Spectrally-normalized margin bounds for neural networks

Peter Bartlett* Dylan J. Foster† Matus Telgarsky‡

Abstract

This paper presents a margin-based multiclass generalization bound for neural networks which scales with their margin-normalized *spectral complexity*: their Lipschitz constant, meaning the product of the spectral norms of the weight matrices, times a certain correction factor. This bound is empirically investigated for a standard AlexNet network on the `mnist` and `cifar10` datasets, with both original and random labels, where it tightly correlates with the observed excess risks.

1 Overview

Neural networks owe their astonishing success not only to their ability to fit any data set: they also *generalize well*, meaning they provide a close fit on unseen data. A classical statistical adage is that models capable of fitting too much will generalize poorly; what's going on here?

Let's navigate the many possible explanations provided by statistical theory. A first observation is that any analysis based solely on the number of possible labellings on a finite training set — as is the case with VC dimension — is doomed: if the function class can fit all possible labels (as is the case with neural networks in standard configurations (Zhang et al., 2017)), then this analysis can not distinguish it from the collection of all possible functions!

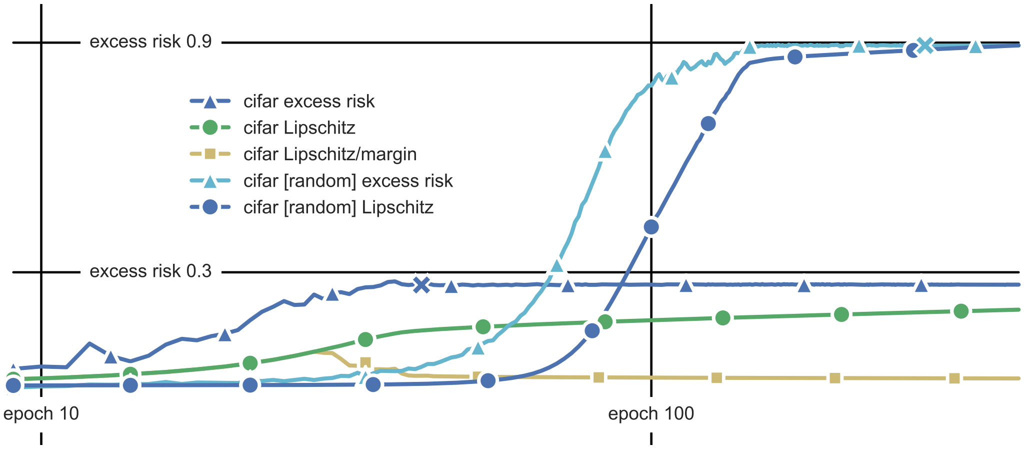


Figure 1: An analysis of AlexNet (Krizhevsky et al., 2012) trained on `cifar10`, both with original and with random labels. Triangle-marked curves track excess risk across training epochs (on a log scale), with an 'x' marking the earliest epoch with zero training error. Circle-marked curves track Lipschitz constants, normalized so that the two curves for random labels meet. The Lipschitz constants tightly correlate with excess risk, and moreover normalizing them by *margins* (resulting in the square-marked curve) neutralizes growth across epochs.

*`<peter@berkeley.edu>`; University of California, Berkeley; work performed while visiting the Simons Institute.

†`<djf244@cornell.edu>`; Cornell University; work performed while visiting the Simons Institute.

‡`<mjt@illinois.edu>`; University of Illinois, Urbana-Champaign; work performed while visiting the Simons Institute.

Next let's consider *scale-sensitive* measures of complexity, such as Rademacher complexity and metric entropy, which work directly with real-valued function classes, and moreover are sensitive to their magnitudes. Figure 1 plots the excess risk (the test error minus the training error) across training epochs against one candidate scale-sensitive complexity measure, the Lipschitz constant of the network (the product of the spectral norms of their weight matrices), and demonstrates that they are tightly correlated (which is not the case for, say, the l_2 norm of the weights). The data considered in Figure 1 is the standard `cifar10` dataset, both with original and with random labels, which has been used as a sanity check when analyzing neural network generalization (Zhang et al., 2017).

There is still an issue with basing a complexity measure purely on the Lipschitz constant (although it has already been successfully employed to regularize neural networks (Cisse et al., 2017)): as depicted in Figure 1, the measure grows over time, despite the excess risk plateauing. Fortunately, there is a standard resolution to this issue: investigating the *margins* (a precise measure of confidence) of the outputs of the network. This tool has been used to study the behavior of 2-layer networks, boosting methods, SVMs, and many others (Bartlett, 1996; Schapire et al., 1997; Boucheron et al., 2005); in boosting, for instance, there is a similar growth in complexity over time (each training iteration adds a weak learner), whereas margin bounds correctly stay flat or even decrease. This behavior is recovered here: as depicted in Figure 1, even though standard networks exhibit growing Lipschitz constants, normalizing these Lipschitz constants by the margin instead gives a decaying curve.

1.1 Contributions

This work investigates a complexity measure for neural networks which is based on the Lipschitz constant, but normalized by the margin of the predictor. The two central contributions are as follows.

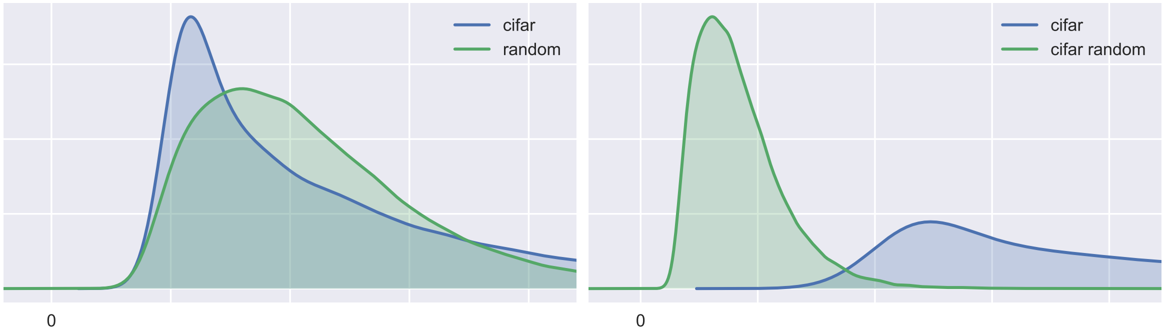
- Theorem 1.1 below will give the rigorous statement of the generalization bound which is the basis of this work. In contrast to prior work, this bound: **(a)** scales with the Lipschitz constant (product of spectral norms of weight matrices) divided by the margin; **(b)** has no dependence on combinatorial parameters (e.g., number of layers or nodes) outside of log factors; **(c)** is multiclass (with no explicit dependence on the number of classes); **(d)** measures complexity against a *reference network* (e.g., for the ResNet (He et al., 2016), the reference network has identity mappings at each layer). The bound is stated below, with a general form and analysis summary appearing in Section 3, the full details relegated to the appendix.

- An empirical investigation, in Section 2, of neural network generalization on the standard datasets `cifar10`, `cifar100`, and `mnist` using the preceding bound. Rather than using the bound to provide a single number, it can be used to form a *margin distribution* as in Figure 2. These margin distributions will illuminate the following intuitive observations: **(a)** `cifar10` is harder than `mnist`; **(b)** random labels make `cifar10` and `mnist` much more difficult; **(c)** the margin distributions (and bounds) converge during training, even though the weight matrices continue to grow; **(d)** l_2 regularization ("weight decay") does not significantly impact margins or generalization.

A more detailed description of the margin distributions is as follows. Suppose a neural network computes a function $f : \mathbb{R}^d \to \mathbb{R}^k$, where k is the number of classes; the most natural way to convert this to a classifier is to select the output coordinate with the largest magnitude, meaning $x \mapsto \arg\max_j f(x)_j$. The *margin*, then, is to measure the gap between the output for the correct label and other labels, meaning $f(x)_y - \max_{j \neq y} f(x)_j$.

Unfortunately, margins alone do not seem to say much; see for instance Figure 2a, where the collections of all margins for all data points — the *unnormalized margin distribution* — are similar for `cifar10` with and without random labels. What is missing is an appropriate *normalization*, as in Figure 2b. This normalization is provided by Theorem 1.1, which can now be explained in detail.

To state the bound, a little bit of notation is necessary. The networks will use L fixed nonlinearities $(\sigma_1, \ldots, \sigma_L)$, where σ_i is ρ_i-Lipschitz (e.g., as with coordinate-wise ReLU, and max-pooling, as discussed in Appendix A.1); occasionally, it will also hold that $\sigma_i(0) = 0$. Given L weight matrices $\mathcal{A} = (A_1, \ldots, A_L)$



(a) Margins.

(b) Normalized margins.

Figure 2: Margin distributions at the end of training AlexNet on `cifar10`, with and without random labels. With proper normalization, random labels demonstrably correspond to a harder problem.

let $F_{\mathcal{A}}$ denote the function computed by the corresponding network:

$$F_{\mathcal{A}}(x) := \sigma_L(A_L\sigma_{L-1}(A_{L-1}\cdots\sigma_1(A_1x)\cdots)).$$

Whenever data $(x_1,\ldots,x_n)$ are given, collect them as rows of a matrix $X \in \mathbb{R}^{n\times d}$. Occasionally, notation will be overloaded to discuss $F_{\mathcal{A}}(X^T)$, a matrix whose i^{th} column is $F_{\mathcal{A}}(x_i)$. The l_2 norm $\|\cdot\|_2$ is always computed entry-wise; thus, for a matrix, it corresponds to the Frobenius norm.

Next, define a collection of *reference matrices* $(M_1,\ldots,M_L)$ with each dimension at most W; for instance, to obtain a good bound for ResNet (He et al., 2016), it is sensible to set $M_i := I$, the identity map, and the bound below will worsen as the network moves farther from the identity map; for AlexNet (Krizhevsky et al., 2012), the simple choice $M_i = 0$ suffices. Finally, letting $\|\cdot\|_\sigma$ and $\|\cdot\|_1$ respectively denote spectral norm and the unrolled l_1 vector norm, the *spectral complexity* $R_{F_{\mathcal{A}}} = R_{\mathcal{A}}$ of a network $F_{\mathcal{A}}$ with weights $\mathcal{A}$ is

$$R_{\mathcal{A}} := \left(\prod_{i=1}^{L}\rho_i\|A_i\|_\sigma\right)\left(\sum_{i=1}^{L}\frac{\|A_i - M_i\|_1^{2/3}}{\|A_i\|_\sigma^{2/3}}\right)^{3/2}. \tag{1.1}$$

The following theorem provides a generalization bound for neural networks whose nonlinearities are fixed but whose weight matrices $\mathcal{A}$ have bounded spectral complexity $R_{\mathcal{A}}$.

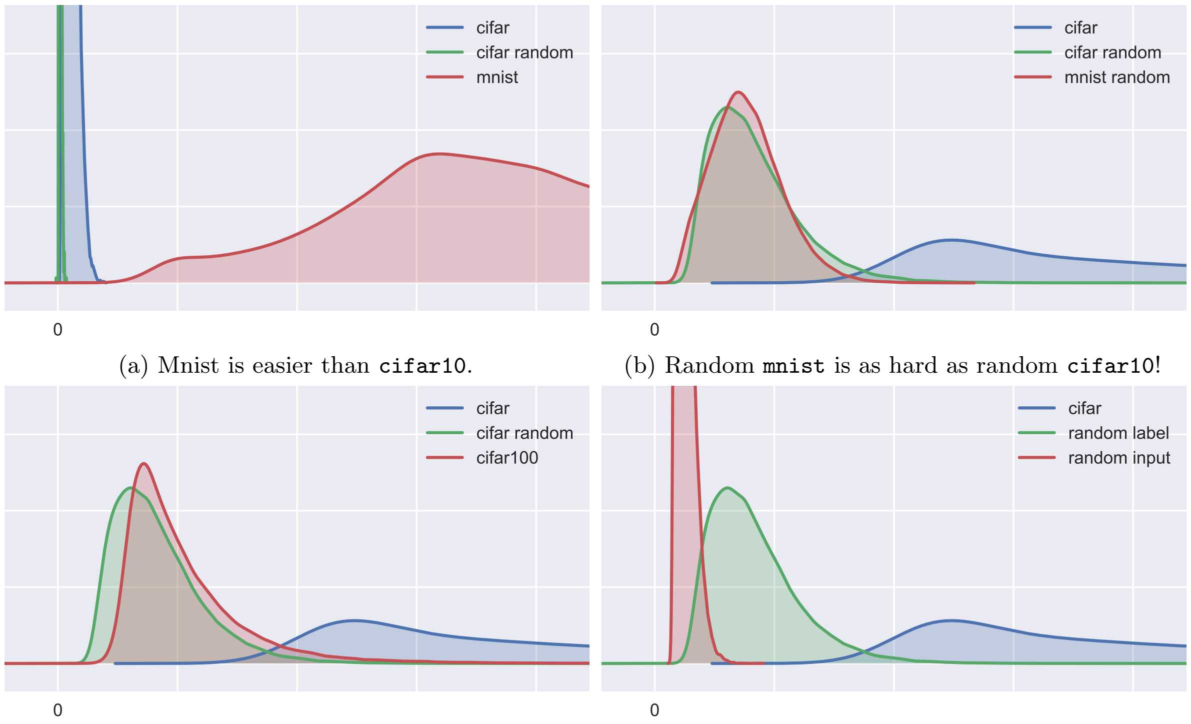
Theorem 1.1. *Let nonlinearities $(\sigma_1,\ldots,\sigma_L)$ and reference matrices $(M_1,\ldots,M_L)$ be given as above (i.e., σ_i is ρ_i-Lipschitz and $\sigma_i(0)=0$). Then with probability at least $1-\delta$ over an iid draw of n examples $((x_i,y_i))_{i=1}^n$, every margin $\gamma > 0$ and network $F_{\mathcal{A}} : \mathbb{R}^d \to \mathbb{R}^k$ with weight matrices $\mathcal{A} = (A_1,\ldots,A_L)$ satisfy*

$$\Pr\left[\arg\max_j F_{\mathcal{A}}(x)_j \neq y\right] \le \widehat{\mathcal{R}}_\gamma(F_{\mathcal{A}}) + \widetilde{\mathcal{O}}\left(\frac{\|X\|_2 R_{\mathcal{A}}}{\gamma n}\ln(n)\ln(W) + \sqrt{\frac{\ln(1/\delta)}{n}}\right),$$

where $\widehat{\mathcal{R}}_\gamma(f) \le n^{-1}\sum_i \mathbb{1}\left[f(x_i)_{y_i} \le \gamma + \max_{j\neq y_i} f(x_i)_j\right]$ and $\|X\|_2 = \sqrt{\sum_i \|x_i\|_2^2}$.

The full proof (based on metric entropy) is relegated to the appendix, but a sketch is provided in Section 3, along with a more general form (not limited to spectral norms), along with a (non-matching!) lower bound. Section 3 also gives a discussion of related work, but briefly it's essential to note that margin and Lipschitz-sensitive bounds have a long history in the neural networks literature (Bartlett, 1996; Anthony and Bartlett, 1999; Neyshabur et al., 2015); the distinction here is the sensitivity to specifically the spectral norm, as well as no explicit appearance of combinatorial quantities such as numbers of parameters or layers (outside of log terms, and indices to summations and products).

To close, miscellaneous observations and open problems are collected in Section 4.



(a) Mnist is easier than `cifar10`.

(b) Random `mnist` is as hard as random `cifar10`!

(c) `cifar100` is as hard as `cifar10` with random labels!

(d) Random inputs are harder than random labels.

Figure 3: A variety of margin distributions. Axes are re-scaled in Figure 3a, but identical in the other subplots; the `cifar10` (blue) and random `cifar10` (green) distributions are the same each time.

2 Generalization case studies via margin distributions

This section will now use the core generalization bound, stated in Theorem 1.1, to empirically study generalization behavior of neural networks via margin distributions.

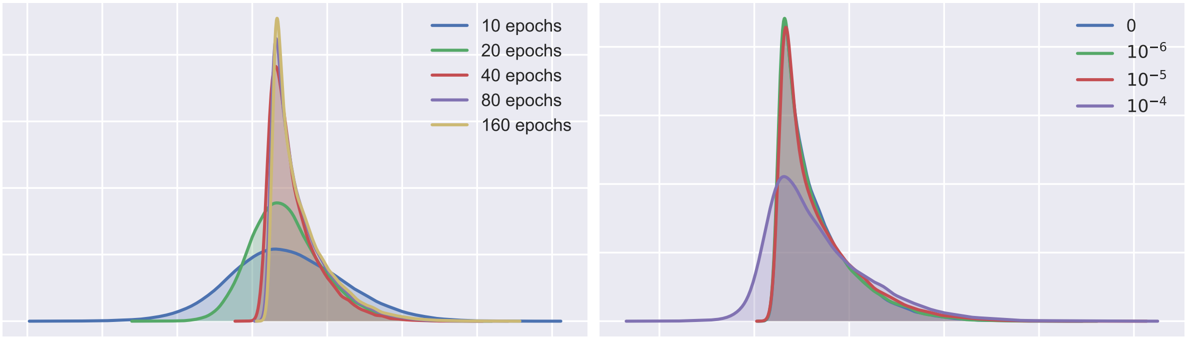
Before proceeding with the plots, it's a good time to give a more refined description of the margin distribution, one that is suitable for comparisons across datasets. Given n data points $((x_i, y_i))_{i=1}^n$ as rows of matrix $X \in \mathbb{R}^{n \times d}$ and predictor $F_\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^k$ with spectral complexity $R_\mathcal{A}$ (cf. eq. (1.1)), the (normalized) margin distribution is the univariate empirical distribution of the data points each transformed into a single scalar according to

$$(x, y) \mapsto \frac{F_\mathcal{A}(x)_y - \max_{i \neq y} F_\mathcal{A}(x)_i}{R_\mathcal{A} \|X\|_2 / n},$$

where spectral complexity $R_\mathcal{A}$ is from eq. (1.1). The normalization is thus derived from the bound in Theorem 1.1, but ignoring log terms.

Taken this way, the two margin distributions for two datasets can be interpreted as follows. Considering any fixed point on the horizontal axis, if the *cumulative* distribution of one density is lower than the other, then it corresponds to a lower right hand side in Theorem 1.1. For no reason other than visual interpretability, the plots here will instead depict a density estimate of the margin distribution. The vertical and horizontal axes are rescaled in different plots, but the random and true `cifar10` margin distributions are always the same.

A little more detail about the experimental setup is as follows. All experiments were implemented in Keras (Chollet et al., 2015). In order to minimize conflating effects of optimization and regularization, the optimization method was vanilla sgd with step size 0.01, and all regularization (weight decay, batch normalization, etc.) were disabled. "`cifar`" in general refers to `cifar10`, however `cifar100` will also be



(a) Margins across epochs.

(b) Various levels of l_2 regularization.

Figure 4

explicitly mentioned. The network architecture is essentially AlexNet (Krizhevsky et al., 2012) with all normalization/regularization removed, and with no adjustments of any kind (even to the learning rate) across the different experiments.

Comparing datasets. A first comparison is of `cifar10` and the standard `mnist` digit data. `mnist` is considered "easy", by which it is typically meant that any of a variety of methods can achieve roughly 1% test error. The "easiness" is corroborated by Figure 3a, where the margin distribution for `mnist` places all its mass far to the right of the mass for `cifar10`. Interestingly, randomizing the labels of `mnist`, as in Figure 3b, results in a margin distribution to the left of not only `cifar10`, but also slightly to the left of (but close to) `cifar10` with randomized labels.

Next, Figure 3c compares `cifar10` and `cifar100`, where `cifar100` uses the same input images as `cifar10`; indeed, `cifar10` is obtained from `cifar100` by collapsing the original 100 categories into 10 groups. Interestingly, `cifar100`, from the perspective of margin bounds, is just as difficult as `cifar10` with random labels. This is consistent with the large test observed error on `cifar100` (which has not been "optimized" in any way via regularization).

Lastly, Figure 3d replaces the `cifar10` *input images* with random images sampled from Gaussians matching the first- and second-order image statistics (see (Zhang et al., 2017) for similar experiments).

Convergence of margins. As was pointed out in Section 1, the weights of the neural networks do not seem to converge in the usual sense during training (the norms grow continually). However, as depicted in Figure 4a, the sequence of (normalized!) margin distributions is itself converging.

Regularization. As remarked in (Zhang et al., 2017), regularization only seems to bring minor benefits to test error (though adequate to be employed in all cutting edge results). This observation is certainly consistent with the margin distributions in Figure 4b, which do not improve (e.g., by shifting to the right) in any visible way under regularization. An open question, discussed further in Section 4, is to design regularization that improves margins.

3 Analysis of margin bound

This section will sketch the proof of Theorem 1.1, state a few generalizations, give a lower bound, and discuss related work.

3.1 Multiclass margin bound

The starting point of this analysis is a margin-based bound for multiclass prediction. To state the bound, first recall the *margin operator* $\mathcal{M}(v,y) := v_y - \max_{i \neq y} v_i$, and define the *ramp loss* as

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

and *ramp risk* as $\mathcal{R}_\gamma(f) := \mathbb{E}(\ell_\gamma(-\mathcal{M}(f(x), y)))$. Given a sample $S := ((x_1, y_1), \ldots, (x_n, y_n))$, define an empirical counterpart $\widehat{\mathcal{R}}_\gamma$ of $\mathcal{R}_\gamma$ as $\widehat{\mathcal{R}}_\gamma(f) := n^{-1} \sum_i \ell_\gamma(-\mathcal{M}(f(x_i), y_i))$; note that $\mathcal{R}_\gamma$ and $\widehat{\mathcal{R}}_\gamma$ respectively upper bound the probability and fraction of error on the source distribution and training set. Lastly, given a set of real-valued functions $\mathcal{H}$, define the *Rademacher complexity* as $\mathfrak{R}(\mathcal{H}_{|S}) := \mathbb{E} \sup_{h \in \mathcal{H}} \sum_{i=1}^n \epsilon_i h(x_i, y_i)$ where the expectation is over the Rademacher random variables $(\epsilon_1, \ldots, \epsilon_n)$ (meaning $\Pr[\epsilon_i = +1] = \Pr[\epsilon_i = -1] = 1/2$).

With this notation in place, the basic bound is as follows.

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

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

Then, with probability at least $1 - \delta$ over a sample S of size n, every $f \in \mathcal{F}$ satisfies

$$\Pr[\arg\max_i f(x)_i \neq y] \leq \widehat{\mathcal{R}}_\gamma(f) + 2\mathfrak{R}((\mathcal{F}_\gamma)_{|S}) + 3\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

This bound is a direct consequence of standard tools in Rademacher complexity. In order to instantiate this bound, covering numbers will be used to directly upper bound the Rademacher complexity term $\mathfrak{R}((\mathcal{F}_\gamma)_{|S})$. Interestingly, the choice of directly working in terms of covering numbers seems essential to providing a bound with no explicit dependence on k; by contrast, prior work primarily (solely?) handles multiclass via a Rademacher complexity analysis on each coordinate of a k-tuple of functions, and pays a factor $\sqrt{k}$ (Zhang, 2004).

3.2 Covering number complexity upper bounds

This subsection proves Theorem 1.1 via Lemma 3.1 by controlling the Rademacher complexity $\mathfrak{R}((\mathcal{F}_\gamma)_{|S})$ for networks with bounded spectral complexity via covering numbers. This first step is to prove a more general covering number bound (in terms of general operator norms) which will eventually be specialized with l_2 data norms and spectral operator norms to prove Theorem 1.1. A tantalizing direction for future work is to specialize the general bound in other ways, namely ones that are better adapted to the geometry of neural networks as encountered in practice.

The structure of the networks is the same as before; namely, given matrices $\mathcal{A} = (A_1, \ldots, A_L)$, define a mapping $F_\mathcal{A}$ as

$$F_\mathcal{A}(Z) := \sigma_L(A_L \sigma_{L-1}(A_{L_1} \cdots \sigma_1(A_1 Z) \cdots)),$$

with the convention $F_\emptyset(Z) = Z$.

- The input $Z \in \mathcal{V}_1$ is associated with some norm $|Z|_1 \leq B$. The subscript merely indicates an index, and does not refer to any l_1 norm. The vector space $\mathcal{V}_1$, and moreover the eventual collection of vector spaces $(\mathcal{V}_1, \ldots, \mathcal{V}_L)$, have no fixed meaning and are simply abstract vector spaces. However, when using these tools to prove Theorem 1.1, $\mathcal{V}_1 = \mathbb{R}^{d \times n}$ and $Z \in \mathcal{V}_1$ is formed by collecting the n data points into its columns; that is, $Z = X^\top$.

- The linear operators $A_i : \mathcal{V}_i \to \mathcal{W}_{i+1}$ are associated with some operator norm $|A_i|_{i\to i+1} \le c_i$:

  $$|A_i|_{i\to i+1} := \sup_{|Z|_i \le 1} |||A_i Z|||_{i+1} = c_i,$$

  where once again $|||\cdot|||_{i+1}$ is an abstract norm introduced purely to lend this proof more flexibility. As stated before, these linear operators $\mathcal{A} = (A_1, \dots, A_L)$ vary across functions $F_{\mathcal{A}}$. When used to prove Theorem 1.1, Z is a matrix (the forward image of data matrix $X^\top$ across layers), and these norms are all matrix norms.

- The ρ_i-Lipschitz mappings $\sigma_i : \mathcal{W}_{i+1} \to \mathcal{V}_{i+1}$ have ρ_i measured with respect to norms $|\cdot|_{i+1}$ and $|||\cdot|||_{i+1}$: for any $z, z' \in \mathcal{W}_{i+1}$,

  $$\left|\sigma_i(z) - \sigma_i(z')\right|_{i+1} \le \rho_i |||z - z'|||_{i+1}.$$

  These Lipschitz mappings are considered fixed within $F_{\mathcal{A}}$. Note again that these operations, when applied to prove Theorem 1.1, operate on matrices which represent the forward images of all data points together. Lipschitz properties of the standard coordinate-wise ReLU and max-pooling operators can be found in Appendix A.1.

Lastly, just before giving the bound, the notation for (proper) covering numbers is as follows. The notation $\mathcal{N}(U, \epsilon, \|\cdot\|)$ means the least cardinality of any subset $V \subseteq U$ which *covers* U at scale ϵ with norm $\|\cdot\|$, meaning

$$\sup_{A\in U} \min_{B\in V} \|A - B\| \le \epsilon.$$

Choices of U that will be used in the present work include both the image $\mathcal{F}_{|S}$ of data S under some function class F, as well as the conceptually simpler choice of a family of matrix products.

Lemma 3.2. *Let $(\epsilon_1, \dots, \epsilon_L)$ be given, along with fixed Lipschitz mappings $(\sigma_1, \dots, \sigma_L)$ (where σ_i is ρ_i-Lipschitz), and operator norm bounds constants $(c_1, \dots, c_L)$. Suppose the matrices $\mathcal{A} = (A_1, \dots, A_L)$ lie within $\mathcal{B}_1 \times \cdots \times \mathcal{B}_L$ where $A_i \in \mathcal{B}_i$ has $|A_i|_{i\to i+1} \le c_i$. Lastly, let data Z be given with $|Z|_1 \le B$. Then the neural net images $\mathcal{H}_Z := \{F_{\mathcal{A}}(Z) : \mathcal{A} \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_L\}$ have covering number bound*

$$\mathcal{N}\left(\mathcal{H}_Z, \tau, |\cdot|_{L+1}\right) \le \prod_{i=1}^{L} \sup_{\substack{(A_1,\dots,A_{i-1}) \\ \forall j<i\text{.}\, A_j\in\mathcal{B}_j}} \mathcal{N}\left(\left\{A_i F_{(A_1,\dots,A_{i-1})}(Z) : A_i \in \mathcal{B}_i\right\}, \epsilon_i, |||\cdot|||_i\right)$$

$$\text{where} \quad \tau := \sum_{j\le L} \epsilon_j \rho_j \prod_{l=j+1}^{L} \rho_l c_l.$$

The method of proof is to recursively build a cover with each layer: first a cover $\mathcal{F}_1$ of $\{A_1 Z : A_1 \in \mathcal{B}_1\}$ is produced, then for each $F \in \mathcal{F}_1$ a cover of $\{A_2 \sigma_1(F) : A_2 \in \mathcal{B}_2\}$ is produced and these are unioned together across all $F \in \mathcal{F}_1$; continuing this procedure arrives at a final cover at layer L.

It remains to show that the resulting set is indeed a cover. To this end, fix a particular $\mathcal{A} = (A_1, \dots, A_L)$ and Z, and for convenience define mapped elements F_{i+1} and G_i via $F_{i+1} = A_{i+1} G_i = A_{i+1}\sigma_i(A_i \cdots \sigma_1(A_1 Z) \cdots)$ which represent states of the network at layer $i+1$ while evaluating Z. Suppose inductively that there is a cover element $\widehat{G}_i$ close to G_i, in the sense that $|G_i - \widehat{G}_i|_{i+1}$ is small.

Now choose $\widehat{F}_{i+1}$ close to $A_{i+1}\widehat{G}_i$, meaning $|||A_{i+1}\widehat{G}_i - \widehat{F}_{i+1}|||_{i+2} \le \epsilon_{i+1}$. The essential chain of inequalities is

$$\begin{aligned}
|G_{i+1} - \widehat{G}_{i+1}|_{i+2} &\le \rho_{i+1} |||F_{i+1} - \widehat{F}_{i+1}|||_{i+2} \\
&\le \rho_{i+1} |||F_{i+1} - A_{i+1}\widehat{G}_i|||_{i+2} + \rho_{i+1} |||A_{i+1}\widehat{G}_i - \widehat{F}_{i+1}|||_{i+2} && \because \text{triangle inequality} \\
&\le \rho_{i+1} |A_{i+1}|_{i+1\to i+2} \left|G_i - \widehat{G}_i\right|_{i+1} + \rho_{i+1}\epsilon_{i+1}, && \because \text{choice of } \widehat{F}_{i+1}
\end{aligned}$$

where the term $|G_i - \widehat{G}_i|_{i+1}$ is handled by induction. These inequalities are similar to those in an existing covering number proof (Anthony and Bartlett, 1999, Chapter 12) (itself rooted in the earlier work of Bartlett (1996)); however (a) that proof operates node by node, and can not take advantage of special norms on $\mathcal{A}$, and (b) that proof does not maintain an empirical cover across layers, instead explicitly covering the parameters of all weight matrices, which incurs the number of parameters as a multiplicative factor. The idea here to push an empirical cover through layers, meanwhile, is reminiscent of VC dimension proofs for neural networks (Anthony and Bartlett, 1999, Chapter 8).

Returning to the task of proving Theorem 1.1, what remains is to instantiate the general covering bound, Lemma 3.2, in a way that gives rise to spectral norms.

Theorem 3.3. *Let fixed nonlinearities $(\sigma_1, \ldots, \sigma_L)$ and reference matrices $(M_1, \ldots, M_L)$ be given, where σ_i is ρ_i-Lipschitz and $\sigma_i(0) = 0$. Let spectral norm bounds $(s_1, \ldots, s_L)$, and matrix l_1 norm bounds $(b_1, \ldots, b_L)$ be given. Let data matrix $X \in \mathbb{R}^{n \times d}$ be given, where the n rows correspond to data points. Let $\mathcal{H}_X$ denote the family of matrices obtained by evaluating X with all choices of network $F_\mathcal{A}$:*

$$\mathcal{H}_X := \left\{ F_\mathcal{A}(X^T) \ : \ \mathcal{A} = (A_1, \ldots, A_L), \ \|A_i\|_\sigma \le s_i, \ \|A_i - M_i\|_1 \le b_i \right\},$$

where each matrix has dimension at most W along each axis. Then for any $\epsilon > 0$,

$$\ln \mathcal{N}(\mathcal{H}_X, \epsilon, \|\cdot\|_2) \le \frac{\|X\|_2^2 \ln(2W^2)}{\epsilon^2} \left(\prod_{j=1}^{L} s_j^2 \rho_j^2 \right) \left(\sum_{i=1}^{L} \left(\frac{b_i}{s_i} \right)^{2/3} \right)^3 .$$

The key to the proof of Theorem 3.3 via Lemma 3.2 is the following matrix covering lemma.

Lemma 3.4. *Let positive reals a, b, ϵ and positive integer p be given, along with matrix $X \in \mathbb{R}^{n \times d}$ with $\max_i \|X\mathbf{e}_i\|_2 \le b$. Then*

$$\ln \mathcal{N}\left(\left\{ XA : A \in \mathbb{R}^{d \times p}, \|A\|_1 \le a \right\}, \epsilon, \|\cdot\|_2 \right) \le \left\lceil \frac{a^2 b^2}{\epsilon^2} \right\rceil \ln(2dp).$$

The proof of Lemma 3.4 relies upon the *Maurey sparsification lemma* (Pisier, 1980) which is stated in terms of sparsifying convex hulls, and in its use here is inspired by covering number bounds for linear predictors (Zhang, 2002). Following those techniques, it is possible to produce a bound that scales with $\|A\|_2$ and $\|X\|_2$, but even for the case of the identity matrix $X = I$, this incurs an extra dimension factor, thus the use of $\|A\|_1$ here as a simplifying choice which helps Theorem 1.1 avoid any appearance of W and L outside of log terms.

The path to prove Theorem 1.1 is now clear. Thanks to standard conversions between Rademacher complexity and covering numbers, the Rademacher complexity term in the generic margin bound Lemma 3.1 can be controlled via the covering number estimate from Theorem 3.3.

Lemma 3.5. *Let fixed nonlinearities $(\sigma_1, \ldots, \sigma_L)$ and reference matrices $(M_1, \ldots, M_L)$ be given where σ_i is ρ_i-Lipschitz and $\sigma_i(0) = 0$. Further let margin $\gamma > 0$, data bound B, spectral norm bounds $(s_i)_{i=1}^L$, and l_1 norm bounds $(b_i)_{i=1}^L$ be given. Then with probability at least $1 - \delta$ over an iid draw of n examples $((x_i, y_i))_{i=1}^n$ with $\sqrt{\sum_i \|x_i\|_2^2} \le B$, every network $F_\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^k$ whose weight matrices $\mathcal{A} = (A_1, \ldots, A_L)$ obey $\|A_i\|_\sigma \le s_i$ and $\|A_i - M_i\|_1 \le b_i$ satisfies*

$$\Pr\left[\arg\max_j F_\mathcal{A}(x)_j \ne y \right] \le \widehat{\mathcal{R}}_\gamma(f) + \frac{8}{n} + \frac{72 B \ln(2W) \ln(n)}{\gamma n} \left(\prod_{i=1}^{L} s_i \rho_i \right) \left(\sum_{i=1}^{L} \frac{b_i^{2/3}}{s_i^{2/3}} \right)^{3/2} + 3\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Theorem 1.1 follows from Lemma 3.5 by union bounding over various choices of γ, B, $(s_1, \ldots, s_L)$, $(b_1, \ldots, b_L)$.

3.3 Rademacher complexity lower bounds

By reduction to the linear case (i.e., removing all nonlinearities), it is easy to provide a lower bound on the Rademacher complexity of the networks studied here. Unfortunately, this bound only scales with the product of spectral norms, and not the other terms in $R_{\mathcal{A}}$ (cf. eq. (1.1)).

Theorem 3.6. *Consider the setting of Theorem 3.3, but all nonlinearities are the ReLU $z \mapsto \max\{0, z\}$, and all non-output dimensions are at least 2 (meaning $W \geq 2$). Let data $S := (x_1, \ldots, x_n)$ be collected into data matrix $X \in \mathbb{R}^{n \times d}$. Then, for any scalar $r > 0$,*

$$\mathfrak{R}\left(\left\{F_{\mathcal{A}} : \mathcal{A} = (A_1, \ldots, A_L), \prod_i \|A_i\|_\sigma \leq r\right\}_{|S}\right) \geq \Omega(r \cdot \|X\|_2). \tag{3.1}$$

Note that, due to the nonlinearity, the lower bound should indeed depend on $\prod_i \|A_i\|_\sigma$ and not $\|\prod_i A_i\|_\sigma$; as a simple sanity check, there exist networks for which the latter quantity is 0, but the network does not compute the zero function.

3.4 Related work

To close this section on proofs, it is a good time to summarize connections to existing literature.

Margin theory originally arose to analyze 2-layer networks (Bartlett, 1996), indeed with a proof technique that inspired the layer-wise induction used to prove Theorem 1.1 in the present work. Margin theory was quickly extended to many other settings (see for instance the survey by Boucheron et al. (2005)), one major success being an explanation of the generalization ability of boosting methods, which exhibit an explicit growth in the size of the function class over time, but a stable excess risk (Schapire et al., 1997). The contribution of the present work is to provide a margin bound (and corresponding Rademacher analysis) which can be adapted to various operator norms at each layer. Additionally, the present work operates in the multiclass setting, and avoids an explicit dependence on the number of classes k, which seems to appear in prior work (Zhang, 2004; Tewari and Bartlett, 2007).

There are numerous generalization bounds for neural networks, including VC-dimension and fat-shattering bounds (many of these can be found in (Anthony and Bartlett, 1999)). Scale-sensitive analysis of neural networks started with (Bartlett, 1996), which can be interpreted in the present setting as utilizing data norm $\|\cdot\|_\infty$ and operator norm $\|\cdot\|_{\infty\to\infty}$ (equivalently, the $\|\cdot\|_{1,\infty}$ matrix norm). This analysis can be adapted to give a Rademacher complexity analysis (Bartlett and Mendelson, 2002), and has been adapted to other norms (Neyshabur et al., 2015), although the preceding $\|\cdot\|_\infty$ setting is still needed to avoid extra combinatorial factors. More work is still needed to develop complexity analyses that have matching upper and lower bounds, and also to determine which norms are well-adapted to neural networks as used in practice.

The present analysis utilizes covering numbers, and is most closely connected to earlier covering number bounds (Anthony and Bartlett, 1999, Chapter 12), themselves based on the earlier fat-shattering analysis (Bartlett, 1996), however the technique here of pushing an empirical cover through layers is akin to VC dimension proofs for neural networks (Anthony and Bartlett, 1999). The use of Maurey's sparsification lemma was inspired by linear predictor covering number bounds (Zhang, 2002).

4 Further observations and open problems

Adversarial examples. Adversarial examples are a phenomenon where the neural network predictions can be altered by adding seemingly imperceptible noise to an input (Goodfellow et al., 2014). This phenomenon can be connected to margins as follows. The margin is nothing more than the distance an input must traverse before its label is flipped; consequently, low margin points are more susceptible to adversarial noise than high margin points. Concretely, taking the 100 lowest margin inputs from `cifar10` and adding uniform noise at scale 0.15 yielded flipped labels on 5.86% of the images, whereas the same level of noise on high margin points yielded 0.04% flipped labels. Can the bounds here suggest a way to defend against adversarial examples?

Regularization. It was observed in (Zhang et al., 2017) that explicit regularization contributes little to the generalization performance of neural networks. In the margin framework, standard weight decay (l_2) regularization seemed to have little impact on margin distributions in Section 2. On the other hand, in the boosting literature, special types of regularization were developed to maximize margins (Shalev-Shwartz and Singer, 2008); perhaps a similar development can be performed here?

SGD. The present analysis applies to predictors that have large margins; what is missing is an analysis verifying that sgd applied to standard neural networks returns large margin predictors! Indeed, perhaps sgd returns not simply large margin predictors, but predictors that are well-behaved in a variety of other ways which can be directly translated into refined generalization bounds.

Improvements to Theorem 1.1. There are many way to improve Theorem 1.1. One is simply to determine better lower bounds. Another question is whether there can be a better choice of layer geometries (norms) which yields better bounds on practical networks. Lastly, is there a way to replace the nonlinearities' worst-case Lipschitz constant with an (empirically) averaged quantity?

Rademacher vs. covering. Is it possible to prove Theorem 1.1 solely via Rademacher complexity, with no invocation of covering numbers?

Acknowledgements

The authors thank Srinadh Bhojanapalli, Ryan Jian, Behnam Neyshabur, Maxim Raginsky, Andrew J. Risteski, and Belinda Tzen for useful conversations and feedback. The authors thank Ben Recht for giving a provocative lecture at the Simons Institute, stressing the need for understanding of both generalization and optimization of neural networks. M.T. and D.F. acknowledge the use of a GPU machine provided by Karthik Sridharan and made possible by a NVIDIA GPU grant. D.F. acknowledges the support of the NDSEG fellowship. Lastly, the authors are grateful to La Burrita (both the north and the south Berkeley campus locations) for upholding the glorious tradition of the California Burrito.

References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Peter L. Bartlett. For valid generalization the size of the weights is more important than the size of the network. In *NIPS*, 1996.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, Nov 2002.

Stéphane Boucheron, Olivier Bousquet, and Gabor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *ICML*, 2017.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2014. `arXiv:1412.6572 [stat.ML]`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffery Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2012.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *COLT*, 2015.

Gilles Pisier. Remarques sur un résultat non publié de b. maurey. *Séminaire Analyse fonctionnelle (dit)*, pages 1–12, 1980.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML*, pages 322–330, 1997.

Shai Shalev-Shwartz and Yoram Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *COLT*, 2008.

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.

A Proofs

This appendix collects various proofs omitted from the main text.

A.1 Lipschitz properties of ReLU and max-pooling nonlinearities

The standard *ReLU* ("Rectified Linear Unit") is the univariate mapping

$$\sigma_{\mathrm{r}}(r) := \max\{0, r\}.$$

When applied to a vector or a matrix, it operates coordinate-wise. While the ReLU is currently the most popular choice of univariate nonlinearity, another common choice is the *sigmoid* $r \mapsto 1/(1+\exp(-r))$. More generally, these univariate nonlinearities are Lipschitz, and this carries over to their vector and matrix forms as follows.

Lemma A.1. *If $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ is ρ-Lipschitz along every coordinate, then it is ρ-Lipschitz according to $\|\cdot\|_p$ for any $p \geq 1$.*

Proof. for any $z, z' \in \mathbb{R}^d$,

$$\|\sigma(z) - \sigma(z')\|_p = \left(\sum_i |\sigma(z)_i - \sigma(z')_i|^p\right)^{1/p} \leq \left(\sum_i \rho^p |z_i - z'_i|^p\right)^{1/p} = \rho\|z - z'\|_p.$$

□

Define a *max-pooling operator* $\mathcal{P}$ as follows. Given an input and output pair of finite-dimensional vector spaces $\mathcal{T}$ and $\mathcal{T}'$ (possibly arranged as matrices or tensors), the max-pooling operator iterates over a collection of sets of indices $\mathcal{Z}$ (whose cardinality is equal to the dimension of $\mathcal{T}$'), and for each element of $Z_i \in \mathcal{Z}$ sets the corresponding coordinate i in the output to the maximum entry of the input over Z_i: given $T \in \mathcal{T}$,

$$\mathcal{P}(T)_i := \max_{j \in Z_i} T_j.$$

The following Lipschitz constant of pooling operators will depend on the number of times each coordinate is accessed across elements of $\mathcal{Z}$; when this operator is used in computer vision, the number of times is typically a small constant, for instance 5 or 9 (Krizhevsky et al., 2012).

Lemma A.2. *Suppose that each coordinate j of the input appears in at most m elements of the collection $\mathcal{Z}$. Then the max-pooling operator $\mathcal{P}$ is $m^{1/p}$-Lipschitz wrt $\|\cdot\|_p$ for any $p \geq 1$. In particular, the max-pooling operator is 1-Lipschitz whenever $\mathcal{Z}$ forms a partition.*

Proof. Let $T, T' \in \mathcal{T}$ be given. First consider any fixed set of indices $Z \in \mathcal{Z}$, and suppose without loss of generality that $\mathcal{P}(T)_Z = \max_{j \in Z} T_j \geq \max_{j \in Z} T'_j$. Then

$$|\mathcal{P}(T)_Z - \mathcal{P}(T')_Z|^p = \left(\min_{j' \in Z} \max_{j \in Z} T_j - T'_{j'}\right)^p = \max_{j \in Z}\left(T_j - T'_j\right)^p \leq \sum_{j \in Z}\left|T_j - T'_j\right|^p.$$

Consequently,

$$\begin{aligned}
\|\mathcal{P}(T) - \mathcal{P}(T')\|_p &= \left(\sum_i |\mathcal{P}(T)_i - \mathcal{P}(T')_i|^p\right)^{1/p} = \left(\sum_{Z \in \mathcal{Z}} |\mathcal{P}(T)_Z - \mathcal{P}(T')_Z|^p\right)^{1/p} \\
&\leq \left(\sum_{Z \in \mathcal{Z}} \sum_{j \in Z} |T_j - T'_j|^p\right)^{1/p} = \left(\sum_j \sum_{Z \in \mathcal{Z} : j \in Z} |T_j - T'_j|^p\right)^{1/p} \\
&\leq \left(m \sum_j |T_j - T'_j|^p\right)^{1/p} = m^{1/p}\|T - T'\|_p.
\end{aligned}$$

□

A.2 Margin properties in Section 3.1

The goal of this subsection is to prove the general margin bound in Lemma 3.1. To this end, it is first necessary to establish a few properties of the margin operator $\mathcal{M}(v,j) := v_j - \max_{i \neq j} v_i$ and of the ramp loss ℓ_λ.

Lemma A.3. *For every j and every $p \geq 1$, $\mathcal{M}(\cdot, j)$ is 2-Lipschitz wrt $\|\cdot\|_p$.*

Proof. Let v, v', j be given, and suppose (without loss of generality) $\mathcal{M}(v,j) \geq \mathcal{M}(v',j)$. Choose coordinate $i \neq j$ so that $\mathcal{M}(v',j) = v'_j - v'_i$. Then

$$\begin{aligned}\mathcal{M}(v,j) - \mathcal{M}(v',j) &= \left(v_j - \max_{l \neq j} v_j\right) - \left(v'_j - v'_i\right) = v_j - v'_j + v'_i + \min_{l \neq j}(-v_l) \\ &\leq \left(v_j - v'_j\right) + \left(v'_i - v_i\right) \leq 2\|v - v'\|_\infty \leq 2\|v - v'\|_p.\end{aligned}$$

□

Next, recall the definition of the ramp loss

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

and of the ramp risk

$$\mathcal{R}_\gamma(f) := \mathbb{E}(\ell_\gamma(-\mathcal{M}(f(x), y))).$$

(These quantities are standard; see for instance (Boucheron et al., 2005; Zhang, 2004; Tewari and Bartlett, 2007).)

Lemma A.4. *For any $f : \mathbb{R}^d \to \mathbb{R}^k$ and every $\gamma > 0$,*

$$\Pr[\arg\max_i f(x)_i \neq y] \leq \Pr[\mathcal{M}(f(x), y) \geq 0] \leq \mathcal{R}_\gamma(f),$$

where the $\arg\max$ *follows any deterministic tie-breaking strategy.*

Proof.

$$\begin{aligned}\Pr[\arg\max_i f(x)_i \neq y] &\leq \Pr[\max_{i \neq y} f(x)_i \geq f(x)_y] \\ &= \Pr[-\mathcal{M}(f(x), y) \geq 0] \\ &= \mathbb{E}\mathbb{1}[-\mathcal{M}(f(x), y) \geq 0] \\ &\leq \mathbb{E}\ell_\gamma(-\mathcal{M}(f(x), y))\end{aligned}$$

□

With these tools in place, the proof of Lemma 3.1 is straightforward.

Proof of Lemma 3.1. Since ℓ_γ has range $[0,1]$, it follows by standard properties of Rademacher complexity (Mohri et al., 2012, Theorem 3.1) that with probability at least $1 - \delta$, every $f \in \mathcal{F}$ satisfies

$$\mathcal{R}_\gamma(f) \leq \widehat{\mathcal{R}}_\gamma(f) + 2\mathfrak{R}((\mathcal{F}_\gamma)_{|S}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

The bound now follows by applying Lemma A.4 to the left hand side. □

A.3 Proof of general covering bound (Lemma 3.2)

Proof of Lemma 3.2. Inductively construct covers $\mathcal{F}_1, \ldots, \mathcal{F}_L$ of $\mathcal{W}_2, \ldots, \mathcal{W}_{L+1}$ as follows.

- Choose an ϵ_1-cover $\mathcal{F}_1$ of $\{A_1 Z : A_1 \in \mathcal{B}_1\}$, thus
$$|\mathcal{F}_1| \leq \mathcal{N}(\{A_1 Z : A_1 \in \mathcal{B}_1\}, \epsilon_1, |||\cdot|||_2) =: N_1.$$
- For every element $F \in \mathcal{F}_i$, construct an ϵ_{i+1}-cover $\mathcal{G}_{i+1}(F)$ of
$$\left\{A_{i+1}\sigma_i(F) : A_{i+1} \in \mathcal{B}_{i+1}\right\}.$$
Since the covers are proper, meaning $F = A_i F_{(A_1,\ldots,A_{i-1})}(Z)$ for some matrices $(A_1, \ldots, A_i) \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_i$, it follows that
$$\left|\mathcal{G}_{i+1}(F)\right| \leq \sup_{\substack{(A_1,\ldots,A_i)\\ \forall j\leq i\centerdot A_j \in \mathcal{B}_j}} \mathcal{N}\left(\left\{A_{i+1}F_{A_1,\ldots,A_i}(Z) : A_{i+1} \in \mathcal{B}_{i+1}\right\}, \epsilon_{i+1}, |||\cdot|||_{i+2}\right) =: N_{i+1}.$$
Lastly form the cover
$$\mathcal{F}_{i+1} := \bigcup_{F \in \mathcal{F}_i} \mathcal{G}_{i+1}(F),$$
whose cardinality satisfies
$$|\mathcal{F}_{i+1}| \leq |\mathcal{F}_i| \cdot N_{i+1} \leq \prod_{l=1}^{i+1} N_l.$$

Define $\mathcal{F} := \left\{\sigma_L(F) : F \in \mathcal{F}_L\right\}$; by construction, $\mathcal{F}$ satisfies the desired cardinality constraint. to show that it is indeed a cover, fix any $(A_1, \ldots, A_L)$ satisfying the above constraints, and for convenience define recursively the mapped elements
$$F_1 = A_1 X \in \mathcal{W}_2, \qquad G_i = \sigma_i(F_i) \in \mathcal{V}_{i+1} \qquad F_{i+1} = A_{i+1}G_i \in \mathcal{W}_{i+2}.$$
The goal is to exhibit $\widehat{G}_L \in \mathcal{F}$ satisfying $|G_L - \widehat{G}_L|_{L+1} \leq \tau$. To this end, inductively construct approximating elements $(\widehat{F}_i, \widehat{G}_i)$ as follows.

- Base case: set $\widehat{G}_0 = X$.
- Choose $\widehat{F}_i \in \mathcal{F}_i$ with $|||A_i\widehat{G}_{i-1} - \widehat{F}_i|||_{i+1} \leq \epsilon_i$, and set $\widehat{G}_i := \sigma_i(\widehat{F}_i)$.

To complete the proof, it will be shown inductively that
$$|G_i - \widehat{G}_i|_{i+1} \leq \sum_{1\leq j\leq i} \epsilon_j \rho_j \prod_{l=j+1}^{i} \rho_l c_l.$$
For the base case,
$$|G_0 - \widehat{G}_0|_1 = 0.$$
For the inductive step,
$$\begin{aligned}
|G_{i+1} - \widehat{G}_{i+1}|_{i+2} &\leq \rho_{i+1}|||F_{i+1} - \widehat{F}_{i+1}|||_{i+2}\\
&\leq \rho_{i+1}|||F_{i+1} - A_{i+1}\widehat{G}_i|||_{i+2} + \rho_{i+1}|||A_{i+1}\widehat{G}_i - \widehat{F}_{i+1}|||_{i+2}\\
&\leq \rho_{i+1}|A_{i+1}|_{i+1\to i+2}\left|G_i - \widehat{G}_i\right|_{i+1} + \rho_{i+1}\epsilon_{i+1}\\
&\leq \rho_{i+1}c_{i+1}\left(\sum_{j\leq i}\epsilon_j\rho_j\prod_{l=j+1}^{i}\rho_l c_l\right) + \rho_{i+1}\epsilon_{i+1}\\
&= \sum_{j\leq i+1}\epsilon_j\rho_j\prod_{l=j+1}^{i+1}\rho_l c_l.
\end{aligned}$$

□

A.4 Proof of spectral covering bound (Theorem 3.3)

The first step is to establish the matrix covering bound in Lemma 3.4. This proof is inspired by covering number bounds for linear predictors due to Zhang (2002), which centrally rely upon the following sparsification lemma due to Maurey.

Lemma A.5 (Maurey; cf. (Pisier, 1980), (Zhang, 2002, Lemma 1))**.** *Fix Hilbert space $\mathcal{H}$ with norm $\|\cdot\|$. Let $U \in \mathcal{H}$ be given with representation $U = \sum_{i=1}^{d} \alpha_i V_i$ where $V_i \in \mathcal{H}$ and $\alpha \in \mathbb{R}^d_{\geq 0} \setminus \{0\}$. Then for any positive integer k, there exists a choice of nonnegative integers $(k_1, \dots, k_d)$, $\sum_i k_i = k$, such that*

$$\left\| U - \frac{\|\alpha\|_1}{k} \sum_{i=1}^{d} k_i V_i \right\|^2 \leq \frac{\|\alpha\|_1}{k} \sum_{i=1}^{d} \alpha_i \|V_i\|^2 \leq \frac{\|\alpha\|_1^2}{k} \max_i \|V_i\|^2.$$

Proof. Set $\beta := \|\alpha\|_1$ for convenience, and let $(W_1, \dots, W_k)$ denote k iid random variables where $\Pr[W_1 = \beta V_i] := \alpha_i/\beta$. Define $W := k^{-1} \sum_{i=1}^{k} W_i$, whereby

$$\mathbb{E}W = \mathbb{E}W_1 = \sum_{i=1}^{d} \beta V_i \left(\frac{\alpha_i}{\beta} \right) = U.$$

Consequently

$$\begin{aligned}
\mathbb{E}\|U - W\|^2 &= \frac{1}{k^2} \mathbb{E} \left\| \sum_i (U - W_i) \right\|^2 = \frac{1}{k^2} \mathbb{E} \left(\sum_i \|U - W_i\|^2 + \sum_{i \neq j} \langle U - W_i, U - W_j \rangle \right) \\
&= \frac{1}{k} \mathbb{E}\|U - W_1\|^2 = \frac{1}{k} \left(\mathbb{E}\|W_1\|^2 - \|U\|^2 \right) \leq \frac{1}{k} \mathbb{E}\|W_1\|^2 \\
&= \frac{1}{k} \sum_{i=1}^{d} \frac{\alpha_i}{\beta} \|\beta V_i\|^2 = \frac{\beta}{k} \sum_{i=1}^{d} \alpha_i \|V_i\|^2 \\
&\leq \frac{\beta^2}{k} \max_i \|V_i\|^2.
\end{aligned}$$

To finish, by the probabilistic method, there exists integers $(j_1, \dots, j_k) \in \{1, \dots, d\}^k$ and an assignment $\widehat{W}_i := \beta V_{j_i}$ and $\widehat{W} := k^{-1} \sum_{i=1}^{k} \widehat{W}_i$ such that

$$\left\| U - \widehat{W} \right\|^2 \leq \mathbb{E}\|U - W\|^2.$$

The result now follows by defining integers $(k_1, \dots, k_d)$ according to $k_i := \sum_{l=1}^{k} \mathbb{1}[j_l = i]$. □

The Maurey sparsification lemma easily gives the matrix covering bound in Lemma 3.4.

Proof of Lemma 3.4. Let matrix $X \in \mathbb{R}^{n \times d}$ be given, set $N := 2dp$ and $k := \lceil a^2 b^2 / \epsilon^2 \rceil$, and define

$$\begin{aligned}
\{V_1, \dots, V_N\} &:= \left\{ sX\mathbf{e}_i\mathbf{e}_j^\top : s \in \{-1, +1\}, i \in \{1, \dots, d\}, j \in \{1, \dots, p\} \right\}, \\
\mathcal{C} &:= \left\{ \frac{a}{k} \sum_{i=1}^{N} k_i V_i : k_i \geq 0, \sum_{i=1}^{N} k_i = k \right\} = \left\{ \frac{a}{k} \sum_{j=1}^{k} V_{i_j} : (i_1, \dots, i_k) \in [N]^k \right\},
\end{aligned}$$

where the k_i's are integers, and $\|V_i\|_2 \leq \max_i \|X\mathbf{e}_i\|_2 \leq b$ by construction. It will now be shown that $\mathcal{C}$ is the desired cover. Firstly, $|\mathcal{C}| \leq N^k$ by construction, namely the final equality above. Secondly, let A with $\|A\|_1 \leq a$ be given, and note that

$$XA = X \sum_{i=1}^{d} \sum_{j=1}^{p} A_{ij} \mathbf{e}_i \mathbf{e}_j^\top = \|A\|_1 \sum_{i=1}^{d} \sum_{j=1}^{p} \frac{A_{ij}}{\|A\|_1} \left(X\mathbf{e}_i\mathbf{e}_j^\top \right) \in a \cdot \mathrm{conv}(\{V_1, \dots, V_N\}),$$

where $\mathrm{conv}(\{V_1, \ldots, V_n\})$ is the convex hull of $\{V_1, \ldots, V_N\}$. Consequently, by Lemma A.5, there exist nonnegative integers $(k_1, \ldots, k_N)$ with $\sum_i k_i = k$ with

$$\left\| XA - \frac{a}{k} \sum_{i=1}^{N} k_i V_i \right\|_2^2 \leq \frac{a^2 b^2}{k} \leq \epsilon^2.$$

□

Together, the preceding pieces give the proof of Theorem 3.3.

Proof of Theorem 3.3. First dispense with the parenthetical statement regarding coordinate-wise ReLU and max-pooling operaters, which are Lipschitz by Lemmas A.1 and A.2. The rest of the proof is now a consequence of Lemma 3.2 with all data norms set to the l_2 norm ($|\cdot|_i = |||\cdot|||_i = \|\cdot\|_2$), all operator norms set to the spectral norm ($|\cdot|_{i\to i+1} = \|\cdot\|_\sigma$), the matrix constraint sets set to $\mathcal{B}_i = \left\{A_i : \|A_i\|_\sigma \leq s_i, \|A_i - M_i\|_1 \leq b_i\right\}$, and lastly the per-layer cover resolutions $(\epsilon_1, \ldots, \epsilon_L)$ set according to

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

By this choice, it follows that the final cover resolution τ provided by Lemma 3.2 satisfies

$$\tau \leq \sum_{j\leq L} \epsilon_j \rho_j \prod_{l=j+1}^{L} \rho_l s_l = \sum_{j\leq L} \alpha_j \epsilon = \epsilon.$$

The key technique in the remainder of the proof is to apply Lemma 3.2 with the covering number estimate from Lemma 3.4, but centering the covers at M_i (meaning the cover at layer i is of matrices $\mathcal{B}_i$ where $A_i \in \mathcal{B}_i$ satisfies $\|A_i - M_i\|_1 \leq b_i$), and collecting $(x_1, \ldots, x_n)$ as rows of matrix $X \in \mathbb{R}^{n\times d}$.

To start, the covering number estimate from Lemma 3.2 can be combined with Lemma 3.4 to give

$$\begin{aligned}
\ln \mathcal{N}(\mathcal{H}_{|S}, \epsilon, \|\cdot\|_2) &\leq \sum_{i=1}^{L} \sup_{\substack{(A_1,\ldots,A_{i-1})\\ \forall j<i\centerdot A_j \in \mathcal{B}_j}} \ln \mathcal{N}\left(\left\{A_i F_{(A_1,\ldots,A_{i-1})}(X^\top) : A_i \in \mathcal{B}_i\right\}, \epsilon_i, \|\cdot\|_2\right) \\
&\overset{(*)}{=} \sum_{i=1}^{L} \sup_{\substack{(A_1,\ldots,A_{i-1})\\ \forall j<i\centerdot A_j \in \mathcal{B}_j}} \ln \mathcal{N}\left(\left\{F_{(A_1,\ldots,A_{i-1})}(X^\top)^\top (A_i - M_i)^\top : \|A_i - M_i\|_1 \leq b_i, \|A_i\|_\sigma \leq s_i\right\}, \epsilon_i, \|\cdot\|_2\right) \\
&\leq \sum_{i=1}^{L} \sup_{\substack{(A_1,\ldots,A_{i-1})\\ \forall j<i\centerdot A_j \in \mathcal{B}_j}} \ln \mathcal{N}\left(\left\{F_{(A_1,\ldots,A_{i-1})}(X^\top)^\top (A_i - M_i)^\top : \|A_i - M_i\|_1 \leq b_i\right\}, \epsilon_i, \|\cdot\|_2\right) \\
&\leq \sum_{i=1}^{L} \sup_{\substack{(A_1,\ldots,A_{i-1})\\ \forall j<i\centerdot A_j \in \mathcal{B}_j}} \frac{b_i^2 \max_j \|F_{(A_1,\ldots,A_{i-1})}(X^\top)^\top \mathbf{e}_j\|_2^2}{\epsilon_i^2} \ln(2W^2), \qquad \text{(A.1)}
\end{aligned}$$

where $(*)$ follows first since l_2 covering a matrix and its transpose is the same, and secondly since the cover can be translated by $F_{(A_1,\ldots,A_{i-1})}(X^\top)^\top M_i^\top$ without changing its cardinality. In order to simplify this expression, note for any $(A_1, \ldots, A_{i-1})$ that

$$\begin{aligned}
\max_j \|F_{(A_1,\ldots,A_{i-1})}(X^\top)^\top \mathbf{e}_j\|_2 &\leq \|F_{(A_1,\ldots,A_{i-1})}(X^\top)^\top\|_2 \\
&= \|F_{(A_1,\ldots,A_{i-1})}(X^\top)\|_2 \\
&= \|\sigma_{i-1}(A_{i-1} F_{(A_1,\ldots,A_{i-2})}(X^\top) - \sigma_{i-1}(0)\|_2 \\
&\leq \rho_{i-1} \|A_{i-1} F_{(A_1,\ldots,A_{i-2})}(X^\top) - 0\|_2 \\
&\leq \rho_{i-1} \|A_{i-1}\|_\sigma \|F_{(A_1,\ldots,A_{i-2})}(X^\top)\|_2,
\end{aligned}$$

which by induction gives

$$\max_j \|F_{(A_1,\ldots,A_{i-1})}(X^\top)^\top \mathbf{e}_j\|_2 \le \|X\|_2 \prod_{j=1}^{i-1} \rho_j \|A_j\|_\sigma. \tag{A.2}$$

Combining eqs. (A.1) and (A.2), then expanding the choice of ϵ_i and collecting terms,

$$\begin{aligned}
\ln \mathcal{N}(\mathcal{H}_{|S}, \epsilon, \|\cdot\|_2) &\le \sum_{i=1}^{L} \sup_{\substack{(A_1,\ldots,A_{i-1}) \\ \forall j<i\text{.} A_j \in \mathcal{B}_j}} \frac{b_i^2 \|X\|_2^2 \prod_{j<i} \rho_j^2 \|A_j\|_\sigma^2}{\epsilon_i^2} \ln(2W^2) \\
&\le \sum_{i=1}^{L} \frac{b_i^2 B^2 \prod_{j<i} \rho_j^2 s_j^2}{\epsilon_i^2} \ln(2W^2) \\
&= \frac{B^2 \ln(2W^2) \prod_{j=i}^{L} \rho_j^2 s_j^2}{\epsilon^2} \sum_{i=1}^{L} \frac{b_i^2}{\alpha_i^2 s_i^2} \\
&= \frac{B^2 \ln(2W^2) \prod_{j=i}^{L} \rho_j^2 s_j^2}{\epsilon^2} \left(\bar{\alpha}^3\right).
\end{aligned}$$

□

A.5 Proof of Theorem 1.1

The first step is to prove Lemma 3.5, which, in contrast to Theorem 1.1, has matrix and data constraints given before the data is seen.

Proof of Lemma 3.5. Consider the class of networks $\mathcal{F}_\lambda$ obtained by affixing the ramp loss ℓ_γ and the negated margin operator $-\mathcal{M}$ to the output of the provided network class:

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

Since $(z,y) \mapsto \ell_\gamma(-\mathcal{M}(z,y))$ is $2/\gamma$-Lipschitz wrt $\|\cdot\|_2$ by Lemma A.3 and definition of ℓ_γ, the function class $\mathcal{F}_\gamma$ still falls under the setting of Theorem 3.3, and gives

$$\ln \mathcal{N}\left((\mathcal{F}_\gamma)_{|S}, \epsilon, \|\cdot\|_2\right) \le \frac{4B^2 \ln(2W^2)}{\gamma^2 \epsilon^2} \left(\prod_{j=1}^{L} s_j^2 \rho_j^2\right) \left(\sum_{i=1}^{L} \left(\frac{b_i}{s_i}\right)^{2/3}\right)^3 =: \frac{R}{\epsilon^2}.$$

What remains is to relate covering numbers and Rademacher complexity via a Dudley entropy integral; note that most presentations of this technique place $1/n$ inside the covering number norm, and thus the application here is the result of a tiny amount of massaging. Continuing with this in mind, the Dudley entropy integral bound on Rademacher complexity grants

$$\mathfrak{R}((\mathcal{F}_\gamma)_{|S}) \le \inf_{\alpha>0} \left(4\alpha + \frac{12}{n} \int_\alpha^{\sqrt{n}} \sqrt{\frac{R}{\epsilon^2}}\, \mathrm{d}\epsilon\right) = \inf_{\alpha>0} \left(4\alpha + \ln(\sqrt{n}/\alpha) \frac{12\sqrt{R}}{n}\right).$$

The inf is uniquely minimized at $\alpha := 3\sqrt{R}/n$, but the desired bound may be obtained by the simple choice $\alpha := 1/n$, and plugging the resulting Rademacher complexity estimate into Lemma 3.1. □

The proof of Theorem 1.1 now follows by instantiating Lemma 3.5 for many choices of its various parameters, and applying a union bound. There are many ways to cut up this parameter space and organize the union bound; the following lemma makes one such choice, whereby Theorem 1.1 is easily proved. A slightly better bound is possible by invoking positive homogeneity of $(\sigma_1, \ldots, \sigma_L)$ to balance the spectral norms of the matrices $(A_1, \ldots, A_L)$, however these rebalanced matrices are then used in the comparison to $(M_1, \ldots, M_L)$, which is harder to interpret when $M_i \neq 0$.

Lemma A.6. *Suppose the setting and notation of Theorem 1.1. With probability at least $1-\delta$, every network $F_{\mathcal{A}} : \mathbb{R}^d \to \mathbb{R}^k$ with weight matrices $\mathcal{A} = (A_1, \ldots, A_L)$ and every $\gamma > 0$ satisfy*

$$\begin{aligned}
&\Pr\left[\arg\max_j F_{\mathcal{A}}(x)_j \neq y\right] \\
&\quad \leq \widehat{\mathcal{R}}_\gamma(F_{\mathcal{A}}) + \frac{8}{n} \\
&\quad + \frac{144\ln(n)\ln(2W)}{\gamma n}\left(\prod_i \rho_i\right)\left(1+\|X\|_2\right)\left(\sum_{i=1}^{L}\left(\left(\frac{1}{L}+\|A_i-M_i\|_1\right)\prod_{j\neq i}\left(\frac{1}{L}+\|A_j\|_\sigma\right)\right)^{2/3}\right)^{3/2} \\
&\quad + \sqrt{\frac{9}{2n}}\sqrt{\ln(1/\delta)+\ln(2n/\gamma)+2\ln(2+\|X\|_2)+2\sum_{i=1}^{L}\ln(2+L\|A_i-M_i\|_1)+2\sum_{i=1}^{L}\ln(2+L\|A_i\|_\sigma)}.
\end{aligned} \tag{A.3}$$

Proof. Given positive integers $(\vec{j}, \vec{k}, \vec{l}) = (j_1, j_2, j_3, k_1, \ldots, k_L, l_1, \ldots, l_L)$, define a set of instances (a set of triples $(\gamma, X, \mathcal{A})$)

$$\begin{aligned}
\mathcal{B}(\vec{j}, \vec{k}, \vec{l}) &= \mathcal{B}(j_1, j_2, j_3, k_1, \ldots, k_L, l_1, \ldots, l_L) \\
&:= \left\{(\gamma, X, \mathcal{A}) \;:\; 0 < \frac{1}{\gamma} < \frac{2^{j_1}}{n},\ \|X\|_2 < j_2,\ \|A_i - M_i\|_1 < \frac{k_i}{L},\ \|A_i\|_\sigma < \frac{l_i}{L}\right\}.
\end{aligned}$$

Correspondingly subdivide δ as

$$\begin{aligned}
\delta(\vec{j}, \vec{k}, \vec{l}) &= \delta(j_1, j_2, j_3, k_1, \ldots, k_L, l_1, \ldots, l_L) \\
&:= \frac{\delta}{2^{j_1}\cdot j_2(j_2+1)\cdot k_1(k_1+1)\cdots k_L(k_L+1)\cdot l_1(l_1+1)\cdots l_L(l_L+1)}.
\end{aligned}$$

Fix any $(\vec{j}, \vec{k}, \vec{l})$. By Lemma 3.5, with probability at least $1-\delta(\vec{j}, \vec{k}, \vec{l})$, every $(\gamma, X, \mathcal{A}) \in \mathcal{B}(\vec{j}, \vec{k}, \vec{l})$ satisfies

$$\begin{aligned}
\Pr\left[\arg\max_j F_{\mathcal{A}}(x)_i \neq y\right] \;\leq\; & \widehat{\mathcal{R}}_\gamma(f) + \frac{8}{n} \\
& + \underbrace{\frac{72\cdot 2^{j_1}\cdot j_2\ln(2W)\ln(n)}{n^2}\left(\prod_{i=1}^{L}\rho_i\right)\left(\sum_{i=1}^{L}\left(\frac{k_i}{L}\prod_{j\neq i}\frac{l_j}{L}\right)^{2/3}\right)^{3/2}}_{=:\heartsuit} \\
& + \underbrace{3\sqrt{\frac{\ln(1/\delta)+\ln(2^{j_1})+2\ln(1+j_2)+2\sum_{i=1}^{L}\ln(1+k_i)+2\sum_{i=1}^{L}\ln(1+l_i)}{2n}}}_{=:\clubsuit}.
\end{aligned} \tag{A.4}$$

Since $\sum_{\vec{j},\vec{k},\vec{l}}\delta(\vec{j}, \vec{k}, \vec{l}) = \delta$, by a union bound, the preceding bound holds simultaneously over all $\mathcal{B}(\vec{j}, \vec{k}, \vec{l})$ with probability at least $1-\delta$.

Thus, to finish the proof, discard the preceding failure event, and let an arbitrary $(\gamma, X, \mathcal{A})$ be given. Choose the smallest $(\vec{j}, \vec{k}, \vec{l})$ so that $(\gamma, X, \mathcal{A}) \in \mathcal{B}(\vec{j}, \vec{k}, \vec{l})$; by the preceding union bound, eq. (A.4) holds for this $(\vec{j}, \vec{k}, \vec{l})$. The remainder of the proof will massage eq. (A.4) into the form in the statement of Theorem 1.1.

As such, first consider the case $j_1 = 1$, meaning $\gamma < 2/n$; then

$$\Pr\left[\arg\max_j F_{\mathcal{A}}(x)_j \neq y\right] \leq 1 < \frac{1}{\gamma n},$$

where the last expression lower bounds the right hand side of eq. (A.3), thus completing the proof in the case $j_1 = 1$. Suppose henceforth that $j_1 \geq 2$ (and $\gamma \geq 2/n$).

Combining the preceding bound $j_2 \geq 2$ with the definition of $\mathcal{B}(\vec{j}, \vec{k}, \vec{l})$, the elements of $(\vec{j}, \vec{k}, \vec{l})$ satisfy

$$\begin{aligned} & 2^{j_1} \leq \frac{2n}{\gamma}, \\ & j_2 \leq 1 + \|X\|_2, \\ \forall i \,\textbf{.} \quad & k_i \leq 1 + L\|A_i - M_i\|_1, \\ \forall i \,\textbf{.} \quad & l_i \leq 1 + L\|A_i\|_\sigma. \end{aligned}$$

For the term $\heartsuit$, the factors with $(\vec{j}, \vec{k}, \vec{l})$ are bounded as

$$\begin{aligned} & 2^{j_1} \cdot j_2 \left(\sum_{i=1}^{L} \left(k_i \prod_{j \neq i} l_j \right)^{2/3} \right)^{3/2} \\ & \leq \frac{2n}{\gamma} \left(1 + \|X\|_2\right) \left(\sum_{i=1}^{L} \left((L^{-1} + \|A_i - M_i\|_1) \prod_{j \neq i} (L^{-1} + \|A_i\|_\sigma) \right)^{2/3} \right)^{3/2}. \end{aligned}$$

For the term $\clubsuit$, the factors with $(\vec{j}, \vec{k}, \vec{l})$ are bounded as

$$\begin{aligned} & \ln(2^{j_1}) + 2\ln(1 + j_2) + 2\sum_{i=1}^{L} \ln(1 + k_i) + 2\sum_{i=1}^{L} \ln(1 + l_i) \\ & \leq \ln(2n/\gamma) + 2\ln(2 + \|X\|_2) + 2\sum_{i=1}^{L} \ln(2 + L\|A_i - M_i\|_1) + 2\sum_{i=1}^{L} \ln(2 + L\|A_i\|_\sigma). \end{aligned}$$

Plugging these bounds on $\heartsuit$ and $\clubsuit$ into eq. (A.4) gives eq. (A.3). □

The proof of Theorem 1.1 is now a consequence of Lemma A.6, simplifying the bound with a $\widetilde{\mathcal{O}}(\cdot)$. Before proceeding, it is useful to pin down the asymptotic notation $\widetilde{\mathcal{O}}(\cdot)$, as it is not completely standard in the multivariate case. The notation can be understood via the lim sup view of $\mathcal{O}(\cdot)$; namely, $f = \widetilde{\mathcal{O}}(g)$ if there exists a constant C so that any sequence $((n^{(j)}, \gamma^{(j)}, X^{(j)}, A_1^{(j)}, \ldots, A_L^{(j)}))_{j=1}^{\infty}$ with $n^{(j)} \to \infty$, $\gamma^{(j)} \to \infty$, $\|X^{(j)}\|_2 \to \infty$, $\|A_i^{(j)}\|_1 \to \infty$ satisfies

$$\limsup_{j \to \infty} \frac{f(n^{(j)}, \gamma^{(j)}, X^{(j)}, A_1^{(j)}, \ldots, A_L^{(j)})}{g(n^{(j)}, \gamma^{(j)}, X^{(j)}, A_1^{(j)}, \ldots, A_L^{(j)}) \operatorname{poly}\log(g(n^{(j)}, \gamma^{(j)}, X^{(j)}, A_1^{(j)}, \ldots, A_L^{(j)}))} \leq C.$$

Proof of Theorem 1.1. Let $f = f_0 + f_1 + f_2$ denote the three excess risk terms of the upper bound from Lemma A.6, and $g = g_1 + g_2$ denote the two excess risk terms of the upper bound from Theorem 1.1; as discussed above, the goal is to show that there exists a universal constant C so that for any sequence of tuples $((n^{(j)}, \gamma^{(j)}, X^{(j)}, A_1^{(j)}, \ldots, A_L^{(j)}))_{j=1}^{\infty}$ increasing as above, $\limsup_{j \to \infty} f/(g \operatorname{poly}\log(g)) \leq C$.

It is immediate that $\limsup_{j \to \infty} f_0/g = 0$ and $\limsup_{j \to \infty} f_1/(g_1 \ln(g)) \leq 144$. The only trickiness arises when studying $f_2/(g_2 \ln(g))$, namely the term $\sum_i \ln(2 + L\|A_i - M_i\|_1)$, since g_2 instead has the term $\ln(\sum_i \|A_i - M_i\|_1^{2/3})$, and the ratio of these two can scale with L. A solution however is to compare to $\ln(\prod_i \|A_i\|_\sigma)$, noting that $\|A_i\|_1 \leq W\|A_i\|_2 \leq W^{3/2}\|A_i\|_\sigma$:

$$\limsup_{j \to \infty} \frac{\sum_i \ln(2 + L\|A_i^{(j)} - M_i\|_1)}{\ln(\prod_i \|A_i^{(j)}\|_\sigma)} \leq \limsup_{j \to \infty} \frac{\sum_i \ln(2 + L\|A_i^{(j)}\|_1 + L\|M_i\|_1)}{\sum_i \ln(\|A_i^{(j)}\|_1/W^{3/2})} = 1.$$

□

A.6 Proof of lower bound (Theorem 3.6)

Proof of Theorem 3.6. Define

$$\mathcal{F}(r) := \left\{ A_L\sigma_{L-1}(A_{L-1}\cdots\sigma_2(A_2\sigma_1(A_1x)) : \prod_{i=1}^{L}\|A_i\|_\sigma \le r \right\},$$

where each $\sigma_i = \sigma$ is the ReLU and each $A_k \in \mathbb{R}^{d_k\times d_{k-1}}$, with $d_0 = d$ and $d_L = 1$, and let $S := (x_1, \ldots, x_n)$ denote the sample.

Define a new class $\mathcal{G}(r) = \big\{x \mapsto \langle a, x\rangle \mid \|w\|_2 \le r\big\}$. It will be shown that $\mathcal{G}(r) \subseteq \mathcal{F}(C\cdot r)$ for some $C > 0$, whereby the result easily follows from a standard lower bound on $\mathfrak{R}(\mathcal{G}(r)_{|S})$.

Given any linear function $x \mapsto \langle a, x\rangle$ with $\|a\|_2 \le r$, construct a network $f = A_L\sigma_{L-1}(A_{L-1}\cdots\sigma_2(A_2\sigma_1(A_1x)))$ as follows:

- $A_1 = (\mathbf{e}_1 - \mathbf{e}_2)a^\top$.
- $A_k = \mathbf{e}_1\mathbf{e}_1^\top + \mathbf{e}_2\mathbf{e}_2^\top$ for each $k \in \{2, \ldots, L-1\}$.
- $A_L = \mathbf{e}_1 - \mathbf{e}_2$.

It is now shown that $f(x) = \langle a, x\rangle$ pointwise. First, observe $\sigma(A_1x) = (\sigma(\langle a, x\rangle), \sigma(-\langle a, x\rangle), 0, \ldots, 0)$. Since σ is positive homogeneous, $\sigma_{L-1}(A_{L_1}\cdots\sigma_2(A_2y) = A_{L-1}A_{L-2}\cdots A_2y = (y_1, y_2, 0, \ldots, 0)$ for any y in the non-negative orthant. Because $\sigma(A_1x)$ lies in the non-negative orthant, this means $\sigma_{L-1}(A_{L-1}\cdots\sigma_2(A_2\sigma_1(A_1x))) = (\sigma(\langle a, x\rangle), \sigma(-\langle a, x\rangle), 0, \ldots, 0)$. Finally, the choice of $A_L = \mathbf{e}_1 - \mathbf{e}_2$ gives $f(x) = \sigma(\langle a, x\rangle) - \sigma(-\langle a, x\rangle) = \langle a, x\rangle$.

Observe that for all $k \in \{2, \ldots, L-1\}$, $\|A_k\|_\sigma = 1$. For the other layers, $\|A_L\|_\sigma = \|A_L\|_2 = \sqrt{2}$ and $\|A_1\|_\sigma = \sqrt{2}\cdot r$, which implies $f \in \mathcal{F}(2r)$.

Combining the pieces,

$$\mathfrak{R}(\mathcal{F}(2r)_{|S}) \ge \mathfrak{R}(\mathcal{G}(r)_{|S}) = \mathbb{E}\sup_{a:\|a\|_2\le r}\sum_{t=1}^{n}\epsilon_t\langle a, x_t\rangle = r\cdot\mathbb{E}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_2.$$

Finally, by the Khintchine-Kahane inequality there exists $c > 0$ such that

$$\mathbb{E}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|_2 \ge c\cdot\sqrt{\sum_{t=1}^{n}\|x_t\|_2^2} = c\|X\|_2.$$

□