Biometrika (2022), **109**, 2, *pp.* 295–315 doi: 10.1093/biomet/asab026
Advance Access publication 22 October 2021

Confidence regions in Wasserstein distributionally robust estimation

By JOSE BLANCHET

Department of Management Science and Engineering, Stanford University, Huang Engineering Center, 475 Via Ortega, Stanford, California 94305, U.S.A.
jose.blanchet@stanford.edu

KARTHYEK MURTHY

Engineering Systems and Design pillar, Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372
karthyek_murthy@sutd.edu.sg

and NIAN SI

Department of Management Science and Engineering, Stanford University, Huang Engineering Center, 475 Via Ortega, Stanford, California 94305, U.S.A.
niansi@stanford.edu

Summary

Estimators based on Wasserstein distributionally robust optimization are obtained as solutions of min-max problems in which the statistician selects a parameter minimizing the worst-case loss among all probability models within a certain distance from the underlying empirical measure in a Wasserstein sense. While motivated by the need to identify optimal model parameters or decision choices that are robust to model misspecification, these distributionally robust estimators recover a wide range of regularized estimators, including square-root lasso and support vector machines, among others. This paper studies the asymptotic normality of these distributionally robust estimators as well as the properties of an optimal confidence region induced by the Wasserstein distributionally robust optimization formulation. In addition, key properties of min-max distributionally robust optimization problems are also studied; for example, we show that distributionally robust estimators regularize the loss based on its derivative, and we also derive general sufficient conditions which show the equivalence between the min-max distributionally robust optimization problem and the corresponding max-min formulation.

Some key words: Asymptotic normality; Confidence region; Distributionally robust optimization; Wasserstein distance.

1. Introduction

In recent years, distributionally robust optimization formulations based on Wasserstein distances have sparked a substantial amount of interest. One reason for this interest, as demonstrated by a range of examples in statistical learning and operations research, is that these formulations

provide a flexible way to quantify and hedge against the impact of model misspecification. Motivated by those applications, this paper aims to understand their fundamental statistical properties, such as asymptotic normality of the distributionally robust estimators and the associated confidence regions deemed optimal in a suitable sense to be described shortly.

Before providing a review of Wasserstein distributionally robust optimization and its connections to several areas, such as artificial intelligence, machine learning and operations research, we set the stage by first introducing the elements of a typical data-driven distributionally robust estimation problem.

Suppose that $\{X_k : 1 \leqslant k \leqslant n\} \subset \mathbb{R}^m$ are independent and identically distributed samples from an unknown distribution P_*. A typical nonrobust stochastic optimization formulation informed by P_n focuses on minimizing empirical expected loss of the form $E_{P_n}\{\ell(X;\beta)\} = n^{-1}\sum_{i=1}^n \ell(X_i;\beta)$ over the parameter choices $\beta \in B \subseteq \mathbb{R}^d$. In this paper we take B to be a closed, convex subset of $\mathbb{R}^d$. Let the empirical risk minimization estimators be

$$\beta_n^{\mathrm{ERM}} \in \arg\min_{\beta\in B} E_{P_n}\{\ell(X;\beta)\}. \tag{1}$$

On the other hand, a distributionally robust formulation recognizes the distributional uncertainty inherent in P_n being a noisy representation of an unknown distribution. Therefore, it enriches the empirical risk minimization (1) by considering an estimator of the form

$$\beta_n^{\mathrm{DRO}}(\delta) \in \arg\min_{\beta\in B} \sup_{P\in\mathcal{U}_\delta(P_n)} E_P\{\ell(X;\beta)\}, \tag{2}$$

where the set $\mathcal{U}_\delta(P_n)$ is called the distributional uncertainty region and δ is the size of the distributional uncertainty. Here, given a measurable function $f(\cdot)$, the notation $E_P\{f(X)\}$ denotes expectation with respect to a probability distribution P. Wasserstein distributionally robust formulations advocate choosing

$$\mathcal{U}_\delta(P_n) = \{P \in \mathcal{P}(\Omega) : W(P_n,P) \leqslant \delta^{1/2}\},$$

where $W(P_n,P)$ is the Wasserstein distance between distributions P_n and P defined below, and $\mathcal{P}(\Omega)$ is the set of probability distributions supported on a closed set $\Omega \subseteq \mathbb{R}^m$.

DEFINITION 1 (WASSERSTEIN DISTANCES). *Given a lower semicontinuous function $c : \Omega\times\Omega \to [0,\infty]$, the optimal transport cost $D_c(P,Q)$ between any two distributions $P, Q \in \mathcal{P}(\Omega)$ is defined as*

$$D_c(P,Q) = \min_{\pi\in\Pi(P,Q)} E_\pi\{c(X,Y)\},$$

where $\Pi(P,Q)$ denotes the set of all joint distributions of the random vector (X,Y) with marginal distributions P and Q, respectively. If we specifically take $c(x,y) = d(x,y)^2$, where $d(\cdot)$ is a metric, we obtain the Wasserstein distance of order 2 by setting $W(P,Q) = \{D_c(P,Q)\}^{1/2}$.

The quantity $W(P_n,P)$ may be interpreted as the cheapest way to transport mass from the distribution P_n to the mass of another probability distribution P while measuring the cost of transportation from location $x \in \Omega$ to location $y \in \Omega$ in terms of the squared distance between x and y. In this paper we shall work with Wasserstein distances of order 2, which explains why it is natural to use $\delta^{1/2}$ to specify the distributional uncertainty region $\mathcal{U}_\delta(P_n)$ as above. Since

$W(P_n, P_n) = 0$, the empirical risk minimizing estimator in (1) can be seen as a special case of the formulation (2) by setting $\delta = 0$.

The need for selecting model parameters or making decisions using a data-driven approach which is robust to model uncertainties has sparked a rapidly growing literature on Wasserstein distributionally robust optimization, via formulations such as (2); see, for example, Gao & Kleywegt (2016), Chen et al. (2018), Gao et al. (2018), Mohajerin Esfahani & Kuhn (2018), Zhao & Guan (2018) and Blanchet & Murthy (2019) for applications in operations research, and Yang (2017, 2020) for examples specifically in stochastic control.

In principle, the min-max formulation (2) is distributionally robust in the sense that its solution guarantees a uniform performance over all probability distributions in $\mathcal{U}_{\delta_n}(P_n)$. Roughly speaking, for every choice of parameter or decision β, the min-max game type formulation in (2) introduces an adversary that chooses the most adversarial distribution from a class of distributions $\mathcal{U}_{\delta_n}(P_n)$. The goal of the procedure is to then choose a decision that also hedges against these adversarial perturbations, thus introducing adversarial robustness into settings where the quality of optimal solutions are sensitive to incorrect model assumptions.

Interestingly, the min-max formulation (2), which is derived from the above robustness viewpoint, has been shown to recover many machine learning estimators when applied to suitable loss functions $\ell(\cdot)$; some examples include the square-root lasso and support vector machines (Blanchet et al., 2019a), the group lasso (Blanchet & Kang, 2017), adaptive regularization (Volpi et al., 2018; Blanchet et al., 2019b), among others (Shafieezadeh-Abadeh et al., 2015; Chen & Paschalidis, 2018; Duchi et al., 2020) and Gao et al., 2020. The utility of the distributionally robust formulation (2) has also been explored in adversarial training of neural networks; see, for example, Staib & Jegelka (2017) and Sinha et al. (2018).

Generic formulations such as (2) are becoming increasingly tractable; see, for example, Luo & Mehrotra (2017) and Mohajerin Esfahani & Kuhn (2018) for convex programming based approaches, and Sinha et al. (2018) and Blanchet et al. (2020) for stochastic gradient descent based iterative schemes.

Motivated by this wide range of applications, we investigate the asymptotic behaviour of the optimal value and optimal solutions of (2). In order to specifically describe the contributions, we introduce the following notation. For any positive integer n and $\delta_n > 0$, let

$$\Psi_n(\beta) = \sup_{P \in \mathcal{U}_{\delta_n}(P_n)} E_P\{\ell(X;\beta)\}$$

denote the distributionally robust objective function in (2). Suppose that β_* uniquely minimizes the population risk. According to (1) and (2), we have β_n^{DRO} and β_n^{ERM} minimize, respectively, the distributionally robust loss $\Psi_n(\beta)$ and the empirical loss in (1). Next, let

$$\Lambda_{\delta_n}(P_n) = \left\{\beta \in B : \beta \in \arg\min_{\beta \in B} E_P\{\ell(X;\beta)\} \text{ for some } P \in \mathcal{U}_{\delta_n}(P_n)\right\} \tag{3}$$

denote the set of choices of $\beta \in B$ that are compatible with the distributional uncertainty region, in the sense that for every $\beta \in \Lambda_{\delta_n}(P_n)$ there exists a probability distribution $P \in \mathcal{U}_{\delta_n}(P_n)$ for which β is optimal. In other words, if $\mathcal{U}_{\delta_n}(P_n)$ represents the set of probabilistic models which are, based on the empirical evidence, plausible representations of the underlying phenomena, then each such representation induces an optimal decision and $\Lambda_{\delta_n}(P_n)$ encodes the set of plausible decisions. Let $\Lambda^+_{\delta_n}(P_n)$ be the closure of $\cap_{\epsilon>0}\Lambda_{\delta_n+\epsilon}(P_n)$. Typically, $\Lambda^+_{\delta_n}(P_n) = \Lambda_{\delta_n}(P_n)$, but this is not always true as illustrated in Example 1. Asymptotically, as δ_n decreases to zero, the distinction is negligible. However, choosing a set such as $\Lambda^+_{\delta_n}(P_n)$ as a natural set of plausible

decisions is sensible because we guarantee that a distributionally robust solution belongs to this region. Our main result also implies that all distributionally robust solutions are asymptotically equivalent within $o_p(n^{-1/2})$ distance from each other.

With the above notation, the key contributions of this article can be described as follows. We first establish the convergence in distribution of the triplet

$$\left[n^{1/2}\{\beta_n^{\mathrm{ERM}} - \beta_*\},\ n^{\bar{\gamma}/2}\{\beta_n^{\mathrm{DRO}}(\delta_n) - \beta_*\},\ n^{1/2}\left\{\Lambda^+_{\delta_n}(P_n) - \beta_*\right\}\right] \tag{4}$$

for a suitable $\bar{\gamma} \in (0, 1/2]$ that depends on the rate at which the size of the distributional uncertainty, δ_n, is decreased to zero; see Theorem 1. We identify the joint limiting distributions of the triplet (4). The third component of the triplet in (4), $n^{1/2}\{\Lambda^+_{\delta_n}(P_n) - \beta_*\}$, considers a suitably scaled and centred version of the choices of $\beta \in B$ which are compatible with the respective distributional uncertainty region $\mathcal{U}_{\delta_n}(P_n)$ in the sense described above. Therefore, $\Lambda^+_{\delta_n}(P_n)$ is a natural choice of the confidence region. We further develop an approximation for $\Lambda^+_{\delta_n}(P_n)$.

Second, we utilize the limiting result of (4) to examine how the choice of the size of distributional ambiguity, δ_n, affects the qualitative properties of the distributionally robust estimators and the induced confidence regions. Specifically, choosing $\delta_n = \eta n^{-\gamma}$, we characterize the behaviour of the solutions for different choices of $\eta, \gamma \in (0, \infty)$ as $n \to \infty$. It emerges that the canonical, $O(n^{-1/2})$, rate of convergence is achieved only if $\gamma \leqslant 1$, and the limiting distribution corresponding to the distributionally robust estimator and that of the empirical risk minimizer are different only if $\gamma \geqslant 1$. Hence, to both obtain the canonical rate and tangible benefits from the distributionally robust optimization formulation we must choose $\gamma = 1$, which corresponds to the resulting $\bar{\gamma}$ in (4) being equal to 1. Moreover, given any $\alpha \in (0, 1)$, utilizing the limiting distribution of the triplet in (4) we are able to identify a positive constant $\eta_\alpha \in (0, +\infty)$ such that whenever $\eta \geqslant \eta_\alpha$ in the choice $\delta_n = \eta/n$, the set $\Lambda^+_{\delta_n}(P_n)$ is an asymptotic $(1 - \alpha)$-confidence region for β_*.

Finally, we establish the existence of an equilibrium game value. The distributionally robust optimization formulation assumes that the adversary selects a probability model after the statistician chooses a parameter. The equilibrium value of the game is attained if inf sup equals sup inf in (2), namely, if we allow the statistician to choose a parameter optimally after the adversary selects a probability model. We show in great generality that the equilibrium value of the game exists.

We end the introduction with a discussion of related statistical results. The asymptotic normality of M-estimators which minimize an empirical risk of the form $E_{P_n}\{\ell(X; \beta)\}$ was first established in the pioneering work of Huber (1967). Subsequent asymptotic characterizations in the presence of constraints on the choices of parameter vector β have been developed in Dupacova & Wets (1988) and Shapiro (1989, 1991, 1993, 2000), again in the standard M-estimation setting. Our work here is different because of the presence of the adversarial perturbation to the loss represented by the inner maximization in (2).

Asymptotic normality in the related context of regularized estimators for least squares regression was established in Knight & Fu (2000). As mentioned earlier, distributionally robust estimators of the form (2) recover lasso-type estimators as particular examples (Blanchet et al., 2019a). In these cases, the inner max problem involving the adversary can be solved in closed form, resulting in the presence of regularization. However, our results can be applied even in the general context in which no closed-form solution to the inner maximization can be obtained. Therefore, our results in this paper can be seen as extensions of the results by Knight & Fu (2000), from a distributionally robust optimization perspective.

We comment that some of our results involving convergence of sets may be of interest to applications in the area of empirical likelihood (Owen, 1988, 1990, 2001). This is because $\Lambda_{\delta_n}(P_n)$ can be characterized in terms of a function, namely, the robust Wasserstein profile function, which resembles the definition of the empirical likelihood profile function. We refer the reader to Blanchet et al. (2019a) for more discussion on the robust Wasserstein profile function and its connections to empirical likelihood. We also refer to Cisneros-Velarde et al. (2020) for additional applications, including graphical lasso, which could benefit from our results.

2. Preliminaries and assumptions

2.1. *Convergence of closed sets*

For a sequence $\{A_k : k \geqslant 1\}$ of closed subsets of $\mathbb{R}^d$, the inner and outer limits are defined, respectively, by

$$\begin{aligned}\mathrm{Li}_{n\to\infty} A_n &= \big\{z \in \mathbb{R}^d : \text{ there exists a sequence } (a_n)_{n\geqslant 1} \text{ with } a_n \in A_n \text{ convergent to } z\big\}, \text{ and}\\ \mathrm{Ls}_{n\to\infty} A_n &= \big\{z \in \mathbb{R}^d : \text{ there exist positive integers } n_1 < n_2 < n_3 < \cdots \text{ and } a_k \in A_{n_k}\\ &\qquad\qquad \text{such that the sequence } (a_k)_{k\geqslant 1} \text{ is convergent to } z\big\}.\end{aligned}$$

We clearly have $\mathrm{Li}_{n\to\infty} A_n \subseteq \mathrm{Ls}_{n\to\infty} A_n$. The sequence $\{A_n : n \geqslant 1\}$ is said to converge to a set A in the Painlevé–Kuratowski sense if

$$A = \mathrm{Li}_{n\to\infty} A_n = \mathrm{Ls}_{n\to\infty} A_n.$$

Since $\mathbb{R}^d$ is a locally compact Hausdorff space, the topology induced by Painlevé–Kuratowski convergence on the space of closed subsets of $\mathbb{R}^d$ is completely metrizable, separable, and coincides with the well-known topology of closed convergence, also known as Fell topology; see Molchanov (2005, Ch. 1). The notion of convergence of sets we utilize here will be the above-defined Painlevé–Kuratowski convergence. After equipping the space of closed subsets with the Borel σ-algebra, we are able to define probability measures and further define the usual weak convergence of measures; see, for example, Billingsley (2013, Ch. 1).

2.2. *Notation and assumptions*

Throughout the paper, we use $A \succ 0$ to denote that a given symmetric matrix A is positive definite, and the notation C° and $\mathrm{cl}(C)$ to denote the interior and closure of a subset C of Euclidean space, respectively. In the case of taking expectations with respect to the data-generating distribution P_*, we drop the subindex in the expectation operator as in $E_{P_*}\{f(X)\} = E\{f(X)\}$. We use $\Rightarrow$ to denote weak convergence and $\to$ to denote convergence in probability. We let $\mathbb{I}(\cdot)$ be the indicator function. Let $\|\cdot\|_p$ be the dual norm of $\|\cdot\|_q$, where $1/p + 1/q = 1$ for $q \in (1, \infty)$, and $p = \infty$ or 1 for $q = 1$ or ∞, respectively.

As mentioned in § 1, suppose that Ω is a closed subset of $\mathbb{R}^m$ and B is a closed, convex subset of $\mathbb{R}^d$. Assumptions 1 and 2 below are taken to be satisfied throughout the development, unless indicated otherwise.

Assumption 1. The transportation cost $c : \Omega \times \Omega \to [0, \infty]$ is of the form $c(u, w) = \|u - w\|_q^2$.

Assumption 2. The function $\ell : \Omega \times B \to \mathbb{R}$ satisfies the following properties:

(a) The loss function $\ell(\cdot)$ is twice continuously differentiable, and for each x, $\ell(x,\cdot)$ is convex.

(b) Let $h(x,\beta) = D_\beta \ell(x,\beta)$, and assume there exists $\beta_* \in B^\circ$ satisfying the optimality condition $E\{h(X,\beta_*)\} = 0$. In addition, $E\{\|h(X,\beta_*)\|_2^2\} < \infty$, the matrix $C = E\left\{D_\beta h(X,\beta_*)\right\} \succ 0$, $E\left\{D_x h(X,\beta_*) D_x h(X,\beta_*)^{\mathrm{T}}\right\} \succ 0$, and $\mathrm{pr}\{\|D_x \ell(X,\beta_*)\|_p > 0\} > 0$.

(c) For every $\beta \in \mathbb{R}^d$, $\|D_{xx}\ell(\cdot\,;\beta)\|_p$ is uniformly continuous and bounded by a continuous function $M(\beta)$. Further, there exists a positive constant $M' < \infty$ such that $\|D_x h(x,\beta)\|_q \leqslant M'(1 + \|x\|_q)$ for β in a neighbourhood of β_*. In addition, $D_x h(\cdot)$ and $D_\beta h(\cdot)$ satisfy the following locally Lipschitz continuity:

$$\|D_x h(x+\Delta,\beta_*+u) - D_x h(x,\beta_*)\|_q \leqslant \kappa'(x)\big(\|\Delta\|_q + \|u\|_q\big),$$
$$\big\|D_\beta h(x+\Delta,\beta_*+u) - D_\beta h(x,\beta_*)\big\|_q \leqslant \bar{\kappa}(x)\big(\|\Delta\|_q + \|u\|_q\big),$$

for $\|\Delta\|_q + \|u\|_q \leqslant 1$, where $\kappa', \bar{\kappa} : \mathbb{R}^m \to [0,\infty)$ are such that $E[\{\kappa'(X_i)\}^2] < \infty$ and $E\{\bar{\kappa}^2(X_i)\} < \infty$.

Assumption 1 covers most of the cases in the literature described in § 1. One exception that does not immediately satisfy Assumption 1, but which can be easily adapted after a simple change of variables, is the weighted l_2 norm, also known as the Mahalanobis distance, namely $c(x,y) = (x-y)^{\mathrm{T}} A (x-y)$, where $A \succ 0$; see Blanchet et al. (2020). The requirement that $\ell(\cdot)$ is twice differentiable in Assumption 2(a) is useful in the analysis to identify a second-order expansion for the objective in (2), which helps quantify the impact of adversarial perturbations. The convexity of $\ell(x,\cdot)$, together with C being positive definite in Assumption 2(b), implies the uniqueness of β_*. The uniqueness of β_* is a standard assumption in the derivation of rates of convergence for estimators; see, for example, Huber (1967) and van der Vaart & Wellner (1996, § 3.2.2). Assumption 2(b) also allows us to rule out redundancies in the underlying source of randomness, e.g., collinearity in the setting of linear regression. The first part of Assumption 2(c) ensures that the inner maximization in (2) is finite by controlling the magnitude of the adversarial perturbations. The local Lipschitz continuity requirement in x arises with the optimal transportation analysis technique in Blanchet et al. (2019a, cf. their Assumption A6). Analogous regularity in β is useful in proving the confidence region limit theorem; see the discussion following Theorem 3. Limiting results which study the impact of relaxing some of these assumptions are given immediately after describing the main result in § 3.1.

3. MAIN RESULTS

3.1. *The main limit theorem*

In order to state our main results we introduce more definitions. Define

$$\varphi(\xi) = 4^{-1} E\big[\big\|\big\{D_x h(X,\beta_*)\big\}^{\mathrm{T}}\xi\big\|_p^2\big],$$

and its convex conjugate, $\varphi^*(\zeta) = \sup_{\xi\in\mathbb{R}^d}\left\{\xi^{\mathrm{T}}\zeta - \varphi(\xi)\right\}$. In addition, define

$$S(\beta) = \left[E\left\{\|D_x \ell(X;\beta)\|_p^2\right\}\right]^{1/2}, \tag{5}$$

$$f_{\eta,\gamma}(x) = x\mathbb{I}(\gamma \geqslant 1) - \eta^{1/2} D_\beta S(\beta_*)\mathbb{I}(\gamma \leqslant 1) \tag{6}$$

for $\eta \geqslant 0$, $\gamma \geqslant 0$. By Assumption 2(b), we have that $S(\beta)$ is differentiable at β_*. Recalling the matrix $C = E\left\{D_\beta h(X, \beta_*)\right\}$ introduced in Assumption 2(b), we define

$$\Lambda^+_{\delta_n}(P_n) = \mathrm{cl}\left\{\cap_{\epsilon>0}\Lambda_{\delta_n+\epsilon}(P_n)\right\},$$

which is the right limit of $\Lambda_{\delta_n}(P_n)$ defined in (3). Finally, define the sets

$$\Lambda_\eta = \left\{u : \varphi^*(Cu) \leqslant \eta\right\}, \qquad \Lambda_{\eta,\gamma} = \begin{cases} \Lambda_\eta & \text{if } \gamma = 1, \\ \mathbb{R}^d & \text{if } \gamma < 1, \\ \{0\} & \text{if } \gamma > 1. \end{cases} \tag{7}$$

We now state our main result.

THEOREM 1. *Suppose that Assumptions 1 and 2 are satisfied with $q \in (1, \infty)$, $\Omega = \mathbb{R}^m$ and $E\left(\|X\|_2^2\right) < \infty$. If $H \sim \mathcal{N}[0, \mathrm{cov}\{h(X, \beta_*)\}]$ and $\delta_n = n^{-\gamma}\eta$ for some $\gamma, \eta \in (0, \infty)$, then we have the following joint convergence in distribution:*

$$\left[n^{1/2}\left(\beta_n^{\mathrm{ERM}} - \beta_*\right),\ n^{\bar\gamma/2}\left\{\beta_n^{\mathrm{DRO}}(\delta_n) - \beta_*\right\},\ n^{1/2}\left\{\Lambda^+_{\delta_n}(P_n) - \beta_*\right\}\right] \Rightarrow \left[C^{-1}H,\ C^{-1}f_{\eta,\gamma}(H),\ \Lambda_{\eta,\gamma} + C^{-1}H\right],$$

where $\bar\gamma = \min\{\gamma, 1\}$ and $\Lambda_{\eta,\gamma}$ is defined as in (7).

The proof of Theorem 1 is presented in § 5.3. For $q = 1$ or ∞, which corresponds to $p = \infty$ or 1, $S(\beta)$ may not be differentiable at β_*, in which case the limited distribution presents a discontinuity which makes it difficult to use in practice. Hence, we prefer not to cover this here. Theorem 1 can be used as a powerful conceptual tool. For example, let us examine how a sensible choice for the parameter δ_n can be obtained as an application of Theorem 1 by considering the following cases.

Case 1, $\gamma > 1$: If $n\delta_n \to 0$, corresponding to the case $\gamma > 1$, we have $f_{0,\gamma}(H) = H$ from the definition of the parametric family in (6). Therefore, from Theorem 1,

$$\left[n^{1/2}(\beta_n^{\mathrm{ERM}} - \beta_*),\ n^{\bar\gamma/2}\{\beta_n^{\mathrm{DRO}}(\delta_n) - \beta_*\},\ n^{1/2}\left\{\Lambda^+_{\delta_n}(P_n) - \beta_*\right\}\right] \Rightarrow \left[C^{-1}H, C^{-1}H, \{C^{-1}H\}\right],$$

which implies that the influence of the robustification vanishes in the limit when $\delta_n = o(n^{-1})$.

Case 2, $\gamma < 1$: If $n\delta_n \to \infty$, corresponding to the case $\gamma < 1$, the rate of convergence for the distributionally robust estimator is slower than the canonical $O(n^{-1/2})$ rate:

$$\beta_n^{\mathrm{DRO}}(\delta_n) = \beta_* - \eta^{1/2}n^{-\gamma/2}C^{-1}D_\beta S(\beta_*) + o_p\left(n^{-\gamma/2}\right), \tag{8}$$

where $n^{\gamma/2}o_p(n^{-\gamma/2}) \to 0$ in probability as $n \to \infty$. The relationship (8) reveals an uninteresting limit, $n^{1/2}\{\Lambda^+_{\delta_n}(P_n) - \beta_*\} \Rightarrow \mathbb{R}^d$, exposing a slower than $O(n^{-1/2})$ rate of convergence $\Lambda^+_{\delta_n}(P_n)$. In fact, (8) indicates that an $O(n^{-\gamma/2})$ scaling will result in a nondegenerate limit.

Case 3, $\gamma = 1$: when $\delta_n = \eta/n$, we have that all components in the triplet in Theorem 1 have nontrivial limits.

Theorem 2 provides a geometric insight relating $\beta_n^{\mathrm{DRO}}(\delta_n)$, β_n^{ERM} and $\Lambda^+_{\delta_n}(P_n)$, which justifies a picture describing $\Lambda^+_{\delta_n}(P_n)$ as a set containing both $\beta_n^{\mathrm{DRO}}(\delta_n)$ and β_n^{ERM}. The observation that

$\beta_n^{\mathrm{ERM}} \in \Lambda_{\delta_n}(P_n)$ is immediate because $\Lambda_\delta(P_n)$ is increasing in δ, so $\beta_n^{\mathrm{ERM}} \in \Lambda_0(P_n) \subset \Lambda_{\delta_n}^+(P_n)$. On the other hand, the observation that $\beta_n^{\mathrm{DRO}}(\delta_n) \in \Lambda_{\delta_n}^+(P_n)$ is nontrivial and it relies on the exchangeability of inf and sup in Theorem 2. An appropriate choice of η which results in the set $\Lambda_{\delta_n}^+(P_n)$ also possessing desirable coverage for β_* is prescribed in § 3.2.

THEOREM 2. *Suppose that Assumption 1 is enforced. We further assume that the loss function $\ell(\cdot)$ is continuous and nonnegative, $\ell(x,\cdot)$ is convex for each x, and $E_{P_*}\{\ell(X,\beta)\}$ has a unique optimizer $\beta_* \in B^\circ$. Then, for any $\delta > 0$,*

$$\inf_{\beta\in B} \sup_{P\in\mathcal{U}_\delta(P_n)} E_P\{\ell(X;\beta)\} = \sup_{P\in\mathcal{U}_\delta(P_n)} \inf_{\beta\in B} E_P\{\ell(X;\beta)\}, \tag{9}$$

and there exists a distributionally robust estimator choice $\beta_n^{\mathrm{DRO}}(\delta) \in \Lambda_\delta^+(P_n)$.

The proof of Theorem 2 is presented in the Supplementary Material. Example 1 demonstrates that the set of minimizers of the distributionally robust formulation (2) is not necessarily unique, and that the set $\Lambda_\delta(P_n)$ may not contain distributionally robust solutions. Theorem 2 indicates that the right limit $\Lambda_\delta^+(P_n)$ contains a distributionally robust solution. Theorem 1 implies that the minimizers of (2) differ by at most $o_p(n^{-1/2})$ in magnitude, which indicates that they are asymptotically equivalent and the inclusion of one solution of (2) in $\Lambda_\delta^+(P_n)$ is sufficient for the scaling considered.

Example 1. Let the loss function be

$$\ell(x,\beta) = f(\beta) + \{x^2 - \log(x^2+1)\}f(\beta-4),$$

where $f(\beta) = 3\beta^2/4 - 1/8\beta^4 + 3/8$ for $\beta \in [-1,1]$, and $f(\beta) = |\beta|$ otherwise. We have that $\ell(x,\beta)$ is twice differentiable and convex, satisfying Assumptions 1 and 2. Then, if the empirical measure P_n is a Dirac measure centred at zero with $n = 1$, and $\delta = 1$, we have the distributionally robust estimators $\beta_n^{\mathrm{DRO}}(\delta) \in [1,3]$. Further, $[1,3] \subset \Lambda_\delta^+(P_n)$, but $[1,3] \cap \Lambda_\delta(P_n) = \varnothing$.

Next, we turn to the relationship between β_n^{ERM} and $\beta_n^{\mathrm{DRO}}(\delta_n)$ when $\delta_n = \eta/n$. From the first two terms in the triplet, we have

$$\begin{aligned}\beta_n^{\mathrm{DRO}}(\delta_n) &= \beta_n^{\mathrm{ERM}} - \eta^{1/2}C^{-1}D_\beta S(\beta_*)n^{-1/2} + o_p\big(n^{-1/2}\big)\\ &= \beta_n^{\mathrm{ERM}} - \delta_n^{1/2}C^{-1}D_\beta S(\beta_n^{\mathrm{ERM}}) + o_p\big(\delta_n\big).\end{aligned} \tag{10}$$

The right-hand side of (10) points to the canonical $O\left(n^{-1/2}\right)$ rate of convergence of the Wasserstein distributionally robust estimator, and it can readily be used to construct confidence regions, as we shall explain in § 3.2.

The relation (10) also exposes the presence of an asymptotic bias term, namely, $S(\beta) = [E\{\|D_x\ell(X;\beta)\|_p^2\}]^{1/2}$, which points towards selection of optimizers possessing reduced sensitivity with respect to perturbations in data. A precise mathematical statement of this sensitivity-reduction property is given in Corollary 1, and its proof is in the Supplementary Material.

COROLLARY 1. *Suppose that Assumptions 1 and 2 are in force, and consider*

$$\bar\beta_n^{\mathrm{DRO}} \in \arg\min_{\beta\in B}\left(E_{P_n}\{\ell(X;\beta)\} + n^{-1/2}\left[\eta E_{P_n}\left\{\|D_x\ell(X;\beta)\|_p^2\right\}\right]^{1/2}\right). \tag{11}$$

Then, if $\delta_n = \eta/n$, $\beta_n^{\mathrm{DRO}}(\delta_n) = \bar\beta_n^{\mathrm{DRO}} + o_p(n^{-1/2})$.

While the formulation on the right-hand side of (11) is conceptually appealing, it may not be desirable from an optimization point of view due to the potentially nonconvex nature of the objective involved. On the other hand, under Assumption 2, the distributionally robust objective $\Psi_n(\beta)$ is convex; see, for example, the reasoning in Blanchet et al. (2020, Theorem 2a), while also enjoying the sensitivity-reduction property of the formulation in (11).

A result of a similar type to Corollary 1 is given in Gao et al. (2020), but the focus there is on the objective function of (2) being approximated by a suitable regularization. The difference between this type of result and Corollary 1 is that our focus is on the asymptotic equivalence of the actual optimizers. Behind a result such as Corollary 1, it is key to have a more nuanced approximation which precisely characterizes the second-order term of size $O(\delta_n)$; see the Supplementary Material.

We conclude this section with results which examine the effects of relaxing some assumptions made in the statement of Theorem 1. Proposition 1 asserts that convergence of the natural confidence region $\Lambda^+_{\delta_n}(P_n)$, as identified in Theorem 1 holds even if the support of the probability distributions in the uncertainty region $\mathcal{U}_{\delta_n}(P_n)$ is constrained to be a strict subset Ω of $\mathbb{R}^d$. For this purpose, we introduce the following notation. For any set $C \in \mathbb{R}^m$, let $C^\epsilon = \{x \in C : B_\epsilon(x) \subset C\}$, where $B_\epsilon(x)$ is the neighbourhood around x defined as $B_\epsilon(x) = \{y : \|y - x\|_2 \leqslant \epsilon\}$. Thus, for any probability measure P, we have $\lim_{\epsilon \to 0} P(C^\epsilon) = P(C^\circ)$.

PROPOSITION 1. *Suppose that Assumptions 1 and 2 are satisfied with* $q \in [1, \infty]$ *and* $E\left(\|X\|_2^2\right) < \infty$. *In addition, suppose that the data-generating measure* P_* *satisfies* $P_*(\Omega^\circ) = 1$. *If we take* $H \sim \mathcal{N}(0, \mathrm{cov}\{h(X, \beta_*)\})$ *and* $\delta_n = n^{-\gamma}\eta$ *for some* $\gamma, \eta \in (0, \infty)$, *then the following convergence holds as* $n \to \infty$: $n^{1/2}\{\Lambda_{\delta_n}(P_n) - \beta_*\} \Rightarrow \Lambda_{\eta,\gamma} + C^{-1}H$.

The steps involved in proving Proposition 1 are presented in § 5.1. A discussion on the validity of a central limit theorem for the estimator β_n^{DRO} in the presence of constraints restricting transportation within the support set Ω is presented in § 6.

In the case where the unique minimizer β_* may not necessarily lie in the interior of the set B, as opposed to the requirement in Assumption 2(b), one may obtain the extension in Proposition 2 as the limiting result for the estimator $\beta_n^{\mathrm{DRO}}(\delta_n)$. As in the previous results, we take $h(x, \beta) = D_\beta \ell(x; \beta)$. The proofs of all the subsequent propositions are given in the Supplementary Material.

PROPOSITION 2. *Suppose that Assumptions 1,* 2(a) *and* 2(c) *are satisfied, and that* β_* *is the unique minimizer of* $\min_{\beta \in B} E\{\ell(X, \beta)\}$. *Suppose that the set B is compact, and there exist* $\varepsilon > 0$ *and twice continuously differentiable functions* $g_i(\beta)$ *such that*

$$B \cap B_\varepsilon(\beta_*) = \left\{\beta \in B_\varepsilon(\beta_*) : g_i(\beta) = 0, i \in I,\ g_j(\beta) \leqslant 0, j \in J\right\},$$

where I, J *are finite index sets and* $g_i(\beta_*) = 0$ *for all* $i \in J$. *With this identification of the set B, suppose that the following so-called Mangasarian–Fromovitz constraint qualification is satisfied at* β_*: *the gradient vectors* $\{Dg_i(\beta_*) : i \in I\}$ *are linearly independent, and there exists a vector* w *such that* $w^{\mathrm{T}} Dg_i(\beta_*) = 0$ *for all* $i \in I$ *and* $w^{\mathrm{T}} Dg_j(\beta_*) < 0$ *for all* $j \in J$.

Suppose that Λ_0 *is the set of Lagrange multipliers satisfying the first-order optimality conditions and the following second-order sufficient conditions:* $\lambda \in \Lambda_0$ *if and only if* $D_\beta L(\beta_*, \lambda) = 0$, $\lambda_i \geqslant 0$ *for* $i \in J$ *and* $\max_{\lambda \in \Lambda_0} w^{\mathrm{T}} D_{\beta\beta} L(\beta_*, \lambda) w > 0$ *for all* $w \in \mathcal{C}$, *where* $L(\beta, \lambda) = E\{\ell(X, \beta)\} + \sum_{i \in I \cup J} \lambda_i g_i(\beta)$ *is the Lagrangian function associated with the minimization* $\min_{\beta \in B} E\{\ell(X, \beta)\}$, *and*

$$\mathcal{C} = \left[w : w^{\mathrm{T}} Dg_i(\beta_*) = 0, i \in I,\ w^{\mathrm{T}} Dg_j(\beta_*) \leqslant 0, j \in J,\ w^{\mathrm{T}} E\{h(X, \beta_*)\} \leqslant 0\right]$$

is the nonempty cone of critical directions. In addition, suppose that $\omega(\xi)$ *is the unique minimizer of* $\min_{u\in\mathcal{C}}\{\xi^{\mathrm{T}}u + 2^{-1}q(u)\}$, *where* $q(u) = \max\{u^{\mathrm{T}}D_{\beta\beta}L(\beta_*,\lambda)u : \lambda\in\Lambda_0\}$. *Then, if* $\delta_n = \eta n^{-1}$ *for* $\eta\in(0,\infty)$, $E\{\|h(X,\beta_*)\|_2^2\} < \infty$ *and* $E\{D_\beta h(X,\beta_*)\} \succ 0$, *we have the following convergence as* $n\to\infty$:

$$n^{1/2}\{\beta_n^{\mathrm{DRO}}(\delta_n) - \beta_*\} \Rightarrow \omega\{-H + \eta^{1/2}D_\beta S(\beta_*)\},$$

where $H\sim\mathcal{N}[0,\mathrm{cov}\{h(X,\beta_*)\}]$.

The Mangasarian–Fromovitz constraint qualification conditions and the necessary and sufficient conditions in the statement of Proposition 2 are standard in the literature if the optimal β_* lies on the boundary of the set B; see, for example, Shapiro (1989). Please refer to the discussion following Theorem 3.1 in Shapiro (1989) for sufficient conditions under which $\omega(\xi)$ is unique.

Proposition 3 extends the sensitivity reduction property in Corollary 1 to settings where the minimizer for $\min_{\beta\in B}E_{P_*}\{\ell(X;\beta)\}$ is not unique.

PROPOSITION 3. *Suppose that Assumptions* 1, 2(a) *and* 2(c) *are satisfied, the set* B *is compact and the choice of the radii* $(\delta_n : n\geqslant 1)$ *is such that* $n\delta_n\to\eta\in(0,\infty)$. *Let the set* B_* *be* $\arg\min_{\beta\in B}E_{P_*}\{\ell(X;\beta)\}$. *Then, the distributionally robust optimization objective* $\Psi_n(\beta)$ *satisfies*

$$n^{1/2}[\Psi_n(\beta) - E\{\ell(X;\beta)\}] \Rightarrow Z(\beta) + \eta^{1/2}S(\beta),$$

where $Z(\cdot)$ *is a zero-mean Gaussian process with covariance function* $\mathrm{cov}\{Z(\beta_1),Z(\beta_2)\} = \mathrm{cov}\{\ell(X,\beta_1),\ell(X,\beta_2)\}$. *The above weak convergence holds, as* $n\to\infty$, *on the space of continuous functions equipped with the uniform topology on compact sets. Consequently, if* $\arg\min_{\beta\in B_*}\{Z(\beta) + \eta^{1/2}S(\beta)\}$ *is singleton with probability one, we have, as* $n\to\infty$,

$$\beta_n^{\mathrm{DRO}}(\delta_n) \Rightarrow \arg\min_{\beta\in B_*}\{Z(\beta) + \eta^{1/2}S(\beta)\}.$$

3.2. *Construction of Wasserstein distributionally robust confidence regions*

As mentioned in § 1, for suitably chosen δ_n, the set $\Lambda_{\delta_n}^+(P_n)$ represents a natural confidence region. In particular, $\Lambda_{\delta_n}^+(P_n)$ possesses an asymptotically desired coverage, say at least level $1-\alpha$, if and only if

$$1-\alpha \leqslant \lim_{n\to\infty}\mathrm{pr}\{\beta_*\in\Lambda_{\delta_n}^+(P_n)\} = \mathrm{pr}[-C^{-1}H\in\{u:\varphi^*(Cu)\leqslant\eta\}],$$

or, equivalently, if $\eta\geqslant\eta_\alpha$, where η_α is the $(1-\alpha)$-quantile of the random variable $\varphi^*(H)$.

Recall the earlier geometric insight describing $\Lambda_{\delta_n}^+(P_n)$ as a set containing both $\beta_n^{\mathrm{DRO}}(\delta_n)$ and β_n^{ERM}, as a consequence of Theorem 2. Following this, if we let $\eta\geqslant\eta_\alpha$, we then have

$$\lim_{n\to\infty}\mathrm{pr}\{\beta_*\in\Lambda_{\delta_n}^+(P_n),\ \beta_n^{\mathrm{DRO}}\in\Lambda_{\delta_n}^+(P_n),\ \beta_n^{\mathrm{ERM}}\in\Lambda_{\delta_n}^+(P_n)\} = \lim_{n\to\infty}\mathrm{pr}\{\beta_*\in\Lambda_{\delta_n}^+(P_n)\} \geqslant 1-\alpha,$$

which presents the picture of $\Lambda_{\delta_n}^+(P_n)$ as a confidence region simultaneously containing β_*, β_n^{ERM} and $\beta_n^{\mathrm{DRO}}(\delta_n)$ with a desired level of confidence.

The function $\varphi^*(H)$ can be computed in closed form in some settings. But, typically, computing $\varphi^*(\cdot)$ may be challenging. We now describe how to obtain a consistent estimator for η_α. Define the empirical version of $\varphi(\xi)$, namely

$$\varphi_n(\xi) = \frac{1}{4} E_{P_n}\left[\left\|\{D_x h(X,\beta_*)\}^{\mathrm{T}}\xi\right\|_p^2\right] = \frac{1}{4n}\sum_{i=1}^{n}\left\|\{D_x h(X,\beta_*)\}^{\mathrm{T}}\xi\right\|_p^2,$$

and the associated empirical convex conjugate, $\varphi_n^*(\zeta) = \sup_{\xi\in\mathbb{R}^d}\left\{\xi^{\mathrm{T}}\zeta - \varphi_n(\xi)\right\}$. Proposition 4 provides a basis for computing a consistent estimator for η_α.

PROPOSITION 4. *Let Ξ_n be any consistent estimator of* $\mathrm{cov}\{h(X,\beta)\}$*, and write $\bar{\Xi}_n$ for any factorization of Ξ_n such that $\bar{\Xi}_n\bar{\Xi}_n^{\mathrm{T}} = \Xi_n$. Let Z be a d-dimensional standard Gaussian random vector independent of the sequence $(X_n : n \geqslant 1)$. Then, (i) the distribution of $\varphi^*(Z)$ is continuous, (ii) $\varphi_n^*(\cdot) \Rightarrow \varphi^*(\cdot)$ as $n \to \infty$ uniformly on compact sets, and (iii) $\varphi_n^*(\bar{\Xi}_n Z) \Rightarrow \varphi^*(H)$.*

Given the collection of samples $\{X_i\}_{i=1}^n$, we can generate independent and identically distributed copies of Z and use Monte Carlo to estimate the $(1-\alpha)$-quantile, $\eta_\alpha(n)$, of $\varphi_n^*(\bar{\Xi}_n Z)$. The previous proposition implies that $\eta_\alpha(n) = \eta_\alpha + o_p(1)$ as $n \to \infty$. This is sufficient to obtain an implementable expression for $\beta_n^{\mathrm{DRO}}\{\eta_\alpha(n)/n\}$ which is asymptotically equivalent to (10), as it differs only by an error of maginutude $o_p(n^{-1/2})$.

Next, we provide rigorous support for the approximation $\Lambda_{\delta_n}^+(P_n) \approx \beta_n^{\mathrm{ERM}} + n^{-1/2}\Lambda_\eta$, which can be used to approximate $\Lambda_{\delta_n}^+(P_n)$, providing we can estimate Λ_η.

COROLLARY 2. *Under the assumptions of Theorem 1, and with $\gamma = 1$,*

$$n^{1/2}\left\{\Lambda_{\delta_n}^+(P_n) - \beta_n^{\mathrm{ERM}}\right\} \Rightarrow \Lambda_\eta.$$

Moreover, if $\eta(n) = \eta + o(1)$ and $C_n \to C$, then $\Lambda_{\eta(n)}^n = \{u : \varphi_n^*(C_n u) \leqslant \eta(n)\} \to \Lambda_\eta$.

Proof of Corollary 2. Following directly from Theorem 1 and an application of the continuous mapping theorem,

$$\begin{aligned} n^{1/2}\left\{\Lambda_{\delta_n}^+(P_n) - \beta_n^{\mathrm{ERM}}\right\} &= n^{1/2}\left\{\Lambda_{\delta_n}^+(P_n) - \beta_*\right\} - n^{1/2}\left\{\beta_n^{\mathrm{ERM}} - \beta_*\right\} \\ &\Rightarrow \Lambda_\eta + C^{-1}H - C^{-1}H. \end{aligned}$$

The second part of the result follows from the regularity results in Proposition 4. □

The next result, as we shall explain, allows us to obtain computationally efficient approximations of the set Λ_η. A completely analogous result can be used to estimate $\Lambda_{\eta(n)}^n$, simply replacing $\varphi^*(\cdot)$, $\varphi(\cdot)$ and C by $\varphi_n^*(\cdot)$, $\varphi_n(\cdot)$ and C_n.

PROPOSITION 5. *The support function of the convex set $\Lambda_\eta = \{u : \varphi^*(Cu) \leqslant \eta\}$ is $h_{\Lambda_\eta}(v) = 2\{\eta\varphi(C^{-1}v)\}^{1/2}$, where the support function of a convex set A is defined as $h_A(x) = \sup\{x \cdot a : a \in A\}$.*

Remark 1. Proposition 5 can be used to obtain a tight envelope of the set Λ_η by evaluating an intersection of hyperplanes that enclose Λ_η. Recall from the definition of a support function

that $\Lambda_\eta = \cap_u\{v : u \cdot v \leqslant h_{\Lambda_\eta}(u)\}$. Therefore, for any $u_1, \ldots, u_m$, we have that Λ_η is contained in $\cap_{u_1,\ldots u_m}\{v : u_i \cdot v \leqslant h_{\Lambda_\eta}(u_i)\}$ and that $\Lambda^n_{\eta(n)}$ is contained in $\cap_{u_1,\ldots u_m}\{v : u_i \cdot v \leqslant h_{\Lambda^n_{\eta(n)}}(u_i)\}$.

4. NUMERICAL EXAMPLES: GEOMETRY AND COVERAGE PROBABILITIES

4.1. *Distributionally robust linear regression*

We first offer a brief introduction to the distributionally robust version of the linear regression problem considered in Blanchet et al. (2019a). Specifically, the data is generated by $Y = \beta_*^{\mathrm{T}} X + \epsilon$, where $X \in \mathbb{R}^d$ and ϵ are independent, $C = E(XX^{\mathrm{T}})$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We consider square loss $\ell(x, y; \beta) = 1/2(y - \beta^{\mathrm{T}} x)^2$ and take the cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0, \infty]$ to be

$$c\{(x, y), (u, v)\} = \begin{cases} \|x - u\|_q^2 & \text{if } y = v, \\ \infty & \text{otherwise.} \end{cases} \tag{12}$$

Then, from Blanchet et al. (2019a, Theorem 1), we have

$$\min_{\beta \in \mathbb{R}^d} \sup_{P : D_c(P, P_n) \leqslant \delta_n} E_P\left[\ell(X, Y; \beta)\right] = \frac{1}{2} \min_{\beta \in \mathbb{R}^d} \left[E_{P_n}\left\{(Y - \beta^{\mathrm{T}} X)^2\right\}^{1/2} + \delta_n^{1/2} \|\beta\|_p\right]^2, \tag{13}$$

where p satisfies $1/p + 1/q = 1$. Following Corollary 2, an approximate confidence region is $\Lambda^+_{\delta_n}(P_n) \approx n^{-1/2}\Lambda_{\eta_\alpha} + \beta_n^{\mathrm{ERM}}$, where $\Lambda_{\eta_\alpha} = \{\theta : \varphi^*(C\theta) \leqslant \eta_\alpha\}$, $\varphi(\xi) = 4^{-1}E\{\|e\xi - (\xi^{\mathrm{T}} X)\beta_*\|_p^2\}$, the constant η_α is such that $\mathrm{pr}\{\varphi^*(H) \leqslant 1 - \alpha\} = \eta_\alpha$ for $H \sim \mathcal{N}(0, C\sigma^2)$ and $\delta_n = \eta_\alpha/n$. By performing a change of variables via linear transformation in the analysis of the case $c(x, y) = \|x - y\|_2^2$, Theorem 1 can be directly adapted to the choice $c(x, y)$ being a Mahalanobis metric as in

$$c\,(x, y) = (x - y)^{\mathrm{T}} A\,(x - y), \tag{14}$$

for some matrix $A \succ 0$. The respective $\Lambda_{\eta_\alpha} = \{\theta : \varphi^*(C\theta) \leqslant \eta_\alpha\}$ is computed in terms of $\varphi(\xi) = 4^{-1}E\left\{\|\xi^{\mathrm{T}} D_x h(X, \beta_*) A^{-1/2}\|_2^2\right\}$. For the choice $c\,(x, y) = (x - y)^{\mathrm{T}} A\,(x - y)$, the relationship between distributionally robust and regularized estimators, as in (13), is

$$\min_{\beta \in \mathbb{R}^d} \sup_{P : D_c(P, P_n) \leqslant \delta_n} E_P\left\{l(X, Y; \beta)\right\} = \frac{1}{2} \min_{\beta \in \mathbb{R}^d} \left[E_{P_n}\left\{(Y - \beta^{\mathrm{T}} X)^2\right\}^{1/2} + \delta_n^{-1/2} \left\|A^{-1/2}\beta\right\|_2\right]^2.$$

See Blanchet et al. (2019b) for an account of improved out-of-sample performance resulting from Mahalanobis cost choices.

4.2. *Shape of confidence regions*

The goal of this section is to provide some numerical implementations to gain intuition about the geometry of the set Λ_η for different transportation cost choices. We use the empirical set $\Lambda^n_{\eta_\alpha} = \{\theta : \varphi_n^*(C_n\theta) \leqslant n^{-1/2}\tilde{\eta}_\alpha\}$, to approximate the desired confidence region as in Corollary 2. In the above expression, $\varphi_n(\xi) = 4^{-1}E_{P_n}\{\|e\xi - (\xi^{\mathrm{T}} X)\beta_n^{\mathrm{ERM}}\|_p^2\}$, $\eta_\alpha(n)$ is such that $\mathrm{pr}(\tilde{\varphi}_n^*(H) \leqslant 1 - \alpha) = \eta_\alpha(n)$ for $H \sim \mathcal{N}(0, C_n\sigma_n^2)$, $C_n = E_{P_n}\left[XX^{\mathrm{T}}\right]$, and $\sigma_n^2 = E_{P_n}[\{(Y - (\beta_n^{\mathrm{ERM}})^{\mathrm{T}} X)\}^2]$.

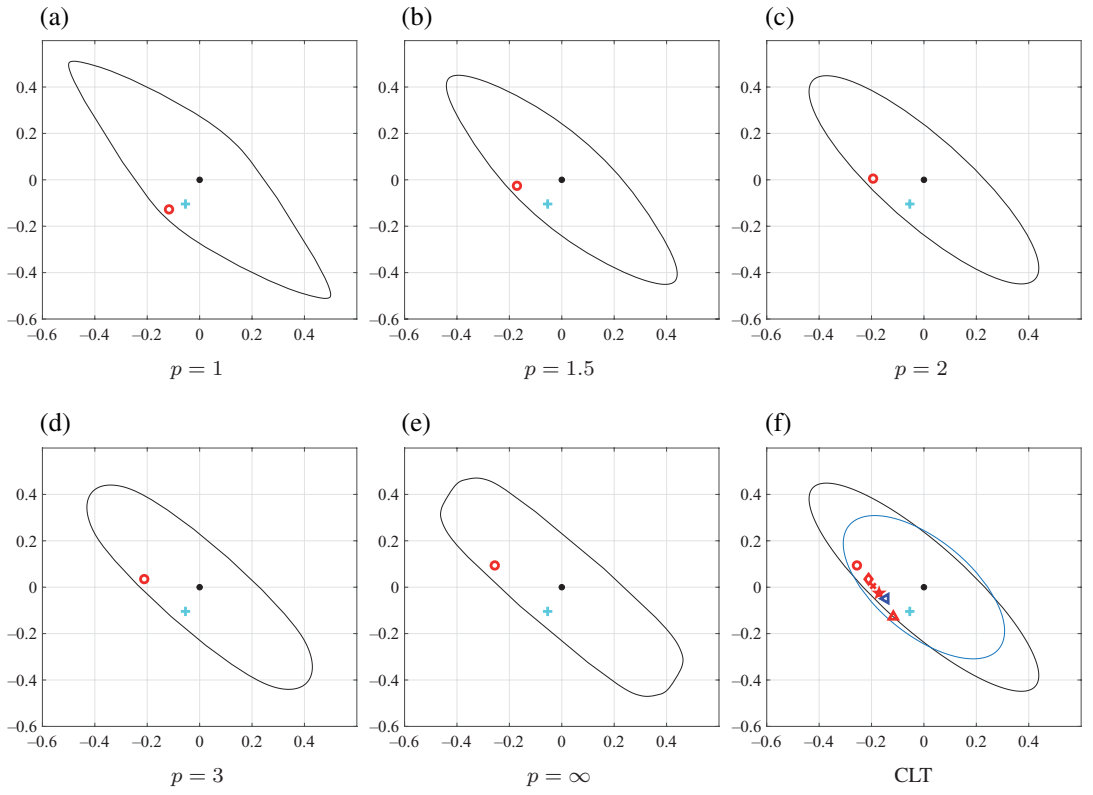

Fig. 1. Confidence regions for different norm choices and the central limit theorem based confidence region plotted together with the respective β_n^{DRO} (red circle) estimators and β_n^{ERM} (black dot). The blue cross represents the true value. In panel (f) the following estimators are included: ℓ_1 DRO (red triangle), $\ell_3/2$ DRO (red star), ℓ_2 DRO (red cross) and ℓ_3 DRO (red diamond).

In the following numerical experiments, the data is sampled from a linear regression model with parameters $\sigma^2 = 1$, $\beta_* = [0.5, 0.1]^{\mathrm{T}}$, $n = 100$ and

$$X \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \tag{15}$$

with $\rho = 0.7$. In Figs. 1(a)–1(e) we show the 95% confidence region corresponding to the choices $p = 1, 3/2, 2, 3, \infty$, $q = \infty, 3, 2, 3/2$ by means of support functions defined in Proposition 5. In addition, a confidence region for β_* resulting from the asymptotic normality of the least-squares estimator, $n^{1/2}(\beta_n^{\mathrm{ERM}} - \beta^*) \Rightarrow \mathcal{N}(0, C^{-1}\sigma^2)$, is

$$\Lambda_{\mathrm{CLT}}(P_n) = n^{-1/2}\{\theta : \theta^{\mathrm{T}} C\theta/\sigma^2 \leqslant \chi^2_{1-\alpha}(d)\} + \beta_n^{\mathrm{ERM}},$$

where $\chi^2_{1-\alpha}(d)$ is the $(1-\alpha)$-quantile of the chi-squared distribution with d degrees of freedom. One can select the matrix A in the Mahalanobis metric (14) such that the resulting confidence region coincides with $\Lambda_{\mathrm{CLT}}(P_n)$. Namely, A is chosen by solving the equation

$$E\left[\left\{e\xi - \left(\xi^{\mathrm{T}}X\right)\beta_*\right\} A^{-1} \left\{e\xi - \left(\xi^{\mathrm{T}}X\right)\beta_*\right\}^{\mathrm{T}}\right] = C\sigma^2.$$

Table 1. *Coverage probability*

β_0	ρ	ℓ_2-confidence region		CLT confidence region	
		Coverage for β_n^{DRO}	Coverage for β_*	Coverage for β_n^{DRO}	Coverage for β_*
$\begin{bmatrix}0.5\\0.5\end{bmatrix}$	0.95	100.0%	94.5%	99.4%	94.6%
	0	100.0%	94.0%	97.1%	93.5%
	−0.95	100.0%	94.8%	75.8%	94.4%
$\begin{bmatrix}1.0\\0.0\end{bmatrix}$	0.95	100.0%	94.6%	93.7%	95.4%
	0	100.0%	94.6%	100%	94.1%
	−0.95	100.0%	95.3%	91.2%	94.9%

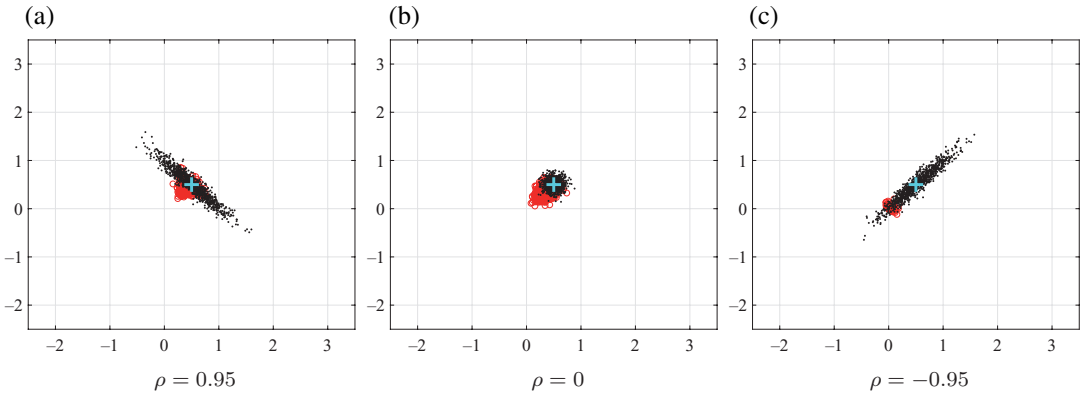

Fig. 2. Scatter plots of β_n^{ERM} (black circles) and β_n^{DRO} (red circles) for $\beta_0 = [0.5, 0.5]^{\mathrm{T}}$. The blue cross represents the true value.

Figure 1(f) gives the confidence region for the choice $p = 2$ and $\Lambda_{\mathrm{CLT}}(P_n)$ superimposed with various distributionally robust minimizers along with the empirical risk minimizer. It is evident from the figures that $p = 1$ gives a diamond shape, $p = 2$ gives an elliptical shape and $p = \infty$ gives a rectangular shape. Furthermore, we see that the distributionally robust optimization solutions all reside in their respective confidence regions, but may lie outside of the confidence regions of other norms.

We find that the induced confidence regions constructed by the Wasserstein distributionally robust optimization formulations are somewhat similar across the various l_p norms, but they are all different to the standard central limit theorem based confidence region. The Mahalanobis cost can be calibrated to exactly match the standard central limit theorem confidence region.

4.3. *Coverage probabilities and distributionally robust optimization solutions*

We now test the scenario in which the covariates are highly correlated. Specifically, the data is sampled from a linear regression model with parameters $\sigma^2 = 1$, $n = 100$, $p = 2$. The random vector X is taken to be distributed in (15), considering three different values for ρ: $\rho = 0.95, 0, -0.95$. We consider the following two cases for the underlying parameter β_*: $\beta_* = [0.5, 0.5]^{\mathrm{T}}$ and $\beta_* = [1, 0]^{\mathrm{T}}$. In Table 1 we report the coverage probabilities of the underlying β_* and $\beta_n^{\mathrm{DRO}}(\delta_n)$ in both the ℓ_2 confidence region and the central limit theorem based confidence regions. Specifically, we report the following four probabilities: $\mathrm{pr}\{\beta_n^{\mathrm{DRO}} \in \Lambda_{\delta_n}^+(P_n)\}$, $\mathrm{pr}\{\beta_* \in \Lambda_{\delta_n}^+(P_n)\}$, $\mathrm{pr}\{\beta_n^{\mathrm{DRO}} \in \Lambda_{\mathrm{CLT}}(P_n)\}$ and $\mathrm{pr}\{\beta_* \in \Lambda_{\mathrm{CLT}}(P_n)\}$.

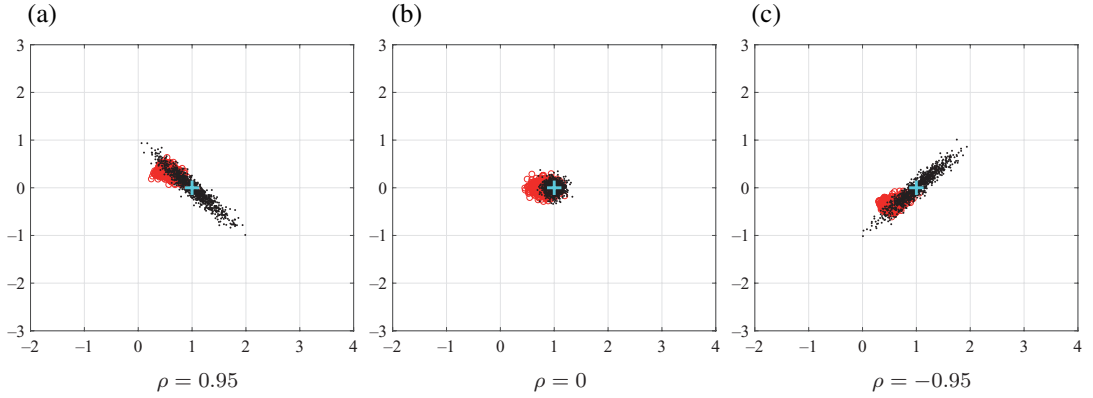

Fig. 3. Scatter plots of β_n^{ERM} (black circles) and β_n^{DRO} (red circles) for $\beta_0 = [1.0, 0.0]^{\mathrm{T}}$. The blue cross represents the true value.

Figures 2 and 3 show scatterplots of the estimators β_n^{ERM} and β_n^{DRO} when the underlying β_* takes the values $[0.5, 0.5]^{\mathrm{T}}$ and $[1, 0]^{\mathrm{T}}$, respectively. In the near-collinearity cases where $\rho = 0.95$ or -0.95, the lower spreads for the distributionally robust estimators reveal their better performance over the empirical risk-minimizing solutions. The utility of the proposed ℓ_2 confidence region emerges in light of the better performance of the distributionally robust estimator β_n^{DRO} and its aforementioned lack of membership in $\Lambda_{\mathrm{CLT}}(P_n)$.

We sample 1000 datasets and report the coverage probabilities in Table 1. We observe that for β_*, both the ℓ_2 confidence region and the central limit theorem based confidence region achieve the target 95% coverage. Furthermore, the coverage for the distributionally robust estimator of the ℓ_2 confidence region is 100%, which validates our theory. However, when $\rho = -0.95$ and $\beta_* = [0.5, 0.5]^{\mathrm{T}}$, the coverage for the distributionally robust estimator in the central limit theorem based confidence region is only 75.8%. In this example, the asymptotic results developed indicate that this coverage probability converges to zero, when n tends to infinity.

5. Proofs of main results

5.1. *Preliminaries*

Theorem 1 is obtained by considering appropriate level sets involving auxiliary functionals. Following Blanchet et al. (2019a), we define the robust Wasserstein profile function, associated with the estimation of β_* by solving $E_{P_n}\{D_\beta h(X, \beta)\} = 0$, as follows:

$$R_n(\beta) = \inf_{P \in \mathcal{P}(\Omega)} \big[D_c(P, P_n) : \beta \in \arg\min_{\beta \in B} E_P \{\ell(X; \beta)\} \big].$$

This definition, as noted in Blanchet et al. (2019a), allows us to characterize the set $\Lambda_\delta^+(P_n)$ in terms of an associated level set; in particular, we have

$$\Lambda_\delta^+(P_n) = \mathrm{cl}\{\beta : R_n(\beta) \leqslant \delta\},$$

where $\mathrm{cl}(\cdot)$ denotes closure. Indeed, this is because

$$\Lambda_\delta^+(P_n) = \mathrm{cl}\big[\cap_{\epsilon > 0} \big\{ \beta \in B : \beta \in \arg\min_{\beta \in B} E_P\{\ell(X; \beta)\} \text{ for some } P \in \mathcal{U}_{\delta_n + \epsilon}(P_n) \big\} \big].$$

If $\beta \in B^\circ$, we have $R_n(\beta) = \inf_{P \in \mathcal{P}(\Omega)}[D_c(P, P_n) : E_P\{h(X, \beta)\} = 0]$.

Next, for the sequence of radii $\delta_n = n^{-\gamma}\eta$, for some positive constants η, γ, define functions $V_n^{\mathrm{DRO}} : \mathbb{R}^d \to \mathbb{R}$ and $V_n^{\mathrm{ERM}} : \mathbb{R}^d \to \mathbb{R}$, as below, by considering suitably scaled versions of the distributionally robust and empirical risk objective functions, namely

$$V_n^{\mathrm{DRO}}(u) = n^{\bar{\gamma}}\{\Psi_n(\beta_* + n^{-\bar{\gamma}/2}u) - \Psi_n(\beta_*)\} \text{ and}$$
$$V_n^{\mathrm{ERM}}(u) = n[E_{P_n}\{\ell(X;\beta_* + n^{-1/2}u)\} - E_{P_n}\{\ell(X;\beta_*)\}],$$

where $\bar{\gamma} = \min\{\gamma, 1\}$ is defined in Theorem 1. Moreover, define $V : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ via $V(x,u) = x^{\mathrm{T}}u + 2^{-1}u^{\mathrm{T}}Cu$. The following result, as we shall see, can be used to establish Theorem 1 directly.

THEOREM 3. *Suppose that the assumptions made in Theorem 1 hold. Then,*

$$\{V_n^{\mathrm{ERM}}(\cdot),\ V_n^{\mathrm{DRO}}(\cdot),\ nR_n(\beta_* + n^{-1/2} \times \cdot)\}$$
$$\Rightarrow \{V(-H,\cdot),\ V\{-f_{\eta,\gamma}(H),\cdot\},\ \varphi^*(H - C \times \cdot)\}$$

on the space $C(\mathbb{R}^d;\mathbb{R})^3$ *equipped with the topology of uniform convergence in compact sets.*

Ensuring smoothness of $D_\beta h(x+\Delta,\beta)$ and $D_x h(x+\Delta,\beta)$ around $\beta = \beta_*$ as in Assumption 2(c) is useful towards investigating the behaviour of $nR_n(\cdot)$ in the neighbourhood of β^*, as required in the third component in the triplet in Theorem 3.

5.2. *Proof of Theorem 3*

Throughout this section we suppose that the assumptions imposed in Theorem 1 hold. Let $H_n = n^{-1/2}\sum_{i=1}^n h(X_i,\beta_*)$. The following sequence of results will be useful in proving Theorem 3 and Proposition 1. Propositions 6 and 7 hold true for $\Omega = \mathbb{R}^d$, while Propositions 8–12 hold true for general Ω under the assumption $P_*(\Omega^\circ) = 1$ in Proposition 1.

PROPOSITION 6. *Fix* $\alpha \in [0,1]$. *Given* $\varepsilon, \varepsilon', K > 0$, *there exists a positive integer* n_0 *such that*

$$\mathrm{pr}\left[\left|n^{\alpha-1}V_n^{\mathrm{ERM}}\{n^{(1-\alpha)/2}u\} - n^{\alpha/2}H_n^{\mathrm{T}}u - 2^{-1}u^{\mathrm{T}}Cu\right| \leqslant \varepsilon'\right] \geqslant 1-\varepsilon$$

for every $n > n_0$ *and* $\|u\|_2 \leqslant K$. *Specifically, if* $\alpha = 1$,

$$\mathrm{pr}\left\{\left|V_n^{\mathrm{ERM}}(u) - H_n^{\mathrm{T}}u - 2^{-1}u^{\mathrm{T}}Cu\right| \leqslant \varepsilon'\right\} \geqslant 1-\varepsilon. \tag{16}$$

PROPOSITION 7. *Given* $\varepsilon, \varepsilon', K > 0$, *there exists a positive integer* n_0 *such that*

$$\mathrm{pr}\left\{\left|V_n^{\mathrm{DRO}}(u) + f_{\eta,\gamma}(-H_n)^{\mathrm{T}}u - 2^{-1}u^{\mathrm{T}}Cu\right| \leqslant \varepsilon'\right\} \geqslant 1-\varepsilon \tag{17}$$

for every $n > n_0$ *and* $\|u\|_2 \leqslant K$.

PROPOSITION 8. *Define the set* $\Theta \subset \mathbb{R}^d$ *as* $\Theta = \{\beta \in B^\circ : 0 \in \mathrm{conv}[\{h(x,\beta) \mid x \in \Omega\}]^\circ\}$, *where* $\mathrm{conv}(S)$ *denotes the convex hull of the set* S. *For* $\beta_* + n^{-1/2}u \in \Theta$, $nR_n(\beta_* + n^{-1/2}u) =$

$\max_\xi \{-\xi^{\mathrm{T}} H_n - M_n(\xi, u)\}$, *where*

$$M_n(\xi,u) = \frac{1}{n}\sum_{i=1}^{n} \max_{\Delta: X_i + n^{-1/2}\Delta \in \Omega} \Big\{ \xi^{\mathrm{T}} \int_0^1 D_x h\big(X_i + n^{-1/2} t\Delta, \beta_* + n^{-1/2} tu\big)\Delta \,\mathrm{d}t + \xi^{\mathrm{T}} \int_0^1 D_\beta h\big(X_i + n^{-1/2} t\Delta, \beta_* + n^{-1/2} tu\big) u \,\mathrm{d}t - \|\Delta\|_q^2 \Big\}.$$

Furthermore, there exists a neighbourhood of $\beta*$, $B_\epsilon(\beta_*)$, *such that* $B_\epsilon(\beta_*) \subset \Theta$.

PROPOSITION 9. *Consider any* $\varepsilon, \varepsilon', K > 0$. *Then there exists* $b_0 \in (0, \infty)$ *such that, for any* $b \geqslant b_0$, $c_0 > 0$, $\epsilon_0 > 0$, *we have a positive integer* n_0 *such that*

$$\mathrm{pr}\left[\sup_{\|u\|_2 \leqslant K} \big\{ nR_n\big(\beta_* + n^{-1/2}u\big) - f_{\mathrm{up}}(H_n, u, b, c) \big\} \leqslant \varepsilon' \right] \geqslant 1 - \varepsilon$$

for all $n \geqslant n_0$, *and* $f_{\mathrm{up}}(H_n, u, b, c)$ *equals*

$$\max_{\|\xi\|_p \leqslant b} \big(-\xi^{\mathrm{T}} H_n - E\big[4^{-1} \|\{D_x h(X, \beta_*)\}^{\mathrm{T}} \xi\|_p^2 + \xi^{\mathrm{T}} D_\beta h(X, \beta_*) u\big] \mathbb{I}(X \in C_0^{\epsilon_0}) \big),$$

with $C_0 = \{x \in \Omega : \|x\|_p \leqslant c_0\}$.

PROPOSITION 10. *For any* $\varepsilon, \varepsilon', K, b > 0$, *there exists a positive integer* n_0 *such that*

$$\mathrm{pr}\left[\sup_{\|u\|_2 \leqslant K} \big\{ nR_n\big(\beta_* + n^{-1/2}u\big) - f_{\mathrm{low}}(H_n, u, b) \big\} \geqslant -\varepsilon' \right] \geqslant 1 - \varepsilon$$

for all $n > n_0$, *where*

$$f_{\mathrm{low}}(H_n, u, b) = \max_{\|\xi\|_p \leqslant b} \left(-\xi^{\mathrm{T}} H_n - E\left[4^{-1} \|\{D_x h(X, \beta_*)\}^{\mathrm{T}} \xi\|_p^2 + \xi^{\mathrm{T}} D_\beta h(X, \beta_*) u \right] \right).$$

PROPOSITION 11. *For any* $\varepsilon > 0$ *there exist constants* $a, n_0 > 0$ *such that, for every* $n \geqslant n_0$, $\mathrm{pr}\{nR_n(\beta_*) \leqslant a\} \geqslant 1 - \varepsilon$.

PROPOSITION 12. *For any* $\varepsilon, \varepsilon', K > 0$, *there exist positive constants* n_0, δ *such that*

$$\sup_{\substack{\|u_1 - u_2\|_2 \leqslant \delta \\ \|u_i\|_2 \leqslant K}} \big| nR_n\big(\beta_* + n^{-1/2}u_1\big) - nR_n\big(\beta_* + n^{-1/2}u_2\big) \big| \leqslant \varepsilon'$$

with probability exceeding $1 - \varepsilon$ *for every* $n > n_0$.

With the statements of these results, we proceed with the proof of Theorem 3.

Proof of Theorem 3. Since $E\{h(X, \beta_*)\} = 0$, it follows from the central limit theorem that $H_n \Rightarrow -H$, where $H \sim \mathcal{N}[0, E\{h(X, \beta_*)h(X, \beta_*)^{\mathrm{T}}\}]$. Since the inequalities (17) and (16) are associated with the same H_n, it follows from Propositions 6 and 7 that

$$V_n^{\mathrm{ERM}}(\cdot) \Rightarrow V^{\mathrm{ERM}}(\cdot) = V(-H, \cdot), \qquad V_n^{\mathrm{DRO}}(\cdot) \Rightarrow V^{\mathrm{DRO}}(\cdot) = V\{-f_{\eta,\gamma}(H), \cdot\} \tag{18}$$

jointly on the space topologized by uniform convergence on compact sets.

To prove convergence of the third component of the triplet considered in Theorem 3, observe from the definitions of $\varphi^*(\cdot)$ and C that

$$\varphi^*(H - Cu) = \max_{\xi} \left(\xi^{\mathrm{T}}[H - E\{D_\beta h(X, \beta_*)\}u] - 4^{-1}E\|\{D_x h(X, \beta_*)\}^{\mathrm{T}}\xi\|_p^2\right). \tag{19}$$

Consider any fixed $K \in (0, +\infty)$. Due to the weak convergence $H_n \Rightarrow -H$, applications of the continuous mapping theorem to the bounds in Propositions 9 and 10 result in

$$f_{\mathrm{up}}(H_n, u, b, c) \Rightarrow \max_{\|\xi\|_p \leqslant b} \left(\xi^{\mathrm{T}}H - E\left[4^{-1}\left\|\left\{D_x h(X, \beta_*)\right\}^{\mathrm{T}}\xi\right\|_p^2 + \xi^{\mathrm{T}}D_\beta h(X, \beta_*)u\right]\mathbb{I}(X \in C_0^{\epsilon_0})\right), \tag{20}$$

$$f_{\mathrm{low}}(H_n, u, b) \Rightarrow \max_{\|\xi\|_p \leqslant b} \left(\xi^{\mathrm{T}}H - E\left[4^{-1}\|\{D_x h(X, \beta_*)\}^{\mathrm{T}}\xi\|_p^2 + \xi^{\mathrm{T}}D_\beta h(X, \beta_*)u\right]\right) \tag{21}$$

for any u satisfying $\|u\|_2 \leqslant K$. Since the bounds in Propositions 9 and 10 hold for arbitrarily large choices of the constants b, c, and an arbitrarily small choice for constant ϵ_0, combining with the assumption that $P_*(\Omega^\circ) = 1$, we conclude from (19), (20) and (21) that

$$nR_n\left(\beta_* + n^{-1/2}u\right) \Rightarrow \varphi^*(H - Cu) \tag{22}$$

for any u satisfying $\|u\|_2 \leqslant K$. Finally, from Propositions 11 and 12, the collection $\{nR_n(\beta_* + n^{-1/2} \times \cdot)\}$ is tight; see, for example, Billingsley (2013, Theorem 7.4). As a consequence of this tightness and the finite-dimensional convergence in (22), we have $nR_n\left(\beta_* + n^{-1/2} \times \cdot\right) \Rightarrow \varphi^*(H - C \times \cdot)$. Combining this observation with those in (18), we obtain the desired convergence result in Theorem 3. Furthermore, since $f_{\mathrm{low}}(H_n, u, b)$ and $f_{\mathrm{up}}(H_n, u, b)$ are associated with the same H_n as inequalities (17) and (16), we have the three terms converging jointly. □

5.3. *Proof of Theorem 1*

Theorem 1 is proved by considering suitable level sets of the component functions in the triplet $\{V_n^{\mathrm{ERM}}(\cdot),\ V_n^{\mathrm{DRO}}(\cdot),\ nR_n(\beta_* + n^{-1/2} \times \cdot)\}$ considered in Theorem 3. To reduce the clutter in expressions, from here onwards we refer to the distributionally robust estimator (2) simply as β_n^{DRO}, with the dependence on the radius δ_n to be understood from the context. To begin, consider the following tightness result.

PROPOSITION 13. *The sequences* $\{\arg\min_u V_n^{\mathrm{ERM}}(u) : n \geqslant 1\}$ *and* $\{\arg\min_u V_n^{\mathrm{DRO}}(u) : n \geqslant 1\}$ *are tight.*

Observe that $V_n^{\mathrm{ERM}}(\cdot)$ and $V_n^{\mathrm{DRO}}(\cdot)$ are minimized, respectively, at $n^{1/2}(\beta_n^{\mathrm{ERM}} - \beta_*)$ and $n^{\bar{\gamma}/2}(\beta_n^{\mathrm{DRO}} - \beta_*)$. Furthermore, due to the positive definiteness of C in Assumption 2(b), we have that $V^{\mathrm{ERM}}(\cdot)$ and $V^{\mathrm{DRO}}(\cdot)$ are strongly convex with respect to u and have unique minimizers, with probability 1. Therefore, due to the tightness of the sequences $\{n^{1/2}(\beta_n^{\mathrm{ERM}} - \beta_*)\}_{n\geqslant 1}$ and $\{n^{\bar{\gamma}/2}(\beta_n^{\mathrm{DRO}} - \beta_*)\}_{n\geqslant 1}$, see Proposition 13, and the weak convergence of $V_n^{\mathrm{ERM}}(\cdot)$ and $V_n^{\mathrm{DRO}}(\cdot)$ in Theorem 3, we have the following convergences:

$$\begin{aligned} n^{1/2}(\beta_n^{\mathrm{ERM}} - \beta_*) &\Rightarrow \arg\min_u V(-H, u) = C^{-1}H, \\ n^{\bar{\gamma}/2}(\beta_n^{\mathrm{DRO}} - \beta_*) &\Rightarrow \arg\min_u V^{\mathrm{DRO}}(u) = C^{-1}f_{\eta,\gamma}(H). \end{aligned} \tag{23}$$

Finally, to prove the convergence of the sets $\Lambda^+_{\delta_n}(P_n)$, we proceed as follows. Define $G_n(u) = nR_n(\beta_* + n^{-1/2}u)$, $G(u) = \varphi^*(H - Cu)$ and $\alpha_n = n\delta_n$. For any function $f : B \to \mathbb{R}$ and $\alpha \in [0, +\infty]$, let $\mathrm{lev}(f, \alpha)$ denote the level set $\{x \in \mathbb{R}^d : f(x) \leqslant \alpha\}$.

PROPOSITION 14. *If* $\delta_n = n^{-1}\eta$, *then* $\mathrm{cl}\{\mathrm{lev}(G_n, \alpha_n)\} \Rightarrow \mathrm{lev}(G, \eta)$.

PROPOSITION 15. *If* $\delta_n = n^{-\gamma}\eta$ *for some* $\gamma > 1$, *then* $\mathrm{cl}\{\mathrm{lev}(G_n, \alpha_n)\} \Rightarrow \{C^{-1}H\}$.

PROPOSITION 16. *If* $\delta_n = n^{-\gamma}\eta$ *for some* $\gamma < 1$, *then* $\mathrm{cl}\{\mathrm{lev}(G_n, \alpha_n)\} \Rightarrow \mathbb{R}^d$.

Propositions 14–16 allow us to complete the proof of Theorem 1 as follows. From the definition of $R_n(\beta)$, $\Lambda^+_{\delta_n}(P_n) = \{\beta : R_n(\beta) \leqslant \delta_n\} = \beta_* + n^{-1/2}\{u : G_n(u) \leqslant \alpha_n\}$. From Propositions 14–16,

$$n^{1/2}\left\{\Lambda^+_{\delta_n}(P_n) - \beta_*\right\} = \{u : G_n(u) \leqslant \alpha_n\} \Rightarrow \begin{cases} \mathrm{lev}(G, \eta) & \text{if } \gamma = 1, \\ \mathbb{R}^d & \text{if } \gamma < 1, \\ \{C^{-1}H\} & \text{if } \gamma > 1. \end{cases}$$

Observe that $\varphi^*(u) = \varphi^*(-u)$. Therefore, $\mathrm{lev}(G, \eta) = \{u : \varphi^*(H - Cu) \leqslant \eta\} = C^{-1}H + \{u : \varphi^*(Cu) \leqslant \eta\}$. Since the three terms in Theorem 3 converge jointly, we have the three terms in Theorem 1 also converging jointly. This completes the proof of Theorem 1.

Proposition 1 follows by adopting exactly the same steps used to establish the convergence of $n^{1/2}\left\{\Lambda^+_{\delta_n}(P_n) - \beta_*\right\}$ in the proof of Theorem 1.

6. DISCUSSION

We discuss the subtleties in deriving a limit theorem for the distributionally robust estimator β_n^{DRO} when the support of the random vector X is a strict subset of $\mathbb{R}^m$. Suppose that the support of X is constrained to be contained in the set $\Omega = \{x \in \mathbb{R}^m : Ax \leqslant b\}$ specified in terms of linear constraints involving an $l \times m$ matrix A and $b \in \mathbb{R}^l$. For the sake of clarity, we discuss here only the nondegenerate case where $\delta_n = \eta/n$.

Considering the transportation cost $c(x, y) = \|x - y\|_2^2$ in Definition 1, we demonstrate in the Supplementary Material that the central limit theorem, $n^{1/2}\{\beta_n^{\mathrm{DRO}}(\delta_n) - \beta_*\} \Rightarrow C^{-1}H - \eta^{1/2}C^{-1}D_\beta S(\beta_*)$, continues to hold, for example, in the elementary case where the matrix A has linearly independent rows, X has a probability density which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$ and the support Ω is compact. A key element which emerges in the verification is that the fraction of samples which get transported to the boundary of the set Ω stays at $O_p(n^{-1/2})$ as $n \to \infty$.

On the other hand, when the set $\Omega = \{x \in \mathbb{R}^m : Ax \leqslant b\}$ has equality constraints, as in, for example, $\Omega = \{(x_1, x_2, \ldots, x_m) \in \mathbb{R}^2 : x_1 - x_2 = 0\}$, the bias term in the limit theorem gets affected due to the constraint binding all the samples $\{X_1, \ldots, X_n\}$, and the fraction of samples which get transported to the boundary of the set Ω is 1. This can be easily seen in the linear regression example in § 4, where $\ell(x, y; \beta) = (y - \beta^{\mathrm{T}}x)^2$ and the support is taken as $\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = x_2\}$. For this elementary example, we instead have

$$n^{1/2}\{\beta_n^{\mathrm{DRO}}(\delta_n) - \beta_*\} \Rightarrow C^{-1}H - \eta^{1/2}C^{-1}D_\beta\tilde{S}(\beta_*), \tag{24}$$

where $\tilde{S}(\beta)$ is different from the term $S(\beta)$, as in $\tilde{S}(\beta_*) = 2^{1/2-1/q}|\beta^{\mathrm{T}}1|\|\beta\|_p^{-1}S(\beta)$. Here, recall the earlier definition $S(\beta) = [E\{\|D_x\ell(X;\beta)\|_p^2\}]^{1/2}$ in (5) for the unconstrained support case. The computations required to arrive at the above conclusion are presented in the Supplementary Material. In the presence of general support constraints of the form $\Omega = \{x \in \mathbb{R}^m : Ax = b\}$, we show that (24) holds with $\tilde{S}(\beta) = \|P_{\mathcal{N}(A)}\beta\|_2$ for quadratic losses of the form $\ell(x;\beta) = a + \beta^{\mathrm{T}}x + \beta^{\mathrm{T}}C\beta$; here, A is taken to be a matrix with linearly independent rows and $P_{\mathcal{N}(A)}$ denotes the projection operator onto the null space of A. The bias term here is again different when compared to the term resulting from $S(\beta) = \|\beta\|_2$ exhibited in Theorem 1. As reasoned above, the presence of equality constraints for the support Ω introduces new challenges to be tackled in another study.

ACKNOWLEDGEMENT

This work was supported by the Air Force Office of Scientific Research (FA9550-20-1-0397), the National Science Foundation (1915967, 1820942 and 1838576) and the Ministry of Education Singapore (2018 134).

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online includes the technical proofs.

REFERENCES

BILLINGSLEY, P. (2013). *Convergence of Probability Measures*. New York: John Wiley and Sons.

BLANCHET, J. & KANG, Y. (2017). Distributionally robust groupwise regularization estimator. *Proc. Mach. Learn. Res.* **77**, 97–112.

BLANCHET, J., KANG, Y. & MURTHY, K. (2019a). Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Prob.* **56**, 830–57.

BLANCHET, J., KANG, Y., MURTHY, K. & ZHANG, F. (2019b). Data-driven optimal transport cost selection for distributionally robust optimization. In *Proc. Winter Simulation Conf. (WSC)*, pp. 3740–51.

BLANCHET, J. & MURTHY, K. (2019). Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* **44**, 565–600.

BLANCHET, J., MURTHY, K. & ZHANG, F. (2020). Optimal transport based distributionally robust optimization: Structural properties and iterative schemes. *arXiv:*1810.02403v2.

CHEN, R. & PASCHALIDIS, I. C. (2018). A robust learning approach for regression models based on distributionally robust optimization. *J. Mach. Learn. Res.* **19**, 517–64.

CHEN, Z., KUHN, D. & WIESEMANN, W. (2018). Data-driven chance constrained programs over Wasserstein balls. *arXiv:*1809.00210.

CISNEROS-VELARDE, P., PETERSEN, A. & OH, S.-Y. (2020). Distributionally robust formulation and model selection for the graphical lasso. *Proc. Mach. Learn. Res.* **108**, 756–65.

DUCHI, J. C., HASHIMOTO, T. & NAMKOONG, H. (2020). Distributionally robust losses against mixture covariate shifts. *arXiv:*2007.13982.

DUPACOVA, J. & WETS, R. (1988). Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Ann. Statist.* **16**, 1517–49.

GAO, R., CHEN, X. & KLEYWEGT, A. J. (2020). Wasserstein distributional robustness and regularization in statistical learning. *arXiv:*1712.06050v3.

GAO, R. & KLEYWEGT, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. *arXiv:*1604.02199v2.

GAO, R., XIE, L., XIE, Y. & XU, H. (2018). Robust hypothesis testing using Wasserstein uncertainty sets. In *Proc. 32nd Int. Conf. Neural Information Processing Systems*, pp. 7913–23.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1. Berkeley, CA: University of California Press.

KNIGHT, K. & FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–78.

LUO, F. & MEHROTRA, S. (2017). Decomposition algorithm for distributionally robust optimization using Wasserstein metric. *arXiv:*1704.03920.

MOHAJERIN ESFAHANI, P. & KUHN, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.* **171**, 115–66.

Molchanov, I. S. (2005). *Theory of Random Sets*, vol. 19. Berlin: Springer.
Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–49.
Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90–120.
Owen, A. B. (2001). *Empirical Likelihood*. London: Chapman and Hall/CRC.
Shafieezadeh-Abadeh, S., Esfahani, P. & Kuhn, D. (2015). Distributionally robust logistic regression. In *Proc. 28th Conf. Advances in Neural Information Processing Systems*, pp. 1576–84.
Shapiro, A. (1989). Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.* **17**, 841–58.
Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Ann. Oper. Res.* **30**, 169–86.
Shapiro, A. (1993). Asymptotic behavior of optimal solutions in stochastic programming. *Math. Oper. Res.* **18**, 829–45.
Shapiro, A. (2000). On the asymptotics of constrained local *m*-estimators. *Ann. Statist.* **28**, 948–60.
Sinha, A., Namkoong, H. & Duchi, J. (2018). Certifiable distributional robustness with principled adversarial training. In *Proc. Int. Conf. Learning Representations*.
Staib, M. & Jegelka, S. (2017). Distributionally robust deep learning as a generalization of adversarial training. In *Proc. NIPS Workshop on Machine Learning and Computer Security*.
van der Vaart, A. & Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer.
Volpi, R., Namkoong, H., Sener, O., Duchi, J., Murino, V. & Savarese, S. (2018). Generalizing to unseen domains via adversarial data augmentation. In *Proc. 32nd Int. Conf. Neural Information Processing Systems*, pp. 5339–49.
Yang, I. (2017). A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance. *IEEE Control Syst. Lett.* **1**, 164–9.
Yang, I. (2020). Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Trans. Automatic Control*, doi:10.1109/TAC.2020.3030884.
Zhao, C. & Guan, Y. (2018). Data-driven risk-averse stochastic optimization with Wasserstein metric. *Oper. Res. Lett.* **46**, 262–7.

[*Received on* 5 *June* 2019. *Editorial decision on* 19 *March* 2021]