A. The Rademacher Complexity of Linear Classes [Proof of Theorem 3]

In this section, we provide a proof of Theorem 3 and present improved bounds for the Rademacher complexity of linear hypotheses. We will analyze each of the three sub-cases namely, $p \in (1, 2]$, $p > 1$, and $p = 1$ separately in the subsections that follow. Recall that the group norm $\| \cdot \|_{p_1, p_2}$ of matrix $\mathbf{X}$ is defined by

$$\|\mathbf{X}\|_{p_1,p_2} = \|(\|\mathbf{x}_1\|_{p_1}, \cdots, \|\mathbf{x}_m\|_{p_1})\|_{p_2},$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are the columns of $\mathbf{X}$. For $p_1, p_2 \leq \infty$, this group-norm can be rewritten as follows:

$$\|\mathbf{X}\|_{p_1,p_2} = \left[\sum_{i=1}^{m} \left(\sum_{j=1}^{d} |X_{j,i}|^{p_1} \right)^{\frac{p_2}{p_1}} \right]^{\frac{1}{p_2}}.$$

A.1. Case $p \in (1, 2]$

For convenience, we will use the shorthand $\mathbf{u}_{\boldsymbol{\sigma}} = \sum_{i=1}^{m} \sigma_i \mathbf{x}_i$. By definition of the dual norm, we can write:

$$\begin{aligned}
\mathfrak{R}_S(\mathcal{F}_p) &= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\|\mathbf{w}\|_p \leq W} \mathbf{w} \cdot \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right] \\
&= \frac{W}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\|\mathbf{u}_{\boldsymbol{\sigma}}\|_{p^*} \right] && \text{(dual norm property)} \\
&\leq \frac{W}{m} \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\|\mathbf{u}_{\boldsymbol{\sigma}}\|_{p^*}^2 \right]}. && \text{(Jensen's inequality)}
\end{aligned}$$

Now, for $p^* \geq 2$, $\Psi \colon \mathbf{u} \mapsto \frac{1}{2}\|\mathbf{u}\|_{p^*}^2$ is $(p^* - 1)$-smooth with respect to $\| \cdot \|_{p^*}$, that is, the following inequality holds for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$\Psi(\mathbf{y}) \leq \Psi(\mathbf{x}) + \nabla\Psi(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{p^* - 1}{2} \|\mathbf{y} - \mathbf{x}\|_{p^*}^2$$

In view of that, by successively applying the $(p^* - 1)$-smoothness inequality, we can write:

$$2\Psi(\mathbf{u}_\sigma) \leq 2 \sum_{k=1}^{m} \left\langle \nabla\Psi\Big(\sum_{i=1}^{k-1} \sigma_i \mathbf{x}_i \Big), \sigma_k \mathbf{x}_k \right\rangle + (p^* - 1) \sum_{i=1}^{m} \|\sigma_i \mathbf{x}_i\|_{p^*}^2.$$

Conditioning on $\sigma_1, \ldots, \sigma_{k-1}$ and taking expectation gives:

$$2 \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\Psi(\mathbf{u}_\sigma)] \leq (p^* - 1) \sum_{i=1}^{m} \|\mathbf{x}_i\|_{p^*}^2.$$

Thus, the following upper bound holds for the empirical Rademacher complexity:

$$\mathfrak{R}_S(\mathcal{F}_p) \leq \frac{W}{m} \sqrt{(p^* - 1) \sum_{i=1}^{m} \|\mathbf{x}_i\|_{p^*}^2}.$$

A.2. General case $p > 1$

Here again, we use the shorthand $\mathbf{u}_{\boldsymbol{\sigma}} = \sum_{i=1}^{m} \sigma_i \mathbf{x}_i$. By definition of the dual norm, we can write:

$$\begin{aligned}
\mathfrak{R}_S(\mathcal{F}_p) &= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\|\mathbf{w}\|_p \leq W} \mathbf{w} \cdot \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right] \\
&= \frac{W}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\|\mathbf{u}_{\boldsymbol{\sigma}}\|_{p^*} \right] && \text{(dual norm property)} \\
&\leq \frac{W}{m} \Big[\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\|\mathbf{u}_{\boldsymbol{\sigma}}\|_{p^*}^{p^*} \right] \Big]^{\frac{1}{p^*}}. && \text{(Jensen's inequality, } p^* \in [1, +\infty)) \\
&= \frac{W}{m} \Big[\sum_{j=1}^{d} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[|\mathbf{u}_{\boldsymbol{\sigma},j}|^{p^*} \right] \Big]^{\frac{1}{p^*}}.
\end{aligned}$$

Next, by Khintchine's inequality (Haagerup, 1981), the following holds:

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[|\mathbf{u}_{\boldsymbol{\sigma},j}|^{p^*}\right] \le B_{p^*}\left[\sum_{i=1}^{m} x_{i,j}^2\right]^{\frac{p^*}{2}},$$

where $B_{p^*} = 1$ for $p^* \in [1,2]$ and

$$B_{p^*} = 2^{\frac{p^*}{2}}\frac{\Gamma\left(\frac{p^*+1}{2}\right)}{\sqrt{\pi}},$$

for $p \in [2,+\infty)$. This yields the following bound on the Rademacher complexity:

$$\mathfrak{R}_S(\mathcal{F}_p) \le \begin{cases} \frac{W}{m}\|\mathbf{X}^\top\|_{2,p^*} & \text{if } p^* \in [1,2], \\ \frac{\sqrt{2}W}{m}\left[\frac{\Gamma\left(\frac{p^*+1}{2}\right)}{\sqrt{\pi}}\right]^{\frac{1}{p^*}}\|\mathbf{X}^\top\|_{2,p^*} & \text{if } p^* \in [2,+\infty). \end{cases}$$

A.3. Case $p = 1$

The bound on the Rademacher complexity for $p = 1$ was previously known but we reproduce the proof of this theorem for completeness. We closely follow the proof given in (Mohri et al., 2018).

Proof. For any $i \in [m]$, x_{ij} denotes the jth component of $\mathbf{x}_i$.

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_1) &= \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_1 \le W} \mathbf{w}\cdot\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\right] \\
&= \frac{W}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\right\|_\infty\right] && \text{(by definition of the dual norm)} \\
&= \frac{W}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\max_{j\in[d]}\left|\sum_{i=1}^{m}\sigma_i x_{ij}\right|\right] && \text{(by definition of } \|\cdot\|_\infty) \\
&= \frac{W}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\max_{j\in[d]}\max_{s\in\{-1,+1\}} s\sum_{i=1}^{m}\sigma_i x_{ij}\right] && \text{(by definition of } |\cdot|) \\
&= \frac{W}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{z}\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i z_i\right],
\end{aligned}$$

where $\mathcal{A}$ denotes the set of d vectors $\{s(x_{1j},\ldots,x_{mj})^\top : j \in [d], s \in \{-1,+1\}\}$. For any $\mathbf{z} \in A$, we have $\|\mathbf{z}\|_2 \le \sup_{\mathbf{z}\in A}\|\mathbf{z}\|_2 = \|\mathbf{X}^\top\|_{2,\infty}$. Further, $\mathcal{A}$ contains at most $2d$ elements. Thus, by Massart's lemma (Mohri et al., 2018),

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_1) \le W\|\mathbf{X}^\top\|_{2,\infty}\frac{\sqrt{2\log(2d)}}{m},$$

which concludes the proof. □

A.4. Comparing $\|\mathrm{M}^\top\|_{p,q}$ and $\|\mathrm{M}\|_{q,p}$ [Proof of Proposition 1]

In this section, we prove Proposition 1. This proposition implies that for $p \in (1,2)$, the group norm $\|\mathbf{X}^\top\|_{2,p^*}$, is always a lower bound on the term $\|\mathbf{X}\|_{p^*,2}$. These two norms are a major component of the Rademacher complexity of linear classes.

Proof. First, (11) follows from (10) by substituting $\mathbf{M} = \mathbf{A}^\top$ for a matrix $\mathbf{A}$: For $q \le p$,

$$\min(m,d)^{\frac{1}{p}-\frac{1}{q}}\|\mathbf{A}\|_{p,q} \le \|\mathbf{A}^\top\|_{q,p} \le \|\mathbf{A}\|_{p,q}$$

which implies that

$$\|\mathbf{A}^\top\|_{q,p} \le \|\mathbf{A}\|_{p,q} \le \min(m,d)^{\frac{1}{q}-\frac{1}{p}}\|\mathbf{A}^\top\|_{q,p}$$

However, now p and q are swapped in comparison to (11). Now after swapping them again, for $p \leq q$,

$$\|\mathbf{A}^\top\|_{p,q} \leq \|\mathbf{A}\|_{q,p} \leq \min(m,d)^{\frac{1}{p}-\frac{1}{q}} \|\mathbf{A}^\top\|_{p,q}$$

The rest of this proof will be devoted to showing (10).

Next, if $p = q$, then $\|\mathbf{M}\|_{q,p} = \|\mathbf{M}^\top\|_{p,q}$. For the rest of the proof, we will assume that $q < p$. Specifically, $q < +\infty$ which allows us to consider fractions like $\frac{p}{q}$.

We will show that for $q < p$, the following inequality holds: $\|\mathbf{M}\|_{q,p} \leq \|\mathbf{M}^\top\|_{p,q}$, or equivalently, $\|\mathbf{M}\|_{q,p}^q \leq \|\mathbf{M}^\top\|_{p,q}^q$.

We will use the shorthand $\alpha = \frac{p}{q} > 1$. By definition of the group norm and using the notation $\mathbf{U}_{ij} = |\mathbf{M}_{ij}|^q$, we can write

$$\begin{aligned}\|\mathbf{M}\|_{q,p}^q = \left[\sum_{i=1}^m \left[\sum_{j=1}^d |\mathbf{M}_{ij}|^q\right]^{\frac{p}{q}}\right]^{\frac{q}{p}} = \left[\sum_{i=1}^m \left[\sum_{j=1}^d \mathbf{U}_{ij}\right]^{\alpha}\right]^{\frac{1}{\alpha}} &= \left\| \begin{bmatrix} \sum_{j=1}^d \mathbf{U}_{1j} \\ \vdots \\ \sum_{j=1}^d \mathbf{U}_{mj} \end{bmatrix} \right\|_\alpha \\ &\leq \sum_{j=1}^d \left\| \begin{bmatrix} \mathbf{U}_{1j} \\ \vdots \\ \mathbf{U}_{mj} \end{bmatrix} \right\|_\alpha = \sum_{j=1}^d \left[\sum_{i=1}^m |\mathbf{M}_{ij}|^p\right]^{\frac{q}{p}} = \|\mathbf{M}^\top\|_{p,q}^q.\end{aligned}$$

To show that this inequality is tight, note that equality holds for an all-ones matrix. Next, we prove the inequality

$$\min(m,d)^{\frac{1}{q}-\frac{1}{p}} \|\mathbf{M}^\top\|_{p,q} \leq \|\mathbf{M}\|_{q,p},$$

for $q \leq p$. Applying Lemma 1 twice gives

$$\|\mathbf{M}^\top\|_{p,q} \leq \|\mathbf{M}^\top\|_{q,q} = \|\mathbf{M}\|_{q,q} \leq d^{\frac{1}{q}-\frac{1}{p}} \|\mathbf{M}\|_{p,q}. \tag{18}$$

Again applying Lemma 1 twice gives

$$\|\mathbf{M}^\top\|_{p,q} \leq m^{\frac{1}{q}-\frac{1}{p}} \|\mathbf{M}^\top\|_{p,p} = m^{\frac{1}{q}-\frac{1}{p}} \|\mathbf{M}\|_{p,p} \leq m^{\frac{1}{q}-\frac{1}{p}} \|\mathbf{M}\|_{p,q}. \tag{19}$$

(Lemma 1 was presented in Section 3.3 and is proved in Appendix B.) Next, we show that (18) is tight if $d \leq m$ and that (19) is tight if $d \geq m$. If $d \leq m$, the bound is tight for the block matrix $\mathbf{M} = \left[\, \mathbf{I}_{d\times d} \mid \mathbf{0} \,\right]$, and, if $d \geq m$, then the bound is tight for the block matrix $\mathbf{M} = \left[\frac{\mathbf{I}_{d\times d}}{\mathbf{0}}\right]$. □

A.5. Constant Analysis

In this section, we study the constants in the two known bounds on the Rademacher complexity of linear classes for $1 < p \leq 2$. Specifically,

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) \leq \begin{cases} \dfrac{W}{m}\sqrt{p^*-1}\|\mathbf{X}\|_{p^*,2} & (20) \\[2ex] \dfrac{\sqrt{2}W}{m}\left[\dfrac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\right]^{\frac{1}{p^*}} \|\mathbf{X}^\top\|_{2,p^*} & (21) \end{cases}$$

We will compare the constants in equations (20) and (21), namely $\frac{\sqrt{2}W}{m}\left(\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\right)^{\frac{1}{p^*}}$ and $\frac{W}{m}\sqrt{p^*-1}$. Since $\frac{W}{m}$ divides both of these constants, we drop this factor and work with the expressions $c_1(p) := \sqrt{p^*-1}$ and $c_2(p) := \sqrt{2}\left(\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\right)^{\frac{1}{p^*}}$. To start, we first establish upper and lower bound on $c_2(p)$.

Lemma 3. *Let $c_2(p) = \sqrt{2}\left(\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\right)^{\frac{1}{p^*}}$. Then the following inequalities hold:*

$$e^{-\frac{1}{2}}\sqrt{p^*} \leq c_2(p) \leq e^{-\frac{1}{2}}\sqrt{p^*+1}.$$

Proof. For convenience, we set $q = p^*$, $f_1(q) = c_1(p)$, $f_2(q) = c_2(p)$. Next, we recall a useful inequality (Olver et al., 2010) bounding the gamma function:

$$1 < (2\pi)^{-\frac{1}{2}} x^{\frac{1}{2}-x} e^{x} \Gamma(x) < e^{\frac{1}{12x}}. \tag{22}$$

We start with the upper bound. If we apply the right-hand side inequality of (22) to $\Gamma(\frac{q+1}{2})$ we get the following bound on $f_2(q)$:

$$f_2(q) \leq 2^{\frac{1}{2q}} e^{-\frac{1}{2}} \sqrt{q+1} e^{-\frac{1}{2q}+\frac{1}{6(q+1)q}} \tag{23}$$

It is easy to verify that,

$$2^{\frac{1}{2q}} e^{-\frac{1}{2q}+\frac{1}{6q(q+1)}} = e^{\frac{1}{q}(\frac{\ln 2-1}{2}+\frac{1}{6q(q+1)})}. \tag{24}$$

Furthermore, the expression $(\frac{\ln 2-1}{2} + \frac{1}{6q(q+1)})$ decreases with increasing q. At $q = 2$, it is negative, which implies that (24) is less than 1 for $q \geq 2$. Hence

$$f_2(q) \leq e^{-\frac{1}{2}} \sqrt{q+1}$$

Next, we prove the lower bound. Applying the lower bound of (22) to $\Gamma(\frac{q+1}{2})$ results in

$$f_2(q) \geq e^{-\frac{1}{2}} \sqrt{q} \left(e^{-\frac{1}{2q}(\log 2-1)} \sqrt{1+\frac{1}{q}} \right).$$

We will establish that $\left(e^{-\frac{1}{2q}(\log 2-1)} \sqrt{1+\frac{1}{q}} \right) \geq 1$, which will complete the proof of the lower bound. We prove this statement by showing that

$$\left(e^{-\frac{1}{2q}(\log 2-1)} \sqrt{1+\frac{1}{q}} \right)^2 = e^{-\frac{1}{q}(\log 2-1)} \left(1+\frac{1}{q}\right) \geq 1.$$

By applying some elementary inequalities

$$\begin{aligned} e^{-\frac{1}{q}(\log 2-1)} \left(1+\frac{1}{q}\right) &\geq \left(\frac{1}{q}(\log 2-1)+1\right)\left(1+\frac{1}{q}\right) && (\text{using } e^x \geq 1+x) \\ &= 1+\frac{1}{q}\left(\log(2) - \frac{1-\log(2)}{q}\right) \\ &\geq 1 \end{aligned}$$

The last inequality follows since $\left(\log(2) - \frac{1-\log(2)}{q}\right)$ increases with q, and is positive at $q = 2$. □

Lastly, we establish our main claim that $c_2(p) \leq c_1(p)$.

Lemma 4. *Let* $c_1(p) = \sqrt{p^*-1}$ *and* $c_2(p) = \sqrt{2}\left(\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\right)^{\frac{1}{p^*}}$. *Then*

$$c_2(p) \leq c_1(p),$$

for all $1 \leq p \leq 2$.

Proof. For convenience, set $q = p^*$, $f_1(q) = c_1(p)$, and $f_2(q) = c_2(p)$. First note that $f_1(2) = f_2(2)$. Next, we claim $\frac{d}{dq} f_1(q) \geq \frac{d}{dq} f_2(q)$ for $q \geq 2$, and this implies that $c_2(p) \leq c_1(p)$ for $1 \leq p \leq 2$.

The rest of this proof is devoted to showing that $\frac{d}{dq} f_1(q) \geq \frac{d}{dq} f_2(q)$. Upon differentiating we get that $f_1'(q) = \frac{1}{2\sqrt{q-1}}$. Next, we will differentiate f_2. To start, we recall that the digamma function ψ is defined as the logarithmic derivative of the gamma function, $\psi(x) = \frac{d}{dx}(\log \Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}$.

Now we state a useful inequality (see Equation 2.2 in Alzer (1997)) bounding the digamma function, $\psi(x)$.

$$\psi(x) \leq \log(x) - \frac{1}{2x} \tag{25}$$

Now we differentiate $\ln f_2$:

$$\begin{aligned}
\frac{d}{dq}(\ln f_2(q)) &= \frac{\frac{q}{2}\psi(\frac{q+1}{2}) - (\ln(\Gamma(\frac{q+1}{2})) - \ln(\sqrt{\pi}))}{q^2} \\
&\leq \frac{\frac{q}{2}(\log(\frac{q+1}{2} - \frac{1}{q+1}) - (\ln(\Gamma(\frac{q+1}{2})) - \ln\sqrt{\pi})}{q^2} && \text{(by (25))} \\
&\leq \frac{\frac{q}{2}(\log\frac{q+1}{2} - \frac{1}{q+1}) - (\frac{1}{2}\ln 2 + \frac{q}{2}\log\frac{q+1}{2} - \frac{q+1}{2})}{q^2} && \text{(by the left-hand equality in (22))} \\
&= \frac{1}{2q} + \frac{1}{q^2}\Big(\frac{1}{2(q+1)} - \frac{1}{2}\log 2\Big) \\
&\leq \frac{1}{2q}.
\end{aligned}$$

The last line follows since we only consider $q \geq 2$ and $\frac{1}{2(q+1)} - \frac{1}{2}\ln 2 \leq 0$ in this range. Finally, the fact that $\frac{d}{dq}(\ln f_2(q)) = f_2'(q)/f_2(q)$ implies

$$\begin{aligned}
f_2'(q) &= f_2(q)\frac{d}{dq}(\ln f_2(q)) \\
&\leq \frac{1}{2q} f_2(q) && \text{(by } \frac{d}{dq}(\ln f_2(q)) \leq \frac{1}{2q}\text{)} \\
&\leq \frac{e^{-\frac{1}{2}}\sqrt{q+1}}{2q} && \text{(by applying the upper bound in Lemma 3)} \\
&= \frac{1}{2\sqrt{q-1}}\,\frac{e^{-\frac{1}{2}}\sqrt{(q+1)(q-1)}}{q} \\
&\leq e^{-\frac{1}{2}}\frac{1}{2\sqrt{q-1}} && \text{(using } q^2 - 1 \leq q^2\text{)} \\
&\leq \frac{1}{2\sqrt{q-1}} = f_1'(q) && \text{(using } e^{-\frac{1}{2}} < 1\text{).}
\end{aligned}$$

□

B. Proof of Theorem 4

In this section, we give a detailed proof of Theorem 4. We start with the following lemma that characterizes the nature of adversarial perturbations.

Lemma 5. *Let g be a nondecreasing function, $\mathbf{x}, \mathbf{w} \in \mathbb{R}^d$, and $y \in \{\pm 1\}$. Then*

$$\inf_{\|\mathbf{x}-\mathbf{x}'\|_r \leq \epsilon} yg(\mathbf{w}\cdot\mathbf{x}) = yg(\mathbf{w}\cdot\mathbf{x} - \epsilon y\|\mathbf{w}\|_{r^*})$$

Proof. First note that

$$\inf_{\|\mathbf{x}-\mathbf{x}'\|_r \leq \epsilon} yg(\mathbf{w}\cdot\mathbf{x}) = \inf_{\|\mathbf{s}\|_r \leq 1} yg(\mathbf{w}\cdot\mathbf{x} + \epsilon\mathbf{w}\cdot\mathbf{s})$$

If $y = 1$,

$$\begin{aligned}
\inf_{\|\mathbf{s}\|_r \leq 1} g(\mathbf{w}\cdot\mathbf{x} + \epsilon\mathbf{w}\cdot\mathbf{s}) &= g(\mathbf{w}\cdot\mathbf{x} + \inf_{\|\mathbf{s}\|_r \leq 1}\epsilon\mathbf{w}\cdot\mathbf{s}) && (g \text{ is nondecreasing}) \\
&= g(\mathbf{w}\cdot\mathbf{x} - \epsilon\|\mathbf{w}\|_{r^*}) && \text{(definition of dual norm)} \\
&= yg(\mathbf{w}\cdot\mathbf{x} - \epsilon y\|\mathbf{w}\|_{r^*}) && (y = 1)
\end{aligned}$$

Similarly, if $y = -1$,

$$\begin{aligned}\inf_{\|\mathbf{s}\|_r \le 1} -g(\mathbf{w}\cdot\mathbf{x} + \epsilon\mathbf{w}\cdot\mathbf{s}) &= -g(\mathbf{w}\cdot\mathbf{x} + \sup_{\|\mathbf{s}\|_r \le 1} \epsilon\mathbf{w}\cdot\mathbf{s}) && (-g \text{ is non-increasing})\\ &= -g(\mathbf{w}\cdot\mathbf{x} + \epsilon\|\mathbf{w}\|_{r^*}) && (\text{definition of dual norm})\\ &= yg(\mathbf{w}\cdot\mathbf{x} - \epsilon y\|\mathbf{w}\|_{r^*}) && (y = -1)\end{aligned}$$

□

Before proceeding to the proof of Theorem 4, we formally establish Lemma 1 and Lemma 2 from Section 3.

Proof of Lemma 1. We prove that if $p \ge r^*$, then

$$\sup_{\|\mathbf{w}\|_p \le 1} \|\mathbf{w}\|_{r^*} = d^{1-\frac{1}{r}-\frac{1}{p}}$$

and otherwise,

$$\sup_{\|\mathbf{w}\|_p \le 1} \|\mathbf{w}\|_{r^*} = 1.$$

If $p \ge r^*$, by Hölder's generalized inequality with $\frac{1}{r^*} = \frac{1}{p} + \frac{1}{s}$,

$$\sup_{\|\mathbf{w}\|_p \le 1} \|\mathbf{w}\|_{r^*} \le \sup_{\|\mathbf{w}\|_p \le 1} \|\mathbf{1}\|_s \|\mathbf{w}\|_p = \|\mathbf{1}\|_s = d^{\frac{1}{s}} = d^{\frac{1}{r^*}-\frac{1}{p}} = d^{1-\frac{1}{r}-\frac{1}{p}}.$$

Equality holds at the vector $\frac{1}{d^{\frac{1}{p}}}\mathbf{1}$, and this implies that the inequality in the line above is an equality. Now for $p \le r^*$, $\|\mathbf{w}\|_p \ge \|\mathbf{w}\|_{r^*}$, implying that $\sup_{\|\mathbf{w}\|_p \le 1} \|\mathbf{w}\|_{r^*} \le 1$. Here, equality is achieved at a unit vector $\mathbf{e}_1$. □

Proof of Lemma 2. Recall that $v_{\boldsymbol{\sigma}} = \frac{1}{m}\sum_{i=1}^m \sigma_i$. Then, in view of the symmetry $v_{-\boldsymbol{\sigma}} = -v_{\boldsymbol{\sigma}}$, we can write

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \epsilon v_{\boldsymbol{\sigma}}\|\mathbf{w}\|_{r^*}\right] = \epsilon W \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le 1} v_{\boldsymbol{\sigma}}\|\mathbf{w}\|_{r^*}\right] = \frac{\epsilon W}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le 1} |v_{\boldsymbol{\sigma}}|\|\mathbf{w}\|_{r^*}\right].$$

By Lemma 1, we have

$$\frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le 1} |v_{\boldsymbol{\sigma}}|\|\mathbf{w}\|_{r^*}\right] = \frac{1}{2}\max(d^{1-\frac{1}{p}-\frac{1}{r}}, 1)\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[|v_{\boldsymbol{\sigma}}|]. \tag{26}$$

Now, by Jensen's inequality and $\mathbb{E}[\sigma_i\sigma_j] = \mathbb{E}[\sigma_i]\mathbb{E}[\sigma_j] = 0$ for $i \ne j$, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[|\mathbf{v}_{\boldsymbol{\sigma}}|] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left|\sum_{i=1}^m \sigma_i\right|\right] \le \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left(\sum_{i=1}^m \sigma_i\right)^2\right]} = \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[m + \sum_{i \ne j}\sigma_i\sigma_j\right]} = \sqrt{m}.$$

Furthermore, by Khintchine's inequality (Haagerup, 1981), the following lower bound holds:

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left|\sum_{i=1}^m \sigma_i\right|\right] \ge \sqrt{\frac{m}{2}}.$$

Substituting these upper and the lower bounds into (26) completes the proof. □

We now proceed to prove Theorem 4. Recall from Section 3.3 that we seek to analyze

$$\begin{aligned}\mathfrak{R}_S(\widetilde{\mathcal{F}}_p) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \frac{1}{m}\sum_{i=1}^m \sigma_i \inf_{\|\mathbf{x}_i - \mathbf{x}_i'\|_r \le \epsilon} y_i\langle\mathbf{w}, \mathbf{x}_i'\rangle\right]\\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \frac{1}{m}\sum_{i=1}^m \sigma_i(y_i\langle\mathbf{w}, \mathbf{x}_i\rangle - \epsilon\|\mathbf{w}\|_{r^*})\right] && \text{[by Lemma 5]}\\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle\mathbf{w}, \mathbf{u}_{\boldsymbol{\sigma}}\rangle - \epsilon v_{\boldsymbol{\sigma}}\|\mathbf{w}\|_{r^*}\right],\end{aligned} \tag{27}$$

where we used the shorthand $\mathbf{u}_{\boldsymbol{\sigma}} = \frac{1}{m}\sum_{i=1}^{m} y_i \sigma_i \mathbf{x}_i$ and $v_{\boldsymbol{\sigma}} = \frac{1}{m}\sum_{i=1}^{m}\sigma_i$. The next two theorems give upper and lower bounds on $\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}_p)$, thereby proving Theorem 4.

Theorem 11. *Let* $\mathcal{F}_p = \{\mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{w}\rangle : \|\mathbf{w}\|_p \le W\}$ *and* $\widetilde{\mathcal{F}}_p = \{\inf_{\|\mathbf{x}'-\mathbf{x}\|_r \le \epsilon} f(\mathbf{x}') : f \in \mathcal{F}_p\}$*. Then, the following upper bound holds:*

$$\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}_p) \le \mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) + \epsilon \frac{W}{2\sqrt{m}} d^{1-\frac{1}{r}-\frac{1}{p}}$$

Proof. Using (27) and the sub-additivity of supremum we can write:

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}_p) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle \mathbf{w}, \mathbf{u}_{\sigma}\rangle - \epsilon v_{\boldsymbol{\sigma}} \|\mathbf{w}\|_{r^*}\right] \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle \mathbf{w}, \mathbf{u}_{\boldsymbol{\sigma}}\rangle\right] + \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} -\epsilon v_{\boldsymbol{\sigma}} \|\mathbf{w}\|_{r^*}\right] \\
&= \mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) + \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \epsilon v_{\boldsymbol{\sigma}} \|\mathbf{w}\|_{r^*}\right] \\
&= \mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) + \frac{1}{2}\epsilon \frac{W}{\sqrt{m}} d^{1-\frac{1}{r}-\frac{1}{p}} && \text{[by Lemma 2]},
\end{aligned}$$

which completes the proof. □

Theorem 12. *Let* $\mathcal{F}_p = \{\mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{w}\rangle : \|\mathbf{w}\|_p \le W\}$ *and* $\widetilde{\mathcal{F}}_p = \{\inf_{\|\mathbf{x}'-\mathbf{x}\|_r \le \epsilon} f(\mathbf{x}') : f \in \mathcal{F}_p\}$*. Then, the following lower bound holds:*

$$\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}_p) \ge \max\left(\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p), W\frac{\epsilon d^{1-\frac{1}{r}-\frac{1}{p}}}{2\sqrt{2m}}\right)$$

Proof. The proof involves two symmetrization arguments. Since $-\boldsymbol{\sigma}$ follows the same distribution as $\boldsymbol{\sigma}$, we have the equality

$$\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}_p) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle \mathbf{w}, \mathbf{u}_{-\sigma}\rangle - \epsilon v_{-\sigma} \|\mathbf{w}\|_{r^*}\right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} -\langle \mathbf{w}, \mathbf{u}_{\sigma}\rangle + \epsilon v_{\boldsymbol{\sigma}} \|\mathbf{w}\|_{r^*}\right]. \tag{28}$$

Similarly, $\mathbf{w}$ can be replaced with $-\mathbf{w}$, thus we have

$$\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}_p) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle \mathbf{w}, \mathbf{u}_{\boldsymbol{\sigma}}\rangle + \epsilon v_{\boldsymbol{\sigma}} \|\mathbf{w}\|_{r^*}\right]. \tag{29}$$

Averaging (27) and (29) and using the sub-additivity of the supremum gives

$$\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}_p) = \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle \mathbf{w}, \mathbf{u}_{\boldsymbol{\sigma}}\rangle - \epsilon v_{\boldsymbol{\sigma}} \|\mathbf{w}\|_{r^*}\right] + \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle \mathbf{w}, \mathbf{u}_{\boldsymbol{\sigma}}\rangle + \epsilon v_{\boldsymbol{\sigma}} \|\mathbf{w}\|_{r^*}\right] \ge \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle \mathbf{w}, \mathbf{u}_{\boldsymbol{\sigma}}\rangle\right] = W\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p).$$

Now, averaging (28) and (29), and using the sub-additivity of supremum give:

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}_p) &= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} -\langle \mathbf{w}, \mathbf{u}_{\boldsymbol{\sigma}}\rangle + \epsilon v_{\boldsymbol{\sigma}} \|\mathbf{w}\|_{r^*}\right] + \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle \mathbf{w}, \mathbf{u}_{\boldsymbol{\sigma}}\rangle + \epsilon v_{\boldsymbol{\sigma}} \|\mathbf{w}\|_{r^*}\right] \\
&\ge \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} v_{\boldsymbol{\sigma}} \|\mathbf{w}\|_{r^*}\right] \ge \frac{1}{2\sqrt{2m}}\epsilon d^{1-\frac{1}{p}-\frac{1}{r}}, && \text{[from Lemma 2]}.
\end{aligned}$$

which completes the proof. □

C. Adversarial Rademacher Complexity of ReLU

In this section, we prove upper and lower bounds on the Rademacher complexity of the ReLU unit. We will use the notation $z_+ = \max(z, 0)$, for any $z \in \mathbb{R}$. We use the family of functions $\mathcal{G}_p$ defined in (15) with the corresponding adversarial class $\widetilde{\mathcal{G}}_p$:

$$\widetilde{\mathcal{G}}_p = \Big\{(\mathbf{x}, y) \mapsto \inf_{\|\mathbf{s}\|_r \le \epsilon} y(\mathbf{w}\cdot(\mathbf{x}+\mathbf{s}))_+ \colon \|\mathbf{w}\|_p \le W, y \in \{-1, +1\}\Big\}.$$

Since $z \mapsto z_+$ is non-decreasing, by Lemma 5, $\widetilde{\mathcal{G}}_p$ can be equivalently expressed as follows:

$$\widetilde{\mathcal{G}}_p = \{(\mathbf{x}, y) \mapsto y(\mathbf{w}\cdot\mathbf{x} - \epsilon y\|\mathbf{w}\|_{r^*})_+ \colon \|\mathbf{w}\|_p \le W, y \in \{-1, 1\}\}.$$

In view of that, the adversarial Rademacher complexity of the ReLU unit can be written as follows:

$$\widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G}_p) = \mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{G}}_p) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \frac{1}{m}\sum_{i=1}^m \sigma_i y_i(\mathbf{w}\cdot\mathbf{x}_i - y_i\epsilon\|\mathbf{w}\|_{r^*})_+\right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \frac{1}{m}\sum_{i=1}^m \sigma_i(\mathbf{w}\cdot\mathbf{x}_i - y_i\epsilon\|\mathbf{w}\|_{r^*})_+\right]. \quad (30)$$

C.1. Upper Bounds

Theorem 5. *Let $\mathcal{G}_p$ the class defined in (15) and let $\mathcal{F}_p$ be the linear class as defined in (8). Then, given a sample $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, the adversarial Rademacher complexity of $\mathcal{G}_p$ can be bounded as follows:*

$$\widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G}_p) \le \mathfrak{R}_{T_\epsilon}(\mathcal{F}_p) + \epsilon\frac{W}{2\sqrt{m}}\max(1, d^{1-\frac{1}{r}-\frac{1}{p}}),$$

where $T_\epsilon = \{i \colon y_i = -1$ or $(y_i = 1$ and $\|\mathbf{x}_i\|_r > \epsilon)\}$.

Proof. Consider an index $i \in [m]$ such that $i \notin T_\epsilon$, so that $\|\mathbf{x}_i\|_r \le \epsilon$ and $y_i = 1$. Then, by Hölder's inequality, we have

$$y_i\mathbf{w}\cdot\mathbf{x}_i - y_i\epsilon\|\mathbf{w}\|_{r^*} = \|\mathbf{w}\|_{r^*}\left(\frac{\mathbf{w}}{\|\mathbf{w}\|_{r^*}}\cdot\mathbf{x}_i - \epsilon\right) \le \|\mathbf{w}\|_{r^*}(\|\mathbf{x}_i\|_r - \epsilon) \le 0,$$

and therefore $(\mathbf{w}\cdot\mathbf{x}_i - \epsilon\|\mathbf{w}\|_{r^*})_+ = 0$ for all $\mathbf{w}$ with $\|\mathbf{w}\|_p \le W$. Thus, using the expression (30), we can write:

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{G}}_p) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \frac{1}{m}\sum_{i \in T_\epsilon} \sigma_i(y_i\mathbf{w}\cdot\mathbf{x}_i - \epsilon\|\mathbf{w}\|_{r^*})_+\right] \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \frac{1}{m}\sum_{i \in T_\epsilon} \sigma_i(y_i\mathbf{w}\cdot\mathbf{x}_i - \epsilon\|\mathbf{w}\|_{r^*})\right] && \text{(1-Lipschitzness of } z \mapsto z_+) \\
&= \frac{|T_\epsilon|}{m}\mathfrak{R}_{T_\epsilon}(\widetilde{\mathcal{F}}_p) \\
&\le \mathfrak{R}_{T_\epsilon}(\mathcal{F}_p) + \epsilon\frac{W}{2\sqrt{m}}\max(1, d^{1-\frac{1}{r}-\frac{1}{p}}), && \text{(Theorem 4)}
\end{aligned}$$

which completes the proof. □

C.2. Lower Bounds

Theorem 6. *Let $\mathcal{G}_p$ be the class as defined in (15). Then it holds that*

$$\widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G}_p) \ge \frac{W}{2\sqrt{2}m}\sup_{\|\mathbf{s}\|_p = 1}\left(\sum_{i \in T_{\epsilon,\mathbf{s}}}(\langle\mathbf{s}, \mathbf{x}_i\rangle - \epsilon y_i\|\mathbf{s}\|_{r^*})^2\right)^{\frac{1}{2}}$$

where $T_{\epsilon,\mathbf{s}} = \{i \colon \langle\mathbf{s}, \mathbf{x}_i\rangle - y_i\epsilon\|\mathbf{s}\|_{r^*} > 0\}$.

Proof. By definition of the supremum, we can write:

$$\mathfrak{R}_S(\mathcal{G}_p) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \leq W} \frac{1}{m}\sum_{i=1}^m \sigma_i y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle - y_i\epsilon\|\mathbf{w}\|_{r^*})_+\right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{B\leq W\\ \|\mathbf{s}\|_p=1}} \frac{B}{m}\sum_{i=1}^m \sigma_i(\langle \mathbf{s}, \mathbf{x}_i\rangle - \epsilon y_i\|\mathbf{s}\|_{r^*})_+\right].$$

Now, for a fixed $\mathbf{s}$, it is straightforward to take the supremum over B: if the quantity $\sum_{i=1}^m \sigma_i(\langle \mathbf{s}, \mathbf{z}_i\rangle - \epsilon\|\mathbf{s}\|_{r^*})_+$ is positive, the expression is maximized by taking $B = W$; otherwise it is maximized by $B = 0$. Thus, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{B< W\\ \|\mathbf{s}\|_p=1}} \frac{B}{m}\sum_{i=1}^m \sigma_i(\langle \mathbf{s}, \mathbf{x}_i\rangle - y_i\epsilon\|\mathbf{s}\|_{r^*})_+\right] &= \frac{W}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{s}\|_p=1} \max\left(0, \sum_{i=1}^m \sigma_i(\langle \mathbf{s}, \mathbf{x}_i\rangle - \epsilon\|\mathbf{s}\|_{r^*})_+\right)\right]\\
&\geq \frac{W}{m}\sup_{\|\mathbf{s}\|_p=1}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\max\left(0, \sum_{i=1}^m \sigma_i(\langle \mathbf{s}, \mathbf{x}_i\rangle - \epsilon y_i\|\mathbf{s}\|_{r^*})_+\right)\right]\\
&= \frac{W}{2m}\sup_{\|\mathbf{s}\|_p=1}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left|\sum_{i=1}^m \sigma_i(\langle \mathbf{s}, \mathbf{x}_i\rangle - \epsilon y_i\|\mathbf{s}\|_{r^*})_+\right|\right]\\
&= \frac{W}{2m}\sup_{\|\mathbf{s}\|_p=1}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left|\sum_{i\in T_{\epsilon,\mathbf{s}}} \sigma_i(\langle \mathbf{s}, \mathbf{x}_i\rangle - y_i\epsilon\|\mathbf{s}\|_{r^*})\right|\right].
\end{aligned}$$

Next, by the Khintchine-Kahane inequality (Haagerup, 1981), the following lower bound holds:

$$\begin{aligned}
\frac{W}{2m}\sup_{\|\mathbf{s}\|_p=1}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left|\sum_{i\in T_{\epsilon,\mathbf{s}}} \sigma_i(\langle \mathbf{s}, \mathbf{x}_i\rangle - \epsilon y_i\|\mathbf{s}\|_{r^*})\right|\right] &\geq \frac{W}{2\sqrt{2}m}\sup_{\|\mathbf{s}\|_p=1}\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left(\sum_{i\in T_{\epsilon,\mathbf{s}}} \sigma_i(\langle \mathbf{s}, \mathbf{x}_i\rangle - y_i\epsilon\|\mathbf{s}\|_{r^*})\right)^2\right]\right)^{\frac{1}{2}}\\
&= \frac{W}{2\sqrt{2}m}\sup_{\|\mathbf{s}\|_p=1}\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum_{i,j\in T_{\epsilon,\mathbf{s}}} \sigma_i\sigma_j(\langle \mathbf{s}, \mathbf{x}_i\rangle - \epsilon y_i\|\mathbf{s}\|_{r^*})(\langle \mathbf{s}, \mathbf{x}_j\rangle - \epsilon y_i\|\mathbf{s}\|_{r^*})\right]\right)^{\frac{1}{2}}\\
&= \frac{W}{2\sqrt{2}m}\sup_{\|\mathbf{s}\|_p=1}\left(\sum_{i\in T_{\epsilon,\mathbf{s}}} (\langle \mathbf{s}, \mathbf{x}_i\rangle - y_i\epsilon\|\mathbf{s}\|_{r^*})^2\right)^{\frac{1}{2}},
\end{aligned}$$

which completes the proof. □

D. Adversarial Rademacher for Neural Nets with One Hidden Layer with a Lipschitz Activation Function

In this section, we present an upper bound on the adversarial Rademacher complexity of one-layer neural networks with an activation function satisfying some reasonable requirements. Our analysis uses the notion of coverings.

Definition 2 (ϵ-covering)**.** *Let $\epsilon > 0$ and let $(V, \|\cdot\|)$ be a normed space. $\mathcal{C} \subseteq V$ is an* ϵ-covering of V *if for any $v \in V$, there exists $v' \in \mathcal{C}$ such that $\|v - v'\| \leq \epsilon$.*

In particular, we will use the following lemma regarding the size of coverings of balls of a certain radius in a normed space.

Lemma 6. *(Mohri et al., 2018) Fix an arbitrary norm $\|\cdot\|$ and let $\mathcal{B}$ be the ball radius R in this norm. Let $\mathcal{C}$ be a smallest possible ϵ-covering of $\mathcal{B}$. Then*

$$|\mathcal{C}| \leq \left(\frac{3R}{\epsilon}\right)^d$$

Next, we give the proof of the main theorem of this section.

Theorem 7. *Let ρ be a function with Lipschitz constant L_ρ satisfying $\rho(0) = 0$ and consider perturbations in r-norm. Then, the following upper bound holds for the adversarial Rademacher complexity of $\mathcal{G}_p^n$:*

$$\widetilde{\mathfrak{R}}_S(\mathcal{G}_p^n) \leq L_\rho\left[\frac{W\Lambda\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|\mathbf{X}\|_{r,\infty} + \epsilon)}{\sqrt{m}}\right]\left(1 + \sqrt{d(n+1)\log(9m)}\right).$$

Proof. Let $\mathcal{C}_1$ be a covering of the ℓ_1 ball of radius Λ with ℓ_1 balls of radius δ_1 and $\mathcal{C}_2$ a covering of the ℓ_p ball of radius W with ℓ_p balls of radius δ_2. We will later choose δ_1 and δ_2 as functions of m, W, and Λ. For any $\mathbf{x}$, define $\widetilde{f}(\mathbf{x})$ and $\widetilde{f}^c(\mathbf{x})$ as follows:

$$\widetilde{f}(\mathbf{x}) = \inf_{\|\mathbf{x}'-\mathbf{x}\|_r \le \epsilon} y \sum_{j=1}^{n} u_j \rho(\mathbf{w}_j \cdot \mathbf{x}') \quad \text{and} \quad \widetilde{f}^c(\mathbf{x}) = \inf_{\|\mathbf{x}'-\mathbf{x}\|_r \le \epsilon} y \sum_{j=1}^{n} u_j^c \rho(\mathbf{w}_j^c \cdot \mathbf{x}'),$$

where $\mathbf{u}^c$ is the closest element to $\mathbf{u}$ in $\mathcal{C}_1$ and $\mathbf{w}^c$ is the closest element to $\mathbf{w}$ in $\mathcal{C}_2$. Define ϵ' as follows:

$$\epsilon' = \sup_{i \in [m]} \sup_{\substack{\|\mathbf{u}\|_1 \le \Lambda \\ \|\mathbf{w}\|_p \le W}} |\widetilde{f}(\mathbf{x}_i) - \widetilde{f}^c(\mathbf{x}_i)|.$$

One can bound the Rademacher complexity of the whole class $\mathcal{G}_p^n$ in terms of the Rademacher complexity of this same class restricted to $\mathbf{u} \in \mathcal{C}_1$ and $\mathbf{w}_j \in \mathcal{C}_2$.

$$\begin{aligned} \widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G}_p^n) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\substack{\|\mathbf{u}\|_1 \le \Lambda \\ \|\mathbf{w}\|_j \le W}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \inf_{\|\mathbf{x}_i - \mathbf{x}_i'\|_r \le \epsilon} y_i \sum_{j=1}^{n} u_j \rho(\mathbf{w}_j \cdot \mathbf{x}_i') \right] \\ &\le \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\substack{||\mathbf{u}||^c \in \mathcal{C}_1 \\ \mathbf{w}_j^c \in \mathcal{C}_2}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \inf_{\|\mathbf{x}_i - \mathbf{x}_i'\|_r \le \epsilon} y_i \sum_{j=1}^{n} u_j^c \rho(\mathbf{w}_j^c \cdot \mathbf{x}_i') \right] + \epsilon' \end{aligned} \tag{31}$$

Then, by Massart's lemma, the first term in (31) can be bounded as follows:

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\substack{||\mathbf{u}||^c \in \mathcal{C}_1 \\ \mathbf{w}_j^c \in \mathcal{C}_2}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \inf_{\|\mathbf{x}_i - \mathbf{x}_i'\|_r \le \epsilon} y_i \sum_{j=1}^{n} u_j^c \rho(\mathbf{w}_j^c \cdot \mathbf{x}_i') \right] \le \frac{K\sqrt{2\log(|\mathcal{C}_1||\mathcal{C}_2|^n)}}{m} \tag{32}$$

with

$$K^2 = \sup_{\substack{\mathbf{w}_j^c \in \mathcal{C}_2 \\ \mathbf{u}^c \in \mathcal{C}_1}} \sum_{i=1}^{m} \left(\inf_{\|\mathbf{x}_i - \mathbf{x}_i'\|_r \le \epsilon} y_i \sum_{j=1}^{n} u_j^c \rho(\mathbf{w}_j^c \cdot \mathbf{x}_i') \right)^2 .$$

We will show the following upper bound for K:

$$K \le \sqrt{m}\Lambda W \max\Big(1, d^{1-\frac{1}{r}-\frac{1}{p}}(\|\mathbf{X}\|_{r,\infty} + \epsilon)\Big). \tag{33}$$

Let $\mathbf{x}_*^c$ be the minimizer of $f^c(\mathbf{x})$ within an ϵ-ball around $\mathbf{x}$. Since $\widetilde{f}^c$ is continuous and the closed unit r-ball is compact, the extreme value theorem implies that $\mathbf{x}_*^c$ exists. Then

$$\widetilde{f}^c(\mathbf{x}) = y \sum_{j=1}^{n} u_j^c \rho(\mathbf{w}_j^c \cdot \mathbf{x}_*^c) \tag{34}$$

We then apply the following inequalities:

$$\begin{aligned} \left| y_i \sum_{j=1}^{n} u_j^c \rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c) \right| &\le \sum_{j=1}^{n} |u_j^c||\rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c)| && \text{(triangle inequality)} \\ &= \sum_{j=1}^{n} |u_j^c||\rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c) - \rho(0)| && (\rho(0) = 0 \text{ assumption}) \\ &\le L_\rho \sum_{j=1}^{n} |u_j^c||\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c| && \text{(Lipschitz property)} \\ &\le L_\rho \sum_{j=1}^{n} |u_j^c|\|\mathbf{w}_j^c\|_p \|\mathbf{x}_{i*}^c\|_{p^*} && \text{(Hölder's inequality)} \\ &\le L_\rho \sum_{j=1}^{n} |u_j^c| W \|\mathbf{x}_{i*}^c\|_{p^*} && (\|\mathbf{w}_j\| \le W) \\ &\le L_\rho \Lambda W \|\mathbf{x}_{i*}^c\|_{p^*} && (\|\mathbf{u}\| \le \Lambda) \\ &\le L_\rho \Lambda W (\|\mathbf{X}\|_{r,\infty} + \epsilon) \max(1, d^{1-\frac{1}{r}-\frac{1}{p}}). \end{aligned} \tag{35}$$

The last inequality is justified by the following, where we use the triangle inequality and Lemma 1:

$$\begin{aligned}
\|\mathbf{x}_{i*}^c\|_p &\le \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})\|\mathbf{x}_{i*}^c\|_r \\
&\le \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|\mathbf{x}_i\|_r + \|\mathbf{x}_{i*}^c - \mathbf{x}_i\|_r) \\
&\le \max(1, d^{1-\frac{1}{r}-\frac{1}{p}})(\max_{i\in[m]}\|\mathbf{x}_i\|_r + \epsilon) \\
&\le \max(1, d^{1-\frac{1}{r}-\frac{1}{p}})(\|\mathbf{X}\|_{r,\infty} + \epsilon).
\end{aligned} \tag{36}$$

Equation (35) implies the desired bound (33) on K. Next, plugging in the bound from Lemma 6 in (32), we obtain

$$\widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G}_p^n) \le \frac{L_\rho \Lambda W \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|\mathbf{X}\|_{r,\infty} + \epsilon)}{\sqrt{m}} \sqrt{2d\log\left(\frac{3\Lambda}{\delta_1}\right) + 2nd\log\left(\frac{3W}{\delta_2}\right)} + \epsilon'. \tag{37}$$

We now turn our attention to estimating ϵ'. Similar to (34), we define $\mathbf{x}_*$ as the minimizer of $\widetilde{f}(\mathbf{x})$ within an ϵ-ball around $\mathbf{x}$ where

$$\widetilde{f}(\mathbf{x}) = y\sum_{j=1}^n u_j \rho(\mathbf{w}_j \cdot \mathbf{x}_*).$$

We decompose the difference between $\widetilde{f}(\mathbf{x}_i)$ and $\widetilde{f}^c(\mathbf{x}_i)$ and bound each piece separately:

$$\widetilde{f}(\mathbf{x}_i) - \widetilde{f}^c(\mathbf{x}_i) = \left(y\sum_{j=1}^n u_j\rho(\mathbf{w}_j \cdot \mathbf{x}_{i*}) - y\sum_{j=1}^n u_j\rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c)\right) + \left(y\sum_{j=1}^n u_j\rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c) - y\sum_{j=1}^n u_j^c\rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c)\right). \tag{38}$$

The first term above can be bounded as follows:

$$y\sum_{j=1}^n u_j\rho(\mathbf{w}_j \cdot \mathbf{x}_{i*}) - y\sum_{j=1}^n u_j\rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c) \tag{39}$$

$$\begin{aligned}
&\le y\sum_{i=1}^n u_j\rho(\mathbf{w}_j \cdot \mathbf{x}_{i*}^c) - y\sum_{j=1}^n u_j\rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c) && \text{(infimum of first sum at } \mathbf{x}_*\text{)} \\
&\le \sum_{j=1}^n |u_j||\rho(\mathbf{w}_j \cdot \mathbf{x}_{i*}^c) - \rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c)| && \text{(triangle inequality)} \\
&\le L_\rho\sum_{j=1}^n |u_j||(\mathbf{w}_j - \mathbf{w}_j^c)\cdot \mathbf{x}_{i*}^c| && \text{(Lipschitz property)} \\
&\le L_\rho\sum_{j=1}^n |u_j|\|\mathbf{w}_j - \mathbf{w}_j^c\|_p\|\mathbf{x}_{i*}^c\|_{p^*} && \text{(Hölder's inequality)} \\
&\le L_\rho\sum_{j=1}^n |u_j|\delta_2\|\mathbf{x}_{i*}^c\|_{p^*} && (\|\mathbf{w}_j - \mathbf{w}_j^c\| \le \delta_2) \\
&\le L_\rho\sum_{j=1}^n |u_j|\delta_2\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|\mathbf{X}\|_{r,\infty} + \epsilon) && \text{(equation (36))} \\
&\le L_\rho\Lambda\delta_2(\|\mathbf{X}\|_{r,\infty} + \epsilon)\max(1, d^{1-\frac{1}{p}-\frac{1}{r}}). && (\|\mathbf{u}\|_1 \le \Lambda)
\end{aligned} \tag{40}$$

Similarly we can bound the second term in (38) as follows:

$$
\begin{aligned}
& y\sum_{j=1}^{n} u_j \rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c) - y\sum_{j=1}^{n} u_j^c \rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c) && && (41)\\
&\leq \sum_{j=1}^{n} |u_j - u_j^c||\rho(\mathbf{w}_j \cdot \mathbf{x}_{i*}^c)| && \text{(triangle inequality)}\\
&= \sum_{j=1}^{n} |u_j - u_j^c||\rho(\mathbf{w}_j \cdot \mathbf{x}_{i*}^c) - \rho(0)| && (\rho(0) = 0 \text{ assumption})\\
&\leq L_\rho \sum_{j=1}^{n} |u_j - u_j^c||\mathbf{w}_j \cdot \mathbf{x}_{i*}^c| && \text{(Lipschitz property)}\\
&\leq L_\rho \sum_{j=1}^{n} |u_j - u_j^c|\|\mathbf{w}_j\|_p\|\mathbf{x}_{i*}^c\|_{p^*} && \text{(Hölder's inequality)}\\
&\leq L_\rho \sum_{j=1}^{n} |u_j - u_j^c|W\|\mathbf{x}_{i*}^c\|_{p^*} && (\|\mathbf{w}_j\| \leq W)\\
&\leq L_\rho \sum_{j=1}^{n} |u_j - u_j^c|W(\|\mathbf{X}\|_{r,\infty} + \epsilon)\max(1, d^{1-\frac{1}{p}-\frac{1}{r}}) && \text{(equation (36))}\\
&\leq L_\rho \delta_1 W(\|\mathbf{X}\|_{r,\infty} + \epsilon)\max(1, d^{1-\frac{1}{p}-\frac{1}{r}}) && (\|\mathbf{u} - \mathbf{u}^c\|_1 \leq \delta_1) && (42)
\end{aligned}
$$

Combining equations (40) and (42) results in

$$\widetilde{f}(\mathbf{x}_i) - \widetilde{f}^c(\mathbf{x}_i) \leq L_\rho(\|\mathbf{X}\|_{r,\infty} + \epsilon)\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(W\delta_1 + \Lambda\delta_2).$$

By a similar analysis, one can also show that $\widetilde{f}^c(\mathbf{x}_i) - \widetilde{f}(\mathbf{x}_i) \leq L_\rho(\|\mathbf{X}\|_{r,\infty} + \epsilon)\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(W\delta_1 + \Lambda\delta_2)$. Therefore

$$\epsilon' \leq L_\rho(\|\mathbf{X}\|_{r,\infty} + \epsilon)\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(W\delta_1 + \Lambda\delta_2) \tag{43}$$

Combining equations (43) and (37) and choosing $\delta_1 = \frac{\Lambda}{2\sqrt{m}}$ and $\delta_2 = \frac{W}{2\sqrt{m}}$ yield

$$\widetilde{\mathfrak{R}}_S(\mathcal{G}_p^n) \leq \left(\frac{L_\rho W\Lambda \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|\mathbf{X}\|_{r,+\infty} + \epsilon)}{\sqrt{m}}\right)\left(1 + \sqrt{2d(n+1)\log(6\sqrt{m})}\right),$$

which completes the proof. □

E. Characterizing adversarial perturbations for ReLU neural networks

E.1. Condition for adversarial perturbations to be on the r-sphere (proof of Theorem 8)

In this section we provide the proof of Theorem 8 which characterizes adversarial perturbations to a one-layer neural net. First, by the extreme value theorem, (17) achieves its minimum on $\|\mathbf{s}\|_r \leq 1$. Thus we can restate (17) as

$$\min_{\|\mathbf{s}\|_r \leq 1} f(\mathbf{s}) = \sum_{j=1}^{n} u_j(\mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}))_+. \tag{44}$$

Theorem 8. *Let d be the dimension and n the number of neurons. Consider (44) as defined above. If either $\|\mathbf{x}\|_r \geq \epsilon$ or $n < d$, an optimum is attained on the sphere $\{\mathbf{s}\colon \|\mathbf{s}\|_r = 1\}$. Otherwise, an optimum is attained either at $\mathbf{s} = -\frac{1}{\epsilon}\mathbf{x}$ or on $\|\mathbf{s}\|_r = 1$.*

The proof of this theorem relies on two important lemmas stated below. We defer the proofs of these lemmas to the end of the section.

Lemma 7. *Consider (44). Then an optimum is obtained in either*

1. $S_1 \coloneqq \{\mathbf{s}\colon \|\mathbf{s}\|_r = 1\}$
2. $S_2 = \{\mathbf{s}\colon \mathbf{w}_{j_k} \cdot (\mathbf{x} + \epsilon\mathbf{s}) = 0$ *for linearly independent* $\mathbf{w}_{j_1} \dots \mathbf{w}_{j_d}\}$

Lemma 8. *Consider the intersection of* d *linearly independent hyperplanes defined by*

$$\mathbf{v}_k \cdot (\mathbf{x} + \epsilon\mathbf{s}) = 0\colon k = 1 \dots d \tag{45}$$

for a fixed $\mathbf{x}$*. They intersect at a single point given by* $\mathbf{s} = -\frac{1}{\epsilon}\mathbf{x}$*.*

Next we use lemmas 7 and 8 to prove Theorem 8.

Proof of Theorem 8. By Lemma 7, there exists a point $\mathbf{s}^*$ with

$$f(\mathbf{s}^*) = \min_{\|\mathbf{s}\|_r \leq 1} f(\mathbf{s})$$

for which either $\|\mathbf{s}^*\|_r = 1$ or

$$\{\mathbf{s}^*\colon \mathbf{w}_{j_k} \cdot (\mathbf{x} + \epsilon\mathbf{s}^*) = 0 \text{ for some linearly independent } \mathbf{w}_{j_1} \dots \mathbf{w}_{j_d}\}$$

If $n < d$, then there aren't d linearly independent w_is, and thus $\mathbf{s}^*$ satisfies $\|\mathbf{s}^*\|_r = 1$.

Now assume that $n \geq d$ and $\|\mathbf{s}^*\|_r \neq 1$. Lemma 8 implies that $\mathbf{s}^* = -\frac{1}{\epsilon}\mathbf{x}$ and hence $\|\mathbf{x}\|_r < \epsilon$. Taking the contrapositive of this statement results in

$$n \geq d \text{ and } \|\mathbf{x}\|_r \geq \epsilon \Rightarrow \|\mathbf{s}^*\|_r = 1$$

□

We end the subsection with the proofs of lemmas 7 and 8. Before we prove Lemma 7 we state and prove a simpler statement that will be used in its proof.

Lemma 9. *Consider (44). Then an optimum is obtained at either*

1. $S_1 \coloneqq \{\mathbf{s}\colon \|\mathbf{s}\|_r = 1\}$
2. $S_2 = \{\mathbf{s}\colon \mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}) = 0$ *for some* $\mathbf{w}_j\}$

Proof. We know from calculus that every extreme point of f is obtained either on the boundary of the optimization region, at a point where the function isn't differentiable, or where the derivative is zero. First, observe that at any non-differentiable point with $\|\mathbf{s}\|_r < 1$, some $\mathbf{w}_j$ must satisfy $\mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}) = 0$. Now we'll consider the third case, points where $\nabla f(\mathbf{s}) = 0$. Assume that $\mathbf{s}^*$ is an extreme point for which f is differentiable (and with derivative zero). Then we claim that there is another point in either S_1 or S_2 that achieves the same objective value. Let $P = \{j : \mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}^*) > 0\}$ Then

$$f(\mathbf{s}^*) = \sum_{j \in P} u_j(\mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}^*))$$

Fix this set P. Note that the region where

$$f(\mathbf{s}) = \sum_{j \in P} u_j(\mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}))$$

is defined by

$$R = \left\{\mathbf{s}\colon \|\mathbf{s}\|_r \leq 1, \mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}) \geq 0 \text{ for } j \in P, \mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}) \leq 0 \text{ for } j \in P^C\right\} \tag{46}$$

By assumption,

$$\nabla f(\mathbf{s}^*) = \epsilon \sum_{j \in P} u_j \mathbf{w}_j = 0$$

However, for any other $\mathbf{s}$ in the region defined by (46)

$$\nabla f(\mathbf{s}) = \epsilon \sum_{j \in P} u_j \mathbf{w}_j = f(\mathbf{s}^*) = 0$$

Hence, f is constant on the interior of the region defined by (46). By continuity, it is constant on the closure of this region as well. Hence an optimum of the same value is obtained in either S_1 or S_2. □

Proof of Lemma 7. This will be a proof by induction. Let $\mathbf{s}^*$ be an optimum. Define $Z_0^{\mathbf{s}} = \{\mathbf{w}_j : \mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}) = 0\}$ and let k be the dimension of $\text{span}(Z_0^{\mathbf{s}^*})$. The induction will be on k.

Base Case: By the previous lemma, when looking for the optimum, we only need to consider $\mathbf{s}$ for which $\|\mathbf{s}\|_r = 1$ or $\mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}) = 0$ for some j. Assume that we have an extreme point $\mathbf{s}^*$ for which $\|\mathbf{s}^*\| < 1$. Then $k \geq 1$.

Inductive Step: Let $\mathbf{s}^*$ be our extreme point and assume that $\|\mathbf{s}^*\|_r < 1$. Our induction hypothesis is that $\dim(\text{span}(Z_0^{\mathbf{s}^*})) = k < d$. We will show that there is another point $\mathbf{t}$ that achieves the same objective value satisfying either $\|\mathbf{t}\|_r = 1$ or $\dim(\text{span}(Z_0^{\mathbf{t}})) = k + 1$.

Let Z be any linearly independent subset of $Z_0^{\mathbf{s}^*}$. We can parameterize $\mathbf{s}$ to be in the intersection of the hyperplanes that define Z. Formally, let $\mathbf{v} \in \text{span}(Z)$ with $\mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{v}) = 0$ for all $\mathbf{w}_j \in Z$, and let $\mathbf{A}: \mathbb{R}^{d-k} \to \mathbb{R}^d$ be a matrix whose columns span $Z^\perp$. Take $\mathbf{s} = \mathbf{v} + \mathbf{A}\mathbf{s}'$, $P = \{j : \mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}^*) > 0\}$, and $N = \{j : \mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}^*) < 0\}$. Then by continuity,

$$f(\mathbf{s}) = \sum_{j \in P} u_j \mathbf{w}_j \cdot (\mathbf{x} + \epsilon(\mathbf{v} + \mathbf{A}\mathbf{s}'))$$

holds on the region defined by

$$R = \{\mathbf{s}' : \|\mathbf{v} + \mathbf{A}\mathbf{s}'\|_r \leq 1, \mathbf{w}_j \cdot (\mathbf{x} + \epsilon(\mathbf{v} + \mathbf{A}\mathbf{s}')) \geq 0 \text{ for } j \in P, \mathbf{w}_j \cdot (\mathbf{x} + \epsilon(\mathbf{v} + \mathbf{A}\mathbf{s}')) \leq 0 \text{ for } j \in N\} \tag{47}$$

For convenience, set

$$g(\mathbf{s}') := f(\mathbf{v} + \mathbf{A}\mathbf{s}')$$

We assumed that our optimum $\mathbf{s}^*$ satisfied $\|\mathbf{s}^*\|_r < 1$ and $\mathbf{w}_j \cdot (x + \epsilon\mathbf{s}) \neq 0$ for $j \in P \cup N$, which entails that our critical point is in the interior of R. On the interior of this region, to find all critical points, we can differentiate g in $\mathbf{s}'$:

$$\nabla g(\mathbf{s}') = \mathbf{A}^\top \sum_{j \in P} u_j \mathbf{w}_j$$

and set $\nabla g(\mathbf{s}')$ equal to zero. This expression is independent of $\mathbf{s}' \in R$. Let $\mathbf{z}$ be a critical point of g in $\text{int}(R)$. Then $\nabla g(\mathbf{z}) = 0$ implies that $\nabla g(\mathbf{s}') = 0$ for all $\mathbf{s}' \in \text{int}(R)$. Hence, g is constant on R. This implies that there is another point $\mathbf{s}'$ with the same objective value on ∂R. For this point, either $\|\mathbf{v} + \mathbf{A}\mathbf{s}'\|_r = 1$, or $\|\mathbf{v} + \mathbf{A}\mathbf{s}'\|_r < 1$ and $\mathbf{w}_j \cdot (\mathbf{x} + \epsilon(\mathbf{v} + \mathbf{A}\mathbf{s}')) = 0$ for some $j \in P \cup N$. If the second option holds, $j \in P \cup N$ means that $\mathbf{w}_j \notin \text{span}\, Z_0^{\mathbf{s}^*}$. It follows that $\text{span}(Z_0^{\mathbf{s}^*} \cup \{\mathbf{w}_j\})$ is dimension $k + 1$ and this completes the induction step. □

Finally we prove Lemma 8.

Proof of Lemma 8. By substitution $\mathbf{s} = -\frac{1}{\epsilon}\mathbf{x}$ is a solution to the system of equations (45). Since d linearly independent equations intersect at a point, it is the only solution to these equations. □

E.2. A Necessary Condition

In this subsection we present a necessary condition at the optimum when perturbations are measured in any general r-norm. Throughout this subsection, $\mathbf{u} \odot \mathbf{v}$ will be the elementwise product of $\mathbf{u}$ an $\mathbf{v}$, $\mathbf{u}^r$ will be elementwise exponentiation, $\|\mathbf{v}\|$ will be elementwise absolute value, and $\text{sgn}(\mathbf{v})$ will be the vector of signs of the components of $\mathbf{v}$. We adopt the convention $\text{sgn}(0) = 0$. Recall the definition of dual norm:

$$\|\mathbf{u}\|_{r^*} = \sup_{\|\mathbf{v}\|_r \leq 1} \mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|_{r^*}$$

Equality holds at the vector $\mathbf{v} = \frac{1}{\|\mathbf{u}\|_r^{r-1}} |\mathbf{u}|^{r-1} \odot \text{sgn}(\mathbf{u})$, which has unit r^*-norm. For convenience we, define

$$\text{dual}_r(\mathbf{u}) = (\text{sgn}\, \mathbf{u}) \odot \frac{|\mathbf{u}|^{r-1}}{\|\mathbf{u}\|_r^{r-1}}$$

which gives

$$\mathbf{u} \cdot \text{dual}_r(\mathbf{u}) = \|\mathbf{u}\|_r^r = 1.$$

Below we state and prove the main theorem of this section.

Theorem 13. *Let $1 < r < \infty$. Take*

$$f(\mathbf{s}) = \sum_{j=1}^{n} u_j(\mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}))_+ \tag{48}$$

Assume that either $\|\mathbf{x}\|_r \geq \epsilon$ or $n < d$. Let $\mathbf{s}^*$ is a minimizer of f on the unit r-sphere. Define the following sets:

$$P = \{j \colon \mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}^*) > 0\}$$

$$Z = \{j \colon \mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}^*) = 0\}$$

$$N = \{j \colon \mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}^*) < 0\}$$

Let P_Z be the orthogonal projection onto the subspace spanned by the vectors in Z, and P_{Z^C} be the projection onto the complement of this subspace. Then the following holds: If $P \neq \varnothing$

$$\mathbf{s}^* = -\frac{\epsilon}{\lambda} \left| \left(\sum_{j \in P} u_j \mathbf{w}_j + \sum_{j \in Z} t_j u_j \mathbf{w}_j \right) \right|^{r-1} \odot \operatorname{sgn} \left(\sum_{j \in P} u_j \mathbf{w}_j + \sum_{j \in Z} t_j u_j \mathbf{w}_j \right) \tag{49}$$

where the constants t_j, λ are given by the equations

$$\|\mathbf{s}^*\|_r = 1 \tag{50}$$

$$P_Z \mathbf{s}^* = -\frac{1}{\epsilon} P_Z \mathbf{x} \tag{51}$$

Further, if $P = \varnothing$,

$$\mathbf{s}^* = -\frac{P_Z \mathbf{x}}{\|P_Z \mathbf{x}\|} \tag{52}$$

Using the dual_r *notation,* $\mathbf{s}^*$ *can be expressed as*

$$\mathbf{s}^* = \operatorname{dual}_r \left(\left| \sum_{j \in P} u_j \mathbf{w}_j + \sum_{j \in Z} t_j u_j \mathbf{w}_j \right| \right) \odot \operatorname{sgn}(\sum_{j \in P} u_j \mathbf{w}_j + \sum_{j \in Z} t_j u_j \mathbf{w}_j) \tag{53}$$

Notice that for $r = 2$, $\operatorname{dual}_r(\mathbf{s}^*) = \mathbf{s}^*$ and then we can write $\mathbf{s}^*$ explicitly:

$$\mathbf{s}^* = -\left(\sqrt{1 - \frac{\|P_Z \mathbf{x}\|_2^2}{\epsilon^2}} \frac{P_{Z^C} \sum_{j \in P} u_j \mathbf{w}_j}{\left\| P_{Z^C} \sum_{j \in P} u_j \mathbf{w}_j \right\|_2} + \frac{\|P_Z \mathbf{x}\|_2}{\epsilon} \frac{P_Z \mathbf{x}}{\|P_Z \mathbf{x}\|} \right)$$

Before proceeding with the proof of this theorem, we state a useful definition and lemma. Recall the definition of the subgradient of a convex function:

Definition 3. *The subdifferential of a convex function f_1 is the set*

$$\partial f_1(\mathbf{x}) = \{\mathbf{v} \colon f_1(\mathbf{y}) - f_1(\mathbf{x}) \geq \mathbf{v} \cdot (\mathbf{y} - \mathbf{x})\}$$

while the subdifferential of a concave function f_2 is the set

$$-\partial(-f_2(\mathbf{x})) = \{\mathbf{v} : f_2(\mathbf{y}) - f_2(\mathbf{x}) \leq \mathbf{v} \cdot (\mathbf{y} - \mathbf{x})\}$$

For a function $f = f_1 + f_2$ that is the sum of a convex function f_1 and a concave function f_2, the following observation from (Polyakova, 1986) shows why these definitions are useful for us.

Lemma 10. *Let $f = f_1 + f_2$ with f_1 convex and f_2 concave. Assume that f has a local minimum at x^*. Then*

$$\mathbf{0} \in \partial f_1(\mathbf{x}^*) + \partial f_2(\mathbf{x}^*)$$

Note that the same statement holds for local maxima of f. We defer the proof of this lemma to the end of this subsection.

To prove Theorem 13, we form a Lagrangian for computing the optimum of (48). Lemma 10 gives a necessary condition in terms of the subgradient of this Lagrangian. Subsequently, we use information about the dual variables obtained via Theorem 8 and convexity to show (49), (50), and (52). (Note that either $\|\mathbf{x}\|_r \geq \epsilon$ or $n < d$ are precisely the conditions for Theorem 8). After that, standard linear algebra shows (51).

Proof of Theorem 13. **Establishing Equations (49) and (50):** First note that the objective f is the sum of a convex and a concave function: take

$$f_1(\mathbf{s}) = \sum_{j:u_j>0} u_j\left(\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s})\right)_+ \quad f_2(\mathbf{s}) = \sum_{j:u_j<0} u_j\left(\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s})\right)_+$$

f_1 is convex because it is the sum of convex functions and f_2 is concave because it is the sum of concave functions. This observation will allow us the apply Lemma 10. We form the corresponding Lagrangian:

$$L(\mathbf{s}) = \sum_{j=1}^{n} u_j(\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s}))_+ + \frac{\lambda}{r}(\|\mathbf{s}\|_r^r - 1)$$

L is convex in an open set around every local minimum. On this set, since we are optimizing over $\|\mathbf{s}\|_r \leq 1$, we know that $\lambda \geq 0$. Further, Theorem 8 shows that there must be an optimum on the unit r-sphere for $\|\mathbf{x}\|_r \geq \epsilon$.

By Lemma 10, we want to find a condition when $\mathbf{0}$ is in the subdifferential. We use the following two facts:

1. $$\partial(x)_+ = \begin{cases}\{0\} & \text{if } x<0\\ [0,1] & \text{if } x=0\\ \{1\} & \text{if } x>0\end{cases}$$

2. For $1 < r < \infty$, the r norm is differentiable. Hence we can write:

$$\nabla\|\mathbf{s}\|_r^r = |\mathbf{s}|^{r-1}\odot \operatorname{sgn}\mathbf{s} = \|\mathbf{s}\|_r^{r-1}\operatorname{dual}_{r^*}(\mathbf{s})$$

Hence, if $\|\mathbf{s}\| = 1$, $\partial\|\mathbf{s}\|_r^r = \operatorname{dual}_{r^*}(\mathbf{s}) = \operatorname{sgn}\mathbf{s}\odot|\mathbf{s}|^{r-1}$.

Then applying Lemma 10, we need

$$\mathbf{0}\in\epsilon\partial\sum_{j\in P}u_j(\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s})_+ + \epsilon\partial\sum_{j\in Z}u_j(\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s}))_+ + \epsilon\partial\sum_{j\in N}u_j(\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s})_+) + \partial\frac{\lambda}{r}(\|\mathbf{s}\|_r^r - 1).$$

Hence for some $t_j\in[0,1]$,

$$\mathbf{0} = \epsilon\sum_{j\in P}u_j\mathbf{w}_j + \epsilon\sum_{j\in Z}t_ju_j\mathbf{w}_j + \frac{\lambda}{r}\partial\|\mathbf{s}\|_r^r \tag{54}$$

Using Theorem 8, we choose an optimum on the boundary $\|\mathbf{s}\|_r = 1$. First we consider $\mathbf{s}^*$ with $\lambda \neq 0$. This allows for solving for $\partial\|\mathbf{s}\|_r^r$:

$$\operatorname{dual}_r(\mathbf{s}^*) = \mathbf{s}^*\odot|\mathbf{s}^*|^{r-1} = -\frac{\epsilon}{\lambda}\left(\sum_{u\in P}u_j\mathbf{w}_j + \sum_{j\in Z}t_ju_j\mathbf{w}_j\right)$$

Now since $\operatorname{dual}_r(\mathbf{s}^*)$ has r^*-norm 1, this allows us to solve for $|\lambda|$. Further recall that at a local minimum, $\lambda \geq 0$ which tells us $\operatorname{sgn}\lambda$. Using this information, we can solve for λ which establishes (50). Since $1 < r < \infty$, this equation further establishes (49).

Establishing Equation (52): Now we consider the case where $\lambda = 0$ or $P = \varnothing$. For $\lambda = 0$, we will show by contradiction that P must be empty. Assume that $P \neq \varnothing$. Equation (54) then simplifies to

$$\mathbf{0} = \epsilon \sum_{j \in P} u_j \mathbf{w}_j + \epsilon \sum_{u \in Z} t_j u_j \mathbf{w}_j$$

which implies that

$$\sum_{j \in P} u_j \mathbf{w}_j = - \sum_{j \in Z} t_j u_j \mathbf{w}_j$$

However, if we take the dot product with $\mathbf{x} + \epsilon \mathbf{s}^*$,

$$\sum_{j \in P} u_j \mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}^*) = - \sum_{j \in Z} \mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}^*) = 0$$

and therefore, $\mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}^*) \leq 0$ for some $j \in P$ which contradicts the definition of P. Therefore, P must be empty.

Now we assume that $\mathbf{s}^*$ has $P = \varnothing$ and we show that there is a point $\mathbf{z}^*$ that achieves the same objective value as $\mathbf{s}^*$ but has $N = \varnothing$. This will be proved by induction on the size of $N_{\mathbf{z}}$. This will then imply that we can take $\mathbf{s}^* = -\frac{P_Z \mathbf{x}}{\|P_Z \mathbf{x}\|}$.

Denote by $Z_{\mathbf{s}}, N_{\mathbf{s}}$

$$P_{\mathbf{s}} = \{j : \mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}) > 0\}$$
$$Z_{\mathbf{s}} = \{j : \mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}) = 0\}$$
$$N_{\mathbf{s}} = \{j : \mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}) < 0\}$$

For the base case, we use a point $\mathbf{s}$ that achieves the optimal value and has $P_{\mathbf{s}} = \varnothing$. If $N_{\mathbf{s}} = \varnothing$, we are done. Otherwise, for the induction step, we assume $N_{\mathbf{s}} \neq \varnothing$. We will find a vector $\mathbf{z}$ that achieves these same objective value as s, but $N_{\mathbf{s}} \supsetneq N_{\mathbf{z}}$. Pick a vector $\mathbf{v}$ perpendicular to $\operatorname{span}\{\mathbf{w}_j\}_{j \in Z_{\mathbf{s}}}$ but not perpendicular to $\operatorname{span}\{\mathbf{w}_j\}_{j \in N_s}$. Such a vector must exist because if $\mathbf{w}_k \in \operatorname{span}\{\mathbf{w}_j\}_{j \in Z_{\mathbf{s}}}$, then $\mathbf{w}_k \in \mathbf{Z}_s$. We now consider

$$\mathbf{z}(\delta) = \frac{\mathbf{s} + \delta \mathbf{v}}{\|\mathbf{s} + \delta \mathbf{v}\|}$$

Note that

$$\mathbf{z}(\delta) \cdot \mathbf{w}_j = 0$$

for each $j \in Z_{\mathbf{s}}$ for all δ. Because the strict inequality

$$\mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{z}(\delta)) < 0 \; j \in N_{\mathbf{s}}$$

is satisfied for $\delta = 0$, it is also satisfied for some small $\delta \neq 0$. We can now increase or decrease δ until

$$\mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{z}(\delta)) = 0 \text{ for some } j \in N_{\mathbf{s}}$$

and $\mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{z}(\delta)) < 0$ for the remaining js in N. We then have $N_{\mathbf{s}} \supsetneq N_{\mathbf{z}(\delta)}$. Furthermore, $f(\mathbf{s}) = f(\mathbf{z}(\delta))$ because the set P is still empty.

Establishing Equation (51): Let $\{\mathbf{f}_k\}_{k=1}^{d_Z}$ be an orthonormal basis of $\operatorname{span}\{\mathbf{w}_j\}_{j \in Z}$. We will show that $\mathbf{x} \cdot \mathbf{f}_k = -\epsilon \mathbf{s}^* \cdot \mathbf{f}_k$. Since $P_Z \mathbf{x}$ and $-\epsilon P_Z \mathbf{s}^*$ are contained in the subspace spanned by the vectors in Z, this would imply that $P_Z \mathbf{s}^* = -\frac{1}{\epsilon} P_Z \mathbf{x}$. Let

$$\mathbf{f}_k = \sum_{j \in Z} a_{kj} \mathbf{w}_j \tag{55}$$

for some constants a_{kj}. Recall that for all $j \in Z$,

$$\mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}^*) = 0.$$

We then use the above equation and (55) to take the dot product of $\mathbf{x}$ and $\mathbf{f}_k$:

$$\mathbf{x} \cdot \mathbf{f}_k = \mathbf{x} \cdot \sum_{j \in Z} a_{kj} \mathbf{w}_j = \sum_{j \in Z} a_{kj} \mathbf{x} \cdot \mathbf{w}_j = -\epsilon \sum_{j \in Z} a_{kj} \mathbf{s}^* \cdot \mathbf{w}_j = -\epsilon \mathbf{f}_k \cdot \mathbf{s}^*.$$

The above establishes equation (51) and completes the proof of the theorem. □

We end the section by proving Lemma 10.

Proof of Lemma 10. We will show that

$$-\partial f_2(\mathbf{x}^*) \subset \partial f_1(\mathbf{x}^*) \tag{56}$$

This implies

$$\mathbf{0} \in \partial f_1(\mathbf{x}^*) + \partial f_2(\mathbf{x}^*).$$

We prove (56) by contrapositive. We pick a point $\mathbf{x}^*$ and assume that (56) does not hold. Then we show that $\mathbf{x}^*$ cannot be a minimum. Assume (56) does not hold. This assumption implies that for some vector $\mathbf{c}$, $\mathbf{c} \in \partial f_2(\mathbf{x}^*)$ but $-\mathbf{c} \notin \partial f_1(\mathbf{x}^*)$. Then there exists an $\mathbf{x}$ for which

$$f_2(\mathbf{x}) - f_2(\mathbf{x}^*) \leq \mathbf{c}^\top(\mathbf{x} - \mathbf{x}^*)$$

$$f_1(\mathbf{x}) - f_1(\mathbf{x}^*) < -\mathbf{c}^\top(\mathbf{x} - \mathbf{x}^*)$$

Summing the above inequalities, we get:

$$f_1(\mathbf{x}) + f_2(\mathbf{x}) < f_1(\mathbf{x}^*) + f_2(\mathbf{x}^*)$$

so $\mathbf{x}^*$ cannot be a local minimum. □

F. Towards Dimension-Independent Bounds for Neural Networks

F.1. Proof of Theorem 10

Recall from Section 6.2 that given a sample $\mathcal{S}$, $C_\mathcal{S}$ denotes the set of all possible partitions of points in $\mathcal{S}$ that can be obtained based on the sign pattern they induced over the set of weight vectors $\mathbf{u}, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$. For a given partition $\mathcal{C} \in \mathcal{C}_\mathcal{S}$, we denote by $n_\mathcal{C}$ the number of parts in $\mathcal{C}$. Furthermore, we define $C^*_\mathcal{S}$ to be the size of the set $C_\mathcal{S}$ and $\Pi^*_\mathcal{S} = \max_\mathcal{C} n_\mathcal{C}$. We now proceed to prove Theorem 10 that establishes a data dependent bound on the Rademacher complexity of neural networks with one hidden layer.

Theorem 10. *Consider the family of functions $\mathcal{G}^n_p$ with $p \in [1, \infty]$, activation function $\rho(z) = (z)_+$, and perturbations in r-norm for $1 < r < \infty$. Assume that for our sample $\|\mathbf{x}_i\|_r \geq \epsilon$. Then, the following upper bound on the Rademacher complexity holds:*

$$\widetilde{\mathfrak{R}}_\mathcal{S}(\mathcal{G}^n_p) \leq \left[\frac{W\Lambda \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(K(p,d)\|\mathbf{X}^\top\|_{\infty,p^*} + \epsilon)}{\sqrt{m}}\right] C^*_\mathcal{S}\sqrt{\Pi^*_\mathcal{S}},$$

where $K(p,d)$ is defined as

$$K(p,d) = \begin{cases} \sqrt{2\log(2d)} & \textit{if } p = 1 \\ \sqrt{2}\left[\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\right]^{\frac{1}{p^*}} & \textit{if } 1 < p \leq 2 \\ 1 & \textit{if } p \geq 2 \end{cases} \tag{57}$$

Proof of Theorem 10. Let $\mathcal{C}_t$ denote a partition in partitions $\mathcal{C}$. Furthermore, define $\mathbf{s}_t = \operatorname{argmin}_{\|\mathbf{s}\|_r \leq 1} \sum_{j=1}^n u_j \mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s})_+$

for $\mathbf{x} \in \mathcal{C}_t$ and $P_t = \{j : \mathbf{w}_j \cdot (\mathbf{x} + \epsilon \mathbf{s}_t) > 0\}$. The Rademacher complexity of the network can be bounded as

$$\begin{aligned}
\widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G}_p^n) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \frac{1}{m}\sum_{i=1}^m \sigma_i \inf_{\|\mathbf{s}\|_r \le 1} y_i \sum_{j=1}^n u_j(\mathbf{w}_j \cdot (\mathbf{x}_i + \epsilon \mathbf{s}))_+\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \frac{1}{m}\sum_{i=1}^m \sigma_i y_i \sum_{j=1}^n u_j(\mathbf{w}_j \cdot (\mathbf{x}_i + \epsilon \mathbf{s}_i))_+\right] && \text{(definition of } \mathbf{s}_i\text{)} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}}\sum_{i \in \mathcal{C}_t} \sigma_i y_i \sum_{j=1}^n u_j(\mathbf{w}_j \cdot (\mathbf{x}_i + \epsilon \mathbf{s}_t))_+\right] && \text{(definition of } \mathcal{C}_t\text{)} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}}\sum_{i \in \mathcal{C}_t} \sigma_i y_i \sum_{j \in P_t} u_j(\mathbf{w}_j \cdot (\mathbf{x}_i + \epsilon \mathbf{s}_t))\right] && \text{(definition of } P_t\text{)} \\
&\le \left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}}\sum_{i \in \mathcal{C}_t} \sigma_i y_i \sum_{j \in P_t} u_j \mathbf{w}_j \cdot \mathbf{x}_i\right] + \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}}\sum_{i \in \mathcal{C}_t} \sigma_i y_i \sum_{j \in P_t} \epsilon u_j \mathbf{w}_j \cdot \mathbf{s}_t))\right]\right) && (58)
\end{aligned}$$

Next we bound each term in equation (58) separately. For the first term we can write:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}}\sum_{i \in \mathcal{C}_t} \sigma_i y_i \sum_{j \in P_t} u_j \mathbf{w}_j \cdot \mathbf{x}_i\right] &= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \left|\frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}}\sum_{i \in \mathcal{C}_t} \sigma_i y_i \sum_{j \in P_t} u_j \mathbf{w}_j \cdot \mathbf{x}_i\right|\right] && \text{(sign symmetry)} \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \left|\frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}}\sum_{j \in P_t} u_j \mathbf{w}_j \cdot \sum_{i \in \mathcal{C}_t} \sigma_i y_i \mathbf{x}_i\right|\right] && \text{(reordering summations)} \\
&\le \frac{W}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}}\sum_{j \in P_t} |u_j| \left\|\sum_{i \in \mathcal{C}_t} \sigma_i y_i \mathbf{x}_i\right\|_{p^*}\right] && \text{(dual norm definition)}
\end{aligned}$$

Using the bound on the ℓ_1 norm of $\mathbf{u}$ we get:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \le W \\ \|\mathbf{u}\|_1 \le \Lambda}} \frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}}\sum_{i \in \mathcal{C}_t} \sigma_i y_i \sum_{j \in P_t} u_j \mathbf{w}_j \cdot \mathbf{x}_i\right] &\le \frac{W}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{W},\mathbf{u}} \frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}} \Lambda \left\|\sum_{i \in \mathcal{C}_t} \sigma_i y_i \mathbf{x}_i\right\|_{p^*}\right] && \text{(dual norm definition)} \\
&\le \frac{1}{m}\frac{\Lambda W}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{W},\mathbf{u}} \sum_{t=1}^{n_{\mathcal{C}}} \left\|\sum_{i \in \mathcal{C}_t} \sigma_i \mathbf{x}_i\right\|_{p^*}\right] && (\sigma_i \text{ distributed like } y_i\sigma_i) \\
&\le \frac{1}{m}\frac{\Lambda W}{2}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum_{\mathcal{C}} \sum_{t=1}^{n_{\mathcal{C}}} \left\|\sum_{i \in \mathcal{C}_t} \sigma_i \mathbf{x}_i\right\|_{p^*}\right] && \text{(summing over all partitions)} \\
&= \frac{1}{m}\frac{\Lambda W}{2}\sum_{\mathcal{C}} \sum_{t=1}^{n_{\mathcal{C}}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i \in \mathcal{C}_t} \sigma_i \mathbf{x}_i\right\|_{p^*}\right].
\end{aligned}$$

Next, note that

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i \in \mathcal{C}_t} \sigma_i \mathbf{x}_i\right\|_{p^*}\right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le 1} \sum_{i \in \mathcal{C}_t} \sigma_i \mathbf{w} \cdot \mathbf{x}_i\right] = |\mathcal{C}_t| \mathfrak{R}_{\mathcal{C}_t}(\mathcal{F}_p)$$

where $\mathcal{F}_p$ is the linear function class defined in (8) with $W = 1$. Hence, applying Theorem 3,

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i \in \mathcal{C}_t} \sigma_i \mathbf{x}_i\right\|_{p^*}\right] \le K(p,d)\|\mathbf{X}_t^\top\|_{2,p^*} \tag{59}$$

with $K(p,d)$ as defined in (57). $\mathbf{X}_t$ is the matrix with data points in $\mathcal{C}_t$ as columns. Furthermore, we can write:

$$\begin{aligned}
\|\mathbf{X}_t^\top\|_{2,p^*} &= \left(\sum_{j=1}^{d} \|\mathbf{X}_t(j)\|_2^{p^*}\right)^{\frac{1}{p^*}} \quad [\mathbf{X}_t(j) \text{ denotes } j\text{th row of } \mathbf{X}] \\
&\leq \sqrt{|\mathcal{C}_t|}\left(\sum_{j=1}^{d} \|\mathbf{X}(j)\|_\infty^{p^*}\right)^{\frac{1}{p^*}} \\
&= \sqrt{|\mathcal{C}_t|}\|\mathbf{X}^\top\|_{\infty,p^*}.
\end{aligned}$$

Using the above bound we can write:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \leq W \\ \|\mathbf{u}\|_1 \leq \Lambda}} \frac{1}{m}\sum_{t=1}^{n_\mathcal{C}}\sum_{i\in\mathcal{C}_t}\sigma_i y_i \sum_{j\in P_t} u_j \mathbf{w}_j \cdot \mathbf{x}_i\right] &\leq \frac{K(p,d)\Lambda W}{m}\sum_{\mathcal{C}}\sum_{t=1}^{n_\mathcal{C}}\sqrt{|C_t|}\|\mathbf{X}^\top\|_{\infty,p^*} \\
&\leq \frac{K(p,d)\Lambda W}{\sqrt{m}}|\mathcal{C}_\mathcal{S}^*|\sqrt{\Pi_\mathcal{S}^*}\|\mathbf{X}^\top\|_{\infty,p^*}. \qquad (60)
\end{aligned}$$

Here the last inequality follows from the fact that $\sum_{t=1}^{n_\mathcal{C}}|\mathcal{C}_t| = m$ and $\sum_{t=1}^{n_\mathcal{C}}\sqrt{|\mathcal{C}_t|}$ is maximized when $|\mathcal{C}_t| = m/n_\mathcal{C}$ for all t. Now for the second term in (58) we can write:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \leq W \\ \|\mathbf{u}\|_1 \leq \Lambda}} \frac{1}{m}\sum_{t=1}^{n_\mathcal{C}}\sum_{i\in\mathcal{C}_t}\sigma_i y_i \sum_{j\in P_t}\epsilon u_j \mathbf{w}_j\cdot\mathbf{s}_t\right] &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \leq W \\ \|\mathbf{u}\|_1 \leq \Lambda}} \frac{1}{m}\sum_{t=1}^{n_\mathcal{C}}\sum_{i\in\mathcal{C}_t}\sigma_i \sum_{j\in P_t}\epsilon u_j \mathbf{w}_j\cdot\mathbf{s}_t\right] && (y_i\sigma_i \text{ distributed like } \sigma_i) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \leq W \\ \|\mathbf{u}\|_1 \leq \Lambda}} \frac{1}{m}\sum_{t=1}^{n_\mathcal{C}}\sum_{j\in P_t}\epsilon u_j \mathbf{w}_j \sum_{i\in\mathcal{C}_t}\sigma_i\cdot\mathbf{s}_t\right] && \text{(reorder summations)} \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \leq W \\ \|\mathbf{u}\|_1 \leq \Lambda}} \frac{1}{m}\sum_{t=1}^{n_\mathcal{C}}\sum_{j\in P_t}\epsilon|u_j|W\left\|\sum_{i\in\mathcal{C}_t}\sigma_i\cdot\mathbf{s}_t\right\|_{p^*}\right] && \text{(dual norm)} \\
&\leq \frac{\epsilon W\Lambda}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p \leq W \\ \|\mathbf{u}\|_1 \leq \Lambda}} \sum_{t=1}^{n_\mathcal{C}}\left\|\sum_{i\in\mathcal{C}_t}\sigma_i\cdot\mathbf{s}_t\right\|_{p^*}\right] && \text{(dual norm)} \\
&\leq \frac{\epsilon W\Lambda}{m}\sup_{\|\mathbf{s}_t\|_{r^*}\leq 1}\|\mathbf{s}_t\|_{p^*}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Big[\sup_{\substack{\|\mathbf{w}_j\|_p \leq W \\ \|\mathbf{u}\|_1 \leq \Lambda}} \sum_{t=1}^{n_\mathcal{C}}\Big|\sum_{i\in\mathcal{C}_t}\sigma_i\Big|\Big] && (\mathbf{s}_i \text{ constraint}) \\
&= \frac{\epsilon W\Lambda}{m}\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Big[\sup_{\substack{\|\mathbf{w}_j\|_p \leq W \\ \|\mathbf{u}\|_1 \leq \Lambda}} \sum_{t=1}^{n_\mathcal{C}}\Big|\sum_{i\in\mathcal{C}_t}\sigma_i\Big|\Big] && \text{(Lemma 1)} \\
&\leq \frac{\epsilon W\Lambda}{m}\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Big[\sum_{\mathcal{C}}\sum_{t=1}^{n_\mathcal{C}}\Big|\sum_{i\in\mathcal{C}_t}\sigma_i\Big|\Big] && \text{(sum over all classes)} \\
&= \frac{\epsilon W\Lambda}{m}\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})\sum_{\mathcal{C}}\sum_{t=1}^{n_\mathcal{C}}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Big[\Big|\sum_{i\in\mathcal{C}_t}\sigma_i\Big|\Big].
\end{aligned}$$

By Jensen's inequality, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Big[|\sum_{i\in\mathcal{C}_t}\sigma_i|\Big] \leq \sqrt{|\mathcal{C}_t|}.$$

Substituting this bound above we get that

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|\mathbf{w}_j\|_p\leq W\\ \|\mathbf{u}\|_1\leq\Lambda}}\frac{1}{m}\sum_{t=1}^{n_{\mathcal{C}}}\sum_{i\in\mathcal{C}_t}\sigma_i y_i\sum_{j\in P_t}\epsilon u_j\mathbf{w}_j\cdot\mathbf{s}_t\right] &\leq \frac{\epsilon W\Lambda}{m}\max(1,d^{1-\frac{1}{p}-\frac{1}{r}})\sum_{\mathcal{C}}\sum_{t=1}^{n_{\mathcal{C}}}\sqrt{|\mathcal{C}_t|}\\ &\leq \frac{\epsilon\Lambda W}{\sqrt{m}}\max(1,d^{1-\frac{1}{p}-\frac{1}{r}})|\mathcal{C}^*_{\mathcal{S}}|\sqrt{\Pi^*_{\mathcal{S}}}.\end{aligned} \tag{61}$$

Combining (60) and (61) completes the proof. □

We would like to point out that in the above analysis one can replace the dependence on $\|\mathbf{X}\|_{\infty,p^*}$ with a dependence on $\|\mathbf{X}\|_{2,p^*}$ at the expense of a slower rate of convergence (in terms of m). In order to do this we use Proposition 1 to bound the right hand side of (59) as:

$$\|\mathbf{X}_t^\top\|_{2,p^*}\leq\max(1,m^{\frac{1}{p^*}-\frac{1}{2}})\|\mathbf{X}\|_{p^*,2}.$$

Substituting the above bound into the analysis we get the following corollary.

Corollary 1. *Consider the family of functions $\mathcal{G}_p^n$ with $p\in[1,\infty)$, activation function $\rho(z)=(z)_+$, and perturbations in r-norm for $1<r<\infty$. Assume that for our sample $\|\mathbf{x}_i\|_r\geq\epsilon$. Then, the following upper bound on the Rademacher complexity holds:*

$$\widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G}_p^n)\leq\left[\frac{W\Lambda\max(1,d^{1-\frac{1}{p}-\frac{1}{r}})\Big(K(p,d)\max(1,m^{\frac{1}{p^*}-\frac{1}{2}})\|\mathbf{X}\|_{p^*,2}+\epsilon\Big)}{\sqrt{m}}\right]C^*_{\mathcal{S}}\Pi^*_{\mathcal{S}},$$

F.2. Bounding $\Pi^*_{\mathcal{S}}$.

Notice that a key data dependent quantity that controls the Rademacher complexity bound in the previous analysis is $\Pi^*_{\mathcal{S}}$, i.e., the maximum number of partitions that $\mathcal{S}$ can induce on the weights $\mathbf{w}_1,\ldots,\mathbf{w}_k$. As mentioned in Section 6.2 our notion of ϵ-adversarial shattering provides a general way to bound $\Pi^*_{\mathcal{S}}$. We restate the definition of ϵ-adversarial shattering here and then discuss its implications.

Definition 4. *Fix the sample $\mathcal{S}=((\mathbf{x}_1,y_1)\ldots(\mathbf{x}_m,y_m))$ and $(\mathbf{w}_1,\ldots,\mathbf{w}_n)$. Let $\mathbf{s}_i=\operatorname{argmin}_{\|\mathbf{s}\|_r\leq1}y_i\sum_{j=1}^n u_j(\mathbf{w}_j\cdot(\mathbf{x}_i+\epsilon\mathbf{s}))_+$, and define the following three sets:*

$$\begin{aligned}P_i&=\{j\colon\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s}_i)>0\}\\ Z_i&=\{j\colon\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s}_i)=0\}\\ N_i&=\{j\colon\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s}_i)<0\}.\end{aligned}$$

Let $\Pi_{\mathcal{S}}(\mathbf{W})$ be the number of distinct (P_i,Z_i,N_i)s that are induced by $\mathcal{S}$, where $\mathbf{W}$ is a matrix that admits the $\mathbf{w}_j$s as columns. We call $\Pi_{\mathcal{S}}(\mathbf{W})$ the ϵ-adversarial growth function. *We say that $\mathbf{W}$ is* ϵ-adversarially shattered *if every $P\subset[n]$ is possible.*

We will further study the above notion of ϵ-adversarial shattering to bound $\Pi^*_{\mathcal{S}}$ under assumptions on the weight matrix $\mathbf{W}$. In particular, we will be interested in vectors $\mathbf{w}_1,\ldots,\mathbf{w}_n$ such that for all $i\in[n]$, the set Z_i is empty. In this case we say that $\mathbf{W}$ is ϵ-adversarially shattered if every partition of the weights into sets P_i,N_i is possible. For this setting, we state below a lemma that is analogous to Sauer's lemma in statistical learning theory (Sauer, 1972; Shelah, 1972) and helps us bound the ϵ-adversarial growth function $\Pi_{\mathcal{S}}(\mathbf{W})$.

Lemma 11. *Fix an integer $t\geq1$. Fix a sample $\mathcal{S}=((\mathbf{x}_1,y_1)\ldots(\mathbf{x}_m,y_m))$ and weights $\mathbf{w}_1,\ldots,\mathbf{w}_n$ such that for all $i\in[n]$, $Z_i=\varnothing$, and no subset of the weights of size more than t can be ϵ-adversarially shattered by $\mathcal{S}$. Then it holds that*

$$\Pi_{\mathcal{S}}(\mathbf{W})\leq\sum_{i=0}^t\binom{n}{i}. \tag{62}$$

Proof. The proof is similar to the proof of Sauer's lemma (Sauer, 1972; Shelah, 1972) and use an induction on $n+t$.

Base Case. We first show that for $n = 0$ and any t,

$$\Pi_{\mathcal{S}}(\mathbf{W}) \leq \sum_{i=0}^{t} \binom{0}{i} = 1.$$

This easily follows since if $n = 0$, there is no set to shatter. Next, we show that for $t = 0$ and any n,

$$\Pi_{\mathcal{S}}(\mathbf{W}) \leq \sum_{i=0}^{0} \binom{n}{i} = 1.$$

The above holds since if no set of size one can be shattered, then all the points in $\mathcal{S}$ fall in a single part of the partition.

Inductive Step. Let $n + t = k$ and assume that (62) holds for all n, t with $n + t < k$. Notice that $\Pi_{\mathcal{S}}(\mathbf{W})$ is simply the maximum number of labelings of W that can be induced by $\mathcal{S}$. Let A be the set of all such labelings and let A' be the smallest subset of A that induces the maximal number of different labelings on $\mathbf{w}_2, \ldots, \mathbf{w}_n$. Notice that A' cannot shatter more than t of the weights in $\mathbf{w}_2, \ldots, \mathbf{w}_n$. Furthermore, $A \smallsetminus A'$ cannot shatter more than $t - 1$ of the weights, since any labeling in $A \smallsetminus A'$ has a corresponding labeling in A with opposite label on $\mathbf{w}_1$. Hence, if $A \smallsetminus A'$ shatters more than $t - 1$ of the weights in $\mathbf{w}_2, \ldots, \mathbf{w}_n$ then we get that A shatters more than t of the weights in $\mathbf{w}_1, \ldots, \mathbf{w}_n$. Finally, using the induction hypothesis we get that

$$\begin{aligned}
\Pi_{\mathcal{S}}(\mathbf{W}) &= |A| \\
&= |A'| + |A \smallsetminus A'| \\
&\leq \sum_{i=0}^{t} \binom{n-1}{i} + \sum_{i=0}^{t-1} \binom{n-1}{i} \\
&= \sum_{i=0}^{t} \binom{n}{i}.
\end{aligned}$$

□

Finally, we end the section by demonstrating that the notion of ϵ-adversarial shattering can lead to dimension independent bounds on $\Pi^*_{\mathcal{S}}$ under certain assumptions. We believe that this notion warrants further investigation and is key in deriving dimension independent bounds for more general setting. Below we analyze a special case of orthogonal vectors.

Lemma 12. *Fix $p > 1$. Let $\mathcal{S} = ((\mathbf{x}_1, y_1) \ldots (\mathbf{x}_m, y_m))$ be a sample and $\mathbf{w}_1, \ldots \mathbf{w}_t$ be a set of weight vectors. Let $\mathbf{W}$ be the matrix with $\mathbf{w}_i$s as columns. Furthermore, we make the following assumptions*

1. $\|\mathbf{w}_j\|^2 \geq w^2_{\mathit{min}}$ *for all* $j \in [t]$.
2. $\mathbf{w}_j \cdot \mathbf{w}_k = 0$ *for all* $j \neq k$.
3. $\|\mathbf{W}^\top\|_{2,p^*} \leq \tau$.
4. $u_j = 1$.

If $\mathcal{S}$ ϵ-adversarially shatters $\mathbf{w}_1, \ldots \mathbf{w}_t$ with perturbations measured in $r = 2$ norm then it holds that

$$t \leq \frac{4\tau^2 c_2^2(p^*) \|\mathbf{X}\|^2_{p,\infty}}{\epsilon^2 w^2_{\mathit{min}}},$$

where the constant $c_2(p^*)$ (as in Lemma 3) is defined as,

$$c_2(p^*) := \sqrt{2}\Big(\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\Big)^{\frac{1}{p^*}}.$$

Proof. For orthogonal $\mathbf{w}_j$'s, Theorem 9 implies that $Z_i = \varnothing$. Thus, the optimal perturbation is characterized by

$$\mathbf{s}_i^* = -\frac{\sum_{j\in P_i}\mathbf{w}_j}{\|\sum_{j\in P}\mathbf{w}_i\|_2}$$

In the following, it will be more convenient to work with the negative of this quantity, so we define

$$\mathbf{s}_i = -\mathbf{s}_i^* = \frac{\sum_{j\in P_i}\mathbf{w}_j}{\|\sum_{j\in P_i}\mathbf{w}_j\|_2}.$$

For a given shattering P_i, N_i by an example $\mathbf{x}_i$ the following holds:

$$\forall j \in P_i, (\mathbf{w}_j\cdot\mathbf{x}_i - \epsilon\mathbf{w}_j\cdot\mathbf{s}_i) > 0 \tag{63}$$

$$\forall j \in N_i, (\mathbf{w}_j\cdot\mathbf{x}_i - \epsilon\mathbf{w}_j\cdot\mathbf{s}_i) < 0. \tag{64}$$

Next, we define $\mathbf{W}^+$ and $\mathbf{W}^-$ as follows:

$$\mathbf{W}^+ = \sum_{j\in P_i}\mathbf{w}_j$$
$$\mathbf{W}^- = \sum_{j\in N_i}\mathbf{w}_j.$$

Furthermore, let $\Delta\mathbf{W} = \mathbf{W}^+ - \mathbf{W}^-$. Then summing over the inequalities in (63) and (64) we can write:

$$\begin{aligned}\Delta\mathbf{W}\cdot\mathbf{x}_i &> \epsilon\Delta\mathbf{W}\cdot\mathbf{s}_i\\ &= \epsilon\frac{\Delta\mathbf{W}\cdot\mathbf{W}^+}{\|\mathbf{W}^+\|_2}\end{aligned}$$

Using the fact that $|\Delta\mathbf{W}\cdot\mathbf{x}_i| \le \|\Delta\mathbf{W}\|_{p^*}\|\mathbf{X}\|_{p,\infty}$ we can write:

$$\|\mathbf{W}^+\|_2\|\mathbf{W}\|_{p^*}\|\mathbf{X}\|_{p,\infty} > \epsilon\Delta\mathbf{W}\cdot\mathbf{W}^+. \tag{65}$$

Since $\mathcal{S}$ ϵ-adversarially shatters $\mathbf{W}$, (65) must hold for every partition P_i, N_i, and hence must hold in expectation over the random partition as well. Hence, introducing Rademacher random variables $\sigma_1, \ldots, \sigma_t$ we can write:

$$\underset{\boldsymbol{\sigma}}{\mathbb{E}}\big[\|\mathbf{W}^+\|_2\|\Delta\mathbf{W}\|_{p^*}\|\mathbf{X}\|_{p,\infty}\big] > \epsilon\underset{\boldsymbol{\sigma}}{\mathbb{E}}\big[\Delta\mathbf{W}\cdot\mathbf{W}^+\big], \tag{66}$$

where $\mathbf{W}^+ = \sum_{j=1}^t 1_{\sigma_j>0}\mathbf{w}_j$ and $\Delta\mathbf{W} = \sum_{j=1}^t \sigma_j\mathbf{w}_j$. We bound the right-hand side in (66) above as

$$\epsilon\underset{\boldsymbol{\sigma}}{\mathbb{E}}\big[\Delta\mathbf{W}\cdot\mathbf{W}^+\big] = \epsilon\underset{\boldsymbol{\sigma}}{\mathbb{E}}\Big[\Big(\sum_{j=1}^t\sigma_j\mathbf{w}_j\Big)\Big(\sum_{j=1}^t\sigma_j 1_{\sigma_j>0}\mathbf{w}_j\Big)\Big] \tag{67}$$

$$= \epsilon\sum_{j,k=1}^t\mathbb{E}[1_{\sigma_j>0}\sigma_k]\mathbf{w}_j\cdot\mathbf{w}_k$$

$$= \epsilon\Big(\sum_{j\neq k}\mathbb{E}[1_{\sigma_j>0}]\,\mathbb{E}[\sigma_k]\mathbf{w}_j\cdot\mathbf{w}_k + \sum_{j=1}^t\mathbb{E}[1_{\sigma_j>0}]\mathbf{w}_j\cdot\mathbf{w}_j\Big)$$

$$= \frac{\epsilon}{2}\sum_{j=1}^t\|\mathbf{w}_j\|^2. \tag{68}$$

Next, using Cauchy-Schwarz inequality we upper bound the left hand side of (66) as

$$\begin{aligned}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\big[\|\mathbf{W}^+\|_2\|\Delta\mathbf{W}\|_{p^*}\|\mathbf{X}\|_{p,\infty}\big] &\le \sqrt{\underset{\boldsymbol{\sigma}}{\mathbb{E}}[\|\mathbf{W}^+\|_2^2]}\sqrt{\underset{\boldsymbol{\sigma}}{\mathbb{E}}[\|\Delta\mathbf{W}\|_{p^*}^2]}\|\mathbf{X}\|_{p,\infty}\\ &\le \sqrt{\sum_{j=1}^t\mathbb{E}[1_{\sigma_j>0}]\|\mathbf{w}_j\|^2}\sqrt{\underset{\boldsymbol{\sigma}}{\mathbb{E}}[\|\Delta\mathbf{W}\|_{p^*}^2]}\|\mathbf{X}\|_{p,\infty} \text{ [Using orthogonality of the } \mathbf{w}_j \text{ vectors.]}\\ &= \sqrt{\frac{1}{2}\sum_{j=1}^t\|\mathbf{w}_j\|^2}\sqrt{\underset{\boldsymbol{\sigma}}{\mathbb{E}}[\|\Delta\mathbf{W}\|_{p^*}^2]}\|\mathbf{X}\|_{p,\infty}.\end{aligned} \tag{69}$$

Furthermore, since $p^* > 1$, using the analysis in Section A and the Khintchine-Kahane inequality (Haagerup, 1981):

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\|\Delta\mathbf{W}\|_{p^*}^2] &\leq 2\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\|\Delta\mathbf{W}\|_{p^*}]^2 \\ &= 2\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\|\sum_{j=1}^{t}\sigma_j\mathbf{w}_j\|_{p^*}]^2 \\ &\leq 2c_2^2(p^*)\|\mathbf{W}^\top\|_{2,p^*}^2 \\ &\leq 2c_2^2(p^*)\tau^2. \end{aligned} \tag{70}$$

Combining (68), (69) and (70) we can write:

$$\epsilon\sqrt{\frac{1}{2}\sum_{j=1}^{t}\|\mathbf{w}_j\|^2} < \sqrt{2}c_2(p^*)\tau\|\mathbf{X}\|_{p,\infty}.$$

From our assumption we also have that $\|\mathbf{w}_j\|^2 \geq w_{\min}^2$ for all $j \in [t]$. Substituting above we get

$$\epsilon \cdot w_{\min}\sqrt{\frac{t}{2}} < \sqrt{2}c_2(p^*)\tau\|\mathbf{X}\|_{p,\infty}.$$

Rearranging, we get that

$$t \leq \frac{4c_2^2(p^*)\tau^2\|\mathbf{X}\|_{p,\infty}^2}{\epsilon^2 w_{\min}^2}.$$

□