Appendix: Principled Learning Method for Wasserstein Distributionally Robust Optimization with Local Perturbations

A. Proofs

When $M = 0$ and $\beta_n = 0$ for all n, a β_n-locally perturbed data distribution is the empirical data distribution, *i.e.*, $\mathbb{P}'_n = \mathbb{P}_n$. Therefore, Theorem 1 is a special case of Theorem 4. Also, in such cases, $R^{\mathrm{prop}}_{\alpha_n,p}(\mathbb{P}_n, h) = R^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, h)$ and $\hat{h}^{\mathrm{prop}}_{\alpha_n,p} = \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}$, and Theorems 2 and thus 3 are a special case of Theorems 5 and 6, respectively. In this respect, we omit proofs for Theorems 1, 2, and 3.

A.1. Proof of Proposition 1

Proof of Proposition 1. Since $\mathbb{P}_n \in \mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)$, we have

$$R(\mathbb{P}_n, h) \le R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h).$$

Let $\mathbb{Q}^*$ be such that $R(\mathbb{Q}^*, h) = \sup_{\mathbb{Q}\in\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)} R(\mathbb{Q}, h) = R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h)$. Since h is Lipschitz continuous, the Kantorovich-Rubinstein duality (Villani, 2008, Remark 6.5) gives

$$\begin{aligned} R(\mathbb{Q}^*, h) - R(\mathbb{P}_n, h) &\le \mathrm{Lip}(h)\mathcal{W}_1(\mathbb{Q}^*, \mathbb{P}_n) \\ &\le \mathrm{Lip}(h)\mathcal{W}_p(\mathbb{Q}^*, \mathbb{P}_n) \\ &\le \mathrm{Lip}(h)\alpha_n. \end{aligned}$$

Here, the second inequality is due to $\mathcal{W}_1(\mathbb{Q}^*, \mathbb{P}_n) \le \mathcal{W}_p(\mathbb{Q}^*, \mathbb{P}_n)$ for $p \in [1, \infty)$ (Villani, 2008, Remark 6.6). Thus,

$$\left| R(\mathbb{P}_n, h) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) \right| \le \mathrm{Lip}(h)\alpha_n. \tag{9}$$

Write $\mathbb{P}'_n = \frac{1}{n}\sum_{i=1}^n \delta_{z'_i}$ for some $\{z'_1, \dots, z'_n\}$ such that $\left\| z'_i - z_i \right\| \le \beta_n$ for all $i \in [n]$. Then, we have $z'_i \in \mathcal{Z} + \mathcal{B}(M)$ and $h(z'_i)$'s are well defined. By the Lipschitz continuity of h and the definition of $\mathbb{P}'_n$, we have

$$\begin{aligned} \left| R(\mathbb{P}_n, h) - R(\mathbb{P}'_n, h) \right| &= \left| \frac{1}{n}\sum_{i=1}^n (h(z_i) - h(z'_i)) \right| \\ &\le \frac{1}{n}\sum_{i=1}^n \mathrm{Lip}(h)\left\| z'_i - z_i \right\| \\ &\le \mathrm{Lip}(h)\beta_n. \end{aligned}$$

Therefore, we have

$$\left| R(\mathbb{P}'_n, h) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) \right| \le (\alpha_n + \beta_n)\mathrm{Lip}(h).$$

This concludes the proof. □

A.2. Proof of Theorem 4

Proof of Theorem 4. Write $\mathbb{P}'_n = \frac{1}{n}\sum_{i=1}^n \delta_{z'_i}$ for some $\{z'_1, \dots, z'_n\}$ such that $\left\| z'_i - z_i \right\| \le \beta_n$ for all $i \in [n]$. Then, we have $z'_i \in \mathrm{Conv}(\mathcal{Z}) + \mathcal{B}(M)$ and $h(z'_i)$'s are well defined.

[Step 1] In this step we first establish an upper bound for the local worst-case risk $R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, h)$. Since h is well defined and differentiable on $\operatorname{Conv}(\mathcal{Z}) + \mathcal{B}(M)$, we can apply the mean value theorem. Due to the (C_{H}, k)-Hölder continuity of $\nabla_z h$, for any $i \in [n]$ and $\tilde{z}_i \in \mathcal{Z}$, we have

$$\begin{aligned}
h(\tilde{z}_i) &= h(z_i') + \langle \nabla_z h(c_i), \tilde{z}_i - z_i' \rangle \\
&= h(z_i') + \langle \nabla_z h(z_i'), \tilde{z}_i - z_i' \rangle + \langle \nabla_z h(c_i) - \nabla_z h(z_i'), \tilde{z}_i - z_i' \rangle \\
&\leq h(z_i') + \big\|\nabla_z h(z_i')\big\|_* \big\|\tilde{z}_i - z_i'\big\| + C_{\mathrm{H}} \big\|\tilde{z}_i - z_i'\big\|^{1+k},
\end{aligned}$$

where $c_i = \tau_i z_i' + (1 - \tau_i)\tilde{z}_i$ for some $\tau_i \in [0, 1]$. By the triangle inequality and Jensen's inequality, $(a + b)^{1+k} \leq 2^k(a^{1+k} + b^{1+k})$ for any $a, b \geq 0$, we have

$$\begin{aligned}
&h(z_i') + \big\|\nabla_z h(z_i')\big\|_* \big\|\tilde{z}_i - z_i'\big\| + C_{\mathrm{H}} \big\|\tilde{z}_i - z_i'\big\|^{1+k} \\
&\leq h(z_i') + \big\|\nabla_z h(z_i')\big\|_* (\beta_n + \|\tilde{z}_i - z_i\|) + C_{\mathrm{H}} 2^k (\|\tilde{z}_i - z_i\|^{1+k} + \beta_n^{1+k}) \\
&= \beta_n \left(\big\|\nabla_z h(z_i')\big\|_* + C_{\mathrm{H}} 2^k \beta_n^k \right) + h(z_i') + \big\|\nabla_z h(z_i')\big\|_* \|\tilde{z}_i - z_i\| + C_{\mathrm{H}} 2^k \|\tilde{z}_i - z_i\|^{1+k}.
\end{aligned}$$

To this end, we set $C_{\mathrm{H},k} := C_{\mathrm{H}} 2^k$ and $t_i := \|\tilde{z}_i - z_i\|$. By Gao et al. (2017, Lemma 2), for any $\eta > 0$ and $\lambda \geq 0$, we have

$$\begin{aligned}
&\big\|\nabla_z h(z_i')\big\|_* t_i + C_{\mathrm{H},k} t_i^{1+k} - \lambda t_i^p \\
&\leq \left(\big\|\nabla_z h(z_i')\big\|_* + \frac{p-k-1}{p-1} C_{\mathrm{H},k} \eta \right) t_i - \left(\lambda - \frac{k}{p-1} C_{\mathrm{H},k} \eta^{-\frac{p-k-1}{k}} \right) t_i^p.
\end{aligned}$$

By substituting η with α_n^k,

$$\begin{aligned}
&\big\|\nabla_z h(z_i')\big\|_* t_i + C_{\mathrm{H},k} t_i^{1+k} - \lambda t_i^p \\
&\leq \left(\big\|\nabla_z h(z_i')\big\|_* + \frac{p-k-1}{p-1} C_{\mathrm{H},k} \alpha_n^k \right) t_i - \left(\lambda - \frac{k}{p-1} C_{\mathrm{H},k} \alpha_n^{-(p-k-1)} \right) t_i^p \\
&=: h_{\alpha_n}(z_i') t_i - \left(\lambda - C_{\alpha_n} \right) t_i^p. \qquad (10)
\end{aligned}$$

Since $\mathcal{Z}$ is bounded, there exists a constant $D_{\mathcal{Z}}$ such that $\sup_{z, \tilde{z} \in \mathcal{Z}} \|z - \tilde{z}\| \leq D_{\mathcal{Z}}$. Then,

$$\sup_{0 \leq t \leq D_{\mathcal{Z}}} \{ h_{\alpha_n}(z_i') t - (\lambda - C_{\alpha_n}) t^p \} = \begin{cases} h_{\alpha_n}(z_i') D_{\mathcal{Z}} - (\lambda - C_{\alpha_n}) D_{\mathcal{Z}}^p & \text{if } 0 \leq \lambda \leq C_{\alpha_n}, \\ h_{\alpha_n}(z_i') t_*(\lambda) - (\lambda - C_{\alpha_n}) t_*^p(\lambda) & \text{if } C_{\alpha_n} < \lambda, \end{cases}$$

where $t_*(\lambda) = \min\left\{ \left(\frac{h_{\alpha_n}(z_i')}{(\lambda - C_{\alpha_n})p} \right)^{1/(p-1)}, D_{\mathcal{Z}} \right\}$. Here,

$$\left(\frac{h_{\alpha_n}(z_i')}{(\lambda - C_{\alpha_n})p} \right)^{1/(p-1)} < D_{\mathcal{Z}} \Leftrightarrow C_{\alpha_n} + \frac{h_{\alpha_n}(z_i')}{p D_{\mathcal{Z}}^{p-1}} < \lambda.$$

Thus,

$$\sup_{0 \leq t \leq D_{\mathcal{Z}}} \{ h_{\alpha_n}(z_i') t - (\lambda - C_{\alpha_n}) t^p \} = \begin{cases} h_{\alpha_n}(z_i') D_{\mathcal{Z}} - (\lambda - C_{\alpha_n}) D_{\mathcal{Z}}^p, & \text{if } 0 \leq \lambda \leq C_{\alpha_n} + \frac{h_{\alpha_n}(z_i')}{p D_{\mathcal{Z}}^{p-1}}, \\ p^{-p^*} (p-1) (\lambda - C_{\alpha_n})^{-\frac{1}{p-1}} \big\| h_{\alpha_n}(z_i') \big\|_*^{p^*}, & \text{if } C_{\alpha_n} + \frac{h_{\alpha_n}(z_i')}{p D_{\mathcal{Z}}^{p-1}} < \lambda. \end{cases}$$

Note that $\big\|h_{\alpha_n}(z_i')\big\|_* = h_{\alpha_n}(z_i')$. Let $\lambda_* := C_{\alpha_n} + \frac{\max_{i \in [n]} \{ h_{\alpha_n}(z_i') \}}{p D_{\mathcal{Z}}^{p-1}}$. Using the triangle inequality and the Hölder continuity of $\nabla_z h$, for any $z \in \mathcal{Z}$ and some point $z_0 \in \mathcal{Z}$, we have

$$\begin{aligned}
\big\|\nabla_z h(z)\big\|_* &\leq \big\|\nabla_z h(z_0)\big\|_* + \big\|\nabla_z h(z) - \nabla_z h(z_0)\big\|_* \\
&\leq \big\|\nabla_z h(z_0)\big\|_* + C_{\mathrm{H}} \|z - z_0\|^k \\
&\leq \big\|\nabla_z h(z_0)\big\|_* + C_{\mathrm{H}} D_{\mathcal{Z}}^k.
\end{aligned}$$

This implies $\left\|\nabla_z h(z)\right\|_*$ is bounded for all $z \in \mathrm{Conv}(\mathcal{Z}) + \mathcal{B}(M)$. We denote the upper bound by L_∇, *i.e.*, $\left\|\nabla_z h(z)\right\|_* \leq L_\nabla < \infty$ for all $z \in \mathrm{Conv}(\mathcal{Z}) + \mathcal{B}(M)$. Then, we have

$$\frac{\max_{i\in[n]}\{h_{\alpha_n}(z_i')\}}{pD_{\mathcal{Z}}^{p-1}} \leq \frac{L_\nabla + \frac{p-k-1}{p-1}C_{\mathrm{H},k}\alpha_n^k}{pD_{\mathcal{Z}}^{p-1}} < \infty. \tag{11}$$

At the same time, by the definition of $\|h_{\alpha_n}\|_{\mathbb{P}'_n,1}$, we have

$$\frac{0 + \frac{p-k-1}{p-1}C_{\mathrm{H},k}\alpha_n^k}{p\alpha_n^{p-1}} \leq \frac{\|h_{\alpha_n}\|_{\mathbb{P}'_n,1}}{p\alpha_n^{p-1}}, \tag{12}$$

and the left-hand side diverges to infinity as n increases due to $p > 1 + k$. Since $\|h_{\alpha_n}\|_{\mathbb{P}'_n,1} \leq \|h_{\alpha_n}\|_{\mathbb{P}'_n,p^*}$ and by the inequalities (11) and (12) give for a large enough n,

$$\lambda_* < C_{\alpha_n} + \frac{\|h_{\alpha_n}\|_{\mathbb{P}'_n,p^*}}{p\alpha_n^{p-1}}.$$

Therefore, for a large enough n,

$$\begin{aligned}
\inf_{\lambda_*<\lambda}\left\{\lambda\alpha_n^p + \frac{1}{n}\sum_{i=1}^n \sup_{0\leq t\leq D_{\mathcal{Z}}}\{h_{\alpha_n}(z_i')t - (\lambda - C_{\alpha_n})t^p\}\right\} &= C_{\alpha_n}\alpha_n^p + \alpha_n\|h_{\alpha_n}\|_{\mathbb{P}'_n,p^*} \\
&\leq C_{\alpha_n}\alpha_n^p + \alpha_n\left\{\|\nabla_z h\|_{\mathbb{P}'_n,p^*} + \frac{p-k-1}{p-1}C_{\mathrm{H},k}\alpha_n^k\right\} \\
&= \alpha_n\|\nabla_z h\|_{\mathbb{P}'_n,p^*} + C_{\mathrm{H},k}\alpha_n^{1+k}.
\end{aligned} \tag{13}$$

The inequality is due to the Minkowski inequality. By arranging all the results, for a large enough n, we have

$$\begin{aligned}
&R_{\alpha_n,p}^{\mathrm{worst}}(\mathbb{P}_n, h) - R(\mathbb{P}'_n, h) \\
&\overset{(4)}{=} \min_{\lambda\geq 0}\left\{\lambda\alpha_n^p + \frac{1}{n}\sum_{i=1}^n \sup_{\tilde{z}\in\mathcal{Z}}\left\{h(\tilde{z}) - h(z_i') - \lambda\|\tilde{z} - z_i\|^p\right\}\right\} \\
&\leq \beta_n(\|\nabla_z h\|_{\mathbb{P}'_n,1} + C_{\mathrm{H},k}\beta_n^k) + \min_{\lambda\geq 0}\left\{\lambda\alpha_n^p + \frac{1}{n}\sum_{i=1}^n \sup_{\tilde{z}\in\mathcal{Z}}\left\{\left\|\nabla_z h(z_i')\right\|_*\|\tilde{z}_i - z_i\| + C_{\mathrm{H},k}\|\tilde{z}_i - z_i\|^{1+k} - \lambda\|\tilde{z} - z_i\|^p\right\}\right\} \\
&\leq \beta_n(\|\nabla_z h\|_{\mathbb{P}'_n,1} + C_{\mathrm{H},k}\beta_n^k) + \min_{\lambda\geq 0}\left\{\lambda\alpha_n^p + \frac{1}{n}\sum_{i=1}^n \sup_{0\leq t\leq D_{\mathcal{Z}}}\left\{\left\|\nabla_z h(z_i')\right\|_* t + C_{\mathrm{H},k}t^{1+k} - \lambda t^p\right\}\right\} \\
&\overset{(10)}{\leq} \beta_n(\|\nabla_z h\|_{\mathbb{P}'_n,1} + C_{\mathrm{H},k}\beta_n^k) + \min_{\lambda\geq\lambda_*}\left\{\lambda\alpha_n^p + \frac{1}{n}\sum_{i=1}^n \sup_{0\leq t_i\leq D_{\mathcal{Z}}}\left\{h_{\alpha_n}\left(z_i'\right)t_i - \left(\lambda - C_{\alpha_n}\right)t_i^p\right\}\right\} \\
&\overset{(13)}{\leq} \beta_n(\|\nabla_z h\|_{\mathbb{P}'_n,1} + C_{\mathrm{H},k}\beta_n^k) + \alpha_n\|\nabla_z h\|_{\mathbb{P}'_n,p^*} + C_{\mathrm{H},k}\alpha_n^{1+k} \\
&= O(\beta_n + \alpha_n^{1+k}) + \alpha_n\|\nabla_z h\|_{\mathbb{P}'_n,p^*}.
\end{aligned}$$

Thus, we have

$$R_{\alpha_n,p}^{\mathrm{worst}}(\mathbb{P}_n, h) - R(\mathbb{P}'_n, h) - \alpha_n\|\nabla_z h\|_{\mathbb{P}'_n,p^*} = O(\beta_n + \alpha_n^{1+k}). \tag{14}$$

[Step 2] In this step, we establish a lower bound for the local worst-case risk $R_{\alpha_n,p}^{\mathrm{worst}}(\mathbb{P}_n, h)$. By the definition of the Wasserstein ball $\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)$, we have

$$\begin{aligned}
&R_{\alpha_n,p}^{\mathrm{worst}}(\mathbb{P}_n, h) - R(\mathbb{P}'_n, h) \\
&\geq \sup_{\tilde{z}_i\in\mathcal{Z}}\left\{\frac{1}{n}\sum_{i=1}^n\{h(\tilde{z}_i) - h(z_i')\} \mid \left(\frac{1}{n}\sum_{i=1}^n\|\tilde{z}_i - z_i\|^p\right)^{1/p} \leq \alpha_n\right\}.
\end{aligned}$$

Again, the mean value theorem and the Hölder continuity assumption on $\nabla_z h$ give

$$\begin{aligned} h(\tilde{z}_i) &= h(z_i') + \langle \nabla_z h(c_i), \tilde{z}_i - z_i' \rangle \\ &= h(z_i') + \langle \nabla_z h(z_i'), \tilde{z}_i - z_i' \rangle + \langle \nabla_z h(c_i) - \nabla_z h(z_i'), \tilde{z}_i - z_i' \rangle \\ &\geq h(z_i') + \langle \nabla_z h(z_i'), \tilde{z}_i - z_i' \rangle - C_{\mathrm{H}} \left\| \tilde{z}_i - z_i' \right\|^{1+k} \\ &\geq h(z_i') + \langle \nabla_z h(z_i'), (\tilde{z}_i - z_i) + (z_i - z_i') \rangle - C_{\mathrm{H},k} \left(\left\| \tilde{z}_i - z_i \right\|^{1+k} + \beta_n^{1+k} \right) \\ &\geq h(z_i') + \langle \nabla_z h(z_i'), \tilde{z}_i - z_i \rangle - \left\| \nabla_z h(z_i') \right\|_* \beta_n - C_{\mathrm{H},k} \left(\left\| \tilde{z}_i - z_i \right\|^{1+k} + \beta_n^{1+k} \right), \end{aligned}$$

where $c_i = tz_i + (1-t)\tilde{z}_i$ for some $t \in [0,1]$. Thus, we have

$$\begin{aligned} & R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) - R(\mathbb{P}'_n, h) \\ \geq & -\beta_n (\|\nabla_z h\|_{\mathbb{P}'_n,1} + C_{\mathrm{H},k}\beta_n^k) \\ & + \sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n \{ \langle \nabla_z h(z_i'), \tilde{z}_i - z_i \rangle - C_{\mathrm{H},k} \|\tilde{z}_i - z_i\|^{1+k} \} \mid \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{z}_i - z_i\|^p \right)^{1/p} \leq \alpha_n \right\} \\ \geq & -\beta_n (\|\nabla_z h\|_{\mathbb{P}'_n,1} + C_{\mathrm{H},k}\beta_n^k) \\ & + \sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n \langle \nabla_z h(z_i'), \tilde{z}_i - z_i \rangle \mid \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{z}_i - z_i\|^p \right)^{1/p} \leq \alpha_n \right\} \\ & - \sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n C_{\mathrm{H},k} \|\tilde{z}_i - z_i\|^{1+k} \mid \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{z}_i - z_i\|^p \right)^{1/p} \leq \alpha_n \right\} \\ =: & -\beta_n \left(\|\nabla_z h\|_{\mathbb{P}'_n,1} + C_{\mathrm{H},k}\beta_n^k \right) + S_1 - S_2. \end{aligned}$$

As for the term S_1, by the definition of the dual norm we have

$$S_1 \leq \sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n \left\| \nabla_z h(z_i') \right\|_* \|\tilde{z}_i - z_i\| \mid \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{z}_i - z_i\|^p \right)^{1/p} \leq \alpha_n \right\},$$

and by the Hölder inequality,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \left\| \nabla_z h(z_i') \right\|_* \|\tilde{z}_i - z_i\| \leq & \left(\frac{1}{n} \sum_{i=1}^n \left\| \nabla_z h(z_i') \right\|_*^{p^*} \right)^{1/p^*} \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{z}_i - z_i\|^p \right)^{1/p} \\ \leq & \; \alpha_n \|\nabla_z h\|_{\mathbb{P}'_n,p^*}, \end{aligned}$$

where the inequalities hold with equalities when for all $i \in [n]$

$$\|\tilde{z}_i - z_i\| = \alpha_n \left(\frac{\left\| \nabla_z h(z_i') \right\|_*^{p^*}}{\frac{1}{n} \sum_{j=1}^n \left\| \nabla_z h(z_j') \right\|_*^{p^*}} \right)^{1/p}.$$

Here,

$$\alpha_n \left(\frac{\left\| \nabla_z h(z_i') \right\|_*^{p^*}}{\frac{1}{n} \sum_{j=1}^n \left\| \nabla_z h(z_j') \right\|_*^{p^*}} \right)^{1/p} = \alpha_n \left(\frac{\left\| \nabla_z h(z_i') \right\|_*}{\|\nabla_z h\|_{\mathbb{P}'_n,p^*}} \right)^{p^*/p} \leq \alpha_n \left(\frac{\left\| \nabla_z h(z_i') \right\|_*}{\|\nabla_z h\|_{\mathbb{P}'_n,1}} \right)^{p^*/p}.$$

Since α_n vanishes and $\mathcal{Z}$ is an open set, $\tilde{z}_i \in \mathcal{Z}$ if the term $\frac{\|\nabla_z h(z_i')\|_*}{\|\nabla_z h\|_{\mathbb{P}_n',1}}$ is bounded. That is, the boundedness of $\frac{\|\nabla_z h(z_i')\|_*}{\|\nabla_z h\|_{\mathbb{P}_n',1}}$ is a sufficient condition to achieve $S_1 = \alpha_n \|\nabla_z h\|_{\mathbb{P}_n',p^*}$. It is noteworthy that the numerator $\|\nabla_z h(z_i')\|_*$ is bounded by L_∇, and due to the local perturbation, we have

$$\begin{aligned}
\|\nabla_z h(z_i')\|_* &\geq \|\nabla_z h(z_i')\|_* - \|\nabla_z h(z_i) - \nabla_z h(z_i')\|_* \\
&\geq \|\nabla_z h(z_i)\|_* - C_\mathrm{H} \|z_i' - z_i\|^{1+k} \\
&\geq \|\nabla_z h(z_i)\|_* - C_\mathrm{H} \beta_n^{1+k}.
\end{aligned}$$

Thus it is enough to show that the denominator $\|\nabla_z h\|_{\mathbb{P}_n,1}$ has a lower bound.

By the assumption $\mathbb{E}_\mathrm{data}(\|\nabla_z h\|_*) \geq C_\nabla$ and the fact $\|\nabla_z h(z)\|_* \leq L_\nabla$ for all $z \in \mathrm{Conv}(\mathcal{Z}) + \mathcal{B}(M)$, the McDiarmid inequality (Devroye et al., 1996, pages 136-137) implies that for a fixed $\delta > 0$, the following holds with probability at least $1 - \delta$.

$$\|\nabla_z h\|_{\mathbb{P}_n,1} \geq \mathbb{E}_\mathrm{data}(\|\nabla_z h\|_*) - L_\nabla \sqrt{\frac{2}{n} \log(\frac{1}{\delta})}. \tag{15}$$

Therefore, for a large enough n, $\|\nabla_z h\|_{\mathbb{P}_n,1}$ is strictly greater than zero with high probability, and this implies that $S_1 = \alpha_n \|\nabla_z h\|_{\mathbb{P}_n',p^*}$ with high probability.

As for the term S_2, we note the fact $(\frac{1}{n}\sum_{i=1}^n \|\tilde{z}_i - z_i\|^{1+k})^{\frac{1}{1+k}} \leq \left(\frac{1}{n}\sum_{i=1}^n \|\tilde{z}_i - z_i\|^p\right)^{1/p}$ as $p > 1 + k$. Since the equality holds when $\|\tilde{z}_i - z_i\| = \alpha_n$ for all $i \in [n]$, we have

$$\sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n C_{\mathrm{H},k} \|\tilde{z}_i - z_i\|^{1+k} \mid \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{z}_i - z_i\|^p \right)^{1/p} \leq \alpha_n \right\} \leq C_{\mathrm{H},k} \alpha_n^{1+k}.$$

Thus, combining the terms S_1 and S_2 shows that for a large enough n and a fixed $\delta > 0$, the following holds with probability at least $1 - \delta$.

$$R_{\alpha_n,p}^\mathrm{worst}(\mathbb{P}_n, h) - R(\mathbb{P}_n', h) - \alpha_n \|\nabla_z h\|_{\mathbb{P}_n',p^*} \geq -\beta_n (\|\nabla_z h\|_{\mathbb{P}_n',1} + C_{\mathrm{H},k} \beta_n^k) - C_{\mathrm{H},k} \alpha_n^{1+k}. \tag{16}$$

[Step 3] By the inequalities (14) and (16), we have the following.

$$\left| R(\mathbb{P}_n', h) + \alpha_n \|\nabla_z h\|_{\mathbb{P}_n',p^*} - R_{\alpha_n,p}^\mathrm{worst}(\mathbb{P}_n, h) \right| = O_p(\beta_n + \alpha_n^{1+k}).$$

This concludes the proof. □

Remark 6. *The inequality (15) shows that $\|\nabla_z h\|_{\mathbb{P}_n,1}$ has a lower bound with high probability. To appropriately use the result of Theorem 4 to Theorems 5 and 6, we need a uniform bound result of $\|\nabla_z h\|_{\mathbb{P}_n,1}$. Note that the inequality (15) does not hold when the loss h depends on data. We use the same $\mathcal{H}$ as in Theorems 5 and 6 and give a uniform bound result in the following proposition.*

Proposition 2. *Let $\mathcal{Z}$ be an open and bounded subset of $\mathbb{R}^d$. For constants $C_\mathrm{H}, C_\nabla, L > 0$, $k \in (0, 1]$, and $M \geq \sup_{n \in \mathbb{N}} \beta_n$, we let $\mathcal{H}$ be a uniformly bounded set of differentiable functions $h : \mathrm{Conv}(\mathcal{Z}) + \mathcal{B}(M) \to \mathbb{R}$ such that its gradient $\nabla_z h$ is (C_H, k)-Hölder continuous, $\mathbb{E}_\mathrm{data}(\|\nabla_z h\|_*) \geq C_\nabla$, and $\mathrm{Lip}(h) \leq L$. Then, for $\delta > 0$ and a large enough n, the following holds with probability at least $1 - \delta$.*

$$\|\nabla_z h\|_{\mathbb{P}_n,1} \geq \mathbb{E}_\mathrm{data}(\|\nabla_z h\|_*) - 2\sqrt{2} \left(L C_{\mathrm{H},k,2} + \frac{k}{d L C_{\mathrm{H},k,2}} \right) n^{-\frac{k}{2k+d}} - L \sqrt{\frac{2}{n} \log(\frac{2}{\delta})},$$

for some constant $C_{\mathrm{H},k,2} > 0$.

Proof. By the McDiarmid inequality (Devroye et al., 1996, pages 136-137) and symmetrization arguments (van der Vaart & Wellner, 1996, Lemma 2.3.1), for $\delta > 0$, the following holds with probability at least $1 - \delta$.

$$\sup_{h \in \mathcal{H}} \left| \|\nabla_z h\|_{\mathbb{P}_n, 1} - \mathbb{E}_{\text{data}}(\|\nabla_z h\|_*) \right| \leq 2\mathfrak{R}_n(\nabla\tilde{\mathcal{H}}) + L\sqrt{\frac{2}{n}\log(\frac{2}{\delta})},$$

where $\nabla\tilde{\mathcal{H}} := \{\|\nabla_z h\|_* \mid h \in \mathcal{H}\}$. By the assumption $\mathbb{E}_{\text{data}}(\|\nabla_z h\|_*) \geq C_\nabla$ and the fact that $L\sqrt{\frac{2}{n}\log(\frac{2}{\delta})}$ converges to zero as n increases, $\|\nabla_z h\|_{\mathbb{P}_n,1}$ is strictly greater than zero if $\mathfrak{R}_n(\nabla\tilde{\mathcal{H}})$ vanishes. Therefore, it is enough to show that $\mathfrak{R}_n(\nabla\tilde{\mathcal{H}})$ vanishes.

We denote a set of (C_{H}, k)-Hölder continuous functions by $\mathcal{G}_{\mathrm{H},k} := \{g : \mathcal{Z} \to \mathbb{R} \mid g$ is (C_{H}, k)-Hölder continuous and $\|g\|_\infty \leq L.\}$. Then for all $\left\|\nabla_z \tilde{h}\right\|_* \in \nabla\tilde{\mathcal{H}}$, $\left\|\nabla_z \tilde{h}\right\|_*$ is (C_{H}, k)-Hölder continuous because

$$\begin{aligned}\left|\left\|\nabla_z \tilde{h}(z_{[1]})\right\|_* - \left\|\nabla_z \tilde{h}(z_{[2]})\right\|_*\right| &\leq \left\|\nabla_z \tilde{h}(z_{[1]}) - \nabla_z \tilde{h}(z_{[2]})\right\|_* \\ &\leq C_{\mathrm{H}}\left\|z_{[1]} - z_{[2]}\right\|^k,\end{aligned}$$

for all $z_{[1]}, z_{[2]} \in \mathrm{Conv}(\mathcal{Z}) + \mathcal{B}(M)$. Further, because of the differentiability and Lipschitz continuity of $\tilde{h} \in \mathcal{H}$, we have $\left\|\left\|\nabla_z \tilde{h}\right\|_*\right\|_\infty \leq L$. Thus $\nabla\tilde{\mathcal{H}} \subseteq \mathcal{G}_{\mathrm{H},k}$, which implies $\mathfrak{R}_n(\nabla\tilde{\mathcal{H}}) \leq \mathfrak{R}_n(\mathcal{G}_{\mathrm{H},k})$.

For $u > 0$, let $N_u := \mathcal{N}(u, \mathcal{G}_{\mathrm{H},k}, \|\cdot\|_\infty)$ be the u-covering number of $\mathcal{G}_{\mathrm{H},k}$ with respect to $\|\cdot\|_\infty$ and let $\tilde{\mathcal{G}}_u := \{\tilde{g}_1, \ldots, \tilde{g}_{N_u}\}$ be the corresponding u-cover. For a set $\{\sigma_i\}_{i=1}^n$ of independent Rademacher random variables, for some $j \in [N_u]$,

$$\begin{aligned}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i g(z_i)\right| &\leq \frac{1}{n}\left|\sum_{i=1}^n \sigma_i \tilde{g}_j(z_i)\right| + \frac{1}{n}\left|\sum_{i=1}^n \sigma_i (g(z_i) - \tilde{g}_j(z_i))\right| \\ &\leq \frac{1}{n}\left|\sum_{i=1}^n \sigma_i \tilde{g}_j(z_i)\right| + u.\end{aligned}$$

The second inequality is due to the Cauchy–Schwarz inequality. Then by the Massart's lemma for a bounded and finite function space, we have

$$\sup_{g \in \mathcal{G}_{\mathrm{H},k}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i g(z_i)\right| \leq \sup_{\tilde{g} \in \tilde{\mathcal{G}}_u} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i \tilde{g}(z_i)\right| + u \leq L\sqrt{\frac{2\log N_u}{n}} + u.$$

Therefore,

$$\begin{aligned}\mathfrak{R}_n(\mathcal{G}_{\mathrm{H},k}) &\leq \inf_{u>0}\left\{u + L\sqrt{\frac{2\log\mathcal{N}(u, \mathcal{G}_{\mathrm{H},k}, \|\cdot\|_\infty)}{n}}\right\} \\ &\leq \inf_{u>0}\left\{u + L\sqrt{2(1 + C_{\mathrm{H},k,2})}\sqrt{\frac{u^{-d/k}}{n}}\right\} \\ &= \left(L\sqrt{2(1 + C_{\mathrm{H},k,2})}\right)^{\frac{2k}{2k+d}}\left(\left(\frac{d}{2k}\right)^{\frac{2k}{2k+d}} + \left(\frac{d}{2k}\right)^{-\frac{d}{2k+d}}\right)n^{-\frac{k}{2k+d}},\end{aligned}$$

for some constant $C_{\mathrm{H},k,2} > 0$. Here, the second inequality is due to Lorentz (1962, Theorem 2):

$$C_{\mathrm{H},k,1} \leq \lim_{u \to 0} \frac{\log\mathcal{N}(u, \mathcal{G}_{\mathrm{H},k}, \|\cdot\|_\infty)}{u^{-d/k}} \leq C_{\mathrm{H},k,2},$$

for some constant $C_{\mathrm{H},k,1} > 0$.[1] Therefore, $\mathfrak{R}_n(\mathcal{G}_{\mathrm{H},k})$ vanishes with high probability. □

[1] Lorentz (1962, Theorem 2) considers the uniform norm $\|\cdot\|_\infty$ on $\mathcal{Z}$, but any norm gives the same conclusion because any two norms are equivalent on the finite dimensional space $\mathbb{R}^d$.

A.3. Proof of Theorem 5

Proof. Let $h^{\mathrm{worst}}_{\alpha_n,p,\mathcal{H}} = \mathrm{argmin}_{h\in\mathcal{H}} R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, h)$. Since $\mathcal{Z}$ is bounded and $\mathcal{H}$ is uniformly bounded, there exist constants $D_{\mathcal{Z}}$ and $C_{\mathcal{H},\infty}$ such that $\sup_{z_1,z_2\in\mathcal{Z}}\|z_1 - z_2\| \le D_{\mathcal{Z}}$ and $\sup_{h\in\mathcal{H}}\sup_{z\in\mathcal{Z}}|h(z)| \le C_{\mathcal{H},\infty}$, respectively. As for the outline, we decompose an excess risk as follows.

$$\begin{aligned}
R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, h^{\mathrm{worst}}_{\alpha_n,p,\mathcal{H}}) &= \underbrace{R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p})}_{\text{(T1)}} \\
&+ \underbrace{R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p})}_{\text{(T2)}} \\
&+ \underbrace{R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h^{\mathrm{worst}}_{\alpha_n,p,\mathcal{H}})}_{\text{(T3)}} \\
&+ \underbrace{R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h^{\mathrm{worst}}_{\alpha_n,p,\mathcal{H}}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, h^{\mathrm{worst}}_{\alpha_n,p,\mathcal{H}})}_{\text{(T4)}}.
\end{aligned}$$

As for the term (T3), by the definition of $\hat{h}^{\mathrm{worst}}_{\alpha_n,p}$,

$$\text{(T3)} = R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h^{\mathrm{worst}}_{\alpha_n,p,\mathcal{H}}) \le 0.$$

[Step 1] In this step, we obtain an upper bound of the term (T2). By Theorem 4, for any fixed $\delta > 0$, there exists finite constants $\tilde{M}_1 > 0, \tilde{N}_1 \in \mathbb{N}$ such that the following holds with probability at least $1 - \delta/2$.[2]

$$\frac{|R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p})|}{\beta_n + \alpha_n^{1+k}} \le \tilde{M}_1, \tag{17}$$

for any $n \ge \tilde{N}_1$. Similarly, there exists finite constants $\tilde{M}_2 > 0, \tilde{N}_2 \in \mathbb{N}$ such that the following holds with probability at least $1 - \delta/2$.

$$\frac{|R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) - R^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p})|}{\beta_n + \alpha_n^{1+k}} \le \tilde{M}_2, \tag{18}$$

for any $n \ge \tilde{N}_2$. Choose $\varepsilon_n > 0$ so that $\varepsilon_n = \Theta(\log(n)(\beta_n + \alpha_n^{1+k}))$.[3] Then there exists $\tilde{N} \ge \max\{\tilde{N}_1, \tilde{N}_2\}$ such that for all $n \ge \tilde{N}$, we have $\varepsilon_n - (\tilde{M}_1 + \tilde{M}_2)(\beta_n + \alpha_n^{1+k}) > 0$. Fix such n. Under the product of the above two events (17) and (18), assume that $R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) > R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) + \varepsilon_n$. Then

$$\begin{aligned}
R^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) &\ge R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - \tilde{M}_1(\beta_n + \alpha_n^{1+k}) \\
&> R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) + \varepsilon_n - \tilde{M}_1(\beta_n + \alpha_n^{1+k}) \\
&\ge R^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) + \varepsilon_n - (\tilde{M}_1 + \tilde{M}_2)(\beta_n + \alpha_n^{1+k}) \\
&> R^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}),
\end{aligned}$$

which contradicts the definition of $\hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}$. Thus, with probability at least $1 - \delta$, we have

$$\text{(T2)} = R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) \le \varepsilon_n = \Theta(\log(n)(\beta_n + \alpha_n^{1+k})).$$

for sufficiently large n, or

$$\text{(T2)} = R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) = O(\log(n)(\beta_n + \alpha_n^{1+k})). \tag{19}$$

[2] We refer Remark 6 and Proposition 2.

[3] For positive sequences (a_n) and (b_n), $b_n = \Theta(a_n)$ indicates that there exist $C_1 > 0, C_2 > 0, n_0 \in \mathbb{N}$ such that $C_1 a_n \le b_n \le C_2 a_n$ for all $n \ge n_0$.

[Step 2] This step is based on proof of Lee & Raginsky (2018, Theorem 3). As for the term $(\mathrm{T1})$, by the inequality (C.4) and Lemma 5 of Lee & Raginsky (2018), we have

$$(\mathrm{T1}) = R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) \leq \frac{48\mathfrak{C}(\mathcal{H})}{\sqrt{n}} + \frac{48LD^p_{\mathcal{Z}}}{\sqrt{n}\alpha_n^{p-1}} + C_{\mathcal{H},\infty}\sqrt{\frac{2}{n}\log(\frac{2}{\delta})}, \tag{20}$$

with probability at least $1 - \delta/2$.

As for the term $(\mathrm{T4})$, by the inequality (C.5) of Lee & Raginsky (2018), the following holds with probability at least $1-\delta/2$.

$$(\mathrm{T4}) = R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h^{\mathrm{worst}}_{\alpha_n,p,\mathcal{H}}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, h^{\mathrm{worst}}_{\alpha_n,p,\mathcal{H}}) \leq C_{\mathcal{H},\infty}\sqrt{\frac{2}{n}\log(\frac{2}{\delta})}. \tag{21}$$

Therefore, by combining all the inequalities (19), (20), and (21), the following holds with probability at least $1 - 2\delta$,

$$\begin{aligned}
&R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, h^{\mathrm{worst}}_{\alpha_n,p,\mathcal{H}}) \\
&\leq \frac{48\mathfrak{C}(\mathcal{H})}{\sqrt{n}} + \frac{48LD^p_{\mathcal{Z}}}{\sqrt{n}\alpha_n^{p-1}} + 2C_{\mathcal{H},\infty}\sqrt{\frac{2}{n}\log(\frac{2}{\delta})} + O(\log(n)(\beta_n + \alpha_n^{1+k})) \\
&= O(n^{-1/2}(\mathfrak{C}(\mathcal{H}) + \alpha_n^{1-p}) + \log(n)(\beta_n + \alpha_n^{1+k})).
\end{aligned}$$

This concludes the proof. □

A.4. Proof of Theorem 6

Proof of Theorem 6. Let $h_{\mathcal{H}} = \mathrm{argmin}_{h\in\mathcal{H}} R(\mathbb{P}_{\mathrm{data}}, h)$. Since $\mathcal{H}$ is uniformly bounded, there exists a constant $C_{\mathcal{H},\infty}$ such that $\sup_{h\in\mathcal{H}}\sup_{z\in\mathcal{Z}}|h(z)| \leq C_{\mathcal{H},\infty}$. Now decompose the excess risk as follows.

$$\begin{aligned}
R(\mathbb{P}_{\mathrm{data}}, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R(\mathbb{P}_{\mathrm{data}}, h_{\mathcal{H}}) &= \underbrace{R(\mathbb{P}_{\mathrm{data}}, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p})}_{(\mathrm{T1})} \\
&+ \underbrace{R(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p})}_{(\mathrm{T2})} \\
&+ \underbrace{R^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p})}_{(\mathrm{T3})} \\
&+ \underbrace{R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{prop}}_{(\alpha_n,\beta_n),p}) - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p})}_{(\mathrm{T4})} \\
&+ \underbrace{R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) - R(\mathbb{P}_n, \hat{h}^{\mathrm{ERM}}_n)}_{(\mathrm{T5})} \\
&+ \underbrace{R(\mathbb{P}_n, \hat{h}^{\mathrm{ERM}}_n) - R(\mathbb{P}_{\mathrm{data}}, h_{\mathcal{H}})}_{(\mathrm{T6})}.
\end{aligned}$$

[Step 1] In this step, we obtain an upper bound of the term $(\mathrm{T5})$. For all $h \in \mathcal{H}$ and small enough α_n, we have

$$R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) \leq R(\mathbb{P}_n, h) + \mathrm{Lip}(h)\alpha_n \leq R(\mathbb{P}_n, h) + L\alpha_n. \tag{22}$$

The first inequality is due to the inequality (9), the second inequality is due to the assumption. Applying the infimum operator to the inequality (22) gives

$$R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) \leq R(\mathbb{P}_n, \hat{h}^{\mathrm{ERM}}_n) + L\alpha_n = R(\mathbb{P}_n, \hat{h}^{\mathrm{ERM}}_n) + O(\alpha_n).$$

Therefore,

$$(\mathrm{T5}) = R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\mathrm{worst}}_{\alpha_n,p}) - R(\mathbb{P}_n, \hat{h}^{\mathrm{ERM}}_n) = O(\alpha_n). \tag{23}$$

[Step 2] In this step, we obtain an upper bound for the terms (T2), (T3), and (T4). For any fixed $\delta > 0$, the following holds with probability at least $1 - \delta$.

$$R(\mathbb{P}_n, h) \leq R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) \leq R^{\text{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, h) + O(\beta_n + \alpha_n^{1+k}).$$

The first inequality is due to $\mathbb{P}_n \in \mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)$ and the second inequality is due to Theorem 4. Thus,

$$(\text{T2}) = R(\mathbb{P}_n, \hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p}) - R^{\text{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, \hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p}) = O_p(\beta_n + \alpha_n^{1+k}).$$

As for the term (T3), by Theorem 4, we have

$$(\text{T3}) = R^{\text{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, \hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p}) - R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p}) = O_p(\beta_n + \alpha_n^{1+k}).$$

As for the term (T4), the inequality (19) gives

$$(\text{T4}) = R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p}) - R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, \hat{h}^{\text{worst}}_{\alpha_n,p}) = O_p(\log(n)(\beta_n + \alpha_n^{1+k})).$$

Therefore,

$$(\text{T2}) + (\text{T3}) + (\text{T4}) = O_p(\log(n)(\beta_n + \alpha_n^{1+k})). \tag{24}$$

[Step 3] In this step, we obtain an upper bound for the terms (T1) and (T6). Note that the term (T1) is bounded by $\sup_{h\in\mathcal{H}} |R(\mathbb{P}_n, h) - R(\mathbb{P}_{\text{data}}, h)|$. As for the term (T6), we have

$$\begin{aligned}
R(\mathbb{P}_n, \hat{h}^{\text{ERM}}_n) - R(\mathbb{P}_{\text{data}}, h_{\mathcal{H}}) &= R(\mathbb{P}_n, \hat{h}^{\text{ERM}}_n) - R(\mathbb{P}_n, h_{\mathcal{H}}) + R(\mathbb{P}_n, h_{\mathcal{H}}) - R(\mathbb{P}_{\text{data}}, h_{\mathcal{H}}) \\
&\leq 0 + R(\mathbb{P}_n, h_{\mathcal{H}}) - R(\mathbb{P}_{\text{data}}, h_{\mathcal{H}}) \\
&\leq \sup_{h\in\mathcal{H}} |R(\mathbb{P}_n, h) - R(\mathbb{P}_{\text{data}}, h)|.
\end{aligned}$$

The first inequality is due to the definition of $\hat{h}^{\text{ERM}}_n$. Thus, the sum of the terms (T1) and (T6) is bounded by $2\sup_{h\in\mathcal{H}} |R(\mathbb{P}_n, h) - R(\mathbb{P}_{\text{data}}, h)|$. The McDiarmid inequality (Devroye et al., 1996, pages 136-137) and symmetrization arguments (van der Vaart & Wellner, 1996, Lemma 2.3.1) provide

$$\sup_{h\in\mathcal{H}} |R(\mathbb{P}_n, h) - R(\mathbb{P}_{\text{data}}, h)| \leq 2\mathfrak{R}_n(\mathcal{H}) + C_{\mathcal{H},\infty}\sqrt{\frac{2}{n}\log(\frac{2}{\delta})}, \tag{25}$$

with probability at least $1 - \delta$.

Lastly, by aggregating the inequalities (23), (24) and (25),

$$\begin{aligned}
&R(\mathbb{P}_{\text{data}}, \hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p}) - \inf_{h\in\mathcal{H}} R(\mathbb{P}_{\text{data}}, h) \\
&= O_p(\mathfrak{R}_n(\mathcal{H}) + n^{-1/2} + \alpha_n + \log(n)(\beta_n + \alpha_n^{1+k})).
\end{aligned}$$

This concludes the proof. □

A.5. Details for Section 3.3

We first define some notations. Let $\mathcal{X} \subseteq [-1, 1]^{d-1}$ and $\mathcal{Y} = \{\pm 1\}$ be open sets with respect to the ℓ_2-norm and the discrete norm $I(\cdot \neq 0)$, respectively. We set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\left\|(x, y)\right\| = \|x\|_2 + 4I(y \neq 0)$. Note that $\mathcal{X} \times \mathcal{Y}$ is clearly open and bounded with respect to $\left\|(x, y)\right\|$. For a matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{\tilde{d}_1 \times \tilde{d}_2}$, its Frobenius norm is defined as $\left\|\tilde{\mathbf{A}}\right\|_{\text{F}} = \sqrt{\sum_{i=1}^{\tilde{d}_2} \sum_{j=1}^{\tilde{d}_1} \tilde{\mathbf{A}}_{ij}^2}$ and the matrix ℓ_p-norm $\left\|\tilde{\mathbf{A}}\right\|_p := \sup_{\|u\|_p=1} \left\|\tilde{\mathbf{A}}u\right\|_p$ for $p \in [1, \infty]$. Now we define the space of deep neural networks. For an integer J and a set of integers $\mathbf{d} := \{d_0, \ldots, d_J\}$ such that $d_0 = d - 1$ and $d_J = 1$, we let $\mathcal{A} = \{\mathbf{A}_1, \ldots, \mathbf{A}_J\}$ be J weight matrices such that $\mathbf{A}_i \in \mathbb{R}^{d_i \times d_{i-1}}$. For a constant $\gamma > 0$ and a set of positive constants $\mathbf{M} := \{M_1, \ldots, M_J\}$, define

$$\mathcal{F}^{\mathcal{X}\times\mathcal{Y}}_{\mathbf{d},\mathbf{M},\gamma} := \{yf(x) = y\phi_J(\mathbf{A}_J\phi_{J-1}(\mathbf{A}_{J-1} \ldots \phi_1(\mathbf{A}_1 x) \ldots)) \mid \|\mathbf{A}_i\|_{\text{F}} \leq M_i, i \in [J], \gamma \leq \prod_{i\in[J]} \|\mathbf{A}_i\|_2\}$$

$$\mathcal{F}_{\mathbf{d},\mathbf{M},\gamma} := \{f(x) = \phi_J(\mathbf{A}_J\phi_{J-1}(\mathbf{A}_{J-1}\dots\phi_1(\mathbf{A}_1x)\dots)) \mid \|\mathbf{A}_i\|_{\mathrm{F}} \le M_i, i\in[J], \gamma \le \prod_{i\in[J]}\|\mathbf{A}_i\|_2\},$$

where $\phi_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ is a 1-Lipschitz activation function and satisfies $\phi_i(\mathbf{0}_{d_i}) = \mathbf{0}_{d_i}$ for all $i \in [J]$, and $\mathbf{0}_{d_i}$ is the vector of d_i zeros. Note that we omit intercepts here for notational simplicity. For $\phi_1, \dots, \phi_{J-1}$, we employ the hyperbolic tangent function and ϕ_J is the identity function.[4] Lastly, for a positive constants s, we define

$$\begin{aligned}\mathcal{F}^{\mathcal{X}\times\mathcal{Y}}_{\mathbf{d},\mathbf{M},\gamma,s} &:= \{yf(x) \in \mathcal{F}^{\mathcal{X}\times\mathcal{Y}}_{\mathbf{d},\mathbf{M},\gamma} \mid \sum_{i\in[J]}\|\mathbf{A}_i\|_0 \le s\} \\ \mathcal{F}_{\mathbf{d},\mathbf{M},\gamma,s} &:= \{f(x) \in \mathcal{F}_{\mathbf{d},\mathbf{M},\gamma} \mid \sum_{i\in[J]}\|\mathbf{A}_i\|_0 \le s\},\end{aligned} \tag{26}$$

where $\|\mathbf{A}\|_0$ is the number of non-zero entries of a matrix $\mathbf{A}$. To this ends, we will set $\mathbf{M} = \mathbf{1}_J$, the vector of J ones.

Corollary 2 (A formal statement of Corollary 1)**.** *Let $\mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}$ be a set of sparse deep neural networks, defined in* (26)*. For some constant $C_\nabla > 0$, let $\mathcal{H} = \{h(x,y) \mid h(x,y) = \log(1+\exp(-yf(x)))$ and $\mathbb{E}_{\text{data}}\left(\|\nabla_x f(x)\|_2\right) > C_\nabla$ for $f \in \mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}\}$.*[5] *Then the excess worst-case risks of $\hat{h}^{\text{prop}}_{\alpha_n,p}$ and $\hat{h}^{\text{ERM}}_n$ are*

$$\begin{aligned}\mathcal{E}^{\text{worst}}_{\alpha_n,p}(\hat{h}^{\text{prop}}_{\alpha_n,p}) &= O_p(n^{-1/2}\alpha_n^{1-p} \vee \log(n)\alpha_n^{1+k}), \\ \mathcal{E}^{\text{worst}}_{\alpha_n,p}(\hat{h}^{\text{ERM}}_n) &= O_p(n^{-1/2} \vee \alpha_n).\end{aligned}$$

Furthermore, the excess risks of $\hat{h}^{\text{prop}}_{\alpha_n,p}$ and $\hat{h}^{\text{ERM}}_n$ are

$$\begin{aligned}\mathcal{E}(\hat{h}^{\text{prop}}_{\alpha_n,p}) &= O_p(n^{-1/2} \vee \alpha_n \vee \log(n)(\alpha_n^{1+k})), \\ \mathcal{E}(\hat{h}^{\text{ERM}}_n) &= O_p(n^{-1/2}).\end{aligned}$$

Proof. [Step 1] Clearly, $\mathcal{X} \times \mathcal{Y}$ is open and bounded. In addition, the domain $\mathcal{X}$ is bounded and weights $\|\mathbf{A}_i\|_{\mathrm{F}}$ are bounded for all $i \in [J]$, for all $f \in \mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}$, we have $\sup_{x\in\mathcal{X}}|f(x)| \le C_{\mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}}$ for some constant $C_{\mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}} > 0$. In short, $\mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}$ is uniformly bounded, and $\mathcal{H}$ is uniformly bounded as well. In addition, for all $f \in \mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}$, due to the differentiability of the hyperbolic tangent function, f is twice continuously differentiable, and this implies that for all $h \in \mathcal{H}$, h is twice continuously differentiable. Uniformly boundedness of $\mathcal{A}$ and the boundedness of $\mathcal{Z}$ implies that uniformly boundedness of $\|\nabla_z h\|_*$ and the Frobenius norm of the Hessian matrix of h. This provides existence of constants C_{H} and L such that $\nabla_z h$ is $(C_{\mathrm{H}}, 1/2)$-Hölder continuous for all $h \in \mathcal{H}$ and $\mathrm{Lip}(h) \le L$.

Lastly, by the definition of the dual norm and the discrete norm,

$$\|\nabla_z h\|_* = \sup_{\|u\|\le 1}\langle\nabla_z h, u\rangle = \sup_{\|s\|_2\le 1}\langle\nabla_x h, s\rangle = \|\nabla_x h\|_2\,. \tag{27}$$

Since $\nabla_x h = \frac{\exp(-yf(x))}{1+\exp(-yf(x))}(-y)\nabla_x f(x)$, we have

$$\|\nabla_x h\|_2 = \left|\frac{1}{1+\exp(yf(x))}\right|\|\nabla_x f\|_2 \ge \frac{1}{1+\exp(C_{\mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}})}\|\nabla_x f\|_2\,.$$

Therefore, all the conditions in Theorems 2 and 3 are satisfied.

[Step 2] Since $\ell_{\log}(z) := \log(1+\exp(-z))$ is continuously differentiable on $(-2B_{\mathcal{F}}, 2B_{\mathcal{F}})$, $\ell_{\log}(z)$ is Lipschitz continuous on $[-B_{\mathcal{F}}, B_{\mathcal{F}}]$. It implies that there exists a finite Lipschitz constant. Let $L_{\log}$ be a Lipschitz constant on $[-B_{\mathcal{F}}, B_{\mathcal{F}}]$. Due to Talagrand's lemma (Mohri et al., 2018, Lemma 5.7), we have

$$\mathfrak{R}_n(\mathcal{H}) = \mathfrak{R}_n(\ell_{\log}\circ\mathcal{F}^{\mathcal{X}\times\mathcal{Y}}_{\mathbf{d},\mathbf{1}_J,\gamma,s}) \le L_{\log}\mathfrak{R}_n(\mathcal{F}^{\mathcal{X}\times\mathcal{Y}}_{\mathbf{d},\mathbf{1}_J,\gamma,s}).$$

[4] We may employ other differentiable activation functions with Lipschitz constant less than or equal to one. The differentiability of activation functions is required to satisfy the conditions of Theorem 1. However, it can be easily shown that this condition can be relaxed to hold only $\mathbb{P}_{\text{data}}$-almost surely, so that the ReLU function can be employed, by re-stating Theorem 1 with $\mathbb{P}_{\text{data}}$-almost sure conditions. Here, for the sake of simplicity, we simply use the hyperbolic tangent function, which is differentiable.

[5] The sufficient condition for $\mathbb{E}_{\text{data}}(\|\nabla_x f(x)\|_2) > C_\nabla$ may not be obvious, but it is assumed to be held based on Figures 2 and 3.

Due to Lemma 7 below, we have

$$\mathfrak{R}_n(\mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}^{\mathcal{X}\times\mathcal{Y}}) \leq \mathfrak{R}_n(\mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma}^{\mathcal{X}\times\mathcal{Y}}) = \mathfrak{R}_n(\mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma}) \leq O(n^{-1/2}).$$

The equality is due to for all $i \in [J]$, $\sigma_i \overset{d}{=} \sigma_i y_i$ for the Rademacher random variables σ_i. Therefore, by Mohri et al. (2018, Theorem 11.3) and Theorem 3, $\mathcal{E}(\hat{h}_n^{\mathrm{ERM}}) = O_p(n^{-1/2})$ and $\mathcal{E}(\hat{h}_{\alpha_n,p}^{\mathrm{prop}}) = O_p(n^{-1/2} \vee \alpha_n \vee \log(n)(\alpha_n^{1+k}))$ are obtained.

[Step 3] Here we prove the excess worst-case risk bound for $\hat{h}_n^{\mathrm{ERM}}$. An essentially the same argument as (22) yields that for all $h \in \mathcal{H}$,

$$R_{\alpha_n,p}^{\mathrm{worst}}(\mathbb{P}_{\mathrm{data}}, h) \leq R(\mathbb{P}_{\mathrm{data}}, h) + \mathrm{Lip}(h)\alpha_n \leq R(\mathbb{P}_{\mathrm{data}}, h) + L\alpha_n, \tag{28}$$

Applying the infimum operator on $R(\mathbb{P}_{\mathrm{data}}, h) \leq R_{\alpha_n,p}^{\mathrm{worst}}(\mathbb{P}_{\mathrm{data}}, h)$ gives

$$\inf_{h\in\mathcal{H}} R(\mathbb{P}_{\mathrm{data}}, h) \leq \inf_{h\in\mathcal{H}} R_{\alpha_n,p}^{\mathrm{worst}}(\mathbb{P}_{\mathrm{data}}, h). \tag{29}$$

Therefore, the inequalities (28) and (29) give

$$\begin{aligned}\mathcal{E}_{\alpha_n,p}^{\mathrm{worst}}(h) &= R_{\alpha_n,p}^{\mathrm{worst}}(\mathbb{P}_{\mathrm{data}}, h) - \inf_{h\in\mathcal{H}} R_{\alpha_n,p}^{\mathrm{worst}}(\mathbb{P}_{\mathrm{data}}, h)\\ &\leq R(\mathbb{P}_{\mathrm{data}}, h) + L\alpha_n - \inf_{h\in\mathcal{H}} R(\mathbb{P}_{\mathrm{data}}, h)\\ &= \mathcal{E}(h) + L\alpha_n.\end{aligned}$$

By Theorem 3 we conclude that $\mathcal{E}_{\alpha_n,p}^{\mathrm{worst}}(\hat{h}_n^{\mathrm{ERM}}) = O_p(n^{-1/2} \vee \alpha_n)$.

[Step 4] We now prove that $\mathcal{E}_{\alpha_n,p}^{\mathrm{worst}}(\hat{h}_{\alpha_n,p}^{\mathrm{prop}}) = O_p(n^{-1/2}\alpha_n^{1-p} \vee \log(n)\alpha_n^{1+k})$. By Theorem 2, it is enough to show that $\mathfrak{C}(\mathcal{H}) := \int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{H}, \|\cdot\|_\infty)}du$ is finite.

For all $(x, y) \in \mathcal{Z}$ and $f_1, f_2 \in \mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}$, we have

$$|\ell_{\log}(yf_1(x)) - \ell_{\log}(yf_2(x))| \leq L_{\log}|yf_1(x) - yf_2(x)| = L_{\log}|f_1(x) - f_2(x)|.$$

Therefore, $\mathcal{N}(u, \mathcal{H}, \|\cdot\|_\infty) \leq \mathcal{N}(\frac{u}{L_{\log}}, \mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}, \|\cdot\|_\infty)$, and thus by Lemma 8 below we have

$$\log \mathcal{N}(\frac{u}{L_{\log}}, \mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}, \|\cdot\|_\infty) \leq (s+1)\log\left(\frac{2JV^2L_{\log}}{u}\right).$$

Therefore, an integration by substitution gives

$$\begin{aligned}\int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{H}, \|\cdot\|_\infty)}du &\leq \int_0^\infty \sqrt{\mathcal{N}(\frac{u}{L_{\log}}, \mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}, \|\cdot\|_\infty)}du\\ &= \sqrt{(s+1)}\int_0^\infty \sqrt{\log\left(\frac{2JV^2L_{\log}}{u}\right)}du\\ &= \sqrt{(s+1)}\int_0^{2JV^2L_{\log}} \sqrt{\log\left(\frac{2JV^2L_{\log}}{u}\right)}du\\ &= \sqrt{(s+1)}\int_0^\infty (4JV^2L_{\log})y^2\exp(-y^2)dy.\end{aligned}$$

Since

$$\int_0^\infty y^2\exp(-y^2)dy = -\frac{1}{2}\int_0^\infty y(-2y\exp(-y^2))dy = \frac{1}{2}\int_0^\infty \exp(-y^2)dy = \frac{\sqrt{\pi}}{4},$$

we have $\int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{H}, \|\cdot\|_\infty)} < \infty$ and this concludes the proof. □

Remark 7 (Different hypothesis spaces)**.** *In essence, the results of Corollary 2 hold if for a hypothesis space $\mathcal{F}$, the Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ is $O(n^{-1/2})$ and the entropy integral $\int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{F}, \|\cdot\|_\infty)}$ is bounded. It is well known that these conditions hold for a reproducing kernel Hilbert space and a linear hypothesis space under mild conditions.*

Remark 8 (When α_n vanishes fast)**.** *Consider the logistic regression setting, i.e., $P(Y = 1 \mid X = x) = \exp(\beta_*^T x)/(1 + \exp(\beta_*^T x))$ for some $\beta_* \in \mathbb{R}^d$. Blanchet & Murthy (2019, Theorem 1) showed that $\mathcal{W}_p(\mathbb{P}_{\rm data}, \mathbb{P}_n) \leq \frac{1}{\sqrt{n}}$ holds with high probability, under mild conditions on $\mathbb{P}_{\rm data}$. In this case, we choose $\alpha_n = (n^{1/2}\log(n))^{-\frac{1}{p+k}}$. Then the proposed excess worst-case risk bound is $\mathcal{E}^{\rm worst}_{\alpha_n,p}(\hat{h}^{\rm prop}_{\alpha_n,p}) = O_p(n^{-\frac{1+k}{2(p+k)}}\log(n)^{\frac{p-1}{p+k}})$. By setting $p = \frac{1+k^2}{1-k}$, $\mathcal{E}^{\rm worst}_{\alpha_n,p}(\hat{h}^{\rm prop}_{\alpha_n,p}) = O_p(n^{-\frac{1}{2}(1-k)}\log(n)^k)$. We can choose arbitrary small $k > 0$, and thus the convergence rate is near $O(n^{-1/2})$.*[6] *Similar results hold for the excess risk bound.*

Remark 9 (Regression)**.** *For a constant $B > 0$, we let $\mathcal{X} \times \mathcal{Y} \subseteq [-1,1]^{d-1} \times [-B, B]$ be an open set with respect to the ℓ_2-norm. We set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\|(x,y)\| = \sqrt{\|x\|_2^2 + y^2}$. We let $\mathcal{H} = \{h(x,y) \mid h(x,y) = |y - f(x)| \text{ for } f \in \mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}\}$.*[7] *Then similar results hold.*

With the notations defined in the front of this section, we quote the following two lemmas: the Rademacher complexity bound of $\mathcal{F}_{\mathbf{d},\mathbf{M},\gamma}$ by Golowich et al. (2018, Corollary 1) and the covering number bound of $\mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}$ by Schmidt-Hieber (2017, Lemma 5).

Lemma 7 (Rademacher complexity bound)**.** *Assume that $\|x\|_2 \leq C_\mathcal{X}$. Then*

$$\mathfrak{R}_n(\mathcal{F}_{\mathbf{d},\mathbf{M},\gamma}) \leq C_\mathcal{X}\left(\prod_{i=1}^J M_i\right)\min\left(\bar{\log}^{3/4}(n)\sqrt{\frac{\bar{\log}(\gamma^{-1}\prod_{i=1}^J M_i)}{\sqrt{n}}}, \sqrt{\frac{J}{n}}\right),$$

where $\bar{\log}(z) := 1 \vee \log(z)$.

Lemma 8 (Covering number bound)**.** *Let $V := \prod_{i=0}^J (d_i + 1)$, then for any $u > 0$,*

$$\log \mathcal{N}(u, \mathcal{F}_{\mathbf{d},\mathbf{1}_J,\gamma,s}, \|\cdot\|_\infty) \leq (s+1)\log\left(\frac{2JV^2}{u}\right).$$

[6] For $h : \mathcal{Z} \to \mathbb{R}$, $\nabla_z h$ is $(C_{\rm H}, k_1)$-Hölder continuous, $\sup_{h\in\mathcal{H}}\sup_{z\in\mathcal{Z}}\|\nabla_z h(z)\|_* \leq L$ and any $k_2 \leq k_1$,

$$\begin{aligned}\sup_{z,\tilde{z}\in\mathcal{Z}} \frac{\|\nabla_z h(z_1) - \nabla_z h(z_2)\|_*}{\|z_1 - z_2\|^{k_2}} &\leq \sup_{\|z-\tilde{z}\|\leq 1} \frac{\|\nabla_z h(z_1) - \nabla_z h(z_2)\|_*}{\|z_1 - z_2\|^{k_2}} + \sup_{\|z-\tilde{z}\|>1} \frac{\|\nabla_z h(z_1) - \nabla_z h(z_2)\|_*}{\|z_1 - z_2\|^{k_2}} \\ &\leq \sup_{\|z-\tilde{z}\|\leq 1} \frac{\|\nabla_z h(z_1) - \nabla_z h(z_2)\|_*}{\|z_1 - z_2\|^{k_1}} + \sup_{\|z-\tilde{z}\|>1} \frac{\|\nabla_z h(z_1) - \nabla_z h(z_2)\|_*}{\|z_1 - z_2\|^{k_2}} \\ &\leq C_{\rm H} + 2L.\end{aligned}$$

Thus $\nabla_z h$ is $(C_{\rm H} + 2L, k_2)$-Hölder continuous.

[7] Since $\nabla_z h(z) = \mathrm{Sign}(y - f(x))[\nabla_x f(x), 1]^T$, $\mathbb{E}_{\rm data}(\|\nabla_z h(z)\|_2) \geq 1 =: C_\nabla$.

B. Implementation Details

In this section, we provide implementation details including the used algorithm and hyper-parameters. Our algorithm is presented in Algorithm 1. Tensorflow implementation for experiments is available at `https://github.com/ykwon0407/wdro_local_perturbation`.

Algorithm 1 Principled learning method for WDRO when data are perturbed in classification settings

```
**Input:** training dataset $\mathcal{Z}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$, a (deep neural network) model $f_\theta$ parametrized by $\theta$, batch size $B$, hyper-parameters $\tilde{\gamma}_1, \tilde{\gamma}_2, \lambda_{\text{grad}} > 0$, optimization algorithm $\mathfrak{A}$.
Initialize parameters $\theta$ in $f_\theta$
**while** until a convergent condition is met **do**
    Sample $\{(x_{[1]}, y_{[1]}), \dots, (x_{[B]}, y_{[B]})\}$ from $\mathcal{Z}_n$
    **for** $b = 1$ **to** $B$ **do**
        **if** WDRO+MIX **then**
            Sample $\gamma$ from $\mathrm{Beta}(\tilde{\gamma}_1, \tilde{\gamma}_2)$
            $x'_{[b]} = \gamma x_{[b]} + (1-\gamma) x_{[B+1-b]}$
            $y'_{[b]} = \gamma y_{[b]} + (1-\gamma) y_{[B+1-b]}$        ▷ Mixup
        **end if**
        $h_\theta(x'_{[b]}, y'_{[b]}) =$ Cross-entropy loss $\left[y_{[b]}, f_\theta(x_{[b]})\right]$        ▷ calculate loss per observation
    **end for**
    $\mathcal{L} = B^{-1} \sum_{b=1}^{B} h_\theta(x'_{[b]}, y'_{[b]}) + \lambda_{\text{grad}} \left\| \nabla_x h_\theta(x'_{[b]}, y'_{[b]}) \right\|_2^2$        ▷ calculate the objective function
    $\theta \leftarrow \mathfrak{A}(\mathcal{L}, \theta)$        ▷ update parameters
**end while**
```

B.1. Objective function

The sample space of the CIFAR-10 and CIFAR-100 datasets can be written as $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq [-1, 1]^{3072}$ and $\mathcal{Y} = \{1, \dots, k\} \subseteq \mathbb{R}$. In this space, we define the norm by $\left\|(x, y)\right\| = \|x\|_2 + 4 \cdot I(y \neq 0)$. This gives $\left\|\nabla_z h(x', y')\right\|_* = \left\|\nabla_x h(x', y')\right\|_2$ for any $(x', y') \in \mathcal{X} \times \mathcal{Y}$, as in (27). Therefore, when $p = p^* = 2$, the penalty term in (8) is $\alpha_n \|\nabla_z h\|_{\mathbb{P}'_n, p^*} = \alpha_n \sqrt{n^{-1} \sum_{i=1}^{n} \left\|\nabla_x h(x'_i, y'_i)\right\|_2^2}$. Instead of this term, we use $\lambda_{\text{grad}} \left(n^{-1} \sum_{i=1}^{n} \left\|\nabla_x h(x'_i, y'_i)\right\|_2^2\right)$ for computational convenience.

B.2. Hyper-parameter settings

We set the penalty parameter $\lambda_{\text{grad}} = 0.004$ and the batch size $B = 64$. For MIXUP and WDRO+MIX, the interpolation with hyper-parameters $\tilde{\gamma}_1 = \tilde{\gamma}_2 = 0.5$ is applied.

For the model architecture, we use the Wide ResNet model with depth 28 and width 2 including the batch normalization and the leaky ReLU activation as in Oliver et al. (2018) and Berthelot et al. (2019). Our implementation of the model and training hyper-parameters closely matches that of Berthelot et al. (2019).

For the optimization algorithm $\mathfrak{A}$, we choose Adam optimizer with the learning rate fixed as 0.002. Instead of decaying the learning rate, we use an exponential moving average of the parameters with a decay of 0.999, and apply a weight decay of 0.02 at each update for the model as in Berthelot et al. (2019). We train the model with 100×2^{16} images.

C. Additional experiment: selection of the penalty parameter

In this section, we compare the accuracy of WDRO and WDRO+MIX with various penalty parameters λ_{grad} using the contaminated CIFAR-10 and CIFAR-100 datasets. The penalty parameters vary as 0.004, 0.016, and 0.064. The training sample size is 50000 and we apply the salt and pepper noise to 1% pixels of 10000 test images for the contaminated datasets. We train the model five times.

Table 4 compares accuracy as the penalty parameter changes. In all cases, a significantly higher accuracy is attained when $\lambda_{\text{grad}} = 0.016$ than other λ_{grad} values. With this result, we anticipate that our proposed methods can achieve higher accuracy than the one in Section 5, by carefully selecting the penalty parameter λ_{grad}.

Table 4. Accuracy comparison WDRO and WDRO+MIX with various penalty parameter λ_{grad}. Other details are given in Table 2.

METHODS	λ_{grad}		
	0.004	0.016	0.064
CIFAR-10			
WDRO	87.4 ± 0.4	$\mathbf{87.9 \pm 0.2}$	86.2 ± 0.2
WDRO+MIX	87.3 ± 0.4	$\mathbf{88.2 \pm 0.3}$	86.8 ± 0.2
CIFAR-100			
WDRO	62.1 ± 0.4	$\mathbf{64.1 \pm 0.3}$	62.6 ± 0.4
WDRO+MIX	60.6 ± 0.7	$\mathbf{62.2 \pm 0.2}$	61.3 ± 0.2

References

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5050–5060, 2019.

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

Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 1996.

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

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299, 2018.

Lee, J. and Raginsky, M. Minimax statistical learning with Wasserstein distances. In *Advances in Neural Information Processing Systems*, pp. 2687–2696, 2018.

Lorentz, G. Metric entropy, widths, and superpositions of functions. *The American Mathematical Monthly*, 69(6):469–485, 1962.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 3235–3246, 2018.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.

van der Vaart, A. W. and Wellner, J. A. *Weak convergence and empirical processes*. Springer, 1996.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.