\right)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp(-Y_i\beta^T X_i)\right) + \delta\left\|\beta\right\|_q,
\end{aligned}$$

which is equivalent to regularized logistic regression in the theorem statement.

We proved for logistic regression while considering SVM with Hinge loss function, let us apply the duality formulation in Proposition 4 to write the worst case expected Hinge loss function as:

$$\sup_{P:\ D_c(P,P_n)\leq\delta} E_P\left[\left(1+Y\beta^T X\right)^+\right] = \inf_{\lambda\geq 0}\left\{\delta\lambda + \frac{1}{n}\sum_{i=1}^{n}\sup_x\left\{\left(1-Y_i\beta^T x\right)^+ - \lambda\left\|x-X_i\right\|_p\right\}\right\}.$$

For each i, let us consider the the maximization problem and for simplicity we denote $\Delta u_i = x - X_i$

$$\begin{aligned}
&\sup_{\Delta u_i}\left\{\left(1-Y_i\beta^T\left(X_i+\Delta u_i\right)\right)^+ - \lambda\left\|\Delta u_i\right\|_p\right\} \\
&= \sup_{\Delta u_i}\sup_{0\leq\alpha_i\leq 1}\left\{\alpha_i\left(1-Y_i\beta^T\left(X_i+\Delta u_i\right)\right) - \lambda\left\|\Delta u_i\right\|_p\right\} \\
&= \sup_{0\leq\alpha_i\leq 1}\sup_{\Delta u_i}\left\{\alpha_i Y_i\beta^T\Delta u_i - \lambda\left\|\Delta u_i\right\|_p + \alpha_i\left(1-Y_i\beta^T X_i\right)\right\} \\
&= \sup_{0\leq\alpha_i\leq 1}\sup_{\Delta u_i}\left\{\alpha_i\left\|\beta\right\|_q\left\|\Delta u_i\right\|_p - \lambda\left\|\Delta u_i\right\|_p + \alpha_i\left(1-Y_i\beta^T X_i\right)\right\} \\
&= \begin{cases} \left(1-Y_i\beta^T X_i\right)^+ & \text{if } \left\|\beta\right\|_q \leq \lambda \\ +\infty & \text{if } \left\|\beta\right\|_q > \lambda \end{cases}
\end{aligned}$$

The first equality is due to $x^+ = \sup_{0\leq\alpha\leq 1} x$; second equality is because the function is concave in Δu_i linear in α and α is in a compact set, we can apply minimax theorem to switch the order of maximals; third equality is due to applying Holder inequality to the first term and since the

second term only depends on the norm of Δu_i we can argue the equality also holds for this maximization problem. We notice the objective function is a minimization problem thus we will require $\lambda \geq \|\beta\|_q$. Then we have

$$\inf_{\lambda \geq \|\beta\|_q} \left\{ \delta\lambda + \frac{1}{n}\sum_{i=1}^{n} \left(1 - Y_i\beta^T X_i\right)^+ \right\} = \frac{1}{n}\sum_{i=1}^{n} \left(1 - Y_i\beta^T X_i\right)^+ + \delta \|\beta\|_q .$$

□

5.3. **Proofs of RWP function limit theorems for linear and logistic regression examples.** We first obtain the dual formulation of the respective RWP functions for linear and logistic regressions using Proposition 1. Let $E[h(x, y; \beta)] = \mathbf{0}$ be the estimating equation under consideration ($h(x, y; \beta) = (y - \beta^T x)x$ for linear regression and $h(x, y; \beta)$ as in (25) for logistic regression). Recall that the cost function is $c(\cdot) = N_q(\cdot)$. Due to the duality result in Proposition 1, we obtain

$$\begin{aligned} R_n(\beta_*) &= \inf \left\{ D_c(P, P_n) : \ E_P[h(X, Y; \beta_*)] = \mathbf{0} \right\} \\ &= \sup_{\lambda} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \sup_{(x', y')} \left\{ \lambda^T h(x', y'; \beta_*) - N_q\big((x', y'), (X_i, Y_i)\big) \right\} \right\}. \end{aligned}$$

As $N_q((x', y'), (X_i, Y_i)) = \infty$ when $y' \neq Y_i$, the above expression simplifies to,

$$R_n(\beta_*) = \sup_{\lambda} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \sup_{x'} \left\{ \lambda^T h(x', Y_i; \beta_*) - \|x' - X_i\|_q^{\rho} \right\} \right\}, \tag{45}$$

where $\rho = 2$ for the case of linear regression (Theorem 4) and $\rho = 1$ for the case of logistic regression (Theorem 5). As RWP function here is similar to the RWP function for general estimating equation in Section 2.4, a similar limit theorem holds. We state here the precise assumptions required to prove RWP limit theorems for the dual formulation in (45).

Assumptions:

A2') Suppose that $\beta_* \in \mathbb{R}^d$ satisfies $E[h(X, Y; \beta_*)] = \mathbf{0}$ and $E\|h(X, Y; \beta_*)\|_2^2 < \infty$ (While we do not assume that β_* is unique, the results are stated for a fixed β_* satisfying $E[h(X, Y; \beta_*)] = \mathbf{0}$.)

A4') Suppose that for each $\xi \neq \mathbf{0}$, the partial derivative $D_x h(x, y; \beta_*)$ satisfies,

$$P\left(\left\| \xi^T D_x h(X, Y; \beta_*) \right\|_p > 0 \right) > 0.$$

A6') Assume that there exists $\bar{\kappa} : \mathbb{R}^m \to \infty$ such that

$$\|D_x h(x + \Delta, y; \beta_*) - D_x h(x, y; \beta_*)\|_p \leq \bar{\kappa}(x, y)\|\Delta\|_q,$$

for all $\Delta \in \mathbb{R}^d$, and $E[\bar{\kappa}(X, Y)^2] < \infty$.

Lemma 4. *If $\rho \geq 2$, under Assumptions A2'), A4') and A6'), we have,*

$$nR_n(\beta_*; \rho) \Rightarrow \bar{R}(\rho),$$

where

$$\bar{R}(\rho) = \sup_{\xi \in \mathbb{R}^d} \left\{ \rho\xi^T H - (\rho - 1)E \left\| \xi^T D_x h(X, Y; \beta_*) \right\|_p^{\rho/(\rho-1)} \right\},$$

with $H \sim \mathcal{N}(\mathbf{0}, Cov[h(X, Y; \beta_*])$ and $1/p + 1/q = 1$.

Lemma 5. *If $\rho = 1$, in addition to assuming A2'), A4'), suppose that $D_x h(\cdot, y; \beta_*)$ is continuous for every y in the support of probability distribution of Y. Also suppose that X has a positive probability density (almost everywhere) with respect to the Lebesgue measure. Then,*

$$nR_n(\beta_*; 1) \Rightarrow \bar{R}(1),$$

where

$$\bar{R}(1) = \sup_{\xi: P(\|\xi^T D_x h(X,Y;\beta_*)\|_p > 1) = 0} \{\xi^T H\},$$

with $H \sim \mathcal{N}(\mathbf{0}, Cov[h(X, Y; \beta_*])$.

The proof of Lemma 4 and 5 follows closely the proof of our results in Section 2 and therefore it is omitted. We prove Theorem and 4 and 5 as a quick application of these lemmas.

Proof of Theorem 4. To show that the RWP function dual formulation in (45) converges in distribution, we verify the assumptions of Lemma 4 with $h(x, y; \beta) = (y - \beta^T x)x$. Under the null hypothesis H_0, $Y - \beta_*^T X = e$ is independent of X, has zero mean and finite variance σ^2. Therefore,

$$\begin{aligned} E\left[h(X, Y; \beta)\right] &= E\left[eX\right] = 0, \text{ and} \\ E\|h(X, Y; \beta)\|_2^2 &= E\left[e^2 X^T X\right] = \sigma^2 E\|X\|_2^2, \end{aligned}$$

which is finite, because trace of the covariance matrix Σ is finite. This verifies Assumption A2'). Further,

$$D_x h(X, Y; \beta_*) = \left(y - \beta_*^T X\right) I_d - X\beta_*^T = eI_d - X\beta_*^T,$$

where I_d is the $d \times d$ identity matrix. For any $\xi \neq \mathbf{0}$,

$$P\left(\|\xi^T D_x h(X, Y; \beta_*)\|_p = 0\right) = P\left(e\xi = (\xi^T X)\beta\right) = 0,$$

thus satisfying Assumption A4') trivially. In addition,

$$\|D_x h(x + \Delta, y; \beta_*) - D_x h(x, y; \beta_*)\|_p = \left\|\beta_*^T \Delta I_d - \Delta \beta_*^T\right\|_p \leq c\|\Delta\|_q,$$

for some positive constant c. This verifies Assumption A6'). As all the assumptions imposed in Lemma 4 are easily satisfied, using $\rho = 2$, we obtain the following convergence in distribution as a consequence of Lemma 4.

$$R_n(\beta_*) \Rightarrow \sup_{\xi \in \mathbb{R}^d} \left\{2\xi^T H - E\left\|e\xi - (\xi^T X)\beta_*\right\|_p^2\right\},$$

as $n \to \infty$. Here, $H \sim \mathcal{N}(\mathbf{0}, \text{Cov}[h(X, Y; \beta_*)]$. As $\text{Cov}[h(X, Y; \beta_*)] = E\left[e^2 XX^T\right] = \sigma^2\Sigma$, if we let $Z = H/\sigma$, we obtain the limit law,

$$L_1 = \sup_{\xi \in \mathbb{R}^d} \left\{2\sigma\xi^T Z - E\left\|e\xi - (\xi^T X)\beta_*\right\|_p^2\right\},$$

where $Z = \mathcal{N}(\mathbf{0}, \Sigma)$, as in the statement of the theorem.

Proof of the stochastic upper bound in Theorem 4: For the stochastic upper bound, let us consider the asymptotic distribution L_1 and rewrite the maximization problem as,

$$\begin{aligned} L_1 &= \sup_{\|\xi\|_p=1} \sup_{\alpha\geq 0} \left\{ 2\sigma\alpha\xi^T Z - \alpha^2 E \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2 \right\} \\ &\leq \sup_{\|\xi\|_p=1} \sup_{\alpha\geq 0} \left\{ 2\sigma\alpha \left\| Z \right\|_q - \alpha^2 E \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2 \right\}, \end{aligned}$$

because of Hölder's inequality. By solving the inner optimization problem in α, we obtain

$$L_1 \leq \sup_{\|\xi\|_p=1} \frac{\sigma^2 \left\| Z \right\|_q^2}{E \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2} = \frac{\sigma^2 \left\| Z \right\|_q^2}{\inf_{\|\xi\|_p=1} E \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2}. \tag{46}$$

Next, consider the minimization problem in the denominator: Due to triangle inequality,

$$\begin{aligned} \inf_{\|\xi\|_p=1} E \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2 &\geq \inf_{\|\xi\|_p=1} E \left(|e| \left\| \xi \right\|_p - \left| \xi^T X \right| \left\| \beta_* \right\|_p \right)^2 \\ &= E\left|e\right|^2 + \inf_{\|\xi\|_p=1} \left\{ \left\| \beta_* \right\|_p^2 E \left| \xi^T X \right|^2 - 2 \left\| \beta_* \right\|_p E\left|e\right| E \left| \xi^T X \right| \right\} \\ &\geq E\left|e\right|^2 + \inf_{\|\xi\|_p=1} \left\{ \left\| \beta_* \right\|_p^2 \left(E \left| \xi^T X \right| \right)^2 - 2 \left\| \beta_* \right\|_p E\left|e\right| E \left| \xi^T X \right| \right\} \\ &= E\left|e\right|^2 - \left(E\left|e\right|\right)^2 + \inf_{\|\xi\|_p=1} \left(\left\| \beta_* \right\|_p E \left| \xi^T X \right| - E\left|e\right| \right)^2 \\ &\geq E\left|e\right|^2 - \left(E\left|e\right|\right)^2 = \mathrm{Var}\left[|e|\right]. \end{aligned}$$

Combining the above inequality with (46), we obtain,

$$\sup_{\xi\in\mathbb{R}^d} \left\{ \sigma^2 \xi^T Z - E \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2 \right\} \leq \frac{\sigma^2 \left\| Z \right\|_q^2}{\mathrm{Var}\left|e\right|}$$

Consequently,

$$nR_n(\beta_*) \xrightarrow{D} L_1 := \max_{\xi\in\mathbb{R}^d} \left\{ \sigma\xi^T Z - E \left\| e\xi - (\xi^T X)\beta_* \right\|_p^2 \right\} \overset{D}{\leq} \frac{E[e^2]}{E[e^2] - \left(E\left|e\right|\right)^2} \|Z\|_q^2.$$

If random error e is normally distributed, then

$$nR_n(\beta_*) \lesssim_D \frac{\pi}{\pi - 2} \|Z\|_q^2,$$

thus establishing the desired upper bound. □

Proof of Theorem 5. Under null hypothesis H_0, the training samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are produced from the logistic regression model with parameter β_*. As β_* minimizes the expected log-exponential loss $l(x, y; \beta)$, the corresponding optimality condition is $E[h(X, Y; \beta_*)] = \mathbf{0}$, where

$$h(x, y; \beta_*) = \frac{-yx}{1 + \exp(y\beta_* x)}.$$

As $E\|h(X, Y; \beta_*)\|_2^2 \leq E\|X\|_2^2$ is finite, Assumption A2') is satisfied. Let I_d denote $d \times d$ identity matrix. While

$$D_x h(x, y; \beta_*) = \frac{-yI_d}{1 + \exp(y\beta_*^T x)} + \frac{x\beta_*^T}{(1 + \exp(y\beta_*^T x))(1 + \exp(-y\beta_*^T x))}.$$

is continuous (as a function of x) for every y, it is also true that

$$P\left(\left\|\xi^T D_x h(X,Y;\beta_*)\right\|_p = 0\right) = P\left(Y\left(1+\exp(-Y\beta_*^T X)\right)\xi = (\xi^T X)\beta\right) = 0,$$

for any $\xi \neq \mathbf{0}$, thus satisfying Assumption A4'). As all the conditions required for the convergence in distribution in Lemma 5 are satisfied, we obtain,

$$\sqrt{n}R_n(\beta_*) \Rightarrow \sup_{\xi\in A} \xi^T Z,$$

where $Z \sim \mathcal{N}(\mathbf{0}, E[XX^T/(1+\exp(Y\beta_*^T X))^2])$ as a consequence of Lemma 5. Here, the set $A = \{\xi \in \mathbb{R}^d : \text{ess sup}\|\xi^T D_x h(X,Y;\beta_*)\| \leq 1\}$.

Proof of the stochastic upper bound in Theorem 5: First, we claim that A is a subset of the norm ball $\{\xi \in \mathbb{R}^d : \|\xi\|_p \leq 1\}$. To establish this, we observe that,

$$\begin{aligned}\left\|\xi^T D_x h(X,Y;\beta_*)\right\|_p &\geq \left\|\frac{-Y\xi}{1+\exp(Y\beta_*^T X)}\right\|_p - \left\|\frac{(\xi^T X)\beta_*}{\left(1+\exp(Y\beta_*^T X)\right)\left(1+\exp(Y\beta_*^T X)\right)}\right\|_p \\ &\geq \left(\frac{1}{1+\exp(Y\beta_*^T X)} - \frac{\|X\|_q\|\beta_*\|_p}{\left(1+\exp(Y\beta_*^T X)\right)\left(1+\exp(-Y\beta_*^T X)\right)}\right)\|\xi\|_p, \qquad (47)\end{aligned}$$

because $Y \in \{+1,-1\}$, and due to Hölder's inequality $|\xi^T X| \leq \|\xi\|_p\|X\|_q$. If $\xi \in \mathbb{R}^d$ is such that $\|\xi\|_p = (1-\epsilon)^{-2} > 1$ for a given $\epsilon > 0$, then following (47), $\|\xi^T D_x h(X,Y)\| > 1$, whenever

$$(X,Y) \in \Omega_\epsilon := \left\{(x,y) : \frac{\|x\|_q\|\beta_*\|_p}{1+\exp(-y\beta_*^T x)} \leq \frac{\epsilon}{2},\ \frac{1}{1+\exp(y\beta_*^T x)} \geq 1-\frac{\epsilon}{2}\right\}.$$

Since X has positive density almost everywhere, the set Ω_ϵ has positive probability for every $\epsilon > 0$. Thus, if $\|\xi\|_p > 1$, $\|\xi^T D_x h(X,Y;\beta_*)\| > 1$ with positive probability. Therefore, A is a subset of $\{\xi : \|\xi\|_p \leq 1\}$. Consequently,

$$L_3 := \sup_{\xi\in A} \xi^T Z \overset{D}{\leq} \sup_{\xi:\|\xi\|_p\leq 1} \xi^T Z = \|Z\|_q.$$

If we let $\tilde{Z} \sim \mathcal{N}(\mathbf{0}, E[XX^T])$, then $\text{Cov}[\tilde{Z}] - \text{Cov}[Z]$ is positive definite. As a result, L_3 is stochastically dominated by $L_4 := \|\tilde{Z}\|_q$, thus verifying the desired stochastic upper bound in the statement of Theorem 5. □

Proof of Theorem 6. Instead of characterizing the exact weak limit, we will find a stochastic upper bound for $R_n(\beta_*)$. The RWP function, as in the proof of Theorem 4, admits the following dual representation (see (45)):

$$\begin{aligned}R_n(\beta_*) &= \sup_\lambda \left\{-\frac{1}{n}\sum_{i=1}^n \sup_{x'}\left\{\lambda^T(Y_i - \beta_*^T x')x' - \|x'-X_i\|_\infty^2\right\}\right\} \\ &= \sup_\lambda \left\{-\lambda^T\frac{Z_n}{\sqrt{n}} - \frac{1}{n}\sum_{i=1}^n \sup_\Delta\left\{e_i\lambda^T\Delta - (\beta_*^T\Delta)(\lambda^T X_i) - \left(\|\Delta\|_\infty^2 + (\beta_*^T\Delta)(\lambda^T\Delta)\right)\right\}\right\},\end{aligned}$$

where $Z_n = n^{-1/2}\sum_{i=1}^n e_i X_i$, $e_i = Y_i - \beta_*^T X_i$. In addition, we have changed the variable from $x' - X_i = \Delta$. If we let $\zeta = \sqrt{n}\lambda$, then

$$nR_n(\beta_*) = \sup_{\zeta}\left\{-\zeta^T Z_n - \frac{1}{\sqrt{n}}\sum_{i=1}^n \sup_{\Delta}\left\{e_i\zeta^T\Delta - (\beta_*^T\Delta)(\zeta^T X_i) - \left(\sqrt{n}\|\Delta\|_\infty^2 + (\beta_*^T\Delta)(\zeta^T\Delta)\right)\right\}\right\}$$
$$\leq \sup_{\zeta}\left\{-\zeta^T Z_n - \frac{1}{\sqrt{n}}\sum_{i=1}^n \sup_{\|\Delta\|_\infty}\left\{\left\|e_i\zeta^T - (\zeta^T X_i)\beta_*^T\right\|_1\|\Delta\|_\infty - \sqrt{n}\left(1 - \frac{\|\beta_*\|_1\|\zeta\|_1}{\sqrt{n}}\right)\|\Delta\|_\infty^2\right\}\right\},$$

where we have used Hölder's inequality thrice to obtain the upper bound. If we solve the inner supremum over the variable $\|\Delta\|$, we obtain,

$$nR_n(\beta_*) \leq \sup_{\zeta}\left\{-\zeta^T Z_n - \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\left\|e_i\zeta - (\zeta^T X_i)\beta_*\right\|_1^2}{4\sqrt{n}\left(1 - \|\beta_*\|_1\|\zeta\|_1 n^{-1/2}\right)}\right\}$$
$$\leq \sup_{a\geq 0}\sup_{\zeta:\|\zeta\|_1=1}\left\{-a\zeta^T Z_n - \frac{a^2}{4\left(1 - a\|\beta_*\|_1 n^{-1/2}\right)}\frac{1}{n}\sum_{i=1}^n\left\|e_i\zeta - (\zeta^T X_i)\beta_*\right\|_1^2\right\},$$

where we have split the optimization into two parts: one over the magnitude (denoted by a), and another over all unit vectors ζ. Further, due to Hölder's inequality, we have $|\zeta^T Z_n| \leq \|Z_n\|_\infty$ as $\|\zeta\|_1 = 1$. Therefore,

$$nR_n(\beta_*) \leq \sup_{a\geq 0}\left\{c_1(n)a - \frac{a^2}{\left(1 - c_2(n)a\right)}c_3(n)\right\},$$

where we have let

$$c_1(n) = \|Z_n\|_\infty^2, \quad c_2(n) = \|\beta_*\|_1 n^{-1/2}, \quad \text{and} \quad c_3(n) = \inf_{\zeta:\|\zeta\|_1=1}\frac{1}{4n}\sum_{i=1}^n\left\|e_i\zeta - (\zeta^T X_i)\beta_*\right\|_1^2.$$

As $\|\beta_*\|_1 n^{-1/2} \to 0$ when $n \to \infty$, we have $c_2(n) \to 0$. Therefore, the above supremum over a is attained at $a = c_1(n)/2c_3(n) + o(1)$ when $n \to \infty$. Consequently,

$$nR_n(\beta_*) \leq \frac{\|Z_n\|_\infty^2}{\inf_{\{\zeta:\|\zeta\|_1=1\}}\frac{1}{n}\sum_{i=1}^n\|e_i\zeta - (\zeta^T X_i)\beta_*\|_1^2} + o(1). \tag{48}$$

The infimum in the denominator can be lower bounded as in the proof of Theorem 4. In particular, due to triangle inequality,

$$\frac{1}{n}\sum_{i=1}^n\left\|e_i\zeta - (\zeta^T X_i)\beta_*\right\|_1^2 \geq \frac{1}{n}\sum_{i=1}^n\left(|e_i|\,\|\zeta\|_1 - \left|\zeta^T X_i\right|\|\beta_*\|_1\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^n|e_i|^2 + \|\beta_*\|_1^2\frac{1}{n}\sum_{i=1}^n\left|\zeta^T X_i\right|^2 - 2\|\beta_*\|_1\frac{1}{n}\sum_{i=1}^n|e_i|\left|\zeta^T X_i\right|$$
$$= \frac{1}{n}\sum_{i=1}^n|e_i|^2 + \|\beta_*\|_1^2\frac{1}{n}\sum_{i=1}^n\left|\zeta^T X_i\right|^2 - 2\|\beta_*\|_1 E|e_i|\frac{1}{n}\sum_{i=1}^n\left|\zeta^T X_i\right| - \epsilon_n(\zeta),$$

where $\epsilon_n(\zeta) = 2\|\beta_*\|_1\frac{1}{n}\sum_{j=1}^n(|e_i| - E|e_i|)|\zeta^T X_i|$. If we let $\tilde{e}_i = |e_i| - E|e_i|$, then $E[\tilde{e}_i] = 0$ and $\mathrm{Var}[\tilde{e}_i] \leq \mathrm{Var}[e_i]$. As $\tilde{e}_i$ is independent of X_i, $E[\tilde{e}_i|\zeta^T X_i|] = 0$. In addition,

$$\mathrm{Var}\left[\tilde{e}_i|\zeta^T X_i|\right] = \mathrm{Var}[\tilde{e}_i]\zeta^T\Sigma\zeta \leq \mathrm{Var}[e_i]\zeta^T\Sigma\zeta,$$

where we recall that $\Sigma = \mathrm{Cov}[X]$. With the assumption that on the largest eigen value of Σ, denoted by $\lambda_{\max}(\Sigma)$, is $o(nC(n,d)^2)$, we have

$$\sup_{\|\zeta\|_1=1} \zeta^T \Sigma \zeta \leq \sup_{|\zeta|_1=1} \lambda_{\max}(\Sigma) \|\zeta\|_2^2 \leq \lambda_{\max}(\Sigma) = o(nC(n,d)^2).$$

Consequently, the variance of $\frac{1}{n}\sum_{j=1}^n \tilde{e}_i \left|\zeta^T X_i\right|$ is of order $o\left(C(n,d)^2\right)$ uniformly in ζ for $\|\zeta\|_1 = 1$. Combining this with the assumption that $\|\beta_*\|_1 = o(1/C(n,d))$ we have

$$\epsilon_n(\zeta) = 2\|\beta_*\|_1 \frac{1}{n}\sum_{j=1}^n (|e_i| - E|e_i|)\left|\zeta^T X_i\right| = o_p(1).$$

Since the bound is uniformly in ζ such that $\|\zeta\|_1 = 1$, we have for sufficiently large n,

$$\begin{aligned}
&\inf_{\zeta:\|\zeta\|_1=1} \frac{1}{n}\sum_{i=1}^n \left\|e_i\zeta - \zeta^T X_i \beta_*\right\|_1^2 \\
&\geq \frac{1}{n}\sum_{i=1}^n |e_i|^2 + \inf_{\zeta:\|\zeta\|_1=1} \left\{ \|\beta_*\|_1^2 \frac{1}{n}\sum_{i=1}^n \left|\zeta^T X_i\right|^2 - 2\|\beta_*\|_1 E|e_i| \frac{1}{n}\sum_{i=1}^n \left|\zeta^T X_i\right| \right\} + o_p(1) \\
&\geq \frac{1}{n}\sum_{i=1}^n |e_i|^2 - \left(E|e_i|\right)^2 + \inf_{\zeta:\|\zeta\|_1=1} \left(\|\beta_*\|_1 \frac{1}{n}\sum_{i=1}^n \left|\zeta^T X_i\right| - E|e_i| \right)^2 + o_p(1) \\
&\geq \mathrm{Var}\,|e_i| + o_p(1).
\end{aligned}$$

Then, as $n \to \infty$, it follows from (48),

$$nR_n(\beta_*) \leq \frac{\|Z_n\|_\infty^2}{\mathrm{Var}\,|e|} + o_p(1).$$

The second claim is a direct consequence of Corollary 2.1 in [10] when X has sub-Gaussian tails. Finally, the last claim is the special example of computing the $(1-\alpha)$ quantile of $\|Z\|_\infty$ for $Z \sim \mathcal{N}(0, I_d)$. Here, the distribution of maximum of d i.i.d. standard normal random variables have $\Phi^{-1}(1-\alpha/2d)$ as its $(1-\alpha)$ quantile, and $E[e^2]/(E[e^2] - (E|e|)^2) = \pi/(\pi-2)$ when the additive error e is normally distributed. □

6. Numerical Examples

In this section, we consider two examples that demonstrate the numerical performance of the generalized LASSO algorithm (see Example 2) when the regularization parameter λ is selected as described in Section 4.1 using a suitable quantile of the RWPI limiting distribution.

Example 4. Consider the linear model $Y = 3X_1 + 2X_2 + 1.5X_4 + e$ where the vector of predictor variables $X = (X_1, \ldots, X_d)$ is distributed according to the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma_{k,j} = 0.5^{|k-j|}$ and additive error e is normally distributed with mean 0 and standard deviation $\sigma = 10$. Letting n denote the number of training samples, we illustrate the effectiveness of the RWPI based generalized LASSO procedure for various values of d and n by computing the mean square loss / error (MSE) over a simulated test data set of size $N = 10000$. Specifically, we take the number of predictors to be $d = 300$ and 600, the number of standardized i.i.d. training samples to range from $n = 350, 700, 3500, 10000$, and the desired confidence level to be 95%, that is, $1-\alpha = 0.95$. In each instance, we run the generalized LASSO algorithm using the 'flare' package proposed in [19] (available as a library in R) with regularization parameter λ chosen as prescribed in Section 4.1.

Repeating each experiment 100 times, we report the average training and test MSE in Tables 1 and 2, along with the respective results for ordinary least squares regression (OLS) and generalized LASSO algorithm with regularization parameter chosen as prescribed by cross-validation (denoted as G-LASSO CV in the tables.) We also report the average ℓ_1 and ℓ_2 error of the regression coefficients in Tables 1 and 2. In addition, we report the empirical coverage probability that the optimal error $E[(Y - \beta_*^T X)^2] = \sigma^2 = 100$ is smaller than the worst case expected loss computed by the DRO formulation (15). As this empirical coverage probability reported in Table 3 is closer to the desired confidence $1 - \alpha = 0.95$, the worst case expected loss computed by (15) can be seen as a tight upper bound of the optimal loss $E[l(X, Y; \beta_*)]$ (thus controlling generalization) with probability at least $1 - \alpha = 0.95$.

Example 5. Consider the diabetes data set from the 'lars' package in R (see [11]), where there are 64 predictors (including 10 baseline variables and other 54 possible interactions) and 1 response. After standardizing the variables, we split the entire data set of 442 observations into $n = 142$ training samples (chosen uniformly at random) and the remaining $N = 300$ samples as test data for each experiment, in order to compute training and test mean square errors using the generalized LASSO algorithm with regularization parameter picked as in Section 4.1. After repeating the experiment 100 times, we report the average training and test errors in Table 4, and compare the performance of RWPI based regularization parameter selection with other standard procedures such as OLS and generalized LASSO algorithm with regularization parameter chosen according to cross-validation.

Training data size, n	Method	Training Error	Test Error	ℓ_1 loss $\|\beta - \beta_*\|_1$	ℓ_2 loss $\|\beta - \beta_*\|_2$
350	RWPI	101.16(±8.11)	122.59(±6.64)	4.08(±0.69)	5.23(±0.76)
	G-LASSO CV	92.23(±7.91)	117.25(±6.07)	3.91(±0.42)	5.02(±1.28)
	OLS	13.95(±2.63)	702.73(±188.05)	31.59(±3.64)	436.19(±50.55)
700	RWPI	101.81(±3.01)	117.96(±4.80)	3.31(±0.40)	4.38(±0.48)
	G-LASSO CV	99.66(±4.64)	115.46(±4.36)	2.96(±0.37)	3.98(±0.66)
	OLS	56.82(±3.94)	178.44(±21.74)	10.99(±0.57)	152.04(±8.25)
3500	RWPI	102.55(±2.39)	108.44(±2.54)	2.18(±0.16)	3.28(±1.66)
	G-LASSO CV	100.74(±2.35)	113.83(±2.33)	2.66(±0.14)	3.91(±2.18)
	OLS	90.37(±2.17)	114.78(±5.50)	3.96(±0.20)	54.67(±3.09)
10000	RWPI	102.12(±8.11)	105.97(±0.88)	1.13(±0.08)	1.63(±0.11)
	G-LASSO CV	100.69(±7.91)	112.82(±0.71)	1.15(±0.07)	1.94(±0.12)
	OLS	95.91(±1.11)	107.74(±2.96)	2.23(±0.10)	30.91(±1.43)

TABLE 1. Sparse linear regression for $d = 300$ predictor variables in Example 4. The training and test mean square errors of RWPI based generalized LASSO regularization parameter selection is compared with ordinary least squares estimator (written as OLS) and cross-validation based generalized LASSO estimator (written as G-LASSO CV)

ACKNOWLEDGEMENT

Support from NSF through Award 1436700 and the Norges Bank is gratefully acknowledged.

Training data size, n	Method	Training Error	Test Error	ℓ_1 loss $\|\beta-\beta_*\|_1$	ℓ_2 loss $\|\beta-\beta_*\|_2$
350	RWPI	108.05(±8.38)	109.46(±4.68)	4.02(±0.71)	4.08(±0.70)
	G-LASSO CV	93.17(±10.83)	104.51(±4.76)	2.23(±0.38)	6.89(±2.35)
	OLS	–	–	–	–
700	RWPI	104.33(±5.03)	103.18(±2.14)	2.91(±0.42)	2.99(±0.43)
	G-LASSO CV	100.50(±4.70)	99.92(±2.18)	1.45(±0.28)	2.82(±0.64)
	OLS	14.27(±2.02)	699.06(±137.45)	31.66(±2.21)	518.02(±44.87)
3500	RWPI	101.52(±2.52)	96.38(±0.80)	1.23(±0.24)	1.32(±0.24)
	G-LASSO CV	102.58(±2.49)	98.55(±0.94)	1.18(±0.15)	1.94(±0.24)
	OLS	82.22(±2.31)	102.01(±6.14)	6.76(±0.23)	114.05(±5.73)
10000	RWPI	101.36(±1.11)	94.86(±0.36)	0.75(±0.13)	0.81(±0.14)
	G-LASSO CV	103.00(±1.11)	98.55(±0.49)	1.16(±0.08)	1.94(±0.13)
	OLS	95.11(±1.10)	99.53(±4.83)	3.26(±0.11)	63.67(±2.16)

TABLE 2. Sparse linear regression for $d = 600$ predictor variables in Example 4. The training and test mean square errors of RWPI based generalized LASSO regularization parameter selection is compared with ordinary least squares estimator (written as OLS) and cross-validation based generalized LASSO estimator (written as G-LASSO CV). As $n < d$ when $n = 350$, OLS estimation is not applicable in that case (denoted by a blank)

No. of predictors	Training sample size			
d	350	700	3500	10000
300	0.974	0.977	0.975	0.969
600	0.963	0.966	0.970	0.968

TABLE 3. Coverage Probability of empirical worst case expected loss in Example 4

	Training Error	Testing Error
RWPI	0.58(±0.05)	0.60(±0.04)
G-LASSO CV	0.44(±0.06)	0.57(±0.03)
OLS	0.26(±0.05)	1.38(±0.68)

TABLE 4. Linear Regression for Diabetes data in Example 5 with 142 training samples and 300 test samples. The training and test mean square errors of RWPI based generalized LASSO regularization parameter selection is compared with ordinary least squares estimator (written as OLS) and cross-validation based generalized LASSO estimator (written as G-LASSO CV).

REFERENCES

[1] Pierre Alquier. LASSO, iterative feature selection and the correlation selector: Oracle inequalities and numerical performances. *Electronic Journal of Statistics*, 2:1129–1152, 2008.

[2] Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization. In *Advances in Neural Information Processing Systems 27*, pages 1556–1564. 2014.

[3] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root LASSO: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[4] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[5] Jose Blanchet and Yang Kang. Sample-out-of-sample inference based on Wasserstein distance. *arXiv preprint arXiv: 1605.01340, 2016.*

[6] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *arXiv preprint arXiv:1604.01446*, 2016.
[7] Francesco Bravo. Empirical likelihood based inference with applications to some econometric models. *Econometric Theory*, 20(02):231–264, 2004.
[8] Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 12 2007.
[9] Song Xi Chen and Peter Hall. Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 21(3):1166–1181, September 1993.
[10] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 12 2013.
[11] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
[12] Peyman Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.
[13] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.
[14] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems 28*, pages 2053–2061. 2015.
[15] Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199v1*, 2016.
[16] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.
[17] Nils Lid Hjort, Ian McKeague, and Ingrid Van Keilegom. Extending the scope of empirical likelihood. *The Annals of Statistics*, pages 1079–1111, 2009.
[18] Keiiti Isii. On sharpness of Tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14(1):185–197, 1962.
[19] Xingguo Li, Tuo Zhao, Xiaoming Yuan, and Han Liu. The flare package for high dimensional linear regression and precision matrix estimation in R. *The Journal of Machine Learning Research*, 16(1):553–557, 2015.
[20] Sahand Negahban, Pradeep Ravikumar, Martin Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-Estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.
[21] Whitney Newey and Richard Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.
[22] XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.
[23] Art Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.
[24] Art Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, pages 90–120, 1990.
[25] Art Owen. Empirical likelihood for linear models. *The Annals of Statistics*, pages 1725–1747, 1991.
[26] Art Owen. *Empirical likelihood*. CRC press, 2001.
[27] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized LASSO: A precise analysis. In *In proceedings of the 51st Annual Allerton Conference of Communication, Control, and Computing*, pages 1002–1009, October 2013.
[28] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *ICML*, 2016.
[29] Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, pages 300–325, 1994.
[30] Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems. Volume I: Theory*. Springer Science & Business Media, 1998.
[31] Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems. Volume II: Applications*. Springer Science & Business Media, 1998.
[32] Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 630–638, 2016.
[33] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[34] Vivien Seguy and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems 28*, pages 3312–3320. 2015.
[35] Soroosh Shafieezadeh-Abadeh, Peyman Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584. 2015.
[36] James Smith. Generalized Chebychev inequalities: Theory and applications in decision analysis. *Operations Research*, 43(5):807–825, 1995.
[37] Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, July 2015.
[38] Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Earth mover's distances on discrete surfaces. *ACM Trans. Graph.*, 33(4):67:1–67:12, July 2014.
[39] Sanvesh Srivastava, Volkan Cevher, Quoc Tran-Dinh, and David B Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *AISTATS*, 2015.
[40] Frode Terkelsen. Some minimax theorems. *Mathematica Scandinavica*, 31(2):405–413, 1973.
[41] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
[42] Cédric Villani. *Optimal transport: old and new*. Springer Science & Business Media, 2008.
[43] Changbao Wu. Weighted empirical likelihood inference. *Statistics & Probability Letters*, 66(1):67–79, January 2004.
[44] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and LASSO. In *Advances in Neural Information Processing Systems 21*, pages 1801–1808. 2009.
[45] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *J. Mach. Learn. Res.*, 10:1485–1510, December 2009.
[46] Mai Zhou. *Empirical likelihood method in survival analysis*, volume 79. CRC Press, 2015.

Appendix

Appendix A. Strong duality of the linear semi-infinite programs in the paper

In the main body of the paper, we have utilized strong duality of linear semi-infinite programs in two contexts: 1) to derive a dual representation of the RWP function in order to perform asymptotic analysis (see Proposition 1), and 2) to derive distributional robust representations (see Proposition 4). Establishing these strong dualities rely on the following well-known result on problem of moments ([18, 21]).

The problem of moments. Let Ω be a nonempty Borel measurable subset of $\mathbb{R}^m$, which, in turn, is endowed with the Borel sigma algebra $\mathcal{B}_\Omega$. Let X be a random vector taking values in the set Ω, and $f = (f_1, \ldots, f_k) : \Omega \to \mathbb{R}^k$ be a vector of moment functionals. Let $\mathcal{P}_\Omega$ and $\mathcal{M}_\Omega^+$ denote, respectively, the set of probability and non-negative measures, respectively on $(\Omega, \mathcal{B}_\Omega)$ such that the Borel measurable functionals $\phi, f_1, f_2, \ldots, f_k$, defined on Ω, are all integrable. Given a real vector $q = (q_1, \ldots, q_k)$, the objective of the problem of moments is to find the worst-case bound,

$$v(q) := \sup \left\{ E_\mu[\phi(X)] : E_\mu[f(X)] = q,\ \mu \in \mathcal{P}_\Omega \right\}. \tag{49}$$

If we let $f_0 = \mathbf{1}_\Omega$, it is convenient to add the constraint, $E_\mu[f_0(X)] = 1$, by appending $\tilde{f} = (f_0, f_1, \ldots, f_k)$, $\tilde{q} = (1, q_1, \ldots, q_k)$, and consider the following reformulation of the above problem:

$$v(q) := \sup \left\{ \int \phi(x) d\mu(x) : \int \tilde{f}(x) d\mu(x) = \tilde{q},\ \mu \in \mathcal{M}_\Omega^+ \right\}. \tag{50}$$

Then, under the assumption that a certain Slater's type of condition is satisfied, one has the following equivalent dual representation for the moment problem (50). See Theorem 1 (and the discussion of Case [I] following Theorem 1) in [18] for a proof of the following result:

Proposition 6. *Let* $\mathcal{Q}_{\tilde{f}} = \left\{ \int \tilde{f}(x) d\mu(x) : \mu \in \mathcal{M}_{\Omega}^{+} \right\}$. *If* $\tilde{q} = (1, q_1, \ldots, q_k)$ *is an interior point of* $\mathcal{Q}_{\tilde{f}}$, *then*

$$v(q) = \inf \left\{ \sum_{i=0}^{k} a_i q_i : \ a_i \in \mathbb{R}, \ \sum_{i=0}^{k} a_i \tilde{f}_i(x) \geq \phi(x) \text{ for all } x \in \Omega \right\}.$$

In the rest of this section, we recast the dual reformulation of RWP function (in (1)) and the dual reformulation of the distributional representation in Proposition 4 as particular cases of the dual representation of the problem of moments in Proposition 6.

Dual representation of RWP function Recall from Section 2.3 that W is a random vector taking values in $\mathbb{R}^m$ and $h(\cdot, \theta)$ is Borel measurable.

Proof of Proposition 1. For simplicity, we do not write the dependence on parameter θ in $h(u, \theta)$ and $R_n(\theta)$ in this proof; nevertheless, we should keep in mind that the RWP function is a function of parameter θ. Given estimating equation $E[h(W)] = \mathbf{0}$. Recall the definition of the corresponding RWP function,

$$\begin{aligned} R_n &:= \inf \left\{ D_c(P, P_n) : \ E_P\big[h(W)\big] = \mathbf{0} \right\} \\ &= \inf \left\{ E_\pi\big[c(U, W)\big] : \ E_\pi\big[h(U)\big] = \mathbf{0}, \ \pi_{_W} = P_n, \ \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m) \right\}. \end{aligned}$$

where $\pi_{_W}$ denotes the marginal distribution of W. To recast this as a problem of moments as in (49), let $\Omega = \{(u, w) \in \mathbb{R}^m \times \mathbb{R}^m : c(u, w) < \infty\}$,

$$f(u, w) = \begin{bmatrix} \mathbf{1}_{\{w = W_1\}}(u, w) \\ \mathbf{1}_{\{w = W_2\}}(u, w) \\ \vdots \\ \mathbf{1}_{\{w = W_n\}}(u, w) \\ h(u) \end{bmatrix} \qquad \text{and} \qquad q = \begin{bmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \\ \mathbf{0} \end{bmatrix}.$$

Further, let $\phi(u, w) = -c(u, w)$, for all $(u, w) \in \Omega$. Then,

$$R_n = -\sup \left\{ E_\pi\big[\phi(U, W)\big] : \ E_\pi\big[f(U, W)\big] = q, \ \pi \in \mathcal{P}_\Omega \right\},$$

is of the same form as (49). Let H denote the convex hull of the range $\{h(u) : (u, w) \in \Omega\}$. Then, following the definition of $Q_{\tilde{f}}$ in the abstract formulation in Proposition 6, we obtain $Q_{\tilde{f}} = \mathbb{R}_{+}^{n+1} \times H$. As $\{\mathbf{0}\}$ lies in the interior of convex hull H, it is immediate that the Slater's condition, $\tilde{q} = (1, q)$ lying in the interior of the $Q_{\tilde{f}}$, is satisfied. Consequently, we obtain the following dual representation of R_n due to Proposition 6:

$$\begin{aligned} R_n \ &= \ -\inf_{a_i \in \mathbb{R}} \left\{ a_0 + \frac{1}{n} \sum_{i=1}^{n} a_i : \ a_0 + \sum_{i=1}^{n} a_i \mathbf{1}_{\{w = W_i\}}(u, w) \right. \\ &\qquad \left. + \sum_{i=n+1}^{k} a_i h_i(u) \geq -c(u, w), \ \text{for all } (u, w) \in \Omega \right\} \\ &= \ -\inf_{a_i \in \mathbb{R}} \left\{ a_0 + \frac{1}{n} \sum_{i=1}^{n} a_i : \ a_0 + a_i \geq \sup_{u : c(u, W_i) < \infty} \left\{ -c(u, W_i) - \sum_{i=n+1}^{k} a_i h_i(u) \right\} \right\}. \end{aligned}$$

As the inner supremum is not affected even if we take supremum over $\{u : c(u, W_i) = \infty\}$, after letting $\lambda = (a_{n+1}, \ldots, a_k)$ for notational convenience, we obtain

$$R_n = \sup_{\lambda} \left\{ \frac{1}{n} \sum_{i=1}^{n} \inf_{u \in \mathbb{R}^m} \left\{ c(u, W_i) + \lambda^T h(u) \right\} \right\} \tag{51}$$

As λ is a free variable, we flip the sign of λ to arrive at the statement of Proposition 1. This completes the proof. □

Dual representation of the DRO formulation in (15) Here, we provide a proof for the dual representation in Proposition 4 that has been instrumental in establishing the distributional robust representations of LASSO and regularized logistic regression.

Proof of Proposition 4. Given a Borel measurable g, our first objective is to prove that the worst-case loss $\sup\{E_P[g(W)] : D_c(P, P_n) \leq \delta\}$ admits the dual representation,

$$\sup\{E_P[g(W)] : D_c(P, P_n) \leq \delta\} = \inf_{\lambda \geq 0} \left\{ \lambda\delta + \frac{1}{n} \sum_{i=1}^{n} \phi_\lambda(W_i) \right\}, \tag{52}$$

with $\phi_\lambda(W_i) = \sup_u\{g(u) - \lambda c(u, w)\}$. This would essentially prove Proposition 4 if we let $W = (X, Y)$, $g(W) = l(X, Y; \beta)$ and $\phi_\lambda(X, Y; \beta) = \phi_\lambda(W)$.

Since the problem $\sup\{E_P[g(W)] : D_c(P, P_n) \leq \delta\}$ has inequality constraints, one way is to proceed exactly as in RWP dual formulation above except for restricting the Lagrange multiplier corresponding to the equality constraint to be non-negative. Alternatively, one can recast the problem as in (49) with the introduction of a slack variable S as below:

$$\begin{aligned} & \sup\left\{E_P[g(W)] : D_c(P, P_n) \leq \delta\right\} \\ = \; & \sup\left\{E_\pi\big[g(U)\big] : \; E_\pi\big[c(U, W) + S\big] = \delta, \; \pi_{_W} = P_n, \; \pi(S = v) = 1, \; \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}_+)\right\}. \end{aligned}$$

In the context of notation introduced for the problem of moments described at the beginning of this appendix, let $\Omega = \{(u, w, s) : c(u, w) < \infty, s \geq 0\}$,

$$f(u, w, s) = \begin{bmatrix} \mathbf{1}_{\{w = W_1\}}(u, w, s) \\ \mathbf{1}_{\{w = W_2\}}(u, w, s) \\ \vdots \\ \mathbf{1}_{\{w = W_n\}}(u, w, s) \\ \mathbf{1}_{\{s = v\}} \\ c(u, w) + s \end{bmatrix} \qquad \text{and} \qquad q = \begin{bmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \\ 1 \\ \delta \end{bmatrix}.$$

In addition, if we let $\phi(u, w, s) = g(u)$, then

$$\sup\left\{E_P[g(W)] : \; D_c(P, P_n) \leq \delta\right\} = \sup\left\{E_\pi[\phi(U, W, S)] : \; E_\pi[f(U, W, S)] = q, \; \pi \in \mathcal{P}_\Omega\right\},$$

is a problem of moments of the form (49). Similar to the RWP dual formulation discussed earlier in the section, $\tilde{q} = (1, q)$ lies in the interior of $Q_{\tilde{f}} = \mathbb{R}_+^{n+3}$, thus satisfying Slater's condition for all $\delta > 0$. Then, due to Proposition 6, we obtain

$$\sup\left\{E_P[g(W)] : \; D_c(P, P_n) \leq \delta\right\} = \inf_{a \in \mathbb{A}} \left\{ a_0 + \frac{1}{n} \sum_{i=1}^{n} a_i + a_{n+1} + a_{n+2}\delta \right\}$$

where the set $\mathbb{A}$ is the collection of $a = (a_0, a_1, \ldots, a_{n+1}) \in \mathbb{R}^{n+3}$ such that

$$a_0 + \sum_{i=1}^{n} a_i \mathbf{1}_{\{w=W_i\}}(u,w,s) + a_{n+1}\mathbf{1}_{\{s=v\}}(u,w,s) + a_{n+2}\big(c(u,w)+s\big) \geq g(u),$$

for all $(u,w,s) \in \Omega$. Further, observe that the value of the optimization problem above does not change, even if we consider only the following constraints:

$$\begin{aligned} a_0 + a_i + a_{n+1} &\geq \sup\left\{ g(u) - a_{n+2}\big(c(u,W_i)+s\big) : u \in \mathbb{R}^m,\ s \geq 0 \right\} \\ &= \begin{cases} \sup_{u\in\mathbb{R}^m}\left\{ g(u) - a_{n+2}c(u,W_i)\right\} & \text{if } a_{n+2} \geq 0, \\ \infty & \text{if } a_{n+2} < 0. \end{cases} \end{aligned}$$

If we recall the notation that $\phi_\lambda(W_i) = \sup_{u\in\mathbb{R}^m}\left\{g(u) - \lambda c(u,W_i)\right\}$, then

$$\begin{aligned} & \sup\left\{E_P[g(W)] :\ D_c(P,P_n) \leq \delta\right\} \\ = & \inf_{\substack{a_{n+2}\geq 0 \\ a_i\in\mathbb{R},}} \left\{ a_0 + \frac{1}{n}\sum_{i=1}^{n} a_i + a_{n+1} + a_{n+2}\delta : a_0 + a_i + a_{n+1} \geq \phi_{a_{n+2}}(W_i) \right\} \\ = & \inf_{a_{n+2}\geq 0}\left\{\phi_{a_{n+2}}(W_i) + a_{n+2}\delta\right\}, \end{aligned}$$

thus proving (52). As explained earlier, letting $W = (X,Y)$, $g(W) = l(X,Y;\beta)$ and $\phi_\lambda(X,Y;\beta)$ in (52) verifies the proof of Proposition 4. □

Appendix B. Exchange of sup and inf in the DRO formulation (15)

Proposition 7. *Let us write*

$$\mathcal{U}_\delta = \left\{P : \mathcal{D}_c\big(P,P_n\big) \leq \delta\right\},$$

and define

$$g\left(\beta\right) = \sup_{P\in\mathcal{U}_\delta} E_P\left[l\big(X,Y;\beta\big)\right].$$

Suppose that $g\left(\cdot\right)$ is convex and assume that there exists $b \in \mathbb{R}$ such that $\kappa_b = \{\beta : g\left(\beta\right) \leq b\}$ is compact and non-empty. In addition, suppose that $E_P\left[l\big(X,Y;\beta\big)\right]$ is lower semi-continuous and convex as a function of β throughout κ_b. Then,

$$\min_{\beta\in\mathbb{R}^d} \sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta} E_P\left[l\big(X,Y;\beta\big)\right] = \sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta} \min_{\beta\in\mathbb{R}^d} E_P\left[l\big(X,Y;\beta\big)\right].$$

Proof. By definition of κ_b we have that

$$\min_{\beta\in\mathbb{R}^d} \sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta} E_P\left[l\big(X,Y;\beta\big)\right] = \min_{\beta\in\kappa_b} \sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta} E_P\left[l\big(X,Y;\beta\big)\right].$$

By a min-max result of Terkelsen (see Corollary 2 in [40]), since both $\mathcal{U}_\delta$ and κ_b are convex, κ_b is compact, $E_P\left[l\big(X,Y;\beta\big)\right]$ is lower semi-continuous and convex throughout κ_b as a function of β, and $E_P\left[l\big(X,Y;\beta\big)\right]$ is concave as a function of P, then

$$\min_{\beta\in\kappa_b} \sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta} E_P\left[l\big(X,Y;\beta\big)\right] = \sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta} \min_{\beta\in\kappa_b} E_P\left[l\big(X,Y;\beta\big)\right].$$

The proof is complete if are able to argue the identity

$$\sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta}\min_{\beta\in\kappa_b}E_P\left[l\left(X,Y;\beta\right)\right]=\sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta}\min_{\beta\in\mathbb{R}^d}E_P\left[l\left(X,Y;\beta\right)\right].$$

To see this, note that we always have

$$\sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta}\min_{\beta\in\kappa_b}E_P\left[l\left(X,Y;\beta\right)\right]\geq\sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta}\min_{\beta\in\mathbb{R}^d}E_P\left[l\left(X,Y;\beta\right)\right]. \tag{53}$$

Let us assume that the strict inequality holds. If this is the case then we most have that there exists $\beta'\notin\kappa_b$ such that

$$\begin{aligned}b &< g\left(\beta'\right)=\sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta}E_P\left[l\left(X,Y;\beta'\right)\right]\\ &< \sup_{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta}\min_{\beta\in\kappa_b}E_P\left[l\left(X,Y;\beta\right)\right]\\ &\leq b\end{aligned}$$

where the second inequality follows because we are assuming that (53) holds with strict inequality, and the last inequality follows because for each $\beta\in\kappa_b$

$$\min_{\beta\in\kappa_b}E_P\left[l\left(X,Y;\beta\right)\right]\leq E_P\left[l\left(X,Y;\beta\right)\right].$$

We therefore contradict the assumption that the strict inequality in (53) holds. Hence, the proof is complete. □

Proof of Lemma 1. Let us consider linear regression loss function first. Under null hypothesis, $E\|X\|_2^2<\infty$ and $E[e^2]<\infty$. Therefore, for any $\beta\in\mathbb{R}^d$, $E[l(X,Y;\beta)]=E[(Y-\beta^TX)^2]<\infty$. Further, as the loss function $l(x,y;\beta)$ is a convex function of β,

$$g(\beta)=\sup_{P\in\ \mathcal{U}_\delta}E_P[l(X,Y;\beta)]=\left(\sqrt{E_{P_n}[(Y-\beta^TX)^2]}+\sqrt{\delta}\|\beta\|_p\right)^2$$

is convex as well and finite for all β in $\mathbb{R}^d$ (the second equality follows from the distributional robust representation in Theorem 2). Further, as $g(\beta)\to\infty$ when $\|\beta\|_p\to\infty$ and $g(\beta)$ is convex and continuous in $\mathbb{R}^d$, the level sets $\kappa_b=\{\beta:g(\beta)\leq b\}$ are compact and nonempty as long as $b>(\sqrt{E_{P_n}[(Y-\beta_*^TX)^2]}+\sqrt{\delta}\|\beta_*\|)^2$. Finally, due to the convexity and finiteness of $E[l(X,Y;\beta)]$ lower semi-continuity of $E[l(X,Y;\beta)]$ is immediate as well. As all the conditions in Proposition 7 are satisfied, the sup and inf in the DRO formulation (15) can be exchanged in the linear regression example as a consequence of Proposition 7.

It is straightforward to check that exactly same argument applies for logistic regression loss function as well when $E\|X\|_2^2$ is finite. □

E-mail address: jose.blanchet@columbia.edu, yangkang@stat.columbia.edu, karthyek.murthy@columbia.edu