SUPPLEMENTARY MATERIAL TO THE PAPER,

"Robust Wasserstein Profile Inference And Applications to Machine Learning"

JOSE BLANCHET, YANG KANG, AND KARTHYEK MURTHY

This supplementary material to the paper "Robust Wasserstein Profile Inference and Applications to Machine Learning" is organized as follows: Proofs of all the main results in the paper are furnished in Section A. As some of the main results in our paper utilize strong duality for problems of moments, a quick introduction to problem of moments along with a well-known strong duality result that is useful in our context is provided in Section B. A technical result on exchange of sup and inf in the DRO formulation (8) is presented in Section C. Relevant bibliography utilized in this supplementary material is available at the end of this supplementary material.

Appendix A. Proofs of main results

This section, comprising the proofs of the main results, is organized as follows: Subsection A.1 is devoted to derive the results on distributionally robust representations presented in Section 2.4. The proofs of results on coverage properties are presented in Section A.2. Subsection A.3 contains the proofs of stochastic upper and lower bounds (and hence weak limits) presented in Section 3.3. Subsection A.4 contains the proofs of Theorems 5 and 6 as applications of the stochastic upper and lower bounds presented in Section 3.3. Some of the useful technical results that are not central to the argument are presented in Sections B and C.

A.1. **Proofs of the distributionally robust representations in Section 2.4.**

Here we provide proofs for results in Sections 2.3, 2.4 that recover various norm regularized regressions as a special cases of distributionally robust regression (Proposition 2, Theorems 1 and 2).

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

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

(A1) Management Science and Engineering, Stanford University
(A2) Department of Statistics, Columbia University
(A3) Engineering Systems & Design, Singapore University of Technology & Design
E-mail addresses: jose.blanchet@stanford.edu, yangkang@stat.columbia.edu, karthyek_murthy@sutd.edu.sg.

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

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

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

$$\begin{aligned}\phi_\gamma(\bar{X}_i;\beta) &= \left(\bar{\beta}^T\bar{X}_i\right)^2 + \sup_{\Delta\in\mathbb{R}^{d+1}} \left\{ -\left(\gamma - \left\|\bar{\beta}\right\|_p^2\right)\|\Delta\|_q^2 + 2\left|\bar{\beta}^T\bar{X}_i\right|\left\|\bar{\beta}\right\|_p \|\Delta\|_q \right\} \\ &= \begin{cases} \left(\bar{\beta}^T\bar{X}_i\right)^2 \gamma/(\gamma - \left\|\bar{\beta}\right\|_p^2) & \text{if } \gamma > \left\|\bar{\beta}\right\|_p^2, \\ +\infty & \text{if } \gamma \leq \left\|\bar{\beta}\right\|_p^2. \end{cases} \end{aligned} \tag{1}$$

With this expression for $\phi_\gamma(X_i, Y_i; \beta)$, we next investigate the right hand side of the duality relation in Proposition 1. As $\phi_\gamma(x, y; \beta) = \infty$ when $\gamma \leq \|\beta\|_p^2$, we obtain from the dual formulation in Proposition 1 that

$$\begin{aligned}\sup_{\mathbb{P}:\mathcal{D}_c(\mathbb{P},\mathbb{P}_n)\leq\delta} \mathbb{E}_{\mathbb{P}}\left[l(X,Y;\beta)\right] &= \inf_{\gamma\geq 0}\left\{\gamma\delta + \frac{1}{n}\sum_{i=1}^n \phi_\gamma(X_i,Y_i;\beta)\right\} \\ &= \inf_{\gamma>\|\beta\|_p^2}\left\{\gamma\delta + \frac{\gamma}{\gamma - \|\bar{\beta}\|_p^2}\frac{1}{n}\sum_{i=1}^n (\bar{\beta}^T\bar{X}_i)^2\right\}. \end{aligned} \tag{2}$$

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

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

This completes the proof of Proposition 2. □

Outline of a proof of Theorem 1. The proof of Theorem 1 is essentially the same as the proof of Proposition 2, except for adjusting for ∞ in the definition of cost function $N_q((x,y),(u,v))$ when $y \neq v$ (as in the derivation leading to $\phi_\gamma(X_i, Y_i; \beta)$ defined in (11)). First, see that

$$\phi_\gamma(X_i,Y_i;\beta) = \sup_{x'\in\mathbb{R}^d, y'\in\mathbb{R}} \left\{(y'^T x'^2 - \gamma N_q\big((x',y'),(X_i,Y_i)\big)\right\}.$$

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

$$\begin{aligned}\phi_\gamma(X_i,Y_i;\beta) &= \sup_{x'\in\mathbb{R}^d} \left\{(Y_i - \beta^T x'^2 - \gamma N_q\big((x',Y_i),(X_i,Y_i)\big)\right\}. \\ &= \sup_{x'\in\mathbb{R}^d} \left\{(Y_i - \beta^T x'^2 - \gamma\|x' - X_i\|_q^2\big)\right\}.\end{aligned}$$

Now, following same lines of reasoning as in the proof of Theorem 2 and the derivation leading to (1), we obtain

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

The rest of the proof is same as in the proof of Proposition 2.

Proof of Theorem 2. As in the proof of Proposition 2, we apply the duality formulation in Proposition 1 to write the worst case expected log-exponential loss function as:

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

For each (X_i, Y_i), following Lemma 1 in [5], we obtain

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

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

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

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

For SVM with hinge loss function, let us apply the duality formulation in Proposition 1 to write the worst case expected Hinge loss function as:

$$\sup_{\mathbb{P}:\ D_c(\mathbb{P},\mathbb{P}_n)\le\delta} \mathbb{E}_{\mathbb{P}}\big[\ \left(1 - Y\beta^T X\right)^+\ \big] = \inf_{\lambda\ge 0}\left\{\delta\lambda + \frac{1}{n}\sum_{i=1}^n \sup_x \left\{\left(1 - Y_i\beta^T x\right)^+ - \lambda\left\|x - X_i\right\|_p\right\}\right\}.$$

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

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

The first equality follows from the observation that $x^+ = \sup_{0 \le \alpha \le 1} x$; second equality is because the function is concave in Δ_i, linear in α; as α is in a compact set, we can apply minimax theorem to switch the order of maxima; third equality is due to applying Hölder inequality to the first term, and since the second term only depends on the norm of Δ_i, the equality holds for this maximization problem. For the outer minimization, it is sufficient to restrict to $\lambda \ge \left\| \beta \right\|_q$. As a result, we obtain

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

This completes the proof. □

A.2. **Proofs of results on coverage properties.**

Proof of Proposition 6. Let $\hat{\mathbb{P}}$ be a probability measure from the set,

$$\{\mathbb{P} : D_c(\mathbb{P}, \mathbb{P}_n) \le \delta,\ \mathbb{E}_{\mathbb{P}}\big[D_\beta l(X, Y; \beta_*)\big] = \mathbf{0}\},$$

which is non-empty, because $\delta > R_n(\beta_*)$. Then,

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{P}: D_c(\mathbb{P}, \mathbb{P}_n) \le \delta} \mathbb{E}_{\mathbb{P}} \left[l(X, Y; \beta) \right] \ge \inf_{\beta \in \mathbb{R}^d} \mathbb{E}_{\hat{\mathbb{P}}} \left[l(X, Y; \beta) \right] = \mathbb{E}_{\hat{\mathbb{P}}} \left[l(X, Y; \beta_*) \right].$$

Moreover, since $D_c(\cdot)$ is symmetric in its arguments, we have $D_c(\hat{\mathbb{P}}, \mathbb{P}_n) \le \delta$. As a result,

$$\mathbb{E}_{\mathbb{P}_n} \left[l(X, Y; \beta_*) \right] - \inf_{\beta} \sup_{\mathbb{P} \in \mathcal{U}_\delta(\mathbb{P}_n)} \mathbb{E}_{\mathbb{P}} \left[l(X, Y; \beta) \right] \le \sup_{\mathbb{P}: D_c(\hat{\mathbb{P}}, \mathbb{P}) \le \delta} \mathbb{E}_{\mathbb{P}} \left[l(X, Y; \beta_*) \right] - \mathbb{E}_{\hat{\mathbb{P}}} \left[l(X, Y; \beta_*) \right]. \tag{3}$$

On the other hand,

$$\inf_{\beta} \sup_{\mathbb{P} \in \mathcal{U}_\delta(\mathbb{P}_n)} \mathbb{E}_{\mathbb{P}} \left[l(X, Y; \beta) \right] - \mathbb{E}_{\mathbb{P}_n} \left[l(X, Y; \beta_*) \right] \le \sup_{\mathbb{P}: D_c(\mathbb{P}_n, \mathbb{P}) \le \delta} \mathbb{E}_{\mathbb{P}} \left[l(X, Y; \beta_*) \right] - \mathbb{E}_{\mathbb{P}_n} \left[l(X, Y; \beta_*) \right],$$

which can be bounded from above to result in the desired bound, $C_1 \delta + C_2(n) \mathbf{1}_{\rho=2} \sqrt{\delta}$, by substituting the regularized regression estimators derived in Theorem 1 (when $\rho = 2$) and Theorem 2 (when $\rho = 1$). Likewise, repeating the proofs of Theorems 1 and 2 for the case where the baseline distribution is set to be $\hat{\mathbb{P}}$ (instead of $\mathbb{P}_n$), we obtain for any $\beta \in \mathbb{R}^d$ that

$$\sup_{\mathbb{P}: D_c(\hat{\mathbb{P}}, \mathbb{P}) \le \delta} \mathbb{E}_{\mathbb{P}} \left[l(X, Y; \beta) \right] - \mathbb{E}_{\hat{\mathbb{P}}} \left[l(X, Y; \beta) \right] = \delta \|\beta\|_p,$$

for the logistic regression example in Theorem 2; and

$$\begin{aligned}\sup_{\mathbb{P}:D_c(\hat{\mathbb{P}},\mathbb{P})\leq\delta} \mathbb{E}_{\mathbb{P}}\left[l(X,Y;\beta)\right] - \mathbb{E}_{\hat{\mathbb{P}}}\left[l(X,Y;\beta)\right] &= 2\sqrt{\delta}\|\beta\|_p\sqrt{\mathbb{E}_{\hat{\mathbb{P}}}\left[(Y-\beta^T X)^2\right]} + \delta\|\beta\|_p^2 \\ &\leq 2\sqrt{\delta}\|\beta\|_p\sqrt{\sup_{\mathbb{P}\in\mathcal{U}_\delta(\mathbb{P}_n)}\mathbb{E}_{\mathbb{P}}\left[(Y-\beta^T X)^2\right]} + \delta\|\beta\|_p^2, \\ &= 2\sqrt{\delta}\|\beta\|_p\sqrt{\mathbb{E}_{\mathbb{P}_n}\left[(Y-\beta^T X)^2\right]} + 3\delta\|\beta\|_p^2,\end{aligned}$$

for the linear regression example in Theorem 1. This verifies the upper bound for (3). □

Proof of Theorem 4. Since $\delta = n^{-\rho/2}\eta$ for some $\eta \geq \eta_\alpha$, we have from the definition of η_α that,

$$\lim_{n\to\infty}\mathbb{P}(R_n(\beta_*) > \delta) = \lim_{n\to\infty}\mathbb{P}(n^{\rho/2}R_n(\beta_*) > \eta) \leq \alpha,$$

as $n\to\infty$. Then it follows from Proposition 6 that,

$$\left|\mathbb{E}_{\mathbb{P}_n}[l(X,Y;\beta_*)] - \inf_{\beta\in\mathbb{R}^d}\sup_{\mathbb{P}\in\mathcal{U}_\delta(\mathbb{P}_n)}\mathbb{E}_{\mathbb{P}}\left[l(X,Y;\beta)\right]\right| \leq C_1\eta n^{-\rho/2} + C_2(n)\sqrt{\eta}\mathbf{1}_{\{\rho=2\}}n^{-\rho/4},$$

with probability greater than or equal to $1-\alpha$, as $n\to\infty$. Moreover, due to Chebyshev's inequality, we obtain,

$$|\mathbb{E}_{\mathbb{P}_n}[l(X,Y;\beta_*)] - \mathbb{E}_{\mathbb{P}_*}[l(X,Y;\beta_*)]| \leq \sqrt{\frac{\mathrm{Var}_{\mathbb{P}_*}[l(X,Y;\beta_*)]}{\alpha n}},$$

and subsequently, $C_2(n)/(2\|\beta_*\|_p) \leq \sqrt{\mathbb{E}_{\mathbb{P}_*}[l(X,Y;\beta_*)]} + (\alpha^{-1}n^{-1}\mathrm{Var}_{\mathbb{P}_*}[l(X,Y;\beta_*)])^{1/4}$, with probability exceeding $1-\alpha$. Since $\mathbb{E}_{\mathbb{P}_*}[l(X,Y;\beta_*)] = \inf_\beta \mathbb{E}_{\mathbb{P}_*}[l(X,Y;\beta)]$, the desired convergence in the statement of Theorem 4 follows from triangle inequality and an application of union bound to the above two inequalities. □

A.3. **Proofs of asymptotic stochastic upper and lower bounds of RWP function in Section 3.3.** We first use Proposition 3 to derive a dual formulation for $n^{\rho/2}R_n(\theta_*)$ which will be the starting point of our analysis. Due to Assumption A2), $\mathbb{E}[h(W,\theta_*)] = \mathbf{0}$. Combining this observation with the positive definiteness in Assumption A4), we have that $\mathbf{0}$ lies in the interior of convex hull of $\{h(u,\theta_*) : u\in\mathbb{R}^m\}$ by using a supporting hyperplane argument as in the proof of [1, Proposition 8]. Then, due to Proposition 3,

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

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

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

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

Due to the fundamental theorem of calculus (using Assumption A3)), we have that

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

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

$$n^{\rho/2} R_n(\theta_*) = \sup_{\zeta} \left\{ -\zeta^T H_n - M_n(\zeta) \right\}, \tag{4}$$

where

$$M_n(\zeta) = \frac{1}{n} \sum_{i=1}^{n} \sup_{\Delta} \left\{ \zeta^T \int_0^1 Dh\left(W_i + n^{-1/2} \Delta u \right) \Delta du - \|\Delta\|_q^\rho \right\}. \tag{5}$$

The reformulation in (4) is our starting point of the analysis.

To proceed further, we first state a result which will allow us to apply a localization argument in the representation of $n^{\rho/2} R_n(\theta_*)$ in (4). Recall the definition of M_n above in (5) and that $H_n = n^{-1/2} \sum_{i=1}^n h(W_i)$.

Lemma 1. *Suppose that the Assumptions A2) to A4) are in force. Then, for every $\varepsilon > 0$, there exists $n_0 > 0$ and $b \in (0, \infty)$ such that*

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

for all $n \geq n_0$.

Proof of Lemma 1. Recall that $q > 1$ and $p = q/(q-1)$. For $\zeta \neq 0$, we write $\bar{\zeta} = \zeta / \|\zeta\|_p$. Let us define the vector $V_i\left(\bar{\zeta}\right) = Dh\left(W_i\right)^T \bar{\zeta}$, and put

$$\Delta_i' = \Delta_i'\left(\bar{\zeta}\right) = \left| V_i\left(\bar{\zeta}\right) \right|^{p/q} sgn\left(V_i\left(\bar{\zeta}\right) \right). \tag{6}$$

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

$$\begin{aligned}
&\sup_{\Delta} \left\{ \zeta^T \int_0^1 Dh(W_i + n^{-1/2}\Delta u)\Delta du - \|\Delta\|_q^\rho \right\} \\
&\quad = \sup_{\Delta} \left\{ \zeta^T Dh(W_i)\Delta - \|\Delta\|_q^\rho + \zeta^T \int_0^1 \left[Dh(W_i + n^{-1/2}\Delta u) - Dh(W_i) \right] \Delta du \right\} \\
&\quad \geq \max \Bigg\{ c \left\| \zeta^T Dh(W_i) \right\|_p \left\| \Delta_i' \right\|_q - c^\rho \left\| \Delta_i' \right\|_q^\rho \\
&\qquad\qquad + c\zeta^T \int_0^1 \left[Dh(W_i + cn^{-1/2}\Delta_i' u) - Dh\left(W_i\right) \right] \Delta_i' du, 0 \Bigg\} I\left(W_i \in C_0\right).
\end{aligned} \tag{7}$$

Due to Hölder's inequality,

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

Because of continuity $Dh(\cdot)$ and the fact that $W_i \in C_0$ (so the integrand is bounded), we have that the previous expression converges to zero as $n \to \infty$. Therefore, for given positive constants ε', c (note than convergence is uniform on $W_i \in C_0$), there exists n_0 such that for all $n \geq n_0$

$$cI\left(W_i \in C_0\right) \left| \zeta^T \int_0^1 \left[Dh(W_i + cn^{-1/2}\Delta_i' u) - Dh(W_i) \right] \Delta_i' du \right| \leq c\varepsilon' \|\zeta\|_p. \tag{8}$$

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

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

Consequently, it follows from (7) and (8) that

$$M_n(\zeta) \geq \frac{1}{n}\sum_{i=1}^n \left\{ c \|\zeta\|_p \|\bar{\zeta}^T Dh(W_i)\|_p^p - c^\rho \|\bar{\zeta}^T Dh(W_i)\|_p^{\rho\frac{p}{q}} - c\varepsilon' \|\zeta\|_p \right\} I\left(W_i \in C_0\right). \tag{9}$$

Now, since the map $\bar{\zeta} \hookrightarrow \left\|\bar{\zeta}^T Dh(W_i)\right\|_p^p$ is Lipschitz continuous on $\left\|\bar{\zeta}\right\|_p = 1$, we conclude that,

$$\frac{1}{n}\sum_{i=1}^n \left\|\bar{\zeta}^T Dh(W_i)\right\|_p^p I\left(W_i \in C_0\right) \to \mathbb{E}\left[\left\|\bar{\zeta}^T Dh\left(W\right)\right\|_p^p I\left(W \in C_0\right)\right], \tag{10}$$

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

$$\inf_{\bar{\zeta}} \mathbb{E}\left\|\bar{\zeta}^T Dh\left(W\right)\right\|_p^p > \delta. \tag{11}$$

Consecutively, by selecting $c_0 > 0$ large enough, we conclude from (10) that for $n \geq N'\left(\delta\right)$,

$$\frac{1}{n}\sum_{i=1}^n \left\|\bar{\zeta}^T Dh(W_i)\right\|_p^p I\left(W_i \in C_0\right) > \frac{\delta}{2}. \tag{12}$$

Further, if we let $c_1 := \sup_{w \in C_0} \|\bar{\zeta}^T Dh(w)\|_p^{p/q} < \infty$, then

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

for all $n > N'(\delta)$. As a consequence, if $n \geq N'\left(\delta\right)$, it follows from (9) and (12) that

$$\begin{aligned}
\sup_{\|\zeta\|_p > b} \left\{-\zeta^T H_n - M_n\left(\zeta\right)\right\} &\leq \sup_{\|\zeta\|_p > b} \left\{ -\zeta^T H_n - \left(\frac{c\delta\|\zeta\|_p}{2} - (cc_1)^\rho - c\varepsilon'\|\zeta\|_p \right) \right\} \\
&\leq \sup_{\|\zeta\|_p > b} \left\{ -\zeta^T H_n - \|\zeta\|_p \left\{ c\left(\frac{\delta}{2} - \varepsilon'\right) - \frac{(cc_1)^\rho}{b} \right\} \right\}.
\end{aligned}$$

Consequently, on the set $\|H_n\|_q \leq b'$, we obtain

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

Now, if we take $c > 4(b'+1)/\delta$, $\varepsilon' = \delta/4$ and b to be large enough such that $b > (cc_1)^\rho$ then

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

Therefore, if $n \geq n_0$ (see (8)), then

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

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

Lemma 2. *For any $b > 0$ and $c_0 \in (0, \infty)$,*

$$\frac{1}{n}\sum_{i=1}^{n} \left\|\zeta^T Dh\left(W_i\right)\right\|_p^{\rho/(\rho-1)} I\left(\left\|W_i\right\|_p \leq c_0\right) \to \mathbb{E}\left[\left\|\zeta^T Dh\left(W\right)\right\|_p^{\rho/(\rho-1)} I(\left\|W\right\|_p \leq c_0)\right],$$

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

Proof of Lemma 2. We first argue a suitable Lipschitz property for the map $\zeta \hookrightarrow \left\|\zeta^T Dh\left(W_i\right)\right\|_p^{\rho/(\rho-1)}$. It is elementary that for any $0 \leq a_0 < a_1$ and $\gamma > 1$

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

Applying this observation with

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

and using that $\left\|\zeta^T Dh\left(W_i\right)\right\|_p \leq b\left\|Dh\left(W_i\right)\right\|_p$ for $\left\|\zeta\right\|_p \leq b$, we obtain

$$\left|\left\|\zeta_0^T Dh\left(W_i\right)\right\|_p^{\rho/(\rho-1)} - \left\|\zeta_1^T Dh\left(W_i\right)\right\|_p^{\rho/(\rho-1)}\right| \leq \frac{\rho}{\rho-1} b^{1/(\rho-1)} \left\|Dh\left(W_i\right)\right\|_p^{\rho/(\rho-1)} \left\|\zeta_0 - \zeta_1\right\|_p.$$

Consequently, we have that

$$\left|\frac{1}{n}\sum_{i=1}^{n} \left\|\zeta_0^T Dh\left(W_i\right)\right\|_p^{\frac{\rho}{\rho-1}} - \frac{1}{n}\sum_{i=1}^{n} \left\|\zeta_1^T Dh\left(W_i\right)\right\|_p^{\frac{\rho}{\rho-1}}\right| \leq \frac{\rho}{\rho-1}\left\|\zeta_0 - \zeta_1\right\|_p \frac{b^{\frac{1}{\rho-1}}}{n} \sum_{i=1}^{n} \left\|Dh\left(W_i\right)\right\|_p^{\frac{\rho}{\rho-1}}.$$

Since $Dh(\cdot)$ is continuous, $\mathbb{E}\left[\left\|Dh\left(W\right)\right\|_p^{\rho/(\rho-1)} I(\left\|W\right\|_p \leq c_0)\right] < \infty$, thus yielding the tightness of

$$\frac{1}{n}\sum_{i=1}^{n} \left\|\zeta^T Dh(W_i)\right\|_p^{\rho/(\rho-1)} I\left(\left\|W_i\right\|_p \leq c_0\right),$$

under the uniform topology on compact sets. The Strong Law of Large Numbers guarantees that finite dimensional distributions converge (for any choice of $\zeta_1, \ldots, \zeta_k, k \geq 1$), and, since the limit is deterministic, we obtain the desired convergence in probability. □

Proof of Theorem 3. Let us first observe that $R_n(\theta_*) \geq 0$ (choosing $\zeta = 0$). Then, as a consequence of Lemma 1, there exists $b > 0$ such that the event

$$\mathcal{A}_n = \left\{ n^{\rho/2} R_n(\theta_*) = \max_{\left\|\zeta\right\|_p \leq b} \left\{-\zeta^T H_n - M_n\left(\zeta\right)\right\} \right\}, \tag{13}$$

where the outer supremum is attained at some $\left\|\zeta_*\right\|_p \leq b$, occurs with probability at least $1 - \varepsilon$, as long as $n \geq n_0$. In other words, $\mathbb{P}(\mathcal{A}_n) \geq 1 - \varepsilon$ when $n \geq n_0$.

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

$$\Delta_i' = \Delta_i'\left(\bar{\zeta}\right) = \left|V_i\left(\bar{\zeta}\right)\right|^{p/q} sgn\left(V_i\left(\bar{\zeta}\right)\right). \tag{14}$$

Next, let $\bar{\Delta}_i = c_i \Delta_i'$ with c_i chosen so that

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

In such case we have that

$$\begin{aligned}\max_{\Delta}\left\{\zeta^T Dh\left(W_i\right)\Delta - \|\Delta\|_q^\rho\right\} &= \max_{\|\Delta\|_q \geq 0}\left\{\left\|\zeta^T Dh\left(W_i\right)\right\|_p \|\Delta\|_q - \|\Delta\|_q^\rho\right\} \\ &= \zeta^T Dh\left(W_i\right)\bar{\Delta}_i - \left\|\bar{\Delta}_i\right\|_q^\rho \\ &= \left\|\zeta^T Dh\left(W_i\right)\right\|_p^{\rho/(\rho-1)}\left(\frac{1}{\rho}\right)^{1/(\rho-1)}\left(1-\frac{1}{\rho}\right). \end{aligned} \tag{15}$$

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

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

where

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

Therefore

$$\max_{\|\zeta\|_p \leq b}\left\{-\zeta^T H_n - M_n(\zeta)\right\} \leq \max_{\|\zeta\|_p \leq b}\left\{-\zeta^T H_n - M_n'(\zeta, c_0)\right\}. \tag{16}$$

Define

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

where the equality follows from (15). We then claim that

$$\sup_{\|\zeta\|_q \leq b}\left|\widehat{M}_n(\zeta, c_0) - M_n'(\zeta, c_0)\right| \to 0. \tag{17}$$

In order to verify (17), note, using the continuity of $Dh(\cdot)$, that for any $\varepsilon' > 0$ there exists n_0 such that if $n \geq n_0$ then (uniformly over $\|\zeta\|_p \leq b$),

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

Therefore, if $n \geq n_0$,

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

Since $\varepsilon' > 0$ is arbitrary, (17) stands verified. Then, applying Lemma 2 we obtain

$$\widehat{M}_n(\zeta, c_0) \to \mathbb{E}\left(\zeta^T Dh\left(W_i\right)\bar{\Delta}_i du - \left\|\bar{\Delta}_i\right\|_q^\rho\right)^+ I\left(W_i \in C_0\right),$$

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

$$\max_{\|\zeta\|_p \leq b} \left\{-\zeta^T H_n - M_n'(\zeta, c_0)\right\}$$
$$\Rightarrow \max_{\|\zeta\|_p \leq b} \left\{-\zeta^T H - \kappa(\rho)\, \mathbb{E}\left[\left\|\zeta^T Dh(W)\right\|_p^{\rho/(\rho-1)} I\left(\|W\|_p \leq c_0\right)\right]\right\}, \tag{18}$$

as $n \to \infty$, where

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

and $H \sim \mathcal{N}(0, Cov[h(W, \theta_*)])$. From (16) and the construction of (13), we can easily obtain that $n^{\rho/2} R_n(\theta_*)$ is stochastically bounded (asymptotically) by

$$\max_{\zeta} \left\{-\zeta^T H - \kappa(\rho)\, \mathbb{E}\left[\left\|\zeta^T Dh(W)\right\|_p^{\rho/(\rho-1)}\right]\right\},$$

which verifies the first part of the theorem when $\rho > 1$.

Now, for $\rho = 1$, we will follow very similar steps. Again, due to Lemma 1 we concentrate on the region $\|\zeta\|_p \leq b$ for some $b > 0$. For the upper bound, define Δ_i' as in (14). Using a localization technique similar to that described in the proof of Lemma 1 in which the set C_0 as introduced we might assume that $\|W_i\|_p \leq c_0$ for some $c_0 > 0$. Then, for a given constant $c > 0$, setting $\Delta_i = c\Delta_i'$, we obtain that

$$\max_{\|\zeta\|_p \leq b} \left\{-\zeta^T H_n - \frac{1}{n}\sum_{i=1}^n \sup_{\Delta_i} \left\{\zeta^T \int_0^1 Dh(W_i + \Delta_i u/n^{1/2})\Delta_i du - \|\Delta_i\|_q\right\}\right\}$$
$$\leq \max_{\|\zeta\|_p \leq b} \left\{-\zeta^T H_n - \frac{1}{n}\sum_{i=1}^n \left(c\zeta^T \int_0^1 Dh(W_i + c\Delta_i' u/n^{1/2})\Delta_i' du - c\left\|\Delta_i'\right\|_q\right) I(W_i \in C_0)\right\}.$$

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

$$\frac{1}{n}\sum_{i=1}^n I(W_i \in C_0) \int_0^1 \zeta^T \left[Dh(W_i + c\Delta_i' u/n^{1/2}) - Dh(W_i)\right] \Delta_i' du \to 0$$

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

$$\max_{\|\zeta\|_p \leq b} \left\{-\zeta^T H_n - \frac{1}{n}\sum_{i=1}^n \left(c\zeta^T Dh(W_i)\Delta_i' du - c\left\|\Delta_i'\right\|_q\right) I(W_i \in C_0)\right\}$$
$$= \max_{\|\zeta\| \leq b} \left\{-\zeta^T H_n - \frac{1}{n}\sum_{i=1}^n c\left(\left\|\zeta^T Dh(W)\right\|_p - 1\right)^+ \|\Delta_i'\|_q I(\|W_i\|_p \leq c_0)\right\}$$
$$\Rightarrow \max_{\|\zeta\| \leq b} \left\{-\zeta^T H - c\mathbb{E}\left[\left(\left\|\zeta^T Dh(W)\right\|_p - 1\right)^+ \|\bar{\zeta}^T Dh(W)\|_p^{p/q} I(\|W\|_p \leq c_0)\right]\right\},$$

because $\|\Delta'\|_q^q = \|\bar{\zeta}^T Dh(W_i)\|_p^p$. Next, as the constant c can be arbitrarily large, we obtain a stochastic upper bound of the form

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

This completes the proof of Theorem 3. □

Proof of Proposition 4. We follow the notation introduced in the proof of Theorem 3. Recall from (4) and (5) that

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

Let $A := \{\zeta : \operatorname{esssup}\left\|\zeta^T Dh(w)\right\|_p \leq 1\}$, where the essential supremum is taken with respect to the Lebesgue measure. Then, due to Hölder's inequality, if $\zeta \in A$,

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

Consequently,

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

Letting $n \to \infty$ we conclude that

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

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

Finally, we provide the proof of Proposition 5.

Proof of Proposition 5. Recall from (4) and (5) that

$$n^{1/2}R_n(\theta_*) = \sup_{\zeta}\left\{\zeta^T H_n - \frac{1}{n}\sum_{k=1}^{n}\sup_{\Delta}\left\{\int_0^1 \zeta^T Dh\left(W_i + \Delta u/n^{1/2}\right)\Delta du - \|\Delta\|_q^\rho\right\}\right\}. \quad (19)$$

As in the proof of Theorem 3, due to Lemma 1, we might assume that $\|\zeta\|_p \leq b$ for some $b > 0$.

The strategy will be to split the inner supremum in values of $\|\Delta\|_q \leq \delta n^{1/2}$ and values $\|\Delta\|_q > \delta n^{1/2}$ for a suitably small positive constant δ. In Step 1, we shall show that the supremum is achieved with high probability in the former region. Then, in Step 2, we analyze the region in which $\|\Delta\|_q \leq \delta n^{1/2}$ and argue that the integrals inside the summation in (19) can be replaced by $\zeta^T Dh(W_i)\Delta$. Once this substitution is performed we can solve the inner maximization problem explicitly in Step 3 and, finally, we will apply a weak convergence result on ζ-compact sets to conclude the result. We now proceed to execute this strategy.

Execution of Step 1: Pick $\delta > 0$ small, to be chosen in the sequel, then note that A5) implies (by redefining κ if needed, due to the continuity of $Dh(\cdot)$) that

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

Therefore, for ζ such that $\|\zeta\|_p \leq b$,

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

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

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

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

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

yields

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

So, in both cases we can write

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

Next, as $\mathbb{E}\|W_n\|^\rho < \infty$, we have that for any $\varepsilon' > 0$,

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

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

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

Similarly, choosing n sufficiently large we can guarantee that

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

Therefore, we conclude that for any fixed $\delta > 0$,

$$\sup_{\|\Delta\|_q \geq \delta\sqrt{n}} \left\{ \int_0^1 \left| \zeta^T Dh(W_i + \Delta u/n^{1/2}) \Delta \right| du - \|\Delta\|_q^\rho \right\} \leq 0 \tag{20}$$

provided n is large enough, thus achieving the desired result over the region $\|\Delta\|_q \geq \delta\sqrt{n}$.

Execution of Step 2: Next, we let $\varepsilon'' > 0$, and note that

$$\begin{aligned}
&\sup_{\|\Delta\|_q \le \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T Dh(W_i + \Delta u/n^{1/2})\Delta du - \|\Delta\|_q^\rho \right\} \qquad (21)\\
&\le \sup_{\|\Delta\|_q \le \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T \left[Dh(W_i + \Delta u/n^{1/2}) - Dh(W_i) \right] \Delta du - \varepsilon'' \|\Delta\|_q^\rho \right\}\\
&\qquad + \sup_{\|\Delta\|_q \le \delta\sqrt{n}} \left\{ \zeta^T Dh\left(W_i\right)\Delta - (1-\varepsilon'')\|\Delta\|_q^\rho \right\}.
\end{aligned}$$

We now argue locally, using A6), a bound for the first term in the right hand side of (21):

$$\begin{aligned}
&\sup_{\|\Delta\|_q \le \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T \left[Dh(W_i + \Delta u/n^{1/2}) - Dh(W_i) \right] \Delta du - \varepsilon'' \|\Delta\|_q^\rho \right\} \qquad (22)\\
&\qquad \le \sup_{\|\Delta\|_q \le \delta\sqrt{n}} \left\{ \|\zeta\|_p \bar\kappa\left(W_i\right) \|\Delta\|_q^2 / n^{1/2} - \varepsilon'' \|\Delta\|_q^\rho \right\}\\
&\qquad \le \sup_{\|\bar\Delta\|_q \le 1} \left\{ b\bar\kappa\left(W_i\right) \left\|\bar\Delta\right\|_q^2 \delta^2 n^{1/2} - \varepsilon'' \left\|\bar\Delta\right\|_q^\rho \left(\delta n^{1/2}\right)^\rho \right\}.
\end{aligned}$$

As $\sup_{x\in[0,1]} \left\{ a_n x^2 - b_n x^\rho \right\} \le (\rho-2)^+ (a_n^\rho / b_n^2)^{1/(\rho-2)}/\rho$ when $b_n > a_n$, we have, for all n sufficiently large, that

$$\sup_{\|\Delta\|_q \le \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T \left[Dh(W_i + \Delta u/n^{1/2}) - Dh(W_i) \right] \Delta du - \varepsilon'' \|\Delta\|_q^\rho \right\} \le \frac{(\rho-2)^+}{\rho} \left(\frac{b\bar\kappa(W_i)}{\varepsilon''\sqrt{n}} \right)^{\rho/(\rho-2)}.$$

Since $\mathbb{E}[\bar\kappa(W)^2] < \infty$ (from Assumption A6)), we have that $\mathbb{P}(\bar\kappa(W_i) > \varepsilon'''\sqrt{i} \text{ i.o.}) = 0$ for any $\varepsilon''' > 0$. Consecutively, $\bar\kappa(W_i) < \varepsilon'''\sqrt{i}$ for all i large enough, and therefore,

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

which can be made arbitrarily small by choosing ε''' arbitrarily small. Therefore, for any fixed $\varepsilon'', \delta > 0$,

$$\overline{\lim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \sup_{\|\Delta\|_q \le \delta\sqrt{n}} \left\{ \int_0^1 \zeta^T \left[Dh\left(W_i + \Delta u/n^{1/2}\right) - Dh\left(W_i\right) \right] \Delta du - \varepsilon'' \|\Delta\|_q^\rho \right\} = 0. \qquad (23)$$

Execution of Step 3: Next, it follows from (20), (21) and (23) that for any fixed $\varepsilon'', \delta > 0$, there exists N_0 such that if $n \geq N_0$,

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

where

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

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

Next, note that A5) implies that

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

and, therefore, since $Dh\left(\cdot\right)$ is continuous (therefore locally bounded) and $\mathbb{E}\|W_i\|_q^\rho < \infty$ also by A5), we have that

$$\mathbb{E}\|Dh\left(W\right)\|_p^{\rho/(\rho-1)} < \infty.$$

Then, an argument similar to Lemma 2 shows that

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

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

A.4. **Proofs of RWP function limit theorems for linear and logistic regression examples.** We first obtain the dual formulation of the respective RWP functions for linear and logistic regressions using Proposition 3. Let $\mathbb{E}[h(x, y; \beta)] = \mathbf{0}$ be the estimating equation under consideration ($h(x, y; \beta) = (y - \beta^T x)x$ for linear regression and $h(x, y; \beta)$ as in (27) for logistic regression). Recall that the cost function is $c(\cdot) = N_q(\cdot)$. Due to the duality result in Proposition 3, we obtain

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

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

$$R_n(\beta_*) = \sup_{\lambda} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \sup_{x'} \left\{ \lambda^T h(x', Y_i; \beta_*) - \|x' - X_i\|_q^\rho \right\} \right\}, \tag{24}$$

where $\rho = 2$ for the case of linear regression (Theorem 5) and $\rho = 1$ for the case of logistic regression (Theorem 6). As RWP function here is similar to the RWP function for general

estimating equation in Section 3.3, a similar limit theorem holds. We state here the assumptions for proving RWP limit theorems for the dual formulation in (24).

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

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

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

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

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

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

Lemma 3. *If $\rho \geq 2$, under Assumptions A2'), A4') and A6'), we have,*

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

where

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

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

Lemma 4. *If $\rho = 1$, in addition to assuming A2'), A4'), suppose that $D_x h(\cdot,y;\beta_*)$ is continuous for every y in the support of probability distribution of Y. Also suppose that X has a positive probability density (almost everywhere) with respect to the Lebesgue measure. Then,*

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

where

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

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

The proof of Lemma 3 and 4 follows closely the proof of our results in Section 3 and therefore it is omitted. We prove Theorem 5 and 6 as a quick application of these lemmas.

Proof of Theorem 5. To show that the RWP function dual formulation in (24) converges in distribution, we verify the assumptions of Lemma 3 with $h(x,y;\beta) = (y - \beta^T x)x$. Under the null hypothesis H_0, $Y - \beta_*^T X = e$ is independent of X, has zero mean and finite variance σ^2. Therefore,

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

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

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

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

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

thus satisfying Assumption A4') trivially. In addition,

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

for some positive constant c. This verifies Assumption A6'). As all the assumptions imposed in Lemma 3 are easily satisfied, using $\rho = 2$, we obtain the following convergence in distribution as a consequence of Lemma 3.

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

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

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

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

Proof of the stochastic upper bound in Theorem 5: For the stochastic upper bound, let us consider the asymptotic distribution L_1 and rewrite the maximization problem as,

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

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

$$L_1 \leq \sup_{\|\xi\|_p=1} \frac{\sigma^2 \left\|Z\right\|_q^2}{\mathbb{E}\left\|e\xi - (\xi^T X)\beta_*\right\|_p^2} = \frac{\sigma^2 \left\|Z\right\|_q^2}{\inf_{\|\xi\|_p=1} \mathbb{E}\left\|e\xi - (\xi^T X)\beta_*\right\|_p^2}. \tag{25}$$

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

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

Combining the above inequality with (25), we obtain,

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

Consequently,

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

If random error e is normally distributed, then

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

thus establishing the desired upper bound. □

Proof of Theorem 6. Under null hypothesis H_0, the training samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are produced from the logistic regression model with parameter β_*. As β_* minimizes the expected log-exponential loss $l(x, y; \beta)$, the corresponding optimality condition is $\mathbb{E}[h(X, Y; \beta_*)] = \mathbf{0}$, where

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

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

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

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

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

for any $\xi \neq \mathbf{0}$, thus satisfying Assumption A4'). As all the conditions required for the convergence in distribution in Lemma 4 are satisfied, we obtain,

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

where $Z \sim \mathcal{N}(\mathbf{0}, \mathbb{E}[XX^T/(1 + \exp(Y\beta_*^T X))^2])$ as a consequence of Lemma 4. Here, the set $A = \{\xi \in \mathbb{R}^d : \text{ess sup}\|\xi^T D_x h(X, Y; \beta_*)\| \leq 1\}$.

Proof of the stochastic upper bound in Theorem 6: First, we claim that A is a subset of the norm ball $\{\xi \in \mathbb{R}^d : \|\xi\|_p \leq 1\}$. To establish this, we observe that,

$$\begin{aligned}\left\|\xi^T D_x h(X, Y; \beta_*)\right\|_p &\geq \left\|\frac{-Y\xi}{1 + \exp(Y\beta_*^T X)}\right\|_p - \left\|\frac{(\xi^T X)\beta_*}{\left(1 + \exp(Y\beta_*^T X)\right)\left(1 + \exp(Y\beta_*^T X)\right)}\right\|_p \\ &\geq \left(\frac{1}{1 + \exp(Y\beta_*^T X)} - \frac{\|X\|_q\|\beta_*\|_p}{\left(1 + \exp(Y\beta_*^T X)\right)\left(1 + \exp(-Y\beta_*^T X)\right)}\right)\|\xi\|_p,\end{aligned} \tag{26}$$

because $Y \in \{+1, -1\}$, and due to Hölder's inequality $|\xi^T X| \leq \|\xi\|_p\|X\|_q$. If $\xi \in \mathbb{R}^d$ is such that $\|\xi\|_p = (1 - \epsilon)^{-2} > 1$ for a given $\epsilon > 0$, then following (26), $\|\xi^T D_x h(X, Y)\|_p > 1$, whenever

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

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

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

If we let $\tilde{Z} \sim \mathcal{N}(\mathbf{0}, \mathbb{E}[XX^T])$, then $\mathrm{Cov}[\tilde{Z}] - \mathrm{Cov}[Z]$ is positive definite. As a result, L_3 is stochastically dominated by $L_4 := \|\tilde{Z}\|_q$, thus verifying the desired stochastic upper bound in the statement of Theorem 6. □

Proof of Theorem 7. Instead of characterizing the exact weak limit, we will find a stochastic upper bound for $R_n(\beta_*)$. The RWP function, as in the proof of Theorem 5, admits the following dual representation (see (24)):

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

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

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

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

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

where we have split the optimization into two parts: one over the magnitude (denoted by a), and another over all unit vectors ζ. Further, due to Hölder's inequality, we have $|\zeta^T Z_n| \leq \|Z_n\|_\infty$ as $\|\zeta\|_1 = 1$. Therefore, letting $c_1(n) = \|Z_n\|_\infty$, $c_2(n) = \inf_{\zeta : \|\zeta\|_1 = 1} \frac{1}{n} \sum_{i=1}^n \left\| e_i \zeta - (\zeta^T X_i) \beta_* \right\|_1^2$ and $c_3(n) = 1 + a \|\beta\|_1^2 n^{-1/2}$, observe that

$$nR_n(\beta_*) \leq \sup_{a \geq 0} \left\{ c_1(n) a - \frac{c_2(n)}{4 c_3(n)} a^2 \right\} = \frac{c_1^2(n)}{c_2(n)} (1 + o(1)) = \frac{\|Z_n\|_\infty^2 (1 + o(1))}{\inf_{\{\zeta : \|\zeta\|_1 = 1\}} \frac{1}{n} \sum_{i=1}^n \|e_i \zeta - (\zeta^T X_i) \beta_*\|_1^2}.$$

Since $\left\| e_i \zeta - (\zeta^T X_i) \beta_* \right\|_1^2 \geq \left(|e_i| \|\zeta\|_1 - \left| \zeta^T X_i \right| \|\beta_*\|_1 \right)^2$, the denominator, $c_2(n)$, can be lower bounded as follows:

$$\begin{aligned} c_2(n) &:= \inf_{\zeta : \|\zeta\|_1 = 1} \mathbb{E}_{\mathbb{P}_n} \left\| e\zeta - (\zeta^T X) \beta_* \right\|_1^2 \geq \inf_{\zeta : \|\zeta\|_1 = 1} \mathbb{E}_{\mathbb{P}_n} \left[\left(|e| - |\zeta^T X| \|\beta_*\|_1 \right)^2 \right] \\ &\geq \mathbb{E}_{\mathbb{P}_n} \left[\inf_{\zeta : \|\zeta\|_1 = 1} \mathbb{E}_{\mathbb{P}_n} \left[\left(|e| - |\zeta^T X| \|\beta_*\|_1 \right)^2 \mid X \right] \right] \geq \mathbb{E}_{\mathbb{P}_n} \left[\inf_{c \in \mathbb{R}} \mathbb{E}_{\mathbb{P}_n} \left[(|e| - c)^2 \mid X \right] \right]. \end{aligned}$$

Since e_i and X_i are independent and $\min_c \mathbb{E}[(Z - c)^2] = \mathrm{Var}[Z]$ for any random variable Z, we obtain that $c_2(n) \geq \mathrm{Var}_n |e|$. Therefore $nR_n(\beta_*) \leq \|Z_n\|_\infty^2 (1 + o(1)) / \mathrm{Var}_n |e|$. □

Appendix B. Strong duality for the linear semi-infinite program resulting from the RWP function

In the main body of the paper, we have utilized strong duality of linear semi-infinite programs to derive a dual representation of the RWP function in order to perform asymptotic analysis (see Proposition 3). Establishing strong duality in this context relies on the following well-known result on problem of moments ([2, 3]).

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

$$v(q) := \sup \big\{ \mathbb{E}_\mu[\phi(X)] : \mathbb{E}_\mu[f(X)] = q,\ \mu \in \mathcal{P}_\Omega \big\}. \tag{27}$$

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

$$v(q) := \sup \left\{ \int \phi(x) d\mu(x) : \int \tilde{f}(x) d\mu(x) = \tilde{q},\ \mu \in \mathcal{M}_\Omega^+ \right\}. \tag{28}$$

Then, under the assumption that a certain Slater's type of condition is satisfied, one has the following equivalent dual representation for the moment problem (28). See Theorem 1 (and the discussion of Case [I] following Theorem 1) in [2] for a proof of the following result:

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

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

In the rest of this section, we recast the dual reformulation of RWP function (in (3)) and the dual reformulation of the distributional representation in Proposition 1 as particular cases of the dual representation of the problem of moments in Proposition 1.

Dual representation of RWP function. Recall from Section 3.2 that W is a random vector taking values in $\mathbb{R}^m$ and $h(\cdot, \theta)$ is Borel measurable.

Proof of Proposition 3. For simplicity, we do not write the dependence on parameter θ in $h(u, \theta)$ and $R_n(\theta)$ in this proof; nevertheless, we should keep in mind that the RWP function is a function of parameter θ. Given estimating equation $\mathbb{E}[h(W)] = \mathbf{0}$, recall the definition of the corresponding RWP function,

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

where $\pi_{_W}$ denotes the marginal distribution of W and $\mathbb{P}_n$ is the empirical distribution formed from distinct samples $\{W_1, \ldots, W_n\}$. To recast this as a problem of moments as in (27), let

$\Omega = \{(u,w) \in \mathbb{R}^m \times \{W_1, \ldots, W_n\} : c(u,w) < \infty\}$,

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

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

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

is of the same form as (27). Since the constraints $\mathbb{E}_\pi[\mathbf{1}_{\{w=W_i\}}(U,W)] = 1/n$, for $i = 1, \ldots, n$, together specify that $\mathbb{P}_\pi(\Omega) = 1$, the constraint that $E_\pi[\mathbf{1}_\Omega(U,W)] = 1$ is redundant. Moreover, as $\{\mathbf{0}\}$ lies in the interior of convex hull of the range $\{h(u) : (u,w) \in \Omega\}$, observe that the set $\mathcal{Q}_f := \{\int f d\mu : \mu \in \mathcal{M}_\Omega^+\}$ is simply $\mathbb{R}_+^n \times \mathbb{R}$. Then it is immediate that the Slater's condition $q \in \text{int}(\mathcal{Q}_f)$ is satisfied for the moment problem,

$$R_n = -\sup\left\{\int \phi(u,w) d\mu(u,w) : \int f(u,w) d\mu(u,w) = q, \mu \in \mathcal{M}_\Omega^+\right\}.$$

Consequently, we obtain the following dual representation of R_n due to Proposition 1:

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

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

$$R_n = \sup_\lambda\left\{\frac{1}{n}\sum_{i=1}^n \inf_{u \in \mathbb{R}^m}\left\{c(u,W_i) + \lambda^T h(u)\right\}\right\}. \tag{29}$$

As λ is a free variable, we flip the sign of λ to arrive at the statement of Proposition 3. This completes the proof. □

Appendix C. Exchange of sup and inf in the DRO formulation (8)

The inf-sup exchange in Proposition 2 below is obtained by suitably modifying the inf-sup exchange in [1, Theorem 2] and its proof to accommodate more relaxed assumptions than in [1]. The sequence of steps in the proof of Proposition 2 is similar to that of [1, Theorem 2] and is given here for completeness.

Proposition 2. *For a given probability distribution $\mathbb{Q}$, define*

$$g(\beta) := \sup_{\mathbb{P}:\, \mathcal{D}_c(\mathbb{P},\mathbb{Q}) \leq \delta} \mathbb{E}_\mathbb{P}\left[l\big(X,Y;\beta\big)\right],$$

for $\beta \in \mathbb{R}^d$. Suppose that $g(\cdot)$ is real-valued and the level set $\{\beta \in \mathbb{R}^d : g(\beta) \leq b\}$ is bounded for every $b \in \mathbb{R}$. In addition, suppose that $\mathbb{E}_\mathbb{P}\left[l\big(X,Y;\beta\big)\right]$ is convex and lower semicontinuous in the variable β, for every $\mathbb{P} \in \mathcal{U}_\delta(\mathbb{Q}) := \{\mathbb{P} : \mathcal{D}_c(\mathbb{P},\mathbb{Q}) \leq \delta\}$. Then,

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

Proof. We begin by defining the sequence of approximation problems,

$$g_N(\beta) := \sup_{\mathbb{P}\in\mathcal{U}_\delta^N(\mathbb{Q})} \mathbb{E}_{\mathbb{P}}\left[l\big(X,Y;\beta\big)\right],$$

where $N = 1, 2, \ldots$, and

$$\mathcal{U}_\delta^N(\mathbb{Q}) = \{\mathbb{P} \in \mathcal{P}(\mathcal{K}_N) : \mathcal{D}_c(\mathbb{P},\mathbb{Q}) \leq \delta\},$$

with $\mathcal{P}(\mathcal{K}_N)$ denoting the set of probability distributions over the set $\mathcal{K}_N := \{x : \|x\|_2 \leq N\}$. Then, due to the compactness of the set $\mathcal{U}_\delta^N(\mathbb{Q})$, we obtain

$$\inf_{\beta\in\mathbb{R}^d} g_N(\beta) = \inf_{\beta\in\mathbb{R}^d} \sup_{\mathbb{P}\in\mathcal{U}_\delta^N(\mathbb{Q})} \mathbb{E}_{\mathbb{P}}\left[l\big(X,Y;\beta\big)\right] = \sup_{\mathbb{P}\in\mathcal{U}_\delta^N(\mathbb{Q})} \inf_{\beta\in\mathbb{R}^d} \mathbb{E}_{\mathbb{P}}\left[l\big(X,Y;\beta\big)\right],$$

as a consequence of Sion's minimax theorem [6]. Therefore, with $g_N(\cdot)$ being an increasing sequence of functions, we have

$$\begin{aligned}\lim_{N\to\infty}\inf_{\beta\in\mathbb{R}^d} g_N(\beta) &= \sup_{N\geq 1}\inf_{\beta\in\mathbb{R}^d} g_N(\beta) = \sup_{N\geq 1}\sup_{\mathbb{P}\in\mathcal{U}_\delta^N(P_n)}\inf_{\beta\in\mathbb{R}^d}\mathbb{E}_{\mathbb{P}}\left[l\big(X,Y;\beta\big)\right]\\ &\leq \sup_{\mathbb{P}\in\mathcal{U}_\delta(\mathbb{Q})}\inf_{\beta\in\mathbb{R}^d}\mathbb{E}_{\mathbb{P}}\left[l\big(X,Y;\beta\big)\right] \leq \inf_{\beta\in\mathbb{R}^d}\sup_{\mathbb{P}\in\mathcal{U}_\delta(\mathbb{Q})}\mathbb{E}_{\mathbb{P}}\left[l\big(X,Y;\beta\big)\right] \qquad (30)\\ &= \inf_{\beta\in\mathbb{R}^d} g(\beta).\end{aligned}$$

The rest of the proof is divided into three technical steps:

Step 1: In this step, we show that the sequence of functions $\{g_N(\cdot) : N \geq 1\}$ converges pointwise to the function $g(\cdot)$, as $N \to \infty$. Since $g_N(\beta)$ is increasing in N, we have that $g_N(\beta)$ converges as $N \to \infty$, for every β. Let the function $g^*(\cdot)$ denote the pointwise limit, $g^*(\cdot) = \lim_{N\to\infty} g_N(\cdot)$. With $g_N(\cdot) \leq g(\cdot)$ for every N, we have $g^*(\beta) \leq g(\beta)$. Since $g(\cdot)$ is real-valued, we consequently have $g^*(\beta) \leq g(\beta) < +\infty$, for every $\beta \in \mathbb{R}^d$.

To show that $g^*(\beta)$ necessarily equals $g(\beta)$ for every β, we argue via contradiction as follows: Suppose that $\varepsilon := g(\beta) - g^*(\beta) > 0$ for some $\beta \in \mathbb{R}^d$. Consider any $\mathbb{P}' \in \mathcal{U}_\delta(P_n)$ such that $\mathbb{E}_{\mathbb{P}'}[l(X,Y;\beta)] \in (g(\beta) - \varepsilon/2, g(\beta)]$. With $g(\beta)$ being finite, there exists N_0 sufficiently large such that

$$\mathbb{E}_{\mathbb{P}'}\left[l(X,Y;\beta)\mathbb{I}(\|X\|_2 > N)\right] < \varepsilon/4 \quad \text{and} \quad \left[1-\mathbb{P}'(\mathcal{K}_N)\right]\mathbb{E}_{\mathbb{Q}}\left[l(X,Y;\beta)\mathbb{I}(\|X\|_2 \leq N)\right] > -\varepsilon/4,$$

for all $N > N_0$. From $\mathbb{P}'$, we construct a measure $\mathbb{P}'_N \in \mathcal{U}_\delta^N(\mathbb{Q})$ by letting,

$$\mathbb{P}'_N(\cdot) = \mathbb{P}'(\cdot) + \left[1-\mathbb{P}'(\mathcal{K}_N)\right]\frac{\mathbb{Q}(\cdot)}{\mathbb{Q}(\mathcal{K}_N)},$$

for all N large enough such that $\mathbb{Q}(\mathcal{K}_N) > 0$. Then,

$$g^*(\beta) \geq g_N(\beta) \geq \mathbb{E}_{\mathbb{P}'_N}\left[l(X,Y;\beta)\right] > \mathbb{E}_{\mathbb{P}'}\left[l(X,Y;\beta)\right] - \varepsilon/2,$$

for all $N > N_0$. With $\mathbb{E}_{\mathbb{P}'}[l(X,Y;\beta)] \in (g(\beta)-\varepsilon/2, g(\beta)]$, we then have $g^*(\beta) > g(\beta) - \varepsilon$, which leads to a contradiction to the assumption that $\varepsilon := g(\beta) - g^*(\beta) > 0$. This verifies that the pointwise limit $g^*(\cdot) = g(\cdot)$.

Step 2: In this next step, we show that the sequence of functions $\{g_N(\cdot) : N \geq 1\}$ epiconverges to the function $g(\cdot)$, as $N \to \infty$. See, for example, [4, Definition 7.1] for a definition of epiconvergence. To accomplish this step, we first see that for every sequence $\{\beta_N : N \geq 1\}$ satisfying $\beta_N \to \beta \in \mathbb{R}^d$,

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

for any positive integer M. Indeed, this is because $g_N(\cdot)$ is an increasing sequence of functions and $g_M(\cdot)$, being pointwise maxima of lower semicontinuous functions, is lower semicontinuous. Letting $M \to \infty$, we then have

$$\liminf_{N\to\infty} g_N(\beta_N) \geq g(\beta),$$

due to the pointwise convergence concluded in Step 1. Next, for any $\beta \in \mathbb{R}^d$, if we pick the sequence $\beta_N = \beta$, we have $\lim_{N\to\infty} g_N(\beta_N) = \lim_{N\to\infty} g_N(\beta) = g(\beta)$. We therefore have from the epiconvergence characterization in [4, Proposition 7.1] that the sequence $\{g_N : N \geq 1\}$ epiconverges to the function $g(\cdot)$.

Step 3: In this final step, we show that the optimal values $\inf_{\beta\in\mathbb{R}^d} g_N(\beta)$ converge to $\inf_{\beta\in\mathbb{R}^d} g(\beta)$, as $N \to \infty$. With $\mathbb{E}_{\mathbb{P}}[l(X, Y; \beta)]$ being convex in the variable β, we have that the pointwise maximum $g(\cdot)$ is convex. Combining this observation with the level-boundedness of the limiting function $g(\cdot)$, we have from [4, Exercise 7.32(c)] that the sequence $\{g_N(\beta) : N \geq 1\}$ is eventually level-bounded. Further, since the functions $g_N(\cdot), g(\cdot)$ are lower semicontinuous and proper, we obtain the desired optimal value convergence,

$$\inf_{\beta\in\mathbb{R}^d} g_N(\beta) \to \inf_{\beta\in\mathbb{R}^d} g(\beta),$$

as a consequence of [4, Theorem 7.33].

The conclusion in Step 3 forces the inequalities in (30) to be equalities, thus rendering the desired inf-sup interchange in the statement of Proposition 2. □

Proof of Lemma 1. Let us consider linear regression loss function first. Under the null hypothesis, $\mathbb{E}\|X\|_2^2 < \infty$ and $\mathbb{E}[e^2] < \infty$. Therefore, for any $\beta \in \mathbb{R}^d$, $\mathbb{E}[l(X, Y; \beta)] = \mathbb{E}[(Y - \beta^T X)^2] < \infty$. Further, as the loss function $l(x, y; \beta)$ is a convex and continuous in the variable β, we have that $\mathbb{E}_{\mathbb{P}}[l(X, Y; \beta)]$ is convex and lower semicontinuous for any $\mathbb{P} \in \mathcal{U}_\delta(\mathbb{P}_n)$. Next, the distributionally robust representation in Theorem 1,

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

allows us to conclude that $g(\beta)$ is finite for every $\beta \in \mathbb{R}^d$. Further, as $g(\beta) \to \infty$ when $\|\beta\|_p \to \infty$ and $g(\beta)$ is convex and continuous in $\mathbb{R}^d$, the level sets $\{\beta : g(\beta) \leq b\}$ are compact and nonempty for every $b > (\sqrt{\mathbb{E}_{\mathbb{P}_n}[(Y - \beta_*^T X)^2]} + \sqrt{\delta}\|\beta_*\|)^2$. This verifies the level-boundedness requirement in the statement of Proposition 2. As all the conditions in Proposition 2 are satisfied, the sup and inf in the DRO formulation (8) can be exchanged in the linear regression example as a consequence of Proposition 2. Exactly similar reasoning applies for logistic regression loss function when $\mathbb{E}\|X\|_2^2$ is finite. □

References

[1] Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in Wasserstein distributionally robust estimation. *arXiv preprint arXiv:1906.01614*, 2019.

[2] Keiiti Isii. On sharpness of Tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14(1):185–197, 1962.

[3] Whitney Newey and Richard Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

[4] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[5] Soroosh Shafieezadeh-Abadeh, Peyman Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584. 2015.

[6] Maurice Sion and others. On general minimax theorems. *Pacific J. Math*, 8(1):171–176, 1958.
[7] Frode Terkelsen. Some minimax theorems. *Mathematica Scandinavica*, 31(2):405–413, 1973.