
Generalization Bounds for Rank-sparse Neural Networks

Antoine Ledent

School of Computing and Information Systems (SCIS)
Singapore Management University (SMU)
aledent@smu.edu.sg

Rodrigo Alves

Department of Applied Mathematics
Czech Technical University in Prague (CTU)
rodrigo.alves@fit.cvut.cz

Yunwen Lei

Department of Mathematics
The University of Hong Kong
leiyw@hku.hk

Abstract

It has been recently observed in much of the literature that neural networks exhibit a bottleneck rank property: for larger depths, the activation and weights of neural networks trained with gradient-based methods tend to be of approximately low rank. In fact, the rank of the activations of each layer converges to a fixed value referred to as the “bottleneck rank”, which is the minimum rank required to represent the training data. This perspective is in line with the observation that regularizing linear networks (without activations) with weight decay is equivalent to minimizing the Schatten p quasi norm of the neural network. In this paper we investigate the implications of this phenomenon for generalization. More specifically, we prove generalization bounds for neural networks which exploit the approximate low rank structure of the weight matrices if present. The final results rely on the Schatten p quasi norms of the weight matrices: for small p , the bounds exhibit a sample complexity $\tilde{O}(WrL^2)$ where W and L are the width and depth of the neural network respectively and where r is the rank of the weight matrices. As p increases, the bound behaves more like a norm-based bound instead.

1 Introduction

The success of neural networks in many applications during the most recent artificial intelligence boom has driven substantial research efforts into understanding their generalization behavior [1, 2, 3, 4, 5, 6]. Whilst early work [7] focused on traditional approaches such as VC dimensions and parameter count, the increasingly evident ability of DNNs to generalize well (even in massively *overparametrized* regimes) has shifted the community’s focus beyond purely architectural bounds [8, 9]. It is now accepted that a satisfactory explanation of Neural Networks’ generalizing abilities must incorporate more refined information about the training process and the properties of the representation functions, which spontaneously arise from training DNNs on *natural data* [10, 11].

One of the simplest ways to characterize the ‘simplicity’ of functions learnt by neural networks through the geometry of weight space is by considering *norms* of the *weight matrices*. Indeed, several works have provided various bounds with reduced or absent explicit dependence on architectural quantities, preferring to express neural network capacity through various functions of the weight matrices. Such generalization error bounds, which primarily rely on the norms of the weight matrices to express function class capacity are generally referred to as ‘*norm-based bounds*’. The earliest attempts include many bounds obtained through various vector contraction inequalities and the

‘peeling technique’ [12, 13]. In this subcategory of bounds, the Rademacher complexity is bounded iteratively by peeling off layers, resulting in a product of norms of the weight matrices. For instance, the Rademacher complexity bound in [13], which extends earlier work from [12], scales as ¹:

$$\left(\sqrt{\frac{\mathfrak{B}^2 L \prod_{\ell=1}^L \|A_\ell\|_{\text{Fr}}^2}{N}} \right). \quad (1.1)$$

Here, N is the number of samples, \mathfrak{B} is an upper bound on the norms of the samples and A_1, \dots, A_L are the network’s weight matrices. A remarkable feat of this bound is the complete lack of explicit dependence on architectural parameters (i.e., depth, width, etc.). However, the implicit dependence introduced by the product of Frobenius norm is large: even assuming that each weight matrix has unit spectral norm, if the rank of each A_ℓ is r and each matrix has homogeneous spectrum, then the bound will still correspond to a sample complexity of r^L , an exponential dependence on the depth L .

Later norm-based bounds have largely favored alternative approaches via covering numbers, which yield a less dramatic exponential depth dependence. The most recognizable examples of such results include the bounds of [3] and [1], which were independently discovered through two approaches:

$$\tilde{O} \left(\frac{L\sqrt{W}}{\sqrt{N}} \left(\prod_{\ell=1}^L \|A_\ell\| \right) \left(\sum_{\ell=1}^L \frac{\|A_\ell - M_\ell\|_{\text{Fr}}^2}{\|A_\ell\|_\sigma^2} \right)^{\frac{1}{2}} \right), \quad (1.2)$$

$$\tilde{O} \left(\frac{1}{\sqrt{N}} \prod_{\ell=1}^L \|A_\ell\| \left(\sum_{\ell=1}^L \frac{\|(A_\ell - M_\ell)^\top\|_{2,1}^{\frac{2}{3}}}{\|A_\ell\|^{\frac{2}{3}}} \right)^{\frac{3}{2}} \right). \quad (1.3)$$

Here the M_ℓ s are arbitrary *reference matrices* chosen in advance ². The implicit exponential dependence in depth is reduced to a product of *spectral* norms which, at least in theory, could be expected to scale as $O(1)$ or forced to be a unit value [14]. With this interpretation, it becomes clearer that the bounds improve when the networks stay close to initialization, especially if the number of updated weights is sparse, which would make the norm $\|(A_\ell - M_\ell)^\top\|_{2,1}$ small.

Such observations can begin to explain some of DNNs’ successes, in particular with respect to the lottery ticket hypothesis [15]. However, they do not paint a complete picture, and are largely vacuous by large margins. In fact, the most impressive recent bounds in the theory of neural networks tend to implicitly incorporate data dependence. For instance, in [5, 6, 16] various upper bounds on norms of activations of intermediary layers and Lipschitz constants of gradients are replaced by empirical equivalents through a powerful technique called loss function augmentation, greatly attenuating the implicit exponential dependency on depth. Pushing it to the extreme, [11] prove bounds which incorporate a notion of confidence for intermediary concepts learnt at each layer.

The discovery of the Neural Tangent Kernel [17, 18] is certainly one of the biggest breakthroughs in deep learning theory in recent years. Indeed, the underlying theory demonstrates that Neural Networks behave approximately like kernels at the overparametrized regime, and are able to characterize the model’s generalization behavior directly in terms of the alignment between the labels and the NTK features of the data distribution. However, such bounds rely on very strict overparametrization assumptions. Furthermore, it has been shown that the NTK regime does not outperform the moderate overparametrization regime [19, 20].

Another tremendous advance towards understanding the generalization power of neural networks occurred with the gradual understanding that *depth induces low-rank representations* [21, 22, 23, 24, 25, 26]. In other words, the training dynamic of DNNs through gradient based methods naturally encourages low-rank representations, converging to a bottleneck intrinsic dimension which captures the data’s features parsimoniously. In particular, the connection between weight decay and implicit Schatten quasi-norm regularization has been made in [26, 24]. There has been a lot of recent interest in this phenomenon, which has been demonstrated experimentally in various contexts spanning both DNNs, CNNs and even LLMs [27]. The phenomenon can most easily be understood by initially considering the case of *linear neural networks*. In this case, we have the following powerful result:

¹For simplicity, we assume that the nonlinearities are 1-Lipschitz in this related works section.

²For instance, they can be set to zero or the initialized values of the weights

Theorem 1.1 (cf. [24], Theorem 1). *Let $w \geq \min(\mathcal{C}, d)$ and let B_1, \dots, B_L be matrices such that $B_1 \in \mathbb{R}^{w \times d}$, $B_L \in \mathbb{R}^{\mathcal{C} \times w}$, $B_2, \dots, B_{L-1} \in \mathbb{R}^{w \times w}$. For any matrix $A \in \mathbb{R}^{\mathcal{C} \times d}$ we have*

$$L\|A\|_{\text{sc}, \frac{2}{L}}^{\frac{2}{L}} = \min \sum_{\ell=1}^L \|B_\ell\|^2 \quad \text{subject to} \quad B_L B_{L-1} \dots B_1 = A. \quad (1.4)$$

For any matrix Z and any $0 < p \leq 2$, our notation $\|Z\|_{\text{sc}, p}$ refers to the Schatten p quasi norm of Z ³:

$$\|Z\|_{\text{sc}, p}^p = \sum_i \sigma_i(Z)^p, \quad (1.5)$$

where $\sigma_i(Z)$ refers to the i th singular value of Z (listed in decreasing order). Thus, as p approaches zero, the quantity $\|Z\|_{\text{sc}, p}^p$ approaches the rank of the matrix Z . Slightly larger values of p induce a softer form of rank sparsity. Thus, Theorem 1.1 implies that a simple weight decay constraint on factor matrices B_1, \dots, B_L is equivalent to a rank-sparsity inducing constraint on $\|A\|_p^p$ where $p = \frac{2}{L} \rightarrow 0$ when $L \rightarrow \infty$. Deep results [24, 22] in optimization theory demonstrate that under specific circumstances, the above described implicit rank-sparsity inducing effect of weight decay at large depth *occurs even in the presence of activation functions*: the ranks of each layer’s *post activation representations*⁴ converges to a single quantity referred to as the “bottleneck rank”, which can be understood as the minimum dimension required to effectively represent the data.

In this work, we characterize the effects of rank-sparsity in the weight matrices in terms of *generalization*. We provide bounds which *interpolate between the norm-based and parametric regimes* at each layer with the *parametric interpolation* technique originally developed for matrix completion [28]. This allows us to control the L^∞ (cf. Prop. F.2) or L^2 (cf. Prop. I.3) covering number of each layer in term of the Schatten p quasi norms of the weight matrices. Treating Lipschitz and boundedness conditions on the loss function as constant, our contributions are summarized as follows.

- We prove new generalization bounds for **linear networks** which incorporate the implicit low-rank effects of depth by incorporating the Schatten p quasi norm of the weight matrix. As $p \rightarrow 0$, the bound provides a sample complexity of $\tilde{O}([\mathcal{C} + d]\text{rank}(A))$. To the best of our knowledge, this is a new characterization of the generalization behavior of multi-class linear classification with low-rank dependencies between the classes.
- For fully-connected neural networks, we prove (cf. Theorem 3.2) bounds of the order of $\tilde{O}\left(\sqrt{\frac{L}{N}} \left[\sum_{\ell=1}^L \left[\mathfrak{B} \prod_{i=1}^L \rho_i \|A_i\| \right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\|A_\ell - M_\ell\|_{\text{sc}, p_\ell}^{p_\ell}}{\|A_\ell\|^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [w_\ell + w_{\ell-1}]^{1+\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}}\right)$.
- Our results hold simultaneously over every sequence of Schatten indices p_ℓ s, which we tune to balance the effects of the norm-based factor $\left[\mathfrak{B} \prod_{i=1}^L \rho_i \|A_i\| \right]^{2p_\ell/[p_\ell+2]}$ and the low rank structure in the term $\|A_\ell\|_{\text{sc}, p_\ell}^{p_\ell}/\|A_\ell\|^{p_\ell}$ (for $M_\ell = 0$).
- We also provide extensions of those results relying on Loss Function Augmentation to replace the norm-based factor $\mathfrak{B} \prod_{i=1}^L \rho_i \|A_i\|$ by $\mathfrak{B}_{\ell-1, A}^{\mathfrak{C}} \mathfrak{B} \prod_{i=\ell}^L \rho_i \|A_i\|$ where $\mathfrak{B}_{\ell-1, A}^{\mathfrak{C}}$ is the maximum norm of a convolutional patch at layer $\ell - 1$.
- To the best of our knowledge, the proof technique, *parametric interpolation*, has only been previously used in the context of matrix completion [28], a substantially different setting.

2 Related Works

Deep Learning Theory is a vast and active area of research spanning various topics from optimization [29, 30, 31] to generalization [1, 32] and the intersection between both [17, 33]. In this section, we focus on norm-based and parameter-counting bounds for neural networks satisfying certain norm

³For $p \geq 1$, $\|\cdot\|_{\text{sc}, p}$ is a norm, but we refer to it as a quasi-norm to cover the case $p < 1$ and $p \geq 1$.

⁴Whilst the low-rank properties of the *activations* and *weight matrices* are not exactly equivalent, they are closely related. Indeed, consider N input activations x_1, \dots, x_N lying on a subspace of dimension r , then for any weight matrix $W \in \mathbb{R}^{m \times n}$, there exists a matrix $\tilde{W} \in \mathbb{R}^{m \times d}$ of rank $\leq r$ s.t. $\tilde{W}x = Wx$ for all i .

constraints. Indeed, this is the branch of literature closest to our work. A more mathematically detailed description of the related works is available in Sections B and C.

Norm-based Bounds by peeling techniques were established for Deep Neural Networks in [12] and [13], resulting in bounds involving products of Frobenius norms of the weights (cf. equation (1.1)). In [34], further norm-based bounds were proved involving the products of the L^1 norms of the columns of the weight matrices. Recently, the technique has also been ingeniously extended to **Convolutional Neural Networks (CNNs)** in [32], which elegantly exploits the sparsity of connections to achieve in bounds of the order $\frac{L_1 \prod_{\ell=1}^L \|A_\ell\|_{\text{Fr}} \sqrt{L} \prod_{\ell=1}^L \underline{w}_\ell \underline{B}}{\sqrt{N}}$, where \underline{w}_ℓ is the spatial dimension of the convolutional patches at layer ℓ and \underline{B} is the maximum L^2 norm of any convolutional patch. The results in [32] are not directly comparable to ours, and are especially well adapted to the case where the spatial dimension is very small. Indeed, as long as the convolutional filters are not one dimensional, the bound depends exponentially on depth. [16] also proves generalization bounds for CNNs which take weight sharing into account, and the bounds form the basis for the treatment of the norm based components in our proofs. However, the bounds in those works cannot incorporate rank-sparsity into the analysis and scale unfavorably with depth. There is also a variety of works which applies generalization analyses to non-standard learning settings or loss functions, including pairwise [35, 36, 37] learning or triplet-based learning and contrastive learning [38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48]. However, such works usually focus on the effect of the loss function. In addition to the above fundamental works, many works provide non-vacuous bounds for practical architectures through various post processing techniques including data-dependent priors [49], **optimization of the bound** or network compression/discretization [50, 51, 52, 53]. A particularly notable contribution is made in [54], which provides PAC Bayesian generalization bounds for two-layer networks where the main complexity term involves Frobenius norms of the weight matrices, remarkably, without an additional factor of the width, achieving non vacuous bounds in various cases. However, the bounds are PAC Bayesian and not directly comparable to ours since the training error is replaced by an expectation over a posterior distribution (derandomized analogues such as [55] reintroduce a factor of width), only GELU and ERF activation functions are covered and the number of layers is limited to 2. The **gradient dynamic** of deep neural networks and deep linear networks was widely studied from an optimization perspective to demonstrate neural rank collapse, with some works also covering networks with activations [24, 56, 27, 57, 58, 59, 60, 61, 62, 63]. In terms of the implications of such results for **linear networks** on *generalization*, one can of course consider the results which apply to neural networks with activations and apply them to this case [1, 3] (see Corollary C.1 for a transfer of such results to the linear case), or construct Rademacher complexity bounds specifically for linear networks as in [64]. However, such approaches involve an exponential dependence on depth which make it more difficult to capture any low-rank structure in the classes, since such structure appears for larger values of L (smaller values of p). One can also consider all existing bounds for the *multi-class linear classification problem*, where the state of the art is provided by [65] which proves a bound of $\tilde{O}\left(\sqrt{\frac{\|A\|_{\text{Fr}}^2 \mathfrak{B}^2}{N}}\right)$, improving earlier results from [66, 67, 68] in terms of implicit dependence on \mathcal{C} . Since then, the result has been extended to many variants of the learning problem, including multi-label learning, structured output prediction and ranking problems [69, 70, 71, 72, 73].

Parameter counting bounds such as those of [4] and [74] apply both to CNNs and DNNs (in fact, [74] proves both norm based and parameter counting bounds for DNNs, CNNs and ResNets). Taking into account implicit factors of L coming through logarithmic factors, the tightest parameter counting bound in both cases scales like $\tilde{O}\left(\sqrt{\frac{\mathcal{W}L}{N}}\right)$ where \mathcal{W} is the number of parameters in the network. Our own result in Corollary G.6 additionally takes the low-rank structure into account. Norm-based [75] and parameter counting [76] bounds have also been recently provided for transformers [77].

To the best of our knowledge, there are only two works which have utilized the *low-rank structure* in neural networks: [13], relies on Schatten p norms of weight matrices to improve on the bound (1.1) for matrices with low Schatten p quasi-norm. The argument relies on Lipschitz continuity arguments and yields bounds with a decay in N worse than $O(1/\sqrt{N})$. The recent work [64] manages to achieve tighter bounds taking explicit low-rank structure into account. The resulting bound scales as $L_1 C_1^L \mathfrak{B} \left[\prod_{\ell=1}^L \|A_\ell\| \right] Lr \sqrt{\frac{\underline{W}}{N}}$. One of the consequences of the proof strategy is the presence of

explicit exponential dependence on L through the term C_1^L where C_1 is an intractable constant⁵. In addition, even ignoring the factor of C_1^L , the bound still involves a product of spectral norms (with an exponent of 1), whereas our bounds involve a product of spectral norms with a potentially smaller, tunable exponent of $\frac{2p\epsilon}{p\epsilon+2}$. However, the elegant and short proof in [64] does imply that there are far fewer polylogarithmic factors compared to our bounds. See Appendix C for detailed formulae.

Our proof technique draws some inspiration from [28], which introduced the *parametric interpolation technique* to estimate the complexity of **matrix completion** with Schatten p quasi norm constraints. However, the derivations are radically different. Indeed, in the matrix completion literature on *approximate recovery* [79, 80, 81], the *observations* are *entries* of an unknown matrix A : the independent variable $x = (i, j)$ in the supervised learning formulation is drawn from an arbitrary distribution over the discrete set of entries $[m] \times [n]$. In contrast, in our work, the *observations* are modelled as *outputs of a neural network*, i.e., they take the form $F(x_i)$ where F is a neural network as defined in equation (3.6). Even for a single linear layer, the learning setting is different and the outputs take the form Ax_i instead of $A_{i,j}$. The analysis in [28] relies on bounding Rademacher complexities directly through elaborate chaining arguments, bypassing the need for covering numbers and interpolating between $p = 0$ and $p = 1$. In fact, the authors mention that any readily obtainable covering number would yield vacuous bounds in the matrix completion context. In contrast, our analysis interpolates between $p = 0$ and $p = 2$ and produces chainable covers for weight spaces with Schatten constraints. A more detailed discussion of related works is provided in Appendix C.

3 Main Results

We consider a learning problem using a loss function $l : \mathbb{R}^C \times \mathcal{Y} \rightarrow \mathbb{R}^+$, where C is the number of classes (in classification) or 1 (in regression). In this work, we assume loss function is L_1 -Lipschitz with respect to the L^∞ norm and uniformly bounded by \mathcal{B} . In regression, the output space is $\mathcal{Y} = \mathbb{R}$ whilst in classification, it is $\mathcal{Y} = \{1, \dots, C\}$. In the case of regression, an example of such a loss function is the truncated square loss

$$l(\hat{y}, y) = \min(\mathcal{B}, |\hat{y} - y|^2). \quad (3.1)$$

Indeed, in this case, the loss function is uniformly bounded by \mathcal{B} and uniformly $2\mathcal{B}$ -Lipschitz. For classification, a typical example used in much of the literature is the following margin loss [1, 4]:

$$l(\hat{y}, y) = \begin{cases} 1, & \text{if } \arg \max_i \hat{y}_i \neq y, \\ 1 - \frac{\hat{y}_y - \max_{i \neq y} \hat{y}_i}{\gamma}, & \text{if } 0 \leq \hat{y}_y - \max_{i \neq y} \hat{y}_i \leq \gamma, \\ 0, & \text{if } \hat{y}_y \geq \max_{i \neq y} \hat{y}_i + \gamma. \end{cases} \quad (3.2)$$

In this case, the loss function is uniformly bounded by $\mathcal{B} = 1$ and uniformly $2/\gamma$ -Lipschitz with respect to the L^∞ norm. We assume the learner is provided with N i.i.d. samples $x_1, \dots, x_N \in \mathcal{X} = \mathbb{R}^d$ and the associated labels $y_1, \dots, y_N \in \mathcal{Y} = \{1, 2, \dots, C\}$, each of which is drawn from a joint distribution $\mathcal{D} \sim \mathcal{X} \times \mathcal{Y}$. We are interested in high probability bounds on the *generalization gap*:

$$\text{GAP} := \mathbb{E}[l(F_A(x), y)] - \widehat{\mathbb{E}}[l(F_A(x), y)] = \mathbb{E}[l(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N l(F_A(x_i), y_i), \quad (3.3)$$

where $\widehat{\mathbb{E}}$ denotes the empirical expectation over the sample $(x_1, y_1), \dots, (x_N, y_N)$, and $F_A = F_{(A_1, \dots, A_L)}$ is our trained network. For classification, an advantage of the margin loss described in Eq. (3.2) is that a bound on GAP immediately translates to a bound on the probability of misclassification: let $I_{\gamma, X} = \frac{|\{i : \hat{y}_y - \max_{i \neq y} \hat{y}_i \leq \gamma\}|}{N}$ denote the proportion of samples which are either classified incorrectly or correctly but with margin $\leq \gamma$. The misclassification probability [1, 4, 74] satisfies

$$\mathbb{P}\left(\arg \max_c [F_A(x_i)]_c \neq y\right) \leq \mathbb{E}[l(F_A(x), y)] \leq \text{GAP} + \widehat{\mathbb{E}}[l(F_A(x), y)] \leq \text{GAP} + I_{\gamma, X}. \quad (3.4)$$

⁵The constant is described as ‘rather large’ in [78], since it accumulates factors of Talagrand’s majorizing measure theorem and generic chaining, suggesting at an absolute minimum $C_1 \geq 11$.

We provide results for three types of models: Linear Networks, Deep Neural Networks, and Convolutional Neural Networks (CNN). In all the results below, the matrices M_ℓ are reference matrices which can take any fixed value as long as it is chosen in advance. Thus, they play an analogous role to a prior in PAC-Bayesian bounds [49]. Typical choices for M_ℓ include $M_\ell = 0$ or setting M_ℓ to the initialization when training a neural networks with gradient based methods. Whilst traditional norm-based bounds such as (1.3) typically benefit from setting M_ℓ to the initialization, our bounds typically perform better by setting $M_\ell = 0$, since this is the configuration which allows the quantity $\frac{\|A_\ell - M_\ell\|_{\text{sc},p}^p}{\|A_\ell\|^p} = \frac{\|A_\ell\|_{\text{sc},p}^p}{\|A_\ell\|^p}$ to approach $\text{rank}(A_\ell)$. We provide an experimental evaluation of the behavior of our bounds on MNIST and CIFAR-10 for both DNNs and CNNs in Appendix D.

3.1 Linear Networks

We first present our results for linear networks, which amount to generalization bounds for multi-class *linear classification* with a bounded loss in the presence of a low-rank structure over the classes. However, the true value of those results is to illustrate the potential gains in generalization ability which arise from the implicit low-rank structure which appears in neural networks trained with gradient-based methods. In this section, we consider classifiers defined as follows:

$$F : \mathbb{R}^d \rightarrow \mathbb{R}^C : x \rightarrow Ax = B_L B_{L-1} \dots B_1 x \quad (3.5)$$

for some matrices $B_L \in \mathbb{R}^{C \times w}$, $B_{L-1} \in \mathbb{R}^{w \times w}$, \dots , $B_2 \in \mathbb{R}^{w \times w}$, $B_1 \in \mathbb{R}^{w \times d}$.

We have the following result, which shows the implications of Theorem 1.1 for generalization:

Theorem 3.1. *W.p. $\geq 1 - \delta$ over the draw of the training set, every linear network as defined in equation (3.5) satisfies the following generalization bound: $\text{GAP} - O\left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}}\right) \leq$*

$$\tilde{O}\left(\sqrt{\frac{[\mathcal{L}_1 \mathfrak{B}]^{\frac{2p}{2+p}} \|A\|_{\text{sc},p}^{\frac{2p}{2+p}} \min(\mathcal{C}, d)^{\frac{p}{p+2}} [\mathcal{C} + d]^{\frac{2}{p+2}}}{N}}\right) \leq \tilde{O}\left(\sqrt{\frac{[\mathcal{L}_1 \mathfrak{B}]^{\frac{2}{L+1}} [\sum \|B_\ell\|_{\text{Fr}}^2]^{\frac{L}{L+1}} [\mathcal{C} + d]}{NL^{\frac{L}{L+1}}}}\right)$$

where $\mathfrak{B} = \sup_{i=1}^N \|x_i\|$, and the \tilde{O} hides polylogarithmic factors of $\|A\|$, $\|A\|_{\text{sc},p}$, \mathcal{C} , d , N , $p = \frac{2}{L}$.

The idea of the proof is to separate the function class into two components by applying a tunable thresholding operator in the singular values of the matrix A . A proof sketch is provided in subsection 3.2.1 below and the full proof is included in the appendix. Our full results in the appendix (cf. Theorem E.2) incorporate a finer dependency on architectural quantities such as the number of classes through a more refined analysis over different norms at each activation space, though we omit such subtleties from the main paper to reduce confusion. Assuming $L \gg 1$ and $\mathcal{L}_1, \mathfrak{B} \in O(1)$, $\|B_\ell\| = 1 \forall \ell$, our Theorem 3.1 scales like $\tilde{O}\left(\frac{\sum_\ell \|B_\ell\|_{\text{Fr}}^2 [\mathcal{C} + d]}{L}\right)$ in sample complexity, whereas the known result (1.2) scales as $\tilde{O}\left(\frac{L^3 \min(\mathcal{C}, d) \sum_\ell \|B_\ell\|_{\text{Fr}}^2}{L}\right)$ in [3]. Thus, the simplified analysis in the linear case removes a factor of L^3 , which allows taking the limit as $L \rightarrow \infty$: in this case, our bound behaves more and more like a parameter-counting bound since $\frac{\sum \|B_\ell\|_{\text{Fr}}^2}{L} \rightarrow \text{rank}(A)$. On the other hand, the dependency on the input space dimension d is absent in (1.2). Whilst it may be possible to improve the dependency on d in Theorem 3.1 assuming the input data lies on a low-dimensional subspace, this would require significant modifications to the proofs, and is best left to future work.

3.2 Fully Connected Neural Networks

We consider neural networks of the following form

$$x \rightarrow F_A(x) := A_L \sigma_L(A^{L-1} \sigma_{L-1}(\dots \sigma_1(A^1 x) \dots)), \quad (3.6)$$

where the matrices $A^\ell \in \mathbb{R}^{w_\ell \times w_{\ell-1}}$ are the weight matrices, $A = (A^L, A^{L-1}, \dots, A^1)$ denotes the set of weight matrices considered together, w_ℓ denotes the width at layer ℓ and $\sigma_1, \dots, \sigma_L$ denote elementwise activation functions with Lipschitz constants ρ_ℓ for $\ell = 1, \dots, L$. For any pair of layers indices $\ell_1 < \ell_2$ we also use the notation $F^{\ell_1 \rightarrow \ell_2}$ for the function defined by $F^{\ell_1 \rightarrow \ell_2}(x) = A_{\ell_2} \sigma_{\ell_2}(A^{\ell_2-1} \sigma_{\ell_2-1}(\dots \sigma_{\ell_1}(A^{\ell_1} x) \dots))$. The result below follows from Theorem E.2 in \tilde{O} notation:

Theorem 3.2. Fix reference matrices M_1, \dots, M_L . For every $\delta > 0$, w.p. $\geq 1 - \delta$, the following generalization bound holds simultaneously over all values of $p_\ell \in [0, 2]$ for $\ell = 1, \dots, L$:

$$\text{GAP} \leq \tilde{O} \left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{B} \sqrt{\frac{L^3}{N}} + \mathcal{B} [L \mathcal{L}_1]^{\frac{p}{2+p}} \sqrt{\frac{L}{N}} \mathcal{R}_{F_A} \right), \quad \text{where} \quad (3.7)$$

$$\mathcal{R}_{F_A} := \left[\sum_{\ell=1}^L \left[\mathfrak{B} \prod_{i=1}^L \rho_i \|A_i\| \right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\|A_\ell - M_\ell\|_{\text{sc}, p_\ell}^{p_\ell}}{\|A_\ell\|^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [w_\ell + w_{\ell-1}]^{1+\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}}$$

and the \tilde{O} absorbs polylog factors of $\mathfrak{B}, \bar{W}, L \max_i \|A_i\|, \mathcal{L}_1$ and N . Here, $p := \max_{\ell=1}^L p_\ell$ denotes the maximum of all the indices p_ℓ and $\mathfrak{B} := \sup_{i=1}^N \|x_i\|$ is the maximum input L^2 norm.

Next, using Loss Augmentation [5, 6], we obtain the following extension where norm-based terms are replaced by milder analogues relying on empirical estimates of intermediary activations. Since the product of spectral norms estimate $\|x\| \prod_{i=1}^\ell \|A^i\|$ is usually large compared to the activation norm $\|F^{0 \rightarrow \ell}(x)\|$, this can often lead to substantial numerical improvements, as seen in Section D.1.

Theorem 3.3 (Cf. Theorem E.6). Fix reference matrices M_1, \dots, M_L . For every $\delta > 0$, w.p. $\geq 1 - \delta$, the following bound holds simultaneously over all values of $p_\ell \in [0, 2]$ for $\ell = 1, \dots, L$:

$$\text{GAP} \leq \tilde{O} \left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{B} \sqrt{\frac{L^3}{N}} + \mathcal{B} [L \mathcal{L}_1]^{\frac{p}{2+p}} \sqrt{\frac{L}{N}} \mathcal{R}_{F_A} \right), \quad (3.8)$$

where $\mathfrak{B}_{\ell-1, A} := \max(\max_{i \leq N} \|F_{A^1, \dots, A^{\ell-1}}(x_i)\|, 1)$,

$$\mathcal{R}_{F_A}^{\text{emp}} := \left[\sum_{\ell=1}^L \left[\mathfrak{B}_{\ell-1, A} \prod_{i=\ell}^L \rho_i \|A_i\| \right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\|A_\ell - M_\ell\|_{\text{sc}, p_\ell}^{p_\ell}}{\|A_\ell\|^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [w_\ell + w_{\ell-1}]^{1+\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}},$$

and the \tilde{O} notation absorbs polylogarithmic factors⁶ of $\mathfrak{B}_{\ell-1, A}, \bar{W}, L, \max_i \|A_i\|, \mathcal{L}_1$ and N . Here, $p := \max_{\ell=1}^L p_\ell$ is the maximum of index p_ℓ and $\mathfrak{B} := \sup_{i=1}^N \|x_i\|$ is the maximum input L^2 norm.

Theorem 3.2 holds for all values of p_ℓ simultaneously: they can be optimized after training. For $p_\ell = 0 \quad \forall \ell$, Thm 3.3 yields a sample complexity of $\tilde{O} \left(L^3 + L \sum_{\ell=1}^L [w_\ell + w_{\ell-1}] \text{rank}(A_\ell) \right)$. However, the additive term of L^3 can easily be removed with a simpler argument dedicated to the situation where $p_\ell = 0$ (cf. Thm. E.8). Thus, our results provide a parametric complexity estimate for neural networks whose weight matrices satisfy a low rank property: for a neural network with fixed width \bar{W} , the sample complexity is $\tilde{O}(\bar{W} L^2 r)$ rather than $\tilde{O}(\bar{W}^2 L^2)$ in a pure parameter counting bound: when applying Corollary 3.3 (or Thm. E.8) to a network with fixed width \bar{W} , each weight matrix $A_\ell \in \mathbb{R}^{\bar{W} \times \bar{W}}$ only contributes $\tilde{O}(\bar{W} r)$ ‘parameters’ to the sample complexity, instead of the full \bar{W}^2 . Theorem 3.2 further incorporates *approximate low rank structure*. Indeed, in equation (3.7), the indices p_ℓ interpolate between the parametric and non parametric regimes.

When p increases, the term $\frac{\|A_\ell\|_{\text{sc}, p_\ell}^{p_\ell}}{\|A\|^{p_\ell}}$ captures low-rank structure inherent in very small singular values which are not quite equal to zero. However, the threshold p_ℓ must be carefully tuned due to a tradeoff with the scaling factor of $\left[\mathfrak{B} \prod_{i=1}^L \rho_i \|A_i\| \right]^{\frac{2p_\ell}{p_\ell+2}}$. As a partial limitation, we note that the bound does not exactly coincide with the norm-based result (1.3) when p is set to 2, since there is always a parametric dependence on the input dimension from the term $[\mathcal{C} + d]^{\frac{2}{p+2}}$: whilst the bound is an interpolation between norm-based and parameter counting bounds, it maintains a slight bias towards parameter counting. However, this is achieved without being uniformly inferior

⁶In this paper, our \tilde{O} notation absorbs factors of $\max_{\ell=1}^L \|A_\ell\|$, which is stricter than the alternative of accepting polylogarithmic factors of $\prod_{\ell=1}^L \|A_\ell\|$. In particular, one of the multiplicative factors of L in our bound arises from a logarithmic factor in $\prod_{\ell=1}^L \|A_\ell\|$. We find this approach more natural as it makes the full-rank parameter counting complexity $\tilde{O}(\mathcal{W}L) = \tilde{O}(\bar{W}L^2)$ match classic *lower-bounds* on VC dimension [7, 82].

to either. To see this, consider the one-layer case in an idealized situation where all the weights take binary values $\in \{1, -1\}$. Here, the norm-based bound (1.3) yields a sample complexity of $\tilde{O}(\mathcal{C}d)$, which agrees with parameter-counting. Theorem 3.2 also yields a sample complexity of $\tilde{O}\left(\mathcal{M}^{\frac{2p}{p+2}} \min(\mathcal{C}, d)^{\frac{p}{2+p}} \max(\mathcal{C}, d)^{\frac{2}{2+p}}\right) = \tilde{O}\left(\sqrt{\mathcal{C}d} \|A\|_{\text{Fr}}\right) = \tilde{O}(\mathcal{C}d)$.

3.2.1 Proof Sketches

To highlight our proof techniques, we provide a proof sketch for simplified versions of our results.

Proposition 3.4 (Simplified form of Proposition F.2: covering number bound for one layer). *Consider the following function class of linear maps from $\mathbb{R}^d \rightarrow \mathbb{R}^m$: $\mathcal{F}^p := \{M \in \mathbb{R}^{m \times d}, \|M\|_p \leq \mathcal{M}; \|M\| \leq s\}$. For any $\epsilon > 0$ and for any dataset $x_1, \dots, x_N \in \mathbb{R}^d$ such that $\|x_i\| \leq B$ for all i , we have the following bound on the L^∞ covering number (cf Prop. F.2 for formal definition) of the class:*

$$\log(\mathcal{N}_\infty(\mathcal{F}_r^p, \epsilon)) \lesssim [m + d] \left[\frac{\mathcal{M}B}{\epsilon} \right]^{\frac{2p}{p+2}} \log\left(\frac{mdN\mathcal{M}}{\epsilon}\right).$$

Proof Sketch. For simplicity we assume $B = 1$ in the sketch. For any matrix $M \in \mathcal{F}^p$, write $\sum_{i=1}^{\min(m,d)} \rho_i u_i v_i^\top$ for its singular value decomposition (with the singular values ordered in (any) decreasing order, including zeros). For a threshold τ to be determined later, we decompose every $M \in \mathcal{F}^p$ into $M = M_1 + M_2$ where $M_1 = \sum_{i \leq T} \rho_i u_i v_i^\top$ and $M_2 = \sum_{i \geq T+1} \rho_i u_i v_i^\top$ where T is the last index such that $\rho_T > \tau$. By Markov's inequality, since $\|M\|_p^p = \sum_i \rho_i^p \leq \mathcal{M}^p$, we have $\text{rank}(M_1) \leq \tau \leq \frac{\mathcal{M}^p}{\tau^p}$. In addition, the spectral norm of M_1 is bounded as follows: $\|M_1\| = \rho_1 = \|M\| \leq \|M\|_{\text{sc},p} \leq \mathcal{M}$. Thus, M_1 belongs to the set $\mathcal{F}_1^p := \{Z \in \mathcal{F}^p : \text{rank}(Z) \leq \frac{\mathcal{M}^p}{\tau^p}, \|Z\| \leq \mathcal{M}\}$ (a function class with few parameters its members are *low-rank*). Furthermore, it is clear that

$$\|M_2\|_{\text{Fr}}^2 = \sum_{k=T+1}^{\min(m,d)} \rho_k^2 \leq \tau^2 \min(m, d).$$

Therefore M_2 belongs to the class $\mathcal{F}_2^p := \{Z \in \mathcal{F}^p : \|Z\|_{\text{Fr}}^2 \leq \tau^2 \min(m, d)\}$ (whose members have *small norms*). By a parameter counting argument, we can bound the L^∞ covering number of \mathcal{F}_1^p :

$$\mathcal{N}_\infty(\mathcal{F}_1^p, \epsilon/2) \lesssim [m + d] [\text{rank}] \log\left(\frac{\mathcal{M}}{\epsilon}\right) \lesssim [m + d] \frac{\mathcal{M}^p}{\tau^p} \log\left(\frac{\mathcal{M}}{\epsilon}\right).$$

By classic norm-based arguments (Thm. 4 in [83]), we can bound the covering number of \mathcal{F}_2^p as:

$$\log(\mathcal{N}(\mathcal{F}_2^p, \epsilon/2)) \lesssim [\text{max Frob norm}] \log\left(\frac{\tau^2 mdN}{\epsilon}\right) = \frac{\tau^2 \min(m, d)}{\epsilon^2} \log\left(\frac{\tau^2 mdN}{\epsilon}\right).$$

Combining both bounds gives $\log(\mathcal{N}_\infty(\mathcal{F}_r^p)) \leq \log(\mathcal{N}(\mathcal{F}_1^p, \epsilon/2)) + \log(\mathcal{N}(\mathcal{F}_2^p, \epsilon/2))$

$$\lesssim \frac{\tau^2 \min(m, d)}{\epsilon^2} \log\left(\frac{\tau^2 mdN}{\epsilon}\right) + [m + d] \frac{\mathcal{M}^p}{\tau^p} \log\left(\frac{\mathcal{M}}{\epsilon}\right)$$

Setting $\tau = \mathcal{M}^{\frac{p}{p+2}} \epsilon^{\frac{2}{p+2}}$ gives $\log(\mathcal{N}_\infty(\mathcal{F}_r^p)) \lesssim [m + d] \left[\frac{\mathcal{M}}{\epsilon} \right]^{\frac{2p}{p+2}} \log\left(\frac{mdN\mathcal{M}}{\epsilon}\right)$ as expected. \square

We then show how to extend the proof to the two-layer case. Here, we consider networks of the form $F_A : x \rightarrow A^2 \sigma(A^1 x)$ where σ is an elementwise 1-Lipschitz activation function (e.g. ReLU) and $A_1 \in \mathbb{R}^{m \times d}, A^2 \in \mathbb{R}^{C \times m}$ are weight matrices. We fix $p_1 = p_2 = p \in [0, 2]$ and $s_1, s_2, \mathcal{M}_1, \mathcal{M}_2 > 0$, and let \mathcal{F} denote the class of such networks which further satisfy for all $i \in \{1, 2\}$: $\|A^i\| \leq s_i$; $\|A^i\|_p \leq \mathcal{M}_i$ (or, by abuse of notation, the corresponding matrices (A^2, A^1)). For simplicity, the loss $\ell : \mathbb{R}^C \times [C] \rightarrow \mathbb{R}^+$ is assumed to be 1-Lipschitz w.r.t. the L^∞ norm.

Proposition 3.5 (Simplified form of Proposition E.1). *Suppose we are given an i.i.d. training set $(x_1, y_1), \dots, (x_N, y_N)$ from a joint distribution over $\mathbb{R}^d \times [C]$ such that $\|x_i\| \leq B$ w.p. 1. For any $\delta > 0$, w.p. $\geq 1 - \delta$, the generalization gap of any network $F_A \in \mathcal{F}$ is bounded by:*

$$\tilde{O}\left(\left[s_1 s_2\right]^{\frac{p}{2+p}}\left[r_1[m+d]^{1+\frac{p}{p+2}}+r_2[m+C]\right]^{\frac{1}{2}} \frac{1}{\sqrt{N}}+\sqrt{\frac{\log (1 / \delta)}{N}}\right), \quad(3.9)$$

where for $i=1,2, r_i:=\left[\frac{\mathcal{M}_i^p}{s_i^p}\right]^{\frac{2}{p+2}}$ is a soft analogue of the rank and the \tilde{O} notation incorporates polylogarithmic factors of $N, m, d, C, s_1, s_2, \mathcal{M}_1, \mathcal{M}_2$.

Proof Sketch. Since the loss function is 1 Lipschitz, we only need to find an L^∞ cover of \mathcal{F} , i.e., a set $\mathcal{C} \subset \mathcal{F}$ such that for any $A = (A^1, A^2) \in \mathcal{F}$, there exists a cover element $(\bar{A}^1, \bar{A}^2) \in \mathcal{C}$ such that

$$\forall i, \|F_A(x_i) - F_{\bar{A}}(x_i)\|_\infty \leq \epsilon$$

(here, recall $F_A(x_i) \in \mathbb{R}^C$ is a vector of scores for all classes). We achieve this by adapting standard chaining techniques (see [3,5]), with the caveat that we need to change the norm of the cover at the intermediary layer, incurring a factor of \sqrt{m} . More precisely, let $\mathcal{F}^1, \mathcal{F}^2$ denote the set of matrices $A^1 \in \mathcal{R}^{m \times d}$ and $A^2 \in \mathcal{R}^{C \times m}$ satisfying the relevant constraints above. First, we can apply Proposition 3.4 above, with ϵ set to $\epsilon_1 := \epsilon/[2\sqrt{m}s_2]$, to achieve a $\mathcal{C}_1 \subset \mathcal{F}_1$ such that for all $A \in \mathcal{F}$ there exists a $\bar{A} \in \mathcal{C}$ such that $\forall i, \|\sigma(A^1(x_i)) - \sigma(\bar{A}^1(x_i))\| \leq \|A^1 x_i - \bar{A}^1 x_i\| \leq \sqrt{m} \leq \|A^1 x_i - \bar{A}^1 x_i\| \leq \sqrt{m} \epsilon_1 = \frac{\epsilon}{2s_2}$, and $\log(|\mathcal{C}|) \in \tilde{O}\left([m+d]\left[\frac{\mathcal{M}_1 B}{\epsilon_1}\right]^{\frac{2p}{p+2}}\right) = \tilde{O}\left([m+d]\left[s_1 s_2 B\right]^{\frac{2p}{p+2}}\left[\frac{\mathcal{M}_1}{s_2 \epsilon}\right]^{\frac{2p}{p+2}}\right)$. Similarly, for any $\bar{A}^1 \in \mathcal{C}_1$, $\|\sigma(\bar{A}^1 x_i)\| \leq s_1 B \quad \forall i$, and therefore another application of Prop. 3.4 with B replaced by $\tilde{B} s_1 B$ and ϵ replaced by $\epsilon_2 := \epsilon/2$ yields a cover $\mathcal{C}_2(\bar{A}^1)$ with size $\log(|\mathcal{C}_2(\bar{A}^1)|) \in \tilde{O}\left(\left[\frac{\mathcal{M}_2[B s_1]}{\epsilon_2}\right]^{\frac{2p}{p+2}}\right) = \tilde{O}\left([B s_1 s_2]\left[\frac{\mathcal{M}_2}{\epsilon s_2}\right]^{\frac{2p}{p+2}}\right)$.

The cover $\mathcal{C} := \bigcup_{\bar{A}^1 \in \mathcal{C}_1} \mathcal{C}_2(\bar{A}^1)$ is now an ϵ cover of \mathcal{F} . Indeed, for any $(A^1, A^2) \in \mathcal{F}$, we can define \bar{A}^1 to be the cover element associated to A^1 in \mathcal{C}_1 and subsequently \bar{A}^2 to be the cover element in $\mathcal{C}_2(\bar{A}^1)$ associated to A^2 , which yields for any $i \leq N$:

$$\begin{aligned} \|A^2 \sigma(A^1 x_i) - \bar{A}^2 \sigma(\bar{A}^1 x_i)\|_\infty &\leq \|A^2 \sigma(A^1 x_i) - A^2 \sigma(\bar{A}^1 x_i)\|_\infty + \|A^2 \sigma(\bar{A}^1 x_i) - \bar{A}^2 \sigma(\bar{A}^1 x_i)\|_\infty \\ &\leq \epsilon_2 + \|A^2\| \|\sigma(A^1(x_i)) - \sigma(\bar{A}^1(x_i))\| \leq \epsilon/2 + s_2 \epsilon/2s_2 = \epsilon. \end{aligned}$$

Furthermore, the resulting cover has cardinality bounded as

$$\log(|\mathcal{C}|) \in \tilde{O}\left(\frac{[B s_1 s_2]^{\frac{2p}{p+2}}}{\epsilon}\left[[m+d]^{1+\frac{p}{p+2}}\left[\frac{\mathcal{M}_1}{\epsilon s_1}\right]^{\frac{2p}{p+2}}+[m+C]\left[\frac{\mathcal{M}_2}{\epsilon s_2}\right]^{\frac{2p}{p+2}}\right]\right),$$

which yields eq (3.9) after substituting for r_i and calculations based on Dudley's entropy integral. \square

Thm E.2 is then derived from a union bound to ensure uniformity over $\mathcal{M}_1, \mathcal{M}_2$ etc. This step is standard for $s_1, s_2, \mathcal{M}_1, \mathcal{M}_2$ but involves a more tedious continuity argument for the parameter p .

3.3 Convolutional Neural Networks (CNNs)

In this section, we present an extension of our results to Convolutional Neural Networks (CNNs). Since CNNs are a well-known class of models, we leave a more thorough description of our notation for the forward pass to Appendix G. We denote the filters of each layer by A_ℓ (which incorporates each weight only once), whilst we use $\text{op}(A_\ell)$ to denote matricization of the layer, i.e., the matrix which represents the linear operation performed by the convolution at layer ℓ with the weights A_ℓ . We use U_ℓ for the number of channels and d_ℓ to refer to the dimension of the input patches. This includes both spatial and channel dimensions: if the input is an RGB image and the filter has spatial dimension $(2, 2)$, $d_0 = 2 \times 2 \times 3 = 12$. Thus, the number of trainable parameters at layer ℓ is $d_{\ell-1} \times U_\ell$, and write w_ℓ for the spatial dimension at layer ℓ (for instance, a 28×28 image will have $w_0 = 28 \times 28 = 784$). Thus, $U_\ell \times w_\ell$ is the number of preactivations at layer ℓ .

Theorem 3.6. Fix reference matrices M_1, \dots, M_L . For every $\delta \in (0, 1)$, w.p. $\geq 1 - \delta$, the following generalization bound holds simultaneously over all values of $0 \leq p_\ell \leq 2$ for $\ell = 1, \dots, L$:

$$\text{GAP} \leq \tilde{O} \left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{B} \sqrt{\frac{L^3}{N}} + \mathcal{B} [L L_1]^{\frac{p}{2+p}} \sqrt{\frac{L}{N}} \mathcal{R}_{F_A}^{\mathfrak{C}} \right), \quad \text{where} \quad (3.10)$$

$$\mathcal{R}_{F_A}^{\mathfrak{C}} := \left[\sum_{\ell=1}^L \left[\mathfrak{B} \prod_{i=1}^L \rho_i \|\text{op}(A_i)\| \right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\|A_\ell - M_\ell\|_{\text{sc}, p_\ell}^{p_\ell}}{\|\text{op}(A_\ell)\|^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [U_\ell + d_{\ell-1}] W_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}},$$

$\mathfrak{B} := \sup_{i=1}^N \|x_i\|$ and the \tilde{O} notation absorbs multiplicative factors of logarithms of the quantities $\mathfrak{B}, \bar{W}, \mathcal{A}, \mathcal{W}, L, \max_i \|\text{op}(A_i)\|, L_1$ and N . Here, $W_\ell := U_\ell \times w_\ell$ (for $\ell \neq L$) and $W_L := 1$.

This result incorporates weight sharing in the same sense as other CNN bounds [4, 16]; the term $\|A_\ell - M_\ell\|_{\text{sc}, p_\ell}^{p_\ell}$ only involves the weights once, irrespective of the number of patches.

Theorem 3.7 (Cf. Theorem G.10). For every $\delta \in (0, 1)$, with probability greater than $1 - \delta$, the following generalization bound holds simultaneously over all values of $0 \leq p_\ell \leq 2$ for $\ell = 1, \dots, L$:

$$\text{GAP} \leq \tilde{O} \left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{B} \sqrt{\frac{L^3}{N}} + \mathcal{B} [L L_1]^{\frac{p}{2+p}} \sqrt{\frac{L}{N}} \mathcal{R}_{F_A}^{\text{emp}, \mathfrak{C}} \right), \quad \text{where}$$

$$\mathcal{R}_{F_A}^{\text{emp}, \mathfrak{C}} := \left[\sum_{\ell=1}^L \left[\mathfrak{B}_{\ell-1, A}^{\mathfrak{C}} \prod_{i=\ell}^L \rho_i \|\text{op}(A_i)\| \right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\|A_\ell - M_\ell\|_{\text{sc}, p_\ell}^{p_\ell}}{\|\text{op}(A_\ell)\|^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [U_\ell + d_{\ell-1}] W_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}},$$

$\mathfrak{B}_{\ell-1, A}^{\mathfrak{C}} := \max(\max_{i \leq N, o} \|[F_{A^1, \dots, A^{\ell-1}(x_i)}]_{S_{\ell-1, o}}\|, 1)$ denotes the maximum norm of a convolutional patch at layer $\ell - 1$ and the \tilde{O} notation absorbs multiplicative factors of logarithms of the quantities $\mathfrak{B}, \bar{W}, \mathcal{A}, \mathcal{W}, L, \max_i \|\text{op}(A_i)\|, L_1, N$. Here, $W_\ell := U_\ell \times w_\ell$ (for $\ell \neq L$) and $W_L := 1$.

Compared to [4], Thm. 3.7 incorporates a low-rank structure: each layer's contribution to sample complexity is reduced from $L \times U_\ell \times d_\ell$ (the number of parameters in the layer) to $L[U_\ell + d_\ell]\text{rank}(A_\ell)$ (U_ℓ is the number of channels at layer ℓ and $d_{\ell-1}$ is the dimension of patches at layer $\ell - 1$).

4 Conclusion and Future Directions

For linear networks, deep neural networks and convolutional neural networks, we have shown generalization bounds which capture the implications of approximate low-rank structure in weight matrices in terms of sample complexity. The bounds are expressed in terms of the Schatten p quasi norms of the weight matrices, and are, to the best of our knowledge, the first bounds for deep learning which incorporate both norm-based and parameter counting qualities. When $p \rightarrow 0$, the bounds behave like a parameter-counting sample complexity of $\tilde{O}(L^3 + L \sum_{\ell=1}^L [w_\ell + w_{\ell-1}]\text{rank}(A_\ell))$, taking into account (exact) low-rank structure in the weight matrices. For more moderate values of p , the bounds incorporate norm-based quantities and incorporate *approximate* low-rank structure. Even in the case of linear maps, Subsection 3.1 provides original bounds for multi-class classification in the presence of *low-rank structure over the classes*, offering an alternative to classic results such as [65].

A limitation of our work is that we require the *weight matrices* to be low-rank, whereas most research on neural rank collapse [21, 27] demonstrates instead that the *activations* at each layer are low-rank. The two are very closely connected, but are not completely equivalent: it is possible that capturing the effect of low-rank *activations* could lead to further improvements and insights. Crucially, we believe a substantial modification of our proofs could tackle such situations. However, this would require a much more challenging network-wide approach to the simultaneous tuning of the parametric interpolation thresholds and is left to future work. Like other uniform convergence results which are agnostic to the training procedure, our bounds are vacuous at for large-scale networks in the absence of aggressive bound optimization, a limitation which may be addressed in future work by better controlling the norm-based components during training, manually truncating the ranks or incorporating data-dependent priors with a Bayesian approach. Lastly, we note that the low-rank assumption requires overparametrization and is not always satisfied for smaller architectures or layers (cf. Appendix D). Finally, other tantalizing future directions include proving fast rates in N and the extension of our work to other architectures such as Resnets and transformers.

Acknowledgements

Rodrigo Alves thanks Recombee for their support. The work of Yunwen Lei is partially supported by the Research Grants Council of Hong Kong [Project Nos. 17302624]. We thank the anonymous reviewers, Felix Biggs and Nadav Timor for their comments on our manuscript.

References

- [1] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30:6240–6249, 2017.
- [2] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. 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 5947–5956. Curran Associates, Inc., 2017.
- [3] 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.
- [4] Philip M. Long and Hanie Sedghi. Size-free generalization bounds for convolutional neural networks. In *International Conference on Learning Representations*, 2020.
- [5] Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9725–9736. Curran Associates, Inc., 2019.
- [6] Vaishnav Nagarajan and J. Zico Kolter. Deterministic pac-bayesian generalization bounds for deep networks via generalizing noise-resilience. *CoRR*, abs/1905.13344, 2019.
- [7] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK, 1999.
- [8] Vaishnav Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [9] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2017.
- [10] Giacomo De Palma, Bobak Kiani, and Seth Lloyd. Random deep neural networks are biased towards simple functions. *Advances in Neural Information Processing Systems*, 32, 2019.
- [11] Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. In *International Conference on Learning Representations*, 2020.
- [12] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grunwald, 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.
- [13] 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.
- [14] Pedro Cisneros-Velarde, Zhijie Chen, Sanmi Koyejo, and Arindam Banerjee. Optimization and generalization guarantees for weight normalization, 2025.
- [15] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

[16] Antoine Ledent, Waleed Mustafa, Yunwen Lei, and Marius Kloft. Norm-based generalisation bounds for deep multi-class convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8279–8287, May 2021.

[17] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[18] Sanjeev Arora, S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *ICML*, 2019.

[19] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. The curse of depth in kernel regime. In Melanie F. Pradier, Aaron Schein, Stephanie Hyland, Francisco J. R. Ruiz, and Jessica Z. Forde, editors, *Proceedings on "I (Still) Can't Believe It's Not Better!" at NeurIPS 2021 Workshops*, volume 163 of *Proceedings of Machine Learning Research*, pages 41–47. PMLR, 13 Dec 2022.

[20] Philip M. Long. Properties of the after kernel. *arXiv preprint arXiv:2105.10585*, 2021.

[21] Arthur Jacot. Bottleneck structure in learned features: Low-dimension vs regularity tradeoff. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[22] Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions, 2023.

[23] Zihan Wang and Arthur Jacot. Implicit bias of SGD in $\$1_2\$$ -regularized linear DNNs: One-way jumps from high to low rank. In *The Twelfth International Conference on Learning Representations*, 2024.

[24] Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[25] Seijin Kobayashi, Yassir Akram, and Johannes Von Oswald. Weight decay induces low-rank attention layers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[26] Suzanna Parkinson, Greg Ongie, and Rebecca Willett. Relu neural networks with linear layers are biased towards single-and multi-index models. *SIAM Journal on Mathematics of Data Science*, 7(3):1021–1052, 2025.

[27] Laura Balzano, Tianjiao Ding, Benjamin D. Haeffele, Soo Min Kwon, Qing Qu, Peng Wang, Zhangyang Wang, and Can Yaras. An overview of low-rank structures in the training and adaptation of large models, 2025.

[28] Antoine Ledent and Rodrigo Alves. Generalization analysis of deep non-linear matrix completion. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26290–26360. PMLR, 21–27 Jul 2024.

[29] 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.

[30] Yuan Cao, Difan Zou, Yuanzhi Li, and Quanquan Gu. The implicit bias of batch normalization in linear models and two-layer linear convolutional neural networks. In *Annual Conference on Learning Theory (COLT)*, 2023.

[31] Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions, 2022.

[32] Tomer Galanti, Mengjia Xu, Liane Galanti, and Tomaso Poggio. Norm-based generalization bounds for compositionally sparse neural networks. In *Advances in Neural Information Processing Systems*, 2023.

- [33] Spencer Frei, Yuan Cao, and Quanquan Gu. Provable generalization of sgd-trained neural networks of any width in the presence of adversarial label noise. In *International Conference on Machine Learning*, pages 3427–3438. PMLR, 2021.
- [34] Henry Gouk, Timothy M Hospedales, and Massimiliano Pontil. Distance-based regularisation of deep networks for fine-tuning. *arXiv preprint arXiv:2002.08253*, 2020.
- [35] Stéphan Cléménçon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*.
- [36] Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. *Advances in Neural Information Processing Systems*, 33:21236–21246, 2020.
- [37] Yunwen Lei, Mingrui Liu, and Yiming Ying. Generalization guarantee of sgd for pairwise learning. *Advances in neural information processing systems*, 34:21216–21228, 2021.
- [38] Sanjeev Arora. A theoretical analysis for contrastive representation learning. In *International Conference on Machine Learning*, 2019.
- [39] Yunwen Lei, Tianbao Yang, Yiming Ying, and Ding-Xuan Zhou. Generalization analysis for contrastive representation learning. In *International Conference in Machine Learning*, 2023.
- [40] Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, 2021.
- [41] Nong Minh Hieu, Antoine Ledent, Yunwen Lei, and Cheng Yeaw Ku. Generalization analysis for deep contrastive representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17186–17194, 2025.
- [42] Nong Minh Hieu and Antoine Ledent. Generalization analysis for supervised contrastive representation learning under non-IID settings. In *Forty-second International Conference on Machine Learning*, 2025.
- [43] Rodrigo Alves and Antoine Ledent. Context-aware representation: Jointly learning item features and selection from triplets. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6492–6502, 2025.
- [44] Naghmeh Ghanooni, Dennis Wagner, Waleed Mustafa, Anthony Widjaja Lin, Sophie Fellenz, and Marius Kloft. Spectral dynamics of contrastive learning with spurious features. In *High-dimensional Learning Dynamics 2025*, 2025.
- [45] Naghmeh Ghanooni, Waleed Mustafa, Yunwen Lei, Anthony Widjaja Lin, and Marius Kloft. Generalization bounds with logarithmic negative-sample dependence for adversarial contrastive learning. *Transactions on Machine Learning Research*, 2024.
- [46] Naghmeh Ghanooni, Barbod Pajoum, Harshit Rawal, Sophie Fellenz, Vo Nguyen Le Duy, and Marius Kloft. Multi-level supervised contrastive learning. *arXiv preprint arXiv:2502.02202*, 2025.
- [47] Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang. Towards the generalization of contrastive self-supervised learning. In *International Conference on Learning Representation*, 2023.
- [48] Xin Zou and Weiwei Liu. Generalization bounds for adversarial contrastive learning. *Journal of Machine Learning Research*, 2023.
- [49] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.
- [50] 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. *arXiv preprint arXiv:1804.05862*, 2018.

[51] 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.

[52] Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G Wilson. Pac-bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

[53] Waleed Mustafa, Philipp Liznerski, Antoine Ledent, Dennis Wagner, Puyu Wang, and Marius Kloft. Non-vacuous generalization bounds for adversarial risk in stochastic neural networks. In *International conference on artificial intelligence and statistics*, pages 4528–4536. PMLR, 2024.

[54] Felix Biggs and Benjamin Guedj. Non-vacuous generalisation bounds for shallow neural networks. In *International conference on machine learning*, pages 1963–1981. PMLR, 2022.

[55] Felix Biggs and Benjamin Guedj. On margins and derandomisation in pac-bayes. In *International Conference on Artificial Intelligence and Statistics*, pages 3709–3731. PMLR, 2022.

[56] Pierfrancesco Beneventano and Blake Woodworth. Gradient descent converges linearly to flatter minima than gradient flow in shallow linear networks, 2025.

[57] Tomer Galanti, Zachary S Siegel, Aparna Gupte, and Tomaso Poggio. Characterizing the implicit bias of regularized sgd in rank minimization. *arXiv e-prints*, pages arXiv–2206, 2022.

[58] Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *arXiv preprint arXiv:2103.10427*, 2021.

[59] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. volume 30, 2017.

[60] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

[61] Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. In *International Conference on Algorithmic Learning Theory*, pages 1429–1459. PMLR, 2023.

[62] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[63] Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A Poggio. Feature learning in deep classifiers through intermediate neural collapse. In *International conference on machine learning*, pages 28729–28745. PMLR, 2023.

[64] Andrea Pinto, Akshay Rangamani, and Tomaso A Poggio. On generalization bounds for neural networks with low rank layers. In Gautam Kamath and Po-Ling Loh, editors, *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pages 921–936. PMLR, 24–27 Feb 2025.

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

[66] Yann Guermeur. Lp-norm Sauer–Shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences*, 89:450 – 473, 2017.

[67] Khadija Musayeva, Fabien Lauer, and Yann Guermeur. Rademacher complexity and generalization performance of multi-category margin classifiers. *Neurocomputing*, 342:6 – 15, 2019. Advances in artificial neural networks, machine learning and computational intelligence.

[68] Mehryar Mohri and Andres Munoz Medina. Learning theory and algorithms for revenue optimization in second price auctions with reserve. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 262–270, Bejing, China, 22–24 Jun 2014. PMLR.

[69] Liang Wu, Antoine Ledent, Yunwen Lei, and Marius Kloft. Fine-grained generalization analysis of vector-valued learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10338–10346, 2021.

[70] Waleed Mustafa, Yunwen Lei, Antoine Ledent, and Marius Kloft. Fine-grained generalization analysis of structured output prediction. *arXiv preprint arXiv:2106.00115*, 2021.

[71] Zi-Hao Zhou, Jun-Jie Wang, Tong Wei, and Min-Ling Zhang. Weakly-supervised contrastive learning for imprecise class labels. *arXiv preprint arXiv:2505.22028*, 2025.

[72] Yifan Zhang and Min-Ling Zhang. Generalization analysis for multi-label learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60220–60243. PMLR, 21–27 Jul 2024.

[73] Jin Zhang, Ze Liu, Defu Lian, and Enhong Chen. Generalization error bounds for two-stage recommender systems with tree structure. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[74] Florian Graf, Sebastian Zeng, Bastian Rieck, Marc Niethammer, and Roland Kwitt. On measuring excess capacity in neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10164–10178. Curran Associates, Inc., 2022.

[75] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5793–5831. PMLR, 17–23 Jul 2022.

[76] Yufeng Zhang, Boyi Liu, Qi Cai, Lingxiao Wang, and Zhaoran Wang. An analysis of attention via the lens of exchangeability and latent variable models. *arXiv preprint arXiv:2212.14852*, 2022.

[77] Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2024.

[78] Andreas Maurer. A chain rule for the expected suprema of gaussian processes. In *Proceedings of the 27th Annual Conference on Learning Theory (COLT)*, pages 245–257. PMLR, 2014.

[79] Rina Foygel, Ruslan Salakhutdinov, Ohad Shamir, and Nathan Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. In *Advances in Neural Information Processing Systems*, pages 2133–2141, 2011.

[80] Nathan Srebro and Russ R Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2056–2064. Curran Associates, Inc., 2010.

[81] Ohad Shamir and Shai Shalev-Shwartz. Collaborative filtering with the trace norm: Learning, bounding, and transducing. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 661–678. PMLR, 2011.

[82] Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

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

[84] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018.

[85] Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Multi-class deep boosting. In *Advances in Neural Information Processing Systems*, pages 2501–2509, 2014.

[86] Yann Guermeur. Combining discriminant models with new multi-class svms. *Pattern Analysis & Applications*, 5(2):168–179, Jun 2002.

[87] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 10 2011.

[88] Yunwen Lei, Urun Dogan, Alexander Binder, and Marius Kloft. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[89] Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, and Tuo Zhao. On tighter generalization bound for deep neural networks: Cnns, resnets, and beyond. *arXiv preprint arXiv:1806.05159*, 2018.

[90] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In Peter Auer and Ron Meir, editors, *Learning Theory*, pages 545–560, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[91] Robert A Vandermeulen and Antoine Ledent. Beyond smoothness: Incorporating low-rank analysis into nonparametric density estimation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12180–12193. Curran Associates, Inc., 2021.

[92] Chen Xu, Zhouchen Lin, and Hongbin Zha. A unified convex surrogate for the schatten-p norm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[93] Fanhua Shang, Yuanyuan Liu, and James Cheng. Tractable and scalable schatten quasi-norm approximations for rank minimization. In *Artificial Intelligence and Statistics*, pages 620–629. PMLR, 2016.

[94] Fanhua Shang, Yuanyuan Liu, and James Cheng. Unified scalable equivalent formulations for schatten quasi-norms. *arXiv preprint arXiv:1606.00668*, 2016.

[95] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, pages 224–240, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

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

[97] Yann Guermeur. Rademacher complexity of margin multi-category classifiers. *Neural Computing and Applications*, 32:17995–18008, 2020.

[98] R.M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[99] Antoine Ledent, Rodrigo Alves, Yunwen Lei, and Marius Kloft. Fine-grained generalization analysis of inductive matrix completion. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25540–25552. Curran Associates, Inc., 2021.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss limitations of the results in multiple places in the main paper and appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Computing the bounds over many different runs is computationally prohibitive. In addition, similarly to norm-based bounds, our results demonstrate the bounds’ ability to capture the relationship between each trained network’s weight configurations and the test performance. This cannot be achieved if the results are averaged over many runs. We do run our experiments over many width configurations on both DNNs and CNNs for many width configurations and observe similar behavior. Lastly, the focus of our work is mostly theoretical, and our experimental setup is in line with comparable literature.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[NA\]](#)

Justification: this work is of purely theoretical nature and there are no negative possible societal consequences to consider.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Table of Notation

Notation	Meaning
$\ \cdot\ , \ \cdot\ _\sigma$	Spectral norm of a matrix
$\ \cdot\ , \ \cdot\ _2$	(L2) Norm of a vector
$\ \cdot\ _{\text{Fr}}$	Frobenius norm of a matrix
$\ \cdot\ _{\text{sc},p}$	Schatten p quasi norm of a matrix
$\ \cdot\ _*$	Nuclear norm
$\ Z\ _{2,1}$	$\sum_i \sqrt{\sum_j Z_{i,j}^2}$
\mathcal{C}	Number of classes
l	Loss function
L_l	Lipschitz constant of l (w.r.t. L^∞ norm)
\mathcal{B}	upper bound on l
$A_\ell \in \mathbb{R}^{w_\ell \times w_{\ell-1}}$	Weight matrix at layer ℓ
$M_\ell \in \mathbb{R}^{w_\ell \times w_{\ell-1}}$	Reference matrix at layer ℓ
σ_ℓ	(Elementwise) Activation function at layer ℓ
ρ_ℓ	Lipschitz constant of σ_ℓ
$F_A(x) :=$ $A_L \sigma_L(A^{L-1} \sigma_{L-1}(\dots \sigma_1(A^1 x) \dots))$	NN with weight matrices A_1, \dots, A_L (Cf. Equation (3.6))
$F_A^{\ell_1 \rightarrow \ell_2}(\tilde{x})$	$\sigma_{\ell_2}(A_{\ell_2} \sigma_{\ell_2-1}(\dots \sigma_{\ell+1}(A_{\ell+1} \tilde{x}) \dots))$ Map from layer ℓ_1 to layer ℓ_2
Architectural Quantities	
L	Depth
N	Number of samples
w_ℓ	Width at layer ℓ
\bar{W}	$\max_{\ell=0}^L w_\ell$ (Maximum width of the network)
\underline{w}_ℓ	$\min(w_\ell, w_{\ell-1})$
\bar{w}_ℓ	$[w_\ell + w_{\ell-1}]$
\tilde{w}_ℓ	$[w_\ell \underline{w}_\ell]$ if $\ell \neq L$ \underline{w}_L if $\ell = L$
d	input dimension
Function Classes	
$\mathcal{F}_{\mathcal{M}_2}^2$	$\{Z \in \mathbb{R}^{m \times d} : \ Z\ _{\text{Fr}} \leq \mathcal{M}\}$
\mathcal{F}_r^p	$\{Z \in \mathbb{R}^{m \times d} : \ Z\ _{\text{sc},p}^p \leq \mathcal{M}^p = r s^p; \ Z\ \leq s\}$
$\mathcal{E}_{r,s}$	$\{A \in \mathbb{R}^{m \times d} : \text{rank}(A) \leq r, \ A\ \leq s\}$
\mathcal{B}_ℓ	$\{A_\ell \in \mathbb{R}^{w_\ell \times w_{\ell-1}} : \ A_\ell - M_\ell\ _{\text{sc},p_\ell} \leq \mathcal{M}_\ell; \ A_\ell\ \leq s_\ell\}$
$\mathcal{F}_\mathcal{B}$	$\{F_A = F_{(A_1, \dots, A_L)} : A_\ell \in \mathcal{B}_\ell \quad \forall \ell \leq L\}$ (cf. Eq. (E.1))
$\mathcal{L}^{\lambda_{aug}}$	$= \{F_A = F_{(A_1, \dots, A_L)} : \ A_\ell - M_\ell\ _{\text{sc}} \leq \mathcal{M}_\ell \quad \ A\ \leq s_\ell \quad \forall \ell \leq L\}$
$\mathcal{L}^{\mathfrak{C}, \lambda_{aug}}$	$\{\lambda_{aug}(F_A, x, y) : F_A \in \mathcal{F}_\mathcal{B}\}$ $\{\lambda_{aug}(F_A, x, y) : F_A \in \mathcal{F}_\mathcal{B}^{\mathfrak{C}}\}$
Constraints/scaling factors	
p_ℓ	Schatten index for layer ℓ
$p = \max_\ell p_\ell$	Maximum Schatten index over all layers

s	Upper bound on $\ Z\ $
s_ℓ	Upper bound on $\ A_\ell\ $ (fully connected)
\mathcal{M}	Upper bound on $\ Z\ _{\text{sc},p}$
\mathcal{M}_ℓ	Upper bound on $\ A - M_\ell\ _{\text{sc},p_\ell}$
r	$\frac{\mathcal{M}^p}{s^p}$ (rank proxy)
r_ℓ	$\frac{\mathcal{M}_\ell^p}{s_\ell^p}$ (rank proxy)
\mathfrak{B}	$\sup_{i=1}^N \ x_i\ $
\tilde{r}_ℓ	a priori upper bound on $\text{rank}(A_\ell)$
$\mathfrak{B}_{\ell-1,A}$	$\max(\max_{i \leq N} \ F_{A^1, \dots, A^{\ell-1}}(x_i)\ , 1)$
Constants/log factors	
$\Gamma_{\mathcal{F}_r^p}$	$\left(\frac{16[\mathcal{M}^p + 1][b+1][1+s][m+d]}{\epsilon^{\frac{p}{p+2}}} + 7 \right) mN$
$\Gamma'_{\mathcal{F}_r^p}$	$\left(\frac{16[\mathcal{M}^p + 1][b+1][1+s][m+d]^2}{\epsilon^{\frac{p}{p+2}}} + 7 \right) mN$
$\Gamma_{\mathcal{F}_r^p, \ell}$	$\left[\left(\frac{16[\mathcal{M}_\ell^{p\ell} + 1][b \prod_{i=1}^\ell \rho_i s_i + 1][1+s_\ell][w_\ell + w_{\ell-1}]}{\epsilon^{\frac{p_\ell}{p_\ell+2}}} + 7 \right) w_\ell N \right]$
$\Gamma_{\mathcal{F}_r^p, L}$	$\left[\left(\frac{16[\mathcal{M}_L^{pL} + 1][b \prod_{i=1}^L \rho_i s_i + 1][1+\sqrt{w_L} s_L][w_L + w_{L-1}]}{\epsilon_L^{\frac{p_L}{p_L+2}}} + 7 \right) w_L N \right]$
$\bar{\Gamma}$	$\left[128 \prod_{i=1}^L [(\rho_i + 1)s_i + 1]^3 [b + 1][L_L L + 1] \bar{W}^4 N^2 \right]$
$\underline{\Gamma}$	$\left[[b + 1][L L_L + 1] \bar{W} N \prod_{i=1}^L [(\rho_i + 1)s_i + 1] \right]$ ($\bar{\Gamma} = 40 \underline{\Gamma}^6$, cf. eq (E.141))
γ_{FC}	$12 \left[[b + 1][L L_L + 1] \bar{W} N \prod_{i=1}^L [\rho_i \ A_i\ + 1] \right] [4\bar{W} + b \prod_i \rho_i \ A_i\]$
Θ_{\log}	$2 \log [4\bar{W} + b \prod_i \rho_i \ A_i\] \left[\log(b) + L \left[\sum_{\ell=1}^L \log(\ A_\ell\) + 2 \log(4\bar{W}) \right] \right]$
Γ_R	$[\mathfrak{B} + 2][L L_L + 1] \bar{W} N \prod_{i=1}^L [(\rho_i + 1)[\ A_i\ + 2] + 1]$
$\Theta_{\log}^{\lambda_{aug}}$	$2 \log [4\bar{W} + b \prod_i \rho_i \ A_\ell\] \times$ $\left[\log(b) + L \left[\sum_{\ell=1}^L \log(\ A_\ell\) + \sum_{\ell=1}^{L-1} \log(\mathfrak{B}_{\ell, A}) + 2 \log(4\bar{W}) \right] \right]$
Convolutional Neural Networks	
O_ℓ	Number of convolutional patches at layer ℓ
w_ℓ	Spatial dimension at layer ℓ
$O_{\ell-1}$	Number of spatial dimensions before pooling at layer ℓ
U_ℓ	Number of channel dimensions at layer ℓ
d_ℓ	Dimension of input patches at layer ℓ
$S^{\ell,o} \in [U_\ell] \times [w_\ell]$	o th Convolutional Patch at layer ℓ
$A_\ell \in \mathbb{R}^{U_\ell \times d_\ell}$	Weight matrix at layer ℓ
$M_\ell \in \mathbb{R}^{U_\ell \times d_\ell}$	Initialized weight matrix at layer ℓ
Λ_{A^ℓ}	Convolution operation associated to A_ℓ
σ_ℓ	Represented by the matrix $\text{op}(A_\ell)$ $: \mathbb{R}^{U_\ell \times O_\ell} \rightarrow \mathbb{R}^{U_\ell \times w_\ell}$
$F_{A^1, A^2, \dots, A^L}^{\mathfrak{C}}$	Activation function (including pooling) $(\sigma_L \circ \Lambda_{A^L} \circ \sigma_{L-1} \circ \Lambda_{A^{L-1}} \circ \dots \sigma_1 \circ \Lambda_{A^1})(x)$ (Cf. Equation (G.1))

\mathcal{A}	$\sum_{\ell=0}^L U_\ell \times O_{\ell-1}$ (Total number of preactivations)
\mathcal{W}	$\sum_{\ell=1}^L d_{\ell-1} \times U_\ell$ (Total number of parameters in the network)
$\mathcal{F}_{\mathcal{B}}^{\mathfrak{C}}$	$\left\{ F_A = F_{(A_1, \dots, A_L)}^{\mathfrak{C}} : \ A_\ell - M_\ell\ _{\text{sc}} \leq \mathcal{M}_\ell, \right. \\ \left. \ \text{op}(A_\ell)\ \leq s_\ell \quad \forall \ell < L, \ A_L^\top\ _{2,\infty} \leq s_L \right\}$
\underline{w}_ℓ	Spatial patch size at layer ℓ
\underline{B}	$\max_{o,i} \ (x_i)_{S^{0,o}}\ $ Maximum norm of a convolutional patch at input layer
Constants/log factors for CNNs	
$\gamma_{\mathfrak{C}}$	$12 \left[[\mathfrak{B} + 1][L L_l + 1] \mathcal{W} \mathcal{A} N \prod_{i=1}^L [\rho_i \ \text{op}(A_i)\ + 1] \right] \\ \times [4\bar{W} + \mathfrak{B} \prod_i \rho_i \ \text{op}(A_i)\]$
$\gamma_{\mathfrak{C}, \lambda_{aug}}$	$12 \left[[\mathfrak{B} + 1][L L_l + 1] \mathcal{W} \mathcal{A} N \prod_{i=1}^L [\rho_i \ \text{op}(A_i)\ + 1] \right] \times \\ [4\bar{W} + \mathfrak{B} \prod_i \rho_i \ \text{op}(A_i)\]$
$\Gamma_{\mathfrak{C}_r^p}$	$\left(\frac{16[\mathcal{M}^p + 1][b+1][1+s][U'+d]}{\epsilon^{\frac{p}{p+2}}} + 7 \right) U' NO$
$\Gamma_{\mathfrak{C}_r^p, \ell}^{\mathfrak{C}}$	$\left[\left(\frac{16[\mathcal{M}_{\ell}^{p_\ell} + 1][b \prod_{i=1}^{\ell} \rho_i s_i + 1][1 + \sqrt{A} s_\ell] \mathcal{W} \mathcal{A}}{\epsilon_{\ell}^{\frac{p_\ell}{p_\ell+2}}} + 7 \right) N \mathcal{A} \right]$
$\Gamma_{\mathfrak{C}_r^p}'$	$\left(\frac{16[\mathcal{M}^p + 1][b+1][1+s][U'+d]U'w'}{\epsilon^{\frac{p}{p+2}}} + 7 \right) U' NO$
$\Gamma_R^{\mathfrak{C}}$	$\left[[\mathfrak{B} + 2][L L_l + 1] \mathcal{W} \mathcal{A} N \prod_{i=1}^L [(\rho_i + 1)[\ A_i\ + 2] + 1] \right]$
$\Gamma_{\mathfrak{C}, \lambda_{aug}}$	$[b + 1][L(L_l + 1)) \mathcal{W} \mathcal{A} N \prod_{i=1}^L [(\rho_i + 1)s_i + 1]$

Table 1: Table of notations for quick reference

B Table of Comparison to Existing Literature

Linear Classification Linear Networks		
Proof Technique	Bound (excluding polylog factors)	Reference
Parameter Counting	$L_1 \sqrt{\frac{Cd}{N}}$	[4]
Norm Based	$L_1 \mathfrak{B} \sqrt{\mathcal{C} \frac{\ A\ _{\text{Fr}}^2}{N}}$	[66], [67], [84], [64]
Norm-based/peeling from Linear Networks	$L_1 \mathfrak{B} \sqrt{\frac{\mathcal{C} \min_{\ell}(\text{Rank}(B_{\ell}))}{N}} \prod_{\ell} \ B_{\ell}\ $	[64]
Norm-based	$L_1 \mathfrak{B} \sqrt{\frac{\ A\ _{\text{Fr}}^2}{N}}$	[65] Deduced from [3] or [1] cf. Corollary C.1
Norm-based	$L_1 \mathfrak{B} L^{\frac{3}{2}} \ A\ ^{\frac{L-1}{L}} \sqrt{\frac{\mathcal{C} \ A\ _{\text{sc}, \frac{2}{L}}^{\frac{2}{L}}}{N}}$	[64]
Parametric Interpolation	$\mathcal{B} L_1^{\frac{p}{2+p}} \sqrt{\frac{\mathfrak{B}^{\frac{2p}{2+p}} \ A\ _{\text{sc}, p}^{\frac{2p}{2+p}} [\mathcal{C} + d]}{N}}$ $\mathcal{B} \sqrt{\frac{\text{rank}(A)[\mathcal{C} + d]}{N}}$	cf. Theorem 3.1 $p = 0$ in above line
DNNs, norm-based		
Peeling	$L_1 \sqrt{\frac{\mathfrak{B}^2 L \prod_{\ell=1}^L \ A_{\ell}\ _{\text{Fr}}^2}{N}}$	[13]
Peeling	$\frac{L_1 \mathfrak{B} \max_{\ell} \frac{\ A_{\ell}^T\ _{2,1}}{\ A_{\ell}\ } \left[\prod_{\ell=1}^L \ A_{\ell}\ \right]}{N^{\frac{1}{2+3p}}} \left[\frac{L}{p} \right]^{\frac{1}{\frac{2}{3}+p}}$	[13]
Peeling	$L_1 \frac{C_1 \sqrt{w_1} \frac{\ A_1\ _{\text{Fr}}}{\ A_1\ } + \sum_{\ell=2}^L C_1^{L-\ell} C_2 \sqrt{w_{\ell} \tilde{r}_{\ell}} \frac{\ A_{\ell}\ _{\text{Fr}}}{\ A_{\ell}\ }}{\sqrt{N}} \times \prod_{\ell=1}^L \ A_{\ell}\ \mathfrak{B}$	[64]
(Peeling)	$L_1 C_1^L \mathfrak{B} \left[\prod_{\ell=1}^L \ A_{\ell}\ \right] L r \sqrt{\frac{W}{N}}$	Particular case of above
Pac Bayes	$L_1 \frac{L \sqrt{W}}{\sqrt{N}} \left(\prod_{\ell=1}^L \ A_{\ell}\ \right) \left(\sum_{\ell=1}^L \frac{\ A_{\ell}\ _{\sigma}^2}{\ A_{\ell}\ _{\sigma}^2} \right)^{\frac{1}{2}}$	[3]
Covering Numbers	$L_1 \frac{1}{\sqrt{N}} \prod_{\ell=1}^L \ A_{\ell}\ \left(\sum_{\ell=1}^L \frac{\ (A_{\ell} - M_{\ell})^T\ _{2,1}^{\frac{2}{3}}}{\ A_{\ell}\ ^{\frac{2}{3}}} \right)^{\frac{3}{2}}$	[1]
DNNs, Parameter Counting		
	$\mathcal{B} \sqrt{\frac{W S L}{N}}$	[4]
	$\mathcal{B} \sqrt{\frac{W L}{N}}$	[74]
	$\mathcal{B} \sqrt{\frac{L \left[\sum_{\ell=1}^L [w_{\ell} + w_{\ell-1}] \text{rank}(A_{\ell}) \right]}{N}}$	Theorem E.8

DNNs Parametric Interpolation		
	$\mathcal{B}[L L_1]^{\frac{p}{2+p}} \sqrt{\frac{L}{N}} \left[\sum_{\ell=1}^L \left[\mathfrak{B} \prod_{i=1}^L \rho_i \ A_i\ \right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\ A_\ell - M_\ell\ _{sc, p_\ell}^{p_\ell}}{\ A_\ell\ ^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [w_\ell + w_{\ell-1}]^{1+\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}}$	Theorem 3.2
Parametric Interpolation & Loss augmentation	$\mathcal{B}[L L_1]^{\frac{p}{2+p}} \left[\sum_{\ell=1}^L \left[\mathfrak{B}_{\ell-1, A} \prod_{i=\ell}^L \rho_i \ A_i\ \right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\ A_\ell - M_\ell\ _{sc, p_\ell}^{p_\ell}}{\ A_\ell\ ^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [w_\ell + w_{\ell-1}]^{1+\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}}$	Theorem 3.3
CNNs, Parameter Counting		
	$\mathcal{B} \sqrt{\frac{\mathcal{W}SL}{N}}$ $\mathcal{B} \sqrt{\frac{\mathcal{W}L}{N}}$ $\mathcal{B} \sqrt{\frac{L[\sum_{\ell=1}^L [U_\ell + d_{\ell-1}] \text{rank}(A_\ell)]}{N}}$	[4] [74] Corollary G.6
CNNs, Norm-based		
Covering Numbers	$\frac{\mathbb{L}_1 \prod_{\ell=1}^L \ \text{op}(A_\ell)\ }{\sqrt{N}} \left[\sum_{\ell=1}^L \left(\frac{U_\ell w_\ell \ [A_\ell - M_\ell]^\top \ _{2,1}}{\ \text{op}(A_\ell)\ } \right)^{\frac{2}{3}} \right]^{\frac{3}{2}}$	[16]
Peeling	$\frac{\mathbb{L}_1 \prod_{\ell=1}^L \ A_\ell\ _{\text{Fr}} \sqrt{L} \prod_{\ell=1}^L w_\ell B}{\sqrt{N}}$	[32]
CNNs Parametric Interpolation		
	$\mathcal{B}[L L_1]^{\frac{p}{2+p}} \sqrt{\frac{L}{N}} \left[\sum_{\ell=1}^L \left[\mathfrak{B} \prod_{i=1}^L \rho_i \ \text{op}(A_i)\ \right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\ A_\ell - M_\ell\ _{sc, p_\ell}^{p_\ell}}{\ \text{op}(A_\ell)\ ^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [U_\ell + d_{\ell-1}] W_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}}$	Theorem G.5 Note: $W_\ell := U_\ell \times w_\ell$ (for $\ell \neq L$) and $W_L := 1$.
Parametric Interpolation and Loss augmentation	$\mathcal{B}[L L_1]^{\frac{p}{2+p}} \left[\sum_{\ell=1}^L \left[\mathfrak{B}_{\ell-1, A}^c \prod_{i=\ell}^L \rho_i \ \text{op}(A_i)\ \right]^{\frac{2p_\ell}{p_\ell+2}} \times \left[\frac{\ A_\ell - M_\ell\ _{sc, p_\ell}^{p_\ell}}{\ \text{op}(A_\ell)\ ^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [U_\ell + d_{\ell-1}] W_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}}$	Theorem 3.7

Table 2: Summary of results and previous State of the Art. By convention, all results are expressed in \tilde{O} notation, which hides polylogarithmic factors of all relevant architectural ($\mathcal{W}, \mathcal{A}, \bar{W}, L$) and scaling ($\max_\ell \|A_\ell\|, \|A\|_{\text{sc}}$, etc.) quantities. In addition, we only consider the dominant Rademacher complexity term: for simplicity we ignore both the missing terms of $O\left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}}\right)$ and additional terms of the form $\tilde{O}\left(\sqrt{\frac{L}{N}}\right)$ which would occur from making the bounds post hoc w.r.t. to the norm constraints (some works do this explicitly whilst some do not).

C In-depth Comparison to Existing Works

In all the descriptions below, the \tilde{O} notation absorbs polylogarithmic factors in $\mathcal{B}, \mathcal{W}, \max_i \|A_\ell\|, N, L_1$ as well as the Lipschitz constant L_1 of the loss or the margin γ .

Recall that in [3], it was shown that

$$\text{GAP} - O\left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}}\right) \leq \tilde{O}\left(L_1 \frac{L\sqrt{W}}{\sqrt{N}} \left(\prod_{\ell=1}^L \|A_\ell\|\right) \left(\sum_{\ell=1}^L \frac{\|A_\ell\|_{\text{Fr}}^2}{\|A_\ell\|_\sigma^2}\right)^{\frac{1}{2}}\right). \quad (\text{C.1})$$

In [1], it was shown that with probability greater than $1 - \delta$ over the draw of the training set we have

$$\text{GAP} - \mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} \leq \tilde{O}\left(L_1 \frac{1}{\sqrt{N}} \prod_{\ell=1}^L \|A_\ell\| \left(\sum_{\ell=1}^L \frac{\|(A_\ell - M_\ell)^\top\|_{2,1}^{\frac{2}{3}}}{\|A_\ell\|^{\frac{2}{3}}}\right)^{\frac{3}{2}}\right). \quad (\text{C.2})$$

Note that equation (C.1) was discovered concurrently with (C.2) with a different approach ([3] relied on a PAC Bayes approach).

C.1 Bounds for Linear Neural Networks

In this subsection, we briefly list generalization bounds which can be obtained for linear maps in the multi-class classification case, focussing especially on the ones which apply to linear networks of the form $B_L \dots B_1$.

Most of the early literature on generalization bounds for linear multi-class classification focused on kernels, which are a slightly more general setting.

For a single output ($\mathcal{C} = 1$), it is well known from early results [68, 83] that the Rademacher complexity⁷ of linear classifiers is bounded as $\tilde{O}\left(\sqrt{\frac{\|a\|^2 \sum_i \|x_i\|^2}{N^2}}\right) \leq \tilde{O}\left(\mathcal{B} \sqrt{\frac{\|A\|^2}{N}}\right)$. Early work on generalization bounds for multi-class classification [85, 68, 66, 67, 86, 87] used inequalities for Rademacher complexities of composite function classes to obtain bounds of the form $\tilde{O}\left(\mathcal{B} \sqrt{\frac{\mathcal{C} \|A\|_{\text{Fr}}^2}{N}}\right)$. Later in [88, 65], the bound was refined to $\tilde{O}\left(\mathcal{B} \sqrt{\frac{\|A\|_{\text{Fr}}^2}{N}}\right)$ by exploiting the L^∞ continuity of common loss functions.

In [64], the authors proved the following bound specifically targetted at linear networks of the form $A = B_L \dots B_2 B_1$:

$$\tilde{O}\left(\mathcal{B} \sqrt{\frac{\mathcal{C} \min_\ell(\text{Rank}(B_\ell))}{N}} \prod_\ell \|B_\ell\|\right). \quad (\text{C.3})$$

Note that $\prod_\ell \|B_\ell\| \geq \|A\|$ and $\|A\| \min_\ell \sqrt{\text{rank}(B_\ell)} \geq \|A\| \sqrt{\text{rank}(A)} \geq \|A\|_{\text{Fr}}$. Thus, the bound in (C.3) behaves at least as $\tilde{O}\left(\sqrt{\frac{\mathcal{C} \|A\|_{\text{Fr}}^2}{N}}\right)$, which exhibits the same behavior as the results in [86, 66].

The results are expressed in terms of vector output Gaussian complexity, which has the advantage of yielding bounds for L2 Lipschitz loss functions. When the loss function is L^∞ Lipschitz, the bound from [65] will be tighter by a factor of $\sqrt{\mathcal{C}}$ (ignoring any logarithmic factors). The main advantage of the bound in equation (C.3) is the introduction of the novel proof technique which also allows the authors to derive the bound (C.9) for neural networks with activation functions.

Comparison between Bound (C.1) and Theorem 3.1 for linear networks:

It is worth noting that the Bound (C.1) can be instantiated as follows:

⁷For simplicity, we use \tilde{O} notation, however, many of the result in this section do not require logarithmic factors when expressed as a function of Rademacher complexity only. However, making them post hoc w.r.t. to the constraints requires additional logarithmic factors.

Corollary C.1. *With probability greater than $1 - \delta$ over the draw of the training set, every matrix A satisfies the following generalization bound*

$$\text{GAP} \leq \tilde{O} \left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} + L_1 \mathfrak{B} \|A\|^{\frac{L-1}{L}} L^{\frac{3}{2}} \sqrt{\frac{\mathcal{C} \|A\|_{\text{sc}, \frac{2}{L}}^{\frac{2}{L}}}{N}} \right). \quad (\text{C.4})$$

Proof. By the proof of Theorem 1.1, the values of B_1, \dots, B_L which minimize the weight decay regulariser $\sum_{\ell=1}^L \|B_\ell\|_{\text{Fr}}^2$ satisfy $\|B_\ell\| = \|A\|^{\frac{1}{L}}$ and even $\sigma_i(B_\ell) = \sigma_i(A)^{\frac{1}{L}}$ for all ℓ, i where $\sigma_i(Z)$ denotes the i th singular value of the matrix Z . It follows that for this choice of B_ℓ s, we have $\prod_{i \neq \ell} \|B_i\| = \|A\|^{\frac{L-1}{L}}$ and $\|B_\ell\|_{\text{Fr}}^2 = \|A\|_{\text{sc}, \frac{2}{L}}^{\frac{2}{L}}$. The theorem follows upon replacing the values into equation (C.1). \square

In contrast, for $p = \frac{2}{L}$, our Theorem 3.1 scales as

$$\begin{aligned} & \tilde{O} \left(L_1^{\frac{p}{2+p}} \sqrt{\frac{\mathfrak{B}^{\frac{2p}{2+p}} \|A\|_{\text{sc}, p}^{\frac{2p}{2+p}} \min(\mathcal{C}, d)^{\frac{p}{p+2}} \max(\mathcal{C}, d)^{\frac{2}{p+2}}}{N}} \right) \\ & \leq \tilde{O} \left(L_1^{\frac{1}{L+1}} \sqrt{\frac{\mathfrak{B}^{\frac{2}{L+1}} \|A\|_{\text{sc}, p}^{\frac{2}{L+2}} [\mathcal{C} + d]}{N}} \right). \end{aligned} \quad (\text{C.5})$$

The two are not directly comparable: for one thing, the bound in (C.4) is purely norm-based, which means that it enjoys more obvious scaling properties. This also makes it more vulnerable to large inputs. It is worth noting that both bounds potentially improve when L grows, as this influences the schatten quasi norm index. However, this benefit is strongly mitigated in the case of (C.1) by the multiplicative factors of L , which are not present in our case. The explicit dependency on \mathcal{C} is also worse in (C.1), but we believe this is an artefact of the L^2 based proof technique and might be improvable with an argument more carefully dedicated to the linear case.

C.2 Norm-based Bounds for Fully Connected Neural Networks

In [13], extending work from [12], Rademacher complexity bounds of the following order were obtained:

$$\tilde{O} \left(L_1 \sqrt{\frac{\mathfrak{B}^2 L \prod_{\ell=1}^L \|A_\ell\|_{\text{Fr}}^2}{N}} \right). \quad (\text{C.6})$$

In addition, the same paper [13] contains other variants, which mostly consist of products of various norms of the weight matrices. For instance (Corollary 2 page 13) for any fixed value p : $\text{GAP} \leq$

$$\tilde{O} \left(\mathcal{B} \sqrt{\frac{\log(\frac{1}{\delta})}{N}} + L_1 \mathfrak{B} \mathcal{U} \left[\prod_{\ell=1}^L \|A_\ell\| \right] \left[\frac{L}{p} \right]^{\frac{1}{\frac{2}{3}+p}} \frac{1}{N^{\frac{1}{2+3p}}} \right), \quad (\text{C.7})$$

where \mathcal{U} is an apriori upper bound on $\max_\ell \frac{\|A_\ell^\top\|_{2,1}}{\|A_\ell\|}$. Note: we have simplified the expression to adapt it to our \tilde{O} notation assuming that $\frac{M_p(A)}{M(A)} \leq O(\bar{W}^{\frac{1}{p}})$, where $M_p(A)$ (resp. $M(A)$) are upper bounds on the Schatten p and spectral norms of A respectively (such an assumption makes sense since for any matrix with dimensions bounded by \bar{W} , we have $\frac{\|A\|_{\text{sc}, p}}{\|A\|} \leq \bar{W}^{\frac{1}{p}}$). It is not possible to directly compare the factors of L in our bounds and in (C.7) since a fully post hoc version of that result would likely involve further factors of L as in our own proof.

Note that $\mathcal{U} \leq \bar{W}$, thus setting $p = \frac{1}{3}$ we obtain for instance

$$\text{GAP} - O \left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} \right) \leq \tilde{O} \left(L_1 \mathfrak{B} \prod_{\ell=1}^L \|A_\ell\| \frac{\bar{W} L}{N^{\frac{1}{3}}} \right). \quad (\text{C.8})$$

Still in the category of ‘peeling’ based bounds relying on vector contraction inequalities, the recent work of [64] (cf. Theorem 7), which appeared independently of the present work, has established bounds of the following form for the Gaussian complexity of neural networks with fixed norm and rank constraints

$$O\left(\prod_{\ell=1}^L \|A_\ell\| \mathfrak{B} \frac{C_1 \sqrt{w_1} \frac{\|A_1\|_{\text{Fr}}}{\|A_1\|} + \sum_{\ell=2}^L C_1^{L-\ell} C_2 \sqrt{w_\ell} \tilde{r}_\ell \frac{\|A_\ell\|_{\text{Fr}}}{\|A_\ell\|}}{\sqrt{N}}\right), \quad (\text{C.9})$$

where \tilde{r}_ℓ is an upper bound on the rank of layer ℓ . In fact, a notable property of the bound is that there are no logarithmic factors in the Rademacher complexity bound: however, note the Rademacher complexity bound into an excess risk bound would incur at least a logarithmic factor of N , and making the bound post hoc w.r.t. the norm constraints would incur additional logarithmic factors of the norms, as well as a non logarithmic factor of \sqrt{L} . Here, C_1, C_2 are unspecified constants inherited from [78] (which are described as ‘rather large’ on page 4 of [78]). However, the authors in [64] do conjecture that the constants and the accompanying explicitly exponential dependence on depth could be removed with a more careful analysis

$$\mathfrak{G} \leq O\left(C_1^L \mathfrak{B} \left[\prod_{\ell=1}^L \|A_\ell\|\right] L \underline{r} \sqrt{\frac{\bar{W}}{N}}\right), \quad (\text{C.10})$$

where \mathfrak{G} denotes the Gaussian complexity, and \underline{r} the minimum rank of the weight matrices A_1, \dots, A_L .

Comparison between (C.10) in [64] and (E.125)

Both results take the low-rank structure of the weight matrices into account and improve when neural collapse occurs. However, both the proof techniques and the final results differ significantly. Whilst translating the bound (C.10) into an *excess risk* bound would imply at least a logarithmic factor of N , and making the bound post hoc w.r.t. the norm constraints would incur additional logarithmic factors of the norms, as well as a non logarithmic factor of \sqrt{L} , the bound for the Gaussian complexity is free of any logarithmic factors (in contrast, even our Rademacher complexity bound from the calculation in Proposition E.1 includes logarithmic factors of the norms arising from the application of L^∞ results such at Lemma I.5), which could give it an advantage when L is small. On the other hand, the bound C.10 and C.9 exhibit both explicit and implicit exponential dependence on depth through the factors C_1^L and $\prod \|A_\ell\|$ respectively. In contrast, the result in G.7 has no non logarithmic dependency on norm based quantities. In fact, it belongs to the family of parameter counting bounds, whilst (C.10) is a purely norm based. In the case where all the spectral norms and the norms of the datapoints are considered to be $O(1)$, the bound (C.10) amounts to a sample complexity of $\tilde{O}(C_1^{2L} \underline{r}^2 L^2 \bar{W})$, whilst Theorem G.7 amounts to a sample complexity of $\tilde{O}(\underline{r} L^2 \bar{W})$: thus, the bounds are nearly identical in this particular situation apart from the removal of a factor of $\underline{r} C_1^L$ (which removes exponential dependence on depth) in our work at the cost of additional logarithmic factors. It is worth noting that the proofs are radically different: we rely on layer-wise covering numbers analogous to [1, 4, 74, 16] whilst [64] relies on vector contraction inequalities and is closer to the works of [12, 13]. In particular, one of the factors of L inside the square root in Theorem E.8 arises from making the bounds post hoc w.r.t. the spectral norms. It would be interesting to investigate whether this factor can be removed when assuming that $\|A_\ell\| = 1$ for all L .

Several other comparable bounds are proved in [64], which appeared independently of the present work. The bounds only apply to the exactly low-rank case, whilst we provide a more refined analysis based on the singular value spectrum decay implicit in the low Schatten p quasi norm constraints. In addition, we also consider convolutional neural networks, and the proof techniques are radically different.

C.3 Norm-based Bounds for Convolutional Neural Networks:

In [16], it was shown that ⁸

$$\text{GAP} - O \left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} \right) \leq \tilde{O} \left(\sqrt{L} \frac{\left[\sum_{\ell=1}^L (T_\ell)^{2/3} \right]^{\frac{3}{2}}}{\sqrt{N}} \right) \quad (\text{C.11})$$

where if $\ell \leq L-1$,

$$T_\ell = B_{\ell-1}(X) \|(A_\ell - M_\ell)^\top\|_{2,1} \sqrt{w_\ell} \max_{U \leq L} \frac{\prod_{u=\ell+1}^U \|\text{op}(\tilde{A}_u)\|}{B_U(X)}$$

and if $\ell = L$,

$$T_L = \frac{B_{L-1}(X)}{\gamma} \|A^L - M^L\|_{\text{Fr}}.$$

Here, we denote $B_\ell(X) = \max_i \|F_A^{0 \rightarrow \ell}(x)\|_\ell$, where the norm $\|\cdot\|_\ell$ denotes the maximum L^2 norm of a convolutional patch for a vector of activations at layer ℓ .

The above bound relies on loss function augmentation to control the L2 norm of patches at each layer. However, the following simpler bound (Theorem E.2) can also be of interest:

$$\text{GAP} - O \left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} \right) \leq \tilde{O} \left(\frac{L_1 \sqrt{L} \prod_{\ell=1}^L \|\text{op}(A_\ell)\|}{\sqrt{N}} \left[\sum_{\ell=1}^L \left(\frac{U_\ell w_\ell \|(A_\ell - M_\ell)^\top\|_{2,1}}{\|\text{op}(A_\ell)\|} \right)^{\frac{2}{3}} \right]^{\frac{3}{2}} \right). \quad (\text{C.12})$$

Recently, in [32] an ingenious approach was used to provide bounds for convolutional neural networks in terms of products of the Frobenius norms of the weight matrices. The bounds and proof techniques are closer to the school of [12, 13] than those of [1], and rely on vector concentration inequalities and direct calculations of Rademacher complexities rather than covering numbers, whilst incorporating many novel elements specific to the CNN situation. We have, considering a post hoc version of Theorem 3.2 in [32] and translating to our notation, $\text{GAP} - O \left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} \right) \leq$

$$\tilde{O} \left(\mathcal{B} \sqrt{\frac{L^2}{N}} + \frac{L_1 \prod_{\ell=1}^L \|A_\ell\|_{\text{Fr}} \sqrt{L \prod_{\ell=1}^L w_\ell B}}{\sqrt{N}} \right), \quad (\text{C.13})$$

where w_ℓ denotes the spatial size of the convolutional patches at layer ℓ and $B = \max_{o,i} \|(x_i)_{S^{0,o}}\|$ is the maximum L^2 norm of a convolutional patch in the input.

The bounds (C.12) and (C.13) exploit the sparsity of connections at the first layer to obtain bounds which only depend on the maximum norm of an input convolutional patch (as opposed to the full norm of the input), a characteristic which is shared by our result (cf. theorem 3.6. In addition, the results in [16] exploit the weight sharing to involve only the norms of the weight matrices (rather than the linear operators $\text{op}(A_\ell)$) in the numerator. Quite remarkably, the bound in [32] achieves a similar effect in the context of peeling type results by relying exclusively on the sparsity of connections. Our results for CNNs can be interpreted as a low-rank generalization of those in [16] incorporating low-rank structure in the weight matrices, just as our results for DNNs generalize those of [1]. However, the results in [32] are not directly comparable, since they belong to the peeling family, affording them a more favorable dependence on width but a poorer dependence on depth if the norms $\|A_\ell\|_{\text{Fr}}$ are moderately large.

⁸As explained in the main paper, our convention on \tilde{O} notation is stricter than that in [16]: we allow the \tilde{O} to absorb factors of $\log(\max_{\ell=1}^L \|A_\ell\|)$, but we treat $\log(\prod_{\ell=1}^L \|A_\ell\|)$ as $\tilde{O}(L)$, which explains the additional multiplicative factor of \sqrt{L} . This factor is absent in [74] due to the use of a purely L^2 covering number approach, though the implicit dependency on the number of classes at the last layer is stronger.

C.4 Parameter Counting Bounds

In [7], generalization bounds based on VC dimensions were established for fully connected neural networks. In more recent research the focus of parameter counting bounds has been on different architectures such as Convolutional Neural Networks or ResNets. We focus on CNNs below as this work focuses on DNNs and CNNs.

In [4] (Theorem 3.1), the following bound is proved for CNNs with fixed norm constraints:

$$\text{GAP} \leq \tilde{O} \left(\mathcal{B} \sqrt{\frac{\mathcal{W}SL}{N}} + \mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} \right), \quad (\text{C.14})$$

where $\mathcal{W} := \sum_{\ell=1}^L d_{\ell-1} \times U_{\ell}$ denotes the total number of parameters in the network and \mathcal{S} is an a priori upper bound on $[\max_{\ell} \|A_{\ell}\|] - 1$.

Later in [74] (Theorem 3.4), the following result was shown⁹:

$$\text{GAP} \leq \tilde{O} \left(\mathcal{B} \sqrt{\frac{\mathcal{W}L}{N}} + \mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} \right). \quad (\text{C.15})$$

Note: in [4], the regime of interest is that where the spectral norms of the weight matrices approach 1, which explains the appearance of the quantity \mathcal{S} . In fact, the quantity appears through an upper bound on the product of spectral norms as $(1 + \nu)^L \leq \exp(\nu L)$, thus, the multiplicative factor of $[LS]$ can relatively straightforwardly be replaced by $\tilde{O}(L)$. This results in a bound which looks identical to equation (C.15) when expressed in \tilde{O} notation. However, the bound in [74] presents many other advantages, including applicability to Resnets and much smaller constants and logarithmic factors. We also note that [89] provide rank-sensitive bounds for neural networks. However, the bounds include a multiplicative factor of the Jacobian of the full network (both outside and inside the logarithmic factors): the bounds will coincide (up to log factors) with our pure parameter counting bounds in the particular case when all the spectral norms of the weights are exactly equal to 1, but behave less favorably in other cases.

D Experiments

In this section, we present some experiments to demonstrate that our bound can successfully capture the low rank structure in the weights to improve upon competing bounds. Although the numerical advantage of our bounds over the best compared method (the parameter counting bounds of [74]), the bounds generally show a much milder growth as the width of the layers grows compared to all existing bounds when the generalization gap is kept relatively constant. Thus, our bounds are able to adapt to capture the presence of unused function class capacity in the form of low-rank weights. We present results for both DNNs on the MNIST dataset and CNNs on the CIFAR-10 dataset. All experiments were run with 128 CPUs with 500GB ram and DGX A100.

D.1 Fully-connected Neural Networks (MNIST)

To evaluate the bounds' behavior for fully connected neural networks, we conducted experiments on the MNIST dataset. We set the number of non-linear hidden layers to $L = 3$ and varied the width of the hidden layers w_{ℓ} in $\{300, 400, 500, 700\}$ to explore different regimes. To guarantee that we could analyze the phenomenon studied in this research, we enforced rank-sparsity through two complementary approaches: (1) implicitly, by preceding each non-linear hidden layer with four linear layers. (2) explicitly, by regularizing the neural network through weight decay. We selected the regularization parameter from $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, \dots, 10^2\}$ and analyzed models with the

⁹Note, keeping in line with the convention used for the \tilde{O} notation in this paper, we interpret $[\max_{\ell} \|A_{\ell}\|]$ as a constant. This implies that the presence of the quantities $C_{i,j}$ in [74] introduces a logarithmic dependency in $[\max_{\ell} \|A_{\ell}\|]^L$ inside the square root, corresponding to a multiplicative factor of L inside the square root.

highest weight decay that achieved an accuracy of at least 0.9. We then maximize the margin γ subject to $I_{\gamma, X} \leq 0.1$.

The results can be seen in Figure D.1. To facilitate comparison, denser bars are lower than less dense bars when compared to their counterparts, i.e, by considering models with the same width. In addition, we only plot the bounds in \tilde{O} terms, ignoring all polylogarithmic factors. This facilitates the comparison since some competing baselines are post hoc whilst some are not: our bounds hold uniformly over all values of p_ℓ , $\|A_\ell - M_\ell\|_{scc, p_\ell}$, γ etc., whilst some of the related works do not explicitly perform the required union bound to achieve this. The practice of evaluating the numerical values of generalization bounds only up to polylogarithmic factors [74, 64] or not at all [3, 4] is in line with the literature. When comparing to [64], we evaluate the bound for two values of the intractable constant $C_1 = 1, 2$. This is to account for the the authors' statement that the factor of C_1^L could potentially be removed with a more refined proof, and certainly underestimates the numerical value of the established bound, since the constant C_1 is ‘rather large’ [78]. We include both versions of our bound: Theorem 3.2 (without loss function augmentation) and Theorem 3.3 (with loss function augmentation). A key observation is that as the width increases, our bounds do not increase as fast as the competing results, suggesting a successful use of the rank sparsity of the problem: in this experiment, the accuracy and generalization gap are constant (overparametrized regime), therefore, the more modest growth exhibited by our model indicates a better ability to capture true generalization error than the competing methods, which show a much sharper growth with width.

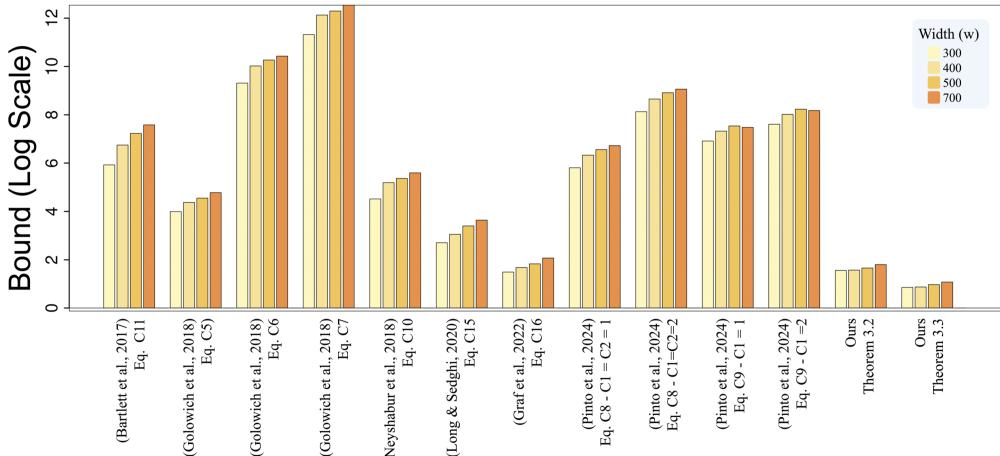


Figure D.1: Empirical comparison between our generalization results and previous works.

Further Validation of the Low-rank Condition Moreover, the singular value spectra plots from Figure D.6 below demonstrates the presence of rank sparsity in our trained models. This confirms the observations made by a multitude of recent papers in the optimization literature [21, 22, 27], which further explains why our bounds’ advantage improves when the width increases. In this example, even the value $p_\ell = 0$ yields an advantage over classic parameter counting bounds due to the exact low-rank structure.

D.2 Convolutional Neural Networks (CIFAR-10)

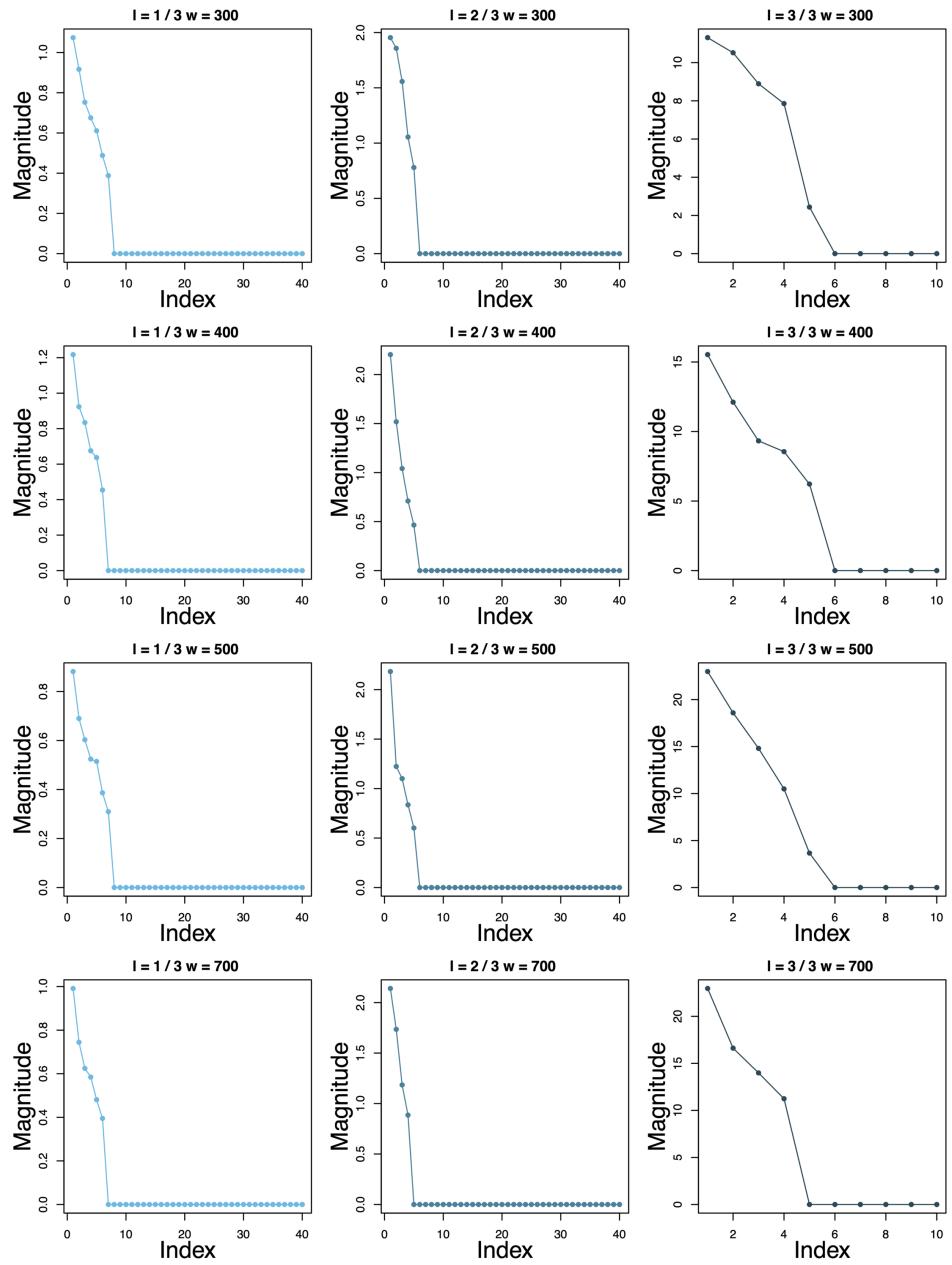


Figure D.2: Spectral Decay of Intermediary Layers in the MNIST Dataset (Fully Connected).

D.3 Convolutional Neural Networks (CIFAR-10)

In this section, we present experiments on the CIFAR-10 dataset. We train a neural network with 3 convolutional layers with kernel size $(3, 3)$ each followed by 5 fully connected layers. The last (output) layer has width 10 since there are 10 classes, and the width of the intermediary fully connected layers is varied between 100 and 1000. Due to the additional computational burden of the task and the more pronounced overparametrization, we evaluate the bounds on a single trained network for each architecture. To moderate the effect of norm-based factors, we employ spectral regularization, constraining the spectral norms of all weight matrices to 1. We set the margin dynamically to ensure a loss of only 1 percent in accuracy compared to a margin of zero. As in the fully connected case, we note that as most of the bounds are asymptotic in nature (and many of the sources do not calculate the explicit constants) their ‘practical evaluation’ challenging. For this reason, we only evaluate dominant term of each bound (ignoring the constants and logarithmic factors) and we employ a logarithmic scale, which is better suited to this asymptotic analysis since the bounds differ from each other by very large factors. This is also standard practice in the related literature [74, 4, 1] whenever bounds are evaluated. As in the fully connected case, our bounds outperform the competing ones. However, more interesting is the fact that our bounds exhibit a milder dependency on the width, which better aligns with the true generalization error.

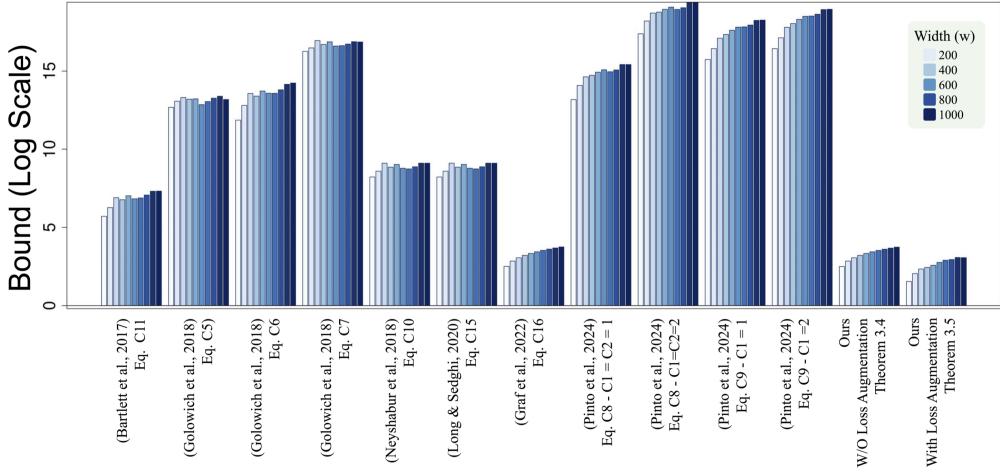


Figure D.3: Numerical Comparison of our Bounds with other Bounds from the Literature (CNNs, CIFAR-10).

Furthermore, to provide further evidence of spectral decay and strong *approximate low-rank structure* in our trained models, we provide the following plots of the singular spectral of the weights of our trained models. The plots confirm that the approximate low-rank structure is stronger for larger widths, which also explains the fact that our bounds’ advantage over existing approaches widens in such overparametrization regimes.

We also provide a few additional comparisons to validate our conclusion that our bounds respond to varying degrees of overparametrization better than pure parameter counting bounds. As mentioned in the main paper, the choice of p_ℓ can be made after training, as the bounds hold uniformly over their choice. In fact, the choice of p_ℓ made after optimizing the bounds is a relevant factor in the interpretation of the bound in terms of overparametrization regime: a larger but moderate value of p (in the range $[0.2, 1]$) indicates that the norm-based factors are moderate enough to be involved in the bound and that the low-rank structure in the weight matrices is sufficiently pronounced to be exploited. On the other hand, a value of $p = 0$ can indicate one of two things: either (1) the norm-based factors are too large for the low rank structure to be exploited, or (2) the weight matrices are exactly low-rank. In our experiments, we observe the optimization of the bound often results in the value $p_\ell = 0$ being chosen for convolutional layers, whilst the fully connected layers tend to correspond to more moderate values of p_ℓ . This can be seen in Figure D.5 below, which corresponds

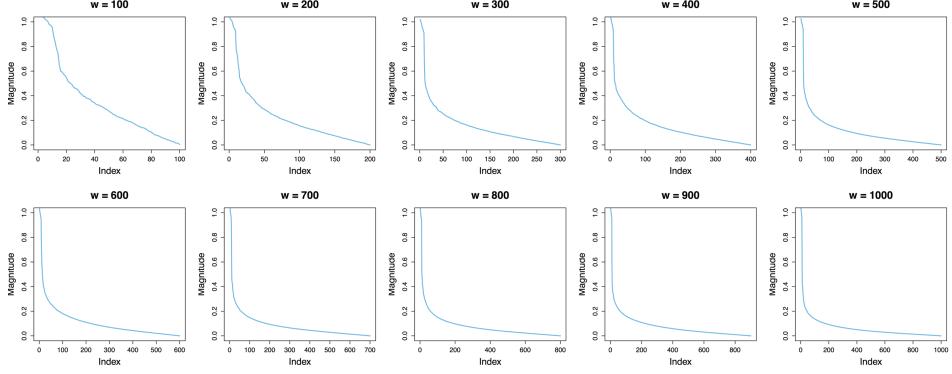


Figure D.4: Illustration of the Spectral Decay of the Trained Models on the Layer fc2.

to the most overparametrized CNN model ($w = 1000$) where the p_ℓ s are selected based on the optimization of the bounds from Theorem 3.7 (with loss function augmentation).

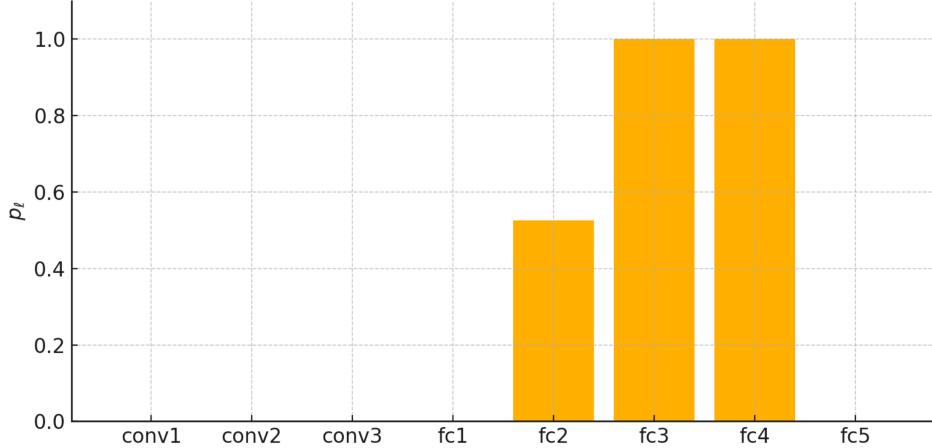


Figure D.5: Chosen Values of p_ℓ (with Loss Function Augmentation) for the most Overparametrized CNN Model (width = 1000)

The results are somewhat expected: convolutional layers are already quite parameter efficient, making the exploitation of the low-rank structure comparatively less necessary. Furthermore, the norm-based factors arising from the product $\prod_{i=\ell}^L s_i$ are larger since convolutional layers have a larger spectral norm due to the repetition over patches. On the other hand, the fully connected layers are able to select larger values of p_ℓ due to a more pronounced low-rank structure and moderate norm-based factors achieved through a combination of loss function augmentation and spectral regularization during training.

We note that the choice of p_ℓ generally introduces a tradeoff which is only imperfectly captured by the regime corresponding to the value of p_ℓ chosen when optimizing the bound. Indeed, lower values of p_ℓ correspond both to a bias towards the parameter-counting component of our bounds and stronger exploitation of the approximate low-rank structure (assuming the norm based factors are not too large). Thus, more moderate values of p_ℓ are preferable in terms of the bounds' responsiveness to changes in architecture, whilst setting very close to zero can sometimes be beneficial in terms of overall numerical performance due to the convergence to a parameter counting result. To better isolate the bounds' ability to capture low-rank structure and its indirect effect on function class capacity, we evaluate the bounds when fixing the value of p_ℓ to various values for all layers. We then compare the behavior of our results from Theorem G.5 (without loss function augmentation, in blue) and Theorem 3.7 (with loss function augmentation, in turquoise) to the parameter counting baseline of [74] (in yellow). To isolate the behavior as the overparametrization increases, we plot the

ratio between the value of the bounds compared to the value of the same bound for a width of 100. Similarly, we also plot the ratio of the *test error* (in red) compared to the test error for a width of 100 (i.e. 0.2793). The results are in Figure D.6 below.

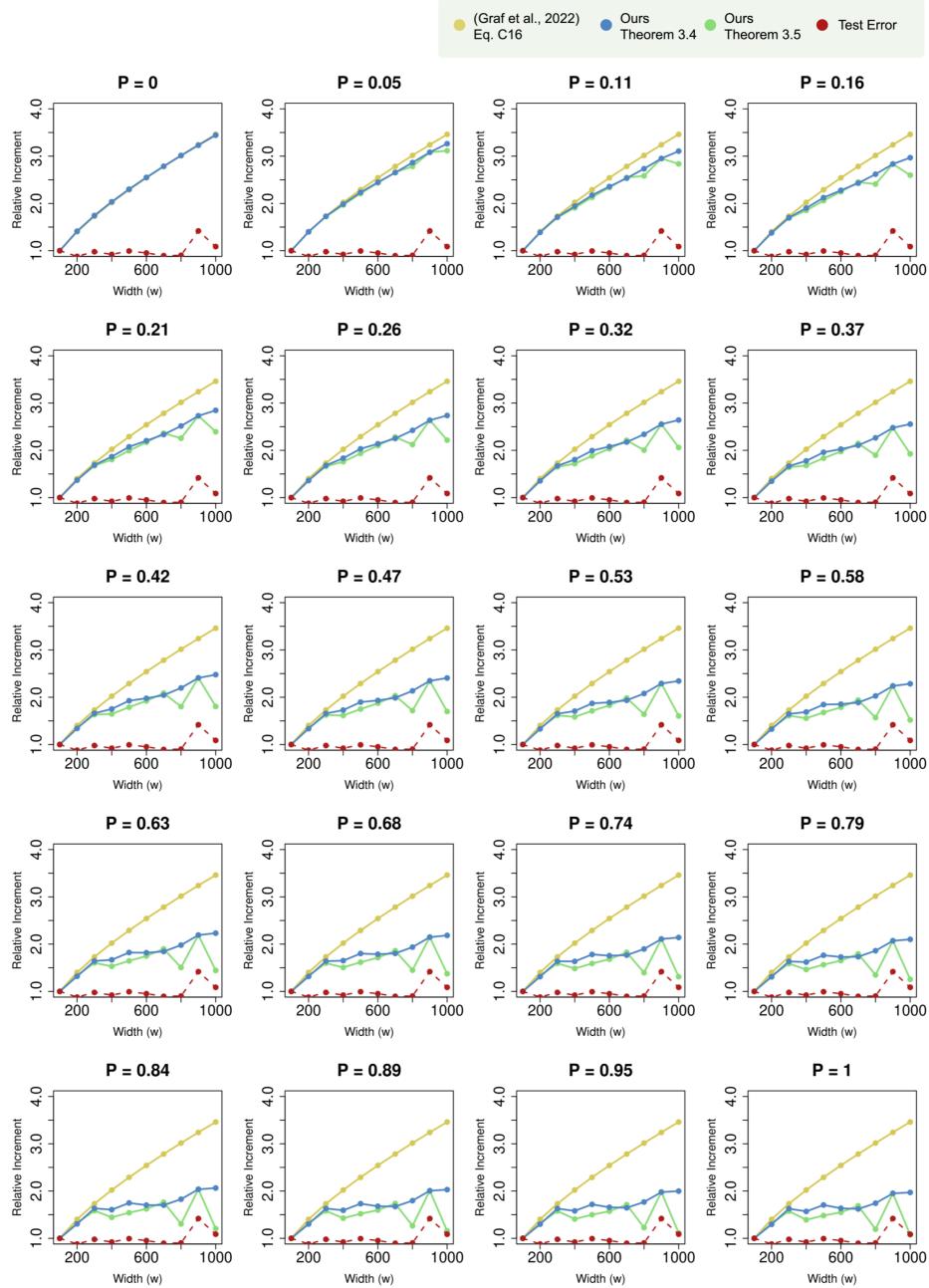


Figure D.6: Relative Growth of our Main Results as a Function of Width

We can directly observe that for moderate values of p_ℓ , both our bounds are successful at capturing the implicit function class restriction that occurs through both norm-based effects and spontaneous rank-sparsity in the overparametrized regime: for small widths (100 to 300), the bounds all grow linearly, which correspond to a parameter-counting behavior (indeed, the parameter count grows quadratically in the width, which results in linear growth of the term $\sqrt{\frac{D}{N}}$ where D is the number of parameters), at the regime of moderate overparametrization (width 300 to 700), our bounds maintain

relatively constant behavior, which matches the behavior of the test error. Then, at the extremely overparametrized regime (widths 800 to 1000), the bounds begin to grow again and exhibit less stable behavior, which also better matches the behavior of the test error compared to the parameter counting baseline, which continues to grow linearly.

E Generalization Bounds for Fully Connected Networks

In this section, we prove our generalization bounds for classes of neural networks with Schatten p quasi norm constraints on the weight matrices. For each layer ℓ , we have a set of admissible matrices $\mathbb{R}^{w_\ell \times w_{\ell-1}} \ni \mathcal{B}_\ell := \{A_\ell \in \mathbb{R}^{w_\ell \times w_{\ell-1}} : \|A_\ell - M_\ell\|_{\text{sc},p_\ell} \leq \mathcal{M}_\ell; \|A_\ell\| \leq s_\ell\}$ and

$$\begin{aligned} \mathcal{F}_\mathcal{B} &:= \{F_A = F_{(A_1, \dots, A_L)} : A_\ell \in \mathcal{B}_\ell \quad \forall \ell \leq L\} \\ &= \left\{ F_A = F_{(A_1, \dots, A_L)} : \|A_\ell - M_\ell\|_{\text{sc}} \leq \mathcal{M}_\ell, \|A\| \leq s_\ell \quad \forall \ell \leq L \right\}. \end{aligned} \quad (\text{E.1})$$

We define similarly:

$$\mathcal{F}_\mathcal{B}^{0 \rightarrow \ell} := \{F_A^{0 \rightarrow \ell} = F_{(A_1, \dots, A_\ell)} : A_\ell \in \mathcal{B}_\ell \quad \forall l \leq \ell\}, \quad (\text{E.2})$$

where we define

$$\mathcal{B}_\ell = \{A_\ell \in \mathbb{R}^{w_\ell \times w_{\ell-1}} : \|A_\ell - M_\ell\|_{\text{sc},p_\ell} \leq \mathcal{M}_\ell; \|A_\ell\| \leq s_\ell\}.$$

Proposition E.1. *Fix some reference/initialization matrices $M^1 \in \mathbb{R}^{w_1 \times w_0} = \mathbb{R}^{w_1 \times d}, \dots, M^L \in \mathbb{R}^{w_L \times w_{L-1}} = \mathbb{R}^{C \times w_{L-1}}$. Assume also that the inputs $x \in \mathbb{R}^d$ satisfy $\|x\| \leq b$ w.p. 1. Fix a set of constraint parameters $0 \leq p_\ell \leq 2, s_\ell, \mathcal{M}_\ell$ (for $\ell = 1, \dots, L$).*

With probability greater than $1 - \delta$ over the draw of an i.i.d. training set x_1, \dots, x_N , every neural network $F_A(x) := A_L \sigma_L(A^{L-1} \sigma_{L-1}(\dots \sigma_1(A^1 x) \dots))$ satisfying the following conditions:

$$\begin{aligned} \|A^\ell - M^\ell\|_{\text{sc},p_\ell} &\leq \mathcal{M}_\ell & \forall 1 \leq \ell \leq L \\ \|A\| &\leq s_\ell & \forall 1 \leq \ell \leq L-1 \\ \|A_L^\top\|_{2,\infty} &\leq s_L, \end{aligned} \quad (\text{E.3})$$

also satisfies the following generalization bound:

$$\begin{aligned} \mathbb{E} [1(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N 1(F_A(x_i), y_i) \\ \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta})}{N}} + \frac{240\sqrt{\log_2(4\Gamma)} \log(N)[\mathcal{B}+1]}{\sqrt{N}} (1 + L L_1)^{\frac{p}{2+p}} \mathcal{R}_{\mathcal{M},s,p,b} \end{aligned} \quad (\text{E.4})$$

where $\mathcal{R}_{\mathcal{M},s,p,b} :=$

$$\left[\sum_{\ell=1}^{L-1} \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} [w_\ell \underline{w}_\ell]^{\frac{p_\ell}{p_\ell+2}} + \left[b \prod_i \rho_i s_i \right]^{\frac{2p_L}{p_L+2}} r_\ell^{\frac{2}{p_L+2}} \bar{w}_L^{\frac{2}{p_L+2}} \underline{w}_L^{\frac{p_L}{p_L+2}} \right]^{\frac{1}{2}}, \quad (\text{E.5})$$

$\underline{\Gamma} := \left[[b+1][L L_1+1] \bar{W} N \prod_{i=1}^L [(\rho_i + 1)s_i + 1] \right]$ and we use the shorthands $r_\ell := \frac{\mathcal{M}_\ell^p}{s_\ell^p}$, $\underline{w}_\ell := \min(w_\ell, w_{\ell-1})$, $\bar{w}_\ell := [w_\ell + w_{\ell-1}]$.

Proof. Write $\mathcal{F}_\mathcal{B}$ for the class of neural networks satisfying conditions (E.3). By Lemma I.7 with $\alpha = \frac{1}{N}$ and Proposition E.11, we have the following bound on the empirical Rademacher complexity of $\mathcal{F}_\mathcal{B}$

$$\begin{aligned}
& \Re_S(l \circ \mathcal{F}_B) \\
& \leq \frac{4}{N} + \frac{12}{\sqrt{N}} \int_{\frac{1}{N}}^B \sqrt{96 \log_2(4\Gamma) \left[\frac{(1+L\mathbf{L}_1)}{\min(\varepsilon, 1)} \right]^{\frac{2p}{2+p}} \sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} d\epsilon} \\
& \leq \frac{4}{N} + \frac{120\sqrt{\log_2(4\Gamma)}}{\sqrt{N}} (1+L\mathbf{L}_1)^{\frac{p}{2+p}} \left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}} \left[\int_{\frac{1}{N}}^1 \frac{1}{\epsilon} d\epsilon + \mathcal{B} - 1 \right] \\
& \leq \frac{4}{N} + \frac{120\sqrt{\log_2(4\Gamma)}[\log(N) + \mathcal{B}]}{\sqrt{N}} (1+L\mathbf{L}_1)^{\frac{p}{2+p}} \left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}}. \tag{E.6}
\end{aligned}$$

Next, by Theorem I.6, we have w.p. $\geq 1 - \delta$,

$$\begin{aligned}
& \mathbb{E} [l(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N l(F_A(x_i), y_i) \leq 6\mathcal{B} \sqrt{\frac{\log(4/\delta)}{2N}} + \frac{8}{N} \\
& + \frac{240\sqrt{\log_2(4\Gamma)}[\log(N) + \mathcal{B}]}{\sqrt{N}} (1+L\mathbf{L}_1)^{\frac{p}{2+p}} \left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}} \\
& \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta})}{N}} \\
& + \frac{240\sqrt{\log_2(4\Gamma)}\log(N)[\mathcal{B}+1]}{\sqrt{N}} (1+L\mathbf{L}_1)^{\frac{p}{2+p}} \left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}},
\end{aligned} \tag{E.7}$$

where the last inequality holds for $N \geq 9$ (if this doesn't hold, inequality (E.4) holds trivially). Plugging equation (E.7) into equation (E.6) yields the result from equation (E.4) as expected. \square

Using Proposition E.1 we can prove the following post hoc version, where the values of p_ℓ can be optimized.

Theorem E.2. *With probability greater than $1 - \delta$, for any $b \in \mathbb{R}^+$, every trained neural network F_A and sampling distribution with $\|x\| \leq b$ w.p. 1 satisfy the following generalization gap for all possible values of the p_ℓ s:*

$$\begin{aligned}
& \mathbb{E} [l(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N l(F_A(x_i), y_i) \\
& \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(1/\delta)}{N}} + 6\mathcal{B} \sqrt{\frac{5 + \Theta_{\log}}{N}} + 480[\mathcal{B}+1](1+L\mathbf{L}_1)^{\frac{p}{2+p}} \frac{\sqrt{\log(\gamma_{FC}) \log(N)}}{\sqrt{N}} \tilde{\mathcal{R}}_{F_A},
\end{aligned}$$

where $\Theta_{\log} := 2 \log [4\bar{W} + b \prod_i \rho_i \|A_i\|] \left[|\log(b)| + L \left[\sum_{\ell=1}^L |\log(\|A_\ell\|)| + 2 \log(4\bar{W}) \right] \right]$, $\gamma_{FC} := 12 \left[[b+1][L\mathbf{L}_1+1]\bar{W}N \prod_{i=1}^L [\rho_i \|A_i\| + 1] \right] [4\bar{W} + b \prod_i \rho_i \|A_i\|]$, b is an a priori bound on the input samples' L^2 norm and where

$$\begin{aligned} \tilde{\mathcal{R}}_{F_A} := & \left[\sum_{\ell=1}^{L-1} \left[b \rho_L \|A_L^\top\|_{2,\infty} \prod_{i=1}^{L-1} \rho_i \|A_i\| \right]^{\frac{2p_\ell}{p_\ell+2}} \left[\frac{\|A_\ell - M_\ell\|_{\text{sc},p_\ell}^{p_\ell}}{\|A_\ell\|^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} [w_\ell w_\ell]^{\frac{p_\ell}{p_\ell+2}} \right. \\ & \left. + \left[b \rho_L \|A_L^\top\|_{2,\infty} \prod_{i=1}^{L-1} \rho_i \|A_i\| \right]^{\frac{2p_L}{p_L+2}} \left[\frac{\|A_L - M_L\|_{\text{sc},p_L}^{p_L}}{\|A_L^\top\|_{2,\infty}^{p_L}} \right]^{\frac{2}{p_L+2}} \bar{w}_L^{\frac{2}{p_L+2}} w_L^{\frac{p_L}{p_L+2}} \right]^{\frac{1}{2}}. \end{aligned}$$

Furthermore, the result also holds with b replaced by the empirical quantity $\tilde{b} = \max_{i \leq N} \|x_i\|$.

Proof. Let $\Delta_1, \Delta_2 > 0$ be positive real numbers to be determined later. For every set of values $k^{(b)}, k^{(\mathcal{M}_\ell)}, k^{(s_\ell)} \in \mathbb{Z}$ and $1 \leq k^{(p_\ell)} \leq \frac{2}{\Delta}$ (where $\ell = 1, \dots, L$) we apply proposition E.11 for the following values of $b, k^{(s_\ell)}, \mathcal{M}_\ell, p_\ell, \delta$:

$$b \leftarrow \tilde{b} = \exp(\Delta_1 k^{(b)}) \quad (\text{E.8})$$

$$s_\ell \leftarrow \tilde{s}_\ell = \exp(\Delta_1 k^{(s_\ell)}/L) \quad (\text{E.9})$$

$$\mathcal{M}_\ell \leftarrow \tilde{\mathcal{M}}_\ell = \tilde{s}_\ell \tilde{r}_\ell^{\frac{1}{p}} \quad \text{with} \quad (\text{E.10})$$

$$\tilde{r}_\ell := \exp(\Delta_1 k^{(\mathcal{M}_\ell)}) \quad (\text{E.11})$$

$$p_\ell \leftarrow \tilde{p}_\ell := \Delta_2 k^{(p_\ell)} \quad (\text{E.12})$$

$$\delta_{k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}} = \delta \Delta_2^L \frac{1}{3^{3L+1}} 2^{-|k^{(b)}| - \sum_\ell |k^{(s_\ell)}| - \sum_\ell |k^{(\mathcal{M}_\ell)}|}. \quad (\text{E.13})$$

This implies that with probability greater than $1 - \delta_{k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}}$,

$$\begin{aligned} \mathbb{E} [\mathbb{I}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbb{I}(F_A(x_i), y_i) & \leq 6 \mathcal{B} \sqrt{\frac{\log(\frac{4}{\delta_{k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}}})}{N}} \\ & + \frac{240 \sqrt{\log_2(4\Gamma_{k^{(b)}, k^{(s_\ell)}})} \log(N)[\mathcal{B}+1]}{\sqrt{N}} (1 + L L_1)^{\frac{p}{2+p}} \mathcal{R}_{k^{(p_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(b)}, k^{(s_\ell)}}, \end{aligned} \quad (\text{E.14})$$

where $\mathcal{R}_{k^{(p_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(b)}, k^{(s_\ell)}}$ is defined as the value of $\mathcal{R}_{\mathcal{M}, s, p, b}$ for $b, s_\ell, \mathcal{M}_\ell, p_\ell$ taking the values defined in equation (E.8) and $\Gamma_{k^{(b)}, k^{(s_\ell)}} = \left[[\exp(k^{(b)}\Delta) + 1][L_1 + 1]\bar{W}N \prod_{\ell=1}^L [\rho_\ell \exp(k^{(s_\ell)}\Delta) + 1] \right]$.

Note that

$$\sum_{k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}} \delta_{k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}} = \delta \Delta_2^L \frac{1}{3^{3L+1}} \left[\frac{2}{\Delta_2} \right]^L \left[\sum_{i \in \mathbb{Z}} \frac{1}{2^i} \right]^{2L+1} \leq \frac{2}{3} \delta < \delta. \quad (\text{E.15})$$

Thus, equation (E.14) holds simultaneously over all admissible values of $k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}$ with probability $\geq 1 - \delta$.

For a given neural network F_A and choice of p_ℓ s, we utilise the result above for the following values of $k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}$:

$$\begin{aligned} k^{(b)} &:= \lceil \log(b)/\Delta_1 \rceil \\ k^{(s_\ell)} &:= \lceil \log(s_\ell)L/\Delta_1 \rceil \\ k^{(\mathcal{M}_\ell)} &:= \lceil \log(\mathcal{M}_\ell)/\Delta_1 \rceil \\ k^{(p_\ell)} &:= \left\lceil \frac{p_\ell}{\Delta_2} \right\rceil \end{aligned} \quad (\text{E.16})$$

where $b \leftarrow \max_i \|x_i\|$, $s_\ell \leftarrow \|A_\ell\|$, $r_\ell \leftarrow \|A_\ell - M_\ell\|_{\text{sc}, p_\ell}^{p_\ell} / \|A\|^{p_\ell} = \sum_i \sigma_i (A - M_\ell)^{p_\ell}$ (for $\ell \neq L$) and $r_L \leftarrow \|A_L - M_L\|_{\text{sc}, p_L}^{p_L} / \|A_L^\top\|_{2, \infty}^{p_L} \leq \sqrt{w_L} \sum_i \sigma_i (A_L - M_L)^{p_L}$.

For these values, we certainly have:

$$\underline{\Gamma}_{k^{(b)}, k^{(s_\ell)}} \leq \left[[\exp(\Delta_1)b + 1][L_1 + 1]\bar{W}N \prod_{i=1}^L [\rho_i s_i \exp(\Delta_1/L) + 1] \right] \quad (\text{E.17})$$

$$\leq \exp(2\Delta_1) \left[[b + 1][L L_1 + 1]\bar{W}N \prod_{i=1}^L [(\rho_i + 1)s_i + 1] \right] = \exp(2\Delta_1)\underline{\Gamma}, \quad (\text{E.18})$$

because $\underline{\Gamma} := [[b + 1][L L_1 + 1]\bar{W}N \prod_{i=1}^L [(\rho_i + 1)s_i + 1]]$. Further, we have

$$\tilde{r}_\ell \leq r_\ell e^{\Delta_1}. \quad (\text{E.19})$$

Note also that

$$\frac{2\tilde{p}_\ell}{\tilde{p}_\ell + 2} - \frac{2p_\ell}{p_\ell + 2} \leq \frac{2[\tilde{p}_\ell - p_\ell]}{\tilde{p}_\ell + 2} \leq 2[\tilde{p}_\ell - p_\ell]/2 \leq \Delta_2 \quad (\text{E.20})$$

and

$$\frac{2}{\tilde{p}_\ell + 2} - \frac{2}{p_\ell + 2} = \frac{2[\tilde{p}_\ell - p_\ell]}{[\tilde{p}_\ell + 2][p_\ell + 2]} \leq \frac{2\Delta_2}{4} \leq \Delta_2/2. \quad (\text{E.21})$$

Thus, we can continue:

$$\mathcal{R}_{k^{(p_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(b)}, k^{(s_\ell)}} \leq \left[\sum_{\ell=1}^L \left[\exp(\Delta_1)b \prod_i \rho_i s_i \exp(\Delta_1/L) \right]^{\frac{2\tilde{p}_\ell}{\tilde{p}_\ell + 2}} \tilde{r}_\ell^{\frac{2}{\tilde{p}_\ell + 2}} \bar{w}_\ell^{\frac{2}{\tilde{p}_\ell + 2}} \tilde{w}_\ell^{\frac{\tilde{p}_\ell}{\tilde{p}_\ell + 2}} \right]^{\frac{1}{2}} \quad (\text{E.22})$$

$$\leq \exp(\Delta_1) \left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2\tilde{p}_\ell}{\tilde{p}_\ell + 2}} \tilde{r}_\ell^{\frac{2}{\tilde{p}_\ell + 2}} \bar{w}_\ell^{\frac{2}{\tilde{p}_\ell + 2}} \tilde{w}_\ell^{\frac{\tilde{p}_\ell}{\tilde{p}_\ell + 2}} \right]^{\frac{1}{2}} \quad (\text{E.23})$$

$$\leq \left[\exp(\Delta_1) \left[b \prod_i \rho_i s_i \right]^{\Delta_2/2} \right] \left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell + 2}} \tilde{r}_\ell^{\frac{2}{p_\ell + 2}} \bar{w}_\ell^{\frac{2}{p_\ell + 2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell + 2}} \right]^{\frac{1}{2}} \quad (\text{E.24})$$

$$\leq \left[e^{\Delta_1} \left[b \prod_i \rho_i s_i \right]^{\frac{\Delta_2}{2}} e^{\Delta_1/2} \right] \left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell + 2}} \left[r_\ell \bar{w}_\ell \right]^{\frac{2}{p_\ell + 2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell + 2}} \right]^{\frac{1}{2}} \quad (\text{E.25})$$

$$\leq \left[e^{\Delta_1} \left[b \prod_i \rho_i s_i \right]^{\frac{\Delta_2}{2}} e^{\Delta_1/2} \right] \left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell + 2}} \left[r_\ell \bar{w}_\ell \right]^{\frac{2}{p_\ell + 2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell + 2}} \right]^{\frac{1}{2}} \quad (\text{E.26})$$

$$\leq \left[e^{1.5\Delta_1} \left[b \prod_i \rho_i s_i \right]^{\frac{\Delta_2}{2}} \bar{w}_\ell^{\frac{\Delta_2}{2}} \right] \left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell + 2}} \left[r_\ell \bar{w}_\ell \right]^{\frac{2}{p_\ell + 2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell + 2}} \right]^{\frac{1}{2}} \quad (\text{E.27})$$

$$\leq \left[e^{1.5\Delta_1} \left[b \prod_i \rho_i s_i \right]^{\frac{\Delta_2}{2}} \bar{w}_\ell^{\frac{\Delta_2}{2}} \right] \left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell + 2}} \left[r_\ell \bar{w}_\ell \right]^{\frac{2}{p_\ell + 2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell + 2}} \right]^{\frac{1}{2}} \quad (\text{E.28})$$

$$\leq \left[e^{1.5\Delta_1} \left[b \prod_i \rho_i s_i \right]^{\frac{\Delta_2}{2}} \bar{w}_\ell^{\frac{\Delta_2}{2}} \right] \tilde{\mathcal{R}}_{F_A} \quad (\text{E.29})$$

$$\leq \left[4\bar{W} + b \prod_i \rho_i s_i \right]^{\max(\Delta_1, \Delta_2)} \tilde{\mathcal{R}}_{F_A} \quad (\text{E.30})$$

where at equation (E.25) we have used equation (E.20), at equation (E.26) we have used equation (E.19), at equation (E.27) we have used the trivial fact that $\frac{p_\ell}{p_\ell+2} \geq \frac{p_\ell}{2+p_\ell}$ and at equation (E.28) we have used equation (E.21) and the fact that $r_\ell \leq \bar{w}_\ell$.

Equation (E.30) motivates the choice

$$\Delta_1 = \Delta_2 = \frac{\log(2)}{\log[4\bar{W} + b \prod_i \rho_i s_i]}, \quad (\text{E.31})$$

which ensures that

$$\mathcal{R}_{k^{(p_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(b)}, k^{(s_\ell)}} \leq 2\tilde{\mathcal{R}}_{F_A}. \quad (\text{E.32})$$

Note that for this choice of Δ_1 and Δ_2 , we can write first (from equation (E.17)),

$$\underline{\Gamma}_{k^{(b)}, k^{(s_\ell)}} \leq \underline{\Gamma} 4 \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \quad (\text{E.33})$$

and then:

$$\begin{aligned} & \log \left(\frac{1}{\delta_{k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}}} \right) \\ & \leq \log\left(\frac{1}{\delta}\right) + [3L + 1] \log(3) + \left[|k^{(b)}| + \sum_\ell |k^{(s_\ell)}| + \sum_\ell |k^{(\mathcal{M}_\ell)}| \right] + L \log(1/\Delta_2) \\ & \leq \log\left(\frac{1}{\delta}\right) + L \left[4.3 + \log \log \left[4\bar{W} + b \prod_i \rho_i s_i \right] \right] + \left[|k^{(b)}| + \sum_\ell |k^{(s_\ell)}| + \sum_\ell |k^{(\mathcal{M}_\ell)}| \right] \quad (\text{E.34}) \end{aligned}$$

$$\leq \log\left(\frac{1}{\delta}\right) + L \left[4.3 + \log \log \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \right] + 2 \quad (\text{E.35})$$

$$+ \left[|\log(b)| + L \left[\sum_{\ell=1}^L |\log(\|A_\ell\|)| + \log(\bar{W}) \right] \right] / \Delta_1 \quad (\text{E.36})$$

$$\leq \log\left(\frac{1}{\delta}\right) + L \left[4.3 + \log \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \right] + 2 \quad (\text{E.37})$$

$$+ 2 \log \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \left[|\log(b)| + L \left[\sum_{\ell=1}^L |\log(\|A_\ell\|)| + \log(\bar{W}) \right] \right]$$

$$\leq \log\left(\frac{1}{\delta}\right) + 2 + 2 \log \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \left[|\log(b)| + L \left[\sum_{\ell=1}^L |\log(\|A_\ell\|)| + 2 \log(4\bar{W}) \right] \right]$$

$$= \log\left(\frac{1}{\delta}\right) + 2 + \Theta_{\log} \quad (\text{E.38})$$

where at equation (E.34) we have used the definition of Δ_2 (i.e. equation (E.31)) and the fact $\log(\log(2)) \geq -1$, at equation (E.36) we have used equations (E.16), and at equation (E.37) we have used the fact that $\log(2) \geq \frac{1}{2}$ and the inequality $\log \log(x) \leq \log(x)$ and at equation (E.38) we have used the definition of Θ_{\log} .

We now plug equations (E.30), (E.32), (E.38) and (E.33) into equation (E.14) to obtain:

$$\begin{aligned}
& \mathbb{E}[\mathbf{l}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbf{l}(F_A(x_i), y_i) \leq 6[\mathcal{B}+1] \sqrt{\frac{\log\left(\frac{4}{\delta_{k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}}}\right)}{N}} \\
& + \frac{240 \sqrt{\log_2(2\Gamma_{k^{(b)}, k^{(s_\ell)}}) \log(N)} [\mathcal{B}+1]}{\sqrt{N}} (1 + L \mathbf{L}_1)^{\frac{p}{2+p}} \mathcal{R}_{k^{(p_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(b)}, k^{(s_\ell)}} \\
& \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(1/\delta)}{N}} + 6[\mathcal{B}+1] \sqrt{\frac{2 + \Theta_{\log}}{N}} \\
& + 480 \mathcal{B} (1 + L \mathbf{L}_1)^{\frac{p}{2+p}} \frac{\sqrt{\log_2(8\Gamma[4\bar{W} + b \prod_i \rho_i \|A_\ell\|]) \log(N)}}{N} \tilde{\mathcal{R}}_{F_A} \tag{E.39}
\end{aligned}$$

$$\begin{aligned}
& \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(1/\delta)}{N}} + 6[\mathcal{B}+1] \sqrt{\frac{2 + \Theta_{\log}}{N}} + 480 \mathcal{B} (1 + L \mathbf{L}_1)^{\frac{p}{2+p}} \times \\
& \sqrt{\frac{\log(12[b+1][L \mathbf{L}_1+1]\bar{W}N \prod_{i=1}^L [\rho_i + 1] s_i + 1] [4\bar{W} + b \prod_i \rho_i \|A_\ell\|]) \log(N)}{N}} \tilde{\mathcal{R}}_{F_A} \\
& \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(1/\delta)}{N}} + 6[\mathcal{B}+1] \sqrt{\frac{2 + \Theta_{\log}}{N}} \\
& + 480 \mathcal{B} (1 + L \mathbf{L}_1)^{\frac{p}{2+p}} \frac{\sqrt{\log(\gamma_{FC}) \log(N)}}{N} \tilde{\mathcal{R}}_{F_A}, \tag{E.40}
\end{aligned}$$

where at line (E.39) we have used equation (E.32), and at the last line we have used the definition

$$\gamma_{FC} := 12 \left[[b+1][L \mathbf{L}_1+1]\bar{W}N \prod_{i=1}^L [\rho_i \|A_i\| + 1] \right] \left[4\bar{W} + b \prod_i \rho_i \|A_i\| \right].$$

This concludes the proof of the first statement.

For the second statement, it suffices to use an elementary form of loss function augmentation by considering the following augmented loss for any value of b greater than 1: $l'(x, y) = \max(\mathbf{l}(F_A(x), y), \mathcal{B} \mathbf{1}_{\|x\| > b})$. Indeed, given any training set x_1, \dots, x_N (not necessarily satisfying $\|x_i\| \leq b$), we can apply Proposition E.11 to the training set $\{x_i : \|x_i\| \leq b\}$, which results in a cover of $l' \circ \mathcal{F}_{\mathcal{B}}$ with respect to the whole set $\{x_i\}$ since $l' = \mathcal{B}$ whenever $\|x_i\| \geq b$. The cover is bounded above by the RHS of equation (E.131) and we can continue the argument as in the proof above and conclude that

$$\mathbb{E}[l'(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N l'(F_A(x_i), y_i) \leq \mathcal{Q}, \tag{E.41}$$

where \mathcal{Q} is the right hand side of equation E.40. Note that the argument above already uses a posthoc argument to show that the bound holds for all values of b simultaneously with a single failure probability. Thus, the same is true for equation (E.41). Note that it is not necessary for l' to be Lipschitz in the input x , since no cover of input space is required.

Next, we have:

$$\begin{aligned}
& \mathbb{E}[\mathbf{l}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbf{l}(F_A(x_i), y_i) \\
& \leq \mathbb{E}[l'(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N l'(F_A(x_i), y_i) \tag{E.42}
\end{aligned}$$

$$\leq \mathcal{Q} + \frac{1}{N} \sum_{i=1}^N l'(F_A(x_i)) - \frac{1}{N} \sum_{i=1}^N \mathbf{l}(F_A(x_i), y_i) \tag{E.43}$$

$$\leq \mathcal{Q}. \tag{E.44}$$

where at the second line E.42, we have used the fact that $l' \geq 1$, at the third line (E.43) we have used equation (E.41) and at the fourth line (E.44) we have used the fact that $b = \max_i \|x_i\|$. This completes the proof. \square

E.1 Tighter Results with Loss Function Augmentation

In this section, we use the technique of Loss Function Augmentation to improve the scaling of the bound with respect to norm based quantities.

We first recall the following definition of the ramp loss (defined for any B), which we use in controlling the intermediary activations of the neural network:

$$\lambda_B(x) = \min(B \frac{(x - B)_+}{B}, B). \quad (\text{E.45})$$

We now consider the following augmented loss function:

$$l_{\text{aug}}(F_A, x) := \max \left(\max_{\ell \leq L-1} \left[\lambda_{b_\ell}(\|F_A^{0 \rightarrow \ell}(x)\|) \right], l(F_A(x), y), B 1_{\|x\| > b_0} \right). \quad (\text{E.46})$$

We first prove the following extension of Lemma E.12 to incorporate the loss function augmentation over activations.

Lemma E.3. *Assume $s_\ell \geq 1$ and let b_ℓ (for $\ell = 0, 1, \dots, L$) be some positive real numbers with $b_L = 1/L$. Define $\hat{\rho}_\ell = \sup_{u \geq l} \rho_{l \rightarrow u} / b_u$ where as before, $\rho_{l \rightarrow u} = \rho_l \prod_{m=l+1}^u s_m \rho_m$.*

Assume as in Lemma E.12 that we have a fixed dataset x_1, \dots, x_N with $\|x_i\| \leq b_0$ for all $i \leq N$. Let $\mathcal{B}_1, \dots, \mathcal{B}_L$ be arbitrary subsets of the spaces $\mathbb{R}^{w_\ell \times w_{\ell-1}}$ and instantiate the rest of the notations from Lemma E.12, and explicitly set $\epsilon_\ell = \frac{b_\ell \epsilon}{L \hat{\rho}_\ell}$.

There exist covers $\mathcal{C}_1 \subset \mathcal{B}_1, \dots, \mathcal{C}_{0 \rightarrow \ell} \subset \mathcal{B}_1 \times \dots \mathcal{B}_\ell, \dots, \mathcal{C}_L \subset \mathcal{B}_1 \times \dots \mathcal{B}_L$ such that for all $A = (A_1, \dots, A_L) \in \mathcal{B}_1 \times \dots \mathcal{B}_L$, there exist $\bar{A}^1, \dots, \bar{A}^L$ such that for all $\ell \leq L$, $\bar{A} = \bar{A}^{0 \rightarrow L} := (\bar{A}^1, \dots, \bar{A}^L) \in \mathcal{C}_{0 \rightarrow L}$ and for all $i \leq N, \ell \leq L$:

$$|\lambda_{\text{aug}}(F_A, x_i, y) - \lambda_{\text{aug}}(\bar{A}, x_i, y)| \leq \epsilon. \quad (\text{E.47})$$

Furthermore the covers satisfy the following covering number bound:

$$\log(|\mathcal{C}_{0 \rightarrow L}|) \leq \sum_{\ell=1}^L \log(\mathcal{N}_{\infty, \ell}(\mathcal{B}_\ell, \epsilon_\ell, N, 3b_{\ell-1})), \quad (\text{E.48})$$

where as before, for any $\bar{b} > 0$ and $\bar{N} \in \mathbb{N}$, $\mathcal{N}_{\infty, \ell}(\mathcal{B}_\ell, \epsilon_\ell, N, \bar{b})$ is defined as the minimum number \mathcal{N} such that for any dataset $x_1^{\ell-1}, \dots, x_{\bar{N}}^{\ell-1}$ with $\bar{N} \leq N$ and $\|x_i^{\ell-1}\|_\ell \leq \bar{b}$ for all $i \leq \bar{N}$, where $\|\cdot\|_\ell = \|\cdot\|$ for $\ell \neq L$ and $\|\cdot\|_L = \|\cdot\|_\infty$.

Proof. The proof is quite similar to the proof of Proposition 8 in [42], itself inspired from [74, 5, 16].

We will construct the covers such that they satisfy the following claim:

Claim 1: There exists a cover $\mathcal{C}_{0 \rightarrow L}$ satisfying equation (E.48) such that for any $A = (A^1, \dots, A^L) \in \mathcal{B}_1 \times \dots \mathcal{B}_L$, there exists a $\bar{A}^1, \dots, \bar{A}^L$ such that for each $x \in X$ and each $\ell \leq L$, one of the following two conditions is satisfied:

$$\forall \ell_1 \leq \ell, \|F_{A^1, \dots, \bar{A}^{\ell_1-1}, A^{\ell_1}}^{0 \rightarrow \ell_1}(x) - F_{\bar{A}^1, \dots, \bar{A}^{\ell_1-1}, \bar{A}^{\ell_1}}^{0 \rightarrow \ell}(x)\|_\ell \leq \epsilon_\ell \rho_\ell \quad (\text{E.49})$$

or there exists $\ell_1 < \ell$ such that

$$\|F_{\bar{A}^1, \dots, \bar{A}^{\ell_1}}^{0 \rightarrow \ell_1}(x)\|_\ell > 3b_{\ell_1} \quad (\text{E.50})$$

(Here as usual $\|\cdot\|_\ell = \|\cdot\|_2$ if $\ell \neq L$ and $\|\cdot\|_L = \|\cdot\|_\infty$).

Proof of Claim 1:

For the **first layer**, writing $X := \{x_1, \dots, x_N\}$, we know by assumption that there exists a cover $\mathcal{C}_1(X) = \mathcal{C}_{0 \rightarrow 1} \subset \mathcal{B}_1$ such that for all $A_1 \in \mathcal{B}_1$, there exists a $\bar{A}_1 \in \mathcal{C}_1(X_0)$ such that for all $i \leq N$,

$$\|(A^1 - \bar{A}^1)(x_i)\| \leq \epsilon_1, \quad (\text{E.51})$$

from which it also follows that

$$\|F_{A^1}^{0 \rightarrow 1}(x_1) - F_{\bar{A}^1}^{0 \rightarrow 1}(x_1)\| = \|\sigma_1(A^1 x_i) - \sigma_1(\bar{A}^1(x_i))\| \leq \epsilon_1 \rho_1. \quad (\text{E.52})$$

It follows that $\{F_{A^1}^{0 \rightarrow 1} : A^1 \in \mathcal{B}_1\} \subset \mathcal{F}_B^{0 \rightarrow 1}$ is an $L^{\infty, 2}$ cover of $\mathcal{F}_B^{0 \rightarrow 1}$ as required: equation (E.49) holds for $\ell = 1$.

For the **inductive case**:

Assume that the covers $\mathcal{C}_{0 \rightarrow 1}, \dots, \mathcal{C}_{0 \rightarrow \ell}$ have been constructed and satisfy the claim up to and including layer ℓ .

For each element $\bar{A}^{0 \rightarrow \ell}$ of $\mathcal{C}_{0 \rightarrow \ell}$, we can split the dataset X into two sets $X_{\ell, \bar{A}^{0 \rightarrow \ell}} \subset X$ and $[X_{\ell, \bar{A}^{0 \rightarrow \ell}}]^c$ as follows:

$$X_{\ell, \bar{A}^{0 \rightarrow \ell}} = \left\{ x \in X : \forall \ell_1 \leq \ell : \|F_{\bar{A}^1, \dots, \bar{A}^{\ell_1}}^{0 \rightarrow \ell_1}(x_i)\|_{\ell_1} \leq 3b_{\ell_1} \right\}. \quad (\text{E.53})$$

Next, by the definition of $\mathcal{N}_{\infty, \ell}(\mathcal{B}_\ell, \epsilon_\ell, N, 3b_{\ell-1})$, there exists a cover $\mathcal{C}_{\ell+1, \bar{A}^{0 \rightarrow \ell}} \subset \mathcal{B}_{\ell+1}$ such that for any $A^{\ell+1} \in \mathcal{B}_{\ell+1}$, there exists $\bar{A}^{\ell+1} \in \mathcal{C}_{\ell, \bar{A}^{0 \rightarrow \ell}}$ such that for any $x \in X_{\ell, \bar{A}^{0 \rightarrow \ell}}$,

$$\|A^{\ell+1} F_{\bar{A}^{0 \rightarrow \ell}}^{0 \rightarrow \ell}(x) - \bar{A}^{\ell+1} F_{\bar{A}^{0 \rightarrow \ell}}^{0 \rightarrow \ell}(x)\|_\ell \leq \epsilon_\ell. \quad (\text{E.54})$$

Accordingly, we define the cover

$$\mathcal{C}_{0 \rightarrow \ell+1} = \bigcup_{\bar{A}^{0 \rightarrow \ell} \in \mathcal{C}_\ell} \{\bar{A}^{0 \rightarrow \ell}\} \times \mathcal{C}_{\ell+1, \bar{A}^{0 \rightarrow \ell}}. \quad (\text{E.55})$$

For any $A^{0 \rightarrow \ell+1} = (A^1, \dots, A^{\ell+1}) \in \mathcal{B}_1 \times \dots \times \mathcal{B}_{\ell+1}$, the associated cover element $\bar{A}^{0 \rightarrow \ell+1}$ is defined by $(\bar{A}^1, \dots, \bar{A}^\ell, \bar{A}^{\ell+1})$ where $(\bar{A}^1, \dots, \bar{A}^\ell)$ is the cover element of $\mathcal{C}_{0 \rightarrow \ell}$ associated to (A^1, \dots, A^ℓ) defined by the induction hypothesis, and $\bar{A}^{\ell+1}$ is the cover element of $\mathcal{C}_{\ell+1}$ associated to $A^{\ell+1}$ by the definition of $\mathcal{N}_{\infty, \ell}(\mathcal{B}_\ell, \epsilon_\ell, N, 3b_{\ell-1})$ to satisfy Condition (E.54).

Equation (E.54) implies that for any $x \in X_{\ell, \bar{A}^{0 \rightarrow \ell}}$, condition (E.49) holds for the value $\ell_1 = \ell$. In addition, we already know from the induction hypothesis that the same condition holds for $\ell_1 < \ell$. Therefore, the condition holds for all $\ell_1 < \ell + 1$ for all $x \in X_{\ell, \bar{A}^{0 \rightarrow \ell}}$. Furthermore, by definition of $X_{\ell, \bar{A}^{0 \rightarrow \ell}}$, any $x \in [X_{\ell, \bar{A}^{0 \rightarrow \ell}}]^c$ has to satisfy equation (E.50) for some $\ell_1 < \ell + 1$. Thus, every $x \in X$ indeed satisfies either condition (E.49) or condition (E.50), as desired.

This concludes the proof of the *claim*.

To finish the proof of the *lemma*, it only remains to prove equation E.47. Consider an arbitrary $A \in \mathcal{B}_1 \times \dots \times \mathcal{B}_L$ and the associated cover element $\bar{A} = (\bar{A}^1, \bar{A}^2, \dots, \bar{A}^L)$.

Let $x \in X$ be an arbitrary sample. Let ℓ^* be the smallest number less than $L - 1$ such that

$$\|F_{\bar{A}^1, \dots, \bar{A}^{\ell^*}}^{0 \rightarrow \ell^*}(x)\|_\ell > 3b_\ell. \quad (\text{E.56})$$

and let $\ell^* = L$ if the above equation doesn't hold for any value of ℓ .

By instantiating the claim for any $\ell \leq \ell^*$, we know that equation (E.49) holds for all $\ell_1 \leq \ell$. Thus,

$$|\lambda_{b_\ell}(\|F_{A^0 \rightarrow \ell}^{0 \rightarrow \ell}(x)\|) - \lambda_{b_\ell}(\|F_{\bar{A}^0 \rightarrow \ell}^{0 \rightarrow \ell}(x)\|)| \quad (\text{E.57})$$

$$\leq \frac{1}{b_\ell} \|F_{A^0 \rightarrow \ell}^{0 \rightarrow \ell}(x) - F_{\bar{A}^0 \rightarrow \ell}^{0 \rightarrow \ell}(x)\| \quad (\text{E.58})$$

$$\leq \frac{1}{b_\ell} \sum_{\ell_1=1}^{\ell} \left\| F_{\bar{A}^1, \dots, \bar{A}^{\ell_1}, A^{\ell_1+1}, \dots, A^\ell}^{0 \rightarrow \ell}(x) - F_{\bar{A}^1, \dots, \bar{A}^{\ell_1-1}, A^{\ell_1}, \dots, A^\ell}^{0 \rightarrow \ell}(x) \right\| \quad (\text{E.59})$$

$$\leq \frac{1}{b_\ell} \sum_{\ell_1=1}^{\ell} \epsilon_{\ell_1} \rho_{\ell_1 \rightarrow \ell} \quad (\text{E.60})$$

$$= \frac{1}{b_\ell} \sum_{\ell_1=1}^{\ell} \frac{\epsilon}{L \rho_{\hat{\rho}_{\ell_1}}} \rho_{\ell_1 \rightarrow \ell} \quad (\text{E.61})$$

$$\leq \frac{1}{b_\ell} \sum_{\ell_1=1}^{\ell} \frac{\epsilon}{L \rho_{\ell_1 \rightarrow \ell} / b_\ell} \rho_{\ell_1 \rightarrow \ell} \leq \frac{\ell}{L} \epsilon \leq \epsilon, \quad (\text{E.62})$$

where at equation (E.60) we have used the claim, at line (E.61) we have used the definition of ϵ_{ℓ_1} , and at line (E.62) we have used the definition of $\hat{\rho}_{\ell_1}$.

In particular, it certainly follows that if $\ell \leq L-1$,

$$\|F_{A^0 \rightarrow \ell}^{0 \rightarrow \ell}(x)\| \geq \|F_{\bar{A}^0 \rightarrow \ell}^{0 \rightarrow \ell}(x)\| - \|F_{A^0 \rightarrow \ell}^{0 \rightarrow \ell}(x) - F_{\bar{A}^0 \rightarrow \ell}^{0 \rightarrow \ell}(x)\| > 3b_\ell - b_\ell = 2b_\ell, \quad (\text{E.63})$$

which implies that $\lambda_{\text{aug}}(F_A, x, y) = \lambda_{\text{aug}}(F_{\bar{A}}, x, y) = 0$ and equation (E.47) indeed holds. On the other hand, if $\ell^* = L$, then since $b_L = 1/L$, equation (E.62) instantiated for $\ell = \ell^* = L$ also shows that

$$|\text{l}(F_A(x), y) - \text{l}(F_{\bar{A}}(x), y)| \leq \epsilon, \quad (\text{E.64})$$

whilst (E.62) instantiated for any $\ell \leq L-1$ also shows that

$$\left| \max_{\ell \leq L-1} [\lambda_{b_\ell}(\|F_A^{0 \rightarrow \ell}(x)\|)] - \max_{\ell \leq L-1} [\lambda_{b_\ell}(\|F_{\bar{A}}^{0 \rightarrow \ell}(x)\|)] \right| \leq \epsilon. \quad (\text{E.65})$$

Together, equations (E.64) and (E.65) imply that equation (E.47) holds, as expected. \square

Proposition E.4. *For any granularity $\epsilon \geq \frac{1}{N}$ and any values of $1 \leq b_{\ell_1}, \dots, b_{L-1}, s_\ell, \rho_\ell$, there exists a cover \mathcal{C} of the augmented loss class $\{\lambda_{\text{aug}}(F_A, x, y) : A = (A^1, \dots, A^L) \in \mathcal{B}_1 \times \dots \times \mathcal{B}_L\}$ with cardinality bounded as follows:*

$$\log(|\mathcal{C}|) \leq 288 \log_2(4L) \left[\frac{L(1+L)}{\min(\epsilon, 1)} \right]^{\frac{2p}{2+p}} \sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L s_i \rho_i \right]^{\frac{2p_\ell}{p_\ell+2}} r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}}. \quad (\text{E.66})$$

Proof. The proof is analogous to that of proposition E.11 and consists in plugging in the one layer covering number estimates (Propositions (F.4) and (F.2)) into equation (E.48), taking into account our new definition of $\epsilon_\ell = \frac{b_\ell \epsilon}{L \hat{\rho}_\ell}$:

$$\begin{aligned} \log(|\mathcal{C}|) &\leq \sum_{\ell=1}^L \log(\mathcal{N}_{\infty, \ell}(\mathcal{B}_\ell, \epsilon_\ell, N, 3b_{\ell-1})) \\ &\leq 72 \left[\sum_{\ell=1}^{L-1} \left[\frac{\mathcal{M}_\ell b_{\ell-1}}{\epsilon_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} [w_\ell \underline{w}_\ell]^{\frac{p_\ell}{p_\ell+2}} \log_2(\Gamma_{\mathcal{F}_r^p, \ell}) \right. \\ &\quad \left. + \left[\frac{\mathcal{M}_L b_{L-1}}{\epsilon_L} \right]^{\frac{2p_L}{p_L+2}} \bar{w}_L^{\frac{2}{p_L+2}} \underline{w}_L^{\frac{p_L}{p_L+2}} \log_2(\Gamma_{\mathcal{F}_r^p, L}) \right]. \end{aligned} \quad (\text{E.67})$$

Note also that since we are assuming that $s_\ell \geq 1$ for all ℓ and $b_\ell \geq 1$ for all $\ell \neq L$, we also have $\epsilon_\ell = \frac{b_\ell \epsilon}{L \hat{\rho}_\ell} \geq \epsilon \rho_{\ell \rightarrow L}$, thus, the bounds on the logarithmic factors $\Gamma_{\mathcal{F}_r^p, L}, \Gamma_{\mathcal{F}_r^p, \ell}$ derived in equations (E.141) still hold. Thus, we can continue:

$$\log(|\mathcal{C}|) \leq 72 \log_2(\bar{\Gamma}) \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b_{\ell-1}}{\epsilon_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \quad (\text{E.69})$$

$$\leq 72 \log_2(\bar{\Gamma}) \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b_{\ell-1} \rho_\ell L [1 + L_1] \prod_{i=\ell+1}^L s_i \rho_i}{\epsilon} \right]^{\frac{2p_\ell}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \quad (\text{E.70})$$

$$\leq 72 \log_2(\bar{\Gamma}) \left[\frac{L(1 + L_1)}{\min(\epsilon, 1)} \right]^{\frac{2p}{2+p}} \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell}{s_\ell} b_{\ell-1} \prod_{i=\ell}^L s_i \rho_i \right]^{\frac{2p_\ell}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \quad (\text{E.71})$$

$$= 288 \log_2(4\bar{\Gamma}) \left[\frac{L(1 + L_1)}{\min(\epsilon, 1)} \right]^{\frac{2p}{2+p}} \sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L s_i \rho_i \right]^{\frac{2p_\ell}{p_\ell+2}} r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}}, \quad (\text{E.72})$$

as expected. \square

Then, we can proceed with the following straightforward analogue of Proposition E.1:

Proposition E.5. *Fix some reference/initialization matrices $M^1 \in \mathbb{R}^{w_1 \times w_0} = \mathbb{R}^{w_1 \times d}, \dots, M^L \in \mathbb{R}^{w_L \times w_{L-1}} = \mathbb{R}^{\mathcal{C} \times w_{L-1}}$. Assume also that the inputs $x \in \mathbb{R}^d$ satisfy $\|x\| \leq b$ w.p. 1. Fix a set of constraint parameters $0 \leq p_\ell \leq 2, 1 \leq s_\ell, \mathcal{M}_\ell$ (for $\ell = 1, \dots, L$). Also fix some numbers $1 \leq b_{\ell_1}, \dots, b_{L-1}$ and $b_L = \frac{1}{L_1}$.*

With probability greater than $1 - \delta$ over the draw of an i.i.d. training set x_1, \dots, x_N , every neural network $F_A(x) := A_L \sigma_L(A^{L-1} \sigma_{L-1}(\dots \sigma_1(A^1 x) \dots))$ satisfying the following conditions:

$$\begin{aligned} \|A^\ell - M^\ell\|_{\text{sc}, p_\ell} &\leq \mathcal{M}_\ell & \forall 1 \leq \ell \leq L \\ \|A\| &\leq s_\ell & \forall 1 \leq \ell \leq L-1 \end{aligned} \quad (\text{E.73})$$

$$\|A_L^\top\|_{2, \infty} \leq s_L \quad (\text{E.74})$$

also satisfies the following generalization bound:

$$\begin{aligned} \mathbb{E}[\text{l}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \text{l}(F_A(x_i), y_i) \\ \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta})}{N}} + \frac{416\sqrt{\log_2(4\bar{\Gamma})} \log(N)[\mathcal{B}+1]}{\sqrt{N}} [L(1 + L_1)]^{\frac{p}{2+p}} \mathcal{R}_{\mathcal{M}, s, p, b_\ell} \end{aligned} \quad (\text{E.75})$$

where $\mathcal{R}_{\mathcal{M}, s, p, b_\ell} :=$

$$\left[\sum_{\ell=1}^{L-1} \left[b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{p_\ell}{p_\ell+2}} [w_\ell \underline{w}_\ell]^{\frac{p_\ell}{p_\ell+2}} + [b_{L-1} \rho_L]^{\frac{2p_L}{p_L+2}} r_\ell^{\frac{2}{p_L+2}} \bar{w}_L^{\frac{2}{p_L+2}} \underline{w}_L^{\frac{p_L}{p_L+2}} \right]^{\frac{1}{2}}, \quad (\text{E.76})$$

and as usual, $\underline{\Gamma} := \left[[b+1][L L_1 + 1] \bar{W} N \prod_{i=1}^L [\rho_i + 1] s_i + 1 \right]$ and we use the shorthands $r_\ell := \frac{\mathcal{M}_\ell p}{s_\ell^p}, \underline{w}_\ell := \min(w_\ell, w_{\ell-1}), \bar{w}_\ell := [w_\ell + w_{\ell-1}]$.

Proof. We first note that with a calculation nearly identical to that of the proof of Proposition E.1 relying on Proposition E.4 instead of Proposition E.11 with an additional factor of $\sqrt{3}$ and $(1 + L L_1)$

replaced by $L(1 + L_1)$ as well as $b \prod_{i=1}^L s_i \rho_i$ replaced by $b_{\ell-1} \prod_{i=\ell}^L s_i \rho_i$, we obtain the following bound on the augmented loss class λ_{aug} :

$$\begin{aligned} & \mathbb{E} [\lambda_{aug}(F_A, x, y)] - \frac{1}{N} \sum_{i=1}^N \lambda_{aug}(F_A, x_i, y_i) \\ & \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta})}{N}} + \frac{832\sqrt{\log_2(4\Gamma)} \log(N)[\mathcal{B}+1]}{\sqrt{N}} [L(1 + L_1)]^{\frac{p}{2+p}} \mathcal{R}_{\mathcal{M},s,p,b_\ell}. \end{aligned} \quad (\text{E.77})$$

However, we know by definition of λ_{aug} that

$$\mathbb{E} [\lambda_{aug}(F_A, x, y)] \geq \mathbb{E} l(F_A(x), y). \quad (\text{E.78})$$

Furthermore, since the last condition in (E.73) holds, we certainly also have:

$$\frac{1}{N} \sum_{i=1}^N \lambda_{aug}(F_A, x_i, y_i) = \frac{1}{N} l(F_A(x_i), y_i). \quad (\text{E.79})$$

The result now follows. \square

We can now apply the above results to obtain a post hoc version using argument similar to those used in the proof of Theorem E.2

Theorem E.6. *With probability greater than $1 - \delta$, for any $b \in \mathbb{R}^+$, every trained neural network F_A satisfies the following generalization gap for all possible values of the p_ℓ s:*

$$\begin{aligned} & \mathbb{E} [l(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N l(F_A(x_i), y_i) \\ & \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(1/\delta)}{N}} + 6\mathcal{B} \sqrt{\frac{1.5 + \Theta_{\log}^{\lambda_{aug}}}{N}} \\ & \quad + 832[\mathcal{B}+1] [L(1 + L_1)]^{\frac{p}{2+p}} \frac{\sqrt{\log(\gamma_{FC})} \log(N)}{\sqrt{N}} \tilde{\mathcal{R}}_{F_A, \lambda_{aug}}, \end{aligned} \quad (\text{E.80})$$

where $\Theta_{\log}^{\lambda_{aug}} :=$

$$2 \log \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \left[|\log(b)| + L \left[\sum_{\ell=1}^L |\log(\|A_\ell\|)| + \sum_{\ell=1}^{L-1} \log(\mathfrak{B}_{\ell,A}) + 2 \log(4\bar{W}) \right] \right],$$

$\gamma_{FC} := 12 \left[[b+1][L L_1 + 1] \bar{W} N \prod_{i=1}^L [\rho_i \|A_i\| + 1] \right] [4\bar{W} + b \prod_i \rho_i \|A_i\|], b = \max_{i=1}^N \|x_i\|$, and where

$$\tilde{\mathcal{R}}_{F_A, \lambda_{aug}} := \quad (\text{E.81})$$

$$\left[\sum_{\ell=1}^{L-1} \left[\mathfrak{B}_{\ell-1,A} \rho_L \|A_L^\top\|_{2,\infty} \prod_{i=\ell}^{L-1} \rho_i \|A_i\| \right]^{\frac{2p_\ell}{p_\ell+2}} \left[\frac{\|A_\ell - M_\ell\|_{\text{sc},p_\ell}^{p_\ell}}{\|A_\ell\|^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} [w_\ell \underline{w}_\ell]^{\frac{p_\ell}{p_\ell+2}} \right] \quad (\text{E.82})$$

$$+ \left[\mathfrak{B}_{L-1,A} \rho_L \|A_L^\top\|_{2,\infty}^{\frac{2p_L}{p_L+2}} \left[\frac{\|A_L - M_L\|_{\text{sc},p_L}^{p_L}}{\|A_L^\top\|_{2,\infty}^{p_L}} \right]^{\frac{2}{p_L+2}} \bar{w}_L^{\frac{2}{p_L+2}} \underline{w}_L^{\frac{p_L}{p_L+2}} \right]^{\frac{1}{2}}. \quad (\text{E.83})$$

where $\mathfrak{B}_{\ell-1,A} := \max(\max_{i \leq N} \|F_{A^1, \dots, A^{\ell-1}}(x_i)\|, 1)$.

Proof. Let $\Delta_1, \Delta_2 > 0$ be positive real numbers to be determined later. For every set of values $k^{(b)}, k^{(b_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(s_\ell)} \in \mathbb{Z}$ and $0 \leq k^{(p_\ell)} \leq \frac{2}{\Delta_2}$ (where $\ell = 1, \dots, L$) we apply proposition E.11

for the following values of $b, k^{(s_\ell)}, \mathcal{M}_\ell, p_\ell, \delta$:

$$b \leftarrow \tilde{b} = \exp(\Delta_1 k^{(b)}) \quad (\text{E.84})$$

$$b_\ell \leftarrow \tilde{b}_\ell = \exp(\Delta_1 k^{(b_\ell)})$$

$$s_\ell \leftarrow \tilde{s}_\ell = \exp(\Delta_1 k^{(s_\ell)}/L)$$

$$\mathcal{M}_\ell \leftarrow \tilde{\mathcal{M}}_\ell = \tilde{s}_\ell \tilde{r}_\ell^{\frac{1}{p}} \quad \text{with}$$

$$\tilde{r}_\ell := \exp(\Delta_1 k^{(\mathcal{M}_\ell)})$$

$$p_\ell \leftarrow \tilde{p}_\ell := \Delta_2 k^{(p_\ell)}$$

$$\delta_{k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}} = \delta \Delta_2^L \frac{1}{3^{4L+1}} 2^{-|k^{(b)}| - \sum_\ell |k^{(b_\ell)}| - \sum_\ell |k^{(s_\ell)}| - \sum_\ell |k^{(\mathcal{M}_\ell)}|}. \quad (\text{E.85})$$

Using Proposition E.5, this implies that with probability greater than $1 - \delta_{k^{(b)}, k^{(b_\ell)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}}$,

$$\begin{aligned} \mathbb{E}[\mathbb{I}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbb{I}(F_A(x_i), y_i) &\leq 6 \mathcal{B} \sqrt{\frac{\log(\frac{4}{\delta_{k^{(b)}, k^{(b_\ell)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}}})}{N}} \\ &+ \frac{416 \sqrt{\log_2(4 \underline{\Gamma}_{k^{(b)}, k^{(s_\ell)}}) \log(N) [\mathcal{B} + 1]}}{\sqrt{N}} [L(1 + \mathbb{L}_1)]^{\frac{p}{2+p}} \mathcal{R}_{k^{(p_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(b)}, k^{(s_\ell)}}, \end{aligned} \quad (\text{E.86})$$

where $\mathcal{R}_{k^{(b)}, k^{(b_\ell)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}}$ is defined as the value of $\mathcal{R}_{\mathcal{M}, s, p, b_\ell}$ for $b, b_\ell, s_\ell, \mathcal{M}_\ell, p_\ell$ taking the values defined in equations (E.84) and $\underline{\Gamma}_{k^{(b)}, k^{(s_\ell)}} = \left[[\exp(k^{(b)} \Delta) + 1] [\mathbb{L}_1 + 1] \bar{W} N \prod_{\ell=1}^L [\rho_\ell \exp(k^{(s_\ell)} \Delta) + 1] \right]$.

Note that

$$\sum_{k^{(b)}, k^{(b_\ell)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}} \delta_{k^{(b)}, k^{(b_\ell)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}} = \delta \Delta_2^L \frac{1}{3^{4L+1}} \left[\frac{2}{\Delta_2} \right]^L \left[\sum_{i \in \mathbb{Z}} \frac{1}{2^i} \right]^{3L+1} \leq \delta. \quad (\text{E.87})$$

Thus, equation (E.86) holds simultaneously over all admissible values of $k^{(b)}, k^{(b_\ell)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}$ with probability $\geq 1 - \delta$.

For a given neural network F_A and choice of p_ℓ s, we utilise the result above for the following values of $k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}$:

$$k^{(b)} := \lceil \log(b)/\Delta_1 \rceil \quad (\text{E.88})$$

$$k^{(b_\ell)} = \lceil \log(\mathfrak{B}_{\ell, A})/\Delta_1 \rceil \quad (\text{E.89})$$

$$k^{(s_\ell)} = \lceil \log(s_\ell) L/\Delta_1 \rceil$$

$$k^{(\mathcal{M}_\ell)} = \lceil \log(r_\ell)/\Delta_1 \rceil$$

$$k^{(p_\ell)} := \left\lceil \frac{p_\ell}{\Delta_2} \right\rceil,$$

where $b \leftarrow \max_i \|x_i\|$, $b_\ell \leftarrow \mathfrak{B}_{\ell, A}$, $s_\ell \leftarrow \|A_\ell\|$, $r_\ell \leftarrow \|A_\ell - M_\ell\|_{\text{sc}, p_\ell}^{p_\ell} / \|A\|^{p_\ell} = \sum_i \sigma_i (A - M_\ell)^{p_\ell}$ (for $\ell \neq L$) and $r_L \leftarrow \|A_L - M_L\|_{\text{sc}, p_L}^{p_L} / \|A_L^\top\|_{2, \infty}^{p_L} \leq \sqrt{w_L} \sum_i \sigma_i (A_L - M_L)^{p_L}$.

As in the case of Theorem E.2, the following equation still holds as it doesn't involve $k^{(b_\ell)}$ (cf. Eq. (E.17))

$$\underline{\Gamma}_{k^{(b)}, k^{(s_\ell)}} \leq \left[[\exp(\Delta_1) b + 1] [\mathbb{L}_1 + 1] \bar{W} N \prod_{i=1}^L [\rho_i s_i \exp(\Delta_1/L) + 1] \right] \quad (\text{E.90})$$

$$\leq \exp(2\Delta_1) \left[[b + 1] [\mathbb{L}_1 + 1] \bar{W} N \prod_{i=1}^L [[\rho_i + 1] s_i + 1] \right] = \exp(2\Delta_1) \underline{\Gamma}, \quad (\text{E.91})$$

because $\Gamma := \left[[b+1][L \cdot L_1 + 1] \bar{W} N \prod_{i=1}^L [[\rho_i + 1] s_i + 1] \right]$. For the same reason, we still have (cf. equations (E.19), (E.20) and (E.21)):

$$\tilde{r}_\ell \leq r_\ell e^{\Delta_1}, \quad (\text{E.92})$$

$$\frac{2\tilde{p}_\ell}{\tilde{p}_\ell + 2} - \frac{2p_\ell}{p_\ell + 2} \leq \frac{2[\tilde{p}_\ell - p_\ell]}{\tilde{p}_\ell + 2} \leq 2[\tilde{p}_\ell - p_\ell]/2 \leq \Delta_2 \quad (\text{E.93})$$

and

$$\frac{2}{\tilde{p}_\ell + 2} - \frac{2}{p_\ell + 2} = \frac{2[\tilde{p}_\ell - p_\ell]}{[\tilde{p}_\ell + 2][p_\ell + 2]} \leq \frac{2\Delta_2}{4} \leq \Delta_2/2. \quad (\text{E.94})$$

Thus, plugging in the new values of $k^{(b)}, k^{(b_\ell)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}$ into the definition of $\mathcal{R}_{\mathcal{M},s,p,b_\ell} = \left[\sum_{\ell=1}^L [b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i] \frac{2p_\ell}{p_\ell + 2} r_\ell \frac{2}{p_\ell + 2} \bar{w}_\ell \frac{2}{p_\ell + 2} \tilde{w}_\ell \frac{p_\ell}{p_\ell + 2} \right]^{\frac{1}{2}}$ we now obtain:

$$\mathcal{R}_{k^{(b)}, k^{(b_\ell)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}} \quad (\text{E.95})$$

$$\leq \left[\sum_{\ell=1}^L \left[\exp(\Delta_1) b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i \exp(\Delta_1/L) \right]^{\frac{2\tilde{p}_\ell}{p_\ell + 2}} \tilde{r}_\ell \frac{2}{p_\ell + 2} \bar{w}_\ell \frac{2}{p_\ell + 2} \tilde{w}_\ell \frac{p_\ell}{p_\ell + 2} \right]^{\frac{1}{2}} \quad (\text{E.96})$$

$$\leq \left[\sum_{\ell=1}^L \left[[b_{\ell-1} \exp(\Delta_1)] \prod_{i=\ell}^L \rho_i [s_i \exp(\Delta_1/L)] \right]^{\frac{2\tilde{p}_\ell}{p_\ell + 2}} \tilde{r}_\ell \frac{2}{p_\ell + 2} \bar{w}_\ell \frac{2}{p_\ell + 2} \tilde{w}_\ell \frac{p_\ell}{p_\ell + 2} \right]^{\frac{1}{2}} \quad (\text{E.97})$$

$$\leq \exp(\Delta_1) \left[\sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i \right]^{\frac{2\tilde{p}_\ell}{p_\ell + 2}} \tilde{r}_\ell \frac{2}{p_\ell + 2} \bar{w}_\ell \frac{2}{p_\ell + 2} \tilde{w}_\ell \frac{p_\ell}{p_\ell + 2} \right]^{\frac{1}{2}} \quad (\text{E.98})$$

$$\leq [\exp(\Delta_1)] \left[\sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i \right]^{\frac{\Delta_2}{2}} \left[b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell + 2}} \tilde{r}_\ell \frac{2}{p_\ell + 2} \bar{w}_\ell \frac{2}{p_\ell + 2} \tilde{w}_\ell \frac{p_\ell}{p_\ell + 2} \right]^{\frac{1}{2}} \quad (\text{E.99})$$

$$\leq \left[\exp(\Delta_1) \left[b \prod_i \rho_i s_i \right]^{\frac{\Delta_2}{2}} \right] \left[\sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell + 2}} \tilde{r}_\ell \frac{2}{p_\ell + 2} \bar{w}_\ell \frac{2}{p_\ell + 2} \tilde{w}_\ell \frac{p_\ell}{p_\ell + 2} \right]^{\frac{1}{2}} \quad (\text{E.100})$$

$$\leq \left[e^{\Delta_1} \left[b \prod_i \rho_i s_i \right]^{\frac{\Delta_2}{2}} e^{\Delta_1/2} \right] \left[\sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell + 2}} [r_\ell \bar{w}_\ell]^{\frac{2}{p_\ell + 2}} \tilde{w}_\ell \frac{p_\ell}{p_\ell + 2} \right]^{\frac{1}{2}} \quad (\text{E.101})$$

$$\leq \left[e^{1.5\Delta_1} \left[b \prod_i \rho_i s_i \right]^{\frac{\Delta_2}{2}} \right] \left[\sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell + 2}} r_\ell \frac{2}{p_\ell + 2} \bar{w}_\ell \frac{2}{p_\ell + 2} \tilde{w}_\ell \frac{p_\ell}{p_\ell + 2} \right]^{\frac{1}{2}} \quad (\text{E.102})$$

$$\leq \left[e^{1.5\Delta_1} \left[b \prod_i \rho_i s_i \right]^{\frac{\Delta_2}{2}} \bar{w}_\ell^{\frac{\Delta_2}{2}} \right] \left[\sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell + 2}} [r_\ell \bar{w}_\ell]^{\frac{2}{p_\ell + 2}} \tilde{w}_\ell \frac{p_\ell}{p_\ell + 2} \right]^{\frac{1}{2}} \quad (\text{E.103})$$

$$\leq \left[e^{1.5\Delta_1} \left[b \prod_i \rho_i s_i \right]^{\frac{\Delta_2}{2}} \bar{w}_\ell^{\frac{\Delta_2}{2}} \right] \tilde{\mathcal{R}}_{F_A, \lambda_{aug}} \quad (\text{E.104})$$

$$\leq \left[4\bar{W} + b \prod_i \rho_i s_i \right]^{\max(\Delta_1, \Delta_2)} \tilde{\mathcal{R}}_{F_A, \lambda_{aug}}, \quad (\text{E.105})$$

where at equation (E.99) we have used equation (E.93), at equation (E.100) we have used the fact that $b_\ell \leq b \prod_{i=1}^\ell \rho_i s_i$ (differently from the proof of Theorem E.2). The rest of the calculation is analogous

to the proof of Theorem E.2: at equation (E.101) we have used equation (E.92), at equation (E.102) we have used the trivial fact that $\frac{\hat{p}_\ell}{\hat{p}_\ell+2} \geq \frac{p_\ell}{2+p_\ell}$ and at equation (E.103) we have used equation (E.94) and the fact that $r_\ell \leq \bar{w}_\ell$.

Thus, since the multiplicative error term is unchanged, we can proceed, as in the proof of Theorem E.2, to set

$$\Delta_1 = \Delta_2 = \frac{\log(2)}{\log[4\bar{W} + b \prod_i \rho_i s_i]}, \quad (\text{E.106})$$

ensuring

$$\mathcal{R}_{k^{(p_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(b)}, k^{(s_\ell)}} \leq 2\tilde{\mathcal{R}}_{F_A}. \quad (\text{E.107})$$

Similarly, as before, we can write first (from equation (E.90))

$$\underline{\Gamma}_{k^{(b)}, k^{(s_\ell)}} \leq \underline{\Gamma} 4 \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \quad (\text{E.108})$$

and then:

$$\log \left(\frac{1}{\delta_{k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}}} \right) \quad (\text{E.109})$$

$$\leq \log\left(\frac{1}{\delta}\right) + [4L + 1] \log(3) + \left[|k^{(b)}| + \sum_\ell |k^{(s_\ell)}| + \sum_\ell |k^{(b_\ell)}| + \sum_\ell |k^{(\mathcal{M}_\ell)}| \right] + L \log(1/\Delta_2) \quad (\text{E.110})$$

$$\leq \log\left(\frac{1}{\delta}\right) + L \left[5.5 + \log \log \left[4\bar{W} + b \prod_i \rho_i s_i \right] \right] \quad (\text{E.111})$$

$$+ \left[|k^{(b)}| + \sum_\ell |k^{(s_\ell)}| + \sum_\ell |k^{(b_\ell)}| + \sum_\ell |k^{(\mathcal{M}_\ell)}| \right] \quad (\text{E.112})$$

$$\leq \log\left(\frac{1}{\delta}\right) + L \left[5.5 + \log \log \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \right] + 1.5 \quad (\text{E.113})$$

$$+ \left[|\log(b)| + L \left[\sum_{\ell=1}^L |\log(\|A_\ell\|)| + \sum_{\ell=1}^{L-1} \log(\mathfrak{B}_{\ell,A}) + \log(\bar{W}) \right] \right] / \Delta_1 \quad (\text{E.114})$$

$$\leq \log\left(\frac{1}{\delta}\right) + L \left[5.5 + \log \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \right] + 1.5 \quad (\text{E.115})$$

$$+ 2 \log \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \left[|\log(b)| + L \left[\sum_{\ell=1}^L |\log(\|A_\ell\|)| + \sum_{\ell=1}^{L-1} \log(\mathfrak{B}_{\ell,A}) + \log(\bar{W}) \right] \right]$$

$$\leq \log\left(\frac{1}{\delta}\right) + 1.5 + \quad (\text{E.116})$$

$$2 \log \left[4\bar{W} + b \prod_i \rho_i \|A_\ell\| \right] \left[|\log(b)| + L \left[\sum_{\ell=1}^L |\log(\|A_\ell\|)| + \sum_{\ell=1}^{L-1} \log(\mathfrak{B}_{\ell,A}) + 2 \log(4\bar{W}) \right] \right]$$

$$= \log\left(\frac{1}{\delta}\right) + 1.5 + \Theta_{\log}^{\lambda_{aug}}, \quad (\text{E.117})$$

where at equation (E.111) we have used the definition of Δ_2 (i.e. equation (E.106)) and the fact that $\log(\log(2)) \geq -1$, at equation (E.114) we have used equations (E.88), and at equation (E.115) we

have used the fact that $\log(2) \geq \frac{1}{2}$ and the inequality $\log \log(x) \leq \log(x)$, at equation (E.116) we have used the fact that $4 \log(4) \geq 5.5$, and at equation (E.117) we have used the definition of $\Theta_{\log}^{\lambda_{\text{aug}}}$.

We now plug equations (E.105), (E.107), (E.117) and (E.108) into equation (E.86) to obtain:

$$\begin{aligned} \mathbb{E}[\mathbb{I}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbb{I}(F_A(x_i), y_i) &\leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta_{k^{(b)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}}})}{N}} \\ &+ \frac{832 \sqrt{\log_2(2\Gamma_{k^{(b)}, k^{(s_\ell)}})} \log(N)[\mathcal{B}+1]}{\sqrt{N}} [L(1 + L_1)]^{\frac{p}{2+p}} \mathcal{R}_{k^{(b)}, k^{(b_\ell)}, k^{(s_\ell)}, k^{(\mathcal{M}_\ell)}, k^{(p_\ell)}} \end{aligned} \quad (\text{E.118})$$

$$\leq 6[\mathcal{B}+1] \sqrt{\frac{\log(1/\delta)}{N}} + 6[\mathcal{B}+1] \sqrt{\frac{1.5 + \Theta_{\log}^{\lambda_{\text{aug}}}}{N}} \quad (\text{E.119})$$

$$+ 832 \mathcal{B} L[(1 + L_1)]^{\frac{p}{2+p}} \frac{\sqrt{\log_2(8\Gamma[4\bar{W} + b \prod_i \rho_i \|A_\ell\|])} \log(N)}{N} \tilde{\mathcal{R}}_{F_A, \lambda_{\text{aug}}} \quad (\text{E.120})$$

$$\leq 6[\mathcal{B}+1] \sqrt{\frac{\log(1/\delta)}{N}} + 6[\mathcal{B}+1] \sqrt{\frac{1.5 + \Theta_{\log}}{N}} + 832 \mathcal{B} [L[1 + L_1]]^{\frac{p}{2+p}} \times \quad (\text{E.121})$$

$$\frac{\sqrt{\log(12[b+1][L L_1+1]\bar{W}N \prod_{i=1}^L [\rho_i + 1]s_i + 1] [4\bar{W} + b \prod_i \rho_i \|A_\ell\|])} \log(N)}{N} \tilde{\mathcal{R}}_{F_A, \lambda_{\text{aug}}} \quad (\text{E.122})$$

$$\leq 6[\mathcal{B}+1] \sqrt{\frac{\log(1/\delta)}{N}} + 6[\mathcal{B}+1] \sqrt{\frac{1.5 + \Theta_{\log}}{N}} \quad (\text{E.123})$$

$$+ 832 \mathcal{B} [L(1 + L_1)]^{\frac{p}{2+p}} \frac{\sqrt{\log(\gamma_{FC})} \log(N)}{N} \tilde{\mathcal{R}}_{F_A, \lambda_{\text{aug}}}, \quad (\text{E.124})$$

where at line (E.120) we have used equation (E.107), and at the last line we have used the definition

$$\gamma_{FC} := 12 \left[[b+1][L L_1+1]\bar{W}N \prod_{i=1}^L [\rho_i \|A_i\| + 1] \right] \left[4\bar{W} + b \prod_i \rho_i \|A_i\| \right].$$

This concludes the proof. \square

E.2 The Case $p_\ell = 0$ for all ℓ

We note that both of the above results hold for all values of the p_ℓ s, including 0. However, the fact that Theorem E.2 is posthoc with respect to all combinations of values of p_ℓ introduces an additional term of $\sqrt{\frac{L^3}{N}}$. In this subsection, we show that this can be avoided with a more dedicated proof for the case $p_\ell = 0$.

We first note the following immediate corollary of Proposition E.1:

Proposition E.7. *Instate the assumptions of Proposition E.1, and assume additionally that $p_\ell = 0$ for all ℓ . With probability at least $1 - \delta$ over the draw of the training set,*

$$\begin{aligned} \mathbb{E}[\mathbb{I}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbb{I}(F_A(x_i), y_i) \\ \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta})}{N}} + \frac{240 \sqrt{\log_2(4\Gamma)} \log(N)[\mathcal{B}+1]}{\sqrt{N}} \left[\sum_{\ell=1}^L [w_\ell + w_{\ell-1}] \tilde{r}_\ell \right], \end{aligned} \quad (\text{E.125})$$

where \tilde{r}_ℓ is an a priori upper bound on $\text{rank}(A_\ell)$.

Next, we consider the following post hoc version.

Theorem E.8. *W.p. $\geq 1 - \delta$ over the draw of the training set, every neural network satisfies:*

$$\begin{aligned} & \mathbb{E} [\mathbb{1}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbb{1}(F_A(x_i), y_i) \\ & \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta})}{N}} + \frac{240\sqrt{\log_2(4\Gamma_R)} \log(N)[\mathcal{B}+1]}{\sqrt{N}} \left[\sum_{\ell=1}^L [w_\ell + w_{\ell-1}] \text{rank}(A_\ell) \right] \\ & \quad + \sqrt{\frac{\log(4/\delta) + 2 \log \left[[\mathfrak{B}+2] \prod_{\ell=1}^L (\|A_\ell\| + 2) \right] + L \log(\bar{W})}{N}}, \end{aligned}$$

where $\Gamma_R = [\mathfrak{B}+2][L \text{L}_1 + 1] \bar{W} N \prod_{i=1}^L [[\rho_i + 1][\|A_i\| + 2] + 1]$.

Proof. We apply Proposition E.7 simultaneously for all the values of $b, s_\ell \in \mathbb{N}_+$ and $\tilde{r}_\ell \in [\bar{W}]$ after setting

$$\delta_{b,s_\ell,\tilde{r}_\ell} = \frac{\delta}{[b+1]^2 \prod_\ell [s_\ell + 1]^2 \bar{W}^L}. \quad (\text{E.126})$$

Note that

$$\sum_{b \in \mathbb{N}_+, s_\ell \in \mathbb{N}_+, \tilde{r}_\ell \leq \bar{W}} \delta_{b,s_\ell,\tilde{r}_\ell} \leq \frac{\delta \bar{W}^L [\pi^2/6 - 1]^{1+L}}{\bar{W}^L} \leq \delta. \quad (\text{E.127})$$

Thus, inequality (E.125) holds simultaneously over all values of $b, s_\ell \in \mathbb{N}_+$ and $\tilde{r}_\ell \in [\bar{W}]$ with failure probability δ with the loss function $\mathbb{1}$ replaced by l_{aug} as in the proof of Theorem E.2.

We apply this to $s_\ell = \lceil \|A_\ell\| \rceil$, $b = [\mathfrak{B}]$, $\tilde{r}_\ell = \text{rank}(A_\ell)$. This immediately yields

$$\begin{aligned} & \mathbb{E} [\mathbb{1}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbb{1}(F_A(x_i), y_i) \\ & \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta_{b,s_\ell,\tilde{r}_\ell}})}{N}} + \frac{240\sqrt{\log_2(4\Gamma_R)} \log(N)[\mathcal{B}+1]}{\sqrt{N}} \left[\sum_{\ell=1}^L [w_\ell + w_{\ell-1}] \text{rank}(A_\ell) \right], \end{aligned} \quad (\text{E.128})$$

where

$$\begin{aligned} \Gamma_R & \leq \left[[b+1][L \text{L}_1 + 1] \bar{W} N \prod_{i=1}^L [[\rho_i + 1] s_i + 1] \right] \\ & \leq [\mathfrak{B}+2][L \text{L}_1 + 1] \bar{W} N \prod_{i=1}^L [[\rho_i + 1][\|A_i\| + 2] + 1]. \end{aligned}$$

Next, we note that

$$\log\left(\frac{4}{\delta_{b,s_\ell,\tilde{r}_\ell}}\right) \leq \log(4/\delta) + 2 \log \left[[\mathfrak{B}+2] \prod_{\ell=1}^L (\|A_\ell\| + 2) \right] + L \log(\bar{W}). \quad (\text{E.129})$$

Plugging this back into equation (E.128) yields the result as expected. \square

E.3 Covering Number Bound for Fully Connected Networks

In this subsection, we use the results of Section F to construct a covering number bound for the class of fully connected neural networks. First, we briefly recall the following definitions.

Definition E.9 (Covering number). Let $V \subset \mathbb{R}^n$ and $\|\cdot\|$ be a norm in \mathbb{R}^n . The covering number w.r.t. $\|\cdot\|$, denoted by $\mathcal{N}(V, \epsilon, \|\cdot\|)$, is the minimum cardinality m of a collection of vectors $\mathbf{v}^1, \dots, \mathbf{v}^m \in \mathbb{R}^n$ such that $\sup_{\mathbf{v} \in V} \min_{j=1, \dots, m} \|\mathbf{v} - \mathbf{v}^j\| \leq \epsilon$. In particular, if $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ is a function class and $X = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ are data points, $\mathcal{N}(\mathcal{F}(X), \epsilon, (1/\sqrt{n})\|\cdot\|_2)$ is the minimum cardinality m of a collection of functions $\mathcal{F} \ni f^1, \dots, f^m : \mathcal{X} \rightarrow \mathbb{R}$ such that for any $f \in \mathcal{F}$, there exists $j \leq m$ such that $\sum_{i=1}^n (1/n) |f^j(x_i) - f(x_i)|^2 \leq \epsilon^2$. Similarly, $\mathcal{N}(\mathcal{F}(X), \epsilon, \|\cdot\|_\infty)$ is the minimum cardinality m of a collection of functions $\mathcal{F} \ni f^1, \dots, f^m : \mathcal{X} \rightarrow \mathbb{R}$ such that for any $f \in \mathcal{F}$, there exists $j \leq m$ such that $i \leq n, |f^j(x_i) - f(x_i)| \leq \epsilon$.

Definition E.10. Let \mathcal{F} be a class of real-valued functions with range X . Let also $S = (x_1, x_2, \dots, x_n) \in X$ 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, \dots, x_n is defined by

$$\mathfrak{R}_S(\mathcal{F}) := \mathbb{E}_\delta \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \delta_i f(x_i), \quad (\text{E.130})$$

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

We now have the following result on the covering number of the class $\mathcal{F}_{\mathcal{B}}$.

Proposition E.11. let $x_1, \dots, x_N \in \mathbb{R}^d$ be an arbitrary set of inputs satisfying $\|x_i\| \leq b$ for all i . We have the following bound on the ℓ^∞ covering number of the class $\mathcal{F}_{\mathcal{B}}$ for any granularity $\frac{1}{N} \leq \epsilon$:

$$\begin{aligned} & \log(\mathcal{N}_\infty(\ell \circ \mathcal{F}_{\mathcal{B}}, \epsilon)) \\ & \leq 24 \log_2(\bar{\Gamma}) \left[\frac{(1 + L L_1)}{\min(\epsilon, 1)} \right]^{\frac{2p}{2+p}} \sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} \left[\frac{\mathcal{M}_\ell^{p_\ell}}{s_\ell^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \\ & \leq 96 \log_2(4\Gamma) \left[\frac{(1 + L L_1)}{\min(\epsilon, 1)} \right]^{\frac{2p}{2+p}} \sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}}, \end{aligned} \quad (\text{E.131})$$

where $p = \max_\ell(p_\ell)$, $\underline{w}_\ell = \min(w_\ell, w_{\ell-1})$, $\bar{w}_\ell = [w_\ell + w_{\ell-1}]$, $r_\ell = \frac{\mathcal{M}_\ell^{p_\ell}}{s_\ell^{p_\ell}}$ and $\tilde{w}_\ell := [w_\ell \underline{w}_\ell]$ if $\ell \neq L$ \underline{w}_L if $\ell = L$. Here we also have $\bar{\Gamma} = 128\Gamma^4$ and $\Gamma = \left[[b+1][L L_1+1] \bar{W} N \prod_{i=1}^L [[\rho_i+1]s_i+1] \right]$ where $\bar{W} := \max_{\ell=0,1,\dots,L} w_\ell$.

Before proving Proposition E.11, we will need the following Lemma, which we will later apply to the following choice: $\mathcal{B}_\ell := \{A_\ell \in \mathbb{R}^{w_\ell \times w_{\ell-1}} : \|A_\ell - M_\ell\|_{\text{sc}, p_\ell} \leq \mathcal{M}_\ell; \|A_\ell\| \leq s_\ell\}$.

Lemma E.12. Let $\epsilon_1, \dots, \epsilon_L > 0$ be some arbitrary positive real numbers, and assume we have a fixed dataset x_1, \dots, x_N with $\|x_i\| \leq b$ for all $i \leq N$. Let $\mathcal{B}_1, \dots, \mathcal{B}_L$ be arbitrary subsets of the spaces $\mathbb{R}^{w_\ell \times w_{\ell-1}}$.

There exist covers $\mathcal{C}_1 \subset \mathcal{B}_1, \dots, \mathcal{C}_{0 \rightarrow \ell} \subset \mathcal{B}_1 \times \dots \times \mathcal{B}_\ell, \dots, \mathcal{C}_L \subset \mathcal{B}_1 \times \dots \times \mathcal{B}_L$ such that for all $(A_1, \dots, A_L) \in \mathcal{B}_1 \times \dots \times \mathcal{B}_L$, there exist $\bar{A}^1, \dots, \bar{A}^L$ such that for all $\ell \leq L$, $\bar{A}^{0 \rightarrow \ell} := (\bar{A}^1, \dots, \bar{A}^\ell) \in \mathcal{C}_{0 \rightarrow \ell}$ and for all $i \leq N$,

$$\begin{aligned} \|F_{A^1, \dots, A^\ell}^{0 \rightarrow \ell}(x_i) - F_{\bar{A}^1, \dots, \bar{A}^\ell}^{0 \rightarrow \ell}(x_i)\|_2 & \leq \sum_{l=1}^{\ell} \epsilon_l \rho_{l \rightarrow \ell} \quad \text{if } \ell \neq L \\ \|F_{A^1, \dots, A^\ell}^{0 \rightarrow L}(x_i) - F_{\bar{A}^1, \dots, \bar{A}^\ell}^{0 \rightarrow L}(x_i)\|_\infty & \leq \sum_{l=1}^L \epsilon_l \rho_{l \rightarrow L}. \end{aligned} \quad (\text{E.132})$$

Here $\rho_{\ell_1 \rightarrow \ell_2} := \rho_{\ell_1} \prod_{\ell=\ell_1+1}^{\ell_2} \rho_\ell s_\ell$ where $\rho_0 = 1$ by convention.

Furthermore the covers satisfy the following covering number bound:

$$\log(|\mathcal{C}_{0 \rightarrow L}|) \leq \sum_{\ell=1}^L \log(\mathcal{N}_{\infty, \ell}(\mathcal{B}_\ell, \epsilon_\ell, N, b\rho_{0 \rightarrow \ell-1})), \quad (\text{E.133})$$

where for any $\bar{b} > 0$ and $\bar{N} \in \mathbb{N}$, $\mathcal{N}_{\infty, \ell}(\mathcal{B}_\ell, \epsilon_\ell, N, \bar{b})$ is defined as the minimum number \mathcal{N} such that for any dataset $x_1^{\ell-1}, \dots, x_{\bar{N}}^{\ell-1}$ with $\bar{N} \leq N$ and $\|x_i^{\ell-1}\|_\ell \leq \bar{b}$ for all $i \leq \bar{N}$, where $\|\cdot\|_\ell = \|\cdot\|$ for $\ell \neq L$ and $\|\cdot\|_L = \|\cdot\|_\infty$.

Proof. The proof consists in stringing together the covering numbers of each individual layer with arguments analogous to those in the comparable literature on generalization bounds for neural networks [1, 16, 74].

For the **first layer**, writing $X := \{x_1, \dots, x_N\}$, we know by assumption that there exists a cover $\mathcal{C}_1(X) = \mathcal{C}_{0 \rightarrow 1} \subset \mathcal{B}_1$ such that for all $A_1 \in \mathcal{B}_1$, there exists a $\bar{A}_1 \in \mathcal{C}_1(X_0)$ such that for all $i \leq N$,

$$\|(A^1 - \bar{A}^1)(x_i)\| \leq \epsilon_1, \quad (\text{E.134})$$

from which it also follows that

$$\|F_{A^1}^{0 \rightarrow 1}(x_1) - F_{\bar{A}^1}^{0 \rightarrow 1}(x_1)\| = \|\sigma_1(A^1 x_i) - \sigma_1(\bar{A}^1(x_i))\| \leq \epsilon_1 \rho_1. \quad (\text{E.135})$$

It follows that $\{F_{A^1}^{0 \rightarrow 1} : A^1 \in \mathcal{B}_1\} \subset \mathcal{F}_{\mathcal{B}}^{0 \rightarrow 1}$ is an $L^{\infty, 2}$ cover of $\mathcal{F}_{\mathcal{B}}^{0 \rightarrow 1}$ as required: equation E.132 holds for $\ell = 1$.

For the **inductive case**:

Assume that the covers $\mathcal{C}_{0 \rightarrow 1}, \dots, \mathcal{C}_{0 \rightarrow \ell}$ have been constructed and satisfy equation (E.132). For each element $\bar{A}^{0 \rightarrow \ell}$ of $\mathcal{C}_{0 \rightarrow \ell}$, we have

$$\|F_{\bar{A}^{0 \rightarrow \ell}}^{0 \rightarrow \ell}(x_i)\|_2 \leq \rho_{0 \rightarrow \ell} b.$$

Thus, we can construct a cover $\mathcal{C}_{\ell+1}(F_{\bar{A}^{0 \rightarrow \ell}}^{0 \rightarrow \ell}(X)) \subset \mathcal{B}_{\ell+1}$ such that for all $A^{\ell+1} \in \mathcal{B}_{\ell+1}$ there exists a $\bar{A}^{\ell+1} \in \mathcal{B}_{\ell+1}$ such that for all $i \leq N$, we have

$$\|F_{(\bar{A}^1, \dots, \bar{A}^\ell, A^{\ell+1})}^{0 \rightarrow \ell+1}(x_i) - F_{(\bar{A}^1, \dots, \bar{A}^\ell, \bar{A}^{\ell+1})}^{0 \rightarrow \ell+1}(x_i)\|_{\ell+1} \leq \epsilon_{\ell+1} \quad (\text{E.136})$$

and

$$\log(|\mathcal{C}_{\ell+1}(F_{\bar{A}^{0 \rightarrow \ell}}^{0 \rightarrow \ell}(X))|) \leq \log(\mathcal{N}_{\infty, \ell+1}(\mathcal{B}_{\ell+1}, \epsilon_{\ell+1}, N, b\rho_{0 \rightarrow \ell})). \quad (\text{E.137})$$

We now construct the cover

$$\mathcal{C}_{0 \rightarrow \ell+1} := \bigcup_{\bar{A}^{0 \rightarrow \ell} \in \mathcal{C}_{0 \rightarrow \ell}} \mathcal{C}_{\ell+1}(F_{\bar{A}^{0 \rightarrow \ell}}^{0 \rightarrow \ell}(X)). \quad (\text{E.138})$$

For any $A^1, \dots, A^{\ell+1}$, by the induction hypothesis ¹⁰, there exists $(\bar{A}^1, \dots, \bar{A}^\ell) \in \mathcal{C}_{0 \rightarrow \ell}$ such that for all $i \leq N$,

$$\|F_{A^1, \dots, A^\ell}^{0 \rightarrow \ell}(x_i) - F_{\bar{A}^1, \dots, \bar{A}^\ell}^{0 \rightarrow \ell}(x_i)\| \leq \sum_{l=1}^{\ell} \epsilon_l \rho_{l \rightarrow \ell}.$$

Thus we have

$$\begin{aligned} & \|F_{A^1, \dots, A^{\ell+1}}^{0 \rightarrow \ell+1}(x_i) - F_{\bar{A}^1, \dots, \bar{A}^{\ell+1}}^{0 \rightarrow \ell+1}(x_i)\| \\ & \leq \|F_{A^1, \dots, A^{\ell+1}}^{0 \rightarrow \ell+1}(x_i) - F_{\bar{A}^1, \dots, \bar{A}^\ell, A^{\ell+1}}^{0 \rightarrow \ell+1}(x_i)\| + \|F_{\bar{A}^1, \dots, \bar{A}^\ell, A^{\ell+1}}^{0 \rightarrow \ell+1}(x_i) - F_{\bar{A}^1, \dots, \bar{A}^{\ell+1}}^{0 \rightarrow \ell+1}(x_i)\| \\ & \leq \rho_{\ell+1} s_{\ell+1} \|F_{A^1, \dots, A^\ell}^{0 \rightarrow \ell}(x_i) - F_{\bar{A}^1, \dots, \bar{A}^\ell}^{0 \rightarrow \ell}(x_i)\| + \rho_{\ell+1} \|A^{\ell+1}(F_{\bar{A}^{0 \rightarrow \ell}}^{0 \rightarrow \ell}(x_i)) - \bar{A}^{\ell+1}(F_{\bar{A}^{0 \rightarrow \ell}}^{0 \rightarrow \ell}(x_i))\| \\ & \leq \rho_{\ell+1} s_{\ell+1} \sum_{l=1}^{\ell} \epsilon_l \rho_{l \rightarrow \ell} + \rho_{\ell+1} \epsilon_{\ell+1} = \sum_{l=1}^{\ell+1} \epsilon_l, \end{aligned}$$

¹⁰if $\ell = L$ the theorem is proved, so we can assume $\ell \neq L$ in which case $\|\cdot\|_\ell = \|\cdot\|$

as expected. Furthermore, by construction (E.138), the cardinality of $\mathcal{C}_{0 \rightarrow \ell+1}$ certainly satisfies:

$$\begin{aligned}
|\mathcal{C}_{0 \rightarrow \ell+1}| &\leq |\mathcal{C}_{0 \rightarrow \ell}| \times \mathcal{N}_{\infty, \ell+1}(\mathcal{B}_{\ell+1}, \epsilon_{\ell+1}, N, b\rho_{0 \rightarrow \ell+1}) \\
&\leq \prod_{l=1}^{\ell} \mathcal{N}_{\infty, l}(\mathcal{B}_{\ell+1}, \epsilon_l, N, b\rho_{0 \rightarrow l}) \times \mathcal{N}_{\infty, \ell+1}(\mathcal{B}_{\ell+1}, \epsilon_{\ell+1}, N, b\rho_{0 \rightarrow \ell+1}) \\
&= \prod_{l=1}^{\ell+1} \mathcal{N}_{\infty, l}(\mathcal{B}_{\ell+1}, \epsilon_l, N, b\rho_{0 \rightarrow l}),
\end{aligned}$$

where at the last line we have used the induction hypothesis. \square

We can now proceed with the Proof of Proposition E.11.

Proof of Proposition E.11. For any choice of $\epsilon_1, \dots, \epsilon_L$, by Lemma E.12, we know that $l \circ \mathcal{F}_B$ admits an L^∞ cover \mathcal{C} with granularity $\epsilon = \sum_{\ell=1}^L \epsilon_\ell \rho_{\ell \rightarrow L}$ and cardinality

$$\log(|\mathcal{C}|) \leq \sum_{\ell=1}^L \log(\mathcal{N}_{\infty, \ell}(\mathcal{B}_\ell, \epsilon_\ell, N, b\rho_{0 \rightarrow \ell-1})). \quad (\text{E.139})$$

Let us denote $\bar{w}_\ell := [w_\ell + w_{\ell-1}]$, $\underline{w}_\ell = \min(w_\ell, w_{\ell-1})$ and we further define $\tilde{w}_\ell = [w_\ell \underline{w}_\ell]$ if $\ell \neq L$, \underline{w}_L if $\ell = L$.

Further, by Propositions (F.4) (for $\ell \neq L$) and (F.2) (for layer $\ell = L$), we can further continue from equation (E.139) as follows:

$$\begin{aligned}
\log(|\mathcal{C}|) &\leq \sum_{\ell=1}^L \log(\mathcal{N}_{\infty, \ell}(\mathcal{B}_\ell, \epsilon_\ell, N, b\rho_{0 \rightarrow \ell-1})) \\
&\leq 24 \left[\sum_{\ell=1}^{L-1} \left[\frac{\mathcal{M}_\ell b\rho_{0 \rightarrow \ell-1}}{\epsilon_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} [w_\ell \underline{w}_\ell]^{\frac{p_\ell}{p_\ell+2}} \log_2(\Gamma_{\mathcal{F}_r^p, \ell}) \right. \\
&\quad \left. + \left[\frac{\mathcal{M}_L b\rho_{0 \rightarrow L-1}}{\epsilon_L} \right]^{\frac{2p_L}{p_L+2}} \bar{w}_L^{\frac{2}{p_L+2}} \underline{w}_L^{\frac{p_L}{p_L+2}} \log_2(\Gamma_{\mathcal{F}_r^p, L}) \right] \\
&\leq 24 \log_2(\bar{\Gamma}) \left[\sum_{\ell=1}^{L-1} \left[\frac{\mathcal{M}_\ell b\rho_{0 \rightarrow \ell-1}}{\epsilon_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} [w_\ell \underline{w}_\ell]^{\frac{p_\ell}{p_\ell+2}} + \left[\frac{\mathcal{M}_L b\rho_{0 \rightarrow L-1}}{\epsilon_L} \right]^{\frac{2p_L}{p_L+2}} \bar{w}_L^{\frac{2}{p_L+2}} \underline{w}_L^{\frac{p_L}{p_L+2}} \right] \\
&\leq 24 \log_2(\bar{\Gamma}) \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b\rho_{0 \rightarrow \ell-1}}{\epsilon_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}}, \quad (\text{E.140})
\end{aligned}$$

where $\Gamma_{\mathcal{F}_r^p, \ell} := \left[\left(\frac{16[\mathcal{M}_\ell^{p_\ell} + 1][b \prod_{i=1}^\ell \rho_i s_i + 1][1+s_\ell][w_\ell + w_{\ell-1}]^2}{\epsilon_\ell^{\frac{p_\ell}{p_\ell+2}}} + 7 \right) w_\ell N \right]$, $\Gamma_{\mathcal{F}_r^p, L} := \left[\left(\frac{16[\mathcal{M}_L^{p_L} + 1][b \prod_{i=1}^L \rho_i s_i + 1][1 + \sqrt{w_L} s_L][w_L + w_{L-1}]}{\epsilon_L^{\frac{p_L}{p_L+2}}} + 7 \right) w_L N \right]$ (indeed, note that since $\|A_L^\top\|_{2, \infty} \leq s_L$, we certainly have $\|A_L\| \leq w_L s_L$). Next, we upper bound the logarithmic factors arising from $\Gamma_{\mathcal{F}_r^p, \ell}$ and $\Gamma_{\mathcal{F}_r^p, L}$ as follows:

$$\begin{aligned}
\max_\ell \Gamma_{\mathcal{F}_r^p, \ell} &\leq \left[\left(\frac{64[\max_\ell \frac{\mathcal{M}_\ell^{p_\ell}}{s_\ell^{p_\ell}} + 1][b + 1] \prod_{i=1}^L [(\rho_i + 1)s_i + 1]^3 \bar{W}^{1.5} [L_L L + 1]}{\min(1, \varepsilon)} + 7 \right) \bar{W} N \right] \\
&\leq \left[128 \prod_{i=1}^L [(\rho_i + 1)s_i + 1]^3 [b + 1] [L_L L + 1] \bar{W}^4 N^2 \right] := \bar{\Gamma} = 128 \underline{\Gamma}^4, \quad (\text{E.141})
\end{aligned}$$

where $\underline{\Gamma} := \left[[b + 1][L_L L + 1] \bar{W} N \prod_{i=1}^L [(\rho_i + 1)s_i + 1] \right]$ and we have used the fact that $\max_\ell \frac{\mathcal{M}_\ell^{p_\ell}}{s_\ell^{p_\ell}} \leq \bar{W} \sqrt{w_L}$.

We now set $\epsilon_\ell := \frac{\alpha_\ell \varepsilon}{\rho_{\ell \rightarrow L}} = \frac{\alpha_\ell \varepsilon}{\prod_{i=\ell+1}^L s_i \rho_i}$ for some α_ℓ such that $\sum_\ell \alpha_\ell = 1/L_1$. Note that this ensures that the granularity of the cover \mathcal{C} is

$$\epsilon = L_1 \sum_{\ell=1}^L \epsilon_\ell \rho_{\ell \rightarrow L} = L_1 \sum_{\ell=1}^L \frac{\alpha_\ell \varepsilon}{\rho_{\ell \rightarrow L}} \rho_{\ell \rightarrow L} = \varepsilon. \quad (\text{E.142})$$

Then, we can plug in the value of ε into equation (E.140) to obtain:

$$\begin{aligned} \log(|\mathcal{C}|) &\leq 24 \log_2(\bar{\Gamma}) \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b \rho_{0 \rightarrow \ell-1}}{\epsilon_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \\ &\leq 24 \log_2(\bar{\Gamma}) \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b \prod_{i \neq \ell} \rho_i s_i}{\varepsilon \alpha_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \\ &\leq 24 L_1 \log_2(\bar{\Gamma}) \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b \prod_i \rho_i s_i}{\varepsilon s_\ell \alpha_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}} \\ &\leq 24 \log_2(\bar{\Gamma}) \left[\frac{(1+L_1 L_1)}{\min(\varepsilon, 1)} \right]^{\frac{2p}{2+p}} \sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} \left[\frac{\mathcal{M}_\ell^{p_\ell}}{s_\ell^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}}, \end{aligned} \quad (\text{E.143})$$

where at the last line, we have set $\alpha_\ell = \frac{1}{L_1 L}$. \square

F Covering Numbers for Linear Maps with Schatten Quasi-norm Constraints (One Layer Case)

In this section, we prove covering number bounds for classes of linear maps with bounded Schatten p quasi norms, which corresponds to the one layer case in our analysis. Thus, the results in this section form the basic ingredient used for the proofs in all other sections. To achieve this, we begin with the following parameter counting argument to bound the complexity of low rank matrices. Proposition F.1 is known, but we reproduce the proof for completeness. The other results in this section are original to the best of our knowledge.

Proposition F.1 (Cf. [28] (Lemma D1 page 30), cf. also [90, 91]). *Let $\mathcal{E}_{r,s}$ denote the set $\{A \in \mathbb{R}^{m \times d} : \text{rank}(A) \leq r, \|A\|_\sigma \leq s\}$. There exists a cover $\mathcal{C} \subset \mathcal{E}_{r,s}$ with respect to the spectral norm such that*

$$\log(|\mathcal{C}|) \leq [m+d]r \log \left[1 + \frac{6s}{\epsilon} \right] = \tilde{O}([m+d]r). \quad (\text{F.1})$$

Proof. First, note that every matrix $A \in \mathcal{E}_{r,s}$ can be written as $A = WV^\top$ for $W \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{d \times r}$ with $\|W\| \leq \sqrt{s}$ and $\|V\| \leq \sqrt{s}$. Then, by Lemma (I.4), there are $\epsilon/2\sqrt{s}$ covers \mathcal{C}_1 and \mathcal{C}_r of the balls of radius \sqrt{s} in $\mathbb{R}^{m \times r}$ and $\mathbb{R}^{d \times r}$ (w.r.t. the spectral norm) with cardinalities

$$\begin{aligned} \log(|\mathcal{C}_1|) &\leq [mr] \log \left[1 + \frac{6s}{\epsilon} \right] \\ \log(|\mathcal{C}_r|) &\leq [dr] \log \left[1 + \frac{6s}{\epsilon} \right]. \end{aligned} \quad (\text{F.2})$$

Let $A = WV^\top \in \mathcal{E}_{r,s}$, we define $\bar{A} = \bar{W}\bar{V}^\top$ where \bar{W} (resp. \bar{V}) is the cover element in \mathcal{C}_1 (resp. \mathcal{C}_r) associated to W (resp. V). By definition of the covers via Lemma (I.4) we have

$$\|W - \bar{W}\|_\sigma, \|V - \bar{V}\|_\sigma \leq \frac{\epsilon}{2\sqrt{s}}. \quad (\text{F.3})$$

Using this, we certainly have

$$\begin{aligned}
\|A - \bar{A}\|_\sigma &= \|WV^\top - \bar{W}\bar{V}^\top\|_\sigma \\
&\leq \|W(V^\top - \bar{V}^\top)\|_\sigma + \|(W - \bar{W})\bar{V}^\top\|_\sigma \\
&\leq \|W\|_\sigma \|(V^\top - \bar{V}^\top)\|_\sigma + \|(W - \bar{W})\|_\sigma \|\bar{V}^\top\|_\sigma \\
&< s \frac{\epsilon}{2\sqrt{s}} + \frac{\epsilon}{2\sqrt{s}} s = \epsilon.
\end{aligned} \tag{F.4}$$

Thus, $\mathcal{C} := \{A \in \mathbb{R}^{m \times d} : A = WV^\top, W \in \mathcal{C}_1, V \in \mathcal{C}_r\}$ is a valid cover of $\mathcal{E}_{r,s}$ w.r.t. the spectral norm. Furthermore, by equations (F.2) we certainly have

$$\log(|\mathcal{C}|) \leq \log(|\mathcal{C}_1|) + \log(|\mathcal{C}_r|) \tag{F.5}$$

$$\leq [m+d]r \log \left[1 + \frac{6s}{\epsilon} \right], \tag{F.6}$$

as expected. \square

Proposition F.2 (L^∞ cover of \mathcal{F}_r^p). *Consider the set of matrices with a bounded Schatten norm as follows: $\mathcal{F}_r^p := \{Z \in \mathbb{R}^{m \times d} : \|Z - M\|_{sc,p}^p \leq \mathcal{M}^p, \|Z\| \leq s\}$ (where M is a fixed reference matrix), viewed as linear maps from \mathbb{R}^d to \mathbb{R}^m , and assume as usual that we have a training set $x_1, \dots, x_N \in \mathbb{R}^d$ such that $\|x_i\| \leq b$ for all $i \leq N$.*

There exists a cover $\mathcal{C} \subset \mathcal{F}_r^p$ such that for all $Z \in \mathcal{F}_r^p$, there exists a $\bar{Z} \in \mathcal{C}$ such that for all $i \leq N$, we have

$$\|(Z - \bar{Z})x_i\|_\infty \leq \epsilon \tag{F.7}$$

and

$$\begin{aligned}
\log(|\mathcal{C}|) &\leq 24 \left[\frac{\mathcal{M}b}{\epsilon} \right]^{\frac{2p}{p+2}} [m+d]^{\frac{2}{p+2}} \min(m, d)^{\frac{p}{p+2}} \log_2(\Gamma_{\mathcal{F}_r^p}) \\
&\leq 24 \left[\frac{bs}{\epsilon} \right]^{\frac{2p}{p+2}} [m+d] r^{1-\frac{p}{p+2}} \log_2(\Gamma_{\mathcal{F}_r^p}),
\end{aligned}$$

where $\Gamma_{\mathcal{F}_r^p} := \left(\frac{16[\mathcal{M}^p+1][b+1][1+s][m+d]}{\epsilon^{\frac{p}{p+2}}} + 7 \right) mN$.

Remark F.3. Note that in particular, the proposition holds for $p = 0$, replacing \mathcal{M}^p with a constraint on the rank. In that case, the proposition follows immediately from Proposition F.1. Furthermore, it also follows from the limiting case as $p \rightarrow 0$. This is only true thanks to a careful management of the logarithmic factor $\Gamma_{\mathcal{F}_r^p}$: indeed, in addition to the constraint $\|Z\|_{sc,p}^p \leq \mathcal{M}^p$, we also independently require the constraint $\|Z\| \leq s$. This second constraint is only necessary to maintain good scaling as $p \rightarrow 0$. Indeed, we could have bounded $\|Z\|$ via $\|Z\| \leq \|Z\|_{sc,p} \leq \mathcal{M}$. However, this would result in a factor of \mathcal{M} (without an exponent of p) inside the logarithmic term, making the bound blow up at a rate of $1/p$ as $p \rightarrow 0$. To see this, note that as $p \rightarrow 0$,

$$\log(\|Z\|_{sc,p}) = \frac{\log(\|Z\|_{sc,p}^p)}{p} = O\left(\frac{\log(\text{rank}(Z))}{p}\right) \rightarrow \infty. \tag{F.8}$$

Proof. The proof relies on the general "parametric interpolation" technique developed in [28]. After a simple translation argument, we can assume without loss of generality that $M = 0$. Let τ be a threshold to be determined later. We first prove the following claim:

Claim 1: Any matrix $Z \in \mathcal{F}_r^p$ can be decomposed into a sum $Z_1 + Z_2$ where $Z_1 \in \mathcal{E}_{r_1, \mathcal{M}}$ and $Z_2 \in \mathcal{F}_{\mathcal{M}_2}^2$ where $r_1 = \frac{\mathcal{M}^p}{\tau^p}$ and $\mathcal{M}_2^2 = \tau^2 \min(m, d)$.

Proof of Claim 1:

Let $\rho_1, \dots, \rho_{\min(m,d)}$ denote the singular values of Z , so that the singular value decomposition of Z is $Z = \sum_k \rho_k v_k w_k^\top$ for some unit vectors $v_k \in \mathbb{R}^m$ and $w_k \in \mathbb{R}^d$.

Let $T = \max(k : \rho_k > \tau)$ and write $Z_1 = \sum_{k \leq T} \rho_k v_k w_k^\top$ and $Z_2 = Z - Z_1 = \sum_{k > T} \rho_k v_k w_k^\top$.

Note that by Markov's inequality, since $\|Z\|_{\text{sc},p}^p \leq \mathcal{M}^p$, we have that $\text{rank}(Z_1) \leq T \leq \frac{\mathcal{M}^p}{\tau^p}$. Furthermore, we clearly have $\|Z_1\| = \rho_1 = \|Z\| \leq \|Z\|_{\text{sc},p} \leq \mathcal{M}$. Thus, $Z_1 \in \mathcal{E}_{r_1, \mathcal{M}}$ as expected. Next, it is clear that

$$\|Z_2\|_{\text{Fr}}^2 = \sum_{k=T+1}^{\min(m,d)} \rho_k^2 \leq \tau^2 \min(m, d). \quad (\text{F.9})$$

Thus, $Z_2 \in \mathcal{F}_{\mathcal{M}_2}^2$ as expected, which *concludes the proof of the claim*.

Thus, it follows that for any choice of τ and granularity ϵ , we can apply Propositions F.1 and I.3 to obtain $\epsilon/2$ covers $\mathcal{C}_1 \subset \mathcal{E}_{r_1, \mathcal{M}}$ and $\mathcal{C}_2 \in \mathcal{F}_{\mathcal{M}_2}^2$ such that for any $Z_1 \in \mathcal{E}_{r_1, \mathcal{M}}$ and for any $Z_2 \in \mathcal{F}_{\mathcal{M}_2}^2$ there exist $\bar{Z}_1 \in \mathcal{C}_1$ and $\bar{Z}_2 \in \mathcal{C}_2$ such that for all $i \leq N$, $\|(Z_1 - \bar{Z}_1)x_i\|_\infty \leq \epsilon/2$ and $\|(Z_2 - \bar{Z}_2)x_i\|_\infty \leq \epsilon/2$. It is then clear that for any $Z \in \mathcal{F}_r^p$, the element $\bar{Z} = \bar{Z}_1 + \bar{Z}_2 \in \mathcal{E}_{r_1, \mathcal{M}} + \mathcal{F}_{\mathcal{M}_2}^2$ satisfies $\|(Z - \bar{Z})x_i\|_\infty \leq \epsilon$. Thus, $\mathcal{C} = \mathcal{C}_1 + \mathcal{C}_2$ is the required ϵ cover of \mathcal{F}_r^p (which satisfies inequality F.7), and by Propositions F.1 (with $\epsilon \leftarrow \epsilon/(2b)$ since the required granularity is $\epsilon/2$ and Proposition F.1 gives a cover w.r.t. the spectral norm) and I.3 (with $\epsilon \leftarrow \epsilon/2$) its cardinality satisfies:

$$\begin{aligned} \log(|\mathcal{C}|) &\leq \log(|\mathcal{C}_1|) + \log(|\mathcal{C}_2|) \\ &\leq [m+d]r_1 \log\left[1 + \frac{12bs}{\epsilon}\right] + \frac{144\mathcal{M}_2^2 b^2}{\epsilon^2} \log_2\left[\left(\frac{16\mathcal{M}_2 b}{\epsilon} + 7\right)md\right] \\ &= [m+d]\frac{\mathcal{M}^p}{\tau^p} \log\left[1 + \frac{12s b}{\epsilon}\right] + \frac{144\tau^2 \min(m, d)b^2}{\epsilon^2} \log_2\left[\left(\frac{16\tau \min(m, d)b}{\epsilon} + 7\right)mN\right], \end{aligned} \quad (\text{F.10})$$

where at Equation (F.10) we have substituted the values of r_1 and \mathcal{M}_2 . This motivates the following choice of τ :

$$\tau = 144^{-\frac{1}{p+2}} \mathcal{M}^{\frac{p}{p+2}} \epsilon^{\frac{2}{p+2}} b^{-\frac{2}{p+2}} \left[\frac{[m+d]}{\min(m, d)} \right]^{\frac{1}{p+2}}, \quad (\text{F.11})$$

which leads to the following covering number bound:

$$\begin{aligned} \log(|\mathcal{C}|) &\leq 144^{\frac{p}{p+2}} \left[\frac{\mathcal{M} b}{\epsilon} \right]^{\frac{2p}{p+2}} \frac{[m+d]^{\frac{2}{p+2}}}{\min(m, d)^{-\frac{p}{p+2}}} \left[\log\left[1 + \frac{12s b}{\epsilon}\right] \right. \\ &\quad \left. + \log_2\left[\left[\frac{16[\frac{\mathcal{M}^p \epsilon}{b}]^{\frac{2}{p+2}} \min(m, d)^{\frac{p+1}{p+2}} [m+d]^{\frac{1}{p+2}} b}{\epsilon} + 7\right] md\right]\right] \\ &\leq 24 \left[\frac{\mathcal{M} b}{\epsilon} \right]^{\frac{2p}{p+2}} [m+d]^{\frac{2}{p+2}} \min(m, d)^{\frac{p}{p+2}} \log_2\left[\left(\frac{16[\mathcal{M}^p + 1][b+1][1+s][m+d]}{\epsilon^{\frac{p}{p+2}}} + 7\right)mN\right], \end{aligned}$$

where at the last line, we have used the fact that $p \leq 2$ and therefore $144^{\frac{p}{p+2}} \leq 12$. The proof is complete. \square

We have the following very direct $L2$ analogue, which follows directly by applying Proposition F.2 with $\epsilon \leftarrow \epsilon/\sqrt{m}$.

Proposition F.4 (L^2 cover of \mathcal{F}_r^p). *Consider the set of matrices with a bounded Schatten norm as follows: $\mathcal{F}_r^p := \{Z \in \mathbb{R}^{m \times d} : \|Z - M\|_{\text{sc},p}^p \leq \mathcal{M}^p = r s^p; \|Z\| \leq s\}$ (where M is a fixed reference matrix) viewed as linear maps from \mathbb{R}^d to \mathbb{R}^m , and assume as usual that we have a training set $x_1, \dots, x_N \in \mathbb{R}^d$ such that $\|x_i\| \leq b$ for all $i \leq N$.*

There exists a cover $\mathcal{C} \subset \mathcal{F}_r^p$ such that for all $Z \in \mathcal{F}_r^p$, there exists a $\bar{Z} \in \mathcal{C}$ such that for all $i \leq N$, we have

$$\|(Z - \bar{Z})x_i\|_2 \leq \epsilon. \quad (\text{F.12})$$

and

$$\begin{aligned} \log(|\mathcal{C}|) &\leq 24 \left[\frac{\mathcal{M} b}{\epsilon} \right]^{\frac{2p}{p+2}} [m+d]^{\frac{2}{p+2}} \min(m, d)^{\frac{p}{p+2}} [m]^{\frac{p}{p+2}} \log_2(\Gamma'_{\mathcal{F}_r^p}) \\ &\leq 24 \left[\frac{bs}{\epsilon} \right]^{\frac{2p}{p+2}} [m+d]^{1+\frac{p}{p+2}} r^{1-\frac{p}{p+2}} \log_2(\Gamma'_{\mathcal{F}_r^p}), \end{aligned}$$

$$\text{where } \Gamma'_{\mathcal{F}_r^p} := \left(\frac{16[\mathcal{M}^p + 1][b+1][1+s][m+d]^2}{\epsilon^{\frac{p}{p+2}}} + 7 \right) mN.$$

G Extension to Convolutional Neural Networks (CNN)

In this section, we generalize our results to Convolutional Neural Networks. In particular, this section culminates in Theorem G.5, which provides a post hoc generalization bound for convolutional neural networks whose weight matrices have low Schatten p quasi norm.

We first introduce some more detailed notation, which generally follows the existing literature on norm based bounds [16] and is reintroduced for completeness. One difference is that we assume that the pooling operation only operates on the spatial dimension (which is standard in real life CNNs applications), which results in the single notation U_ℓ for the channel dimension at layer ℓ , which replaces both notations m_ℓ and U_ℓ from [16].

Let $x \in \mathbb{R}^{U \times w}$ denote an input ‘image’, $A \in \mathbb{R}^{U \times d}$ and S^1, S^2, \dots, S^O be O ordered subsets of $(\{1, 2, \dots, w\} \times \{1, 2, \dots, U\})$ each of cardinality d corresponding to the convolutional ‘patches’. We will write $\Lambda_A(x) \in \mathbb{R}^{U \times O}$ for the image of x under the convolutional operation associated with A : $\Lambda_A(x)_{j,o} = \sum_{i=1}^d X_{S_i^o} A_{j,i}$. The sets S^1, S^2, \dots, 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 (or “torch.nn.functional.conv2d” in Pytorch). To represent the patches at a given layer, we add a layer index ℓ . For instance, if the input is a $28 \times 28 \times 1$ image and the convolutional filters are of size 2×2 with stride 2, $S^{\ell-1=0,o=1} = S^{0,1} = \{(1,1), (1,2), (2,1), (2,2)\}$ represents the first convolutional patch, and $S^{0,2} = \{(3,1), (3,2), (4,1), (4,2)\}$ represents the second convolutional patch.

We will also write $\text{op}(A^\ell)$ for the matrix in $\mathbb{R}^{U_\ell O_{\ell-1} \times U_{\ell-1} w_{\ell-1}}$ that represents the convolution operation Λ_{A^ℓ} (after unravelling all the dimensions), as we will rely on its spectral norm $\|\text{op}(A^\ell)\|$ in calculations. To represent the full architecture, we use the index ℓ to represent the ℓ ’th layer: for instance, the ℓ ’th layer’s convolutional operation is denoted by $\Lambda_{A^\ell} : \mathbb{R}^{U_{\ell-1} \times w_{\ell-1}} \rightarrow \mathbb{R}^{U_\ell \times O_{\ell-1}}$ and acts on the $\ell - 1$ th layer’s activations in the space. It is also assumed that the last layer is fully-connected.

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

$$F_{A^1, A^2, \dots, A^L}^{\mathcal{C}} : \mathbb{R}^{U_0 \times w_0} \rightarrow \mathbb{R}^{U_L \times w_L} : x \mapsto (\sigma_L \circ \Lambda_{A^L} \circ \sigma_{L-1} \circ \Lambda_{A^{L-1}} \circ \dots \circ \sigma_1 \circ \Lambda_{A^1})(x), \quad (\text{G.1})$$

where for each $\ell \leq L$, the weight A^ℓ is a matrix in $\mathbb{R}^{w \times d_\ell}$ and for each ℓ , $\sigma_\ell : \mathbb{R}^{U_\ell \times O_\ell} \rightarrow \mathbb{R}^{U_\ell \times w_\ell}$ is L^∞ Lipschitz with constant ρ_ℓ and represents both the elementwise activation operation and any spatial pooling.

G.1 One Layer Case

We now consider the covering number of the class of functions corresponding to a layer’s convolutional operation with a matrix A such that $\|A - M\|_{\text{sc},p} \leq \mathcal{M}$ for some number \mathcal{M} .

Proposition G.1 (L^∞ cover of \mathcal{C}_r^p for a convolutional layer). *Assume we are given a training set $x_1, \dots, x_N \in \mathbb{R}^{U_0 \times w_0}$ such that $\sup_{i,o} \|(x_i)_{S^0,o}\| \leq b$. Consider the set of matrices with a bounded Schatten norm as follows: $\mathcal{F}_r^p := \{Z \in \mathbb{R}^{U' \times d} : \|Z - M\|_{\text{sc},p}^p \leq \mathcal{M}^p = r s^p; \|\text{op}(Z)\| \leq s\}$ (where M is an arbitrary reference matrix) and the associated class $\mathcal{C}_r^p := \{\Lambda_A : A \in \mathcal{F}_r^p\}$ of linear maps representing the associated convolution operations from $\mathbb{R}^{U \times w}$ to $\mathbb{R}^{U' \times w'}$.*

There exists a cover $\mathcal{C} \subset \mathcal{C}_r^p$ such that for all $Z \in \mathcal{C}_r^p$, there exists a $\bar{Z} \in \mathcal{C}$ such that for all $i \leq N$, we have

$$\|(\Lambda_Z - \Lambda_{\bar{Z}})(x_i)\|_\infty \leq \epsilon \quad (\text{G.2})$$

and

$$\begin{aligned}\log(|\mathcal{C}|) &\leq 24 \left[\frac{\mathcal{M} b}{\epsilon} \right]^{\frac{2p}{p+2}} [U' + d]^{\frac{2}{p+2}} \min(U', d)^{\frac{p}{p+2}} \log_2(\Gamma_{\mathfrak{C}_r^p}) \\ &\leq 24 \left[\frac{b s}{\epsilon} \right]^{\frac{2p}{p+2}} [U' + d] \min(U', d)^{\frac{p}{p+2}} r^{1-\frac{p}{p+2}} \log_2(\Gamma_{\mathfrak{C}_r^p}),\end{aligned}$$

where $\Gamma_{\mathfrak{C}_r^p} := \left(\frac{16[\mathcal{M}^p+1][b+1][1+s][U'+d]}{\epsilon^{\frac{p}{p+2}}} + 7 \right) U' NO$.

Proof. Note that $\|Z\| \leq \|\text{op}(Z)\| \leq s$. Thus, the proposition follows immediately from Proposition F.2 applied to the auxiliary dataset collecting each input convolutional patch as a sample and replacing $N \leftarrow NO$ and $m \leftarrow U'$.

□

We also have the following immediate L^2 version, which follows from the previous proposition with $\epsilon \leftarrow \frac{\epsilon}{U'w'}$ and utilizing the L^∞ Lipschitzness of the nonlinearity σ .

Proposition G.2 (L^2 cover of \mathfrak{C}_r^p for a convolutional layer). *Instate the notation of Proposition G.1 and consider a ρ -Lipschitz activation function $\sigma : \mathbb{R}^{U' \times O} \rightarrow \mathbb{R}^{U' \times w'}$. There exists a cover \mathcal{C} such that for all $Z \in \mathfrak{C}_r^p$, there exists a $\bar{Z} \in \mathcal{C}$ such that for all $i \leq N$, we have*

$$\|(\Lambda_Z - \Lambda_{\bar{Z}})(x_i)\|_{\text{Fr}} \leq \epsilon \rho \quad (\text{G.3})$$

and

$$\begin{aligned}\log(|\mathcal{C}|) &\leq 24 \left[\frac{\mathcal{M} b}{\epsilon} \right]^{\frac{2p}{p+2}} [U' + d]^{\frac{2}{p+2}} \min(U', d)^{\frac{p}{p+2}} [U'w']^{\frac{p}{p+2}} \log_2(\Gamma'_{\mathfrak{C}_r^p}) \\ &\leq 24 \left[\frac{b s}{\epsilon} \right]^{\frac{2p}{p+2}} [U' + d] r^{1-\frac{p}{p+2}} [U'w']^{\frac{p}{p+2}} \log_2(\Gamma'_{\mathfrak{C}_r^p}),\end{aligned}$$

where $\Gamma'_{\mathfrak{C}_r^p} := \left(\frac{16[\mathcal{M}^p+1][b+1][1+s][U'+d]U'w'}{\epsilon^{\frac{p}{p+2}}} + 7 \right) U' NO$.

We now move on to our covering number bound for the class of CNNs with fixed norm constraints: we write

$$\mathcal{F}_{\mathcal{B}}^{\mathfrak{C}} := \left\{ F_A = F_{(A_1, \dots, A_L)}^{\mathfrak{C}} : \|A_\ell - M_\ell\|_{\text{sc}} \leq \mathcal{M}_\ell \quad \|\text{op}(A_\ell)\| \leq s_\ell \quad \forall \ell < L \quad \|A_L^\top\|_{2,\infty} \leq s_L \right\}$$

for the class of all neural networks whose matrices satisfy the norm constraints with fixed constraints \mathcal{M}_ℓ, s_ℓ etc.

G.2 Covering Number For the Class $\mathcal{F}_{\mathcal{B}}^{\mathfrak{C}}$

We have the following analogue of Proposition E.11 for the convolutional case:

Proposition G.3. *Let x_1, \dots, x_N be a given training set satisfying $\|x_i\| \leq b$ for all i . Consider the class $\mathcal{F}_{\mathcal{B}}^{\mathfrak{C}} := \left\{ F_A = F_{(A_1, \dots, A_L)}^{\mathfrak{C}} : \|A_\ell - M_\ell\|_{\text{sc}} \leq \mathcal{M}_\ell \quad \|\text{op}(A_\ell)\| \leq s_\ell \quad \forall \ell < L \quad \|A_L^\top\|_{2,\infty} \leq s_L \right\}$ of all CNNs satisfying a set of fixed norm constraints for a given set of $p_\ell, \mathcal{M}_\ell, s_\ell$. We have the following covering number bound for the loss class $1 \circ \mathcal{F}_{\mathcal{B}}^{\mathfrak{C}}$ for any $\epsilon \geq \frac{1}{N}$:*

$$\begin{aligned}\log(\mathcal{N}_\infty(1 \circ \mathcal{F}_{\mathcal{B}}^{\mathfrak{C}}, \epsilon)) &\leq 72 \log_2(4 \underline{\Gamma}^{\mathfrak{C}}) \left[\frac{(1 + L L_1)}{\min(\epsilon, 1)} \right]^{\frac{2p}{2+p}} \times \\ &\quad \sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} \left[\frac{\mathcal{M}_\ell^{p_\ell}}{s_\ell^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}},\end{aligned} \quad (\text{G.4})$$

where $\underline{\Gamma}^{\mathfrak{C}} := [b+1][L L_1+1] \mathcal{W} \mathcal{A} \mathcal{N} \prod_{i=1}^L [[\rho_i+1]s_i+1]$.

Proof. The proof is similar to that of Proposition E.11: most differences are in the detailed calculations.

As before, for any choice of $\epsilon_1, \dots, \epsilon_L$, by Lemma E.12, we know that $1 \circ \mathcal{F}_B^C$ admits an L^∞ cover \mathcal{C} with granularity $\epsilon = \sum_{\ell=1}^L \epsilon_\ell \rho_{\ell \rightarrow L}$ and cardinality

$$\log(|\mathcal{C}|) \leq \sum_{\ell=1}^L \log(\mathcal{N}_{\infty, \ell}(\mathcal{B}_\ell^C, \epsilon_\ell, N, b\rho_{0 \rightarrow \ell-1})), \quad (\text{G.5})$$

where $\mathcal{B}_\ell^C := \{A_\ell \in \mathbb{R}^{U_\ell \times d_{\ell-1}} : \|A_\ell - M_\ell\|_{\text{sc}, p_\ell} \leq \mathcal{M}_\ell; \|\text{op}(A_\ell)\| \leq s_\ell\}$ (for $\ell \neq L$) and $\mathcal{B}_L^C := \{A_\ell \in \mathbb{R}^{U_\ell \times d_{\ell-1}} : \|A_\ell - M_L\|_{\text{sc}, p_L} \leq \mathcal{M}_L; \|\text{op}(A_\ell)\| \leq s_L\}$.

We now use the following notation: we write U_ℓ for the number of channels at layer ℓ and d_ℓ for the number of pixels in each convolutional patch at layer ℓ , so that $\min(U_\ell, d_{\ell-1})$ is the minimum dimension of the filter applied at layer ℓ . we also write $W_\ell := U_\ell \times w_\ell$ with the convention that $W_L = 1$. We also write $\mathcal{W} := \sum_{\ell=1}^L d_{\ell-1} \times U_\ell$ for the total number of parameters in the network and $\mathcal{A} := \sum_{\ell=0}^L U_\ell \times O_{\ell-1}$ for the total number of preactivations/input dimensions including both the input and the final layer output. Note that by Propositions G.1 and G.2, we can continue from equation (G.5) as follows:

$$\log(|\mathcal{C}|) \leq 24 \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b\rho_{0 \rightarrow \ell-1}}{\epsilon_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}} \log_2(\Gamma_{\mathcal{F}_r^P, \ell}^C), \quad (\text{G.6})$$

where for all $\ell \leq L$, $\Gamma_{\mathcal{F}_r^P, \ell}^C := \left[\left(\frac{16[\mathcal{M}_\ell p_\ell + 1][b \prod_{i=1}^\ell \rho_i s_i + 1][1 + \sqrt{\mathcal{A}} s_\ell] \mathcal{W} \mathcal{A}}{\epsilon_\ell^{\frac{p_\ell}{p_\ell+2}}} + 7 \right) N \mathcal{A} \right]$ (indeed, since $\|A_L^\top\|_{2, \infty} \leq s_L$, we certainly have $\|A_L\| \leq s_L \mathcal{A}$).

We now set $\epsilon_\ell = \frac{\epsilon}{(L L_1 + 1) \rho_{\ell \rightarrow L}}$ as before and bound all the logarithmic factors as follows:

$$\begin{aligned} \max_\ell \Gamma_{\mathcal{F}_r^P, \ell}^C &\leq \left[\left(\frac{16[\max_\ell \frac{\mathcal{M}_\ell p_\ell}{s_\ell p_\ell} + 1] \prod_{i=1}^L [[\rho_i + 1] s_i + 1]^3 [b + 1] [L L_1 + 1] \mathcal{W} \mathcal{A}^{1.5}}{\min(\epsilon, 1)} + 7 \right) N \mathcal{A} \right] \\ &\leq \left[64 \prod_{i=1}^L [[\rho_i + 1] s_i + 1]^3 [b + 1] [L_1 + 1] \mathcal{W} \mathcal{A}^3 N^2 \right] \leq [4 \underline{\Gamma}^C]^3 \end{aligned} \quad (\text{G.7})$$

where $\underline{\Gamma}^C := [b + 1] [L L_1 + 1] \mathcal{W} \mathcal{A} N \prod_{i=1}^L [[\rho_i + 1] s_i + 1]$. Plugging the value of epsilon into equation (G.6) we can continue with a calculation similar to that of equations (E.143):

$$\begin{aligned} \log(|\mathcal{C}|) &\leq 72 \log_2(4 \underline{\Gamma}^C) \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b\rho_{0 \rightarrow \ell-1}}{\epsilon_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}} \\ &\leq 72 \log_2(4 \underline{\Gamma}^C) \left[\frac{(1 + L L_1)}{\min(\epsilon, 1)} \right]^{\frac{2p_\ell}{2+p_\ell}} \times \\ &\quad \sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} \left[\frac{\mathcal{M}_\ell p_\ell}{s_\ell p_\ell} \right]^{\frac{2}{p_\ell+2}} [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}}, \end{aligned}$$

as expected. \square

G.3 Generalization Bound with Fixed Constraints (CNNs)

Based on the above, we obtain the following analogue of Proposition E.1:

Proposition G.4. Fix some reference/initialization matrices $M^1 \in \mathbb{R}^{w_1 \times w_0} = \mathbb{R}^{w_1 \times d}, \dots, M^L \in \mathbb{R}^{w_L \times w_{L-1}} = \mathbb{R}^{C \times w_{L-1}}$. Assume also that the inputs x satisfy $\|x\| \leq b$ w.p. 1. Fix a set of constraint parameters $0 \leq p_\ell \leq 2, s_\ell, \mathcal{M}_\ell$ (for $\ell = 1, \dots, L$).

With probability greater than $1 - \delta$ over the draw of an i.i.d. training set x_1, \dots, x_N , every neural network $F_A^{\mathfrak{C}}(x)$ defined with equation (G.1) satisfying the following conditions:

$$\begin{aligned} \|A^\ell - M^\ell\|_{\text{sc}, p_\ell} &\leq \mathcal{M}_\ell & \forall 1 \leq \ell \leq L \\ \|\text{op}(A)_\ell\| &\leq s_\ell & \forall 1 \leq \ell \leq L-1 \\ \|A_L^\top\|_{2,\infty} &\leq s_L, \end{aligned} \quad (\text{G.8})$$

also satisfies the following generalization bound:

$$\begin{aligned} \mathbb{E} [\mathbb{I}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbb{I}(F_A(x_i), y_i) \\ \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta})}{N}} + \frac{208\sqrt{\log_2(4\Gamma^{\mathfrak{C}})} \log(N)[\mathcal{B}+1]}{\sqrt{N}} (1 + L \mathbb{L}_1)^{\frac{p}{2+p}} \mathcal{R}_{\mathcal{M}, \text{s}, p, b}^{\mathfrak{C}}, \end{aligned} \quad (\text{G.9})$$

where $\mathcal{R}_{\mathcal{M}, \text{s}, p, b}^{\mathfrak{C}} :=$

$$\left[\sum_{\ell=1}^L \left[b \prod_i \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} \left[\frac{\mathcal{M}_\ell^{p_\ell}}{s_\ell^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} r_{\mathfrak{C}, \ell}^{\frac{2}{p_\ell+2}} [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}} \quad (\text{G.10})$$

with $r_{\mathfrak{C}, \ell} := \frac{\mathcal{M}_\ell^{p_\ell}}{s_\ell^{p_\ell}}$ and $\Gamma^{\mathfrak{C}} := [b+1][L \mathbb{L}_1 + 1] \mathcal{WAN} \prod_{i=1}^L [\rho_i + 1] s_i + 1$.

Proof. The proof is exactly the same as that of Proposition E.11 with $r_\ell^{\frac{2}{p_\ell+2}} \bar{w}_\ell^{\frac{2}{p_\ell+2}} \tilde{w}_\ell^{\frac{p_\ell}{p_\ell+2}}$ replaced by $r_{\mathfrak{C}, \ell}^{\frac{2}{p_\ell+2}} [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}}$ and the initial constant of 98 replaced by 72. \square

G.4 Post Hoc Result for CNNs

Based on the above, we are finally in a position to present the following analogue of Theorem E.2 for the convolutional case:

Theorem G.5. *With probability greater than $1 - \delta$, for any $b \in \mathbb{R}^+$, every trained neural network F_A and sampling distribution with $\|x\| \leq b$ w.p. 1 satisfy the following generalization gap for all possible values of the p_ℓ s:*

$$\begin{aligned} \mathbb{E} [\mathbb{I}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbb{I}(F_A(x_i), y_i) \\ \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(1/\delta)}{N}} + 6\mathcal{B} \sqrt{\frac{2 + \Theta_{\log}^{\mathfrak{C}}}{N}} + 416[\mathcal{B}+1] (1 + L \mathbb{L}_1)^{\frac{p}{2+p}} \frac{\sqrt{\log(\gamma_{\mathfrak{C}})} \log(N)}{\sqrt{N}} \mathcal{R}_{F_A}^{\mathfrak{C}}, \end{aligned} \quad (\text{G.11})$$

where

$$\Theta_{\log}^{\mathfrak{C}} :=$$

$$2 \log \left[4\mathcal{W}\mathcal{A} + \sup_{i=1}^N \|x_i\| \prod_i \rho_i \|\text{op}(A_i)\| \right] \left[|\log(\sup_i \|x_i\|)| + L \left[\sum_{\ell=1}^L |\log(\|\text{op}(A_\ell)\|)| + 2 \log(4\mathcal{A}\mathcal{W}) \right] \right]$$

$$\gamma_{\mathfrak{C}} :=$$

$$12 \left[[\mathfrak{B}+1][L \mathbb{L}_1 + 1] \mathcal{WAN} \prod_{i=1}^L [\rho_i \|\text{op}(A_i)\| + 1] \right] \left[4\bar{W} + \mathfrak{B} \prod_i \rho_i \|\text{op}(A_i)\| \right].$$

and where

$$\begin{aligned} \mathcal{R}_{FA}^{\mathfrak{C}} := & \left[\sum_{\ell=1}^{L-1} \left[\sup_i \|x_i\| \rho_L \|A_L^{\top}\|_{2,\infty} \prod_{i=1}^{L-1} \rho_i \|\operatorname{op}(A_i)\| \right]^{\frac{2p_{\ell}}{p_{\ell}+2}} \left[\frac{\|A_{\ell} - M_{\ell}\|_{\text{sc},p_{\ell}}^{p_{\ell}}}{\|\operatorname{op}(A_{\ell})\|^{p_{\ell}}} \right]^{\frac{2}{p_{\ell}+2}} \right. \\ & \left. + \left[b \rho_L \|A_L^{\top}\|_{2,\infty} \prod_{i=1}^{L-1} \rho_i \|A_i\| \right]^{\frac{2p_L}{p_L+2}} \left[\frac{\|A_L - M_L\|_{\text{sc},p_L}^{p_L}}{\|A_L^{\top}\|_{2,\infty}^{p_L}} \right]^{\frac{2}{p_L+2}} \bar{w}_L^{\frac{2}{p_L+2}} \underline{w}_L^{\frac{p_L}{p_L+2}} \right]^{\frac{1}{2}}. \end{aligned}$$

Proof. The proof is the same as that of Theorem E.2 with the constant 240 replaced by 208, with $\|A_{\ell}\|$ replaced by $\|\operatorname{op}(A_{\ell})\|$, and \bar{W} replaced by \mathcal{AW} .

□

G.5 The case $p_{\ell} = 0$ (CNN)

For $p_{\ell} = 0$ we obtain the following immediate corollary from Proposition G.4.

Corollary G.6. *Instate the notation and assumptions of Proposition G.4, and assume additionally that $p_{\ell} = 0$ for all ℓ . We have*

$$\begin{aligned} & \mathbb{E} [\ell(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \ell(F_A(x_i), y_i) \\ & \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta})}{N}} + \frac{208\sqrt{\log_2(4\Gamma^{\mathfrak{C}})} \log(N)[\mathcal{B}+1]}{\sqrt{N}} \left[\sum_{\ell=1}^L [U_{\ell} + d_{\ell-1}] \tilde{r}_{\ell} \right], \end{aligned} \quad (\text{G.12})$$

where \tilde{r}_{ℓ} is an a priori upper bound on $\operatorname{rank}(A_{\ell})$ and as usual, $\Gamma^{\mathfrak{C}} = [b+1][L\mathcal{L}_1+1]\mathcal{WAN} \prod_{i=1}^L [\rho_i+1]s_i+1$.

With a union bound, we achieve the following post hoc version:

Theorem G.7. *W.p. $\geq 1 - \delta$ over the draw of the training set, every neural network satisfies:*

$$\mathbb{E} [\ell(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \ell(F_A(x_i), y_i) \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta})}{N}} + \frac{208\sqrt{\log_2(4\Gamma_R^{\mathfrak{C}})} \log(N)[\mathcal{B}+1]}{\sqrt{N}} \quad (\text{G.13})$$

where $\Gamma_R^{\mathfrak{C}} = [\mathcal{B}+2][L\mathcal{L}_1+1]\mathcal{WAN} \prod_{i=1}^L [\rho_i+1][\|A_i\|+2]+1$.

Proof. The proof is nearly exactly the same as that of Theorem E.8. □

G.6 Results for CNNs with Loss Function Augmentation

Proposition G.8. *For any granularity $\epsilon \geq \frac{1}{N}$ and any values of $1 \leq b_{\ell_1}, \dots, b_{L-1}, s_{\ell}, \rho_{\ell}$, there exists a cover \mathcal{C} of the augmented loss class $\mathcal{L}^{\mathfrak{C}, \lambda_{\text{aug}}} := \{\lambda_{\text{aug}}(F_A, x, y) : F_A \in \mathcal{F}_{\mathcal{B}}^{\mathfrak{C}}\}$ with cardinality bounded as follows:*

$$\begin{aligned} \log(|\mathcal{C}|) & \leq 216 \log(4\Gamma^{\mathfrak{C}, \lambda_{\text{aug}}}) \times \\ & \left[\frac{L(1+\mathcal{L}_1)}{\min(\epsilon, 1)} \right]^{\frac{2p}{2+p}} \sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L s_i \rho_i \right]^{\frac{2p_{\ell}}{p_{\ell}+2}} [U_{\ell} + d_{\ell-1}]^{\frac{2}{p_{\ell}+2}} \min(U_{\ell}, d_{\ell-1})^{\frac{p_{\ell}}{p_{\ell}+2}} W_{\ell}^{\frac{p_{\ell}}{p_{\ell}+2}}. \end{aligned} \quad (\text{G.14})$$

Proof. The proof is similar to that of Proposition G.3 and E.4 with the main difference only in the calculation of the log terms. By Lemma E.3 and Propositions G.1 and G.2, there exists a cover \mathcal{C} of

the loss class $\lambda_{aug}(F_A)$ with cardinality satisfying

$$\begin{aligned} \log(|\mathcal{C}|) &\leq \sum_{\ell=1}^L \log(\mathcal{N}_{\infty,\ell}(\mathcal{B}_\ell^{\mathfrak{C}}, \epsilon_\ell, N, 3b_{\ell-1})) \\ &\leq 72 \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b_{\ell-1}}{\epsilon_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}} \log_2(\Gamma_{\mathcal{F}_r^{\mathfrak{C}},\ell}^{\mathfrak{C},\lambda_{aug}}). \end{aligned} \quad (\text{G.15})$$

Next, recall that

$$\epsilon_\ell = \frac{b_\ell \epsilon}{L \hat{\rho}_\ell} \quad (\text{G.16})$$

and $\hat{\rho}_{\ell_1} = \max_\ell \rho_{\ell_1 \rightarrow \ell} / b_\ell \leq \max_\ell \rho_{\ell_1 \rightarrow \ell} / (\min(1, L_1^{-1})) \leq \rho_{\ell_1 \rightarrow \ell} \rho_i s_i [1 + L_1] \leq \rho_{\ell_1 \rightarrow L} \rho_i s_i [1 + L_1] = \rho_i \prod_{i=\ell_1+1}^L L \rho_i s_i [1 + L_1]$, thus

$$\epsilon_\ell \geq \frac{\epsilon}{L \rho_{\ell \rightarrow L} [1 + L_1]}. \quad (\text{G.17})$$

Thus, by a nearly identical calculation to that in equation (G.7) we obtain

$$\begin{aligned} \max_\ell \Gamma_{\mathcal{F}_r^{\mathfrak{C}},\ell}^{\mathfrak{C},\lambda_{aug}} &\leq \left[\left(\frac{16 [\max_\ell \frac{\mathcal{M}_\ell p_\ell}{s_\ell p_\ell} + 1] \prod_{i=1}^L [[\rho_i + 1] s_i + 1]^3 [b + 1] [L L_1 + 1] \mathcal{W} \mathcal{A}^{1.5}}{\min(\epsilon, 1)} + 7 \right) N \mathcal{A} \right] \\ &\leq \left[64 \prod_{i=1}^L [[\rho_i + 1] s_i + 1]^3 [b + 1] [L_1 + 1] \mathcal{W} \mathcal{A}^3 N^2 \right] \leq [4 \underline{\Gamma}^{\mathfrak{C},\lambda_{aug}}]^3 \end{aligned} \quad (\text{G.18})$$

where $\underline{\Gamma}^{\mathfrak{C},\lambda_{aug}} := [b + 1] [L (L_1 + 1)] \mathcal{W} \mathcal{A} N \prod_{i=1}^L [[\rho_i + 1] s_i + 1]$. Thus, we can continue from equation (G.15): $\log(|\mathcal{C}|) \leq$

$$\begin{aligned} &216 \log(4 \underline{\Gamma}^{\mathfrak{C},\lambda_{aug}}) \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b_{\ell-1}}{\epsilon_\ell} \right]^{\frac{2p_\ell}{p_\ell+2}} [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}} \\ &\leq 216 \log(4 \underline{\Gamma}^{\mathfrak{C},\lambda_{aug}}) \sum_{\ell=1}^L \left[\frac{\mathcal{M}_\ell b_{\ell-1} \rho_\ell L [1 + L_1] \prod_{i=\ell+1}^L s_i \rho_i}{\epsilon} \right]^{\frac{2p_\ell}{p_\ell+2}} \\ &\quad \times [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}} \\ &\leq 216 \log(4 \underline{\Gamma}^{\mathfrak{C},\lambda_{aug}}) \left[\frac{L (1 + L_1)}{\min(\epsilon, 1)} \right]^{\frac{2p}{2+p}} \sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L s_i \rho_i \right]^{\frac{2p_\ell}{p_\ell+2}} \\ &\quad \times [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}}, \end{aligned}$$

as expected. \square

Next, similarly to the proof of Proposition G.4, we can obtain:

Proposition G.9. Fix some reference/initialization matrices $M^1 \in \mathbb{R}^{w_1 \times w_0} = \mathbb{R}^{w_1 \times d}, \dots, M^L \in \mathbb{R}^{w_L \times w_{L-1}} = \mathbb{R}^{\mathcal{C} \times w_{L-1}}$. Assume also that the inputs x satisfy $\|x\| \leq b$ w.p. 1. Fix a set of constraint parameters $0 \leq p_\ell \leq 2, s_\ell, \mathcal{M}_\ell$ (for $\ell = 1, \dots, L$).

With probability greater than $1 - \delta$ over the draw of an i.i.d. training set x_1, \dots, x_N , every neural network $F_A^{\mathfrak{C}}(x)$ defined with equation (G.1) satisfying the following conditions:

$$\begin{aligned} \|A^\ell - M^\ell\|_{\text{sc}, p_\ell} &\leq \mathcal{M}_\ell & \forall 1 \leq \ell \leq L \\ \|\text{op}(A)_\ell\| &\leq s_\ell & \forall 1 \leq \ell \leq L-1 \\ \|A_L^\top\|_{2,\infty} &\leq s_L, \end{aligned} \quad (\text{G.19})$$

also satisfies the following generalization bound:

$$\begin{aligned} \mathbb{E} [\mathbf{l}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbf{l}(F_A(x_i), y_i) \\ \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(\frac{4}{\delta})}{N}} + \frac{361 \sqrt{\log_2(4\Gamma^{\mathfrak{C}, \lambda_{aug}}) \log(N)[\mathcal{B}+1]}}{\sqrt{N}} (1 + L \mathbf{L}_1)^{\frac{p}{2+p}} \mathcal{R}_{\mathcal{M}, \mathbf{s}, p, b}^{\mathfrak{C}, \lambda_{aug}} \end{aligned} \quad (\text{G.20})$$

where $\mathcal{R}_{\mathcal{M}, \mathbf{s}, p, b}^{\mathfrak{C}, \lambda_{aug}} :=$

$$\left[\sum_{\ell=1}^L \left[b_{\ell-1} \prod_{i=\ell}^L \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} \left[\frac{\mathcal{M}_\ell^{p_\ell}}{s_\ell^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} \mathbf{r}_{\mathfrak{C}, \ell}^{\frac{2}{p_\ell+2}} [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}}$$

with $\mathbf{r}_{\mathfrak{C}, \ell} := \frac{\mathcal{M}_\ell^{p_\ell}}{s_\ell^{p_\ell}}$ and $\Gamma^{\mathfrak{C}, \lambda_{aug}} := [b+1][L(\mathbf{L}_1+1)]\mathcal{WAN} \prod_{i=1}^L [\rho_i + 1] s_i + 1$.

Proof. The proof is the same as that of Proposition G.4 with an additional factor of $\sqrt{3}$, relying on Proposition G.8 instead of Proposition G.3. \square

Finally, we can obtain the following post hoc version from the above.

Theorem G.10. *With probability greater than $1 - \delta$, for any $b \in \mathbb{R}^+$, every trained neural network F_A satisfies the following generalization gap for all possible values of the p_ℓ s:*

$$\begin{aligned} \mathbb{E} [\mathbf{l}(F_A(x), y)] - \frac{1}{N} \sum_{i=1}^N \mathbf{l}(F_A(x_i), y_i) \leq 6[\mathcal{B}+1] \sqrt{\frac{\log(1/\delta)}{N}} + \\ 6\mathcal{B} \sqrt{\frac{2 + \Theta_{\log}^{\mathfrak{C}, \lambda_{aug}}}{N}} + 722[\mathcal{B}+1][L(1 + \mathbf{L}_1)]^{\frac{p}{2+p}} \frac{\sqrt{\log(\gamma_{\mathfrak{C}, \lambda_{aug}})} \log(N)}{\sqrt{N}} \mathcal{R}_{F_A}^{\mathfrak{C}, \lambda_{aug}}, \end{aligned} \quad (\text{G.21})$$

where

$$\begin{aligned} \Theta_{\log}^{\mathfrak{C}, \lambda_{aug}} &:= 2 \log \left[4\mathcal{WAN} + \sup_{i=1}^N \|x_i\| \prod_i \rho_i \|\text{op}(A_i)\| \right] \times \\ &\left[|\log(\sup_i \|x_i\|)| + L \left[\sum_{\ell=1}^L |\log(\|\text{op}(A_\ell)\|)| + \sum_{\ell=1}^{L-1} \log(|\mathfrak{B}_{\ell-1, A}|) + 2 \log(4\mathcal{AW}) \right] \right], \\ \gamma_{\mathfrak{C}, \lambda_{aug}} &:= 12 \left[\left[[\mathfrak{B}+1][L\mathbf{L}_1+1]\mathcal{WAN} \prod_{i=1}^L [\rho_i \|\text{op}(A_i)\| + 1] \right] \right] \left[4\bar{W} + \mathfrak{B} \prod_i \rho_i \|\text{op}(A_i)\| \right], \end{aligned}$$

and where

$$\begin{aligned} \mathcal{R}_{F_A}^{\mathfrak{C}, \lambda_{aug}} &:= \left[\sum_{\ell=1}^L \left[\sup_i \|x_i\| \|A_L^\top\|_{2,\infty} \mathfrak{B}_{\ell-1, A} \rho_\ell \prod_{i=\ell+1}^L \rho_i s_i \right]^{\frac{2p_\ell}{p_\ell+2}} \times \right. \\ &\left. \left[\frac{\|A_\ell - M_\ell\|_{\text{sc}, p_\ell}^{p_\ell}}{\|\text{op}(A_\ell)\|^{p_\ell}} \right]^{\frac{2}{p_\ell+2}} [U_\ell + d_{\ell-1}]^{\frac{2}{p_\ell+2}} \min(U_\ell, d_{\ell-1})^{\frac{p_\ell}{p_\ell+2}} W_\ell^{\frac{p_\ell}{p_\ell+2}} \right]^{\frac{1}{2}}. \end{aligned} \quad (\text{G.22})$$

Proof. The proof is the same as that of Theorem E.6 with the original constant of 416 replaced by 361 and $\|A_\ell\|$ replaced by $\|\text{op}(A_\ell)\|$. \square

H Generalization Bound for Linear Networks with Schatten Constraints

In this section, we briefly mention that it is possible to prove alternative bounds for linear networks assuming the presence of Schatten quasi norm constraints on the factor matrices (instead of L^2 constraints). Indeed, a generalization of Theorem 3.1 (Theorem I.2) allows also one to translate such constraints into a single Schatten quasi norm constraint on the linear map A .

Theorem H.1. *For any $0 \leq p_\ell \leq 1$, write p for the conjugate such that $\frac{1}{p} = \sum_{\ell=1}^L \frac{1}{p_\ell}$. W.p. $\geq 1 - \delta$ over the draw of the training set, every linear network as defined in equation (3.5) satisfies the following generalization bound:*

$$\text{GAP} - O \left(\mathcal{B} \sqrt{\frac{\log(1/\delta)}{N}} \right) \leq \tilde{O} \left(\sqrt{\left[L_1 \mathfrak{B} \right]^{\frac{2p}{2+p}} \frac{\left[\sum_{\ell=1}^L \frac{p \|B_\ell\|_{\text{sc}, p_\ell}^{p_\ell}}{p_\ell} \right]^{\frac{2}{2+p}}}{N} [\mathcal{C} + d]} \right) \quad (\text{H.1})$$

where $\mathfrak{B} = \sup_{i=1}^N \|x_i\|$, the \tilde{O} notation hides polylogarithmic factors.

I Known Results

This section collects some well-known results which are key to our proof, and which we include for the benefit of self-completeness.

Theorem I.1 ([24], Theorem 1). *Let $w \geq \min(\mathcal{C}, d)$ and let B_1, \dots, B_L be matrices such that $B_1 \in \mathbb{R}^{w \times d}, B_L \in \mathbb{R}^{\mathcal{C} \times w}, B_2, \dots, B_{L-1} \in \mathbb{R}^{w \times w}$. For any matrix $A \in \mathbb{R}^{w \times d}$ we have*

$$L \|A\|_{\text{sc}, \frac{2}{L}}^{\frac{2}{L}} = \min \sum_{\ell=1}^L \|B_\ell\|^2 \quad \text{subject to } B_L B_{L-1} \dots B_1 = A. \quad (\text{I.1})$$

Theorem I.2 ([92, 93, 94]). *Let $w \geq \min(\mathcal{C}, d)$ and let B_1, \dots, B_L be matrices such that $B_1 \in \mathbb{R}^{w \times d}, B_L \in \mathbb{R}^{\mathcal{C} \times w}, B_2, \dots, B_{L-1} \in \mathbb{R}^{w \times w}$. Fix p_1, \dots, p_L such that $\frac{1}{p} = \sum \frac{1}{p_i}$.*

For any matrix $A \in \mathbb{R}^{w \times d}$ we have

$$\|A\|_{\text{sc}, p}^p = \min \sum_{\ell=1}^L \left[\frac{p \|B_\ell\|_{\text{sc}, p_\ell}^{p_\ell}}{p_\ell} \right] \quad \text{subject to } B_L B_{L-1} \dots B_1 = A. \quad (\text{I.2})$$

Proposition I.3 (Cf. Proposition 5 in [16] (with $U = 1$), see also [65]). *Let $x_1, \dots, x_N \in \mathbb{R}^d$ be such that $\|x_i\| \leq b$ for all $i \leq N$. For any fixed choice of reference matrix $M \in \mathbb{R}^{m \times d}$, consider the set \mathcal{F}_M of matrices $A \in \mathbb{R}^{m \times d}$ such that $\|A - M\|_{\text{Fr}} \leq a$. For any $\epsilon > 0$ there exists a cover $\mathcal{C} \subset S_M$ such that for all $A \in \mathcal{F}_M$, there exists $\tilde{A} \in \mathcal{C}$ such that for all $i \leq N$, $\|(A - M)x_i\|_\infty \leq \epsilon$ and*

$$\log(|\mathcal{C}|) \leq \frac{36a^2b^2}{\epsilon^2} \log_2 \left[\left(\frac{8ab}{\epsilon} + 7 \right) mN \right]. \quad (\text{I.3})$$

Lemma I.4 (Lemma 8 in [4]). *The (internal) covering number \mathcal{N} of the ball of radius κ in dimension d (with respect to any norm $\|\cdot\|$) can be bounded by:*

$$\mathcal{N} \leq \left[\frac{3\kappa}{\epsilon} \right]^d \leq \left(\frac{3\kappa}{\epsilon} + 1 \right)^d \quad (\text{I.4})$$

Recall also the following lemma from [83]:

Proposition I.5. *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 $\|X_{i,\cdot}\|_2 \leq b, \forall i = 1, \dots, n$. For $U_{a,b}(X) = \{X\alpha : \|\alpha\|_2 \leq a, \alpha \in \mathbb{R}^d\}$, we have*

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

Theorem I.6 (Generalization bound from Rademacher complexity, cf. e.g., [95], [96], [97] etc.).
Let Z, Z_1, \dots, Z_N be i.i.d. 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

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(f(Z)) \leq \frac{1}{N} \sum_{i=1}^N f(z_i) + 2 \mathfrak{R}_S(\mathcal{F}) + 3 \sqrt{\frac{\log(4/\delta)}{2N}},$$

where $\mathfrak{R}_S(\mathcal{F})$ can be either the empirical or expected Rademacher complexity. In particular, if $f^* \in \arg \min_{f \in \mathcal{F}} \mathbb{E}(f(Z))$ and $\hat{f} \in \arg \min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N f(z_i)$, then

$$\mathbb{E}(\hat{f}(Z)) \leq \mathbb{E}(f^*(Z)) + 4 \mathfrak{R}_S(\mathcal{F}) + 6 \sqrt{\frac{\log(2/\delta)}{2N}}.$$

Lemma I.7 (Dudley's entropy theorem [98], this version from [16, 99], see also [1, 74, 4]). *Let \mathcal{F} be a real-valued function class taking values in $[0, \mathcal{B}]$. Let S be a finite sample of size N . We have the following relationship between the Rademacher complexity $\mathfrak{R}(\mathcal{F}|_S)$ and the L^2 covering number $\mathcal{N}(\mathcal{F}|_S, \epsilon, \|\cdot\|_2)$.*

$$\mathfrak{R}(\mathcal{F}|_S) \leq \inf_{\alpha > 0} \left(4\alpha + \frac{12}{\sqrt{N}} \int_{\alpha}^{\mathcal{B}} \sqrt{\log \mathcal{N}(\mathcal{F}|_S, \epsilon, \|\cdot\|_2)} d\epsilon \right),$$

where the norm $\|\cdot\|_2$ on \mathbb{R}^N is defined by $\|x\|_2^2 = \frac{1}{N} (\sum_{i=1}^N |x_i|^2)$.