ECE543 Project Paper: Spectrally-normalized margin bounds for neural networks

Forest Yang

May 2019

1 Introduction

This paper [Bartlett et al., 2017] proves a generalization bound for neural networks free of explicit "combinatorial parameters" like depth or number of parameters other than a log factor. This is achieved via a covering number bound for the set of matrices achieved by a fixed matrix times another matrix with bounded norm, which is then applied to bound the Rademacher complexity via chaining. It is shown empirically that the *normalized margin*, which consists of the margin of the network (akin to $yf(x)$ for binary classification, but modified for multiclass) divided by the Rademacher complexity (in other words, the inverse of the Rademacher complexity of the function class composed with ramp loss with corresponding margin parameter) captures how difficult a dataset is to learn. More precisely, by calculating the normalized margin of each input point and plotting the resulting distribution, one can gauge the difficulty of the classification problem. A hard problem corresponds to a distribution further to the left.

In this report I will present the proof of the paper's result in a self-contained manner, leaving some proofs to the appendix. The proofs are essentially the same as in the paper, except Lemma 3.2 is very slightly generalized with a negligible modification to the proof ($\|X\|_p$ is allowed to be $\|X\|_{p,p'}$ with (p', q) conjugate exponents). Also, because the empirical Rademacher complexity is used there is no need to state Lemma A.8 in terms of a bound on the data norm $\|X\|_2$ like in the paper, and thus no need to include data norm bounds in the union bound in Lemma A.9. Thus, unlike the paper $\|X\|_2$ is included in the bounds immediately. Also, there seems to be a slight bug in the paper when doing the union bound in Lemma A.9 – they assume that $j_1 \geq 2$, i.e. $\gamma \leq \frac{n}{2}$ without giving a proper justification. Instead I increased the range of j_1 to all integers.

Then, I will compare the paper's result with the neural network generalization bounds from the course notes, after translating the framework of the course notes to a more typical neural net framework. I will briefly discuss how Golowich et al. [2017], which a bound in the notes is derived from, mentions this paper in an arguably wrong way. Finally, I will discuss how it seems inevitable that the matrix covering lemma was used in the way it was by the paper.

2 The result

Preliminaries. Consider a multiclass classification problem with inputs from $\mathbb{R}^d$ and k classes, $[k]$. We use the following notation for a neural network. Consider L fixed nonlinearities $(\sigma_1, \ldots, \sigma_L)$ $\sigma_i : \mathbb{R} \to \mathbb{R}$ which, when applied to a vector, are applied component-wise. Denote the Lipschitz

constant of σ_i as ρ_i. Fix $d = d_0, d_1, d_2, \ldots, d_L = k$. Given L weight matrices $\mathbf{A} := (A_1, \ldots, A_L)$, the neural network with $\mathbf{A}$ for weights is the function

$$F_{\mathbf{A}}(x) = \sigma_L(A_L \sigma_{L-1}(A_{L-1} \ldots \sigma_1(A_1 x) \ldots)).$$

Denote the dataset as $Z^n = ((x_1, y_1), \ldots, (x_n, y_n)) \subset \mathbb{R}^d \times [k], z_i := (x_i, y_i)$ (can be reasoned about as a random variable) and denote $X \in \mathbb{R}^{n \times d}$ as the matrix whose ith row is x_i. Define $W = \max_{0 \le i \le L} d_i$. For matrices $\|\cdot\|_p$ is norm obtained by vectorizing the matrix and taking the $\|\cdot\|_p$ norm of the vector ($p = 2$ corresponds to Frobenius norm), $\|\cdot\|_\sigma$ is the spectral or largest singular value norm, and $\|\cdot\|_{p,q}$ is the q norm of the vector holding the p norms of the columns, i.e. $\|\cdot\|_{p,q} = \|(\|A_{:,1}\|_p, \ldots, \|A_{:,m}\|_p)\|_q$. Now we can give the *spectral complexity* $R_{\mathbf{A}}$ of the network $F_{\mathbf{A}}$, and present the main result.

$$R_{\mathbf{A}} = \left(\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma \right) \left(\sum_{i=1}^{L} \frac{\|A_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}} \right)^{3/2}.$$

Theorem 1.1 (Main result.)**.** Let nonlinearities $(\sigma_i)_{i=1}^L$ be given where each σ_i is ρ_i-Lipschitz and $\sigma_i(0) = 0$ for all $i \in [L]$. Then for $(x, y), (x_1, y_1), \ldots, (x_n, y_n)$ drawn iid from any probability distribution over $\mathbb{R}^d \times \{1, \ldots, k\}$, with probability at least $1 - \delta$ over $((x_i, y_i))_{i=1}^n$, for every $\gamma > 0$ and $\mathbf{A} = (A_i)_{i=1}^L$ s.t. $A_i \in \mathbb{R}^{d_i \times d_{i-1}}$,

$$\mathcal{R}(F_{\mathbf{A}}) := \Pr\left[\arg\max_j F_{\mathbf{A}}(x)_j \neq y \right] \le \widehat{\mathcal{R}}_{\gamma,n}(F_{\mathbf{A}}) + \tilde{O}\left(\frac{\|X\|_F R_{\mathbf{A}}}{\gamma n} \ln(W) + \sqrt{\frac{\ln(1/\delta)}{n}} \right),$$

where $\widehat{\mathcal{R}}_\gamma(f) := \frac{1}{n} \sum_{i=1}^n \ell_\gamma(-\mathcal{M}(f(x_i), y_i)) \le \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{f(x_i)_{y_i} \le \gamma + \max_{j \neq y_i} f(x_i)_j\}$.

Furthermore, in bounding the covering number of neural networks, it will help to denote the product of sets of matrices as the set of products, i.e. if $\mathcal{A} \subset \mathbb{R}^{d \times m}$ and $\mathcal{X} \subset \mathbb{R}^{m \times n}$, $\mathcal{A}\mathcal{X} := \{AX \in \mathbb{R}^{d \times n} : A \in \mathcal{A}, X \in \mathcal{X}\}$. Denote $\sigma \circ \mathcal{A} = \{\sigma(A) : A \in \mathcal{A}\}$.

3 The proof

3.1 Rademacher complexity

Define the margin $\mathcal{M} : \mathbb{R}^k \times [k] \to \mathbb{R}$ by $\mathcal{M}(v, y) = v_y - \max_{j \neq y} v_j$, the *ramp loss*

$$\ell_\gamma(r) := \begin{cases} 0 & r < -\gamma \\ 1 + \frac{r}{\gamma} & r \in [-\gamma, 0] \\ 1 & r > 0. \end{cases}$$

and the *ramp risk* $\mathcal{R}_\gamma(f) := \mathbb{E}\, \ell_\gamma(\mathcal{M}(f(x), y))$, where the expectation is over x, y drawn from some underlying probability distribution over $\mathbb{R}^d \times [k]$. For a set of real-valued functions $\mathcal{H}$, define $\mathcal{H}_{|z^n} = \{(h(x_i, y_i))_{i=1}^n \mid h \in \mathcal{H}\}$. Finally, given a set $A \subset \mathbb{R}^n$ define the *Rademacher complexity* as

$$\mathsf{Rad}(A) := \frac{1}{n} \mathbb{E}_\varepsilon \sup_{a \in A} \sum_{i=1}^n \varepsilon_i a_i, \quad \frac{\varepsilon_i + 1}{2} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(0.5).$$

By standard techniques, we have the following bound:

Lemma 3.1. Given functions $\mathcal{F}$ from $\mathbb{R}^d \to \mathbb{R}^k$ and any $\gamma > 0$, define

$$\mathcal{F}_\gamma := \{(x, y) \mapsto \ell_\gamma(-\mathcal{M}(f(x), y)) : f \in \mathcal{F}\}.$$

Then, with probability at least $1 - \delta$ over z^n, every $f \in \mathcal{F}$ satisfies

$$\mathcal{R}(f) \leq \widehat{\mathcal{R}}_{n,\gamma}(f) + 2\mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Proof. See appendix. □

3.2 Matrix covering number

We define the proper covering number as follows.

$$\mathcal{N}(U, \epsilon, \|\cdot\|) = \min_{V \subset U}\{|V| : \sup_{A \in U} \min_{B \in V} \|A - B\| \leq \epsilon\}.$$

A set $V \subset U$ which satisfies the condition in the above set construction is said to *cover* U at scale ϵ with norm $\|\cdot\|$, meaning for any element of U, there is an element of V that is ϵ close. For the upcoming inductive proof in the covering of neural nets, the fact that V is a proper cover, i.e. $V \subset U$, will be important.
Cover numbers are how one controls Rademacher complexity in the case of continuous valued loss (c.f. VC dimension). By covering a continuous valued set, one has essentially "turned the set finite" modulo some error based on ϵ, so that one can still apply the finite class lemma to bound the Rademacher complexity. In fact, the Dudley entropy integral obtained by the method of chaining will provide an even tighter bound than the finite class lemma. The other approach in class we saw for bounding the Rademacher complexity of a continuous set was direct calculation, which was doable in the RKHS framework.
The valuable covering number bound for a fixed matrix times a variable one is as follows:

Lemma 3.2. Let conjugate exponents (p, q) and (r, s) be given with $p \leq 2$, as well as positive reals (a, ϵ) and positive integer m. Let matrix $X \in \mathbb{R}^{n \times d}$ be given with $\|X\|_p = b$. Then

$$\mathcal{N}\left(\left\{XA : A \in \mathbb{R}^{d \times m}, \|A\|_{q,s} \leq a\right\}, \epsilon, \|\cdot\|_2\right) \leq \left\lceil \frac{a^2 b^2 m^{2/r}}{\epsilon^2} \ln(2dm) \right\rceil.$$

Actually, it is possible to generalize this lemma slightly to the following, with a very small modification to the proof.

Lemma 3.2.2. Let conjugate exponents (p', q) and (r, s) be given as well as $p \leq 2$, positive reals (a, ϵ) and positive integer m. Let matrix $X \in \mathbb{R}^{n \times d}$ be given with $\|X\|_{p,p'} = b$. Then

$$\mathcal{N}\left(\left\{XA : A \in \mathbb{R}^{d \times m}, \|A\|_{q,s} \leq a\right\}, \epsilon, \|\cdot\|_2\right) \leq \left\lceil \frac{a^2 b^2 m^{2/r}}{\epsilon^2} \ln(2dm) \right\rceil.$$

Proof. See appendix. □

3.3 Neural net covering number

We would like to cover the set (recall $X \in \mathbb{R}^{n\times d}$ has data points for rows)

$$\begin{aligned}\mathcal{F}_{\mathcal{A}}(X^\top) &:= \{F_{\mathbf{A}}(X^\top) : \forall i \in \{1,\dots,L\},\ \|A_i\|_\sigma \le s_i,\ \|A_i^\top\|_{2,1} \le b_i\} \\ &= \{\sigma_L(A_L\sigma_{L-1}(A_{L-1}\dots\sigma_1(A_1X^\top)\dots)) : \forall i \in [L],\ A_i \in \mathcal{A}_i\},\end{aligned}$$

where $\mathcal{A}_i := \{A_i \in \mathbb{R}^{d_i\times d_{i-1}} : \|A_i\|_\sigma \le s_i,\ \|A_i^\top\|_{2,1}\}$. Let $\mathcal{X}_0 = \{X^\top\}$, $\mathcal{A} := \mathcal{A}_L \times \dots \times \mathcal{A}_1$. For $i \in [L]$, define

$$\begin{aligned}\mathcal{X}_i :&= \{\sigma_i(A_i\dots\sigma_1(A_1X^\top)\dots) \in \mathbb{R}^{d_i\times n} : \forall\, 1 \le j \le i,\ A_j \in \mathcal{A}_j\} \\ &= \{\sigma_i(A_i(X_{i-1})) \in \mathbb{R}^{d_i\times n} : X_{i-1} \in \mathcal{X}_{i-1}, A_i \in \mathcal{A}_i\} =: \sigma \circ \mathcal{A}_i\mathcal{X}_{i-1}.\end{aligned}$$

In other words, $\mathcal{X}_i$ is the set of possible outputs at the ith layer with the dataset $X^\top$ as input and supposing $A_j \in \mathcal{A}_j$ for each $j \le i$. Notice that $\mathcal{X}_L = \mathcal{F}_{\mathcal{A}}(X^\top)$, and that $\mathcal{X}_0$ is a cover for itself of size 1. Thus, if we can inductively construct a cover $\widehat{\mathcal{X}}_{i+1}$ for $\mathcal{X}_{i+1}$ using a cover $\widehat{\mathcal{X}}_i$ for $\mathcal{X}_i$, we will construct a cover $\mathcal{X}_L = \mathcal{F}_{\mathcal{A}}(X)$.

Lemma A.7. Let $\epsilon_1,\dots,\epsilon_L > 0$. Let $N_i = \sup_{X_{i-1}\in\mathcal{X}_{i-1}} \mathcal{N}(\mathcal{A}_iX_{i-1}, \epsilon_i, \|\cdot\|_2)$. Then, there is a proper cover of $\mathcal{F}_{\mathcal{A}}(X^\top)$ of size at most $\prod_{i=1}^L N_i$ at scale $\sum_{j=1}^L \left[\rho_j\epsilon_j \prod_{l=j+1}^L \rho_l s_l\right]$.

Proof. Base case: $\widehat{\mathcal{X}}_0 = \mathcal{X}_0 = \{X^\top\}$.
Inductive step: Assume $\widehat{\mathcal{X}}_i$ is a proper cover for $\mathcal{X}_i$ of size $\le \prod_{j=1}^i N_i$ at scale $\sum_{j\le i}\left[\rho_j\epsilon_j\prod_{l=j+1}^i\rho_l s_l\right]$. We have $\mathcal{X}_{i+1} = \sigma_{i+1}\circ\mathcal{A}_{i+1}\mathcal{X}_i$, where the product of two sets is the set of products of their elements. For each $\widehat{X}_i \in \widehat{\mathcal{X}}_i$, there is a proper cover of $\mathcal{A}_{i+1}\widehat{X}_i$ at scale ϵ_{i+1} of size at most N_{i+1}. We may denote the cover $\widehat{\mathcal{A}}_{i+1}(\widehat{X}_i)\widehat{X}_i$ because each element has the form $\widehat{A}_{i+1}\widehat{X}_i$ for some $\widehat{A}_{i+1} \in \mathcal{A}_{i+1}$, due to being a proper cover. Consider the union

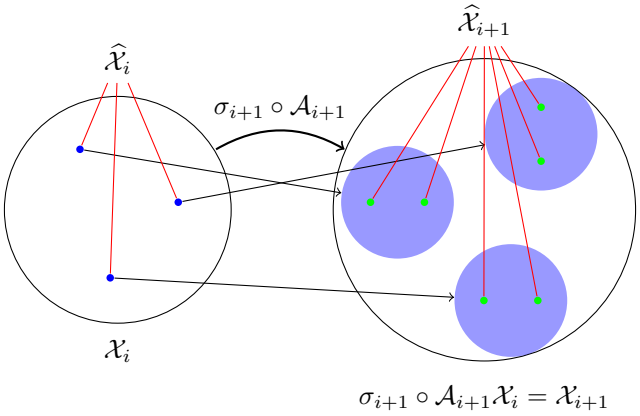
$$\widehat{\mathcal{A}}_{i+1}(\widehat{\mathcal{X}}_i)\widehat{\mathcal{X}}_i := \bigcup_{\widehat{X}_i\in\widehat{\mathcal{X}}_i} \widehat{\mathcal{A}}_{i+1}(\widehat{X}_i)\widehat{X}_i.$$

Set $\widehat{\mathcal{X}}_{i+1} = \{\sigma_{i+1}(\hat{A}_i\hat{X}_i) : \hat{A}_i\hat{X}_i \in \widehat{\mathcal{A}}_{i+1}(\widehat{\mathcal{X}}_i)\mathcal{X}_i\} = \sigma_{i+1}\circ\widehat{\mathcal{A}}_{i+1}(\widehat{\mathcal{X}}_i)\mathcal{X}_i$.
The size $|\widehat{\mathcal{X}}_{i+1}| \le |\widehat{\mathcal{X}}_i|N_{i+1} \le \prod_{j\le i+1} N_j$. Let's compute the scale of this cover. Given $X_{i+1} = \sigma_{i+1}(A_{i+1}X_i) \in \mathcal{X}_{i+1}$, by the above, there is $\widehat{X}_{i+1} = \sigma_{i+1}(\widehat{A}_{i+1}\widehat{X}_i) \in \widehat{\mathcal{X}}_{i+1}$ such that

$$\begin{aligned}\|X_{i+1} - \widehat{X}_{i+1}\| &= \|\sigma_{i+1}(A_{i+1}X_i) - \sigma_{i+1}(\widehat{A}_{i+1}\widehat{X}_i)\| \le \rho_{i+1}\|A_{i+1}X_i - \widehat{A}_{i+1}\widehat{X}_i\| \\ &\le \rho_{i+1}\left(\|A_{i+1}X_i - A_{i+1}\widehat{X}_i\| + \|A_{i+1}\widehat{X}_i - \widehat{A}_{i+1}\widehat{X}_i\|\right) \\ &\le \rho_{i+1}s_{i+1}\|X_i - \widehat{X}_i\| + \rho_{i+1}\epsilon_{i+1} \\ &\le \rho_{i+1}s_{i+1}\sum_{j\le i}\left[\rho_j\epsilon_j\prod_{l=j+1}^i\rho_l s_l\right] + \rho_{i+1}\epsilon_{i+1} \\ &= \sum_{j\le i+1}\rho_j\epsilon_j\prod_{l=j+1}^{i+1}\rho_l s_l.\end{aligned}$$

□

HELPFUL PICTURE



Now we can plug in values of ϵ_i to the above proposition to obtain a covering number bound for neural networks.

Theorem 3.3.

$$\ln \mathcal{N}(\mathcal{F}_{\mathcal{A}}(X^\top), \epsilon, \|\cdot\|_2) \leq \frac{\|X\|_2^2 \ln(2W^2)}{\epsilon^2} \left(\prod_{i=1}^{L} \rho_i^2 s_i^2\right) \left(\sum_{i=1}^{L} \left(\frac{b_i}{s_i}\right)^{2/3}\right)^3.$$

Proof. Recall that by Lemma A.7, if $N_i = \sup_{X_{i-1} \in \mathcal{X}_{i-1}} \mathcal{N}(\mathcal{A}_i X_{i-1}, \epsilon_i, \|\cdot\|_2)$ then there is a proper cover of $\mathcal{F}_{\mathcal{A}}(X^\top)$ of size at most $\prod_{i=1}^{L} N_i$ at scale $\tau = \sum_{j=1}^{L} \left[\rho_j \epsilon_j \prod_{l=j+1}^{L} \rho_l s_l\right]$. We set

$$\epsilon_i = \frac{\alpha_i \epsilon}{\rho_i \prod_{j>i} \rho_j s_j}, \quad \alpha_i = \frac{1}{\bar{\alpha}} \left(\frac{b_i}{s_i}\right)^{2/3}, \quad \bar{\alpha} = \sum_{i=1}^{L} \left(\frac{b_i}{s_i}\right)^{2/3}.$$

At first, this choice may seem out of the blue, but in retrospect it makes sense. For the scale at the ith layer, we divide by $\rho_i \prod_{j>i} \rho_i s_i$ because that is roughly how much the space has blown up. ϵ_i is roughly proportional to $\frac{b_i}{s_i}$ because if $b_i \geq \|A_i^\top\|_{2,1}$ is large compared to s_i, the difficulty b_i of covering the ith layer at a small scale is not worth counteracting the blowup s_i.
The scale τ becomes

$$\tau = \sum_{i=1}^{L} \left[\rho_i \epsilon_i \prod_{j>i} \rho_j s_j\right] = \sum_{i=1}^{L} \left[\rho_i \frac{\alpha_i \epsilon}{\rho_i \prod_{j>i} \rho_j s_j} \prod_{j>i} \rho_j s_j\right] = \epsilon \sum_{i=1}^{L} \alpha_i = \epsilon.$$

Therefore, the cover of $\mathcal{F}_{\mathcal{A}}(X^\top)$ corresponding to these choices of ϵ_i is at scale ϵ. Now we compute the size. By the matrix covering lemma, Lemma 3.2, $(p = q = 2, r = \infty, s = 1, \max\{d, m\} \leq W)$

$$\ln N_i = \sup_{X_{i-1} \in \mathcal{X}_{i-1}} \ln \mathcal{N}(\mathcal{A}_i X_{i-1}, \epsilon_i, \|\cdot\|_2) \leq \left\lceil \frac{\sup_{X_{i-1} \in \mathcal{X}_{i-1}} \|X_{i-1}\|_2^2 b_i^2}{\epsilon_i^2} \right\rceil \ln(2W^2).$$

Now for $X_i = \sigma_i(A_i X_{i-1}) \in \mathcal{X}_i$, using $\sigma_i(0) = 0$,

$$\|X_i\|_2 = \|\sigma_i(A_i X_{i-1})\|_2 \le \rho_i \|A_i X_{i-1}\|_2 \le \rho_i s_i \|X_{i-1}\|_2 \le \|X\|_2 \prod_{1 \le j \le i} \rho_i s_i,$$

where the last part is by induction. Thus,

$$\begin{aligned}
\ln \mathcal{N}(\mathcal{F}_{\mathcal{A}}(X^\top), \epsilon, \|\cdot\|_2) &\le \sum_{i=1}^{L} \ln N_i \le \ln(2W^2) \sum_{i=1}^{L} \left\lceil \frac{\sup_{X_{i-1} \in \mathcal{X}_{i-1}} \|X_{i-1}\|_2^2 b_i^2}{\epsilon_i^2} \right\rceil \\
&\le \ln(2W^2) \sum_{i=1}^{L} \left\lceil \frac{\|X\|_2^2 (\prod_{1 \le j < i} \rho_j^2 s_j^2) b_i^2}{\alpha_i^2 \epsilon^2 (\rho_i \prod_{j>i} \rho_i s_i)^{-2}} \right\rceil \\
&\le \ln(2W^2) \left\lceil \frac{\|X\|_2^2 (\prod_{i=1}^{L} \rho_i^2 s_i^2)}{\epsilon^2} \right\rceil \sum_{i=1}^{L} \left\lceil \frac{b_i^2}{s_i^2 \alpha_i^2} \right\rceil \\
&= \ln(2W^2) \left\lceil \frac{\|X\|_2^2 (\prod_{i=1}^{L} \rho_i^2 s_i^2)}{\epsilon^2} \right\rceil \left\lceil \left(\sum_{i=1}^{L} \left(\frac{b_i}{s_i} \right)^{2/3} \right)^3 \right\rceil.
\end{aligned}$$

□

3.4 Plugging into Rademacher complexity

Lemma A.5 (Dudley integral)**.** Let $\mathcal{F}$ be a real valued function class taking values in $[-1, 1]$ containing $\mathbf{0}$. Then,

$$\mathsf{Rad}(\mathcal{F}_{|S}) \le \inf_{\alpha > 0} \left[\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \varepsilon, \|\cdot\|_2)}\, \mathrm{d}\varepsilon \right].$$

Proof. See appendix. □

We use Lemma A.5 and Theorem 3.3 to prove Lemma A.8:

Lemma A.8. Let $(\sigma_1, \ldots, \sigma_L)$ be given where σ_i is ρ_i-Lipschitz, and let margin $\gamma > 0$. Let spectral norm bounds $\mathbf{s} = (s_i)_{i=1}^L$, and $\|\cdot\|_{2,1}$ norm bounds $\mathbf{b} = (b_i)_{i=1}^L$ be given. Define $\mathcal{A}_i^{\mathbf{s},\mathbf{b}}, \mathcal{A}^{\mathbf{s},\mathbf{b}}$ accordingly. Then with probability at least $1 - \delta$ over i.i.d. samples $z^n = ((x_i, y_i))_{i=1}^n$, $X^\top = [x_1, \ldots, x_n]$, every set of weight matrices $\mathbf{A} \in \mathcal{A}^{\mathbf{s},\mathbf{b}}$ satisfies

$$\mathcal{R}(F_{\mathbf{A}}) \le \widehat{\mathcal{R}}_{n,\gamma}(F_{\mathbf{A}}) + \frac{8}{n} + \frac{72 \|X\|_2 \ln(2W) \ln(n)}{\gamma n} \left(\prod_{i=1}^{L} s_i \rho_i \right) \left(\sum_{i=1}^{L} \frac{b_i^{2/3}}{s_i^{2/3}} \right)^{3/2} + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Proof. By Lemma 3.1, with probability at least $1 - \delta$, for any $\mathbf{A} \in \mathcal{A}$,

$$\mathcal{R}(F_{\mathbf{A}}) \le \widehat{\mathcal{R}}_{n,\gamma}(F_{\mathbf{A}}) + 2\mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}, \quad \mathcal{F}_\gamma = \ell_\gamma \circ -\mathcal{M} \circ \mathcal{F}_{\mathcal{A}^{\mathbf{b},\mathbf{s}}}.$$

Note that $\mathcal{M}(\cdot, y)$ is 2-Lipschitz for every $y \in [k]$:

$$\begin{aligned}|\mathcal{M}(z,y) - \mathcal{M}(x,y)| &= |z_y - x_y - \max_{j\neq y} z_j + \max_{j\neq y} x_j| \leq |z_y - x_y| + |\max_{j\neq y} z_j - \max_{j\neq y} x_j| \\ &\leq |z_y - x_y| + \max_j |z_j - x_j| \leq 2\|z-y\|_\infty \leq 2\|z-y\|_2.\end{aligned}$$

Furthermore, ℓ_γ is $\frac{1}{\gamma}$ Lipschitz. Thus, by considering $\ell_\gamma(-\mathcal{M}(\sigma_L(\cdot), y_i))$ as the final nonlinearity with Lipschitz contant $\frac{2\rho_L}{\gamma}$, we may apply Theorem 3.3 (theorem holds with different nonlinearities for each point, which is the case here since each point has a different y_i, as long as their Lipschitz constants are the same) to obtain that

$$\ln \mathcal{N}((\mathcal{F}_\gamma)_{|z^n}, \epsilon, \|\cdot\|_2) \leq \ln(2W^2) \frac{4\|X\|_2^2 (\prod_{i=1}^L \rho_i^2 s_i^2)}{\epsilon^2} \left(\sum_{i=1}^L \left(\frac{b_i}{s_i}\right)^{2/3}\right)^3 =: \frac{R^2_{\mathcal{A}^{\mathbf{s},\mathbf{b}}}}{\epsilon^2}.$$

By the Dudley entropy integral,

$$\mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n}) \leq \inf_{\alpha>0} \frac{4\alpha}{\sqrt{n}} + \frac{12}{n}\int_\alpha^{\sqrt{n}} \frac{R_{\mathcal{A}^{\mathbf{s},\mathbf{b}}}}{\epsilon}\,\mathrm{d}\epsilon = \inf_{\alpha>0} \frac{4\alpha}{\sqrt{n}} + \frac{12 R_{\mathcal{A}^{\mathbf{s},\mathbf{b}}}}{n} \ln(\sqrt{n}/\alpha).$$

If we choose $\alpha = \frac{1}{\sqrt{n}}$, we obtain that for all $\mathbf{A} \in \mathcal{A}^{\mathbf{s},\mathbf{b}}$

$$\begin{aligned}\mathcal{R}(F_{\mathbf{A}}) &\leq \widehat{\mathcal{R}}_{n,\gamma}(F_{\mathbf{A}}) + 2\mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\ &\leq \widehat{\mathcal{R}}_{n,\gamma}(F_{\mathbf{A}}) + \frac{8}{n} + \frac{24\ln(n) R_{\mathcal{A}^{\mathbf{s},\mathbf{b}}}}{n} + 3\sqrt{\frac{\ln(2/\delta)}{2n}} \\ &= \widehat{\mathcal{R}}_{n,\gamma}(F_{\mathbf{A}}) + \frac{8}{n} + \frac{48\|X\|_2 \ln(n)\sqrt{\ln(2W^2)}}{\gamma n}\left(\prod_{i=1}^L \rho_i s_i\right)\left(\sum_{i=1}^L \left(\frac{b_i}{s_i}\right)^{2/3}\right)^{3/2} + 3\sqrt{\frac{\ln(2/\delta)}{2n}},\end{aligned}$$

which implies the bound since $\sqrt{\ln(2W^2)} \leq 1.5\ln(2W)$. □

3.5 Union bound over a countable set

Lemma A.9. Suppose the setting and notation of Theorem 1.1. With probability at least $1-\delta$, every network $F_{\mathbf{A}} : \mathbb{R}^d \to \mathbb{R}^k$ with weight matrices $\mathbf{A} = (A_1, \ldots, A_L)$ and every $\gamma > 0$ satisfy

$$\begin{aligned}\mathcal{R}(F_{\mathbf{A}}) \leq \widehat{\mathcal{R}}_{n,\gamma}(F_{\mathbf{A}}) + \frac{8}{n} + \frac{144\|X\|_2 \ln(n)\ln(2W)}{\gamma n}\left(\prod_{i=1}^L \rho_i \left(\|A_i\|_\sigma + \frac{1}{L}\right)\right)\left[\sum_{i=1}^L \left(\frac{\|A_i^\top\|_{2,1} + \frac{1}{L}}{\|A_i\|_\sigma + \frac{1}{L}}\right)^{2/3}\right]^{3/2} \\ + \frac{3}{\sqrt{2n}}\sqrt{\ln\left(\frac{2}{\delta}\right) + \ln\left(\max\left\{2\gamma, \frac{1}{\gamma}\right\}\right) + \sum_{i=1}^L 2\ln(L\|A_i\|_\sigma + 2) + 2\ln(L\|A_i^\top\|_{2,1} + 2)}.\end{aligned}$$

The additive $\frac{1}{L}$ terms are a bit of a nuisance, but seem necessary to avoid a dependence on L like $\sqrt{L}$ or 2^L, as other ways of cutting up the space would give. Furthermore, one may justify their presence by saying that if the matrix norms are all less than $\frac{1}{L}$, then $\sqrt{\frac{\ln(1/\delta)}{n}}$ dominates. This expression is $\tilde{O}$ the expression in Theorem 1.1.

Proof. See appendix. □

4 Comparisons

A paper Neyshabur et al. [2017] obtains a similar looking bound with a different method (PAC-Bayes method). Another paper Arora et al. [2018] is said to obtain stronger results, using a compression method. Here I will focus on comparing the bound on the Rademacher complexity obtained with the matrix covering method with bounds in the course notes.

4.1 Course notes comparison

First of all, the setting of the notes when talking about neural nets is different in a subtle yet significant way. The notes define certain function classes $\mathcal{F}_0, \ldots, \mathcal{F}_\ell$ representing the set of possible hidden units at each layer. It seems to always be the case that the set of base classifiers in the setting of the notes is the first layer of the actual neural net, i.e. the mappings $x \mapsto \sigma_1(A_1 x)$, so I will use the notation $\mathcal{H}_1, \ldots, \mathcal{H}_L$, with H standing for hidden unit.
Let us be more precise. We define $\mathcal{B}_1 = \{A \in \mathbb{R}^{d_1 \times d} : \|A^\top\|_{2,\infty} \leq B_1\}$, for $i \geq 2$, $\mathcal{B}_i = \{A \in \mathbb{R}^{d_i \times d_{i-1}} : \|A^\top\|_{1,\infty} \leq B_i\}$, and $\mathcal{B} = \mathcal{B}_1 \times \ldots \times \mathcal{B}_L$. Define $\mathcal{H}_1 = \{x \mapsto \sigma_1(a^\top x) : \|a\|_2 \leq B_1\}$. Then, for $i \geq 2$, we define

$$\mathcal{H}_i = \{x \mapsto \sigma_i(a_i^\top \sigma_{i-1}(A_{i-1}(\ldots \sigma_1(A_1 x)\ldots))) : \forall 1 \leq j \leq i,\, A_j \in \mathcal{B}_j,\, a_i \in \mathbb{R}^{d-1},\, \|a_i\| \leq B_i\}.$$

In other words, $\mathcal{H}_i$ is the set of possible hidden unit functions that can be computed at the ith layer, since $\|A^\top\|_{1,\infty} \leq B$ expresses the condition that each row of A has 1-norm $\leq B$, and the ith layer can be viewed as a tuple of d_{i-1} functions. Notice that if we let $F_{\mathbf{A}_{:i}} : \mathbb{R}^d \to \mathbb{R}^{d_i}$ denote the function the neural net computes at the ith layer, and $\mathcal{B}_{:i} = \mathcal{B}_1 \times \ldots \times \mathcal{B}_i$, we may also write $\mathcal{H}_i$ as

$$\mathcal{H}_i = \left\{ \sigma_i(\sum_{l=1}^{d_{i-1}} a_l f_l) : \exists\, (A_1, \ldots, A_{i-1}) \in \mathcal{B}_{:i-1},\ (f_1, \ldots, f_{d_{i-1}}) = F_{\mathbf{A}_{:i-1}},\, \|a\|_1 \leq B_i \right\}.$$

From this, we have that $\mathcal{H}_i \subset \sigma_i \circ B_i \operatorname{abs\,conv}(\mathcal{H}_{i-1})$, but is not necessarily equal to $\sigma_i \circ B_i \operatorname{abs\,conv}(\mathcal{H}_{i-1})$, for two reasons.
First, d_{i-1} is fixed from the beginning, i.e. $\mathcal{H}_i$ does not contain combinations of elements from $\mathcal{H}_{i-1}$ which require more than d_{i-1} elements. Second, and perhaps more subtly, even though $f_1, \ldots, f_{d_{i-1}} \in \mathcal{H}_{i-1}$, it is not necessary that for any $(h_1, \ldots, h_{d_{i-1}})$, there exists $\mathbf{A}$ such that $F_{\mathbf{A}_{:i-1}} = (h_1, \ldots, h_{d_{i-1}})$. This is because once you fix the matrices $(A_1, \ldots, A_{i-2}) \in \mathcal{B}_{:i-2}$, certain functions in $\mathcal{H}_{i-1}$ may become unavailable. While it's true by definition that for any $h \in \mathcal{H}_{i-1}$, there is some $F_{\mathbf{A}_{:i-1}}$ with h as a component function, $F_{\mathbf{A}_{:i-1}}$ may not compute arbitrary tuples of functions from $\mathcal{H}_{i-1}$.
With this in mind, we can still use the analysis in the course notes. It just may not be as tight, since $\mathcal{H}_i \subset \sigma_i \circ B_i \operatorname{abs\,conv}(\mathcal{H}_{i-1})$ may be proper.

$$\mathsf{Rad}((\mathcal{H}_i)_{|X^n}) \leq \mathsf{Rad}(\sigma_i \circ B_i \operatorname{abs\,conv}((\mathcal{H}_{i-1})_{|X^n})) \leq 2\rho_i B_i\, \mathsf{Rad}((\mathcal{H}_{i-1})_{|X^n}).$$

With the base case $\mathsf{Rad}((\mathcal{H}_1)_{|X^n}) \leq \frac{2\rho_1 B_1 \|X\|_2}{n}$ we obtain

$$\mathsf{Rad}((\mathcal{H}_L)_{|X^n}) \leq 2^L \left(\prod_{i=1}^{L} \rho_i B_i\right) \frac{\|X\|_2}{n}.$$

Unfortunately, this bound has a 2^L in it. However, we may apply the result from the notes by GRS:

$$\begin{aligned}\mathsf{Rad}((\mathcal{H}_L)_{|X^n}) &\le \left(\prod_{i=2}^{L}\rho_i B_i\right)\left(\mathsf{Rad}((\mathcal{H}_1)_{|X^n}) + \frac{2}{n}\sqrt{(\log 2)\sum_{i=1}^{n}\sup_{\|a\|\le B_1}\sigma_1(a^\top X_i)^2}\right)\\ &\le \left(\prod_{i=2}^{L}\rho_i B_i\right)\left(\frac{2\rho_1 B_1\|X\|_2}{n} + \frac{2}{n}\sqrt{L(\log 2)\rho_1^2 B_1^2\|X\|_2^2}\right)\\ &= \frac{2\|X\|_2}{n}\left(\prod_{i=1}^{L}\rho_i B_i\right)\left(1+\sqrt{L\log 2}\right).\end{aligned}$$

The ratio of this bound to the above bound is $\frac{1+\sqrt{L\log 2}}{2^{L-1}}$. This bound is much better. If we replaced the bounds B_i with the norms they represent, we would get

$$\text{complexity}(F_{\mathbf{A}}) \lessapprox \frac{\|X\|_2\sqrt{L}}{n}\left(\rho_1\|A_1^\top\|_{2,\infty}\prod_{i=2}^{L}\rho_i\|A_i^\top\|_{1,\infty}\right),$$

where the $\lessapprox$ acknowledges some imprecision in the statement. The matrix covering bound gives

$$\text{complexity}(F_{\mathbf{A}}) \lessapprox \frac{\|X\|_2\ln(n)\ln(2W)}{n}\left(\prod_{i=1}^{L}\rho_i\|A_i\|_\sigma\right)\left(\sum_{i=1}^{L}\left(\frac{\|A_i^\top\|_{2,1}}{\|A_i\|_\sigma}\right)^{2/3}\right)^{3/2}.$$

Let us compute the ratio of the matrix covering bound to the GRS bound, assuming each matrix is the same:

$$r = \frac{L^{3/2}\|A\|_\sigma^{L-1}\|A^\top\|_{2,1}\ln(n)\ln(2W)}{L^{1/2}\|A^\top\|_{1,\infty}^{L-1}\|A^\top\|_{2,\infty}}.$$

In the worst case for the matrix covering lemma, $\|A^\top\|_{1,\infty} = \|A\|_\sigma/W$, as in the case where A has one nonzero column with all ones. Furthermore, $\|A^\top\|_{2,1} \le W\|A^\top\|_{2,\infty}$. Then,

$$r \approx L(\sqrt{W})^{L-1}W\ln(n)\ln(2W),$$

i.e. the GRS bound is exponentially better than the matrix covering bound.
On the other hand, it is possible for $\|A\|_\sigma \le \|A\|_{1,\infty}/\sqrt{W}$, like in the case where A has a row with identical entries and the rest are zero. Thus, even if we assume that $\|A^\top\|_{2,1} = W\|A^\top\|_{2,\infty}$ which is worst case, we obtain

$$r \approx \frac{LW\ln(n)\ln(2W)}{(\sqrt{W})^{L-3}}.$$

In this regime, aside from the pesky $\ln(n)$ factor which the GRS bound has the advantage of not having, the matrix covering bound is exponentially better.
It seems that the $\|\cdot\|_{1,\infty}$ is advantageous, compared to an analogous bound using Frobenius norm constraints presented in the slides, in that there are cases where $\|\cdot\|_{1,\infty} < \|\cdot\|_\sigma$ (this is not true if $\|\cdot\|_{1,\infty}$ is replaced with $\|\cdot\|_2$. It is nice that the GRS/notes method applied to ℓ_1 norms and absolute convex hulls can allow us to use this norm, and the analysis is rather clean.

4.2 GRS paper's mentioning

Golowich et al. [2017] say that since $\frac{\|A_i^\top\|_{2,1}}{\|A_i\|_\sigma} \geq 1$, the matrix covering derived bound

$$\frac{\|X\|_2 \ln(n) \ln(2W)}{n} \left(\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma \right) \left[\sum_{i=1}^{L} \left(\frac{\|A_i^\top\|_{2,1}}{\|A_i\|_\sigma} \right)^{2/3} \right]^{3/2} = CL^{3/2},$$

i.e. it has a $L^{3/2}$ dependence on L, and is not size independent. Quote: "there is still a strong and unavoidable polynomial dependence... (many pages later) hence the bound scales at least as $\sqrt{L^3/n}$." But, this is not exactly true, because

$$\frac{\|X\|_2 \ln(\ldots)}{n} \left(\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma \right) \left[\sum_{i=1}^{L} \left(\frac{\|A_i^\top\|_{2,1}}{\|A_i\|_\sigma} \right)^{2/3} \right]^{3/2} = \frac{\|X\|_2 \ln(\ldots)}{n} \left(\prod_{i=1}^{L} \rho_i \right) \left[\sum_{i=1}^{L} \left(\|A_i^\top\|_{2,1} \prod_{j \neq i} \|A_i\|_\sigma \right)^{2/3} \right]^{3/2}$$

and the "unavoidable" $L^{3/2}$ is gone.

5 Inevitability discussion

If we attempt to apply the slightly generalized matrix covering lemma:

$$\begin{aligned} \ln \mathcal{N}(\mathcal{X}_L, \tau_L, \|\cdot\|_2) &\leq \sum_{i=1}^{L} \sup_{X_{i-1} \in \mathcal{X}_{i-1}} \ln \mathcal{N}(\mathcal{A}_i X_{i-1}, \epsilon_i, \|\cdot\|) \\ &\leq \sum_{i=1}^{L} \frac{(\sup_{X_{i-1} \in \mathcal{X}_{i-1}} \|X_{i-1}^\top\|_{p,p'}^2) \|A_i^\top\|_{q,s}^2 W^{2/r}}{\epsilon_i^2} \ln(2W^2). \end{aligned}$$

Our choices are restricted here. Firstly, the covering works with the $\|\cdot\|_2$ norm, due to Dudley applying the finite class lemma to the increment classes, which requires $\|\cdot\|_2$ norms. Therefore, it seems inevitable that the scale of our cover is

$$\tau_L = \sum_{i=1}^{L} \left[\rho_i \epsilon_i \left(\prod_{j=i+1}^{L} \rho_j \|A_j\|_\sigma \right) \right].$$

Specifically, $\|A_i(X_{i-1} - \hat{X}_{i-1})\|_2 \leq \|A_i\|_\sigma \|X_{i-1} - \hat{X}_{i-1}\|_2$ seems forced, i.e. covering elements with respect to $\|\cdot\|_2$ forces the appearance of the spectral norms, $\|A_i\|_\sigma$. This means that we should choose $\epsilon_i = \frac{C_i \epsilon}{\rho_i \prod_{j=i+1}^{L} (\rho_j \|A_j\|_\sigma)}$ for some $\sum_{i=1}^{L} C_i = 1$ for the scales ϵ_i, to make $\tau_L = \epsilon$. In order to get the nice final expression for the covering number, the fact that

$$\|X_{i-1}^\top\|_2 \leq (\prod_{j=1}^{i-1} \rho_j \|A_j\|_\sigma) \|X\|_2$$

was used. Then, the $\|X\|_2 \prod_{j=1}^{i-1} \rho_j \|A_j\|_\sigma$ is merged with the $\rho_i \prod_{j=i+1}^{L} \rho_j \|A_j\|_\sigma$ to get the factor $\|X\|_2 \prod_{j=1}^{L} \rho_j \|A_j\|_\sigma$ that can be pulled out of all terms of the sum, modulo leaving behind $\|A_i\|_\sigma$

in each term.
Thus, by the above, since ϵ_i is forced to carry $\rho_i \prod_{j=i+1}^{L} \rho_j \|A_j\|_\sigma$, in order to get a nice expression, it seems like $\|X_{i-1}^\top\|_{p,p'}$ must satisfy

$$\|X_{i-1}^\top\|_{p,p'} \leq \left(\prod_{j=1}^{i-1} \rho_j \|A_j\|_\sigma \right) \|X^\top\|_{p,p'},$$

that is if we want to replace the $\|X\|_2$ in the covering number bound with $\|X^\top\|_{p,p'}$. Unfortunately, $X_i^\top \in \mathbb{R}^{n \times d_i}$, which means that rows represent sample points. If each layer is then applying a transformation to each row, which should make the spectral norm pop out, and p' corresponds to the rows, we should take $p' = 2$. Furthermore, there is no benefit to having p equal anything but 2, because the other parameters $(p', q), (r, s)$ are unaffected by p, and p must be ≤ 2, $\|\cdot\|_p \leq \|\cdot\|_l$ for $l \leq p$. Then, $q = 2$, and we take $s = 1$ so that $r = \infty$, making the $W^{2/r}$ go away. In other words, what is done in the paper seems like the only viable application of the matrix covering lemma.

References

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv e-prints*, art. arXiv:1802.05296, Feb 2018.

Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv e-prints*, art. arXiv:1706.08498, Jun 2017.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-Independent Sample Complexity of Neural Networks. *arXiv e-prints*, art. arXiv:1712.06541, Dec 2017.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. *arXiv e-prints*, art. arXiv:1707.09564, Jul 2017.

6 Appendix

Lemma 3.1. Given functions $\mathcal{F}$ from $\mathbb{R}^d \to \mathbb{R}^k$ and any $\gamma > 0$, define

$$\mathcal{F}_\gamma := \{(x, y) \mapsto \ell_\gamma(-\mathcal{M}(f(x), y)) : f \in \mathcal{F}\}.$$

Then, with probability at least $1 - \delta$ over z^n, every $f \in \mathcal{F}$ satisfies

$$\mathcal{R}(f) \leq \widehat{\mathcal{R}}_{n,\gamma}(f) + 2\mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Proof. Define $\Delta_n(z^n) := \sup_{f \in \mathcal{F}} \widehat{\mathcal{R}}_{n,\gamma}(f) - \mathcal{R}_\gamma(f)$. We use the symmetrization argument (without absolute values); let $z^{n'}$ be an i.i.d. sample separate from z^n but drawn from the same distribution.

Let $\widehat{\mathcal{R}}'_{n,\gamma}$ denote the empirical ramp risk on the $z^{n'}$ dataset. Further, define the function h by $h(f, z_i) = \ell_\gamma(-\mathcal{M}(f(x_i), y_i))$. We have

$$\begin{aligned}
\mathbb{E}_{z^n}\Delta_n(z^n) &= \mathbb{E}_{z^n}\sup_{f\in\mathcal{F}}\left[\mathcal{R}_\gamma(f) - \widehat{\mathcal{R}}_{n,\gamma}(f)\right] = \mathbb{E}_{z^n}\sup_{f\in\mathcal{F}}\left[\mathbb{E}_{z^{n'}}\widehat{\mathcal{R}}'_{n,\gamma}(f) - \widehat{\mathcal{R}}_{n,\gamma}(f)\right] \\
&\le \mathbb{E}_{z^n}\mathbb{E}_{z^{n'}}\sup_{f\in\mathcal{F}}\left[\widehat{\mathcal{R}}'_{n,\gamma}(f) - \widehat{\mathcal{R}}_{n\gamma}(f)\right] = \frac{1}{n}\mathbb{E}_\varepsilon\mathbb{E}_{z^n}\mathbb{E}_{z^{n'}}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i(h(f, z'_i) - h(f, z_i)) \\
&\le \frac{1}{n}\mathbb{E}_{z^{n'}}\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i h(f, z'_i) + \frac{1}{n}\mathbb{E}_{z^n}\mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i h(f, z_i) = 2\mathbb{E}_{z^n}\mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n}).
\end{aligned}$$

Now we observe that $\Delta_n(z^n)$ and $\mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n})$ both satisfy bounded differences as functions of z^n with constant $\frac{1}{n}$, because $\ell_\gamma \in [0, 1]$. Therefore by McDiarmid's,

$$\begin{aligned}
&\text{w.p. } \ge 1 - \frac{\delta}{2}, \qquad \Delta_n(z^n) \le \mathbb{E}_{z^n}\Delta(z^n) + \sqrt{\frac{\log(2/\delta)}{2n}}, \\
&\text{w.p. } \ge 1 - \frac{\delta}{2}, \qquad \mathbb{E}_{z^n}\mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n}) \le \mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n}) + \sqrt{\frac{\log(2/\delta)}{2n}}.
\end{aligned}$$

Noting that $\mathcal{R}(f) \le \mathcal{R}_\gamma(f)$, because $\mathbb{1}\{y \ne \arg\max_i f(x)_i\} \le \ell_\gamma(-\mathcal{M}(f(x), y))$, and summing the two inequalities which w.p. $\ge 1 - \delta$ both hold, we get

$$\begin{aligned}
\mathcal{R}(f) \le \mathcal{R}_\gamma(f) &\le \Delta(z^n) + \widehat{\mathcal{R}}_{n,\gamma}(f) \le \mathbb{E}\Delta(z^n) + \sqrt{\frac{\log(2/\delta)}{2n}} \\
&\le \widehat{\mathcal{R}}_{n,\gamma}(f) + 2\mathbb{E}\mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n}) + \sqrt{\frac{\log(2/\delta)}{2n}} \\
&\le \widehat{\mathcal{R}}_{n,\gamma}(f) + 2\mathsf{Rad}((\mathcal{F}_\gamma)_{|z^n}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.
\end{aligned}$$

□

Lemma 3.2.2. Let conjugate exponents (p', q) and (r, s) be given as well as $p \le 2$, positive reals (a, ϵ) and positive integer m. Let matrix $X \in \mathbb{R}^{n\times d}$ be given with $\|X\|_{p,p'} = b$. Then

$$\mathcal{N}\left(\left\{XA : A \in \mathbb{R}^{d\times m}, \|A\|_{q,s} \le a\right\}, \epsilon, \|\cdot\|_2\right) \le \left\lceil \frac{a^2b^2m^{2/r}}{\epsilon^2}\ln(2dm)\right\rceil.$$

Proof. The proof of this lemma uses the Maurey sparsification lemma:

Lemma A.6. Fix Hilbert space $\mathcal{H}$ with norm $\|\cdot\|$. If $U = \sum_{i=1}^d \alpha_i V_i$ with $\alpha \in \mathbb{R}^d_{\ge 0} \ne 0$ and $V_i \in \mathcal{H}$, given positive integer k, there are nonnegative integers $(k_1, \ldots, k_d)$ s.t. $\sum_{i=1}^d k_i = k$ and

$$\left\|U - \frac{\|\alpha\|_1}{k}\sum_{i=1}^d k_i V_i\right\|^2 \le \frac{\|\alpha\|_1}{k}\sum_{i=1}^d \alpha_i\|V_i\|^2 \le \frac{\|\alpha\|_1^2}{k}\max_{i\in[d]}\|V_i\|^2.$$

Proof Of Maurey's Lemma. First assume $\sum_{i=1}^d \alpha_i = 1$, i.e. $\alpha \in \Delta_d$. Then if we define k i.i.d. random variables $I_1, \ldots, I_k$ s.t. $\Pr[I_j = i] = \alpha_i$ for all $i \in [d]$,

$$\mathbb{E}\left[\frac{1}{k}\sum_{j=1}^k V_{I_j}\right] = \mathbb{E}V_{I_1} = \sum_{i=1}^d \alpha_i V_i = U.$$

By independence, the variance of the sum of the V_{I_j}'s is the sum of the variances (identify $*$ with $\langle\cdot,\cdot\rangle_{\mathcal{H}}$:

$$\mathbb{E}\left(U - \frac{1}{k}\sum_{j=1}^k V_{I_j}\right)^2 = \operatorname{Var}\left(\frac{1}{k}\sum_{j=1}^k V_{I_j}\right) = \frac{k}{k^2}\operatorname{Var}(V_{I_1}) \le \frac{1}{k}\mathbb{E}\|V_{I_1}\|^2 = \frac{1}{k}\sum_{i=1}^d \alpha_i\|V_i\|^2.$$

By the probabilistic argument, there is some $i_1, i_2, \ldots, i_k$ such that

$$\left\|U - \frac{1}{k}\sum_{j=1}^k V_{i_k}\right\|^2 = \left\|U - \frac{1}{k}\sum_{i=1}^d k_i V_i\right\|^2 \le \frac{1}{k}\sum_{i=1}^d \alpha_i\|V_i\|^2,$$

where $k_i := |\{j : i_j = i\}|$. Note that if $\alpha \neq 1$, we may replace U with $U/\|\alpha\|_1$ and α with $\alpha/\|\alpha\|_1$. Then, multiply both sides by $\|\alpha\|_1^2$ to get the original statement. □

Now we prove Lemma 3.2.2. Recall (p', q) and (r, s) are conjugate exponents and $p \ge 2$. Denote $b := \|X\|_{p,p'}$ and $Y = \begin{bmatrix}\frac{X_{:,1}}{\|X_{:,1}\|_p} & \ldots & \frac{X_{:,d}}{\|X_{:,d}\|_p}\end{bmatrix} \in \mathbb{R}^{n\times d}$ as the matrix with the columns of X scaled by their norms for its columns. Furthermore, define $S \in \mathbb{R}^{d\times m}$ by $S_{ij} = \|X_{:,i}\|_p$, i.e. each column of S has d entries which equal the d norms of X's columns. We will consider the setting of Maurey's Lemma with $\mathcal{H} = \mathbb{R}^{n\times m}$ and norm $\|\cdot\|_2$ induced by the Frobenius inner product. Let $A \in \mathbb{R}^{d\times m}$ be a matrix with $\|A\|_{q,s} \le a$. Notice that ($\odot$ denotes elementwise product)

$$XA = Y(S \odot A) = Y\sum_{i=1}^d\sum_{j=1}^m A_{ij}\|X_{:,i}\|_p \mathbf{e}_i\mathbf{e}_j^\top = Y\sum_{i=1}^d\sum_{j=1}^m A_{ij}S_{ij}\mathbf{e}_i\mathbf{e}_j^\top.$$

In other words, XA is a nonnegative combination of the $N = 2dm$ basis elements $\{Y\mathbf{e}_i\mathbf{e}_j^\top\}_{i,j=1}^{d,m} \cup \{-Y\mathbf{e}_i\mathbf{e}_j^\top\}_{i,j=1}^{d,m}$, placing it in the setting of Maurey's Lemma. The sum of the weights is $\|S \odot A\|_1$. Using the conjugacy of (p', q) and (r, s):

$$\begin{aligned}\|S \odot A\|_1 &\le \left\| \begin{bmatrix}\|S_{:,1}\|_{p'}\|A_{:,1}\|_q & \ldots & \|S_{:,m}\|_{p'}\|A_{:,m}\|_q\end{bmatrix}\right\|_1 \\ &\le \left\|\begin{bmatrix}\|S_{:,1}\|_{p'} & \ldots & \|S_{:,m}\|_{p'}\end{bmatrix}\right\|_r \left\|\begin{bmatrix}\|A_{:,1}\|_q & \ldots & \|A_{:,m}\|_q\end{bmatrix}\right\|_s = \|S\|_{p',r}\|A\|_{q,s}.\end{aligned}$$

Furthermore, notice that each column of S is identically equal to the $\|\cdot\|_p$-norms of the columns of X, so $\|S\|_{p',r} = m^{1/r}\|X\|_{p,p'} = m^{1/r}b$. Denote $\bar{a} = abm^{1/r}$. Then, by Maurey's Lemma, given

$k \in \mathbb{Z}_+$, there exist $k_1, \dots k_{2dm}$ summing to k where

$$\left\| \frac{\bar{a}}{k} \sum_{i=1}^{d} \sum_{j=1}^{m} k_{d(j-1)+i} Y \mathbf{e}_i \mathbf{e}_j^\top - \frac{\bar{a}}{k} \sum_{i=1}^{d} \sum_{j=1}^{m} k_{d(m-1+j)+i} Y \mathbf{e}_i \mathbf{e}_j^\top - XA \right\|_2^2$$
$$\leq \frac{\bar{a}^2}{k} \max_{ij} \|Y \mathbf{e}_i \mathbf{e}_j^\top\|^2 \leq \frac{\bar{a}^2}{k} \max_{i} \|Y \mathbf{e}_i\|_2^2 = \frac{\bar{a}^2}{k} \max_{i} \frac{\|Xe_i\|_2^2}{\|Xe_i\|_p^2} \leq \frac{a^2 b^2 m^{2/r}}{k}.$$

Thus, if we choose $k = \left\lceil \frac{a^2 b^2 m^{2/r}}{\epsilon^2} \right\rceil$, the RHS is $\leq \epsilon$. Notice there is a subtlety; $\|S \odot A\|_1 \leq \bar{a}$ but may not equal $\bar{a}$; Maurey's Lemma says we need $\frac{\|\alpha\|_1}{k}$ outside the sums in the first line where $\|\alpha\|_1$ is the sum of the weights of the basis elements when they sum to XA. What if $\|\alpha\|_1 = \|S \odot A\|_1 < \bar{a}$? In that case, since we have each element and their negative in the basis, we can artificially increase the sum of the weights by adding weight to an element and its negative without changing XA until it equals $\bar{a}$. This lets us put $\frac{\bar{a}}{k}$ in the first line.
Defining $\{V_i\}_{i=1}^N = \{Y \mathbf{e}_i \mathbf{e}_j^\top\}_{i,j=1}^{d,m} \cup \{-Y \mathbf{e}_i \mathbf{e}_j^\top\}_{i,j=1}^{d,m}$ by numbering the elements arbitrarily, we have

$$\left\{ \frac{1}{k} \sum_{i=1}^{N} k_i V_i : k_i \in \mathbb{N}, \sum_{i=1}^{N} k_i = k \right\} \text{ is proper cover for } \{XA : \|A\|_{q,s} \leq a\} \text{ at scale } \epsilon \text{ for } \|\cdot\|_2.$$

The size of this cover is less than N^k, because there are k units of mass, each having N choices for which element to go to. Therefore, $\mathcal{N}(\{XA : \|A\|_{q,s} \leq a\}, \epsilon, \|\cdot\|_2) \leq k \ln N = \left\lceil \frac{a^2 b^2 m^{2/r}}{\epsilon^2} \right\rceil \ln(2dm)$. □

Comment: the point of the scaling is to deal with arbitrary norms. If you do not scale, then you get $\frac{\|X\|_{2,\infty}^2 \|A\|_1^2}{\epsilon^2} \ln(2dm)$.

Lemma A.5 (Dudley integral)**.** Let $\mathcal{F}$ be a real valued function class taking values in $[-1, 1]$ containing $\mathbf{0}$. Then,

$$\mathsf{Rad}(\mathcal{F}_{|S}) \leq \inf_{\alpha > 0} \left[\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \varepsilon, \|\cdot\|_2)} \, \mathrm{d}\varepsilon \right].$$

Proof. Let $N \in \mathcal{N}$ be arbitrary and define $\varepsilon_i = \sqrt{n} 2^{-(i-1)}$; in particular, $\varepsilon_1 = \sqrt{n} 2^{-(1-1)} = \sqrt{n}$. Furthermore, define V_i as a ε_i-cover of $\mathcal{F}_{|S}$ of size $|V_i| = \mathcal{N}(\mathcal{F}_{|S}, \varepsilon_i, \|\cdot\|_2)$. But, take V_1 to be the specific $\sqrt{n}$-cover, $V_1 = \{\mathbf{0}\}$. Note we've used the assumption that $f \in \mathcal{F}$ has output in $[-1, 1]$ to say that $\{\mathbf{0}\}$ is indeed a $\sqrt{n}$-cover. For any $f \in \mathcal{F}$, denote, $v^i[f]$ as the element of V_i which covers $f_{|S} := (f(x_1), \dots, f(x_n))$, i.e. $\sqrt{\sum_{t=1}^n (v^i[f]_t - f(x_t))^2} \leq \varepsilon_i$. Now,

$$\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \langle \epsilon, f_{|S} \rangle = \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left\langle \epsilon, f_{|S} - v^N[f] + \sum_{i=1}^{N-1} (v^{i+1}[f] - v^i[f]) + v^1[f] \right\rangle$$
$$\leq \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \langle \epsilon, f - v^N[f] \rangle + \sum_{i=1}^{N-1} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \langle \epsilon, v^{i+1}[f] - v^i[f] \rangle + \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \langle \epsilon, v^1[f] \rangle.$$

Bounding the first term is simple Cauchy-Schwarz: $\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\langle \epsilon, f_{|S} - v^N[f]\rangle \le \sqrt{n}\varepsilon^N$. The last term, since $V_1 = \{\mathbf{0}\}$, is 0. Now for the intermediate terms, notice

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\langle \epsilon, v^{i+1}[f] - v^i[f]\rangle = \mathsf{URad}(W_i), \qquad W_i := \{v^{i+1}[f] - v^i[f] : f \in \mathcal{F}\}.$$

Since $v^{i+1}[f] \in V_{i+1}$ and $v^i[f] \in V_i$, there are at most $|V_{i+1}||V_i| \le |V_{i+1}|^2$ elements of $|W_i|$. Furthermore, the norm of an element is at most $\|v^{i+1}[f] - v^i[f]\| \le \varepsilon^{i+1} + \varepsilon^i = 3\varepsilon^{i+1}$. Thus, by the finite class lemma,

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\langle \epsilon, v^{i+1}[f] - v^i[f]\rangle \le \sup_{w\in W_i}\|w\|_2\sqrt{2\log|W_i|} \le 6\varepsilon_{i+1}\sqrt{\log|V_{i+1}|}.$$

We're almost done. We just need to plug these in and relate the final quantity to an integral.

$$\begin{aligned}
\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\langle \epsilon, f_{|S}\rangle &\le \sqrt{n}\varepsilon_N + 6\sum_{i=1}^{N-1}\varepsilon_{i+1}\sqrt{\log|V_{i+1}|} \\
&= \sqrt{n}\varepsilon_N + 12\sum_{i=1}^{N-1}(\varepsilon_{i+1} - \varepsilon_{i+2})\sqrt{\log|V_{i+1}|} \\
&= \sqrt{n}\varepsilon_N + 12\sum_{i=2}^{N}(\varepsilon_i - \varepsilon_{i+1})\sqrt{\log\mathcal{N}(\mathcal{F}_{|S}, \varepsilon_i, \|\cdot\|)} \\
&\le \sqrt{n}\varepsilon_N + 12\int_{\varepsilon_{N+1}}^{\sqrt{n}/2}\sqrt{\log\mathcal{N}(\mathcal{F}_{|S}, \varepsilon, \|\cdot\|_2)}\,\mathrm{d}\varepsilon.
\end{aligned}$$

Now given $\alpha > 0$, pick the minimum N such that $\varepsilon_{N+1} > \alpha$. Then, $\alpha \ge \varepsilon_{N+2} = \varepsilon_N/4$, so $\varepsilon_N \le 4\alpha$. Thus,

$$\mathsf{URad}(\mathcal{F}_{|S}) = \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\langle \epsilon, f_{|S}\rangle \le 4\sqrt{n}\alpha + 12\int_{\alpha}^{\sqrt{n}/2}\sqrt{\log\mathcal{N}(\mathcal{F}_{|S}, \varepsilon, \|\cdot\|_2)}\,\mathrm{d}\varepsilon.$$

We obtain the exact statement in the lemma by infing over $\alpha > 0$ and dividing by n to normalize the URad to get Rad. □

Lemma A.9. Suppose the setting and notation of Theorem 1.1. With probability at least $1-\delta$, every network $F_{\mathbf{A}} : \mathbb{R}^d \to \mathbb{R}^k$ with weight matrices $\mathbf{A} = (A_1, \ldots, A_L)$ and every $\gamma > 0$ satisfy

$$\begin{aligned}
\mathcal{R}(F_{\mathbf{A}}) \le \widehat{\mathcal{R}}_{n,\gamma}(F_{\mathbf{A}}) + \frac{8}{n} + \frac{144\|X\|_2\ln(n)\ln(2W)}{\gamma n}\left(\prod_{i=1}^{L}\rho_i\left(\|A_i\|_\sigma + \frac{1}{L}\right)\right)\left[\sum_{i=1}^{L}\left(\frac{\|A_i^\top\|_{2,1} + \frac{1}{L}}{\|A_i\|_\sigma + \frac{1}{L}}\right)^{2/3}\right]^{3/2} \\
+ \frac{3}{\sqrt{2n}}\sqrt{\ln\left(\frac{2}{\delta}\right) + \ln\left(\max\left\{2\gamma, \frac{1}{\gamma}\right\}\right) + \sum_{i=1}^{L}2\ln(L\|A_i\|_\sigma + 2) + 2\ln(L\|A_i^\top\|_{2,1} + 2)}.
\end{aligned}$$

Proof. Recall Lemma A.8 says that with probability at least $1-\delta$,

$$\mathcal{R}(F_{\mathbf{A}}) \le \widehat{\mathcal{R}}_{n,\gamma}(F_{\mathbf{A}}) + \frac{8}{n} + \frac{72\|X\|_2\ln(2W)\ln(n)}{\gamma n}\left(\prod_{i=1}^{L}s_i\rho_i\right)\left(\sum_{i=1}^{L}\frac{b_i^{2/3}}{s_i^{2/3}}\right)^{3/2} + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

We divide up the space as follows. We instantiate the bound of Lemma A.8 for the following values of $\gamma, \mathbf{s}, \mathbf{b}$:

$$\gamma = 2^j, \quad s_i = \frac{k_i}{L}, \quad b_i = \frac{l_i}{L}, \qquad j \in \mathbb{Z}, \quad k_i \in \mathbb{Z}_+, \quad l_i \in \mathbb{Z}_+, \qquad i = 1, \ldots, L.$$

Accordingly, to each $j, \mathbf{k}, \mathbf{l}$ we associate a probability:

$$\delta(j, \mathbf{k}, \mathbf{l}) = \frac{\delta}{2^{|j|} \prod_{i=1}^{L} k_i(k_i+1) l_i(l_i+1)}.$$

It can be checked that

$$\sum_{j \in \mathbb{Z},\ \mathbf{k}, \mathbf{l} \in (\mathbb{Z}_+)^L} \delta(j, \mathbf{k}, \mathbf{l}) = 3\delta.$$

Thus, the bound of Lemma A.8 with $\gamma, \mathbf{s}, \mathbf{b}$ determined by $j, \mathbf{k}, \mathbf{l}$ as above and McDiarmid term $3\sqrt{\frac{\log(2/\delta(j,\mathbf{k},\mathbf{l}))}{2n}}$ holds for every $j \in \mathbb{Z}$, $\mathbf{k}, \mathbf{l} \in (Z_+)^L$ with probability at least $1 - 3\delta$.
Now assume the bounds mentioned above all hold. Given $\mathbf{A}$, pick the smallest $j \in \mathbb{Z}$ such that $\gamma \le 2^j$ and for each $i = 1, \ldots, L$, the smallest $k_i, l_i \in \mathbb{Z}_+$ such that $\|A_i\|_\sigma \le \frac{k_i}{L}$ and $\|A_i^\top\|_{2,1} \le \frac{l_i}{L}$. Let $C = \frac{72\|X\|_2 \ln(n) \ln(2W)}{n} \prod_{i=1}^{L} \rho_i$. Since $2^{j-1} < \gamma$, $\frac{k_i}{L} < \|A_i\|_\sigma + \frac{1}{L}$, $\frac{l_i}{L} < \|A_i^\top\|_{2,1} + \frac{1}{L}$,

$$\begin{aligned}
\mathcal{R}(F_{\mathbf{A}}) &\le \widehat{\mathcal{R}}_{n,2^{j-1}}(F_{\mathbf{A}}) + \frac{8}{n} + C\frac{1}{2^{j-1}} \left[\sum_{i=1}^{L} \left(\frac{l_i}{L} \left(\prod_{m \ne i} \frac{k_m}{L} \right) \right)^{2/3} \right]^{3/2} + 3\sqrt{\frac{\ln(2/\delta(j,\mathbf{k},\mathbf{l}))}{2n}} \\
&\le \widehat{\mathcal{R}}_{n,\gamma}(F_{\mathbf{A}}) + \frac{8}{n} + C\frac{2}{\gamma} \left[\sum_{i=1}^{L} \left(\left(\|A_i^\top\|_{2,1} + \frac{1}{L} \right) \prod_{m \ne i} \left(\|A_i\|_\sigma + \frac{1}{L} \right) \right)^{2/3} \right]^{3/2} + 3\sqrt{\frac{\ln(2/\delta(j,\mathbf{k},\mathbf{l}))}{2n}}.
\end{aligned}$$

Now we bound $\ln(2/\delta(j, \mathbf{k}, \varepsilon l))$. Notice that if $j \ge 0$, then $2^{|j|} = 2^j \le 2\gamma$, and if $j \le 0$, then $2^{|j|} = 2^{-j} \le \frac{1}{\gamma}$, so that $2^{|j|} \le \max\{2\gamma, \gamma^{-1}\}$.

$$\begin{aligned}
\ln\left(\frac{2}{\delta(j,\mathbf{k},\mathbf{l})}\right) &\le \ln\left(\frac{2}{\delta}\right) + \ln(2^{|j|}) + \sum_{i=1}^{L} \ln\left[k_i(k_i+1)\right] + \ln\left[l_i(l_i+1)\right] \\
&\le \ln\left(\frac{2}{\delta}\right) + \ln\left(\max\left\{2\gamma, \frac{1}{\gamma}\right\}\right) + \sum_{i=1}^{L} 2\ln(k_i+1) + 2\ln(l_i+1) \\
&\le \ln\left(\frac{2}{\delta}\right) + \ln\left(\max\left\{2\gamma, \frac{1}{\gamma}\right\}\right) + \sum_{i=1}^{L} 2\ln(L\|A_i\|_\sigma + 2) + 2\ln(L\|A_i^\top\|_{2,1} + 2).
\end{aligned}$$

Therefore, with probability at least $1 - 3\delta$, for any $\gamma > 0$ and $\mathbf{A}$,

$$\begin{aligned}
\mathcal{R}(F_{\mathbf{A}}) \le \widehat{\mathcal{R}}_{n,\gamma}(F_{\mathbf{A}}) + \frac{8}{n} + \frac{144\|X\|_2 \ln(n)\ln(2W)}{\gamma n} \left(\prod_{i=1}^{L} \rho_i\right) \left[\sum_{i=1}^{L} \left(\left(\|A_i^\top\|_{2,1} + \frac{1}{L} \right) \prod_{m \ne i} \left(\|A_i\|_\sigma + \frac{1}{L} \right) \right)^{2/3} \right]^{3/2} \\
+ \frac{3}{\sqrt{2n}} \sqrt{\ln\left(\frac{2}{\delta}\right) + \ln\left(\max\left\{2\gamma, \frac{1}{\gamma}\right\}\right) + \sum_{i=1}^{L} 2\ln(L\|A_i\|_\sigma + 2) + 2\ln(L\|A_i^\top\|_{2,1} + 2)}.
\end{aligned}$$

□