12-2019

Improved generalisation bounds for deep learning through L∞ covering numbers

Antoine LEDENT
Singapore Management University, aledent@smu.edu.sg

Yunwen LEI

Marius KLOFT

Improved Generalisation Bounds for Deep Learning Through L^∞ Covering Numbers

Anonymous Author(s)
Affiliation
Address
email

Abstract

By replacing L^2 covering number approaches in Rademacher analysis with an analysis based on L^∞ covering numbers, we show generalisation error bounds for deep learning with two main improvements over the state of the art. First, our bounds have no explicit dependence on the number of classes except for logarithmic factors. This holds even when formulating the bounds in terms of the L^2-norm of the weight matrices, while previous bounds exhibit at least a square-root dependence on the number of classes in this case. Second, we adapt the Rademacher analysis of DNNs to incorporate weight sharing—a task of fundamental theoretical importance which was previously attempted only under very restrictive assumptions. In our results, each convolutional filter contributes only once to the bound, regardless of how many times it is applied.

1 Introduction

The statistical theory of deep learning has enjoyed a revival since 2017 with the advent of learning guarantees for deep neural networks expressed in terms of various norms of the weight matrices and classification margins [1, 2, 3, 4]. Many improvements have surfaced to make bounds non-vacuous at realistic scales, including better depth dependence, bounds that apply to ResNets [5] and PAC-Bayesian bounds using network compression.

Yet, several questions of fundamental theoretical importance remain unsolved. (1) How can we account for weight sharing in convolutional neural networks (CNNs)? So far, the best bound [4] accounting for weight sharing is valid only if, in each layer, the convolutional filters are orthonormal. (2) How can we remove or decrease the dependence of bounds on the number of classes? This question is of central importance in extreme classification [6]. In [2], the authors show a bound that has no explicit class dependence (except for log terms). However, this bound is formulated in terms of the $L^{2,1}$ norms of the network's weight matrices. If we convert the occurring $L^{2,1}$ norms into L^2 norms, we obtain a square-root dependence on the number of classes.

In this paper, we provide, up to only logarithmic terms, a complete solution to both of the above questions. Our bound relies only on L^2 norms. Although, in the hidden layers, it scales as the square root of the maximum network width (as other L^2 bounds for DNNs), it has no explicit (non-logarithmic) dependence on the width of the output layer, that is, the number of classes. Furthermore, our bound accounts for weight sharing: the Frobenius norm of the weight matrix of each convolutional filter contributes only once to the bound, regardless of how many times it is applied, and regardless of any orthogonality conditions and how many filters a layer contains.

Submitted to NeurIPS 2019 Workshop on Machine Learning with Guarantees. Do not distribute.

2 Related Work

The now often cited paper [2] provides the following bound:

Theorem 2.1 (Bartlett et al., 2017)**.** *Assume that* $(x, y), (x_1, y_1), \dots, (x_n, y_n)$ *are drawn iid from any probability distribution over* $\mathbb{R}^d \times \{1, 2, \dots, K\}$*. Denote by* $F_{\mathcal{A}}$ *the function represented by the network with weights* $\mathcal{A} = \{A^1, A^2, \dots, A^L\}$ *and involving the nonlinearities* $\sigma_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ *(where* $d_0 = d$ *is the input dimension and* $d_L = K$ *is the number of classes) so that* $F_{\mathcal{A}}(x) = \sigma_L\left(A^L \sigma_{L-1}\left(A^{L-1} \dots \sigma_1\left(A^1 x\right)\right)\right)$.

The final layer of the network is translated into a class prediction by taking the argmax over components, with an arbitrary rule for breaking ties. For any classifier $f : \mathbb{R}^d \to \mathbb{R}^h$ *and any real number* $\gamma > 0$*, write also*

$$\widehat{R}_\gamma(f) = \frac{\sum_{i=1}^n 1\left[f(x_i)_{y_i} \le \gamma + \max_{j \ne y_i} f(x_i)_j\right]}{n},$$

$\|X\|_{\mathrm{Fr}}$ *for the Frobenius norm of the data matrix* $X \in \mathbb{R}^{n \times d}$*, as well as* $\|X\|_{2,2}^2$ *for the quantity* $\frac{1}{n}\sum_{i=1}^n (\sum_{j=1}^d X_{ij}^2) = \frac{\|X\|_{\mathrm{Fr}}^2}{n}$.

For $(x, y), (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ *drawn iid from any probability distribution over* $\mathbb{R}^d \times \{1, 2, \dots, K\}$*, with probability at least* $1 - \delta$*, every network* $F_{\mathcal{A}}$ *with weight matrices* $\mathcal{A}$ *and every margin* $\gamma > 0$ *satisfy:*

$$\mathbb{P}(\arg\max_j(F_{\mathcal{A}}(x)_j) \ne y) \le \widehat{R}_\gamma(F_{\mathcal{A}}) + \widetilde{\mathcal{O}}\left(\frac{\|X\|_{2,2} M_{\mathcal{A}}}{\gamma\sqrt{n}} \log(\bar{W}) + \sqrt{\frac{\log(1/\delta)}{n}}\right), \tag{1}$$

where $\bar{W} = \max_{i=1}^L d_i$ *is the maximum width of the network, and*

$$M_{\mathcal{A}} = \left(\prod_{i=1}^L \rho_i \|A^i\|_\sigma\right)\left(\sum_{i=1}^L \frac{\|(A^i)^\top\|_{2,1}^{\frac{2}{3}}}{\|A^i\|_\sigma^{\frac{2}{3}}}\right)^{\frac{3}{2}}. \tag{2}$$

Here $\|\cdot\|_\sigma$ *denotes the spectral norm, and for any matrix* $A \in \mathbb{R}^{a \times b}$*,* $\|A\|_{2,1} = \sum_{j=1}^b \sqrt{\sum_{i=1}^a A_{i,j}^2}$*.*

Around the same time as the above result appeared, the authors in [1] used a PAC Bayesian approach to prove an analogous result with $M_{\mathcal{A}}$ replaced by the quantity below[1]:

$$M_{\mathcal{A},2} := L\sqrt{\bar{W}}\left(\prod_{i=1}^L \rho_i \|A^i\|_\sigma\right)\left(\sum_{i=1}^L \frac{\|A^i\|_2^2}{\|A^i\|_\sigma^2}\right)^{\frac{1}{2}}. \tag{3}$$

The above bounds are fully post hoc, scale-sensitive and have the further satisfying property of taking the classification margins into account. However, they apply generally to fully connected networks and take very little architectural information into account. In particular, if the above bounds are applied to a convolutional neural network, when calculating the squared Frobenius norms $\|A^i\|_2^2$, the matrix A^i is the matrix representing the linear operation performed by the convolution, which implies that the weights of each filter will be summed as many times as it is applied. This effectively adds a dependence on the square root of the size of the corresponding activation map at each term of the sum. Furthermore, the L^2 version of the above includes a dependence on the square root of the number of classes through the maximum width W of the network.

In late 2017 and 2018, there was a spur of research effort on the question of fine-tuning the analyses that provided the above bounds, with improved dependence on depth [7], and some bounds for recurrent neural networks [8, 3]. Notably, in [4], the authors provided an analogue of Theorem 2.1 for convolutional networks, but only under some very specific assumptions.

Since then, other lines of research (especially the PAC Bayesian school building on [1]) have focused on obtaining more meaningful bounds at realistic scales using various techniques including model

[1] Note that the result using formula 3 can also be derived from expressing 1 in terms of L^2 norms and using Jensen's inequality

compression, as well as understanding any implicit restriction on the function class imposed by the optimisation procedure [9, 10, 11, 12, 13].

Still, the fundamental question of taking weight sharing into account in the Rademacher analysis of DNNs was left unsolved until the first version of our work, and an independent solution [14] simultaneously appeared on arXiv. In this note, we present our solution to the weight sharing problem. Furthermore, we present our solution to the multiclass problem in the L^2 theory, which corresponds to a improvement of a factor of $\sqrt{C}$ compared to the state of the art.

3 Informal Outline of Contributions

In this section, we state our main results which can be considered as specific examples of our general results in Section A.

Theorem 3.1 (Multi-class, fully connected)**.** *Assume that $(x, y), (x_1, y_1), \ldots, (x_n, y_n)$ are drawn iid from any probability distribution over $\mathbb{R}^d \times \{1, 2, \ldots, K\}$, and let us use the notation of [2]. Write $W_1, W_2, \ldots, W_L$ for the width of each layer. With probability at least $1 - \delta$, every network $F_\mathcal{A}$ with weight matrices $\mathcal{A}$ and every margin $\gamma > 0$ satisfy:*

$$\mathbb{P}(\arg\max_j (F_\mathcal{A}(x)_j) \neq y) \leq \widehat{R}_\gamma(F_\mathcal{A}) + \widetilde{\mathcal{O}}\left(\frac{\max_{i=1}^n \|x_i\|_2 R_\mathcal{A}}{\gamma\sqrt{n}} \log(W) + \sqrt{\frac{\log(1/\delta)}{n}}\right), \quad (4)$$

where $W = \bar{W} = \max_{i=1}^L W_i$ is the maximum width of the network, and

$$R_\mathcal{A} := L\rho_L \max_i \|A^L_{i,\bullet}\|_2 \left(\prod_{i=1}^{L-1} \rho_i \|A^i\|_\sigma\right) \left(\sum_{i=1}^{L-1} \frac{(\sqrt{W_i}\|A^i\|_2)^2}{\|A^i\|_\sigma^2} + \frac{\|A^L\|_2^2}{\max_i \|A^L_{i,\bullet}\|_2^2}\right)^{\frac{1}{2}},$$

and $\widehat{R}_\gamma(F_\mathcal{A})$ is defined as in Theorem 2.1.

Proof. The result follows directly from Theorem A.1, which is presented in Section A. □

Note that the last term of the sum does not explicitly contain architectural information, and the bound only depends on W_i for $i \leq L - 1$, but not on W_L (the number of classes). This means the above is a class-size free generalisation bound (up to a logarithmic factor) with L^2 norms of the last layer weight matrix. This improves on the earlier $L^{2,1}$ norm result in [2]. To see this, let us consider a standard situation where the rows of the matrix A^L have approximately the same L^2 norm, i.e., $\|A^L_{i,\bullet}\|_2 \asymp a$. In this case, our bound involves $\|A^L\|_{\mathrm{Fr}} \asymp \sqrt{W_L} a$, which incurs a square-root dependency on the number of classes. As a comparison, the bound in [2] involves $\|(A^L)^\top\|_{2,1} \asymp W_L a$, which incurs a linear dependency on the number of classes. If we further impose an L_2-constraint on the last layer as $\|A^L\|_{\mathrm{Fr}} \leq a$ as in the SVM case for a constant a [15], then our bound would enjoy a logarithmic dependency while the bound in [2] enjoys a square-root dependency.

Suppose now we have a *convolutional architecture* where we collect the weights in matrices A^1, A^2,…, and A^L, with $A^l \in \mathbb{R}^{m_l \times d_l}$ (here m_l is the number of filters at layer l, and d_l is the size of the filters in that layer) each row being a filter (represented only once), so that the i^{th} row of A^l represents the i^{th} convolutional filter of layer l. For $l \leq L$ and a weight matrix A^l, we will also write $\tilde{A}^l$ for the matrix representing the linear operation that consists in applying each of the filters over each of the patches of the previous layer [2]. Thus the full network can be represented in matrix form as $F_\mathcal{A}(x) = \sigma_L\big(\tilde{A}^L \sigma_{L-1}\big(\tilde{A}^{L-1} \ldots \sigma_1\big(\tilde{A}^1 x\big)\big)\big)$. We have the following result, which follows directly from our general Theorem A.1 below.

Theorem 3.2. *With probability at least $1 - \delta$ over the draw of the training data, every network $F_\mathcal{A}$ with weight matrices $\mathcal{A} = \{A^1, A^2, \ldots, A^L\}$ and every margin $\gamma > 0$ satisfy:*

$$\mathbb{P}\left(\arg\max_j (F_\mathcal{A}(x)_j) \neq y\right) \leq \widehat{R}_\gamma(F_\mathcal{A}) + \widetilde{\mathcal{O}}\left(\frac{\max_{i=1}^n \|x_i\|_2 R_\mathcal{A}}{\gamma\sqrt{n}} \log(W) + \sqrt{\frac{\log(1/\delta)}{n}}\right), \quad (5)$$

[2] The dimensions of this matrix depend on the stride and on the size of the previous layer

where W is the maximum number of neurons in a single layer (after pooling) and

$$R_{\mathcal{A}} := L\left(\rho_L \max_i \|A^L_{i,\boldsymbol{\cdot}}\|_2 \prod_{l=1}^{L-1} \rho_l \|\tilde{A}^l\|_\sigma\right)\left(\sum_{l=1}^{L-1} \frac{(\sqrt{W_l}\|A^l\|_2)^2}{\|\tilde{A}^l\|_\sigma^2} + \frac{\|A^L\|_2^2}{\max_i \|A^L_{i,\boldsymbol{\cdot}}\|_2^2}\right)^{\frac{1}{2}},$$

$A^L_{i,\boldsymbol{\cdot}}$ denotes the i'th row of A^L, and for all $\|\cdot\|_\sigma$ and $\|\cdot\|_2$ denote the standard spectral and Frobenius norms respectively.

While we still have to use the spectral norm of the complete convolution operation represented by $\tilde{A}^l$ in the first factor, the Frobenius norm involved is that of the matrix A^l (the filter) instead of $\tilde{A}^l$ (the matrix representing the full convolutional operation), which means we are only summing the square norms of each filter once, regardless of how many time it is used. As a comparison, applying the result in [2] to CNN's yields a bound involving the whole matrix $\widetilde{A}$ ignoring the structure of CNNs. This means that through exploiting weight sharing, we remove a factor of $\sqrt{O_{l-1}}$ in the l^{th} term of the sum compared to a standard the result in [2], where O_l denotes the number of convolutional patches in layer l. We have also replaced the width dependence by a dependence on the width after pooling by exploiting the L^∞-continuity of the pooling operation.

Remark: Note that while for simplicity we presented our results with the Frobenius norms of the filter matrices A^l in the numerators of $r_{\mathcal{A}}$, our proof also allows us to replace these quantities by $\|A^l - M^l\|_{\mathrm{Fr}}$, for some arbitrary matrices M^l chosen in advance (typically the initialised weights).

4 Main ideas of proof

Obtaining PAC guarantees go through bounding the covering numbers of the function class considered. In the case of neural networks, the first step is then to provide a bound on the covering numbers of individual layers. If we apply classical results on linear classifiers as is done in [2] (where results on L^2 covering numbers are used) by viewing a convolutional layer as a linear map, we cannot take advantage of weight sharing. In this work, we circumvent this difficulty by applying results on the L^∞ covering numbers of classes of linear classifiers to a different problem where each "(convolutional patch, sample, output channel)" combination is viewed as a single data point. More precisely, we will make use of the following proposition from [16] (Theorem 4, page 537).

Proposition 4.1. *Let $n, d \in \mathbb{N}$, $a, b > 0$. Suppose we are given n data points collected as the rows of a matrix $X \in \mathbb{R}^{n\times d}$, with $\|X_{i,\boldsymbol{\cdot}}\|_2 \le b, \forall i = 1, \dots, n$. For $U_{a,b}(X) = \{X\alpha : \|\alpha\|_2 \le a, \alpha \in \mathbb{R}^d\}$, we have*

$$\log \mathcal{N}\left(U_{a,b}(X), \epsilon, \|\cdot\|_\infty\right) \le \frac{36a^2b^2}{\epsilon^2}\log_2\left(\frac{8abn}{\epsilon} + 6n + 1\right).$$

With convolutional layers in mind, we now consider the problem of bounding the L^∞ covering number of $\{(v_i^\top X^j)_{i\le I, j\le J} : \sum_{i\le I}\|v_i\|_2^2 \le a^2\}$ (where $X^j \in \mathbb{R}^{d\times n}$ for all j) with only logarithmic dependence on n, I, J. Here, I plays the role of the number of output channels, while J plays the role of the number of convolutional patches. To do so, we apply the above result 4.1 on the $nIJ \times dI$ matrix constructed as follows:

$$\begin{pmatrix} X^1 & 0 & \dots & 0 & X^2 & 0 & \dots & 0 & \dots & X^J & \dots & \dots & 0 \\ 0 & X^1 & \dots & 0 & 0 & X^2 & 0 & \dots & \dots & 0 & X^J & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & X^1 & 0 & 0 & \dots & X^2 & \dots & 0 & \dots & \dots & X^J \end{pmatrix},$$

with the corresponding vectors being constructed as $(v_1, v_2, \dots, v_I) \in \mathbb{R}^{dI}$.

If we compose the linear map on $\mathbb{R}^{n\times d}$ represented by $(v_1, v_2, \dots, v_I)^\top$ with k real-valued functions with L^∞ Lipschitz constant 1, the above argument yields comparable bounds on the $\|\cdot\|_{\infty,2}$ covering number of the composition, loosing a factor of $\sqrt{k}$ only (for the last layer, $k = 1$, and for convolutional layers, k is the number of neurons in the layer left after pooling).

This solves the problem for a single layer network. Once this is taken care of, the rest of the proof consists in adaptation of classic chaining arguments and a union bound on probabilities of events [2, 4, 5].

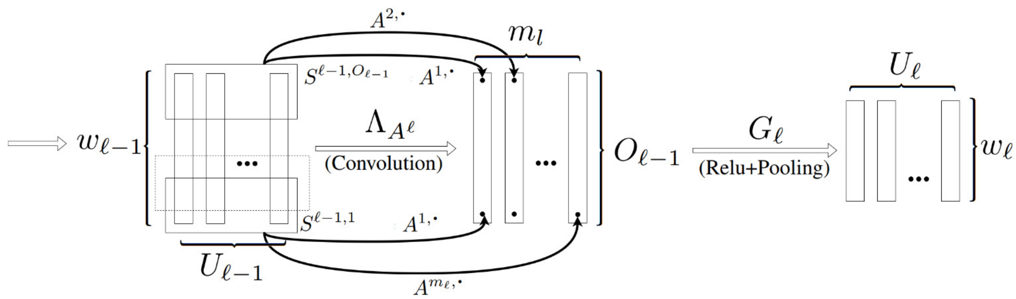

Figure 1 – Illustration of architecture for one layer

A Precise Notation and Results

A.1 Notation

We use the following notation to represent linear layers with weight sharing such as convolution. Let $x \in \mathbb{R}^{U\times w}$, $A \in \mathbb{R}^{m\times d}$ and $S^1, S^2, \ldots, S^O$ be O ordered subsets of $(\{1,2,\ldots,w\}\times\{1,2,\ldots,U\})$ each of cardinality d[3], where we will denote by S^o_i the i^{th} element of S^o. We will denote by $\Lambda_A(x)$ the element of $\mathbb{R}^{m\times O}$ such that $\Lambda_A(x)_{j,o} = \sum_{i=1}^d X_{S^o_i} A_{j,i}$. In a typical example the sets $S^1, S^2, \ldots, S^O$ represent the image patches where the convolutional filters are applied, and Λ would be represented via the "tf.nn.conv2d" function in Tensorflow. See We will also write $\tilde{A}^l$ for the matrix in $\mathbb{R}^{(U_{l-1}w_{l-1})\times(O_{l-1}m_l)}$ that represents the convolution operation Λ_{A^l}.

To represent a full network, we suppose that we are given a number $L \in \mathbb{N}$ of layers, $7L+2$ numbers $m_1, m_2, \ldots, m_L, d_1, d_2, \ldots, d_L, \rho_1, \rho_2, \ldots, \rho_L, w_0, w_1, \ldots, w_L, U_0, U_1, \ldots, U_L, O_1, O_2, \ldots, O_L,$ and $k_1, k_2, \ldots, k_L$, as well as $\sum_{l=0}^L O_l$ ordered sets $S^{l,o} \subset \{1,2,\ldots,U_l\}\times\{1,2,\ldots,w_l\}$ (for $l \le L$, $o \le O_l$), and $L-1$ functions $G_l : \mathbb{R}^{m_l\times O_{l-1}} \to \mathbb{R}^{U_l\times w_l}$ (for $l = 1,2,\ldots,L$) satisfying the following conditions.

1. For all $l \in \{1,2,\ldots,L-1\}$, G_l is ρ_l Lipschitz (component-wise) with respect to the L_∞ norm.
2. For all $l \in \{1,2,\ldots,L-1\}$, and for each $o \le O_l$, $S^{l,o}$ has cardinality d_l.

The architecture above can help us represent a feedforward neural network involving possible (intra-layer) weight sharing as

$$F_{A^1,A^2,\ldots,A^L} : \mathbb{R}^{U_0\times w_0} \to \mathbb{R}^{U_L\times w_l} : x \mapsto (G_L \circ \Lambda_{A^L} \circ G_{L-1} \circ \Lambda_{A^{L-1}} \circ \ldots G_1 \circ \Lambda_{A^1})(x),$$

where for each $l \le L$, the weight A^l is a matrix in $\mathbb{R}^{m_l\times d_l}$. Note that as usual, offset terms can be accounted for by adding a dummy dimension of constants at each layer (this dimension must belong to $S^{l,o}$ for each o).

To aid understanding, we provide a quick table of notations in Figure 1.

Throughout the text, we also fix some norms $|\cdot|_{\mathcal{L}_0}$,$|\cdot|_{\mathcal{L}_1}$,..., and $|\cdot|_{\mathcal{L}_L}$ on the spaces $\mathbb{R}^{U_0\times w_0}$, $\mathbb{R}^{U_1\times w_1}$, ..., and $\mathbb{R}^{U_L\times w_L}$, some functions $|\cdot|_{\mathcal{L}^*_l}$ on $\mathbb{R}^{m_l\times d_l}$ for $1 \le l \le L$, and some numbers $k_1, k_2, \ldots, k_L \in \mathbb{N}$ such that the following three properties are satisfied:

1. For all $l \le L$ and all $\xi \in \mathbb{R}^{U_l\times w_l}$, if $|\xi|_{\mathcal{L}_l} \le 1$, then $\forall o \le O_l, \quad \sum_{\delta\in S^{l,o}} (\xi_\delta)^2 \le 1$.
2. For all $l \in \{1,\ldots,L\}$, all $a > 0$ and all $\xi_1, \xi_2 \in \mathbb{R}^{U_{l-1}\times w_{l-1}}$, if $|\xi_1 - \xi_2|_{\mathcal{L}_{l-1}} \le a$, then
$$|(G_l \circ \Lambda_{A^l})(\xi_1) - (G_l \circ \Lambda_{A^l})(\xi_2)|_{\mathcal{L}_{l-1}} \le a|A^l|_{\mathcal{L}^*_l}.$$
3. For any $\xi \in \mathbb{R}^{U_l\times w_l}$, $|\xi|^2_{\mathcal{L}_l} \le k_l\|\xi\|^2_\infty$

[3]We suppose for notational simplicity that all convolutional filters at a given layer are of the same size. It is clear that the proof applies to the general case as well.

Notation	Meaning
G_l	Activation functions + pooling at layer l
A^l	Filter matrix at layer l
Λ_{A^l}	Convolution operation relative to filter matrix A^l
$\tilde{A}^l$	Matrix representing Λ_{A^l} (Has repeated weights in conv. net)
O_l	Number of convolutional patches at layer l
m_l	# of channels at layer l before nonlinearity (=# of output channels at layer $l-1$)
$S^{l,o}$	o^{th} convolutional patch at layer l
w_l	Number of spatial dimensions at layer l
U_l	Number of channels after nonlinearity
ρ_l	Lipschitz constant of G_l
$W_l = U_l w_l$	Width (after pooling) at layer l
$W = \max_l W_l$	Maximum network width (after any pooling)
$\bar{W} = \max_l O_{l-1} m_l$	Maximum network width (before any pooling)
$\mathcal{W}$	Total number of parameters
d_l	Size of convolutional patches corresponding to the operation Λ_{A^l}
k_l	Smallest integer such that $\|\cdot\|_{\mathcal{L}_l} \leq \sqrt{k_l}\|\cdot\|_\infty$, $k_L = 1$, $k_l = W_l$ if $\lvert\cdot\rvert_{\mathcal{L}_l} = \|\cdot\|_2$ and $k_l = d_l$ if $\|x\|^2_{\mathcal{L}_l} = \max_{o \leq O^l} \sum_{\delta \in S^{l,o}} (x_\delta)^2$
$K = W_L$	Number of classes

Table 1 – Table of notations for quick reference

4. For all l, there exist real numbers $\mathcal{D}_l$ and $\mathcal{E}_l$ such $\forall A \in \mathbb{R}^{m_l \times d_l}$,

$$\frac{\|A\|^2_{\mathcal{L}_l^*}}{\mathcal{D}_l} \leq \|A\|_2^2 \leq \mathcal{E}_l \|A\|^2_{\mathcal{L}_l^*}.$$

The two main examples of suitable such norms are the following.

The standard L^2 and spectral norms. We can set $|A|_{\mathcal{L}_l} = |A|_{\mathrm{Fr}}$ for all l, $|A|_{\mathcal{L}_l^*} = \rho_l |\tilde{A}|_\sigma$ for all $l \leq L-1$ and $|A|_{\mathcal{L}_L^*} = \rho_L \max_i \|A_{i,\cdot}\|_2$, where $|\cdot|_\sigma$ denotes the usual spectral norm for matrices, and $\tilde{A}$ is the circulant matrix that represents the convolution operation performed by Λ_A. This choice satisfies the conditions on the norms $|\cdot|_{\mathcal{L}_0}, \ldots, |\cdot|_{\mathcal{L}_L}$ with $\mathcal{D}_l = w_l$ and $\mathcal{E}_l = m_l$, and $k_l = W_l$.

Through Lipschitz constants. First, for all $l \leq L$ and all $x \in \mathbb{R}^{U_l \times w_l}$, define $\|x\|^2_{\mathcal{L}_l} = \max_{o \leq O^l} \sum_{\delta \in S^{l,o}} (x_\delta)^2$. For each $A^l \in \mathbb{R}^{m_l \times d_l}$, we can then simply define $\|A^l\|_{\mathcal{L}_l^*}$ as the Lipschitz constant of $G \circ \Lambda_A : \mathbb{R}^{U_{l-1} \times w_{l-1}} \to \mathbb{R}^{U_l \times w_l}$ with respect to the distances induced by the norms $\|\cdot\|_{\mathcal{L}_{l-1}}$ and $\|\cdot\|_{\mathcal{L}_l}$. This satisfies the above conditions with k_l being the maximum number of active neurons in a single convolutional patch of layer l.

Mix of the above To obtain the results 3.1 and 3.2 with the dividend $\max_i \|A^L_{i,\cdot}\|_2^2$ in the last term of the sum, we use the spectral norms up to layer $L-1$ and the Lipschitz one for the last layer.

A.2 General Results

We can now formulate our main Theorems. We always assume that we are given a classification problem with i.i.d. data-points $(x, y), (x_1, y_1), \ldots, (x_n, y_n)$ with $y, y_1, \ldots, y_n \in \{1, 2, \ldots, K\}$.

Theorem A.1 (Post-hoc asymptotic result)**.** *Assume we are given an architecture and classification problem as described in section A. For all $\delta > 0$, with probability $> 1 - \delta$ over the draw of the training set it holds that every network as described in section A, and every margins $\gamma > 0$ satisfy:*

$$\mathbb{P}\left(\arg\max_j (F_{\mathcal{A}}(x)_j) \neq y\right) \leq \widehat{R}_\gamma(F_{\mathcal{A}}) + \widetilde{\mathcal{O}}\left(\frac{\|X\|_{(\mathcal{L}_0,\infty)^\top} R_{\mathcal{A}}}{\gamma\sqrt{n}} \log(\bar{W}) + \sqrt{\frac{\log(1/\delta)}{n}}\right), \quad (6)$$

where $\|X\|_{(\mathcal{L}_0,\infty)^\top} := \max_{i \leq n} |x_i|_{\mathcal{L}_0}$, $\bar{W} = \max_{l=0}^L O_{l-1} m_l$, and

$$R_{\mathcal{A}}^2 = L^2 \sum_{l=1}^L k_l \rho_l^2 \|A^l\|_2^2 \prod_{i \neq l} \|A^i\|^2_{\mathcal{L}_i^*}.$$

The more precise non-asymptotic result from which Theorem A.1 can be deduced is the following.

Theorem A.2 (Post-hoc result)**.** *Assume we are given an architecture and classification problem as described in Section A. For all $\delta > 0$, with probability $> 1-\delta$ over the draw of the training set it holds that every network as described in section A, and every margins $\gamma > 0$ satisfy:*

$$\begin{aligned}
&\mathbb{P}_{(x,y)}\left(\arg\max_j(F_{\mathcal{A}}(x)_j) \neq y\right) \\
&\leq \widehat{\mathcal{R}}_n + \frac{8}{n} + \frac{576(\|X\|_{(\mathcal{L}_0,\infty)^\top}+1)}{\gamma\sqrt{n}}\sqrt{\bar{R}}\left[\log_2(32n^2\bar{\Gamma}/\gamma + 7\bar{W}n)\right]^{\frac{1}{2}}\log(n) \\
&+ 3\sqrt{\frac{\log\left(\frac{4n}{\delta\gamma}\right)}{2n} + \frac{1}{n}\log(2+\|X\|_{(\mathcal{L}_0,\infty)^\top})} + 3\sqrt{\frac{1}{n}\left(\sum_{l=1}^{L}\log\left[(2+L\|A^l\|_2)(2+L\|\tilde{A}^l\|_\sigma)\right]\right)},
\end{aligned} \tag{7}$$

where

$$\widehat{\mathcal{R}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[(F_L(x_i))_{y_i} - \max_{j\leq K, j\neq y_i}(F_L(x_i))_j \leq \gamma\right],$$

$$\bar{R} = L^2\sum_{l=1}^{L}k_l\rho_l^2\left(\frac{1}{L}+\|A^l\|_2\right)^2\prod_{i\neq l}\left(\frac{1}{L}+\|\tilde{A}^i\|_\sigma\right)^2,$$

and

$$\bar{\Gamma} = \max_{l=0}^{L}\left[\left(\|X\|_{(\mathcal{L}_0,\infty)^\top}+1\right)e\left(\|A^l\|_2+\frac{1}{L}\right)O_{l-1}m_l\prod_{i=1}^{l-1}\left(\frac{1}{L}+\|\tilde{A}^i\|_\sigma\right)\right].$$

B Proofs

Let us first make the following important points about one of our notational choices.

Important remarks :

1. Throughout the proofs, we will be using mixed $L^{p,q,r}$ norms. Importantly, any sample/batch dimension will always be averaged instead of summed! This convention helps reduce the number of unnecessary factors of n to drag along. Thus if $X \in \mathbb{R}^n$, n is the sample dimension and $p \geq 1$

$$\|X\|_p := \left(\frac{1}{n}\sum_{i=1}^{n}|X_i|^p\right)^{\frac{1}{p}}.$$

   Similarly, if $X \in \mathbb{R}^{I\times n\times J}$, n is the sample dimension and $1 \leq p,q,r \leq \infty$

$$\|X\|_{p,q,r}^r = \sum_{k=1}^{J}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{I}|X_{i,j,k}|^p\right)^{\frac{q}{p}}\right)^{\frac{r}{q}} \tag{8}$$

   This notation involving mixed norms will also (in fact, mostly) be used when some or all of p,q,r are infinite, in which case the factor of $1/n$ is irrelevant. For instance, if $X \in \mathbb{R}^{I\times n\times J}$ and n is the sample dimension, we will write

$$\|X\|_{(2,\infty,\infty)} = \max_{j_2\leq n}\max_{j_3\leq J}\sqrt{\sum_{j_1=1}^{I}(X_{j_1.j_2,j_3})^2}.$$

2. We interpret 'tensor multiplication' for tensors as contracting the last slice of the first tensor with the first slice of the second one, when the dimensions match. For instance, if $A \in \mathbb{R}^{a\times b\times c}$ and $B \in \mathbb{R}^{c\times d}$, $AB \in \mathbb{R}^{a\times b\times d}$ is defined by $(AB)_{i,j,k} = \sum_{l=1}^{c}A_{i,j,l}B_{lk}$.

3. The transpose of a tensor is defined by completely swapping the order of the dimensions, and we sometimes put the transpose in the index when referring to norms. Thus if $Y \in \mathbb{R}^{J\times n\times I}$,

$$\|Y\|_{(\infty,\infty,2)^\top} = \|Y^\top\|_{(\infty,\infty,2)} = \max_{j_2\le n}\max_{j_3\le J}\sqrt{\sum_{j_1=1}^{I}(Y_{j_1.j_2,j_3})^2}.$$

B.1 Size-independent covering number bounds for a single convolutional layer

A key aspect of the proof is that we can use proposition 4.1 to obtain an L^∞-covering of the map represented by a convolutional layer. Indeed, by viewing each (sample, convolutional patch, output channel) trio as an individual data point, we can, for each ϵ, find $\mathcal{N}_\epsilon$ filters $f_1,\ldots,f_{\mathcal{N}_\epsilon}$ with $\|f_i\|_{Fr} \le a \quad \forall i$ such for any convolutional map represented by the filter f (with $\|f\|_{Fr}\le a$), there exists a $u_f \in \{1,2,\ldots,\mathcal{N}_\epsilon\}$ such that for any input x_i, any convolutional patch S, and any output channel j, the outputs of f and f_{u_f} corresponding to this (input, patch, channel) combination differ by less than ϵ.

More precisely, we have the following result:

Proposition B.1. *Let positive reals (a,b,ϵ) and positive integer m be given. Let the tensor $X \in \mathbb{R}^{n\times U\times d}$ be given with $\forall i\in\{1,2,\ldots,n\}, \forall u \in \{1,2,\ldots,U\}, \quad \|X_{i,u,\boldsymbol{\cdot}}\|_2 \le b$. We have*

$$\log\mathcal{N}\left(\{XA : A\in\mathbb{R}^{d\times m}, \|A\|_{\mathrm{Fr}}\le a\},\epsilon,\|\boldsymbol{\cdot}\|_{\infty,\infty,\infty}\right) \le \frac{36a^2b^2}{\epsilon^2}\log_2\left[\left(\frac{8ab}{\epsilon}+6\right)nmU+1\right], \tag{9}$$

where the norm $\|\boldsymbol{\cdot}\|_{\infty,\infty,\infty}$ is over the space $\mathbb{R}^{n\times U\times m}$ and XA is defined by $(XA)_{u,i,j} = \sum_{o=1}^{d} X_{u,i,o}A_{o,j}$.

Proof. This follows immediately from Lemma 4.1 applied to the following nmU modified data points in $\mathbb{R}^{d\times m}$ (considered as a simple vector space with the inner product being applied after broadcasting) and function class: for all $\delta \in \{1,2,\ldots,d\}\times\{1,2,\ldots,m\}$, for all $i\le n$, $u \le U$ and $j\le m$, $(x_{i,u,j})_\delta = X_{i,u,\delta_1}$ for $\delta_2 = j$ and $(x_{i,u,j})_\delta = 0$ otherwise. I.e., for all (sample, patch, output channel) combination (i,u,j) (with $i\le n, u\le U, j\le m$), the corresponding data point is a matrix in $\mathbb{R}^{d\times m}$ whose j^{th} column is the corresponding convolutional patch in X, and the the other columns are 0.

The function class is defined by

$$\{F_A : \mathbb{R}^{d\times m}\to\mathbb{R} : x\mapsto\langle x, A\rangle; \|A\|_2\le a\},$$

where $\langle\boldsymbol{\cdot},\boldsymbol{\cdot}\rangle$ denotes the inner product after broadcasting: for $A,B\in\mathbb{R}^{n_1\times n_2}$, $\langle A,B\rangle := \mathrm{Tr}(AB^\top)$. □

Definition B.2. *Let $\rho > 0$, and $\tilde{G} : \mathbb{R}^m\to\mathbb{R}^m$ be such that for all $i\in\{1,2,\ldots,m\}$, $\tilde{G}_i$ is ρ Lipschitz with respect to the L^∞ norm. Next, define G as a truncation of $\tilde{G}$ where only the top k values are retained, with an arbitrary tie-breaking strategy, so that*

$$\begin{aligned}&\forall i\in\{1,2,\ldots,m\},\\ &G_i = \tilde{G}_i \quad \textit{if} \quad \#\left(\left\{j\in\{1,2,\ldots,m\} : \tilde{G}_j > \tilde{G}_i \vee (\tilde{G}_j = \tilde{G}_i \wedge j > i)\right\}\right) < k\\ &G_i = 0 \quad \textit{otherwise}.\end{aligned} \tag{10}$$

We will call any function G that can be represented in this way a k-sparse ρ-Lipschitz function (with respect to the L^∞ norm).

Next, we have the following key steps in our analysis.

Corollary B.3. *Let n, O, m be natural numbers, $\mathcal{Y}$ be a finite dimensional vector space endowed with the norm $|\boldsymbol{\cdot}|_{\mathcal{L}}$ and let $G:\mathbb{R}^{O\times m}\to\mathcal{Y}$ be ρ-Lipschitz with respect to the L_∞ norm. Assume also* *that there exists a number $k > 0$ such for any $y \in \mathcal{Y}$, $|y|_{\mathcal{L}} \leq k\|y\|_\infty$. For any $X \in \mathbb{R}^{n\times O\times d}$ such that $\|X^{i,o,\cdot}\|_2^2 \leq b^2$ ($\forall i, o$), we have*

$$\log \mathcal{N}\left(\left\{G(XA) : A \in \mathbb{R}^{d\times m}, \|A\|_2 \leq a\right\}, \epsilon, \|\cdot\|_{(|\cdot|_{\mathcal{L}},\infty)^\top}\right) \leq \frac{36ka^2b^2}{\epsilon^2\rho^2}\log_2\left[\left(\frac{8ab}{\epsilon\rho\sqrt{k}}+7\right)mnO\right] \tag{11}$$

where for a tensor $B \in \mathbb{R}^{n\times H}$,

$$\|B\|_{(2,\infty)^\top} = \|B^\top\|_{(2,\infty)} = \max_{i=1}^{n} |B_{i,\cdot}|_{\mathcal{L}}.$$

In particular, if $\mathcal{Y} = \mathbb{R}^{h_1\times h_1}$ and $G : \mathbb{R}^{O\times m} \to \mathbb{R}^{h_1\times h_1}$ is k-sparse the above result holds with $|\cdot|_{\mathcal{L}} = \|\cdot\|_2$ and $|\cdot|_{(\mathcal{L},\infty)^\top} = |\cdot|_{(2,\infty)^\top}$, and for $|\cdot|_{\mathcal{L}} = A$ similar result holds with the norms $|\cdot|_{\mathcal{L}_l}$ defined as maxima of L^2 norms over individual patches. Note that G need not be continuous. Possible choices of G include component-wise Relu followed be replacing the $m-k$ smallest activations by zero, or explicitly defining k entries of $G(x)$ as maxima or averages of given subsets of the entries of x.

Proof. This follows immediately from Proposition B.1 the fact that if $\mathcal{A} \subset \mathbb{R}^{d\times m}$ is such that $X\mathcal{A}$ is an $(\epsilon, \|\cdot\|_{\infty,\infty,\infty})$-cover of

$$\left\{XA : A \in \mathbb{R}^{d\times m}, \|A\|_2 \leq a\right\},$$

then $G(X\mathcal{A})$ is a $(\sqrt{k}\epsilon\rho, \|\cdot\|_{(2,\infty,\infty)^\top})$-cover of

$$\left\{G(XA) : A \in \mathbb{R}^{d\times m}, \|A\|_2 \leq a\right\}.$$

□

B.2 Covering number bound for networks with fixed norm constraints

With this result in our toolkit, we can prove a first covering number result about neural networks.

We have the following result.

Theorem B.1. *Suppose we are given an architecture as described in section A, a design matrix $X \in \mathbb{R}^{n\times U_0\times w_0}$, and numbers $0 < a_1, a_2, \ldots, a_l, s_1, s_2, \ldots, s_l$. Define the family of tensors obtained by applying the network $F_{A^1,A^2,\ldots,A^L}$ for values of $A^1, A^2, \ldots, A^L$ satisfying norm constraints as follows*

$$\mathcal{H}_X := \left\{F_{A^1,A^2,\ldots,A^L}\left(X_{i,\cdot,\cdot}\right) : \|\tilde{A}^l\|_\sigma \leq s_i \,\wedge\, \|A^l\|_2 \leq a_l\right\}.$$

Suppose also that $\forall i, \|x_i\|_{\mathcal{L}_0}^2 \leq b^2$ for some $b > 0$. We have

$$\log \mathcal{N}\left(\mathcal{H}, \epsilon, \|\cdot\|_{(\infty,\mathcal{L}_0)^\top}\right) \leq L^2b^2\prod_{i=1}^{L}s_i^2\rho_i^2\sum_{l=1}^{L}\frac{36k_la_l^2}{s_l^2\epsilon^2}\log_2\left(\frac{8\left(b\prod_{i=1}^{l-1}\rho_is_i\right)na_lO_{l-1}m_l}{\epsilon}+7\bar{W}n\right),$$

where as usual, W is the maximum width of the network.

Proof. Note that for any $x \in \mathbb{R}^{U_0\times w_0}$ with $\|x\|_2 \leq b$ and any $A^1, A^2, \ldots, A^l$ satisfying the conditions, we have $\|F_{A^1,A^2,\ldots,A^l}(x)\|_2 \leq \prod_{i=1}^{l-1}\rho_is_i$. Hence, by proposition C.1, it suffices to prove the result for $L = 1$.

The case $L = 1$ follows from Corollary B.3 applied to $\bar{O}, \bar{d}, \bar{m}$ and $\bar{X} \in \mathbb{R}^{\bar{O}\times n\times\bar{d}}$ where $\bar{O} = O_0$, $\bar{d} = d_1$, $\bar{m} = m_1$ and for $u \leq \bar{O} = O_0$, $i \leq n$ and $j \leq d$, $\bar{X}_{u,i,j} = X^{i,S_j^{1,u}}$. Note here that $S_j^{1,u} \in \{1,2,\ldots,U_0\}\times\{1,2,\ldots,w_0\}$. □

B.3 Joint generalisation bound for fixed norm constraints

The next step is to use the above, together with the classic Rademacher theorem E.1 and Dudley's Entropy integral, to obtain a result about large margin multi-class classifiers.

Theorem B.2. *Suppose we have a K class classification problem and are given n i.i.d. observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in \mathbb{R}^{U_0 \times w_0} \otimes \{1, 2, \ldots, K\}$ drawn from our ground truth distribution (X, Y), as well as a fixed architecture as described in Section A, where we assume the last layer is fully connected and has width K and corresponds to scores for each class. Suppose also that with probability one $\|x\|_{\mathcal{L}_o} \leq b$. Suppose we are given $2L$ numbers $a_1, a_2, \ldots, a_L$ and $s_1, s_2, \ldots, s_L$. For any $\delta > 0$ and any margin $\gamma > 0$, with probability $> 1 - \delta$ over the draw of the training set, for any network $\mathcal{A} = (A^1, A^2, \ldots, A^L)$ satisfying $\forall l : \|A^l\|_2 \leq a_l \wedge \|\tilde{A}^l\|_\sigma \leq s_i$, we have*

$$\begin{aligned}
&\mathbb{P}\left(\underset{j \in \{1,2,\ldots,K\}}{\arg\max} (F_L(x))_j \neq y\right) \\
&\leq \widehat{R}_\gamma + \frac{8}{n} + \frac{288}{\gamma\sqrt{n}}\sqrt{R}\left[\log_2(\Gamma n^2/\gamma + 7\bar{W}n)\right]^{\frac{1}{2}}\log(n) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}},
\end{aligned} \tag{12}$$

where

$$\widehat{R}_\gamma \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left((F_{A^1,A^2,\ldots,A^L}(x_i))_{y_i} - \max_{j \neq y_i}(F_{A^1,A^2,\ldots,A^L}(x_i))_j \leq \gamma\right),$$

$$\begin{aligned}
R &:= L^2 b^2 \prod_{i=1}^{L} s_i^2 \rho_i^2 \sum_{l=1}^{L} \frac{k_l a_l^2}{s_l^2}, \quad \textit{and} \\
\Gamma &:= \max_{l=1}^{L}\left(b\prod_{i=1}^{l-1}\rho_i s_i a_l O_{l-1} m_l\right).
\end{aligned} \tag{13}$$

Proof. We will apply the classic Rademacher theorem to the function $l_\gamma(-M(x, y))$, where $M(x, y) = (F_{A^1,A^2,\ldots,A^L}(x))_y - \max_{j \neq y}(F_{A^1,A^2,\ldots,A^L}(x))_j$, and for any $\theta > 0$ the *ramp loss* λ_θ is defined by

$$\lambda_\theta(x) := \begin{cases} 0 & x \leq -\theta \\ 1 + x/\theta & x \in [-\theta, 0] \\ 1 & \text{otherwise.} \end{cases}$$

Let us define

$$\widehat{R}_\gamma = \frac{1}{n}\sum_{i=1}^{n} l_\gamma(-M(x_i, y_i)).$$

Using this, note that we have immediately for any $\delta > 0$, that with probability greater than $1 - \delta$ over the training set:

$$\begin{aligned}
&\mathbb{P}\left(\underset{j \in \{1,2,\ldots,K\}}{\arg\max} (F_L(x))_j \neq y\right) \leq \mathbb{E}\left(l_\gamma(-M(x, y))\right) \\
&\leq \widehat{R}_\gamma + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} + 2\hat{\mathfrak{R}}_n(l_\gamma(-M(x, y))).
\end{aligned} \tag{14}$$

Applying Theorem B.1 (with $S_j^L = \{j\}$ for each $j \in \{1, 2, \ldots, K\}$ so that $\|\cdot\|_{\mathcal{L}_L} = \|\cdot\|_\infty$) to F and noting that any $(\epsilon, \|\cdot\|_\infty)$-covering of $F(X)$ (where X is the design matrix) is a $(2\epsilon/\gamma, \|\cdot\|_\infty)$ -covering of $l_\gamma(-M(x_i, y_i))$ $(i = 1, 2, \ldots, n)$, we obtain that

$$\log \mathcal{N}\left(\mathcal{H}_k, |\cdot|, \epsilon\right) \leq L^2 b^2 \prod_{i=1}^{L} s_i^2 \rho_i^2 \sum_{i=1}^{L} \frac{36 k_l a_l^2 4}{\gamma^2 s_l^2 \epsilon^2} \log_2\left(\frac{8\left(b\prod_{i=1}^{l}\rho_i s_i\right) n a_l O_{l-1} m_l}{\epsilon\gamma/2} + 7\bar{W}n\right), \tag{15}$$

where $\mathcal{H}_k$ is the function class of networks of the form $F_L(x)$ with weight matrices satisfying $\forall l : \|A^l\|_2 \leq a_l \wedge \|\tilde{A}^l\|_\sigma \leq s_i$, and $k_L = 1$. Applying Dudley's entropy formula (31) with $\alpha = \frac{1}{n}$, we then obtain, for all k:

$$\begin{aligned}
\hat{\mathfrak{R}}_n(l_\gamma(-M(x,y))) &\leq 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{\log \mathcal{N}(\mathcal{F}|S, \epsilon, \|\cdot\|_p)} \\
&\leq \frac{4}{n} + 2\sqrt{36} \frac{12}{\sqrt{n}\gamma} \sqrt{R} \int_{\frac{1}{n}}^1 \frac{\sqrt{\log_2(16\Gamma n/(\epsilon\gamma) + 7\bar{W}n)}}{\epsilon\gamma} d\epsilon \\
&\leq \frac{4}{n} + \frac{144}{\gamma\sqrt{n}} \sqrt{R} \int_{\frac{1}{n}}^1 \frac{\sqrt{\log_2(16\Gamma n^2/\gamma + 7\bar{W}n)}}{\epsilon} d\epsilon \\
&= \frac{4}{n} + \frac{144}{\gamma\sqrt{n}} \sqrt{R} \sqrt{\log_2(16\Gamma n^2/\gamma + 7\bar{W}n)} \log(n)
\end{aligned} \tag{16}$$

Plugging this back into equation (14), we obtain that for every $\delta > 0$ and every k (with $k_L = 1$ as usual) we have with probability $> 1 - \delta$ over the training set:

$$\mathbb{P}\left(\arg\max_{j \in \{1,2,\ldots,K\}} (F_L(x))_j \neq y \right) \tag{17}$$

$$\leq \widehat{R}_\gamma + \frac{8}{n} + \frac{288}{\gamma\sqrt{n}} \sqrt{R_\kappa} \left[\log_2(16\Gamma n^2/\gamma + 7\bar{W}n)\right]^{\frac{1}{2}} \log(n) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}, \tag{18}$$

as expected. □

B.4 Proof of main Theorems A.2 and A.1

All the pieces are now in place to present the

Proof of Theorem A.2. The general proof technique is similar to the proof of the main theorem in [2] and further references, the main differences being that we must use our stronger Theorem B.2 to take width reduction and weight sharing into account.

For each choice of positive integers $G, B_1, B_2, \ldots, B_L, S_1, S_2, \ldots, S_L, b$, define

$$\delta(G, B, S, b) = \frac{\delta}{2^G \prod_{l=1}^L B_l S_l (B_l + 1)(S_l + 1) b(b+1)}. \tag{19}$$

Let also

$$\mathcal{S}(G, B, S, b) = \left\{ (X, \gamma, \mathcal{A}) : \frac{1}{\gamma} \leq \frac{2^G}{n}, \forall l \leq L, \|A^l\|_2 \leq \frac{B_l}{L} \wedge \|\tilde{A}^l\|_\sigma \leq \frac{S_l}{L}, \|X\|_{(\infty, \mathcal{L}_0)^\top} \leq b \right\}.$$

Apply Theorem B.2 for $\gamma^{-1} = \frac{2^G}{n}$, $a_l = B_l$, $s_l = S_l$, $b = b$, we see that with probability $> 1 - \delta(G, B, S, b)$ over the draw of the training set, every (data, network, margin) combination $(X, \gamma, \mathcal{A}) \in \mathcal{S}(G, B, S, b)$ satisfies

$$\begin{aligned}
&\mathbb{P}_{(x,y)}\left(E_L(x,y)\right) \\
&\leq \frac{1}{n} \sum_{i=1}^n \mathbb{I}\left(M_L(x_i, y_i) \leq \frac{n}{2^G} \right) + \frac{8}{n} + 3\sqrt{\frac{\log\left(\frac{2}{\delta(G,B,S,b)}\right)}{2n}} \\
&+ \frac{288 \times 2^G}{n\sqrt{n}} \sqrt{L^2 b^2 \prod_{i=1}^L \frac{S_i^2}{L^2} \rho_i^2 \sum_{i=1}^L \frac{k_l B_l^2}{S_l^2}} \left[\log_2(16\Gamma n^2/\gamma + 7\bar{W}n)\right]^{\frac{1}{2}} \log(n)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(M_L(x_i,y_i)\leq \frac{n}{2^G}\right)+\frac{8}{n}\\
&+3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}+\frac{1}{2n}\sum_{l=1}^{L}\log(B_l(B_l+1))+\log(S_l(S_l+1))+\frac{1}{2n}\log(b(b+1))+\frac{1}{2n}\log(2^G)}\\
&+\frac{288\times 2^G}{n\sqrt{n}}\sqrt{L^2b^2\prod_{i=1}^{L}\frac{S_i^2}{L^2}\rho_i^2\sum_{i=1}^{L}\frac{k_lB_l^2}{S_l^2}}\left[\log_2(16\Gamma n^2/\gamma+7\bar{W}n)\right]^{\frac{1}{2}}\log(n)
\end{aligned}
\tag{20}
$$

where $\Gamma=\max_{l=1}^{L}\left(b\frac{S^l}{L}O_{l-1}m_l\prod_{i=1}^{l-1}\rho_i\frac{B_i}{L}\right)$,

$$
M_L(x,y):=(F_{A^1,A^2,\ldots,A^L}(x))_y-\max_{j\neq y}(F_{A^1,A^2,\ldots,A^L}(x))_j,
$$

and $E_L(x,y):=\{M_L(x,y)\leq 0\}$. Since $\sum_{(G,B,S,b)}\delta(G,B,S,b)=\delta$, we have that with probability $>1-\delta$ over draw of the training set, the above inequality holds where (G,B,S,b) are the smallest integers such that $(X,\gamma,\mathcal{A})\in(G,B,S,b)$. In this case, note that we have

$$
\begin{aligned}
\frac{B_l}{L}&\leq\|A^l\|_2+\frac{1}{L}\quad\forall l\leq L\\
\frac{S_l}{L}&\leq\|\tilde{A}^l\|_\sigma+\frac{1}{L}\quad\forall l\leq L\\
\frac{2^{G-1}}{n}&<\frac{1}{\gamma}\leq\frac{2^G}{n}\\
\|X\|_{(\infty,\mathcal{L}_0)^\top}&\leq b\leq\|X\|_{(\infty,\mathcal{L}_0)^\top}+1
\end{aligned}
\tag{21}
$$

This allows us to conclude, plugging equation (21) into equation (20) that w.p. $>1-\delta$, we have:

$$
\begin{aligned}
&\mathbb{P}_{(x,y)}\left(E_L(x,y)\right)\\
&\leq\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(M_L(x_i,y_i)\leq\frac{n}{2^G}\right)+\frac{8}{n}\\
&+3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}+\frac{1}{2n}\sum_{l=1}^{L}\log(B_l(B_l+1))+\log(S_l(S_l+1))+\frac{1}{2n}\log(b(b+1))+\log(2^G)}\\
&+\frac{2^G288}{n\sqrt{n}}\sqrt{L^2b^2\prod_{i=1}^{L}\frac{S_i^2}{L^2}\rho_i^2\sum_{i=1}^{L}\frac{k_lB_l^2}{S_l^2}}\left[\log_2\left(16n^22^G\max_{l=1}^{L}\left(be\frac{S^l}{L}O_{l-1}m_l\prod_{i=1}^{l-1}\rho_i\frac{B_i}{L}\right)+7\bar{W}n\right)\right]^{\frac{1}{2}}\log(n)\\
&\leq\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(M_L(x_i,y_i)\leq\gamma\right)+\frac{8}{n}\\
&+3\sqrt{\frac{\log\left(\frac{4n}{\delta\gamma}\right)}{2n}+\frac{1}{n}\left((2+\|X\|_{(\infty,\mathcal{L}_0)^\top})+\sum_{l=1}^{L}\log\left[(2+L\|A^l\|_2)(2+L\|\tilde{A}^l\|_\sigma)\right]\right)}\\
&+\frac{576}{\gamma\sqrt{n}}\sqrt{L^2(\|X\|_{(\infty,\mathcal{L}_0)^\top}+1)^2\prod_{i=1}^{L}\rho_i^2\sum_{i=1}^{L}k_l(\|A^l\|_2+1/L)^2\prod_{i\neq l}\left(\|\tilde{A}^l\|_\sigma+1/L\right)^2}\\
&\left[\log_2\left(\frac{32n^2}{\gamma}\max_{l=0}^{L}\left((1+\|X\|_{(\infty,\mathcal{L}_0)^\top})\prod_{i=1}^{l-1}\rho_i(\|\tilde{A}^i\|_\sigma+1/L)(\|A^l\|_2+1/L)O_{l-1}m_l\right)+7\bar{W}n\right)\right]^{\frac{1}{2}}\log(n),
\end{aligned}
\tag{22}
$$

as expected.

$\square$

Armed with this, the proof of Theorem A.1 is just a matter of simplifying into $\widetilde{\mathcal{O}}$ notation:

Proof of Theorem A.1. The proof is a matter of simplifying theorem A.2 into the $\widetilde{\mathcal{O}}$ notation. Recall that if $f, g : \mathbb{R}^m \to \mathbb{R}$, $f = \widetilde{\mathcal{O}}(x)$ iff $\lim_{n\to\infty} \frac{f(x_n)}{g\,\mathrm{Polylog}(g(x_n))} < C$ for any choice of sequence $x_1, x_2, \dots$ such that $\lim_{n\to\infty} x_n = \infty$ for some absolute constant C. Let f_0, f_1, f_2 be the three excess risk terms in Theorem A.2, it is clear that $f_0 = \frac{8}{n} = \widetilde{\mathcal{O}}\left(\frac{\sqrt{R}}{\gamma\sqrt{n}}\log\left(\max_{l\leq L} O_{l-1}m_l\right)\right)$. As for f_1, note that $\log(n)$ and $\log(\gamma)$ are both $O\left(\frac{\sqrt{R}}{\gamma\sqrt{n}}\right)$, and *be* $\prod_{i=1}^{l-1}\rho_i\left(\frac{1}{L}+\|A^i\|_{\mathcal{L}_i}\right)\left(\frac{1}{L}+\|A^l\|_2\right)$ is $o(R)$. Finally, since $\frac{\|A\|^2_{\mathcal{L}_l^*}}{\mathcal{D}_l} \leq \|A\|_2^2 \leq \mathcal{E}_l\|A\|^2_{\mathcal{L}_l^*}$, we have for large enough $\|A^l\|_2, \|\tilde{A}^l\|_{\mathcal{L}_l^*}$:

$$
\begin{aligned}
&2\sum_{l=1}^{L}\log\left[(2+L\|A^l\|_2)(2+L\|\tilde{A}^l\|_{\mathcal{L}_l^*})\right] \leq 5\left[L\log(L)+\max_{l\leq L}\log(\mathcal{E}_l)+\log\left(\prod_{i=1}^{L}\|\tilde{A}^i\|_{\mathcal{L}_i^*}\right)\right]\\
&\leq 5L\left(\log(L)+\max_{l\leq L}\log(\mathcal{E}_l)\right)+5\max_{\tilde{l}}\log\left(\frac{\|A^{\tilde{l}}\|_2}{\sqrt{\mathcal{D}_{\tilde{l}}}}\prod_{i\neq\tilde{l}}\|\tilde{A}^i\|_{\mathcal{L}_i^*}\right)\\
&\leq 5L\left(\log(L)+\max_{l\leq L}\log(\mathcal{E}_l)\right)-5\max_{\tilde{l}}\log\left(\sqrt{\mathcal{D}_{\tilde{l}}}\right)+5\log\left(\sqrt{R}\right)\\
&= O\left(\log\left(\gamma\sqrt{n}\frac{\sqrt{R}}{\gamma\sqrt{n}}\right)\right)=\widetilde{\mathcal{O}}\left(\frac{\sqrt{R}}{\gamma\sqrt{n}}\right),
\end{aligned}
$$

where $\tilde{l} = \arg\min(k_i : i \leq L)$, and at the last step, we used again the fact that $\log(n)$ and $\log(\gamma)$ are both $O\left(\frac{\sqrt{R}}{\gamma\sqrt{n}}\right)$, as well as the fact that $L\log(L)$ is $\widetilde{O}(\sqrt{R})$.

$\square$

C Chaining covering number bounds.

In this section, we state and prove a general result about the covering numbers of functions obtained through function composition. This result is mostly a combination of lemma A.7 in [2] and the beginning of the proof of Theorem 3.3 in the same reference.

Proposition C.1. *Let L be a natural number and $a_1, \dots, a_L > 0$ be real numbers. Let $\mathcal{V}_0, \mathcal{V}_1, \dots, \mathcal{V}_L$ be $L+1$ vector spaces, with arbitrary norms $|\cdot|_0, |\cdot|_1, \dots, |\cdot|_L$, let $B_1, B_2, \dots, B_L$ be L vector spaces with norms $\|\cdot\|_1, \|\cdot\|_2, \dots, \|\cdot\|_L$ and $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_L$ be the balls of radii $a_1, a_2, \dots, a_L$ in the spaces $B_1, B_2, \dots, B_L$ with the norms $\|\cdot\|_1, \|\cdot\|_2, \dots, \|\cdot\|_L$ respectively*[4]*. Suppose also that for each $l \in \{1, 2, \dots, L\}$ we are given an operator $F^l : \mathcal{V}_{l-1} \times B_l \to \mathcal{V}_l : (x, A) \to F^l_A(x)$. Suppose also that there exist real numbers $\rho_1, \rho_2, \dots, \rho_L > 0$ such that the following properties are satisfied.*

1. *For all $l \in \{1, 2, \dots, L\}$ and for all $A \in \mathcal{B}_l$, the Lipschitz constant of the operator F^l_A with respect to the norms $|\cdot|_{l-1}$ and $|\cdot|_l$ is less than ρ_l.*

2. *For all $l \in \{1, 2, \dots, L\}$, all $b > 0$, and all $\epsilon > 0$, there exists a subset $\mathcal{C}_l(b, \epsilon) \subset \mathcal{B}_l$ such that*

$$\log(\#\left(\mathcal{C}_l(b,\epsilon)\right)) \leq \frac{C_{l,\epsilon}a_l^2b^2}{\epsilon^2}, \tag{23}$$

where $C_{l,\epsilon}$ is some function of l, ϵ and, and, for all $A \in \mathcal{B}_l$ and all $X \in \mathcal{V}_{l-1}$ such that $|X|_{i-1} \leq b$, there exists an $\bar{A} \in \mathcal{C}_l(b, \epsilon)$ such that

$$\left|F^l_A(X) - F^l_{\bar{A}}(X)\right|_l \leq \epsilon. \tag{24}$$

[4] The proof works with $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_L$ being arbitrary sets, but we formulate the problem as above to aid the intuitive comparison with the areas of application of the Proposition.

For each l and each $\mathcal{A}^l = (A^1, A^2, \dots, A^l) \in \mathcal{B}^l := \mathcal{B}_1 \times \mathcal{B}_2 \times \dots, \mathcal{B}_l$, let us define

$$F^l_{\mathcal{A}^l} : \mathcal{V}_0 \to \mathcal{V}_L : x \to F^l_{\mathcal{A}^l}(x) = F^l_{A^l} \circ \dots \circ F^2_{A^2} \circ F^1_{A^1},$$

and $F_{\mathcal{A}} = F^L_{\mathcal{A}^L}$. For each $\epsilon > 0$, there exists a subset $\mathcal{C}_\epsilon$ of $\mathcal{B}^L$ such that for all $\mathcal{A} = (A^1, A^2, \dots, A^L) \in \mathcal{B} := \mathcal{B}^L$, there exists an $\bar{\mathcal{A}} \in \mathcal{C}_\epsilon$ such that the following two conditions are satisfied.

$$\left|F^l_{\mathcal{A}^l}(X) - F^l_{\bar{\mathcal{A}}^l}(X)\right|_l \leq \frac{\epsilon}{\prod_{j=l+1}^L \rho_j} \qquad (\forall l \leq L), \quad \textit{and} \tag{25}$$

$$\log \#(\mathcal{C}) \leq \frac{|X|_1^2}{\epsilon^2} \prod_{i=1}^L \rho_i^2 \left[\sum_{l=1}^L \left(\frac{C_{l,\epsilon}^{\frac{1}{2}} a_l}{\rho_l}\right)^{\frac{2}{3}}\right]^3 \leq L^2 \frac{|X|_1^2}{\epsilon^2} \prod_{i=1}^L \rho_i^2 \sum_{l=1}^L \left(\frac{C_{l,\epsilon}^{\frac{1}{2}} a_l}{\rho_l}\right)^2 .$$

In particular, for any $X \in \mathcal{V}_0$ and any $\epsilon > 0$, the following bound on the $(\epsilon, |\cdot|_L)$-covering number of $\{F_{\mathcal{A}}(X) : \mathcal{A} \in \mathcal{B}^L\}$ holds.

$$\log \mathcal{N}\left(\{F_{\mathcal{A}}(X) : \mathcal{A} \in \mathcal{B}\}, \epsilon, |\cdot|_L\right) \leq L^2 \frac{|X|_0^2}{\epsilon^2} \prod_{i=1}^L \rho_i^2 \sum_{i=1}^L \left(\frac{C_{l,\epsilon}^{\frac{1}{2}} a_l}{\rho_i}\right)^2 . \tag{26}$$

Proof. The proof draws inspiration from the ideas in [2]. However, we must keep the generality of the norms $|\cdot|_0, |\cdot|_1, \dots, |\cdot|_L$ until further into the proof, and we also keep track of the errors at the intermediary layers, yielding a stronger result.

For $l = 1, \dots, L$, let $\epsilon_l = \frac{\epsilon \alpha_l}{\prod_{i=l+1}^L \rho_i}$, where the $\alpha_l > 0$ will be determined later satisfying $\sum_{l=1}^L \alpha_l = 1$.

Using the second assumption, let us pick for each l the subset $\mathcal{C}_l = \mathcal{C}_l\left(|X|_0 \prod_{i=1}^{l-1} \rho_i, \epsilon_l\right)$ satisfying the assumption. Let us define also the set $\mathcal{C} := \mathcal{C}_1 \times \mathcal{C}_2 \times \dots \times \mathcal{C}_L \subset \mathcal{B}$.

Claim 1

For all $A \in \mathcal{B}$, there exists a $\bar{\mathcal{A}} \in \mathcal{C}$ such that for all $l \leq L$,

$$\left|F^l_{\mathcal{A}}(X) - F^l_{\bar{\mathcal{A}}}(X)\right|_l \leq \frac{\epsilon}{\prod_{j=l+1}^L \rho_j}. \tag{27}$$

Proof of Claim 1

To show this, observe first that for any $1 \leq l \leq L$ and for any $A^1, A^2, \dots, A^l$,

$$\left|F^{l-1} \circ \dots \circ F^2 \circ F^1(X)\right|_l \leq |X|_0 \prod_{i=1}^{l-1} \rho_i, \tag{28}$$

and therefore, by definition of $\mathcal{C}_l$, we have that for any $A^1, A^2, \dots, A^{l-1}$, $\{F_{A^1, A^2, \dots, A^{l-1}, A^l}(X) : A^l \in \mathcal{C}_l\}$ is an $(\epsilon_l, |\cdot|_l)$ cover of $\{F_{A^1, A^2, \dots, A^{l-1}, A^l}(X) : A^l \in \mathcal{B}_l\}$.

Let us now fix $A^1, A^2, \dots, A^L$ and define $\bar{A}_l \in \mathcal{C}_l$ inductively so that $F^l_{\bar{A}_l}(F_{\bar{A}_1, \bar{A}_2, \dots, \bar{A}_{l-1}}(X))$ is an element of $\{F^l_A(F_{\bar{A}_1, \bar{A}_2, \dots, \bar{A}_{l-1}}(X)) : A \in \mathcal{C}_l\}$ minimising the distance to $F_{\bar{A}_1, \bar{A}_2, \dots, \bar{A}_{l-1}, A_l}(X)$ in terms of the $|\cdot|_l$ norm.

We now have for all $l \leq L$:

$$\begin{aligned} |F_{\mathcal{A}}(X) - F_{\bar{\mathcal{A}}}(X)|_l &\leq \sum_{i=1}^l \left|F_{\left(\bar{A}_1, \bar{A}_2, \dots, \bar{A}_{i-1}, A^i, \dots, A^l\right)}(X) - F_{\left(\bar{A}_1, \bar{A}_2, \dots, \bar{A}_i, A^{i+1}, \dots, A^l\right)}(X)\right|_l \\ &\leq \sum_{i=1}^l \prod_{j=i+1}^l \rho_j \left|F_{(\bar{A}_1, \bar{A}_2, \dots, \bar{A}_{i-1}, A^i)}(X) - F_{\left(\bar{A}_1, \bar{A}_2, \dots, \bar{A}_i\right)}(X)\right|_l \end{aligned}$$

$$\leq \sum_{i=1}^{l} \prod_{j=i+1}^{l} \rho_j \epsilon_i = \frac{1}{\prod_{j=l+1}^{L} \rho_j} \sum_{i=1}^{l} \epsilon \alpha_i \leq \frac{\epsilon}{\prod_{j=l+1}^{L} \rho_j}, \tag{29}$$

as expected.

This concludes the proof of the claim.

To prove the proposition, we now simply need to calculate the cardinality of $\mathcal{C}$:

$$\begin{aligned}
&\log \mathcal{N}\left(\{F_{\mathcal{A}}(X) : \mathcal{A} \in \mathcal{B}\}, \epsilon, |\cdot|_L\right) \leq \log(\#(\mathcal{C})) \leq \sum_{l=1}^{L} \log(\#(\mathcal{C}_l)) \\
&= \sum_{l=1}^{L} \frac{C_{l,\epsilon} a_l^2 \left(|X|_0 \prod_{i=1}^{l-1} \rho_i\right)^2}{\epsilon_l^2} \leq \frac{1}{\epsilon^2} \sum_{l=1}^{L} \frac{C_{l,\epsilon} a_l^2 \left(|X|_0 \prod_{i=1}^{l-1} \rho_i\right)^2 \left(\prod_{i=l+1}^{L} \rho_i\right)^2}{\alpha_l^2} \\
&= \frac{|X|_0^2 \prod_{i=1}^{L} \rho_i^2}{\epsilon^2} \sum_{l=1}^{L} \frac{C_{l,\epsilon} a_l^2}{\rho_l^2 \alpha_l^2}.
\end{aligned} \tag{30}$$

Optimizing over the α_l's subject to $\sum_{l=1}^{L} \alpha_l = 1$, we find the Lagrangian condition

$$\left(-\frac{2C_{l,\epsilon} a_l^2/\rho_l^2}{\alpha_l^3}\right)_{l=1}^{L} \propto (1)_{l=1}^{L},$$

yielding

$$\alpha_l = \frac{(\sqrt{C_{l,\epsilon}} a_l/\rho_l)^{\frac{2}{3}}}{\sum_{i=1}^{L} (\sqrt{C_i} a_i/\rho_i)^{\frac{2}{3}}}.$$

Substituting back into equation (30), we obtain

$$\begin{aligned}
\log \mathcal{N}\left(\{F_{\mathcal{A}}(X) : \mathcal{A} \in \mathcal{B}\}, \epsilon, |\cdot|_L\right) &\leq \frac{|X|_0^2 \prod_{i=1}^{L} \rho_i^2}{\epsilon^2} \left[\sum_{i=1}^{L} \left(\frac{\sqrt{C_i} a_i}{\rho_i}\right)^{\frac{2}{3}}\right]^2 \sum_{l=1}^{L} \left(\frac{\sqrt{C_{l,\epsilon}} a_l}{\rho_l}\right)^{2-4/3} \\
&\leq \frac{|X|_0^2 \prod_{i=1}^{L} \rho_i^2}{\epsilon^2} \left[\sum_{l=1}^{L} \left(\frac{\sqrt{C_{l,\epsilon}} a_l}{\rho_l}\right)^{2/3}\right]^3,
\end{aligned}$$

as expected. The second inequality follows by Jensen's inequality. □

D Dudley's entropy formula

For completeness, we include a proof of (a variant of) the classic Dudley's entropy formula. To enable a comparison with the results used in [2], we write the result with arbitrary L^p norms. We will, however, only use the L^∞ version, as in [15].

Proposition D.1. *Let $\mathcal{F}$ be a real-valued function class taking values in $[0, 1]$, and assume that $0 \in \mathcal{F}$. Let S be a finite sample of size n. For any $2 \leq p \leq \infty$, we have the following relationship between the Rademacher complexity $\mathfrak{R}(\mathcal{F}|_S)$ and the covering number $\mathcal{N}(\mathcal{F}|S, \epsilon, \|\cdot\|_p)$.*

$$\mathfrak{R}(\mathcal{F}|_S) \leq \inf_{\alpha > 0} \left(4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}(\mathcal{F}|S, \epsilon, \|\cdot\|_p)}\right),$$

where the norm $\|\cdot\|_p$ on $\mathbb{R}^m$ is defined by $\|x\|_p^p = \frac{1}{n}(\sum_{i=1}^{m} |x_i|^p)$.

Proof. Let $N \in \mathbb{N}$ be arbitrary and let $\epsilon_i = 2^{-(i-1)}$ for $i = 1, 2, \ldots, N$. For each i, let V_i denote the cover achieving $\mathcal{N}\left(\mathcal{F}|S, \epsilon_i, \|\cdot\|_p\right)$, so that

$$\forall f \in \mathcal{F} \quad \exists v \in V_i \quad \left(\frac{1}{n} \sum_{t=1}^{n} (f(x_t) - v_t)^p\right)^{\frac{1}{p}} \leq \epsilon_i, \tag{31}$$

and $\#(V_i) = \mathcal{N}(\mathcal{F}|S, \epsilon_i, \|\cdot\|_p)$. For each $f \in \mathcal{F}$,let $v^i[f]$ denote the nearest element to k in V_i. Then we have, where $\sigma_1, \sigma_2, \ldots, \sigma_n$ are n i.i.d. Rademacher random variables,

$$\begin{aligned}
&\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^n \sigma_t f(x_t) \\
&= \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^n \sigma_t \left(f_t(x_t) - v_t^N[f]\right) - \sum_{i=1}^{N-1}\frac{1}{n}\sum_{t=1}^n \sigma_t \left(v_t^i[f] - v_t^{i+1}[f]\right) + \frac{1}{n}\sum_{t=1}^n \sigma_t v_t^1[f]\right] \\
&\leq \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^n \sigma_t \left(f_t(x_t) - v_t^N[f]\right)\right] + \sum_{i=1}^{N-1} \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^n \sigma_t \left(v_t^i[f] - v_t^{i+1}[f]\right)\right] \\
&\qquad + \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^n \sigma_t v_t^1[f]\right].
\end{aligned}$$

For the third term, pick $V_1 = \{0\}$, so that

$$\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^n \sigma_t v_t^1[f]\right] = 0.$$

For the first term, we use Hölder's inequality to obtain, where q is the conjugate of p,

$$\begin{aligned}
\sum_{i=1}^{N-1} \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^n \sigma_t \left(f_t(x_t) - v_t^N[f]\right)\right] &\leq \mathbb{E}_\sigma \left(\frac{1}{n}\sum_{t=1}^n |\sigma_t|^q\right)^{\frac{1}{q}} \left(\frac{1}{n}\sum_{t=1}^n \left|f_t(x_t) - v_t^N[f]\right|^p\right)^{\frac{1}{p}} \\
&\leq \epsilon_N.
\end{aligned}$$

Next, for the remaining terms, we define $W_i = \{v^i[f] - v^{i+1}[f] | f \in \mathcal{F}\}$. Then note that we have $|W_i| \leq |V_i||V_{i+1}| \leq |V_{i+1}|^2$, and then

$$\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^n \sigma_t \left(v_t^i[f] - v_t^{i+1}[f]\right)\right] \leq \mathbb{E}_\sigma \sup_{w\in W_i} \left[\frac{1}{n}\sum_{t=1}^n \sigma_t w_t\right].$$

Next,

$$\begin{aligned}
&\sup_{w\in W_i} \sqrt{\frac{1}{n}\sum_{t=1}^n w_t^2} = \sup_{f\in\mathcal{F}} \left\|v^i[f] - v^{i+1}[f]\right\|_2 \\
&\leq \sup_{f\in\mathcal{F}} \left\|v^i[f] - (f(x_1), \ldots, f(x_n))\right\|_2 + \sup_{f\in\mathcal{F}} \left\|(f(x_1), \ldots, f(x_n)) - v^{i+1}[f]\right\|_2 \\
&\leq \sup_{f\in\mathcal{F}} \left\|v^i[f] - (f(x_1), \ldots, f(x_n))\right\|_p + \sup_{f\in\mathcal{F}} \left\|(f(x_1), \ldots, f(x_n)) - v^{i+1}[f]\right\|_p \\
&\leq \epsilon_i + \epsilon_{i+1} = 3\epsilon_{i+1},
\end{aligned}$$

where at the third line, we have used the fact that $p \geq 2$. Using this, as well as Massart's lemma, we obtain

$$\mathbb{E}_\sigma \sup_{w\in W_i} \left[\frac{1}{n}\sum_{t=1}^n \sigma_t w_t\right] \leq \frac{1}{\sqrt{n}}\sqrt{2\sup_{w\in W_i}\frac{1}{n}\sum_{t=1}^n w_t^2 \log|W_i|} \leq \frac{3\epsilon_{i+1}}{\sqrt{n}}\sqrt{2\log|W_i|} \leq \frac{6}{\sqrt{n}}\epsilon_{i+1}\sqrt{\log|V_{i+1}|}.$$

Collecting all the terms, we have

$$\begin{aligned}
\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^n \sigma_t f(x_t) &\leq \epsilon_N + \frac{6}{\sqrt{n}}\sum_{i=1}^{N-1} \epsilon_{i+1}\sqrt{\log\mathcal{N}(\mathcal{F}_S, \epsilon_{i+1}, \|\cdot\|_p)} \\
&\leq \epsilon_N + \frac{12}{\sqrt{n}}\sum_{i=1}^{N} (\epsilon_i - \epsilon_{i+1})\sqrt{\log\mathcal{N}(\mathcal{F}_S, \epsilon_i, \|\cdot\|_p)}
\end{aligned}$$

$$\leq \epsilon_N + \frac{12}{\sqrt{n}} \int_{\epsilon_{N+1}}^{1} \sqrt{\log \mathcal{N}\left(\mathcal{F}_S, \epsilon, \|\cdot\|_p\right)} d\epsilon.$$

Finally, select any $\alpha > 0$ and take N to be the largest integer such that $\epsilon_{N+1} > \alpha$. Then $\epsilon_N = 4\epsilon_{N+2} \leq 4\alpha$, and therefore

$$\epsilon_N + \frac{12}{\sqrt{n}} \int_{\epsilon_{N+1}}^{1} \sqrt{\log \mathcal{N}\left(\mathcal{F}_S, \epsilon, \|\cdot\|_p\right)} d\epsilon \leq 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}\left(\mathcal{F}|_S, \epsilon, \|\cdot\|_p\right)} d\epsilon,$$

as expected. □

E Rademacher Theorem

Recall the definition of the Rademacher complexity of a function class $\mathcal{F}$:

Definition E.1. *Let $\mathcal{F}$ be a class of real-valued functions with range X. Let also $S = (x_1, x_2, \ldots, x_n) \in X$ be n samples from the domain of the functions in $\mathcal{F}$. The empirical Rademacher complexity $\mathfrak{R}_S(\mathcal{F})$ of $\mathcal{F}$ with respect to $x_1, x_2, \ldots, x_n$ is defined by*

$$\mathfrak{R}_S(\mathcal{F}) := \mathbb{E}_\delta \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \delta_i f(x_i), \tag{32}$$

where $\delta = (\delta_1, \delta_2, \ldots, \delta_n) \in \{\pm 1\}^n$ is a set of n iid Rademacher random variables (which take values 1 or -1 with probability 0.5 each).

Recall the following classic theorem([17]):

Theorem E.1. *Let $Z, Z_1, \ldots, Z_n$ be iid random variables taking values in a set $\mathcal{Z}$. Consider a set of functions $\mathcal{F} \in [0,1]^{\mathcal{Z}}$. $\forall \delta > 0$, we have with probability $\geq 1 - \delta$ over the draw of the sample S that*

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(f(Z)) \leq \frac{1}{n} \sum_{i=1}^{n} f(z_i) + 2\,\mathfrak{R}_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

References

[1] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*. openreview.net, 2018.

[2] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6240–6249. Curran Associates, Inc., 2017.

[3] Jiong Zhang, Qi Lei, and Inderjit S. Dhillon. Stabilizing gradients for deep neural networks via efficient SVD parameterization. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 5801–5809. PMLR, 2018.

[4] Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, and Tuo Zhao. On tighter generalization bounds for deep neural networks: CNNs, resnets, and beyond, 2019.

[5] Fengxiang He, Tongliang Liu, and Dacheng Tao. Why ResNet Works? Residuals Generalize. *arXiv e-prints*, page arXiv:1904.01367, Apr 2019.

[6] Yashoteja Prabhu and Manik Varma. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 263–272, New York, NY, USA, 2014. ACM.

[7] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 297–299. PMLR, 06–09 Jul 2018.

[8] Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks, 2019.

[9] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 254–263, Stockholm, Sweden, 10–15 Jul 2018. PMLR.

[10] Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a PAC-bayesian compression approach. In *International Conference on Learning Representations*. openreview.net, 2019.

[11] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019, to appear.

[12] Taiji Suzuki. Fast generalization error bound of deep learning from a kernel perspective. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1397–1406, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[13] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[14] Philip M. Long and Hanie Sedghi. Size-free generalization bounds for convolutional neural networks. *arXiv e-prints*, page arXiv:1905.12600, May 2019.

[15] Yunwen Lei, Ürün Dogan, Ding-Xuan Zhou, and Marius Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Trans. Information Theory*, 65(5):2995–3021, 2019.

[16] Tong Zhang. Covering number bounds of certain regularized linear function classes. *J. Mach. Learn. Res.*, 2:527–550, March 2002.

[17] Clayton Scott. Rademacher complexity. *Lecture Notes*, Statistical Learning Theory, 2014.