Confidence Regions in Wasserstein Distributionally Robust Estimation

JOSE BLANCHET, KARTHYEK MURTHY, AND NIAN SI

ABSTRACT. Wasserstein distributionally robust optimization estimators 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 (in a Wasserstein sense) from the underlying empirical measure. 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, as particular cases. This paper studies the asymptotic normality of these distributionally robust estimators as well as the properties of an optimal (in a suitable sense) 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.

1. INTRODUCTION

In recent years, distributionally robust optimization formulations based on Wasserstein distances have sparked a substantial amount of interest in the literature. 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 the 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 $\{X_k : 1 \leq k \leq n\} \subset \mathbb{R}^m$ are independent and identically distributed samples from an unknown distribution P_*. A typical non-robust 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

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

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

(A1) MANAGEMENT SCIENCE AND ENGINEERING, STANFORD UNIVERSITY
(A2) ENGINEERING SYSTEMS & DESIGN, SINGAPORE UNIVERSITY OF TECHNOLOGY & DESIGN
(A3) MANAGEMENT SCIENCE AND ENGINEERING, STANFORD UNIVERSITY
E-mail addresses: jose.blanchet@stanford.edu, karthyek_murthy@sutd.edu.sg, niansi@stanford.edu.

risk minimization (1) by considering an estimator of the form,

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

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,

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

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,

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

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, Mohajerin Esfahani and Kuhn [2018], Zhao and Guan [2018], Blanchet and Murthy [2019], Gao and Kleywegt [2016], Gao et al. [2018], Chen et al. [2018] for applications in operations research and Yang [2017, 2018] 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., 2019], the group lasso [Blanchet and Kang, 2017], adaptive regularization [Volpi et al., 2018, Blanchet et al., 2019], among others [Shafieezadeh-Abadeh et al., 2015, Gao et al., 2017, Duchi et al., 2019, Chen and Paschalidis, 2018]. The utility of the distributionally robust formulation (2) has also been explored in adversarial training of Neural Networks; see, for example Sinha et al. [2018], Staib and Jegelka [2017].

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

Motivated by these 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, let us 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) - (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

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

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 of such representations 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,

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

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), namely, $n^{1/2}\{\Lambda^+_{\delta_n}(P_n) - \beta_*\}$, considers a suitably scaled and centered 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 $\Lambda^+_{\delta_n}(P_n)$; see Section 3.2.

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 $\gamma \leq 1$ and the limiting distribution corresponding to the distributionally robust estimator and that of the empirical risk minimizer are different only if $\gamma \geq 1$. 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) to be 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 $\eta \geq \eta_\alpha$ 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; see Theorem 2.

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 and Wets [1988], 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 has been established in Knight et al. [2000]. As mentioned earlier, distributionally robust estimators of the form (2) recover lasso-type estimators as particular examples [Blanchet et al., 2019]. 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 et al. [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. [2019] for more discussion on the robust Wasserstein profile function and its connections to empirical likelihood. We also refer to Cisneros-Velarde et al. [2019] for additional applications, including graphical lasso, which could benefit from our results.

2. Preliminaries and Assumptions

2.1. **Convergence of closed sets.** We begin with a brief introduction to the notion of convergence of closed sets before introducing the assumptions required to state our main results. For a sequence $\{A_k : k \geq 1\}$ of closed subsets of $\mathbb{R}^d$, the inner and outer limits are defined, respectively, by

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

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

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

in which case we write PK-$\lim_n A_n = A$. 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 metrisable, separable and coincides with the well-known topology of closed convergence, also known as Fell topology; see Molchanov [2005, Chapter 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, Chapter 1].

2.2. **Assumptions and notation.** 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 Section 1, suppose that Ω is a closed subset of $\mathbb{R}^m$ and B is a closed, convex subset of $\mathbb{R}^d$. Assumptions A1 and A2 below are taken to be satisfied throughout the development, unless indicated otherwise.

(A1) The transportation cost $c : \Omega \times \Omega \to [0, \infty]$ is of the form $c(u, w) = \|u - w\|_q^2$.

(A2) 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 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 symmetric matrix $C = E\{D_\beta h(X, \beta_*)\} \succ 0$, $E\{D_x h(X, \beta_*) D_x h(X, \beta_*)^{\mathrm{T}}\} \succ 0$, 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 $\|D_x h(x, \beta)\|_q \le M'(1 + \|x\|_q)$ for β in a neighborhood of β_*. In addition, $D_x h(\cdot)$ and $D_\beta h(\cdot)$ satisfy the following locally Lipschitz continuity:

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

for $\|\Delta\|_q + \|u\|_q \le 1$, 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 A1 covers most of the cases in the literature described in Section 1. One exception that does not immediately satisfy Assumption A1, but which can be easily adapted after a simple change-of-variables, is the weighted l_2 norm (also known as Mahalanobis distance), namely $c(x, y) = (x - y)^{\mathrm{T}} A(x - y)$, where $A \succ 0$, see Blanchet et al. [2018]. The requirement that $\ell(\cdot)$ is twice differentiable in Assumption A2.a is useful in the analysis to identify a second-order expansion for the objective in (2), which helps quantify the the impact of adversarial perturbations. Convexity of $\ell(x, \cdot)$, together with C being positive definite in A2.b, implies 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], van der Vaart et al. [1996, Section 3.2.2]. Assumption A2.b also allows us to rule out redundancies in the underlying source of randomness (e.g. colinearity in the setting of linear regression). The first part of Assumption A2.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. [2019], c.f. 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 Section 3.1 below.

3. MAIN RESULTS

3.1. **The main limit theorem.** In order to state our main results we introduce a few more definitions. Define

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

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

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

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

for $\eta \geq 0, \gamma \geq 0$. By Assumption A2.b, we have $S(\beta)$ is differentiable at β_*. Recall the matrix $C = E\{D_\beta h(X,\beta_*)\}$ introduced in Assumption A2.b and

$$\Lambda^+_{\delta_n}(P_n) = \mathrm{cl}\left\{\cap_{\epsilon>0}\Lambda_{\delta_n+\epsilon}(P_n)\right\}, \tag{7}$$

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

$$\Lambda_\eta = \{u : \varphi^*(Cu) \leq \eta\}, \qquad \Lambda_{\eta,\gamma} = \begin{cases}\Lambda_\eta & \text{if } \gamma = 1,\\ \mathbb{R}^d & \text{if } \gamma < 1,\\ \{0\} & \text{if } \gamma > 1.\end{cases} \tag{8}$$

We now state our main result.

Theorem 1. *Suppose that Assumptions A1 - A2 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:*

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

where $\bar\gamma = \min\{\gamma, 1\}$ and $\Lambda_{\eta,\gamma}$ is defined as in (8).

The proof of Theorem 1 is presented in Section 5.2. For $q = 1$ or ∞, which corresponds to $p = \infty$ or 1, $S(\beta)$ may not be differentiable at β_*, then the limiting 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, where $\gamma > 1$: 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,

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

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

Case 2, where $\gamma < 1$: 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 than $O(n^{-1/2})$ rate:

$$\beta_n^{DRO}(\delta_n) = \beta_* - \eta^{1/2}n^{-\gamma/2}C^{-1}D_\beta S(\beta_*) + o_p\big(n^{-\gamma/2}\big), \tag{9}$$

where $n^{\gamma/2}o_p(n^{-\gamma/2}) \to 0$, in probability, as $n \to \infty$. The relationship (9) 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, (9) indicates that $O(n^{-\gamma/2})$ scaling will result in a non-degenerate limit.

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

Theorem 2 below provides a geometric insight relating $\beta_n^{DRO}(\delta_n)$, β_n^{ERM} and $\Lambda_{\delta_n}^+(P_n)$, which justifies a picture describing $\Lambda_{\delta_n}^+(P_n)$ as a set containing both $\beta_n^{DRO}(\delta_n)$ and β_n^{ERM}. The observation that $\beta_n^{ERM} \in \Lambda_{\delta_n}(P_n)$ is immediate because $\Lambda_\delta(P_n)$ is increasing in δ, so $\beta_n^{ERM} \in \Lambda_0(P_n) \subset \Lambda_{\delta_n}^+(P_n)$. On the other hand, the observation that $\beta_n^{DRO}(\delta_n) \in \Lambda_{\delta_n}^+(P_n)$ is non-trivial and it relies on the exchangeability of inf and sup in Theorem 2 below. An appropriate choice of η which results in the set $\Lambda_{\delta_n}^+(P_n)$ also possessing desirable coverage for β_* is prescribed in Section 3.2.

Theorem 2. *Suppose that Assumption A1 is enforced. We further assume the loss function $\ell(\cdot)$ is continuous and non-negative, for each x, $\ell(x, \cdot)$ is convex, and $E_{P_*}\{\ell(X, \beta)\}$ has a unique optimizer $\beta_* \in B^\circ$. Then for any $\delta > 0$,*

$$\inf_{\beta \in B} \sup_{P \in \mathcal{U}_\delta(P_n)} E_P\{\ell(X; \beta)\} = \sup_{P \in \mathcal{U}_\delta(P_n)} \inf_{\beta \in B} E_P\{\ell(X; \beta)\}, \tag{10}$$

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

The proof of Theorem 2 is presented in Section D of the supplementary material. Example 1 below 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. $\ell(x, \beta)$ is twice-differentiable and convex satisfying Assumptions A1 - A2. Then, if the empirical measure P_n is a Dirac measure centered at zero with $n = 1$, and $\delta = 1$, we have the distributionally robust estimators $\beta_n^{DRO}(\delta) \in [1, 3]$. 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 $\beta_n^{DRO}(\delta_n)$, when $\delta_n = \eta/n$. From the first two terms in the triplet, we have,

$$\begin{aligned} \beta_n^{DRO}(\delta_n) &= \beta_n^{ERM} - \eta^{1/2} C^{-1} D_\beta S(\beta_*) n^{-1/2} + o_p\big(n^{-1/2}\big) \\ &= \beta_n^{ERM} - \delta_n^{1/2} C^{-1} D_\beta S(\beta_n^{ERM}) + o_p\big(\delta_n\big). \end{aligned} \tag{11}$$

The right hand side of (11) 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 Section 3.2 below.

Relation (11) 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 below and its proof is presented in Section D of the supplementary material.

Corollary 1. *Suppose that A1 - A2 are in force and consider*

$$\bar{\beta}_n^{DRO} \in \arg\min_{\beta \in B} \left[E_{P_n}\{\ell(X; \beta)\} + n^{-1/2} \left[\eta E_{P_n}\left\{ \|D_x \ell(X; \beta)\|_p^2 \right\} \right]^{1/2} \right]. \tag{12}$$

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

While the formulation on the right-hand side of (12) 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 A2, the distributionally robust objective $\Psi_n(\beta)$ is convex; see, for example, the reasoning in Blanchet et al. [2018, Theorem 2a] while also enjoying the sensitivity-reduction property of the formulation in (12).

A similar type of result to Corollary 1 is given in Gao et al. [2017], 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 Proposition A1 in the supplementary material.

We conclude this section with results which examine the effects of relaxing some assumptions made in the statement of Theorem 1 above. Proposition 1 below 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 neighborhood around x defined as $B_\epsilon(x) = \{y : \|y - x\|_2 \leq \epsilon\}$. Thus, for any probability measure P, we have $\lim_{\epsilon \to 0} P(C^\epsilon) = P(C^\circ)$.

Proposition 1. *Suppose that Assumptions A1 - A2 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 Section 5. 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 Section 6.

In the case where the unique minimizer β_* may not necessarily lie in interior of the set B (as opposed to the requirement in Assumption A2.b, one may obtain the extension in Proposition 2 as the limiting result for the estimator $\beta_n^{DRO}(\delta_n)$. As in the previous results, we take $h(x, \beta) = D_\beta \ell(x; \beta)$. The proof of Proposition 2 is in Section A.2 of the supplementary material.

Proposition 2. *Suppose that Assumptions A1, A2.a, A2.c are satisfied and 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,*

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

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 $D_\beta L(\beta_*, \lambda) = 0$, $\lambda_i \geq 0$ 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*

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

is the non-empty cone of critical directions. In addition, suppose that $\omega(\xi)$ *is the unique minimizer of* $\min_{u \in \mathcal{C}} \left\{\xi^{\mathrm{T}} u + 2^{-1} q(u)\right\}$, *where* $q(u) = \max\left\{u^{\mathrm{T}} D_{\beta\beta} L(\beta_*, \lambda) u : \lambda \in \Lambda_0\right\}$. *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* $n \to \infty$*:*

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

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 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. The proof of Proposition 3 is presented in Section A.2 of the supplementary material.

Proposition 3. *Suppose that Assumptions A1, A2.a and A2.c are satisfied, the set* B *is compact, and the choice of the radii* $(\delta_n : n \geq 1)$ *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,*

$$n^{1/2}\left[\Psi_n(\beta) - E\{\ell(X;\beta)\}\right] \Rightarrow Z(\beta) + \eta^{1/2} S(\beta), \tag{13}$$

where $Z(\cdot)$ *is a zero mean Gaussian process with covariance function* $cov\{Z(\beta_1), Z(\beta_2)\} = cov\{\ell(X,\beta_1), \ell(X,\beta_2)\}$. *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* $n \to \infty$,

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

3.2. **Construction of Wasserstein distributionally robust confidence regions.** As mentioned in the Introduction, 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 level at least $1-\alpha$, if and only if

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

or, equivalently, if $\eta \geq \eta_\alpha$, 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 $\beta_n^{DRO}(\delta_n)$ and β_n^{ERM}, as a consequence of Theorem 2. Following this, if we let $\eta \geq \eta_\alpha$, we then have,

$$\begin{aligned}\lim_{n\to\infty} \mathrm{pr}\left\{\beta_* \in \Lambda^+_{\delta_n}(P_n),\ \beta_n^{DRO} \in \Lambda^+_{\delta_n}(P_n),\ \beta_n^{ERM} \in \Lambda^+_{\delta_n}(P_n)\right\} &= \lim_{n\to\infty} \mathrm{pr}\left\{\beta_* \in \Lambda^+_{\delta_n}(P_n)\right\} \\ &\geq 1-\alpha,\end{aligned}$$

which presents the picture of $\Lambda^+_{\delta_n}(P_n)$ as a confidence region simultaneously containing β_*, β_n^{ERM} and $\beta_n^{DRO}(\delta_n)$, 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 below, whose proof is in Section E of the supplementary material, provides a basis for computing a consistent estimator for η_α.

Proposition 4. *Let Ξ_n be any consistent estimator of 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 $(X_n : n \geq 1)$. 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 quantile $(1-\alpha)$-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 $\beta_n^{DRO}\{\eta_\alpha(n)/n\}$ which is asymptotically equivalent to (11), as it differs only by an error of maginutude $o_p(n^{-1/2})$.

Next, we provide rigorous support for the approximation

$$\Lambda_{\delta_n}^+(P_n) \approx \beta_n^{ERM} + n^{-1/2}\Lambda_\eta,$$

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 $\gamma = 1$, we have*

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

Moreover, if $\eta(n) = \eta + o(1)$, and $C_n \to C$, then

$$\Lambda_{\eta(n)}^n = \{u : \varphi_n^*(C_n u) \leq \eta(n)\} \to \Lambda_\eta.$$

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

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

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 $\Lambda_\eta = \{u : \varphi^*(Cu) \leq \eta\}$ 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 $h_A(x) = \sup\{x\cdot a : a \in A\}$.

The proof of Proposition 5 is in Section E of the supplementary material.

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 support function that

$$\Lambda_\eta = \cap_u \{v : u \cdot v \leq h_{\Lambda_\eta}(u)\}.$$

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

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. [2019]. 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

$$c\{(x, y), (u, v)\} = \begin{cases} \|x - u\|_q^2 & \text{if } y = v, \\ \infty & \text{otherwise.} \end{cases} \tag{14}$$

Then, from Blanchet et al. [2019, Theorem 1], we have

$$\min_{\beta \in \mathbb{R}^d} \sup_{P: D_c(P, P_n) \leq \delta_n} E_P[\ell(X, Y; \beta)] = \frac{1}{2} \min_{\beta \in \mathbb{R}^d} \left[E_{P_n}\left\{(Y - \beta^{\mathrm{T}} X)^2\right\}^{1/2} + \delta_n^{1/2} \|\beta\|_p \right]^2, \tag{15}$$

where p satisfies $1/p + 1/q = 1$. Following Corollary 2, an approximate confidence region is

$$\Lambda^+_{\delta_n}(P_n) \approx n^{-1/2} \Lambda_{\eta_\alpha} + \beta_n^{ERM},$$

where $\Lambda_{\eta_\alpha} = \{\theta : \varphi^*(C\theta) \leq \eta_\alpha\}$, $\varphi(\xi) = 4^{-1} E\{\|e\xi - (\xi^{\mathrm{T}} X)\beta_*\|_p^2\}$, the constant η_α is such that $\mathrm{pr}\{\varphi^*(H) \leq 1 - \alpha\} = \eta_\alpha$ 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,

$$c(x, y) = (x - y)^{\mathrm{T}} A (x - y), \tag{16}$$

for some matrix $A \succ 0$. The respective $\Lambda_{\eta_\alpha} = \{\theta : \varphi^*(C\theta) \leq \eta_\alpha\}$ is computed in terms of

$$\varphi(\xi) = 4^{-1} E\{\|\xi^{\mathrm{T}} D_x h(X, \beta_*) A^{-1/2}\|_2^2\}.$$

For the choice $c(x, y) = (x - y)^{\mathrm{T}} A (x - y)$, the relationship between distributionally robust and regularized estimators, as in (15), is

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

See Blanchet et al. [2019] 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

$$\Lambda^n_{\eta_\alpha} = \{\theta : \varphi_n^*(C_n\theta) \leq n^{-1/2} \tilde{\eta}_\alpha\},$$

to approximate the desired confidence region as in Corollary 2. In the above expression, $\varphi_n(\xi) = 4^{-1} E_{P_n}\{\|e\xi - (\xi^{\mathrm{T}} X)\beta_n^{ERM}\|_p^2\}$, $\eta_\alpha(n)$ is such that $\mathrm{pr}(\tilde{\varphi}_n^*(H) \leq 1 - \alpha) = \eta_\alpha(n)$ for $H \sim \mathcal{N}(0, C_n \sigma_n^2)$, $C_n = E_{P_n}[XX^{\mathrm{T}}]$, and $\sigma_n^2 = E_{P_n}[\{(Y - (\beta_n^{ERM})^{\mathrm{T}} X\}^2]$.

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

$$X \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \tag{17}$$

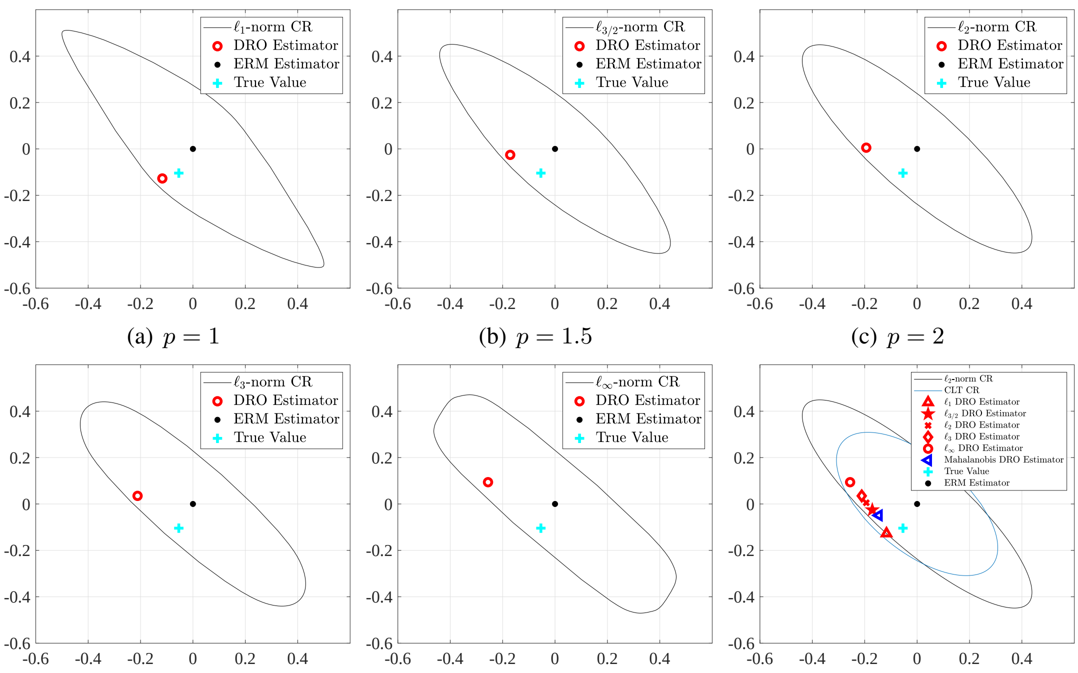
with $\rho = 0.7$. In Figures 1(a)-1(e), we draw the 95% confidence region corresponding to the choices $p = 1, 3/2, 2, 3, \infty$, ($q = \infty, 3, 2, 3/2$, respectively) by means of support functions defined in Proposition 5. In addition, a confidence region for β_* resulting from the asymptotic normality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

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

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 (16) such that the resulting confidence region coincides with $\Lambda_{CLT}(P_n)$. Namely, A is chosen by solving the equation

$$E\left[\{e\xi - (\xi^{\mathrm{T}} X)\,\beta_*\}\, A^{-1}\,\{e\xi - (\xi^{\mathrm{T}} X)\,\beta_*\}^{\mathrm{T}}\right] = C\sigma^2. \tag{18}$$

Figure 1(f) gives the confidence region for the choice $p = 2$ and $\Lambda_{CLT}(P_n)$ superimposed with various distributionally robust minimizer 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



(a) $p = 1$ (b) $p = 1.5$ (c) $p = 2$

(d) $p = 3$ (e) $p = \infty$ (f) CLT

FIGURE 1. Confidence regions for different norm choices and central limit theorem based confidence region plotted together with the respective β_n^{DRO} estimators and β_n^{ERM}

.

find 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. As noted, 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.** In this section, we 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 (17), considering three different values for ρ : we choose $\rho = 0.95, 0, -0.95$. 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 below, we report the coverage probabilities of the underlying β_* and $\beta_n^{DRO}(\delta_n)$ in both the ℓ_2-confidence region and the central limit theorem based confidence regions. Specifically, we report the following four probabilities:

$$\mathrm{pr}\{\beta_n^{DRO} \in \Lambda^+_{\delta_n}(P_n)\}, \quad \mathrm{pr}\{\beta_* \in \Lambda^+_{\delta_n}(P_n)\}, \quad \mathrm{pr}\{\beta_n^{DRO} \in \Lambda_{CLT}(P_n)\}, \quad \mathrm{pr}\{\beta_* \in \Lambda_{CLT}(P_n)\}.$$

We sample 1000 datasets and report the coverage probabilities in Table 1. From 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.

TABLE 1. Coverage Probability

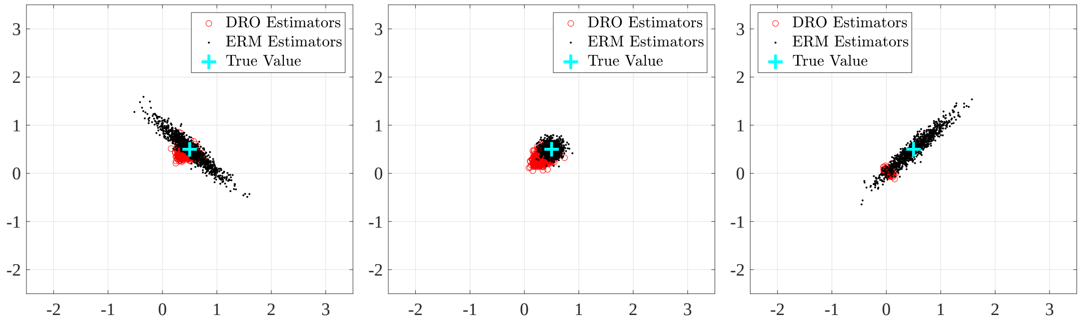
β_0	ρ	ℓ_2-confidence region		CLT confidence region	
		Coverage for β_n^{DRO}	Coverage for β_*	Coverage for β_n^{DRO}	Coverage for β_*
$\begin{bmatrix}0.5\\0.5\end{bmatrix}$	0.95	100.0%	94.5%	99.4%	94.6%
	0	100.0%	94.0%	97.1%	93.5%
	-0.95	100.0%	94.8%	75.8%	94.4%
$\begin{bmatrix}1.0\\0.0\end{bmatrix}$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

Figures 2 and 3 show the scatter plots 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 $\Lambda_{CLT}(P_n)$.

5. PROOFS OF MAIN RESULTS

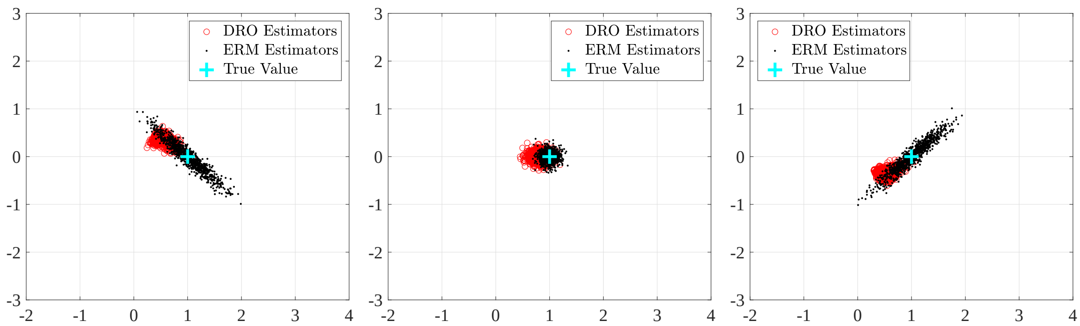
Theorem 1 is obtained by considering appropriate level sets involving auxiliary functionals which we define next. Following Blanchet et al. [2019], we define the robust Wasserstein profile function, associated with the estimation of β_* by solving $E_{P_n}\{D_\beta h(X, \beta)\} = 0$, as follows:

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



(a) $\rho = 0.95$ (b) $\rho = 0$ (c) $\rho = -0.95$

FIGURE 2. Scatter plots of β_n^{ERM} (black circles) and β_n^{DRO} (red circles) for $\beta_0 = [0.5, 0.5]^{\mathrm{T}}$.



(a) $\rho = 0.95$ (b) $\rho = 0$ (c) $\rho = -0.95$

FIGURE 3. Scatter plots of β_n^{ERM} (black circles) and β_n^{DRO} (red circles) for $\beta_0 = [1.0, 0.0]^{\mathrm{T}}$.

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

$$\Lambda_\delta^+(P_n) = \mathrm{cl}\{\beta : R_n(\beta) \le \delta\}, \tag{19}$$

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 $V_n^{DRO} : \mathbb{R}^d \to \mathbb{R}$ and $V_n^{ERM} : \mathbb{R}^d \to \mathbb{R}$, as below, by considering suitably scaled versions of the distributionally robust and empirical risk objective functions, namely

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

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 we have,*

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

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 A2.c, is useful towards investigating the behavior of $nR_n(\,\cdot\,)$ in the neighborhood of β^*, as required in the third component in the triplet in Theorem 3.

5.1. **Proof of Theorem 3.** Throughout this section, we suppose that the assumptions imposed in Theorem 1 hold. Let

$$H_n = n^{-1/2}\sum_{i=1}^{n} h\left(X_i,\beta_*\right)$$

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*

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

for every $n > n_0$ and $\|u\|_2 \le K$. Specifically, if $\alpha = 1$, we have

$$\mathrm{pr}\left\{\left|V_n^{ERM}(u) - H_n^{\mathrm{T}}u - 2^{-1}u^{\mathrm{T}}Cu\right| \le \varepsilon'\right\} \ge 1-\varepsilon. \tag{20}$$

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

$$\mathrm{pr}\left\{\left|V_n^{DRO}(u) + f_{\eta,\gamma}(-H_n)^{\mathrm{T}}u - 2^{-1}u^{\mathrm{T}}Cu\right| \le \varepsilon'\right\} \ge 1-\varepsilon, \tag{21}$$

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

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 $\beta_* + n^{-1/2}u \in \Theta$, We have,

$$nR_n\big(\beta_* + n^{-1/2}u\big) = \max_{\xi}\left\{-\xi^{\mathrm{T}}H_n - M_n(\xi,u)\right\},$$

where

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

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

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

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

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

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

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

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

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

for all $n > n_0$, where

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

Proposition 11. *For any $\varepsilon > 0$, there exist constants $a, n_0 > 0$ such that for every $n \geq n_0$,*

$$\mathrm{pr}\left\{nR_n(\beta_*) \leq a\right\} \geq 1 - \varepsilon,$$

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

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

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

Proofs of Propositions 6 - 12 are furnished in Section B in the supplementary material. With the statements of these results, we proceed with the proof of Theorem 3 as follows.

Proof of Theorem 3. Since $E\{h(X,\beta_*)\} = 0$, it follows from central limit theorem that $H_n \Rightarrow -H$, where $H \sim \mathcal{N}(0, E\{h(X,\beta_*)h(X,\beta_*)^{\mathrm{T}}\})$. Since inequalities (21) and (20) are associated with the same H_n , it follows from Propositions 6 and 7 that,

$$V_n^{ERM}(\cdot) \Rightarrow V^{ERM}(\cdot) = V(-H, \cdot) \quad \text{and} \quad V_n^{DRO}(\cdot) \Rightarrow V^{DRO}(\cdot) = V\{-f_{\eta,\gamma}(H), \cdot\} \tag{22}$$

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,

$$\varphi^*(H - Cu) = \max_{\xi} \big(\xi^{\mathrm{T}}[H - E\{D_\beta h(X,\beta_*)\}u] - 4^{-1}E\|\{D_x h(X,\beta_*)\}^{\mathrm{T}}\xi\|_p^2\big). \tag{23}$$

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

$$f_{up}(H_n, u, b, c) \Rightarrow \max_{\|\xi\|_p \leq b} \big\{\xi^{\mathrm{T}} H - E\big[4^{-1}\|\big\{D_x h(X,\beta_*)\big]^{\mathrm{T}}\xi\|_p^2 + \xi^{\mathrm{T}} D_\beta h(X,\beta_*)u\big]\mathbb{I}(X \in C_0^{\epsilon_0})\big\}, \tag{24}$$

$$f_{low}(H_n, u, b) \Rightarrow \max_{\|\xi\|_p \leq b} \big\{\xi^{\mathrm{T}} H - E\big[4^{-1}\|\{D_x h(X,\beta_*)\}^{\mathrm{T}}\xi\|_p^2 + \xi^{\mathrm{T}} D_\beta h(X,\beta_*)u\big]\big\}, \tag{25}$$

for any u satisfying $\|u\|_2 \leq K$. Since the bounds in Propositions 9, 10 hold for arbitrarily large choices for constants b, c, and arbitrarily small choice for constant ϵ_0 combining with the assumption $P_*(\Omega^\circ) = 1$, we conclude from the observations (23), (24), and (25) that

$$nR_n\big(\beta_* + n^{-1/2}u\big) \Rightarrow \varphi^*(H - Cu), \tag{26}$$

for any u satisfying $\|u\|_2 \leq K$. Finally, we have from Propositions 11 and 12 that 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 (26), we have that,

$$nR_n\big(\beta_* + n^{-1/2} \times \cdot\big) \Rightarrow \varphi^*(H - C \times \cdot).$$

Combining this observation with those in (22), we obtain the desired convergence result in Theorem 3. Furthermore, since $f_{low}(H_n, u, b)$ and $f_{up}(H_n, u, b)$ are associated with the same H_n with inequalities (21) and (20), we have the three terms converge jointly. □

5.2. **Proof of Theorem 1.**

Proof of Theorem 1. Theorem 1 is proved by considering suitable level sets of the component functions in the triplet, $\{V_n^{ERM}(\cdot),\ V_n^{DRO}(\cdot),\ nR_n(\beta_* + n^{-1/2} \times \cdot\ )\}$, considered in Theorem 3. To reduce clutter in expressions, from here-onwards we refer 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 whose proof is provided in Section C.

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

Observe that $V_n^{ERM}(\cdot)$ and $V_n^{DRO}(\cdot)$ are minimized, respectively, at $n^{1/2}(\beta_n^{ERM} - \beta_*)$ and $n^{\bar{\gamma}/2}(\beta_n^{DRO} - \beta_*)$. Furthermore, due to the positive definiteness of C in Assumption A2.b, we have that $V^{ERM}(\cdot)$ and $V^{DRO}(\cdot)$ are strongly convex with respect to u and have unique minimizers, with probability 1. Therefore, due to the tightness of the sequences $\{n^{1/2}(\beta_n^{ERM} - \beta_*)\}_{n\geq 1}$ and $\{n^{\bar{\gamma}/2}(\beta_n^{DRO} - \beta_*)\}_{n\geq 1}$; see Proposition 13 and the weak convergence of $V_n^{ERM}(\cdot)$ and $V_n^{DRO}(\cdot)$ in Theorem 3, we have the following convergences:

$$\begin{aligned} n^{1/2}(\beta_n^{ERM} - \beta_*) &\Rightarrow \arg\min_u\ V(-H, u) = C^{-1}H, \\ n^{\bar{\gamma}/2}(\beta_n^{DRO} - \beta_*) &\Rightarrow \arg\min_u\ V^{DRO}(u) = C^{-1}f_{\eta,\gamma}(H) \end{aligned} \tag{27}$$

Finally, to prove the convergence of the sets $\Lambda_{\delta_n}^+(P_n)$, we proceed as follows. Define

$$G_n(u) = nR_n(\beta_* + n^{-1/2}u), \quad G(u) = \varphi^*(H - Cu), \quad \text{and} \quad \alpha_n = n\delta_n.$$

For any function $f : B \to \mathbb{R}$ and $\alpha \in [0, +\infty]$, let lev(f, α) denote the level set $\{x \in \mathbb{R}^d : f(x) \leq \alpha\}$.

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 above, whose proofs are furnished in Section C, allow us to complete the proof of Theorem 1 as follows. It follows from the definition of $R_n(\beta)$ that,

$$\Lambda_{\delta_n}^+(P_n) = \{\beta : R_n(\beta) \leq \delta_n\} = \beta_* + n^{-1/2}\{u : G_n(u) \leq \alpha_n\}.$$

We have from Propositions 14 - 16 that

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

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

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

6. DISCUSSIONS

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, denoted by Ω, is constrained to be a strict subset of $\mathbb{R}^m$. Suppose that the support of X is constrained to be contained in the set $\Omega = \{x \in \mathbb{R}^m : Ax \leq b\}$ 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 non-degenerate case where $\delta_n = \eta/n$.

Considering the transportation cost $c(x, y) = \|x - y\|_2^2$ in Definition 1, we demonstrate in Section A.3 of the Supplementary material that the central limit theorem, $n^{1/2}\{\beta_n^{DRO}(\delta_n) - \beta_*\} \Rightarrow C^{-1}H - \eta^{1/2}C^{-1}D_\beta S(\beta_*)$, 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 (offered in Proposition 18 in Section A.3 of the supplementary material) is that the fraction of samples which get transported to the boundary of the set Ω stays $O_p(n^{-1/2})$, as $n \to \infty$.

On the other hand, when the set $\Omega = \{x \in \mathbb{R}^m : Ax \leq b\}$ 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 is easily seen in the linear regression example in Section 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,

$$n^{1/2}\{\beta_n^{DRO}(\delta_n) - \beta_*\} \Rightarrow C^{-1}H - \eta^{1/2}C^{-1}D_\beta\tilde{S}(\beta_*), \tag{28}$$

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 Example A1 in Section A.3 of the supplementary material. In the presence of general support constraints of the form $\Omega = \{x \in \mathbb{R}^m : Ax = b\}$, we show with Example A2 in Section A.3 that (28) 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.

ACKNOWLEDGEMENTS

Material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support is gratefully acknowledged from NSF grants 1915967, 1820942 and 1838576 and MOE SRG ESD 2018 134.

REFERENCES

G. Beer and R. Lucchetti. Minima of quasi-convex functions. *Optimization*, 20(5):581–596, 1989. doi: 10.1080/02331938908843480. URL `https://doi.org/10.1080/02331938908843480`.

G. Beer, R. T. Rockafellar, and R. J.-B. Wets. A characterization of epi-convergence in terms of convergence of level sets. *Proceedings of the American Mathematical Society*, 116(3):753–761, 1992. ISSN 00029939, 10886826. URL `http://www.jstor.org/stable/2159443`.

P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

J. Blanchet and Y. Kang. Distributionally robust groupwise regularization estimator. *arXiv preprint arXiv:1705.04241*, 2017.

J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019. doi: 10.1287/moor.2018.0936.

J. Blanchet, K. Murthy, and F. Zhang. Optimal transport based distributionally robust optimization: Structural properties and iterative schemes. *arXiv preprint arXiv:1810.02403*, 2018.

J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56.3, September 2019.

J. Blanchet, Y. Kang, K. Murthy, and F. Zhang. Data-driven optimal transport cost selection for distributionally robust optimization. In *2019 Winter Simulation Conference (WSC)*, pages 3740–3751, 2019.

S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

R. Chen and I. C. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *The Journal of Machine Learning Research*, 19(1):517–564, 2018.

Z. Chen, D. Kuhn, and W. Wiesemann. Data-driven chance constrained programs over Wasserstein balls. *arXiv preprint arXiv:1809.00210*, 2018.

P. Cisneros-Velarde, S.-Y. Oh, and A. Petersen. Distributionally robust formulation and model selection for the graphical lasso. *arXiv preprint arXiv:1905.08975*, 2019.

J. C. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses against mixture covariate shifts. *preprint*, 2019.

J. Dupacova and R. Wets. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Ann. Statist.*, 16(4):1517–1549, 12 1988. doi: 10.1214/aos/1176351052. URL `https://doi.org/10.1214/aos/1176351052`.

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

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

R. Gao, L. Xie, Y. Xie, and H. Xu. Robust hypothesis testing using Wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pages 7902–7912, 2018.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233, Berkeley, Calif., 1967. University of California Press. URL `https://projecteuclid.org/euclid.bsmsp/1200512988`.

K. Knight, W. Fu, et al. Asymptotics for lasso-type estimators. *The Annals of statistics*, 28(5):1356–1378, 2000.

F. Luo and S. Mehrotra. Decomposition algorithm for distributionally robust optimization using Wasserstein metric. *arXiv preprint arXiv:1704.03920*, 2017.

P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, Sep 2018. ISSN 1436-4646. doi: 10.1007/s10107-017-1172-1. URL `https://doi.org/10.1007/s10107-017-1172-1`.

I. S. Molchanov. *Theory of random sets*, volume 19. Springer, 2005.

A. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2): 237–249, 1988.

A. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, pages 90–120, 1990.

A. B. Owen. *Empirical likelihood*. Chapman and Hall/CRC, 2001.

R. T. Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

S. Shafieezadeh-Abadeh, P. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584. 2015.

S. Shalev-Shwartz and Y. Singer. Online learning: Theory, algorithms, and applications. 2007.

A. Shapiro. Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.*, 17(2):841–858, 06 1989. doi: 10.1214/aos/1176347146. URL `https://doi.org/10.1214/aos/1176347146`.

A. Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186, Dec 1991. ISSN 1572-9338. doi: 10.1007/BF02204815. URL `https://doi.org/10.1007/BF02204815`.

A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845, 1993. doi: 10.1287/moor.18.4.829. URL `https://doi.org/10.1287/moor.18.4.829`.

A. Shapiro. On the asymptotics of constrained local m-estimators. *Ann. Statist.*, 28(3):948–960, 05 2000. doi: 10.1214/aos/1015952006. URL `https://doi.org/10.1214/aos/1015952006`.

A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Hk6kPgZA-`.

M. Sion et al. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

M. Staib and S. Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, 2017.

A. van der Vaart, A. van der Vaart, A. van der Vaart, and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. ISBN 9780387946405. URL `https://books.google.com.sg/books?id=OCenCW9qmp4C`.

R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018.

R. A. Wijsman. Convergence of sequences of convex sets, cones and functions. ii. *Transactions of the American Mathematical Society*, 123(1):32–45, 1966. ISSN 00029947. URL `http://www.jstor.org/stable/1994611`.

I. Yang. A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance. *IEEE control systems letters*, 1(1):164–169, 2017.

I. Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *arXiv preprint arXiv:1812.09808*, 2018.

C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262 – 267, 2018. ISSN 0167-6377. doi: https://doi.org/10.1016/j.orl.2018.01.011. URL `http://www.sciencedirect.com/science/article/pii/S0167637718300506`.

Supplementary material

APPENDIX A. PROOFS PERTAINING TO LIMIT THEOREMS OF β_n^{ERM} AND $\beta_n^{DRO}(\delta_n)$

In this section we first present the proofs of Propositions 6 - 7 which are useful towards establishing convergences of the first two components of the triple considered in Theorem 3. Following these, we present the proofs of Propositions 2 - 3, both pertaining to limit theorems for the distributionally robust estimator under relaxed assumptions. Towards the end of this section, we also provide the proofs of statements made in Section 6.

Proof of Proposition 6. Recall that $h(x;\beta) = D_\beta \ell(x;\beta)$. For n sufficiently large, we have $\beta_* + n^{-\alpha/2}u \in B^\circ$ for $\|u\| \leq 2$. With $\ell(\cdot)$ being twice continuously differentiable, employing Taylor expansion up to the quadratic term, we obtain,

$$\begin{aligned} n^{\alpha-1} V_n^{ERM}\{n^{(1-\alpha)/2}u\} &= n^{\alpha}\left[E_{P_n}\left\{\ell\left(X;\beta_* + n^{-\alpha/2}u\right)\right\} - E_{P_n}\left\{\ell\left(X;\beta_*\right)\right\}\right] \\ &= n^{\alpha/2}E_{P_n}\{h(X;\beta_*)\}^{\mathrm{T}}u + \frac{1}{2}u^{\mathrm{T}}E_{P_n}\{D_\beta h(X;\beta_*)\}u + o(1), \end{aligned}$$

as $n \to \infty$, uniformly over u in compact sets. With this expansion, the statement of Proposition 6 follows as a direct consequence of the definitions, $H_n = n^{-1/2}\sum_{i=1}^n h(X_i,\beta_*)$, $C = E\{D_\beta h(X;\beta_*)\}$ and an application of the law of large numbers, $\lim_{n\to\infty} E_{P_n}\{D_\beta h(X,\beta_*)\} \to C$ almost surely. □

A.1. **Proof of Proposition 7.** This subsection is devoted to the proof of Proposition 7 taking $\Omega = \mathbb{R}^m$. The following notation will be used in the sequence of results below used to prove Proposition 7. Given $q \in (1,\infty)$, let $D_q(v), H_q(v)$ denote the first derivative (gradient) and second derivative (Hessian) of the function $f(\Delta) = \|\Delta\|_q^2$ evaluated at $\Delta = v$. Recall that we take p to be such that $p^{-1} + q^{-1} = 1$. We also define the map $T_p : \mathbb{R}^m \to \mathbb{R}^m$ as,

$$T_p(v) = \|v\|_p^{1-p/q}\mathrm{sgn}(v)|v|^{p/q},$$

where $\mathrm{sgn}(\cdot)$ denotes the sign function.

Proposition 7 is proved via the sequence of results below.

Lemma 1. *Letting $\eta_n = \delta_n n^\gamma$, we have, for $\beta \in B$*

$$n^{\gamma/2}\left[\Psi_n(\beta) - E_{P_n}\{\ell(X;\beta)\}\right] = \inf_{\lambda\geq 0}\left[\lambda\eta_n + \frac{1}{4\lambda}E_{P_n}\left\{\|D_x\ell(X;\beta)\|_p^2\right\} + e_n(\beta,\lambda)\right],$$

where the function $e_n(\beta,\lambda)$ is $e_n(\beta,\lambda) = E_{P_n}\{f_n(X,\beta,\lambda)\}$, with $f_n(\cdot)$ defined as,

$$f_n(x,\beta,\lambda) = \sup_{\Delta\in\mathbb{R}^m}\left[n^{\gamma/2}\left\{\ell\big(x + n^{-\gamma/2}\Delta;\beta\big) - \ell(x;\beta)\right\} - \lambda\|\Delta\|_q^2\right] - \frac{1}{4\lambda}\|D_x\ell(x;\beta)\|_p^2.$$

Proof of Lemma 1. It follows from Blanchet and Murthy [2019, Theorem 1] that

$$\Psi_n(\beta) = \inf_{\lambda\geq 0}\left[n^{\gamma/2}\lambda\delta_n + E_{P_n}\{\phi_\lambda(X;\beta,\lambda)\}\right], \quad \text{where}$$

$$\phi_\lambda(x;\beta,\lambda) = \sup_{\Delta\in\{\Delta\in\mathbb{R}^m | x+n^{-\gamma/2}\Delta\in\Omega\}}\left\{\ell\big(x+n^{-\gamma/2}\Delta;\beta\big) - \lambda n^{-\gamma/2}\|\Delta\|_q^2\right\}.$$

With $\Omega = \mathbb{R}^m$, it follows from the definition of $f_n(\cdot)$ that,

$$\begin{aligned} n^{\gamma/2}\{\phi_\lambda(x;\beta) - \ell(x;\beta)\} &= \sup_{\Delta\in\mathbb{R}^m}\left[n^{\gamma/2}\left\{\ell\big(x+n^{-\gamma/2}\Delta;\beta\big) - \ell(x;\beta)\right\} - \lambda\|\Delta\|_q^2\right] \\ &= f_n(x,\beta,\lambda) + (4\lambda)^{-1}\|D_x\ell(x;\beta)\|_p^2. \end{aligned}$$

Then, since $e_n(\beta,\lambda) = E_{P_n}[f_n(X,\beta,\lambda)]$ and $\delta_n n^\gamma = \eta_n$, we obtain,

$$n^{\gamma/2}\left[\Psi_n(\beta) - E_{P_n}\left\{\ell(x;\beta)\right\}\right] = \inf_{\lambda\geq 0}\left[\lambda\eta_n + \frac{1}{4\lambda}E_{P_n}\left\{\|D_x\ell(X;\beta)\|_p^2\right\} + e_n(\beta,\lambda)\right].$$

This completes the verification of the statement of Lemma 1. □

Lemma 2. *For any $\Delta, \Delta_* \in \mathbb{R}^d$, letting $\xi = \Delta - \Delta_*$, we have the following inequalities:*

a) *if $q \in (1, 2]$, then $\|\Delta\|_q^2 \geq \|\Delta_*\|_q^2 + D_q(\Delta_*)^{\mathrm{T}}\xi + (q-1)\|\xi\|_q^2$; and*
b) *if $q > 2$, then $\|\Delta\|_q^2 \geq \|\Delta_*\|_q^2 + D_q(\Delta_*)^{\mathrm{T}}\xi + C\min\left\{\|\xi\|_2^2, \|\xi\|_q^q\|\Delta_*\|_{q-2}^{-(q-2)}\right\}$, where C is a positive constant which depends only on d and q.*

The proof of Lemma 2 is technical in nature and is provided in Section F.

Lemma 3. *For any $v \in \mathbb{R}^d, \lambda > 0, \varepsilon > 0$ and $d \times d$ symmetric matrix B, we have that the value of optimization,*

$$\sup_{\Delta\in\mathbb{R}^d}\left(v^{\mathrm{T}}\Delta - \lambda\|\Delta\|_q^2 + \varepsilon\Delta^{\mathrm{T}}B\Delta\right) \tag{29}$$

is upper bounded and lower bounded as follows:

$$0 \leq \sup_{\Delta\in\mathbb{R}^d}\left\{v^{\mathrm{T}}\Delta - \lambda\|\Delta\|_q^2 + \varepsilon\Delta^{\mathrm{T}}B\Delta\right\} - \frac{\|v\|_p^2}{4\lambda} - \frac{\varepsilon}{4\lambda^2}T_p(v)^{\mathrm{T}}BT_p(v) \leq \frac{c_0\varepsilon^{\bar{q}}\|\Delta_v\|_2^2}{\min\{(\lambda - c_1\varepsilon)^+, \lambda^{\frac{1}{q-1}}\}}$$

where $\Delta_v = (2\lambda)^{-1}T_p(v)$, $\bar{q} = \min\{2, q/(q-1)\}$, and c_0, c_1 are positive constants which depends only on d, q and the Frobenius norm of the matrix B.

Proof of Lemma 3. First, we consider the case where B is the zero matrix. When $B = 0$, we have

$$\sup_{\Delta\in\mathbb{R}^d}\left(v^{\mathrm{T}}\Delta - \lambda\|\Delta\|_q^2\right) = (4\lambda)^{-1}\|v\|_p^2,$$

in which the maximum is attained at $\Delta = \Delta_v$; here, recall that Δ_v is $\Delta_v = (2\lambda)^{-1}T_p(v)$. The corresponding optimality condition is

$$v - \lambda D_q(\Delta_v) = 0, \tag{30}$$

where $D_q(\Delta_v)$ is the first derivative of the function $\|\Delta\|_q^2$ evaluated at $\Delta = \Delta_v$. Next, for the case where the matrix B is not zero, we proceed by changing the variable from Δ to ξ with the relationship, $\Delta = \Delta_v + \varepsilon\xi$. Then the objective $f(\Delta) = \Delta^{\mathrm{T}}v - \lambda\|\Delta\|_q^2 + \varepsilon\Delta^{\mathrm{T}}B\Delta$ is rewritten in terms of the variable ξ as follows: $f\left(\Delta_v + \varepsilon\xi\right)$ equals

$$\begin{aligned}&\left(v^{\mathrm{T}}\Delta_v - \lambda\|\Delta_v\|_q^2\right) + \varepsilon\Delta_v^{\mathrm{T}}B\Delta_v + \varepsilon\xi^{\mathrm{T}}\left(v + 2\varepsilon B\Delta_v\right) - \lambda\left(\|\Delta_v + \varepsilon\xi\|_q^2 - \|\Delta_v\|_q^2\right) + \varepsilon^3\xi^{\mathrm{T}}B\xi\\ &\quad = (4\lambda)^{-1}\|v\|_p^2 + \varepsilon\Delta_v^{\mathrm{T}}B\Delta_v + \varepsilon\xi^{\mathrm{T}}\left\{v - \lambda D_q(\Delta_v) + 2\varepsilon B\Delta_v\right\}\\ &\qquad - \lambda\left\{\|\Delta_v + \varepsilon\xi\|_q^2 - \|\Delta_v\|_q^2 - \varepsilon\xi^{\mathrm{T}}D_q(\Delta_v)\right\} + \varepsilon^3\xi^{\mathrm{T}}B\xi.\end{aligned}$$

Then, we have from (30) that,

$$\begin{aligned}&\varepsilon^{-2}\left\{f(\Delta) - (4\lambda)^{-1}\|v\|_p^2 - \varepsilon\Delta_v^{\mathrm{T}}B\Delta_v\right\}\\ &\qquad = 2\xi^{\mathrm{T}}B\Delta_v - \lambda\varepsilon^{-2}\left\{\|\Delta_v + \varepsilon\xi\|_q^2 - \|\Delta_v\|_q^2 - \varepsilon\xi^{\mathrm{T}}D_q(\Delta_v)\right\} + \varepsilon\xi^{\mathrm{T}}B\xi.\end{aligned} \tag{31}$$

For deriving the upper bound in the statement of Lemma 3, we proceed by utilizing the bound, $\|\Delta_v + \varepsilon\xi\|_q^2 - \|\Delta_v\|_q^2 - \varepsilon\xi^{\mathrm{T}} D_q(\Delta_v) \geq P_v(\varepsilon\xi)$ from Lemma 2, where $P_v : \mathbb{R}^d \to \mathbb{R}_+$ defined as,

$$P_v(x) = \tilde{c} \begin{cases} \|x\|_2^2 & \text{if } q \in (1,2] \\ \min\left(\|x\|_2^2, \|x\|_2^q \|\Delta_v\|_2^{2-q}\right) & \text{if } q > 2 \end{cases}$$

for a suitable positive constant $\tilde{c}$ that depends only on d and q; indeed, the existence of consant $\tilde{c}$ satisfying this requirement follows from the observation that $\|x\|_q \geq \hat{c}\|x\|_2$ for a suitable positive constant $\hat{c}$ which depends only upon d and q. Then we have the following upper bound from (31):

$$\varepsilon^{-2}\left\{\sup_{\Delta} f(\Delta) - (4\lambda)^{-1}\|v\|_p^2 - \varepsilon\Delta_v^{\mathrm{T}} B\Delta_v\right\} \leq \sup_{\xi}\left\{2\xi^{\mathrm{T}} B\Delta_v - \lambda\varepsilon^{-2} P_v(\varepsilon\xi) + \varepsilon\xi^{\mathrm{T}} B\xi\right\}. \tag{32}$$

The following observations are useful in simplifying the right hand side of (32). With $\|B\|$ denoting the Frobenius norm of the matrix B, we have $\|B\xi\|_2 \leq \|B\|\|\xi\|_2$ and $\xi^{\mathrm{T}} B\xi \leq \|B\|\|\xi\|_2^2$. As a consquence, when $q \in (1,2]$,

$$\sup_{\xi}\left\{2\xi^{\mathrm{T}} B\Delta_v - \lambda\varepsilon^{-2} P_v(\varepsilon\xi) + \varepsilon\xi^{\mathrm{T}} B\xi\right\} \leq \sup_{\xi}\left\{2\xi^{\mathrm{T}} B\Delta_v - (\lambda\tilde{c} - \varepsilon\|B\|)\|\xi\|_2^2\right\} = \frac{\|B\|^2\|\Delta_v\|_2^2}{(\lambda\tilde{c} - \varepsilon\|B\|)^+}.$$

In the above expression, $x^+ = \max\{x, 0\}$ denotes the positive part of any real number x. Next, when $q > 2$, we have the following as a consequence of Cauchy-Schwarz inequality:

$$\begin{aligned}
&\sup_{\xi}\left(2\xi^{\mathrm{T}} B\Delta_v - \lambda\tilde{c}\varepsilon^{-2}\|\varepsilon\xi\|_2^q\|\Delta_v\|_2^{2-q} + \varepsilon\xi^{\mathrm{T}} B\xi\right) \\
&\quad\leq \sup_{C\geq 0}\left(2\|B\|\|\Delta_v\|_2 C - \lambda\tilde{c}\|\Delta_v\|_2^{2-q}\varepsilon^{q-2} C^q + \varepsilon\|B\|C^2\right) \\
&\quad\leq \sup_{C\geq 0}\left(2\|B\|\|\Delta_v\|_2 C - 2^{-1}\lambda\tilde{c}\|\Delta_v\|_2^{2-q}\varepsilon^{q-2} C^q\right) + \sup_{C\geq 0}\left(\varepsilon\|B\|C^2 - 2^{-1}\lambda\tilde{c}\|\Delta_v\|_2^{2-q}\varepsilon^{q-2} C^q\right) \\
&\quad\leq c\varepsilon^{-\frac{q-2}{q-1}}\lambda^{-\frac{1}{q-1}}\|\Delta_v\|_2^2\left\{1 + (\varepsilon\lambda^{-1})^{\frac{q}{(q-1)(q-2)}}\right\},
\end{aligned}$$

where c is a suitable positive constant which depends only upon d, q and $\|B\|$. Then letting $\bar{q} = \min\{2, q/(q-1)\}$, we obtain from (32) and the above two upper bounds that,

$$\sup_{\Delta} f(\Delta) - (4\lambda)^{-1}\|v\|_p^2 - \varepsilon\Delta_v^{\mathrm{T}} B\Delta_v \leq \frac{c_0\varepsilon^{\bar{q}}\|\Delta_v\|_2^2}{\min\{(\lambda - c_1\varepsilon)^+, \lambda^{\frac{1}{q-1}}\}}$$

where c_0, c_1 are positive constants which depends only on d, q and $\|B\|$. With this conclusion proving the upper bound, the lower bound is obtained by letting $\Delta = \Delta_v$ in the evaluation of $f(\Delta)$. This concludes the proof of Lemma 3. □

Lemma 4. *For any $\beta \in B$ and $\lambda > 0$, we have the following approximation for the term $e_n(\beta, \lambda)$ identified in Lemma 1: As $n \to \infty$,*

$$e_n(\beta, \lambda) = 8^{-1}\lambda^{-2} n^{-\gamma/2} a_n(\beta) + O_p(n^{-\bar{q}\gamma/2}),$$

where

$$a_n(\beta) = E_{P_n}\left[T_p\{D_x\ell(X;\beta)\}^{\mathrm{T}} D_{xx}\ell(X;\beta) T_p\{D_x\ell(X;\beta)\}\right],$$

and the convergence is uniform over β in compact subsets of B and λ bounded away from zero. Moreover, the $O_p(n^{-\bar{q}\gamma/2})$ term is such that $\sup_{\lambda>0}\lambda^2 O_p(n^{-\bar{q}\gamma/2})$ is bounded from below by an integral random variable.

Proof of Lemma 4. Consider any β satisfying $\|\beta\|_2 \leq b$ and $\lambda > \lambda_0 \in (0,1]$. For any $\varepsilon' > 0$, we have as a consequence of Taylor expansion and uniform continuity of $D_{xx}\ell(\cdot;\beta)$ in Assumption A2.c that,

$$\left|\ell(x+\Delta n^{-\gamma/2};\beta) - \ell\left(x;\beta\right) - n^{-\gamma/2}D_x\ell(x;\beta)^{\mathrm{T}}\Delta - 2^{-1}n^{-\gamma}\Delta^{\mathrm{T}}D_{xx}\ell(x;\beta)\Delta\right| \leq \varepsilon' n^{-\gamma}\|\Delta\|_q^2,$$

for all $n \geq n_0$ where n_0 is sufficiently large. Then it follows from the definition of $f_n(\cdot)$ and Assumption A2.c that $f_n(x,\beta,\lambda)$ is upper and lower bounded, respectively, by,

$$\sup_{\Delta}\left\{D_x\ell(x;\beta)^{\mathrm{T}}\Delta + 2^{-1}n^{-\gamma/2}\Delta^{\mathrm{T}}D_{xx}\ell(x;\beta)\Delta - \left(\lambda - \varepsilon' n^{-\gamma/2}\right)\|\Delta\|_q^2\right\} - \frac{1}{4\lambda}\|D_x\ell(x,\beta)\|_p^2 \text{ and}$$
$$\sup_{\Delta}\left\{D_x\ell(x;\beta)^{\mathrm{T}}\Delta + 2^{-1}n^{-\gamma/2}\Delta^{\mathrm{T}}D_{xx}\ell(x;\beta)\Delta - \left(\lambda + \varepsilon' n^{-\gamma/2}\right)\|\Delta\|_q^2\right\} - \frac{1}{4\lambda}\|D_x\ell(x,\beta)\|_p^2.$$

Letting $\varepsilon = n^{-\gamma/2}$, $v = D_x\ell(x;\beta)$, $\bar{q} = \min\{2, q/(q-1)\}$ and $B = D_{xx}\ell(x;\beta)$, we obtain from the bounds derived for (29) in Lemma 3 that,

$$f_n(x,\beta,\lambda) - 8^{-1}\lambda^{-2}n^{-\gamma/2}T_p\left\{D_x\ell(x;\beta)\right\}^{\mathrm{T}}D_{xx}\ell(x;\beta)T_p\left\{D_x\ell(x;\beta)\right\}$$

is upper bounded by,

$$c_u(1+\varepsilon')n^{-\bar{q}\gamma/2}\|T_p\left\{D_x\ell(x;\beta)\right\}\|_2^2\,\lambda_0^{-\max\{2,\frac{1}{q-1}\}}$$

and likewise, lower bounded by,

$$-c_l\varepsilon' n^{-\gamma}\|T_p\left\{D_x\ell(x;\beta)\right\}\|_2^2\lambda^{-2} \tag{33}$$

for suitable positive constants c_l, c_u which are, in turn, determined by the constants b, d and q. Since $e_n(\beta,\lambda)$ is defined to equal $E_{P_n}[f_n(X,\beta,\lambda)]$, due to the finiteness of the second moment of $\sup[T_p\{D_x\ell(X;\beta)\} : \|\beta\|_2 \leq b]$, we have that

$$e_n(\beta,\lambda) = 8^{-1}\lambda^{-2}n^{-\gamma/2}E_{P_n}\left[T_p\left\{D_x\ell(X;\beta)\right\}^{\mathrm{T}}D_{xx}\ell(X;\beta)T_p\left\{D_x\ell(X;\beta)\right\}\right] + O_p(n^{-\bar{q}\gamma/2}),$$

where the convergence is uniform over (β,λ) such that $\|\beta\|_2 \leq b$ and $\lambda > \lambda_0$. The observation that the $O_p(n^{-\bar{q}\gamma/2})$ term satisfies $\lambda^2 O_p(n^{-\bar{q}\gamma/2})$ is bounded from below by an integral random variable, uniformly over all $\lambda > 0$ and $\|\beta\| \leq b$, follows from the lower bound in (33). □

Proposition 17. *As $n \to \infty$, we have,*

$$\Psi_n(\beta) = E_{P_n}[\ell(X;\beta)] + \delta_n^{1/2}\left[E_{P_n}\{\,\|D_x\ell(X;\beta)\|_p^2\,\}\right]^{1/2} + \delta_n\frac{a_n(\beta)}{2E_{P_n}\{\,\|D_x\ell(X;\beta)\|_p^2\,\}} + o_p\left(\delta_n\right),$$

uniformly over β in compact sets.

Proof of Proposition 17. We have from Lemma 1 and 4 that $n^{\gamma/2}\left(\Psi_n(\beta) - E_{P_n}[\ell(X;\beta)]\right)$ equals,

$$\lim_{\lambda_0\downarrow 0}\inf_{\lambda\geq\lambda_0}\left[\lambda\eta_n + \frac{1}{4\lambda}E_{P_n}\left\{\|D_x\ell(X;\beta)\|_p^2\right\} + 8^{-1}\lambda^{-2}n^{-\gamma/2}a_n(\beta) + O_p(n^{-\bar{q}\gamma/2})\right], \tag{34}$$

where the $O_p(n^{-\bar{q}\gamma/2})$ term in the above equation is uniform over $\{(\beta,\lambda) : \|\beta\|_2 \leq b, \lambda > \lambda_0\}$, for any $b, \lambda_0 > 0$, and $\sup_{\lambda>0}\lambda^2 O_p(n^{-\bar{q}\gamma/2})$ is bounded from below by an integral random variable. To solve this minimization, we begin by understanding the solution to the problem $\inf_{\lambda\geq 0} g_1(\lambda)$, where

$$g_1(\lambda) = a\lambda + b/\lambda + c\varepsilon/\lambda^2,$$

where a, b, ε are positive constants and c is non-negative. Changing variable as in $\lambda = (b/a)^{1/2}(1 + \varepsilon u a^{1/2})$ results in,

$$\inf_{\lambda\geq 0} g_1(\lambda) = 2(ab)^{1/2} + \varepsilon acb^{-1} + \varepsilon^2 a^{3/2}\inf_{u\geq -\varepsilon^{-1}a^{-1/2}} g_2(u), \tag{35}$$

where

$$g_2(u) = \frac{b^{1/2}u^2}{1+\varepsilon u a^{1/2}} - \frac{c}{b}\frac{u(2+\varepsilon u a^{1/2})}{(1+\varepsilon u a^{1/2})^2}.$$

Since

$$g_2(u) \geq \frac{b^{1/2}u^2}{1+\varepsilon u a^{1/2}} - \frac{c}{b}\frac{2u}{(1+\varepsilon u a^{1/2})} = \frac{b^{1/2}u^2 - (2c/b)u}{1+\varepsilon u a^{1/2}},$$

for $u \geq 0$, we have that, $\inf_{u\geq 0} g_2(u) = 0$ if $c = 0$ and $\inf_{u \geq 0} g_2(u) < 0$ if $c > 0$. For the case $c > 0$, for all values of $u > 0$ such that $b^{1/2}u^2 - (2c/b)u < 0$ we have $g_2(u) > b^{1/2}u^2 - (2c/b)u$. Since $b^{1/2}u^2 - (2c/b)u$ is lower bounded by $-c^2 b^{-5/2}$ irrespective of the value of u, we have,

$$-c^2 b^{-5/2} \leq \inf_{u \geq -1/(\varepsilon a^{1/2})} g_3(u) \leq 0,$$

for all sufficiently small ε. Moreover, the infimum is attained at $u \geq 0$. Combining this observation with (35), we obtain that,

$$\left| \inf_{\lambda \geq 0} g_1(\lambda) - 2(ab)^{1/2} - \varepsilon ac/b \right| \leq \varepsilon^2 c^2 a^{3/2} b^{-5/2}, \tag{36}$$

for all sufficiently small ε, and the infimum is attained at a choice of $\lambda \geq (b/a)^{1/2}$. Letting $a = \eta_n$, $b = 4^{-1}E_{P_n}\{\|D_x\ell(X;\beta)\|_p^2\}, \varepsilon = n^{-\gamma/2}$ and $c = 8^{-1}a_n(\beta) \geq 0$, we obtain from (36) that,

$$\begin{aligned}
&\inf_{\lambda\geq 0}\left[\lambda\eta_n + \frac{1}{4\lambda}E_{P_n}\left\{\|D_x\ell(X;\beta)\|_p^2\right\} + 8^{-1}\lambda^{-2}n^{-\gamma/2}a_n(\beta)\right] \\
&\quad = \left[\eta_n E_{P_n}\left\{\|D_x\ell(X;\beta)\|_p^2\right\}\right]^{1/2} + 2^{-1}n^{-\gamma/2}\eta_n a_n(\beta)\left[E_{P_n}\left\{\|D_x\ell(X;\beta)\|_p^2\right\}\right]^{-1} + O_p(n^{-\gamma}),
\end{aligned}$$

as $n \to \infty$, and that the limit supremum of the sequence of minimizers which attain the above infimum is positive. Consequently, as $n \to \infty$, we have that (34) equals,

$$\left[\eta_n E_{P_n}\left\{\|D_x\ell(X;\beta)\|_p^2\right\}\right]^{1/2} + 2^{-1}n^{-\gamma/2}b_n a_n(\beta)\left[E_{P_n}\left\{\|D_x\ell(X;\beta)\|_p^2\right\}\right]^{-1} + O_p(n^{-\bar{q}\gamma/2}),$$

due to $\bar{q} \leq 2$ and the tightness of the collection $\{\lambda^2 O_p(n^{-\bar{q}\gamma/2}) : \lambda > 0\}$. Since (34) in turn equals $n^{\gamma/2}(\Psi_n(\beta) - E_{P_n}\{\ell(X;\beta)\})$, we obtain the claim in Proposition 17 by substituting $\eta_n = \delta_n n^\gamma$. □

Proof of Proposition 7. For ease of notation, define $S_n(\beta) = [E_{P_n}\{\|D_x\ell(X;\beta)\|_p^2\}]^{1/2}$. Then it follows from the definitions of $V_n^{DRO}(\cdot), V_n^{ERM}(\cdot)$ and the conclusion in Lemma 17 that,

$$\begin{aligned}
V_n^{DRO}(u) = n^{\bar{\gamma}-1}V_n^{ERM}\{n^{(1-\bar{\gamma})/2}u\} &+ n^{\bar{\gamma}}\delta_n^{1/2}\left\{S_n(\beta_* + n^{-\bar{\gamma}/2}u) - S_n(\beta_*)\right\} \\
&+ \frac{n^{\bar{\gamma}}\delta_n}{2}\left\{\frac{a_n(\beta_* + n^{-\bar{\gamma}/2}u)}{S_n(\beta_* + n^{-\bar{\gamma}/2}u)} - \frac{a_n(\beta_*)}{S_n(\beta_*)}\right\} + o\left(1\right),
\end{aligned} \tag{37}$$

as $n \to \infty$, uniformly over u in compact sets. Since $\bar{\gamma} = \min\{\gamma, 1\}$, due to the twice continuous differentiability of $\ell(\cdot)$, we have that,

$$\begin{aligned}
V_n^{DRO}(u) &= n^{\bar{\gamma}-1}V_n^{ERM}\{n^{(1-\bar{\gamma})/2}u\} + \eta^{1/2}n^{\bar{\gamma}-\gamma/2}\left\{S_n(\beta_* + n^{-\bar{\gamma}/2}u) - S_n(\beta_*)\right\} + o(1) \\
&= n^{(\bar{\gamma}-1)/2}H_n^{\mathrm{T}}u + \frac{1}{2}u^{\mathrm{T}}Cu + \eta^{1/2}n^{(\bar{\gamma}-\gamma)/2}D_\beta S_n(\beta_*)^{\mathrm{T}}u + o(1),
\end{aligned}$$

as $n \to \infty$, uniformly over u in compact sets. Since $D_\beta S_n(\beta_*)$ converges to $D_\beta S(\beta_*)$, combining the above observation with the statement of Proposition 6, we obtain the conclusion of Proposition 7. □

A.2. **Proofs of Propositions 2 - 3.**

Proof of Proposition 2. First, consider the Lagrangian function,

$$L_n(\beta,\lambda) = \Psi_n(\beta) + \sum_{i \in I \cup J} \lambda_i g_i(\beta),$$

and the pointwise maximum function,

$$\Phi_n(\beta) = \max\left\{L_n(\beta,\lambda) : \lambda \in \Lambda_0\right\}.$$

Under the stated Mangasarian-Fromovitz constraint qualification conditions, we have that the set Λ_0 is nonempty, bounded convex polytope; see the discussion following Assumption B.3 in Shapiro [1989]. Therefore, Λ_0 is a convex hull of a finite set of extreme points denoted by Λ_e. Then from the definition of $L(\beta,\lambda)$,

$$\begin{aligned}\Phi_n(\beta) &= \max\left[\Psi_n(\beta) + L(\beta,\lambda) - E\{\ell(X;\beta)\} : \lambda \in \Lambda_e\right] \\ &= \left[\Psi_n(\beta) - E_{P_n}\{\ell(X;\beta)\}\right] + \left[E_{P_n}\{\ell(X;\beta)\} - E\{\ell(X;\beta)\}\right] + \max\left[L(\beta,\lambda) : \lambda \in \Lambda_e\right].\end{aligned}$$

Letting $H_n = -n^{1/2}\left[E_{P_n}\{h(X;\beta_*)\} - E\{h(X;\beta_*)\}\right]$ and taking $S_n(\beta)$ as in the proof of Proposition 7, we obtain the following from the smoothness properties of $\ell(\cdot)$ in Assumptions A2.a, A2.c, expansion for $\Psi_n(\beta)$ in Proposition 17, and its subsequent application in Proposition 7:

$$n\left\{\Phi_n(\beta_* + n^{-1/2}u) - \Phi_n(\beta_*)\right\} = I_n(u) + J_n(u) + K_n(u),$$

where

$$\begin{aligned}I_n(u) &= n\delta_n^{1/2}\left\{S_n(\beta_* + n^{-1/2}u) - S_n(\beta_*)\right\} + o_p(n\delta_n) = \eta^{1/2} D_\beta S_n(\beta_*)^{\mathrm{T}} u + o_p(1), \\ J_n(u) &= n\left[E_{P_n}\{\ell(X;\beta_* + n^{-1/2}u)\} - E_{P_n}\{\ell(X;\beta_*)\}\right] + n\left[E_P\{\ell(X;\beta_* + n^{-1/2}u)\} - E\{\ell(X;\beta)\}\right] \\ &= -H_n^{\mathrm{T}} u + o_p(1) \\ K_n(u) &= \max_{\lambda \in \Lambda_e} L(\beta_* + n^{-1/2}u, \lambda) - \max_{\lambda \in \Lambda_e} L(\beta_*, \lambda) = 2^{-1} q(u) + o(1),\end{aligned}$$

uniformly over compact sets of the variable u. While the simplications for terms $I_n(u), J_n(u)$ are following the obtained same reasoning in the proofs of Propositions 6 - 7, the last equality pertaining to $K_n(u)$ follows from the finiteness of the set Λ_e, taylor expansion for $\max_{\lambda \in \Lambda_e} L(\beta_* + n^{-1/2}u)$ around $u = 0$, and the Kuhn-Tucker optimality condition that $D_\beta L(\beta_*, \lambda) = 0$ for all $\lambda \in \Lambda_e$. Thus,

$$n\left\{\Phi_n(\beta_* + n^{-1/2}u) - \Phi_n(\beta_*)\right\} = \left\{-H_n + \eta^{1/2} D_\beta S_n(\beta_*)\right\}^{\mathrm{T}} u + 2^{-1} q(u) + o_p(1), \tag{38}$$

uniformly in compact sets over the variable u.

Next, we observe that the cone $\mathcal{C}$ of critical directions is nonempty under the second-order sufficient conditions stated in Proposition 2 (see the discussion following Theorem 3.1 in Shapiro [1989]. Following the same lines of the reasoning in [Shapiro, 1989, Lemma 3.1 - 3.3], we have a neighborhood $\mathcal{N}$ of β_* such that if $\beta_n^{DRO}(\delta_n) \in \mathcal{N}$, then

$$\min_{\beta \in B} \Psi_n(\beta) = \min_{\beta \in \mathcal{C}} \Phi_n(\beta), \tag{39}$$

where $\mathcal{C}$ is the critical cone of directions given in the statement of Proposition 2; here, the conditions which are required for applying these results in Shapiro [1989] are verified as follows: The conditions stated in Assumptions A.1, A.4 - A.5, B.4, C.4 and D are direct consequences of the continuous differentiability properties of $\ell(\cdot)$, compactness of B, and finite moments assumed in the statement of Proposition 2 and Assumptions A2.a and A2.c in Section 1. While the conditions in [Shapiro, 1989, Assumptions A.2, A.6] follow from the compactness and aforementioned continuous differentiability properties, the conditions stated in Assumptions A.3, B.1 - B.3, C.5 and D of Shapiro [1989] are explicitly mentioned in the statement of Proposition 2. Now, with the tightness of the collection $n^{1/2}\{\beta_n^{DRO}(\delta_n)-\beta_*\}$ verified as in Proposition 13, we have that the probability of the event $\{\beta_n^{DRO}(\delta_n) \in \mathcal{N}\}$ is $1-o_p(1)$. Therefore, we have from (39) and (38) that,

$$\begin{aligned} n^{1/2}\left\{\beta_n^{DRO}(\delta_n)-\beta_*\right\} &= \arg\min_{u\in\mathcal{C}} \left\{\Phi_n(\beta_*+n^{-1/2}u)-\Phi_n(\beta_*)\right\} \\ &= \arg\min_{u\in\mathcal{C}} \left[\left\{-H_n+\eta^{1/2}D_\beta S_n(\beta_*)\right\}^{\mathrm{T}} u + 2^{-1}q(u) + o_p(1)\right], \end{aligned}$$

with probability $1-o_p(1)$, as $n\to\infty$. As noted earlier, the small $o_p(1)$ term is uniform over compact sets of the variable u. Due to central limit theorem, we have $H_n \Rightarrow H$, where $H \sim \mathcal{N}[0, \mathrm{cov}\{h(X,\beta_*)\}]$. We also have $D_\beta S_n(\beta_*) \to D_\beta S(\beta_*)$, as $n\to\infty$. With the cone $\mathcal{C}$ being nonempty as reasoned above and $\omega(\xi) = \arg\min_{u\in\mathcal{C}}\{u^{\mathrm{T}}\xi + 2^{-1}q(u)\}$ unique, we then obtain

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

as a consequence of argmax/argmin continuous mapping theorem; see van der Vaart et al. [1996, Corollary 3.2.3a]. □

Proof of Proposition 3. Due to the continuous differentiability properties of $\ell(\cdot)$ in Assumption A2.c and the compactness of the set B, we have from [van der Vaart et al., 1996, Theorems 2.7.11 and 2.5.6] that the class $\{\ell(X;\beta) : \beta \in B\}$ is P_*-Donsker. Consequently, we have the uniform central limit theorem that,

$$n^{1/2}\left[E_{P_n}\{\ell(X;\beta)\} - E\{\ell(X;\beta)\}\right] \Rightarrow Z(\beta),$$

as $n\to\infty$, uniformly over continuous functions defined on the set B. Similarly, applying the continuity properties of $\|D_x\ell(X;\beta)\|_p^2$ in Assumption A2.c, we have from van der Vaart et al. [1996, Theorems 2.7.11 and 2.4.1] that

$$\sup_{\beta\in B} |S_n(\beta) - S(\beta)| \to 0$$

as $n\to\infty$. Since $n\delta_n \to \eta$, we obtain by combining the above two convergences that,

$$\delta_n^{-1/2}\left[E_{P_n}\{\ell(X;\beta)\} + \delta_n^{1/2}S_n(\beta) - E\{\ell(X;\beta)\}\right] \Rightarrow \eta^{-1/2}Z(\beta) + S(\beta),$$

uniformly. On the other hand, we have from Proposition 17 that the DRO objective $\Psi_n(\beta)$ and $E_{P_n}[\ell(X;\beta)] + \delta_n^{1/2}S_n(\beta)$ differ only by $O_p(\delta_n)$. Therefore,

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

uniformly. Recall that B_* is the set of minimizers of $\min_{\beta\in B} E\{\ell(X;\beta)\}$. Let us denote the optimal value $\min_{\beta\in B} E\{\ell(X;\beta)\}$ as m. Due to the above uniform convergence and almost sure finiteness of $\sup_{\beta\in B}\left|Z(\beta)+\eta^{1/2}S(\beta)\right|$, given $\varepsilon > 0$, there exists N large enough such that $\min_{\beta\in B_*}\Psi_n(\beta) < m+\varepsilon$ for all $n > N$. Therefore, if the right hand side is singleton almost surely, we have

$$\arg\min_{\beta\in B}\Psi_n(\beta) \Rightarrow \arg\min_{\beta\in B_*}\left\{Z(\beta)+\eta^{1/2}S(\beta)\right\},$$

as $n\to\infty$, as a consequence of the Argmin/argmax continuous mapping theorem; see van der Vaart et al. [1996, Corollary 3.2.3a]. □

A.3. **Statements and proofs of the results in Section 6.**

Proposition 18. *Suppose that the support of X is constrained to be contained in the set $\Omega = \{x \in \mathbb{R}^m : Ax \leq b\}$ specified in terms of linear constraints involving an $l \times m$ matrix A with linearly independent rows and $b \in \mathbb{R}^l$. Consider the Wasserstein distance defined as in Definition 1 with the transportation cost $c(x, y) = \|x - y\|_2^2$. Suppose that $\delta_n = \eta n^{-1}$, 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. Then we have,*

$$n^{1/2}\{\beta_n^{DRO}(\delta_n) - \beta_*\} \Rightarrow C^{-1}H - \eta^{1/2}C^{-1}D_\beta S(\beta_*), \tag{40}$$

as $n \to \infty$.

As in the proof of Proposition 7, we first present the constrained counterpart to Lemma 3 which is useful for the setting considered in Proposition 18.

Lemma 5. *For any $x, v \in \mathbb{R}^m, \lambda > 0, \varepsilon > 0$, $d \times d$ symmetric matrix B, $l \times m$ matrix A, and $b \in \mathbb{R}^l$, we have*

$$\sup_{x:A(x+\varepsilon\Delta)\leq b} \left\{v^{\mathrm{T}}\Delta - \lambda\|\Delta\|_2^2 + \varepsilon\Delta^{\mathrm{T}}B\Delta\right\} = \frac{\|v\|_2^2}{4\lambda} - \frac{1}{4\lambda}\xi^{\mathrm{T}}H\xi, \tag{41}$$

where $\xi = \{2\lambda\varepsilon^{-1}(Ax - b) + A\tilde{B}v\}^+$, $\tilde{B}$ is an $m \times m$ matrix given by the inverse of $(I_m - \varepsilon\lambda^{-1}B)$ with I_m denoting the identity matrix, and H is an $l \times l$ matrix given by the inverse of $A\tilde{B}A^{\mathrm{T}}$.

Proof of Lemma 5. For any $x \in \Omega$, we have $Ax \leq b$. Consequently, the constrained optimization in (41) is feasible for the choice $\Delta = 0$. Then, due to Lagrange's theorem for convex duality, we have that the objective in (41) equals

$$\inf_{\mu\geq 0}\sup_{\Delta\in\mathbb{R}^d}\left[v^{\mathrm{T}}\Delta - \lambda\|\Delta\|_2^2 + \varepsilon\Delta^{\mathrm{T}}B\Delta - \mu^{\mathrm{T}}\left\{A(x+\varepsilon\Delta) - b\right\}\right]$$

$$= \inf_{\mu\geq 0}\left\{-\mu^{\mathrm{T}}(Ax - b) + \sup_{\Delta\in\mathbb{R}^d} f(\Delta,\mu)\right\} \tag{42}$$

where, for any $\mu \geq 0, \Delta \in \mathbb{R}^d$ we define $f(\Delta, \mu)$ as,

$$f(\Delta,\mu) = (v - \varepsilon A^{\mathrm{T}}\mu)^{\mathrm{T}}\Delta - \lambda\|\Delta\|_2^2 + \varepsilon\Delta^{\mathrm{T}}B\Delta. \tag{43}$$

Utilizing the optimality condition that $v - \varepsilon A^{\mathrm{T}}\mu = 2(\lambda + \varepsilon B\Delta)$, we obtain

$$\sup_{\Delta\in\mathbb{R}^d} f(\Delta,\mu) = \frac{1}{4\lambda}(v - \varepsilon A^{\mathrm{T}}\mu)^{\mathrm{T}}\tilde{B}(v - \varepsilon A^{\mathrm{T}}\mu).$$

Then, we obtain from (42) that

$$\inf_{\mu\geq 0}\left\{-\mu^T(Ax - b) + \sup_{\Delta\in\mathbb{R}^d} f(\Delta,\mu)\right\} = \frac{\|v\|_2^2}{4\lambda} + \inf_{\mu\geq 0}\left\{-\mu^{\mathrm{T}}\left(Ax - b + \frac{\varepsilon}{2\lambda}A\tilde{B}v\right) + \frac{\varepsilon}{4\lambda}\mu^{\mathrm{T}}A\tilde{B}A^{\mathrm{T}}\mu\right\}$$

$$= \frac{\|v\|_2^2}{4\lambda} + \frac{\varepsilon}{2\lambda}\inf_{\mu\geq 0}\left(-\mu^{\mathrm{T}}\xi + \frac{\varepsilon}{2}\mu^{\mathrm{T}}A\tilde{B}A^{\mathrm{T}}\mu\right).$$

where $\xi = \{2\lambda\varepsilon^{-1}(Ax - b) + A\tilde{B}v\}^+$ denotes the component-wise positive part. This is because, for any $\mu = (\mu_1, \ldots, \mu_l)$ which attains the infimum in the above left hand side, it is necessarily the case that $\mu_i = 0$ whenever the respective $\xi_i < 0$ for any $i = 1, \ldots, l$. Moreover,

$$\inf_{\mu\geq 0}\left(-\mu^{\mathrm{T}}\xi + \frac{\varepsilon}{2}\mu^{\mathrm{T}}A\tilde{B}A^{\mathrm{T}}\mu\right) = \inf_{\mu\in\mathbb{R}^l}\left(-\mu^{\mathrm{T}}\xi + \frac{\varepsilon}{2}\mu^{\mathrm{T}}A\tilde{B}A^{\mathrm{T}}\mu\right),$$

because of the following reasoning: $\xi \geq 0$ component-wise and if any $\mu = (\mu_1, \ldots, \mu_l)$ which attains the optimum in the right-hand side is such that $\mu_i < 0$ for some i, then one can strictly decrease the objective by increasing μ_i if the respective $\xi_i > 0$, (or) not change the objective by making $\mu_i = 0$. Consequently,

$$\inf_{\mu \geq 0} \left(-\mu^{\mathrm{T}} \nu + \frac{\varepsilon}{2} \mu^{\mathrm{T}} A \tilde{B} A^{\mathrm{T}} \mu \right) = \inf_{\mu \in \mathbb{R}^m} \left(-\mu^{\mathrm{T}} \xi + \frac{\varepsilon}{2} \mu^{\mathrm{T}} A \tilde{B} A^{\mathrm{T}} \mu \right)$$
$$= -2^{-1} \varepsilon^{-1} \xi^{\mathrm{T}} \left(A \tilde{B} A^{\mathrm{T}} \right)^{-1} \xi,$$

because A is taken to have linearly independent rows and the respective optimality condition is $\xi - \varepsilon A \tilde{B} A^T \mu = 0$. Therefore, we have from the Lagrange duality, (42) and the above simplication that the objective in (41) equals $(4\lambda)^{-1}(\|v\|_2^2 - \xi^{\mathrm{T}} H \xi)$, thus concluding the proof. □

Proof of Proposition 18. Due to the presence of the constraints $\Omega = \{x \in \mathbb{R}^m : Ax \leq b\}$, we have $\Psi_n(\beta)$ as in the statement of Lemma 1 with $e_n(\beta, \lambda) = E_{P_n}[f_n(X, \beta, \lambda)]$ and

$$f_n(x, \beta, \lambda) = \sup_{x: A(x + n^{-1/2}\Delta) \leq b} \left[n^{1/2} \left\{ \ell\big(x + n^{-1/2}\Delta; \beta\big) - \ell(x; \beta) \right\} - \lambda \|\Delta\|_q^2 \right] - \frac{1}{4\lambda} \|D_x \ell(x; \beta)\|_2^2.$$

Fixing $b > 0$ and $\lambda_0 \in (0, 1)$, consider any β such that $\|\beta\|_2 \leq b$ and $\lambda > \lambda_0$. To apply Lemma 5 for evaluating $f_n(x, \beta, \lambda)$ as in the proof of Lemma 4, we identify the respective quantities in (41) in the statement of Lemma 5 as follows: Letting $\varepsilon = n^{-1/2}$, $v = D_x \ell(x; \beta)$, $\bar{q} = \min\{2, q/(q-1)\}$, $B = D_{xx} \ell(x; \beta)$, $H_n(x, \beta, \lambda)$ be the inverse of $A \left\{ I_m - n^{-1/2} \lambda^{-1} D_{xx} \ell(X; \beta) \right\}^{-1} A^{\mathrm{T}}$ and

$$\xi_n(x, \beta, \lambda) = \left[2\lambda n^{1/2}(Ax - b) + A \left\{ I_m - n^{-1/2} \lambda^{-1} D_{xx} \ell(X; \beta) \right\}^{-1} D_x \ell(x; \beta) \right]^+$$

we have that $f_n(x, \beta, \lambda) - (4\lambda)^{-1} \xi_n(x, \beta, \lambda)^{\mathrm{T}} H_n(x, \beta, \lambda) \xi_n(x, \beta, \lambda)$ is upper and lower bounded, respectively, by

$$c_u (1 + \varepsilon') n^{-\bar{q}/2} \|\xi_n(x, \beta, \lambda)\|_2^2 \, \lambda_0^{-\max\{2, \frac{1}{q-1}\}} \quad \text{and} \quad -c_l \varepsilon' n^{-1} \|\xi_n(x, \beta, \lambda)\|_2^2 \lambda^{-2},$$

for suitable positive constants c_l, c_u which are, in turn, determined by the constants b, d and q.

Next, with Ω being compact, we have from the expression for $\xi_n(\cdot)$ and the uniform boundedness of $D_x \ell(x, \beta), D_{xx} \ell(x, \beta)$ (over the set $x \in \Omega, \|\beta\| \leq b$) that,

$$\mathrm{pr}\left\{ \|\xi_n(X, \beta, \lambda)\|_2 > 0 \right\} \leq \mathrm{pr}\left[\min_{i=1,\ldots,l} \{b_i - (Ax)_i\} < M \lambda^{-1} n^{-1/2} \right],$$

for some suitably large constant M. The above right hand side is $O_p(\lambda n^{-1/2})$, as $n \to \infty$, since the distribution X is absolutely continuous and satisfies $\mathrm{pr}(AX \leq b) = 1$. Then, letting

$$a_n(\beta, \lambda) = \lambda^{-1} n^{1/2} E_{P_n} \left\{ \xi_n(X, \beta, \lambda)^{\mathrm{T}} H_n(X, \beta, \lambda) \xi_n(X, \beta, \lambda) \right\},$$

we have $\sup_{n, \lambda > \lambda_0, \|\beta\|_2 \leq b} a_n(\beta, \lambda) < \infty$ due to the uniform boundedness of $\xi_n(x, \beta, \lambda)$ over $n \geq 1, x \in \Omega, \|\beta\| \leq b$ and $\lambda > \lambda_0$. With $e_n(\beta, \lambda)$ defined to equal $E_{P_n}[f_n(X, \beta, \lambda)]$, we therefore obtain,

$$e_n(\beta, \lambda) = 4^{-1} \lambda^{-2} n^{-1/2} a_n(\beta, \lambda) + O_p(n^{-\bar{q}/2}),$$

where the convergence pertaining to the $O_p(\cdot)$ term is uniform over (β, λ) such that $\|\beta\|_2 \leq b$ and $\lambda > \lambda_0$. Likewise, due to the above lower bound for $f_n(\cdot)$, the $O_p(n^{-\bar{q}1/2})$ term $\lambda^2 O_p(n^{-\bar{q}/2})$ is bounded from below by an integral random variable, uniformly over all $\lambda > 0$ and $\|\beta\| \leq b$. Furthermore, due to continuous differentiability of $\ell(\cdot)$ over compact Ω, we have that $a_n(\beta, \lambda)$ is lipschitz over $\lambda >$

$\lambda_0, \|\beta\|_2 \leq b$. Combining this with the above expression for $e_n(\beta,\lambda)$ and that of $\Psi_n(\beta)$ derived from Lemma 1, we have, $n^{1/2}\left[\Psi_n(\beta) - E_{P_n}\{\ell(X;\beta)\}\right]$ equals,

$$\inf_{\lambda\geq 0}\left[\lambda\eta + \frac{1}{4\lambda}E_{P_n}\left\{\|D_x\ell(X;\beta)\|_p^2\right\} + \frac{a_n(\beta,\lambda)}{4\lambda^2 n^{1/2}} + O_p(n^{-\bar{q}/2})\right].$$

The desired conclusion then follows by utilizing the uniform boundedness, lipschitzness of $a_n(\beta,\lambda)$ and proceeding as in the proofs of Propositions 17 and 7 given earlier in this supplementary material. □

The following examples show that the convergence (40) may not hold if the set $\Omega = \{x \in \mathbb{R}^m : Ax \leq b\}$ has equality constraints.

Example 2. *For the linear regression example in Section 4, suppose that the support for X, denoted by the set $\Omega = \{x \in \mathbb{R}^2 : Ax \leq b\}$, where the matrix A and vector b are such that*

$$\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 - x_2 = 0\}.$$

Suppose that $\delta_n = \eta n^{-1}$. With the loss $\ell(x,y;\beta) = (y-\beta^{\mathrm{T}}x)^2$ and the transportation cost $c(\cdot)$ given as in (14)*, we have the following from the definition of $\phi_\lambda(\cdot)$ in the proof of Lemma 1: for any $x = (x_1,x_2) \in \Omega$, with x_1 being equal to x_2,*

$$\begin{aligned}\phi_\lambda(x;\beta,\lambda) &= \sup_{\Delta\in\mathbb{R}}\left\{\left(y - \beta^{\mathrm{T}}x - n^{-1/2}\Delta\beta^{\mathrm{T}}1\right)^2 - \lambda n^{-1/2}2^{2/q}\Delta^2\right\}\\ &= (y-\beta^{\mathrm{T}}x)^2 + n^{-1/2}\sup_{\Delta\in R}\left[-2(y-\beta^{\mathrm{T}}x)\beta^{\mathrm{T}}1\Delta - \left\{\lambda 2^{2/q} - (\beta^{\mathrm{T}}1)^2 n^{-1/2}\right\}\Delta^2\right]\\ &= (y-\beta^{\mathrm{T}}x)^2 + n^{-1/2}\frac{(y-\beta^{\mathrm{T}}x)^2}{\lambda 2^{2/q}(\beta^{\mathrm{T}}1)^{-2} - n^{-1/2}} = \frac{(y-\beta^{\mathrm{T}}x)^2}{1-\lambda^{-1}2^{-2/q}(\beta^{\mathrm{T}}1)^2 n^{-1/2}}.\end{aligned}$$

For the choice $\delta_n = \eta n^{-1}$, the distributionally robust optimization objective simplifies as below by exploiting the dual representation for $\Psi_n(\beta)$ in Lemma 1:

$$\begin{aligned}\Psi_n(\beta) &= \inf_{\lambda\geq 0}\left\{\lambda\eta n^{-1/2} + \frac{E_{P_n}(Y-\beta^{\mathrm{T}}X)^2}{1-\lambda^{-1}2^{-2/q}(\beta^{\mathrm{T}}1)^2 n^{-1/2}}\right\}\\ &= E_{P_n}(Y-\beta^{\mathrm{T}}X)^2 + n^{-1/2}\inf_{\mu\geq 0}\left\{\eta\mu + \mu^{-1}2^{-2/q}(\beta^{\mathrm{T}}1)^2 E_{P_n}(Y-\beta^{\mathrm{T}}X)^2\right\} + n^{-1}\eta 2^{-2/q}(\beta^{\mathrm{T}}1)^2\\ &= E_{P_n}(Y-\beta^{\mathrm{T}}X)^2 + n^{-1/2}2^{1-1/q}\eta^{1/2}|\beta^{\mathrm{T}}1|\{E_{P_n}(Y-\beta^{\mathrm{T}}X)^2\}^{1/2} + n^{-1}\eta 2^{-2/q}(\beta^{\mathrm{T}}1)^2.\end{aligned}$$

Suppose that β_*, denoting an optimal parameter minimizing $E\{(Y-\beta^{\mathrm{T}}X)^2\}$, is such that $\beta_*^{\mathrm{T}}1 \neq 0$. Then

$$n\left\{\Psi_n(\beta_* + n^{-1/2}u) - \Psi_n(\beta_*)\right\} = H_n^{\mathrm{T}}u + u^T E_{P_n}(XX^{\mathrm{T}})\,u + \eta^{1/2}D_\beta\tilde{S}(\beta_*)^{\mathrm{T}}u + \eta 2^{-2/q}(\beta_*^{\mathrm{T}}1)^2 + o(1),$$

where $H_n = -n^{1/2}E_{P_n}\{2(Y-\beta_*^{\mathrm{T}}X)X\}$ and $\tilde{S}(\beta) = 2^{1-1/q}|\beta^{\mathrm{T}}1|\{E_{P_n}(Y-\beta^{\mathrm{T}}X)^2\}^{1/2}$. The above convergence happens uniformly in compact sets over u and as $n\to\infty$. Consequently, when $C = E[XX^{\mathrm{T}}]$ is positive definite, we have the the following central limit theorem for the distributionally robust estimator $\beta_n^{DRO}(\delta_n)$ incorporating support constraint: As $n\to\infty$,

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

where H is normally distributed as in Theorem 1. Comparing this limiting result with that in Theorem 1, we see that the limit has changed with the introduction of support constraints via the term $D_\beta\tilde{S}(\beta)$, *instead of $D_\beta S(\beta)$ appearing in Theorem 1. In particular, we see that the terms $S(\beta)$ and $\tilde{S}(\beta)$ differ as in,*

$$\tilde{S}(\beta_*) = 2^{1/2-1/q}\frac{|\beta^T 1|}{\|\beta\|_p}S(\beta).$$

Example 3. *Suppose that $\ell(x;\beta) = a + \beta^{\mathrm{T}}x + \beta^{\mathrm{T}}C\beta$ for some $a \in \mathbb{R}$ and positive semi-definite C. Let $r \le m$ be a positive integer and the support for X be given by $\Omega = \{x \in \mathbb{R}^m : Ax = b\}$, where the matrix A is an $(r \times m)$ matrix with linearly independent rows and $b \in \mathbb{R}^r$. Suppose that $\delta_n = \eta n^{-1}$. With the transportation cost $c(\cdot)$ given by $c(x,x') = \|x - x'\|_2^2$, we have the following from the definition of $\phi_\lambda(\cdot)$ in the proof of Lemma 1: for any $x \in \Omega$, we have $Ax = b$ and*

$$\begin{aligned}\phi_\lambda(x;\beta,\lambda) &= \ell(x;\beta) + n^{-1/2}\sup_{\Delta}\left\{\beta^{\mathrm{T}}\Delta - \lambda\|\Delta\|_2^2 : A(x + n^{-1/2}\Delta) = b\right\}\\ &= \ell(x;\beta) + n^{-1/2}\sup_{\Delta}\left\{\beta^{\mathrm{T}}\Delta - \lambda\|\Delta\|_2^2 : A\Delta = 0\right\}\\ &= \ell(x;\beta) + n^{-1/2}\inf_{\mu\in\mathbb{R}^r}\sup_{\Delta}\left\{(\beta - A^{\mathrm{T}}\mu)^{\mathrm{T}}\Delta - \lambda\|\Delta\|_2^2\right\},\end{aligned}$$

as a consequence of convex duality. Then

$$\phi_\lambda(x;\beta,\lambda) = \ell(x;\beta) + \frac{n^{-1/2}}{4\lambda}\inf_{\mu\in\mathbb{R}^r}\|\beta - A^{\mathrm{T}}\mu\|_2^2 = \ell(x;\beta) + \|(\mathbb{I}_m - A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}A)\beta\|_2^2,$$

where I_m is the $m \times m$ identity matrix. For the choice $\delta_n = \eta n^{-1}$, we obtain the following from the dual representation in Lemma 1:

$$\begin{aligned}\Psi_n(\beta) &= E_{P_n}\{\ell(X;\beta)\} + \inf_{\lambda\ge 0}\left\{\lambda\eta n^{-1/2} + \frac{n^{-1/2}}{4\lambda}\|(\mathbb{I}_m - A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}A)\beta\|_2^2\right\}\\ &= E_{P_n}\{\ell(X;\beta)\} + \eta^{1/2}n^{-1/2}\|(\mathbb{I}_m - A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}A)\beta\|_2\\ &= E_{P_n}\{\ell(X;\beta)\} + \delta_n^{1/2}\|P_{\mathcal{N}(A)}\beta\|_2,\end{aligned}$$

where $P_{\mathcal{N}(A)} = I_m - A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}A$ is the matrix for projecting onto the null space of A. Letting $H_n = n^{1/2}E_{P_n}\{h(X;\beta)\}$ and $\tilde{S}(\beta) = E_{P_n}\left\{\|P_{\mathcal{N}(A)}\beta\|_2^2\right\}^{1/2}$,

$$n\left\{\Psi_n(\beta_* + n^{-1/2}u) - \Psi_n(\beta_*)\right\} = H_n^{\mathrm{T}}u + u^TCu + \eta^{1/2}D_\beta\tilde{S}(\beta_*)^{\mathrm{T}}u + +o(1),$$

as $n \to \infty$ and uniformly in compact sets over u. Consequently,

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

where H is normally distributed as in Theorem 1. With $S(\beta) = \|\beta\|_2$ in this example, we see that the introduction of support constraint results in a bias term that differs from that in Theorem 1 by,

$$\tilde{S}(\beta) = \frac{\|P_{\mathcal{N}(A)}\beta\|}{\|\beta\|_2}S(\beta),$$

where $P_{\mathcal{N}(A)}$ is the projection matrix for projecting onto the null space of the matrix A.

APPENDIX B. PROOFS OF PROPOSITIONS 8 - 12

In this section we present the proofs of Propositions 8 - 12, which are useful towards establishing the convergence of the last component of the triple considered in Theorem 3.

Proof of Proposition 8. By utilizing the duality for linear semi-infinite programs as in the proof of Proposition 3 of Blanchet et al. [2019], for $\beta_* + n^{-1/2}u \in \Theta$, we obtain that

$$nR_n(\beta_* + n^{-1/2}u) = \max_{\xi} \Bigg(-\sum_{i=1}^{n} \xi^{\mathrm{T}} h(X_i, \beta_* + n^{-1/2}u) - \sum_{i=1}^{n} \max_{\Delta: X_i+\Delta \in \Omega} \left[\xi^{\mathrm{T}} \left\{ h(X_i + \Delta, \beta_* + n^{-1/2}u) - h(X_i, \beta_* + n^{-1/2}u) \right\} - \|\Delta\|_q^2 \right] \Bigg).$$

As a result,

$$nR_n(\beta_* + n^{-1/2}u) = \max_{\xi} \left[-\sum_{i=1}^{n} \max_{\Delta: X_i+\Delta \in \Omega} \left\{ \xi^{\mathrm{T}} h(X_i + \Delta, \beta_* + n^{-1/2}u) - \|\Delta\|_q^2 \right\} \right]$$
$$= \max_{\xi} \Bigg(-\sum_{i=1}^{n} \xi^{\mathrm{T}} h(X_i, \beta_*) - \sum_{i=1}^{n} \max_{X_i+\Delta \in \Omega} \left[\xi^{\mathrm{T}} \left\{ h(X_i + \Delta, \beta_* + n^{-1/2}u) - h(X_i, \beta_*) \right\} - \|\Delta\|_q^2 \right] \Bigg).$$

By rescaling $\xi = n^{1/2}\xi, \Delta = n^{1/2}\Delta$ and letting $H_n = n^{-1/2} \sum_{i=1}^{n} h(X_i, \theta_*)$, we obtain,

$$nR_n(\beta_* + n^{-1/2}u) = \max_{\xi} \left\{ -\xi^{\mathrm{T}} H_n - M_n(\xi, u) \right\},$$

where

$$\begin{aligned} &M_n(\xi, u) \\ =&\frac{1}{n} \sum_{i=1}^{n} \max_{\Delta: X_i+n^{-1/2}\Delta \in \Omega} \left[n^{1/2} \xi^{\mathrm{T}} \left\{ h(X_i + n^{-1/2}\Delta, \beta_* + n^{-1/2}u) - h(X_i, \beta_*) \right\} - \|\Delta\|_q^2 \right] \qquad (44) \\ =&\frac{1}{n} \sum_{i=1}^{n} \max_{\Delta: X_i+n^{-1/2}\Delta \in \Omega} \left\{ \xi^{\mathrm{T}} \int_0^1 D_x h \left(X_i + n^{-1/2} t\Delta, \beta_* + n^{-1/2} tu \right) \Delta \mathrm{d}t \right. \\ &\left. + \xi^{\mathrm{T}} \int_0^1 D_\beta h \left(X_i + n^{-1/2} t\Delta, \beta_* + n^{-1/2} tu \right) u \mathrm{d}t - \|\Delta\|_q^2 \right\}, \end{aligned}$$

where the latter equality follows from the fundamental theorem of calculus. This completes the proof of the first part of Proposition 8.

For the second part, we first show $\beta_* \in \Theta$. For any non-zero $\xi \in \mathbb{R}^d$, we have $E\left\{\xi^{\mathrm{T}} h(X, \beta_*)\right\} = 0$, due to Assumption A2.b. We claim 0 lies in the interior of $\operatorname{conv}\left(\{\xi^{\mathrm{T}} h(x, \beta_*), x \in \Omega\}\right)$. Otherwise, we must have $h(X, \beta^*) = 0$, almost surely, and $\xi^{\mathrm{T}} h(x, \beta^*)$ have the same sign, for all $x \in \Omega$. Without loss of generality, we assume $\xi^{\mathrm{T}} h(x, \beta^*) \geq 0$ for all $x \in \Omega$. Then, we have $D_x \left\{\xi^{\mathrm{T}} h(x, \beta^*)\right\} = 0$, almost surely, which leads to a contradiction to $E\left\{D_x h(X, \beta_*) D_x h(X, \beta_*)^{\mathrm{T}}\right\} \succ 0$. Therefore, there exists $\underline{x}_\xi, \overline{x}_\xi \in \Omega$ such as

$$\xi^{\mathrm{T}} h(\underline{x}_\xi, \beta_*) < 0 < \xi^{\mathrm{T}} h(\overline{x}_\xi, \beta_*).$$

If $\beta_* \notin \Theta$, which means 0 lies on the boundary of $\operatorname{conv}\left[\{h(x, \beta_*), x \in \Omega\}\right]$, by applying the supporting hyperplane theorem; see, for example, Boyd et al. [2004, section 2.5.2], there exists a non-zero ξ such that for all $x \in \Omega$,

$$\xi^{\mathrm{T}} h(x, \beta_*) \leq 0,$$

which leads to a contradiction.

Since $\beta_* \in \Theta$, there exists $\epsilon > 0$ such as $B_\epsilon(0) \subset \operatorname{conv}[\{h(x,\beta_*), x \in \Omega\}]$. Consider basis points $e_i = (0, \ldots, 1, \ldots, 0)^{\mathrm{T}}$, whose coordinates are all zero, except the i-th entry that equals one. So, $\operatorname{conv}\left[\{\epsilon e_i\}_{i=1}^d \cup \{-\epsilon e_i\}_{i=1}^d\right] \subset \operatorname{conv}[\{h(x,\beta_*), x \in \Omega\}]$ is a neighborhood of 0. To simplify the notation, let $y_i = \epsilon e_i$ for $i = 1, 2, \ldots, d$, and $y_i = -\epsilon e_{i-d}$ for $i = d+1, d+2, \ldots, 2d$. By Carathéodory's theorem; see, for example, Rockafellar [1970, Theorem 17.1], we have for each y_i, there exists $x_{i,1} \ldots x_{i,d+1}$ such that y_i is a convex combination of $h(x_{i,1}, \beta) \ldots h(x_{i,d+1}, \beta)$. Then, due to the continuity of $D_\beta h(x,\beta)$ around β_* in Assumption A2.c, there exists a neighborhood of β_*, $B_\epsilon(\beta_*)$, such that for all $\beta \in B_\epsilon(\beta_*)$, for $i = 1, 2, \ldots, 2d$ and $j = 1, 2, \ldots, d+1$,

$$\|h(x_{i,j}, \beta) - h(x_{i,j}, \beta)\|_2 < \epsilon/2.$$

Then by applying the same convex combination to obtain y_i^β, we have for all $i = 1, 2, \ldots, 2d$, $\left\|y_i^\beta - y_i\right\|_2 < \epsilon/2$. Therefore, $\operatorname{conv}\left(\left\{y_i^\beta\right\}_{i=1}^{2d}\right) \subset \operatorname{conv}[\{h(x,\beta), x \in \Omega\}]$ is a neighborhood of 0, which completes the proof. □

A key component of the proofs of the upper and lower bounds for $nR_n(\beta_* + n^{-1/2}u)$ is the following tightness result.

Lemma 6. *For any $\varepsilon, K > 0$, there exists $n_0 > 0$ and $b \in (0, \infty)$ such that*

$$\operatorname{pr}\left[\max_{\|\xi\|_q \geq b} \{-\xi^{\mathrm{T}} H_n - M_n(\xi, u)\} > 0\right] \leq \varepsilon,$$

for all $n \geq n_0$ and uniformly over u such that $\|u\|_2 \leq K$.

Lemma 7. *For any positive constants b, c_0 and any bounded set $C \in \mathbb{R}^d$, we have*

$$\frac{1}{n}\sum_{i=1}^n \left[\left\|\{D_x h(X,\beta_*)\}^{\mathrm{T}} \xi\right\|_p^2 + \xi^{\mathrm{T}} D_\beta h(X_i,\beta_*) u\right] \mathbb{I}(X_i \in C)$$
$$\to E\left(\left[\left\|\{D_x h(X,\beta_*)\}^{\mathrm{T}} \xi\right\|_p^2 + \xi^{\mathrm{T}} D_\beta h(X_i,\beta_*) u\right] \mathbb{I}(X_i \in C)\right),$$

uniformly over $\|\xi\|_q \leq b$ and $\|u\|_2 \leq K$ in probability as $n \to \infty$.

Proofs of Lemmas 6 and 7 are presented in Section F. The following definitions are useful in the proofs of Proposition 9 and Lemma 6. For a fixed u, Δ, let

$$I(X_i, \Delta, u) = I_1(X_i, \Delta, u) + I_2(X_i, \Delta, u), \tag{45}$$

where $i \in \{1, \ldots, n\}$,

$$I_1(X_i, \Delta, u) = \int_0^1 \left\{D_x h\left(X_i + n^{-1/2} t\Delta, \beta_* + n^{-1/2} tu\right) - D_x h(X_i, \beta_*)\right\} \Delta \mathrm{d}t \quad \text{and}$$
$$I_2(X_i, \Delta, u) = \int_0^1 \left\{D_\beta h\left(X_i + n^{-1/2} t\Delta, \beta_* + n^{-1/2} tu\right) - D_\beta h(X_i, \beta_*)\right\} u \mathrm{d}t.$$

Then, we have

$$M_n(\xi, u) = \frac{1}{n}\sum_{i=1}^n \left[\xi^{\mathrm{T}} D_\beta h(X_i, \beta_*) u + \max_{\Delta: X_i + n^{-1/2}\Delta \in \Omega} \left\{\xi^{\mathrm{T}} D_x h(X_i, \beta_*) \Delta + \xi^{\mathrm{T}} I(X_i, \Delta, u) - \|\Delta\|_q^2\right\}\right].$$

In addition, for $\xi \neq 0$, we write $\bar{\xi} = \xi / \|\xi\|_p$. Let us define the vector $V_i(\bar{\xi}) = D_x h(X_i, \beta_*)^{\mathrm{T}} \bar{\xi}$ and put

$$\Delta_i' = \Delta_i'(\bar{\xi}) = \begin{cases} |V_i(\bar{\xi})|^{p/q} \mathrm{sgn}\{V_i(\bar{\xi})\} & q \in (1, \infty) \\ V_i(\bar{\xi}) \mathbb{I}[|V_i(\bar{\xi})| = \max_j \{|V_j(\bar{\xi})|\}] & q = 1 \\ \mathrm{sgn}\{V_i(\bar{\xi})\} & q = \infty. \end{cases} \tag{46}$$

Proof of Proposition 9. First observe that $R_n(\cdot) \geq 0$ (consider the choice $\xi = 0$). Given $K, \varepsilon > 0$, define the event,

$$\mathcal{A}_n = \left\{ nR_n\left(\beta_* + n^{-1/2}u\right) = \max_{\|\xi\|_p \leq b} \{-\xi^{\mathrm{T}} H_n - M_n(\xi, u)\} \text{ for all } u \text{ such that } \|u\|_2 \leq K \right\}.$$

where $b > 0$ is such that $\mathrm{pr}(\mathcal{A}_n) \geq 1 - \varepsilon$ for $n \geq n'$. Such a $b \in (0, \infty)$ exists because of Lemma 6 and the fact that the set $\{\beta_* + n^{-1/2}u \mid \|u\|_2 \leq K\}$ will eventually become a subset of Θ when n is sufficiently large.

Next, for any $c_0 > 0, \epsilon_0 > 0$ define

$$M_n'(\xi, u, c_0, \epsilon_0) = \frac{1}{n} \sum_{i=1}^n \{\xi^{\mathrm{T}} D_x h(X, \beta_*) \bar{\Delta}_i - \|\bar{\Delta}_i\|_q^2 + \xi^{\mathrm{T}} I(X_i, \bar{\Delta}_i, u) + \xi^{\mathrm{T}} D_\beta h(X, \beta_*) u\} \, \mathbb{I}\left(X_i \in C_0^{\epsilon_0}\right),$$

where $C_0 = \{w \in \Omega : \|w\|_p \leq c_0\}$, $I(X_i, \Delta, u)$ is defined as in (45) and $\bar{\Delta}_i = c_i \Delta_i'$, which is defined in (46) with c_i chosen so that

$$\left\|\bar{\Delta}_i\right\|_q = \frac{1}{2} \left\|D_x h(X_i, \beta_*)^{\mathrm{T}} \xi\right\|_p.$$

Since $D_x h(X_i, \beta_*)$ is continuous, $\|\xi\|_p$ is bounded, and C_0 is compact, we have

$$\sup_{x \in C_0} \left\{ \frac{1}{2} \left\|D_x h(X_i, \beta_*)^{\mathrm{T}} \xi\right\|_p \right\} < \infty.$$

Therefore, there exists $n_1 > 0$ such that for all $n \geq n_1$ and $X_i \in C_0^{\epsilon_0}$, we have $X_i + n^{-1/2} \bar{\Delta}_i \in C_0$, and thus $M_n(\xi, u) \geq M_n'(\xi, u, c_0, \epsilon_0)$, for every u and $n \geq n_1$. With these definitions, observe that

$$\begin{aligned} \max_{X_i + n^{-1/2}\Delta \in \Omega} \left\{ \xi^{\mathrm{T}} D_x h(X_i, \beta_*) \Delta - \|\Delta\|_q^2 \right\} &= \xi^{\mathrm{T}} D_x h(X_i, \beta_*) \bar{\Delta}_i - \|\bar{\Delta}_i\|_q^2 \\ &= \frac{1}{4} \left\| \{D_x h(X, \beta_*)\}^{\mathrm{T}} \xi \right\|_p^2 \end{aligned} \tag{47}$$

and

$$\max_{\|\xi\|_q \leq b} \{-\xi^{\mathrm{T}} H_n - M_n(\xi, u)\} \leq \max_{\|\xi\|_q \leq b} \left\{-\xi^{\mathrm{T}} H_n - M_n'(\xi, u, c_0, \epsilon_0)\right\}. \tag{48}$$

Next, define

$$\begin{aligned} \hat{M}_n(\xi, u, c_0, \epsilon_0) &= \frac{1}{n} \sum_{i=1}^n \left\{ \xi^{\mathrm{T}} D_x h(X_i, \beta_*) \bar{\Delta}_i - \left\|\bar{\Delta}_i\right\|_q^2 + \xi^{\mathrm{T}} D_\beta h(X_i, \beta_*) u \right\} \mathbb{I}(X_i \in C_0^{\epsilon_0}) \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{4} \left\|D_x h(X_i, \beta_*)^{\mathrm{T}} \xi\right\|_p^2 + \xi^{\mathrm{T}} D_\beta h(X_i, \beta_*) u \right\} \mathbb{I}(X_i \in C_0^{\epsilon_0}), \end{aligned}$$

where the equality follows from (47). Due to Lemma 7, we have

$$\hat{M}_n(\xi, u, c_0, \epsilon_0) \to E\left(\left[\frac{1}{4} \left\| \{D_x h(X, \beta_*)\}^{\mathrm{T}} \xi \right\|_p^2 + \xi^{\mathrm{T}} D_\beta h(X, \beta_*) u \right] \mathbb{I}(X \in C_0^{\epsilon_0}) \right).$$

in probability, uniformly over $\|\xi\|_p \leq b$ and $\|u\|_2 \leq K$. Furthermore,

$$\sup_{\|\xi\|_p \leq b} \left| \hat{M}_n(\xi, u, c_0, \epsilon_0) - M_n'(\xi, u, c_0, \epsilon_0) \right| \to 0, \tag{49}$$

because, from the uniform continuity of $D_\beta h(\cdot)$ and $D_x h(\cdot)$ in compact sets, we have that

$$|\xi^{\mathrm{T}} I(X_i, \bar{\Delta}_i, u)| \mathbb{I}(X_i \in C_0^{\epsilon_0}) \to 0, \tag{50}$$

uniformly over $\|\xi\|_p \leq b$ and $\|u\|_2 \leq K$. Combining the observations in (49) and (50), we obtain that for any $\varepsilon' > 0$ there exists $n_0 \geq n_1$ sufficiently large such that,

$$\begin{aligned} &\max_{\|\xi\|_p \leq b} \left\{ -\xi^{\mathrm{T}} H_n - M_n'(\xi, u, c_0, \epsilon_0) \right\} \\ &\quad \leq \max_{\|\xi\|_p \leq b} \left\{ -\xi^{\mathrm{T}} H_n - E\left(\left[\frac{1}{4} \left\| \{D_x h(X, \beta_*)\}^{\mathrm{T}} \xi \right\|_p^2 + \xi^{\mathrm{T}} D_\beta h(X, \beta_*) u \right] \mathbb{I}(X \in C_0^{\epsilon_0}) \right) \right\} + \varepsilon'. \end{aligned}$$

Then the statement of Proposition 9 follows from (48), the definition of the event $\mathcal{A}_n$ and the observation that $\mathrm{pr}(\mathcal{A}_n) \geq 1 - \varepsilon$. □

Proof of Proposition 10. For the lower bound, we reexpress the expression for $M_n(\xi, u)$ in (44) as follows:

$$\begin{aligned} M_n(\xi, u) \leq& \frac{1}{n} \sum_{i=1}^n \max_{\Delta \in \mathbb{R}^d} \left[n^{1/2} \xi^{\mathrm{T}} \left\{ h(X_i + n^{-1/2}\Delta, \beta_* + n^{-1/2}u) - h(X_i, \beta_* + n^{-1/2}u) \right\} - \|\Delta\|_q^2 \right] \\ &+ \frac{1}{n} \sum_{i=1}^n n^{1/2} \xi^{\mathrm{T}} \left\{ h(X_i, \beta_* + n^{-1/2}u) - h(X_i, \beta_*) \right\}. \end{aligned} \tag{51}$$

Employing the fundamental theorem of calculus, we obtain that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n n^{1/2} \xi^{\mathrm{T}} \left\{ h(X_i, \beta_* + n^{-1/2}u) - h(X_i, \beta_*) \right\} &= \frac{1}{n} \sum_{i=1}^n \int_0^1 \xi^{\mathrm{T}} D_\beta h(X_i, \beta_* + t n^{-1/2} u) u \mathrm{d}t \\ = \xi^{\mathrm{T}} \left\{ \frac{1}{n} \sum_{i=1}^n D_\beta h(X_i, \beta_*) \right\} u &+ \frac{1}{n} \sum_{i=1}^n \int_0^1 \xi^{\mathrm{T}} \left\{ D_\beta h(X_i, \beta_* + t n^{-1/2} u) - D_\beta h(X_i, \beta_*) \right\} u \mathrm{d}t \\ \leq \xi^{\mathrm{T}} \left\{ \frac{1}{n} \sum_{i=1}^n D_\beta h(X_i, \beta_*) \right\} u &+ \|\xi\|_p \frac{1}{n} \sum_{i=1}^n \int_0^1 \left\| \left\{ D_\beta h(X_i, \beta_* + t n^{-1/2} u) - D_\beta h(X_i, \beta_*) \right\} u \right\|_q \mathrm{d}t. \end{aligned}$$

Then, given $\varepsilon, \varepsilon' > 0$, due to continuity of $D_\beta h(\cdot)$ in Assumption A2.c, finiteness $E\{\bar{\kappa}(X_i)\}$ and the law of large numbers, there exists n_0 sufficiently large such that for all $n \geq n_0$, $\|\xi\|_p \leq b$, $\|u\|_2 \leq K$, we have

$$\frac{1}{n} \sum_{i=1}^n n^{1/2} \xi^{\mathrm{T}} \left\{ h(X_i, \beta_* + n^{-1/2}u) - h(X_i, \beta_*) \right\} \leq \xi^{\mathrm{T}} E\{D_\beta h(X, \beta_*)\} u + \varepsilon'/2, \tag{52}$$

with probability exceeding $1 - \varepsilon/2$.

Next, given $\nu, \varepsilon'', b, K \in (0, \infty)$, it follows from Assumption A2 and the same line of reasoning in the proof of Proposition 5 in Blanchet et al. [2019] that there exists n_0 such that,

$$\sup_{\|\Delta\|_q \geq \nu n^{1/2}} \left[n^{1/2} \xi^{\mathrm{T}} \left\{ h(X_i + n^{-1/2}\Delta, \beta_* + n^{-1/2}u) - h(X_i, \beta_* + n^{-1/2}u) \right\} - \|\Delta\|_q^2 \right] \leq 0, \tag{53}$$

for all $n \geq n_0$, $\|\xi\|_p \leq b$, $\|u\|_2 \leq K$, and consequently, the first term in the right hand side of (51) is bounded from above by

$$\frac{1}{n}\sum_{i=1}^{n} \min\left\{\frac{1}{4\,(1-\varepsilon'')}\|\xi^{\mathrm{T}} D_x h(X_i,\beta_* + n^{-1/2}u)\|_p^2, c_n\right\} + \nu,$$

for some sequence $(c_n : n \geq 1)$ satisfying $c_n \to \infty$ as $n \to \infty$ (the exact value of c_n is not important). It follows from Assumption A2.c that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \min&\left[\frac{1}{4\,(1-\varepsilon'')}\left\|\left\{D_x h(X_i,\beta_* + n^{-1/2}u)\right\}^{\mathrm{T}}\xi\right\|_p^2, c_n\right] + \nu \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \min\left[\frac{1}{4\,\{1-\varepsilon''\}}\|\left(D_x h(X,\beta_*)\right)^{\mathrm{T}}\xi\|_p^2, c_n\right] + n^{-1/2}\|\xi\|_p\|u\|_q \frac{1}{n}\sum_{i=1}^{n}\kappa'(X_i) + \nu.\end{aligned}$$

Then, for n_0 suitably large, a similar application of Lemma 7 as in Proposition 9 results in,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \min&\left[\frac{1}{4\,(1-\varepsilon'')}\left\|\left\{D_x h(X_i,\beta_* + n^{-1/2}u)\right\}^{\mathrm{T}}\xi\right\|_p^2, c_n\right] + \nu \\ &\leq \frac{1}{4\,(1-\varepsilon'')} E\left\|\left\{D_x h(X,\beta_*)\right\}^{\mathrm{T}}\xi\right\|_p^2 + \frac{\varepsilon'}{4} + \nu,\end{aligned}$$

for all $n \geq n_0$, $\|\xi\|_p \leq b$, $\|u\|_2 \leq K$, with probability exceeding $1-\varepsilon/2$. Choosing ν, ε'' suitably small, we combine the above observation with that in (52) to obtain that,

$$\begin{aligned}nR_n\left(\beta_* + n^{-1/2}u\right) &= \max_{\xi}\left\{-\xi^{\mathrm{T}} H_n - M_n(\xi,u)\right\} \\ &\geq \max_{\substack{\|\xi\|_p \leq b \\ \|u_i\|_2 \leq K}}\left[-\xi^{\mathrm{T}} H_n - E\left\{\frac{1}{4}\|D_x h(X,\beta_*)^{\mathrm{T}}\xi\|_p^2 + \xi^{\mathrm{T}} D_\beta h(X,\beta_*)u\right\}\right] - \varepsilon',\end{aligned}$$

for all $n \geq n_0$, $\|u\|_2 \leq K$, with probability exceeding $1-\varepsilon$. □

Proof of Proposition 11. Due to equation (24), given $\varepsilon > 0$, there exists a and n_1 such as

$$\mathrm{pr}\{f_{up}(H_n,u,b,c) > a\} < \varepsilon/2.$$

for all $n > n_1$. Recall Proposition 9 and apply specifically with $u = 0$, we have

$$\sup_{n \geq \max\{n_0,n_1\}} \mathrm{pr}\{nR_n(\beta_*) \geq a+1\} \leq \varepsilon.$$

□

Lemma 8 below is useful to prove Proposition 12. Proof of Lemma 8 is presented in Section F.

Lemma 8. *Given any $K, b \in (0,\infty)$ and $\varepsilon \in (0,1)$, there exist positive constants n_0, L such that,*

$$\sup_{\|\xi\|_p \leq b} |M_n(\xi,u_1) - M_n(\xi,u_2)| \leq L\|u_1 - u_2\|_q,$$

with probability exceeding $1-\varepsilon$.

Proof of Proposition 12. For any $u_j, j = 1,2$, satisfying $\|u_j\|_2 \leq K$, let ξ_j attain the supremum in the relation $n^{\rho/2}R_n(\beta_* + n^{-1/2}u_j) = \sup_\xi\left\{-\xi^{\mathrm{T}} H_n - M_n(\xi,u_j)\right\}$. Then we have,

$$\left|nR_n(\beta_* + n^{-1/2}u_1) - nR_n(\beta_* + n^{-1/2}u_2)\right| \leq \max_{j=1,2}|M_n(\xi_j,u_1) - M_n(\xi_j,u_2)|\,. \tag{54}$$

For the given choices of ε, K, we have from Lemma 6 that there exist positive constants b and n_0 such that the optimal choices $\xi_j,\ j = 1, 2$, satisfy $\|\xi_j\|_p \leq b$, each with probability exceeding $1 - \varepsilon/3$. Consequently, we have from (54) and Lemma 8 that

$$\sup_{\|u_j\|_2 \leq K, j=1,2} \left| nR_n(\beta_* + n^{-1/2}u_1) - nR_n(\beta_* + n^{-1/2}u_2) \right| \leq L\|u_1 - u_2\|_q,$$

with probability exceeding $1-\varepsilon$, for all n suitably large. This completes the proof of Proposition 12. □

APPENDIX C. PROOFS OF PROPOSITIONS 13 - 16

In this section, we provide proofs of Propositions 13 - 16, which are key in the proof of Theorem 1.

Proof of Proposition 13. Due to the convexity of $\ell(\cdot)$, we have that $V_n^{ERM}(\cdot)$ is convex. In addition, for $\beta_1, \beta_2 \in \mathbb{R}^d$ and $\alpha \in (0, 1)$, we have

$$\begin{aligned}
\Psi_n\{\alpha\beta_1 + (1-\alpha)\beta_2\} &= \sup_{P: D_c(P,P_n) \leq \delta_n} E_P\left\{\ell(X; \alpha\beta_1 + (1-\alpha)\beta_2)\right\} \\
&\leq \sup_{P: D_c(P,P_n) \leq \delta_n} \left[\alpha E_P\left\{\ell(X;\beta_1)\right\} + (1-\alpha)E_P\left\{\ell(X;\beta_2)\right\}\right] \\
&\leq \alpha \sup_{P: D_c(P,P_n) \leq \delta_n} E_P\left\{\ell(X;\beta_1)\right\} + (1-\alpha) \sup_{P: D_c(P,P_n) \leq \delta_n} E_P\left\{\ell(X;\beta_2)\right\} \\
&= \alpha\Psi_n(\beta_1) + (1-\alpha)\Psi_n(\beta_2).
\end{aligned}$$

Due to the convexity of $\Psi_n(\cdot)$, we have that $V_n^{DRO}(\cdot)$ is also convex. Furthermore, due to the positive definiteness of $C = E\left\{D_\beta h(X, \beta_*)\right\}$ in Assumption A2.b, the smallest eigen value of C, denoted by $\lambda_{\min}(C)$, is positive. Equipped with these observations, we proceed as follows:

For a given $\varepsilon, \varepsilon' > 0$, let K_1 be such that $\sup_n \mathrm{pr}(\|H_n\|_2 > K_1) \leq \varepsilon$,

$$K_2 = \max\left\{K_1, \sup_{v:\|v\|_2 \leq K_1} f_{\eta,\gamma}(v)\right\}, \quad \text{and} \quad K_3 = 2\frac{2K_2 + \{2\varepsilon'\lambda_{\min}(C)\}^{1/2}}{\lambda_{\min}(C)}.$$

Observe from the definition of $f_{\eta,\gamma}(\cdot)$ that $K_2 \in (0, +\infty)$. Due to Propositions 6 and 7, there exists n_0 such that,

$$\begin{aligned}
V_n^{ERM}(u) &\geq H_n^{\mathrm{T}}u + \frac{1}{2}u^{\mathrm{T}}Cu - \varepsilon' \text{ and} \\
V_n^{DRO}(u) &\geq f_{\eta,\gamma}(H_n)^{\mathrm{T}}u + \frac{1}{2}u^{\mathrm{T}}Cu - \varepsilon',
\end{aligned}$$

for all u such that $\|u\|_2 \leq K_3$ and $n \geq n_0$. On the event, $\|H_n\|_2 \leq K_1$, we have that both $V_n^{ERM}(u)$ and $V_n^{DRO}(u)$ are bounded from below by,

$$V_l(u) = -K_2\|u\|_2 + \frac{\lambda_{\min}(C)}{2}\|u\|_2^2 - \varepsilon',$$

when $n \geq n_0$. Since $V_l(u) > 0$ for all u such that $\|u\|_2 \geq K_3/2$, we have that

$$\sup_{n \geq n_0} \mathrm{pr}\left\{V_n^{ERM}(u) > 0\right\} \geq 1 - \varepsilon \quad \text{and} \quad \sup_{n \geq n_0} \mathrm{pr}\left\{V_n^{DRO}(u) > 0\right\} \geq 1 - \varepsilon,$$

for all u such that $\|u\|_2 \in [K_3/2, K_3]$. Define $U = \{u \in \mathbb{R}^d : \|u\|_2 \leq K_3/2\}$. As $\min_u V_n^{ERM}(u) \leq 0$ and $\min_u V_n^{DRO}(u) \leq 0$, it follows from the convexity of $V_n^{ERM}(\cdot)$ and $V_n^{DRO}(\cdot)$ that,

$$\sup_{n \geq n_0} \mathrm{pr}\left\{\arg\min_u V_n^{ERM}(u) \subseteq U\right\} \geq 1 - \varepsilon$$

$$\text{and} \qquad \sup_{n \geq n_0} \text{pr}\big\{ \arg\min_u V_n^{DRO}(u) \subseteq U \big\} \geq 1 - \varepsilon,$$

thus verifying the claim. □

Proof of Proposition 14. Due to Theorem 3, we have that $G_n(\cdot) \Rightarrow G(\cdot)$, uniformly in compact sets. Then it follows from Skorokhod representation theorem that there exists a probability space where the convergence,

$$\sup_{\|u\| \leq K} |G_n(u) - G(u)| \to 0, \tag{55}$$

happen almost surely, for every $K \in (0, \infty)$. Since $G(\cdot)$ is continuous, we have from Theorem 7.14 of Rockafellar and Wets [2009] that G_n converges continuously to G, almost surely. A simple consequence of this observation; see Rockafellar and Wets [2009, Theorem 7.11], is that the epigraphs of G_n converges to the epigraph of G (alternatively, G_n epiconverges to G) almost surely. Observe that $G(\cdot)$ is convex and $\inf_u G(u) = 0$; this is because $\varphi^*(0) = 0$. Moreover, since $\alpha_n = n\delta_n = \eta \in (0, +\infty)$, we have that

$$\text{lev}(G_n, \alpha_n) = \text{lev}(G_n, \eta) \to \text{lev}(G, \eta),$$

almost surely, in the Painelevé-Kuratowski sense; see, for example, Beer and Lucchetti [1989, Theorem 5.1], Wijsman [1966, Theorem 7.1], or Beer et al. [1992]. Then, by Rockafellar and Wets [2009, Proposition 4.4], we have $\text{cl}\{\text{lev}(G_n, \alpha_n)\} \to \text{lev}(G, \eta)$. Consequently, $\text{cl}\{\text{lev}(G_n, \alpha_n)\} \Rightarrow \text{lev}(G, \eta)$. □

Proof of Proposition 15. Following the same reasoning used in the proof of Proposition 14 to arrive at (55), we have a probability space where the convergence,

$$\sup_{\|u\| \leq K} |G_n(u) - G(u)| \to 0 \quad \text{and} \quad n^{1/2}\left(\beta_n^{ERM} - \beta_*\right) \to C^{-1}H, \tag{56}$$

happen almost surely, for every $K \in (0, \infty)$; here, the latter convergence follows from (27).

Next, observe that $\alpha_n = n\delta_n \to 0$. Then, we have from Rockafellar and Wets [2009, Proposition 7.7a] that

$$\text{Ls}_{n\to\infty} \text{lev}(G_n, \alpha_n) \subseteq \text{lev}(G, 0) = \{u : \varphi^*(H - Cu) = 0\} = \{C^{-1}H\}, \tag{57}$$

where the latter equality follows from the strict convexity of $\varphi^*(\cdot)$ and the positive definiteness of C in Assumption A2.b.

Furthermore, since $R_n(\beta_n^{ERM}) = 0$, we have,

$$n^{1/2}(\beta_n^{ERM} - \beta_*) \in \text{lev}\left\{ nR_n(\beta_* + n^{-1/2} \times \cdot\,),\ \alpha_n \right\} = \text{lev}(G_n, \alpha_n),$$

for every n. Therefore, from the second convergence in (56), we obtain

$$C^{-1}H \in \text{Li}_{n\to\infty} \text{lev}(G_n, \alpha_n).$$

Combining this observation with that in (57) and Rockafellar and Wets [2009, Proposition 4.4], we obtain that PK-$\lim_n$ $\text{cl}\{\text{lev}(G_n, \alpha_n)\} = \{C^{-1}H\}$ almost surely. As a result, $\text{cl}\{\text{lev}(G_n, \alpha_n)\} \Rightarrow \{C^{-1}H\}$. □

Proof of Proposition 16. Following the same reasoning in the proof of Proposition 14, we have a probability space where the convergence in (55) happen almost surely, for every $K \in (0, +\infty)$. Consider any fixed $u \in \mathbb{R}^d$. Since $G_n(u) \to G(u) < \infty$ almost surely and $\alpha_n = n\delta_n \to \infty$, there exists a random variable N_u, defined on the same probability space, such that, $G_n(u) < \alpha_n$, with probability 1, for all $n \geq N_u$. As a result, we have $u \in \text{lev}(G_n, \alpha_n)$, for all but finitely many n, with probability 1. Then it follows from the definition of inner limit (Li_n) of sets that $u \in \liminf_n \text{lev}(G_n, \alpha_n)$. Since the choice of $u \in \mathbb{R}^d$ is arbitrary, we have that

$$\mathbb{R}^d \subseteq \text{Li}_{n\to\infty} \text{lev}(G_n, \alpha_n).$$

As $\limsup_n \text{lev}(G_n, \alpha_n)$ is essentially a subset of $\mathbb{R}^d$, it follows that PK-$\lim_n \text{cl}\{\text{lev}(G_n, \alpha_n)\} = \mathbb{R}^d$, almost surely. Consequently, $\text{cl}\{\text{lev}(G_n, \alpha_n)\} \Rightarrow \text{lev}(G, \alpha)$. □

APPENDIX D. PROOFS OF THEOREM 2 AND COROLLARY 1

Proof of Theorem 2. We define

$$\mathcal{K}_N = \Omega \cap \{x \in \mathbb{R}^m : \|x\|_2 \le N\}, \ \mathcal{U}_\delta^N(P_n) = \left\{P \in \mathcal{P}(\mathcal{K}_N) : W(P, P_n) \le \delta^{1/2}\right\},$$
$$g_N(\beta) = \sup_{P \in \mathcal{U}_\delta^N(P_n)} E_P\{\ell(X, \beta)\}, \text{ and } g(\beta) = \sup_{P \in \mathcal{U}_\delta(P_n)} E_P\{\ell(X, \beta)\}.$$

As slight abuse of the notation, we define $\ell(x, \beta) = +\infty$ for $\beta \notin B$, and thus $g(\beta) = g_N(\beta) = +\infty$ for $\beta \notin B$. Since B is closed, we have $g(\beta)$ and $g_N(\beta)$ are lower semi-continuous.

Now, we divide the proof of equation (10) into three steps.

Step 1: we show that $g_N(\beta) \to g(\beta)$ pointwisely as $N \to \infty$.

Step 2: we show that the sequence $\{g_N(\beta)\}$ epi-converges to $g(\beta)$; see, for example, Rockafellar and Wets [2009, Definition 7.1].

Step 3: we finally show that $\inf_{\beta\in B} \lim_{N\to\infty} g_N(\beta) = \lim_{N\to\infty} \inf_{\beta\in B} g_N(\beta)$.

By Fatou's lemma and the non-negativity of $\ell(\cdot)$, we have $E_P\{\ell(X, \cdot)\}$ is lower semi-continuous on B for any $P \in \mathcal{U}_\delta^N$. By the weak convergence of probability measure, we have $E_\cdot\{\ell(X, \beta)\}$ is continuous in weak topology on $\mathcal{U}_\delta^N(P_n)$ for any $\beta \in B$. By Sion's minimax theorem [Sion et al., 1958] and the compactness of $\mathcal{U}_\delta^N(P_n)$ in weak topology, we have

$$\lim_{N\to\infty} \inf_{\beta\in B} g_N(\beta) = \sup_{N\ge1} \inf_{\beta\in B} g_N(\beta) = \sup_{N\ge1} \inf_{\beta\in B} \sup_{P\in\mathcal{U}_\delta^N(P_n)} E_P\{\ell(X, \beta)\} = \sup_{N\ge1} \sup_{P\in\mathcal{U}_\delta^N(P_n)} \inf_{\beta\in B} E_P\{\ell(X, \beta)\}.$$

By the weak duality, we have

$$\begin{aligned}
\inf_{\beta\in B} \lim_{N\to\infty} g_N(\beta) &= \inf_{\beta\in B} \sup_{P\in\mathcal{U}_\delta(P_n)} E_P\{\ell(X, \beta)\} \\
&\ge \sup_{P\in\mathcal{U}_\delta(P_n)} \inf_{\beta\in B} E_P\{\ell(X, \beta)\} \\
&\ge \sup_{N\ge1} \sup_{P\in\mathcal{U}_\delta^N(P_n)} \inf_{\beta\in B} E_P\{\ell(X, \beta)\} \\
&= \lim_{N\to\infty} \inf_{\beta\in B} g_N(\beta).
\end{aligned}$$

Therefore, all the inequalities above should be equalities, which completes the proof.

Now, we execute proofs of Steps 1 - 3.

Proof of Step 1: Since $g_N(\beta)$ is increasing, we have $g_N(\beta)$ converges. Assume $\lim_{N\to\infty} g_N(\beta) = g^*(\beta)$, where $g^*(\beta)$ could be $+\infty$. We have $g^*(\beta) \le g(\beta)$. We use proof by contradiction. If $g^*(\beta) < g(\beta)$, we consider two cases: $g(\beta) < +\infty$ and $g(\beta) = +\infty$.

Case 1: $g(\beta) < \infty$. Let $\epsilon = g(\beta) - g^*(\beta) > 0$. Let $P' \in \mathcal{U}_\delta(P_n)$ such as $E_{P'}\{\ell(X,\beta)\} > g(\beta) - \epsilon/2$. There exists N sufficiently large such that

$$E_{P'}\{\ell(X,\beta)\mathbb{I}(\|X\|_2 > N)\} < \epsilon/2.$$

Then, we construct a measure $P'_N \in \mathcal{U}_\delta^N(P_n)$ that for any Borel set $A \subset \mathcal{K}_N$,

$$P'_N(A) = P'(A) + \left\{1 - P'(\mathcal{K}_N)\right\} P_n(A).$$

Therefore, we have

$$g^*(\beta) \geq g_N(\beta) \geq E_{P'_N}\{\ell(X,\beta)\} > E_{P'}\{\ell(X,\beta)\} - \epsilon/2 > g(\beta) - \epsilon,$$

which leads a contradiction.

Case 2: $g(\beta) = +\infty$ and $g^*(\beta) < +\infty$. Let $P' \in \mathcal{U}_\delta(P_n)$ such as $E_{P'}\{\ell(X,\beta)\} > g^*(\beta) + 1$. There exists N sufficiently large such that

$$E_{P'}\{\ell(X,\beta)\mathbb{I}(\|X\|_2 > N)\} < 1.$$

Then, we construct a measure $P'_N \in \mathcal{U}_\delta^N(P_n)$ that for any Borel set $A \subset \mathcal{K}_N$,

$$P'_N(A) = P'(A) + \left\{1 - P'(\mathcal{K}_N)\right\} P_n(A).$$

Therefore, we have

$$g^*(\beta) \geq g_N(\beta) \geq E_{P'_N}\{\ell(X,\beta)\} > E_{P'}\{\ell(X,\beta)\} - \epsilon/2 > g^*(\beta),$$

which leads a contradiction.

Proof of Step 2: By Rockafellar and Wets [2009, Proposition 7.2], we need to check two conditions:

(i) For every $\beta \in \mathbb{R}^d$ and for every sequence $\{\beta_N\}_{N=1}^\infty$ converging to β, we claim $\liminf_{N\to\infty} g_N(\beta_N) \geq g(\beta)$. Recalling $g_N(\beta_N) \geq g_M(\beta_N)$ by monotonicity for $N > M$ and the lower semi-continuity of $g_M(\beta)$, we have

$$\liminf_{N\to\infty} g_N(\beta_N) \geq \liminf_{N\to\infty} g_M(\beta_N) \geq g_M(\beta).$$

By taking M to the infinity, we have $\liminf_{N\to\infty} g_N(\beta_N) \geq g(\beta)$.

(ii) for every $\beta \in \mathbb{R}^d$, we pick a sequence $\beta_N = \beta$, then

$$\lim_{N\to\infty} g_N(\beta_N) = \lim_{N\to\infty} g_N(\beta) = g(\beta).$$

Proof of Step 3: We claim $g(\cdot)$ is level-bounded. Since $E_{P_*}\{\ell(X,\beta)\}$ has a unique minimizer, its level set $\left\{\beta \in \mathbb{R}^d : E_{P_*}\{\ell(X,\beta)\} \leq b\right\}$ is bounded. For every P_n and δ, there always exists $\epsilon \in (0,1)$, such that

$$(1-\epsilon)P_n + \epsilon P_* \in \mathcal{U}_\delta(P_n).$$

Then, we have $g(\beta) \geq \epsilon E_{P_*}\{\ell(X,\beta)\}$. Therefore, the level set $\left\{\beta \in \mathbb{R}^d : g(\beta) \leq b\right\} \subset \left\{\beta \in \mathbb{R}^d : E_{P_*}\{\ell(X,\beta)\} \leq b/\epsilon\right\}$ is bounded.

By Rockafellar and Wets [2009, Exercise 7.32(c)], we have the sequence $\{g_N(\cdot)\}$ is eventually level-bounded. Further, since g_N, g are lower semi-continuous and proper, by Rockafellar and Wets [2009, Theorem 7.33], we have the desired result.

Then, we proceed with the claim that there exists $\beta_n^{DRO}(\delta) \in \Lambda_\delta^+(P_n)$. First, $\mathcal{U}_\delta^N(P_n)$ is compact and $\inf_{\beta\in B} g_N(\beta)$ is upper-semicontinuous on P, and thus there exists P_N such that

$$P_N \in \arg\max_{P\in\mathcal{U}_\delta^N(P_n)} \inf_{\beta\in B} g_N(\beta).$$

Further, for any $\beta_N \in \arg\min_{\beta\in B} g_N(\beta)$, when N is sufficiently large, whose existence is guaranteed by Rockafellar and Wets [2009, Theorem 7.33]. Then, we have

$$\inf_{\beta\in B} \sup_{P\in\mathcal{U}_\delta^N(P_n)} E_P\left\{\ell\left(X,\beta\right)\right\} \geq E_{P_N}\left\{\ell(X,\beta_N)\right\} \geq \sup_{P\in\mathcal{U}_\delta^N(P_n)} \inf_{\beta\in B} E_P\left\{\ell\left(X,\beta\right)\right\}.$$

By Sion's minimax theorem, we have all the inequalities above are equalities. Therefore, $\beta_N \in \arg\min_\beta E_{P_N}\left\{\ell(X;\beta)\right\}$ and thus $\beta_N \in \Lambda_\delta(P_n)$. Finally, since the sequence $\{\beta_N\}_{N=1}^\infty$ is bounded and all its cluster points belong to $\arg\min g(\beta)$ by Rockafellar and Wets [2009, Theorem 7.33], combining with the closedness of $\Lambda_\delta^+(P_n)$, we have the desired result. □

Proof of Corollary 1. Define $\bar{\Psi}_n(\beta) = E_{P_n}[\ell(X;\beta)] + \eta^{1/2}n^{-1/2}\{E_{P_n}\|D_x\ell(X;\beta)\|^2\}^{1/2}$ and

$$\bar{V}_n(u) = n^{1/2}\left\{\bar{\Psi}_n\left(\beta_* + n^{-1/2}u\right) - \bar{\Psi}_n(\beta_*)\right\}.$$

Following the lines of the proof of Proposition 7, we have $\bar{V}_n(u) = V_n^{DRO}(u)+o_p(n^{-1/2})$. Consequently, since the collection $\{V_n^{DRO}(\cdot)\}_{n\geq 1}$ is tight and strongly convex (see the proof of Proposition 13), we have that the sequences $\{\bar{V}_n(\cdot)\}_{n\geq 1}$ and $\{\arg\min_u \bar{V}_n(u) : n \geq 1\}$ are tight. Then, as a consequence of Theorem 3, we have that $\bar{V}_n(\cdot) \Rightarrow V\{-f_{\eta,1}(H),\cdot\}$, uniformly in compact sets. Since the functions $\bar{V}_n(\cdot)$ and $V\{-f_{\eta,1}(H),\cdot\}$ are minimized, respectively, at $n^{1/2}\left(\bar{\beta}_n^{DRO} - \beta_*\right)$ and $C^{-1}f_{\eta,1}(H)$, we have that

$$n^{1/2}\left(\bar{\beta}_n^{DRO} - \beta_*\right) \Rightarrow C^{-1}f_{\eta,1}(H),$$

as $n \to \infty$. Then the conclusion that

$$n^{1/2}\left(\bar{\beta}_n^{DRO} - \beta_n^{DRO}\right) \to 0,$$

in probability, follows automatically from the convergence $n^{1/2}\left(\beta_n^{DRO} - \beta_*\right) \Rightarrow C^{-1}f_{\eta,1}(H)$; see Theorem 1 as a consequence of the continuous mapping theorem. This verifies the statement of Corollary 1. □

Appendix E. Proofs of Proposition 4 and Proposition 5

Proof of Proposition 4. i). Since $E\left[D_xh(X,\beta_*)D_xh(X,\beta_*)^{\mathrm{T}}\right] \succ 0$, we have $\varphi(\xi) \geq c\left\|\xi\right\|_2^2$ for some numerical constant $c > 0$ and thus $\varphi^*(\cdot)$ is continuous.

ii). Since $\beta_n^{ERM} \to \beta_*$ almost surely, we have $\varphi_n(\xi) \overset{a.s.}{\to} \varphi(\xi)$ for any ξ. Then, since $\|\xi^{\mathrm{T}}D_xh(X,\beta_*)\|_p^2$ is Lipschitz in ξ for $\left\|\xi\right\|_p \leq b$, i.e., for $\left\|\xi_1\right\|_p \leq b$, $\left\|\xi_2\right\|_p \leq b$,

$$\left|\left\|\{D_xh(X,\beta_*)\}^{\mathrm{T}}\xi_1\right\|_p^2 - \left\|\{D_xh(X,\beta_*)\}^{\mathrm{T}}\xi_2\right\|_p^2\right| \leq 2b\left\|D_xh(X,\beta_*)\right\|_q^2\left\|\xi_1 - \xi_2\right\|_p$$

and $E\left[\left\|D_xh(X,\beta_*)\right\|_q^2\right] < \infty$, we have the uniform law of large numbers that

$$\sup_{\|\xi\|_p\leq b}\left|\varphi_n(\xi) - \varphi(\xi)\right| \to 0$$

almost surely, uniformly over $\left\|\xi\right\|_p \leq b$. Then, we have

$$\sup_{\|\xi\|_p\leq b}\left\{\xi^{\mathrm{T}}\zeta - \varphi_n(\xi)\right\} \to \sup_{\|\xi\|_p\leq b}\left\{\xi^{\mathrm{T}}\zeta - \varphi(\xi)\right\}$$

almost surely, uniformly over ζ in compact sets as $n \to \infty$. Finally, since b is chosen arbitrarily, we conclude $\varphi_n^*(\cdot) \overset{p}{\to} \varphi^*(\cdot)$ as $n \to \infty$ uniformly on compact sets.

iii). Observe that $\varphi_n^*\left(\bar{\Xi}_n Z\right)=\varphi_n^*\left(\bar{\Xi}_n Z\right)-\varphi^*\left(\bar{\Xi}_n Z\right)+\varphi^*\left(\bar{\Xi}_n Z\right)$. The continuous mapping theorem and $\bar{\Xi}_n Z \Rightarrow H$ give us $\varphi^*\left(\bar{\Xi}_n Z\right) \Rightarrow \varphi^*(H)$. And ii) gives us $\varphi_n^*\left(\bar{\Xi}_n Z\right)-\varphi^*\left(\bar{\Xi}_n Z\right) \xrightarrow{p} 0$. □

Proof of Proposition 5. For any convex function $f(\cdot)$ with $\inf f<0$, it is well-known [Rockafellar and Wets, 2009, Exercise 11.6] that the support function of the level set $A=\{u: f(u) \leq 0\}$ is $h_A(v)=\inf_{\lambda>0} \lambda f^*(\lambda^{-1} v)$, where f^* is the convex conjugate of f. Since the convex conjugate of $\varphi^*(C \times \cdot\,)-\eta$ is $\varphi(C^{-1} \times \cdot\,)+\eta$, the support function of $\Lambda_\eta=\{u: \varphi^*(C u)-\eta \leq 0\}$ is

$$h_{\Lambda_\eta}(v)=\inf_{\lambda>0} \lambda\left\{\varphi\left(\lambda^{-1} C^{-1} v\right)+\eta\right\}=\inf_{\lambda>0}\left\{\lambda^{-1} \varphi\left(C^{-1} v\right)+\lambda \eta\right\}=2\{\eta \varphi(C^{-1} v)\}^{1/2}.$$

This completes the proof of Proposition 5. □

APPENDIX F. PROOFS OF TECHNICAL RESULTS

Proof of Lemma 2. A proof of the conclusion in Part a) of Lemma 2 can be found in Appendix A of Shalev-Shwartz and Singer [2007]. For the proof of Part b), we proceed as follows.

For brevity, let D denote the derivative of the function $\|\Delta\|_q^2$ evaluated at $\Delta=\Delta_*$ and $H(\Delta)$ denote the hessian matrix of function $\frac{1}{2}\|\Delta\|_q^2$. Then, for any $x \in R^d$,

$$\begin{aligned}
& x^T H(\Delta) x \\
& =\frac{1}{q}\left(\frac{2}{q}-1\right)\left(\sum_{i=1}^d|\Delta_i|^q\right)^{2/q-2}\left\{q \sum_{i=1}^d \operatorname{sgn}(\Delta_i)|\Delta_i|^{q-1} x_i\right\}^2 \\
& +(q-1)\left(\sum_{i=1}^d|\Delta_i|^q\right)^{2/q-1} \sum_{i=1}^d|\Delta_i|^{q-2} x_i^2 \\
& =\left(\sum_{i=1}^d|\Delta_i|^q\right)^{2/q-2}\left[(2-q)\left\{\sum_i \operatorname{sgn}(\Delta_i)|\Delta_i|^{q-1} x_i\right\}^2+(q-1)\left(\sum_{i=1}^d|\Delta_i|^q\right)\left(\sum_i|\Delta_i|^{q-2} x_i^2\right)\right] \\
& =\left(\sum_{i=1}^d|\Delta_i|^q\right)^{2/q-2}\left[\left\{\sum_{i=1}^d \operatorname{sgn}(\Delta_i)|\Delta_i|^{q-1} x_i\right\}^2+(q-1) \sum_{i=1}^d \sum_{j=i+1}^d|\Delta_i|^{q-2}|\Delta_j|^{q-2}(\Delta_i x_j-\Delta_j x_i)^2\right].
\end{aligned}$$

Since $q-1>1$, we obtain that,

$$x^T H(\Delta) x \geq\left(\sum_{i=1}^d|\Delta_i|^q\right)^{2/q-2}\left[\left\{\sum_{i=1}^d \operatorname{sgn}(\Delta_i)|\Delta_i|^{q-1} x_i\right\}^2+\sum_{i=1}^d \sum_{j=i+1}^d|\Delta_i|^{q-2}|\Delta_j|^{q-2}(\Delta_i x_j-\Delta_j x_i)^2\right].$$

Considering only non-zero entries among $\{\Delta_i: i=1, \ldots, d\}$, we re-express the right hand side as,

$$\begin{aligned}
& \left(\sum_{i=1}^d|\Delta_i|^q\right)^{2/q-2}\left\{\left(\sum_{\substack{i=1\\ \Delta_i \neq 0}}^d|\Delta_i|^q \frac{x_i}{\Delta_i}\right)^2+\sum_{\substack{i=1\\ \Delta_i \neq 0}}^d \sum_{\substack{j=i+1\\ \Delta_j \neq 0}}^d|\Delta_i|^q|\Delta_j|^q\left(\frac{x_j}{\Delta_j}-\frac{x_i}{\Delta_i}\right)^2\right\} \\
& \quad=\left(\sum_{i=1}^d|\Delta_i|^q\right)^{2/q-2}\left[\sum_{\substack{j=1\\ \Delta_j \neq 0}}^d|\Delta_j|^q\left\{\sum_{\substack{i=1\\ \Delta_i \neq 0}}^d|\Delta_i|^q\left(\frac{x_i}{\Delta_i}\right)^2\right\}\right]
\end{aligned}$$

$$= \left(\sum_{i=1}^{d} |\Delta_i|^q \right)^{2/q-1} \left\{ \sum_{i=1}^{d} |\Delta_i|^{q-2} (x_i)^2 \right\}$$
$$= \frac{\sum_{i=1}^{d} |\Delta_i|^{q-2} (x_i)^2}{\left\| |\Delta|^{q-2} \right\|_{\frac{q}{q-2}}},$$

where $|\Delta|^{q-2}$ is defined as a vector $(|\Delta_1|^{q-2}, |\Delta_2|^{q-2}, \ldots, |\Delta_d|^{q-2})^{\mathrm{T}}$. Then,

$$x^T H(\Delta) x \geq \frac{\sum_{i=1}^{d} |\Delta_i|^{q-2} (x_i)^2}{\left\| |\Delta|^{q-2} \right\|_{\frac{q}{q-2}}} \geq \frac{\sum_{i=1}^{d} |\Delta_i|^{q-2} (x_i)^2}{\left\| |\Delta|^{q-2} \right\|_{1}} = \frac{\sum_{i=1}^{d} |\Delta_i|^{q-2} (x_i)^2}{\sum_{i=1}^{d} |\Delta_i|^{q-2}}. \tag{58}$$

We can regard the right hand side as the weighted average of $\left\{ x_i^2 \right\}_{i=1}^{d}$.

Next, by applying Taylor's theorem, we have

$$\|\Delta\|_q^2 = \|\Delta^*\|_q^2 + \left(D \|\Delta^*\|_q^2 \right)^{\mathrm{T}} \xi + 2 \int_0^1 (1-t) \xi^{\mathrm{T}} H(\Delta^* + t\xi) \xi \mathrm{d}t.$$

We focus on the last term, which is

$$\int_0^1 (1-t) \xi^{\mathrm{T}} H(\Delta^* + t\xi) \xi \mathrm{d}t \geq \int_0^1 (1-t) \frac{\sum_{i=1}^{d} |\Delta_i^* + t\xi_i|^{q-2} (\xi_i)^2}{\sum_{i=1}^{d} |\Delta_i^* + t\xi_i|^{q-2}} \mathrm{d}t,$$

due to the inequality deduced earlier in (58). As the denominator of the right hand side in the above expression is bounded by,

$$\sum_{i=1}^{d} |\Delta_i^* + t\xi_i|^{q-2} \leq \max(2^{q-3}, 1) \sum_{i=1}^{d} \left(|\Delta_i^*|^{q-2} + |\xi_i|^{q-2} \right),$$

we obtain that,

$$\int_0^1 (1-t) \frac{\sum_{i=1}^{d} |\Delta_i^* + t\xi_i|^{q-2} (\xi_i)^2}{\sum_{i=1}^{d} |\Delta_i^* + t\xi_i|^{q-2}} \mathrm{d}t \geq \frac{1}{\max(2^{q-3}, 1)} \frac{\sum_{i=1}^{d} \left\{ \int_0^1 (1-t) |\Delta_i^* + t\xi_i|^{q-2} \mathrm{d}t \right\} (\xi_i)^2}{\sum_{i=1}^{d} \left(|\Delta_i^*|^{q-2} + |\xi_i|^{q-2} \right)}.$$

Then, we only need to bound the integral

$$\int_0^1 (1-t) |\Delta_i^* + t\xi_i|^{q-2} \mathrm{d}t.$$

If Δ_i^* and ξ_i have the same sign, then

$$\int_0^1 (1-t) |\Delta_i^* + t\xi_i|^{q-2} \mathrm{d}t \geq \int_0^1 (1-t) t^{q-2} |\xi_i|^{q-2} \mathrm{d}t = \frac{1}{(q-1)q} |\xi_i|^{q-2}.$$

On the other hand, if Δ_i^* and ξ_i have different signs, then we obtain that,

$$\int_0^1 (1-t) |\Delta_i^* + t\xi_i|^{q-2} \mathrm{d}t \geq |\xi_i|^{q-2} \left\{ \int_0^a (1-t) (a-t)^{q-2} \mathrm{d}t + \int_a^1 (1-t) (t-a)^{q-2} \mathrm{d}t \right\},$$

where $a = \min \left(\left| \frac{\Delta_i^*}{\xi_i} \right|, 1 \right)$. Computing the integrals in the right hand side of the above inequality, we obtain,

$$\left\{ \int_0^a (1-t) (a-t)^{q-2} \mathrm{d}t + \int_a^1 (1-t) (t-a)^{q-2} \mathrm{d}t \right\} = \frac{(1-a)^q + a^{q-1}(q-a)}{q(q-1)}$$

Since $2a < q$, we have

$$\frac{(1-a)^q + a^{q-1}(q-a)}{q(q-1)} \geq \frac{(1-a)^q + a^q}{q(q-1)} \geq \frac{1}{2^{q-1}q(q-1)}.$$

Then by combining the above observations, we obtain that,

$$2\int_0^1 (1-t)\xi^{\mathrm{T}} H(\Delta^* + t\xi)\xi \mathrm{d}t \geq C' \frac{\sum_{i=1}^d |\xi_i|^q}{\sum_{i=1}^d |\Delta_i^*|^{q-2} + \sum_{i=1}^d |\xi_i|^{q-2}}.$$

where

$$C' = \frac{1}{2^{q-2}q(q-1)\max(2^{q-3},1)}.$$

Moreover, we have,

$$\frac{\sum_{i=1}^d |\xi_i|^q}{\sum_{i=1}^d |\Delta_i^*|^{q-2} + \sum_{i=1}^d |\xi_i|^{q-2}} \geq \frac{1}{2}\min\left(\frac{\sum_{i=1}^d |\xi_i|^q}{\sum_{i=1}^d |\Delta_i^*|^{q-2}}, \frac{\sum_{i=1}^d |\xi_i|^q}{\sum_{i=1}^d |\xi_i|^{q-2}}\right). \tag{59}$$

Due to Chebyshev's sum inequality, we also obtain,

$$d\sum_{i=1}^d |\xi_i|^q \geq \left(\sum_{i=1}^d |\xi_i|^{q-2}\right)\left(\sum_{i=1}^d |\xi_i|^2\right). \tag{60}$$

Letting $C = \frac{1}{2d}C'$, the desired result follows. □

Proof of Lemma 6. For a fixed u, Δ, recall the definitions of $I(X_i, \Delta, u)$, $I_1(X_i, \Delta, u)$, $I_2(X_i, \Delta, u)$, Δ_i' from (45)-(46) and

$$M_n(\xi, u) = \frac{1}{n}\sum_{i=1}^n \left(\xi^{\mathrm{T}} D_\beta h\left(X_i, \beta_*\right) u + \max_{\Delta: X_i + n^{-1/2}\Delta \in \Omega} \left\{\xi^{\mathrm{T}} D_x h\left(X_i, \beta_*\right)\Delta + \xi^{\mathrm{T}} I(X_i, \Delta, u) - \|\Delta\|_q^2\right\}\right). \tag{61}$$

By taking $\Delta = 0$ and recalling Assumption A2.c, we have

$$\begin{aligned} &\max_{\Delta: X_i + n^{-1/2}\Delta \in \Omega} \left\{\xi^{\mathrm{T}} D_x h\left(X_i, \beta_*\right)\Delta + \xi^{\mathrm{T}} I(X_i, \Delta, u) - \|\Delta\|_q^2\right\} \\ \geq\ & \xi^{\mathrm{T}} I_2(X_i, 0, u) \\ \geq\ & -\|\xi\|_p \int_0^1 \left\|D_\beta h\left(X_i, \beta_* + t\frac{u}{n^{1/2}}\right) - D_\beta h\left(X_i, \beta_*\right)\right\|_q \|u\|_q \ \mathrm{d}t \\ \geq\ & -\frac{1}{2}n^{-1/2}\|\xi\|_p \|u\|_q^2 \,\bar{\kappa}(X_i). \end{aligned}$$

Since $\|u\|_2 \leq K$, for any $\epsilon_1 > 0$, there exists $n_1 > 0$ such as for all $n > n_1$, $\frac{1}{2}n^{-1/2}\|u\|_q^2 < 1$.

Then for any $c > 0$, plugging in $\Delta = c\Delta_i'$, we have that $\xi^{\mathrm{T}} D_x h(X_i, \beta_*)\Delta = c\|D_x h(X_i, \beta_*)^{\mathrm{T}}\xi\|_p \|\Delta_i'\|_q$, and thus,

$$\begin{aligned} &\max_{\Delta: X_i + n^{-1/2}\Delta \in \Omega} \left\{\xi^{\mathrm{T}} D_x h\left(X_i, \beta_*\right)\Delta + \xi^{\mathrm{T}} I(X_i, \Delta, u) - \|\Delta\|_q^2\right\} \\ &\geq \left\{c\|D_x h\left(X_i, \beta_*\right)^{\mathrm{T}} \xi\|_p \|\Delta_i'\|_q - c^2\|\Delta_i'\|_q^2 + \xi^{\mathrm{T}} I_1(X_i, c\Delta_i', u)\right\} \mathbb{I}\left(X_i + cn^{-1/2}\Delta_i' \in \Omega\right) - \bar{\kappa}(X_i)\|\xi\|_p. \end{aligned}$$

As a consequence of Hölder's inequality, $|\xi^{\mathrm{T}} I_1(X_i, c\Delta_i', u)|$ is bounded from above by

$$c\|\xi\|_p \int_0^1 \left\|\left\{D_x h\left(X_i + cn^{-1/2}\Delta_i', \beta_* + t\frac{u}{n^{1/2}}\right) - D_x h(X_i, \beta_*)\right\}\Delta_i'\right\|_q \mathrm{d}t.$$

Define the set $C_0 = \{w \in \Omega : \|w\|_p \leq c_0\}$, where c_0 will be chosen large momentarily. Then, due to the continuity of $D_x h(\cdot)$ in Assumption A2.c, we have that

$$\lim_{n\to\infty} |\xi^{\mathrm{T}} I_1(X_i, c\Delta_i', u)|\mathbb{I}(X_i \in C_0) = 0,$$

uniformly over all i such that $X_i \in C_0$, ξ in compact sets, and $\|u\|_2 \leq K$. Therefore, for given positive constants ε', c there exists n_2 such that for all $n \geq n_2$,

$$\sup_i |\xi^{\mathrm{T}} I_1(X_i, c\Delta_i', u)|\mathbb{I}(X_i \in C_0) \leq c\varepsilon' \|\xi\|_p.$$

Further, notice that $\|\Delta_i'\|$ is bounded when $X_i \in C_0$ due to the compactness of C_0 and the continuity of $D_x \{h(X_i, \beta_*)\}$, Let $M_\Delta = \sup_{x\in C_0} \|\Delta_i'(x)\|_2$. As a result, we obtain from (61) that, $M_n(\xi, u)$ is bounded from below by,

$$\begin{aligned}\frac{1}{n} \quad & \xi^{\mathrm{T}} \sum_{i=1}^n D_\beta h(X_i, \beta_*)u - \frac{1}{n}\sum_{i=1}^n \bar{\kappa}(X_i)\|\xi\|_p + \\ & \frac{1}{n}\sum_{i=1}^n \left\{c\|D_x h(X_i, \beta_*)^{\mathrm{T}} \xi\|_p \|\Delta_i'\|_q - c^2\|\Delta_i'\|_q^2 - c\varepsilon'\|\xi\|_p\right\} \mathbb{I}(X_i \in C_0^{cn^{-1/2}M_\Delta}).\end{aligned} \tag{62}$$

As in the proof of Lemma 2 in Blanchet et al. [2019] and $E\left[D_x h(X, \beta_*) D_x h(X, \beta_*)^{\mathrm{T}}\right] \succ 0$ in Assumption A2.b, there exists $\epsilon_0 > 0$, $\delta > 0$, and c_0 sufficiently large such that for all $n \geq N'(\delta)$,

$$\frac{1}{n}\sum_{i=1}^n \|D_x h(X_i, \beta_*)^{\mathrm{T}} \xi\|_p \|\Delta_i'\|_q \mathbb{I}(X_i \in C_0^{\epsilon_0}) > \frac{\delta}{2}\|\xi\|_p.$$

Further, let $c_1 = \sup_{x\in C_0} \|\Delta_i'(x)\|_q^2 < \infty$. By following the proof of Lemma 2 in Blanchet et al. [2019], if $n \geq \max\left\{N'(\delta), n_1, n_2, (M_\Delta c/\epsilon_0)^2\right\}$, we have

$$\sup_{\|\xi\|_p > b} \{\xi^{\mathrm{T}} H_n - M_n(\xi, u)\} \leq \sup_{\|\xi\|_p > b} \|\xi\|_p \left[b' - \left\{c\left(\frac{\delta}{2} - \varepsilon'\right) - \frac{(cc_1)^2}{b}\right\}\right],$$

on the set $\left\{\|H_n\|_q + \left\|\frac{1}{n}\sum_{i=1}^n D_\beta h(X_i, \beta_*)u\right\|_q + \left\|\frac{1}{n}\sum_{i=1}^n \bar{\kappa}(X_i)\right\|_q \leq b'\right\}$. Therefore, we can pick $c = 4(b'+1)/\delta + 1$, $\epsilon' = \delta/4$ and $b = (cc_1)^2 + 1$, then

$$b' - \left\{c\left(\frac{\delta}{2} - \varepsilon'\right) - \frac{(cc_1)^2}{b}\right\} < 0.$$

Notice that, there exists b' and n_3 such that for all $n > n_3$ and $\|u\|_2 \leq K$

$$\mathrm{pr}\left\{\|H_n\|_q + \left\|\frac{1}{n}\sum_{i=1}^n D_\beta h(X_i, \beta_*)u\right\| + \left\|\frac{1}{n}\sum_{i=1}^n \bar{\kappa}(X_i)\right\|_q > b'\right\} < \varepsilon/2. \tag{63}$$

Denote $n_4 = \left(M_\Delta\left(4(b'+1)/\delta + 1\right)/\epsilon_0\right)^2$. Therefore, there exists n_0 such as

$$\mathrm{pr}\left[\max\left\{N'(\delta), n_1, n_2, n_3, n_4\right\} > n_0\right] < \varepsilon/2.$$

Finally, we have the statement of Lemma 6 as a consequence of the union bound. □

Proof of Lemma 7. Lemma 7 follows as a consequence of the continuity properties of $D_x h(\cdot), D_\beta h(\cdot)$ and the strong law of large numbers. The proof of Lemma 7 is similar to the proof of Lemma 3 in Blanchet et al. [2019]. □

Proof of Lemma 8. For $i = 1, \ldots, n$ and $j = 1, 2$, let Δ_{ij} attain the inner supremum in

$$\max_{\Delta} \left[n^{1/2} \xi^{\mathrm{T}} \left\{ h(X_i + n^{-1/2}\Delta, \beta_* + n^{-1/2}u_j) - h(X_i, \beta_*) \right\} - \|\Delta\|_q^2 \right].$$

Then

$$\begin{aligned} &\left| \max_{\Delta} \left[n^{1/2} \xi^{\mathrm{T}} \left\{ h(X_i + n^{-1/2}\Delta, \beta_* + n^{-1/2}u_1) - h(X_i, \beta_*) \right\} - \|\Delta\|_q^2 \right] \right. \\ &\qquad \left. - \max_{\Delta} \left[n^{1/2} \xi^{\mathrm{T}} \left\{ h(X_i + n^{-1/2}\Delta, \beta_* + n^{-1/2}u_2) - h(X_i, \beta_*) \right\} - \|\Delta\|_q^2 \right] \right| \\ &\quad \leq \max_{j=1,2} \left[n^{1/2} \left| \xi^{\mathrm{T}} \left\{ h(X_i + n^{-1/2}\Delta_{ij}, \beta_* + n^{-1/2}u_1) - h(X_i + n^{-1/2}\Delta_{ij}, \beta_* + n^{-1/2}u_2) \right\} \right| \right], \end{aligned}$$

and consequently, it follows from the definition of $M_n(\xi, u)$ that,

$$\begin{aligned} &|M_n(\xi, u_1) - M_n(\xi, u_2)| \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \max_{j=1,2} \left[n^{1/2} \left| \xi^{\mathrm{T}} \left\{ h(X_i + n^{-1/2}\Delta_{ij}, \beta_* + n^{-1/2}u_1) - h(X_i + n^{-1/2}\Delta_{ij}, \beta_* + n^{-1/2}u_2) \right\} \right| \right]. \end{aligned} \tag{64}$$

Next, due to fundamental theorem of calculus, we have that,

$$\begin{aligned} &n^{1/2} \xi^{\mathrm{T}} \left| h(X_i + n^{-1/2}\Delta_{ij},\, \beta_* + n^{-1/2}u_1) - h(X_i + n^{-1/2}\Delta_{ij},\, \beta_* + n^{-1/2}u_2) \right| \\ &\quad = \left| \int_0^1 \xi^{\mathrm{T}} D_\beta h \left[X_i + n^{-1/2}\Delta_{ij},\, \beta_* + n^{-1/2} \left\{ u_1 + (u_2 - u_1)t \right\} \right] (u_2 - u_1) \mathrm{d}t \right| \\ &\quad \leq \|u_1 - u_2\|_q \|\xi\|_p \int_0^1 \left\| D_\beta h \left\{ X_i + n^{-1/2}\Delta_{ij},\ \beta_* + n^{-1/2} \left(u_1 + (u_2 - u_1)t \right) \right\} \right\|_q \mathrm{d}t \\ &\quad \leq \|u_1 - u_2\|_q \|\xi\|_p \left[\|D_\beta h(X_i, \beta_*)\|_q + \bar{\kappa}(X_i) \left\{ n^{-1/2}\|\Delta_{ij}\| + 2c_q n^{-1/2} K \right\} \right], \end{aligned} \tag{65}$$

where c_q is a fixed positive constant such that $\|x\|_q \leq c_q\|x\|_2$. The last inequality follows from Assumption A2.c). Moreover, for a given $b, \nu, K > 0$, we have from (53) that there exists n_0 such that $\Delta_{ij} \leq \nu n^{1/2}$, for all $i \leq n, n \geq n_0$, $\|\xi\|_p \leq b$, $\|u\|_2 \leq K$. Combining this observation with those in (64) and (65), we obtain that

$$\sup_{\|\xi\|_p \leq b} |M_n(\xi, u_1) - M_n(\xi, u_2)| \leq \|u_1 - u_2\|_q b \left\{ E_{P_n} \|D_\beta h(X, \beta_*)\|_q + E_{P_n} \{\bar{\kappa}(X)\} \left(\nu + 2c_q n^{-1/2} K \right) \right\},$$

for all $n \geq n_0$. For any random variable Z, let $\mathrm{CV}(Z) = \mathrm{var}(Z)/E(Z)^2$ denote the coefficient of variation of Z. If n_0 is also taken to be larger than both $2\varepsilon^{-1}\mathrm{CV}\left\{ \|D_\beta h(X, \beta_*)\|_q \right\}$ and $2\varepsilon^{-1}\mathrm{CV}\{\bar{\kappa}(X)\}$, then we have

$$\begin{aligned} \mathrm{pr}\left\{ E_{P_n} \|D_\beta h(X, \beta_*)\|_q \leq 2E \|D_\beta h(X, \beta_*)\|_q \right\} \geq 1 - \varepsilon/2 \text{ and} \\ \mathrm{pr}\left[E_{P_n} \{\bar{\kappa}(X)\} \leq 2E \{\bar{\kappa}(X)\} \right] \geq 1 - \varepsilon/2. \end{aligned}$$

With these observations, if we take $L = 4b \left[E \|D_\beta h(X, \beta_*)\|_q + E\{\bar{\kappa}(X)\}(\nu + 2c_q K) \right]$, then

$$\sup_{\|\xi\|_p \leq b} |M_n(\xi, u_1) - M_n(\xi, u_2)| \leq L\|u_1 - u_2\|_q,$$

with probability exceeding $1 - \varepsilon$. □