How Uniform Random Weights Induce Non-uniform Bias: Typical Interpolating Neural Networks Generalize with Narrow Teachers

Gon Buzaglo [*1] **Itamar Harel** [*1] **Mor Shpigel Nacson** [*1] **Alon Brutzkus** [1] **Nathan Srebro** [2] **Daniel Soudry** [1]

Abstract

Background. A main theoretical puzzle is why over-parameterized Neural Networks (NNs) generalize well when trained to zero loss (i.e., so they interpolate the data). Usually, the NN is trained with Stochastic Gradient Descent (SGD) or one of its variants. However, recent empirical work examined the generalization of a random NN that interpolates the data: the NN was sampled from a seemingly uniform prior over the parameters, conditioned on that the NN perfectly classifies the training set. Interestingly, such a NN sample typically generalized as well as SGD-trained NNs.
Contributions. We prove that such a random NN interpolator typically generalizes well if there exists an underlying narrow "teacher NN" that agrees with the labels. Specifically, we show that such a 'flat' prior over the NN parameterization induces a rich prior over the NN functions, due to the redundancy in the NN structure. In particular, this creates a bias towards simpler functions, which require less relevant parameters to represent — enabling learning with a sample complexity approximately proportional to the complexity of the teacher (roughly, the number of non-redundant parameters), rather than the student's.

1. Introduction

A central theoretical question in deep learning is why Neural Networks (NNs) generalize, despite being over-parameterized, and even when perfectly fitted to the data (Zhang et al., 2017). One of the leading explanations for this phenomenon is that NNs have an "implicit bias" toward generalizing solutions (e.g., Gunasekar et al. (2017); Soudry et al. (2018); Arora et al. (2019); Lyu & Li (2020); Chizat & Bach (2020); Vardi (2023)). This bias stems from underlying interactions between the model and the training method — including the type of optimization step, the initialization, the parameterization, and the loss function.

Previous works (Valle-Perez et al., 2019; Mingard et al., 2021; Chiang et al., 2023) suggested, based on empirical evidence, that a significant part of this implicit bias in NNs is the mapping from the model parameters to the model function. Specifically, suppose we randomly sample the NN parameters from a 'uniform' prior[1], and accept only parameter samples in which the NN perfectly classifies all the training data — i.e., samples from the posterior composed of the same prior and the likelihood of a 0-1 loss function. Then, Chiang et al. (2023) found that the sampled NNs generalize as well as SGD in small-scale experiments.

These results may suggest that such a uniform sampling of the NN parameters induces simple and generalizing NN functions. In this paper, we prove this is indeed the case, and aim to uncover the mechanism behind this phenomenon. In short, we prove typical NN interpolators sampled this way ("students") generalize well, given there exists a "narrow" NN teacher that generates the labels. Next, we explain these results in more detail.

Contributions. In Section 3 we prove that a typical NN sampled from the posterior over interpolators generalizes well (i.e., has a small test error with high probability) with

$$\#\text{samples} = O\left(-\log \tilde{p}\right) , \tag{1}$$

where $\tilde{p}$ is the probability that a random NN (sampled from the 'uniform' prior) is equivalent to the teacher function.[2] Thus, to obtain generalization guarantees for NNs, we proceed to upper bound $(-\log \tilde{p})$.

[*]Equal contribution [1]Technion Institute of Technology, Haifa, Israel [2]Toyota Technological Institute at Chicago, Chicago IL, USA. Correspondence to: Gon Buzaglo <gon.buzaglo@gmail.com>, Itamar Harel <itamarharel01@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]A truly uniform prior does not exist for infinite sets, so the prior is chosen similarly to standard 'uniform-like' initializations: in each layer, the prior is Gaussian (uniform on the ℓ_2 sphere) or uniform in the ℓ_∞ ball.

[2]The proof idea is simple: the number of hypotheses sampled until a successful interpolation is $|\mathcal{H}| \lessapprox 1/\tilde{p}$. Plugging this into the standard sample complexity of a finite hypotheses class $O\left(\log |\mathcal{H}|\right)$, we obtain the result. The actual proof is slightly more complicated since $|\mathcal{H}|$ here weakly depends on the training set.

Next, in Section 4, we examine the case where both the student and teacher parameters are quantized to Q levels, including zero (as in standard numerical formats), and we assume the prior is uniform over all possible quantized values. We examine several architectures:

- For a fully connected multi-layer network with a scalar output, hidden neuron layer widths $\{d_l\}_{l=1}^{L}$ and $\{d_l^\star\}_{l=1}^{L}$ respectively for the student and teacher, input width $d_0 = d_0^\star$, and any activation function σ that satisfies $\sigma(0) = 0$, we prove

$$-\log \tilde{p} \leq \sum_{l=1}^{L} \left(d_l^\star d_{l-1}^\star + 2d_l\right) \log Q. \tag{2}$$

- For convolutional NNs, we obtain analogous results, where channel numbers replace layer widths, with an additional multiplicative factor of the kernel size.

- The proofs in both cases are simple[3], and can be extended for more general architectures.

Lastly, in Section 5 we examine a two-layer neural network with continuous weights and derive similar results, except a margin assumption replaces the quantization assumption, and a margin factor replaces Q in the bound.

Implications. Combining these relatively easy-to-prove results ((1) and (2)), we get a surprisingly novel result: typical NN interpolators have sample complexity approximately proportional to the number of teacher parameters times the number of quantization bits, with only a weak dependence on the student width. Thus, the student generalizes well *if there exists* a teacher that is sufficiently narrow and underparameterized in comparison to the sample number. As a corollary, we show that with high probability over the training set, the volume of interpolators with high generalization error is exponentially small in the size of the training set.

In Section 7 we discuss our assumptions (teacher narrowness and weight quantization), how our results can be straightforwardly extended beyond interplators (to functions with a non-zero training error), whether posterior samping biases us towards sparse represetnations, the effect of parameterization via the minimum description length framework, and the relation of our results to SGD.

[3]Proof idea for two-layer FC nets without biases: The NN function is identical to the teacher NN function, if we set $d_1^\star$ hidden neurons with the same ingoing and outgoing weights as in the teacher; and for the other $d_1 - d_1^\star$ hidden neurons we set the outgoing weights (for the zeroed neurons, the input weights do not matter). This event probability is $\tilde{p} = Q^{-d_0 d_1^\star - d_1}$, satisfying (2).

2. Preliminaries

Notation. We use boldface letters for vectors and matrices. A vector $\mathbf{x} \in \mathbb{R}^d$ is assumed to be a column vector, and we use x_i to denote its i-th coordinate. We denote by $\mathrm{Vec}(\cdot)$ the vectorization operation, which converts a tensor into a column vector by stacking its columns. The indicator function $\mathbb{I}[A]$ is 1 if statement A is true and 0 if statement A is false. Additionally, we use the standard notation $[N] = \{1, \ldots, N\}$ and take $\|\cdot\|$ to be the Euclidean norm. We use the symbols $\odot$ to denote the Hadamard product, i.e. elementwise multiplication, $\otimes$ to denote the Kronecker product, and $*$ to denote the convolution operator. For a pair of vectors $D' = (d_1', \ldots, d_L'), D'' = (d_1'', \ldots, d_L'') \in \mathbb{N}^L$ we denote $D' \leq D''$ if for all $l \in [L]$, $d_l' \leq d_l''$.

Data. Let $\mathcal{D}$ be some data distribution. We consider the problem of binary classification over a finite training set $\mathcal{S}$ that contains N datapoints sampled i.i.d. from $\mathcal{D}$:

$$\mathcal{S} \triangleq \{\mathbf{x}_n\}_{n=1}^{N} \sim \mathcal{D}^N,$$

where $\mathbf{x}_n \in \mathbb{R}^{d_0}$. Since we are interested in realizable models we assume there exists a teacher model, $h^\star$, generating binary labels, i.e. $h^\star(\mathbf{x}) \in \{\pm 1\}$ for any $\mathbf{x} \sim \mathcal{D}$.

Evaluation metrics. For a predictor $h : \mathbb{R}^{d_0} \mapsto \{\pm 1\}$, we define the risk, i.e. the population error $\mathcal{L}_{\mathcal{D}}(h) \triangleq \mathbb{P}_{\mathbf{x}\sim\mathcal{D}}\left(h(\mathbf{x}) \neq h^\star(\mathbf{x})\right)$, and the empirical risk, i.e. the training error $\mathcal{L}_{\mathcal{S}}(h) \triangleq \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left[h(\mathbf{x}_n) \neq h^\star(\mathbf{x}_n)\right]$.

Hypothesis parameterization. We discuss parameterized predictors $\boldsymbol{\theta} \mapsto h_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta} \in \mathbb{R}^M$. Distributions over $\boldsymbol{\theta}$ therefore induce distributions over hypotheses via

$$\mathcal{P}(h) \triangleq \mathbb{P}_{\boldsymbol{\theta}}\left(h_{\boldsymbol{\theta}} = h\right).$$

That is, $\mathcal{P}(h)$ is the probability mass function of sampling parameters $\boldsymbol{\theta}$ mapping to h when the distribution is discrete, or (with a slight abuse of notation) their density when the distribution is continuous.

3. Generalization Bounds for Random Interpolating Hypotheses

In this paper, we study the generalization of interpolating predictors sampled from the posterior of NNs:

$$\mathcal{P}_{\mathcal{S}} \triangleq \mathcal{P}(h \mid \mathcal{L}_{\mathcal{S}}(h) = 0) \propto \mathcal{P}(h)\mathbb{I}\left[\mathcal{L}_{\mathcal{S}}(h) = 0\right], \tag{3}$$

where $\mathcal{P}(h)$ is some prior over the hypotheses class. That is, our "learning rule" amounts to sampling a single predictor from the posterior,

$$\mathcal{A}_{\mathcal{P}}(\mathcal{S}) \sim \mathcal{P}_{\mathcal{S}}, \tag{4}$$

and we would like to analyze the population error $\mathcal{L}_{\mathcal{D}}(\mathcal{A}_{\mathcal{P}}(\mathcal{S}))$ of this sampled predictor.

Samples from the posterior $\mathcal{P}_\mathcal{S}$ can be obtained by the Guess and Check procedure (G&C; Chiang et al. (2023)), defined in Algorithm 1, which can be viewed as a rejection sampling procedure for (3). That is, we can think of drawing a sequence $(h_t)_{t=1}^{\infty}$ of hypotheses i.i.d. from the prior $\mathcal{P}$ and independent of $\mathcal{S}$ (i.e. before seeing the training set). Then, given the training set $\mathcal{S}$, we pick the first hypothesis in the sequence that interpolates the data. We will employ this equivalence in our analysis, and view samples from the posterior $\mathcal{P}_\mathcal{S}$ as if they were generated by this procedure.

Algorithm 1 Guess and Check (G&C)

Input: (1) $\mathcal{P}$, Prior over hypotheses (2) $\mathcal{S}$, Training set.
Output: $\mathcal{A}_\mathcal{P}(S)$
Algorithm:
Draw $h_1, h_2, \ldots \overset{\text{i.i.d.}}{\sim} \mathcal{P}$.
Choose $T \triangleq \min\{t \,|\, \mathcal{L}_\mathcal{S}(h_t) = 0\}$
Return: $\mathcal{A}_\mathcal{P}(S) \triangleq h_T$

We will be particularly interested in the case in which $\mathcal{P}(h)$ is defined through a 'uniform' (or otherwise fairly 'flat' or benign) prior on the parameters $\boldsymbol{\theta}$ in some parameterization $h_{\boldsymbol{\theta}}$. But in this section, we analyze posterior sampling, or equivalently Guess and Check, directly through the induced distribution $\mathcal{P}$ over predictors. In particular, we analyze generalization performance in terms of the probability that a random hypothesis $h \sim \mathcal{P}$ is equivalent to the teacher model. This is formalized in the following definition.

Definition 3.1. We say that a predictor h is *teacher-equivalent* (TE) w.r.t. a data distribution $\mathcal{D}$, and denote $h \equiv h^\star$, if $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}\left(h\left(\mathbf{x}\right) = h^\star\left(\mathbf{x}\right)\right) = 1$, and denote the probability of a random hypothesis to be TE by

$$\tilde{p} \triangleq \mathbb{P}_{h\sim\mathcal{P}}\left(h \equiv h^\star\right) .$$

As we show in the next result, $\tilde{p}$ plays an important role in G&C generalization. Specifically, in Appendix B.2 we derive the following generalization bound.

Lemma 3.2 (G&C (i.e. Posterior Sampling) Generalization)**.** *Let $\varepsilon \in (0,1)$ and $\delta \in \left(0, \frac{1}{5}\right)$, and assume that $\tilde{p} < \frac{1}{2}$. For any N larger than*

$$\frac{-\log\left(\tilde{p}\right) + 3\log\left(\frac{2}{\delta}\right)}{\varepsilon} ,$$

the sample complexity, we have that

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N, h\sim\mathcal{P}_\mathcal{S}}\left(\mathcal{L}_\mathcal{D}(h) < \varepsilon\right) \geq 1-\delta .$$

We observe that the sample complexity required to ensure (ε,δ)-PAC generalization depends on $(-\log\left(\tilde{p}\right))$. Thus, we define the effective sample complexity as

$$\tilde{C} \triangleq -\log\left(\tilde{p}\right) .$$

Moreover, using Markov's inequality, the above lemma implies (see Appendix B.3) the following corollary.

Corollary 3.3 (Volume of Generalizing Interpolators)**.** *For ε, δ as above, and any N larger than*

$$\frac{-\log\left(\tilde{p}\right) + 6\log\left(\frac{2}{\delta}\right)}{\varepsilon} ,$$

the sample complexity, we have that

$$\mathbb{P}_\mathcal{S}\left(\mathbb{P}_{h\sim\mathcal{P}_\mathcal{S}}\left(\mathcal{L}_\mathcal{D}\left(h\right) \geq \varepsilon\right) < \delta\right) \geq 1-\delta .$$

Corollary 3.3 implications. Corollary 3.3 examines, for a single sample of the data $\mathcal{S}$, the relative volume of 'bad' interpolators out of all interpolators — i.e. the probability to sample an interpolator for which $\mathcal{L}_\mathcal{D}\left(h\right) \geq \varepsilon$, given the data $\mathcal{S}$. It states that this relative volume is small (δ) with high probability ($1-\delta$) over the sampling of the data. And δ can be quite small, since for any ε, we have that δ decays exponentially fast in N

$$\delta = 2\exp\left(-\frac{\varepsilon N + \log\left(\tilde{p}\right)}{6}\right) .$$

Proof idea of Lemma 3.2 and relationship to PAC-Bayes. In Appendix B.2, we prove Lemma 3.2 by noting that the expected number of hypotheses we will consider is $1/\tilde{p}$, and so we are essentially selecting an interpolating hypothesis from the effective hypothesis class $\mathcal{H} = \{h_1, \ldots, h_\tau\}$ with $\tau \approx 1/\tilde{p}$, where the hypotheses in this class are chosen before seeing the training set. The sample complexity is thus $\log|\mathcal{H}| = \log\tau \approx -\log\tilde{p}$. The only complication is that the stopping time τ of G&C is random and depends on $\mathcal{S}$. But it is enough to bound τ very crudely (which we do with high probability), as a multiplicative factor to τ results only in an additive logarithmic factor.

Remark 3.4. The same method can be straightforwardly used to extend these results to the case where the teacher NN is not a perfect interpolator, and the G&C algorithm is modified to stop when the training error is bellow some threshold instead of 0 (see Appendix B.4).

The analysis here is similar to PAC-Bayes analysis (McAllester, 1999), which also studies the behavior of a posterior over hypotheses, except that in typical PAC-Bayes analysis the bound is over the *expected* population error $\mathbb{E}_{h\sim\mathcal{P}_\mathcal{S}}\left[\mathcal{L}_\mathcal{D}(h)\right]$ for a sample from the posterior. In fact, noting that $\mathrm{KL}(\mathcal{P}_\mathcal{S}\|\mathcal{P}) = -\log\mathbb{P}_{h\sim\mathcal{P}}(\mathcal{L}_\mathcal{S}(h)=0) \geq -\log\tilde{p}$, a standard PAC-Bayes bound (Langford & Seeger, 2002; McAllester, 2003) will yield that with the same sample complexity as in Lemma 3.2,

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\ \mathbb{E}_{h\sim\mathcal{P}_\mathcal{S}}\left[\mathcal{L}_\mathcal{D}(h)\right] < \varepsilon\ \right) \geq 1-\delta. \tag{5}$$

The difference is that Lemma 3.2 holds with high probability for a single posterior sample, instead of just in the expectation over the posterior. That is, Lemma 3.2 establishes that

not only are random interpolators good on average, but only a small fraction of them are bad.

Using Markov's inequality one can derive from (5) a high probability bound for a for a single draw from the posterior (as in Lemma 3.2), but that bound is less tight then Lemma 3.2 (see Appendix B.5). Also, a variant of the PAC-Bayes theorem that holds with high probability for a single draw from the posterior has also been presented by Alquier (2023, Theorem 2.7). As is, Alquier's result yields a much looser bound, since it is more generic (it applies to any posterior, not just conditioning on interpolation). Thus, Lemma 3.2 can be viewed as a tighter specialization to interpolators.

A sample complexity of $(-\log \tilde{p})$ should not be surprising, and can also be obtained by an Occam Razor / Minimum Description Length learning rule $\mathrm{MDL}_{\mathcal{P}}(\mathcal{S}) = \arg\max_{\mathcal{L}_{\mathcal{S}}(h)=0} \mathcal{P}(h)$ (Blumer et al. (1987), and see also Section 7.3 in Shalev-Shwartz & Ben-David (2014)). Here, we discussed how the same sample complexity is obtained by a single draw from the posterior, as in G&C. More interesting is how, starting from a uniform prior $\mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ over parameters, we end up with an informative induced prior $\mathcal{P}(h)$ over hypotheses, which has high $\tilde{p}$ and thus low sample complexity. The key here is redundancy in the parameterization. In Appendix A we review the general principle of how non-uniform redundancy in the parameterization can induce non-uniform informative priors $\mathcal{P}(h)$ and thus low sample complexity. In the next sections we see how this plays specifically for NNs, analyzing $\tilde{p}$ under the prior induced by a uniform choice of NN parameters.

4. Quantized Nets Sample Complexity

Recall that the generalization bound in Lemma 3.2 depends on the effective sample complexity $\tilde{C} \triangleq -\log(\tilde{p})$. In this section, we derive an upper bound on $\tilde{C}$ for quantized multilayer Fully Connected (FC) Neural Networks (NNs) with a single binary output, with and without additional per-node scaling.

Definition 4.1 (Vanilla FC)**.** For a depth L, widths $D = (d_1, \ldots, d_L)$, and activation function $\sigma : \mathbb{R} \to \mathbb{R}$, a fully connected NN is a mapping $\boldsymbol{\theta} \mapsto h^{FC}_{\boldsymbol{\theta}}$ from parameters

$$\left\{ \boldsymbol{\theta} = \left\{ \mathbf{W}^{(l)}, \mathbf{b}^{(l)} \right\}_{l=1}^{L} \;\middle|\; \mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}, \mathbf{b}^{(l)} \in \mathbb{R}^{d_l} \right\}$$

defined recursively, starting with $f^{(0)}(\mathbf{x}) = \mathbf{x}$, as

$$\forall l \in [L-1]: \; f^{(l)}(\mathbf{x}) = \sigma\left(\mathbf{W}^{(l)} f^{(l-1)}(\mathbf{x}) + \mathbf{b}^{(l)} \right)$$
$$h^{\mathrm{FC}}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathrm{sign}\left(\mathbf{W}^{(L)} f^{(L-1)}(\mathbf{x}) + \mathbf{b}^{(L)} \right).$$

The total parameter count is $M(D) = \sum_{l=1}^{L} d_l(d_{l-1}+1)$. We denote the class of all fully connected NNs as $\mathcal{H}^{\mathrm{FC}}_{D}$.

As we will show, considering NNs in which each neuron is multiplied by a scaling parameter can significantly improve the bound. This architecture modification is common in empirical practices, e.g. batch-normalization (Ioffe & Szegedy, 2015), weight-normalization (Salimans & Kingma, 2016), and certain initializations (Zhang et al., 2019). We formally define this model in the following definition.

Definition 4.2 (Scaled-neuron FC)**.** For a depth L, widths $D = (d_1, \ldots, d_L)$, and activation function $\sigma : \mathbb{R} \to \mathbb{R}$, a scaled neuron fully connected neural network is a mapping $\boldsymbol{\theta} \mapsto h^{\mathrm{SFC}}_{\boldsymbol{\theta}}$ from parameters

$$\boldsymbol{\theta} = \left\{ \mathbf{W}^{(l)}, \mathbf{b}^{(l)}, \boldsymbol{\gamma}^{(l)} \right\}_{l=1}^{L},$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}, \mathbf{b}^{(l)} \in \mathbb{R}^{d_l}, \boldsymbol{\gamma}^{(l)} \in \mathbb{R}^{d_l}$, defined recursively, starting with $f^{(0)}(\mathbf{x}) = \mathbf{x}$, as

$$\forall l \in [L-1] \;\; f^{(l)}(\mathbf{x}) = \sigma\left(\boldsymbol{\gamma}^{(l)} \odot \mathbf{W}^{(l)} f^{(l-1)}(\mathbf{x}) + \mathbf{b}^{(l)} \right)$$
$$h^{\mathrm{SFC}}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathrm{sign}\left(\mathbf{W}^{(L)} f^{(L-1)}(\mathbf{x}) + \mathbf{b}^{(L)} \right).$$

The total parameter count is $M(D) = \sum_{l=1}^{L} d_l(d_{l-1}+2)$. We denote the class of all scaled neuron fully connected NNs as $\mathcal{H}^{\mathrm{SFC}}_{D}$.

We consider **Q-quantized** networks where each of the parameters is chosen from a fixed set $\mathcal{Q} \subset \mathbb{R}$ such that $0 \in \mathcal{Q}$ and $|\mathcal{Q}| \leq Q$. This can be the set of integers $\{-\frac{Q}{2}, -\frac{Q}{2}+1, \ldots, (\frac{Q}{2}-1)\}$ for even Q, or the set of numbers representable as $\log_2 Q$-bit floats (for, e.g. $\log_2 Q = 32$). Fully connected quantized NNs thus have parameters $\boldsymbol{\theta} \in \mathcal{Q}^M$ corresponding to a complexity $C = M \log Q$ (from the classic log cardinality bound, see Appendix B.1).

We consider a teacher $h^\star = h_{\boldsymbol{\theta}^\star}$ that is a Q-quantized network of some depth L and small widths $D^\star = (d^\star_1, \ldots, d^\star_L)$, and a wider student of the same depth L but widths $D > D^\star$. For the student, we consider a **uniform prior over Q-quantized parameterizations**, i.e. $\boldsymbol{\theta} \sim \mathrm{Uniform}\left(Q^{M(D)}\right)$. In other words, to generate $h_{\boldsymbol{\theta}} \sim \mathcal{P}$, each weight (and bias) in the NN is chosen independently and uniformly from $\mathcal{Q}$.

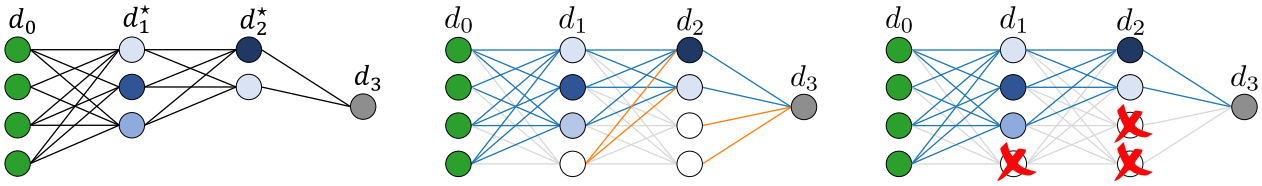
4.1. Main Results

Using the definitions above, we can state the following.

Theorem 4.3 (Main result for fully connected neural networks)**.** *For any activation function such that $\sigma(0) = 0$, depth L, Q-quantized teacher with widths $D^\star$, student with widths $D > D^\star$, $d^\star_0 \triangleq d_0$, and prior $\mathcal{P}$ uniform over Q-quantized parameterizations, we have that:*

1. *For Vanilla Fully Connected Networks:*

$$\tilde{C} \leq \hat{C}^{\mathrm{FC}} \triangleq \left(\sum_{l=1}^{L} \left(d^\star_l d_{l-1} + d^\star_l \right) \right) \log Q \,. \quad (6)$$



(a) Teacher network (b) Vanilla student network (c) Scaled neuron student network

Figure 1. **Illustration of vanilla and scaled neuron three-layer quantized teacher and student neural networks.** Note that the visualization does not show the bias units. The proof of Theorem 4.3 relies on counting student networks which are functionally equivalent to the teacher network. Figure 1(a) depicts a narrow teacher. In Figure 1(b), we visualize a FC student network that replicates the teacher by zeroing out all outgoing weights of any neuron that does not exist in the teacher. Specifically, the blue edges are weights identical to the teacher, and the orange edges are set to zero. Therefore, the white neurons do not affect the network output. In Figure 1(c), we visualize an SFC student network that replicates the teacher by setting the scaling parameter to zero (each zero marked with a red 'x') for any neuron that does not exist in the teacher. In both cases, the gray edges do not affect the function. In this specific example, we can see how the redundancy is higher in SFC than in the vanilla FC network, hinting at better generalization capabilities.

2. *For Scaled Neuron Fully Connected Networks:*

$$\tilde{C} \leq \hat{C}^{\mathrm{SFC}} \triangleq \left(\sum_{l=1}^{L} \left(d_l^\star d_{l-1}^\star + 2d_l \right) \right) \log Q \,. \quad (7)$$

And, by Lemma 3.2, $N = (\tilde{C} + 3\log 2/\delta)/\varepsilon$ samples are enough to ensure that for posterior sampling (i.e. G&C), $\mathcal{L}(\mathcal{A}_{\mathcal{P}}(\mathcal{S})) \leq \varepsilon$ with probability $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^N$ and the sampling.

Remark 4.4. From Corollary 3.3 we deduce that for a given training set, the volume of 'bad' (ε) interpolators is 'small' (δ) with high probability ($1-\delta$) with this sample complexity.

Proof idea. The idea is simple and centers on counting a sufficient number of constraints on the parameters of the student network to ensure it is TE. A fundamental illustration of this concept is provided in the caption of Figure 1. Full proof is in Appendix C.

4.2. Discussion: Comparing Sample Complexities

To understand the sample complexity bound $O(\hat{C}^{\mathrm{FC}})$ and $O(\hat{C}^{\mathrm{SFC}})$ implied by Theorem 4.3, let us first consider the complexities (number of bits, or log cardinalities) of the teacher and student models:

$$C^\star = \left(\sum_{l=1}^{L} \left(d_l^\star d_{l-1}^\star + k d_l^\star \right) \right) \log Q$$

$$C = \left(\sum_{l=1}^{L} \left(d_l d_{l-1} + k d_l \right) \right) \log Q$$

where $k = 1$ for Vanilla FC Networks and $k = 2$ for Scaled FC Networks.

Either way, the dominant term is the quadratic term $\sum_l d_l d_{l-1}$. Lacking other considerations, the sample complexity of learning with the student network would be C. However, we see here that thanks to the parameterization and prior, the student network implicitly adapts to the complexity of the teacher, with $\hat{C} \ll C$ when $D^\star \ll D$ and in any case $C^\star \leq \hat{C} \leq C$. With Vanilla FC Networks, the sample complexity $\hat{C}^{\mathrm{FC}}$, although smaller than C, is still significantly larger than the complexity $C^\star$ of the teacher, which is what we could have hoped for. The quadratic (dominant) terms in $\hat{C}^{\mathrm{FC}}$ are roughly geometric averages of terms from C and $C^\star$, and so we have that, very roughly, $\hat{C}^{\mathrm{FC}} \approx \sqrt{CC^\star}$. We can improve this using scaling, which creates more redundancy since zero scales can deactivate entire units. Indeed, for Scaled Fully Connected Networks, we have that $\hat{C}^{\mathrm{SFC}} = C^\star + \sum_{l=1}^{L} d_l \ll \hat{C}^{\mathrm{FC}} \ll C$ (when $D^\star \ll D$). We still pay a bit for the width of the student, but only linearly instead of quadratically. In particular, even if the width of the student is quadratic in the width of the teacher, the sample complexity of learning by sampling random NNs is almost the same as that of using a much narrower teacher.

Minimum widths. We note that, just as in the minimum description length example in the previous section, an explicit narrowness prior could have of course been fully adaptive to the width of the teacher and ensured learning with the ideal sample complexity $C^\star$. E.g., this is achieved by allowing the student to choose the width of each layer, and using the Occam rule

$$\min_{D' \leq D, \boldsymbol{\theta} \in \mathcal{Q}^{M(D')} \text{ with widths } D'} M(D') \;\; \text{s.t. } \mathcal{L}_S(h_{\boldsymbol{\theta}}) = 0. \quad (8)$$

But from Theorem 4.3 we see that even without such an explicit bias, choosing weights uniformly induces significant inductive bias toward narrow networks.

Maximum sparsity. It is also insightful to compare this to using an explicit sparsity bias, e.g. with an Occam rule of

the form:

$$\min_{\boldsymbol{\theta}\in\mathcal{Q}^{M(D)}} \|\boldsymbol{\theta}\|_0 \quad \text{s.t. } \mathcal{L}_S(h_{\boldsymbol{\theta}}) = 0. \tag{9}$$

The sparsity-inducing rule (9) would have the following bound for the effective sample complexity

$$\hat{C}^{\text{sparse}} \triangleq O\left(M\left(D^\star\right)\left(\log\left(M\left(D^\star\right)+d_0\right)+\log Q\right)\right). \tag{10}$$

See Appendix E for a derivation of this equation. Note that $\hat{C}^{\text{sparse}}$ in (10) does not depend on the size of the student, but rather only on the size of the teacher, which is smaller, that is $M\left(D^\star\right) \ll M\left(D\right)$. For comparison, recall our previous bound for posterior sampling, in the case of FC scaled-neuron networks from (7):

$$\begin{aligned}\hat{C}^{\text{SFC}} &= \left(\sum_{l=1}^{L}\left(d_l^\star d_{l-1}^\star + 2d_l\right)\right)\log Q \\ &= O\left(C^\star + \sum_{l=1}^{L} 2d_l \log Q\right).\end{aligned}$$

When $\log Q = O(1)$, we can rewrite $\hat{C}^{\text{sparse}}$ as follows:

$$\hat{C}^{\text{sparse}} = O\left(C^\star\left(1+\frac{\log M\left(D^\star\right)}{\log Q}\right)\right) = O\left(C^\star \log C^\star\right),$$

that is $\hat{C}^{\text{sparse}}$ has an additional multiplicative factor bounded by $\log C^\star$, and so can be worse than the bounds for posterior sampling. Specifically, in the regime where $2\sum_{l=1}^{L} d_l \log Q \leq \frac{\log M(D^\star)}{\log Q}$, we have that $\hat{C}^{\text{SFC}} < \hat{C}^{\text{sparse}}$. For example, when $\sum_{l=1}^{L} d_l \log Q = O\left(C^\star\right)$, we have that $\hat{C}^{\text{SFC}}$ is suboptimal with a factor of 2 with respect to $C^\star$, whereas $\hat{C}^{\text{sparse}}$ is off by a larger factor of $\log C^\star$.

Minimum norm. One might also ask whether a similar adaptation to teacher width can be obtained by regularizing the norm of the weights. Indeed, (Neyshabur et al., 2015; Golowich et al., 2018) obtained sample complexity bounds that depend only on the ℓ_2 norm $\|\boldsymbol{\theta}\|$ of the learned network, without any dependence on the width of the student (but with an exponential dependence on depth!). These guarantees are not directly applicable in our setting, since applying them to get guarantees on the misclassification error we study here requires bounding the margin. Even with a discrete teacher, without further assumptions on the input distribution, we cannot ensure a margin. If we did consider only integer inputs and integer weights, we could at least ensure a margin of 1. In this case: on one hand, the norm-based sample complexity would scale as $Q^{O(L)}$, i.e. exponential in the $L\log Q$ dependence of $\hat{C}$. On the other hand, the norm-based guarantee would not depend at all on the student widths D, while even $\hat{C}^{\text{SFC}}$ increases linearly with the student widths.

4.3. Extension to convolutional neural networks

In Appendix C, we extend Theorem 4.3 to convolutional neural networks (CNN) and convolutional neural networks where each channel is multiplied by a learned parameter (SCNN). Specifically, we show that for quantized convolutional networks, we get similar bounds on the sample complexity $\hat{C}$ with channel numbers substituting layer widths.

$$\begin{aligned}\hat{C}^{\text{CNN}} &= \left(d_s + 1 + \sum_{l=1}^{L}\left(k_l c_l^\star c_{l-1} + c_l^\star\right)\right)\log Q \\ \hat{C}^{\text{SCNN}} &= \left(d_s^\star + 1 + \sum_{l=1}^{L}\left(k_l c_l^\star c_{l-1}^\star + 2c_l\right)\right)\log Q,\end{aligned}$$

where d_s is the number of neurons in the last convolutional layer (i.e., the width of the last layer which is a fully connected layer with a single output), k_l and c_l are the l^{th} layer's kernel size and number of channels. See Appendix C for precise definitions of the model and statement of the result. Importantly, similar to the fully connected architecture we observe that $\hat{C}^{\text{SCNN}}$ is again close to the teacher's complexity, with only a weak dependence on the student's channel numbers. The proof here is analogous to the proof of Theorem 4.3, where neurons are replaced with channels.

Implications to realistic benchmarks. Currently, no NN has achieved zero test error on real-world datasets, even for simple ones like MNIST. Therefore, the size of the teacher NN is unknown to us when facing practical applications. However, we can use the dimensions of common NN architectures and datasets to approximate the necessary size of the teacher to obtain meaningful generalization bounds using our results. For example, suppose that we substitute the actual sizes of a training set and network into our bound

$$N = \frac{\hat{C} + 3\log\left(\frac{2}{\delta}\right)}{\varepsilon}.$$

We can deduce the size of the teacher required to satisfy this equation. For example, using the size of the ImageNet dataset, $\delta = 0.05$, and the actual test error of some CNNs, we can estimate the required size of the teacher. For simplicity of calculation, we assume that each layer of the teacher NN has exactly α channels compared to the same layer in the student NN, where $0 < \alpha < 1$. In the following table, we show the required width reduction α and the number of parameters in the resulting teacher NN. We use the smallest quantization level (2bit) for which the NN accuracy remains near the FP32 accuracy (less than 0.5% degradation, from (Liu et al., 2022b)).

Table 1 shows that our bound is consistent with a narrow teacher of non-trivial size and widths. For example, in ResNet18 the resulting channel numbers in the teacher layers are $[3, 8\ldots, 8, 16, \ldots 16, 32\ldots 32, 64, \ldots 64]$, where '3'

Architecture	ε	α	#parameters
ResNet18	0.3	0.125	$\sim 241k$
ResNet50	0.25	0.05	$\sim 159k$

Table 1. Approximate relative width reduction and number of parameters in the teacher required to obtain meaningful bounds using standard ResNet architectures and the ImageNet dataset.

counts the input channels and the rest are the following hidden layers. This architecture can implement highly complex non-linear functions. Note that we cannot directly validate the existence of this teacher, as standard optimization methods may not be capable of finding such a solution efficiently, even if it exists (without over-parameterization, it is much harder to find global minima).

5. Continuous Nets Sample Complexity

So far, we focused on quantized uniform priors. However, per-layer continuous spherical priors (e.g., Gaussian) are also quite common in practice and theory. Therefore, in this section, we show how to extend our results beyond the quantized case into a continuous setting, for the special case of two-layer NNs without bias and with the leaky rectifier linear unit (LReLU, Maas et al. (2013)) activation function. Formally, let $h_{\boldsymbol{\theta}}, h_{\boldsymbol{\theta}^\star}$ be fully connected (Definition 4.1) two layer NNs with input dimension d_0, output dimension $d_2 = 1$ and hidden layer dimensions d_1 and $d_1^\star$, respectively. Explicitly:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \operatorname{sign}\left(\mathbf{W}^{(2)}\sigma\left(\mathbf{W}^{(1)}\mathbf{x}\right)\right)$$

$$h_{\boldsymbol{\theta}^\star}(\mathbf{x}) = \operatorname{sign}\left(\mathbf{W}_\star^{(2)}\sigma\left(\mathbf{W}_\star^{(1)}\mathbf{x}\right)\right)$$

where

$$\mathbf{W}^{(1)} = [\mathbf{w}_1, \ldots, \mathbf{w}_{d_1}]^\top \in \mathbb{R}^{d_1 \times d_0}, \ \mathbf{W}^{(2)} \in \mathbb{R}^{d_2 \times d_1},$$

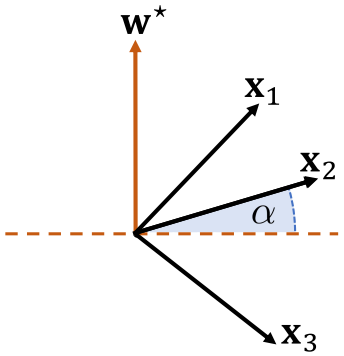
$$\mathbf{W}_\star^{(1)} = \left[\mathbf{w}_1^\star, \ldots, \mathbf{w}_{d_1^\star}^\star\right]^\top \in \mathbb{R}^{d_1^\star \times d_0}, \ \mathbf{W}_\star^{(2)} \in \mathbb{R}^{d_2 \times d_1^\star},$$

and $\sigma(\cdot)$ is the common LReLU with parameter $\rho \notin \{0, 1\}$.

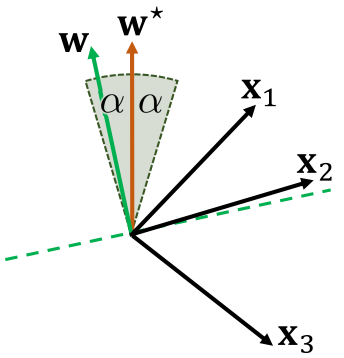
As in the previous sections, our goal is to obtain generalization guarantees by lower bounding $\tilde{p}$ and then combining this results with Lemma 3.2. To this end, we first need to define some prior on the hypotheses.

Assumption 5.1 (Prior over parameters, continuous setting)**.** Suppose that the weights of $h_{\boldsymbol{\theta}}$ are random such that each row of the first layer, $\mathbf{w}_i$, is independently sampled from a uniform distribution on the unit sphere $\mathbb{S}^{d_0-1}$, and the second layer $\mathbf{W}^{(2)}$ is sampled uniformly[4] from $\mathbb{S}^{d_1-1}$. Both $\mathbf{W}^{((1))}$ and $\mathbf{W}^{(2)}$ are independent of the teacher and data.

[4]Sampling $\mathbf{W}^{(2)}$ from the unit sphere is equivalent to sampling it from a Gaussian distribution. In fact, any spherically symmetric distribution in $\mathbb{R}^{d_1}$ will suffice, as it amounts to scaling of the output without affecting the classification.



(a)



(b)

Figure 2. A two-dimensional illustration of the first layer angular margin. In 2(a), we show how the angle α is defined for a single $\mathbf{w}_i^\star$. Note that α is defined as the minimal angle when considering all rows of $\mathbf{W}_\star^{(1)}$. In 2(b), we illustrate how α margin creates a cone around $\mathbf{w}_i^\star$ in which any $\mathbf{w}_i$ results in the same activation pattern as $\mathbf{w}_i^\star$ (i.e., as the teacher) on a training set.

To extend the notion of teacher-equivalence from the quantized setting to the continuous setting, we assume an "angular margin" exists between the training set $\mathcal{S}$ and the teacher, similarly to Soudry & Hoffer (2017).

Definition 5.2 (First layer angular margin)**.** For any training set $\mathcal{S} = \{\mathbf{x}_n\}_{n=1}^N$, we say that $\mathcal{S}$ has *first layer angular margin* α w.r.t. the teacher if

$$\forall i \in [d_1^\star], n \in [N] : \left|\frac{\mathbf{x}_n^\top \mathbf{w}_i^\star}{\|\mathbf{x}_n\|_2 \|\mathbf{w}_i^\star\|_2}\right| > \sin\alpha . \qquad (11)$$

In words, we say that $\mathcal{S}$ has *first layer angular margin* α if all datapoints $\mathbf{x}_n$ are at an angle of at least α from any hyperplane induced by a row $\mathbf{w}_i^\star$ of the matrix $\mathbf{W}_\star^{(1)}$. Here, the rows of $\mathbf{W}_\star^{(1)}$ represent the normals to the hyperplanes. As illustrated in Figure 2, we require some angular margin $\alpha > 0$ to guarantee that the first layer of the student network can be within a certain angular margin of the first layer of the teacher network and still achieve accurate classification on the training set. This assumption prevents degenerate

neurons in the first layer of the teacher network.

Similarly, for the output of the teacher network, we define the second layer angular margin.

Definition 5.3 (Second layer angular margin)**.** For any training set $\mathcal{S} = \{\mathbf{x}_n\}_{n=1}^N$, we say that $\mathcal{S}$ has *second layer angular margin* β w.r.t. the teacher if

$$\forall n \in [N]: \left| \frac{\mathbf{W}_\star^{(2)} \sigma\left(\mathbf{W}_\star^{(1)} \mathbf{x}_n\right)}{\left\|\mathbf{x}_n\right\|_2 \left\|\mathbf{W}_\star^{(2)}\right\|_2} \right| > \sqrt{d_1(1+\rho^2)} \sin \beta. \quad (12)$$

In essence, this ensures some margin in the output of the teacher network. With this definition, our main assumption for the continuous case is stated below.

Assumption 5.4. Let $\alpha < \beta \in \left(0, \frac{\pi}{2}\right)$. There exists $\lambda \in (0,1)$ such that with probability at least $1-\lambda$ over $\mathcal{S} \sim \mathcal{D}^N$, $\mathcal{S}$ has first layer angular margin α (Definition 5.2) and second layer angular margin β (Definition 5.3).

Remark 5.5. Note that α is the minimal margin of all hidden neurons, while β is the margin of the single network output. Thus, intuitively, β is usually larger than α. For Gaussian data, we show empirically in Figure 3 in Appendix D, that the assumption $\beta > \alpha$ holds with high probability.

This assumption allows us to extend the results from the previous section to a continuous setting (proof in Appendix D).

Theorem 5.6 (Interpolation of Continuous Networks)**.** *Assume that $\hat{p}_{\mathcal{S}} < \frac{1}{2}$ a.s. and $d_0 \gg d_1^\star \gg 1$ [5]. Then under Assumption 5.4, for any $\varepsilon, \delta \in (0,1)$ we have*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon\right) \geq 1 - \delta - \lambda,$$

whenever N is larger than the sample complexity

$$\frac{\hat{C}^{\text{cont}} + 2\log\left(\hat{C}^{\text{cont}}\right) + 4\log\left(\frac{8}{\delta}\right)}{\varepsilon},$$

with $\gamma = \arccos\frac{\cos\beta}{\cos\alpha}$ and

$$\hat{C}^{\text{cont}} = -d_1^\star d_0 \log\left(\sin\left(\alpha\right)\right) - d_1 \log\left(\sin\left(\gamma\right)\right) + \frac{1}{2} d_1^\star \log\left(d_0\right) + O\left(d_1^\star + \log\left(d_1\right)\right).$$

Remark 5.7. Since Assumption 5.4 gives a positive margin only in high probability, we use a different generalization guarantee than the one in Lemma 3.2 which takes into account the interpolation probability $\hat{p}_{\mathcal{S}}$, instead of $\tilde{p}$. See Appendix B.7 for more details.

Remark 5.8. Although Assumption 5.4 may be natural in some cases[6], in other cases the margins α, β may decay with N or with the network's dimensions. Therefore, even if the assumption holds, for the generalization bound in Theorem 5.6 to remain meaningful, we need the margin decay rate to be sufficiently slow.[7]

Implications. Our results in this section rely on a margin assumption instead of quantized weights. This margin assumption can improve the quantized bound given the data $\mathcal{S}$ has a large enough margin from the teacher. However, for a generic input distribution without any pre-set margin (e.g., standard Gaussian) we observed empirically the resulting margin is near the numerical precision level, so the resulting bound is not better than the quantized approach.

6. Related work

Random interpolating neural networks. Chiang et al. (2023) empirically studied a few gradient-free algorithms for the optimization of neural networks, suggesting that generalization is not necessarily due to the properties of the SGD. For the G&C algorithm, they also empirically investigated the effect of the loss on the generalization. We focus on the 0-1, rather than some surrogate loss function. Chiang et al. (2023) focused only on small-scale datasets since it is not possible to run their experiments when the training set is larger than a few dozen samples. Theisen et al. (2021) provided a theoretical analysis of random interpolating linear models. In contrast, our work focuses on deeper models with non-linear activation functions, and our teacher assumption relies on depth, such that it is possible to have many zeroed-out hidden neurons and obtain generalization guarantees that stem from the redundancy in parameterized deep networks. Valle-Perez et al. (2019); Mingard et al. (2021) used a connection between neural networks and Gaussian processes to model random interpolating networks, which usually requires the width of the network to be infinite, in contrast with our finite width setting. Teney et al. (2024) empirically investigated random fully connected neural networks without conditioning on the interpolation of a training set. They examined their spectral properties and found a bias towards simple functions (according to various metrics). Berchenko (2024) showed generalization results for uniformly sampled trees learning Boolean functions, and discuss the simplicity bias of random deep NNs. In contrast, we proved an explicit generalization bound for typical interpolating deep neural networks.

Redundancy in neural networks. The Lottery Ticket Hypothesis (Frankle & Carbin, 2019) suggests that for any random neural network, there exists a sparse sub-network capable of competitive generalization with the original. A re-

[5] This assumption is used to simplify the bound. A non-asymptotic version that does not require this assumption is in Appendix D.

[6] Realistic data many times has some intrinsic margin. For example, there is a low probability for (semantic) 'mixings' of dogs and birds from a distribution of natural images of dogs and birds (with a single animal in every image).

[7] For example, for Gaussian data with a random teacher it is possible to derive a lower bound in high probability on the first layer angular margin α and show that it decays as $1/(Nd_0d_1^\star)$, so it adds only a multiplicative $\log(Nd_0d_1^\star)$ factor to the bound.

lated hypothesis was conjectured by Ramanujan et al. (2020) and later proven by Malach et al. (2020) and states that it is possible to prune an initialized neural network without significantly affecting the obtained predictions. In contrast, our work, while also focusing on random neural networks, is oriented towards providing assurances regarding the generalization of these random networks, when conditioned on the event of perfectly classifying the training data. Also, we operate under the assumption of a narrow teacher, which is sparse in the number of neurons — in contrast to the subnetwork in the Lottery Ticket Hypothesis which is sparse in the number of weights.

7. Discussion

Summary. In this work, we examined the generalization of samples from the NN posterior with the 0-1 loss (3). We proved that even when using a uniform (or 'uniform-like') prior for the parameters, typical samples from the NN posterior are biased toward low generalization error — if there exist sufficiently narrow NN teachers.

Implications of narrowness. Assuming a narrow teacher may sometimes limit the class of possible target functions. For example, in a fully connected NN, if the first hidden layer width is smaller than the input width ($d_0 > d_1^\star$) then the teacher NN must be constant on the $d_0 - d_1^*$ nullspace of the first weight layer. However, this is not an issue for a convolutional NN, since typically the number of input channels is small (e.g., $c_0 = 3$ for RGB images), and so $c_0 < c_1^*$ is reasonable. In any case, (1) many realistic datasets are low rank (Udell & Townsend, 2019; Zeno et al., 2024), and (2) in order to represent potentially complex target functions with narrow NNs we require depth (Kidger & Lyons, 2020) — both for student and teacher. This might suggest another reason why deep architectures are more useful in realistic settings.

Quantized models. Our results in Section 4 rely on NNs to have a quantized weights. The sample complexity we obtain is a product of the parameter count of the model (mainly the teacher) and the quantization bits. The simplicity of the bound allows it to be applied to other architectures as well. For example, we can add pooling layers to the CNN models we analyzed. Also, using the same considerations, we may obtain similar results for multi-head attention layers when the student has redundant heads and no LayerNorm. Moreover, using quantization-aware methods (Hubara et al., 2018) one can reduce numerical precision to improve the bound. For example, using 2bits weights (and activations) in ResNet50 on ImageNet results in only 0.5% degradation in accuracy (Liu et al., 2022a)). Such quantization approaches are common in compression-type bounds (e.g. Lotfi et al. (2022)). However, there the goal is to compress the learned model (the student), while here we only need some compressed model to *exist* (i.e. the narrow teacher).

Beyond interpolators. In this work we focus for simplicity on NN interpolators, but in many scenarios, the training loss does not reach zero. In Appendix B.4 we show how to generalize our results to this non-realizable case, where the teacher NN does not reach zero trainning error (i.e., there is some irreducible error). Unfortunately, the use of the non-realizable generalization bound for finite hypothesis classes introduce a quadratic dependence of the sample complexity on the generalization error bound ε. It is interesting to see if this quadratic dependence can be improved.

Does posterior sampling bias towards sparse representations? Our results indicate that posterior sampling biases NNs towards sparse representations. To examine whether this also happens in other models, we empirically examined posterior sampling in a sparse regression setting for linear diagonal networks (a common theoretical model, e.g. Woodworth et al. (2020); Moroshko et al. (2020)) with a Gaussian weights prior, where the ground truth has a single nonzero component. In that case, we do get a bias toward a sparser predictor, but only with sufficient depth. Specifically, we found that a small depth (2 or 3) does not help much to improve generalization compared to depth 1. However, larger depth did seem to help significantly. In contrast, the NN's result in this paper is different, since there depth is not necessary to obtain good generalization results (as our results hold even with depth 2).

Parameterization and minimum description length. The results in this paper rely heavily on the choice of parameterization of the hypotheses. Specifically, as discussed in Section 3, we can obtain results similar to Lemma 3.2 using a Minimum Description Length (MDL)/Occam's Razor learning rule. Choosing a different mapping from parameters (or descriptions) to hypotheses may not benefit from the redundancy and will result in different generalization bounds which may be worse and even trivial. An example for such parameterization and further discussion are presented in Appendix A.

Relation to SGD. As we have mentioned in Remark 4.4, the volume of 'bad' interpolators is exponentially decaying in the size of the training set. Therefore, algorithms with bad generalization properties must have significant probability to sample this 'small' subset of hypotheses. Therefore, it would be interesting to know when practical training algorithms are biased towards this 'bad' domain. For example, the implicit bias of SGD is partially known in some cases (e.g., Lyu & Li (2020)), so it would be interesting to understand the generalization of typical NN interpolators that also obey this implicit bias.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

The authors would like to thank Itay Evron and Yaniv Blumenfeld for valuable comments and discussions. The research of DS was Funded by the European Union (ERC, A-B-C-Deep, 101039436). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (ERCEA). Neither the European Union nor the granting authority can be held responsible for them. DS also acknowledges the support of the Schmidt Career Advancement Chair in AI. Part of this work was done as part of the NSF-Simons funded Collaboration on the Theoretical Foundations of Deep Learning. NS was supported in part by NSF IIS awards.

References

Alquier, P. User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*, 2023.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Berchenko, Y. Simplicity bias in overparameterized machine learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):11052–11060, Mar. 2024. doi: 10.1609/aaai.v38i10.28981. URL https://ojs.aaai.org/index.php/AAAI/article/view/28981.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Occam's razor. *Information processing letters*, 24 (6):377–380, 1987.

Chiang, P., Ni, R., Miller, D. Y., Bansal, A., Geiping, J., Goldblum, M., and Goldstein, T. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1305–1338. PMLR, 09–12 Jul 2020.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., and Bengio, Y. Quantized neural networks: Training neural networks with low precision weights and activations. *journal of machine learning research*, 18(187):1–30, 2018.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Kidger, P. and Lyons, T. Universal approximation with deep narrow networks. In *Conference on learning theory*, pp. 2306–2327. PMLR, 2020.

Langford, J. and Seeger, M. *Bounds for averaging classifiers*. CMU Technical Report CMU-CS-01-102, 2002, 2002.

Liu, Z., Cheng, K.-T., Huang, D., Xing, E. P., and Shen, Z. Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4942–4952, June 2022a.

Liu, Z., Cheng, K.-T., Huang, D., Xing, E. P., and Shen, Z. Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4942–4952, 2022b.

Lotfi, S., Finzi, M., Kapoor, S., Potapczynski, A., Goldblum, M., and Wilson, A. G. Pac-bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Maas, A. L., Hannun, A. Y., Ng, A. Y., et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, pp. 3. Atlanta, GA, 2013.

Malach, E., Yehudai, G., Shalev-Schwartz, S., and Shamir, O. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pp. 6682–6691. PMLR, 2020.

McAllester, D. Simplified pac-bayesian margin bounds. In Schölkopf, B. and Warmuth, M. K. (eds.), *Learning Theory and Kernel Machines*, pp. 203–215, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45167-9.

McAllester, D. A. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, Dec 1999. ISSN 1573-0565. doi: 10.1023/A:1007618624809.

Mingard, C., Valle-Pérez, G., Skalse, J., and Louis, A. A. Is sgd a bayesian sampler? well, almost. *The Journal of Machine Learning Research*, 22(1):3579–3642, 2021.

Moroshko, E., Woodworth, B. E., Gunasekar, S., Lee, J. D., Srebro, N., and Soudry, D. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *Advances in neural information processing systems*, 33: 22182–22193, 2020.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In *Conference on learning theory*, pp. 1376–1401. PMLR, 2015.

Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A., and Rastegari, M. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11893–11902, 2020.

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Soudry, D. and Hoffer, E. Exponentially vanishing sub-optimal local minima in multilayer neural networks. In *ICLR workshop paper*, 2017.

Soudry, D., Hoffer, E., and Srebro, N. The implicit bias of gradient descent on separable data. In *International Conference on Learning Representations*, 2018.

Teney, D., Nicolicioiu, A., Hartmann, V., and Abbasnejad, E. Neural redshift: Random networks are not random functions. *arXiv preprint arXiv:2403.02241*, 2024.

Theisen, R., Klusowski, J., and Mahoney, M. Good classifiers are abundant in the interpolating regime. In *International Conference on Artificial Intelligence and Statistics*, pp. 3376–3384. PMLR, 2021.

TV, M. If x, y are independent χ^2 with m and n degrees of freedom, then $\frac{X}{X+Y} \sim \beta(m/2, n/2)$. Mathematics Stack Exchange, 2017. URL https://math.stackexchange.com/q/2263641. URL:https://math.stackexchange.com/q/2263641 (version: 2017-05-03).

Udell, M. and Townsend, A. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

Valle-Perez, G., Camargo, C. Q., and Louis, A. A. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019.

Vardi, G. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6):86–93, 2023.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Zeno, C., Ongie, G., Blumenfeld, Y., Weinberger, N., and Soudry, D. How do minimum-norm shallow denoisers look in function space? *Advances in Neural Information Processing Systems*, 36, 2024.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Zhang, H., Dauphin, Y. N., and Ma, T. Residual learning without normalization via better initialization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gsz30cKX.

A. Redundancy and Description Length

In this section, we want to give a broader perspective on the effect of the parameterization on $\tilde{p} = \mathbb{P}_{h\sim\mathcal{P}}\left(h \equiv h^\star\right)$. As we saw, the parameterization controls the generalization behavior of G&C and posterior sampling, and how a seemingly uniform prior over parameters can induce a rich prior over hypothesis via redundancy. We first consider a conceptually very direct parameterization.

How parameter redundancy works: a minimal example. Consider a predictor h_σ specified by a description language (i.e. parameterization) using bit strings σ, where descriptions end with the string "END" (encoded in bits), and this string never appears elsewhere in the description[8]. We will consider students that use fixed-length descriptions, i.e., $\sigma \in \{0,1\}^C$, where we only consider the description up to the first "END", i.e. $h_\sigma = h_{\tilde{\sigma}}$ where $\tilde{\sigma}$ is a prefix of σ ending with the first "END".[9] Let us now consider a uniform prior over this parameterization, i.e. $\mathcal{P}(\sigma) = 2^{-C}$ for all $\sigma \in \{0,1\}^C$. Although this prior is uniform over *parameterizations*, it is easy to see that it induces a highly *non-uniform* prior over predictors: for a predictor h_τ with a short description $\tau \in \{0,1\}^{|\tau|}$, $|\tau| \ll C$, it's C-bit description is highly redundant, as the final $C - |\tau|$ bits can be set arbitrarily, as long as the first bits are τ, and so $\mathcal{P}(h_\tau) \geq 2^{-|\tau|}$. For a teacher with a short description $h^\star = h_{\tau^\star}$, $\tau^\star \in \{0,1\}^{C^\star}$, we thus have $-\log\tilde{p} \leq C^\star$. The sample complexity of learning a small teacher thus depends *only* on the complexity of the teacher and not the complexity of the (possibly much larger) student.

Why would MDL fail here? Returning to the MDL / Occam principle mentioned at the end of Section 3, it is important to note that applying this principle to the distribution over *parameters* would not work here. Consider the MDL / Occam / MAP rule which selects the most likely interpolating parameters $\arg\max_{\mathcal{L}_\mathcal{S}(h_\sigma)} \mathcal{P}(\sigma)$. This rule would be useless here, since for all σ in our parameter space $\{0,1\}^C$ have the same prior probability, and its associated sample complexity would be $O(-\log\mathcal{P}(\sigma^\star)) = C$ (not $\mathcal{P}(h_{\sigma^\star})$!), where $\sigma^\star$ is a description of $h^\star$ in our parameter space (e.g. $\sigma^\star = \tau^\star + \text{`0'}\cdot(C - C^\star)$). The important distinction is that Occam considers the probability mass function over *parameters* (roughly, the density under a specific parameterization or base measure). On the other hand, even if we sample h_σ by sampling parameters σ, posterior sampling (i.e. G&C) can be thought of directly in terms of the *distribution over hypotheses*.

When would redundancy fail? To see how changing the parameterization can change the induced distribution over hypothesis, consider instead a non-redundant parameterization, where we take σ to be uniform over valid descriptions, of length at most C, ending with "END" (i.e. strings ending with "END" and that do not otherwise contain "END"). The induced prior over hypothesis is now uniform[10] and even using posterior sampling/G&C would have sample complexity determined by the complexity of the student, rather than the teacher.

When would MDL succeed? Finally, we note that we could of course learn a short teacher with an Occam rule that is *explicitly* biased towards short descriptions, e.g. using the Kraft prior[11] $\mathcal{P}(\sigma) = 2^{-|\sigma|}$. Using such a prior, the Occam rule also enjoys a sample complexity of $O(C^\star)$ that depends only on the length of the teacher. But here the prior over parameters is non-uniform and explicitly biased, and our interest in this paper is in how seemingly uniform priors over parameters can induce non-uniform priors and generalization over predictors due to the choice of parameterization.

[8] This can be a Turing complete programming language, or perhaps better to think of simpler descriptions such as boolean formulas or strings encoding decision trees.

[9] If σ does not contain "END", or $\tilde{\sigma}$ is not a valid description, we can set h_σ to the constant 0 predictor.

[10] We might still have further redundancies in that multiple valid descriptions can describe the same function, again introducing non-uniformity. Consider here a non-redundant description language, e.g. a read-once branching program.

[11] We again absorb any remaining probability in the constant zero predictor. Here we are thinking of an unbounded student and the parameter space being any string ending with the first occurrence of "END". We can also of course bound the size of the student.

B. Generalization Results

This section contains:

- A restatement and discussion on the connection to a well-known result from (Shalev-Shwartz & Ben-David, 2014, Corollary 2.3) in Appendix B.1.
- The proof of our primary generalization result (Lemma 3.2) in Appendix B.2.
- The proof of Corollary 3.3 in Appendix B.3.
- An extension of Lemma 3.2 to non-interpolating solutions in Appendix B.4.
- A further discussion on the relation of Lemma 3.2 to PAC-Bayes is brought in Appendix B.5
- In Appendix B.6 we prove a refined version of Lemma 3.2 and discuss its implications.
- The proof of an alternative generalization result, used for proving generalization in the continuous setting, in Appendix B.7.

B.1. Finite Hypothesis Class PAC Generalization Bound

Here we present a classic generalization bound for finite hypothesis class adapted from (Shalev-Shwartz & Ben-David, 2014), which we will use throughout our paper.

Theorem B.1. [*adapted from (Shalev-Shwartz & Ben-David, 2014, Corollary 2.3)*] *Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$ and $\varepsilon > 0$ and let N be an integer that satisfies*

$$N \geq \frac{\log\left(|\mathcal{H}|/\delta\right)}{\varepsilon}\,.$$

Then, for any realizable data distribution $\mathcal{D}$, with probability of at least $1-\delta$ over the choice of an i.i.d sample S of size N, we have that for every interpolating[12] hypothesis, $h_{\boldsymbol{\theta}}$, it holds that

$$\mathcal{L}_{\mathcal{D}}\left(h_{\mathcal{S}}\right) \leq \varepsilon\,.$$

Applying Theorem B.1 to quantized neural networks. It is possible to directly obtain a generalization bound for any quantized model, such as those presented in Section 4. Note that the hypothesis class of quantized models with a finite number of parameters, for example, neural networks with finite width and depth, with M parameters and Q quantization levels is finite and of size $|\mathcal{H}| = Q^M$. Therefore, the sample complexity from Theorem B.1 becomes

$$N \geq \frac{M\log\left(Q\right) + \log\left(\frac{1}{\delta}\right)}{\varepsilon} \tag{13}$$

Note that Theorem B.1 is used to prove Lemma 3.2.The sample complexity in (13) is dependent on the student parameters, and therefore much worse than the sample complexities we derive throughout our paper.

[12](Shalev-Shwartz & Ben-David, 2014, Corollary 2.3) is more general, but we focus on interpolating hypotheses.

B.2. Proving the Generalization of Algorithm 1

In this section, we rely on the existence of the teacher to provide guarantee on the generalization of Algorithm 1. We will first restate Lemma 3.2.

Lemma B.2 (Lemma 3.2 restated)**.** *Let $\varepsilon \in (0,1)$ and $\delta \in \left(0, \frac{1}{5}\right)$, and assume that $\tilde{p} < \frac{1}{2}$. For any N larger than*

$$\left(-\log\left(\tilde{p}\right) + 3\log\left(\frac{2}{\delta}\right)\right)\frac{1}{\varepsilon}\,,$$

the sample complexity, we have that

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon\right) \geq 1-\delta\,.$$

Proof Outline. The idea is to show that the probability to sample a TE model within a finite number of steps τ is large. Conditioning on that event, we treat those models as a realizable finite hypothesis class, obtaining the sample complexity using Theorem B.1, which is then bounded, and after some technical details the theorem is obtained.

We first recall Def. 3.1 and add some notation:

Definition B.3. [Def. 3.1 extended] For any hypothesis class $\mathcal{H}$, we say that $h \in \mathcal{H}$ is a *teacher-equivalent w.r.t $\mathcal{D}$* (TE) model, and denote $h \equiv h^\star$, if

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}\left(h\left(\mathbf{x}\right) = h^\star\left(\mathbf{x}\right)\right) = 1\,.$$

We denote the probability of a random hypothesis to be TE by

$$\tilde{p} = \mathbb{P}_{h\sim\mathcal{P}}\left(h \equiv h^\star\right)$$

and denote

$$h^\star \in \mathcal{H} \iff \exists h \in \mathcal{H} \;:\; h \equiv h^\star\,.$$

Definition B.4. Recall the sequence $(h_t)_{t=1}^{\infty}$ from Algorithm 1 is sampled such that $h_t \overset{\text{i.i.d.}}{\sim} \mathcal{P}$. For any $\tau \in \mathbb{N}$ we define

$$\mathcal{H}_\tau = \{h_1, \ldots, h_\tau\} \sim \mathcal{P}^\tau$$

as the finite hypothesis class created from the first τ hypotheses.

We first show that the probability of sampling a TE model early enough is large.

Lemma B.5. *For $\tau = \left\lceil \frac{\log(\delta_h)}{\log(1-\tilde{p})} \right\rceil$ it holds that*

$$\mathbb{P}_{\mathcal{H}_\tau}\left(h^\star \in \mathcal{H}_\tau\right) \geq 1-\delta_h\,.$$

That is, the probability to have a TE model within the first τ sampled model is at least $1-\delta_h$.

Proof. Since h_t is sampled $i.i.d$ from $\mathcal{P}$, we have that

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_\tau}\left(h^\star \in \mathcal{H}_\tau\right) &= 1 - \mathbb{P}\left(\forall t = 1,\ldots,\tau \; h_t \not\equiv h^\star\right) \\
&= 1 - \prod_{t=1}^{\tau} \mathbb{P}\left(h_t \not\equiv h^\star\right) \\
&= 1 - \left(1-\tilde{p}\right)^\tau\,.
\end{aligned}$$

Choosing $\tau = \left\lceil \frac{\log(\delta_h)}{\log(1-\tilde{p})} \right\rceil$ we get

$$\begin{aligned}
\left(1-\tilde{p}\right)^\tau &= \left(1-\tilde{p}\right)^{\left\lceil \frac{\log(\delta_h)}{\log(1-\tilde{p})} \right\rceil} \\
&\leq \left(1-\tilde{p}\right)^{\frac{\log(\delta_h)}{\log(1-\tilde{p})}} \\
&= \exp\left(\log\left(1-\tilde{p}\right)\frac{\log\left(\delta_h\right)}{\log\left(1-\tilde{p}\right)}\right) \\
&= \delta_h
\end{aligned}$$

which means

$$\mathbb{P}_{\mathcal{H}_\tau}\left(h^\star \in \mathcal{H}_\tau\right) \geq 1 - \delta_h.$$

□

We now use Theorem B.1 to obtain the sample complexity.

Lemma B.6. *Let $\varepsilon \in (0,1)$, and let $\tau \in \mathbb{N}$ and $\mathcal{H}_\tau$ such that $h^\star \in \mathcal{H}_\tau$. Then for any interpolating model $h_\mathcal{S} \in \mathcal{H}_\tau$, and $N \geq \frac{\log(\tau/\delta_\mathcal{S})}{\varepsilon}$*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_\mathcal{D}\left(h_\mathcal{S}\right) \leq \varepsilon\right) \geq 1 - \delta_\mathcal{S}\,,$$

Proof. If $h^\star \in \mathcal{H}_\tau$ then $\mathcal{H}_\tau$ is realizable, so we get from Theorem B.1 with

$$N \geq \frac{\log\left(\frac{|\mathcal{H}_\tau|}{\delta_\mathcal{S}}\right)}{\varepsilon} = \frac{\log\left(\frac{\tau}{\delta_\mathcal{S}}\right)}{\varepsilon}$$

we get the lemma. □

We now wish to upper bound the sample complexity.

Lemma B.7. *For $\tau = \left\lceil \frac{\log(\delta_h)}{\log(1-\tilde{p})} \right\rceil$, under the assumption that $\tilde{p} < \frac{1}{2}$, it holds that*

$$\log\left(\frac{\tau}{\delta_\mathcal{S}}\right) \leq \log\left(\frac{1}{\tilde{p}}\right) + \log\left(\frac{1}{\delta_\mathcal{S}}\right) + \log\left(\log\left(\frac{1}{\delta_h}\right)\right) + \frac{\tilde{p}}{\log\left(\frac{1}{\delta_h}\right)}$$

Proof. First, we bound

$$\begin{aligned}\log\left(\frac{\tau}{\delta_\mathcal{S}}\right) &= \log\left(\frac{\left\lceil \frac{\log(\delta_h)}{\log(1-\tilde{p})} \right\rceil}{\delta_\mathcal{S}}\right) \\ &\leq \log\left(\frac{\frac{\log(\delta_h)}{\log(1-\tilde{p})} + 1}{\delta_\mathcal{S}}\right),\end{aligned}$$

which, after some simplification, becomes

$$\log\left(\frac{\tau}{\delta_\mathcal{S}}\right) \leq \log\left(\frac{\log\left(\delta_h\right) + \log\left(1-\tilde{p}\right)}{\delta_\mathcal{S}\log\left(1-\tilde{p}\right)}\right). \tag{14}$$

We now recall the Taylor expansion of $\log\left(c+x\right)$ around $x=0$ for some $c>0$,

$$\log\left(c+x\right) = \log\left(c\right) + \frac{x}{c} - \frac{x^2}{2c^2} + O\left(x^3\right), \tag{15}$$

plugging $c=1$ and $x \leftarrow -x$ into (15) we get the following bounds for any $x \in \left[0, \frac{1}{2}\right]$:

$$-x - x^2 \leq \log\left(1-x\right) \leq -x\,. \tag{16}$$

Combining (16) with (14) we get

$$\log\left(\frac{\log\left(\delta_h\right) + \log\left(1-\tilde{p}\right)}{\delta_\mathcal{S}\log\left(1-\tilde{p}\right)}\right) \leq \log\left(\frac{\log\left(\delta_h\right) - \tilde{p}}{\delta_\mathcal{S}\left(-\tilde{p} - p^{\star 2}\right)}\right),$$

which can be written as

$$\log\left(\frac{\log\left(\delta_h\right) + \log\left(1-\tilde{p}\right)}{\delta_\mathcal{S}\log\left(1-\tilde{p}\right)}\right) \leq \log\left(\frac{1}{\delta_\mathcal{S}\left(\tilde{p} + p^{\star 2}\right)}\right) + \log\left(\tilde{p} + \log\left(\frac{1}{\delta_h}\right)\right). \tag{17}$$

We now recall that $\log(c+x)$ is a concave function, and therefore its graph in any x is below the graph of its tangent in $x=0$, that is, from (15),

$$\log(c+x) \leq \log(c) + \left(\frac{d}{dx}\left(\log(c+x)\right)|_{x=0}\right) x = \log(c) + \frac{x}{c}\,. \tag{18}$$

Setting $c = \log\left(\frac{1}{\delta_h}\right) \in \left[0, \frac{1}{2}\right]$ and $x = \tilde{p}$ into (18), we obtain

$$\log\left(\tilde{p} + \log\left(\frac{1}{\delta_h}\right)\right) \leq \log\left(\log\left(\frac{1}{\delta_h}\right)\right) + \frac{\tilde{p}}{\log\left(\frac{1}{\delta_h}\right)}\,. \tag{19}$$

Note that

$$\log\left(\frac{1}{\tilde{p} + p^{\star 2}}\right) \leq \log\left(\frac{1}{\tilde{p}}\right)\,, \tag{20}$$

and therefore combining (20), (19) and (17) together with 14 we finish the proof:

$$\begin{aligned}
\log\left(\frac{\tau}{\delta_{\mathcal{S}}}\right) &\leq \log\left(\frac{\log(\delta_h) + \log(1-\tilde{p})}{\delta_{\mathcal{S}} \log(1-\tilde{p})}\right) \\
&\leq \log\left(\frac{1}{\delta_{\mathcal{S}}\left(\tilde{p} + p^{\star 2}\right)}\right) + \log\left(\tilde{p} + \log\left(\frac{1}{\delta_h}\right)\right) \\
&\leq \log\left(\frac{1}{\delta_{\mathcal{S}}}\right) + \log\left(\frac{1}{\tilde{p}}\right) + \log\left(\log\left(\frac{1}{\delta_h}\right)\right) + \frac{\tilde{p}}{\log\left(\frac{1}{\delta_h}\right)}\,.
\end{aligned}$$

□

We are now ready to prove Theorem B.2:

Proof. (of Theorem B.2) Note that $h \sim \mathcal{P}_{\mathcal{S}}$ is interpolating by definition, and therefore, from Lemma B.6, we have that for any $\tau \in \mathbb{N}$ it holds that

$$\forall\, \varepsilon \geq \frac{\log\left(\frac{\tau}{\delta_{\mathcal{S}}}\right)}{N} \quad \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon | h^\star \in \mathcal{H}_\tau\right) = \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}\left(\bar{h}(S)\right) \leq \varepsilon\right) \geq 1 - \delta_{\mathcal{S}}\,, \tag{21}$$

Where $\bar{h}(S) \in \mathcal{H}_\tau$ is some arbitrary interpolate, since from Lemma B.6 this probability is equal for any interpolator. For $\tau = \left\lceil \frac{\log(\delta_h)}{\log(1-\tilde{p})} \right\rceil$ we similarly get

$$\begin{aligned}
&\mathbb{P}_{S\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon\right) = \\
&\quad [\text{Total Probability}] = \mathbb{P}_{S\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon | h^\star \in \mathcal{H}_\tau\right) \mathbb{P}_{h\sim\mathcal{P}_{\mathcal{S}}}\left(h^\star \in \mathcal{H}_\tau\right) \\
&\qquad + \mathbb{P}_{S\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon | h^\star \notin \mathcal{H}_\tau\right) \mathbb{P}_{h\sim\mathcal{P}_{\mathcal{S}}}\left(h^\star \notin \mathcal{H}_\tau\right) \\
&\quad [\text{Probability is non-negative}] \geq \mathbb{P}_{S\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon | h^\star \in \mathcal{H}_\tau\right) \mathbb{P}_{h\sim\mathcal{P}_{\mathcal{S}}}\left(h^\star \in \mathcal{H}_\tau\right) \\
&\quad [\text{Lemma B.5}] \geq \mathbb{P}_{S\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon | h^\star \in \mathcal{H}_\tau\right) (1-\delta_h) \\
&\quad [(21)] \geq (1-\delta_{\mathcal{S}})(1-\delta_h)\,.
\end{aligned}$$

And as before, this is true for any $N \geq \frac{\log\left(\frac{\tau}{\delta_{\mathcal{S}}}\right)}{\varepsilon}$. specifically, by Lemma B.7, it is true for N greater than

$$\frac{\log\left(\frac{1}{\delta_{\mathcal{S}}}\right) + \log\left(\frac{1}{\tilde{p}}\right) + \log\left(\log\left(\frac{1}{\delta_h}\right)\right) + \frac{\tilde{p}}{\log\left(\frac{1}{\delta_h}\right)}}{\varepsilon}\,,$$

the sample complexity. Now we can choose $\delta_{\mathcal{S}}, \delta_h = \frac{\delta}{2}$, and then for any N greater than

$$\frac{\log\left(\frac{2}{\delta}\right) + \log\left(\frac{1}{\tilde{p}}\right) + \log\left(\log\left(\frac{2}{\delta}\right)\right) + \frac{\tilde{p}}{\log\left(\frac{2}{\delta}\right)}}{\varepsilon},$$

we have that

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon\right) &\geq \left(1 - \frac{\delta}{2}\right)^2 \\ &= 1 - \delta_{\mathcal{S}} + \frac{\delta^2}{4} \\ &\geq 1 - \delta\,.\end{aligned}$$

Note that for any $\delta \in \left(0, \frac{1}{5}\right)$ we have that

$$\log\left(\log\left(\frac{2}{\delta}\right)\right) + \frac{\tilde{p}}{\log\left(\frac{2}{\delta}\right)} \leq 2\log\left(\frac{2}{\delta}\right),$$

bounding the sample complexity for simplification with

$$\left(-\log\left(\tilde{p}\right) + 3\log\left(\frac{2}{\delta}\right)\right)\frac{1}{\varepsilon}\,.$$

□

B.3. Proof for the Volume of Generalizing Interpolators (Corollary 3.3)

We first restate Corollary 3.3, and then we will give its formal proof:

Corollary B.8 (volume of generalizing interpolators restated)**.** *For ε, δ as above, and any N larger than*

$$\frac{-\log\left(\tilde{p}\right)+6\log\left(\frac{2}{\delta}\right)}{\varepsilon}\,,$$

the sample complexity, we have that

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathbb{P}_{h\sim\mathcal{P}_\mathcal{S}}\left(\mathcal{L}_\mathcal{D}\left(h\right)\geq\varepsilon\right)\geq\delta\right)\leq\delta\,.$$

Proof. Using Lemma 3.2, for any

$$N\geq\frac{-\log\left(\tilde{p}\right)+6\log\left(\frac{2}{\delta}\right)}{\varepsilon}\geq\frac{-\log\left(\tilde{p}\right)+3\log\left(\frac{2}{\delta^2}\right)}{\varepsilon}$$

we have that

$$\begin{aligned}\delta^2&\geq\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_\mathcal{S}}\left(\mathcal{L}_\mathcal{D}\left(h\right)\geq\varepsilon\right)\\&=\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_\mathcal{S}}\left[\mathbb{I}\left\{\mathcal{L}_\mathcal{D}\left(h\right)\geq\varepsilon\right\}\right]\\&=\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^N}\left[\mathbb{E}_{h\sim\mathcal{P}_\mathcal{S}}\left[\mathbb{I}\left\{\mathcal{L}_\mathcal{D}\left(h\right)\geq\varepsilon\right\}\right]\right]\\&=\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^N}\left[\mathbb{P}_{h\sim\mathcal{P}_\mathcal{S}}\left(\mathcal{L}_\mathcal{D}\left(h\right)\geq\varepsilon\right)\right]\end{aligned}$$

and then we use Markov's inequality

$$\mathbb{P}_\mathcal{S}\left(\mathbb{P}_{h\sim\mathcal{P}_\mathcal{S}}\left(\mathcal{L}_\mathcal{D}\left(h\right)\geq\varepsilon\right)\geq\delta\right)\leq\frac{\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^N}\left[\mathbb{P}_{h\sim\mathcal{P}_\mathcal{S}}\left(\mathcal{L}_\mathcal{D}\left(h\right)\geq\varepsilon\right)\right]}{\delta}\leq\delta\,,$$

That is,

$$\mathbb{P}_\mathcal{S}\left(\mathbb{P}_{h\sim\mathcal{P}_\mathcal{S}}\left(\mathcal{L}_\mathcal{D}\left(h\right)<\varepsilon\right)\geq1-\delta\right)\geq1-\delta\,.$$

□

B.4. Extension to Non-Interpolators

Let $\varepsilon > 0$. Assume that there exists a narrow teacher NN $h^\star$ s.t $\mathcal{L}_\mathcal{D}(h^\star) = \varepsilon^\star > 0$. Denote by $\tilde{p}$ the teacher equivalence probability, and define the posterior distribution

$$\mathcal{P}_\mathcal{S}(h) = \mathcal{P}_\mathcal{S}^\varepsilon(h) = \mathbb{P}\left(h \,|\, \mathcal{L}_\mathcal{S}(h) \leq \varepsilon^\star + \varepsilon\right) .$$

In the G&C formulation, sampling from $\mathcal{P}_\mathcal{S}^\varepsilon$ is equivalent to stopping at the first model satisfying $\mathcal{L}_\mathcal{S}(h) \leq \varepsilon^\star + \varepsilon$.

Theorem B.9 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_N$ be i.i.d random variables with $\mathbb{E}X_i = \mu$ and $0 \leq X_i \leq 1$ a.s. Then for all $t > 0$*

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^N X_i - \mu \geq t\right) \leq \exp\left(-2Nt^2\right)$$

and

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| \geq t\right) \leq 2\exp\left(-2Nt^2\right) .$$

By definition, $\mathcal{L}_\mathcal{S}(h) = \frac{1}{N}\sum_{(\mathbf{x},y)\in\mathcal{S}} \mathbb{I}\{h(\mathbf{x}) \neq y\}$ so using Hoeffding's inequality

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_\mathcal{S}(h^\star) \geq \varepsilon^\star + \varepsilon\right) \leq e^{-2\varepsilon^2 N} .$$

Lemma B.10. *Let $\varepsilon \in \left(0, \frac{1}{2} - \varepsilon^\star\right)$, and let $\tau \in \mathbb{N}$, and $\mathcal{H}_\tau$ s.t $h^\star \in \mathcal{H}_\tau$. Then for any model $\tilde{h} \in \mathcal{H}_\tau$ satisfying $\mathcal{L}_\mathcal{S}\left(\tilde{h}\right) \leq \varepsilon^\star + \varepsilon$*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_\mathcal{D}\left(\tilde{h}\right) \leq \varepsilon^\star + 2\varepsilon, \mathcal{L}_\mathcal{S}(h^\star) \leq \varepsilon^\star + \varepsilon\right) \geq 1 - (2\tau+1)\, e^{-2\varepsilon^2 N} .$$

Proof. Part 1: We will show that

$$\mathcal{L}_\mathcal{D}\left(\tilde{h}\right) - \varepsilon^\star \leq \max_{h\in\mathcal{H}_\tau} |\mathcal{L}_\mathcal{D}(h) - \mathcal{L}_\mathcal{S}(h)| + \varepsilon .$$

Since $h^\star \in \mathcal{H}_\tau$, w.p. $1 - e^{-2\varepsilon^2 N}$ it holds that $\mathcal{L}_\mathcal{S}(h^\star) \leq \varepsilon^\star + \varepsilon$ so there exists such $\tilde{h} \in \mathcal{H}_\tau$. Then

$$\begin{aligned}
\mathcal{L}_\mathcal{D}\left(\tilde{h}\right) - \varepsilon^\star &= \mathcal{L}_\mathcal{D}\left(\tilde{h}\right) - \mathcal{L}_\mathcal{S}\left(\tilde{h}\right) + \mathcal{L}_\mathcal{S}\left(\tilde{h}\right) - \varepsilon^\star \\
&\leq \left(\mathcal{L}_\mathcal{D}\left(\tilde{h}\right) - \mathcal{L}_\mathcal{S}\left(\tilde{h}\right)\right) + \varepsilon^\star + \varepsilon - \varepsilon^\star \\
&\leq \max_{h\in\mathcal{H}_\tau} |\mathcal{L}_\mathcal{D}(h) - \mathcal{L}_\mathcal{S}(h)| + \varepsilon .
\end{aligned}$$

Part 2: Using the union bound and then Hoeffding's inequality

$$\begin{aligned}
\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_\mathcal{D}\left(\tilde{h}\right) - \varepsilon^\star > 2\varepsilon\right) &\leq \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\max_{h\in\mathcal{H}_\tau} |\mathcal{L}_\mathcal{D}(h) - \mathcal{L}_\mathcal{S}(h)| + \varepsilon > 2\varepsilon\right) \\
&= \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\max_{h\in\mathcal{H}_\tau} |\mathcal{L}_\mathcal{D}(h) - \mathcal{L}_\mathcal{S}(h)| > \varepsilon\right) \\
&= \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\exists h \in \mathcal{H}_\tau \,:\, |\mathcal{L}_\mathcal{D}(h) - \mathcal{L}_\mathcal{S}(h)| > \varepsilon\right) \\
&\leq \sum_{h\in\mathcal{H}_\tau} \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(|\mathcal{L}_\mathcal{D}(h) - \mathcal{L}_\mathcal{S}(h)| > \varepsilon\right) \\
&\leq \sum_{h\in\mathcal{H}_\tau} 2\exp\left(-2N\varepsilon^2\right) \\
&= 2\tau e^{-2N\varepsilon^2} .
\end{aligned}$$

Part 3: Combining the probability lower bounds using the union bound

$$\begin{aligned}\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}\left(\tilde{h}\right)\leq\varepsilon^\star+2\varepsilon,\mathcal{L}_{\mathcal{S}}\left(h^\star\right)\leq\varepsilon^\star+\varepsilon\right)&\geq 1-2\tau e^{-2\varepsilon^2N}-e^{-2\varepsilon^2N}\\&=1-\left(2\tau+1\right)e^{-2\varepsilon^2N}.\end{aligned}$$

□

With Lemma B.5 and Lemma B.10 we deduce the following.

Theorem B.11 (Generalization of non-interpolating student)**.** *Let* $\varepsilon\in\left(0,\frac{1}{2}-\varepsilon^\star\right)$ *and* $\delta\in(0,1)$*. Taking* $\tau=\left\lceil\frac{\log\left(\frac{\delta}{2}\right)}{\log(1-\tilde{p})}\right\rceil$ *and* $N\geq\frac{1}{2\varepsilon^2}\log\left(2\cdot\frac{2\tau+1}{\delta}\right)$ *we get*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}\left(h\right)\leq\varepsilon^\star+2\varepsilon\right)\geq 1-\delta.$$

Proof. Recall that sampling $h\sim\mathcal{P}_{\mathcal{S}}$ is equivalent to sampling an hypothesis with the Guess&Check algorithm. Using $h^\star\in\mathcal{H}_\tau$ to denote the event that there is a teacher equivalent hypothesis sampled within the first τ samples,

$$\begin{aligned}\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}\left(h\right)\leq\varepsilon^\star+2\varepsilon\right)&\geq\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}\left(h\right)\leq\varepsilon^\star+2\varepsilon,\mathcal{L}_{\mathcal{S}}\left(h^\star\right)\leq\varepsilon^\star+\varepsilon,h^\star\in\mathcal{H}_\tau\right)\\&\geq 1-\frac{\delta}{2}-\left(2\tau+1\right)e^{-2\varepsilon^2N}\\&=1-\frac{\delta}{2}-\frac{\delta}{2}\\&=1-\delta.\end{aligned}$$

□

We can simplify the lower bound on N by explicitly writing τ in terms of $\tilde{p}$ in a manner similar to Lemma B.7. Specifically, assuming $\tilde{p}<\frac{1}{2}$

$$\begin{aligned}\log\left(2\tau+1\right)&\leq\log\left(\frac{2\log\left(\frac{\delta}{2}\right)}{\log\left(1-\tilde{p}\right)}+3\right)\\&\leq\log\left(\frac{1}{\tilde{p}+\tilde{p}^2}\right)+\log\left(3\tilde{p}+2\log\left(\frac{2}{\delta}\right)\right)\\&\leq\log\left(\frac{1}{\tilde{p}}\right)+\log\left(2\log\left(\frac{2}{\delta}\right)\right)+\frac{3\tilde{p}}{2\log\left(\frac{2}{\delta}\right)}\end{aligned}$$

and for $\delta<\frac{1}{5}$

$$\log\left(2\tau+1\right)\leq-\log\left(\tilde{p}\right)+2\log\left(\frac{2}{\delta}\right)$$

so the sample complexity is

$$\frac{1}{2\varepsilon^2}\log\left(2\cdot\frac{2\tau+1}{\delta}\right)\leq\frac{\log\left(2\tau+1\right)+\log\left(\frac{2}{\delta}\right)}{2\varepsilon^2}\leq\frac{-\log\left(\tilde{p}\right)+3\log\left(\frac{2}{\delta}\right)}{2\varepsilon^2}$$

Remark B.12. Notice that the sample complexity is quadratically dependent on the population error, as opposed to the linear dependence in the realizable case, i.e. when the teacher is assumed to be a perfect interpolator.

B.5. Relationship to PAC Bayes

As stated in Section 3, a similar result to Lemma 3.2 could be derived with PAC-Bayes analysis (McAllester, 1999), which typically focuses on the *expected* population error $\mathbb{E}_{h\sim\mathcal{P}_\mathcal{S}}[\mathcal{L}_\mathcal{D}(h)]$ of a sample from the posterior. A standard PAC-Bayes bound (Langford & Seeger, 2002; McAllester, 2003) yields the following result:

Proposition B.13. *Let $\varepsilon > 0$ and $\delta \in (0,1)$. Then with sample complexity*

$$O\left(\frac{-\log(\tilde{p}) + \log\left(\frac{1}{\delta}\right)}{\varepsilon}\right)$$

we have that

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\,\mathbb{E}_{h\sim\mathcal{P}_\mathcal{S}}[\mathcal{L}_\mathcal{D}(h)] < \varepsilon\,\right) \geq 1-\delta.$$

We can naively use this bound together with Markov's inequality to get a single sample bound.

Corollary B.14. *Let $\varepsilon > 0$ and $\delta \in (0,1)$. Then with sample complexity*

$$O\left(\frac{-\log(\tilde{p}) + \log\left(\frac{1}{\delta}\right)}{\varepsilon\delta}\right)$$

we have that

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\,\mathbb{P}_{h\sim\mathcal{P}_\mathcal{S}}\left(\mathcal{L}_\mathcal{D}(h) < \varepsilon\right) \geq 1-\delta\,\right) \geq 1-\delta.$$

Proof. Using Prop. B.13,

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\,\mathbb{E}_{h\sim\mathcal{P}_\mathcal{S}}[\mathcal{L}_\mathcal{D}(h)] < \varepsilon\delta\right) \;\geq 1-\delta\,.$$

Then using Markov's inequality, w.p. $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^N$,

$$\mathbb{P}_{h\sim\mathcal{P}_\mathcal{S}}\left(\mathcal{L}_\mathcal{D}(h) \geq \varepsilon\right) \leq \frac{\mathbb{E}_{h\sim\mathcal{P}_\mathcal{S}}[\mathcal{L}_\mathcal{D}(h)]}{\varepsilon} < \frac{\varepsilon\delta}{\varepsilon} = \delta\,.$$

By using the complement probability we get the result. □

Note that the sample complexity in Corollary B.14 is larger than the one in Corollary 3.3. Instead of additive $\log\left(\frac{1}{\delta}\right)$ factors in Corollary 3.3, here we have a multiplicative $\frac{1}{\delta}$ factor.

B.6. Proving a Refined Version of Lemma 3.2

Motivation. Note that two sources of randomness affect the sample complexity in Lemma 3.2: the random sampling of hypotheses from the prior $\mathcal{P}$ in G&C algorithm and the random sampling of the dataset from $\mathcal{D}^N$. To understand how each of these sources affects the obtained random complexity, we derive a refined generalization bound:

Theorem B.15 (*G&C Generalization, refined*). *Let $\varepsilon, \delta_{\mathcal{S}} \in (0,1)$, and $\delta_h \in \left(0, \frac{1}{5}\right)$, and assume that $\tilde{p} < \frac{1}{2}$. For any N larger than*

$$\left(-\log\left(\tilde{p}\right) + \log\left(\frac{1}{\delta_{\mathcal{S}}}\right) + 2\log\left(\log\left(\frac{1}{\delta_h}\right)\right)\right)\frac{1}{\varepsilon},$$

the sample complexity, we have that

$$\mathbb{P}_{\mathcal{A}_{\mathcal{P}}}\left(\mathbb{P}_{S\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(\mathcal{A}_{\mathcal{P}}(S)) < \varepsilon\right) \geq 1-\delta_{\mathcal{S}}\right) \geq 1-\delta_h\,.$$

Remark B.16. Note that the randomness in $\mathcal{A}_{\mathcal{P}}$ is only from the sampling of the sequence $(h_t)_{t=1}^{\infty}$, and not from the dependence of $\mathcal{A}_{\mathcal{P}}\left(\mathcal{S}\right)$ on $\mathcal{S}$.

Discussion. Theorem B.15 guarantees generalization with probability at least $1-\delta_h$ over the hypothesis sampling and probability $\delta_{\mathcal{S}}$ over the data sampling. This separation between δ_h and $\delta_{\mathcal{S}}$ highlights how both sources of randomness play a role in generalization. Interestingly, the sample complexity term N exhibits a logarithmic dependence on $\delta_{\mathcal{S}}$ and only a doubly logarithmic dependence on δ_h. Thus, for any $\delta_{\mathcal{S}}$ and ε, the probability of not sampling a PAC interpolator decays extremely fast (doubly exponential in N):

$$\delta_h = \exp\left(-\exp\left(\frac{1}{2}\left(\varepsilon N + \log\left(\tilde{p}\right) + \log\left(\delta_{\mathcal{S}}\right)\right)\right)\right).$$

In other words, the sampled interpolator is *'typically PAC'*, i.e., PAC with overwhelmingly high probability over the sampled interpolator sequence $(h_t)_{t=1}^{\infty}$.

Proof. Recall that the hypothesis chosen by G&C, $h \sim \mathcal{P}_{\mathcal{S}}$, is interpolating by definition. Set $\tau = \left\lceil \frac{\log(\delta_h)}{\log(1-\tilde{p})} \right\rceil$, then

$$\begin{aligned}
&\mathbb{P}_{\mathcal{A}_{\mathcal{P}}}\left(\mathbb{P}_{S\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(\mathcal{A}_{\mathcal{P}}(S)) \leq \varepsilon\right) \geq 1-\delta_{\mathcal{S}}\right) \\
[\text{Total probability}] =&\mathbb{P}_{\mathcal{A}_{\mathcal{P}}}\left(\mathbb{P}_{S\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(\mathcal{A}_{\mathcal{P}}(S)) \leq \varepsilon\right) \geq 1-\delta_{\mathcal{S}} | h^\star \in \mathcal{H}_\tau\right)\mathbb{P}_{\mathcal{A}_{\mathcal{P}}}\left(h^\star \in \mathcal{H}_\tau\right) \\
&+\mathbb{P}_{\mathcal{A}_{\mathcal{P}}}\left(\mathbb{P}_{S\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(\mathcal{A}_{\mathcal{P}}(S)) \leq \varepsilon\right) \geq 1-\delta_{\mathcal{S}} | h^\star \notin \mathcal{H}_\tau\right)\mathbb{P}_{\mathcal{A}_{\mathcal{P}}}\left(h^\star \notin \mathcal{H}_\tau\right) \\
[\text{Probability is non-negative}] \geq&\mathbb{P}_{\mathcal{A}_{\mathcal{P}}}\left(\mathbb{P}_{S\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(\mathcal{A}_{\mathcal{P}}(S)) \leq \varepsilon\right) \geq 1-\delta_{\mathcal{S}} | h^\star \in \mathcal{H}_\tau\right)\mathbb{P}_{\mathcal{A}_{\mathcal{P}}}\left(h^\star \in \mathcal{H}_\tau\right) \\
[\text{Lemma B.5}] \geq&\mathbb{P}_{\mathcal{A}_{\mathcal{P}}}\left(\mathbb{P}_{S\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(\mathcal{A}_{\mathcal{P}}(S)) \leq \varepsilon\right) \geq 1-\delta_{\mathcal{S}} | h^\star \in \mathcal{H}_\tau\right)\left(1-\delta_h\right) \\
[\text{Lemma B.6}] =&1-\delta_h\,.
\end{aligned}$$

This holds for any $N \geq \frac{\log\left(\frac{\tau}{\delta_{\mathcal{S}}}\right)}{\varepsilon}$. Specifically, by Lemma B.7, it holds for any N larger than

$$\frac{\log\left(\frac{1}{\tilde{p}}\right) + \log\left(\frac{1}{\delta_{\mathcal{S}}}\right) + \log\left(\log\left(\frac{1}{\delta_h}\right)\right)}{\varepsilon} + \frac{\tilde{p}}{\varepsilon\log\left(\frac{1}{\delta_h}\right)},$$

Note that for any $\delta_h \in \left(0, \frac{1}{5}\right)$ it holds that

$$\frac{\tilde{p}}{\log\left(\frac{1}{\delta_h}\right)} \leq \log\left(\log\left(\frac{1}{\delta_h}\right)\right),$$

bounding the sample complexity by

$$\left(\log\left(\frac{1}{\delta_{\mathcal{S}}}\right) + 2\log\left(\log\left(\frac{1}{\delta_h}\right)\right) - \log\left(\tilde{p}\right)\right)\frac{1}{\varepsilon}\,.$$

□

B.7. Proofs Using Nonuniform Learnability

In the next pages, we will show a result that does not use the teacher assumption for the generalization of randomly sampled networks. The result is more general than Lemma 3.2 and B.15, which were both tailored for the teacher assumption. However, the price to pay for relaxing this assumption is that the following result is slightly less tight. Since we do not use the teacher assumption, with some abuse of notation we use $\mathcal{D}$ to denote the joint distribution of feature-label pairs $(\mathbf{x}, y)$.

Instead of the teacher assumption, we rely on the probability of interpolation, defined as:

Definition B.17. For some random hypothesis h from prior $\mathcal{P}$, the interpolation probability is defined as

$$\hat{p}_{\mathcal{S}} \triangleq \mathbb{P}_{h\sim\mathcal{P}}\left(\mathcal{L}_{\mathcal{S}}(h) = 0\right)$$

Theorem B.18 (Generalization of Guess & Check, restated)**.** *Under the assumption that $\hat{p}_{\mathcal{S}} < \frac{1}{2}$ for all $\mathcal{S}$, for any $\delta, \eta \in (0,1)$, and $N \in \mathbb{N}$, we have with probability at least $1-\eta$ over $h \sim \mathcal{P}_{\mathcal{S}}$ that:*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \le \varepsilon\right) \ge 1-\delta\,,$$

where

$$\varepsilon = \varepsilon_{\delta,\eta}(\mathcal{S}) = \frac{\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right) + \log\left(\frac{4}{\delta}\right) + \log\left(\log\left(\frac{2}{\eta}\right)\right) + 2\log\log\left(\frac{\log\left(\frac{2}{\eta}\right)}{\hat{p}_{\mathcal{S}}} + 1\right)}{N}\,. \tag{22}$$

Theorem B.18 proof sketch: We first show that for any sequence of hypotheses h_t

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{S}}(h_t) = 0 \text{ and } \mathcal{L}_{\mathcal{D}}(h_t) > \tilde{\varepsilon}_t\right) \le \delta_t\,,$$

where

$$\tilde{\varepsilon}_t \triangleq \frac{\log 1/\delta_t}{N}\,.$$

Set $\delta_t = \frac{\delta}{4t\log^2(t+1)}$ to obtain

$$\tilde{\varepsilon}_t = \frac{\log t + 2\log\log\left(t+1\right) + \log 4/\delta}{N}\,,$$

and use a union bound, which yields, for any $\delta > 0$ and any sequence $(h_t)_{t\in\mathbb{N}}$ of hypotheses,

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\exists_t\left(\mathcal{L}_{\mathcal{S}}(h_t) = 0 \text{ and } \mathcal{L}_{\mathcal{D}}(h_t) > \tilde{\varepsilon}_t\right)\right) \le \delta\,.$$

Importantly, since this holds for any h_t we can use our h_t sequence from Algorithm 1 and for $t = T$ to get with probability at least $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^N$ that

$$\mathcal{L}_{\mathcal{D}}(\mathcal{A}_{\mathcal{P}}(\mathcal{S})) \le \tilde{\varepsilon}_T\,.$$

Finally, we use the fact that $T \mid \mathcal{S}$ is a geometric random variable with success parameter $\hat{p}_{\mathcal{S}} < \frac{1}{2}$ to obtain

$$\mathbb{P}_{h_t}\left(T > \frac{\log 2/\eta}{\hat{p}_{\mathcal{S}}}\right) \le \eta\,.$$

Taking the complementary of the probability above, combined with the fact that $\tilde{\varepsilon}_T$ is a an increasing function of T concludes the theorem.

For the complete derivation, we proceed with some lemmas before proving Theorem B.18.

Lemma B.19. *For any $\delta > 0$ and any sequence of hypotheses $(h_t)_{t\in\mathbb{N}}$:*

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\exists_t \mathcal{L}_{\mathcal{S}}(h_t) = 0 \textit{ and } \mathcal{L}_{\mathcal{D}}(h_t) > \tilde{\varepsilon}_t\right) \le \delta$$

where

$$\tilde{\varepsilon}_t = \frac{\log t + 2\log\log\left(t+1\right) + \log 4/\delta}{N} \tag{23}$$

Proof. Set $\delta_t = \frac{\delta}{4t\log^2(t+1)}$. We first show that $\sum_t \delta_t < \delta$. Since δ_t is monotonically decreasing

$$\sum_{t=1}^{\infty} \frac{\delta}{4t\log^2(t+1)} \leq \frac{\delta}{4\log^2(2)} + \frac{\delta}{8\log^2(3)} + \int_3^{\infty} \frac{\delta}{4t\log^2(t)}\mathrm{d}t = \frac{\delta}{4}\left(\frac{1}{\log^2(2)} + \frac{0.5}{\log^2(3)} + \frac{1}{\log(3)}\right) \leq \delta\,,$$

where we used the change of variables $u = \log(t)\,, \mathrm{d}u = \frac{\mathrm{d}t}{t}$, to solve the integral

$$\int_3^{\infty} \frac{1}{t\log^2(t)}\mathrm{d}t = \int_{\log(3)}^{\infty} \frac{1}{u^2}\mathrm{d}u = \frac{1}{\log(3)}\,.$$

For each h_t separately, $\mathcal{L}_{\mathcal{D}}(h_t) > \frac{\log 1/\delta_t}{N}$ is a deterministic event so

$$\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{S}}(h_t) = 0 \text{ and } \mathcal{L}_{\mathcal{D}}(h_t) > \frac{\log 1/\delta_t}{N}\right) &= \prod_{n=1}^{N} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left(h_t(\mathbf{x}) = y\right) \\
&= \prod_{n=1}^{N} \left(1 - \mathcal{L}_{\mathcal{D}}(h_t)\right) \\
&= \left(1 - \mathcal{L}_{\mathcal{D}}(h_t)\right)^N \\
&\leq \left(1 - \frac{\log 1/\delta_t}{N}\right)^N \\
&\leq \exp\left(-N\frac{\log 1/\delta_t}{N}\right) \\
&= \delta_t\,.
\end{aligned}$$

Taking a union bound yields the lemma. □

Lemma B.20. *For any $\delta > 0$, and any realization of $(h_t)_{t=1}^{\infty}$, with probability at least $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^N$,*

$$\mathcal{L}_{\mathcal{D}}(h_T) \leq \tilde{\varepsilon}_T\,.$$

where $\tilde{\varepsilon}_t$ is defined in (23) and T is defined in Algorithm 1.

Proof. Lemma B.19 applies to any sequence of hypotheses, and since $\mathcal{S}$ is independent of the sequence h_t, we can also apply it to h_t. Since the Lemma applies to all t, it also applies to any random T, even if it depends on the sample. For the T used by G&C from Algorithm 1, we always have $\mathcal{L}_{\mathcal{S}}(h_t) = 0$, so we get the result from Lemma B.19. □

We now wish to explain the dependence on T in Lemma B.20. We will do so by the following Lemma:

Lemma B.21. *For any η, and any $\mathcal{S}$, under the assumption that $\hat{p}_{\mathcal{S}} < \frac{1}{2}$, we have that $\mathbb{P}_{(h_t)}\left(T > \frac{\log 2/\eta}{\hat{p}_{\mathcal{S}}}\right) \leq \eta$*

Proof. Given $S \sim \mathcal{D}^N$, we observe that T is geometric with parameter $\hat{p}_{\mathcal{S}}$. Using the cumulative distribution function of geometric random variables, we can bound:

$$\mathbb{P}_{(h_t)}\left(T > \frac{\log(2/\eta)}{\hat{p}_{\mathcal{S}}}\right) = (1-\hat{p}_{\mathcal{S}})^{\left\lfloor\frac{\log(2/\eta)}{\hat{p}_{\mathcal{S}}}\right\rfloor} \leq (1-\hat{p}_{\mathcal{S}})^{\frac{\log(2/\eta)}{\hat{p}_{\mathcal{S}}}-1}$$

And now we use a basic property of exponents $(1-x) \leq e^{-x}$, which can be rewritten as $\frac{-1}{\log(1-x)} \leq \frac{1}{x}$, to further bound

$$\begin{aligned}
\mathbb{P}_{(h_t)}\left(T > \frac{\log(2/\eta)}{\hat{p}_{\mathcal{S}}}\right) &\leq (1-\hat{p}_{\mathcal{S}})^{\frac{\log(2/\eta)}{\hat{p}_{\mathcal{S}}}-1} \\
&\leq (1-\hat{p}_{\mathcal{S}})^{\frac{\log\left(\frac{\eta}{2}\right)}{\log(1-\hat{p}_{\mathcal{S}})}-1} \\
&= \exp\left(\left(\frac{\log\left(\frac{\eta}{2}\right)}{\log(1-\hat{p}_{\mathcal{S}})}-1\right)\log(1-\hat{p}_{\mathcal{S}})\right) \\
&\leq \exp\left(\frac{\log\left(\frac{\eta}{2}\right)}{\log(1-\hat{p}_{\mathcal{S}})}\log(1-\hat{p}_{\mathcal{S}}) - \log(1-\hat{p}_{\mathcal{S}})\right) \\
&= \exp\left(\log\left(\frac{\eta}{2}\right) - \log(1-\hat{p}_{\mathcal{S}})\right) \\
&= \frac{\frac{\eta}{2}}{1-\hat{p}_{\mathcal{S}}} \leq \eta
\end{aligned}$$

□

Having proved Lemma B.21 and Lemma B.20 we are ready to prove our main result.

Proof. [of Theorem B.18] From Lemma B.20 we have that

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \tilde{\varepsilon}_T\right) \geq 1-\delta\,. \tag{24}$$

We can now use Lemma B.21 to obtain with probability of at least $1-\eta$ over the sampling of h_t that $T \leq \frac{\log(2/\eta)}{\hat{p}_{\mathcal{S}}}$ whence

$$\begin{aligned}
\tilde{\varepsilon}_T &= \frac{\log T + 2\log\log(T+1) + \log 4/\delta}{N} \\
&\leq \frac{\log\left(\frac{\log(2/\eta)}{\hat{p}_{\mathcal{S}}}\right) + 2\log\log\left(\frac{\log(2/\eta)}{\hat{p}_{\mathcal{S}}}+1\right) + \log 4/\delta}{N} \\
&\leq \frac{\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right) + \log\left(\frac{4}{\delta}\right) + \log\left(\log\left(\frac{2}{\eta}\right)\right) + 2\log\log\left(\frac{\log\left(\frac{2}{\eta}\right)}{\hat{p}_{\mathcal{S}}}+1\right)}{N},
\end{aligned}$$

so with probability of at least $1-\eta$ we have that

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N, h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon_{\eta,\delta}(\mathcal{S})\right) \geq \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \tilde{\varepsilon}_T\right).$$

Finally, from (24) we get

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon_{\eta,\delta}(\mathcal{S})\right) \geq 1-\delta\,.$$

□

Lemma B.22. *Let $x > 1$ and $y > 0$. Then*

$$\log(x+y) \leq \log(x) + \log(1+y)\,.$$

Proof. Let $x, y > 0$.

$$\log(1+x+y) \leq \log(1+x+y+xy) = \log\left((1+x)(1+y)\right) = \log(1+x) + \log(1+y)\,.$$

Now, suppose that $x > 1$ and $y > 0$.

$$\begin{aligned}
\log(x+y) &= \log(1+(x-1)+y) \\
&\leq \log(1+(x-1)) + \log(1+y) \\
&= \log(x) + \log(1+y)\,.
\end{aligned}$$

□

Proposition B.23. *For any $\delta > 0$, with probability at least $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^N$ and $h \sim \mathcal{P}_{\mathcal{S}}$*

$$\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon\,,$$

where

$$\varepsilon = \frac{\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right) + 4\log\left(\frac{8}{\delta}\right) + 2\log\left(\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right)\right)}{N}\,.$$

Proof. Here we take an alternative approach to the one presented in the proof of Theorem B.18. Denote

$$\tilde{\varepsilon}_T = \frac{\log T + 2\log\log\left(T+1\right) + \log 4/\delta}{N}\,.$$

Now note that from the proof of Theorem B.18,

$$T \leq \frac{\log\left(2/\eta\right)}{\hat{p}_{\mathcal{S}}} \Rightarrow \tilde{\varepsilon_T} \leq \varepsilon_{\eta,\delta}(\mathcal{S})\,. \tag{25}$$

We use the law of total probability to write

$$\begin{aligned}
\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon_{\eta,\delta}(\mathcal{S})\right) = \\
&= \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon_{\eta,\delta}(\mathcal{S}) \,\middle|\, T \leq \frac{\log\left(2/\eta\right)}{\hat{p}_{\mathcal{S}}}\right)\mathbb{P}_{h_t}\left(T \leq \frac{\log 2/\eta}{\hat{p}_{\mathcal{S}}}\right) \\
&+ \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon_{\eta,\delta}(\mathcal{S}) \,\middle|\, T > \frac{\log\left(2/\eta\right)}{\hat{p}_{\mathcal{S}}}\right)\mathbb{P}_{h_t}\left(T > \frac{\log 2/\eta}{\hat{p}_{\mathcal{S}}}\right) \\
&\geq \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon_{\eta,\delta}(\mathcal{S}) \,\middle|\, T \leq \frac{\log\left(2/\eta\right)}{\hat{p}_{\mathcal{S}}}\right)\mathbb{P}_{h_t}\left(T \leq \frac{\log 2/\eta}{\hat{p}_{\mathcal{S}}}\right) \\
[(25)] &\geq \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \tilde{\varepsilon_T}\right)\mathbb{P}_{h_t}\left(T \leq \frac{\log 2/\eta}{\hat{p}_{\mathcal{S}}}\right) \\
[\text{Lemma B.20 and Lemma B.21}] &\geq (1-\delta)\,(1-\eta)
\end{aligned}$$

Now we can choose $\delta, \eta \leftarrow \frac{\delta}{2}$, and we obtain

$$\begin{aligned}
\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N}\left(\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon_{\frac{\delta}{2},\frac{\delta}{2}}(\mathcal{S})\right) &\geq \left(1-\frac{\delta}{2}\right)\left(1-\frac{\delta}{2}\right) \\
&= 1-\delta+\frac{\delta^2}{4} \\
&\geq 1-\delta\,.
\end{aligned}$$

Applying Lemma B.22 multiple times we get

$$\begin{aligned}
\varepsilon_{\frac{\delta}{2},\frac{\delta}{2}}(\mathcal{S}) &= \frac{\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right) + \log\left(\frac{8}{\delta}\right) + \log\left(\log\left(\frac{4}{\delta}\right)\right) + 2\log\log\left(\frac{\log\left(\frac{4}{\delta}\right)}{\hat{p}_{\mathcal{S}}} + 1\right)}{N} \\
&\leq \frac{\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right) + 2\log\left(\frac{8}{\delta}\right) + 2\log\left(\log\left(\frac{\log\left(\frac{4}{\delta}\right)+1}{\hat{p}_{\mathcal{S}}}\right)\right)}{N} \\
&= \frac{\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right) + 2\log\left(\frac{8}{\delta}\right) + 2\log\left(\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right) + \log\left(\log\left(\frac{4e}{\delta}\right)\right)\right)}{N} \\
&\leq \frac{\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right) + 2\log\left(\frac{8}{\delta}\right) + 2\log\left(\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right)\right) + \log\left(\log\left(\log\left(\frac{4e}{\delta}\right)\right) + 1\right)}{N} \\
&\leq \frac{\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right) + 4\log\left(\frac{8}{\delta}\right) + 2\log\left(\log\left(\frac{1}{\hat{p}_{\mathcal{S}}}\right)\right)}{N}
\end{aligned}$$

□

C. Proofs for Generalization of Quantized Neural Networks (Section 4)

This section contains the proofs for the results in Section 4, focusing on upper bounding the effective sample complexity C. Specifically,

- In Appendix C.1 we derive the upper bound for vanilla fully connected networks as stated in Theorem 4.3.
- In Appendix C.2 we derive the upper bound for Neuron-Scaled fully-connected networks as stated in Theorem 4.3.
- In Appendix C.3 we derive the upper bound for convolutional neural networks.
- In Appendix C.4 we derive the upper bound for Channel-Scaled convolutional neural networks.

C.1. Vanilla Fully-Connected Neural Networks

Notation Following definition 4.1, for all $l = 1, \ldots, L$, write the weight matrices and bias vectors in block form as

$$\mathbf{W}^{(l)} = \begin{bmatrix} \mathbf{W}_{11}^{(l)} & \mathbf{W}_{12}^{(l)} \\ \mathbf{W}_{21}^{(l)} & \mathbf{W}_{22}^{(l)} \end{bmatrix} \in \mathbb{R}^{d_l \times d_{l-1}}, \mathbf{b}^{(l)} = \begin{bmatrix} \mathbf{b}_1^{(l)} \\ \mathbf{b}_2^{(l)} \end{bmatrix} \in \mathbb{R}^{d_l}, \tag{26}$$

such that

$$\begin{aligned} \mathbf{W}_{11}^{(l)} &\in \mathbb{R}^{d_l^\star \times d_{l-1}^\star}, \\ \mathbf{W}_{12}^{(l)} &\in \mathbb{R}^{d_l^\star \times \left(d_{l-1} - d_{l-1}^\star\right)}, \\ \mathbf{W}_{21}^{(l)} &\in \mathbb{R}^{(d_l - d_l^\star) \times d_{l-1}^\star}, \\ \mathbf{W}_{22}^{(l)} &\in \mathbb{R}^{(d_l - d_l^\star) \times \left(d_{l-1} - d_{l-1}^\star\right)}, \\ \mathbf{b}_1^{(l)} &\in \mathbb{R}^{d_l^\star}, \\ \mathbf{b}_2^{(l)} &\in \mathbb{R}^{d_l - d_l^\star}. \end{aligned}$$

Remark C.1. The blocks have a simple interpretation with reference to Figure 1. $\mathbf{W}_{11}^{(l)}$ represents the blue edges, and $\mathbf{W}_{12}^{(l)}$ represents the orange edges. $\mathbf{W}_{21}^{(l)}, \mathbf{W}_{22}^{(l)}$, both represent gray edges. As for the biases, $\mathbf{b}_1^{(l)}$ corresponds to the bias terms of gray vertices, and $\mathbf{b}_2^{(l)}$ corresponds to the bias terms of white vertices.

Definition C.2. Define the *coordinate-projection* operator $\pi_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l^\star}$ for $d_l \geq d_l^\star$ as

$$\pi_l\left((x_1, \ldots, x_{d_l})^\top\right) = \left(x_1, \ldots, x_{d_l^\star}\right)^\top.$$

Notice that this projection commutes with the component-wise activation function σ. In order to keep the proofs focused, we restate and prove each result in Theorem 4.3 separately. We first restate only the first part of Theorem 4.3.

Theorem C.3. *For any activation function such that $\sigma(0) = 0$, depth L, Q-quantized teacher with widths D^*, student with widths $D > D^*$ and prior $\mathcal{P}$ uniform over Q-quantized parameterizations, we have for Vanilla Fully Connected Networks that:*

$$\tilde{C} \leq \hat{C}^{\mathrm{FC}} \triangleq \left(\sum_{l=1}^{L} \left(d_l^\star d_{l-1} + d_l^\star\right)\right) \log Q. \tag{27}$$

where we defined $d_0^\star \triangleq d_0$. And by Lemma 3.2, $N = (\tilde{C} + 3\log 2/\delta)/\varepsilon$ samples are enough to ensure that for posterior sampling (i.e. G&C), $\mathcal{L}(\mathcal{A}_{\mathcal{P}}(\mathcal{S})) \leq \varepsilon$ with probability $1 - \delta$ over $\mathcal{S} \sim \mathcal{D}^N$ and the sampling.

Proof. The outline of our proof is as follows: we first show a sufficient condition on the parameters of the student to ensure it will be TE, and then count the number of parameter configurations for which this condition holds. This yields a lower bound on the number of TE models, and since the hypothesis class of quantized networks with fixed widths and depths is finite, we are able to calculate the probability of a sampled model to be TE.

We now begin our proof. For all $l = 1, \ldots, L$, recall that

$$\mathbf{W}^{\star(l)} \in \mathbb{R}^{d_l^\star \times d_{l-1}^\star}, \mathbf{b}^{\star(l)} \in \mathbb{R}^{d_l^\star}$$

are the weights and biases of $h_{\boldsymbol{\theta}^\star}$ respectively. Define

$$\mathcal{E} = \left\{ h_{\boldsymbol{\theta}} \in \mathcal{H}_D^{\mathrm{FC}} \middle| \forall l = 1, \ldots, L \; \mathbf{W}_{11}^{(l)} = \mathbf{W}^{\star(l)}, \; \mathbf{W}_{12}^{(l)} = \mathbf{0}_{d_l^\star \times (d_{l-1} - d_{l-1}^\star)}, \; \mathbf{b}_1^{(l)} = \mathbf{b}^{\star(l)} \right\} .$$

We claim that any $h_{\boldsymbol{\theta}} \in \mathcal{E}$ is TE. To prove this, we show by induction over the layer l, that for all $\mathbf{x} \in \mathbb{R}^{d_0}$,

$$\pi_l \left(f_D^{(l)} (\mathbf{x}) \right) = f_{D^\star}^{(l)} (\mathbf{x}) .$$

Let $h_{\boldsymbol{\theta}} \in \mathcal{E}$. Begin from the base case, $l = 1$. Since $d_0 = d_0^\star$,

$$\mathbf{W}^{(1)} = \begin{bmatrix} \mathbf{W}_{11}^{(1)} \\ \mathbf{W}_{21}^{(1)} \end{bmatrix}$$

so using the notation from Def. 4.1 we find that

$$\mathbf{W}^{(1)} f_D^{(0)} (\mathbf{x}) + \mathbf{b}^{(1)} = \mathbf{W}^{(1)} \mathbf{x} + \mathbf{b}^{(1)} = \begin{bmatrix} \mathbf{W}_{11}^{(1)} \\ \mathbf{W}_{21}^{(1)} \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{b}_1^{(1)} \\ \mathbf{b}_2^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{W}_{11}^{(1)} \mathbf{x} + \mathbf{b}_1^{(1)} \\ \mathbf{W}_{21}^{(1)} \mathbf{x} + \mathbf{b}_2^{(1)} \end{bmatrix}$$

and

$$f_D^{(1)} (\mathbf{x}) = \sigma \left(\begin{bmatrix} \mathbf{W}_{11}^{(1)} \mathbf{x} + \mathbf{b}_1^{(1)} \\ \mathbf{W}_{21}^{(1)} \mathbf{x} + \mathbf{b}_2^{(1)} \end{bmatrix} \right) = \begin{bmatrix} \sigma \left(\mathbf{W}_{11}^{(1)} \mathbf{x} + \mathbf{b}_1^{(1)} \right) \\ \sigma \left(\mathbf{W}_{21}^{(1)} \mathbf{x} + \mathbf{b}_2^{(1)} \right) \end{bmatrix} .$$

Therefore, using the definition of π_1, we get

$$\pi_1 \left(f_D^{(1)} (\mathbf{x}) \right) = \sigma \left(\mathbf{W}_{11}^{(1)} \mathbf{x} + \mathbf{b}_1^{(1)} \right) . \tag{28}$$

Since $h_{\boldsymbol{\theta}} \in \mathcal{E}$, we have that $\mathbf{W}_{11}^{(1)} = \mathbf{W}^{\star(1)}$ and $\mathbf{b}_1^{(1)} = \mathbf{b}^{\star(1)}$, so from (28) the coordinate projection is

$$\pi_1 \left(f_D^{(1)} (\mathbf{x}) \right) = \sigma \left(\mathbf{W}^{\star(1)} \mathbf{x} + \mathbf{b}^{\star(1)} \right) = f_{D^\star}^{(1)} (\mathbf{x}) .$$

Next, assume that $\pi_{l-1} \left(f_D^{(l-1)} (\mathbf{x}) \right) = f_{D^\star}^{(l-1)} (\mathbf{x})$ for some $l \leq L - 1$. Since $h_{\boldsymbol{\theta}} \in \mathcal{E}$, for any $l \in [L]$ we have that $\mathbf{W}_{12}^{(l)} = \mathbf{0}$. Therefore, following a similar argument we get

$$\begin{aligned} \pi_l \left(\mathbf{W}^{(l)} f_D^{(l-1)} (\mathbf{x}) + \mathbf{b}^{(l)} \right) &= \pi_l \left(\begin{bmatrix} \mathbf{W}_{11}^{(l)} & \mathbf{W}_{12}^{(l)} \\ \mathbf{W}_{21}^{(l)} & \mathbf{W}_{22}^{(l)} \end{bmatrix} f_D^{(l-1)} (\mathbf{x}) + \begin{bmatrix} \mathbf{b}_1^{(l)} \\ \mathbf{b}_2^{(l)} \end{bmatrix} \right) \\ &= \begin{bmatrix} \mathbf{W}_{11}^{(l)} & \mathbf{0} \end{bmatrix} f_D^{(l-1)} (\mathbf{x}) + \mathbf{b}_1^{(l)} \\ &= \begin{bmatrix} \mathbf{W}^{\star(l)} & \mathbf{0} \end{bmatrix} f_D^{(l-1)} (\mathbf{x}) + \mathbf{b}^{\star(l)} \\ &= \mathbf{W}^{\star(l)} \pi_{l-1} \left(f_D^{(l-1)} (\mathbf{x}) \right) + \mathbf{b}^{\star(l)} \\ &= \mathbf{W}^{\star(l)} f_{D^\star}^{(l-1)} (\mathbf{x}) + \mathbf{b}^{\star(l)} , \end{aligned}$$

that is

$$\pi_l \left(\mathbf{W}^{(l)} f_D^{(l-1)} (\mathbf{x}) + \mathbf{b}^{(l)} \right) = \mathbf{W}^{\star(l)} f_{D^\star}^{(l-1)} (\mathbf{x}) + \mathbf{b}^{\star(l)} . \tag{29}$$

Using the commutativity between σ and π_l, we have

$$\begin{aligned}\pi_l\left(f_D^{(l)}(\mathbf{x})\right) &= \pi_l\left(\sigma\left(\mathbf{W}^{(l)} f_D^{(l-1)}(\mathbf{x}) + \mathbf{b}^{(l)}\right)\right) \\ &= \sigma\left(\pi_l\left(\mathbf{W}^{(l)} f_D^{(l-1)}(\mathbf{x}) + \mathbf{b}^{(l)}\right)\right) \\ &\overset{(29)}{=} \sigma\left(\mathbf{W}^{\star(l)} f_{D^\star}^{(l-1)}(\mathbf{x}) + \mathbf{b}^{\star(l)}\right) \\ &= f_{D^\star}^{(l)}(\mathbf{x})\ .\end{aligned}$$

For the last layer, $l = L$, the proof is identical, except for the application of the activation function σ at the end, so an analogue to (29) is enough. Since we assume that $h_{\boldsymbol{\theta}}$ and $h_{\boldsymbol{\theta}^\star}$ have the same output dimension d_L, this proves that for all $\mathbf{x} \in \mathbb{R}^{d_0}$

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \pi_L\left(h_{\boldsymbol{\theta}}(\mathbf{x})\right) = h_{\boldsymbol{\theta}^\star}(\mathbf{x})\ .$$

That is, $\mathcal{E} \subseteq \{h_{\boldsymbol{\theta}} | h_{\boldsymbol{\theta}} \equiv h_{\boldsymbol{\theta}^\star}\}$ and therefore

$$\mathbb{P}(\mathcal{E}) \leq \mathbb{P}\left(h_{\boldsymbol{\theta}} \equiv h_{\boldsymbol{\theta}^\star}\right)\ .$$

Finally, to calculate the probability to sample $h_{\boldsymbol{\theta}}$ in $\mathcal{E}$ we count the number of constrained parameters - parameters which are either determined by $h_{\boldsymbol{\theta}^\star}$ or are 0 in $\mathcal{E}$. Looking at the dimensions of $\mathbf{W}_{11}^{(l)}$, $\mathbf{W}_{12}^{(l)}$ and $\mathbf{b}_1^{(l)}$, we deduce that there are exactly

$$\mathcal{M} = \sum_{l=1}^{L} d_l^\star \cdot d_{l-1} + d_l^\star = \sum_{l=1}^{L} d_l^\star\left(d_{l-1} + 1\right)$$

such constrained parameters, and denote

$$\hat{C}^{\mathrm{FC}} \triangleq \left(\sum_{l=1}^{L}\left(d_l^\star d_{l-1} + d_l^\star\right)\right)\log Q = \mathcal{M}\log Q\,.$$

Under the uniform prior over parameters $\mathcal{P}$,

$$\tilde{p} \geq \mathbb{P}(\mathcal{E}) = Q^{-\mathcal{M}}\,,$$

so

$$\tilde{C} = -\log(\tilde{p}) \leq \mathcal{M}\log Q = \hat{C}^{\mathrm{FC}}$$

□

C.2. Neuron-Scaled Fully-Connected Neural Networks

We now wish to improve our result, introducing some assumptions on the architecture. We first need to define a new class of architectures.

Definition C.4 (Scaled-neuron FC restated)**.** For a depth L, widths $D = (d_1, \ldots, d_L)$, and activation function $\sigma : \mathbb{R} \to \mathbb{R}$, a scaled neuron fully connected neural network is a mapping $\boldsymbol{\theta} \mapsto h_{\boldsymbol{\theta}}^{FC}$ from parameters

$$\boldsymbol{\theta} = \left(\left\{ \mathbf{W}^{(l)} \right\}_{l=1}^{L}, \left\{ \mathbf{b}^{(l)} \right\}_{l=1}^{L}, \left\{ \boldsymbol{\gamma}^{(l)} \right\}_{l=1}^{L} \right) ,$$

where

$$\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}, \mathbf{b}^{(l)} \in \mathbb{R}^{d_l}, \boldsymbol{\gamma}^{(l)} \in \mathbb{R}^{d_l} ,$$

defined recursively, starting with $f^{(0)}(\mathbf{x}) = \mathbf{x}$, then

$$\forall l \in [L-1] \;\; f^{(l)}(\mathbf{x}) = \sigma\left(\boldsymbol{\gamma}^{(l)} \odot \left(\mathbf{W}^{(l)} f^{(l-1)}(\mathbf{x}) \right) + \mathbf{b}^{(l)} \right)$$
$$h_{\boldsymbol{\theta}}^{FC}(\mathbf{x}) = \operatorname{sign}\left(\mathbf{W}^{(L)} f^{(L-1)}(\mathbf{x}) + \mathbf{b}^{(L)} \right) .$$

The total parameter count is $M(D) = \sum_{l=1}^{L} d_l(d_{l-1} + 2)$. We denote the class of all scaled neuron fully connected neural network as $\mathcal{H}_D^{\mathrm{SFC}}$.

Remark C.5. We use the notation $f_D^{(l)}(\mathbf{x})$ for both $f_{\mathrm{FC}}^{(l)}$ and $f_{\mathrm{SFC}}^{(l)}$ in a fully connected network with hidden dimensions D, as they can be inferred from context.

Theorem C.6. *For any activation function such that $\sigma(0) = 0$, depth L, Q-quantized teacher with widths D^*, student with widths $D > D^*$ and prior $\mathcal{P}$ uniform over Q-quantized parameterizations, we have for Scaled Fully Connected Networks that:*

$$\tilde{C} \leq \hat{C}^{\mathrm{SFC}} \triangleq \left(\sum_{l=1}^{L} \left(d_l^\star d_{l-1}^\star + 2 d_l \right) \right) \log Q , \tag{30}$$

where we defined $d_0^\star \triangleq d_0$. And by Lemma 3.2, $N = (\tilde{C} + 3\log 2/\delta)/\varepsilon$ samples are enough to ensure that for posterior sampling (i.e. G&C), $\mathcal{L}(\mathcal{A}_{\mathcal{P}}(\mathcal{S})) \leq \varepsilon$ with probability $1 - \delta$ over $\mathcal{S} \sim \mathcal{D}^N$ and the sampling.

Proof. The idea is similar to the proof of Theorem C.3, only that now we can define a set $\mathcal{E}$ with even more elements, which will tighten our bound. For any $l \in [L]$, write $\boldsymbol{\gamma}$ as a blocks vector

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_1^{(l)} \\ \boldsymbol{\gamma}_2^{(l)} \end{bmatrix} \in \mathbb{R}^{d_l} ,$$

where

$$\boldsymbol{\gamma}_1^{(l)} \in \mathbb{R}^{d_l^\star} ,$$
$$\boldsymbol{\gamma}_2^{(l)} \in \mathbb{R}^{d_l - d_l^\star} .$$

This time, we are interested in:

$$\mathcal{E} = \left\{ h_{\boldsymbol{\theta}} \in \mathcal{H}_D^{\mathrm{SFC}} \middle| \forall l = 1, \ldots, L \; \mathbf{W}_{11}^{(l)} = \mathbf{W}^{\star(l)}, \; \mathbf{b}_1^{(l)} = \mathbf{b}^{\star(l)}, \; \mathbf{b}_2^{(l)} = \mathbf{0}_{d_l - d_l^\star}, \; \boldsymbol{\gamma}_1^{(l)} = \boldsymbol{\gamma}^{\star(l)}, \boldsymbol{\gamma}_2^{(l)} = \mathbf{0}_{d_l - d_l^\star} \right\} .$$

As in the proof of Theorem C.3, we claim that any $h_{\boldsymbol{\theta}} \in \mathcal{E}$ is TE. This time, we specifically show by induction over the layer l, that for all $\mathbf{x} \in \mathbb{R}^{d_0}$,

$$f_D^{(l)}(\mathbf{x}) = \begin{bmatrix} f_{D^\star}^{(l)}(\mathbf{x}) \\ \mathbf{0}_{d_l - d_l^\star} \end{bmatrix} .$$

Let $h_{\boldsymbol{\theta}} \in \mathcal{E}$. Begin from the base case, $l = 1$. Since $d_0 = d_0^\star$,

$$\mathbf{W}^{(1)} = \begin{bmatrix} \mathbf{W}_{11}^{(l)} \\ \mathbf{W}_{21}^{(l)} \end{bmatrix}$$

so using the notation from Def. 4.1 we find that

$$\boldsymbol{\gamma}^{(1)} \odot \mathbf{W}^{(1)} f_D^{(0)}(\mathbf{x}) + \mathbf{b}^{(1)} = \boldsymbol{\gamma}^{(1)} \odot \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} = \begin{bmatrix} \boldsymbol{\gamma}_1^{(1)} \odot \mathbf{W}_{11}^{(1)}\mathbf{x} + \mathbf{b}_1^{(1)} \\ \boldsymbol{\gamma}_2^{(1)} \odot \mathbf{W}_{21}^{(1)}\mathbf{x} + \mathbf{b}_2^{(1)} \end{bmatrix}$$

and

$$f_D^{(1)}(\mathbf{x}) = \sigma\left(\begin{bmatrix} \boldsymbol{\gamma}_1^{(1)} \odot \mathbf{W}_{11}^{(1)}\mathbf{x} + \mathbf{b}_1^{(1)} \\ \boldsymbol{\gamma}_2^{(1)} \odot \mathbf{W}_{21}^{(1)}\mathbf{x} + \mathbf{b}_2^{(1)} \end{bmatrix}\right) = \begin{bmatrix} \sigma\left(\boldsymbol{\gamma}_1^{(1)} \odot \mathbf{W}_{11}^{(1)}\mathbf{x} + \mathbf{b}_1^{(1)}\right) \\ \sigma\left(\boldsymbol{\gamma}_2^{(1)} \odot \mathbf{W}_{21}^{(1)}\mathbf{x} + \mathbf{b}_2^{(1)}\right) \end{bmatrix}.$$

Therefore, using the definition of π_1, we get

$$\pi_1\left(f_D^{(1)}(\mathbf{x})\right) = \sigma\left(\boldsymbol{\gamma}_1^{(1)} \odot \mathbf{W}_{11}^{(1)}\mathbf{x} + \mathbf{b}_1^{(1)}\right). \tag{31}$$

Since $h_{\boldsymbol{\theta}} \in \mathcal{E}$, we have that $\mathbf{W}_{11}^{(1)} = \mathbf{W}^{\star(1)}$, $\mathbf{b}_1^{(1)} = \mathbf{b}^{\star(1)}$ and $\boldsymbol{\gamma}_1^{(1)} = \boldsymbol{\gamma}^{\star(1)}$, so from (31) the coordinate projection is

$$\pi_1\left(f_D^{(1)}(\mathbf{x})\right) = \sigma\left(\boldsymbol{\gamma}^{\star(1)} \odot \mathbf{W}^{\star(1)}\mathbf{x} + \mathbf{b}^{\star(1)}\right) = f_{D^\star}^{(1)}(\mathbf{x}). \tag{32}$$

In addition, since $\boldsymbol{\gamma}_2^{(l)} = \mathbf{0}_{d_l - d_l^\star}$ and also $\mathbf{b}_2^{(1)} = \mathbf{0}_{d_l - d_l^\star}$, as well as $\sigma(0) = 0$, we have that

$$\sigma\left(\boldsymbol{\gamma}_2^{(1)} \odot \mathbf{W}_{21}^{(1)}\mathbf{x} + \mathbf{b}_2^{(1)}\right) = \mathbf{0}_{d_l - d_l^\star}. \tag{33}$$

Putting 33 and 32 together we have

$$f_D^{(1)}(\mathbf{x}) = \begin{bmatrix} f_{D^\star}^{(1)}(\mathbf{x}) \\ \mathbf{0}_{d_1 - d_1^\star} \end{bmatrix}.$$

Next, assume that $f_D^{(l-1)}(\mathbf{x}) = \begin{bmatrix} f_{D^\star}^{(l-1)}(\mathbf{x}) \\ \mathbf{0}_{d_{l-1} - d_{l-1}^\star} \end{bmatrix}$, for some $l \leq L - 1$. Following a similar argument we get

$$\begin{aligned}
f_D^{(l)}(\mathbf{x}) = & \\
&= \sigma\left(\boldsymbol{\gamma}^{(l)} \odot \mathbf{W}^{(l)} f_D^{(l-1)}(\mathbf{x}) + \mathbf{b}^{(l)}\right) \\
&= \sigma\left(\boldsymbol{\gamma}^{(l)} \odot \begin{bmatrix} \mathbf{W}_{11}^{(l)} & \mathbf{W}_{12}^{(l)} \\ \mathbf{W}_{21}^{(l)} & \mathbf{W}_{22}^{(l)} \end{bmatrix} \begin{bmatrix} f_{D^\star}^{(l-1)}(\mathbf{x}) \\ \mathbf{0}_{d_{l-1} - d_{l-1}^\star} \end{bmatrix} + \begin{bmatrix} \mathbf{b}_1^{(l)} \\ \mathbf{b}_2^{(l)} \end{bmatrix}\right) \\
&= \sigma\left(\boldsymbol{\gamma}^{(l)} \odot \begin{bmatrix} \mathbf{W}_{11}^{(l)} \\ \mathbf{W}_{21}^{(l)} \end{bmatrix} f_{D^\star}^{(l-1)}(\mathbf{x}) + \begin{bmatrix} \mathbf{b}_1^{(l)} \\ \mathbf{b}_2^{(l)} \end{bmatrix}\right) \\
&= \sigma\left(\begin{bmatrix} \boldsymbol{\gamma}_1^{(l)} \odot \mathbf{W}_{11}^{(l)} f_{D^\star}^{(l-1)}(\mathbf{x}) + \mathbf{b}_1^{(l)} \\ \boldsymbol{\gamma}_2^{(l)} \odot \mathbf{W}_{21}^{(l)} f_{D^\star}^{(l-1)}(\mathbf{x}) + \mathbf{b}_2^{(l)} \end{bmatrix}\right) \\
&= \begin{bmatrix} \sigma\left(\boldsymbol{\gamma}_1^{(l)} \odot \mathbf{W}_{11}^{(l)} f_{D^\star}^{(l-1)}(\mathbf{x}) + \mathbf{b}_1^{(l)}\right) \\ \sigma\left(\boldsymbol{\gamma}_2^{(l)} \odot \mathbf{W}_{21}^{(l)} f_{D^\star}^{(l-1)}(\mathbf{x}) + \mathbf{b}_2^{(l)}\right) \end{bmatrix} \\
&= \begin{bmatrix} \sigma\left(\boldsymbol{\gamma}^{\star(l)} \odot \mathbf{W}^{\star(l)} f_{D^\star}^{(l-1)}(\mathbf{x}) + \mathbf{b}^{\star(l)}\right) \\ \mathbf{0}_{d_{l-1} - d_{l-1}^\star} \end{bmatrix} \\
&= \begin{bmatrix} f_{D^\star}^{(l-1)}(\mathbf{x}) \\ \mathbf{0}_{d_{l-1} - d_{l-1}^\star} \end{bmatrix} \\
&= f_{D^\star}^{(l)}(\mathbf{x}).
\end{aligned}$$

For the last layer, $l = L$, the proof is identical, except for the scalar product and the application of the activation function, which are removed. Since we assume that $h_{\boldsymbol{\theta}}$ and $h_{\boldsymbol{\theta}^\star}$ have the same output dimension d_L, this proves that for all $\mathbf{x} \in \mathbb{R}^{d_0}$

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \pi_L\left(h_{\boldsymbol{\theta}}(\mathbf{x})\right) = h_{\boldsymbol{\theta}^\star}(\mathbf{x}) \ .$$

That is, $\mathcal{E} \subseteq \{h_{\boldsymbol{\theta}} | h_{\boldsymbol{\theta}} \equiv h_{\boldsymbol{\theta}^\star}\}$ and therefore

$$\mathbb{P}(\mathcal{E}) \leq \mathbb{P}\left(h_{\boldsymbol{\theta}} \equiv h_{\boldsymbol{\theta}^\star}\right) \ .$$

Finally, to calculate the probability to sample $h_{\boldsymbol{\theta}}$ in $\mathcal{E}$ we count the number of constrained parameters - parameters which are either determined by $h_{\boldsymbol{\theta}^\star}$ or are 0 in $\mathcal{E}$. Looking at the dimensions of $\mathbf{W}_{11}^{(l)}$, $\mathbf{b}_1^{(l)}$, $\mathbf{b}_2^{(l)}$ and $\boldsymbol{\gamma}^{(l)}$, we deduce that there are exactly

$$\mathcal{M} = \sum_{l=1}^{L} \left(d_l^\star d_{l-1}^\star + 2d_l\right)$$

such constrained parameters, and denote

$$\hat{C}^{\mathrm{SFC}} \triangleq \left(\sum_{l=1}^{L} \left(d_l^\star d_{l-1}^\star + 2d_l\right)\right) \log Q = \mathcal{M} \log Q \, .$$

As in the proof of Theorem C.3, under the uniform prior over parameters $\mathcal{P}$,

$$\tilde{p} \geq \mathbb{P}(\mathcal{E}) = Q^{-\mathcal{M}} \, ,$$

so

$$\tilde{C} = -\log(\tilde{p}) \leq \mathcal{M} \log Q = \hat{C}^{\mathrm{SFC}}$$

□

C.3. Convolutional Neural Network

We first restate the definition of a CNN.

Definition C.7 (CNN restated)**.** For multi-channel inputs $\mathbf{x}^j$, $\forall j \in [c_0]$, depth L, activation function $\sigma(\cdot)$, and multi-index channels number $D = (c_1, \ldots, c_L, d_s)$, $\mathbf{h}_{\boldsymbol{\theta}}^{\mathrm{CNN}}$ is a convolutional network (CNN) defined recursively, starting with $\forall j \in [c_0]$, $\mathbf{f}_{\mathrm{CNN}}^{(0),j}(\mathbf{x}) = \mathbf{x}^j$, and then, for all $l \in [L]$ and $i \in [c_l]$, as

$$\begin{aligned} \mathbf{f}_{\mathrm{CNN}}^{(l),i}(\mathbf{x}) =& \sigma\left(\sum_{j=1}^{c_{l-1}} \mathbf{K}_{i,j}^{(l)} * \mathbf{f}_{\mathrm{CNN}}^{(l-1),j}(\mathbf{x}) + b_i^{(l)}\right) \\ \mathbf{h}_{\boldsymbol{\theta}}^{\mathrm{CNN}}(\mathbf{x}) =& \operatorname{sign}\left(\operatorname{Vec}\left(\mathbf{f}_{\mathrm{CNN}}^{(L-1)}(\mathbf{x})\right)^{\top} \mathbf{w}^{(L+1)} + b^{(L+1)}\right), \end{aligned} \tag{34}$$

where for ease of notation the addition of the bias term in (34) is done element-wise. The network parameters are

$$\boldsymbol{\theta} = \left(\left\{\mathbf{K}^{(l)}\right\}_{l=1}^{L}, \left\{\mathbf{b}^{(l)}\right\}_{l=1}^{L}, \mathbf{w}^{(L+1)}, b^{(L+1)}\right),$$

where $\mathbf{K}_{i,j}^{(l)} \in \mathbb{R}^{k_l}$ are convolution operators defined by kernels with k_l parameters, and $\mathbf{b}^{(l)} \in \mathbb{R}^{c_l}$ are bias terms. $\mathbf{w}^{(L+1)} \in \mathbb{R}^{d_s}$ and $b^{(L+1)} \in \mathbb{R}$ are the weights and bias of the convolutional network's last fully connected layer, where d_s is the dimension of $\operatorname{Vec}\left(\mathbf{f}_{\mathrm{CNN}}^{(L)}(\mathbf{x})\right)$. We denote the class of all convolutional networks with multi-index widths D as $\mathcal{H}_D^{\mathrm{CNN}}$.

Remark C.8. As in the case of FCNs, we use the ambiguous $\mathbf{f}_D^{(l)}$ in place of $\mathbf{f}_{\mathrm{CNN}}^{(l)}$ to denote a general convolutional layer in a convolutional neural network with channel pattern D where the specific type can be inferred from context.

Motivation. In order to tackle the analysis of a convolutional neural network at ease, we will define some new operations. We observe that our previous arguments from Section C.1 is agnostic to the convolution itself and relies only on parameter counting. In general, CNNs are analogous to FCNs in the sense that with spatial dimension of 1, a convolutional layer is equivalent to a layer in a FCN, with each channel analogous to a neuron. Therefore, we define:

Definition C.9. For any $l \in [L]$, define

$$\mathbf{K}^{(l)} = \begin{bmatrix} \mathbf{K}_{1,1}^{(l)} & \cdots & \mathbf{K}_{1,c_{l-1}}^{(l)} \\ \vdots & \ddots & \vdots \\ \mathbf{K}_{c_l,1}^{(l)} & \cdots & \mathbf{K}_{c_l,c_{l-1}}^{(l)} \end{bmatrix}$$

and

$$\mathbf{f}_D^{(l)} = \begin{bmatrix} \mathbf{f}_D^{(l),1} \\ \vdots \\ \mathbf{f}_D^{(l),c_l} \end{bmatrix}.$$

We use this notation to concisely write the *multi-channel convolution* operation

$$\left(\mathbf{K}^{(l)} * \mathbf{f}_D^{(l)}\right)_i = \sum_{j=1}^{c_{l-1}} \mathbf{K}_{ij}^{(l)} * \mathbf{f}_D^{(l-1),j}.$$

Notation With these definitions, we define the extension to the block notation from (26):

$$\mathbf{K}^{(l),(11)} = \begin{bmatrix} \mathbf{K}_{1,1}^{(l)} & \cdots & \mathbf{K}_{1,c_{l-1}^\star}^{(l)} \\ \vdots & \ddots & \vdots \\ \mathbf{K}_{c_l^\star,1}^{(l)} & \cdots & \mathbf{K}_{c_l^\star,c_{l-1}^\star}^{(l)} \end{bmatrix}, \mathbf{K}^{(l),(12)} = \begin{bmatrix} \mathbf{K}_{1,c_{l-1}^\star+1}^{(l)} & \cdots & \mathbf{K}_{1,c_{l-1}}^{(l)} \\ \vdots & \ddots & \vdots \\ \mathbf{K}_{c_l^\star,c_{l-1}^\star+1}^{(l)} & \cdots & \mathbf{K}_{c_l^\star,c_{l-1}}^{(l)} \end{bmatrix},$$

$$\mathbf{K}^{(l),(21)} = \begin{bmatrix} \mathbf{K}^{(l)}_{c_l^\star+1,1} & \cdots & \mathbf{K}^{(l)}_{c_l^\star+1,c_{l-1}^\star} \\ \vdots & \ddots & \vdots \\ \mathbf{K}^{(l)}_{c_l,1} & \cdots & \mathbf{K}^{(l)}_{c_l,c_{l-1}^\star} \end{bmatrix}, \mathbf{K}^{(l),(22)} = \begin{bmatrix} \mathbf{K}^{(l)}_{c_l^\star+1,c_{l-1}^\star+1} & \cdots & \mathbf{K}^{(l)}_{c_l^\star+1,c_{l-1}} \\ \vdots & \ddots & \vdots \\ \mathbf{K}^{(l)}_{c_l,c_{l-1}^\star+1} & \cdots & \mathbf{K}^{(l)}_{c_l,c_{l-1}} \end{bmatrix},$$

so

$$\mathbf{K}^{(l)} = \begin{bmatrix} \mathbf{K}^{(l),(11)} & \mathbf{K}^{(l),(12)} \\ \mathbf{K}^{(l),(21)} & \mathbf{K}^{(l),(22)} \end{bmatrix}.$$

Additionally, with

$$\mathbf{f}^{(l)}_{D,1} = \begin{bmatrix} \mathbf{f}^{(l),1}_D \\ \vdots \\ \mathbf{f}^{(l),c_l^\star}_D \end{bmatrix}, \ \mathbf{f}^{(l)}_{D,2} = \begin{bmatrix} \mathbf{f}^{(l),c_l^\star+1}_D \\ \vdots \\ \mathbf{f}^{(l),c_l}_D \end{bmatrix},$$

and Def. C.9 we can write the multi-channel convolution in block form as

$$\begin{aligned} \mathbf{K}^{(l)}\mathbf{f}^{(l)}_D &= \begin{bmatrix} \mathbf{K}^{(l),(11)} & \mathbf{K}^{(l),(12)} \\ \mathbf{K}^{(l),(21)} & \mathbf{K}^{(l),(22)} \end{bmatrix} * \begin{bmatrix} \mathbf{f}^{(l)}_{D,1} \\ \mathbf{f}^{(l)}_{D,2} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{K}^{(l),(11)} * \mathbf{f}^{(l)}_{D,1} + \mathbf{K}^{(l),(12)} * \mathbf{f}^{(l)}_{D,2} \\ \mathbf{K}^{(l),(21)} * \mathbf{f}^{(l)}_{D,1} + \mathbf{K}^{(l),(22)} * \mathbf{f}^{(l)}_{D,2} \end{bmatrix}. \end{aligned}$$

Specifically, we can represent a convolutional layer, as defined in Def. C.7 by

$$\mathbf{f}^{(l)}_D(\mathbf{x}) = \sigma\left(\mathbf{K}^{(l)} * \mathbf{f}^{(l-1)}_D + \mathbf{b}^{(l)}\right). \tag{35}$$

Finally, we denote

$$\mathbf{w}^{(L+1)} = \begin{bmatrix} \mathbf{w}^{(L+1),1} \\ \mathbf{w}^{(L+1),2} \end{bmatrix}$$

where $\mathbf{w}^{(L+1),1} \in \mathcal{Q}^{d_s^\star}$ and $\mathbf{w}^{(L+1),2} \in \mathcal{Q}^{d_s - d_s^\star}$.

Theorem C.10 (Teacher equivalence probability for CNN restated)**.** *For any activation function and any depth L, let $D^\star = (c_1^\star, \ldots, c_L^\star, d_s^\star)$, $D = (c_1, \ldots, c_L, d_s) \in \mathbb{N}^{L+1}$ such that $D^\star \leq D$. If there exists some teacher $h_{\boldsymbol{\theta}^\star} \in \mathcal{H}^{\mathrm{CNN}}_{D^\star}$ and the prior is $\mathcal{P} = \mathcal{P}\left(\mathcal{H}^{\mathrm{CNN}}_D\right)$ then*

$$\tilde{C} \leq \hat{C}^{\mathrm{CNN}} \triangleq \left(d_s + 1 + \sum_{l=1}^{L} k_l c_l^\star c_{l-1} + c_l^\star\right) \log(Q)$$

And by Lemma 3.2, $N = (\tilde{C} + 3\log 2/\delta)/\varepsilon$ samples are enough to ensure that for posterior sampling (i.e. G&C), $\mathcal{L}(\mathcal{A}_{\mathcal{P}}(\mathcal{S})) \leq \varepsilon$ with probability $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^N$ and the sampling.

Proof. We follow the same strategy as in the proof of Theorem C.3. We define a sufficient condition on the parameters of the student network to ensure teacher equivalence, and then find the probability that this condition holds. Recall that for all $l \in [L]$, $\mathbf{K}^{\star(l)}$ and $\mathbf{b}^{\star(l)}$, where $\mathbf{K}^{\star(l)}_{i,j} \in \mathbb{R}^{k_l}$ is a convolution kernel with k_l parameters, and $\mathbf{b}^{\star(l)} \in \mathbb{R}^{c_l}$, are the teacher's l^{th} layer's convolution kernels and bias terms, respectively. Using the notation introduced in Def. C.9, the teacher's convolution kernels can be arranged as

$$\mathbf{K}^{\star(l)} = \begin{bmatrix} \mathbf{K}^{\star(l)}_{1,1} & \cdots & \mathbf{K}^{\star(l)}_{1,c_{l-1}^\star} \\ \vdots & \ddots & \vdots \\ \mathbf{K}^{\star(l)}_{c_l^\star,1} & \cdots & \mathbf{K}^{\star(l)}_{c_l^\star,c_{l-1}^\star} \end{bmatrix}.$$

Define

$$\mathcal{E} = \left\{ h_{\boldsymbol{\theta}} \in \mathcal{H}_D^{\text{CNN}} \;\middle|\; \begin{array}{l} \forall l = 1, \ldots, L \; \mathbf{K}^{(l),(11)} = \mathbf{K}^{\star(l)}, \; \mathbf{K}^{(l),(12)} = \mathbf{0}, \; \mathbf{b}_1^{(l)} = \mathbf{b}^{\star(l)}, \\ \mathbf{w}^{(L+1),1} = \mathbf{w}^{\star(L+1)}, \; \mathbf{w}^{(L+1),2} = \mathbf{0}_{d_s - d_s^\star}, \; b^{(L+1)} = b^{\star(L+1)} \end{array} \right\}.$$

We claim that any $h_{\boldsymbol{\theta}} \in \mathcal{E}$ is TE and can show this by induction over the layer l. With block notation, it is easy to see that this is identical to the proof of Theorem C.3 when substituting $\mathbf{W}^{(l)}$ with $\mathbf{K}^{(l)}$, $f_{\text{FC}}^{(l)}$ with $\mathbf{f}_{\text{CNN}}^{(l)}$, and standard matrix multiplication with convolution. Finally, we turn to finding $\mathbb{P}(\mathcal{E})$. For a CNN $\mathbf{h}_{\boldsymbol{\theta}}^{\text{CNN}} \in \mathcal{E}$, there are $c_l^\star \cdot c_{l-1}^\star$ convolution kernels set to equal the teacher's, $c_l^\star \cdot \left(c_{l-1} - c_{l-1}^\star\right)$ kernels set to 0, and $c_l^\star$ bias terms. Each convolution kernel $\mathbf{K}_{i,j}^{(l)}$ is defined by k_l parameters. From the linear layer, we need to account for d_s weight parameters and one bias term. In total,

$$\tilde{p} \geq \mathbb{P}(\mathcal{E}) = \frac{1}{Q^{\mathcal{M}}}$$

where

$$\mathcal{M} = \sum_{l=1}^{L} c_l^\star \cdot c_{l-1} \cdot k_l + c_l^\star + d_s + 1 = d_s + 1 + \sum_{l=1}^{L} c_l^\star \cdot (c_{l-1} \cdot k_l + 1) \; .$$

so

$$\tilde{C} = -\log(\tilde{p}) \leq \mathcal{M} \log Q = \hat{C}^{\text{CNN}}$$

□

C.4. Channel-Scaled Convolutional Neural Networks

Analogous to the FCN case, we present an additional CNN architecture. For simplicity, we state the definition of the architecture using the previously defined notation from C.9 and (35).

Definition C.11 (Channel Scaled CNN)**.** For multi-channel inputs $\mathbf{x}^j$, $\forall j \in [c_0]$, depth L, activation function $\sigma(\cdot)$, and multi-index channels number $D = (c_1, \ldots, c_L, d_s)$, $\mathbf{h}_{\boldsymbol{\theta}}^{\mathrm{SCN}}$ is a channel scaled convolutional neural network, or *scaled convolutional network* (SCN) for short, defined recursively, starting with $\mathbf{f}_{\mathrm{SCN}}^{(0)}(\mathbf{x}) = \mathbf{x}$, and then, for all $l \in [L]$, as

$$\begin{aligned} \mathbf{f}_{\mathrm{SCN}}^{(l)}(\mathbf{x}) &= \left(\boldsymbol{\gamma}^{(l)} \odot \sigma \mathbf{K}^{(l)} * \mathbf{f}_{\mathrm{SCN}}^{(l-1)} + \mathbf{b}^{(l)}\right) \\ \mathbf{h}_{\boldsymbol{\theta}}^{\mathrm{SCN}}(\mathbf{x}) &= \operatorname{sign}\left(\operatorname{Vec}\left(\mathbf{f}_{\mathrm{SCN}}^{(L)}(\mathbf{x})\right)^{\top} \mathbf{w}^{(L+1)} + b^{(L+1)}\right), \end{aligned} \tag{36}$$

where for ease of notation the addition of the bias term in (36) is done element-wise. The network parameters are

$$\boldsymbol{\theta} = \left(\left\{\mathbf{K}^{(l)}\right\}_{l=1}^{L}, \left\{\mathbf{b}^{(l)}\right\}_{l=1}^{L}, \left\{\boldsymbol{\gamma}^{(l)}\right\}_{l=1}^{L}, \mathbf{w}^{(L+1)}, b^{(L+1)}\right),$$

where $\mathbf{K}_{i,j}^{(l)} \in \mathbb{R}^{k_l}$ are convolution operators defined by kernels with k_l parameters, $\mathbf{b}^{(l)} \in \mathbb{R}^{c_l}$ are bias terms, and $\boldsymbol{\gamma}^{(l)}$ are channel scaling parameters. $\mathbf{w}^{(L+1)} \in \mathbb{R}^s$ and $b^{(L+1)} \in \mathbb{R}$ are the weights and bias of the convolutional network's last fully connected layer, where d_s is the dimension of $\operatorname{Vec}\left(\mathbf{f}_{\mathrm{SCN}}^{(L)}(\mathbf{x})\right)$. We denote the class of all SCNs with multi-index widths D as $\mathcal{H}_D^{\mathrm{SCN}}$.

Theorem C.12 (Teacher equivalence probability for SCN restated)**.** *For any activation function and any depth L, let $D^\star = (c_1^\star, \ldots, c_L^\star, d_s^\star)$, $D = (c_1, \ldots, c_L, d_s) \in \mathbb{N}^{L+1}$ such that $D^\star \leq D$. If there exists some teacher $h_{\boldsymbol{\theta}^\star} \in \mathcal{H}_{D^\star}^{\mathrm{SCN}}$ and the prior is $\mathcal{P} = \mathcal{P}\left(\mathcal{H}_D^{\mathrm{SCN}}\right)$ then*

$$\tilde{C} \leq \hat{C}^{\mathrm{SCN}} \triangleq \left(d_s^\star + 1 + \sum_{l=1}^{L}\left(c_l^\star c_{l-1}^\star k_l + 2c_l\right)\right)\log(Q)$$

And by Lemma 3.2, $N = (\tilde{C} + 3\log 2/\delta)/\varepsilon$ samples are enough to ensure that for posterior sampling (i.e. G&C), $\mathcal{L}(\mathcal{A}_{\mathcal{P}}(\mathcal{S})) \leq \varepsilon$ with probability $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^N$ and the sampling.

Proof. The proof is completely analogous to the proof of Theorem C.6. Specifically, we claim that

$$\mathcal{E} = \left\{ h_{\boldsymbol{\theta}} \in \mathcal{H}_D^{\mathrm{SCN}} \;\middle|\; \begin{array}{l} \forall l = 1, \ldots, L\ \mathbf{K}^{(l),(11)} = \mathbf{K}^{\star(l)},\ \boldsymbol{\gamma}_1^{(l)} = \boldsymbol{\gamma}_1^{\star(l)},\ \boldsymbol{\gamma}_2^{(l)} = \mathbf{0}, \\ \mathbf{b}_1^{(l)} = \mathbf{b}^{\star(l)},\ \mathbf{w}^{(L+1),1} = \mathbf{w}^{\star(L+1)},\ b^{(L+1)} = b^{\star(L+1)} \end{array} \right\}.$$

is a subset of the teacher equivalent SCNs, and therefore

$$\tilde{p} \geq \mathbb{P}(\mathcal{E}) = \frac{1}{Q^{\mathcal{M}}}$$

where

$$\mathcal{M} \triangleq d_s^\star + 1 + \sum_{l=1}^{L}\left(c_l^\star c_{l-1}^\star k_l + 2c_l\right).$$

so

$$\tilde{C} = -\log(\tilde{p}) \leq \mathcal{M}\log Q = \hat{C}^{\mathrm{SCN}}$$

□

D. Proof for the Interpolation Probability in the Continuous Case (Section 5)

D.1. Technical Lemmas

To prove the results about continuous single hidden layer NN, we rely on the following basic Lemma.

Lemma D.1. *For any vector* $\mathbf{y}$ *and random vector* $\mathbf{x} \sim (\mathbf{0}, \mathbf{I}_d)$, $\varepsilon \in \left(0, \frac{\pi}{2}\right)$ *and* $u \in (0,1)$ *we have*

$$\mathbb{P}\left(\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} > \cos(\varepsilon)\right) = \frac{1}{2}\mathbb{P}\left(\left|\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}\right| > \cos(\varepsilon)\right) \geq \frac{\sin(\varepsilon)^{d-1}}{(d_0-1)\, B\left(\frac{1}{2}, \frac{d-1}{2}\right)} \tag{37}$$

$$\mathbb{P}\left(\left|\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}\right| < u\right) \leq \frac{2u}{B\left(\frac{1}{2}, \frac{d-1}{2}\right)}, \tag{38}$$

where we use $B(x,y)$ *to denote the beta function.*

Proof. Since $\mathcal{N}(0, \mathbf{I}_d)$ is spherically symmetric, we have

$$\mathbb{P}\left(\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} > \cos(\varepsilon)\right) = \frac{1}{2}\mathbb{P}\left(\left|\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}\right| > \cos(\varepsilon)\right),$$

and we can set $\mathbf{y} = [1, 0 \ldots, 0]^\top$, without loss of generality. Therefore, as in (TV, 2017)

$$\left|\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}\right|^2 = \frac{x_1^2}{x_1^2 + \sum_{i=2}^d x_i^2} \sim \mathcal{B}\left(\frac{1}{2}, \frac{d-1}{2}\right),$$

where $\mathcal{B}$ denotes the Beta distribution, since $x_1^2 \sim \chi^2(1)$ and $\sum_{i=2}^d x_i^2 \sim \chi^2(d-1)$ are independent chi-square random variables.

Suppose $Z \sim \mathcal{B}(\alpha, \beta)$, $\alpha \in (0,1)$, and $\beta > 1$.

$$\mathbb{P}(Z > u) = \frac{\int_u^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx}{B(\alpha,\beta)} \geq \frac{\int_u^1 1^{\alpha-1}(1-x)^{\beta-1}\, dx}{B(\alpha,\beta)} = \frac{\int_0^{1-u} x^{\beta-1} dx}{B(\alpha,\beta)} = \frac{(1-u)^\beta}{\beta B(\alpha,\beta)}.$$

Therefore, for $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}\right|^2 > \cos^2(\varepsilon)\right) \geq \frac{2\left(1-\cos^2(\varepsilon)\right)^{\frac{d-1}{2}}}{(d-1)\, B\left(\frac{1}{2}, \frac{d-1}{2}\right)} = \frac{2\sin(\varepsilon)^{d-1}}{(d-1)\, B\left(\frac{1}{2}, \frac{d-1}{2}\right)},$$

which proves (37).

Similarly, for $\alpha \in (0,1)$ and $\beta > 1$

$$\mathbb{P}(Z < u) = \frac{\int_0^u x^{\alpha-1}(1-x)^{\beta-1}\, dx}{B(\alpha,\beta)} \leq \frac{\int_0^u x^{\alpha-1} 1^{\beta-1} dx}{B(\alpha,\beta)} = \frac{u^\alpha}{\alpha B(\alpha,\beta)}.$$

Therefore,

$$\mathbb{P}\left(\left|\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}\right|^2 < u^2\right) \leq \frac{2u}{B\left(\frac{1}{2}, \frac{d_0-1}{2}\right)},$$

which proves (38). □

Lemma D.2. *For large* x

$$B\left(\frac{1}{2}, x\right) = \sqrt{\pi/x} + O\left(x^{-3/2}\right). \tag{39}$$

Proof. Using $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, Stirling's approximation for the Gamma function

$$\Gamma(x) = \sqrt{\frac{2\pi}{x}} \left(\frac{x}{e}\right)^x \left(1 + O\left(x^{-1}\right)\right) ,$$

and the definition of the Beta function,

$$\begin{aligned} B\left(\frac{1}{2}, x\right) &= \frac{\Gamma\left(\frac{1}{2}\right)\Gamma(x)}{\Gamma\left(\frac{1}{2}+x\right)} \\ &= \frac{\sqrt{\pi}\sqrt{\frac{2\pi}{x}}\left(\frac{x}{e}\right)^x \left(1 + O\left(x^{-1}\right)\right)}{\sqrt{\frac{2\pi}{x+\frac{1}{2}}}\left(\frac{x+\frac{1}{2}}{e}\right)^{x+\frac{1}{2}} \left(1 + O\left(x^{-1}\right)\right)} \\ &= \sqrt{e\cdot\pi}\sqrt{\frac{1}{x}}\left(\frac{x}{x+\frac{1}{2}}\right)^x \left(1 + O\left(x^{-1}\right)\right) \\ &= \sqrt{e\cdot\pi}\sqrt{\frac{1}{x}}\frac{1}{\left(1+\frac{1}{2x}\right)^x} \left(1 + O\left(x^{-1}\right)\right) \\ &= \sqrt{\frac{\pi}{x}} \left(1 + O\left(x^{-1}\right)\right) . \end{aligned}$$

□

Corollary D.3. *For any vector* $\mathbf{y}$ *and* $\mathbf{x} \sim (\mathbf{0}, \mathbf{I}_d)$*,* $\varepsilon \in \left(0, \frac{\pi}{2}\right)$ *and* $u \in (0,1)$ *we have*

$$\begin{aligned} \mathbb{P}\left(\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} > \cos(\varepsilon)\right) &= \frac{1}{2}\mathbb{P}\left(\left|\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}\right| > \cos(\varepsilon)\right) \\ &\geq \exp\left(d\log(\sin(\varepsilon)) - \frac{1}{2}\log(d) - \frac{1}{2}\log(2\pi)\right)\left(1 + O\left(d^{-1}\right)\right) . \end{aligned}$$

Proof.

$$\begin{aligned} \frac{\sin(\varepsilon)^{d-1}}{(d-1)\,B\left(\frac{1}{2}, \frac{d-1}{2}\right)} &= \frac{\sin(\varepsilon)^{d-1}}{(d-1)\sqrt{\frac{2\pi}{d-1}}\left(1 + O\left(\frac{1}{d-1}\right)\right)} \\ &= \frac{\sin(\varepsilon)^{d-1}}{\sqrt{2\pi(d-1)}}\left(1 + O\left(d^{-1}\right)\right) \\ &= \frac{\sin(\varepsilon)^{d-1}}{\sqrt{2\pi(d-1)}}\left(1 + O\left(d^{-1}\right)\right) \\ &= \exp\left((d-1)\log(\sin(\varepsilon)) - \frac{1}{2}\log(d-1) - \frac{1}{2}d_1^\star \log(2\pi)\right)\left(1 + O\left(d^{-1}\right)\right) \\ &= \exp\left(-(d-1)\left|\log(\sin(\varepsilon))\right| - \frac{1}{2}\log(d-1) - \frac{1}{2}\log(2\pi)\right)\left(1 + O\left(d^{-1}\right)\right) \\ &> \exp\left(-d\left|\log(\sin(\varepsilon))\right| - \frac{1}{2}\log(d) - \frac{1}{2}\log(2\pi)\right)\left(1 + O\left(d^{-1}\right)\right) . \end{aligned}$$

□

For completeness, we follow with the setting and notation introduced in Section 5 with slight modification.

D.2. Setting and Notation

Let $h_{\boldsymbol{\theta}}, h_{\boldsymbol{\theta}^\star}$ be fully connected (Def. 4.1) two layer neural network models with input dimension d_0, output dimension $d_2 = 1$ and hidden layer dimensions d_1 and $d_1^\star$, respectively. To simplify notation, we omit the sign activation from the definition of $h_{\boldsymbol{\theta}}$ and $h_{\boldsymbol{\theta}^\star}$ and denote

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{z}^\top \sigma(\mathbf{W}\mathbf{x})$$
$$h_{\boldsymbol{\theta}^\star}(\mathbf{x}) = \mathbf{z}^{\star\top} \sigma(\mathbf{W}^\star \mathbf{x})$$

where

$$\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{d_1}]^\top \in \mathbb{R}^{d_1 \times d_0}, \ \mathbf{z} \in \mathbb{R}^{d_1},$$
$$\mathbf{W}^\star = \left[\mathbf{w}_1^\star, \ldots, \mathbf{w}_{d_1^\star}^\star, \mathbf{0}, \ldots, \mathbf{0}\right]^\top \in \mathbb{R}^{d_1 \times d_0}, \ \mathbf{z}^\star \in \mathbb{R}^{d_2 \times d_1},$$

and $\sigma(\cdot)$ is the common leaky rectifier linear unit (LReLU, Maas et al. (2013)) with parameter $\rho \notin \{0, 1\}$.

$$\sigma(u) = u a(u) \ \text{ with } a(u) = \begin{cases} 1 & \text{, if } u > 0 \\ \rho & \text{, if } u < 0 \end{cases}, \tag{40}$$

The training set $\mathbf{X} = \left[\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right] \in \mathbb{R}^{d_0 \times N}$ consists of N datapoints. Thus, the output of the FCN on the entire dataset can be written as

$$h_{\boldsymbol{\theta}}(\mathbf{X}) = \sigma(\mathbf{W}\mathbf{X})^\top \mathbf{z} \in \mathbb{R}^N. \tag{41}$$
$$h_{\boldsymbol{\theta}^\star}(\mathbf{X}) = \sigma(\mathbf{W}^\star\mathbf{X})^\top \mathbf{z}^\star \in \mathbb{R}^N. \tag{42}$$

We denote the labels $y^{(n)} = \operatorname{sign}\left(h_{\boldsymbol{\theta}^\star}\left(\mathbf{x}^{(n)}\right)\right)$, and

$$\mathbf{y} = \left[y^{(1)}, \ldots, y^{(N)}\right]^\top \in \{\pm 1\}^N$$

We use the notation $\boldsymbol{a}^{(n)} = a\left(\mathbf{W}\mathbf{x}^{(n)}\right) \in \{\rho, 1\}^{d_1}$ for the activation pattern of the hidden layer of $h_{\boldsymbol{\theta}}$ on the input $\mathbf{x}^{(n)}$. In addition, we define the flattened weights' vectors of $h_{\boldsymbol{\theta}}$ and $h_{\boldsymbol{\theta}^\star}$ as

$$\mathbf{w} = \operatorname{vec}\left(\mathbf{W}^\top \operatorname{diag}(\mathbf{z})\right) \in \mathbb{R}^{d_0 d_1}$$
$$\mathbf{w}^\star = \operatorname{vec}\left(\mathbf{W}^{\star\top} \operatorname{diag}(\mathbf{z}^\star)\right) \in \mathbb{R}^{d_0 d_1},$$

respectively, where $\mathbf{W}^\star$ and $\mathbf{z}^\star$ are padded with 0's to match the dimensions. Let $\boldsymbol{\phi}^{(n)} = \left(\boldsymbol{a}^{(n)} \otimes \mathbf{x}^{(n)}\right) y^{(n)} \in \mathbb{R}^{d_0 d_1}$ where $\otimes$ is the Kronecker product. With this notation, it can be shown that

$$y^{(n)} h\left(\mathbf{x}^{(n)}\right) = \mathbf{w}^\top \phi^{(n)}.$$

Assumption D.4 (Prior over hypotheses, continuous setting, restated)**.** Suppose that the weights of $h_{\boldsymbol{\theta}}$ are random such that each row of the first layer, $\mathbf{w}_i$, is independently sampled from a uniform distribution on the unit sphere $\mathbb{S}^{d_0-1}$, and the second layer $\mathbf{z}$ is sampled uniformly from $\mathbb{S}^{d_1-1}$. Both $\mathbf{w}_i$ and $\mathbf{z}$ are independent of the teacher and data.

D.3. Proof of Theorem 5.6

Definition D.5 (First layer angular margin, restated)**.** For any training set $\mathcal{S} = \{\mathbf{x}_n\}_{n=1}^N$, we say that $\mathcal{S}$ has *first layer angular margin* α w.r.t. the teacher if

$$\forall i \in [d_1^\star], n \in [N] : \left|\frac{\mathbf{x}_n^\top \mathbf{w}_i^\star}{\|\mathbf{x}_n\|_2 \|\mathbf{w}_i^\star\|_2}\right| > \sin\alpha. \tag{43}$$

We denote the event that $h_{\boldsymbol{\theta}^\star}$ and $h_{\boldsymbol{\theta}}$ agree on the activation pattern of the data by

$$\tilde{\mathcal{G}}(\mathbf{X}, \mathbf{W}^\star) \triangleq \left\{\mathbf{W} \in \mathbb{R}^{d_1 \times d_0} | \forall i \in d_1^\star : \ \operatorname{sign}\left(\mathbf{w}_i^\top \mathbf{X}\right) = \operatorname{sign}\left(\mathbf{w}_i^{\star\top} \mathbf{X}\right)\right\}. \tag{44}$$

To bound the probability of this event, we use the following Lemma, adapted from Soudry & Hoffer (2017). For completeness, we write its proof here.

Lemma D.6 (Activation matching probability, adapted from Soudry & Hoffer (2017))**.** *Let* $\mathbf{X}$ *be a dataset with first layer angular margin* α *w.r.t* $h_{\boldsymbol{\theta}^\star}$*. Then*

$$\mathbb{P}_{\mathbf{W}}\left(\tilde{\mathcal{G}}\left(\mathbf{X},\mathbf{W}^\star\right)\right) \geq \left[\frac{\sin\left(\alpha\right)^{d_0-1}}{\left(d_0-1\right)B\left(\frac{1}{2},\frac{d_0-1}{2}\right)}\right]^{d_1^\star},$$

and when $d_0 \gg 1$ *and* $\frac{d_1^\star}{d_0} \ll 1$

$$\mathbb{P}_{\mathbf{W}}\left(\tilde{\mathcal{G}}\left(\mathbf{X},\mathbf{W}^\star\right)\right) \geq \exp\left(d_1^\star d_0 \log\left(\sin\left(\alpha\right)\right) - \frac{1}{2}d_1^\star \log\left(d_0\right) - \frac{1}{2}d_1^\star \log\left(2\pi\right)\right)\left(1+O\left(d_0^{-1}d_1^\star\right)\right).$$

Proof. To bound $\mathbb{P}_{\mathbf{W}}\left(\tilde{\mathcal{G}}\left(\mathbf{X},\mathbf{W}^\star\right)\right)$, we define the event that all weight hyperplanes with normals $\mathbf{w}_i$, have an angle of at most α from the corresponding target hyperplanes with normals $\mathbf{w}_i^\star$.

$$\forall i \in [d_1^\star]\ \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right) \triangleq \left\{\mathbf{W}\in\mathbb{R}^{d_1\times d_0} \middle| \frac{\mathbf{w}_i^\top \mathbf{w}_i^\star}{\|\mathbf{w}_i\|\,\|\mathbf{w}_i^\star\|} > \cos\left(\alpha\right)\right\}.$$

Since $\mathbf{X}$ has first layer angular margin α, in order that $\operatorname{sign}\left(\mathbf{w}_i^\top \mathbf{x}^{(n)}\right) \neq \operatorname{sign}\left(\mathbf{w}_i^{\star\top}\mathbf{x}^{(n)}\right)$, $\mathbf{w}_i$ must be rotated in respect to $\mathbf{w}_i^\star$ by an angle greater then the angular margin α. Therefore, we have that

$$\bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right) \subset \tilde{\mathcal{G}}\left(\mathbf{X},\mathbf{W}^\star\right). \tag{45}$$

And so,

$$\begin{aligned}
\mathbb{P}_{\mathbf{W}}\left(\tilde{\mathcal{G}}\left(\mathbf{X},\mathbf{W}^\star\right)\right) &\overset{(1)}{\geq} \mathbb{P}_{\mathbf{W}}\left(\bigcap_{i=1}^{d_1^\star}\tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right)\right) && (46)\\
&\overset{(2)}{=} \prod_{i=1}^{d_1^\star}\mathbb{P}_{\mathbf{W}}\left(\mathbf{W}\in\tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right)\right) && (47)\\
&= \prod_{i=1}^{d_1^\star}\mathbb{P}_{\mathbf{W}}\left(\frac{\mathbf{w}_i^\top\mathbf{w}_i^\star}{\|\mathbf{w}_i\|\,\|\mathbf{w}_i^\star\|} > \cos\left(\alpha\right)\right) && (48)\\
&\overset{(3)}{\geq} \left[\frac{\sin\left(\alpha\right)^{d_0-1}}{\left(d_0-1\right)B\left(\frac{1}{2},\frac{d_0-1}{2}\right)}\right]^{d_1^\star}, && (49)
\end{aligned}$$

where in (1) we used (45), in (2) we used the independence of $\{\mathbf{w}_i\}_{i=1}^{d_1^\star}$ and in (3) we used (D.1). When $d_0 \gg 1$, we can use Corollary D.3 to get

$$\mathbb{P}_{\mathbf{W}}\left(\mathbf{W}\in\tilde{\mathcal{G}}\left(\mathbf{X},\mathbf{W}^\star\right)\right) \geq \exp\left(d_1^\star d_0 \log\left(\sin\left(\alpha\right)\right) - \frac{1}{2}d_1^\star \log\left(d_0\right) - \frac{1}{2}d_1^\star \log\left(2\pi\right)\right)\left(1+O\left(d_0^{-1}\right)\right)^{d_1^\star}.$$

We can simplify this equation when $d_0 \gg 1$ with the asymptotic expansion of the beta function from Lemma D.2. If $d_1^\star d_0^{-1} \ll 0$ then the error $\left(1+O\left(d_0^{-1}\right)\right)^{d_1^\star} = 1+O\left(d_0^{-1}d_1^\star\right)$. Overall, this means that

$$\begin{aligned}
\mathbb{P}_{\mathbf{W}}\left(\mathbf{W}\in\tilde{\mathcal{G}}\left(\mathbf{X},\mathbf{W}^\star\right)\right) &\geq \left[\frac{\sin\left(\alpha\right)^{d_0-1}}{\sqrt{2\pi\left(d_0-1\right)}}\right]^{d_1^\star}\left(1+O\left(d_1^\star d_0^{-1}\right)\right)\\
&\geq \exp\left(d_1^\star d_0 \log\left(\sin\left(\alpha\right)\right) - \frac{1}{2}d_1^\star \log\left(d_0\right) - \frac{1}{2}d_1^\star \log\left(2\pi\right)\right)\left(1+O\left(d_0^{-1}d_1^\star\right)\right)
\end{aligned}$$

□

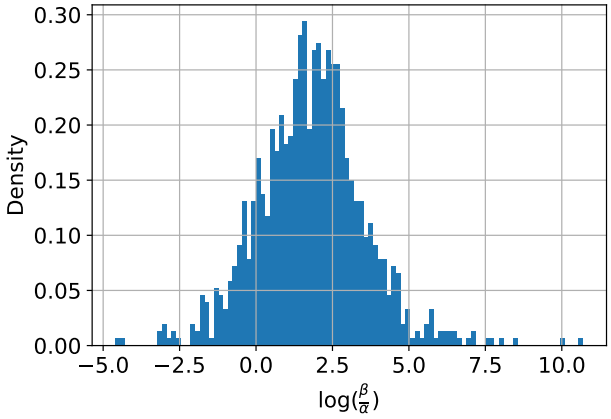


Figure 3. The density of the log of the ratio between β and α, for standard-Gaussian data, and a two-layer neural network with $\rho = 0.01$ and $d_0 = 500$, $d_1 = 10,000$, $d_1^\star = 1,000$. We sampled $50,000$ such datapoints and calculated α, β as the minimal angles as in (11) and (12) for a randomly initialized model, for a total of $1,000$ times.

Definition D.7 (Second layer angular margin, restated)**.** For any training set $\mathcal{S} = \{(\mathbf{x}^{(n)})\}_{n=1}^N$, we say that $\mathcal{S}$ has *second layer angular margin* β w.r.t. the teacher if

$$\forall n \in [N] \quad \left| \frac{h_{\boldsymbol{\theta}^\star}\left(\mathbf{x}^{(n)}\right)}{\left\|\mathbf{x}^{(n)}\right\|_2 \left\|\mathbf{z}^\star\right\|_2} \right| > \sqrt{d_1\left(1+\rho^2\right)} \sin\beta\,. \tag{50}$$

Assumption D.8. Let $\alpha < \beta \in \left(0, \frac{\pi}{2}\right)$. There exists $\lambda \in (0,1)$ such that, with probability at least $1-\lambda$ over the training set $\mathcal{S} = \{(\mathbf{x}_n)\}_{n=1}^N \sim \mathcal{D}^N$, $\mathcal{S}$ has first layer angular margin α (Def. D.5) and second layer angular margin β (Def. D.7).

Theorem D.9 (Interpolation of Continuous Networks, restated)**.** *Under Assumption D.8, with probability at least* $1-\lambda$ *over the dataset* $\mathcal{S}$*, the probability of interpolation is lower-bounded by*

$$\begin{aligned}\hat{p}_{\mathcal{S}} &= \mathbb{P}_{\mathbf{W},\mathbf{z}}\left(\mathcal{L}_S\left(h\right) = 0\right) \\ &\geq \exp\left(d_1^\star d_0 \log\left(\sin\left(\alpha\right)\right) + d_1 \log\left(\sin\left(\gamma\right)\right) - \frac{1}{2} d_1^\star \log\left(d_0\right) + O\left(d_1^\star + \log\left(d_1\right)\right)\right)\end{aligned}$$

where $\gamma = \arccos \frac{\cos\beta}{\cos\alpha}$.

Proof. By Assumption D.8 with probability at least $1-\lambda$, $\mathbf{X}$ has first and second layers angular margins α and β, respectively, w.r.t $h_{\boldsymbol{\theta}^\star}$. We assume that these properties hold for the rest of the proof. Suppose that we condition on the event $\bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right)$ from (45). Then, as in the proof of Lemma D.6, under the margin assumptions, $h_{\boldsymbol{\theta}}$ and $h_{\boldsymbol{\theta}^\star}$ agree on the activation pattern $\mathbf{a}^{(n)}$ and therefore

$$y^{(n)} h^\star\left(\mathbf{x}^{(n)}\right) = \mathbf{w}^{\star\top} \phi^{(n)}$$

in addition to

$$y^{(n)} h\left(\mathbf{x}^{(n)}\right) = \mathbf{w}^\top \phi^{(n)}\,.$$

Using basic properties of the Kronecker product,

$$\left\|\boldsymbol{\phi}^{(n)}\right\|^2 = \left\|\mathbf{a}^{(n)}\right\|^2 \left\|\mathbf{x}^{(n)}\right\|^2\,.$$

Furthermore, since $\mathbf{w}_i, \mathbf{w}_i^\star \in \mathbb{S}^{d_0-1}$,

$$\|\mathbf{w}\|^2 = \sum_{i=1}^{d_1} z_i^2 \|\mathbf{w}_i\|^2 = \|\mathbf{z}\|^2$$

and similarly $\|\mathbf{w}^\star\|^2 = \|\mathbf{z}^\star\|^2$. $\mathbf{a}^{(n)} \in \{1, \rho\}^{d_1}$ so

$$\left\|\mathbf{a}^{(n)}\right\|^2 \leq d_1 \cdot \left(1+\rho^2\right) .$$

With these identities we deduce

$$\left|\frac{\mathbf{w}^{\star\top}\phi^{(n)}}{\|\mathbf{w}^\star\|\left\|\phi^{(n)}\right\|}\right| = \left|\frac{h^\star\left(\mathbf{x}^{(n)}\right)}{\left\|\mathbf{x}^{(n)}\right\|\left\|\mathbf{a}^{(n)}\right\|\|\mathbf{z}^\star\|}\right| \geq \frac{1}{\sqrt{d_1\left(1+\rho^2\right)}}\left|\frac{h^\star\left(\mathbf{x}^{(n)}\right)}{\left\|\mathbf{x}^{(n)}\right\|\|\mathbf{z}^\star\|}\right| .$$

Using the second layer angular margin,

$$\left|\frac{h^\star\left(\mathbf{x}^{(n)}\right)}{\left\|\mathbf{x}^{(n)}\right\|_2\|\mathbf{z}^\star\|_2}\right| > \sqrt{d_1\left(1+\rho^2\right)}\sin\beta \geq \left\|\mathbf{a}^{(n)}\right\|\sin\beta$$

so

$$\left|\frac{\mathbf{w}^{\star\top}\phi^{(n)}}{\|\mathbf{w}^\star\|\left\|\phi^{(n)}\right\|}\right| > \sin\beta\,. \tag{51}$$

Following the same logic as in the proof of Lemma D.6, in order that $\operatorname{sign}\left(y^{(n)}\right) \neq \operatorname{sign}\left(h\left(\mathbf{x}^{(n)}\right)\right)$, $\mathbf{w}$ must be rotated by angle at least β compared to $\mathbf{w}^\star$. That is,

$$\frac{\mathbf{w}^\top\mathbf{w}^\star}{\|\mathbf{w}\|\|\mathbf{w}^\star\|} > \cos\beta \tag{52}$$

implies interpolation of the dataset. From symmetry, the probability of (52) is exactly half that of

$$\frac{\left(\mathbf{w}^\top\mathbf{w}^\star\right)^2}{\|\mathbf{w}\|^2\|\mathbf{w}^\star\|^2} > \cos^2\beta\,.$$

Next,

$$\left(\mathbf{w}^\top\mathbf{w}^\star\right)^2 = \left(\sum_{i=1}^{d_1} z_i z_i^\star \mathbf{w}_i\cdot\mathbf{w}_i^\star\right)^2 ,$$

so (52) is equivalent to

$$\left(\sum_{i=1}^{d_1^\star} z_i z_i^\star \mathbf{w}_i\cdot\mathbf{w}_i^\star\right)^2 > \|\mathbf{z}\|^2\|\mathbf{z}^\star\|^2\cos^2\beta\,.$$

Conditioning on the events $\bigcap_{i=1}^{d_1^\star}\tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right)$, and

$$\forall i = 1,\ldots,d_1^\star\ z_i z_i^\star \geq 0\,,$$

(52) holds if

$$\left(\mathbf{z}\cdot\mathbf{z}^\star\right)^2\cos^2\alpha = \left(\sum_{i=1}^{d_1^\star} z_i z_i^\star\cos\alpha\right)^2 > \left(\|\mathbf{z}\|^2\|\mathbf{z}^\star\|^2\right)\cos^2\beta \tag{53}$$

i.e.

$$\frac{\left(\mathbf{z}\cdot\mathbf{z}^\star\right)^2}{\|\mathbf{z}\|^2\|\mathbf{z}^\star\|^2} > \frac{\cos^2\beta}{\cos^2\alpha}\,.$$

Denote

$$\gamma = \arccos\left(\frac{\cos\beta}{\cos\alpha}\right)$$

then, putting this all together,

$$
\begin{aligned}
&\mathbb{P}_{\mathbf{w}}\left(\forall n \in [N]\ y^{(n)} h\left(\mathbf{x}^{(n)}\right) > 0\right) \\
&\geq \mathbb{P}\left(\forall n \in [N]\ y^{(n)} h\left(\mathbf{x}^{(n)}\right) > 0, \mathbf{W} \in \bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right), \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) \\
&= \mathbb{P}\left(\forall n \in [N]\ y^{(n)} h\left(\mathbf{x}^{(n)}\right) > 0 \middle| \bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right), \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) \mathbb{P}\left(\bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right), \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) \\
&\geq \mathbb{P}\left(\frac{\mathbf{w}^\top \mathbf{w}^\star}{\|\mathbf{w}\| \|\mathbf{w}^\star\|} > \cos\beta \middle| \bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right), \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) \mathbb{P}\left(\bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right), \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) \\
&\geq \frac{1}{2}\mathbb{P}\left(\frac{\left(\mathbf{z} \cdot \mathbf{z}^\star\right)^2}{\|\mathbf{z}\|^2 \|\mathbf{z}^\star\|^2} > \frac{\cos^2\beta}{\cos^2\alpha} \middle| \bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right), \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) \mathbb{P}\left(\bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right), \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) \\
&= \frac{1}{2}\mathbb{P}\left(\frac{\left(\mathbf{z} \cdot \mathbf{z}^\star\right)^2}{\|\mathbf{z}\|^2 \|\mathbf{z}^\star\|^2} > \cos^2\gamma \middle| \bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right), \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) \mathbb{P}\left(\bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right)\right) \mathbb{P}\left(\forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) \\
&= \frac{1}{2}\mathbb{P}\left(\frac{\left(\mathbf{z} \cdot \mathbf{z}^\star\right)^2}{\|\mathbf{z}\|^2 \|\mathbf{z}^\star\|^2} > \cos^2\gamma \middle| \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) && (54) \\
&\quad \cdot \mathbb{P}\left(\bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right)\right) && (55) \\
&\quad \cdot \mathbb{P}\left(\forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right). && (56)
\end{aligned}
$$

Starting from (54),

$$
\mathbb{P}\left(\frac{\left(\mathbf{z} \cdot \mathbf{z}^\star\right)^2}{\|\mathbf{z}\|^2 \|\mathbf{z}^\star\|^2} > \cos^2\gamma \middle| \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) = \mathbb{P}\left(\frac{\left(\sum_{i=1}^{d_1^\star} |z_i z_i^\star|\right)^2}{\|\mathbf{z}\|^2 \|\mathbf{z}^\star\|^2} > \cos^2\gamma\right)
$$

and since $\left(\mathbf{z} \cdot \mathbf{z}^\star\right)^2 \leq \left(\sum_{i=1}^{d_1^\star} |z_i z_i^\star|\right)^2$ almost surely,

$$
\mathbb{P}\left(\frac{\left(\mathbf{z} \cdot \mathbf{z}^\star\right)^2}{\|\mathbf{z}\|^2 \|\mathbf{z}^\star\|^2} > \cos^2\gamma \middle| \forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) \geq \mathbb{P}\left(\frac{\left(\mathbf{z} \cdot \mathbf{z}^\star\right)^2}{\|\mathbf{z}\|^2 \|\mathbf{z}^\star\|^2} > \cos^2\gamma\right).
$$

From Lemma D.1, we obtain

$$
\mathbb{P}\left(\frac{\left(\mathbf{z} \cdot \mathbf{z}^\star\right)^2}{\|\mathbf{z}\|^2 \|\mathbf{z}^\star\|^2} > \cos^2(\gamma)\right) \geq \frac{2\sin(\gamma)^{d_1^\star - 1}}{\left(d_1^\star - 1\right) B\left(\frac{1}{2}, \frac{d_1^\star - 1}{2}\right)}.
$$

As for (55), we know from Lemma D.6 that

$$
\mathbb{P}\left(\bigcap_{i=1}^{d_1^\star} \tilde{\mathcal{G}}_i^\alpha\left(\mathbf{W}^\star\right)\right) \geq \left[\frac{\sin(\alpha)^{d_0 - 1}}{(d_0 - 1) B\left(\frac{1}{2}, \frac{d_0 - 1}{2}\right)}\right]^{d_1^\star}.
$$

For (56), $\mathbb{P}_z(z) = \mathcal{N}(0, 1)$ so

$$
\mathbb{P}\left(\forall i = 1, \ldots, d_1^\star\ z_i z_i^\star > 0\right) = 2^{-d_1^\star}.
$$

Overall,

$$\mathbb{P}_{\mathbf{w}}\left(\forall n\in[N]\ y^{(n)}h\left(\mathbf{x}^{(n)}\right)>0\right)\geq 2^{-d_1^\star}\frac{\sin\left(\gamma\right)^{d_1^\star-1}}{\left(d_1^\star-1\right)B\left(\frac{1}{2},\frac{d_1^\star-1}{2}\right)}\left[\frac{\sin\left(\alpha\right)^{d_0-1}}{\left(d_0-1\right)B\left(\frac{1}{2},\frac{d_0-1}{2}\right)}\right]^{d_1^\star}.$$

When $d_0\gg d_1^\star\gg 1$ we get from Corollary D.3

$$\begin{aligned}\mathbb{P}_{\mathbf{w}}\left(\forall n\in[N]\ y^{(n)}h\left(\mathbf{x}^{(n)}\right)>0\right)\geq 2^{-d_1^\star}\exp\left(d_1^\star d_0\log\left(\sin\left(\alpha\right)\right)-\frac{1}{2}d_1^\star\log\left(d_0\right)-\frac{1}{2}d_1^\star\log\left(2\pi\right)\right)\left(1+O\left(d_0^{-1}d_1^\star\right)\right)\\ \cdot\exp\left(d_1\log\left(\sin\left(\gamma\right)\right)-\frac{1}{2}\log\left(d_1\right)-\frac{1}{2}\log\left(2\pi\right)\right)\left(1+O\left(d_1^{-1}\right)\right).\end{aligned}$$

That is,

$$\begin{aligned}\hat{p}_{\mathcal{S}}&=\mathbb{P}_{\mathbf{w}}\left(\forall n\in[N]\ y^{(n)}h\left(\mathbf{x}^{(n)}\right)>0\right)\\&\geq\exp\left(d_1^\star d_0\log\left(\sin\left(\alpha\right)\right)+d_1\log\left(\sin\left(\gamma\right)\right)-\frac{1}{2}d_1^\star\log\left(d_0\right)+O\left(d_1^\star+\log\left(d_1\right)\right)\right).\end{aligned}$$

□

The following generalization bound follows directly from Prop. B.23 and Theorem 5.6.

Corollary D.10 (Generalization of continuous two layer networks, restated)**.** *Under the assumption that $\hat{p}_{\mathcal{S}}<\frac{1}{2}$, for any $\varepsilon,\delta\in(0,1)$,*

$$\mathbb{P}_{S\sim\mathcal{D}^N,h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h)\leq\varepsilon\right)\geq 1-\delta\ -\lambda,,$$

for

$$N\geq\frac{\hat{C}^{\text{cont}}+4\log\left(\frac{8}{\delta}\right)+2\log\left(\hat{C}^{\text{cont}}\right)}{\varepsilon}$$

with

$$\hat{C}^{\text{cont}}=-d_1^\star d_0\log\left(\sin\left(\alpha\right)\right)-d_1\log\left(\sin\left(\gamma\right)\right)+\frac{1}{2}d_1^\star\log\left(d_0\right)+O\left(d_1^\star+\log\left(d_1\right)\right).$$

Proof. Under Assumption D.8, Theorem 5.6 implies that w.p. at least $1-\lambda$ over $\mathcal{S}\sim\mathcal{D}^N$,

$$\hat{p}_{\mathcal{S}}\geq\exp\left(d_1^\star d_0\log\left(\sin\left(\alpha\right)\right)+d_1\log\left(\sin\left(\gamma\right)\right)+O\left(d_1^\star\log\left(d_0\right)+\log\left(d_1\right)\right)\right).$$

We denote this event by $\mathcal{E}_1$. Recalling that $\hat{C}^{\text{cont}}\geq-\log\left(\hat{p}_{\mathcal{S}}\right)$ when conditioned on $\mathcal{E}_1$, from Prop. B.23 we deduce that,

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h)\leq\varepsilon\right)\geq 1-\delta.$$

We denote this event by $\mathcal{E}_2$. Using the inclusion exclusion principle,

$$\begin{aligned}\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{E}_1\cap\mathcal{E}_2\right)&=\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{E}_1\right)+\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{E}_2\right)-\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{E}_1\cup\mathcal{E}_2\right)\\&\geq 1-\lambda+1-\delta-1\\&=1-\delta-\lambda.\end{aligned}$$

Therefore,

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^N,h\sim\mathcal{P}_{\mathcal{S}}}\left(\mathcal{L}_{\mathcal{D}}(h)\leq\varepsilon\right)\geq 1-\delta\ -\lambda.$$

□

E. Proofs for Sparsest Quantized Interpolator Learning Rule

E.1. Setting and Notation

Given a directed graph $G = (V, E)$ and $x \in V$, we use $d^{\text{in}}(x)$ to denote the in-degree of x, i.e.

$$d^{\text{in}}(x) \triangleq \sum_{y \in V} \mathbb{I}\left[(y, x) \in E\right] .$$

Under the same quantization scheme as in Section 4, consider the following learning rule:

Definition E.1. $\mathcal{A}_0(\mathcal{S}) = h_{\boldsymbol{\theta}_0}$ returns the sparsest quantized interpolator,

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta} \in \mathcal{Q}^M}{\operatorname{argmin}} \|\boldsymbol{\theta}\|_0 \text{ s.t } \forall n \in [N] \;\; y^{(n)} h_{\boldsymbol{\theta}}\left(\mathbf{x}^{(n)}\right) > 0 ,$$

where $\|\boldsymbol{\theta}\|_0$ is the number of nonzero values in $\boldsymbol{\theta}$. With some abuse of notation, we use $\|\mathcal{A}_0(\mathcal{S})\|_0$ and $\|\boldsymbol{\theta}_0\|_0$ interchangeably.

Recall that we denote the total number of parameters in a teacher network $h_{\boldsymbol{\theta}^\star}$ by

$$M(D^\star) = \sum_{l=1}^{L} \left(d_l^\star d_{l-1}^\star + d_l^\star\right) .$$

We additionally denote by

$$W(D^\star) = \sum_{l=1}^{L} d_l^\star d_{l-1}^\star ,$$
$$B(D^\star) = \sum_{l=1}^{L} d_l^\star$$

the maximal number of non-zero weights and biases in $h_{\boldsymbol{\theta}^\star}$, respectively. Denote the class of fully-connected neural networks with at most $W(D^\star)$ non-zero weights and $B(D^\star)$ biases as $\mathcal{H}_{W(D^\star),B(D^\star)}$ [13]. Notice that the number of neural networks in $\mathcal{H}_{W(D^\star),B(D^\star)}$ is bounded by the number of neural networks with $M(D^\star)$ edges and no bias terms. We denote the set of such neural networks by $\mathcal{H}_{M(D^\star)}$, then

$$\left|\mathcal{H}_{W(D^\star),B(D^\star)}\right| \leq \left|\mathcal{H}_{M(D^\star)}\right| . \tag{57}$$

We emphasize that $\mathcal{H}_{W(D^\star),B(D^\star)}$ and $\mathcal{H}_{M(D^\star)}$ do not have fixed depth and hidden layer width, and contains models which do not conform to a specific D.

E.2. Generalization Bound

Lemma E.2. *The number of FCNs with $\|\mathcal{A}_0(\mathcal{S})\|_0$ non-zero parameters is upper bounded by $\left|\mathcal{H}_{M(D^\star)}\right|$.*

Proof. Since $h_{\boldsymbol{\theta}^\star}$ is an interpolating solution, we have that $\|\mathcal{A}_0(\mathcal{S})\|_0 \leq M(D^\star)$. The bound follows from the case in which all neurons have 0 bias, and the number of non-zero weights is equal to the number of non-zero parameters. □

Next, we note that any fully-connected neural network with no bias terms can be represented as a *weighted directed acyclic graph* (WDAG) $G = (V, E, w)$. The vertices V represent the neurons, and the network's non-zero weights are represented as weighted edges. Notice that the input neurons in an FCN are of 0 in-degree, and that all neurons are reachable from some input neuron. This motivates us to define the following.

[13] This is different from $\mathcal{H}_D^{FC}$ as no specific depth and hidden layer widths are assumed for $\mathcal{H}_{W(D^\star),B(D^\star)}$

Definition E.3. Let $\tilde{\mathcal{G}}_{M(D^\star),d_0}$ be the set of DAGs, $G=(V,E)$, containing a subset $\Sigma \subseteq V$ with 0 in-degree. Such that

$$\tilde{\mathcal{G}}_{M(D^\star),d_0} \triangleq \left\{ G=(V,E) \middle| \; G \text{ is a DAG}, \; |E| = M\left(D^\star\right), \; \exists \Sigma \subseteq V \; : \; |\Sigma| = d_0, \; \forall x \in \Sigma \; d^{\text{in}}\left(x\right) = 0 \right\} .$$

We say that a vertex $v \in V$ is reachable from Σ if there exists some directed path from a vertex in Σ to v. With this notion, we further specify

$$\mathcal{G}_{M(D^\star),d_0} \triangleq \left\{ G=(V,E) \in \tilde{\mathcal{G}}_{M(D^\star),d_0} \middle| \; \forall v \in V \; v \text{ is reachable from } \Sigma \right\} .$$

That is, $\mathcal{G}_{M(D^\star),d_0}$ is the subset of $\tilde{\mathcal{G}}_{M(D^\star),d_0}$ in which any node is reachable from some node in Σ.

Clearly, $\left|\mathcal{G}_{M(D^\star),d_0}\right|$ is an upper bound for $\left|\mathcal{H}_{M(D^\star)}\right|$, so

$$\left|\mathcal{H}_{M(D^\star)}\right| \leq Q^{M(D^\star)} \left|\mathcal{G}_{M(D^\star),d_0}\right| . \tag{58}$$

Lemma E.4.

$$\left|\mathcal{G}_{M(D^\star),d_0}\right| \leq \frac{\left(M\left(D^\star\right)\left(M\left(D^\star\right)+d_0\right)\right)^{M(D^\star)}}{M\left(D^\star\right)!} \leq \left(M\left(D^\star\right)+d_0\right)^{2M(D^\star)} .$$

Proof. We start from the basic property, that any directed graph $G=(V,E)$ can be represented as a bipartite undirected graph $\tilde{G} = \left(V' \cup V'', \tilde{E}\right)$ where V', V'' are copies of V and

$$\tilde{E} = \left\{ \{v_1', v_2''\} \in V' \times V'' | (v_1, v_2) \in E \right\} .$$

Let $G=(V,E) \in \mathcal{G}_{M(D^\star),d_0}$, $\Sigma \subseteq V$ the appropriate 0-in-degree subset of nodes in G, and $\tilde{G} = \left(V' \cup V'', \tilde{E}\right)$ its corresponding bipartite representation. By the definition of $\mathcal{G}_{M(D^\star),d_0}$, every vertex $v \in V \setminus \Sigma$ is reachable from some $x \in \Sigma$ and therefore for all $v \in V \setminus \Sigma$, $d^{\text{in}}\left(v\right) \geq 1$ in G, and $\deg\left(v''\right) \geq 1$ in $\tilde{G}$. Since for all $x \in \Sigma$, $d^{\text{in}}\left(x\right) = 0$, and $|E| = \left|\tilde{E}\right| = M\left(D^\star\right)$, we can use the pigeonhole principle to deduce that $V = \Sigma \cup U$ where U is a set of at most $W\left(D^\star\right)$ vertices, $|U| \leq W\left(D^\star\right)$. Hence, any $G \in \mathcal{G}_{M(D^\star),d_0}$ can be represented using an undirected bipartite graph $\hat{G} = \left(\hat{V}' \cup \hat{V}'', \hat{E}\right)$ such that

$$\left|\hat{V}'\right| \leq M\left(D^\star\right) + d_0 ,$$

$$\left|\hat{V}''\right| \leq M\left(D^\star\right) ,$$

$$\left|\hat{E}\right| = M\left(D^\star\right) ,$$

where $\hat{V}'$ is a copy of $\Sigma \cup U$, and $\hat{V}''$ is a copy of U. This means that we can bound $\left|\mathcal{G}_{M(D^\star),d_0}\right|$ with the number of such graphs. The number of possible edges in $\hat{G}$ is

$$\left|\hat{V}'\right| \cdot \left|\hat{V}''\right| = \left(M\left(D^\star\right)+d_0\right) \cdot M\left(D^\star\right) ,$$

so, overall, the number of such bipartite representations for graphs in $\mathcal{G}_{M(D^\star),d_0}$ is

$$\binom{M\left(D^\star\right)\left(M\left(D^\star\right)+d_0\right)}{M\left(D^\star\right)} \leq \frac{\left(M\left(D^\star\right)\left(M\left(D^\star\right)+d_0\right)\right)^{M(D^\star)}}{M\left(D^\star\right)!} \leq \frac{\left(M\left(D^\star\right)+d_0\right)^{2M(D^\star)}}{M\left(D^\star\right)!} \leq \left(M\left(D^\star\right)+d_0\right)^{2M(D^\star)}$$

and

$$|\mathcal{G}| \leq \frac{\left(M\left(D^\star\right)\left(M\left(D^\star\right)+d_0\right)\right)^{M(D^\star)}}{M\left(D^\star\right)!} .$$

□

Collecting the bounds from (57), (58) and Lemma E.4 we find the following corollaries.

Corollary E.5. *The number of $M\left(D^{\star}\right)$-sparse Q-quantized fully-connected neural networks is bounded by*

$$\left|\mathcal{H}_{M\left(D^{\star}\right)}\right| \leq \frac{\left(M\left(D^{\star}\right)\left(M\left(D^{\star}\right)+d_{0}\right)\right)^{M\left(D^{\star}\right)}}{M\left(D^{\star}\right)!} Q^{M\left(D^{\star}\right)} \leq\left(M\left(D^{\star}\right)+d_{0}\right)^{2 M\left(D^{\star}\right)} Q^{M\left(D^{\star}\right)} .$$

Using Theorem B.1 we get,

Corollary E.6. *Let $\varepsilon>0$ and $\delta \in(0,1)$. With probability at least $1-\delta$ over $\mathcal{S} \sim \mathcal{D}^{N}$,*

$$\mathcal{L}_{\mathcal{D}}\left(\mathcal{A}_{0}\left(\mathcal{S}\right)\right) \leq \varepsilon$$

when

$$N \geq \frac{2 M\left(D^{\star}\right) \log\left(M\left(D^{\star}\right)+d_{0}\right)+M\left(D^{\star}\right) \log(Q)+\log\left(\frac{1}{\delta}\right)}{\varepsilon} .$$