Rademacher Complexity for Adversarially Robust Generalization
Supplementary Material

Dong Yin *[1], Kannan Ramchandran †[1], and Peter Bartlett ‡[1,2]

[1]Department of Electrical Engineering and Computer Sciences, UC Berkeley
[2]Department of Statistics, UC Berkeley

A Additional Related Work

A recent line of work analyzes the convergence and generalization problems in distributional robust optimization (DRO) [5, 14, 19]. The notion of DRO differs from ours, since DRO considers the setting where the distribution of the input data is being perturbed, while we consider the perturbation in the feature space. Farnia et al. [6] study the generalization problem when the attack algorithm of the adversary is provided to the learner, which is also a weaker notion than our problem.

A few other lines of work have been trying to conduct theoretical analysis of adversarial examples. Wang et al. [21] analyze the adversarial robustness of nearest neighbors estimator. Papernot et al. [17] try to demonstrate the unavoidable trade-offs between accuracy in the natural setting and the resilience to adversarial attacks, and this trade-off is further studied by Tsipras et al. [20] through some constructive examples of distributions. Fawzi et al. [7] analyze adversarial robustness of fixed classifiers, in contrast to our generalization analysis. Fawzi et al. [8] construct examples of distributions with large latent variable space such that adversarially robust classifiers do not exist; here we argue that these examples may not explain the fact that adversarially perturbed images can usually be recognized by humans. Bubeck et al. [3] try to explain the hardness of learning an adversarially robust model from the computational constraints under the statistical query model. Another recent line of work explains the existence of adversarial examples via high dimensional geometry and concentration of measure [9, 4, 15]. These works provide examples where adversarial examples provably exist as long as the test error of a classifier is non-zero.

Our results show that adding ℓ_1 constraints on the weights of neural networks can improve the generalization gap in the adversarial setting. This is consistent with some recent works which show that sparsified neural networks may improve adversarial robustness [11, 10].

In earlier work, Bagnell proposed a concept of robust supervised learning [1]; robust optimization has been studied in Lasso [23] and SVM [24] problems. Xu and Mannor [22] make the connection between algorithmic robustness and generalization property in the natural setting, whereas our work focus on generalization in the adversarial setting.

B Proof of Theorem 2

First, we have

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}) := \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W}\sum_{i=1}^{n}\sigma_i\langle \mathbf{w}, \mathbf{x}_i\rangle\right] = \frac{W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\|_q\right]. \tag{1}$$

[*] dongyin@berkeley.edu
[†] kannanr@berkeley.edu
[‡] peter@berkeley.edu

We then analyze $\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}})$. Define $\widetilde{f}_{\mathbf{w}}(\mathbf{x}, y) := \min_{\mathbf{x}' \in \mathbb{B}^{\infty}_{\mathbf{x}}(\epsilon)} y\langle \mathbf{w}, \mathbf{x}'\rangle$. Then, we have

$$\widetilde{f}_{\mathbf{w}}(\mathbf{x}, y) = \begin{cases} \min_{\mathbf{x}' \in \mathbb{B}^{\infty}_{\mathbf{x}}(\epsilon)} \langle \mathbf{w}, \mathbf{x}'\rangle & y = 1, \\ -\max_{\mathbf{x}' \in \mathbb{B}^{\infty}_{\mathbf{x}}(\epsilon)} \langle \mathbf{w}, \mathbf{x}'\rangle & y = -1. \end{cases}$$

When $y = 1$, we have

$$\begin{aligned} \widetilde{f}_{\mathbf{w}}(\mathbf{x}, y) = \widetilde{f}_{\mathbf{w}}(\mathbf{x}, 1) &= \min_{\mathbf{x}' \in \mathbb{B}^{\infty}_{\mathbf{x}}(\epsilon)} \langle \mathbf{w}, \mathbf{x}'\rangle = \min_{\mathbf{x}' \in \mathbb{B}^{\infty}_{\mathbf{x}}(\epsilon)} \sum_{i=1}^{d} w_i x_i' \\ &= \sum_{i=1}^{d} w_i \left[\mathbb{1}(w_i \geq 0)(x_i - \epsilon) + \mathbb{1}(w_i < 0)(x_i + \epsilon)\right] = \sum_{i=1}^{d} w_i(x_i - \operatorname{sgn}(w_i)\epsilon) \\ &= \langle \mathbf{w}, \mathbf{x}\rangle - \epsilon\|\mathbf{w}\|_1. \end{aligned}$$

Similarly, when $y = -1$, we have

$$\begin{aligned} \widetilde{f}_{\mathbf{w}}(\mathbf{x}, y) = \widetilde{f}_{\mathbf{w}}(\mathbf{x}, -1) &= -\max_{\mathbf{x}' \in \mathbb{B}^{\infty}_{\mathbf{x}}(\epsilon)} \langle \mathbf{w}, \mathbf{x}'\rangle = -\max_{\mathbf{x}' \in \mathbb{B}^{\infty}_{\mathbf{x}}(\epsilon)} \sum_{i=1}^{d} w_i x_i' \\ &= -\sum_{i=1}^{d} w_i \left[\mathbb{1}(w_i \geq 0)(x_i + \epsilon) + \mathbb{1}(w_i < 0)(x_i - \epsilon)\right] = -\sum_{i=1}^{d} w_i(x_i + \operatorname{sgn}(w_i)\epsilon) \\ &= -\langle \mathbf{w}, \mathbf{x}\rangle - \epsilon\|\mathbf{w}\|_1. \end{aligned}$$

Thus, we conclude that $\widetilde{f}_{\mathbf{w}}(\mathbf{x}, y) = y\langle \mathbf{w}, \mathbf{x}\rangle - \epsilon\|\mathbf{w}\|_1$, and therefore

$$\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_2 \leq W} \sum_{i=1}^{n} \sigma_i (y_i \langle \mathbf{w}, \mathbf{x}_i\rangle - \epsilon\|\mathbf{w}\|_1)\right].$$

Define $\mathbf{u} := \sum_{i=1}^{n} \sigma_i y_i \mathbf{x}_i$ and $v := \epsilon \sum_{i=1}^{n} \sigma_i$. Then we have

$$\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \leq W} \langle \mathbf{w}, \mathbf{u}\rangle - v\|\mathbf{w}\|_1\right]$$

Since the supremum of $\langle \mathbf{w}, \mathbf{u}\rangle - v\|\mathbf{w}\|_1$ over $\mathbf{w}$ can only be achieved when $\operatorname{sgn}(w_i) = \operatorname{sgn}(u_i)$, we know that

$$\begin{aligned} \mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) =& \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \leq W} \langle \mathbf{w}, \mathbf{u}\rangle - v\langle \mathbf{w}, \operatorname{sgn}(\mathbf{w})\rangle\right] \\ =& \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \leq W} \langle \mathbf{w}, \mathbf{u}\rangle - v\langle \mathbf{w}, \operatorname{sgn}(\mathbf{u})\rangle\right] \\ =& \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \leq W} \langle \mathbf{w}, \mathbf{u} - v\operatorname{sgn}(\mathbf{u})\rangle\right] \\ =& \frac{W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\|\mathbf{u} - v\operatorname{sgn}(\mathbf{u})\|_q\right] \\ =& \frac{W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n} \sigma_i y_i \mathbf{x}_i - \big(\epsilon\sum_{i=1}^{n}\sigma_i\big)\operatorname{sgn}(\sum_{i=1}^{n}\sigma_i y_i \mathbf{x}_i)\right\|_q\right]. \end{aligned} \tag{2}$$

Now we prove an upper bound for $\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}})$. By triangle inequality, we have

$$\begin{aligned} \mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) \leq& \frac{W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i y_i \mathbf{x}_i\right\|_q\right] + \frac{\epsilon W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|(\sum_{i=1}^{n}\sigma_i)\operatorname{sgn}(\sum_{i=1}^{n}\sigma_i y_i \mathbf{x}_i)\right\|_q\right] \\ =& \mathfrak{R}_{\mathcal{S}}(\mathcal{F}) + \epsilon W \frac{d^{\frac{1}{q}}}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left|\sum_{i=1}^{n}\sigma_i\right|\right] \\ \leq& \mathfrak{R}_{\mathcal{S}}(\mathcal{F}) + \epsilon W \frac{d^{\frac{1}{q}}}{\sqrt{n}}, \end{aligned}$$

where the last step is due to Khintchine's inequality.

We then proceed to prove a lower bound for $\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}})$. According to (2) and by symmetry, we know that

$$\begin{aligned}\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) =&\frac{W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}(-\sigma_i)y_i\mathbf{x}_i - \big(\epsilon\sum_{i=1}^{n}(-\sigma_i)\big)\operatorname{sgn}(\sum_{i=1}^{n}(-\sigma_i)y_i\mathbf{x}_i)\right\|_q\right] \\ =&\frac{W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i y_i\mathbf{x}_i + \big(\epsilon\sum_{i=1}^{n}\sigma_i\big)\operatorname{sgn}(\sum_{i=1}^{n}\sigma_i y_i\mathbf{x}_i)\right\|_q\right]. \end{aligned} \tag{3}$$

Then, combining (2) and (3) and using triangle inequality, we have

$$\begin{aligned}\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) =&\frac{W}{2n}\mathbb{E}_{\boldsymbol{\sigma}}\Bigg[\left\|\sum_{i=1}^{n}\sigma_i y_i\mathbf{x}_i - \big(\epsilon\sum_{i=1}^{n}\sigma_i\big)\operatorname{sgn}(\sum_{i=1}^{n}\sigma_i y_i\mathbf{x}_i)\right\|_q \\ &+\left\|\sum_{i=1}^{n}\sigma_i y_i\mathbf{x}_i + \big(\epsilon\sum_{i=1}^{n}\sigma_i\big)\operatorname{sgn}(\sum_{i=1}^{n}\sigma_i y_i\mathbf{x}_i)\right\|_q\Bigg] \\ \geq&\frac{W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i y_i\mathbf{x}_i\right\|_q\right] = \mathfrak{R}_{\mathcal{S}}(\mathcal{F}). \end{aligned} \tag{4}$$

Similarly, we have

$$\begin{aligned}\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) \geq&\frac{W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\big(\epsilon\sum_{i=1}^{n}\sigma_i\big)\operatorname{sgn}(\sum_{i=1}^{n}\sigma_i y_i\mathbf{x}_i)\right\|_q\right] \\ =&\frac{W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\epsilon\left|\sum_{i=1}^{n}\sigma_i\right|\left\|\operatorname{sgn}(\sum_{i=1}^{n}\sigma_i y_i\mathbf{x}_i)\right\|_q\right] \\ =&\epsilon W\frac{d^{\frac{1}{q}}}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left|\sum_{i=1}^{n}\sigma_i\right|\right]. \end{aligned}$$

By Khintchine's inequality, we know that there exists a universal constant $c > 0$ such that $\mathbb{E}_{\boldsymbol{\sigma}}[|\sum_{i=1}^{n}\sigma_i|] \geq c\sqrt{n}$. Therefore, we have $\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) \geq c\epsilon W\frac{d^{\frac{1}{q}}}{\sqrt{n}}$. Combining with (4), we complete the proof.

C Multi-class Linear Classifiers

C.1 Proof of Theorem 3

According to the multi-class margin bound in [12], for any fixed γ, with probability at least $1-\delta$, we have

$$\begin{aligned}\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left\{y \neq \arg\max_{y'\in[K]}[f(\mathbf{x})]_{y'}\right\} \leq&\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}([f(\mathbf{x}_i)]_{y_i} \leq \gamma + \max_{y'\neq y}[f(\mathbf{x}_i)]_{y'}) \\ &+\frac{4K}{\gamma}\mathfrak{R}_{\mathcal{S}}(\Pi_1(\mathcal{F})) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}},\end{aligned}$$

where $\Pi_1(\mathcal{F}) \subseteq \mathbb{R}^{\mathcal{X}}$ is defined as

$$\Pi_1(\mathcal{F}) := \{\mathbf{x} \mapsto [f(\mathbf{x})]_k : f \in \mathcal{F}, k \in [K]\}.$$

In the special case of linear classifiers $\mathcal{F} = \{f_{\mathbf{W}}(\mathbf{x}) : \|\mathbf{W}^\top\|_{p,\infty} \leq W\}$, we can see that

$$\Pi_1(\mathcal{F}) = \{\mathbf{x} \mapsto \langle\mathbf{w},\mathbf{x}\rangle : \|\mathbf{w}\|_p \leq W\}.$$

Thus, we have

$$\mathfrak{R}_{\mathcal{S}}(\Pi_1(\mathcal{F})) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\|_q\right],$$

which completes the proof.

C.2 Proof of Theorem 4

Since the loss function in the adversarial setting is

$$\widetilde{\ell}(f_{\mathbf{W}}(\mathbf{x}),y) = \max_{\mathbf{x}'\in\mathbb{B}_{\mathbf{x}}^{\infty}(\epsilon)}\phi_\gamma(M(f_{\mathbf{W}}(\mathbf{x}),y)) = \phi_\gamma(\min_{\mathbf{x}'\in\mathbb{B}_{\mathbf{x}}^{\infty}(\epsilon)}M(f_{\mathbf{W}}(\mathbf{x}),y)).$$

Since we consider linear classifiers, we have

$$\begin{aligned}\min_{\mathbf{x}'\in\mathbb{B}_{\mathbf{x}}^{\infty}(\epsilon)}M(f_{\mathbf{W}}(\mathbf{x}),y) &= \min_{\mathbf{x}'\in\mathbb{B}_{\mathbf{x}}^{\infty}(\epsilon)}\min_{y'\neq y}(\mathbf{w}_y-\mathbf{w}_{y'})^\top\mathbf{x}' \\ &= \min_{y'\neq y}\min_{\mathbf{x}'\in\mathbb{B}_{\mathbf{x}}^{\infty}(\epsilon)}(\mathbf{w}_y-\mathbf{w}_{y'})^\top\mathbf{x}' \\ &= \min_{y'\neq y}(\mathbf{w}_y-\mathbf{w}_{y'})^\top\mathbf{x}-\epsilon\|\mathbf{w}_y-\mathbf{w}_{y'}\|_1 \qquad (5)\end{aligned}$$

Define

$$h_{\mathbf{W}}^{(k)}(\mathbf{x},y) := (\mathbf{w}_y-\mathbf{w}_k)^\top\mathbf{x}-\epsilon\|\mathbf{w}_y-\mathbf{w}_k\|_1+\gamma\mathbb{1}(y=k).$$

We now show that

$$\widetilde{\ell}(f_{\mathbf{W}}(\mathbf{x}),y) = \max_{k\in[K]}\phi_\gamma(h_{\mathbf{W}}^{(k)}(\mathbf{x},y)). \qquad (6)$$

To see this, we can see that according to (5),

$$\min_{\mathbf{x}'\in\mathbb{B}_{\mathbf{x}}^{\infty}(\epsilon)}M(f_{\mathbf{W}}(\mathbf{x}),y) = \min_{k\neq y}h_{\mathbf{W}}^{(k)}(\mathbf{x},y).$$

If $\min_{k\neq y}h_{\mathbf{W}}^{(k)}(\mathbf{x},y)\leq\gamma$, we have $\min_{k\neq y}h_{\mathbf{W}}^{(k)}(\mathbf{x},y) = \min_{k\in[K]}h_{\mathbf{W}}^{(k)}(\mathbf{x},y)$, since $h_{\mathbf{W}}^{(y)}(\mathbf{x},y)=\gamma$. On the other hand, if $\min_{k\neq y}h_{\mathbf{W}}^{(k)}(\mathbf{x},y)>\gamma$, then $\min_{k\in[K]}h_{\mathbf{W}}^{(k)}(\mathbf{x},y)=\gamma$. In this case, we have $\phi_\gamma(\min_{k\neq y}h_{\mathbf{W}}^{(k)}(\mathbf{x},y)) = \phi_\gamma(\min_{k\in[K]}h_{\mathbf{W}}^{(k)}(\mathbf{x},y))=0$. Therefore, we can see that (6) holds.

Define the K function classes $\mathcal{F}_k := \{h_{\mathbf{W}}^{(k)}(\mathbf{x},y) : \|\mathbf{W}^\top\|_{p,\infty}\leq W\}\subseteq\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$. Since $\phi_\gamma(\cdot)$ is $1/\gamma$-Lipschitz, according to the Ledoux-Talagrand contraction inequality [13] and Lemma 8.1 in [16], we have

$$\mathfrak{R}_{\mathcal{S}}(\widetilde{\ell}_{\mathcal{F}})\leq\frac{1}{\gamma}\sum_{k=1}^{K}\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_k). \qquad (7)$$

We proceed to analyze $\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_k)$. The basic idea is similar to the proof of Theorem 2. We define

$\mathbf{u}_y = \sum_{i=1}^n \sigma_i \mathbf{x}_i \mathbb{1}(y_i = y)$ and $v_y = \sum_{i=1}^n \sigma_i \mathbb{1}(y_i = y)$. Then, we have

$$
\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_k) =& \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\|\mathbf{W}^\top\|_{p,\infty}\le W} \sum_{i=1}^n \sigma_i((\mathbf{w}_{y_i}-\mathbf{w}_k)^\top \mathbf{x}_i - \epsilon\|\mathbf{w}_{y_i}-\mathbf{w}_k\|_1 + \gamma\mathbb{1}(y_i = k))\Big] \\
=& \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\|\mathbf{W}^\top\|_{p,\infty}\le W} \sum_{i=1}^n \sum_{y=1}^K \sigma_i((\mathbf{w}_{y_i}-\mathbf{w}_k)^\top \mathbf{x}_i - \epsilon\|\mathbf{w}_{y_i}-\mathbf{w}_k\|_1 + \gamma\mathbb{1}(y_i = k))\mathbb{1}(y_i = y)\Big] \\
=& \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{\|\mathbf{W}^\top\|_{p,\infty}\le W} \sum_{y=1}^K \sum_{i=1}^n \sigma_i((\mathbf{w}_{y}-\mathbf{w}_k)^\top \mathbf{x}_i\mathbb{1}(y_i = y) - \epsilon\|\mathbf{w}_{y}-\mathbf{w}_k\|_1\mathbb{1}(y_i = y) \\
& + \gamma\mathbb{1}(y_i = k)\mathbb{1}(y_i = y))\Big] \\
=& \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\gamma\sum_{i=1}^n \sigma_i\mathbb{1}(y_i = k) + \sup_{\|\mathbf{W}^\top\|_{p,\infty}\le W} \sum_{y\neq k}(\langle \mathbf{w}_y - \mathbf{w}_k, \mathbf{u}_y\rangle - \epsilon v_y\|\mathbf{w}_y - \mathbf{w}_k\|_1)\Big] \\
\le& \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sum_{y\neq k} \sup_{\|\mathbf{w}_k\|_p, \|\mathbf{w}_y\|_p \le W}(\langle \mathbf{w}_y - \mathbf{w}_k, \mathbf{u}_y\rangle - \epsilon v_y\|\mathbf{w}_y - \mathbf{w}_k\|_1)\Big] \\
=& \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sum_{y\neq k} \sup_{\|\mathbf{w}\|_p \le 2W}(\langle \mathbf{w}, \mathbf{u}_y\rangle - \epsilon v_y\|\mathbf{w}\|_1)\Big] \\
=& \frac{2W}{n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sum_{y\neq k} \|\mathbf{u}_y - \epsilon v_y \operatorname{sgn}(\mathbf{u}_y)\|_q\Big],
\end{aligned}
$$

where the last equality is due to the same derivation as in the proof of Theorem 2. Let $n_y = \sum_{i=1}^n \mathbb{1}(y_i = y)$. Then, we apply triangle inequality and Khintchine's inequality and obtain

$$
\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_k) \le \frac{2W}{n}\sum_{y\neq k}\mathbb{E}_{\boldsymbol{\sigma}}[\|\mathbf{u}_y\|_2] + \epsilon d^{\frac{1}{q}}\sqrt{n_y}.
$$

Combining with (7), we obtain

$$
\mathfrak{R}_{\mathcal{S}}(\widetilde{\ell}_{\mathcal{F}}) \le \frac{2WK}{\gamma n}(\sum_{y=1}^K \mathbb{E}_{\boldsymbol{\sigma}}[\|\mathbf{u}_y\|_2] + \epsilon d^{\frac{1}{q}}\sqrt{n_y}) \le \frac{2WK}{\gamma}\left[\frac{\epsilon\sqrt{K}d^{\frac{1}{q}}}{\sqrt{n}} + \frac{1}{n}\sum_{y=1}^K \mathbb{E}_{\boldsymbol{\sigma}}[\|\mathbf{u}_y\|_2]\right],
$$

where the last step is due to Cauchy-Schwarz inequality.

D Neural Network

D.1 Proof of Theorem 6

We first review a Rademacher complexity lower bound in [2].

Lemma 1. *[2] Define the function class*

$$
\widehat{\mathcal{F}} = \{\mathbf{x} \mapsto f_{\mathbf{W}}(\mathbf{x}) : \mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_L), \prod_{h=1}^L \|\mathbf{W}_h\|_\sigma \le r\},
$$

and $\widehat{\mathcal{F}}' = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \|\mathbf{w}\|_2 \le \frac{r}{2}\}$*. Then we have* $\widehat{\mathcal{F}}' \subseteq \widehat{\mathcal{F}}$*, and thus there exists a universal constant* $c > 0$ *such that*

$$
\mathfrak{R}_{\mathcal{S}}(\widehat{\mathcal{F}}) \ge \frac{cr}{n}\|\mathbf{X}\|_F.
$$

According to Lemma 1, in the adversarial setting, by defining

$$
\widetilde{\mathcal{F}}' = \{\mathbf{x} \mapsto \min_{\mathbf{x}'\in\mathbb{B}_{\mathbf{x}}^\infty(\epsilon)} y\langle \mathbf{w}, \mathbf{x}'\rangle : \|\mathbf{w}\|_2 \le \frac{r}{2}\} \subseteq \mathbb{R}^{\mathcal{X}\times\{-1,+1\}},
$$

we have $\widetilde{\mathcal{F}}' \subseteq \widetilde{\mathcal{F}}$. Therefore, there exists a universal constant $c > 0$ such that

$$
\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) \ge \mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}') \ge cr\left(\frac{1}{n}\|\mathbf{X}\|_F + \epsilon\sqrt{\frac{d}{n}}\right),
$$

where the last inequality is due to Theorem 2.

D.2 Proof of Lemma 1

Since $Q(\cdot,\cdot)$ is a linear function in its first argument, we have for any $y, y' \in [K]$,

$$\begin{aligned}
&\max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle Q(\mathbf{w}_{2,y'}-\mathbf{w}_{2,y},\mathbf{W}_1),\mathbf{P}\rangle\\
\leq&\max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle Q(\mathbf{w}_{2,y'},\mathbf{W}_1),\mathbf{P}\rangle+\max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle -Q(\mathbf{w}_{2,y},\mathbf{W}_1),\mathbf{P}\rangle\\
\leq&2\max_{k\in[K],z=\pm 1}\max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle zQ(\mathbf{w}_{2,k},\mathbf{W}_1),\mathbf{P}\rangle.
\end{aligned}\tag{8}$$

Then, for any $(\mathbf{x}, y)$, we have

$$\begin{aligned}
&\max_{\mathbf{x}'\in\mathbb{B}^\infty_{\mathbf{x}}(\epsilon)}\mathbb{1}(y\neq\arg\max_{y'\in[K]}[f_{\mathbf{W}}(\mathbf{x}')]_{y'})\\
\leq&\phi_\gamma(\min_{\mathbf{x}'\in\mathbb{B}^\infty_{\mathbf{x}}(\epsilon)}M(f_{\mathbf{W}}(\mathbf{x}'),y))\\
\leq&\phi_\gamma(\min_{y'\neq y}\min_{\mathbf{x}'\in\mathbb{B}^\infty_{\mathbf{x}}(\epsilon)}[f_{\mathbf{W}}(\mathbf{x}')]_y-[f_{\mathbf{W}}(\mathbf{x}')]_{y'})\\
\leq&\phi_\gamma\Big(\min_{y'\neq y}[f_{\mathbf{W}}(\mathbf{x})]_y-[f_{\mathbf{W}}(\mathbf{x})]_{y'}-\frac{\epsilon}{4}\max_{y'\neq y}\max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle Q(\mathbf{w}_{2,y'}-\mathbf{w}_{2,y},\mathbf{W}_1),\mathbf{P}\rangle\Big)\\
\leq&\phi_\gamma\Big(\min_{y'\neq y}[f_{\mathbf{W}}(\mathbf{x})]_y-[f_{\mathbf{W}}(\mathbf{x})]_{y'}-\frac{\epsilon}{2}\max_{k\in[K],z=\pm 1}\max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle zQ(\mathbf{w}_{2,k},\mathbf{W}_1),\mathbf{P}\rangle\Big)\\
\leq&\phi_\gamma\Big(M(f_{\mathbf{W}}(\mathbf{x}),y)-\frac{\epsilon}{2}\max_{k\in[K],z=\pm 1}\max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle zQ(\mathbf{w}_{2,k},\mathbf{W}_1),\mathbf{P}\rangle\Big):=\widehat{\ell}(f_{\mathbf{W}}(\mathbf{x}),y)\\
\leq&\mathbb{1}\Big(M(f_{\mathbf{W}}(\mathbf{x}),y)-\frac{\epsilon}{2}\max_{k\in[K],z=\pm 1}\max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle zQ(\mathbf{w}_{2,k},\mathbf{W}_1),\mathbf{P}\rangle\leq\gamma\Big),
\end{aligned}$$

where the first inequality is due to the property of ramp loss, the second inequality is by the definition of the margin, the third inequality is due to Theorem 7, the fourth inequality is due to (8), the fifth inequality is by the definition of the margin and the last inequality is due to the property of ramp loss.

D.3 Proof of Theorem 8

We study the Rademacher complexity of the function class

$$\widehat{\ell}_{\mathcal{F}}:=\{(\mathbf{x},y)\mapsto\widehat{\ell}(f_{\mathbf{W}}(\mathbf{x}),y):f_{\mathbf{W}}\in\mathcal{F}\}.$$

Define $M_{\mathcal{F}}:=\{(\mathbf{x},y)\mapsto M(f_{\mathbf{W}}(\mathbf{x}),y):f_{\mathbf{W}}\in\mathcal{F}\}$. Then we have

$$\mathfrak{R}_{\mathcal{S}}(\widehat{\ell}_{\mathcal{F}})\leq\frac{1}{\gamma}\Big(\mathfrak{R}_{\mathcal{S}}(M_{\mathcal{F}})+\frac{\epsilon}{2n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{f_{\mathbf{W}}\in\mathcal{F}}\sum_{i=1}^n\sigma_i\max_{k\in[K],z=\pm 1}\max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle zQ(\mathbf{w}_{2,k},\mathbf{W}_1),\mathbf{P}\rangle\Big]\Big),\tag{9}$$

where we use the Ledoux-Talagrand contraction inequality and the convexity of the supreme operation. For the first term, since we have $\|\mathbf{W}_1\|_1\leq b_1$, we have $\|\mathbf{W}_1^\top\|_{2,1}\leq b_1$. Then, we can apply the Rademacher complexity bound in [2] and obtain

$$\mathfrak{R}_{\mathcal{S}}(M_{\mathcal{F}})\leq\frac{4}{n^{3/2}}+\frac{60\log(n)\log(2d_{\max})}{n}s_1s_2\left((\frac{b_1}{s_1})^{2/3}+(\frac{b_2}{s_2})^{2/3}\right)^{3/2}\|\mathbf{X}\|_F.\tag{10}$$

Now consider the second term in (9). According to [18], we always have

$$\max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle zQ(\mathbf{w}_{2,k},\mathbf{W}_1),\mathbf{P}\rangle\geq 0.\tag{11}$$

In addition, we know that when $\mathbf{P}\succeq 0$ and $\mathrm{diag}(\mathbf{P})\leq 1$, we have

$$\|\mathbf{P}\|_\infty\leq 1.\tag{12}$$

Moreover, we have

$$\|\mathbf{W}_2\|_\infty\leq\|\mathbf{W}_2^\top\|_{2,1}\leq b_2.\tag{13}$$

Then, we obtain

$$\begin{aligned}
&\frac{\epsilon}{2n}\mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{f_{\mathbf{W}}\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i \max_{k\in[K],z=\pm1}\ \max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle zQ(\mathbf{w}_{2,k},\mathbf{W}_1),\mathbf{P}\rangle\Big]\\
\leq&\frac{\epsilon}{2n}\Big(\sup_{f_{\mathbf{W}}\in\mathcal{F}}\max_{k\in[K],z=\pm1}\ \max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle zQ(\mathbf{w}_{2,k},\mathbf{W}_1),\mathbf{P}\rangle\Big)\mathbb{E}_{\boldsymbol{\sigma}}\Big[\big|\sum_{i=1}^{n}\sigma_i\big|\Big]\\
\leq&\frac{\epsilon}{2\sqrt{n}}\sup_{f_{\mathbf{W}}\in\mathcal{F}}\max_{k\in[K],z=\pm1}\ \max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\langle zQ(\mathbf{w}_{2,k},\mathbf{W}_1),\mathbf{P}\rangle\\
\leq&\frac{\epsilon}{2\sqrt{n}}\sup_{f_{\mathbf{W}}\in\mathcal{F}}\max_{k\in[K],z=\pm1}\ \max_{\mathbf{P}\succeq 0,\mathrm{diag}(\mathbf{P})\leq 1}\|zQ(\mathbf{w}_{2,k},\mathbf{W}_1)\|_1\|\mathbf{P}\|_\infty\\
\leq&\frac{2\epsilon}{\sqrt{n}}\sup_{f_{\mathbf{W}}\in\mathcal{F}}\max_{k\in[K]}\|\,\mathrm{diag}(\mathbf{w}_{2,k})^\top\mathbf{W}_1\|_1\\
\leq&\frac{2\epsilon}{\sqrt{n}}\sup_{f_{\mathbf{W}}\in\mathcal{F}}\|\mathbf{W}_1\|_1\|\mathbf{W}_2\|_\infty\\
\leq&\frac{2\epsilon b_1 b_2}{\sqrt{n}},
\end{aligned}\tag{14}$$

where the first inequality is due to (11), the second inequality is due to Khintchine's inequality, the third inequality is due to Hölder's inequality, and the fourth inequality is due to the definition of $Q(\cdot,\cdot)$ and (12), the fifth inequality is a direct upper bound, and the last inequality is due to (13).

Now we can combine (10) and (14) and get an upper bound for $\mathfrak{R}_{\mathcal{S}}(\widehat{\ell}_{\mathcal{F}})$ in (9). Then, Theorem 8 is a direct consequence of Theorem 1 and Lemma 1.

References

[1] J Andrew Bagnell. Robust supervised learning. In *Proceedings of the national conference on artificial intelligence*, volume 20, page 714. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[2] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

[3] Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

[4] Elvis Dohmatob. Limitations of adversarial robustness: strong no free lunch theorem. *arXiv preprint arXiv:1810.04065*, 2018.

[5] Farzan Farnia and David Tse. A minimax approach to supervised learning. In *Advances in Neural Information Processing Systems*, pages 4240–4248, 2016.

[6] Farzan Farnia, Jesse M Zhang, and David Tse. Generalizable adversarial training via spectral normalization. *arXiv preprint arXiv:1811.07457*, 2018.

[7] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, 2016.

[8] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *arXiv preprint arXiv:1802.08686*, 2018.

[9] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

[10] Soorya Gopalakrishnan, Zhinus Marzi, Upamanyu Madhow, and Ramtin Pedarsani. Toward robust neural networks via sparsification. *arXiv preprint arXiv:1810.10625*, 2018.

[11] Yiwen Guo, Chao Zhang, Changshui Zhang, and Yurong Chen. Sparse DNNs with improved adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 242–251, 2018.

[12] Vitaly Kuznetsov, Mehryar Mohri, and U Syed. Rademacher complexity margin bounds for learning with a large number of classes. In *ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels*, 2015.

[13] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[14] Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 2687–2696, 2018.

[15] Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *arXiv preprint arXiv:1809.03063*, 2018.

[16] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[17] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.

[18] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

[19] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[20] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[21] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. *arXiv preprint arXiv:1706.03922*, 2017.

[22] Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

[23] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and Lasso. In *Advances in Neural Information Processing Systems*, 2009.

[24] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(Jul):1485–1510, 2009.