

Confidence Regions in Wasserstein Distributionally Robust Estimation

JOSE BLANCHET, KARTHYEK MURTHY, AND NIAN SI

ABSTRACT. Wasserstein distributionally robust optimization (DRO) 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 model parameters (or) decision choices that are robust to model uncertainties and misspecification, the Wasserstein DRO 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 underlying DRO estimators as well as the properties of an optimal (in a suitable sense) confidence region induced by the Wasserstein DRO formulation.

1. INTRODUCTION

In recent years, distributionally robust optimization (DRO) formulations based on Wasserstein distances (or Wasserstein DRO formulations, for short) 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 (see [35, 8, 41, 43, 24, 29, 42]) and operations research (see [27, 10, 19, 21, 5, 47, 1]), is that these types of formulations provide a flexible way to quantify, and hedge against, the impact of model misspecification.

This paper develops theory of asymptotic normality for Wasserstein DRO estimators and convergence of associated optimal (in a certain sense) confidence regions.

We shall provide a review of Wasserstein DRO and its connections to several areas, such as artificial intelligence, machine learning and operations research, which are of significant interests to various scientific communities. But, first, we would like to set the stage to explain our contributions in this paper by introducing the elements of typical data-driven non-DRO and DRO estimation formulations.

The elements of a typical data-driven (non-DRO) estimation formulation. A typical data-driven formulation considers a data set $\{X_k : 1 \leq k \leq n\} \subseteq \mathbb{R}^m$, which, for simplicity, we will assume to be comprising independent and identically distributed (i.i.d.) samples from an unknown distribution P_* . The random variables X_k are assumed to be constructed in their canonical space (that is, we may assume without loss of generality that $X_k(\omega) = \omega$). Unless otherwise stated, the probability measures that we consider are defined on Borel sets of the Euclidean space. We use $\mathbf{P}$ to denote the distribution of the sequence $\{X_k : k \geq 1\}$, which is the product measure induced by P_* (defined on the associated Kolmogorov σ -field).

(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.

The empirical measure associated with the sample is denoted as P_n . We use $\mathbb{E}_{P_n}[\cdot]$ to denote the expectation with respect to the empirical measure, P_n . Therefore, if $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is any function (say, continuous and bounded, for specificity) then $\mathbb{E}_{P_n}[f(X)] = n^{-1} \sum_{k=1}^n f(X_k)$.

More generally, given a probability model P , we write $\mathbb{E}_P[\cdot]$ to denote the expectation operator associated to P and therefore $\mathbb{E}_P[f(X)] = \int f(x)P(dx)$. Thus, because of our i.i.d. assumption, we have $\mathbb{E}_{P_n}[f(X)] \rightarrow \mathbb{E}_{P_*}[f(X)]$, $\mathbf{P}$ -almost surely, as $n \rightarrow \infty$ for any continuous and bounded function $f(\cdot)$.

In the case of P_* , we will, however, drop the subindex in the expectation operator; for example, we write $\mathbb{E}_{P_*}[f(X)] = \mathbb{E}[f(X)]$.

A typical (non-DRO) data-driven stochastic optimization problem informed by P_n focuses on minimizing empirical expected loss of the form, $\mathbb{E}_{P_n}[\ell(X; \beta)]$, over the decisions (or) parameter choices $\beta \in \mathbb{R}^d$. We may assume that for any given $x \in \mathbb{R}^m$, $\ell(x; \cdot)$ is differentiable with derivative $h(x, \beta) = D_\beta \ell(x; \beta)$ and we can identify a choice,

$$\beta_n^{ERM} \in \arg \min_{\beta} \mathbb{E}_{P_n} [\ell(X; \beta)] \quad (1)$$

via the first order optimality condition $\mathbb{E}_{P_n}[h(\beta_n^{ERM}, X)] = \mathbf{0}$. If we let β_* be the unique minimizer for $\min_{\beta \in \mathbb{R}^d} \mathbb{E}[\ell(X; \beta)]$, then under suitable regularity conditions to be discussed in the sequel, it is well-known that $\beta_n^{ERM} \rightarrow \beta_*$, almost surely. The estimator β_n^{ERM} is called the Empirical Risk Minimization (ERM) estimator.

A typical Wasserstein DRO formulation. A DRO formulation recognizes the distributional uncertainty inherent in P_n being a noisy representation of an unknown distribution; therefore, it is natural to consider an estimator of the form,

$$\beta_n^{DRO}(\delta) \in \arg \min_{\beta} \max_{P \in \mathcal{U}_\delta(P_n)} \mathbb{E}_{P_n} [\ell(X; \beta)], \quad (2)$$

where the set $\mathcal{U}_\delta(P_n)$ is called the distributional uncertainty region and δ is the size of the distributional uncertainty. Wasserstein DRO advocates choosing

$$\mathcal{U}_\delta(P_n) = \{P : W(P, P_n) \leq \sqrt{\delta}\},$$

where $W(P_n, P)$ is the Wasserstein distance between P and P_n , denoted as $W(P_n, P)$. More precisely, we shall work with Wasserstein distances of order 2, or the 2-Wasserstein distance, which explains why it is natural to use $\sqrt{\delta}$ to specify $\mathcal{U}_\delta(P_n)$ as above. Our results can be easily extended to the case of Wasserstein distances of order ρ for $\rho > 1$, we focus on the case $\rho = 2$ to simplify the notation but also because of it being one of the most common cases. While we review this specific notion of Wasserstein distance in Section 2.1.1, it suffices to say that the squared 2-Wasserstein distance $W(P_n, P)$ is interpreted as the cheapest way to transport the mass from P_n to the mass of another probability measure P , while measuring the cost of transportation from location $x \in \mathbb{R}^m$ to location $y \in \mathbb{R}^m$ in terms of the squared distance between x and y . (One may keep in mind the Euclidian distance for simplicity in this discussion, but we will consider, more generally, the p -th norm in our development.)

Motivation behind Wasserstein DRO estimators. With the need to identify model parameters (or) decision choices that are robust to model uncertainties and misspecification, there is a rapidly growing literature on Wasserstein DRO formulations of the form (2) in various quantitative disciplines related to statistics and operations research: see for example, [35, 27, 48, 10, 19, 8, 45, 41, 20, 43, 14, 9]. Wasserstein DRO formulations such as (2) have been shown to recover many machine learning estimators, including square-root LASSO and support vector

machines ([8]), LAD LASSO (see ([13])), and group LASSO ([7]), among others ([35, 18, 16, 43, 13]).

The Wasserstein DRO formulation (2) has also been explored in adversarial training of Neural Networks, see for example [41, 42].

As such, 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/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.

Generic formulations such as (2) are becoming increasingly tractable. See for example, [27, 26] for convex programming based approaches and [41, 11] for stochastic gradient descent based iterative schemes. Algorithmic approaches in specific application settings include portfolio selection [5], covariance estimation [29], Kalman filtering [1], dynamic control [46, 47], hypothesis testing [20] and traveling salesman problems [12]. These algorithmic advances in DRO are accompanied by significant progress in computational optimal transport (see [33, 6, 25] and references therein).

Therefore, from a pragmatic perspective, our quest for understanding the asymptotic behavior of solutions of (2) is motivated by a wide range of applications. Simply put, Wasserstein DRO is becoming an important tool in the arsenal of learning and decision making under uncertainty, and an objective of this article is to investigate the asymptotic behavior of the optimal value and optimal solutions of the DRO formulation (2). We shall also see that the statistical limit theorems lend themselves easily towards new characterizations of confidence regions that are natural from standpoint of the considered distributionally uncertainty region $\mathcal{U}_{\delta}(P_n)$.

An overview of contributions in this article. In order to specifically describe the contributions, let us introduce the following notation. For any positive integer n and $\delta_n > 0$, let

$$\Psi_n(\beta) := \sup_{P \in \mathcal{U}_{\delta_n}(P_n)} \mathbb{E}_P[\ell(X; \beta)] \quad \text{and} \quad L_n(\beta) := \mathbb{E}_{P_n}[\ell(X; \beta)],$$

respectively, denote the DRO objective (2) and the empirical risk. Recall our assumption that β_* uniquely minimizes the population risk. According to (1) - (2), we have $\beta_n^{DRO}(\delta_n) \in \arg \min_{\beta} \Psi_n(\beta)$ and $\beta_n^{ERM} \in \arg \min_{\beta} L_n(\beta)$. Next, let

$$\Lambda_{\delta_n}(P_n) = \left\{ \beta \in \mathbb{R}^d : \beta \in \arg \min_{\beta} \mathbb{E}_P[\ell(X; \beta)] \text{ for some } P \in \mathcal{U}_{\delta_n}(P_n) \right\} \quad (3)$$

denote the set of choices of $\beta \in \mathbb{R}^d$ 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 words, if $\mathcal{U}_{\delta_n}(P_n)$ represents the set of probabilistic models which are, based on the empirical evidence, plausible representations (for the purpose of decision making) of the underlying phenomena, then each of such representations induces an optimal decision and, $\Lambda_{\delta_n}(P_n)$ encodes the corresponding induced set of plausible decisions.

With the above notation, the key contributions of this article can be described as follows:

A) (Joint Asymptotic Limits): *Assuming that the samples $\{X_1, \dots, X_n\}$ are independent and identically distributed, we establish the convergence in distribution of the triplet,*

$$\left(n^{1/2}(\beta_n^{ERM} - \beta_*), n^{\bar{\gamma}/2}(\beta_n^{DRO}(\delta_n) - \beta_*), n^{1/2}(\Lambda_{\delta_n}(P_n) - \beta_*) \right), \quad (4)$$

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 $\beta \in \mathbb{R}^d$ that are compatible with the respective distributional uncertainty region $\mathcal{U}_{\delta_n}(P_n)$ in the sense described above. Under an appropriate notion of convergence of sets (described in Section 2.1.2), the characterization of limit of the sets $n^{1/2}(\Lambda_{\delta_n}(P_n) - \beta_*)$ relies on examining the level sets of a suitably defined profile function (in the spirit of empirical likelihood profile function).

B) (Qualitative and Geometric Insights): *We utilize the limiting result of (4) to examine the effect of the size of distributional ambiguity, δ_n , the qualitative and geometrical properties of the DRO solutions, and geometrical insights of 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 $n \rightarrow \infty$. It emerges that the canonical, $O(n^{-1/2})$, rate of convergence is achieved only if $\gamma \leq 1$ and the limiting distribution of $\beta_n^{DRO}(\delta_n)$ and β_n^{ERM} are different only if $\gamma \geq 1$. Hence to both obtain the canonical rate and tangible benefits from the DRO formulation, we must choose $\gamma = 1$, which corresponds to the resulting $\bar{\gamma}$ in (4) to be equal to 1. Under this choice, an explicit connection between $\beta_n^{DRO}(\delta_n)$ and β_n^{ERM} is obtained in Section 2.2.1 (see equation (12)). Owing to this connection, we can reason that the DRO formulation (2) tends to favor solutions with low sensitivity towards perturbations in data (see Corollary 1 in Section 2.2.1).

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 $\delta_n = \eta n^{-1}$, the set $\Lambda_{\delta_n}(P_n)$ is an asymptotic $(1 - \alpha)$ -confidence region for β_* . For any choice of $\eta \geq \eta_\alpha$ in $\delta_n = \eta n^{-1/2}$, we show that,

$$\lim_{n \rightarrow \infty} P \left(\beta_* \in \beta_n^{ERM} + n^{-1/2} \Lambda_\eta \right) \geq 1 - \alpha, \quad (5)$$

where the set Λ_η is identified from the limiting distribution of $n^{1/2}(\Lambda_{\delta_n}(P_n) - \beta_n^{ERM})$. Explicit construction of the set Λ_η , and hence the confidence region given by $\beta_n^{ERM} + n^{-1/2} \Lambda_\eta$, relies on tools from convex analysis which may be of independent interest; these are discussed in Section 2.2.2.

The choice $\eta = \eta_\alpha$ results in the smallest possible value for η , and subsequently smallest volume set in the family, $\{\Lambda_\eta : \eta > 0\}$, of limiting sets, for which (5) is satisfied. Therefore, we refer to the choice $\delta_n = \eta_\alpha n^{-1/2}$ as the optimal uncertainty size and the set $\beta_n^{ERM} + n^{-1/2} \Lambda_{\eta_\alpha}$ as the optimal confidence region (among those defined using the sets Λ_η).

C) (Technical insights): *Our methodology exposes technical insights, which are likely to be useful in contexts of wider scope than those considered here.*

For example, a key technical result that we obtain, in order to prove asymptotic normality, is a suitable (tractable) approximation, $\hat{\Psi}_n(\beta)$, of the DRO objective $\Psi_n(\beta)$ such that,

$$\left| \hat{\Psi}_n(\beta) - \Psi_n(\beta) \right| = o(\delta_n), \quad (6)$$

as $n \rightarrow \infty$, uniformly over β in any compact subsets of $\mathbb{R}^d$ (see Proposition 15). The approximation $\hat{\Psi}_n(\beta)$ is finer than the existing approximations for the DRO objective in the literature (which is of error $o(\delta_n^{1/2})$; see Theorem 2 in [18]).

To summarize, contributions A) and B) enable the development of confidence regions around the distributionally robust decision, β_n^{DRO} . This work, to the best of our knowledge is the first to characterize asymptotic normality and address the problem of computing confidence regions for DRO solutions. Finally, the methodological ideas in the paper are likely to be helpful in the analysis of Wasserstein DRO formulations.

Connections to related statistical results in the literature. In the context of DRO, the paper [8] takes a different, but related approach to characterize the optimal choice of δ_n as the solution of a certain minimization problem, subject to a coverage constraint. However, asymptotic normality of the optimizer and characterization of optimal confidence regions were well beyond the scope of the treatment in [8].

Asymptotic normality of M-estimators which minimize empirical risk of the form, $E_{P_n}[\ell(X; \beta)]$, was first established in the pioneering work of Huber [22]. Our work here is different because of the presence of the adversarial perturbation to the loss represented by the inner maximization in (2). Subsequent asymptotic characterizations in the presence of constraints on the choices of parameter vector β have been developed in [17, 37, 38, 39, 40], again in the standard M-estimation setting.

Asymptotic normality in the related context of regularized estimators for least squares regression has been established in [23]. As mentioned earlier, DRO estimators of the form (2) recover LASSO-type estimators as particular examples (see [8]). 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 [23], from a DRO perspective.

We comment that some of our results involving convergence of sets may be of interest to applications in the area of Empirical Likelihood (EL) (see [30, 31, 32]). This is because $\Lambda_{\delta_n}(P_n)$ can be characterized in terms of a function, namely, Robust Wasserstein Profile function or RWP function, which resembles the definition of the EL profile function. The RWP function, introduced in [8], has found applications in related statistical problems such as graphical Lasso (see [15]) and the results developed herein might be immediately applicable towards characterizing asymptotic normality and confidence regions in such settings.

Organization of the paper. The rest of the paper is organized as follows. In Section 2, we state the technical assumptions and provide precise statements of our main results. In Section 3, we provide examples of specific confidence regions which are useful to convey geometric insights. Section 4 is devoted to the proofs of main results. Proofs of results, which are technical in nature, are presented in the appendix.

2. ASSUMPTIONS AND MAIN RESULTS

2.1. Preliminaries and Assumptions. We begin with a set of basic definitions and assumptions required to state our main results.

2.1.1. The Wasserstein Distance. First, we review the definition of optimal transport costs, which include Wasserstein distances as special case. Given a lower-semicontinuous cost function $c : \mathbb{R}^m \times \mathbb{R}^m \rightarrow [0, \infty]$, define

$$D_c(P, Q) := \min \{E_\pi [c(X, Y)] : \pi_X = P, \pi_Y = Q\},$$

where the minimum is taken over all (Borel) probability measures on $\mathbb{R}^m \times \mathbb{R}^m$ and given such measure, say π , we write $\pi_X(A) = \pi(A \times \mathbb{R}^m)$ and $\pi_Y(A) = \pi(\mathbb{R}^m \times A)$ (i.e., π_X is the marginal distribution of the first m coordinates under P , the analogous interpretation is understood for π_Y). Using this notation, X encodes a random variable with distribution P and Y encodes a random variable with distribution Q . In words, the quantity $D_c(P, Q)$ is the smallest possible expected value of the random variable $c(X, Y)$ over all joint distributions of the pair (X, Y) which preserve the marginal distributions of X and Y , respectively.

If $d(\cdot)$ is a metric and $r > 0$, then choosing $c(x, y) = d^r(x, y)$ leads to the Wasserstein distance of order r defined via

$$W_r(P, Q) = (D_c(P, Q))^{1/r}.$$

Throughout our development we will work with Wasserstein distances of order $r = 2$ and we will omit the subscript r (simply writing $W(P, Q)$ instead of $W_2(P, Q)$). Moreover, once the metric has been set, we also suppress the dependence on c in $D_c(P, Q)$ by simply writing $D(P, Q)$. Consequently, in our setting $D(P, Q) = W(P, Q)^2$ and therefore,

$$\mathcal{U}_\delta(P_n) = \{P : W(P, P_n) \leq \sqrt{\delta}\} = \{P : D(P, P_n) \leq \delta\}.$$

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

$$\begin{aligned} \text{Li}_{n \rightarrow \infty} A_n &:= \left\{ z \in \mathbb{R}^d : \text{there exists a sequence } \{a_n\}_{n \geq 1} \text{ convergent to } z \right. \\ &\quad \left. \text{for all but finitely many integers } n \right\}, \text{ and} \\ \text{Ls}_{n \rightarrow \infty} A_n &:= \left\{ z \in \mathbb{R}^d : \text{there exist positive integers } n_1 < n_2 < n_3 < \dots \text{ and } a_k \in A_{n_k} \right. \\ &\quad \left. \text{such that the sequence } \{a_k\}_{k \geq 1} \text{ is convergent to } z \right\}. \end{aligned}$$

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

$$A = \text{Li}_{n \rightarrow \infty} A_n = \text{Ls}_{n \rightarrow \infty} A_n,$$

in which case we write $\text{PK-lim}_n A_n = A$. Since $\mathbb{R}^d$ is a locally compact Hausdorff space (LCHS), the topology induced by Painlevé-Kuratowski convergence on the space of closed subsets $2^{\mathbb{R}^d}$ of $\mathbb{R}^d$ is completely metrizable, separable and coincides with the well-known topology of closed convergence (also known as Fell topology, see Chapter 1 in [28]). The notion of convergence of sets we utilize from here onwards will be the above defined Painlevé-Kuratowski convergence.

2.1.3. *Assumptions.* Throughout our development we impose the following assumptions.

A1) The transportation cost $c(\cdot)$ is of the form $c(u, w) = \|u - w\|_q^2$ for some $q \in (1, \infty)$. For the chosen $q \in (1, \infty)$, let p be such that $1/p + 1/q = 1$.

A2) The function $\ell(\cdot)$ satisfies the following properties:

- a) $\ell(\cdot)$ is twice continuously differentiable, non-negative, and for each x , $\ell(x, \cdot)$ is convex.
- b) If we let $h(x, \beta) := D_\beta \ell(x, \beta)$, then there exists $\beta_* \in \mathbb{R}^d$ satisfying $\mathbb{E}[h(X, \beta_*)] = \mathbf{0}$, $\mathbb{E}\|h(X, \beta_*)\|_2^2 < \infty$, $C := \mathbb{E}[D_\beta h(X, \beta_*)] \succ \mathbf{0}$ and $\mathbb{E}[D_x h(X, \beta_*) D_x h(X, \beta_*)^T] \succ \mathbf{0}$. (Throughout the paper, $A \succ \mathbf{0}$ implies that A is strictly positive definite.)
- c) For every $\beta \in \mathbb{R}^d$, $D_{xx}\ell(\cdot; \beta)$ is bounded by a continuous function $M(\beta)$ and uniformly continuous. In addition, $D_x h(\cdot)$ and $D_\beta h(\cdot)$ satisfy the following locally Lipschitz continuity conditions:

$$\begin{aligned} \|D_x h(x + \Delta, \beta_* + u) - D_x h(x, \beta_*)\|_p &\leq \kappa'(x) (\|\Delta\|_q + \|u\|_q), \\ \|D_\beta h(x + \Delta, \beta_* + u) - D_\beta h(x, \beta_*)\|_p &\leq \bar{\kappa}(x) (\|\Delta\|_q + \|u\|_q), \end{aligned}$$

where $\kappa', \bar{\kappa} : \mathbb{R}^m \rightarrow [0, \infty)$ are such that $\mathbb{E}[(\kappa'(X_i))^2] < \infty$ and $\mathbb{E}[\bar{\kappa}^2(X_i)] < \infty$.

Assumption A1) covers most of the cases considered in practice. One case that we do not explicitly consider, but which can be easily adapted after a simple change-of-coordinates transformation, is the weighted l_2 norm case (also known as the Mahalanobis distance), namely $c(x, y) = (x - y)^T A(x - y)$, for a symmetric positive definite matrix A . Efficient iterative procedures for Wasserstein DRO problems using the Mahalanobis distance (and localized variation of it) has been studied in [11]. The limiting distribution in the case $q = \infty$ presents a discontinuity which makes it difficult to use in practice. Hence, we prefer not to cover this here.

The requirement that $\ell(\cdot)$ is twice differentiable in Assumption A2.a) is important in our analysis to study the impact of adversarial perturbations. Convexity of $\ell(x, \cdot)$, together with C being positive definite in A2.b), implies uniqueness of β_* . 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). Assumption A2.c) imposes further regularity conditions on the second derivative of $\ell(\cdot)$, which is again in the spirit of controlling the magnitude of adversarial perturbations to ensure that the inner maximization in (2) is not infinite.

2.2. **Main results.** In order to state our main results we introduce a few more definitions. Let

$$\varphi(\xi) := \frac{1}{4} \mathbb{E} \left(\left\| (D_x h(X, \beta_*))^T \xi \right\|_p^2 \right),$$

and its convex conjugate (alternatively known as Young-Fenchel transform),

$$\varphi^*(\zeta) := \sup_{\xi \in \mathbb{R}^d} \{ \xi^T \zeta - \varphi(\xi) \}.$$

In addition, define

$$S(\beta) := \sqrt{\mathbb{E}_P \|D_x \ell(X; \beta)\|_p^2} \tag{7}$$

and the parametric family of functions,

$$f_{\eta, \gamma}(x) := x \mathbf{1}_{\{\gamma \geq 1\}} - \sqrt{\eta} D_\beta S(\beta_*) \mathbf{1}_{\{\gamma \leq 1\}}, \tag{8}$$

for $\eta \geq 0, \gamma \geq 0$. Recall also the definition of the natural confidence region $\Lambda_{\delta_n}(P_n)$ given in (3). Finally, define

$$\Lambda_{\eta,\gamma} := \begin{cases} \{u : \varphi^*(Cu) \leq \eta\}, & \text{if } \gamma = 1, \\ \mathbb{R}^d, & \text{if } \gamma < 1, \\ \mathbf{0} & \text{if } \gamma > 1. \end{cases} \quad (9)$$

where $C := \mathbb{E}[D_\beta h(X, \beta_*)]$ was introduced in Assumption A2.b).

2.2.1. The main limit theorem. We now state our main result.

Theorem 1. *Suppose that Assumptions A1 - A2 are satisfied, $\delta_n = n^{-\gamma}\eta$ for some $\gamma, \eta \in (0, \infty)$, $\mathbb{E}\|X\|_2^2 < \infty$, and $H \sim \mathcal{N}(\mathbf{0}, \text{Cov}[h(X, \beta_*)])$. Then*

$$\begin{aligned} & \left(n^{1/2}(\beta_n^{ERM} - \beta_*), n^{\bar{\gamma}/2}(\beta_n^{DRO}(\delta_n) - \beta_*), n^{1/2}(\Lambda_{\delta_n}(P_n) - \beta_*) \right) \\ & \Rightarrow (C^{-1}H, C^{-1}f_{\eta,\gamma}(H), \Lambda_{\eta,\gamma} + C^{-1}H), \end{aligned}$$

where $\bar{\gamma} := \min\{\gamma, 1\}$ and $\Lambda_{\eta,\gamma}$ is defined as in (9).

The proof of Theorem 1 is presented in Section 4.2. 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. Following the definition of the parametric family in (8), we have $f_{0,\gamma}(H) = H$ for any $\gamma > 1$. Therefore, we have from Theorem 1 the following distributional limit,

$$\left(n^{1/2}(\beta_n^{ERM} - \beta_*), n^{\bar{\gamma}/2}(\beta_n^{DRO}(\delta_n) - \beta_*), n^{1/2}(\Lambda_{\delta_n}(P_n) - \beta_*) \right) \Rightarrow (C^{-1}H, C^{-1}H, \{C^{-1}H\}),$$

thus allowing us to infer that the influence of the robustification vanishes in the limit when $\delta_n = o(1/n)$.

On the other hand, if $n\delta_n \rightarrow \infty$ corresponding to the case $\gamma < 1$, we have a suboptimal rate (slower than the canonical $O(n^{-1/2})$ rate) of convergence for the DRO estimator as in,

$$\beta_n^{DRO}(\delta_n) = \beta_* - \sqrt{\eta} \frac{C^{-1}D_\beta S(\beta_*)}{n^{\gamma/2}} + o_p(n^{-\gamma/2}), \quad (10)$$

where $n^{\gamma/2}o_p(n^{-\gamma/2}) \rightarrow 0$, in probability, as $n \rightarrow \infty$. The limiting relationship (10) 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 for the sequence $\Lambda_{\delta_n}(P_n)$. (In fact, (10) indicates that scaling of order $O(n^{-\gamma/2})$ will result in a non-degenerate limit).

Finally, when $\delta_n = \eta/n$ (corresponding to the case $\gamma = 1$), we have that all the three components in the limiting result in Theorem 1 have non-trivial limits. The choice of η will be discussed in the context of coverage of the corresponding confidence region (the third term of the triplet) in the next subsection.

Proposition 1 below provides a geometric insight between $\beta_n^{DRO}(\delta_n)$, 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 fact that $\beta_n^{ERM} \in \Lambda_{\delta_n}(P_n)$ is trivial because $\Lambda_\delta(P_n)$ is increasing in δ , so $\beta_n^{ERM} \in \Lambda_0(P_n) \subset \Lambda_{\delta_n}(P_n)$. But the fact that $\beta_n^{DRO}(\delta_n) \in \Lambda_{\delta_n}(P_n)$ is not immediate and it follows a min-max argument as in Lemma 1 in [8].

Proposition 1. *Suppose that Assumptions A1) and A2) are in force, then for any $\delta > 0$,*

$$\min_{\beta} \max_{P: D(P_n, P) \leq \delta} \mathbb{E}_P[\ell(X; \beta)] = \max_{P: D(P_n, P) \leq \delta} \min_{\beta} \mathbb{E}_P[\ell(X; \beta)]. \quad (11)$$

Proposition 1 indicates that $\beta_n^{DRO}(\delta_n) \in \Lambda_{\delta_n}(P_n)$. Because if it was not the case, the left hand side in (11) would be strictly larger than the right hand side of (11).

Now, 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

$$\begin{aligned}\beta_n^{DRO}(\delta_n) &= \beta_n^{ERM} - \sqrt{\eta} \frac{C^{-1} D_{\beta} S(\beta^*)}{\sqrt{n}} + o_p(n^{-1/2}) \\ &= \beta_n^{ERM} - \sqrt{\eta} \frac{C^{-1} D_{\beta} S(\beta_n^{ERM})}{\sqrt{n}} + o_p(n^{-1/2}).\end{aligned}\tag{12}$$

The right hand side of (12) points to the canonical $O(n^{-1/2})$ rate of convergence of the Wasserstein DRO estimator and it can readily be used to construct confidence regions, as we shall explain in the next subsection.

Relation (12) also exposes the presence of an asymptotic bias towards selecting optimizers with a reduction on the size of the sensitivity with respect to the data, namely, $S(\beta) = \sqrt{\mathbb{E} \|D_x \ell(X; \beta)\|_p^2}$, in the neighborhood of β_n^{ERM} . A precise mathematical statement of this sensitivity-reduction property is given in the following result.

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

$$\bar{\beta}_n^{DRO} = \arg \min_{\beta} \left\{ \mathbb{E}_{P_n} [\ell(X; \beta)] + \sqrt{\eta} \frac{\sqrt{\mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2}}{\sqrt{n}} \right\}.$$

Then, if $\delta_n = \eta/n$, we have that $\beta_n^{DRO}(\delta_n) = \bar{\beta}_n^{DRO} + o_p(n^{-1/2})$, where $n^{1/2} o_p(n^{-1/2}) \rightarrow 0$, in probability, as $n \rightarrow \infty$.

Corollary 1 is conceptually useful to understand the Wasserstein DRO estimator as tool to mitigate the sensitivity with respect to perturbations in the data. The formulation in Corollary 1, however, may not be desirable from an optimization point of view, because the objective function may not be convex. On the other hand, under Assumption A2, the DRO objective $\Psi_n(\beta)$ is convex (see, for example, the reasoning in Theorem 2a in [11]).

A similar type of result to Corollary 1 is given in [18], 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 an approximation such as (6), uniformly over β in compact sets.

In the next section, we will discuss the implications of our main limit theorem in choosing the size of uncertainty and the construction of confidence regions.

2.2.2. Construction of Wasserstein DRO 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

$$1 - \alpha \leq \lim_{n \rightarrow \infty} \mathbf{P}(\beta_* \in \Lambda_{\delta_n}(P_n)) = \mathbf{P}(-C^{-1}H \in \{u : \varphi^*(Cu) \leq \eta\}),$$

or, equivalently, if $\eta \geq \eta_{\alpha}$, where η_{α} is the $(1 - \alpha)$ -quantile of the random variable $\varphi^*(H)$.

Remark 1. Here we allow ourselves a small abuse and use $\mathbf{P}(\cdot)$ to denote the distribution of H , hence writing $\mathbf{P}(H \in A)$ to represent the probability of the set A under the distribution of H . Also, if $\gamma = 1$ we will simply write $\Lambda_{\eta, \gamma}$ as Λ_η .

Recall that 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 Proposition 1. Following this, if we let $\eta \geq \eta_\alpha$, we then have,

$$\lim_{n \rightarrow \infty} \mathbf{P}(\beta_* \in \Lambda_{\delta_n}(P_n)) = \lim_{n \rightarrow \infty} \mathbf{P}(\beta_* \in \Lambda_{\delta_n}(P_n), \beta_n^{DRO} \in \Lambda_{\delta_n}(P_n), \beta_n^{ERM} \in \Lambda_{\delta_n}(P_n)) \geq 1 - \alpha,$$

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

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

and the associated empirical convex conjugate,

$$\varphi_n^*(\zeta) := \sup_{\xi \in \mathbb{R}^d} \{ \xi^T \zeta - \varphi_n(\xi) \}.$$

The next result provides a basis for computing a consistent estimator for η_α .

Proposition 2. *Let Ξ_n be any consistent estimator of $\text{Cov}(h(X, \beta))$ and write $\bar{\Xi}_n$ for any factorization of Ξ_n such that $\bar{\Xi}_n \bar{\Xi}_n^T = \Xi_n$. 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 $n \rightarrow \infty$ uniformly on compact sets, iii) $\varphi_n^*(\bar{\Xi}_n Z) \Rightarrow \varphi^*(H)$.*

Given the collection of samples $\{X_i\}_{i=1}^n$, we can generate i.i.d. 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 $n \rightarrow \infty$ which is enough to obtain an implementable expression for $\beta_n^{DRO}(\eta_\alpha(n)/n)$ which is asymptotically equivalent (up to $o_p(n^{-1/2})$ errors) to (12).

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, we have*

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

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

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

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

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

The second part of the result follows from the regularity results in Proposition 2. $\square$

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 3. *The support function of the convex set $\Lambda_\eta := \{u : \varphi^*(Cu) \leq \eta\}$ is given by,*

$$h_{\Lambda_\eta}(v) = 2\sqrt{\eta\varphi(C^{-1}v)},$$

where the support function of a convex set A is defined as $h_A(x) =: \sup\{x \cdot a : a \in A\}$.

Remark 2. Proposition 3 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 $u_1, \dots, u_m$, we have

$$\Lambda_\eta \subset \cap_{u_1, \dots, u_m} \{v : u_i \cdot v \leq h_{\Lambda_\eta}(u_i)\},$$

and

$$\Lambda_{\eta(n)}^n \subset \cap_{u_1, \dots, u_m} \{v : u_i \cdot v \leq h_{\Lambda_{\eta(n)}^n}(u_i)\}.$$

Proof of Proposition 3. For any convex function $f(\cdot)$ with $\inf f < 0$, it is well-known (see, for example, [34, Exercise 11.6]), that the support function of the level set $A := \{u : f(u) \leq 0\}$ is given by $h_A(v) = \inf_{\lambda > 0} \lambda f^*(\lambda^{-1}v)$, where f^* is the convex conjugate of f . Since the convex conjugate of $\varphi^*(C \times \cdot) - \eta$ is given by $\varphi(C^{-1} \times \cdot) + \eta$, the support function of $\Lambda_\eta := \{u : \varphi^*(Cu) - \eta \leq 0\}$ is given by,

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

This completes the proof of Proposition 3. $\square$

3. GEOMETRIC INSIGHTS VIA A NUMERICAL EXAMPLE

The objective of this section is to explore geometrical insights of the confidence regions induced by the Wasserstein DRO formulation and the impact of the cost function on the shape of such confidence regions.

3.1. Distributionally robust linear regression. In this section, we briefly introduce a distributionally robust version of linear regression problem (see [8] for a detailed treatment). Specifically, the data is generated by $Y = \beta_*^T X + \epsilon$ where $X \in \mathbb{R}^d$ and ϵ are independent with $C = \mathbb{E}[XX^T]$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We consider square loss $\ell(x, y; \beta) = (y - \beta^T x)^2$ and take the cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \rightarrow [0, \infty]$ to be

$$c((x, y), (u, v)) = \begin{cases} \|x - u\|_q^2 & \text{if } y = v, \\ \infty & \text{otherwise.} \end{cases} \quad (13)$$

Then, from Theorem 1 of [8], we have

$$\min_{\beta \in \mathbb{R}^d} \sup_{P: D_c(P, P_n) \leq \delta_n} \mathbb{E}_P[\ell(X, Y; \beta)] = \min_{\beta \in \mathbb{R}^d} \left(\sqrt{MSE_n(\beta)} + \sqrt{\delta_n} \|\beta\|_p \right)^2, \quad (14)$$

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

Following Corollary 2, an approximate confidence region is given by,

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

where

$$\Lambda_{\eta_\alpha} = \{\theta : \varphi^*(C\theta) \leq \eta_\alpha\}$$

and $\varphi(\xi) = \frac{1}{4} \mathbb{E} \|e\xi - (\xi^T X) \beta_*\|_p^2$, η_α is such that $\mathbf{P}(\varphi^*(H) \leq 1 - \alpha) = \eta_\alpha$ 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,

$$c(x, y) = (x - y)^T A (x - y), \quad (15)$$

for some positive definite matrix A . In this case, the respective $\Lambda_{\eta_\alpha} = \{\theta : \varphi^*(C\theta) \leq \eta_\alpha\}$ is computed in terms of

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

For the choice $c(x, y) = (x - y)^T A (x - y)$, the relationship between DRO and regularized estimators, as in (14), is given by,

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

See [9] for an account of improved out-of-sample performance resulting from Mahalanobis cost choices.

3.2. Numerical illustration. 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 functions. In practice, we use the empirical set

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

to approximate the confidence region as in Corollary 2, where $\varphi_n(\xi) = \frac{1}{4} \mathbb{E}_{P_n} \|e\xi - (\xi^T X) \beta_n^{ERM}\|_p^2$, $\eta_\alpha(n)$ is such that $\mathbf{P}(\tilde{\varphi}_n^*(H) \leq 1 - \alpha) = \eta_\alpha(n)$ for $H \sim \mathcal{N}(0, C_n \sigma_n^2)$, $\tilde{C}_n = \mathbb{E}_{P_n} [X X^T]$, and $\sigma_n^2 = \mathbb{E}_{P_n} [(Y - (\beta_n^{ERM})^T X)^2]$.

In the following numerical experiments, the data is sampled from a linear regression model with parameters $\sigma^2 = 1$, $\beta_0 = [0.5, 0.1]^T$, $n = 100$ and

$$X \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}\right).$$

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 support functions defined in Proposition 3.

In addition, a confidence region for β_* resulting from the asymptotic normality of the least-squares estimator,

$$\sqrt{n}(\beta_n^{ERM} - \beta_*) \Rightarrow \mathcal{N}(0, C^{-1} \sigma^2),$$

is given by,

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

where $\chi_{1-\alpha}^2(d)$ 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 (15) such that the resulting confidence region coincides with $\Lambda_{CLT}(P_n)$. Namely, A is chosen by solving the equation

$$\mathbb{E} \left[(e\xi - (\xi^T X) \beta_*) A^{-1} (e\xi - (\xi^T X) \beta_*)^T \right] = C\sigma^2. \quad (16)$$

Figure 1(f) gives the confidence region for the choice $p = 2$ and $\Lambda_{CLT}(P_n)$ superimposed with various DRO solutions along with the ERM solution.

From the figures, we can clearly see 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 DRO solutions all reside in their respective confidence regions but may lie outside of the confidence regions of other norms.

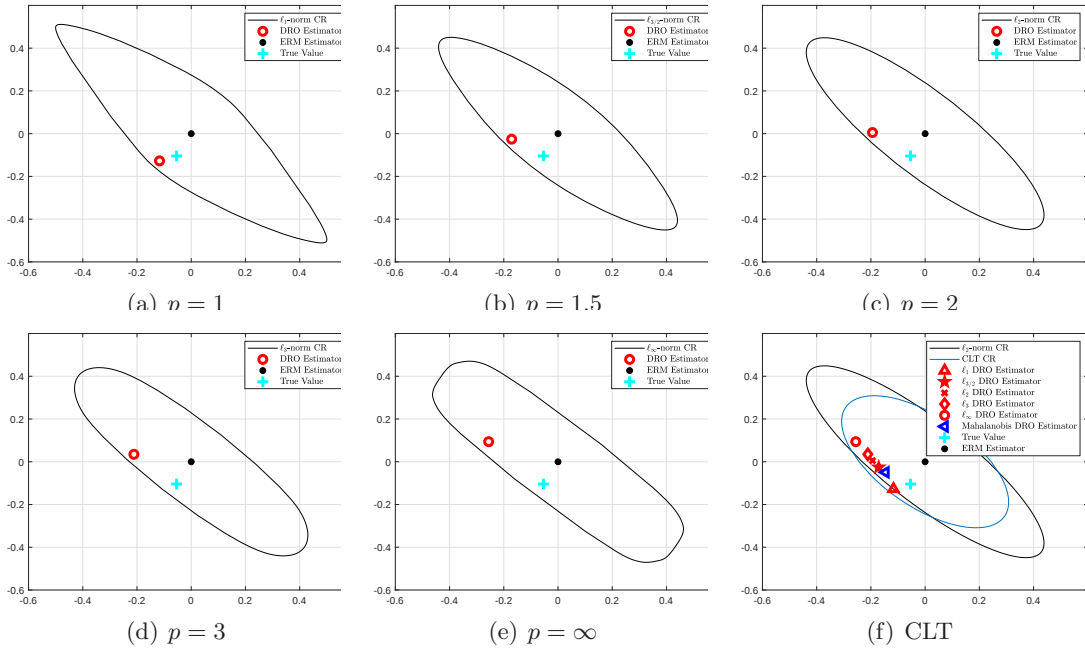


FIGURE 1. Confidence regions for different norms and CLT with DRO solutions centered at the ERM solution

We find the induced confidence regions constructed by the Wasserstein DRO formulations are somewhat similar across the various l_p norms, but they are all different to the standard CLT-confidence region. As noted, the Mahalanobis cost can be calibrated to exactly match the standard CLT confidence region.

4. PROOFS OF MAIN RESULTS

Theorem 1 is obtained by considering appropriate level sets involving auxiliary functionals which we define next.

Following [8], we define the Robust Wasserstein Profile (RWP) function, associated with the estimation of β_* by solving $E_{P_n}[D_\beta h(X, \beta)] = \mathbf{0}$, as follows:

$$R_n(\beta) := \min \{ D(P, P_n) : \mathbb{E}_P[h(X, \beta)] = \mathbf{0} \}.$$

This definition, as noted in [8], allows to characterize the set $\Lambda_\delta(P_n)$ in terms of an associated level set; in particular, we have,

$$\Lambda_\delta(P_n) = \{\beta : R_n(\beta) \leq \delta\}. \quad (17)$$

Note that $\beta \in \Lambda_\delta(P_n)$ if and only if there exists P such that $D(P, P_n) \leq \delta$ and $\mathbb{E}_P[h(X, \beta)] = \mathbf{0}$.

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

$$\begin{aligned} V_n^{DRO}(u) &:= n^{\bar{\gamma}} \left(\Psi_n \left(\beta_* + n^{-\bar{\gamma}/2} u \right) - \Psi_n(\beta_*) \right) \text{ and} \\ V_n^{ERM}(u) &:= n \left(\mathbb{E}_{P_n} \left[\ell(X; \beta_* + n^{-1/2} u) \right] - \mathbb{E}_{P_n} [\ell(X; \beta_*)] \right). \end{aligned}$$

Moreover, define $V : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ via

$$V(x, u) = x^T u + \frac{1}{2} u^T C u.$$

The following result, as we shall see, can be used to establish Theorem 1 directly.

Theorem 2. *Suppose that the assumptions made in Theorem 1 hold. Then we have,*

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

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

4.1. Proof of Theorem 2. For all the results stated in this section, we assume that the assumptions imposed in Theorem 1 are satisfied. Let

$$H_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n h(X_i, \beta_*)$$

The following sequence of results will be useful in proving Theorem 2.

Proposition 4. *Fix $\alpha \in [0, 1]$. Given $\varepsilon', K > 0$, there exists a positive integer n_0 such that*

$$\left| n^{\alpha-1} V_n^{ERM}(n^{(1-\alpha)/2} u) - n^{(\alpha-1)/2} H_n^T u - \frac{1}{2} u^T C u \right| \leq \varepsilon',$$

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

$$\left| V_n^{ERM}(u) - H_n^T u - \frac{1}{2} u^T C u \right| \leq \varepsilon'.$$

Proposition 5. *Given $\varepsilon', K > 0$, there exists a positive integer n_0 such that then*

$$\left| V_n^{DRO}(u) + f_{\eta, \gamma}(-H_n)^T u - \frac{1}{2} u^T C u \right| \leq \varepsilon'.$$

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

Proposition 6. *We have,*

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

where $M_n(\xi, u) :=$

$$\frac{1}{n} \sum_{i=1}^n \max_{\Delta} \left\{ \xi^T \int_0^1 \left(D_x h \left(X_i + t \frac{\Delta}{\sqrt{n}}, \beta_* + t \frac{u}{\sqrt{n}} \right) \Delta + D_{\beta} h \left(X_i + t \frac{\Delta}{\sqrt{n}}, \beta_* + t \frac{u}{\sqrt{n}} \right) u \right) dt - \|\Delta\|_q^2 \right\}.$$

Proposition 7. Consider any $\varepsilon, \varepsilon', K > 0$. Then there exist $b_0, c_0 \in (0, \infty)$ such that for any $b \geq b_0, c \geq c_0$, we have a positive integer n_0 such that,

$$\mathbf{P} \left(\sup_{\|u\|_2 \leq K} \left(nR_n \left(\beta_* + n^{-1/2}u \right) - f_{up}(H_n, u, b, c) \right) \leq \varepsilon' \right) \geq 1 - \varepsilon,$$

for all $n \geq n_0$, where

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

Proposition 8. For any $\varepsilon, \varepsilon', K, b > 0$, there exists a positive integer n_0 such that,

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

for all $n > n_0$, where

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

Proposition 9. For any $\varepsilon > 0$, there exist positive constants n_0, a such that,

$$\mathbf{P} (nR_n(\beta_*) \leq a) \geq 1 - \varepsilon,$$

for every $n \geq n_0$.

Proposition 10. For any $\varepsilon, \varepsilon', K > 0$, there exist positive constants n_0, δ such that,

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

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

Proofs of Propositions 4 - 10 are furnished in Section 4.4. With the statements of these results, we proceed with the proof of Theorem 2 as follows. Since $\mathbb{E}[h(X, \beta_*)] = \mathbf{0}$, it follows from central limit theorem that

$$H_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n h(X_i, \beta_*) \Rightarrow -H,$$

where $H \sim \mathcal{N}(0, \mathbb{E}[h(X, \beta_*)h(X, \beta_*)^T])$. Since $V_n^{ERM}(\cdot)$ and $V_n^{DRO}(\cdot)$ are continuous, it follows from Propositions 4 and 5 that,

$$V_n^{ERM}(\cdot) \Rightarrow V^{ERM}(\cdot) := V(-H, \cdot) \quad \text{and} \quad V_n^{DRO}(\cdot) \Rightarrow V^{DRO}(\cdot) := V(-f_{\eta, \gamma}(H), \cdot), \quad (18)$$

on the space topologized by uniform convergence on compact sets.

To prove convergence of the third component of the triplet considered in Theorem 2, observe from the definitions of $\varphi^*(\cdot)$ and C that,

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

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 7, 8 result in the conclusions that,

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

and

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

for any u satisfying $\|u\|_2 \leq K$. Since the bounds in Propositions 7, 8 hold for arbitrarily large choices for constants b, c , we conclude from the observations (20), (21) and (19) that

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

for any u satisfying $\|u\|_2 \leq K$.

Finally, we have from Propositions 9 and 10 that the collection $\{nR_n(\beta_* + n^{-1/2} \times \cdot)\}$ is tight (see, for example, Theorem 7.4 in [4]). As a consequence of this tightness and the finite dimensional convergence in (22), we have that,

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

Combining this observation with those in (18), we obtain the desired convergence result in Theorem 2. $\square$

4.2. 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 2.

To reduce clutter in expressions, from here-onwards we refer the DRO estimator (2), simply as β_n^{DRO} , with the dependence on the radius δ to be understood from the context.

To begin, consider the following tightness result whose proof is provided in Section 4.3.

Proposition 11. *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 (see 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 11, and the weak convergence of $V_n^{ERM}(\cdot)$ and $V_n^{DRO}(\cdot)$ in Theorem 2, we have that

$$n^{1/2}(\beta_n^{ERM} - \beta_*) \Rightarrow \arg \min_u V(-H, u) = C^{-1}H \quad (23)$$

and

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

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and $\alpha \in [0, +\infty]$, let $\text{lev}(f, \alpha)$ denote the level set $\{x \in \mathbb{R}^d : f(x) \leq \alpha\}$.

Proposition 12. *If $\delta_n = \eta/n$, then $\text{lev}(G_n, \alpha_n) \Rightarrow \text{lev}(G, \eta)$.*

Proposition 13. *If $\delta_n = \eta/n^\gamma$ for some $\gamma > 1$, then $\text{lev}(G_n, \alpha_n) \Rightarrow \{C^{-1}H\}$.*

Proposition 14. *If $\delta_n = \eta/n^\gamma$ for some $\gamma < 1$, then $\text{lev}(G_n, \alpha_n) \Rightarrow \mathbb{R}^d$.*

Propositions 12 - 14 above, whose proofs are furnished in Section 4.3, 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 12 - 14 that

$$n^{1/2}(\Lambda_{\delta_n}(P_n) - \beta_*) = \{u : G_n(u) \leq \alpha_n\} \Rightarrow \begin{cases} \text{lev}(G, \eta) & \text{if } \gamma = 1, \\ \mathbb{R}^d & \text{if } \gamma < 1, \\ \mathbf{0} & \text{if } \gamma > 1. \end{cases}$$

Observe that $\varphi^*(u) = \varphi^*(-u)$. Therefore, $\text{lev}(G, \eta) = \{u : \varphi^*(H - Cu) \leq \eta\} = C^{-1}H + \{u : \varphi^*(Cu) \leq \eta\}$. This completes the proof of Theorem 1. $\square$

Now we provide the proofs of technical propositions involved in Theorem 1.

4.3. Proofs of Propositions 11 - 14. In this section, we provide proofs of Propositions 11 - 14, which are key in the proof of Theorem 1.

Proof of Proposition 11. 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

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

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 $C := E[D_\beta h(X, \beta_*)]$ (see 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 $\sup_n P(\|H_n\|_2 > K_1) \leq \varepsilon$,

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

Observe from the definition of $f_{\eta,\gamma}(\cdot)$ that $K_2 \in (0, +\infty)$. Due to Propositions 4 and 5, there exists n_0 such that,

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

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

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

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,

$$\begin{aligned} \sup_{n \geq n_0} \mathbf{P}\left(\arg \min_u V_n^{ERM}(u) \subseteq U\right) &\geq 1 - \varepsilon \\ \text{and} \quad \sup_{n \geq n_0} \mathbf{P}\left(\arg \min_u V_n^{DRO}(u) \subseteq U\right) &\geq 1 - \varepsilon, \end{aligned}$$

thus verifying the claim. $\square$

Proof of Proposition 12. Due to Theorem 2, 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,

$$\sup_{\|u\| \leq K} |G_n(u) - G(u)| \rightarrow 0, \quad (24)$$

happen almost surely, for every $K \in (0, \infty)$. Since $G(\cdot)$ is continuous, we have from Theorem 7.14 of [34] that G_n converges continuously to G , almost surely. A simple consequence of this observation, see [34, 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 $\varphi^*(\mathbf{0}) = 0$. Moreover, since $\alpha_n := n\delta_n = \eta \in (0, +\infty)$, we have that

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

almost surely, in the Painelevé-Kuratowski sense (see, for example, [2, Theorem 5.1], [44, Theorem 7.1], or [3]). Consequently, $\text{lev}(G_n, \alpha_n) \Rightarrow \text{lev}(G, \eta)$. $\square$

Proof of Proposition 13. Following the same reasoning used in the proof of Proposition 12 to arrive at (24), we have a probability space where the convergence,

$$\sup_{\|u\| \leq K} |G_n(u) - G(u)| \rightarrow 0 \quad \text{and} \quad n^{1/2} (\beta_n^{ERM} - \beta_*) \rightarrow C^{-1}H, \quad (25)$$

happen almost surely, for every $K \in (0, \infty)$; here, the latter convergence follows from (23).

Next, observe that $\alpha_n := n\delta_n \rightarrow 0$. Then, we have from [34, Proposition 7.7a] that

$$\text{Ls}_{n \rightarrow \infty} \text{lev}(G_n, \alpha_n) \subseteq \text{lev}(G, 0) = \{u : \varphi^*(H - Cu) = 0\} = \{C^{-1}H\}, \quad (26)$$

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

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

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

for every n . Therefore, from the second convergence in (25), we obtain

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

Combining this observation with that in (26), we obtain that $\text{PK-lim}_n \text{lev}(G_n, \alpha_n) = \{C^{-1}H\}$ almost surely. As a result, $\text{lev}(G_n, \alpha_n) \Rightarrow \{C^{-1}H\}$. $\square$

Proof of Proposition 14. Following the same reasoning in the proof of Proposition 12, we have a probability space where the convergence in (24) happen almost surely, for every $K \in (0, +\infty)$. Consider any fixed $u \in \mathbb{R}^d$. Since $G_n(u) \rightarrow G(u) < \infty$ almost surely and $\alpha_n = n\delta_n \rightarrow \infty$, 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

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

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

4.4. Proofs of Propositions 4 - 10. In this section we present the proofs of Propositions 4 - 10, which have been crucial in proving Theorem 2.

Proof of Proposition 4. Recall that $h(x; \beta) := D_\beta \ell(x; \beta)$. With $\ell(\cdot)$ being twice continuously differentiable, employing Taylor expansion up to quadratic term, we obtain,

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

as $n \rightarrow \infty$, uniformly over u in compact sets. With this expansion, the statement of Proposition 4 follows as a direct consequence of the definitions, $H_n := n^{-1/2} \sum_{i=1}^n h(X_i, \beta_*)$, $C := E_P[D_\beta h(X; \beta_*)]$ and an application of law of large numbers, $\lim_{n \rightarrow \infty} E_{P_n}[D_\beta h(X, \beta_*)] \rightarrow C$. $\square$

Proposition 15 below is useful to verify Proposition 5. In order to state Proposition 15, define $\tilde{\Delta} := \|D_x \ell(X; \beta)\|_p^{1-p/q} \text{sgn}(D_x \ell(X; \beta)) |D_x \ell(X; \beta)|^{p/q}$ and $a_n(\beta) := E_{P_n} [\tilde{\Delta}^T D_{xx} \ell(X; \beta) \tilde{\Delta}]$. (27)

Proposition 15. *As $n \rightarrow \infty$, we have,*

$$\Psi_n(\beta) = \mathbb{E}_{P_n}[\ell(X; \beta)] + \sqrt{\delta_n} \sqrt{\mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2} + \frac{\delta_n a_n(\beta)}{2 \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2} + o(\delta_n),$$

uniformly over β in compact sets.

Proposition 15 follows as a consequence of the Lemma 1 below, whose proof is technical and is presented in Appendix A.

Lemma 1. *Suppose that $b_n := \delta_n n^\gamma \rightarrow \eta$ as $n \rightarrow \infty$. Then, as $n \rightarrow \infty$,*

$$\Psi_n(\beta) = \mathbb{E}_{P_n}[\ell(X; \beta)] + n^{-\gamma/2} \inf_{\lambda \geq \lambda_0(n)} \left\{ \lambda b_n + \frac{1}{4\lambda} \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2 + \frac{a_n(\beta)}{8\lambda^2 n^{\gamma/2}} \right\} + o(n^{-\gamma}),$$

uniformly over β in compact sets, where $\lambda_0(n) := \sqrt{E_{P_n} \|D_x \ell(X; \beta)\|_p^2 / (2b_n)}$.

Proof of Proposition 15. To prove the statement of Proposition 15, we begin by analyzing the term involving infimum in the statement of Lemma 1. Define,

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

where $a := b_n$, $b := \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2 / 4$, $c := a_n(\beta)/8$ and $\varepsilon = n^{-\gamma/2}$. Changing variables as in $\lambda = \sqrt{b/a}(1 + \varepsilon u \sqrt{a})$ results in,

$$\min_{\lambda \geq \sqrt{b/a}} g_1(\lambda) = \min_{u \geq 0} g_2(u), \quad (28)$$

where

$$g_2(u) := \sqrt{ab}(1 + \varepsilon u \sqrt{a}) + \frac{\sqrt{ab}}{1 + \varepsilon u \sqrt{a}} + \frac{\varepsilon ac}{b(1 + \varepsilon u \sqrt{a})^2}.$$

Then observe that

$$\frac{1}{\varepsilon^2 a^{3/2}} \left(g_2(u) - 2\sqrt{ab} - \varepsilon ac/b \right) = g_3(u), \quad (29)$$

where

$$g_3(u) := \frac{\sqrt{b}u^2}{1 + \varepsilon u \sqrt{a}} - \frac{c}{b} \frac{u(2 + \varepsilon u \sqrt{a})}{(1 + \varepsilon u \sqrt{a})^2}.$$

Since

$$g_3(u) \geq \frac{\sqrt{b}u^2}{1 + \varepsilon u \sqrt{a}} - \frac{c}{b} \frac{2u}{(1 + \varepsilon u \sqrt{a})} = \frac{\sqrt{b}u^2 - (2c/b)u}{1 + \varepsilon u \sqrt{a}},$$

for $u \geq 0$, we have that, $\inf_{u \geq 0} g_3(u) = 0$ if $c = 0$ and $\inf_{u \geq 0} g_3(u) < 0$ if $c > 0$. For the case $c > 0$, when $u > 0$ is such that $\sqrt{b}u^2 - (2c/b)u < 0$ we have $g_3(u) > \sqrt{b}u^2 - (2c/b)u \geq -c^2 b^{-5/2}$. Therefore,

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

Combining this observation with (28) and (29), we obtain that,

$$\left| \min_{\lambda \geq \sqrt{b/a}} g_1(\lambda) - 2\sqrt{ab} - \varepsilon ac/b \right| \leq \varepsilon^2 c^2 a^{3/2} b^{-5/2}.$$

It follows from the definitions of $g_1(\cdot)$, a, b, c, ε that $\min_{\lambda \geq 0} g_1(\lambda)$ is exactly the same as the term involving infimum in Lemma 1. Utilizing the above bound in the statement of Lemma 1, we arrive at the statement of Proposition 15. $\square$

Proof of Proposition 5. For ease of notation, define $S_n(\beta) := \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2$. Then it follows from the definitions of $V_n^{DRO}(\cdot)$, $V_n^{ERM}(\cdot)$ and the conclusion in Lemma 15 that,

$$\begin{aligned} V_n^{DRO}(u) &= n^{\tilde{\gamma}-1} V_n^{ERM}(n^{(1-\tilde{\gamma})/2}u) + n^{\tilde{\gamma}} \sqrt{\delta_n} \left(S_n(\beta_* + n^{-\tilde{\gamma}/2}u) - S_n(\beta_*) \right) \\ &\quad + \frac{n^{\tilde{\gamma}} \delta_n}{2} \left(\frac{a_n(\beta_* + n^{-\tilde{\gamma}/2}u)}{S_n(\beta_* + n^{-\tilde{\gamma}/2}u)} - \frac{a_n(\beta_*)}{S_n(\beta_*)} \right) + o(1), \end{aligned} \quad (30)$$

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

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

as $n \rightarrow \infty$, 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 4, we obtain the conclusion of Proposition 5. $\square$

Proof of Proposition 6. By utilizing the duality for linear semi-infinite programs exactly as in the proof of Proposition 3 of [8], we obtain that

$$\begin{aligned} nR_n(\beta_* + n^{-1/2}u) &= \max_{\xi} \left\{ - \sum_{i=1}^n \xi^T h(X_i, \beta_* + n^{-1/2}u) \right. \\ &\quad \left. - \sum_{i=1}^n \max_{\Delta} \left\{ \xi^T \left(h(X_i + \Delta, \beta_* + n^{-1/2}u) - h(X_i, \beta_* + n^{-1/2}u) \right) - \|\Delta\|_q^2 \right\} \right\}. \end{aligned}$$

As a result,

$$\begin{aligned} nR_n(\beta_* + n^{-1/2}u) &= \max_{\xi} \left\{ - \sum_{i=1}^n \max_{\Delta} \left\{ \xi^T h(X_i + \Delta, \beta_* + n^{-1/2}u) - \|\Delta\|_q^2 \right\} \right\} \\ &= \max_{\xi} \left\{ - \sum_{i=1}^n \xi^T h(X_i, \beta_*) - \sum_{i=1}^n \max_{\Delta} \left\{ \xi^T \left(h(X_i + \Delta, \beta_* + n^{-1/2}u) - h(X_i, \beta_*) \right) - \|\Delta\|_q^2 \right\} \right\}. \end{aligned}$$

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

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

where

$$\begin{aligned} M_n(\xi, u) &= \frac{1}{n} \sum_{i=1}^n \max_{\Delta} \left\{ \sqrt{n} \xi^T \left(h(X_i + n^{-1/2} \Delta, \beta_* + n^{-1/2}u) - h(X_i, \beta_*) \right) - \|\Delta\|_q^2 \right\} \quad (31) \\ &= \frac{1}{n} \sum_{i=1}^n \max_{\Delta} \left\{ \xi^T \int_0^1 \left(D_x h \left(X_i + t \frac{\Delta}{\sqrt{n}}, \beta_* + t \frac{u}{\sqrt{n}} \right) \Delta + D_\beta h \left(X_i + t \frac{\Delta}{\sqrt{n}}, \beta_* + t \frac{u}{\sqrt{n}} \right) u \right) dt - \|\Delta\|_q^2 \right\}, \end{aligned}$$

where the latter equality follows from the fundamental theorem of calculus. This completes the proof of Proposition 6. $\square$

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

$$\mathbf{P} \left(\max_{\|\xi\|_q \geq b} \left\{ -\xi^T H_n - M_n(\xi, u) \right\} > 0 \right) \leq \varepsilon,$$

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

Lemma 3. *For any positive constants b, c_0 , we have*

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left\{ \left\| (D_x h(X, \beta_*)^T \xi \right\|_p^2 + \xi^T D_\beta h(X_i, \beta_*) u \right\} \mathbb{I}(X_i \in C_0) \\ & \rightarrow \mathbb{E} \left[\left\{ \left\| (D_x h(X, \beta_*)^T \xi \right\|_p^2 + \xi^T D_\beta h(X_i, \beta_*) u \right\} \mathbb{I}(X_i \in C_0) \right], \end{aligned}$$

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

Proofs of Lemma 2 - 3 are presented in the appendix. The following definitions are useful in the proofs of Proposition 7 and Lemma 2. For a fixed u, Δ , let

$$I(X_i, \Delta, u) = I_1(X_i, \Delta, u) + I_2(X_i, \Delta, u), \quad (32)$$

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

$$\begin{aligned} I_1(X_i, \Delta, u) &:= \int_0^1 \left(D_x h \left(X_i + t \frac{\Delta}{\sqrt{n}}, \beta_* + t \frac{u}{\sqrt{n}} \right) - D_x h(X_i, \beta_*) \right) \Delta dt \quad \text{and} \\ I_2(X_i, \Delta, u) &:= \int_0^1 \left(D_\beta h \left(X_i + t \frac{\Delta}{\sqrt{n}}, \beta_* + t \frac{u}{\sqrt{n}} \right) - D_\beta h(X_i, \beta_*) \right) u dt. \end{aligned}$$

In addition, recall that $q > 1$ and $p = q/(q-1)$. For $\xi \neq 0$, we write $\bar{\xi} = \xi / \|\xi\|_p$. Let us define the vector $V_i(\bar{\xi}) = D_x h(X_i, \beta_*)^T \bar{\xi}$ and put

$$\Delta'_i = \Delta'_i(\bar{\xi}) = |V_i(\bar{\xi})|^{p/q} \text{sgn}(V_i(\bar{\xi})). \quad (33)$$

Proof of Proposition 7. First observe that $R_n(\cdot) \geq 0$ (consider the choice $\xi = \mathbf{0}$). Given $K, \varepsilon > 0$, define the event,

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

where $b > 0$ is such that $P(\mathcal{A}_n) \geq 1 - \varepsilon$. Such a $b \in (0, \infty)$ exists because of Lemma 2.

Next, for any $c_0 > 0$, define

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

where $C_0 := \{w : \|w\|_p \leq c_0\}$, $I(X_i, \Delta, u)$ is defined as in (32) and $\bar{\Delta}_i = c_i \Delta'_i$, which is defined in (33) with c_i chosen so that

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

Then we have, $M_n(\xi, u) \geq M'_n(\xi, u, c_0)$, for every u . With these definitions, observe that

$$\begin{aligned} \max_{\Delta} \left\{ \xi^T D_x h(X_i, \beta_*) \Delta - \|\Delta\|_q^2 \right\} &= \xi^T D_x h(X_i, \beta_*) \bar{\Delta}_i - \|\bar{\Delta}_i\|_q^2 \\ &= \frac{1}{4} \left\| (D_x h(X, \beta_*)^T \xi \right\|_p^2 \end{aligned} \quad (34)$$

and

$$\max_{\|\xi\|_q \leq b} \{ -\xi^T H_n - M_n(\xi, u) \} \leq \max_{\|\xi\|_q \leq b} \{ -\xi^T H_n - M'_n(\xi, u, c_0) \}. \quad (35)$$

Next, define

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

where the equality follows from (34). Due to Lemma 3, we have

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

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

$$\sup_{\|\xi\|_p \leq b} \left| \hat{M}_n(\xi, u, c_0) - M'_n(\xi, u, c_0) \right| \rightarrow 0, \quad (36)$$

because, as in the proof of Lemma 3, we have from the continuity of $D_\beta h(\cdot)$ and $D_x h(\cdot)$ that

$$|\xi^T I(X_i, \bar{\Delta}_i, u)| \mathbb{I}(X_i \in C_0) \rightarrow 0, \quad (37)$$

uniformly over $\|\xi\|_p \leq b$ and $\|u\|_2 \leq K$. Combining the observations in (36) and (37), we obtain that for any $\varepsilon' > 0$ there exists n_0 sufficiently large such that,

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

Then the statement of Proposition 7 follows from (35), the definition of the event $\mathcal{A}_n$ and the observation that $P(\mathcal{A}_n) \geq 1 - \varepsilon$. $\square$

Proof of Proposition 8. For the lower bound, we reexpress the expression for $M_n(\xi, u)$ in (31) as follows:

$$\begin{aligned}M_n(\xi, u) &= \frac{1}{n} \sum_{i=1}^n \max_{\Delta} \left\{ \sqrt{n} \xi^T \left(h(X_i + n^{-1/2} \Delta, \beta_* + n^{-1/2} u) - h(X_i, \beta_* + n^{-1/2} u) \right) - \|\Delta\|_q^2 \right\} \\ &\quad + \frac{1}{n} \sum_{i=1}^n \sqrt{n} \xi^T \left(h(X_i, \beta_* + n^{-1/2} u) - h(X_i, \beta_*) \right). \quad (38)\end{aligned}$$

Employing fundamental theorem of calculus, we obtain that

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

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

$$\frac{1}{n} \sum_{i=1}^n \sqrt{n} \xi^T \left(h(X_i, \beta_* + n^{-1/2}u) - h(X_i, \beta_*) \right) \leq \xi^T \mathbb{E}[D_\beta h(X, \beta_*)]u + \varepsilon'/2, \quad (39)$$

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 [8] that there exists n_0 such that,

$$\sup_{\|\Delta\|_q \geq \nu\sqrt{n}} \left\{ \sqrt{n} \xi^T \left(h(X_i + n^{-1/2}\Delta, \beta_* + n^{-1/2}u) - h(X_i, \beta_* + n^{-1/2}u) \right) - \|\Delta\|_q^2 \right\} \leq 0, \quad (40)$$

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 (38) is bounded from above by

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

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

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

Then, for n_0 suitably large, a similar application of Lemma 3 as in Proposition 7 results in,

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

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 (39) to obtain that,

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

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

Proof of Proposition 9. Under Assumptions A1 - A2, we have from Proposition 5 of [8] (or equivalently from Propositions 7,8 applied specifically with $u = 0$) that,

$$nR_n(\beta_*) \Rightarrow \psi(H),$$

where $H \sim \mathcal{N}(\mathbf{0}, \text{Cov}[h(X, \beta_*)])$. Given $\varepsilon > 0$, let $a > 0$ be such that $P(\psi(H) \leq x)$ is continuous at $x = a$ and $P(\psi(H) > a) \leq \varepsilon/2$. Due to the above convergence in distribution, there exists n_0 such that,

$$|\mathbf{P}(nR_n(\beta_*) \geq a) - P(\psi(H) \geq a)| \leq \varepsilon/2,$$

for all $n \geq n_0$. Therefore, we have $\sup_{n \geq n_0} P(nR_n(\beta_*) \geq a) \leq \varepsilon$. $\square$

Lemma 4 below is useful to prove Proposition 10. Proof of Lemma 4 is presented in the appendix.

Lemma 4. *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 10. For any $u_j, j = 1, 2$, satisfying $\|u_j\|_2 \leq K$, let ξ_j attain the supremum in the relation $n^{\rho/2}R_n(\beta_* + n^{-1/2}u_j) := \sup_{\xi} \{-\xi^T H_n - M_n(\xi, u_j)\}$. Then we have,

$$\left| nR_n(\beta_* + n^{-1/2}u_1) - nR_n(\beta_* + n^{-1/2}u_2) \right| \leq \max_{j=1,2} |M_n(\xi_j, u_1) - M_n(\xi_j, u_2)|. \quad (41)$$

For the given choices of ε, K , we have from Lemma 2 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 (41) and Lemma 4 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 10. $\square$

4.5. Proofs of Proposition 1 and Corollary 1. To prove Proposition 1, we first restate Proposition 8 in Appendix C of [8] here.

Proposition 16 (Proposition 8, [8]). *Let us define*

$$g(\beta) = \max_{P: D(P_n, P) \leq \delta} \mathbb{E}_P[\ell(X; \beta)].$$

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

$$\min_{\beta} \max_{P: D(P_n, P) \leq \delta} \mathbb{E}_P[\ell(X; \beta)] = \max_{P: D(P_n, P) \leq \delta} \min_{\beta} \mathbb{E}_P[\ell(X; \beta)].$$

Proof of Proposition 1. As the loss function $\ell(x; \beta)$ is a continuous and convex function of β , we have $\mathbb{E}_P[\ell(X; \beta)]$ and $g(\beta)$ are continuous and convex. Then, we pick $\epsilon > 0$, such that $P_n^\epsilon = (1 - \epsilon)P_n + \epsilon P_* \in U_{\delta_n}(P_n)$. Therefore, we have $\kappa_b \subset \{\beta : \mathbb{E}_{P_n^\epsilon}[\ell(X; \beta)] \leq b\} \subset \{\beta : \mathbb{E}_{P_*}[\ell(X; \beta)] \leq b/\epsilon\}$. Since $\mathbb{E}_{P_*}[\ell(X; \beta)]$ is locally strong convex by Assumption A2.b), $\{\beta : \mathbb{E}_{P_*}[\ell(X; \beta)] \leq b/\epsilon\}$ is compact. As all the conditions in Proposition 16 are satisfied, the desired result follows. $\square$

Proof of Corollary 1. Define $\bar{\Psi}_n(\beta) := E_{P_n}[\ell(X; \beta)] + \sqrt{\eta}n^{-1/2}\sqrt{E_{P_n}\|D_x\ell(X; \beta)\|^2}$ 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 5, 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 11), 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 2, 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 $n^{1/2}(\bar{\beta}_n^{DRO} - \beta_*)$ and $C^{-1}f_{\eta,1}(H)$, we have that

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

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

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

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

4.6. Proof of Proposition 2.

Proof of i). Since $\mathbb{E}[D_x h(X, \beta_*) D_x h(X, \beta_*)^T] \succ \mathbf{0}$, we have $\varphi(\xi) \geq c \|\xi\|_2^2$ for some numerical constant $c > 0$ and thus $\varphi^*(\cdot)$ is continuous. $\square$

Proof of ii). Since $\beta_n^{ERM} \xrightarrow{a.s.} \beta_*$, we have $\varphi_n(\xi) \xrightarrow{a.s.} \varphi(\xi)$ for any ξ . Then, since $\|\xi^T D_x h(X, \beta_*)\|_p^2$ is Lipschitz in ξ for $\|\xi\|_p \leq b$, i.e., for $\|\xi_1\|_p \leq b, \|\xi_2\|_p \leq b$,

$$\left| \left\| (D_x h(X, \beta_*))^T \xi_1 \right\|_p^2 - \left\| (D_x h(X, \beta_*))^T \xi_2 \right\|_p^2 \right| \leq 2b \|D_x h(X, \beta_*)\|_q^2 \|\xi_1 - \xi_2\|_p$$

and $\mathbb{E}[\|D_x h(X, \beta_*)\|_q^2] < \infty$, we have the uniform law of large numbers that

$$\sup_{\|\xi\|_p \leq b} |\varphi_n(\xi) - \varphi(\xi)| \xrightarrow{a.s.} 0,$$

uniformly over $\|\xi\|_p \leq b$. Then, we have

$$\sup_{\xi: \|\xi\|_p \leq b} \{\xi^T \zeta - \varphi_n(\xi)\} \xrightarrow{a.s.} \sup_{\xi: \|\xi\|_p \leq b} \{\xi^T \zeta - \varphi(\xi)\}$$

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

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

REFERENCES

- [1] Soroosh Shafieezadeh Abadeh, Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Wasserstein distributionally robust kalman filtering. In *Advances in Neural Information Processing Systems*, pages 8483–8492, 2018.
- [2] G. Beer and R. Lucchetti. Minima of quasi-convex functions. *Optimization*, 20(5):581–596, 1989.
- [3] Gerald Beer, R. T. Rockafellar, and Roger 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.
- [4] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.
- [5] Jose Blanchet, Lin Chen, and Xun Yu Zhou. Distributionally robust mean-variance portfolio selection with Wasserstein distances. *arXiv preprint arXiv:1802.04885*, 2018.
- [6] Jose Blanchet, Arun Jambulapati, Carson Kent, and Aaron Sidford. Towards optimal running times for optimal transport. *arXiv preprint arXiv:1810.07717*, 2018.
- [7] Jose Blanchet and Yang Kang. Distributionally robust groupwise regularization estimator. *arXiv preprint arXiv:1705.04241*, 2017.
- [8] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56.3, September 2019.
- [9] Jose Blanchet, Yang Kang, Fan Zhang, and Karthyek Murthy. Data-driven optimal transport cost selection for distributionally robust optimization. *arXiv preprint arXiv:1705.07152*, 2017.
- [10] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.
- [11] Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport based distributionally robust optimization: Structural properties and iterative schemes. *arXiv preprint arXiv:1810.02403*, 2018.
- [12] John Gunnar Carlsson, Mehdi Behroozi, and Kresimir Mihic. Wasserstein distance and the distributionally robust tsp. *Operations Research*, 66(6):1603–1624, 2018.
- [13] Ruidi Chen and Ioannis Ch Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *The Journal of Machine Learning Research*, 19(1):517–564, 2018.
- [14] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-driven chance constrained programs over wasserstein balls. *arXiv preprint arXiv:1809.00210*, 2018.
- [15] Pedro Cisneros-Velarde, Sang-Yun Oh, and Alexander Petersen. Distributionally robust formulation and model selection for the graphical lasso. *arXiv preprint arXiv:1905.08975*, 2019.
- [16] John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts.
- [17] Jitka Dupacova and Roger Wets. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Ann. Statist.*, 16(4):1517–1549, 12 1988.
- [18] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.
- [19] Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.
- [20] Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. Robust hypothesis testing using wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pages 7902–7912, 2018.
- [21] Grani A Hanasusanto and Daniel Kuhn. Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. *Operations Research*, 2018.
- [22] Peter 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.
- [23] Keith Knight, Wenjiang Fu, et al. Asymptotics for lasso-type estimators. *The Annals of statistics*, 28(5):1356–1378, 2000.
- [24] Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS’18, pages 2692–2701, USA, 2018. Curran Associates Inc.
- [25] Tianyi Lin, Nhat Ho, and Michael I Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. *arXiv preprint arXiv:1901.06482*, 2019.
- [26] Fengqiao Luo and Sanjay Mehrotra. Decomposition algorithm for distributionally robust optimization using Wasserstein metric. *arXiv preprint arXiv:1704.03920*, 2017.

- [27] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, Sep 2018.
- [28] Ilya S Molchanov. *Theory of random sets*, volume 19. Springer, 2005.
- [29] Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *arXiv preprint arXiv:1805.07194*, 2018.
- [30] Art Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.
- [31] Art Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, pages 90–120, 1990.
- [32] Art B Owen. *Empirical likelihood*. Chapman and Hall/CRC, 2001.
- [33] Gabriel Peyr and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.
- [34] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.
- [35] Soroosh Shafieezadeh-Abadeh, Peyman Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584. 2015.
- [36] Shai Shalev-Shwartz and Yoram Singer. Online learning: Theory, algorithms, and applications. 2007.
- [37] Alexander Shapiro. Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.*, 17(2):841–858, 06 1989.
- [38] Alexander Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186, Dec 1991.
- [39] Alexander Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845, 1993.
- [40] Alexander Shapiro. On the asymptotics of constrained local m -estimators. *Ann. Statist.*, 28(3):948–960, 05 2000.
- [41] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.
- [42] Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, 2017.
- [43] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018.
- [44] R. A. Wijsman. Convergence of sequences of convex sets, cones and functions. ii. *Transactions of the American Mathematical Society*, 123(1):32–45, 1966.
- [45] I. Yang. A convex optimization approach to distributionally robust markov decision processes with Wasserstein distance. *IEEE Control Systems Letters*, 1(1):164–169, July 2017.
- [46] Insoon Yang. A convex optimization approach to distributionally robust markov decision processes with Wasserstein distance. *IEEE control systems letters*, 1(1):164–169, 2017.
- [47] Insoon Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *arXiv preprint arXiv:1812.09808*, 2018.
- [48] Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262 – 267, 2018.

Appendix A. TECHNICAL PROOFS OF LEMMAS 1 - 4

The following technical result is useful in proving Lemma 1.

Lemma 5. *Suppose $\Delta, \Delta^* \in R^d$, $\xi = \Delta - \Delta^*$, and D_* denote the derivative of the function $\|\Delta\|_q^2$ evaluated at $\Delta = \Delta_*$. Then,*

a) *for $q \in (1, 2]$, we have*

$$\|\Delta\|_q^2 \geq \|\Delta^*\|_q^2 + D_*^T \xi + (q-1) \|\xi\|_q^2;$$

b) *for $q > 2$, we have*

$$\|\Delta\|_q^2 \geq \|\Delta^*\|_q^2 + D_*^T \xi + C \min \left(\|\xi\|_2^2, \frac{\|\xi\|_q^q}{\|\Delta^*\|_{q-2}^{q-2}} \right),$$

where $C > 0$ is a positive constant that only depends on d and q .

Proof of Lemma 5. A proof of the conclusion in Part a) of Lemma 5 can be found in Appendix A of [36]. For the proof of Part b), we proceed as follows.

Suppose $H(\Delta)$ denotes hessian matrix of function $\frac{1}{2} \|\Delta\|_q^2$. Then, for any $x \in R^d$,

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

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

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

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

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

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

$$x^T H(\Delta) x \geq \frac{\sum_{i=1}^d |\Delta_i|^{q-2} (x_i)^2}{\left\| |\Delta|^{q-2} \right\|_{\frac{q}{q-2}}} \geq \frac{\sum_{i=1}^d |\Delta_i|^{q-2} (x_i)^2}{\left\| |\Delta|^{q-2} \right\|_1} = \frac{\sum_{i=1}^d |\Delta_i|^{q-2} (x_i)^2}{\sum_{i=1}^d |\Delta_i|^{q-2}}. \quad (\text{A.1})$$

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

Next, by applying Taylor's theorem, we have

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

We focus on the last term, which is

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

due to the inequality deduced earlier in (A.1). As the denominator of the right hand side in the above expression is bounded by,

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

we obtain that,

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

Then, we only need to bound the integral

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

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

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

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

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

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,

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

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,

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

where

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

Moreover, we have,

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

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

$$d \sum_{i=1}^d |\xi_i|^q \geq \left(\sum_{i=1}^d |\xi_i|^{q-2} \right) \left(\sum_{i=1}^d |\xi_i|^2 \right). \quad (\text{A.3})$$

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

Proof of Lemma 1. It follows from [10, Theorem 1] that

$$\Psi_n(\beta) = \inf_{\lambda \geq 0} \left\{ n^{\gamma/2} \lambda \delta_n + \mathbb{E}_{P_n} [\phi_\lambda(X; \beta)] \right\}, \quad (\text{A.4})$$

where

$$\phi_\lambda(X; \beta) = \sup_{\Delta} \left\{ \ell \left(X + n^{-\gamma/2} \Delta; \beta \right) - \lambda n^{-\gamma/2} \|\Delta\|_q^2 \right\}. \quad (\text{A.5})$$

For any $\varepsilon > 0$, there exists n_0 sufficiently large such that for all $n \geq n_0$, we have as a consequence of Taylor expansion and uniform continuity of $D_{xx}\ell(\cdot; \beta)$ (see Assumption A2.c)) that,

$$\left| \ell(X + \Delta n^{-\gamma/2}; \beta) - \ell(X; \beta) - n^{-\gamma/2} D_x \ell(X; \beta)^T \Delta - n^{-\gamma} \Delta^T D_{xx} \ell(X; \beta) \Delta \right| \leq \varepsilon n^{-\gamma} \|\Delta\|_q^2$$

uniformly on X . Then, from (A.5),

$$\phi_\lambda(X; \beta) \geq \ell(X; \beta) + n^{-\gamma/2} \sup_{\Delta} \left\{ D_x \ell(X; \beta)^T \Delta + n^{-\gamma/2} \Delta^T D_{xx} \ell(X; \beta) \Delta / 2 - \left(\lambda + \varepsilon n^{-\gamma/2} \right) \|\Delta\|_q^2 \right\}.$$

Next, define,

$$\Delta^* = \frac{\|D_x \ell(X; \beta)\|_p^{1-p/q}}{2\lambda} \text{sign}(D_x \ell(X; \beta)) |D_x \ell(X; \beta)|^{p/q}. \quad (\text{A.6})$$

Then

$$\begin{aligned} \sup_{\Delta} \left\{ D_x \ell(X; \beta)^T \Delta + n^{-\gamma/2} \Delta^T D_{xx} \ell(X; \beta) \Delta / 2 - \left(\lambda + \varepsilon n^{-\gamma/2} \right) \|\Delta\|_q^2 \right\} \\ \geq D_x \ell(X; \beta)^T \Delta^* + n^{-\gamma/2} \Delta^{*T} D_{xx} \ell(X; \beta) \Delta^* / 2 - \left(\lambda + \varepsilon n^{-\gamma/2} \right) \|\Delta^*\|_q^2 \\ = \frac{1}{4\lambda} \|D_x \ell(X; \beta)\|_p^2 + n^{-\gamma/2} \Delta^{*T} D_{xx} \ell(X; \beta) \Delta^* / 2 - \varepsilon n^{-\gamma/2} \|\Delta^*\|_q^2. \end{aligned}$$

Observe that $\|\Delta^*\|_q = \|D_x \ell(X; \beta)\|_p / (2\lambda)$. Further, it follows from the definitions of b_n , Δ^* , $a_n(\beta)$ and the equivalence for representation for $\Psi_n(\beta)$ in (A.4) and (A.5) that,

$$\Psi_n(\beta) \geq \mathbb{E}_{P_n} [\ell(X; \beta)] + n^{-\gamma/2} \inf_{\lambda \geq 0} \left\{ \lambda b_n + \frac{1}{4\lambda} \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2 + \frac{a_n(\beta) - 2\varepsilon \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2}{8\lambda^2 n^{\gamma/2}} \right\}.$$

Since $a_n(\beta) > 2\varepsilon \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2$ for ε suitably small, we have that the infimum above is attained at a $\lambda \geq \lambda_0(n)$, where

$$\lambda_0(n) := \arg \min_{\lambda \geq 0} \left\{ \lambda b_n + \frac{1}{4\lambda} \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2 \right\} = \sqrt{\frac{\mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2}{4b_n}}.$$

As a result,

$$\Psi_n(\beta) \geq \mathbb{E}_{P_n} [\ell(X; \beta)] + n^{-\gamma/2} \inf_{\lambda \geq \lambda_0(n)} \left\{ \lambda b_n + \frac{1}{4\lambda} \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2 + \frac{a_n(\beta) - 2\varepsilon \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2}{8\lambda^2 n^{\gamma/2}} \right\}. \quad (\text{A.7})$$

For an upper bound, we proceed as follows: Given $\varepsilon > 0$, as in the lower bound, we obtain from Taylor's expansion that,

$$\phi_\lambda(X; \beta) \leq \ell(X; \beta) + n^{-\gamma/2} \sup_{\Delta} \left\{ D_x \ell(X; \beta)^T \Delta + n^{-\gamma/2} \Delta^T D_{xx} \ell(X; \beta) \Delta / 2 - \left(\lambda - \varepsilon n^{-\gamma/2} \right) \|\Delta\|_q^2 \right\},$$

for all $n \geq n_0$, where n_0 is suitably large. Changing variables from Δ to $\Delta = \Delta^* + \xi$, we obtain that $\phi_\lambda(X; \beta) - \ell(X; \beta)$ is bounded from above by,

$$n^{-\gamma/2} \left(D_x \ell(X; \beta)^T \Delta^* + n^{-\gamma/2} \Delta^{*T} D_{xx} \ell(X; \beta) \Delta^* / 2 - (\lambda - \varepsilon n^{-\gamma/2}) \|\Delta^*\|_q^2 \right) + n^{-\gamma/2} \sup_{\xi \in \mathbb{R}^d} Q_n(X; \xi),$$

where

$$\begin{aligned} Q_n(X; \xi) &:= D_x \ell(X; \beta)^T \xi + n^{-\gamma/2} \xi^T D_{xx} \ell(X; \beta) \xi / 2 \\ &\quad + n^{-\gamma/2} \xi^T D_{xx} \ell(X; \beta) \Delta^* - \left(\lambda - \varepsilon n^{-\gamma/2} \right) (\|\Delta^* + \xi\|_q^2 - \|\Delta^*\|_q^2). \end{aligned} \quad (\text{A.8})$$

As in the lower bound, it follows from the definition of Δ^* in (A.6) and $\tilde{\Delta}$ and $\varphi_n(\beta)$ in (27) that,

$$\begin{aligned} \mathbb{E}_{P_n} \left[D_x \ell(X; \beta)^T \Delta^* + n^{-\gamma/2} \Delta^{*T} D_{xx} \ell(X; \beta) \Delta^* / 2 - \left(\lambda - \varepsilon n^{-\gamma/2} \right) \|\Delta^*\|_q^2 \right] \\ = \frac{1}{4\lambda} \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2 + \frac{a_n(\beta) + 2\varepsilon \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2}{8\lambda^2 n^{\gamma/2}}. \end{aligned}$$

Combining these observations with those in (A.4) and (A.5), we obtain,

$$\begin{aligned} & \Psi_n(\beta) - \mathbb{E}_{P_n} [\ell(X; \beta)] \\ & \leq n^{-\gamma/2} \inf_{\lambda \geq 0} \left\{ \lambda b_n + \frac{1}{4\lambda} \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2 + \frac{a_n(\beta) + 2\varepsilon \mathbb{E}_{P_n} \|D_x \ell(X; \beta)\|_p^2}{8\lambda^2 n^{\gamma/2}} + \mathbb{E}_{P_n} \left[\sup_{\xi} Q_n(X; \xi) \right] \right\}. \end{aligned} \quad (\text{A.9})$$

With Δ^* defined as in (A.6), we have that the derivative of the function $\|\Delta\|_q^2$, when evaluated at $\Delta = \Delta^*$, equals $D_x \ell(X; \beta)/\lambda$. Let us continue the analysis separately in the following two cases:

Case 1: $q > 2$. It follows from Lemma 5b that,

$$\|\Delta^* + \xi\|_q^2 - \|\Delta^*\|_q^2 \geq \frac{1}{\lambda} \xi^T D_x \ell(X; \beta) + C_q \min \left\{ \|\xi\|_2^2, \frac{\|\xi\|_q^q}{\|\Delta^*\|_{q-2}^{q-2}} \right\}.$$

where C_q is a suitably chosen constant that depends only on q and d . Furthermore, since $\|D_{xx} \ell(\cdot; \beta)\| \leq M(\beta)$ and $M(\beta)$ is continuous (see Assumption A2.c), there exists a large constant $\overline{M}$ such that $\|D_{xx} \ell(\cdot; \beta)\| \leq \overline{M}$ uniformly over β in a compact set.

Then it follows from the definition of $Q_n(X; \xi)$ in (A.8) that,

$$\begin{aligned} Q_n(X; \xi) & \leq n^{-\gamma/2} \left(\frac{1}{2} \xi^T D_{xx} \ell(X; \beta) \xi + \xi^T \left(D_{xx} \ell(X; \beta) \Delta^* + \frac{\varepsilon}{\lambda} D_x \ell(X; \beta) \right) \right) \\ & \quad - \left(\lambda - \varepsilon n^{-\gamma/2} \right) C_q \min \left\{ \|\xi\|_2^2, \frac{\|\xi\|_q^q}{\|\Delta^*\|_{q-2}^{q-2}} \right\} \\ & \leq \max \left\{ Q_n^{(1)}(X; \xi), Q_n^{(2)}(X; \xi) \right\}, \end{aligned}$$

where

$$Q_n^{(1)}(X; \xi) := n^{-\gamma/2} \xi^T \left(D_{xx} \ell(X; \beta) \Delta^* + \frac{\varepsilon}{\lambda} D_x \ell(X; \beta) \right) - \left(\lambda - \frac{\varepsilon + \overline{M}/(2C_q)}{n^{\gamma/2}} \right) C_q \|\xi\|_2^2$$

and

$$\begin{aligned} Q_n^{(2)}(X; \xi) & := n^{-\gamma/2} \xi^T \left(D_{xx} \ell(X; \beta) \Delta^* + \frac{\varepsilon}{\lambda} D_x \ell(X; \beta) \right) + \frac{\overline{M} c_q 2}{n^{\gamma/2}} \|\xi\|_q^2 \\ & \quad - \left(\lambda - \varepsilon n^{-\gamma/2} \right) \frac{C_q}{\|\Delta^*\|_{q-2}^{q-2}} \|\xi\|_q^q, \end{aligned}$$

where $c_q \in [1, \infty)$ is such that $\|x\|_2^2 \leq \|x\|_q^2$ for all $x \in \mathbb{R}^d$.

Next, since $\Delta^* = \tilde{\Delta}/(2\lambda)$, for any a, b such that $1/a + 1/b = 1$, we have,

$$\left| \xi^T \left(D_{xx} \ell(X; \beta) \Delta^* + \frac{\varepsilon}{\lambda} D_x \ell(X; \beta) \right) \right| \leq \frac{\|\xi\|_a}{2\lambda} \left(\|D_{xx} \ell(X; \beta) \tilde{\Delta}\|_b + 2\varepsilon \|D_x \ell(X; \beta)\|_b \right).$$

As a result,

$$\sup_{\xi} Q_n^{(1)}(X; \xi) \leq \frac{\left(\|D_{xx} \ell(X; \beta) \tilde{\Delta}\|_2 + 2\varepsilon \|D_x \ell(X; \beta)\|_2 \right)^2}{16C_q (\lambda n^{\gamma/2})^2 (\lambda - a_1 n^{-\gamma/2})^+}, \quad (\text{A.10})$$

where $a_1 := \varepsilon + \overline{M}/(2C_q)$. Likewise,

$$\sup_{\xi} Q_n^{(2)}(X; \xi) \leq \frac{b_1}{(\lambda n^{\gamma/2})^{q/(q-1)}} \left(\|D_{xx}\ell(X; \beta)\tilde{\Delta}\|_2 + 2\varepsilon \|D_x\ell(X; \beta)\|_2 \right)^{q/(q-1)} + \frac{b_2}{(n^{\gamma/2})^{q/(q-1)}}, \quad (\text{A.11})$$

for some suitable positive constants b_1, b_2 that does not depend on n and X . This follows from the observation that the maximizer of the function $g(x) := an^{-\gamma/2}x + bn^{-\gamma/2}x^2 - cx^q$ is $O((an^{-\gamma/2})^{1/(q-1)})$, when $q > 2$. Furthermore, observe from Assumption A2 that

$$\|D_{xx}\ell(X; \beta)\tilde{\Delta}\|_2^2 \leq \|D_{xx}\ell(X; \beta)\|_2^2 \|\tilde{\Delta}\|_2^2 \leq c_p^{p/q} M^2 \|D_x\ell(X; \beta)\|_p^2,$$

where $c_p \in [1, \infty)$ is such that $\|x\|_2^2 \leq \|x\|_p^2$ for all $x \in \mathbb{R}^d$. Since $\|D_{xx}\ell(x; \beta)\|$ is bounded in x , we have from the assumption $\mathbb{E}\|X\|_2^2 < \infty$ that

$$\mathbb{E}\|D_x\ell(X; \beta)\|_p^2 < \infty.$$

Since $q/(q-1) > 1$, as a consequence of dominated convergence, it follows from (A.10) and (A.11) that

$$\mathbb{E}_{P_n} \left[\sup_{\xi} Q_n(X; \xi) \right] = \mathbb{E}_{P_n} \left[\sup_{\xi} \max \left\{ Q_n^{(1)}(X; \xi), Q_n^{(2)}(X; \xi) \right\} \right] = o(n^{-\gamma/2}),$$

uniformly for all $\lambda \geq \lambda_0$.

Case 2: $q \in (1, 2]$. It follows from Lemma 5a that,

$$\|\Delta^* + \xi\|_q^2 - \|\Delta^*\|_q^2 \geq \frac{1}{\lambda} \xi^T D_x\ell(X; \beta) + (q-1) \|\xi\|_q^2.$$

Then we have,

$$Q_n(X; \xi) \leq n^{-\gamma/2} \left(\frac{1}{2} \xi^T D_{xx}\ell(X; \beta) \xi + \xi^T \left(D_{xx}\ell(X; \beta) \Delta^* + \frac{\varepsilon}{\lambda} D_x\ell(X; \beta) \right) \right) - (\lambda - \varepsilon n^{-\gamma/2}) (q-1) \|\xi\|_q^2.$$

Repeating the steps to arrive at (A.10) that we have that $Q_n(X; \xi) = O(n^{-\gamma})$, uniformly for $\lambda \geq \lambda_0(n)$.

Combining the observation that $Q_n(X; \xi) = o(n^{-1/2})$ from the above two cases, we have from (A.9) that $\Psi_n(\beta) - \mathbb{E}_{P_n} [\ell(X; \beta)]$ is upper bounded by,

$$n^{-\gamma/2} \inf_{\lambda \geq \lambda_0(n)} \left\{ \lambda b_n + \frac{1}{4\lambda} \mathbb{E}_{P_n} \|D_x\ell(X; \beta)\|_p^2 + \frac{a_n(\beta) + 2\varepsilon \mathbb{E}_{P_n} \|D_x\ell(X; \beta)\|_p^2}{8\lambda^2 n^{\gamma/2}} \right\} + o(n^{-\gamma}),$$

as $n \rightarrow \infty$. □

Proof of Lemma 2. For a fixed u, Δ , recall the definitions of $I(X_i, \Delta, u)$, $I_1(X_i, \Delta, u)$, $I_2(X_i, \Delta, u)$ and Δ'_i from (32)-(33).

$$M_n(\xi, u) = \frac{1}{n} \sum_{i=1}^n \left(\xi^T D_{\beta} h(X_i, \beta_*) u + \max_{\Delta} \left\{ \xi^T D_x h(X_i, \beta_*) \Delta + \xi^T I(X_i, \Delta, u) - \|\Delta\|_q^2 \right\} \right). \quad (\text{A.12})$$

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

$$\begin{aligned} & \max_{\Delta} \left\{ \xi^T D_x h(X_i, \beta_*) \Delta + \xi^T I(X_i, \Delta, u) - \|\Delta\|_q^2 \right\} \\ & \geq \max_{c > 0} \left\{ c \|D_x h(X_i, \beta_*)^T \xi\|_p \|\Delta'_i\|_q - c^2 \|\Delta'_i\|_q + \xi^T I(X_i, c\Delta'_i, u), 0 \right\}. \end{aligned}$$

As a consequence of Hölder's inequality, $|\xi^T I(X_i, c\Delta'_i, u)|$ is bounded from above by,

$$\begin{aligned} c\|\xi\|_p \int_0^1 & \left\| \left(D_x h \left(X_i + cn^{-1/2} \Delta'_i, \beta_* + t \frac{u}{\sqrt{n}} \right) - D_x h(X_i, \beta_*) \right) \Delta'_i \right\|_q dt \\ & + \|\xi\|_p \int_0^1 \left\| \left(D_\beta h \left(X_i + cn^{-1/2} \Delta'_i, \beta_* + t \frac{u}{\sqrt{n}} \right) - D_\beta h(X_i, \beta_*) \right) u \right\|_q dt. \end{aligned}$$

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

$$\lim_{n \rightarrow \infty} |\xi^T I(X_i, c\Delta'_i, u)| \mathbb{I}(X_i \in C_0) = 0,$$

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_0 such that for all $n \geq n_0$,

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

As a result, we obtain from (A.12) that, $M_n(\xi, u)$ is bounded from below by,

$$\left[\frac{1}{n} \xi^T \sum_{i=1}^n D_\beta h(X_i, \beta_*) u + \frac{1}{n} \sum_{i=1}^n \max_c \left\{ c \|D_x h(X_i, \beta_*)^T \xi\|_p \|\Delta'_i\|_q - c^2 \|\Delta'_i\|_q - c\varepsilon' \|\xi\|_p, 0 \right\} \right] \mathbb{I}(X_i \in C_0). \quad (\text{A.13})$$

As in the proof of Lemma 2 in [8] and $\mathbb{E} [D_x h(X, \beta_*) D_x h(X, \beta_*)^T] \succ \mathbf{0}$ (see Assumption A2.b)), we have that the second term in the lower bound grows at the rate $\|\xi\|_p^2$ for a suitably large c_0 , uniformly over compact sets of u , as $n, \|\xi\| \rightarrow \infty$. As a result, there exists positive constants n_0, c_1 such that

$$\frac{1}{n} \sum_{i=1}^n \max_c \left\{ c \|D_x h(X_i, \beta_*)^T \xi\|_p \|\Delta'_i\|_q - c^2 \|\Delta'_i\|_q - c\varepsilon' \|\xi\|_p, 0 \right\} \mathbb{I}(X_i \in C_0) \geq c_1 \|\xi\|_p^2, \quad (\text{A.14})$$

for all $n > n_0$, with probability at least $1 - \varepsilon/2$. Since function $\xi^T a - c_1 \|\xi\|_p^2 \rightarrow -\infty$, uniformly in compact sets of a , as $\|\xi\|_p \rightarrow \infty$, there indeed exist positive constants b, b' such that

$$\mathbf{P} \left(\|H_n\|_q + \left\| \frac{1}{n} \sum_{i=1}^n D_\beta h(X_i, \beta_*) u \right\| \leq b' \right) > 1 - \varepsilon/2 \quad (\text{A.15})$$

and $\|\xi\|_p b' - c_1 \|\xi\|^2 < 0$ for all $\|\xi\|_p \geq b, \|u\|_2 \leq K$. Therefore, for every $n > n_0, \|u\|_2 \leq K$, we have from the lower bound in (A.13) that

$$\max_{\|\xi\|_p \geq b} \{-\xi^T H_n - M_n(\xi, u)\} \leq 0,$$

whenever the events (A.14) and (A.15) happen. Since these events happen with probability at least $1 - \varepsilon$, we have the statement of Lemma 2 as a consequence of the union bound. $\square$

Proof of Lemma 3. Lemma 3 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 3 is similar to the proof of Lemma 3 in [8]. $\square$

Proof of Lemma 4. For $i = 1, \dots, n$ and $j = 1, 2$, let Δ_{ij} attain the inner supremum in

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

Then

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

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

$$\begin{aligned} & |M_n(\xi, u_1) - M_n(\xi, u_2)| \\ & \leq \frac{1}{n} \sum_{i=1}^n \max_{j=1,2} \left\{ \sqrt{n} \left| \xi^T \left(h(X_i + n^{-1/2} \Delta_{ij}, \beta_* + n^{-1/2} u_1) - h(X_i + n^{-1/2} \Delta_{ij}, \beta_* + n^{-1/2} u_2) \right) \right| \right\}. \end{aligned} \quad (\text{A.16})$$

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

$$\begin{aligned} & \sqrt{n} \xi^T \left| h(X_i + n^{-1/2} \Delta_{ij}, \beta_* + n^{-1/2} u_1) - h(X_i + n^{-1/2} \Delta_{ij}, \beta_* + n^{-1/2} u_2) \right| \\ & = \left| \int_0^1 \xi^T D_{\beta} h \left(X_i + n^{-1/2} \Delta_{ij}, \beta_* + n^{-1/2} (u_1 + (u_2 - u_1)t) \right) (u_2 - u_1) dt \right| \\ & \leq \|u_1 - u_2\|_q \|\xi\|_p \int_0^1 \left\| D_{\beta} h \left(X_i + n^{-1/2} \Delta_{ij}, \beta_* + n^{-1/2} (u_1 + (u_2 - u_1)t) \right) \right\|_q dt \\ & \leq \|u_1 - u_2\|_q \|\xi\|_p \left(\|D_{\beta} h(X_i, \beta_*)\|_q + \bar{\kappa}(X_i) \left(n^{-1/2} \|\Delta_{ij}\| + 2c_q n^{-1/2} K \right) \right), \end{aligned} \quad (\text{A.17})$$

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 (40) that there exists n_0 such that $\Delta_{ij} \leq \nu \sqrt{n}$, for all $i \leq n, n \geq n_0$, $\|\xi\|_p \leq b$, $\|u\|_2 \leq K$. Combining this observation with those in (A.16) and (A.17), we obtain that

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

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

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

With these observations, if we take $L := 4b(\mathbb{E} \|D_{\beta} h(X, \beta_*)\|_q + \mathbb{E}[\bar{\kappa}(X)](\nu + 2c_q K))$, 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 $1 - \varepsilon$. □