Improved Generalisation Bounds for Deep Learning Through $L^\infty$ Covering Numbers

Antoine Ledent\textsuperscript{*}
Department of Computer Science
TU Kaiserslautern

Yunwen Lei\textsuperscript{†}
Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen, China

Marius Kloft\textsuperscript{‡}
Department of Computer Science
TU Kaiserslautern and University of Southern California

30th May 2019

Abstract

Using proof techniques involving $L^\infty$ 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. Finally we provide a few further technical improvements, including improving the width dependence from before to after pooling. We also examine our bound’s behaviour on artificial data.

1 Introduction

Deep learning has enjoyed an enormous amount of success in a variety of engineering applications in the last decade [1, 2, 3, 4]. However, providing a satisfying explanation to its sometimes surprising generalisation capabilities remains an elusive goal [5, 6, 7, 8]. Previous research splits into several directions depending on a trade-off between abstraction and precise understanding. The more applied branch of the literature focuses on providing new intuitive insights into architectures, resulting in new practical results and better performance. On the other hand, the statistical theory of deep learning provides theoretical bounds on the generalisation performance of deep neural networks (DNNs) by estimating the complexity of the function class corresponding to a given architecture.

\textsuperscript{*}ledent@cs.uni-kl.de
\textsuperscript{†}leiyw@sustc.edu.cn
\textsuperscript{‡}ag-kloft@cs.uni-kl.de
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 [9, 10, 11, 12]. Many improvements have surfaced to make bounds non-vacuous at realistic scales, including better depth dependence, bounds that apply to ResNets [13] 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 accounting for weight sharing is valid only if, in each layer, the convolutional filters are orthonormal and their number is equal to their dimension [12]. This assumption is not met for the commonly used architectures (e.g., Inception, GoogLeNet, AlexNet, ResNet, etc.). (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, where we deal with an extremely high number of classes (e.g., hundreds of thousands) [14]. In [10], 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, in order to make them applicable in practical scenarios, 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.

The rest of this article is organised as follows. In Section 2, we explain in more detail the recent related literature and, in Section 3, summarise our main results in simple terms. In Section 4, we introduce more notation and present our main result in a rigorous mathematical setting. Next, we explain some ideas of the proof in Section 5. Section 6 contains an illustrative example and empirical study on artificial data. We conclude in Section 7. Finally, the appendix contains proofs and some classic results we make use of.

## 2 Related Work

Historically, the early complexity analysis of deep neural networks in the 1990s was mostly based on VC theory [15]. From 2002 [16], more papers focused on analysing the generalisation capabilities of neural networks through Rademacher analysis.

In [17], building on earlier work [18], the authors provide bounds on the Rademacher complexity of the function class consisting of fully connected neural networks of a given fixed width and length, with fixed constraints on the Frobenius norms of the weight matrices, which scale as the product $\prod_{i=1}^{L} M_i$ of the Frobenius norms $M_i$ of the weight matrices of each layer.

The now often cited paper [10], on the other hand, provides the following bound:

**Theorem 2.1** (Bartlett et al., 2017). Assume that $(x, y), (x_1, y_1), \ldots, (x_n, y_n)$ are drawn iid from any probability distribution over $\mathbb{R}^d \times \{1, 2, \ldots, K\}$. Denote by $F_A$ the function represented by the network with weights $A = \{A^1, A^2, \ldots, A^L\}$ 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 $F_A(x) = \sigma_L(A^L \sigma_{L-1}(\cdots \sigma_1(A^1x)))$. 2
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

\[
\hat{R}_\gamma(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[f(x_i)_{y_i} \leq \gamma + \max_{j \neq y_i} f(x_i)_j],
\]

\( \|X\|_{F1} \) for the Frobenius norm of the data matrix \( X \in \mathbb{R}^{n \times d} \), as well as \( \|X\|_2^2 \) for the quantity \( \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{d} X_{i,j}^2) = \frac{\|X\|_F^2}{n} \).

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

\[
\mathbb{P}(\arg \max_j (F_A(x)_j) \neq y) \leq \hat{R}_\gamma(F_A) + \tilde{O} \left( \frac{\|X\|_{2,1} M_A}{\gamma \sqrt{n}} \log(W) + \sqrt{\frac{\log(1/\delta)}{n}} \right),
\]

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

\[
M_A = \left( \prod_{i=1}^{L} \rho_i \|A^i\|_\sigma \right) \left( \sum_{i=1}^{L} \frac{\| (A^i)^\top \|_{2,1}^2}{\|A^i\|_\sigma^2} \right) \frac{1}{2}.
\]

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

Note how the above bounds are fully post hoc, scale-sensitive and have the further satisfying property of taking the classification margins into account. Therefore they represent a breakthrough in the learning theory of neural networks.

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 \( \|A^i\|_F^2 \), 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 [17], and some bounds for recurrent neural networks [19,11]. Notably, in [12], the authors provided an analogue of Theorem 2.1.

\( ^1 \)Note that the result using formula 3 can also be derived from expressing \( \| \cdot \|_{2,1} \) in terms of \( L^2 \) norms and using Jensen’s inequality.
for convolutional networks, but only under some very specific assumptions: orthogonal filters each of unit norm, with the number of filters being equal to the dimension of the filters at each layer. Those conditions are not satisfied by the typical convolutional architectures (GoogLeNet, VGG, Inception, etc.).

Since then, other lines of research (especially the PAC Bayesian school building on [9]) have focused on obtaining more meaningful bounds at realistic scales using various techniques including model compression, as well as understanding any implicit restriction on the function class imposed by the optimisation procedure [20, 21, 22, 23, 6].

Still, the fundamental questions of taking weight sharing into account in the Rademacher analysis of DNNs and reducing the dependence on the number of classes in the $L^2$ theory have so far been left unsolved, and bringing an answer to those questions is the main aim of this paper.

3 Informal Outline of Contributions

In the fully connected case, we have the following bound, which involves only $L^2$-norms of the parameter matrices, but presents no dependence on the number of classes (aside from log terms).

**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 [10]. Write $W_1, W_2, \ldots, W_L$ for the width of each layer. With probability at least $1 - \delta$ every network $F_A$ with weight matrices $A$ and every margin $\gamma > 0$ satisfy:

$$\mathbb{P}(\arg \max_j (F_A(x)_j) \neq y) \leq \hat{R}_\gamma(F_A) + \tilde{O}(\frac{\max_{i=1}^L \|x_i\|_2 R_A}{\gamma \sqrt{n}} \log(W) + \sqrt{\frac{\log(1/\delta)}{n}}),$$

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

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

and $\hat{R}_\gamma(F_A)$ is defined as in Theorem 2.1.

**Proof.** The result follows directly from Theorem 4.1, which is presented in Section 4. \qed

**Remark 3.1.** 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). Furthermore, the dependence on the spectral norm of $A^L$ in the other terms of the sum is reduced to a dependence on $\max_i \|A^L_i\|_2$. Both improvements are based on using the $L^\infty$ continuity of margin-based losses.

Finally, let us informally state the simplest version of our result on convolutional networks. More precise notation as well as technical improvements on the results will be provided in the next section.

Suppose we have a convolutional architecture where we collect the weights in matrices $A^1, A^2, \ldots,$ 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 $\bar{A}^l$ for
the matrix representing the linear operation that consists in applying each of the filters over each of the patches of the previous layer. Thus the full network can be represented in matrix form as \( F_A(x) = \sigma_L \left( \tilde{A}^L \sigma_{L-1} \left( \tilde{A}^{L-1} \ldots \sigma_1 \left( \tilde{A}^1 x \right) \right) \right) \). We have the following result, which follows directly from our general Theorem 4.1 below.

**Theorem 3.2.** For training and testing points \((x, y), (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) as usual drawn iid from any probability distribution over \( \mathbb{R}^d \times \{1, 2, \ldots, K\} \), with probability at least \(1 - \delta\), every network \( F_A \) with weight matrices \( A = \{A^1, A^2, \ldots, A^L\} \) and every margin \( \gamma > 0 \) satisfy:

\[
\mathbb{P} \left( \arg \max_j (F_A(x)_j) \neq y \right) \leq \tilde{R}_\gamma(F_A) + \tilde{O} \left( \frac{\max_{i=1}^n \|x_i\|_2 R_A \log(W) + \sqrt{\log(1/\delta) n}}{\gamma \sqrt{n}} \right),
\]

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

\[
R_A := L \left( \rho_L \max_i \|A_i \|_2 \prod_{l=1}^{L-1} \rho_l \|\tilde{A}^l\|_\sigma \right) \left( \sum_{l=1}^{L-1} \frac{(\sqrt{W_l} \|A_l\|_2)^2}{\|A^l\|_\sigma^2} + \frac{\|A^L\|_\sigma^2}{\max_i \|A_i^L\|_2^2} \right)^{\frac{1}{2}},
\]

\( A^l_{i,j} \) denotes the \( i \)th row of \( A \), and for all \( \|\cdot\|_\sigma \) and \( \|\cdot\|_2 \) denote the standard spectral and Frobenius norms respectively.

**Remark 3.2.** While (unsurprisingly) we still have to use the spectral norm of the complete convolution operation represented by \( \tilde{A}^l \) in the first factor, a notable property of the above bound is that the Frobenius norm involved is that of the matrix \( A^l \) (not \( \tilde{A}^l \)), which means we are only summing the square norms of each filter once. This means we remove a factor of \( \sqrt{O_{l-1}} \) in the \( l \)th term of the sum compared to a standard application Theorem 2.7.

## 4 Precise Notation and Results

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 \( \{1, 2, \ldots, w^1 \times \{1, 2, \ldots, U\} \} \) each of cardinality \( [w] \) and we will denote by \( S_{i,o}^o \) the \( i \)th element of \( S^o \). We will denote by \( \Lambda_A(x) \) the element of \( \mathbb{R}^{m \times O} \) such that \( \Lambda_A(x)_{j,o} = \sum_{i=1}^d X_{S^o} A_{j,i} \). In a typical example the sets \( S^1, S^2, \ldots, S^O \) represent the image patches where the convolutional filters are applied, and \( \Lambda \) would be represented via the "tf.nn.conf2d" function in Tensorflow. See Section 6 for a simple and concrete example which may help in making the definitions clear. We will also write \( \tilde{A}^l \) for the matrix in \( \mathbb{R}^{(U_{l-1} w_{l-1}) \times (O_{l-1} U_l)} \) that represents the convolution operation \( \Lambda_{A^l} \).

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 \( S_{1,o}^l \subseteq \{1, 2, \ldots, U_l\} \times \{1, 2, \ldots, w_l\} \) (for \( l \leq L, o \leq O_l \)), and \( L - 1 \) functions \( G_l : \mathbb{R}^{m_{l+1} \times O_{l+1}} \rightarrow \mathbb{R}^{U_l \times w_l} \) (for \( l = 1, 2, \ldots, L \)) satisfying the following conditions.

1. For all \( l \in \{1, 2, \ldots, L - 1\} \), \( G_l \) is \( \rho_l \) Lipschitz (component-wise) with respect to the \( L_\infty \) norm.

\(^2\)The dimensions of this matrix depend on the stride and on the size of the previous layer

\(^3\)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.
2. For all \( l \in \{1, 2, \ldots, L - 1\} \), and for each \( o \leq O_l \), \( S_l^{l,o} \) has cardinality \( d_l \).

3. For all \( l \leq L - 1 \), any \( o \leq O_l \) and for any \( x \in \mathbb{R}^{U_l \times O_{l-1}} \),

\[
\# \left( \delta \in S_l^{l,o} : G(x)_\delta \neq 0 \right) \leq k_l.
\]

\( k_l \) is the maximum number of non-zero entries in a convolutional patch of the next layer left after
the activations and pooling are applied.

4. By convention, we define \( k_L = 1 \), and \( S_L^{L,k} = \{k\} \) for each \( k \in \{1, 2, \ldots, K\} \).

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

\[
F_{A^1, A^2, \ldots, A^L} : \mathbb{R}^{U_0 \times w_0} \rightarrow \mathbb{R}^{U_L \times w_l} : x \mapsto (G_L \circ \Lambda_{A_L} \circ G_{L-1} \circ \Lambda_{A_{L-1}} \circ \ldots \circ G_1 \circ \Lambda_{A_1})(x),
\]

where for each \( l \leq L \), 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
\( S_l^{l,o} \) for each \( o \)).

Throughout the text, we also fix some norms \( \| \cdot \|_{L^0}, \| \cdot \|_{L^1}, \ldots \) and \( \| \cdot \|_{L^l} \) on the spaces \( \mathbb{R}^{U_0 \times w_0}, \mathbb{R}^{U_1 \times w_1}, \ldots \), and \( \mathbb{R}^{U_L \times w_L} \), as well as some functions \( \| \cdot \|_{L^l_{\gamma}} \) on \( \mathbb{R}^{m_l \times d_l} \) for \( 1 \leq l \leq L \) such that the
following three properties are satisfied:

1. For all \( l \leq L \) and all \( \xi \in \mathbb{R}^{U_l \times w_l} \), if \( |\xi|_{L_l} \leq 1 \), then \( \forall o \leq O_l \), \( \sum_{\delta \in S_l^{l,o}} (x_\delta)^2 \leq 1 \).

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 \( |\xi_1 - \xi_2|_{L_{l-1}} \leq a \), then

\[
|(G_l \circ \Lambda_{A^l})(\xi_1) - (G_l \circ \Lambda_{A^l})(\xi_2)|_{L_{l-1}} \leq a |A|_{L^l_{\gamma}}.
\]

3. For all \( l \), there exist real numbers \( D_l \) and \( E_l \) such \( \forall A \in \mathbb{R}^{m_l \times d_l} \),

\[
\frac{\|A\|^2_{L^2_{\gamma}}}{D_l} \leq \|A\|^2 \leq E_l \|A\|^2_{L^2_{\gamma}}.
\]
The two main examples of suitable such norms are the following.

**The standard \( L^2 \) and spectral norms.** We can set \(|A|_{\mathcal{L}_i} = |A|_{\mathcal{L}}\) for all \( l \), \(|A|_{\mathcal{L}^*} = \rho_l \tilde{A}_l \rho_l \) for all \( l \leq L - 1 \) and \(|A|_{\mathcal{L}_L} = \rho_L \max_i \|A_i\|_2\), where \( \| \cdot \|_2 \) denotes the usual spectral norm for matrices, and \( \tilde{A} \) is the circulant matrix that represents the convolution operation performed by \( \Lambda A \). This choice satisfies the conditions \( \Box \) with \( \mathcal{D}_l = w_l \) and \( \mathcal{E}_l = m_l \).

**Through Lipschitz constants.** First, for all \( l \leq L \) and all \( x \in \mathbb{R}^{U_l \times U_{l-1}} \), define \( \|x\|_{\mathcal{L}_l}^2 = \max_{0 \leq l' \leq l} \sum_{k \in S_{l',o}} (x_k)^2 \). 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 \( G \circ \Lambda A : \mathbb{R}^{U_l \times U_{l-1}} \rightarrow \mathbb{R}^{U_{l-1} \times U_l} \) with respect to the distances induced by the norms \( \| \cdot \|_{\mathcal{L}_l} \) and \( \| \cdot \|_{\mathcal{L}_l^*} \).

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 4.1** (Post-hoc asymptotic result). Assume we are given an architecture and classification problem as described in section 3. For all \( \delta > 0 \), with probability \( 1 - \delta \) over the draw of the training set it holds that every network as described in section 4, and every margins problem as described in section 4. For all \( x, y \).

\[
\mathbb{P} \left( \arg \max_j (F_A(x)_j) \neq y \right) \leq \hat{R}_n \left( F_A \right) + \tilde{O} \left( \frac{\|X\|_{(L_0, \infty)}^\top R_A}{\gamma \sqrt{n}} \log(W) + \sqrt{\frac{\log(1/\delta)}{n}} \right),
\]

where \( \|X\|_{(L_0, \infty)}^\top = \max_{i \leq n} \|x_i\|_{\mathcal{L}_0}, \quad W = \max_{l=0}^{L} O_{l-1} m_l \), and

\[
R_A = L^2 \sum_{l=1}^L k_l \rho_l^2 \|A_l\|^2 \prod_{i \neq l} \|A_i\|^2.
\]

The more precise non-asymptotic result from which Theorem 4.1 can be deduced is the following.

**Theorem 4.2** (Post-hoc result). Assume we are given an architecture and classification problem as described in section 3. For all \( \delta > 0 \), with probability \( 1 - \delta \) over the draw of the training set it holds that every network as described in section 4, and every margins \( \gamma > 0 \) satisfy:

\[
\mathbb{P}_{(x, y)} \left( \arg \max_j (F_A(x)_j) \neq y \right) \leq \hat{n} \left( \frac{\|X\|_{(L_0, \infty)}^\top + 1}{\gamma \sqrt{n}} \right) + \tilde{O} \left( \frac{\log(8n\tilde{\Gamma})}{\tilde{\Gamma}} \right)^{\frac{1}{2}} \log(n)
\]

\[
+ 3 \left( \log \left( \frac{4n}{\delta \gamma} \right) \right) + \frac{1}{n} \left( \log(2 + \|X\|_{(L_0, \infty)}^\top ) + \sum_{l=1}^L \log \left( (2 + L \|A_l\|_2) (2 + L \|A_l\|_{\mathcal{L}_l^*}) \right) \right),
\]

where

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

\[
\hat{R} = L^2 \sum_{l=1}^L k_l \rho_l^2 \left( \frac{1}{L} + \|A_l\|_2 \right)^2 \prod_{i \neq l} \left( \frac{1}{L} + \|A_l\|_{\mathcal{L}_l^*} \right)^2,
\]

and

\[
\bar{\Gamma} = \max_{l=0}^{L} \left[ \left( \|X\|_{(L_0, \infty)}^\top + 1 \right) e \left( \|A_l\|_2 + \frac{1}{L} \right) O_{l-1} m_l \prod_{i=1}^{l-1} \left( \frac{1}{L} + \|A_l\|_{\mathcal{L}_l^*} \right) \right].
\]
5 Main Ideas of the Proof

Let us first recall the following definitions:

Definition 5.1 (Covering number). Let \( \mathcal{F} \) be a class of real-valued functions defined over a space \( \mathcal{Z} \), let \( S = \{z_1, \ldots, z_n\} \subset \mathcal{Z} \), and let \( \| \cdot \| \) be a norm on \( \mathbb{R}^n \). For any \( \epsilon > 0 \), the empirical \( \| \cdot \| \)-norm covering number \( \mathcal{N}(\mathcal{F}, \epsilon, S) \) is defined as the minimal cardinality \( m \) of a collection of vectors \( \mathbf{v}^1, \ldots, \mathbf{v}^m \in \mathbb{R}^n \) such that \( \sup_{f \in \mathcal{F}} \min_{j=1,\ldots,m} \| f(z) - \mathbf{v}^j \| \leq \epsilon \). In this case, we call \( \{ \mathbf{v}^1, \ldots, \mathbf{v}^m \} \) an \( (\epsilon, \| \cdot \|) \)-cover of \( \mathcal{F} \) w.r.t. \( S \). We denote \( \mathcal{N}(\mathcal{F}, \epsilon, \| \cdot \|) = \sup_S \mathcal{N}(\mathcal{F}, \epsilon, S) \).

Definition 5.2 (Fat-Shattering Dimension). Let \( \mathcal{F} \) be a class of real-valued functions defined over a space \( \mathcal{Z} \). We define the Fat-shattering dimension \( \text{fat}_\mathcal{F}(\mathcal{F}) \) at scale \( \epsilon > 0 \) as the largest \( D \in \mathbb{N} \) such that there exist \( D \) points \( z_1, \ldots, z_D \in \mathcal{Z} \) and witnesses \( s_1, \ldots, s_D \in \mathbb{R} \) satisfying: for any \( \delta_1, \ldots, \delta_D \in \{ \pm 1 \} \) there exists \( f \in \mathcal{F} \) with

\[
\delta_i(f(z_i) - s_i) \geq \epsilon/2, \quad \forall i = 1, \ldots, D.
\]

The key to handling both class dependency and weight-sharing in the proof of our results is to obtain bounds on the \( L^\infty \) covering number of \( \delta(X) \) where \( \delta \in [0, 1] \) with logarithmic dependence on \( n \) and \( d \). Once this is taken care of, the rest of the proof follows classic chaining arguments and a union bound on probabilities of events, as in much of the literature [10, 12, 13].

Proposition 5.3. For any \( a, b > 0 \), let \( U_{a,b} = \{ x \mapsto x^\top \alpha : \| \alpha \|_2 \leq a, \| x \|_2 \leq b \} \) be a class of linear functions. For any \( a, b, \epsilon > 0 \), we have

\[
\log \mathcal{N}(U_{a,b}, \epsilon, \| \cdot \|_\infty) \leq \frac{2.7a^2b^2}{\epsilon^2} \log \left( \frac{2eban}{\epsilon} \right).
\]
Drawing inspiration from the work in [24] on SVMs, we begin by bounding the Fat-shattering dimension, and then use this to bound some mixed $L^{\infty,2}$ covering numbers. However, our proof differs in that instead of going through the worst case Rademacher complexity, we prove the bound on the shattering dimension directly, using the key Lemma below. A analogous result on the complexity of linear classifiers appeared in [25], though both the proof and the constant are different. To prove Proposition 5.3, we rely on both the Lemma below, and the classic lemma D.1 (cf. appendix) relating covering numbers and Fat-shattering dimensions. The proof of the Lemma below can be found in the appendix together with the complete proof of our results, and relies mostly on geometric arguments and Stirling’s formula.

**Lemma 5.4.** For any $a, b, \epsilon > 0$, we have $\text{fat}_\epsilon(U_{a,b}) \leq \frac{2.7a^2b^2}{\epsilon^2}.$

### 6 Illustrative Example and Experiments

For our experiments, we consider simple synthetic data constructed as follows. Each data point is a sequence of length $L$ ($L = 200, 1000, 2000, 3000, 4000$ for different experiments) of digits from the set $\{0, 1, 2, 3\}. We fix 20 "signature" sequences $s_1, s_2, \ldots, s_{20}$ of length 15, the first 10 of which (i.e. $\{s_1, s_2, \ldots, s_{10}\}$) are associated with label 0, and the last 10 (i.e. $\{s_{11}, s_{12}, \ldots, s_{20}\}$) of which are associated with label 1. Each data point is created by inserting 5 of the signatures into an originally uniformly random sequence of length $L$ at a uniformly random position. Optionally, we repeat each inserted subsequence a total of $\text{iter}$ times, where $\text{iter}$ is a parameter (duplicate signatures need not all appear consecutively). The label is determined by the majority vote of the signatures present. For instance, if signatures $s_1, s_2$ and $s_{11}$ are present, the label is 0.

We use one-hot encoding (so that each data point is of size $(4, L)$) and employ a two-layer neural network without offset terms composed of one convolutional layer without any padding and one fully connected layer. The filters of the convolutional layer are of size $(4, 15)$ and the stride is 1. We use 50 filters, and pooling is over the whole spacial region, so that the number of hidden neurons is also 50.

In this case, we have $U_0 = 4, w_0 = L, m_1 = U_1 = 50, w_1 = O_1 = 1, d = 15, O_2 = 2, O_0 = L - 15 + 1, m_2 = 2.$ Also, $S_0^{0.1} = \{(0, 0), (0, 1), (0, 2), (0, 3), (1, 0), (1, 1), \ldots, (14, 2), (14, 3)\}$, and $S_{4u+i}^{0,k} = (k - 1 + u, i).$ Finally, $S_{1}^{1,k} = \{k\}$ for $k \leq 50$, and $S_{2}^{2,k} = \{k\}$ for $k = 1, 2$.

To calculate the normalising factor $M_A$ from formula 1 as in [10], we use the standard $\|\cdot\|_{2,1}$ and spectral norms applied to the matrices representing the linear transformations. In particular, $\|X\|_{2,2} = \max_i \|x_i\|_2 = \sqrt{L}$. When applying Theorem 4.1 we use $\|x\|_{L_0}^2 = \max_{\delta \in S_0} \sum_{k \in S_0} (x_k)^2$, so that $\|x_i\|_{L_0} = \sqrt{L}$. Then $\|x\|_{L_2} = \|x\|_\infty$, $\|x\|_{L_2} = \|x\|_\infty$, $\|A^2\|_{L_2} = \|A^2\|_\sigma$, and interestingly, $\|A^1\|_{L_2} = \|A^1\|_2$.

We run the model for both $N = 350$ and $N = 20000$, in both cases for $L$ taking the values 200, 1000, 2000, 3000 and 4000. The parameter iter, which we vary proportionately to the total length (so that iter = 2, 10, 20, 30, 40), appears required for optimisation purposes. Of course, it also has some influence on generalisation, but bridging the data dependency gap is beyond the scope of this work, where we focus on non-local generalisation bounds valid on the whole of weight space.

We illustrate experimental results in Figures 2 and 3 for $L = 1000$ and $4000$. The $R$-normalised margins are orders of magnitude larger than the $M$-normalised ones. Furthermore, in both cases $N = 350$ and $N = 20000$, the value of $L$ has a strong influence on the classically normalised margins, but a milder influence on both our normalised margins and two subjective measures of data insufficiency: the test accuracy and the distribution of the margins. For $N = 20000$ and all values of $L$, the margins are clearly divided into three sets depending on how many inserted signatures in the datapoint are...
associated with the same label\footnote{\{3, 2\} is frequent and difficult to classify, \{4, 1\} is easier and rarer, \{5, 0\} is even easier and very rare.}. For $N = 350$ (all values of $L$), the three groups are still identifiable, but are less well separated, which shows the problem is in a similarly borderline insufficient data regime. In conclusion, classification problems of similar difficulty but different data size lead to similar normalised margins when using $R$ from Theorem 4.1, but very different normalised margins when using $M$ from theorem 2.1.

Figure 3 – Average margins for different values of $L$

7 Conclusion

In this paper, we have proved spectrally normalised generalisation bounds for deep neural networks with significantly reduced dependence on certain parameters and architectural choices. On the issue of generalisation bounds for extreme multi-class classification, we have completely bridged the gap between the states of the art in shallow methods and in deep learning. Furthermore, we have provided the first satisfactory answer to the weight sharing problem in the Rademacher analysis of deep neural networks. Synthetic data experiments show that our bounds are indeed substantially less affected by increases in the size of the problem under the assumption of translation invariance.

Acknowledgements

We thank Yann Guermeur for helpful discussions. This work was partially funded by the German Research Foundation (GRF) award KL 2698/2-1 and by the Federal Ministry of Science and Education (BMBF) awards 031L0023A, 01IS18051A. The work of Yunwen Lei is partially supported by the National Natural Science Foundation of China (Grant No. 61806091).
A 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 $1 \leq p, q, r \leq \infty$

$$\|X\|_{p,q,r}^r = \sum_{k=1}^{J} \left( \frac{1}{n} \sum_{j=1}^{n} \left( \sum_{i=1}^{I} |X_{i,j,k}|^p \right)^{\frac{q}{p}} \right)^{\frac{r}{q}}.$$  \hspace{1cm} (9)

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

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

2. We interpret 'tensor multiplication' for tensors as contracting the dimension in common between the tensors. 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

$$(AB)_{i,j,k} = \sum_{l=1}^{c} A_{i,l} B_{l,k}.$$  

A.1 Proof of key technical Lemmas 5.4 and 5.3

For any $a > 0$, let $A_a = \{ \alpha \in \mathbb{R}^d : \|\alpha\|_2 \leq a \}$ be the ball of radius $a$ in $\mathbb{R}^d$.

Proof of Lemma 5.4 Let $D$ be the Fat shattering dimension. According to the definition of the Fat Shattering dimension and definition of $U_{a,b}$, we know the existence of $z_1, \ldots, z_D \in A_b$ such that for any $\delta \in \{\pm 1\}^D$ we can find $a^\delta \in A_a$ satisfying

$$[(V a^\delta)_i - s_i] \delta_i \geq \epsilon,$$  \hspace{1cm} (10)

where $V \in \mathbb{R}^{D \times d}$ is the matrix with the $i$-th row being $z_i^\top$.

By abuse of notation, we will write $V$ for the linear map from $\mathbb{R}^d$ to $\mathbb{R}^D$ represented by the matrix $V$. Let us also write $B$ for $A_a \cap (\ker(V))^\perp = A_a \cap \text{span}(V_1, V_2, \ldots, V_D)$. Let $\mathbb{R}^D \supset S := \{a^\delta : \delta \in \{\pm 1\}^D \}$. We have $S \subset A_a$. Let $\bar{S}$ be the projection of $S$ on $\ker(V)^\perp$. Note that $V(S) = V(\bar{S})$, and $\bar{S} \in B$.

Let us write $C = V(B_a)$. Note that since $V$ is a linear map, $C$ is convex. Furthermore, it is clear that $V(\bar{S}) \in C$, therefore we can conclude that

$$\text{Conv}(V(\bar{S})) \subset C,$$
where Conv denotes the convex hull.

Note that by the inequalities (10),

\[ [s_1 - \epsilon, s_1 + \epsilon] \times [s_2 - \epsilon, s_2 + \epsilon] \times \cdots \times [s_D - \epsilon, s_D + \epsilon] \subset \text{Conv}(V(\bar{S})) \subset C = V(B). \]  

(11)

Note that we have

\[ \text{Vol}( [s_1 - \epsilon, s_1 + \epsilon] \times [s_2 - \epsilon, s_2 + \epsilon] \times \cdots \times [s_D - \epsilon, s_D + \epsilon] ) = (2\epsilon)^D, \]

and therefore

\[ \text{Vol}(V(B)) \geq (2\epsilon)^D. \]  

(12)

Note also that \( B \) is just a ball inside the space \( \text{span}(V_1, V_2, \ldots, V_D) \), and we have

\[ \left| \det \left( V|_{\text{ker}(V)^{\perp}} \right) \right| \leq b^D. \]  

(13)

Indeed, let us fix an arbitrary orthonormal basis of \( \text{ker}(V)^{\perp} \) and let \( W \) be the matrix representing the operator \( V|_{\text{ker}(V)^{\perp}} \) in this basis, by the fact \( z_i \in \mathcal{A}_b \) and the Cauchy-Scharz inequality, we have

\[ \text{Tr} \left( W^T W \right) \leq \text{Tr} \left( V^T V \right) = \text{Tr} \left( VV^T \right) \leq Db^2, \]

which implies \( \left| \det \left( W^T W \right) \right| \leq b^{2D} \), from which equation (13) follows.

Putting those facts together, we obtain that

\[ \text{Vol}(V(B)) \leq \frac{b^D a^D (\frac{\pi}{2})^{D/2}}{\Gamma(D/2 + 1)}. \]  

(14)

Putting inequalities (12) and (14) together, we obtain

\[ (2\epsilon)^n \leq \frac{b^D a^D (\frac{\pi}{2})^{D/2}}{\Gamma(D/2 + 1)}. \]  

(15)

We will use the following version of Sterling’s identity, which can be found in [26]:

\[ \Gamma(x) \geq \sqrt{2\pi x^{x-1/2}} e^{-x}. \]

Plugging this into the \( (D/2)^{th} \) root of equation (15), we obtain

\[ \frac{a^2 b^2}{4\epsilon^2} \pi / 2 \geq (2\pi)^{\frac{1}{2}} (D/2 + 1)^{1+\frac{1}{D}} e^{-1-\frac{2}{D}} \geq D(2\pi)^{\frac{1}{2}} e^{-1-\frac{2}{D}}. \]

(16)

Finally, rearranging equation (16) yields,

\[ D \leq \frac{a^2 b^2 \epsilon \pi (\epsilon^2 / 2\pi)^{1/D}}{4\epsilon^2} \leq \frac{9(6/5)a^2 b^2}{4\epsilon^2} = \frac{2.7 a^2 b^2}{\epsilon^2}, \]

as expected.
We can now proceed with the

**Proof of Proposition 5.3.** Using this and proposition [D.1] we can prove Lemma 5.3 in two cases.

If \( n \leq \frac{2.7a^2b^2}{\epsilon^2} \), consider the \( \epsilon \) cover \( S \) of \([-ab, ab]^n\) defined by

\[
S = \{-N\epsilon, -(N-1)\epsilon, \ldots, -\epsilon, 0, \ldots, (N-1)\epsilon, N\epsilon\}^n,
\]

where \( N = \left\lfloor \frac{ab}{\epsilon} \right\rfloor \). For each element \( s \in S \), we can define an element \( \alpha(s) \in A_a \) such that \( \|V\alpha(s) - s\|_\infty = \min(\|V\alpha - s\|_\infty : \alpha \in A_a) \). Note that the set \( \alpha(S) \) is an \( \epsilon \) cover of \( \{V\alpha : \alpha \in A_a\} \). Indeed if \( \alpha \in A_a \), let \( s \) be the closest element of \( S \) to \( V\alpha \) in terms of the \( \|\cdot\|_\infty \) norm. We have \( \|s - V\alpha\|_\infty \leq \epsilon/2 \), and therefore \( \|V\alpha(s) - s\|_\infty \leq \epsilon/2 \), from which it follows that \( \|V\alpha(s) - V\alpha\|_\infty \leq \epsilon \). Therefore we have

\[
\log \mathcal{N}(U_{a,b}, \epsilon, \|\cdot\|_\epsilon) \leq n \log \left(2\frac{ab}{\epsilon}\right) \leq \frac{2.7a^2b^2}{\epsilon^2} \log \left(2\frac{ab}{\epsilon}\right) \leq \frac{2.7a^2b^2}{\epsilon^2} \log \left(\frac{2eban}{\epsilon}\right),
\]

as expected.

If \( n > \frac{2.7a^2b^2}{\epsilon^2} \), note that by Lemma 5.4 we have that \( \text{fat}_\epsilon(U_{a,b}) \leq n \), and we can apply proposition [D.1] to obtain:

\[
\log \mathcal{N}(U, \epsilon, \|\cdot\|_\epsilon) \leq \text{fat}_\epsilon(U_{a,b}) \log \left(\frac{2ebn}{\epsilon}\right) \leq \frac{2.7a^2b^2}{\epsilon^2} \log \left(\frac{2eban}{\epsilon}\right),
\]

as expected. \( \square \)

Note first the following immediate consequence of Lemma 5.3.

**Proposition A.1.** Let positive reals \( a, b, \epsilon \) and positive integer \( m \) be given. Let the matrices \( X^u \in \mathbb{R}^{n \times d} \) for \( u \in \{1, 2, \ldots, U\} \) be given with \( \forall u \in \{1, 2, \ldots, U\}, \|X^u\|_{2,\infty} \leq b \). We will also write \( X \in \mathbb{R}^{U \times n \times d} \). We have

\[
\log \mathcal{N}\left(\{X A : A \in \mathbb{R}^{d \times m}, \|A\|_{2,2} \leq a\}, \epsilon, \|\cdot\|_{\infty, \infty, \infty}\right) \leq \frac{2.7a^2b^2}{\epsilon^2} \log \left(\frac{2eabmnU}{\epsilon}\right),
\]

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

**Proof.** This follows immediately from Lemma 5.3 (applied to the \( nmU \) data points in \( \mathbb{R}^{d \times m} \) (considered as a simple vector space with the Hadamard product used as the scalar product) defined by, for all \( \delta \in \{1, 2, \ldots, m\} \times \{1, 2, \ldots, d\} \), \( (x_{u,i,j})_{\delta} = (X^u)_{\delta i} \) for \( \delta_1 = i \) and \( (x_{u,i,j})_{\delta} = 0 \) otherwise, and the function class

\[
\{F_A : \mathbb{R}^{d \times m} \rightarrow \mathbb{R} : x \mapsto x \odot A ; \|A\|_2 \leq a\},
\]

where \( \odot \) denotes the Hadamard product. \( \square \)

**Definition A.2.** Let \( \rho > 0 \), and \( \tilde{G} : \mathbb{R}^m \rightarrow \mathbb{R}^m \) be such that for all \( i \in \{1, 2, \ldots, m\} \), \( \tilde{G}_i \) is \( \rho \) Lipschitz with respect to the \( L^\infty \) 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

\[\forall i \in \{1, 2, \ldots, m\},\]
We will call any function $G$ that can be represented in this way a $k$-sparse $\rho$-Lipschitz function (with respect to the $L^\infty$ norm).

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

**Corollary A.3.** Let $G$ be a $k$-sparse, $\rho$-Lipschitz function. Then for any $X \in \mathbb{R}^{U \times n \times d}$ such that $\|X\|_{(2,\infty,\infty)\top}^2 \leq b^2$,

$$\log N\left(\left\{G(XA) : A \in \mathbb{R}^{d \times m}, \|A\|_2 \leq a\right\}, \epsilon, \|\cdot\|_{(2,\infty,\infty)\top}\right) \leq \frac{2.7ka^2b^2}{\epsilon^2\rho^2} \log \left(\frac{2eabnmU}{\epsilon\rho\sqrt{k}}\right).$$

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 $x$. Here for a tensor $B \in \mathbb{R}^{a \times b \times c}$,

$$\|B\|_{(2,\infty,\infty)\top} = \|B\|_{(2,\infty,\infty)} = \sqrt{\max_{i=1}^a \max_{j=1}^b \sum_{k=1}^c (B_{i,j})^2} = \max_{i=1}^a \max_{j=1}^b \sum_{k=1}^c (B_{i,j,k}).$$

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

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

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

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

\[\square\]

### A.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 A.1.** Suppose we are given an architecture as described in section 4, a design matrix $X \in \mathbb{R}^{n \times U \times d}$ 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

$$\mathcal{H}_X := \left\{F_{A^1,A^2,\ldots,A^L}(X,\ldots) : \|A^l\|_{\sigma} \leq s_l \land \|A^l\|_2 \leq a_l \right\}.$$

Suppose also that $\forall i, \|x_i\|_{L_0}^2 \leq b^2$ for some $b > 0$. We have

$$\log N\left(\mathcal{H}, \epsilon, \|\cdot\|_{(\infty,\mathcal{L}_0)^\top}\right) \leq L^2b^2 \sum_{i=1}^L s_i^2 \rho_i^2 \sum_{l=1}^L \frac{2.7ka_l^2}{s_l^2\epsilon^2} \log \left(\frac{2 \left(b \prod_{i=1}^{l-1} \rho_i s_i \right) ena_i O_{l-1} m_l}{\epsilon}\right).$$

\[^3\text{From here on, we put the sample dimension first.}\]
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 \( \|F_{A_1, A_2, \ldots, A_l}(x)\|_2 \leq \prod_{i=1}^{l-1} \rho_i s_i \). Hence, by proposition [B.1] it suffices to prove the result for \( L = 1 \).

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

### A.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 A.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 4, 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 \( \|x\|_{L_0} \leq b \). 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 \( \forall l : \|A_l\|_2 \leq a_l \land \|\bar{A}_l\|_\sigma \leq s_l \), we have

\[
\mathbb{P} \left( \arg \max_{j \in \{1, 2, \ldots, K\}} (F_L(x))_j \neq y \right) \\
\leq \bar{R}_\gamma + \frac{8}{n} + \frac{80}{\gamma^2 \sqrt{n}} \sqrt{R \left( \log(\Gamma n^2 / \gamma) \right)^{\frac{1}{2}} \log(n)} + 3 \sqrt{\log(\frac{2}{\delta})} / 2n, \tag{20}
\]

where

\[
\bar{R}_\gamma \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( (F_{A_1, A_2, \ldots, A_l}(x_i))_{y_i} - \max_{j \neq y_i} (F_{A_1, A_2, \ldots, A_l}(x_i))_j \leq \gamma \right),
\]

\[
R := L^2 b^2 \prod_{i=1}^{L} \frac{2}{s_i^2} \sum_{i=1}^{L} \frac{k_i a_i^2}{s_i^2}, \quad \text{and}
\]

\[
\Gamma := \max_{l=1}^{L} \left( \frac{b}{\prod_{i=1}^{l-1} \rho_i s_i e a_l O_{l-1} m_l} \right). \tag{21}
\]

**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 \( \lambda_\theta \) 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
\[ \hat{R}_\gamma = \frac{1}{n} \sum_{i=1}^{n} l_\gamma(-M(x_i, y_i)). \]

Using this, note that we have immediately for any \( \delta > 0 \), that with probability greater than \( 1 - \delta \) over the training set:
\[
\mathbb{P} \left( \arg \max_{j \in \{1, 2, \ldots, K\}} (F_L(x))_j \neq y \right) \leq \mathbb{E} (l_\gamma(-M(x, y)))
\]
\[
\leq \hat{R}_\gamma + 3 \sqrt{\frac{\log \left( \frac{2}{\delta} \right)}{2n}} + 2 \hat{R}_n(l_\gamma(-M(x, y))). \quad (22)
\]

Applying Theorem A.1 (with \( S_j^L = \{j\} \) for each \( j \in \{1, 2, \ldots, K\} \) so that \( \| \cdot \|_{L} = \| \cdot \|_{\infty} \) 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
\[
\log \mathcal{N}(\mathcal{H}_k, \| \cdot \|, \epsilon) \leq 2^L b^2 \prod_{i=1}^{L} \sum_{i=1}^{L} \frac{2.7 \log(2)}{\gamma^2 s_i^2 \epsilon^2} \log \left( \frac{2 \left( b \prod_{i=1}^{L} \rho_i s_i \right) \log(4 \Gamma n)}{\epsilon \gamma} \right), \quad (23)
\]
where \( \mathcal{H}_k \) is the function class of networks of the form \( F_L(x) \) with weight matrices satisfying \( \forall l : \| A^l \|_2 \leq a_l \wedge \| A^l \|_\sigma \leq s_i \), and \( k_L = 1 \). Applying Dudley’s entropy formula (38) with \( \alpha = \frac{1}{n} \), we then obtain, for all \( k \):
\[
\hat{R}_n(l_\gamma(-M(x, y))) \leq 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}(F|S, \epsilon, \| \cdot \|)} \, d\epsilon
\]
\[
\leq 4 \frac{\sqrt{2.7}}{\gamma n \sqrt{n}} \sqrt{R} \int_{\alpha}^{1} \sqrt{\log(4 \Gamma n / \epsilon \gamma)} \, d\epsilon
\]
\[
\leq 4 \frac{\sqrt{2.7}}{\gamma n \sqrt{n}} \sqrt{R} \int_{\alpha}^{1} \frac{\sqrt{\log(4 \Gamma n / \epsilon \gamma)}}{\epsilon} \, d\epsilon
\]
\[
= 4 \frac{\sqrt{2.7}}{\gamma n \sqrt{n}} \sqrt{R} \sqrt{\log(4 \Gamma n / \epsilon \gamma)} \log(n)
\]
\[
= 4 \frac{\sqrt{2.7}}{\gamma n \sqrt{n}} \sqrt{R} \left[ \log(4 \Gamma n / \epsilon \gamma) \right]^{1/2} \log(n)
\]
\[
(24)
\]
Plugging this back into equation (22), 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:
\[
\mathbb{P} \left( \arg \max_{j \in \{1, 2, \ldots, K\}} (F_L(x))_j \neq y \right) \leq \hat{R}_\gamma + \frac{80}{\gamma \sqrt{n}} \sqrt{R} \left[ \log(4 \Gamma n / \epsilon \gamma) \right]^{1/2} \log(n) + 3 \sqrt{\frac{\log \left( \frac{2}{\delta} \right)}{2n}},
\]
\[
(25)
\]
as expected. \( \square \)
A.4 Proof of main Theorems 4.2 and 4.1

All the pieces are now in place to present the

Proof of Theorem 4.2. The general proof technique is similar to the proof of the main theorem in [10] and further references, the main differences being that we must use our stronger Theorem A.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

$$\delta(G, B, S, b) = \frac{\delta}{2^G \prod_{l=1}^{L} B_l S_l (B_l + 1) (S_l + 1) b (b + 1)}.$$  \hfill (26)

Let also

$$S(G, B, S, b) = \left\{ (X, \gamma, A) : \frac{1}{\gamma} \leq \frac{2G}{n}, \forall l \leq L, \|A^l\|_2 \leq \frac{B_l}{L} \land \|A_l^l\|_\infty \leq \frac{S_l}{L}, \|X\|_{(\infty, L^\infty)^+} \leq b \right \}.$$  

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

$$\mathbb{P}_{(x,y)}(E_L(x, y))$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \left( M_L(x_i, y_i) \leq \frac{n}{2G} \right) + \frac{8}{n} + \frac{3}{\sqrt{n}} \log \left( \frac{\delta^2}{\delta G, B, S, b} \right)$$  

$$+ \frac{80 \times 2^G}{n \sqrt{n}} \sum_{i=1}^{L} \frac{S_i^2}{\rho_i^2} \sum_{i=1}^{L} \frac{k_i B_i^2}{S_i^4} \left( \log \left( 4 n 2^G \right) \right) \frac{n}{2G} \log(n)$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \left( M_L(x_i, y_i) \leq \frac{n}{2G} \right) + \frac{8}{n}$$

$$+ \frac{3}{\sqrt{n}} \log \left( \frac{\delta^2}{2n} \right) + \frac{1}{2n} \sum_{i=1}^{L} \log(B_i (B_i + 1)) + \log(S_i (S_i + 1)) + \frac{1}{2n} \log(b + 1) + \frac{1}{2n} \log(2^G)$$  

$$+ \frac{80 \times 2^G}{n \sqrt{n}} \sum_{i=1}^{L} \frac{S_i^2}{\rho_i^2} \sum_{i=1}^{L} \frac{k_i B_i^2}{S_i^4} \left( \log \left( 4 n 2^G \right) \right) \frac{n}{2G} \log(n)$$  \hfill (27)

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

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

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 $(X, \gamma, A) \in (G, B, S, b)$. In this case, note that we have

$$\frac{B_l}{L} \leq \|A^l\|_2 + \frac{1}{L} \forall l \leq L,$$
\[
\frac{S_l}{L} \leq \|A'\|_{L_1} + \frac{1}{L} \quad \forall l \leq L
\]

\[
\frac{2^{G-1}}{n} < \frac{1}{\gamma} \leq \frac{2^G}{n}
\]

\[
\|X\|_{(\infty, L_0)^\tau} \leq b \leq \|X\|_{(\infty, L_0)^\tau} + 1
\]  \quad (28)

This allows us to conclude, plugging equation (28) into equation (27) that w.p. $> 1 - \delta$, we have:

\[
\mathbb{P}_{(x,y)}(E_L(x, y)) \\
\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left( M_L(x_i, y_i) \leq \frac{n}{2^G} \right) + \frac{8}{n}
\]

\[
+ 3 \sqrt{\frac{\log \left( \frac{2}{\delta} \right)}{2n}} + \frac{1}{2n} \sum_{i=1}^{L} \log(B_l(B_l + 1)) + \log(S_l(S_l + 1)) + \frac{1}{2n} \log(b(b + 1)) + \frac{1}{2n} \log(2^G)
\]

\[
+ \frac{2^{G-1}}{n} \sqrt{L^2 b^2 \prod_{i=1}^{L} \frac{S_i^2}{L^2} \rho_i^2 \sum_{l=1}^{L} \frac{k_l B_l^2}{S_l^2} \left[ \log \left( 4n2^{G} \max_{l=1}^{L} \left( \frac{b L_{i-1} m_l}{\rho_l B_l} \right) \right) \right] \frac{1}{2} \log(n)}
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left( M_L(x_i, y_i) \leq \gamma \right) + \frac{8}{n}
\]

\[
+ 3 \sqrt{\frac{\log \left( \frac{4n}{\delta} \right)}{2n}} + \frac{1}{n} \left( (2 + \|X\|_{(\infty, L_0)^\tau}) + \sum_{i=1}^{L} \log \left( (2 + L \|A'\|_2)(2 + L \|A'\|_2) \right) \right)
\]

\[
+ \frac{160}{\gamma \sqrt{n}} \sqrt{L^2(\|X\|_{(\infty, L_0)^\tau} + 1)^2 \prod_{i=1}^{L} \rho_i^2 \sum_{i=1}^{L} k_i(\|A'\|_2 + 1/L)^2 \prod_{i \neq l}(\|A'\|_{L_1} + 1/L)^2}
\]

\[
\left[ \log \left( \frac{8en}{\gamma} \max_{l=0}^{L} \left( (1 + \|X\|_{(\infty, L_0)^\tau}) \prod_{i=1}^{l-1} \rho_i(\|A'\|_{L_1} + 1/L)(\|A'\|_2 + 1/L)O_{l-1} m_l \right) \right) \right] \frac{1}{2} \log(n),
\]  \quad (29)

as expected.

Armed with this, the proof of Theorem 4.1 is just a matter of simplifying into $\tilde{O}$ notation:

**Proof of Theorem 4.1.** The proof is a matter of simplifying theorem 4.2 into the $\tilde{O}$ notation. Recall that if $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$, $f = \tilde{O}(g)$ if $\lim_{n \rightarrow \infty} \frac{f(x_n)}{g(\text{Polylog}(f(x_n)))} < C$ for any choice of sequence $x_1, x_2, \ldots$, such that $\lim_{n \rightarrow \infty} x_n = \infty$ for some absolute constant $C$. Let $f_0, f_1, f_2$ be the three excess risk terms in Theorem 4.2, it is clear that $f_0 = \frac{8}{n} = \tilde{O} \left( \frac{n}{\gamma \sqrt{n}} \log(\text{max}_{l \leq L} O_{l-1} m_l) \right)$. As for $f_1$, note that $\log(n)$ and $\log(\gamma)$ are both $O \left( \frac{n}{\gamma \sqrt{n}} \right)$, and $b \prod_{i=1}^{L} \rho_i \left( \frac{1}{L} + \|A'\|_{L_1} \right) \left( \frac{1}{L} + \|A'\|_2 \right)$ is $o(R)$. Finally, since
\[ \frac{{\|A\|_{L_2}^2}}{{\|A\|_{L_2}}} \leq \|A\|_2^2 \leq \mathcal{E}_l \|A\|_{L_2}^2, \] we have for large enough \( \|A^i\|_2, \|\tilde{A}^i\|_{L^*_i} \):

\[
2 \sum_{i=1}^{L} \log \left[ (2 + L \|A^i\|_2)(2 + L \|\tilde{A}^i\|_{L^*_i}) \right] \leq 5 \left[ L \log(L) + \max_{i \leq L} \log(\mathcal{E}_i) + \log \left( \prod_{i=1}^{L} \|\tilde{A}^i\|_{L^*_i} \right) \right]
\]

\[
\leq 5L \left( \log(L) + \max_{i \leq L} \log(\mathcal{E}_i) \right) + 5 \max_{i} \log \left( \frac{{\|A^i\|_2}}{\sqrt{D_i}} \prod_{i \neq i} \|\tilde{A}^i\|_{L^*_i} \right)
\]

\[
\leq 5L \left( \log(L) + \max_{i \leq L} \log(\mathcal{E}_i) \right) - 5 \max_{i} \log \left( \sqrt{D_i} \right) + 5 \log \left( \sqrt{R} \right)
\]

\[
= O \left( \log \left( \frac{{\sqrt{R}}}{\gamma \sqrt{n}} \right) \right) = \tilde{O} \left( \frac{{\sqrt{R}}}{\gamma \sqrt{n}} \right),
\]

where \( \tilde{i} = \arg \min(k_i : i \leq L) \), 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 \( \tilde{O}(\sqrt{R}) \).

\[ \square \]

**B Chain 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 [10] and the beginning of the proof of Theorem 3.3 in the same reference.

**Proposition B.1.** Let \( L \) be a natural number and \( a_1, \ldots, a_L > 0 \) be real numbers. Let \( \mathcal{V}_0, \mathcal{V}_1, \ldots, \mathcal{V}_L \) be \( L + 1 \) vector spaces, with arbitrary norms \( \|\cdot\|_0, \|\cdot\|_1, \ldots, \|\cdot\|_L \), let \( B_1, B_2, \ldots, B_L \) be \( L \) vector spaces with norms \( \|\cdot\|_1, \|\cdot\|_2, \ldots, \|\cdot\|_L \) and \( B_1, B_2, \ldots, B_L \) be the balls of radii \( a_1, a_2, \ldots, a_L \) in the spaces \( B_1, B_2, \ldots, B_L \) with the norms \( \|\cdot\|_1, \|\cdot\|_2, \ldots, \|\cdot\|_L \) respectively.\(^4\) Suppose also that for each \( l \in \{1, 2, \ldots, L\} \) we are given an operator \( F^l_1 : \mathcal{V}_{l-1} \times B_l \to \mathcal{V}_l : (x, A) \to F^l_1(x) \). Suppose also that there exist real numbers \( \rho_1, \rho_2, \ldots, \rho_L > 0 \) such that the following properties are satisfied.

1. For all \( l \in \{1, 2, \ldots, L\} \) and for all \( A \in B_l \), the Lipschitz constant of the operator \( F^l_1 \) with respect to the norms \( \|\cdot\|_{l-1} \) and \( \|\cdot\|_l \) is less than \( \rho_l \).

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

\[
\log(\#(\mathcal{C}_l(b, \epsilon))) \leq \frac{{C_{l,\epsilon} a_l^2 b^2}}{\epsilon^2}, \tag{30}
\]

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

\[
\left| F^l_1(X) - F^l_{\tilde{A}}(X) \right|_l \leq \epsilon. \tag{31}
\]

\(^4\)The proof works with \( B_1, B_2, \ldots, B_L \) 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 $A^l = (A^1, A^2, \ldots, A^l)$ in $B^l := B_1 \times B_2 \times \ldots, B_l$, let us define

$$F^l_{A^l} : V_0 \rightarrow V_L : x \rightarrow F^l_{A^l}(x) = F^l_{A^l} \circ \ldots \circ F^2_{A^2} \circ F^1_{A^1},$$

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

$$\left| F^l_{A^l}(X) - F^l_{\bar{A}^l}(X) \right|_l \leq \frac{\epsilon}{\prod_{j=l+1}^L \rho_j} \quad (\forall l \leq L), \quad \text{and} \quad (32)$$

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

In particular, for any $X \in V_0$ and any $\epsilon > 0$, the following bound on the $(\epsilon, \cdot |_L)$-covering number of $\{F_A(X) : A \in B^L\}$ holds.

$$\log N(\{F_A(X) : A \in B\}, \epsilon, \cdot |_L) \leq L^2 \frac{|X|}{\epsilon^2} \prod_{i=1}^L \rho_i^2 \sum_{i=1}^L \left( \frac{C_{i,\rho_i}}{\rho_i} \right) \cdot \overline{\cdot |_L}.$$

**Proof.** The proof draws inspiration from the ideas in [10]. However, we must keep the generality of the norms $| \cdot |_0, | \cdot |_1, \ldots, | \cdot |_L$ until further into the proof, and we also keep track of the errors at the intermediary layers, yielding a stronger result.

For $l = 1, \ldots, L$, let $\epsilon_l = \frac{\epsilon \alpha_l}{\prod_{j=l+1}^L \rho_j}$, 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 $C_l = C_l \left( |X|_0 \prod_{i=1}^{l-1} \rho_i, \rho_l, \epsilon_l \right)$ satisfying the assumption. Let us define also the set $C := C_1 \times C_2 \times \ldots \times C_L \subset B$.

**Claim 1**

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

$$\left| F^l_A(X) - F^l_{\bar{A}}(X) \right|_l \leq \frac{\epsilon}{\prod_{j=l+1}^L \rho_j}. \quad (34)$$

**Proof of Claim 1**

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

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

and therefore, by definition of $C_l$, we have that for any $A^1, A^2, \ldots, A^{l-1}, \{F_{A_1, A^2, \ldots, A^{l-1}, A^l}(X) : A^l \in C_l\}$ is an $(\epsilon_l, \cdot |_l)$ cover of $\{F_{A_1, A^2, \ldots, A^{l-1}, A^l}(X) : A^l \in B_l\}$.

Let us now fix $A^1, A^2, \ldots, A^L$ and define $\bar{A}^l \in C_l$ inductively so that $F^l_{\bar{A}^l}(F_{\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_{l-1}}(X))$ is an element of $\{F^l_{\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_{l-1}}(X) : A \in C_l\}$ minimising the distance to $F_{\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_{l-1}, A^l}(X)$ in terms of the $| \cdot |_l$ norm.
We now have for all \( l \leq L \):

\[
|F_A(X) - F_A(X)|_l \leq \sum_{i=1}^{l} \left| F(A_1, \ldots, A_{i-1}, A^i) - F(A_1, \ldots, A_{i-1}, A^i) \right|_l
\]

\[
\leq \sum_{i=1}^{l} \prod_{j=i+1}^{l} \rho_j \left| F(A_1, \ldots, A_{i-1}) - F(A_1, A_{i-1}) \right|_l
\]

\[
\leq \sum_{i=1}^{l} \prod_{j=i+1}^{l} \rho_j \varepsilon_i = \frac{1}{\prod_{j=l+1}^{L} \rho_j} \sum_{i=1}^{l} \varepsilon \alpha_i \leq \frac{\varepsilon}{\prod_{j=l+1}^{L} \rho_j},
\]

as expected.

This concludes the proof of the claim.

To prove the proposition, we now simply need to calculate the cardinality of \( C \):

\[
\log N \left( \{ F_A(X) : A \in B \} , \varepsilon , | \cdot |_L \right) \leq \log(\#(C)) \leq \sum_{l=1}^{L} \log(\#(C_l))
\]

\[
= \sum_{l=1}^{L} \frac{C_{l, e} a_l^2 |X|_0 \prod_{i=1}^{l-1} \rho_i}{\varepsilon^2} \leq \frac{1}{\varepsilon^2} \sum_{l=1}^{L} \frac{C_{l, e} a_l^2 |X|_0 \prod_{i=1}^{l-1} \rho_i}{\alpha_l^2} \left( \prod_{i=l+1}^{L} \rho_i \right)^2
\]

\[
= \frac{|X|_0^2 \prod_{i=1}^{L} \rho_i^2}{\varepsilon^2} \sum_{l=1}^{L} \frac{C_{l, e} a_l^2}{\rho_l^2 \alpha_l^2}.
\]

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

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

yielding

\[
\alpha_l = \frac{(\sqrt{C_{l, e} a_l / \rho_l})^{2/3}}{\sum_{i=1}^{L} (\sqrt{C_{l, e} a_l / \rho_l})^{2/3}}.
\]

Substituting back into equation (37), we obtain

\[
\log N \left( \{ F_A(X) : A \in B \} , \varepsilon , | \cdot |_L \right) \leq \frac{|X|_0^2 \prod_{i=1}^{L} \rho_i^2}{\varepsilon^2} \left[ \sum_{l=1}^{L} \left( \sqrt{C_{l, e} a_l} / \rho_l \right)^{2/3} \right]^2 \sum_{l=1}^{L} \left( \sqrt{C_{l, e} a_l} / \rho_l \right)^{2-4/3}
\]

\[
\leq \frac{|X|_0^2 \prod_{i=1}^{L} \rho_i^2}{\varepsilon^2} \left[ \sum_{l=1}^{L} \left( \sqrt{C_{l, e} a_l} / \rho_l \right)^{2/3} \right]^3,
\]

as expected. The second inequality follows by Jensen’s inequality. \( \square \)
C 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 [10], we write the result with arbitrary $L^p$ norms. We will, however, only use the $L^\infty$ version, as in [24].

**Proposition C.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 $\mathcal{R}(\mathcal{F}|S)$ and the covering number $\mathcal{N}(\mathcal{F}|S, \epsilon, \|\cdot\|_p)$.

$$\mathcal{R}(\mathcal{F}|S) \leq \inf_{\alpha > 0} \left( 4\alpha + \frac{12}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}|S, \epsilon, \|\cdot\|_p)} \, \text{d}\alpha \right),$$

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

**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 $\mathcal{N}(\mathcal{F}|S, \epsilon_i, \|\cdot\|_p)$, so that

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

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,

$$\mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \sigma_t f(x_t)$$

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

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

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

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 $p$,

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

$$\leq \epsilon_N.$$
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

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

Next,

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

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

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

Collecting all the terms, we have

\[
\mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \sigma_t f(x_t) \leq \epsilon_N + \frac{6}{\sqrt{n}} \sum_{i=1}^{N-1} \epsilon_{i+1} \sqrt{\log \mathcal{N}(\mathcal{F}_S, \epsilon_{i+1}, \| \cdot \|_p)}
\]
\[
\leq \epsilon_N + \frac{12}{\sqrt{n}} \sum_{i=1}^{N} (\epsilon_i - \epsilon_{i+1}) \sqrt{\log \mathcal{N}(\mathcal{F}_S, \epsilon_i, \| \cdot \|_p)}
\]
\[
\leq \epsilon_N + \frac{12}{\sqrt{n}} \int_{\epsilon_{N+1}}^1 \sqrt{\log \mathcal{N}(\mathcal{F}_S, \epsilon, \| \cdot \|_p)} d\epsilon.
\]

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

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

as expected. \hfill \Box

\section*{D \ Relationship between covering number and fat shattering dimension}

To proceed with the proof of Lemma 5.3, we will need the following classic result (see [24, 27, 28] and [29]).
Proposition D.1 (Control of the covering number by the Fat-shattering dimension). Let $\mathcal{F}$ be a class of real-valued functions defined over a space $\mathcal{S}$ and $\mathcal{S} = \{x_1, x_2, \ldots, x_n\} \in \mathcal{S}^n$ a sample of cardinality $n$. If $\mathcal{F}$ takes values in $[-B, B]$, and $\text{fat}_\varepsilon(\mathcal{F}) < n$ we have
\[
\log \mathcal{N}(\mathcal{F}, \varepsilon, \mathcal{S}) \leq \text{fat}_\varepsilon(\mathcal{F}) \log \left( \frac{2eBn}{\varepsilon} \right).
\] (39)

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 $S = (x_1, x_2, \ldots, x_n) \in X$ 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
\[
\mathfrak{R}_S(\mathcal{F}) := \mathbb{E}_\delta \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \delta_i f(x_i),
\] (40)
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([30]):

Theorem E.1. Let $Z, Z_1, \ldots, Z_n$ be iid random variables taking values in a set $Z$. Consider a set of functions $\mathcal{F} \in [0, 1]^Z$. $\forall \delta > 0$, 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] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc., 2012.
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778. IEEE Computer Society, 2016.
[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. CoRR, abs/1812.04948, 2018.
[4] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science, 362(6419):1140–1144, 2018.
[5] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. 2017.
[6] 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.

[7] Amir Asadi, Emmanuel Abbe, and Sergio Verdu. Chaining mutual information and tightening generalization bounds. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7234–7243. Curran Associates, Inc., 2018.

[8] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

[15] Martin Anthony and Peter Bartlett. *Neural Network Learning: Theoretical Foundations*. 01 2002.

[16] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[17] 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.

[18] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter GrÂNEnwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1376–1401, Paris, France, 03–06 Jul 2015. PMLR.
[19] Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks, 2019.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] G. J. O. Jameson. A simple proof of stirling’s formula for the gamma function. The Mathematical Gazette, 99(544):68–74, 2015.

[27] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. Probability Theory and Related Fields, 161(1):111–153, Feb 2015.

[28] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. 09 2010.

[29] Sham Kakade and Ambuj Tewari. Fat shattering dimension and covering numbers. CMSC 35900 (Spring 2008) Learning Theory, 2008.

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