Adversarial Learning Guarantees for Linear Hypotheses and Neural Networks

Pranjal Awasthi [1] **Natalie S. Frank** [2] **Mehryar Mohri** [3]

Abstract

Adversarial or test time robustness measures the susceptibility of a classifier to perturbations to the test input. While there has been a flurry of recent work on designing defenses against such perturbations, the theory of adversarial robustness is not well understood. In order to make progress on this, we focus on the problem of understanding generalization in adversarial settings, via the lens of Rademacher complexity. We give upper and lower bounds for the adversarial empirical Rademacher complexity of linear hypotheses with adversarial perturbations measured in l_r-norm for an arbitrary $r \geq 1$. We then extend our analysis to provide Rademacher complexity lower and upper bounds for a single ReLU unit. Finally, we give adversarial Rademacher complexity bounds for feed-forward neural networks with one hidden layer.

1. Introduction

Robustness is a key requirement when designing machine learning models and comes in various forms such as robustness to training set corruptions, missing feature values, and model misspecification. In recent years, requiring robustness to *adversarial* or *test time* perturbations has become a key requirement. Starting with the work of Szegedy et al. (2014) it has now been well established that deep neural networks trained via standard gradient descent based algorithms are highly susceptible to imperceptible corruptions to the input at test time (Goodfellow et al., 2014; Chen et al., 2017; Eykholt et al., 2018; Carlini & Wagner, 2018). This has led to a proliferation of work aimed at designing classifiers robust to such perturbations (Madry et al., 2017; Gowal et al., 2018; 2019; Schott et al., 2018) and works aimed at designing more sophisticated attacks to break such classifiers (Athalye et al., 2018; Carlini & Wagner, 2017; Sharma & Chen, 2017)

While the above works have made significant progress in designing practical defenses, theoretical aspects of adversarial robustness are currently poorly understood unlike other notions of training set corruptions that have been widely studied in both the statistics and the computer science communities (Huber, 2011; Kearns & Li, 1993; Kearns et al., 1994). Theoretical understanding of adversarial robustness presents three main challenges. The first is a computational one since even checking the robustness of a given model at a given test input is an NP-hard problem (Awasthi et al., 2019). This has been explored in recent works that construct specific instances of learning problems where standard non-robust learning can be done efficiently, but learning a robust classifier becomes computationally hard (Bubeck et al., 2018b;a; Nakkiran, 2019; Degwekar et al., 2019). The second challenge concerns whether achieving adversarial robustness requires one to compromise on standard accuracy. Recent works have shown specific instance where this tradeoff is inherent (Tsipras et al., 2018; Raghunathan et al., 2019).

Finally, the third challenge, the main focus of this work, is the question of what quantity governs generalization in adversarial settings, and how generalization in adversarial settings compares to its non-adversarial counterpart. The recent work of Schmidt et al. (2018) has shown, via specific constructions, that in some scenarios achieving adversarial generalization requires more data as compared to adversarial generalization. Furthermore, the work of Montasser et al. (2019) casts a shadow of doubt on the use of classical quantities such as the VC-dimension of explain generalization in adversarial settings.

However, generalization for function classes of infinite VC-dimension (such as SVMs with a Gaussian kernel) can be explained via margin-based bounds. Characterizing the Rademacher complexity of the function class is essential in these estimates. In a similar vein, we believe that providing non-trivial bounds on the adversarial Rademacher complexity can help shed light on when generalization is possible in

[1]Google Research and Rutgers University [2]Courant Institute of Math. Sciences [3]Google Research and Courant Institute of Math. Sciences. Correspondence to: Natalie S. Frank <nf1066@nyu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

adversarial settings via similar margin-based bounds. The difficulty is that current bounds on adversarial Rademacher complexity are too loose and vacuous in many settings. This is the barrier that we aim to overcome in this work.

In order to make progress on the mystery of adversarial generalization, a recent line of work (Khim & Loh, 2018; Yin et al., 2019) aims to study the notion of Rademacher complexity for various function classes in the adversarial settings. Focusing mainly on the case of linear models, these works aim to quantify the additional overhead in sample complexity that is incurred when requiring adversarial generalization. Extending the ideas to the case of more general neural networks becomes more challenging and as a result there works instead bound the Rademacher complexity in terms of the Rademacher complexity of an appropriate surrogate. In this work we extend this line of work along several directions.

Our Contributions. We provide a general analysis of the adversarial Rademacher complexity of linear models that holds for perturbations measured in any ℓ_p norm. This extends the prior work of Yin et al. (2019) that applies only to ℓ_∞ adversarial perturbation and provided a finer analysis of linear models as compared to the work of Khim & Loh (2018).

As a consequence of our analysis, we provide a sharp characterization of when the adversarial Rademacher complexity suffers from an additional dimension dependent term as compared to its non-adversarial counterpart. This has algorithmic implications for designing appropriate regularizers for adversarial learning of linear models. As an additional byproduct, we are able to provide improved Rademacher complexity bounds for linear classifiers, even in non-adversarial scenarios!

As a next step towards understanding neural networks, we then extend our analysis to provide data dependent upper and lower bounds on the adversarial Rademacher complexity of a single ReLU unit.

Finally, we provide upper bounds on the adversarial Rademacher complexity of one hidden layer neural networks. In contrast with prior works (Yin et al., 2019; Khim & Loh, 2018), our bounds directly apply to the original network as opposed to a surrogate. Our bounds for neural networks come in two forms. We first provide a general upper bound that applies to any neural network with Lipschitz activations. This bound as a dependence on the underlying dimensionality of the input data. Next, we provide a finer data dependent upper bound that is related to the ϵ-adversarial growth function of the data, a quantity we introduce in this work.

Comparison with Prior Work Yin et al. (2019) and Khim and Loh (2018) previously studied the adversarial Rademacher complexity of linear classifiers and neural networks. Our work adds to this line of research in multiple ways.

Yin et al. (2019) analyzed the adversarial Rademacher complexity of linear models when perturbations are measured in ℓ_∞ norm. They show that in this case the adversarial Rademacher complexity of the loss class is bounded above and below by the sum of its non-adversarial counterpart and a dimension dependent term. Our result is a strict generalization of (Yin et al., 2019) because we provide the analysis of adversarial Rademacher complexity when the perturbations are measured in any general ℓ_r norm.

The recent work of Khim and Loh (2018) also studies the adversarial Rademacher complexity of classes of linear models constrained by the 2-norm under general ℓ_r perturbations. They noted that the upper bound on the adversarial Rademacher complexity involves a constant that depends on r. While the bounds are qualitatively similar, they did not derive the relationship with the dimension d nor did they give a lower bound involving this term. ((Yin et al., 2019) noted the relationship to d only for $r = \infty$.) Further, we consider models bounded by any ℓ_p norm in addition to the 2-norm, and this consideration turns out to have important implications for generalization and model selection. In the process, we improve upon the existing classical analysis of (non-adversarial) Rademacher complexity of linear models, which is of independent interest.

For the case of neural networks, both the works of Yin et al. (2019) and Khim and Loh (2018) replace the adversarial loss defined as $\min_{\mathbf{x}':\|\mathbf{x}'-\mathbf{x}\|\leq\epsilon} \phi(yf(\mathbf{x}'))$, by a surrogate upper bound and analyze the resulting Rademacher complexity of the surrogate. In general, these bounds on the surrogate might not lead to meaningful generalization bounds on the original adversarial loss.

In the work of Yin et al. (2019) the surrogate is chosen to be an upper bound on the adversarial loss based on a semi-definite programming (SDP) based relaxation. However, the relaxation is quite weak and is obtained via taking the worst possible $\mathbf{x}$ in the domain of the input. This approach greatly overestimates the adversarial loss.

In the work of (Khim & Loh, 2018) the surrogate is based on the adversarial loss of another neural network that is derived from the original one via a tree based decomposition. A key issue in adversarial robustness is the analysis of the optimization problem: $\sup_{\mathbf{s}} f(\mathbf{x}+\mathbf{s})$, where $\mathbf{x}$ is a given input and $\mathbf{s}$ is a perturbation chosen in a small ball. Understanding the structure of the optimal $\mathbf{s}$ above is hard in the case of neural networks as various correlations exist among the outputs of different neurons. (Khim & Loh, 2018) get around this by ignoring correlations and assuming that each path in the network can be independently optimized using a different perturbation $\mathbf{s}$, which can often lead to vacuous results. For instance, consider a one-layer network, i.e. a linear combination of neurons. The network as a whole might be a large-margin classifier and hence robust on most inputs. However, if each neuron is a weakly correlated feature, then it can be independently attacked to induce a large loss. As a result, the approach of (Khim & Loh, 2018) greatly overestimates the adversarial loss in this case.

We avoid making such approximations. This requires a deeper analysis as in Sections 4 and 6, finally leading to adversarial shattering that could directly provide dimension-independent bounds. There have also been recent efforts to understand inference and generalization in adversarial settings for finite perturbation sets (Feige et al., 2015; Attias et al., 2019). Finally, the recent work of Wei & Ma (2019) provides generalization bounds for robust classification via studying a notion of layer-wise margin. This result is incomparable to our work as we aim to directly characterize the adversarial Rademacher complexity.

2. Preliminaries

We will denote vectors as lowercase bold letters (e.g., $\mathbf{x}$) and matrices as uppercase bold (e.g., $\mathbf{X}$). The all-ones vector is denoted by $\mathbf{1}$ and Hölder conjugates by a star (e.g., r^*). For a matrix $\mathbf{M}$, the (p,q)-*group norm* is defined as $\|\mathbf{M}\|_{p,q} = \|(\|\mathbf{M}_1\|_1, \ldots, \|\mathbf{M}_d\|_p)\|_q$, where the $\mathbf{M}_i$s are the columns of $\mathbf{M}$.

We focus on binary classification over examples in $\mathbb{R}^d$ and adversarial perturbations measured in ℓ_r-norm for $r \geq 1$. Given a loss function $\ell\colon \mathbb{R} \to [0,c]$, we define the loss of a hypothesis $f\colon \mathbb{R}^d \to \mathbb{R}$ on a pair $(\mathbf{x}, y) \in \mathbb{R}^d \times \{+1,-1\}$ as $\ell_f(\mathbf{x}, y) = \ell(yf(\mathbf{x}))$. As in the familiar classification setting, given a sample $\mathcal{S} = ((\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_m,y_m))$ drawn i.i.d. from a distribution $\mathcal{D}$ over $\mathbb{R}^d \times \{+1,-1\}$, we define the empirical risk and the expected risk of a hypothesis f as

$$R_{\mathcal{S}}(f) = \frac{1}{m}\sum_{i=1}^m \ell_f(\mathbf{x}_i, y_i) \quad R(f) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell_f(\mathbf{x},y)].$$

Let $\mathcal{F}$ be a family of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}$. Then, the *empirical Rademacher complexity* of $\mathcal{F}$ for a sample $\mathcal{S} = (\mathbf{x}_1, \ldots \mathbf{x}_m)$, is defined by

$$\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(\mathbf{x}_i)\right], \tag{1}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)$ is a vector of i.i.d. Rademacher variables, that is independent uniform random variables taking values in $\{-1,+1\}$. The *Rademacher complexity* of $\mathcal{F}$, $\mathfrak{R}_m(\mathcal{F})$, is defined as the expectation of this quantity: $\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}[\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})]$, where $\mathcal{D}$ is a distribution over the input space $\mathbb{R}^d$. The empirical Rademacher complexity is a key data-dependent complexity measure. For a family of functions $\mathcal{F}$ taking values in $[0,1]$, the following learning guarantee holds: for any $\delta > 0$, with probability at least $1-\delta$ over the draw of a sample $S \sim \mathcal{D}^m$, the following inequality holds for all $f \in \mathcal{F}$ (Mohri et al., 2018):

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[f(x)] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{S}}[f(x)] + 2\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}, \tag{2}$$

where we denote by $\mathbb{E}_{x\sim\mathcal{S}}[f(x)]$ the empirical average of f, that is $\mathbb{E}_{x\sim\mathcal{S}}[f(x)] = \frac{1}{m}\sum_{i=1}^m f(x_i)$. A similar inequality holds for the average Rademacher complexity $\mathfrak{R}_m(\mathcal{F}_p) = \mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m}[\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})]$:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[f(x)] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{S}}[f(x)] + 2\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Furthermore, the Rademacher complexity of a hypothesis set also appears as a lower bound in generalization. As an example, for a symmetric family of functions $\mathcal{F}$ taking values in $[-1,+1]$, the following holds (van der Vaart & Wellner, 1996):

$$\frac{1}{2}\left[\mathfrak{R}_m(\mathcal{F}) - \frac{1}{\sqrt{m}}\right] \leq \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\sup_{f\in\mathcal{F}}\left|\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[f(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{S}}[f(x)]\right| \leq 2\mathfrak{R}_m(\mathcal{F}).$$

An important application of these bounds is the derivation of margin bounds, which are crucial in the analysis of classification. Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of a sample $S \sim \mathcal{D}^m$, the following inequality holds for all $f \in \mathcal{F}$ (Koltchinskii & Panchenko, 2002; Mohri et al., 2018):

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[1_{yf(x)\leq 0}] \leq \frac{1}{m}\sum_{i=1}^m 1_{y_i f(x_i)\leq\rho} + \frac{2}{\rho}\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{3}$$

Finer margin guarantees were recently presented by Cortes et al. (2020) in terms of Rademacher complexity and other complexity measures.

Robust Classification. We now extend the definitions above to their adversarial counterparts. In the setting of adversarially robust classification, the loss at $(\mathbf{x}, y)$ is measured in terms of the worst loss incurred over an adversarial perturbation of $\mathbf{x}$ within an ball of a certain radius in a norm $\|\cdot\|$. There are many possible norms we could use to measure a perturbation. A fairly general way to measure perturbations is the ℓ_r norm denoted $\|\cdot\|_r$. We will denote by ϵ the maximum magnitude of the allowed perturbations. Given $\epsilon > 0$, $r \geq 1$, a data point $(\mathbf{x}, y)$, a function $f\colon \mathbb{R}^d \to \mathbb{R}$, and a loss function $\ell\colon \mathbb{R} \to [0,c]$ we define the adversarial loss of f at $(\mathbf{x}, y)$ as

$$\widetilde{\ell}_f(\mathbf{x}, y) = \sup_{\|\mathbf{x}-\mathbf{x}'\|_r \leq \epsilon} \ell(yf(\mathbf{x}')).$$

Similarly, we define the adversarial empirical risk and the adversarial expected risk of a hypothesis f for a sample $\mathcal{S}$ as follows:

$$\widetilde{R}_{\mathcal{S}}(f) = \frac{1}{m}\sum_{i=1}^{m}\tilde{\ell}_f(\mathbf{x}_i, y_i) \quad \widetilde{R}(f) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\tilde{\ell}_f(\mathbf{x}, y)].$$

We also define $\widetilde{\mathfrak{R}}(\mathcal{F})$, the *adversarial Rademacher complexity*, as the adversarial version of Rademacher complexity:

$$\widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i \sup_{\|\mathbf{x}_i-\mathbf{x}_i'\|_r\le\epsilon} f(\mathbf{x}_i')\right].$$

With the above definitions, the following is a consequence of (2) above and Talagrand's contraction lemma (Ledoux & Talagrand, 1991).

Theorem 1. *Let $\mathcal{S}$, $\mathcal{D}$, δ, and $\mathcal{F}$ be as in equation (2). Further let ℓ be a Lipschitz loss function and define the class*

$$\ell_{\mathcal{F}} = \{\ell\circ f : f\in\mathcal{F}\}.$$

Then, with probability at least $1-\delta$ the following holds for all $f\in\mathcal{F}$:

$$\widetilde{R}(f) \le \widetilde{R}_{\mathcal{S}}(f) + 2c\cdot\widetilde{\mathfrak{R}}_{\mathcal{S}}(\ell_{\mathcal{F}}) + 3c\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{4}$$

Throughout the paper, we will assume that the loss function ℓ is non-increasing, a property satisfied by many common loss functions including the hinge loss, logistic loss and the exponential loss. In that case, as pointed out by Yin et al. (2019), the following equality holds:

$$\sup_{\|\mathbf{x}_i-\mathbf{x}_i'\|_r\le\epsilon}\ell(y_i f(\mathbf{x}_i')) = \ell\Big(\inf_{\|\mathbf{x}_i-\mathbf{x}_i'\|_r\le\epsilon} y_i f(\mathbf{x}_i')\Big).$$

Furthermore, when $\ell(\cdot)$ is L-Lipschitz, by Talagrand's Lemma, we have $\widetilde{\mathfrak{R}}_{\mathcal{S}}(\ell_{\mathcal{F}}) \le L\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}})$, where $\widetilde{\mathcal{F}}$ is the class defined by

$$\widetilde{\mathcal{F}} = \Big\{(\mathbf{x}, y)\mapsto \inf_{\|\mathbf{x}-\mathbf{x}'\|_r\le\epsilon} y f(\mathbf{x}') : f\in\mathcal{F}\Big\}. \tag{5}$$

Thus, we obtain the following inequalities:

$$\begin{aligned}\widetilde{\mathfrak{R}}_{\mathcal{S}}(\ell_{\mathcal{F}}) &\le L\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}})\\ &= L\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i \inf_{\|\mathbf{x}_i-\mathbf{x}_i'\|_r\le\epsilon} y_i f(\mathbf{x}_i')\right]. \end{aligned} \tag{6}$$

Providing sharp bounds on $\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}})$ for various function classes $\mathcal{F}$ will be the central focus of this work.

As an application, we can derive *robust margin bounds*:

Theorem 2 (Robust margin bounds). *Let $\mathcal{F}, \mathcal{S}, \mathcal{D}$, and δ be as in equation (3). Further let be $\widetilde{\mathcal{F}}$ as in equation (5). Then, with probability at least $1-\delta$ the following holds for all $f\in\mathcal{F}$:*

$$\widetilde{R}(f) \le \widetilde{R}_{\mathcal{S},\rho}(f) + \frac{2}{\rho}\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{7}$$

3. Adversarial Rademacher Complexity of Linear Hypotheses

In this section, we provide a sharp characterization of the adversarial Rademacher complexity, as defined in (6), for linear function classes with bounded ℓ_p-norm and with perturbations measured in any ℓ_r-norm. Prior work (Yin et al., 2019) studied the case where the perturbations are measured in the ℓ_∞-norm. Our general analysis leads to a deeper understanding of the interplay between the complexity of the hypothesis classes (measured in ℓ_p-norm) and the perturbation set (measured in ℓ_r-norm), and how this dictates whether one can expect an additional dimension dependent penalty in the adversarial case over its non-adversarial counterpart. Furthermore, our analysis explicitly characterizes the dimension dependent term on which the adversarial Rademacher complexity depends on. This provides a finer analysis than the work of Khim & Loh (2018) and also has algorithmic implications. Formally, we study the case when

$$\mathcal{F}_p = \{\mathbf{x}\mapsto\langle\mathbf{w},\mathbf{x}\rangle : \|\mathbf{w}\|_p\le W\}. \tag{8}$$

3.1. Rademacher Complexity of Linear Hypotheses

A crucial aspect of our analysis in the linear case is a more general upper bound on the Rademacher complexity of p-norm bounded linear function classes, in the non-adversarial case. We first state this general bound as it will play an important role in later sections when analyzing the adversarial Rademacher complexity of ReLU functions and more general neural networks.

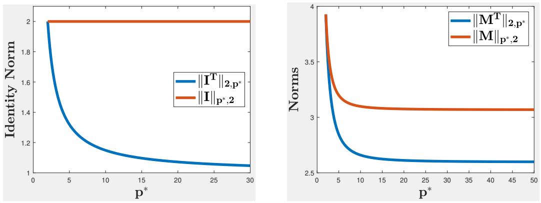
Theorem 3. *Let $\mathcal{F}_p$ be the class of functions defined in (8). Then, given a sample $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ we have*

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) \le \begin{cases}\frac{W}{m}\sqrt{2\log(2d)}\,\|\mathbf{X}^\top\|_{2,p^*} & \textit{if } p=1\\ \frac{\sqrt{2}W}{m}\left[\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\right]^{\frac{1}{p^*}}\|\mathbf{X}^\top\|_{2,p^*} & \textit{if } 1<p\le 2\\ \frac{W}{m}\|\mathbf{X}^\top\|_{2,p^*} & \textit{if } p\ge 2\end{cases}$$

Here, $\mathbf{X}$ is the $d\times m$ matrix with the data points $\mathbf{x}_i$ as columns. We make a few remarks about the theorem above and defer its proof to Appendix A.2. Some well-known bounds on the Rademacher complexity of $\mathcal{F}_p$ are

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) \le \begin{cases}W\sqrt{\frac{2\log(2d)}{m}}\|\mathbf{X}\|_{\max} & \text{if } p=1\\ \frac{W}{m}\sqrt{p^*-1}\|\mathbf{X}\|_{p^*,2} & \text{if } 1<p\le 2\end{cases} \tag{9}$$

Although the case $p\in[1,2]$ in the theorem above is known (Kakade et al., 2008; Mohri et al., 2018), we provide a simpler proof of in Appendix A.1. The inequality for $p=1$ is



(a) (b)

Figure 1: (a) A plot comparing two norms of the 4×4 identity matrix, $\|\mathbf{I}^\top\|_{2,p^*}$ and $\|\mathbf{I}\|_{p^*,2}$; the lower bound on the ratio of the two norms (10) in Proposition 1 holds for this matrix. (b) Same as (a), but for Gaussian matrices.

further reproduced for completeness. Our new bound coincides with (9) when $p = 2$ and is strictly better otherwise. Readers familiar with Rademacher complexity bounds for linear functions will notice that our bound in this case depends on the norm $\|\mathbf{X}^\top\|_{2,p^*}$. In contrast, standard bounds on the Rademacher complexity of linear classes depend on $\|\mathbf{X}\|_{p^*,2}$. In fact one can show that the $\|\mathbf{X}^\top\|_{2,p^*}$ is always smaller than $\|\mathbf{X}\|_{p^*,2}$ for $p \in (1,2]$, that is $p^* \geq 2$, as shown by the last inequality of (10) in the following proposition.

Proposition 1. *Let $\mathbf{M}$ be a $d \times m$ matrix. If $q \leq p$, then*

$$\min(m,d)^{\frac{1}{p}-\frac{1}{q}}\|\mathbf{M}^\top\|_{p,q} \leq \|\mathbf{M}\|_{q,p} \leq \|\mathbf{M}^\top\|_{p,q} \tag{10}$$

If $q \geq p$, then

$$\min(m,d)^{\frac{1}{p}-\frac{1}{q}}\|\mathbf{M}^\top\|_{p,q} \geq \|\mathbf{M}\|_{q,p} \geq \|\mathbf{M}^\top\|_{p,q} \tag{11}$$

These bounds are tight.

The proof is deferred to Appendix A.4. To visualize the ratio between these two norms, we plot the two norms for various values of p^* in Figure 1. For convenience, in the discussion below, we set $c_1(p) = \sqrt{p^*-1}$ and $c_2(p) = \sqrt{2}\big[\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\big]^{\frac{1}{p^*}}$. Regarding the growth of the constant in our bound, one can show that as $p^* \to \infty$, $c_2(p)$ grows asymptotically like $e^{-\frac{1}{2}}\sqrt{p^*}$. In fact one can show that

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

Furthermore, $c_2(p) \leq c_1(p)$ in the relevant region (see Appendix A.5). In Figure 2 we plot $c_1(p), c_2(p)$ and the bounds on $c_2(p)$ to illustrate the growth rate of these constants with p^*. Proposition 1 and that $c_2(p) \leq c_1(p)$ imply that the bounds we give for linear classes are stronger than what was previously known. We believe that the result in Theorem 3 is of wider interest and a recent companion note (Awasthi et al., 2020) provides a more detailed and self contained exposition.

3.2. Adversarial Rademacher Complexity of Linear Hypotheses

We now extend our bounds from the previous section to provide a complete characterization of the adversarial Rademacher complexity of linear function classes under arbitrary r-norm perturbations. These theorems improve upon the recent work of Yin et al. (2019) that studies ∞-norm perturbations and provide a finer analysis, with a matching lower bound, as compared to the recent work of Khim & Loh (2018). Our main result is stated below.

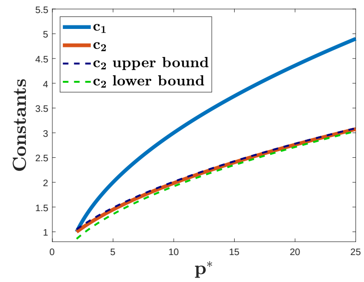


Figure 2: A plot of $c_1(p)$, $c_2(p)$, and the bounds from Lemma 4. Note that $c_1(2) = c_2(2)$ and that the upper and lower bounds on c_2 are tight.

Theorem 4. *Let $\epsilon > 0$, $p, r \geq 1$. Consider a sample $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$. Let $\mathcal{F}_p$ be the class of linear functions defined in (8). Then it holds that*

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

and

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

Notice that when the perturbation is measured in ℓ_∞-norm, i.e. $r = \infty$, we recover the bound of Yin et al. (2019). Hence the theorem above is a strict generalization of the result of Yin et al. (2019). Furthermore, when $\epsilon = 0$, as expected, the adversarial Rademacher complexity equals the standard Rademacher complexity of linear models and we can use our improved bounds from Theorem 3. The theorem above has important implications for the design of regularizers in the context of adversarial learning of linear models. As suggested by the upper bounds above, if $1/r + 1/p \geq 1$, then one can indeed perform adversarially robust learning with minimal statistical overhead in the standard classification setting! More specifically, in this case the upper bound on the adversarial Rademacher complexity has at most $W\epsilon/\sqrt{m}$ overhead on top of the standard bound from Theorem 3 and is dimension independent. Noting that $1 - 1/r = 1/r^*$, we get that for statistical efficiency one should choose a p-norm regularizer on $\mathbf{w}$, where $p \in [1, r^*]$. Our lower bound on the other hand shows that any other choice of a ℓ_p-norm based regularizer will necessarily incur a dimension-dependent penalty.

3.3. Proof sketch of Theorem 4

We provide a brief sketch of the proof of Theorem 4 and provide the details in Appendix B. As a first step, a simple argument shows that

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

Using the above, we can write the adversarial Rademacher complexity as:

$$\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}_p) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle \mathbf{w}, \mathbf{u}_{\boldsymbol{\sigma}}\rangle - \epsilon v_{\boldsymbol{\sigma}}\|\mathbf{w}\|_{r^*}\right] \tag{12}$$

where, for convenience, we set $\mathbf{u}_{\boldsymbol{\sigma}} = \frac{1}{m}\sum_{i=1}^m y_i\sigma_i\mathbf{x}_i$, $v_{\boldsymbol{\sigma}} = \frac{1}{m}\sum_{i=1}^m \sigma_i$. Next, we present two key lemmas.

Lemma 1. *Let $1 \le p, r \le \infty$ and let d be the dimension. Then*

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

Lemma 2. *Let $v_{\boldsymbol{\sigma}} = \frac{1}{m}\sum_{i=1}^m \sigma_i$. Then it holds that*

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} v_{\boldsymbol{\sigma}}\|\mathbf{w}\|_{r^*}\right] \ge \frac{W\epsilon\max(d^{1-\frac{1}{r}-\frac{1}{p}}, 1)}{2\sqrt{2m}},$$

and

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

For the upper bound, using the sub-additivity of supremum and Lemma 2 yields

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

For the lower bound, we apply two symmetrization arguments and show that

$$\begin{aligned}\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}_p) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} -\langle \mathbf{w}, \mathbf{u}_{\sigma}\rangle + \epsilon v_{\boldsymbol{\sigma}}\|\mathbf{w}\|_{r^*}\right] &\quad (13)\\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_p \le W} \langle \mathbf{w}, \mathbf{u}_{\boldsymbol{\sigma}}\rangle + \epsilon v_{\boldsymbol{\sigma}}\|\mathbf{w}\|_{r^*}\right]. &\quad (14)\end{aligned}$$

Averaging equations (12) and (14) and applying the sub-additivity of supremum gives:

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

Now averaging (13) and (14), applying sub-additivity and Lemma 2, the following holds:

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

4. Adversarial Rademacher Complexity of a Rectified Linear Unit

As a first step towards providing a bound for neural networks, in this section we study the adversarial Rademacher complexity of linear functions composed with a rectified linear unit (ReLU). Again we measure the size of functions in p norm and define the function class by

$$\mathcal{G}_p = \{(\mathbf{x}, y) \mapsto (y\langle \mathbf{w}, \mathbf{x}\rangle)_+ : \|\mathbf{w}\|_p \le W, y \in \{-1, 1\}\} \tag{15}$$

where $(z)_+ = \max(z, 0)$. The following theorem presents a data-dependent upper bound on the adversarial Rademacher complexity of the ReLU unit.

Theorem 5. *Let $\mathcal{G}_p$ be the class as defined in (15) and let $\mathcal{F}_p$ be the corresponding 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 $T_\epsilon = \{i : y_i = -1 \text{ or }, y_i = 1 \text{ and } \|\mathbf{x}_i\|_r > \epsilon\}$.

The second term in the bound above is similar to the dimension dependent term that appears in the linear case. The first term is the empirical Rademacher complexity of linear classes with bounded p-norm, but only measured on a carefully chosen subset of the data. This implies that data points with positive labels that have small norm as compared to the perturbation ϵ do not affect the Rademacher complexity. Hence, the guarantee in the theorem treats the two classes $+1$ and -1 asymmetrically.

This phenomenon originates from a property of the function $(z)_+$. Recall that in our setup $(z)_+$ will later be composed with a loss function ℓ. Because $\ell(yz_+)$ is the penalty incurred to the loss, the value yz_+ should be interpreted as a margin. Since the function $\max(0, z)$ is always 0 for $z \le 0$, decreasing z below 0 does not affect the the margin. On the other hand increasing z above zero will increase the margin.

A large margin for a point labeled -1 corresponds to making z_+ as small as possible. As a result, every z with $z \leq 0$ gives the same margin. However, there is no upper bound on the margin for points in the class $+1$. As a result, the classifier $(y, \mathbf{x}) \mapsto y(\langle \mathbf{w}, \mathbf{x} \rangle)_+$ treats all non-negative margins for the class -1 in the same manner, but gives a higher reward for larger margins for the class $+1$.

This observation has implications for adversarial classification; as shown in Appendix C, an adversarially perturbed ReLU is $y \max(\mathbf{w} \cdot \mathbf{x} - \epsilon y \|\mathbf{w}\|_{r^*}, 0)$. If $\mathbf{w} \cdot \mathbf{x}$ is very negative, which corresponds to high confidence for $y = -1$, then the perturbation would not change the value of the loss function. On the other hand, if $\mathbf{w} \cdot \mathbf{x}$ were large and positive, a perturbation would definitely change the value of the margin and then influence the loss. We next complement our upper bound with a data dependent lower bound, stated below, on the adversarial Rademacher complexity.

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

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

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

A natural question that comes to mind is if one can characterize scenarios where the above lower bound leads to a dimension dependent term, as in the lower bound for linear hypotheses. In order to characterize this, for a given s and $\delta > 0$, define the set $T^{\delta}_{\epsilon, \mathbf{s}}$ as

$$T^{\delta}_{\epsilon, \mathbf{s}} = \{i \colon \langle \mathbf{s}, \mathbf{x}_i \rangle - (1 + \delta y_i) y_i \epsilon \|\mathbf{s}\|_{r^*} > 0\}.$$

Notice that $T^{\delta}_{\epsilon, \mathbf{s}}$ is a subset of $T_{\epsilon, \mathbf{s}}$ and contains points in $T^{\delta}_{\epsilon, \mathbf{s}}$ that have a non-trivial margin. Then we get that

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\mathcal{G}_p) &\geq \frac{W}{2\sqrt{2}m} \sup_{\|\mathbf{s}\|_p = 1} \left(\sum_{i \in T^{\delta}_{\epsilon, \mathbf{s}}} (\langle \mathbf{s}, \mathbf{x}_i \rangle - \epsilon y_i \|\mathbf{s}\|_{r^*})^2 \right)^{\frac{1}{2}} \\
&\geq \frac{W}{2\sqrt{2}m} \sup_{\|\mathbf{s}\|_p = 1} \left(\sum_{i \in T^{\delta}_{\epsilon, \mathbf{s}}} (\delta \epsilon \|\mathbf{s}\|_{r^*})^2 \right)^{\frac{1}{2}} \\
&= \frac{W \delta \epsilon}{2\sqrt{2}m} \sup_{\|\mathbf{s}\|_p = 1} |T^{\delta}_{\epsilon, \mathbf{s}}| \|\mathbf{s}\|_{r^*}.
\end{aligned}$$

Denoting $\mathbf{s}^*$ to be the vector that achieves the value $\sup_{\|\mathbf{s}\|_p = 1} \|\mathbf{s}\|_{r^*}$ we get that

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{G}_p) \geq \frac{W \delta \epsilon}{2\sqrt{2}m} |T^{\delta}_{\epsilon, \mathbf{s}^*}| \max(d^{1 - \frac{1}{p} - \frac{1}{r}}, 1).$$

Hence, if for a given constant $\delta > 0$, the size of the set $T^{\delta}_{\epsilon, \mathbf{s}^*}$ is large then we expect a dimension dependent lower bound similar to the linear case.

5. Adversarial Rademacher Complexity of Neural Nets

Building on our analysis for the case of a single ReLU unit, we next give an upper bound on the adversarial Rademacher complexity for the class of one-layer neural networks comprised of a Lipschitz activation ρ with $\rho(0) = 0$. The guarantees of our theorem resemble the bound on the standard Rademacher complexity of neural networks, as provided in (Cortes et al., 2017). An analysis based on other forms of generalization bounds on neural nets is also possible, such as that of Bartlett et al. (2017). The family of functions of such one-layer neural networks is defined as follows:

$$\mathcal{G}^n_p = \Big\{ (\mathbf{x}, y) \mapsto y \sum_{j=1}^{n} u_j \rho(\mathbf{w}_j \cdot \mathbf{x}) \colon \|\mathbf{u}\|_1 \leq \Lambda, \|\mathbf{w}_j\|_p \leq W \Big\}.$$

Our main theorem is stated below.

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

$$\widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G}^n_p) \leq L_\rho \left[\frac{W \Lambda \max(1, d^{1 - \frac{1}{p} - \frac{1}{r}})(\|\mathbf{X}\|_{r, \infty} + \epsilon)}{\sqrt{m}} \right] \times \left(1 + \sqrt{d(n+1) \log(36)} \right).$$

The proof is presented in Appendix D. The only requirements on our activation function ρ is that it is Lipschitz and $\rho(0) = 0$. This stipulation is satisfied by common activation functions like the ReLU, the leaky ReLU, and the hyperbolic tangent, but not the sigmoid or a step function. In comparison to the adversarial Rademacher complexity of linear classifiers, Theorem 7 still includes a $\max(1, d^{1 - \frac{1}{r} - \frac{1}{p}})$ factor, again implying that one should choose a model class with $p \leq r^*$. The complexity of the vector $\mathbf{u}$ is bounded by ℓ_1 norm as that is what turns out to be natural in the proof. However, the dimension dependence is larger by a factor of $\sqrt{d}$. The dependence on the number of neurons $(\sqrt{n})$ is also problematic. This fact is unfortunate since a much larger sample size m would be required for good generalization. In the next section we present a promising approach towards removing the dependence on dimension and the number of neurons in the above bound.

6. Towards Dimension-Independent Bounds

In this section we introduce a new framework for analyzing the adversarial Rademacher complexity of neural networks with ReLU activations. Unlike the case of linear hypotheses, the dimension-dependent term in the upper bound in Theorem 7 cannot be avoided by simply picking the appropriate norm p. In particular, deriving dimension-independent bounds for the adversarial Rademacher complexity of neural networks is a difficult problem. Prior works (Yin et al., 2019; Khim & Loh, 2018) have resorted to bounding the adversarial Rademacher complexity of surrogates that are more tractable. However, it is not clear how those guarantees translate into meaningful bounds on the complexity of the original network. In this section, we present an approach towards obtaining dimension-independent bounds on the adversarial Rademacher complexity of the original network.

A major component of the difficultly in analyzing adversarial Rademacher complexity relates to providing a tight characterization of the optimal adversarial perturbation for a given point $\mathbf{x}_i$, i.e.,

$$\mathbf{s}_i^* = \operatorname*{argmin}_{\mathbf{s}:\|\mathbf{s}\|_r\leq 1} y_i \sum_{j=1}^n u_j(\mathbf{w}_j\cdot(\mathbf{x}_i+\epsilon\mathbf{s}))_+ \tag{16}$$

Thus, to begin, we study properties of such adversarial perturbations to the neural network. Afterwards, we leverage these properties to bound the adversarial Rademacher complexity. Notably, the proofs of these properties heavily rely on the fact that the activation function is ReLU and not any other Lipschitz function. As in the previous section, we will focus on the family of a one layer-network $\mathcal{G}_p^n$ with activation $\rho(z) = z_+$.

6.1. Characterizing Adversarial Perturbations

In this section, we discuss characteristics of adversarial perturbations to neural networks with ReLU activations. The following theorem implies that, if the perturbations are bounded in ℓ_r-norm by ϵ, then typically the optimal adversarial perturbations will have exactly r-norm ϵ.

Theorem 8. *Let d be the dimension and n the number of neurons. Consider the problem*

$$\inf_{\|\mathbf{s}\|_r\leq 1} f(\mathbf{s}) = \sum_{j=1}^n u_j(\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s}))_+. \tag{17}$$

If either $\|\mathbf{x}\|_r \geq \epsilon$ or $n<d$, an optimum is attained on the sphere $\{\mathbf{s}: \|\mathbf{s}\|_r = 1\}$. Otherwise, an optimum is attained either at $\mathbf{s} = -\frac{1}{\epsilon}\mathbf{x}$ or on $\|\mathbf{s}\|_r = 1$.

The proof of the above theorem is deferred to Appendix E.1. Theorem 8 implies that if $n<d$, then the optimal perturbation always has norm ϵ. This result is significant because $n<d$ is a common scenario. At the same time, the theorem also implies that if $\|\mathbf{x}\|_r\geq\epsilon$, then the optimal perturbation still has norm ϵ. In practice, one expects the norm of the data points to be larger than the perturbation. Thus, on real world datasets, one would expect adversarial perturbations to always have norm ϵ.

For $1<r<\infty$, Theorem 8 aids in finding a necessary condition for the optimum. This condition implies that critical points are characterized by specifying which $\mathbf{w}_j$ satisfy $\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s})<0$, $\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s})=0$, and $\mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s})>0$. The exact assertion is fairly involved, so we delay the statement of this theorem to Appendix E.2. However, the theorem simplifies considerably for $r=2$ and we include this case below.

Theorem 9. *Assume that $\|\mathbf{x}\|_r\geq\epsilon$. Let $1<r<\infty$ and take f as in Theorem 8 and s^* as the minimizer of (17). Define the following three sets:*

$$\begin{aligned} N &= \{j: \mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s}^*)<0\}\\ Z &= \{j: \mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s}^*)=0\}\\ P &= \{j: \mathbf{w}_j\cdot(\mathbf{x}+\epsilon\mathbf{s}^*)>0\}. \end{aligned}$$

$\mathbf{s}^*$ is characterized by specifying the sets N, Z, and P. Furthermore, if $r=2$, $\mathbf{s}^*$ can be explicitly expressed in terms of these sets. Let P_Z be the projection onto $\operatorname{span}\{\mathbf{w}_j\}_{j\in Z}$ and P_{Z^C} the projection onto the complement of this subspace. Then, $\mathbf{s}^*$ is given by

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

6.2. Dimension-Independent Bound for ReLU Neural Networks

Observe that, given $\mathbf{u}$ and the weight matrix $\mathbf{W}$ with columns $(\mathbf{w}_1,\dots,\mathbf{w}_n)$, each $\mathbf{x}_i$ partitions these vectors into three sets depending on whether at the optimal $\mathbf{s}_i^*$, $\mathbf{w}_j\cdot(\mathbf{x}_i+\epsilon\mathbf{s}_i^*)$ is positive, zero or negative. As a result, given $\mathbf{W}$ and $\mathbf{u}$, the points in the data set can be partitioned into sets depending on whether they induce the same sign pattern on the columns of $\mathbf{W}$. Let $\mathcal{C}_\mathcal{S}$ denote the set of all such possible partitions and let $\mathcal{C}_\mathcal{S}^*$ be the size of this set. Indexing a particular partition in this set by $\mathcal{C}$, let $n_\mathcal{C}$ be the number of parts in this partition and define $\Pi_\mathcal{S}^* = \max_\mathcal{C} n_\mathcal{C}$. Notice that both $\Pi_\mathcal{S}^*$ and $\mathcal{C}_\mathcal{S}^*$ are data-dependent quantities. We next state a general theorem that does not explicitly depend on the dimension and instead bounds the adversarial Rademacher complexity in terms of the above data-dependent quantities.

Theorem 10. *Consider the family of functions $\mathcal{G}_p^n$ with 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:*

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

Notice that the main difference between the above guarantee and the one from the previous section is that the dimension-dependent term $(1+\sqrt{d(n+1)\log(9m)})$ has been replaced by data-dependent quantities. Next, we discuss how to bound these data-dependent quantities in terms of a notion of *adversarial shattering* that we introduce in this work.

Bounding $\Pi^*_{\mathcal{S}}$ and ϵ-adversarial shattering. A key quantity of interest in understanding the bounds from the above theorem is $\Pi^*_{\mathcal{S}}$. Notice that this corresponds to the maximum number of partitions of the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_j$ that can be induced by the dataset $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$. Viewing the $\mathbf{w}_j$s as examples and the $\mathbf{x}_i$s as hyperplanes, this corresponds to the number of sign patterns on $\mathbf{W}$ that can be induced by $\mathcal{S}$. In standard settings, this would be bounded by the VC-dimension (d in this case). However, we know more about how the $\mathbf{x}_i$s act on these vectors. Notice that at the optimal $\mathbf{s}^*_i$ for a given $\mathbf{x}_i$, for some subset of vectors $\mathbf{w}_j \cdot \mathbf{x}_i + \mathbf{w}_j \cdot \epsilon \mathbf{s}^*_i \geq 0$, and for the rest it must be that $\mathbf{w}_j \cdot \mathbf{x}_i + \mathbf{w}_j \cdot \epsilon \mathbf{s}^*_i \leq 0$. Hence, not only does $\mathbf{x}_i$ induce a sign pattern on the $\mathbf{w}_j$s, it does so with a certain margin. This is reminiscent of the classical notion of *fat shattering* (Mohri et al., 2018) from statistical learning theory. However, in this case, the margin induced could itself depend on the $\mathbf{w}_j$s in a complex manner via the product of $\mathbf{w}_j \cdot \mathbf{s}^*_i$. To formalize this intuition, we define the following notion of ϵ-adversarial shattering.

Definition 1. *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 $\mathbf{s}_i = \operatorname{argmin}_{\|\mathbf{s}\|_r \leq 1} y_i \sum_{j=1}^n u_j (\mathbf{w}_j \cdot (\mathbf{x}_i + \epsilon \mathbf{s}))_+$, 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.*

Under certain assumptions, by carefully studying the above notion of adversarial shattering one can obtain bounds of the form $O(\frac{1}{\epsilon^2})$ on the maximum number of $\mathbf{w}_j$s that can be adversarially shattered by $\mathcal{S}$. This lets us use an argument similar in spirit to Sauer's lemma (Sauer, 1972; Shelah, 1972) to bound $\Pi^*_{\mathcal{S}}$ by $n^{O(1/\epsilon^2)}$, thereby leading to a meaningful bound in Theorem 10. We believe that a further study of the above notion of adversarial shattering is the key to proving general dimension-independent bounds on the adversarial Rademacher complexity of neural networks.

7. Conclusion

In this work we presented a detailed study of the generalization properties of linear models and neural networks under adversarial perturbations. Our bounds for the linear case improve upon prior work and also lead to a novel analysis of the Rademacher complexity of linear hypotheses in non-adversarial settings as well. For the case of a single ReLU unit, while we have upper and lower bounds, it would be interesting to investigate the extent to which they are close to each other. Our analysis for the linear and ReLU hypotheses reveals that by choosing the appropriate norm regularization (ℓ_p) on the weight matrices, one can indeed avoid dimension dependence and achieve generalization in adversarial settings with negligible statistical overhead as compared to the corresponding non-adversarial setting. Our analysis illustrates the importance of choosing p satisfying $\frac{1}{r} + \frac{1}{p} \geq 1$ in algorithms. This relationship further suggests that for robustness to perturbations in an arbitrary norm $\|\cdot\|$, one could regularize by the dual norm of $\|\cdot\|$. Investigating this relationship could be future work. Finally, it would be interesting to use our approach from Section 6.2 based on ϵ-adversarial shattering to provide dimension-independent upper bounds on the adversarial Rademacher complexity of neural networks.

References

Alzer, H. On some inequalities for the Gamma and Psi functions. *Math. Comput.*, 66(217):373–389, 1997.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pp. 162–183, 2019.

Awasthi, P., Dutta, A., and Vijayaraghavan, A. On robustness to adversarial examples and polynomial optimization. In *NeurIPS*, pp. 13737–13747, 2019.

Awasthi, P., Frank, N., and Mohri, M. On the Rademacher complexity of linear hypothesis sets. *CoRR*, abs/2007.11045, 2020. URL `https://arxiv.org/abs/2007.11045`.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *CoRR*, 2017.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. Adversarial examples from cryptographic pseudo-random generators. *arXiv preprint arXiv:1811.06418*, 2018a.

Bubeck, S., Price, E., and Razenshteyn, I. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018b.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Carlini, N. and Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7. IEEE, 2018.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Cortes, C., Gonzalvo, X., Kuznetsov, V., Mohri, M., and Yang, S. AdaNet: Adaptive structural learning of artificial neural networks. In *Proceedings of ICML*, pp. 874–883, 2017.

Cortes, C., Mohri, M., and Suresh, A. T. Relative deviation margin bounds. *CoRR*, abs/2006.14950, 2020.

Degwekar, A., Nakkiran, P., and Vaikuntanathan, V. Computational limitations in robust classification and win-win results. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pp. 994–1028, 2019.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *Proceedings of CVPR*, pp. 1625–1634, 2018.

Feige, U., Mansour, Y., and Schapire, R. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pp. 637–657, 2015.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Gowal, S., Uesato, J., Qin, C., Huang, P.-S., Mann, T., and Kohli, P. An alternative surrogate loss for PGD-based adversarial testing. *arXiv preprint arXiv:1910.09338*, 2019.

Haagerup, U. The best constants in the Khintchine inequality. *Studia Mathematica*, 70:231–283, 1981.

Huber, P. J. *Robust statistics*. Springer, 2011.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proceedings of NIPS*, pp. 793–800, 2008.

Kearns, M. and Li, M. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

Kearns, M. J., Schapire, R. E., and Sellie, L. M. Toward efficient agnostic learning. *Machine Learning*, 17(2-3): 115–141, 1994.

Khim, J. and Loh, P.-L. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, New York, 1991.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. The MIT Press, second edition, 2018.

Montasser, O., Hanneke, S., and Srebro, N. VC classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*, 2019.

Nakkiran, P. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

Olver, F. W. J., , Lozier, D. W., Boisvert, R. F., and Clark, C. W. *The NIST Handbook of Mathematical Functions*. Cambridge Univ. Press, 2010.

Polyakova, L. On minimizing the sum of a convex function and a concave function. *Mathematical Programming Study*, 29, 1986.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Sauer, N. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*, 2018.

Schott, L., Rauber, J., Bethge, M., and Brendel, W. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018.

Sharma, Y. and Chen, P.-Y. Breaking the madry defense model with l_1-based adversarial examples. *arXiv preprint arXiv:1710.10733*, 2017.

Shelah, S. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *Proceedings of ICLR*, 2014.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy, 2018.

van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes*. Springer, 1996.

Wei, C. and Ma, T. Improved sample complexities for deep networks and robust classification via an all-layer margin. *arXiv preprint arXiv:1910.04284*, 2019.

Yin, D., Ramchandran, K., and Bartlett, P. L. Rademacher complexity for adversarially robust generalization. In *Proceedings of ICML*, pp. 7085–7094, 2019.

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 $p^* \geq 2$, $\Psi: \mathbf{u} \mapsto \frac{1}{2}\|\mathbf{u}\|_{p^*}^2$ 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:

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

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

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

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:

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

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 $\mathbf{x}_i$.

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

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 $\|\mathbf{z}\|_2 \leq \sup_{\mathbf{z} \in A}\|\mathbf{z}\|_2 = \|\mathbf{X}^\top\|_{2,\infty}$. Further, $\mathcal{A}$ contains at most $2d$ elements. Thus, by Massart's lemma (Mohri et al., 2018),

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

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 $q \leq p$,

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

which implies that

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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 $\widetilde{\mathcal{G}}_p$:

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

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

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

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

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

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:*

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

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

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:

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

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*

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

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

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

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

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

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

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

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

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 $\|v - v'\| \le \epsilon$.*

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*

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

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 $\mathcal{G}_p^n$:*

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

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 $\mathbf{w}_j \in \mathcal{C}_2$.

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

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

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

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:

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

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

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

We then apply the following inequalities:

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

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:

$$\begin{aligned} & y\sum_{j=1}^n u_j\rho(\mathbf{w}_j \cdot \mathbf{x}_{i*}) - y\sum_{j=1}^n u_j\rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c) && (39)\\ &\le y\sum_{i=1}^n u_j\rho(\mathbf{w}_j \cdot \mathbf{x}_{i*}^c) - y\sum_{j=1}^n u_j\rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c) && \text{(infimum of first sum at } \mathbf{x}_*)\\ &\le \sum_{j=1}^n |u_j||\rho(\mathbf{w}_j \cdot \mathbf{x}_{i*}^c) - \rho(\mathbf{w}_j^c \cdot \mathbf{x}_{i*}^c)| && \text{(triangle inequality)}\\ &\le L_\rho \sum_{j=1}^n |u_j||(\mathbf{w}_j - \mathbf{w}_j^c)\cdot \mathbf{x}_{i*}^c| && \text{(Lipschitz property)}\\ &\le L_\rho \sum_{j=1}^n |u_j|\|\mathbf{w}_j - \mathbf{w}_j^c\|_p \|\mathbf{x}_{i*}^c\|_{p^*} && \text{(Hölder's inequality)}\\ &\le L_\rho \sum_{j=1}^n |u_j|\delta_2 \|\mathbf{x}_{i*}^c\|_{p^*} && (\|\mathbf{w}_j - \mathbf{w}_j^c\| \le \delta_2)\\ &\le L_\rho \sum_{j=1}^n |u_j|\delta_2 \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|\mathbf{X}\|_{r,\infty} + \epsilon) && \text{(equation (36))}\\ &\le L_\rho \Lambda\delta_2(\|\mathbf{X}\|_{r,\infty} + \epsilon)\max(1, d^{1-\frac{1}{p}-\frac{1}{r}}). && (\|\mathbf{u}\|_1 \le \Lambda) \quad (40)\end{aligned}$$

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

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

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* $\mathbf{w}_{j_1} \ldots \mathbf{w}_{j_d}\}$

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

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

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

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

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 $\mathrm{span}(Z_0^{\mathbf{s}^*})$. 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 $\dim(\mathrm{span}(Z_0^{\mathbf{s}^*})) = k < d$. We will show that there is another point $\mathbf{t}$ that achieves the same objective value satisfying either $\|\mathbf{t}\|_r = 1$ or $\dim(\mathrm{span}(Z_0^{\mathbf{t}})) = k + 1$.

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 $\mathbf{v} \in \mathrm{span}(Z)$ with $\mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{v}) = 0$ for all $\mathbf{w}_j \in Z$, and let $\mathbf{A}\colon \mathbb{R}^{d-k} \to \mathbb{R}^d$ be a matrix whose columns span $Z^\perp$. Take $\mathbf{s} = \mathbf{v} + \mathbf{A}\mathbf{s}'$, $P = \{j \colon \mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}^*) > 0\}$, and $N = \{j \colon \mathbf{w}_j \cdot (\mathbf{x} + \epsilon\mathbf{s}^*) < 0\}$. 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

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

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 $\mathrm{int}(R)$. Then $\nabla g(\mathbf{z}) = 0$ implies that $\nabla g(\mathbf{s}') = 0$ for all $\mathbf{s}' \in \mathrm{int}(R)$. 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 $\mathbf{w}_j \notin \mathrm{span}\, Z_0^{\mathbf{s}^*}$. It follows that $\mathrm{span}(Z_0^{\mathbf{s}^*} \cup \{\mathbf{w}_j\})$ is dimension $k + 1$ and this completes the induction step. $\square$

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. $\square$

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 $\mathrm{sgn}(\mathbf{v})$ will be the vector of signs of the components of $\mathbf{v}$. We adopt the convention $\mathrm{sgn}(0) = 0$. 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 $\mathbf{v} = \frac{1}{\|\mathbf{u}\|_r^{r-1}} |\mathbf{u}|^{r-1} \odot \mathrm{sgn}(\mathbf{u})$, which has unit r^*-norm. For convenience we, define

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

which gives

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

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:

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

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

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

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*

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

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

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

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 $t_j \in [0,1]$,

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

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 $\partial\|\mathbf{s}\|_r^r$:

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

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 $\mathbf{s}$, 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

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

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

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

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

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

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

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

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

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

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

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

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

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

Next, note that

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

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

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

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:

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

Using the above bound we can write:

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

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

By Jensen's inequality, we have

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\left|\sum_{i\in\mathcal{C}_t}\sigma_i\right|\right] \le \sqrt{|\mathcal{C}_t|}.$$

Substituting this bound above we get that

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

$$\le\frac{\epsilon\Lambda W}{\sqrt{m}}\max(1,d^{1-\frac{1}{p}-\frac{1}{r}})|\mathcal{C}^*_{\mathcal{S}}|\sqrt{\Pi^*_{\mathcal{S}}}. \tag{61}$$

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:

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

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 $\|\mathbf{x}_i\|_r\ge\epsilon$. Then, the following upper bound on the Rademacher complexity holds:*

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

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 $\mathbf{w}_1,\dots,\mathbf{w}_k$. 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 $\mathcal{S}=((\mathbf{x}_1,y_1)\dots(\mathbf{x}_m,y_m))$ and $(\mathbf{w}_1,\dots,\mathbf{w}_n)$. Let $\mathbf{s}_i=\operatorname{argmin}_{\|\mathbf{s}\|_r\le1}y_i\sum_{j=1}^n u_j(\mathbf{w}_j\cdot(\mathbf{x}_i+\epsilon\mathbf{s}))_+$, and define the following three sets:*

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

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 $\mathbf{w}_1,\dots,\mathbf{w}_n$ 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 $t\ge1$. Fix a sample $\mathcal{S}=((\mathbf{x}_1,y_1)\dots(\mathbf{x}_m,y_m))$ and weights $\mathbf{w}_1,\dots,\mathbf{w}_n$ such that for all $i\in[n]$, $Z_i=\emptyset$, and no subset of the weights of size more than t can be ϵ-adversarially shattered by $\mathcal{S}$. Then it holds that*

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

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. $\|\mathbf{w}_j\|^2 \geq w^2_{min}$ *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

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

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 $Z_i = \emptyset$. 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:

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

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

$$\begin{aligned}\epsilon \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\big[\Delta\mathbf{W} \cdot \mathbf{W}^+\big] &= \epsilon \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Big[\Big(\sum_{j=1}^t \sigma_j \mathbf{w}_j\Big)\Big(\sum_{j=1}^t \sigma_j 1_{\sigma_j > 0} \mathbf{w}_j\Big)\Big] \qquad (67)\\ &= \epsilon \sum_{j,k=1}^t \mathbb{E}[1_{\sigma_j > 0} \sigma_k] \mathbf{w}_j \cdot \mathbf{w}_k \\ &= \epsilon \Big(\sum_{j \neq k} \mathbb{E}[1_{\sigma_j > 0}] \mathbb{E}[\sigma_k] \mathbf{w}_j \cdot \mathbf{w}_k + \sum_{j=1}^t \mathbb{E}[1_{\sigma_j > 0}] \mathbf{w}_j \cdot \mathbf{w}_j\Big) \\ &= \frac{\epsilon}{2} \sum_{j=1}^t \|\mathbf{w}_j\|^2. \qquad (68)\end{aligned}$$

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

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

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

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

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 $\|\mathbf{w}_j\|^2 \ge w_{\text{min}}^2$ for all $j \in [t]$. Substituting above we get

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

Rearranging, we get that

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

□