GENERALIZATION BOUNDS VIA DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper provides a suite of mathematical tools to bound the generalization error of networks that possess low-complexity distillations — that is, when there exist simple networks whose softmax outputs approximately match those of the original network. The primary contribution is the aforementioned bound, which upper bounds the test error of a network by the sum of its training error, the distillation error, and the complexity of the distilled network. Supporting this, secondary contributions include: a generalization bound which can handle convolutions and skip connections, a generalization analysis of the compression step leading to a bound with small width- and depth-dependence via weight matrix stable ranks, and a sampling theorem to sparsify dense networks. The bounds and their behavior are illustrated empirically on the standard `mnist` and `cifar` datasets.

1 OVERVIEW AND MAIN RESULTS

Generalization bounds are statistical tools which take as input various measurements of a predictor on training data, and output a performance estimate for unseen data — that is, they estimate how well the predictor *generalizes* to unseen data. Despite extensive development spanning many decades (Anthony & Bartlett, 1999), there is growing concern that these bounds are not only disastrously loose (Dziugaite & Roy, 2017), but worse that they do not correlate with the underlying phenomena (Jiang et al., 2019b), and even that the basic method of proof is doomed (Zhang et al., 2016; Nagarajan & Kolter, 2019). As a quick calculation backing the issue, the largest network in Figure 1 has a modest size: 6 layers each with 1024 nodes, and generalizes just fine — but the VC dimension is 30 million, which is 500 times larger than the amount of data, giving an upper bound on the probability of misclassification of roughly $\sqrt{500}$.

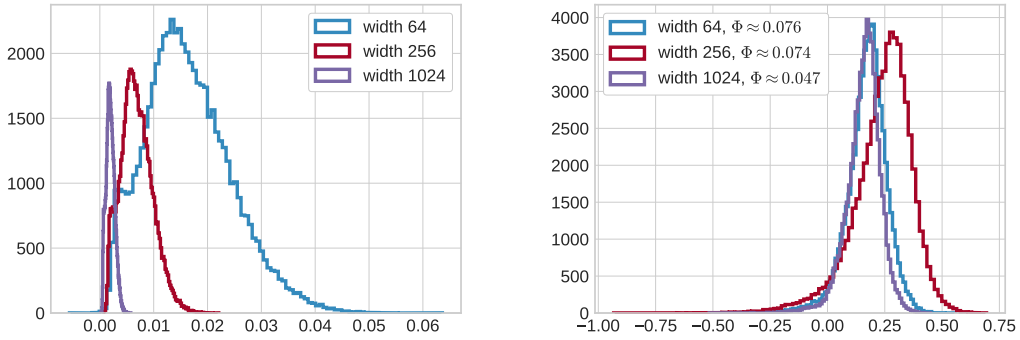
Contrary to this dilemma, there is evidence that these networks can often be compressed or *distilled* into simpler networks — ones, ostensibly, whose generalization is reasonably analyzed. Indeed, this line of thought predates the preceding generalization concerns (Buciluǔ et al., 2006), but has received increasing attention in recent years (Hinton et al., 2015). Indeed, it can be found in many guises, for instance throughout the *pruning* or *lottery ticket* literature, which posits that vast swaths of parameters can simply be deleted with minimal harm to network performance (Frankle & Carbin, 2019; Frankle et al., 2020; Su et al., 2020).

The purpose of this work is to directly apply the distillation insight to tackle the generalization dilemma. Conceptually, this goes beyond the standard distillation perspective, whose goal is to start with a complicated network, but to eventually discard it and utilize a simpler network which has hopefully inherited certain key properties. Instead, the perspective here is that the distilled network absorbs the original network so wholly that it can be used to answer questions about the original network.

Concretely, the questions we answer here about networks via their distillations are generalization upper bounds. Rather than a standard generalization upper bound in terms of network complexity, our core contribution is an elementary and widely-applicable upper bound on generalization via distillation properties:

$$\text{test error} \lesssim \text{training error} + \text{distillation error} + \text{distillation complexity}.$$

As will be outlined momentarily, this work also provides (a) an instantiation of this lemma applicable to residual networks, (b) an explicit compression analysis via pruning with a corresponding generalization bound, (c) empirical support, most notably showing width-independence subsequent to distillation.



(a) Margins before distillation.

(b) Margins after distillation.

Figure 1: **Width independence.** Fully-connected 6-layer networks of widths $\{64, 256, 1024\}$ were trained on `mnist` until training error zero; the margin histograms, normalized by the generalization bound in Lemma 3.2, all differ, and are close to zero. After distillation, the margin distributions are far from zero and nearly the same. In the distillation legend, the second term Φ denotes the distillation error, as defined below. Experiment details may be found in Section 2, which includes a description of margin distributions and ways to interpret them.

1.1 Core distillation lemma, and a convenient corollary

This subsection will present the core distillation lemma in an abstract form, which lacks any mention of deep networks. After some discussion, a corollary for general computation graphs will follow.

To build up to the lemma, first recall the distillation setup. Given a multi-class predictor $f : \mathbb{R}^d \to \mathbb{R}^k$, the goal is to find another predictor $g : \mathbb{R}^d \to \mathbb{R}^k$ which is simpler, but still similar in *distillation error* $\Phi_{\gamma,n}$, defined along with ϕ_γ, *the softmax with temperature* $\gamma > 0$, as

$$\Phi_{\gamma,n}(f,g) := \frac{1}{n}\sum_{i=1}^{n}\left|\phi_\gamma(f(x_i),y_i) - \phi_\gamma(g(x_i),y_i)\right|, \qquad \phi_\gamma(f(x_i),y_i) = \frac{\exp(f(x_i)_{y_i}/\gamma)}{\sum_{j=1}^{k}\exp(f(x_i)_j/\gamma)}.$$

In other words, the softmax outputs for the original predictor f and its distillation g must be close on average along coordinates corresponding to the true label y. The quantity $\gamma > 0$ is sometimes called a *temperature* (Hinton et al., 2015), but in the present work it also plays the role of a *margin*: indeed, the statistical learning theory literature has used such a γ in a softmax-like *loss function* to produce sophisticated generalization bounds via the margins perspective (Srebro et al., 2010, Section 5.1). There are many variants of this scheme, for instance matching the outputs of f and g directly (Buciluǎ et al., 2006), but the softmax is relied upon heavily in the present work.

The core lemma is as follows, but here is its flavor in words. We are handed a function f, and we don't understand its generalization properties. However, we have some simple function classes $(\mathcal{G}_j)_{j\geq 1}$ which we understand pretty well, and we have some prior belief on which $\mathcal{G}_j$ are close to f in the sense of distillation error, namely $\Phi_{\gamma,n}$. We compare f to these $\mathcal{G}_j$ one by one, and when the distillation error is small, we announce that the original f inherits the generalization properties of $\mathcal{G}_j$, despite not necessarily looking anything like the functions in $\mathcal{G}_j$.

Lemma 1.1. *Let temperature (margin) parameter $\gamma > 0$ be given, along with any sets of multiclass predictors $(\mathcal{G}_j)_{j\geq 1}$ with $\mathcal{G}_j \ni g : \mathbb{R}^d \to \mathbb{R}^k$, along with* prior weights *$(p_j)_{j\geq 1}$ on these sets, where $p_j \geq 0$ and $\sum_j p_j = 1$. Then with probability at least $1-\delta$ over an iid draw of data $((x_i,y_i))_{i=1}^n$, every measurable function $f : \mathbb{R}^d \to \mathbb{R}^k$ satisfies*

$$\Pr[f(x) \neq y] \leq \frac{2}{n}\sum_i \left(1 - \phi_\gamma(f(x_i),y_i)\right) + \frac{16}{n} + 6\sqrt{\frac{\ln(2/\delta) + \ln\lg(n)}{n}}$$

$$+ \inf_{\substack{j\geq 1\\ g\in\mathcal{G}_j}}\left[8\Phi_{\gamma,n}(f,g) + \frac{1}{\gamma}\mathrm{Rad}\left(\left\{(x,y)\mapsto g(x)_y - \max_{i\neq y} g(x)_i : g\in\mathcal{G}_j\right\}\right) + 6\sqrt{\frac{\ln(1/p_j)}{n}}\right].$$

As was foreshadowed, the left hand side is the test error, and the right hand side searches over the classes $\mathcal{G}_j$, penalized by the prior $\ln(1/p_j)$, considering the best $g \in \mathcal{G}_j$ within each in terms of distillation measure $\Phi_{\gamma,n}$. One thing that was *not* foreshadowed is that the training error is measured with the softmax as well — this is fine, since, in fact, it is upper bounded by the cross-entropy loss, which is the typical objective function in the setting. The other new element is the clunky distillation complexity term; but in fact, it is just the *Rademacher complexity* of the *multi-class margins* $g(x)_y - \max_{i \neq y} g(x)_i$ *of* $\mathcal{G}_j$, and as is usual in margin analyses is scaled by $1/\gamma$.

Now that the bound has been stated, let us return to the initial context of the paper, namely the difficulty of producing generalization bounds for deep networks. Suppose, as was discussed extensively in (Zhang et al., 2016), that a deep network is trained on random labels, and this is plugged in as f to Lemma 1.1. Achieving good distillation error $\Phi_{\gamma,n}$ will therefore require high complexity, scaling with the number of examples, and it will not be possible to select a simple g. On the other hand, if f is trained on true labels, then finding a simple g should be possible. Indeed, Figure 1 aims to capture this by training 3 different network architectures on `mnist`— each fully-connected with 6 layers and 2 skip connections, the architectures differing by choosing widths $\{64, 256, 1024\}$. Despite the original generalization bounds (and corresponding margin histograms) differing greatly, they all distill to a nearly identical simple predictor with good generalization properties.

As mentioned, and as will be discussed in detail in Section 2, the bounds are normalized by a specific generalization bound, meaning a specific choice of $\mathcal{G}_j$ and Rademacher complexity upper bound. Skip connections have nice properties when combined with distillation, so another contribution of this work is a Rademacher complexity bound which can handle flexible computation graphs including convolutions, multi-variate gates, and skip connections.

Theorem 1.2. *Let temperature (margin) parameter $\gamma > 0$ be given, along with a fixed* computation graph architecture $\mathcal{G}$ *(as detailed in Section 1.3), parameterized by per-layer batch bounds $\vec{b}$, per-layer parameter $(2,1)$-norm bounds $\vec{r}$ layer operator norm bounds $\vec{s}$. With probability at least $1-\delta$ over an iid draw of data $((x_i, y_i))_{i=1}^n$, every measurable function $f : \mathbb{R}^d \to \mathbb{R}^k$ satisfies*

$$\Pr[f(x) \neq y] \leq \frac{2}{n} \sum_i \left(1 - \phi_\gamma(f(x_i), y_i)\right) + 6\sqrt{\frac{\ln(2/\delta)}{n}}$$
$$+ \inf_{(\vec{b},\vec{r},\vec{s}) \in \mathbb{Z}_{\geq 1}^{3L}} \inf_{g \in \mathcal{G}(\vec{b},\vec{r},\vec{s})} \frac{1}{n} \widetilde{\mathcal{O}} \left(\Phi_{\gamma,n}(f,g) + \left(\sum_i \left[r_i b_i \prod_{l=i+1}^{L} s_l \right]^{2/3} \right)^{3/2} \right),$$

where b_i depends on $(x_i)_{i=1}^n$ and should be interpreted as $\mathcal{O}(\sqrt{n})$.

This bound and its proof will be discussed in detail in Section 3. Briefly, it follows from a simplified proof of the covering number-based bound due to Bartlett et al. (2017a); this original proof was also heavily extended to analyze a different computation graph formalism by Wei & Ma (2019). In addition to the generality, the bound also uses b_i to upper bound the Frobenius norm of the input to each layer, which lends significant numerical benefits and again simplifies the proof.

Compared with Lemma 1.1, this bound makes a fixed choice of $(\mathcal{G}_j)$ and their prior probabilities, namely a fixed computation graph, and preferring smaller norms. This bounds was used to produce Figure 1, as will be discussed in Section 2.

1.2 Discussion: *why* can predictors be distilled?

The focus of this paper is on arguing that *if* a predictor can be distilled to a simple network, *then* it generalizes well. A separate question, only scratched here, is *why* networks can be distilled to simpler ones.

This subsection provides a generalization bound where, in a sense, the distillation is explicitly analyzed. Concretely, the bound first simplifies a network by sparsifying its internal connections, and then applies a generalization bound to the sparsified network. The sparsification step can be interpreted as node pruning, and could be of independent interest; it is discussed in Section 3.

Lemma 1.3. *Let data matrix $X \in \mathbb{R}^{n\times d}$ be given. Let $\mathcal{H}$ denote networks of the form $x \mapsto \sigma_L(W_L \cdots \sigma_1(W_1 x))$ with $\|W_i\|_2 \le s_i$ and σ_i is 1-Lipschitz and 1-homogeneous, and $\|W_i\|_{\mathrm{F}} \le R_i$ and width at most m. Then*

$$\mathrm{Rad}(\mathcal{H}) = \widetilde{\mathcal{O}}\left(\frac{\|X\|_{\mathrm{F}}}{n^{3/4}}\left[\prod_j s_j\right]\left[\sum_i (R_i/s_i)^{4/5}\right]^{5/4}\left[\sum_i \ln R_i\right]^{1/4}\right).$$

The term $\|X\|_{\mathrm{F}}/n^{3/4}$ should be interpreted as $\mathcal{O}(n^{-1/4})$, and deserves mention. It arises because the network is first sparsified *with knowledge of the data*, effectively baking data into the network, and making it unclear, at least with the present proof, if the usual $1/\sqrt{n}$ is possible. On the other hand, the bound does not seems to have any explicit appearance of network width: the term R_i/s_i is the square root of the *stable rank*, and there is hope it is insensitive to ambient network width.

While the bound has certain interesting properties, it is empirically usually weaker than Theorem 1.2, though it is slightly better in the early distillation phase (where both bounds are vacuously large).

Is sparsification the only thing that can lead to distillation? Probably not; this question is left to future work.

1.3 Notation

Given data $(z_i)_{i=1}^n$, the *Rademacher complexity* of predictors $\mathcal{H}$ is

$$\mathrm{Rad}(\mathcal{H}) := \mathbb{E}_{\bar\epsilon} \sup_{f\in\mathcal{H}} \frac{1}{n}\sum_i \epsilon_i f(z_i), \qquad \text{where } \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(\{-1,+1\}).$$

The Rademacher complexity is most common tool in generalization theory (Shalev-Shwartz & Ben-David, 2014), and is incorporated in Lemma 1.1 due to its convenience and wide use.

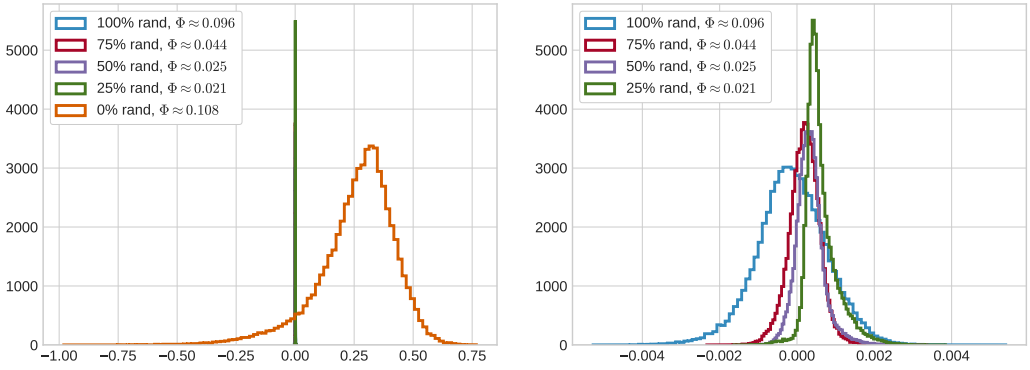
Canonical computation graphs are a flexible way to write down feedforward networks which allow for dense linear layers, convolutional layers, skip connections, and multivariate gates amongst other options, while still having an economical expression. To start the definition, given matrices $A \in \mathbb{R}^{m\times d}$ and $B \in \mathbb{R}^{n\times d}$ let $[A \rangle B] \in \mathbb{R}^{(m+n)\times d}$ denote their *row-wise concatenation*. Given a minibatch $X \in \mathbb{R}^{n\times d}$, a canonical computation graph is recursively defined via

$$X_0^\intercal := X^\intercal, \qquad X_i^\intercal := \sigma_i([W_i\Pi_i D_i \rangle F_i]X_{i-1}^\intercal) = \sigma_i\left(\begin{bmatrix} W_i\Pi_i D_i X_{i-1}^\intercal \\ F_i X_{i-1}^\intercal \end{bmatrix}\right),$$

where: σ_i is a multivariate-to-multivariate 1-Lipschitz function (measured over minibatches on either side with Frobenius norm), where the "1" is for notational convenience; F_i is a *fixed* matrix, for instance an identity mapping as in a residual network's skip connection; D_i is a *fixed* diagonal matrix selecting certain coordinates, for instance the non-skip part in a residual network; Π_i is a Frobenius norm projection of a full minibatch; W_i is a weight matrix, the trainable parameters. In Theorem 1.2, r_i is an upper bound on $\|W_i^\intercal\|_{2,1} \le r_i$ (sum of the $\|\cdot\|_2$-norms of the rows), b_i is the radius of the Frobenius norm ball which Π_i is projecting onto, and s_i is the operator norm of $X \mapsto [W_i\Pi_i D_i X^\intercal \rangle F_i X^\intercal]$. While the definition is intricate, it can not only model basic residual networks, but it is sensitive enough to be able to have $s_i = 1$ and $r_i = 0$ when residual blocks are fully zero'd out: this indeed happens during distillation (cf. Figure 3b), and motivated this definition and its analysis.

A key property or "trick" of this architecture is that it always operates at the level of batches. In particular, since every layer takes as input matrices of examples, it can output matrices with seemingly *more* examples. This is how convolution layers are handled: the input tensor is cut up into many patches, each patch effectively counting as a new example, and this is then fed into the convolution filter which now can look like a regular dense linear layer. Since the activations are arbitrary Lipschitz multivariate gates, they can encode this operation of exploding into patches and combining back into single tensors.

A few more pieces of notation. Throughout, $\|\cdot\| = \|\cdot\|_{\mathrm{F}}$ will denote Frobenius norm, and $\|\cdot\|_2$ will denote spectral norm. The sigmoid will often be used: $\sigma_{\mathrm{s}}(z)) := (1+\exp(-z))^{-1}$.



(a) Permuting different fractions of labels.

(b) Zooming in.

Figure 2: **Label randomization.** Here $\{0\%, 25\%, 50\%, 75\%, 100\%\}$ of the labels were permuted across the respective experiments. In all cases, the margin distribution is collapsed by orders of magnitude. Differences in Figure 2b are an artifact of the way distillation elements were selected.

2 Illustrative Empirical Results

This section will summarize the experimental setup, and then provide further comments on the various experiments backing the figures throughout the paper.

Experimental setup. Experiments were conducted in a simple manner, as follows. The standard `mnist` data was combined with fully-connected networks with skip links in the following architecture: given a width $m \in \{64, 256, 1024\}$, the first layer had shape $m \times d$, and thereafter pairs of layers (each with shape $m \times m$) had parallel skip connections, with ReLUs throughout, with a final fully connected layer mapping to the 10 output labels. The standard `cifar` data used a similar architecture, but with the dense layers in the skip blocks replaced with 2-dimensional convolutions using a kernel width of 3.

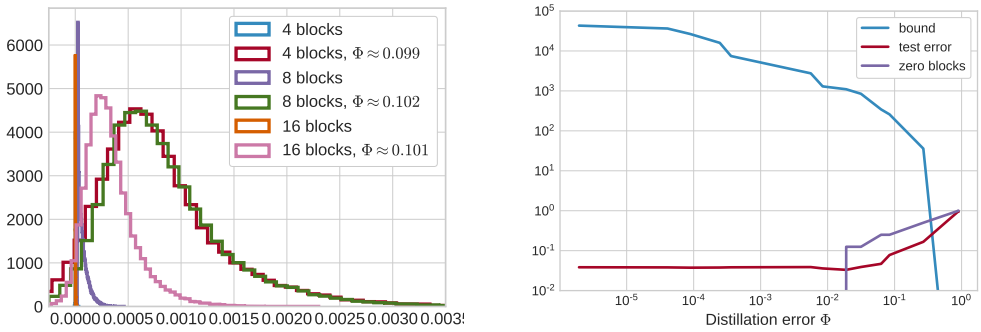
The general training procedure, as follows, was effectively an implementation of Theorem 1.2.

1. A network f was trained to roughly zero training error with mini-batch gradient descent on the cross-entropy loss with with no regularization or other techniques.
2. Thereafter, distilled networks were produced via a *relaxation sequence*: f was duplicated to produce a network g, which was then trained via gradient descent on the objective $\Phi_{1,n}(f, g) + \lambda \sum_j \|W_j^{\mathsf{T}}\|_{2,1}$, where λ was gradually increased from roughly zero until g became the 0 network.
3. To select a single g from this sequence, a few candidates were compared via their normalized margin distribution (to be discussed in detail momentarily), and the g whose 10% margin quantile was largest was selected. This explains why the plots can sometimes differ in $\Phi(f, g)$ while still having margin distributions far to the right of zero.

Margin histograms. These histograms originated in the study of boosting methods (Schapire & Freund, 2012), but have proved useful in studying deep networks as well (Bartlett et al., 2017a; Jiang et al., 2019a). The concrete approach here, as introduced in (Bartlett et al., 2017a), is to normalize the entire margin distribution by (some upper bound on) the Rademacher complexity $\mathrm{Rad}(\mathcal{G})$ for the predictor family: given $g \in \mathcal{G}$,

$$\left\{ \frac{g(x_i)_{y_i} - \max_{j \neq y_i} g(x_i)_j}{\mathrm{Rad}(\mathcal{G})} : i \in \{1, \ldots, n\} \right\}.$$

By this choice, margin distributions can be compared across predictor families and even datasets. Moreover, the margin diagrams have a concrete interpretation: if most of the normalized margins fall to the right of some value c, then a generalization bound of roughly $1/c$ is implied.



(a) Mild depth independence on `cifar` with many-block convolutional networks.

(b) Block deletion, test error, and generalization bound on a large 8-block `mnist` network.

Figure 3: **The role of depth.**

Width indepedence. Now that the experimental setup has been detailed, Figure 1 is saying: starting from networks of different widths and then slowly applying more regularization, the best networks amongst all regularization paths have similar generalization properties. Note that the distilled networks have negative margins on a fraction of the data: this trend was observed across experiments, and seems to imply that distillation quickly gives up on difficult examples in order to quickly decrease complexity.

Depth. Figure 3 shows two effects coming from massive depth. In Figure 3a, there are many blocks, but still distillation can pull the sharply concentrated margin distribution away from the origin towards a nearly common place. Figure 3b shows the generalization bound (from Lemma 3.2), test error, and number of block deletions on a network with 8 blocks (20 layers), each of width 64, all as a function of distillation error $\Phi_{\gamma,n}$. Notably, the test error remains flat until just before $\Phi_{\gamma,n}$ reaches $1/10$. Interestingly, the first block deletion (defined as weight matrix norms falling below $1/100$ has no initial effect on test error.

As Figure 3a is the only appearance of `cifar` and convolutional networks, it is worth pointing out that the bound is much looser in this regime, offering a valuable direction for future work.

Random labels. A standard sanity check for generalization bounds is whether they can reflect the difficulty of fitting random labels (Zhang et al., 2016). While it has been empirically shown that Rademacher bounds do sharply reflect the presence of random labels (Bartlett et al., 2017a, Figures 2 & 3), the effect is extremely clear with distillation: even randomizing just 25% shrinks the margin distribution by multiple orders of magnitude.

3 Analysis Overview and Sketch of Proofs

This section sketches the proofs of all the main results, and provides further context and connections to the literature. Full detailed proofs appear in the appendices.

3.1 Proof Sketch of Core Distillation Lemma 1.1

The core lemma, Lemma 1.1, is a corollary of the following more fundamental core lemma.

Lemma 3.1. *Let $(\mathcal{G}_j)_{j\geq 1}$ be a sequence of classes of bounded functions, meaning $\mathcal{G}_j \ni g : \mathbb{R}^d \to [0,1]$, and let $(p_j)_{j\geq 1}$ with $p_j \geq 0$ and $\sum_j p_j = 1$ be corresponding prior weights. Then with probability at least $1-\delta$ over an iid draw of $S := (z_1, \ldots, z_n)$, every bounded measurable function*

$f : \mathbb{R}^d \to [0,1]$ *satisfies*

$$\mathbb{E}f(z) \le \frac{1}{n}\sum_i f(z_i) + \frac{8}{n} + 3\sqrt{\frac{\ln(2/\delta) + \ln\lg(n)}{n}}$$
$$+ \inf_{\substack{j\ge 1\\ g\in\mathcal{G}_j}} \left[\frac{4}{n}\sum_i |f(z_i) - g(z_i)| + 2\mathrm{Rad}(\mathcal{G}_j) + 3\sqrt{\frac{\ln(1/p_j)}{n}} \right].$$

On the one hand, this bound appears natural, indeed it has the flavor of model selection theorems. On the other hand, the bound is on f, which is *not* chosen via model selection, and instead a model selection penalty is applied after the fact.

In this generality, this tool can be applied in multiple places throughout the literature. For instance, if pruning results chose to minimize distillation error rather than achieving good prediction performance (Frankle & Carbin, 2019; Frankle et al., 2020; Su et al., 2020), they could imply generalization properties of the original network. Secondly, there are many generalization bounds which apply to a compressed or distilled network, but not the original, and Lemma 3.1 can be used to close this gap and provide a generalization bound for the original network. As a concrete example, Arora et al. (2018) provide a procedure to compress/distill a network (cf. their Algorithm 1) and a generalization bound for the distilled network (cf. their Theorems 2.1, 4.1, 5.1, providing our distillation complexity term), but no bound on the test error of the original network (cf. Remark (1) below their Theorem 2.1); consequently, our work can be used to complement theirs by computing the distillation error term of their compression procedure (either empirically or via their Lemma 3) and then applying Lemma 3.1 to obtain a generalization bound for the original uncompressed network.

Turning now to the proof, let $\mathcal{F}$ denote bounded measurable functions, specifically $\mathcal{F} \ni f : \mathbb{R}^d \to [0,1]$. It is now crucial that $(\mathcal{G}_j)_{j\ge 1}$ was chosen before seeing f or data: given a target distillation error τ_k, carve up all of $\mathcal{F}$ into strips based on $\mathcal{G}_j$:

$$\mathcal{F}_{j,k} := \left\{ f \in \mathcal{F} : \inf_{g\in\mathcal{G}_j} \frac{1}{n}\sum_i |f(z_i) - g(z_i)| \le \tau_k \right\}.$$

This quantity is akin to an empirical covering number, however instead of starting with some class of functions and deriving a covering set, we *start* with a covering set $\mathcal{G}_j$, and write down the largest subset of $\mathcal{F}$ which it can cover.

Now that the cover formalism has been pinned down, the rest of the proof is easy. We can get a Rademacher complexity bound on each $\mathcal{F}_{j,k}$ in the standard covering number way (Shalev-Shwartz & Ben-David, 2014). Then unioning over all over all $(\mathcal{G}_j)_{j\ge 1}$ (this is where the prior comes in) and over a grid of choices for τ_k, we obtain a family of bounds which hold simultaneously for all $\mathcal{F}_{j,k}$, and can thus safely apply an infimum, do some algebra, and obtain the desired bound.

Turning back to the proof of Lemma 1.1, consider taking each function class $\mathcal{G}_j$ and composing it with $1-\phi_\gamma$, giving $\mathcal{Q}_j := \left\{(x,y) \mapsto 1 - \phi_\gamma(g(x),y) : g \in \mathcal{G}_j\right\}$, and similarly composing f with $1-\phi_\gamma$. All of these post-composition functions are bounded, and Lemma A.1 can be readily applied, giving roughly Lemma 1.1, albeit with a left hand side in terms of $1-\phi_\gamma$. To lower bound the left hand side with the misclassification probability, we use the fact that $2(1-\phi_\gamma)$ upper bounds the misclassification loss; this is given by Lemma A.1, which can also be used to massage the right hand side into the final form in Lemma 1.1. Full details are given in the appendices; moreover, as detailed in Appendix E.3, this general approach can replace $1-\phi_\gamma$ with any bounded and well-behaved loss, for instance a standard margin loss.

3.2 Proof of Theorem 1.2

The proof follows by plugging the following Rademacher complexity estimate for canonical computation graphs into Lemma 1.1.

Lemma 3.2. *Let data $X \in \mathbb{R}^{n\times d}$ be given. Let computation graph $\mathcal{G}$ be given, where Π_i projects to Frobenius-norm balls of radius b_i, and $\|W_i^\intercal\|_{2,1} \le r_i$, and $\|[W_i\Pi_i D_i \rangle F_i]\|_2 \le s_i$, and $\mathrm{Lip}(\sigma_i) \le 1$*

for convenience, and all layers have width at most m. Then for every $\epsilon > 0$ there exists a covering set $\mathcal{M}$ satisfying

$$\sup_{g\in\mathcal{G}} \min_{\hat{X}\in\mathcal{M}} \left\| g(X^\intercal) - \hat{X} \right\| \le \epsilon \qquad \text{and} \qquad \ln|\mathcal{M}| \le \frac{2^{4/3}\ln(2m^2)}{\epsilon^2} \left[\sum_i \left(r_i b_i \prod_{l=i+1}^{L} s_l \right)^{2/3} \right]^3.$$

Consequently,

$$\mathrm{Rad}(\mathcal{G}) \le \frac{4}{n} + \frac{12\sqrt{\ln(2m^2)}}{n} \left[\sum_i \left(r_i b_i \prod_{l=i+1}^{L} s_l \right)^{2/3} \right]^{3/2}.$$

As discussed before, this proof follows the scheme from (Bartlett et al., 2017a), though simplified in various ways, owing mainly to the omission of general matrix norm penalties on weight matrices, and the omission of the *reference matrices*. The reference matrices were a technique to center the weight norm balls away from the origin; a logical place to center them was at initialization. However, in this distillation setting, it is in fact most natural to center everything at the origin, and apply regularization and shrink to a well-behaved function (rather than shrinking back to the random initialization, which after all defines a complicated function).

Also provided is a short re-proof of the core covering number bound for $(2,1)$-norm bounded matrices (cf. Lemma B.3). Other than that, the generality of computation graphs was in fact conveniently handled in the original proof scheme.

3.3 Proof sketch of Lemma 1.3

As was mentioned before, the first step of the proof is to sparsify the network. The approach followed here is to sparsify each matrix product. Concretely, given weights W_i of layer i, and letting $X_{i-1}^\intercal$ denote the input to this layer, then

$$W_i X_{i-1}^\intercal = \sum_{j=1}^{m} (W_i \mathbf{e}_j)(X_{i-1}\mathbf{e}_j)^\intercal.$$

Written this way, it seems natural that the matrix product should "concentrate", and that considering all m outer products should not be necessary. Indeed, exactly such an approach has been followed before to analyze randomized matrix multiplication schemes (Sarlos, 2006; Hsu et al., 2012). As there is no goal of high probability here, the analysis is simpler, and follows from the Maurey lemma (cf. Lemma B.1), as is used in the $(2,1)$-norm matrix covering lemma in Lemma B.3.

Lemma 3.3. *Let a network be given with ρ_i-Lipschitz homogeneous activations σ_i and weight matrices $(W_1,\dots,W_L)$ of maximum width m, along with data matrix $X \in \mathbb{R}^{n\times d}$ and desired widths $(k_1,\dots,k_L)$ be given. Then there exists a sparsified network output, recursively defined via*

$$\hat{X}_0^\intercal := X^\intercal, \quad \text{and} \quad \hat{X}_i^\intercal := \Pi_i \sigma_i(W_i M_i X_{i-1}^\intercal), \quad \text{where} \quad M_i := \sum_{j\in S_i} \frac{Z_j \mathbf{e}_j \mathbf{e}_j^\intercal}{\|A\mathbf{e}_j\|},$$

where S_i is a multiset of $k_i = |S_i|$ indices, Π_i denotes projection onto the Frobenius-norm ball of radius $\|X\|_{\mathrm{F}} \prod_{j\le i} \|W_j\|_2$, and the scaling term Z_j satisfies $Z_j \le \|W_k\|_{\mathrm{F}}\sqrt{m/k_j}$, and

$$\|\sigma_L(W_L \cdots \sigma_1(W_1 X^\intercal)\cdots) - \hat{X}_L^\intercal\|_{\mathrm{F}} \le \|X\|_{\mathrm{F}} \left[\prod_{i=1}^{L} \|W_i\|_2 \right] \sum_{i=1}^{L} \sqrt{\frac{\mathrm{sr}(W_i)}{k_i}},$$

where $\mathrm{sr}(W_i) := \|W_i\|_{\mathrm{F}}^2 / \|W_i\|_2^2$ *denotes the* stable rank *of W_i.*

As a brief note, it would be wonderful if a generalization bound with this error estimate could be proved, but this is only a sparsification bound.

The statement of this lemma is lengthy and detailed because the exact guts of the construction are needed in the subsequent generalization proof. Specifically, now that there are few nodes, a generalization bound sensitive to narrow networks can be applied. On the surface, it seems reasonable to

apply a VC dimension bound, but this approach did not seem to yield a rate better than $n^{-1/6}$, and also had an explicit dependence on the depth of the network, times other terms visible in Lemma 1.3.

Instead, the approach here, aiming for a better dependence on n and also no explicit dependence on network depth, was to produce an ∞-norm covering number bound (see (Long & Sedghi, 2019) for a related approach), with some minor adjustments (indeed, the ∞-norm parameter covering approach was applied to obtain a Frobenius-norm bound, as in Lemma 3.2). Unfortunately, the magnitudes of weight matrix entries must be controlled for this to work (unlike the VC approach), and this necessitated the detailed form of Lemma 3.3 above.

To close with a few pointers to the literature, as Lemma 3.3 is essentially a pruning bound, it is potentially of independent interest. Secondly, there is already one generalization bound in the literature which exhibits spectral norms, due to (Suzuki et al., 2019); unfortunately, it also has an explicit dependence on network width.

REFERENCES

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

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. 2018. `arXiv:1802.05296 [cs.LG]`.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NIPS*, pp. 6240–6249, 2017a.

Peter L. Bartlett, Nick Harvey, Chris Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. 2017b. `arXiv:1703.02930 [cs.LG]`.

Cristian Buciluǔ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *KDD*, pp. 535–541, 2006.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. 2017. `arXiv:1703.11008 [cs.LG]`.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. 2019. `arXiv:1803.03635 [cs.LG]`.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Pruning neural networks at initialization: Why are we missing the mark? 2020. `arXiv:2009.08576 [cs.LG]`.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *COLT*, 2018.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2015. `arXiv:1503.02531 [stat.ML]`.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Analysis of a randomized approximation scheme for matrix multiplication. 2012. `arXiv:1211.5414 [cs.DS]`.

Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. In *ICLR*, 2019a. `arXiv:1810.00113 [stat.ML]`.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. 2019b. `arXiv:1912.02178 [cs.LG]`.

Philip M. Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks. 2019. `arXiv:1905.12600 [cs.LG]`.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. 2019. arXiv:1902.04742 [cs.LG].

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

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pp. 143–152, 11 2006.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. 2010. arXiv:1009.3896 [cs.LG].

Jingtong Su, Yihang Chen, Tianle Cai, Tianhao Wu, Ruiqi Gao, Liwei Wang, and Jason D. Lee. Sanity-checking pruning methods: Random tickets can win the jackpot. 2020. arXiv:2009.11094 [cs.LG].

Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. 2019. arXiv:1909.11274 [cs.LG].

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 1 edition, 2019.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. 2019. arXiv:1905.03684 [cs.LG].

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. 2016. arXiv:1611.03530 [cs.LG].

A PROOFS OF MAIN DISTILLATION BOUNDS, LEMMAS 1.1 AND 3.1

The proof of the core abstract lemma, Lemma 3.1, is now provided in full.

Proof of Lemma 3.1. For each $j \geq 1$ and $k \in \{1, \ldots, \lg(n)\}$, define

$$\tau_k := 2^{2-k} \qquad \text{and} \qquad \delta_{j,k} := \frac{p_j \delta}{\lg(n)},$$

whereby $\sum_{j,k} \delta_{j,k} = \lg(n) \sum_j p_j \delta / \lg(n) = \delta$. Let $\mathcal{F}$ denote all measurable functions from $\mathbb{R}^d$ to $[0,1]$, and carve it up into strips corresponding to failure probabilities $\delta_{j,k}$: for each (j,k), define

$$\mathcal{F}_{j,k} := \left\{ f \in \mathcal{F} : \inf_{g \in \mathcal{G}_j} \frac{1}{n} \sum_i |f(z_i) - g(z_i)| \leq \tau_k \right\}.$$

By these definitions,

$$\begin{aligned} n\text{Rad}(\mathcal{F}_{j,k}) &= \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}_{j,k}} \sum_i \epsilon_i f(z_i) = \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}_{j,k}} \inf_{g \in \mathcal{G}_j} \sum_i \epsilon_i \left(f(z_i) - g(z_i) + g(z_i) \right) \\ &\leq \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}_{j,k}} \inf_{g \in \mathcal{G}_j} \max_i |\epsilon_i| \sum_i |f(z_i) - g(z_i)| + \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}_j} \sum_i \epsilon_i g(z_i) \\ &\leq n\tau_k + n\text{Rad}(\mathcal{G}_j). \end{aligned}$$

By standard Rademacher complexity bounds (Shalev-Shwartz & Ben-David, 2014), with probability at least $1 - \delta_{j,k}$, for every $f \in \mathcal{F}_{j,k}$,

$$\begin{aligned} \mathbb{E} f(z) &\leq \frac{1}{n} \sum_i f(z_i) + 2\text{Rad}(\mathcal{F}_{j,k}) + 3\sqrt{\frac{\ln(2/\delta_{j,k})}{2n}} \\ &\leq \frac{1}{n} \sum_i f(z_i) + 2\tau_k + 2\text{Rad}(\mathcal{G}_j) + 3\sqrt{\frac{\ln(2/\delta) + \ln(1/p_j) + \ln\lg(n)}{2n}}. \end{aligned} \tag{A.1}$$

To prove the desired statement, discard the union of all preceding failure events, meaning that with probability at least $1 - \delta = 1 - \sum_{j,k} \delta_{j,k}$, eq. (A.1) holds simultaneously for all $f \in \mathcal{F}$ and (j,k). Fixing f, for any j and $g \in \mathcal{G}_j$, then $\sum_i |f(z_i) - g(z_i)|/n \leq 2$, thus there exists k so that $f \in \mathcal{F}_{j,k}$, and we may safely choose k to be the smallest such choice. This smallest k must satisfy

$$\tau_k \geq \frac{1}{n} \inf_{g \in \mathcal{G}_j} \sum_i |f(z_i) - g(z_i)| \geq \frac{1}{2}\tau_k - 2^{1-\lg(n)},$$

where the final slop comes from handling the case $k = \lg(n)$, and thus

$$\tau_k \leq \frac{4}{n} + \frac{2}{n} \inf_{g \in \mathcal{G}_j} \sum_i |f(z_i) - g(z_i)|.$$

Consequently, applying an infimum over (j,k) for all instantiations of eq. (A.1),

$$\begin{aligned} \mathbb{E} f(z) &\leq \frac{1}{n} \sum_i f(z_i) + \inf_{j,k} \left[2\tau_k + 2\text{Rad}(\mathcal{G}_j) + 3\sqrt{\frac{\ln(j/\delta)}{n}} \right] \\ &\leq \frac{1}{n} \sum_i f(z_i) + \frac{8}{n} \\ &\quad + \inf_j \left[\inf_{g \in \mathcal{G}_j} \frac{4}{n} \sum_i |f(z_i) - g(z_i)| + 2\text{Rad}(\mathcal{G}_j) + 3\sqrt{\frac{\ln(2/\delta) + \ln(1/p_j) + \ln\lg(n)}{n}} \right], \end{aligned}$$

and the final bound in the statement comes via the general inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for arbitrary nonnegative (a,b). □

To establish Lemma 1.1, convenient properties of ϕ_γ are first handled.

Lemma A.1. *For any $v \in \mathbb{R}^k$ and $y \in \{1, \ldots, k\}$, letting $\sigma_{\rm s}(z) := (1 + \exp(-z))^{-1}$ denote the sigmoid,*

$$2(1 - \phi_\gamma(v))_y \geq 2\sigma_{\rm s}(-v_y/\gamma + \max_{i \neq y} u_i/\gamma) \geq \mathbb{1}[y \neq \arg\max_i v_i].$$

Proof. First note that the sigmoid $\sigma_{\rm s}(z) := (1 + \exp(-z))^{-1}$ satisfies $2\sigma_{\rm s}(z) \geq \mathbb{1}[z \geq 0]$, since $\sigma_{\rm s}$ is monotone nondecreasing and $\sigma_{\rm s}(0) = 1/2$. Therefore, setting $u := v/\gamma$ for convenience, and recalling that

$$\ln \sum_{i \neq y} \exp u_i \geq \ln \max_{i \neq y} \exp u_i = \max_{i \neq y} u_i,$$

then

$$\begin{aligned}
2(1 - \phi_\gamma(v))_y &= 2(1 - \phi(u))_y \\
&= \frac{2\sum_{i \neq y} \exp(u_i)}{\sum_i \exp(u_i)} = \frac{2}{1 + \frac{\exp(u_y)}{\sum_{i \neq y} \exp(u_i)}} = \frac{2}{1 + \exp(u_y - \ln \sum_{i \neq y} \exp(u_i))} \\
&= 2\sigma_{\rm s}\left(-u_y + \ln \sum_{i \neq y} \exp(u_i))\right) \geq 2\sigma_{\rm s}\left(-u_y + \max_{i \neq y} u_i\right) \\
&\geq \mathbb{1}\left[u_y \leq \max_{i \neq y} u_i\right] \geq \mathbb{1}\left[y \neq \arg\max_i u_i\right].
\end{aligned}$$

□

The proof of Lemma 1.1 is a combination of Lemma A.1 and Lemma 3.1.

Proof of Lemma 1.1. Define $\psi := 1 - \phi_\gamma$ and $\mathcal{Q}_j := \{\psi \circ g : g \in \mathcal{G}_j\}$. Since ψ has range $[0, 1]$, it follows by Lemma 3.1 with probability at least $1 - \delta$ that for any measurable $f : \mathbb{R}^d \to \mathbb{R}^k$,

$$\begin{aligned}
\mathbb{E}\psi(f(x), y) \leq \frac{1}{n}\sum_i \psi(f(x_i), y_i) + \frac{8}{n} + 3\sqrt{\frac{\ln(2/\delta) + \ln\lg(n)}{n}} \\
+ \inf_{\substack{j \geq 1 \\ g \in \mathcal{G}_j}} \left[\frac{4}{n}\sum_i |\psi(f(x_i), y_i) - \psi(g(x_i), y_i)| + 2\mathrm{Rad}(\mathcal{Q}_j) + 3\sqrt{\frac{\ln(1/p_j)}{n}}\right].
\end{aligned}$$

To simplify this expression, note firstly that

$$|\psi(f(x_i), y_i) - \psi(g(x_i), y_i)| = |\phi_\gamma(f(x_i), y_i) - \phi_\gamma(g(x_i), y_i)|$$

for every i. Secondly, letting $\sigma_{\rm s}$ denote the sigmoid (which is $1/4$-Lipschitz), by the properties of ϕ_γ in Lemma A.1 and the Lipschitz peeling property of Rademacher complexity (Shalev-Shwartz & Ben-David, 2014),

$$\begin{aligned}
\mathrm{Rad}(\mathcal{Q}_j) &= \mathrm{Rad}\left(\left\{(x, y) \mapsto 1 - \phi_\gamma(g(x), y) : g \in \mathcal{G}_j\right\}\right) \\
&= \mathrm{Rad}\left(\left\{(x, y) \mapsto \phi_\gamma(g(x), y) : g \in \mathcal{G}_j\right\}\right) \\
&= \mathrm{Rad}\left(\left\{(x, y) \mapsto \sigma_{\rm s}(-g(x)_y/\gamma + \max_{i \neq y} g(x)_i/\gamma) : g \in \mathcal{G}_j\right\}\right) \\
&\leq \frac{1}{4\gamma}\mathrm{Rad}\left(\left\{(x, y) \mapsto \sigma_{\rm s}(g(x)_y - \max_{i \neq y} g(x)_i : g \in \mathcal{G}_j\right\}\right).
\end{aligned}$$

For the final simplification, note again by Lemma A.1 that

$$\Pr[f(x) \neq y] \leq \mathbb{E}2(1 - \phi_\gamma(f(x), y)),$$

thus combining all simplifications and multiplying everything by 2 gives the final bound. □

B SAMPLING TOOLS

The proofs of Lemma 3.2 and Lemma 3.3 both make heavy use of sampling.

Lemma B.1 (Maurey (Pisier, 1980))**.** *Suppose random variable V is almost surely supported on a subset S of some Hilbert space, and let $(V_1, \ldots, V_k)$ be k IID copies of V. Then there exist $(\hat{V}_1, \ldots, \hat{V}_k) \in S^k$ with*

$$\left\| \mathbb{E}V - \frac{1}{k}\sum_i \hat{V}_i \right\|_{\mathrm{F}}^2 \leq \mathop{\mathbb{E}}_{V_1,\ldots,V_k} \left\| \mathbb{E}V - \frac{1}{k}\sum_i V_i \right\|_{\mathrm{F}}^2 = \frac{1}{k}\left[\mathbb{E}\|V\|_{\mathrm{F}}^2 - \|\mathbb{E}V\|_{\mathrm{F}}^2\right] \leq \frac{1}{k}\mathbb{E}\|V\|_{\mathrm{F}}^2 \leq \frac{1}{k}\sup_{\hat{V}\in S}\|\hat{V}\|_{\mathrm{F}}^2.$$

Proof of Lemma B.1. The first inequality is via the probabilistic method. For the remaining inequalities, by expanding the square multiple times,

$$\begin{aligned}\mathop{\mathbb{E}}_{V_1,\ldots,V_k} \left\| \mathbb{E}V - \frac{1}{k}\sum_i V_i \right\|_{\mathrm{F}}^2 &\leq \mathop{\mathbb{E}}_{V_1,\ldots,V_k} \frac{1}{k^2}\left[\sum_i \|\mathbb{E}V - V_i\|_{\mathrm{F}}^2 + \sum_{i\neq j}\left\langle \mathbb{E}V - V_i, \mathbb{E}V - V_j\right\rangle\right] \\ &= \frac{1}{k}\mathbb{E}_{V_1}\|V_1 - \mathbb{E}V\|_{\mathrm{F}}^2 = \frac{1}{k}\left[\mathbb{E}\|V\|_{\mathrm{F}}^2 - \|\mathbb{E}V\|_{\mathrm{F}}^2\right] \leq \frac{1}{k}\mathbb{E}\|V\|_{\mathrm{F}}^2 \leq \frac{1}{k}\sup_{\hat{V}\in S}\|\hat{V}\|_{\mathrm{F}}^2.\end{aligned}$$

□

A first key application of Lemma B.1 is to sparsify products, as used in Lemma 3.3.

Lemma B.2. *Let matrices $A \in \mathbb{R}^{d\times m}$ and $B \in \mathbb{R}^{n\times m}$ be given, along with sampling budget k. Then there exists a selection $(i_1, \ldots, i_k)$ of indices and a corresponding diagonal* sampling matrix *M with at most k nonzero entries satisfying*

$$M := \frac{\|A\|_{\mathrm{F}}^2}{k}\sum_{j=1}^{k}\frac{\mathbf{e}_{i_j}\mathbf{e}_{i_j}^{\mathsf{T}}}{\|A\mathbf{e}_{i_j}\|^2} \qquad \textit{and} \qquad \left\|AB^{\mathsf{T}} - AMB^{\mathsf{T}}\right\|^2 \leq \frac{1}{k}\|A\|^2\|B\|^2.$$

Proof of Lemma B.2. For convenience, define columns $a_i := A\mathbf{e}_i$ and $b_i := B\mathbf{e}_i$ for $i \in \{1, \ldots, m\}$. Define *importance weighting* $\beta_i := (\|a_i\|/\|A\|_{\mathrm{F}})^2$, whereby $\sum_i \beta_i = 1$, and let V be a random variable with

$$\Pr\left[V = \beta_i^{-1}a_i b_i^{\mathsf{T}}\right] = \beta_i,$$

whereby

$$\begin{aligned}\mathbb{E}V &= \sum_{i=1}^{m}\beta_i^{-1}a_i b_i^{\mathsf{T}}\beta_i = \sum_{i=1}^{m}(A\mathbf{e}_i)(B\mathbf{e}_i)^{\mathsf{T}} = A\left[\sum_{i=1}^{m}\mathbf{e}_i\mathbf{e}_i^{\mathsf{T}}\right]B^{\mathsf{T}} = A\left[I\right]B^{\mathsf{T}} = AB, \\ \mathbb{E}\|V\|^2 &= \sum_{i=1}^{m}\beta_i^{-2}\|a_i b_i^{\mathsf{T}}\|_{\mathrm{F}}\beta_i = \sum_{i=1}^{m}\beta_i^{-1}\|a_i\|^2\|b_i\|^2 = \sum_{i=1}^{m}\|A\|_{\mathrm{F}}^2\|b_i\|^2 = \|A\|_{\mathrm{F}}^2\cdot\|B\|_{\mathrm{F}}^2.\end{aligned}$$

By Lemma B.1, there exist indices $(i_1, \ldots, i_k)$ and matrices $\hat{V}_j := \beta_{i_j}^{-1}a_{i_j}b_{i_j}^{\mathsf{T}}$ with

$$\left\|AB^{\mathsf{T}} - \frac{1}{k}\sum_j \hat{V}_j\right\|^2 \leq \left\|\mathbb{E}V - \frac{1}{k}\sum_j \hat{V}_j\right\|^2 = \frac{1}{k}\left[\|A\|_{\mathrm{F}}^2\|B\|_{\mathrm{F}}^2 - \|AB\|_{\mathrm{F}}^2\right] \leq \frac{1}{k}\|A\|_{\mathrm{F}}^2\|B\|_{\mathrm{F}}^2.$$

To finish, by the definition of M,

$$\frac{1}{k}\sum_j \hat{V}_j = \frac{1}{k}\sum_j \beta_{i_j}^{-1}(A\mathbf{e}_{i_j})(B\mathbf{e}_{i_j})^{\mathsf{T}} = A\left[\frac{1}{k}\sum_j \beta_{i_j}^{-1}\mathbf{e}_{i_j}\mathbf{e}_{i_j}^{\mathsf{T}}\right]B^{\mathsf{T}} = A\left[M\right]B^{\mathsf{T}}.$$

□

A second is to cover the set of matrices W satisfying a norm bound $\|W^\intercal\|_{2,1} \le r$. The proof here is more succinct and explicit than the one in (Bartlett et al., 2017a, Lemma 3.2).

Lemma B.3 (See also (Bartlett et al., 2017a, Lemma 3.2))**.** *Let norm bound $r \ge 0$, $X \in \mathbb{R}^{n\times d}$, and integer k be given. Define a family of matrices*

$$\mathcal{M} := \left\{ \frac{r\|X\|_{\mathrm{F}}}{k} \sum_{l=1}^{k} \frac{s_l \mathbf{e}_{i_l}\mathbf{e}_{j_l}^\intercal}{\|X\mathbf{e}_{j_l}\|} : s_l \in \{\pm 1\}, i_l \in \{1,\dots,n\}, j_l \in \{1,\dots,d\} \right\}.$$

Then

$$|\mathcal{M}| \le (2nd)^k, \qquad \sup_{\|W^\intercal\|_{2,1}\le r} \min_{\hat{W}\in\mathcal{M}} \|WX^\intercal - \hat{W}X^\intercal\|_{\mathrm{F}}^2 \le \frac{r^2\|X\|_{\mathrm{F}}^2}{k}.$$

Proof. Let $W \in \mathbb{R}^{m\times d}$ be given with $\|W^\intercal\|_{2,1} \le r$. Define $s_{ij} := W_{ij}/|W_{ij}|$, and note

$$WX^\intercal = \sum_{i,j} \mathbf{e}_i\mathbf{e}_i^\intercal W \mathbf{e}_j\mathbf{e}_j^\intercal X^\intercal = \sum_{i,j} \mathbf{e}_i W_{ij}(X\mathbf{e}_j)^\intercal = \sum_{i,j} \underbrace{\frac{|W_{ij}|\|X\mathbf{e}_j\|_2}{r\|X\|_{\mathrm{F}}}}_{=:q_{ij}} \underbrace{\frac{r\|X\|_{\mathrm{F}} s_{ij}\mathbf{e}_i(X\mathbf{e}_j)^\intercal}{\|X\mathbf{e}_j\|}}_{=:U_{ij}}.$$

Note by Cauchy-Schwarz that

$$\sum_{i,j} q_{ij} \le \frac{1}{r\|X\|_{\mathrm{F}}} \sum_i \sqrt{\sum_j W_{ij}^2}\|X\|_{\mathrm{F}} = \frac{\|W^\intercal\|_{2,1}\|X\|_{\mathrm{F}}}{r\|X\|_{\mathrm{F}}} \le 1,$$

potentially with strict inequality, thus q is not a probability vector. To remedy this, construct probability vector p from q by adding in, with equal weight, some U_{ij} and its negation, so that the above summation form of $WX^\intercal$ goes through equally with p and with q.

Now define IID random variables $(V_1,\dots,V_k)$, where

$$\begin{aligned}
\Pr[V_l = U_{ij}] &= p_{ij}, \\
\mathbb{E}V_l &= \sum_{i,j} p_{ij}U_{ij} = \sum_{i,j} q_{ij}U_{ij} = WX^\intercal, \\
\|U_{ij}\| &= \left\| \frac{s_{ij}\mathbf{e}_i(X\mathbf{e}_j)}{\|X\mathbf{e}_j\|_2} \right\|_{\mathrm{F}} \cdot r\|X\|_{\mathrm{F}} = |s_{ij}| \cdot \|\mathbf{e}_i\|_2 \cdot \left\| \frac{X\mathbf{e}_j}{\|X\mathbf{e}_j\|_2} \right\|_2 \cdot r\|X\|_{\mathrm{F}} = r\|X\|_{\mathrm{F}}, \\
\mathbb{E}\|V_l\|^2 &= \sum_{i,j} p_{ij}\|U_{ij}\|^2 \le \sum_{ij} p_{ij} r^2\|X\|_{\mathrm{F}}^2 = r^2\|X\|_{\mathrm{F}}^2.
\end{aligned}$$

By Lemma B.1, there exist $(\hat{V}_1,\dots,\hat{V}_k) \in S^k$ with

$$\left\| WX^\intercal - \frac{1}{k}\sum_l \hat{V}_l \right\|^2 \le \mathbb{E}\left\| \mathbb{E}V_1 - \frac{1}{k}\sum_l V_l \right\|^2 \le \frac{1}{k}\mathbb{E}\|V_1\|^2 \le \frac{r^2\|X\|_{\mathrm{F}}^2}{k}.$$

Furthermore, the matrices $\hat{V}_l$ have the form

$$\frac{1}{k}\sum_l \hat{V}_l = \frac{1}{k}\sum_l \frac{s_l\mathbf{e}_{i_l}(X\mathbf{e}_{j_l})^\intercal}{\|X\mathbf{e}_{j_l}\|} = \left[\frac{1}{k}\sum_l \frac{s_l\mathbf{e}_{i_l}\mathbf{e}_{j_l}^\intercal}{\|X\mathbf{e}_{j_l}\|} \right] X^\intercal =: \hat{W}X^\intercal,$$

where $\hat{W} \in \mathcal{M}$. Lastly, note $|\mathcal{M}|$ has cardinality at most $(2nd)^k$. □

C Proof of Theorem 1.2 and Lemma 3.2

The first step is the Rademacher complexity bound in Lemma 3.2, which follows the proof scheme laid out in (Bartlett et al., 2017a), with simplifications due to the removal of "reference matrices" and some norm generality.

Proof of Lemma 3.2. Let cover scale ϵ and per-layer scales $(\epsilon_1, \ldots, \epsilon_L)$ be given; the proof will develop a covering number parameterized by these per-layer scales, and then optimize them to derive the final covering number in terms of ϵ. From there, a Dudley integral will give the Rademacher bound.

As in the statement, recursively define

$$X_0^\intercal := X^\intercal, \qquad X_i^\intercal := \sigma_i\left([W_i \Pi_i D_i \natural F_i] X_{i-1}^\intercal\right).$$

The proof will recursively construct an analogous cover via

$$\hat{X}_0^\intercal := X^\intercal, \qquad \hat{X}_i^\intercal := \sigma_i\left([\hat{W}_i \Pi_i D_i \natural F_i] \hat{X}_{i-1}^\intercal\right),$$

where the choice of $\hat{W}_i$ depends on $\hat{X}_{i-1}$, and thus the total cover cardinality will product (and not simply sum) across layers. Specifically, the cover $\mathcal{N}_i$ for $\hat{W}_i$ is given by Lemma B.3 by plugging in $\|\Pi_i D_i \hat{X}_{i-1}^\intercal\|_{\mathrm{F}} \le b_i$, and thus it suffices to choose

$$\text{cover cardinality } k := \frac{r_i^2 b_i^2}{\epsilon_i^2}, \qquad \text{whereby } \min_{\hat{W}_i \in \mathcal{N}_i} \|W_i \Pi_i D_i \hat{X}_{i-1}^\intercal - \hat{W}_i \Pi_i D_i \hat{X}_{i-1}^\intercal\| \le \epsilon_i.$$

By this choice (and the cardinality estimate in Lemma B.3, the full cover $\mathcal{N}$ satisfies

$$\ln|\mathcal{N}| = \sum_i \ln|\mathcal{N}_i| \le \sum_i \frac{r_i^2 b_i^2}{\epsilon_i^2} \ln(2m^2).$$

To optimize the parameters $(\epsilon_1, \ldots, \epsilon_L)$, the first step is to show via induction that

$$\|X_i^\intercal - \hat{X}_i^\intercal\|_{\mathrm{F}} \le \sum_{j \le i} \epsilon_j \prod_{l=j+1}^{i} s_l.$$

The base case is simply $\|X_0^\intercal - \hat{X}^\intercal\| = \|X^\intercal - X^\intercal\| = 0$, thus consider layer $i > 0$. Using the inductive formula for $\hat{X}_i$ and the cover guarantee on $\hat{W}_i$,

$$\begin{aligned}
\left\|X_i^\intercal - \hat{X}_i^\intercal\right\| &= \left\|\sigma_i([W_i \Pi_i D_i \natural F_i] X_{i-1}^\intercal) - \sigma_i([\hat{W}_i \Pi_i D_i \natural F_i] \hat{X}_{i-1}^\intercal)\right\| \\
&\le \left\|[W_i \Pi_i D_i \natural F_i] h X_{i-1}^\intercal - [\hat{W}_i \Pi_i D_i \natural F_i] \hat{X}_{i-1}^\intercal\right\| \\
&\le \left\|[W_i \Pi_i D_i \natural F_i] X_{i-1}^\intercal - [W_i \Pi_i D_i \natural F_i] \hat{X}_{i-1}^\intercal\right\| + \left\|[W_i \Pi_i D_i \natural F_i] \hat{X}_{i-1}^\intercal - [\hat{W}_i \Pi_i D_i \natural F_i] \hat{X}_{i-1}^\intercal\right\| \\
&\le \left\|[W_i \Pi_i D_i \natural F_i]\right\|_2 \left\|X_{i-1}^\intercal - \hat{X}_{i-1}^\intercal\right\| + \left\|[(W_i - \hat{W}_i) \Pi_i D_i \hat{X}_{i-1}^\intercal \natural (F_i - F_i) \hat{X}_{i-1}^\intercal]\right\| \\
&\le s_i \sum_{j \le i-1} \epsilon_j \prod_{l=j+1}^{i-1} s_l + \left\|(W_i - \hat{W}_i) \Pi_i D_i \hat{X}_{i-1}^\intercal\right\| \\
&\le \sum_{j \le i-1} \epsilon_j \prod_{l=j+1}^{i} s_l + \epsilon_i \le \sum_{j \le i} \epsilon_j \prod_{l=j+1}^{i} s_l.
\end{aligned}$$

To balance $(\epsilon_1, \ldots, \epsilon_L)$, it suffices to minimize a Lagrangian corresponding to the cover size subject to an error constraint, meaning

$$L(\vec{\epsilon}, \lambda) = \sum_{i=1}^{L} \frac{\alpha_i}{\epsilon_i^2} + \lambda \left(\sum_{i=1}^{L} \epsilon_i \beta_i - \epsilon\right) \qquad \text{where } \alpha_i := r_i^2 b_i^2 \ln(2m^2),\ \beta_i := \prod_{l=i+1}^{L} s_l,$$

whose unique critical point for $\vec{\epsilon} > 0$ implies the choice

$$\epsilon_i := \frac{1}{Z}\left(\frac{2\alpha_i}{\beta_i}\right)^{1/3} \qquad \text{where } Z := \frac{1}{\epsilon} \sum_i (2\alpha_i \beta_i^2)^{1/3},$$

whereby $\|X_L^\intercal - \hat{X}_L^\intercal\| \leq \epsilon$ automatically, and

$$\begin{aligned}
\ln|\mathcal{N}| &\leq Z^2 \sum_i \frac{r_i^2 b_i^2 \ln(2m^2)}{(2\alpha_i/\beta_i)^{2/3}} \\
&= \frac{1}{\epsilon^2 2^{2/3}} \left[2\sum_i r_i^{2/3} b_i^{2/3} \beta_i^{2/3} \ln(2m^2)^{1/3} \right]^2 \sum_i r_i^{2/3} b_i^{2/3} \ln(2m^2)^{1/3} \beta_i^{2/3} \\
&= \frac{2^{4/3}\ln(2m^2)}{\epsilon^2} \left[\sum_i \left(r_i b_i \prod_{l=i+1}^{L} s_l \right)^{2/3} \right]^3 =: \frac{\tau^2}{\epsilon^2},
\end{aligned}$$

as desired, with τ introduced for convenience in what is to come.

For the Rademacher complexity estimate, by a standard Dudley entropy integral (Shalev-Shwartz & Ben-David, 2014), setting $\hat{\tau} := \max\{\tau, 1/3\}$ for convenience,

$$n\text{Rad}(\mathcal{G}) \leq \inf_\zeta 4\zeta\sqrt{n} + 12\int_\zeta^{\sqrt{n}} \frac{\sqrt{\hat{\tau}}\epsilon}{\mathrm{d}}\epsilon = \inf_\zeta 4\zeta\sqrt{n} + 12\hat{\tau}\,\ln(\epsilon)\Big|_\zeta^{\sqrt{n}} = \inf_\zeta 4\zeta\sqrt{n} + 12\hat{\tau}(\ln\sqrt{n} - \ln\zeta),$$

which is minimized at $\zeta = 3\hat{\tau}/\sqrt{n}$, whereby

$$n\text{Rad}(\mathcal{G}) \leq 12\hat{\tau} + 6\hat{\tau}\ln n - 12\hat{\tau}\ln(3\hat{\tau}/\sqrt{n}) = 12\hat{\tau}(1 - \ln(3\hat{\tau})) \leq 12\hat{\tau} \leq 12\tau + 4.$$

□

Combining Lemma 3.2 with Lemma 1.1 directly gives Theorem 1.2.

Proof of Theorem 1.2. For each parameter to the model, e.g., b_i, organize the corresponding parameter space into shells of radius $j \in \mathbb{Z}_{\geq 1}$, and associate a per-parameter prior weight $q(b_i)_j = (j(j+1))^{-1} = (b_i(b_i+1))^2$, whereby $\sum_j q(b_i)_j = 1$. Consequently, for each choice of parameters $(\vec{r}, \vec{b}, \vec{s}_i)$, the corresponding prior value products the priors for each parameter, and thus again sums to 1 (and let the overall prior $(p_j)_{j\geq 1}$ be any enumeration of these products). Moreover, for any prior value $p(\vec{r}, \vec{b}, \vec{s})$ corresponding to the preceding product for some $(\vec{r}, \vec{b}, \vec{s}_i)$, by Lemma 3.2,

$$\text{Rad}(\mathcal{G}(\vec{r}, \vec{b}, \vec{s}_i)) = \widetilde{\mathcal{O}}\left(\frac{1}{n} \sum_i \left[r_i b_i \prod_{l=i+1}^{L} s_l \right]^{2/3} \right)^{3/2},$$

and moreover note

$$\ln\frac{1}{p(\vec{r}, \vec{b}, \vec{s})} = \mathcal{O}\left(\ln\prod_i r_i^2 b_i^2 s_i^2 \right) = \mathcal{O}\left(\ln\prod_i r_i^{2/3} b_i^{2/3} s_i^{2/3} \right) = \widetilde{\mathcal{O}}\left(\frac{1}{\gamma^2 n} \sum_i \left[r_i b_i \prod_{l=i+1}^{L} s_l \right]^{2/3} \right).$$

Lastly, note that $\psi(v, y) := v_y - \max_{i\neq y} v_i$ is $\sqrt{2}$-Lipschitz, and so can be safely appended to $\mathcal{G}(\vec{r}, \vec{b}, \vec{s}_i))$ and then peeled when calculating Rademacher complexity (Shalev-Shwartz & Ben-David, 2014). Combining these pieces, by Lemma 1.1, with probability at least $1 - \delta$, every mea-

surable $f : \mathbb{R}^d \to \mathbb{R}^k$ satisfies

$$\begin{aligned}
\Pr[f(x) \neq y] &\leq \frac{2}{n}\sum_i \big(1 - \phi_\gamma(f(x_i), y_i)\big) + \frac{16}{n} + 6\sqrt{\frac{\ln(2/\delta) + \ln\lg(n)}{n}} \\
&\quad + \inf_{(\vec{b},\vec{r},\vec{s})\in\mathbb{Z}^{3L}_{\geq 1}} \inf_{g\in\mathcal{G}(\vec{b},\vec{r},\vec{s})} \Bigg[\frac{8}{n}\sum_i \big|\phi_\gamma(f(x_i), y_i) - \phi_\gamma(g(x_i), y_i)\big| + \\
&\qquad \frac{1}{\gamma}\mathrm{Rad}\left(\Big\{\psi \circ g : g \in \mathcal{G}(\vec{b},\vec{r},\vec{s})\Big\}\right) + 6\sqrt{\frac{\ln(1/p(\vec{r},\vec{b},\vec{s})}{n}}\Bigg] \\
&\leq \frac{2}{n}\sum_i \big(1 - \phi_\gamma(f(x_i), y_i)\big) + 6\sqrt{\frac{\ln(2/\delta)}{n}} \\
&\quad + \inf_{(\vec{b},\vec{r},\vec{s})\in\mathbb{Z}^{3L}_{\geq 1}} \inf_{g\in\mathcal{G}(\vec{b},\vec{r},\vec{s})} \left[\frac{8}{n}\sum_i \big|\phi_\gamma(f(x_i), y_i) - \phi_\gamma(g(x_i), y_i)\big| + \widetilde{\mathcal{O}}\left(\sum_i \left[r_i b_i \prod_{l=i+1}^{L} s_l\right]^{2/3}\right)\right].
\end{aligned}$$

□

D PROOF OF STABLE RANK BOUND, LEMMA 1.3

The first step is to establish the sparsification lemma in Lemma 3.3, which in turn sparsifies each matrix product, can not simply invoke Lemma B.2: pre-processing is necessary to control the element-wise magnitudes of the resulting matrix.

Lemma D.1. *Let matrices $A \in \mathbb{R}^{d\times m}$ and $B \in \mathbb{R}^{n\times m}$ be given, along with sampling budget k. Then there exists a selection $(i_1, \ldots, i_k)$ of indices and a corresponding diagonal* sampling matrix *M with at most k nonzero entries satisfying*

$$M := \sum_{j=1}^{k} \frac{Z_{i_j}\mathbf{e}_{i_j}\mathbf{e}_{i_j}^\intercal}{\|a_{i_j}\|} \quad \textit{where} \quad Z_{i_j} \leq \|A\|_{\mathrm{F}}\sqrt{\frac{m}{k}}, \qquad \textit{and} \qquad \left\|AB^\intercal - AMB^\intercal\right\|^2 \leq \frac{4}{k}\|A\|^2\|B\|^2.$$

Proof. Let $\tau > 0$ be a parameter to be optimized later, and define a subset of indices $S := \{i \in \{1, \ldots, m\} : \|A\mathbf{e}_i\| \geq \tau\}$, with $S^c := \{1, \ldots, m\} \setminus S$. Let A_τ denote the matrix obtained by zeroing out columns not in S, meaning

$$A_\tau := \sum_{i\in S}(A\mathbf{e}_i)\mathbf{e}_i^\intercal,$$

whereby

$$\|AB^\intercal - A_\tau B^\intercal\|_{\mathrm{F}} \leq \|A - A_\tau\| \cdot \|B\| \leq \|B\|\sqrt{\sum_{i\in S^c}\|A\mathbf{e}_i\|^2} \leq \tau\sqrt{m}\|B\|.$$

Applying Lemma B.2 to $A_\tau B^\intercal$ gives

$$M := \frac{\|A_\tau\|^2}{k}\sum_{j=1}^{k}\frac{\mathbf{e}_{i_j}\mathbf{e}_{i_j}^\intercal}{\|A_\tau\mathbf{e}_{i_j}\|^2} = \sum_{j=1}^{k}\frac{Z_{i_j}\mathbf{e}_{i_j}\mathbf{e}_{i_j}^\intercal}{\|A_\tau\mathbf{e}_{i_j}\|} \qquad \text{such that} \qquad \|A_\tau B^\intercal - A_\tau MB^\intercal\|^2 \leq \frac{1}{k}\|A_\tau\|^2\|B\|^2,$$

where Z_{i_j} is specified by these equalities. To simplify, note $\|A_\tau\| \leq \|A\|$, and $A_\tau M = AM$. Combining the two inequalities,

$$\|AB^\intercal - AMB^\intercal\| \leq \|AB^\intercal - A_\tau B^\intercal\| + \|A_\tau B^\intercal - A_\tau MB^\intercal\| \leq \tau\sqrt{m}\|B\| + \frac{1}{\sqrt{k}}\|A\|\|B\|.$$

To finish, setting $\tau := \|A\|/\sqrt{mk}$ gives the bound, and ensures that the scaling term Z_{i_j} satisfies, for any $i_j \in S$,

$$Z_{i_j} = \frac{\|A_\tau\|^2}{k\|A_\tau\mathbf{e}_{i_j}\|} \leq \frac{\|A\|_{\mathrm{F}}^2}{k\tau} = \|A\|_{\mathrm{F}}\sqrt{\frac{m}{k}}.$$

□

With this tool in hand, the proof of Lemma 3.3 is as follows.

Proof of Lemma 3.3. Let X_j denote the network output after layer j, meaning

$$X_0^\intercal := X^\intercal, \qquad X_j^\intercal := \sigma_j(W_j X_{j-1}^\intercal),$$

whereby

$$\|X_j^\intercal\|_{\mathrm{F}} = \|\sigma_j(W_j X_{j-1}^\intercal) - \sigma_j(0)\|_{\mathrm{F}} \le \|W_j X_{j-1}^\intercal\|_{\mathrm{F}} \le \|W_j\|_2 \|X_{j-1}^\intercal\|_{\mathrm{F}} \le \|X\|_{\mathrm{F}} \prod_{i\le j} \|W_i\|_2.$$

The proof will inductively choose sampling matrices $(M_1, \ldots, M_L)$ as in the statement and construct

$$\hat{X}_0^\intercal := X^\intercal, \qquad \hat{X}_j^\intercal := \Pi_j \sigma_j(W_j M_j \hat{X}_{j-1}^\intercal),$$

where Π_j denotes projection onto the Frobenius-norm ball of radius $\|X\|_{\mathrm{F}} \prod_{i\le j} \|W_i\|_2$ (whereby $\Pi_j X_j = X_j$), satisfying

$$\left\|X_j - \hat{X}_j\right\|_{\mathrm{F}} \le \|X\|_{\mathrm{F}} \left[\prod_{p=1}^{j} \|W_p\|_2\right] \sum_{i=1}^{j} \sqrt{\frac{\mathrm{sr}(W_i)}{k_i}},$$

which gives the desired bound after plugging in $j = L$.

Proceeding with the inductive construction, the base case is direct since $\hat{X}_0 = X = X_0$ and $\left\|X_0 - \hat{X}_0\right\|_{\mathrm{F}} = 0$, thus consider some $j > 0$. Applying Lemma D.1 to the matrix multiplication $W_j \hat{X}_{j-1}$ with k_j samples, there exists a multiset of S_j coordinates and a corresponding sampling matrix M_j, as specified in the statement, satisfying

$$\left\|W_j \hat{X}_{j-1}^\intercal - W_j M_j \hat{X}_{j-1}^\intercal\right\|_{\mathrm{F}} \le \frac{1}{\sqrt{k_j}} \|W_j\|_{\mathrm{F}} \|\hat{X}_{j-1}\|_{\mathrm{F}} \le \frac{1}{\sqrt{k_j}} \|W_j\|_{\mathrm{F}} \|X\|_{\mathrm{F}} \prod_{i<j} \|W_i\|_2.$$

Using the choice $\hat{X}_j^\intercal := \Pi_j \sigma_j(W_j M_j \hat{X}_{j-1}^\intercal)$,

$$\begin{aligned}
\left\|X_j - \hat{X}_j\right\|_{\mathrm{F}} &= \left\|\sigma_j(W_j X_{j-1}^\intercal) - \Pi_j \sigma_j(W_j M_j \hat{X}_{j-1}^\intercal)\right\|_{\mathrm{F}} \\
&\le \left\|W_j X_{j-1}^\intercal - W_j M_j \hat{X}_{j-1}^\intercal\right\|_{\mathrm{F}} \\
&= \left\|W_j X_{j-1}^\intercal - W_j \hat{X}_{j-1}^\intercal + W_j \hat{X}_{j-1}^\intercal - W_j M_j \hat{X}_{j-1}^\intercal\right\|_{\mathrm{F}} \\
&\le \left\|W_j X_{j-1}^\intercal - W_j \hat{X}_{j-1}^\intercal\right\|_{\mathrm{F}} + \left\|W_j \hat{X}_{j-1}^\intercal - W_j M_j \hat{X}_{j-1}^\intercal\right\|_{\mathrm{F}} \\
&\le \|W_j\|_2 \left\|X_{j-1} - \hat{X}_{j-1}\right\|_{\mathrm{F}} + \frac{1}{\sqrt{k_j}} \|W_j\|_{\mathrm{F}} \|X\|_{\mathrm{F}} \prod_{i<j} \|W_i\|_2 \\
&\le \|W_j\|_2 \left(\|X\|_{\mathrm{F}} \left[\prod_{i<j} \|W_i\|_2\right] \sum_{i<j} \sqrt{\frac{\mathrm{sr}(W_i)}{k_i}} \right) + \sqrt{\frac{\mathrm{sr}(W_j)}{k_j}} \|X\|_{\mathrm{F}} \prod_{i\le j} \|W_i\|_2 \\
&\le \|X\|_{\mathrm{F}} \left[\prod_{i\le j} \|W_i\|_2\right] \sum_{i\le j} \sqrt{\frac{\mathrm{sr}(W_i)}{k_i}}
\end{aligned}$$

as desired. □

To prove Lemma 1.3 via Lemma 3.3, the first step is a quick tool to cover matrices element-wise.

Lemma D.2. *Let $\mathcal{A}$ denote matrices with at most k_2 nonzero rows and k_1 nonzero columns, entries bounded in absolute value by b, and total number of rows and columns each at most m. Then there exists a cover set $\mathcal{M} \subseteq \mathcal{A}$ satisfying*

$$|\mathcal{M}| \le m^{k_1+k_2} \left(\frac{2b\sqrt{k_1 k_2}}{\epsilon}\right)^{k_1 k_2}, \qquad \text{and} \qquad \sup_{A\in\mathcal{A}} \min_{\hat{A}\in\mathcal{M}} \|A - \hat{A}\|_{\mathrm{F}} \le \epsilon.$$

Proof. Consider some fixed set of k_2 nonzero rows and k_1 nonzero columns, and let $\mathcal{M}_0$ denote the covering set obtained by gridding the $k_1 \cdot k_2$ entries at scale $\frac{\epsilon}{\sqrt{k_1 k_2}}$, whereby

$$|\mathcal{M}_0| \leq \left(\frac{2b\sqrt{k_1 k_2}}{\epsilon}\right)^{k_1 k_2}.$$

For any $A \in \mathcal{A}$ with these specific nonzero rows and columns, the $\hat{A} \in \mathcal{M}_0$ obtained by rounding each nonzero entry of A to the nearest grid element gives

$$\|A - \hat{A}\|^2 = \sum_{i,j}(A_{ij} - \hat{A}_{ij})^2 \leq \sum_{i,j}\left(\frac{\epsilon}{\sqrt{k_1 k_2}}\right)^2 = \epsilon^2 \sum_{i,j}\frac{1}{k_1 k_2} = \epsilon^2.$$

The final cover $\mathcal{M}$ is now obtained by unioning copies of $\mathcal{M}_0$ for all $\binom{m}{k_1}\binom{m}{k_2} \leq m^{k_1+k_2}$ possible submatrices of size $k_2 \times k_1$. □

The proof of Lemma 1.3 now carefully combines the preceding pieces.

Proof of Lemma 1.3. The proof proceeds in three steps, as follows.

1. A covering number is estimate for sparsified networks, as output by Lemma 3.3.
2. A covering number for general networks is computed by balancing the error terms from Lemma 3.3 and its cover computed here.
3. This covering number is plugged into a Dudley integral to obtain the desired Rademacher bound.

Proceeding with this plan, let $(\hat{X}_0^\intercal, \ldots, \hat{X}_L^\intercal)$ be the layer outputs (and network input) exactly as provided by Lemma 3.3. Additionally, define diagonal matrices $D_j := \sum_{l \in S_{j+1}}^{!} \mathbf{e}_l \mathbf{e}_l^\intercal$ (with $D_L = I$, where the "!" denotes unique inclusion; these matrices capture the effect of the subsequent sparsification, and can be safely inserted after each W_j without affecting $\hat{X}_j^\intercal$, meaning

$$\hat{X}_j^\intercal = \Pi_j \sigma_j(W_j M_j \hat{X}_{j-1}^\intercal) = \Pi_j \sigma_j(D_j W_j M_j \hat{X}_{j-1}^\intercal).$$

Let per-layer cover precisions $(\epsilon_1, \ldots, \epsilon_L)$ be given, which will be optimized away later. This proof will inductively construct

$$\tilde{X}_0^\intercal := X^\intercal, \qquad \tilde{X}_j^\intercal := \Pi_j \sigma_j(\tilde{W}_j \tilde{X}_{j-1}^\intercal),$$

where $\tilde{W}_j$ is a cover element for $D_j W_j M_j$, and inductively satisfying

$$\|\hat{X}_j^\intercal - \tilde{X}_j^\intercal\| \leq \|X\|_{\mathrm{F}} m^{j/2} \sum_{i \leq j} \epsilon_i \prod_{\substack{l \leq j \\ l \neq i}} \|W_j\|_{\mathrm{F}}.$$

To construct the per-layer cover elements $\tilde{W}_j$, first note by the form of M_j (and the scaling Z_i provided by Lemma 3.3) that

$$b := \max_{i,l}(D_j W_j M_j)_{l,i} \leq \max_i \|W_j M_j \mathbf{e}_i\| \leq Z_i \frac{\|W_j \mathbf{e}_i\|}{\|W_j \mathbf{e}_i\|} \leq \|W_j\|_{\mathrm{F}} \sqrt{\frac{m}{k_{j-1}}}.$$

Consequently, by Lemma D.2, there exists a cover $\mathcal{C}_j$ of matrices of the form $D_j W_j M_j$ satisfying

$$|\mathcal{C}_j| \leq m^{k_j + k_{j-1}} \left(\frac{2b\sqrt{k_j k_{j-1}}}{\epsilon_j}\right)^{k_j k_{j-1}} \leq m^{k_j + k_{j-1}} \left(\frac{2\|W_j\|_{\mathrm{F}}\sqrt{k_j m}}{\epsilon_j}\right)^{k_j k_{j-1}},$$

and the closest cover element $\tilde{W}_j \mathcal{C}_j$ to $D_j W_j M_j$ satisfies $\|D_j W_j M_j - \tilde{W}_j\|_{\mathrm{F}} \leq \epsilon_j$.

Proceeding with the induction, the base case has $\|\hat{X}_0^\intercal - \tilde{X}_0^\intercal\| = \|X^\intercal - X^\intercal\| = 0$, thus consider $j > 0$. The first step is to estimate the spectral norm of $D_j W_j M_j$, which can be coarsely upper bounded via

$$\|D_j W_j M_j\|_2^2 \leq \|D_j W_j M_j\|_{\mathrm{F}}^2 \leq \sum_i \|W_j M_j \mathbf{e}_i\|^2 \leq \sum_i \|W_j\|_{\mathrm{F}}^2 \frac{m}{k_{j-1}} \leq \|W_j\|_{\mathrm{F}}^2 m.$$

By the form of $\hat{X}_j$ and $\tilde{X}_j$,

$$\begin{aligned}
\|\hat{X}_j^\intercal - \tilde{X}_j^\intercal\| &= \|\Pi_j \sigma_j(D_j W_j M_j \hat{X}_{j-1}^\intercal) - \Pi_j \sigma_j(\tilde{W}_j \tilde{X}_{j-1}^\intercal)\| \\
&\leq \|D_j W_j M_j \hat{X}_{j-1}^\intercal - \tilde{W}_j \tilde{X}_{j-1}^\intercal\| \\
&\leq \|D_j W_j M_j \hat{X}_{j-1}^\intercal - D_j W_j M_j \tilde{X}_{j-1}^\intercal\| + \|D_j W_j M_j \tilde{X}_{j-1}^\intercal - \tilde{W}_j \tilde{X}_{j-1}^\intercal\| \\
&\leq \|D_j W_j M_j\|_2 \|\hat{X}_{j-1}^\intercal - \tilde{X}_{j-1}^\intercal\|_{\mathrm{F}} + \|D_j W_j M_j - \tilde{W}_j\|_2 \|\tilde{X}_{j-1}^\intercal\|_{\mathrm{F}} \\
&\leq \sqrt{m} \|W_j\|_{\mathrm{F}} \left[\|X\|_{\mathrm{F}} m^{(j-1)/2} \sum_{i<j} \epsilon_i \prod_{\substack{l<j \\ l \neq i}} \|W_j\|_{\mathrm{F}} \right] + \epsilon_j \|X\|_{\mathrm{F}} \prod_{i<j} \|W_j\|_2 \\
&\leq \|X\|_{\mathrm{F}} m^{j/2} \sum_{i \leq j} \epsilon_i \prod_{\substack{l \leq j \\ l \neq i}} \|W_j\|_{\mathrm{F}},
\end{aligned}$$

which establishes the desired bound on the error.

The next step is to optimize k_j. Let $\epsilon > 0$ be arbitrary, and set $\epsilon_j^{-1} := \epsilon^{-1} 2L\sqrt{m} \|X\|_{\mathrm{F}} \prod_{i \neq j} \|W_i\|_{\mathrm{F}}$, whereby

$$\|\hat{X}_L^\intercal - \tilde{X}_L^\intercal\|_{\mathrm{F}} \leq \frac{\epsilon}{2}, \qquad |\mathcal{C}_j| \leq m^{k_j + k_{j-1}} \left(\frac{4mL\sqrt{k_j} \|X\|_{\mathrm{F}} \prod_i \|W_i\|_{\mathrm{F}}}{\epsilon} \right)^{k_j k_{j-1}}.$$

The overall network cover $\mathcal{N}$ is the product of the covers for all layers, and thus has cardinality satisfying

$$\begin{aligned}
\ln |\mathcal{N}| &\leq \sum_j \ln |\mathcal{C}_j| \leq 2 \sum_j k_j \ln m + \sum_j k_j k_{j-1} \ln \left(\frac{4mL\sqrt{k_j} \|X\|_{\mathrm{F}} \prod_i \|W_i\|_{\mathrm{F}}}{\epsilon_j} \right) \\
&\leq 2 \sum_j k_j \ln m + \sum_j 2k_j^2 \ln \left(\frac{4mL\sqrt{k_j} \|X\|_{\mathrm{F}}}{\epsilon_j} \right) + \left[\sum_j 2k_j^2 \right] \cdot \left[\sum_j \ln \|W_j\|_{\mathrm{F}} \right].
\end{aligned}$$

To choose $(k_1, \ldots, k_L)$, letting $X_L^\intercal$ denote the output of the original unsparsified network, note firstly that the full error bound satisfies

$$\begin{aligned}
\|X_L^\intercal - \tilde{X}_L^\intercal\| &\leq \|X_L^\intercal - \hat{X}_L^\intercal\| + \|\hat{X}_L^\intercal - \tilde{X}_L^\intercal\| \\
&\leq \sum_i \frac{\alpha_i}{\sqrt{k_i}} + \frac{\epsilon}{2}, \qquad \text{where } \alpha_i := \|X\|_{\mathrm{F}} \left[\prod_i \|W_i\|_2 \right] \sqrt{\mathrm{sr}(W_i)}.
\end{aligned}$$

To choose k_i, the approach here is to minimize a Lagrangian corresponding to the cover cardinality, subject to the total cover error being ϵ. Simplifying the previous expressions and noting $2k_j k_{j-1} \leq k_j^2 + k_{j-1}^2$, whereby the dominant term in $\ln |\mathcal{N}|$ is $\sum_j k_j^2$, consider Lagrangian

$$L(k_1, \ldots, k_l, \lambda) := \sum_i k_i^2 + \lambda \left(\sum_i \frac{\alpha_i}{\sqrt{k_i}} - \frac{\epsilon}{2} \right),$$

which has critical points when each k_i satisfies

$$\frac{k_i^{5/2}}{\alpha_i} = \frac{\lambda}{4},$$

thus $k_i := \alpha_i^{2/5}/Z$ with $Z := \epsilon^2/(2\sum_j \alpha_j^{4/5})^2$. As a sanity check (since it was baked into the Lagrangian), plugging this into the cover error indeed gives

$$\|X_L^\intercal - \tilde{X}_L^\intercal\| \leq \sum_i \frac{\alpha_i}{\sqrt{k_i}} + \frac{\epsilon}{2} = \sqrt{Z}\sum_i \alpha_i^{4/5} + \frac{\epsilon}{2} = \epsilon.$$

To upper bound the cover cardinality, first note that

$$\sum_i k_i^2 = \frac{1}{Z^2}\sum_i \alpha_i^{4/5} = \frac{4}{\epsilon^4}\left(\sum_i \alpha_i^{4/5}\right)^5,$$

whereby

$$\begin{aligned}\ln|\mathcal{N}| &= \widetilde{\mathcal{O}}\left(\left[\sum_i k_i^2\right]\cdot\left[\sum_i \ln\|W_i\|_{\mathrm{F}}\right]\right)\\ &= \frac{\beta}{\epsilon^4} \qquad\qquad \text{where } \beta = \widetilde{\mathcal{O}}\left(\|X\|_{\mathrm{F}}^4\left[\prod_j \|W_j\|_2^4\right]\left[\sum_i \mathrm{sr}(W_i)^{2/5}\right]^5\left[\sum_i \ln\|W_i\|_{\mathrm{F}}\right]\right).\end{aligned}$$

The final step is to apply a Dudley entropy integral (Shalev-Shwartz & Ben-David, 2014), which gives

$$n\mathrm{Rad}(\mathcal{H}) = \inf_\zeta\left(4\zeta\sqrt{n} + 12\int_\zeta^{\sqrt{n}}\frac{\sqrt{\beta}}{\epsilon^2}\,\mathrm{d}\epsilon\right) = \inf_\zeta\left(4\zeta\sqrt{n} + 12\left[\frac{1}{\zeta} - \frac{1}{\sqrt{n}}\right]\sqrt{\beta}\right).$$

Dropping the negative term gives an expression of the form $a\zeta + b/\zeta$, which is convex in $\zeta > 0$ and has critical point at $\zeta^2 = b/a$, which after plugging back in gives an upper bound $2\sqrt{ab}$, meaning

$$n\mathrm{Rad}(\mathcal{H}) \leq 2\left(4\sqrt{n}\cdot 12\sqrt{\beta}\right)^{1/2} = 8\sqrt{3}n^{1/4}\beta^{1/4}.$$

Dividing by n and expanding the definition of β gives the final Rademacher complexity bound. □

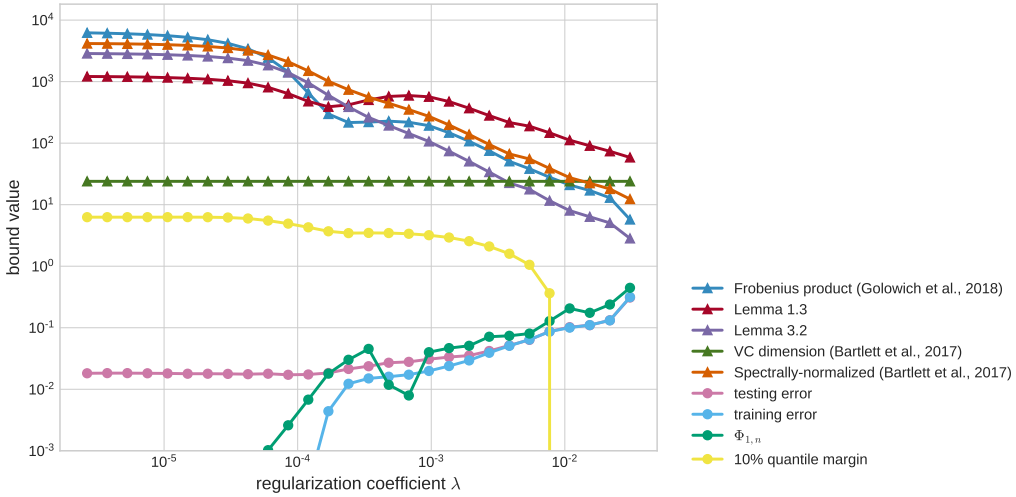


Figure 4: **Comparison of bounds.** See below for a descripion.

E MATERIAL FOR REVIEWERS DURING FEEDBACK PHASE

E.1 COMPARISON OF BOUNDS

In response primarily to comments by AnonReviewer1 and AnonReviewer4, but relevant to all reviewers, we have included Figure 4, which we plan to incorporate in the first nine pages of the paper. Here is a detailed description:

- The figures follows a *single* run of our distillation method as detailed at the start of Section 2. That is, a sequence of distilled functions is produced by gradually increasing the regularization parameter λ. The horizontal axis corresponds to this regularization strength λ, and the vertical axis to various bounds and other values.
- The architecture is fully-connected with three ReLU layers of width 2048, and the data is `mnist`. It is trained to perfect classification accuracy before distillation. Unlike most plots in the paper, there are no skip connections, which allows us to invoke prior bounds from the literature.
- The curves with circle markers denote the training error, testing error, distillation error, and 10% margin quantile of the distilled function g. Note that for regularization strength 10^{-3}, the test error is still small, but the bounds have dropped by an order of magnitude.
- The curves labeled by triangles are bounds.
  - The VC bound stays constant, since the architecture does not change (Bartlett et al., 2017b).
  - The curve for Lemma 3.2 is close to the curve for (Bartlett et al., 2017a), which makes sense since their proofs are similar. It also makes sense that both curves decrease monotonically, since the regularization scheme uses $\|\cdot\|_{2,1}$.
  - The curve for Lemma 1.3 starts the best, but gradually all other curves dominate it; this makes some sense, since the proof itself involves a compression-like step.
  - The Frobenius product bound from (Golowich et al., 2018) looks promising in the middle phase, but becomes worse. It is possible it would remain the best if the regularization scheme replaced the $(2,1)$ norm with a Frobenius norm. We focused on the $(2,1)$ norm, however, since it is used by our Lemma 3.2, which handles skip connections and convolution layers.

The most similar plot in the literature is Figure 4 in (Arora et al., 2018). The main points of contrast are that the curves here include dimension factors for each bound which were dropped there, and of course that the bounds here also bound the original network.

E.2 A SUFFICIENT CONDITION FOR DISTILLATION

This section is in response to a comment by AnonReviewer4, and aims to provide a sufficient condition for distillation. Namely, the sufficient condition is that the weight matrices can be written $W_i + E_i$, where W_i is deterministic and E_i is random; the claim is that we can "distill away" E_i.

This construction is meant more as an illustration than an explanation; it is meant to capture our intuition that distillation can work when there are many useless, canceling weights. Randomness is one way to model such weights.

If the reviewers are satisfied with this material, we will incorporate it in Section 1.2

Proposition E.1. *Let $\delta > 0$ be given, and suppose the data matrix $X \in \mathbb{R}^{n\times d}$ is normalized so that $\|X\|_{\mathrm{F}} \le \sqrt{n}$, and the labels $y_i \in \{\pm 1\}$ are binary for simplicity. Consider a ReLU network $f(x) = A_L\sigma(A_{L-1}\cdots\sigma(A_1x)\cdots)$ where the ith weight matrix can be written $A_i := W_i + E_i$, where W_i satisfies $\|W_i\|_2 \ge 1$, and E_i is random with iid $\mathcal{N}(0, 1/m)$ coordinates and independent of $(W_j)_{j<i}$, and the* width *parameter m satisfies*

$$m \ge \max\left\{ \max_i 4L^2\|W_i\|_{\mathrm{F}}^2 \ , \ \left(3L\sqrt{2\ln(1/\delta)}\prod_{k=1}^{L}\|W_l\|_2\right)^4\right\}.$$

Then, with probability at least $1 - 3L\delta$ over the draw of $(E_L, \ldots, E_1)$, the following statements hold.

1. ***(Original network has bad complexity.)*** *It holds that $\|E_1\|_{\mathrm{F}} = \Theta(\sqrt{d})$, and $\|E_L\|_{\mathrm{F}} = \Theta(1)$, and $\|A_i\|_{\mathrm{F}} = \Theta(\|E_i\|_{\mathrm{F}}) = \Theta(\sqrt{m})$ for $i \in \{2, \ldots, L-1\}$, and the Rademacher complexity upper bound bound from (Golowich et al., 2018, Theorem 1) for ReLU networks with the same weight matrix norms as f is*

$$\frac{\sqrt{L}\prod_{i=1}^{L}\|A_i\|_{\mathrm{F}}}{\sqrt{n}} = \Theta\left(\sqrt{\frac{Ldm^{L-2}}{n}}\right).$$

2. ***(Distilled network has low complexity.)*** *Define the* distilled network *as*

$$g(x) = W_L\sigma(W_{L-1}\cdots\sigma(W_1x)\cdots).$$

The Rademacher complexity bound from (Golowich et al., 2018, Theorem 1) for ReLU networks with the same weight matrix norms as g is

$$\Theta\left(\frac{\sqrt{L}\prod_{i=1}^{L}\|W_i\|_{\mathrm{F}}}{\sqrt{n}}\right).$$

3. ***(Distilled and original networks are close.)*** $\Phi_{1,n}(f, g) \le \frac{1}{2m^{1/4}}$.

The parameter m corresponds to the width of the original network, and is intended to be a free parameter that can be made arbitrarily large. The message of Proposition E.1 is that we can make f have arbitrarily bad complexity, while remaining close to g which has low complexity. From another perspective, Proposition E.1 captures one setting where distillation can vastly reduce the bound from (Golowich et al., 2018, Theorem 1), as depicted in Figure 4.

Proof of Proposition E.1. 1. The first step is to bound $\mathbb{E}\|E_i\|$ for each matrix. By Jensen's inequality,

$$\mathbb{E}\|E_i\| \le \sqrt{\mathbb{E}\|E_i\|^2} = \sqrt{\frac{mm_i}{m}} = \sqrt{m_i},$$

where $m_1 = d$, $m_L = 1$, and $m_i = m$ otherwise. For the lower bound, by the general Taylor lower bound

$$\sqrt{x} \ge \sqrt{r} + \frac{x-r}{2\sqrt{r}} - \frac{(x-r)^2}{2\sqrt{r^3}} \qquad \forall x, r \ge 0,$$

setting $x = \|E_i\|^2$ and $r := \mathbb{E}\|E_i\|^2 = m_i$ and noting $\mathbb{E}\|E_i\|^4 = \mathbb{E}(\sum_{j,k}(E_i)_{j,k}^2)^2$ gives

$$\mathbb{E}(x-r)^2 = \sum_{j,k}\mathbb{E}((E_i)_{j,k}^2 - \mathbb{E}(E_i)_{j,k}^2)^2 = \sum_{j,k}\left(\mathbb{E}(E_i)_{j,k}^4 - (\mathbb{E}(E_i)_{j,k}^2)^2\right) = \sum_{j,k}\frac{3-1}{m^2} = \frac{2m_i}{m},$$

whereby

$$\mathbb{E}\sqrt{\|E_i\|^2} \geq \sqrt{r} + 0 - \frac{\mathbb{E}(x-r)^2}{\sqrt{4r^3}} \geq \sqrt{m_i} - \frac{3m_i}{m\sqrt{4m_i^3}} \geq \sqrt{m_i}\left(1 - \frac{2}{mm_i}\right).$$

Moreover, since the mapping $A \mapsto \|A\|_{\mathrm{F}}$ is 1-Lipschitz, then standard Gaussian concentration (Wainwright, 2019, Theorem 2.26) implies that with probability at least $1 - 2\delta$,

$$\big|\|E_i\| - \mathbb{E}\|E_i\|\big| \leq \sqrt{\frac{2\ln(1/\delta)}{m}}.$$

Combining this with the estimates on $\mathbb{E}\|A_i\|$ for $i \in \{2, \ldots, L\}$, note by the assumptions on m and the preceding derivations that

$$\big|\|A_i\| - \|E_i\|\big| \leq \|W_i\| \leq \Theta\Big(\frac{\|E_i\|}{L}\Big),$$

and in particular

$$\|A_i\| = \Theta\left((1-1/L)\sqrt{m}\right) \quad \text{and} \quad \prod_{i=1}^{L}\|A_i\| = \Theta\left((1-1/L)^{L-2}\sqrt{m^{L-2}d}\right) = \Theta\left(\sqrt{m^{L-2}d}\right).$$

and the lower bounds on m gives the desired bounds on $\|E_i\|$ (after unioning away the combined $2L\delta$ failure probability mass). The complexity estimate is now direct from (Golowich et al., 2018, Theorem 1).

2. This part is direct from the definition of g and (Golowich et al., 2018, Theorem 1).

3. For convenience throughout the proof, define

$$s_i := \|W_i\|_2 \qquad \text{and} \qquad \epsilon := \sqrt{\frac{2\ln(1/\delta)}{m}}.$$

For convenience define a fake identity activation $\sigma_0(z) = z$, which is still 1-homogeneous and 1-Lipschitz, and recursively define

$$Z_0 := X^\intercal, \quad Z_i := A_i\sigma_{i-1}(Z_{i-1}), \qquad U_0 := X^\intercal, \quad U_i := W_i\sigma_{i-1}(U_{i-1}).$$

The first part of the proof will show, inductively, with probability at least $1 - i\delta$, that

$$\|Z_i - U_i\|_{\mathrm{F}} \leq \epsilon\|X\|_{\mathrm{F}}(1+\epsilon)^i\sum_{j=1}^{i}\frac{\prod_{k=1}^{i}s_k}{s_j},$$

which in turn will imply

$$\|Z_L - U_L\|_{\mathrm{F}} \leq \frac{\sqrt{n}}{m^{1/4}}.$$

We'll spend the bulk of the proof establishing this claim; before that, let's see how to complete the proof from there. By (Bartlett et al., 2017a, Lemma A.3), the margin mapping $(\hat{y}, y) \mapsto \hat{y}_y - \max_{i\neq y}\hat{y}_i$ is 2-Lipschitz, and a similar calculation to Lemma A.1 implies that ϕ_1 is $(1/4)$-Lipschitz, therefore

$$\begin{aligned}\Phi_{1,n}(f,g) &= \frac{1}{n}\sum_{i=1}^{n}|\phi_1(f(x_i),y_i) - \phi_1(g(x_i),y_i)| \\ &\leq \frac{1}{2n}\sum_{i=1}^{n}\|f(x_i) - g(x_i)\|_2 \leq \frac{1}{2\sqrt{n}}\|U_L - Z_L\|_{\mathrm{F}} \leq \frac{1}{2m^{1/4}}.\end{aligned}$$

Returning to the induction, the base case is immediate since $Z_0 = X^\intercal = U_0$. For the inductive step, first note that the mapping

$$E \mapsto \|E\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}}$$

is $\|Z_{i-1}\|_{\mathrm{F}}$-Lipschitz, since, for any matrices A, B,

$$\|A\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}} \leq \|(A-B)\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}} + \|B\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}}$$

which can be repeated with A and B swapped, and implies

$$\left|\|A\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}} - \|B\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}}\right| \leq \|(A-B)\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}} \leq \|(A-B)\|_{\mathrm{F}} \cdot \|\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}}.$$

As such, with probability $1-\delta$, by the fact that $\sigma_{i-1}(Z_{i-1})$ and E_i are independent and standard Gaussian concentration (Wainwright, 2019, Theorem 2.26),

$$\|E_i\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}} \leq \epsilon\|\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}},$$

where whereby the 1-homogeneity and 1-Lipschitz properties of σ_{i-1} allow further simplification via

$$\|\sigma_{i-1}(Z_{i-1})\|_{\mathrm{F}} = \|\sigma_{i-1}(Z_{i-1}) - \sigma_{i-1}(0)\|_{\mathrm{F}} \leq \|Z_{i-1}\|_{\mathrm{F}} \leq \|U_{i-1}\|_{\mathrm{F}} + \|U_{i-1} - Z_{i-1}\|_{\mathrm{F}},$$

and similarly

$$\|U_{i-1}\|_{\mathrm{F}} \leq \|X\|_{\mathrm{F}} \prod_{j<i} s_j.$$

Thanks to this concentration inequality, and unioning it with the $(i-1)\delta$ failure probability of the inductive hypothesis, then with probability $1 - i\delta$,

$$\begin{aligned}
\|Z_i - U_i\| &= \|A_i\sigma_{i-1}(Z_{i-1}) - W_i\sigma_{i-1}(U_{i-1})\| \\
&\leq \|A_i\sigma_{i-1}(Z_{i-1}) - W_i\sigma_{i-1}(Z_{i-1})\| + \|W_i\sigma_{i-1}(Z_{i-1}) - W_i\sigma_{i-1}(U_{i-1})\| \\
&\leq \|E_i\sigma_{i-1}(Z_{i-1})\| + \|W_i\|_2 \cdot \|\sigma_{i-1}(Z_{i-1}) - \sigma_{i-1}(U_{i-1})\|_{\mathrm{F}} \\
&\leq \epsilon\|X\|_{\mathrm{F}} \prod_{j<i} s_j + (s_i + \epsilon)\, \|\sigma_{i-1}(Z_{i-1}) - \sigma_{i-1}(U_{i-1})\|_{\mathrm{F}} \\
&\leq \epsilon\|X\|_{\mathrm{F}} \prod_{j<i} s_j + s_i\,(1+\epsilon) \left[\epsilon\|X\|_{\mathrm{F}}(1+\epsilon)^{i-1} \sum_{j=1}^{i-1} \frac{\prod_{k=1}^{i-1} s_k}{s_j}\right] \\
&= \epsilon\|X\|_{\mathrm{F}}\epsilon \prod_{j<i} s_j + \epsilon \left[\|X\|_{\mathrm{F}}(1+\epsilon)^{i} \sum_{j=1}^{i-1} \frac{\prod_{k=1}^{i} s_k}{s_j}\right] \\
&\leq \epsilon\|X\|_{\mathrm{F}}(1+\epsilon)^{i} \sum_{j=1}^{i} \frac{\prod_{k=1}^{i} s_k}{s_j},
\end{aligned}$$

which completes the induction. For the final simplified bound, By the conditions on m, then $(1+\epsilon)^L \leq (1+1/L)^L \leq 3$, and additionally by further using the conditions on m, and since $\min_i \|W_i\| \geq 1$ and $\|X\|_{\mathrm{F}} \leq \sqrt{n}$,

$$\epsilon\|X\|_{\mathrm{F}}\epsilon(1+\epsilon)^L \sum_{j=1}^{i} \frac{\prod_{k=1}^{L} s_k}{s_j} \leq 3L\epsilon\|X\|_{\mathrm{F}} \prod_{k=1}^{L} s_k = \frac{3L\sqrt{2\ln(1/\delta)}\|X\|_{\mathrm{F}} \prod_{k=1}^{L} s_k}{\sqrt{m}} \leq \frac{\sqrt{n}}{m^{1/4}},$$

as desired.

□

E.3 Margin version of Lemma 1.1

To clarify a point by AnonReviewer3 and to provide a useful point of comparison to Lemma 1.1, we may replace the distillation loss with a margin loss and follow nearly exactly the same proof to give a margin-based distillation bound.

To develop this version, define the standard ramp (margin) loss as

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

In the binary case, the corresponding loss is $\ell_\gamma(yf(x))$; for the multiclass case, overload notation so that

$$\ell_\gamma(f(x), y) := \ell_\gamma\left(f(x)_y - \max_i f(x)_i\right).$$

(This multiclass version appears in (Bartlett et al., 2017a), and still upper bounds misclassification error.) The corresponding distillation bound is as follows.

Proposition E.2. *Let temperature (margin) parameter $\gamma > 0$ be given, along with any sets of multiclass predictors $(\mathcal{G}_j)_{j\geq 1}$ with $\mathcal{G}_j \ni g : \mathbb{R}^d \to \mathbb{R}^k$, along with* prior weights $(p_j)_{j\geq 1}$ *on these sets, where $p_j \geq 0$ and $\sum_j p_j = 1$. Then with probability at least $1-\delta$ over an iid draw of data $((x_i, y_i))_{i=1}^n$, every measurable function $f : \mathbb{R}^d \to \mathbb{R}^k$ satisfies*

$$\Pr[f(x) \neq y] \leq \frac{1}{n}\sum_i \ell_\gamma\left(f(x_i), y_i\right) + \frac{8}{n} + 3\sqrt{\frac{\ln(2/\delta) + \ln\lg(n)}{n}}$$
$$+ \inf_{\substack{j\geq 1 \\ g\in\mathcal{G}_j}} \left[\frac{4}{n}\sum_{i=1}^n \left|\ell_\gamma(f(x_i), y_i) - \ell_\gamma(g(x_i), y_i)\right| + \right.$$
$$\left. \frac{2}{\gamma}\mathrm{Rad}\left(\left\{(x,y) \mapsto g(x)_y - \max_{i\neq y} g(x)_i : g \in \mathcal{G}_j\right\}\right) + 3\sqrt{\frac{\ln(1/p_j)}{n}} \right].$$

Proof sketch of Proposition E.2. Since ℓ_γ is bounded, the proof goes through basically identically to the proof of Lemma 1.1 via Lemma 3.1. One key difference is that whereas

$$\Pr[f(x) \neq y] \leq \mathbb{E}2(1 - \phi_\gamma(f(x), y)),$$

we now have

$$\Pr[f(x) \neq y] \leq \mathbb{E}\ell_\gamma(f(x), y),$$

which removes a factor of 2 from most terms in right hand side of the bound as compared with Lemma 1.1.

(Once the reviewer is satisfied with the statement, the proof sketch will be expanded into a full proof.) □