J. Appl. Prob. **56,** 830–857 (2019)
*doi:*10.1017/jpr.2019.49

ROBUST WASSERSTEIN PROFILE INFERENCE AND APPLICATIONS TO MACHINE LEARNING

JOSE BLANCHET,* *Stanford University*
YANG KANG,** *Columbia University*
KARTHYEK MURTHY,*** *Singapore University of Technology and Design*

Abstract

We show that several machine learning estimators, including square-root least absolute shrinkage and selection and regularized logistic regression, can be represented as solutions to distributionally robust optimization problems. The associated uncertainty regions are based on suitably defined Wasserstein distances. Hence, our representations allow us to view regularization as a result of introducing an artificial adversary that perturbs the empirical distribution to account for out-of-sample effects in loss estimation. In addition, we introduce RWPI (robust Wasserstein profile inference), a novel inference methodology which extends the use of methods inspired by empirical likelihood to the setting of optimal transport costs (of which Wasserstein distances are a particular case). We use RWPI to show how to optimally select the size of uncertainty regions, and as a consequence we are able to choose regularization parameters for these machine learning estimators without the use of cross validation. Numerical experiments are also given to validate our theoretical findings.

Keywords: Distributionally robust optimization; Wasserstein distance; regularization; square-root LASSO; logistic regression; support vector machine; limit characterization of optimal Wasserstein ball radius and regularization parameter; empirical likelihood

2010 Mathematics Subject Classification: Primary 60F05
Secondary 62J05; 62J12

1. Introduction

Regularization has become crucial in machine learning practice and the goal of this paper is to revisit the idea of regularization from an optimal transport perspective. Specifically, we show that the role of regularization in machine learning can often be interpreted as the result of optimally transporting mass from the empirical measure in order to maximize a certain loss under a budget constraint. Thus, our results connect directly optimal transport phenomena (a classical concept in probability reviewed in Section 2.1) to regularization (a key tool in machine learning to be discussed in the following subsection).

Moreover, this connection will show that the so-called regularization parameter (i.e. the coefficient of the regularization term) coincides with the size of the budget constraint by

Received 15 February 2018; revision received 25 March 2019.
The supplementary material for this article can be found at http://doi.org/10.1017/jpr.2019.49
* Postal address: Management Science and Engineering, Stanford University, 475 Via Ortega, Stanford, CA 94305, USA.
** Postal address: Columbia University, 1255 Amsterdam Avenue, Rm 1005, New York, NY 10027, USA.
*** Postal address: Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372, Singapore.

which we permit mass transportation to occur. As we shall see, the budget constraint has a natural interpretation based on a distributionally robust optimization (DRO) formulation, which in turn allows us to define a reasonable optimization criterion for the regularization parameter. Thus, our approach uses optimal mass transportation phenomena to explain the nature of regularization and how to select the regularization parameter in several machine learning estimators, including square-root LASSO (least absolute shrinkage and selection) and regularized logistic regression, among others.

The size of the budget constraint is also referred to as the radius (or size) of the uncertainty set in the DRO literature. The method that we develop for optimally choosing this budget constraint can actually be applied to a wide range of inference and decision problems, but we have focused our discussion on machine learning applications because of the substantial amount of activity that the area has generated, and also to demonstrate the utility of the tools that are commonly used in applied probability in this rapidly growing area.

1.1. Regularization in linear regression

In order to introduce the proposed method for optimally choosing the radius of the uncertainty set, let us walk through a simple application in a familiar context, namely, that of linear regression. Throughout the paper any vector is understood to be a column vector and the transpose of x is denoted by $x^\top$. We use the notation $\mathrm{E}_{\mathrm{P}}[\cdot]$ to denote expectation with respect to a probability distribution P.

Example 1. (*Square-root LASSO.*) Consider a training data set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where 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. It is postulated that

$$Y_i = \beta_*^\top X_i + e_i$$

for some $\beta_* \in \mathbb{R}^d$ and errors $\{e_1, \ldots, e_n\}$. Under suitable statistical assumptions we may be interested in estimating β_*. Underlying is a general loss function, $l(x, y; \beta)$, which we shall take for simplicity in this discussion to be the quadratic loss, namely, $l(x, y; \beta) = (y - \beta^\top x)^2$. Let P_n denote the empirical distribution:

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

Over the last two decades, various 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, [47] for an early reference and [21] for a discussion on regularized estimators). One such regularized estimator, implemented, for example in the 'flare' package (see [27]), is the so-called square-root LASSO estimator that is obtained by solving the following convex optimization problem in β:

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

where $\|\beta\|_p$ denotes the ℓ_p-norm. The parameter λ, commonly referred to as the regularization parameter, is crucial for the performance of the algorithm. It is often chosen using cross validation, a procedure that iterates over a multitude of choices of λ in order to choose the best.

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

1.1.1. *DRO representation of square-root LASSO.* One of our contributions in this paper (see Section 2) is a representation of (1) in terms of a distributionally robust optimization formulation. We construct a discrepancy measure, $\mathcal{D}_c(\mathrm{P}, \mathrm{Q})$, corresponding to a Wasserstein-type distance between two probability measures P and Q which is defined in terms of a suitable transportation cost function $c(\cdot)$. If $c(\cdot)$ is based on the ℓ_q-distance (for $q > 1$), we show that

$$\min_{\beta\in\mathbb{R}^d}\{\sqrt{\mathrm{E}_{\mathrm{P}_n}[l(X, Y; \beta)]} + \lambda\|\beta\|_p\}^2 = \min_{\beta\in\mathbb{R}^d} \max_{\mathrm{P}\,:\,\mathcal{D}_c(\mathrm{P},\mathrm{P}_n)\le\delta} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)], \tag{2}$$

where $1/p + 1/q = 1$ and $\lambda = \sqrt{\delta}$. We can gain a great deal of insight from (2). For example, note that the regularization parameter $\lambda = \sqrt{\delta}$ is fully determined by the size (or 'radius') of the uncertainty, δ, in the DRO formulation on the right-hand side of (2). In addition, we can interpret (2) as a game in which an artificial adversary is introduced in order to explore and quantify out-of-sample effects in our estimates of the expected loss.

1.1.2. *Optimal choice of the radius δ.* The set $\mathcal{U}_\delta(\mathrm{P}_n) = \{\mathrm{P} : \mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) \le \delta\}$ is called the uncertainty set in the language of DRO, and it represents the class of models that are, in some sense, plausible variations of P_n. Note that $\mathcal{U}_\delta(\mathrm{P}_n)$ is precisely the feasible region over which the maximization is taken in (2). Then we define the collection

$$\Lambda_n(\delta) := \bigcup_{\mathrm{P}\in\mathcal{U}_\delta(\mathrm{P}_n)} \arg\min_{\beta\in\mathbb{R}^d} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)] \tag{3}$$

comprising optimal β for every $\mathrm{P} \in \mathcal{U}_\delta(\mathrm{P}_n)$ to be the set of *plausible* selections of the parameter β_*. For δ chosen sufficiently large, the set $\Lambda_n(\delta)$ is a natural confidence region for β_*. Moreover, we shall see that any β that solves $\inf_\beta \sup_{\mathrm{P}\in\mathcal{U}_\delta(\mathrm{P}_n)} \mathrm{E}[l(X, Y; \beta)]$ is a member of $\Lambda_n(\delta)$.

Given these interpretations, it is natural to select a confidence level, $1 - \alpha$, and then choose $\delta = \delta_n^*$ optimally via

$$\delta_n^* = \min\{\delta > 0 : \mathrm{P}(\beta_* \in \Lambda_n(\delta)) \ge 1 - \alpha\}. \tag{4}$$

In words, the optimization criterion can be stated as finding the smallest δ such that β_* is itself a plausible selection with $1 - \alpha$ confidence. Essentially, given a desired confidence level $1 - \alpha$, we seek to choose a δ just large enough such that $\Lambda_n(\delta)$ is a $(1 - \alpha)$-confidence region for the parameter β_*. As we shall see in Section 4, this choice ensures that any β that minimizes $\inf_\beta \sup_{\mathrm{P}\in\mathcal{U}_\delta(\mathrm{P}_n)} \mathrm{E}[l(X, Y; \beta)]$ is indeed in the confidence region $\Lambda_n(\delta)$. We next explain how to solve the optimization problem in (4) asymptotically as $n \to \infty$.

1.1.3. *The associated Wasserstein profile function.* In order to asymptotically solve (4) we introduce a novel statistical inference methodology, which we call RWPI (robust Wasserstein-distance profile-based inference; pronounced similar to 'rupee'). This can be understood as an extension of empirical likelihood (EL) that uses optimal transport cost rather than the likelihood. The extension is not just a formality, as we shall see, because different phenomena and scalings arise relative to EL.

We next illustrate how δ_n^* in (4) corresponds to 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.

Observe that, by convexity of the loss function, $\beta \in \Lambda_\delta(\mathrm{P}_n)$ if and only if there exists $\mathrm{P} \in \mathcal{U}_\delta(\mathrm{P}_n)$ such that β satisfies the first-order optimality condition, namely

$$D_\beta \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)] = \mathrm{E}_{\mathrm{P}}[(Y - \beta^\top X)X] = \mathbf{0}. \tag{5}$$

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

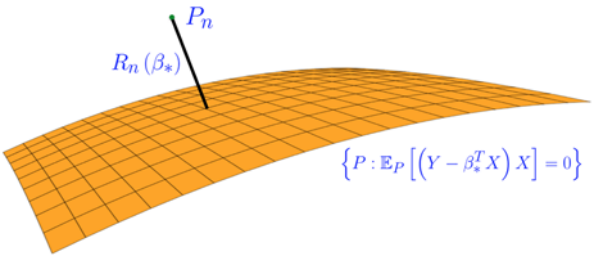

FIGURE 1: Illustration of RWP function evaluated at β_*.

We then introduce the following object, which is the RWP function associated with the estimating equation (5):

$$R_n(\beta) = \inf\{\mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) : \mathrm{E}_{\mathrm{P}}[(Y - \beta^\top X)X] = \mathbf{0}\}. \tag{6}$$

It turns out that the infimum is achieved in the previous expression, so we can write min instead; this is not crucial for our discussion but it is sometimes helpful to keep in mind. Using this definition of $R_n(\beta)$, we can see immediately that the events

$$\{R_n(\beta_*) \le \delta\} = \{\beta_* \in \Lambda_n(\delta)\},$$

which implies that δ_n^* is precisely the $1-\alpha$ quantile, $\chi_{1-\alpha}$, of $R_n(\beta_*)$; that is,

$$\delta_n^* = \chi_{1-\alpha} = \inf\{z : \mathrm{P}(R_n(\beta_*) \le z) \ge 1 - \alpha\}.$$

Moreover, note that $R_n(\beta)$ allows us to provide an explicit characterization of $\Lambda_n(\chi_{1-\alpha})$:

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

So, $\Lambda_n(\chi_{1-\alpha}) = \{\beta : R_n(\beta) \le \chi_{1-\alpha}\}$ is a $(1-\alpha)$-confidence region for β^*.

1.1.4. *Further intuition behind the RWP function.* In order to further explain the role of $R_n(\beta_*)$, let us define $\mathcal{P}_{\mathrm{opt}} := \{\mathrm{P} : \mathrm{E}_{\mathrm{P}}[(Y - \beta_*^\top X)X] = \mathbf{0}\}$. In words, $\mathcal{P}_{\mathrm{opt}}$ is the set of probability measures for which β_* is an optimal risk minimization parameter. Naturally, the distribution of (X, Y), from which the samples are generated, is an element of $\mathcal{P}_{\mathrm{opt}}$. Since $R_n(\beta_*) = \inf\{\mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) : \mathrm{P} \in \mathcal{P}_{\mathrm{opt}}\}$, the set $\{\mathrm{P} : \mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) \le R_n(\beta_*)\}$ denotes the smallest uncertainty region around P_n (in terms of $\mathcal{D}_c$) for which there exists a distribution P satisfying the optimality condition $\mathrm{E}_{\mathrm{P}}[(Y - \beta_*^\top X)X] = \mathbf{0}$. See Figure 1 for a pictorial representation of $\mathcal{P}_{\mathrm{opt}}$ and $R_n(\beta^*)$.

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

$$\{\mathrm{P} : \mathrm{E}_{\mathrm{P}}[(Y - \beta_*^\top X)X] = \mathbf{0}, \mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) \le \delta\}$$

is non-empty.

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

1.2. A broader perspective of our contribution

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. We shall study the RWP function for estimating equations generally and systematically, but we showcase the use of the RWP function only in the context of optimal regularization.

Broadly speaking, RWPI can be seen as a statistical methodology that utilizes a suitably defined RWP function to estimate a parameter of interest. From a philosophical standpoint, RWPI borrows heavily from EL, introduced in the seminal work of Owen [30, 31]. There are important methodological differences, however, as we shall discuss below. In the last three decades there has been a large number of successful applications of EL for inference [11, 22, 32, 35, 52]. In principle all of those applications can be revisited using the RWP function and its ramifications.

We now provide a more precise description of our contributions.

(A) We explain how, by judiciously choosing $\mathcal{D}_c(\cdot)$, we can define a family of regularized regression estimators (see Section 2). In particular, we show how square-root LASSO (see Theorem 1), regularized logistic regression, and support vector machines (see Theorem 2) arise as particular cases of suitable DRO formulations.

(B) We derive general limit theorems for the asymptotic distribution (as the sample size increases) of the RWP function defined for general estimating equations. These limit theorems, derived in Section 3.3, allow us to employ RWPI to perform inference and choose the radius of uncertainty δ in settings that are more general than linear/logistic regression.

(C) We use our results from (B) to obtain prescriptions for regularization parameters in square-root LASSO and regularized logistic regression settings (see Section 4). We also illustrate how coverage results for the optimal risk that demonstrate an $O(n^{-1/2})$ rate of convergence are obtained immediately as a consequence of choosing $\delta \geq R_n(\beta_*)$.

(D) We analyze our regularization selection in the high-dimensional setting for square-root LASSO. Under standard regularity conditions, we show (see Theorem 7) 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 (see Section 4.4 for further details).

(E) We analyze the empirical performance of RWPI-based selection of regularization parameters in the context of square-root LASSO. In Section 5, we compare the performance of RWPI-based optimal regularization with that of a cross-validation-based approach on both simulated and real data. We conclude that the RWPI-based approach yields similar performance, without having to repeat the algorithm over various choices of regularization parameters (as done in cross-validation).

We now provide a discussion of topics related to RWPI.

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

1.3. On related literature

Connections between robust optimization and regularization procedures such as LASSO and support vector machines have been studied in the literature [3, 50, 51]. The methods proposed here differ subtly: While [50] and [51] 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 the empirical distribution, as we shall see, allows us to derive suitable limit laws (or) probabilistic inequalities that can be used to give a systematic prescription for the radius of uncertainty, δ, in the definition of the uncertainty region $\mathcal{U}_\delta(\mathrm{P}_n) = \{\mathrm{P} : \mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) \leq \delta\}$.

It is well understood that as the number of samples n increases, the expected deviation of the empirical distribution from the true distribution decays to zero, as a function of n, at a specific rate. To begin with, as a direct approach towards choosing the size of the uncertainty δ, we can perhaps use a suitable concentration inequality that measures such rate of convergence in terms of Wasserstein distances (see, for example, [17] and references therein). Such a simple specification of the size of the uncertainty, suitably as a function of n, does not arise naturally in the deterministic robust optimization approaches in [50, 51].

For an application of these concentration inequalities to choosing the size of the uncertainty set in the context of distributionally robust logistic regression and data-driven DRO, refer to [28, 40]. The exact representation for regularized logistic regression we derive later, in Section 2.4, can be can be seen as an extension in which the approximate representation described in [40, Remark 1] is made to coincide exactly with the regularized logistic regression estimator that has been widely used in practice. It is important to note that, despite imposing severe tail assumptions, the concentration inequalities used to choose the radius of the uncertainty set in [28, 40] dictate that the size of the uncertainty decay at the rate $O(n^{-1/d})$; unfortunately, this prescription scales non-gracefully as the number of dimensions d increases, and the resulting coverage guarantees suffer from a poor rate of convergence (see, for example, [28, Theorem 3.5], [40, Theorem 2]). Since most of the modern learning and decision problems have huge numbers of covariates, application of such concentration inequalities with poor rates of decay with dimensions may not be suitable for applications.

In contrast to directly using concentration inequalities, as we shall see, the prescription obtained via RWPI typically has a rate of convergence of order $O(n^{-1/2})$ as $n \to \infty$ (for fixed d). In particular, as we discuss in the case of LASSO, according to our results corresponding to contribution (E), RWPI-based prescription of the size of uncertainty can actually be shown (under suitable regularity conditions) to decay at a rate $O(\sqrt{\log d/n})$ (uniformly over d and n such that $\log^2 d \ll n$), which is in agreement with the findings of the high-dimensional statistics literature (see [2, 12, 29] and references therein). A profile-function-based approach towards calibrating the radius of uncertainty in the context of empirical-likelihood-based DRO can be found in [14, 20, 26, 25].

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

$$\min_{\theta\,:\,G(\theta)\leq 0} \ \max_{\mathrm{P}\,:\,\mathcal{D}_c(\mathrm{P},\mathrm{P}_n)\leq\delta} \mathrm{E}_{\mathrm{P}}[H(W,\theta)]$$

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

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

$\{W_1, \ldots, W_n\}$. One can then follow reasoning parallel to what we advocate throughout our LASSO discussion. Argue, by applying the corresponding Karush–Kuhn–Tucker conditions, if possible, that an optimal solution θ_* to the problem

$$\min_{\theta\,:\,G(\theta)\leq 0} \mathrm{E}_{\mathrm{P}_{\mathrm{true}}}[H(W,\theta)]$$

satisfies a system of estimating equations of the form $\mathrm{E}_{\mathrm{P}_{\mathrm{true}}}[h(W,\theta_*)]=0$ for a suitable $h(\cdot)$ (where $\mathrm{P}_{\mathrm{true}}$ is the weak limit of the empirical measure P_n as $n\to\infty$). Then, given a confidence level $1-\alpha$, we should choose δ as the $(1-\alpha)$ quantile of the RWP function

$$R_n(\theta_*)=\inf\{\mathcal{D}_c(\mathrm{P},\mathrm{P}_n):\mathrm{E}_{\mathrm{P}}[h(W,\theta_*)]=\mathbf{0}\}.$$

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 square-root LASSO example, the selection of δ is the smallest possible choice for which θ_* is plausible with $(1-\alpha)$ confidence.

1.4. Connections to related inference literature

We next discuss the connections between RWPI and EL. In EL we build a profile likelihood for an estimating equation. For instance, in the context of EL applied to estimating β satisfying (5) we would build a profile likelihood function in which the optimization object is 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 $\{\mathrm{P}:\mathcal{D}_c(\mathrm{P},\mathrm{P}_n)\leq\delta\}$ 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, we are able, for example, 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 [13, 33, 49]. 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 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 the 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, this 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 using RWPI. In many settings, however, the parameters of the distribution can be easily estimated from the data itself.

Another methodology, strongly related to RWPI, by the name of SOS (sample out-of-sample) inference has been studied recently [6]. A suitable RWP function is built in this

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

setting as well, 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 from independent distributions (defined by the user). The limit 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. As explained in [6, 7], SOS inference is natural in applications such as semi-supervised learning, in which massive amounts of unlabeled data inform the support of the covariates.

1.5. Organization of the paper

The rest of the paper is organized as follows. Section 2 corresponds to contribution (A): we first introduce Wasserstein distances and then discuss distributionally robust representations of popular machine learning algorithms. Section 3 deals with contribution (B): we discuss the RWP function as an inference tool in a way which is parallel to the profile likelihood in EL, and derive the asymptotic distribution of the RWP function for general estimating equations. Section 4 discusses contribution (C), namely the application of the results from (B) for optimal regularization. Our high-dimensional analysis of the RWP function in the case of square-root LASSO is also presented in Section 4. Numerical experiments using both simulated and real data sets are given in Section 5. Proofs of all the results are presented in the supplementary material [9].

2. Optimal transport definitions and DRO representations of machine learning estimators

We begin with definitions of optimal transport costs and Wasserstein distances.

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 $c(u, u) = 0$ for every $u \in \mathbb{R}^m$. Given two probability distributions $\mathrm{P}(\cdot)$ and $\mathrm{Q}(\cdot)$ supported on $\mathbb{R}^m$, the optimal transport cost or discrepancy between P and Q, denoted by $\mathcal{D}_c(\mathrm{P}, \mathrm{Q})$, is defined as

$$\mathcal{D}_c(\mathrm{P}, \mathrm{Q}) = \inf\{\mathrm{E}_\pi[c(U, W)] : \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m),\ \pi_U = \mathrm{P},\ \pi_W = \mathrm{Q}\}. \tag{7}$$

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, π_W denote the marginals of U and W, respectively, under the joint distribution π. Intuitively, the quantity $c(u, w)$ can be interpreted as the cost of transporting unit mass from u in $\mathbb{R}^m$ to another element w in $\mathbb{R}^m$. Then the expectation $\mathrm{E}_\pi[c(U, W)]$ corresponds to the expected transport cost associated with the joint distribution π.

In addition to the stated assumptions on the cost function $c(\cdot)$, if $c^{1/\rho}$ satisfies the properties of a metric for any $\rho > 1$ then $\mathcal{D}_c^{1/\rho}(\mathrm{P}, \mathrm{Q})$ defines a metric between probability distributions (see [48] for a proof and other properties of D_c). For example, if $c(u, w) = \|u - w\|_2^2$ 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,

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

coincides with the Wasserstein distance of order two. More generally, if we choose $c^{1/\rho}(u, w) = \|u - w\|_q$ for some $\rho, q \geq 1$, then $\mathcal{D}_c^{1/\rho}(\cdot)$ is known as the Wasserstein distance of order ρ.

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

Wasserstein distances metrize weak convergence of probability measures under suitable moment assumptions and have received immense attention in probability theory (see [36, 37, 48] for a collection of classical applications). In addition, earth-mover's distance, a particular example of a Wasserstein distance, has been of interest in image processing (see [38, 44]). More recently, optimal transport metrics and Wasserstein distances are being actively investigated for their use in various machine learning applications (see [18, 34, 39, 45] and references therein for a growing list of new applications).

Throughout this paper we consider optimal transport costs $\mathcal{D}_c(\cdot)$ for a judiciously chosen cost function $c(\cdot)$ to result in formulations such as (2). As we shall see in Section 2.4, it is useful to allow $c(\cdot)$ to be lower semi-continuous and potentially be infinite in some region. Thus our setting requires discrepancy choices which are slightly more general than standard Wasserstein distances.

2.2. DRO formulation using optimal transport costs

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 the 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 square-root LASSO in Section 1.1, our first step will be to show that regularized linear and logistic regression estimators admit a DRO formulation of the form

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathrm{P} \,:\, \mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) \le \delta} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)]. \tag{8}$$

In contrast to empirical risk minimization that performs well only on the training data, the DRO problem (8) aims to find 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 (8) is said to be 'distributionally robust', and can be expected to generalize better. See [40], [50], and [51] for earlier works that relate robustness and generalization.

Recasting regularized regression as a DRO problem of the form (8) lets us view these regularized estimators under the lens of distributional robustness. The regularized estimators that we consider in this paper include the regularized logistic regression estimators in Example 2, support vector machines (see [21]), and the family of ℓ_p-norm penalized linear regression estimators of the form

$$\min_{\beta \in \mathbb{R}^d} \{\sqrt{\mathrm{E}_{\mathrm{P}_n}[l(X, Y; \beta)]} + \lambda \|\beta\|_p\} \tag{9}$$

for any $p \in [1, \infty)$. This collection includes the square-root LASSO estimator described in Example 1 as a special case where $p = 1$.

Example 2. (*Regularized logistic regression.*) Consider the context of binary classification in which case the training 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

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

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

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

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

and we are interested in estimating β_* by solving

$$\min_{\beta \in \mathbb{R}^d} \{\mathrm{E}_{\mathrm{P}_n}[l(X, Y; \beta)] + \lambda \|\beta\|_p\} \tag{10}$$

for $p \in [1, \infty)$. Refer to [21] for a more detailed discussion on regularized logistic regression.

2.3. Dual form of the DRO formulation (8)

Though the DRO formulation (8) involves optimizing over uncountably many probability measures, recent strong duality results for Wasserstein DRO (see, for example, Theorem 1 in [8]) ensures that the inner supremum in (8) admits a reformulation which is a simple, univariate optimization problem. Before stating the result, we recall that the definition of discrepancy measure $\mathcal{D}_c$ (see (7)) requires the specification of the cost function $c((x, y), (x', y'))$ between any two predictor–response pairs $(x, y), (x', y') \in \mathbb{R}^{d+1}$.

Proposition 1. *Let* $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0, \infty]$ *be a lower semi-continuous cost function satisfying* $c((x, y), (x', y')) = 0$ *whenever* $(x, y) = (x', y')$. *For* $\gamma \geq 0$ *and loss functions* $l(x, y; \beta)$ *that are upper semi-continuous in* (x, y) *for each* β, *define*

$$\phi_\gamma(X_i, Y_i; \beta) := \sup_{u \in \mathbb{R}^d, v \in \mathbb{R}} \{l(u, v; \beta) - \gamma c((u, v), (X_i, Y_i))\}. \tag{11}$$

Then

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

Consequently, the distributionally robust regression problem (8) reduces to

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathrm{P} : \mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) \leq \delta} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)] = \inf_{\beta \in \mathbb{R}^d} \min_{\gamma \geq 0} \left\{ \gamma \delta + \frac{1}{n} \sum_{i=1}^{n} \phi_\gamma(X_i, Y_i; \beta) \right\}. \tag{12}$$

Proposition 1 follows as a straightforward application of [8, Theorem 1]. As we shall see in Section 2.4, the function $\phi_\gamma(\cdot)$ is explicitly computable for various examples of interest. Of the reformulations in the literature for Wasserstein-distance-based DRO (see [8, 16, 19]), the general cost structure assumed in [8, Theorem 1] is essential for the exact recovery of the machine learning estimators that are presented in Section 2.4.

2.4. Distributionally robust representations

Example 1. (*Continued: Recovering regularized estimators for linear regression.*) We examine the right-hand side of (12) for the square loss function for the linear regression model $Y = \beta^\top X + e$, 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 (7).

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

Proposition 2. (Distributionally robust 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 the ℓ_q-norm. In other words, take $l(x, y; \beta) = (y - \beta^\top x)^2$ and $c((x, y), (u, v)) = \|(x, y) - (u, v)\|_q^2$. Then*

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathrm{P} : \mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) \le \delta} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)] = \inf_{\beta \in \mathbb{R}^d} \{\sqrt{\mathrm{MSE}_n(\beta)} + \sqrt{\delta}\, \|\bar{\beta}\|_p\}^2, \tag{13}$$

where $\mathrm{MSE}_n(\beta) = \mathrm{E}_{\mathrm{P}_n}[(Y - \beta^\top X)^2] = \frac{1}{n} \sum_{i=1}^n (Y_i - \beta^\top X_i)^2$ 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 distributionally robust regression applying a discrepancy measure based on neighborhoods defined using the ℓ_∞-norm:

$$\arg\min_{\beta \in \mathbb{R}^d} \sup_{\mathrm{P} : \mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) \le \delta} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)] = \arg\min_{\beta \in \mathbb{R}^d} \{\sqrt{\mathrm{MSE}_n(\beta)} + \sqrt{\delta}\, \|\bar{\beta}\|_1\}.$$

The right-hand side of (13) resembles ℓ_p-norm regularized regression (except for the fact that we have $\|\bar{\beta}\|_p$ instead of $\|\beta\|_p$). In order to obtain exact equivalence, we 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

$$N_q((x, y), (u, v)) = \begin{cases} \|x - u\|_q & \text{if } y = v, \\ \infty & \text{otherwise,} \end{cases} \tag{14}$$

in order to use $c(\cdot) = N_q(\cdot)$ as the transportation cost instead of the standard ℓ_q-norm $\|(x, y) - (u, v)\|_q$. Subsequently, we can consider modified cost functions of the form $c((x, y), (u, v)) = (N_q((x, y), (u, v)))^\rho$. As this modified cost function assigns infinite cost when $y \neq v$, the infimum in (6) is effectively over joint distributions that do not alter the marginal distribution of Y. As a consequence, the resulting neighborhood set $\{\mathrm{P} : \mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) \le \delta\}$ admits distributional ambiguities only with respect to the predictor variables X.

The following result is essentially the same as Proposition 2 except for the use of the modified cost N_q and the resulting norm regularization of the form $\|\beta\|_p$ (instead of $\|\bar{\beta}\|_p$ as in Proposition 2), thus exactly recovering the regularized regression estimators in (9).

Theorem 1. *Consider the square loss $l(x, y; \beta) = (y - \beta^\top x)^2$ and discrepancy measure $\mathcal{D}_c(\mathrm{P}, \mathrm{P}_n)$ defined as in (7) using the cost function $c((x, y), (u, v)) = (N_q((x, y), (u, v)))^\rho$ with $\rho = 2$. Then*

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathrm{P} : \mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) \le \delta} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)] = \inf_{\beta \in \mathbb{R}^d} \{\sqrt{\mathrm{MSE}_n(\beta)} + \sqrt{\delta}\, \|\beta\|_p\}^2,$$

where $\mathrm{MSE}_n(\beta) = \mathrm{E}_{\mathrm{P}_n}[(Y - \beta^\top X)^2] = n^{-1} \sum_{i=1}^n (Y_i - \beta^\top X_i)^2$ is the mean square error for the coefficient choice β and p is such that $1/p + 1/q = 1$.

Example 2. (*Continued: Recovering regularized estimators for classification.*) Apart from exactly recovering norm-regularized estimators for linear regression, the discrepancy measure $\mathcal{D}_c$ based on the modified norm N_q in (14) 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 2 and also for the widely used support vector machines (see [21]).

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

Theorem 2. (Regularized regression for classification.) *Consider the discrepancy measure* $\mathcal{D}_c(\cdot)$ *defined using the cost function* $c((x, y), (u, v)) = N_q((x, y), (u, v))^\rho$ *with* $\rho = 1$. *Then for logistic regression with a log-exponential loss function and a support vector machine with the hinge loss function, we have*

$$\inf_{\beta\in\mathbb{R}^d} \sup_{\mathrm{P}:\,\mathcal{D}_c(\mathrm{P},\mathrm{P}_n)\le\delta} \mathrm{E}_{\mathrm{P}}[\log(1+\mathrm{e}^{-Y\beta^\top X})] = \inf_{\beta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \log(1+\mathrm{e}^{-Y_i\beta^\top X_i}) + \delta\|\beta\|_p$$

and

$$\inf_{\beta\in\mathbb{R}^d} \sup_{\mathrm{P}:\,D_c(\mathrm{P},\mathrm{P}_n)\le\delta} \mathrm{E}_{\mathrm{P}}[(1-Y\beta^\top X)^+] = \frac{1}{n}\sum_{i=1}^{n} (1-Y_i\beta^\top X_i)^+ + \delta\|\beta\|_p,$$

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

The proofs of all of the results in this subsection are provided in Appendix A.1 in the supplementary material [9]. The example of logistic regression with Wasserstein-distance-based uncertainty sets has been considered in [40]. The representation for regularized logistic regression in Theorem 2 can be seen as an extension in which the approximate representation described in [40, Remark 1] is made to coincide exactly with the regularized logistic regression estimator that has been widely used in practice. The approximate representation for regularized logistic regression in [40] is based on semi-infinite linear programming duality results due to [42]. On the other hand, due to the presence of infinite transportation costs in our DRO formulation that results in the desired exact representation (see Theorem 2), we utilize a different strong duality result, [8, Theorem 1], that is specifically derived for Wasserstein DRO with general cost structures. In addition, other equivalences described for square-root LASSO and support vector machines in terms of Wasserstein DRO, as far as we know, have been reported for the first time in this paper. See [41] for additional examples.

3. The robust Wasserstein profile function

Given an estimating equation $\mathrm{E}_{\mathrm{P}_n}[h(W, \theta)] = \mathbf{0}$, the objective of this section is to study the asymptotic behavior of the associated RWP function $R_n(\theta)$. As discussed in Section 1, this analysis is key in our approach towards constructing the confidence region $\Lambda_n(\theta)$ and choosing the radius of the uncertainty set optimally.

3.1. The RWP function for estimating equations and its use in constructing confidence regions

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 [33]). We provide the definition of the RWP function for estimating $\theta_* \in \mathbb{R}^l$, which we assume satisfies

$$\mathrm{E}[h(W, \theta_*)] = \mathbf{0}, \tag{15}$$

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_* is required to 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 $\{W_1, \ldots, W_n\}$, which are assumed to be independent and identically distributed copies of W, we define the Wasserstein profile function for the estimating equation (15) as

$$R_n(\theta) := \inf\{D_c(\mathrm{P}, \mathrm{P}_n) : \mathrm{E}_{\mathrm{P}}[h(W, \theta)] = \mathbf{0}\}. \tag{16}$$

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

Recall here 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

$$c(u, w) = \|w - u\|_q^\rho, \tag{17}$$

where $\rho \in [1, \infty)$ and $q \in (1, \infty]$. 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, the asymptotic behavior of the RWP function $R_n(\theta)$ is dependent on the exponent ρ in (17), we sometimes write $R_n(\theta; \rho)$ to make this dependence explicit; whenever the context is clear, though, we drop ρ to avoid notational burden. Also, observe that the profile function defined in (6) for the linear regression example is obtained as a particular case by selecting $W = (X, Y)$ and $\beta = \theta$, and defining $h(x, y, \theta) = (y - \theta^\top x)x$.

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

$$n^{\rho/2} R_n(\theta_*; \rho) \Rightarrow \bar{R}(\rho) \tag{18}$$

for a suitably defined random variable $\bar{R}(\rho)$. Throughout this paper, the symbol '$\Rightarrow$' is used to denote convergence in distribution.

As the empirical distribution weakly converges to the underlying probability distribution from which the samples are obtained, it follows from the definition of the RWP function in (18) that $R_n(\theta; \rho) \to 0$ as $n \to \infty$ if and only if θ satisfies $\mathrm{E}[h(W, \theta)] = \mathbf{0}$; for every other θ we have that $n^{\rho/2} R_n(\theta; \rho) \to \infty$. Therefore, the result in (18) can be used to provide confidence regions around θ_* as follows: 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, $\mathrm{P}(\bar{R}(\rho) \le \eta_\alpha) = 1 - \alpha$, then the set

$$\bar{\Lambda}_n(n^{-\rho/2}\eta_\alpha) = \{\theta : R_n(\theta; \rho) \le n^{-\rho/2}\eta_\alpha\}$$

is an approximate $(1 - \alpha)$-confidence region for θ_*. This is because, by definition of $\bar{\Lambda}_n(\cdot)$,

$$\mathrm{P}(\theta_* \in \bar{\Lambda}_n(n^{-\rho/2}\eta_\alpha)) = \mathrm{P}(n^{\rho/2} R_n(\theta_*; \rho) \le \eta_\alpha) \approx \mathrm{P}(\bar{R}(\rho) \le \eta_\alpha) = 1 - \alpha. \tag{19}$$

Throughout the development in this section, the dimension m of the random vector W is kept fixed and the sample size n is sent to infinity; the function $h(\cdot)$ can be quite general.

3.2. The dual formulation of the 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

$$R_n(\theta) = \inf\{\mathrm{E}_\pi[c(U, W)] : \pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m),\ \mathrm{E}_\pi[h(U, \theta)] = \mathbf{0},\ \pi_W = \mathrm{P}_n\},$$

which is a *problem of moments* of the form

$$R_n(\theta) = \inf_{\pi \in \mathcal{P}(\mathbb{R}^m \times \mathbb{R}^m)} \left\{ \mathrm{E}_\pi[c(U, W)] : \mathrm{E}_\pi[h(U, \theta)] = \mathbf{0},\ \mathrm{E}_\pi[\mathbf{1}(W = W_i)] = \frac{1}{n},\ i \le n \right\}. \tag{20}$$

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, [23, 43]).

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

The linear program problem over the variable π in (20) admits a simple dual semi-infinite linear program of the form

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

Proposition 3 states that strong duality holds under mild assumptions, and the dual formulation above indeed equals $R_n(\theta)$.

Proposition 3. *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*

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

A proof of Proposition 3, along with an introduction to the problem of moments, is provided in Appendix B in the supplementary material [9].

3.3. Asymptotic distribution of the RWP function

In order to gain intuition for (18), let us first consider the simple example of estimating the expectation $\theta_*=\mathrm{E}[W]$ of a real-valued random variable W using $h(w,\theta)=w-\theta$.

Example 3. 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 the convex hull of support of W, Proposition (3) implies that

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

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

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

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

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 $\rho = 1$,

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

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

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

So, if indeed $\mathrm{E}[W] = \theta_*$ and $\mathrm{Var}[W] = \sigma_W^2 < \infty$, 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 3 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 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:

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

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

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

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

$$\mathrm{P}(\|\zeta^\top D_w h(W, \theta_*)\|_p > 0) > 0. \tag{21}$$

Assumptions (A1)–(A3) make precise the setting considered. Assumption (A4) is the only assumption which is technical in nature; it can be stated equivalently as

$$\mathrm{E}[D_w h(W, \theta_*) D_w h(W, \theta_*)^\top] \succ 0,$$

where $A \succ 0$ is used to denote that the matrix A is positive definite. Verification of this positive definiteness condition for linear and logistic regression problems is presented in Sections 4.2

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

and 4.3, respectively. In order to state the theorem, let us introduce the notation for the asymptotic stochastic upper bound,

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

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

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

Similarly, we write $\gtrsim_{\mathrm{D}}$ for an asymptotic stochastic lower bound:

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

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, [5]). Now we are ready to state our asymptotic upper bound.

Theorem 3. *Under Assumptions (A1) to (A4) we have, as* $n \to \infty$,

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

where, for $\rho > 1$,

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

if $\rho = 1$,

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

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

We remark that as $\rho \to 1$ we can verify that $\bar{R}(\rho) \Rightarrow \bar{R}(1)$, so formally we can simply keep in mind the expression $\bar{R}(\rho)$ with $\rho > 1$. It is interesting to note that $\bar{R}(\rho)$ resembles the Fenchel transform when viewed as a function of H. Indeed, in the case where $p = q = \rho = 2$ and $\mathrm{E}[D_w h(W,\theta_*)]$ is invertible, the expression for $\bar{R}(\rho)$ simplifies as follows:

$$\bar{R}(\rho) = \max_{\zeta\in\mathbb{R}^r}\{2\zeta^\top H - \zeta^\top \mathrm{E}[D_w h(W,\theta_*)]\zeta\} = H^\top(\mathrm{E}[D_w h(W,\theta_*)])^{-1}H. \tag{22}$$

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 4. *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 square-root LASSO.

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

$$\|D_w h(w,\theta_*)\|_p \le \kappa\|w\|_q^{\rho-1}, \tag{23}$$

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

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

(A6) (*Local 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 $\mathrm{E}[\bar{\kappa}(w)^c] < \infty$ for $c \leq \max\{2, \frac{\rho}{\rho-1}\}$.

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

Proposition 5. *If Assumptions (A1) to (A6) hold 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 3.

As we shall see in the technical development in Appendix A.3 (see supplementary material [9]) where the proofs of the above results are furnished, the dual formulation of the RWP function in Proposition 3 can be reexpressed, assuming only (A1) to (A4), as

$$n^{\rho/2} R_n(\theta_*; \rho) = \sup_{\zeta} \left\{ \zeta^\top H_n - \frac{1}{n} \sum_{k=1}^{n} \sup_{\Delta} \left\{ \int_0^1 \zeta^\top Dh(W_i + \Delta u / n^{1/2}, \theta_*) \Delta \mathrm{d}u - \|\Delta\|_q^\rho \right\} \right\}. \tag{24}$$

In order to make sure that the lower bound asymptotically matches the upper bound obtained in Theorem 3 we need to make sure that we rule out cases in which the inner supremum is infinite in (24) with positive probability in the prelimit.

In Proposition 4 we assume that W has a positive density with respect to the Lebesgue measure because in that case the condition

$$\mathrm{P}(\|\zeta^\top Dh(W, \theta_*)\|_p \leq 1) = 1$$

(which appears in the upper bound obtained in Theorem 3) implies that $\|\zeta^\top Dh(w, \theta_*)\|_p \leq 1$ almost everywhere with respect to the Lebesgue measure. Due to the appearance of the integral in the inner supremum in (24), an upper bound can be obtained for the inner supremum, which translates into a tight lower bound for $n^{\rho/2} R_n(\theta_*)$.

Moving to the case $\rho > 1$ studied in Proposition 5, condition (23) in (A5) guarantees that (for fixed W_i and n)

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

as $\|\Delta\|_q \to \infty$. Therefore, the cost term $-\|\Delta\|_q^\rho$ in (24) will ensure a finite optimum in the prelimit for large n. The condition that $\mathrm{E}\|W\|_q^\rho < \infty$ is natural because we are using an optimal transport cost $c(u, w) = \|u - w\|_q^\rho$. 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 below.

4. Using RWPI for optimal regularization

In this section we aim to utilize the limit theorems for the RWP function derived in Section 3.3 to select the radius of uncertainty, δ, in the DRO formulation (8). Then, from

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

the DRO representations derived in Section 2.4, this would imply an automatic choice of regularization parameter $\lambda = \sqrt{\delta}$ in the square-root LASSO example (following Theorem 1), or $\lambda = \delta$ in the regularized logistic regression (following Theorem 2). In the development below, we follow the logic described in Section 1 for the square-root LASSO setting.

4.1. Selection of δ and coverage properties

Throughout this section, let β_* denote the underlying linear or logistic regression model parameter from which the training samples $\{(X_i, Y_i) : i = 1, \ldots, n\}$ are obtained. Lemma 1 establishes that the infimum and the supremum in the DRO formulation (8) can be exchanged. See Appendix C [9] for a proof of Lemma 1.

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

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathrm{P} \in \mathcal{U}_\delta(\mathrm{P}_n)} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)] = \sup_{\mathrm{P} \in \mathcal{U}_\delta(\mathrm{P}_n)} \inf_{\beta \in \mathbb{R}^d} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)]. \tag{25}$$

Recall the definition of $\Lambda_n(\delta)$ in (3). As a consequence of Lemma 1, the set $\Lambda_n(\delta)$ contains the optimal solution obtained by solving the problem on the left-hand side of (25). Indeed, if this was not the case, the left-hand side of (25) would be strictly smaller than the right-hand side of (25). Recall from Section 1.1.2 that our primary criterion for choosing δ is to choose δ large enough so that $\beta_* \in \Lambda_n(\delta)$ with the desired confidence. The property that the estimator obtained by solving the DRO formulation (8) lies in $\Lambda_n(\delta)$, we believe, makes our selection of δ logically consistent with the ultimate goal of the overall estimation procedure, namely, estimating β_*.

Due to the optimality of β_*, the convexity of the loss $\ell(x, y; \cdot)$ in Examples 1 and 2, and the finiteness of $\mathrm{E}\|X\|_2^2$, we have that $\mathrm{E}[D_\beta l(X, Y; \beta_*)] = \mathbf{0}$. Consider the RWP function with estimating equation $D_\beta l(x, y; \beta) = \mathbf{0}$ given by

$$R_n(\beta) = \inf\{\mathcal{D}_c(\mathrm{P}, \mathrm{P}_n) : D_\beta \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)] = \mathbf{0}\}.$$

Then, as explained in Section 1.1.3, the events $\{R_n(\beta_*) \le \delta\}$ and $\{\beta_* \in \Lambda_n(\delta)\}$ coincide. If δ is selected so that $\delta \ge R_n(\beta_*)$, then the worst-case loss estimated by the DRO formulation (8) can be shown to form an upper bound to the empirical risk evaluated at β_*, thus controlling the bias portion of the generalization error. This is the content of Proposition 6.

Proposition 6. *In the settings of Theorems 1 and 2, if* $\delta \ge R_n(\beta_*)$ *we have*

$$\left| \mathrm{E}_{\mathrm{P}_n}[l(X, Y; \beta_*)] - \inf_{\beta} \sup_{\mathrm{P} \in \mathcal{U}_\delta(\mathrm{P}_n)} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)] \right| \le C_1 \delta + C_2(n) \mathbf{1}_{\{\rho = 2\}} \sqrt{\delta},$$

where $C_1 := (2\rho - 1)\|\beta_*\|^\rho$ *and* $C_2(n) := 2\|\beta_*\|_p \sqrt{\mathrm{E}_{\mathrm{P}_n}[l(X, Y; \beta_*)]}$.

Now, in order to guarantee that $\delta \ge R_n(\beta_*)$ (or, equivalently, $\beta^* \in \Lambda_n(\delta)$) with the desired confidence $1 - \alpha$, it is sufficient to proceed as in Section 3.1: Let η_α be the $(1 - \alpha)$ quantile of the weak limit, $\bar{R}$, resulting from $n^{\rho/2} R_n(\beta_*) \Rightarrow \bar{R}$ as derived in Section 3.3. In light of Theorems 1 and 2 we have $\rho = 2$ for Example 1 and $\rho = 1$ for Example 2. If we take $\eta \ge \eta_\alpha$,

$$\delta = n^{-\rho/2} \eta, \quad \text{and} \quad \Lambda_n(\delta) = \{\beta : R_n(\beta) \le n^{-\rho/2} \eta\}, \tag{26}$$

then $\lim_{n \to \infty} \mathrm{P}(R_n(\beta_*) > n^{-\rho/2} \eta) \le \alpha$. Then, as demonstrated in (19), we have $\lim_{n \to \infty} \mathrm{P}(\beta_* \in \Lambda_n(\delta)) \ge 1 - \alpha$. In Sections 4.2 and 4.3 we illustrate the application of this prescription by deriving upper bounds for $\bar{R}$ that are not dependent on the knowledge of β_*.

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

Theorem 4. *In the settings of Theorems 1 and 2, suppose that the samples* $\{(X_i, Y_i) : i \le n\}$ *are obtained from the distribution* P_* *and* $\mathrm{E}_{\mathrm{P}_*}\|X\|_2^2 < \infty$. *For any* $1-\alpha \in (\frac{1}{2}, 1)$, *if* δ *is chosen to be* $n^{-\rho/2}\eta$ *for some* $\eta \ge \eta_\alpha$ *then we have that*

$$\liminf_{n\to\infty} \mathrm{P}\left(\left| \inf_{\beta\in\mathbb{R}^d} \mathrm{E}_{\mathrm{P}_*}[l(X, Y; \beta)] - \inf_{\beta\in\mathbb{R}^d} \sup_{\mathrm{P}\in\mathcal{U}_\delta(\mathrm{P}_n)} \mathrm{E}_{\mathrm{P}}[l(X, Y; \beta)] \right| < \frac{C}{\sqrt{n}} \right) \ge 1 - 2\alpha$$

for some positive constant C *depending on* ρ, $\mathrm{E}_{\mathrm{P}_*}[\ell(X, Y; \beta_*)]$, *and* $\mathrm{Var}_{\mathrm{P}_*}[\ell(X, Y; \beta_*)]$.

Proofs of Propositon 6 and Theorem 4 are furnished in Appendix A.2 in the supplementary material. Explicit prescriptions for the selection of δ satisfying the conditions of Theorem 4 for the case of linear and logistic regression examples are provided in Sections 4.2 and 4.3.

In contrast to the $O(n^{-1/d})$ rate of convergence for the prescription of δ resulting from concentration inequalities for $D_c(\mathrm{P}_n, \mathrm{P}_*)$ (see, for example, [40, Theorem 2] and [28, Theorem 3.5]), Theorem 4 asserts that the DRO formulation with RWPI-based prescription for δ enjoys the optimal $O(n^{-1/2})$ rate of convergence for the optimal risk. Roughly speaking, this is because the objective of RWPI is to choose the radius δ resulting in good coverage properties for the optimal parameter β_*, which has d degrees of freedom; on the other hand, the objective behind concentration inequalities is to choose δ with good coverage properties for the data-generating probability distribution itself, which is an infinite-dimensional object. It is well known that the distance between a probability distribution and an empirical version of itself constituting n independent samples is $\Omega(n^{-1/d})$ as $n \to \infty$ (see, for example, [46]).

Coverage for the optimal risk, for the particular example of the LASSO estimator, can also be derived, for example, from the limit theorems in [24]. Once δ is chosen using the RWP function, as can be seen from the proofs of Proposition 6 and Theorem 4, the deduction of the rate of convergence and coverage turns out to be fairly intuitive and simple. This serves to illustrate the fundamental role played by the RWP function in determining the radius of the uncertainty set. A unified profile-function-based method to deduce the coverage of optimal risk for regularized estimators is entirely novel. We believe that the approach described here could serve as a template for deducing similar coverage guarantees for more general DRO formulations that are not necessarily amenable to be recast as regularized estimators.

4.2. Linear regression models with squared loss function

In this section we derive the asymptotic limiting distribution of a suitably scaled profile function corresponding to the estimating equation $\mathrm{E}[(Y - \beta^\top X)X] = \mathbf{0}$. The chosen estimating equation describes the optimality condition for the expected loss $\mathrm{E}[(Y - \beta^\top X)^2]$, and therefore the corresponding $R_n(\beta_*)$ is suitable for choosing δ as in (26) and the regularization parameter $\lambda = \sqrt{\delta}$ in Example 1.

4.2.1. *A stochastic upper bound for the RWP limit.* 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 $Y = \beta_*^\top X + e$, where the error term e has zero mean, variance σ^2, and is independent of X. Let $\Sigma = \mathrm{E}[XX^\top]$.

Theorem 5. *Consider the discrepancy measure* $\mathcal{D}_c(\cdot)$ *defined as in (7) using the cost function* $c((x, y), (u, v)) = (N_q((x, y), (u, v)))^2$ *(the function* N_q *is defined in (14)). For* $\beta \in \mathbb{R}^d$, *let*

$$R_n(\beta) = \inf\{D_c(\mathrm{P}, \mathrm{P}_n) : \mathrm{E}_{\mathrm{P}}[(Y - \beta^\top X)X] = \mathbf{0}\}.$$

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

Then, under the null hypothesis H_0,

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

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

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

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

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

In the above theorem, the relationship $L_1 \overset{\mathrm{D}}{\leq} L_2$ denotes that L_1 is stochastically dominated by L_2 in the sense that $\mathrm{P}(L_1 \geq x) \leq \mathrm{P}(L_2 \geq x)$ for all $x \in \mathbb{R}$. Note that this notation for a stochastic upper bound is different from the notation $\lesssim_{\mathrm{D}}$ introduced in Section 3.3 to denote asymptotic stochastic upper bound. A proof of Theorem 5 as an application of Theorem 3 and Proposition 5 is presented in Appendix A.4 (see supplementary material [9]).

4.2.2. *Using Theorem 5 to obtain the regularization parameter for (9).* Let $\eta_{1-\alpha}$ denote the $(1-\alpha)$ quantile of the limiting random variable L_1 in Theorem 5, or its stochastic upper bound L_2. Then, following the prescription in (26) and the DRO equivalence in Theorem 1, the regularization parameter for the ℓ_p-penalized linear regression in (9) can be chosen as follows:

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 5. Let us use $\hat{\eta}_{1-\alpha}$ to denote the estimated quantile. While L_2 is simply the norm of Z, obtaining realizations of the limit law L_1 involves solving an optimization problem for each realization of Z. If $\Sigma = \mathrm{E}[XX^\top]$ is not known, one can use a simple plug-in estimator for $\mathrm{E}[XX^\top]$ 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 the prescription for the regularization parameter obtained by using L_2 does not depend on the variance of e, thus removing the need for estimating the variance of e. This property is a key advantage of using the square-root LASSO estimator over the traditional LASSO (see [2]).

4.2.3. *On the approximation ratio* L_2/L_1 *when* $p = q = 2$. In the case where q is taken to be $q = p = 2$ in Theorem 1 (corresponding to ℓ_2 penalization as in ridge regression), it is possible to obtain an explicit expression for the limit law L_1 as follows: Under the assumptions stated in Theorem 5, we have $\mathrm{E}[(e\mathbf{1}_d - X\beta_*^\top)(e\mathbf{1}_d - X\beta_*^\top)^\top] = \sigma^2 \mathbf{1}_d + \|\beta_*\|^2 \Sigma$. Then, as in (22), we obtain $L_1 = \sigma^2 Z^\top (\sigma^2 \mathbf{1}_d + \|\beta_*\|^2 \Sigma)^{-1} Z$. Suppose that X is centered so that $\mathrm{E}[X] = \mathbf{0}$ and Σ is invertible. Then, if $\Sigma = U \Lambda U^\top$ is the eigendecomposition of Σ we have that $N = \Lambda^{-1/2} U^\top Z$ has a normal distribution with mean $\mathbf{0}$ and covariance $\mathbf{1}_d$. As a result,

$$L_1 = \sigma^2 Z^\top (\sigma^2 \mathbf{1}_d + \|\beta_*\|^2 \Sigma)^{-1} Z = \sum_{i=1}^{d} \frac{\Lambda_{ii}}{1 + \Lambda_{ii} \|\beta_*\|^2 / \sigma^2} N_i^2$$

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

and

$$\frac{\mathrm{E}[e^2]-(\mathrm{E}|e|)^2}{\mathrm{E}[e^2]}L_2=\|Z\|_2^2=\sum_{i=1}^{d}\Lambda_{ii}N_i^2.$$

If we let $c_1=1+\sigma^{-2}\|\beta_*\|^2\max_{i=1,\ldots,d}\Lambda_{ii}$ and $c_2=\mathrm{Var}[e]/\mathrm{Var}|e|$, we arrive at the relationship that $L_1\le L_2\le c_1c_2L_1$.

One could aim to achieve lower bias in estimation by working with the $(1-\alpha)$ quantile of the limit law L_1 (see Proposition 6) instead of that of the stochastic upper bound L_2. In order to do so, we propose to use any consistent estimator for β_* to be plugged into the expression for L_1 to result in an asymptotically optimal prescription for δ. The argument goes as follows: Let us write the limit law L_1 as $L_1(\beta_*)$ in order to make the dependence of the limit law L_1 on β_* explicit. As $L_1(\cdot)$ is a continuous function, if $\beta_n\to\beta_*$ in probability then we have

$$nR_n(\beta_*)-L_1(\beta_n)=(nR_n(\beta_*)-L_1(\beta_*))+(L_1(\beta_*)-L_1(\beta_n))\Rightarrow 0.$$

One could use, for example, sample average approximations (without regularization) to compute β_n. We seek to verify in future research that the estimator obtained via this plug-in approach indeed enjoys better generalization guarantees.

4.3. Logistic regression with a log-exponential loss function

In this section we apply the results in Section 3.3 to prescribe the regularization parameter for the ℓ_p-penalized logistic regression in Example 2.

4.3.1. *A stochastic upper bound for the RWP function.* 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

$$\log\left(\frac{\mathrm{P}(Y=1\mid X=x)}{1-\mathrm{P}(Y=1\mid X=x)}\right)=\beta_*^\top x$$

for predictors $X\in\mathbb{R}^d$ and corresponding responses $Y\in\{-1,1\}$; further, under the 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

$$l(x,y;\beta)=-\log p(y\mid x;\beta)=\log\left(1+\exp\left(-y\beta^\top x\right)\right).$$

If we let

$$h(x,y;\beta)=D_\beta l(x,y;\beta)=\frac{-yx}{1+\exp\left(y\beta^\top x\right)},\tag{27}$$

then the optimal β^* satisfies the first-order condition that $\mathrm{E}[h(x,y;\beta_*)]=\mathbf{0}$.

Theorem 6. *Consider the discrepancy measure $\mathcal{D}_c(\cdot)$ defined as in (7) using the cost function $c((x,y),(u,v))=N_q((x,y),(u,v))$ (the function N_q is defined in (14)). For $\beta\in\mathbb{R}^d$, let*

$$R_n(\beta)=\inf\{D_c(\mathrm{P},\mathrm{P}_n):\mathrm{E}_{\mathrm{P}}[h(x,y;\beta)]=\mathbf{0}\},$$

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

where $h(\cdot)$ is defined in (27). Then, under the null hypothesis H_0,

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

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

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

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

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

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

A proof of Theorem 5 as an application of Theorem 3 and Proposition 4 is presented in Appendix A.4 (see supplementary material [9]).

4.3.2. *Using Theorem 6 to obtain the regularization parameter for (10).* Similar to linear regression, the regularization parameter for regularized logistic regression discussed in Example 2 can be chosen by the following procedure:

1. Estimate the $(1-\alpha)$ quantile of $L_4 := \|\tilde{Z}\|_q$, where $\tilde{Z} \sim \mathcal{N}(\mathbf{0}, \mathrm{E}[XX^\top])$. 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 (10) in Example 2 to be

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

4.4. Optimal regularization in high-dimensional square-root LASSO

In this section, let us restrict our attention to the square-loss function $l(x, y; \beta) = (y - \beta^\top x)^2$ for the linear regression model and the discrepancy measure D_c defined using the cost function $c = N_q$ with $q = \infty$ in (14). Then, due to Theorem 1, this corresponds to the interesting case of square-root LASSO or ℓ_2 LASSO that was a particular example in the class of ℓ_p-norm-penalized linear regression estimators considered in Section 4.2.

As an interesting byproduct of the RWP function analysis, the following theorem presents a prescription for the regularization parameter even in high-dimensional settings where the ambient dimension d is larger than the number of samples n. Given observations $\{(X_i, Y_i) : i = 1, \ldots, n\}$ from the linear model $Y = \beta_*^\top X + e$, let $\tilde{e}_i := (Y_i - \beta_*^\top X_i)/\sigma$ for $i = 1, \ldots, n$. We have that the variance of the normalized error terms $\tilde{e}_i$ does not depend on σ.

Theorem 7. *Suppose that the assumptions imposed in Theorem 5 hold. Then*

$$nR_n(\beta_*) \overset{\mathrm{D}}{\leq} \frac{\|Z_n\|_\infty^2}{\mathrm{Var}_n|\tilde{e}|},$$

where $Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{e}_i X_i$ and $\mathrm{Var}_n|\tilde{e}| := \sum_{i=1}^n (|\tilde{e}_i| - n^{-1} \sum_{k=1}^n |\tilde{e}_i|)^2$.

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

Remark 1. Suppose that the additive error e is normally distributed and the observations $X_i = (X_{i1}, \ldots, X_{id})$ are normalized so that $n^{-1}\sum_{i=1}^{n} X_{ij}^2 = 1$ for $j = 1, \ldots, d$. Then, for any $\alpha < 1/8$, $C > 0$, and $\varepsilon > 0$, due to Lemma 1(iii) of [2], the stochastic bound in Theorem 7 simplifies as follows: Conditional on the observations $\{X_i : i = 1, \ldots, n\}$, we have

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

with probability asymptotically larger than $1 - \alpha$ as $n \to \infty$ uniformly in d such that $\log d \leq Cn^{1/2-\varepsilon}$. Here, $\Phi^{-1}(1 - \alpha)$ denotes the quantile x satisfying $\Phi(x) = 1 - \alpha$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution defined on $\mathbb{R}$. Moreover, if the additive error e is not normally distributed, then under the additional assumption that $\sup_{n\geq 1} \sup_{1\leq j\leq d} \mathrm{E}_{\mathrm{P}_n}|X_j|^a < \infty$ for some $a > 2$, we obtain from Lemma 2(iii) of [2] that

$$\sqrt{R_n(\beta_*)} \leq \frac{\mathrm{E}[e^2]}{\mathrm{E}[e^2] - (\mathrm{E}|e|)^2} \frac{\Phi^{-1}(1 - \alpha/2d)}{\sqrt{n}}$$

with probability asymptotically larger than $1 - \alpha$ as $n \to \infty$ uniformly in d such that $d \leq 0.5\alpha n^{(a-2-\varepsilon)/2}$.

A proof of Theorem 7 is presented in Appendix A.4 (see supplementary material [9]). A commonly adopted approach in the high-dimensional regression literature (see, for example, [1, 2, 4, 29] and references therein) is to start with any choice $\lambda > \|\tilde{S}\|_q$, where $\tilde{S}$ is the score function $D_\beta \mathrm{E}_n[l(X, Y; \beta_*)]$. This choice, in the context of square-root LASSO, results in the regularization parameter being chosen larger than the $(1 - \alpha)$ quantile of $n^{-1/2}\|Z\|_\infty / \sqrt{\mathrm{Var}_n[\tilde{e}]}$ (see (10) in [2]). As observed in Theorem 7, working with an upper bound of the RWP function results in choosing the $(1 - \alpha)$ quantile of $n^{-1/2}\|Z\|_\infty / \sqrt{\mathrm{Var}_n|\tilde{e}|}$. Indeed, this agreement of the regularization parameter with the high-dimensional linear regression literature strengthens the RWPI-based approach for selecting the radius of uncertainty. Since the RWPI-based approach results in a prescription for the regularization parameter that is larger (by a factor $\mathrm{Var}_n[\tilde{e}]/\mathrm{Var}_n|\tilde{e}|$), the generalization error bounds derived in the literature for high-dimensional regularized regression (see, for example, [2, Corollary 1]) hold.

The approach in Theorem 7 is to identify an upper bound that does not depend on β_*. Instead, one could choose $\delta \geq R_n(\hat{\beta}_n)$ by plugging in any consistent estimator $\hat{\beta}_n$. We identify investigating the possibility of obtaining tighter error bounds via this plugin approach as a subject of future research.

5. Numerical examples

In this section we consider two examples that compare the numerical performance of the square-root LASSO algorithm (see Example 1) when the regularization parameter λ is selected in the following two ways: (1) as described in Section 4.2 using a suitable quantile of the RWPI limiting distribution, and (2) using cross-validation. For comparison purposes we also list the performance of the respective ordinary least squares estimator. In both the examples the cross-validation-based approach iterates over a multitude of choices of λ, whereas the optimal regularization via RWPI utilizes the respective square-root LASSO algorithm only once for the prescribed value of λ. This naturally suggests potentially huge savings in computation that could be valuable in large-scale settings.

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

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

TABLE 1: Sparse linear regression for $d = 300$ predictor variables in Example 4. The training and test mean square errors of RWPI-based square-root LASSO regularization parameter selection are compared with the ordinary least squares estimator (OLS) and cross-validation-based square-root LASSO estimator (SQ-LASSO CV) for different values of the training data set size n.

				ℓ_1 loss	ℓ_2 loss
n	Method	Training error	Test error	$\|\beta - \beta_*\|_1$	$\|\beta - \beta_*\|_2$
350	RWPI	101.16(±8.11)	122.59(±6.64)	4.08(±0.69)	5.23(±0.76)
	SQ-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)
	SQ-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)
	SQ-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)
10 000	RWPI	102.12(±8.11)	105.97(±0.88)	1.13(±0.08)	1.63(±0.11)
	SQ-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)

$\mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma_{k,j} = 0.5^{|k-j|}$ and the 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 square-root 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 $N = 10\,000$. Specifically, we take the number of predictors to be $d = 300$ and 600, the number of standardized independent and identically distributed training samples to range over $n = 350$, 700, 3500, 10 000, and the desired confidence level to be 95%, that is, $1 - \alpha = 0.95$. In each instance we run the square-root LASSO algorithm using the 'flare' package proposed in [27] (available as a library in R) with the regularization parameter λ chosen as prescribed in Section 4.2.

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 the square-root LASSO algorithm with regularization parameter chosen as prescribed by cross-validation (denoted as SQ-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 $\mathrm{E}[(Y - \beta_*^\top X)^2] = \sigma^2 = 100$ is smaller than the worst-case expected loss computed by the DRO formulation (8). 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 (8) can be seen as a tight upper bound on the optimal loss $\mathrm{E}[l(X, Y; \beta_*)]$ (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 [15]), where there are 64 predictors (including 10 baseline variables and another 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 $N = 300$

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

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

				ℓ_1 loss	ℓ_2 loss
n	Method	Training error	Test error	$\|\beta - \beta_*\|_1$	$\|\beta - \beta_*\|_2$
350	RWPI	108.05(±8.38)	109.46(±4.68)	4.02(±0.71)	4.08(±0.70)
	SQ-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)
	SQ-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)
	SQ-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)
	SQ-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 3: Coverage probability of empirical worst-case expected loss in Example 4.

	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 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 square-root LASSO regularization parameter selection are compared with the ordinary least squares estimator (OLS) and the cross-validation-based square-root LASSO estimator (SQ-LASSO CV).

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

samples as test data for each experiment, in order to compute training and test mean square errors using the square-root LASSO algorithm with the regularization parameter picked as in Section 4.2. 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 the square-root LASSO algorithm with regularization parameter chosen according to cross-validation.

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

6. Conclusions

We have shown that popular machine learning estimators such as square-root LASSO, regularized logistic regression, support vector machines, etc. can be recast as particular examples of the optimal-transport-based DRO formulation in (8). We introduced a robust Wasserstein profile function and utilized its behavior at the optimal parameter β_* to present a criterion for choosing the radius, δ, in the DRO formulation (8). We illustrated how this translates to choosing regularization parameters and coverage guarantees for optimal risk in the settings of ℓ_p-norm-regularized linear and logistic regression. We observe that the proposed prescriptions for the radius δ for the DRO formulation (8) result in similar prescriptions that arise from independent considerations in the statistics literature. This indeed strengthens the Wasserstein-profile-function-based approach towards choosing the radius, δ, for the DRO formulation (8).

Following the results presented in this paper, we investigate the behavior of the profile function $R_n(\theta)$ in the vicinity of the optimal parameter θ_* in [10] and establish a limiting relationship of the form $n^{\rho/2}R_n(\theta_* + \Delta/\sqrt{n}) \Rightarrow L(\Delta)$ for a continuous $L(\cdot)$. Such a relationship can be used to accomplish the following tasks: (i) construct confidence intervals for the optimal parameter θ_*, (ii) establish error bounds for the solution to the DRO formulation (8), and (iii) systematically establish the validity of plugging in any consistent estimator for θ_* in order to obtain an asymptotically optimal prescription for the radius δ. Such a plugin approach would obviate the need to derive stochastic upper bounds on a case-by-case basis, as is presently required in Section 4.

Supplementary material

Proofs of all the results in this article are furnished in the supplementary material [9].

Acknowledgements

Support from NSF grants 1436700, 1720451, and 1820942, DARPA grant N660011824028, and Norges Bank are gratefully acknowledged by J. Blanchet.

References

[1] Banerjee, A., Chen, S., Fazayeli, F. and Sivakumar, V. (2014). Estimation with norm regularization. In *Proc. Advances in Neural Information Processing Systems 27*, Neural Information Processing Systems Foundation, pp. 1556–1564.

[2] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root LASSO: pivotal recovery of sparse signals via conic programming. *Biometrika* **98,** 791–806.

[3] Bertsimas, D. and Copenhaver, M. S. (2018). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *Europ. J. Operat. Res.* **270,** 931–942.

[4] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of LASSO and Dantzig selector. *Ann. Statist.* **37,** 1705–1732.

[5] Billingsley, P. (2013). *Convergence of Probability Measures*. John Wiley & Sons, Chichester.

[6] Blanchet, J. and Kang, Y. (2016). Sample out-of-sample inference based on Wasserstein distance. Preprint, arXiv:1605.01340.

[7] Blanchet, J. and Kang, Y. (2017). Semi-supervised learning based on distributionally robust optimization. Preprint, arXiv:1702.08848.

[8] Blanchet, J. and Murthy, K. (2016). Quantifying distributional model risk via optimal transport. Preprint, arXiv:1604.01446.

[9] Blanchet, J., Kang, Y. and Murthy, K. (2019). Robust Wasserstein profile inference and applications to machine learning. Supplementary material. Available at http://doi.org/10.1017/jpr.2019.49.

[10] Blanchet, J., Murthy, K. and Si, N. (2018). Confidence regions for optimal transport based distributionally robust optimization problems. In preparation.

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

[11] Bravo, F. (2004). Empirical likelihood based inference with applications to some econometric models. *Econometric Theory* **20,** 231–264.
[12] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35,** 2313–2351.
[13] Chen, S. X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *Ann. Statist.* **21,** 1166–1181.
[14] Duchi, J., Glynn, P. and Namkoong, H. (2016). Statistics of robust optimization: a generalized empirical likelihood approach. Preprint, arXiv:1610.03425.
[15] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32,** 407–499.
[16] Esfahani, P. and Kuhn, D. (2015). Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Preprint, arXiv:1505.05116.
[17] Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Prob. Theory Relat. Fields* **162,** 707–738.
[18] Frogner, C., Zhang, C., Mobahi, H., Araya, M. and Poggio, T. A. (2015). Learning with a Wasserstein loss. In *Proc. Advances in Neural Information Processing Systems 28*, Neural Information Processing Systems Foundation, pp. 2053–2061.
[19] Gao, R. and Kleywegt, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. Preprint, arXiv:1604.02199v1.
[20] Gotoh, J.-Y., Kim, M. J. and Lim, A. E. (2017). Calibration of distributionally robust empirical optimization models. Preprint, arXiv:1711.06565.
[21] Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *Math. Intell.* **27,** 83–85.
[22] Hjort, N. L., McKeague, I. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37,** 1079–1111.
[23] Isii, K. (1962). On sharpness of Tchebycheff-type inequalities. *Ann. Inst. Statist. Math.* **14,** 185–197.
[24] Knight, K. and Fu, W. (2000). Asymptotics for LASSO-type estimators. *Ann. Statist.* **28,** 1356–1378.
[25] Lam, H. (2016). Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. Preprint, arXiv:1605.09349.
[26] Lam, H. and Zhou, E. (2016). The empirical likelihood approach to quantifying uncertainty in sample average approximation. Preprint, arXiv:1604.02573.
[27] Li, X., Zhao, T., Yuan, X. and Liu, H. (2015). The flare package for high dimensional linear regression and precision matrix estimation in R. *J. Mach. Learn. Res.* **16,** 553–557.
[28] Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Math. Program.* **171,** 115–166.
[29] Negahban, S., Ravikumar, P., Wainwright, M. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-Estimators with decomposable regularizers. *Statist. Sci.* **27,** 538–557.
[30] Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75,** 237–249.
[31] Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120.
[32] Owen, A. (1991). Empirical likelihood for linear models. *Ann. Statist.* **19,** 1725–1747.
[33] Owen, A. (2001). *Empirical Likelihood*. CRC Press, Boca Raton, FL.
[34] Peyré, G., Cuturi, M. and Solomon, J. (2016). Gromov–Wasserstein averaging of kernel and distance matrices. In *Proc. Int. Conf. Machine Learning*, Vol. 48. International Machine Learning Society, pp. 2664–2672.
[35] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22,** 300–325.
[36] Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems. Volume II: Applications*. Springer Science & Business Media, New York.
[37] Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems. Volume I: Theory*. Springer Science & Business Media, New York.
[38] Rubner, Y., Tomasi, C. and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *Internat. J. Comput. Vision* **40,** 99–121.
[39] Seguy, V. and Cuturi, M. (2015). Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, Vol. 28. Neural Information Processing Systems Foundation, pp. 3312–3320.
[40] Shafieezadeh-Abadeh, S., Esfahani, P. and Kuhn, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, Vol. 28. Neural Information Processing Systems Foundation, pp. 1576–1584.

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press

[41] SHAFIEEZADEH-ABADEH, S., KUHN, D. AND ESFAHANI, P. M. (2017). Regularization via mass transportation. Preprint, arXiv:1710.10016.
[42] SHAPIRO, A. (2001). On duality theory of conic linear problems. In *Semi-Infinite Programming*, eds M. Á. Goberna and M. A. López, Springer, New York, pp. 135–165.
[43] SMITH, J. (1995). Generalized Chebychev inequalities: theory and applications in decision analysis. *Operat. Res.* **43,** 807–825.
[44] SOLOMON, J., RUSTAMOV, R., GUIBAS, L. AND BUTSCHER, A. (2014). Earth mover's distances on discrete surfaces. *ACM Trans. Graph.* **33,** 67:1–67:12.
[45] SRIVASTAVA, S., CEVHER, V., TRAN-DINH, Q. AND DUNSON, D. B. (2015). WASP: scalable Bayes via barycenters of subset posteriors. In *Proc. Machine Learning Research*, Vol. 38, pp. 912–920.
[46] TALAGRAND, M. (1992). Matching random samples in many dimensions. *Ann. Appl. Probab.* **2,** 846–856.
[47] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. B [Statist. Methodology]* **58,** 267–288.
[48] VILLANI, C. (2008). *Optimal Transport: Old and New*. Springer Science & Business Media, New York.
[49] WU, C. (2004). Weighted empirical likelihood inference. *Statist. Prob. Lett.* **66,** 67–79.
[50] XU, H., CARAMANIS, C. AND MANNOR, S. (2009). Robustness and regularization of support vector machines. *J. Mach. Learn. Res.* **10,** 1485–1510.
[51] XU, H., CARAMANIS, C. AND MANNOR, S. (2009). Robust regression and LASSO. In *Advances in Neural Information Processing Systems*, Vol. 21. Neural Information Processing Systems Foundation, pp. 1801–1808.
[52] ZHOU, M. (2015). *Empirical Likelihood Method in Survival Analysis*. CRC Press, Boca Raton, FL.

https://doi.org/10.1017/jpr.2019.49 Published online by Cambridge University Press