The Size of Teachers as a Measure of Data Complexity: PAC-Bayes Excess Risk Bounds and Scaling Laws

Gintare Karolina Dziugaite
Google DeepMind; Mila; McGill University

Daniel M. Roy
University of Toronto; Vector Institute

Abstract

We study the generalization properties of neural networks through the lens of data complexity. Recent work by Buzaglo et al. (2024) shows that random (nearly) interpolating networks generalize, provided there is a small "teacher" network that achieves small excess risk. We give a short single-sample PAC-Bayes proof of this result and an analogous "fast-rate" result for random samples from Gibbs posteriors. The resulting oracle inequality motivates a new notion of data complexity, based on the minimal size of a teacher network required to achieve any given level of excess risk. We show that polynomial data complexity gives rise to power laws connecting risk to the number of training samples, like in empirical neural scaling laws. By comparing the "scaling laws" resulting from our bounds to those observed in empirical studies, we provide evidence for lower bounds on the data complexity of standard benchmarks.

1 INTRODUCTION

One of the challenges facing deep learning theory is that we know that the generalization performance of deep learning is heavily dependent on the data, yet we have no good way of measuring the complexity of data. At the same time, practicioners have identified empirical "neural scaling laws" that predict the performance of deep learning across wide ranges of data and model sizes after hypertuning. What might neural scaling laws tell us about the complexity of the underlying data?

In this work, we study an idealized model of neural network training, and obtain oracle inequalities relating the risk of the learned neural network ("student") to the risk and size of any smaller network ("teacher"). In particular, the bound is determined by the size of the teacher that optimizes the balance between its risk (ideally small) and its size (ideally small). The bound depends only mildly on the size of the student network that was trained, allowing for a moderate level of overparametrization, which is crucial since the size of the "optimal" teacher (for the number of data at hand) is *a priori* unknown. We obtain this mild dependence building on recent work of Buzaglo et al. (2024).

In order to turn the oracle inequality into an actual rate of decay for excess risk, we ask: for every $\varepsilon > 0$, what is the size (parameter count) $\mathcal{C}(\varepsilon)$ of a (teacher) network necessary to achieve ε excess risk? We propose that the function $\mathcal{C}(\varepsilon)$ is the right architecture-sensitive measure of complexity for data distributions (tasks). For a suite of possible functional forms for $\mathcal{C}(\varepsilon)$, we describe the final excess risk rates for learning.

We are, however, ignorant of the complexity $\mathcal{C}(\varepsilon)$ of real data. To gain some insight, we turn to recent empirical "neural scaling laws", which offer estimated rates. By comparing our rates with those captured in neural scaling laws, we find evidence that $\mathcal{C}(\varepsilon) \in O(\varepsilon^{-p})$ for some $p > 0$. We conclude with a discussion and limitations.

Prior Work Understanding the remarkable generalization ability of deep learning remains a central challenge in machine learning theory. Recent work by Buzaglo et al. (2024) has provided valuable insights by connecting generalization to the existence of a narrow teacher network that generates the labels. Under this realizability assumption, the authors showed that the sample complexity scales with the number of weights in the smallest, consistent teacher network and the number of neurons in the student network, rather than the number of weights in the student network. This is a desirable property, as it suggests that overparameterized networks can generalize well if there exists a simple explanation for the data (i.e., a narrow teacher).

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. © 2025 by the authors.

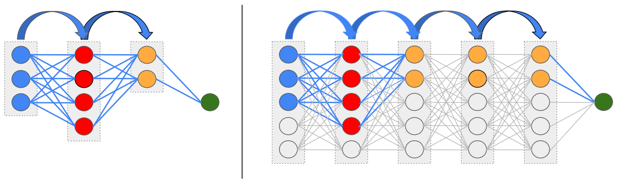

Figure 1: An illustration of a teacher network (left) embedded in a deeper and wider student network (right). The architecture for both networks is a gated activation and residual connection neural network, defined in Definition 4.1. The extra neurons in each student layer are turned off by setting the activation gates $\alpha_i^{(l)}$ in Equation (6) to zero, as in (Buzaglo et al., 2024). Superfluous layers (up to the last one) are made to be identity operators by setting the layer gates $\beta^{(l)}$ to zero. As shown in Appendix D.2, the two networks are functionally equivalent.

Their work highlights the role of implicit bias in model architecture and initialization, suggesting that these factors contribute to finding simpler, more generalizable solutions, even in the heavily overparameterized regime. One significant limitation in their analysis is that the student network has to be of the same depth as the teacher, which would not allow us to explain why slight changes in depth do not affect empirical performance.

To achieve our goals, we relax this assumption. We extend the results to work on student networks of arbitrary width and depth—the smaller the teacher network, the smaller the risk bound (Figure 1). Crucially, we also focus our analysis to consider the situation where no (small) teacher perfectly interpolates the data.[1] Instead of interpolation ("hard" posterior sampling), we consider Gibbs ("soft") posterior sampling, which is a common model for the output of stochastic gradient methods. This allows us to identify that the right measure of data complexity is the function $\mathcal{C}(\varepsilon)$ – the size of network needed to achieve ε excess risk.

1.1 Contributions

- In Section 3, we present a tighter risk bound for predictors sampled from a posterior distribution, conditioned on interpolating the training data. The bound scales with the negative log probability of interpolation, and improves on (Buzaglo et al., 2024, Thm. 3.2).
- In Section 4, we derive a lower bound on the probability of interpolation under the assumption that the labels are generated by a smaller teacher network of the same architecture but with reduced depth and width. The penalty term in this bound depends on the number of parameters in the teacher network and the width and depth of the student network, extending the results of Buzaglo et al. (2024) beyond teacher's and student's of equal depth.
- In Section 5, we relax the realizability assumption (that labels are generated by a small teacher) by studying Gibbs posteriors. [1] We derive an oracle inequality that demonstrates the Gibbs posterior competes with the best teacher plus the penalty term derived above.
- Finally, in Section 6, we introduce a measure of data complexity based on the rate at which the excess risk decreases with increasing (teacher) network size, and present excess risk bounds in this case. Such a model and the corresponding bound captures why, as the student network size grows, the risk may be decreasing, as seen in practical scaling laws. Our bounds offer scaling laws that we then compare to the neural scaling laws presented in empirical studies.

2 PRELIMINARIES

Let Θ index a space of models, let Z denote a space of (labeled) data, and fix a (bounded) loss function $\ell : \Theta \times Z \to [0, 1]$. For $\theta \in \Theta$, let ℓ_θ denote the map $z \mapsto \ell(\theta, z)$, so that $D(\ell_\theta)$ is the risk of θ under $D \in \Delta_1(Z)$, where $\Delta_1(Z)$ denotes the space of probability measures.

We adopt Catoni's notation, which treats measures as linear operators, mapping functions to their integral. For a real-valued, integrable (or nonnegative) function f on Z and $D \in \Delta_1(Z)$, let $D(f) = \int f(z)\, D(\mathrm{d}z)$ denote the integral. Write D^n for the product measure. By a (probability) kernel from, say, Z^n to Θ, we mean a map $\kappa : Z^n \to \Delta(\Theta)$. For a pair of measures μ, π on Θ such that μ is absolutely continuous with respect to π, we write $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} : \Theta \to [0, \infty)$ for (an arbitrary version of) the Radon–Nikodym derivative.

Catoni (2007) proves the following "single-sample" deviation bound, which, unlike traditional PAC-Bayes bounds, yields a high-probability guarantee over both the randomness in the sample and the randomness in a draw $\hat{\theta}$ from a posterior kernel (i.e., data-dependent posterior) distribution on Θ.

Theorem 2.1 (Catoni 2007, Thm. 1.2.7)**.** *For any data distribution $D \in \Delta_1(Z)$, prior distribution $\pi \in \Delta_1(\Theta)$, positive real parameter λ, positive integer n, and kernel $\rho : D^n \to \Delta_1(\Theta)$, with probability at least $1 - \delta$ over $S \sim D^n$ and $\hat{\theta}|S \sim \rho(S)$,*

$$D(\ell_{\hat{\theta}}) \leq \Phi^{-1}_{\lambda/n} \left[\hat{D}_n(\ell_{\hat{\theta}}) + \frac{1}{\lambda} \log \left(\frac{1}{\delta} \frac{\mathrm{d}\rho(S)}{\mathrm{d}\pi}(\hat{\theta}) \right) \right], \quad (1)$$

[1] Concurrent with (Buzaglo et al., 2024), arXiv v2.

where $\hat{D}_n$ is the empirical distribution of S and $\Phi_a^{-1}(q) = (1 - \exp(-aq))/(1 - \exp(-a))$.

The effect of $\Phi^{-1}_{\lambda/n}$ is illuminated by the following inequalities due to (Catoni, 2007, Lem. 3.4.2):

Lemma 2.2. *Let $q, d \geq 0$ and define*

$$B(q,d) = \left(1 + \frac{2d}{n}\right)^{-1}\left[q + \frac{d}{n} + \sqrt{\frac{2dq(1-q)}{n} + \frac{d^2}{n^2}}\right].$$

Then $\inf_{\lambda^*} \Phi^{-1}_{\lambda^*/n}\left[q + d/\lambda\right] \leq q + \sqrt{d/2n}$ and, provided $B(q,d) \leq \frac{1}{2}$, the infimum is even bounded by $B(q,d)$.

Note that λ is nonrandom in Theorem 2.1, while all λ^* achieving (or nearly achieving) the infimum in Lemma 2.2 may depend on q and d, which correspond to random variables in Theorem 2.1. As such, the combination of these two results requires some additional effort, which we make later. For our first application, however, $D_n(\ell_{\hat{\theta}}) = 0$ a.s., in which case, the following result offers a "fast rate" bound. (See Appendix D.1 for its proof.)

Corollary 2.3. *Under the same conditions as Theorem 2.1, if $\hat{D}_n(\ell_{\hat{\theta}}) = 0$ a.s., then, with probability at least $1 - \delta$,*

$$D(\ell_{\hat{\theta}}) \leq \frac{1}{n}\left[\log\left(\frac{1}{\delta}\frac{\mathrm{d}\rho(S)}{\mathrm{d}\pi}(\hat{\theta})\right)\right]. \qquad (2)$$

The appearance of the Radon–Nikodym derivative $\frac{\mathrm{d}\rho(S)}{\mathrm{d}\pi}$ implies that one should aim for the posterior to be absolutely continuous with respect to the prior, i.e., $\rho(S) \ll \pi$, at least with high probability. Note that there is no assumption that $\pi \ll \rho(S)$, despite this additional hypothesis appearing in recent PAC-Bayes surveys.

What is the behaviour of random neural networks that interpolate a training sample? Let d and M be the depth and width of a random multilayer perceptron (MLP), i.e., the "student". In (Buzaglo et al., 2024), the authors present several analyses, one of which makes the assumption that the labels are generated by a "narrow teacher network", i.e., an MLP with the same depth as the student but smaller width $m < M$. Assuming that the teacher network's weights are quantized to Q levels, (Buzaglo et al., 2024) prove that an interpolating (quantized) student has risk less than ε with probability at least $1 - \delta$, given $\Omega((\log(1/\delta) + d(m^2 + M)\log Q)/\varepsilon)$ training samples, i.e., a sample complexity that scales with the number of the teacher parameters $d(m^2)$ and the number of neurons of the student, rather than with the much larger number of student weights.

3 ANALYSIS OF POSTERIOR SAMPLING

Buzaglo et al. (2024) perform a bespoke analysis of posterior sampling, based on bounding the value of a geometric random variable and using generalization bounds for finite hypothesis spaces. In this section we present a much more direct argument based on Theorem 2.1.

We first begin with a presentation of Bayes rule: Fix a data distribution $D \in \Delta_1(Z)$, prior distribution $\pi \in \Delta_1(\Theta)$, and positive integer n. Let $S \sim D^n$ and $\tilde{\theta} \sim \pi$, independently, and let $\hat{D}_n$ be the empirical distribution of S. Define the *probability of interpolation, (conditional on the sample S)* to be $I(S) = \mathbb{P}[\hat{D}_n(\ell_{\tilde{\theta}}) = 0|S] = \pi\{\theta : \hat{D}_n(\ell_\theta) = 0\}$. Let $\rho : Z^n \to \Delta_1(\Theta)$ satisfy $I(S)\,\rho(S)(A) = \mathbb{P}[\tilde{\theta} \in A, \hat{D}_n(\ell_{\tilde{\theta}}) = 0|S]$ for all measurable $A \subseteq \Theta$, a.s. That is, if $\hat{\theta} \sim \pi$, then $\rho(S)$ is the posterior distribution of $\hat{\theta}$ on the event $\hat{D}_n(\ell_{\tilde{\theta}}) = 0$.

Lemma 3.1. *On the event $I(S) > 0$, we have $\frac{\mathrm{d}\rho(S)}{\mathrm{d}\pi} = (I(S))^{-1}$ $\rho(S)$-almost everywhere.*

The following theorem is an immediate consequence of Lemma 3.1 and Corollary 2.3

Theorem 3.2. *Assume $I(S) > 0$ a.s. and let $\hat{\theta}|S \sim \rho(S)$, i.e., we assume interpolation is always possible and we sample $\hat{\theta}$ from the posterior given interpolation. Then, with probability at least $1 - \delta$,*

$$D(\ell_{\hat{\theta}}) \leq \frac{1}{n}\left[\log\frac{1}{\delta} + \log\frac{1}{I(S)}\right]. \qquad (3)$$

In particular, if $I(S) \geq \tilde{p} > 0$ a.s., then

$$D(\ell_{\hat{\theta}}) \leq \frac{1}{n}\left[\log\frac{1}{\delta} + \log\frac{1}{\tilde{p}}\right]. \qquad (4)$$

This results improves on (Buzaglo et al., 2024, Thm. 3.2), removing all spurious constants and placing no restrictions on the values of δ, $I(S)$, and $\tilde{p}$.

The PAC-Bayesian arguments underlying Theorem 3.2 and Equation (4) are also worth highlighting, in light of the claim by Buzaglo et al. that PAC-Bayesian bounds for a random interpolator yield significantly looser risk bounds than their bespoke analysis. While Buzaglo et al. point to a single-sample bound (Alquier, 2024, Thm. 2.7) that delivers the desired high probability bound for a single sample from the posterior, they observe that it leads to a looser, slow-rate bound, but incorrectly claim that the cause for this looseness is that the bound is general and applies to all posteriors. Indeed, Theorem 2.1 is a single-sample bound that applies to all posteriors, yet Corollary 2.3 reveals how it simultaneously yields both slow and fast rate bounds, depending on the scale of the empirical risk. Since interpolators achieve zero risk, it yields a fast rate. Our careful choice of PAC-Bayesian single-sample deviation bounds has thus yielded an improved bound, with a much simpler argument. Since Catoni's bound may not be widely known, our use of it here may be of independent interest.

4 STUDENT–TEACHER EQUIVALENCE AND SAMPLE COMPLEXITY UNDER REALIZABILITY

If we are to understand the generalization properties of neural networks sampled from the posterior, we are motivated by Equation (4) to develop lower bounds on $I(S)$, the probability of interpolation.

In this section, we adopt a strong but useful realizability assumption used by Buzaglo et al. (2024): namely, the labels are assumed to be generated by a (much smaller) "teacher" network, whose parametrization we will denote by θ^*. In Section 5 we drop this assumption, but we will reuse some of the results of this section.

The utility of this realizability assumption is that we can lower bound $I(S)$ by the probability $\tilde{p}$ of the event E_{θ^*} that a "student" network $\theta \sim \pi$ is *functionally equivalent* to θ^*, i.e., $\ell_\theta(x) = \ell_{\theta^*}(x)$ for all x. This works because, if a sampled network $\theta \sim \pi$ computes the same function as θ^*, then θ will interpolate the labels (achieve zero empirical risk). As another motivation for bounding the probability of E_{θ^*}, such bounds will also allow us to derive oracle inequalities *without* making a realizability assumption.

4.1 Functional equivalence in gated networks

Before we discuss posterior distributions, we must discuss prior distributions. What type of networks will our prior assign mass to? In this section, we define a type of residual architecture (He et al., 2016), with skip connections between every pair of adjacent layers. We introduce both gated activations *and* gated residual connections. Provided that we adopt quantized weights, we will be able to easily lower bound the probability of functional equivalence between a fixed teacher network θ^* and a randomly generated, wider and deeper student network θ (Figure 1 captures the key idea).

Definition 4.1. Fix an activation $\sigma : \mathbb{R} \to \mathbb{R}$. The class of *gated activation and residual connection neural networks* of depth L, widths $\mathbf{d} = (d_1, \dots, d_L)$, denoted $\mathrm{NN}_{L,\mathbf{d}}$, is the class of mappings $\{h_\theta\}$ where θ ranges over all tuples $(W^{(l)}, b^{(l)}, \alpha^{(l)}, \beta^{(l)}; l = 1, \dots, L)$, for $W^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$, $b^{(l)} \in \mathbb{R}^{d_l}$, $\alpha^{(l)} \in \mathbb{R}^{d_l}$, $\beta^{(l)} \in \mathbb{R}$, and where

$$h_\theta(x) = \mathrm{sign}(W^{(L)} f^{(L-1)}(x) + b^{(L)}), \tag{5}$$

for mappings $f^{(l)}$ defined recursively as follows:

$$\begin{aligned} &f^{(0)}(x) = x, \\ &f^{(l)}(x) = \beta^{(l)} \sigma(\alpha^{(l)} \odot W^{(l)} f^{(l-1)}(x) + b^{(l)}) + f^{(l-1)}(x), \end{aligned} \tag{6}$$

for $l = 1, ..., L-1$. The number of parameters associated to $\mathrm{NN}_{L,\mathbf{d}}$ is

$$C^{L,\mathbf{d}} = L + \textstyle\sum_{l=1}^{L} (d_l d_{l-1} + 2d_l). \tag{7}$$

In comparison to a standard MLP with skip connections, this parametrization introduces one scalar parameter per activation/neuron and per layer. By construction, if one of these scalars is zero, the associated activation/neuron is turned off (resp., the associated layer becomes an identity). If we remove the skip connections and gates $\beta^{(l)} = 1$, we obtain the scaled-neuron fully connected architecture (Buzaglo et al., 2024).

In order to lower bound the probability of functional equivalence, we will work with quantized weights taking Q possible values, including zero. If the activation σ satisfies $\sigma(0) = 0$, we can show that the negative log probability of a randomly generated network being functionally equivalent to a fixed small teacher network scales only linearly in the depth and width of the student, and linearly with the number of parameters in the teacher, rather than with the total number of parameter in the student. This result extends (Buzaglo et al., 2024, Thm. 5.3) to account for teachers and students of different depths as well as widths. For $(L^*, \mathbf{d}^*)$ and $(L, \mathbf{d})$, write $(L^*, \mathbf{d}^*) \le (L, \mathbf{d})$ if $L^* \le L$ and $d_i^* \le d$ for all $i = 1, \dots, L^*$. (The proof is provided in Appendix D.2.)

Theorem 4.2. *Let $(L^*, \mathbf{d}^*), (L, \mathbf{d})$ satisfy $(L^*, \mathbf{d}^*) \le (L, \mathbf{d})$. Fix $\theta^* \in \mathrm{NN}_{L^*,\mathbf{d}^*}$ and let $\theta \sim \pi$, where π is the uniform prior over $\mathrm{NN}_{L,\mathbf{d}}$, quantized to Q levels, and assume $\sigma(0) = 0$. Then*

$$-\log \pi(E_{\theta^*}) \le \big(C^{L^*,\mathbf{d}^*} + B_{L,\mathbf{d}}^{L^*,\mathbf{d}^*}\big) \log Q, \tag{8}$$

where $B_{L,\mathbf{d}}^{L^*,\mathbf{d}^*} = (L - L^*) + 2\sum_{l=1}^{L^*} (d_l - d_l^*)$.

The quantity $B_{L,\mathbf{d}}^{L^*,\mathbf{d}^*}$ captures the number of parameters needed to turn off all but a teacher-sized subnetwork of the student network. For a student–teacher pair of networks with homogeneous widths $\mathbf{d} = (d, \dots, d)$ and $\mathbf{d}^* = (d^*, \dots, d^*)$, $C^{L^*,\mathbf{d}^*} = \Theta\big(L^*({d^*}^2 + d)\big)$ while $B_{L,\mathbf{d}}^{L^*,\mathbf{d}^*} = \Theta\big(L + L^* d\big)$. Provided $d \in O({d^*}^2)$ and $L \in O(L^* d)$, both these quantities are $O(L^*(d^*)^2)$.

Note that the above lower bound does not attempt to account for the number of distinct embeddings of the teacher network into the larger student network, as the number of such embeddings depends on symmetries in the parameters of the teacher itself. Nonetheless, it is enough to obtain bounds that do not scale linearly with the number of student parameters.

4.2 Risk bound for posterior sampling under realizability

When the data is generated by a small teacher network, we can apply Theorem 4.2 with Theorem 3.2 to derive a bound on the risk of a wider-and-deeper student obtained by sampling from a flat prior over the parameters, conditioned on perfectly fitting the training data.

Recall that $\rho : Z^n \to \Delta_1(\Theta)$ maps a dataset to the posterior distribution under interpolation, i.e., ρ satisfies $\rho(S) = \mathbb{P}[\tilde{\theta} | \hat{D}_n(\ell_{\tilde{\theta}}) = 0, S]$ a.s., where $\tilde{\theta} \sim \pi$.

Theorem 4.3. *Fix σ s.t. $\sigma(0) = 0$. Let S be a size n i.i.d. sample from D and let $\hat{\theta}|S \sim \rho(S)$ be a sample from the posterior under π, conditioned on interpolation, where π is the uniform distribution on $\mathrm{NN}_{L,\mathbf{d}}$, with Q levels of quantization, one of which is zero. Assume that, for some $(L^*, \mathbf{d}^*) \le (L, \mathbf{d})$ and some $\theta^* \in \mathrm{NN}_{L^*,\mathbf{d}^*}$, we have $D(\ell_{\theta^*}) = 0$. Then, with probability $\ge 1 - \delta$,*

$$D(\ell_{\hat{\theta}}) \le \frac{1}{n} \left[\log \frac{1}{\delta} + \big(C^{L^*,\mathbf{d}^*} + B^{L^*,\mathbf{d}^*}_{L,\mathbf{d}} \big) \log Q \right]. \quad (9)$$

5 BEYOND REALIZABILITY: AN (AGNOSTIC) ORACLE INEQUALITY

If we aim to drop the realizability assumption, we must contend with the possibility that $I(S)$ may be zero with positive probability: it may possible that no student network can label the data perfectly. If we cannot reliably sample conditional on interpolation, we might sample conditional on being an empirical risk minimizer. We will not pursue this option directly here. Instead, we perform stochastic optimization by sampling from the so-called Gibbs posterior $\rho(S)$, where ρ is the probability kernel from Z to Θ given by

$$\frac{\mathrm{d}\rho(S)}{\mathrm{d}\pi}(\theta) = \frac{1}{Z_{\beta\pi}(S)} \exp\big(-\beta \hat{D}_n(\ell_\theta) \big) \quad (10)$$

for $Z_{\beta\pi}(S) = \int_\Theta \exp\big(-\beta \hat{D}_n(\ell_\theta) \big) \pi(\mathrm{d}\theta)$. The parameter $\beta > 0$ is the so-called inverse temperature. For *fixed* n and S, the Gibbs posterior concentrates on the empirical risk minimizers as $\beta \to \infty$. If $I(S) > 0$, then the Gibbs posterior converges to the hard posterior studied in previous sections. In this way, the Gibbs posterior can be seen as a relaxation of the hard posterior. In our analysis, β will grow with the number of data n in a precise way to yield nonasymptotic results.

Gibbs posteriors arise as the stationary distributions of Langevin diffusions (Welling and Teh, 2011). The exponentiated, differentiable drift function is the (unnormalized) Radon–Nikodym derivative with respect to Lebesgue measure. Here, the negative empirical risk may not be differentiable, but the Gibbs posterior is nonetheless well defined and represents an approximate minimizer of the empirical risk, where the degree of approximation is determined by β.

We apply Corollary 2.3 to study the properties of a single sample from a Gibbs posterior.

Corollary 5.1. *Consider the setting of Theorem 2.1 but take ρ to satisfy Equation (10). Then, with probability at least $1 - \delta$,*

$$D(\ell_{\hat{\theta}}) \le \Phi^{-1}_{\lambda/n}\Big[(1 - \lambda^{-1}\beta) \hat{D}_n(\ell_{\hat{\theta}}) - \lambda^{-1} \log\big(Z_{\beta\pi}(S) \big) + \lambda^{-1} \log \frac{1}{\delta} \Big].$$

The term $\lambda^{-1} \log\big(Z_{\beta\pi}(S) \big)$ is a so-called local entropy. In machine learning, this entropy term has appeared in optimization algorithms that directly optimize the entropy term (Chaudhari et al., 2019; Dziugaite and Roy, 2018).

As in the realizable case, we can consider lower bounds on the entropy term by identifying the contribution of a teacher network, θ^*. Let $E_{\theta^*} = \{\theta : \ell_\theta = \ell_{\theta^*}\}$. We have

$$\begin{aligned} Z_{\beta\pi}(S) &= \int_\Theta \exp\big(-\beta \hat{D}_n(\ell_\theta) \big) \pi(\mathrm{d}\theta) \\ &\ge \exp\big(-\beta \hat{D}_n(\ell_{\theta^*}) \big) \pi(E_{\theta^*}). \end{aligned} \quad (11)$$

The bound in Equation (11) holds almost surely and uniformly over θ^* and so, taking $\beta = \lambda$, we even have

$$D(\ell_{\hat{\theta}}) \le \inf_{\theta^*} \Phi^{-1}_{\lambda/n} \left[\hat{D}_n(\ell_{\theta^*}) + \frac{\log \frac{1}{\delta\, \pi(E_{\theta^*})}}{\lambda} \right]. \quad (12)$$

We may bound the infimum by taking $\theta^* = \theta^{**}$ for some arbitrary nonrandom $\theta^{**} \in \Theta$. Since θ^{**} is nonrandom, we may replace $\hat{D}_n(\ell_{\theta^{**}})$ with an upper bound in terms of the (nonrandom) risk $D(\ell_{\theta^{**}})$, via Bernstein's inequality. As a result, we have a nonrandom upper bound on $\Phi_{\lambda/n}\big(D(\ell_{\hat{\theta}}) \big)$, and so we may apply Lemma 2.2 to replace $\Phi_{\lambda/n}$. Using the fact that θ^{**} was arbitrary, we arrive at the following oracle inequality. (The detailed bound and its complete derivation appear in Appendix E.)

Theorem 5.2. *For $\theta \in \Theta$, let $r_\theta = D(\ell_\theta)$ and $c_\theta = \log 1/\pi(E_\theta)$. Under the same setting as Corollary 5.1, there exists $\lambda > 0$ such that, letting $\hat{\theta}$ be a sample from the Gibbs posterior with inverse temperature $\beta = \lambda$, with probability at least $1-\delta$, $D(\ell_{\hat{\theta}})$ is bounded by*

$$\inf_{\theta^* \in \Theta} \left\{ r_{\theta^*} + \tilde{\mathcal{O}}\left(\frac{c_{\theta^*}}{n} + \sqrt{\frac{c_{\theta^*}(r_{\theta^*} + \sqrt{r_{\theta^*}/n})}{n}} \right) \right\}.$$

In our neural network setting, the lower bound on $\pi(E_{\theta^*})$ from Theorem 4.2 offers the following immediate corollary:

Corollary 5.3. *Fix a student architecture $(L, \mathbf{d})$. Under the same setting as Theorem 5.2, rewriting the infimum $\inf_{\theta^*}$ by $\inf_{L^*,\mathbf{d}^*} \inf_{\theta^*}$, where the first infimum runs over all (teacher) architectures $(L^*, \mathbf{d}^*) \leq (L, \mathbf{d})$, and the second infimum runs over all (teacher) networks in $\mathrm{NN}_{L^*,\mathbf{d}^*}$, we have*

$$c_{\theta^*} \leq \frac{\log Q}{\sqrt{n}} \big(C^{L^*,\mathbf{d}^*} + B_{L,\mathbf{d}}^{L^*,\mathbf{d}^*} \big). \tag{13}$$

Corollary 5.3 establishes a risk bound for an arbitrary student network without explicit dependence on the number of student parameters. Instead, the bound scales with the number of student neurons and the number of parameters in the teacher that offers the best tradeoff between its risk and parameter count.

5.1 Empirical Proof of Concept with MNIST

While the above results already communicate the key theoretical idea that small teachers with low risk imply generalization, in Appendix C, we look at the consequences of these bounds on MNIST.

6 THE SIZE OF TEACHERS AS A MEASURE OF DATA COMPLEXITY

In light of Corollary 5.3, it is natural to ask when such bounds imply that we can converge to a Bayes optimal prediction and at what rate. Since our oracle inequality makes a comparison to any (small) teacher networks, we propose to characterize data distributions (and architectures) in terms of the minimal complexity $C^{L^*,\mathbf{d}^*}$ of a teacher network required to obtain a set level of excess risk.

Definition 6.1. Fix the data distribution D and let r^* denote the Bayes error (i.e., the expected loss of the best measurable predictor). The complexity of D—with respect to the architecture $\{\mathrm{NN}_{L,\mathbf{d}}\}_{L>0,\mathbf{d}>0}$ and complexity measure $C^{(L^*,\mathbf{d}^*)}$—is measured by the function $\mathcal{C} : [0,\infty) \to [0,\infty]$ given, for $\varepsilon \in (0,1)$, by

$$\begin{aligned} \mathcal{C}(\varepsilon) = \min_{(L^*,\mathbf{d}^*)} \quad & C^{(L^*,\mathbf{d}^*)} \\ \text{s.t.} \quad & \inf_{\theta^* \in \mathrm{NN}_{L^*,\mathbf{d}^*}} D(\ell_{\theta^*}) - r^* \leq \varepsilon. \end{aligned} \tag{14}$$

By definition, $\mathcal{C}(\varepsilon)$ is (minimal) complexity (associated to the "smallest" architecture) necessary and sufficient to achieve an excess risk of ε.[2]

Write $(L^*(\varepsilon), \mathbf{d}^*(\varepsilon))$ for an arbitrary pair achieving the minimum in the definition of $\mathcal{C}(\varepsilon)$. As before, let Q denote a fixed level of quantization. The following result is then immediate:

Theorem 6.2. *Fix a student architecture $(L, \mathbf{d})$ and let $r_\varepsilon = r^* + \varepsilon$ and $c_\varepsilon = \frac{\log Q}{\sqrt{n}} \big(\mathcal{C}(\varepsilon) + B_{L,\mathbf{d}}^{L^*(\varepsilon),\mathbf{d}^*(\varepsilon)} \big)$. There exists $\lambda > 0$ such that, letting $\hat{\theta}$ be a sample from the Gibbs posterior with inverse temperature $\beta = \lambda$, with probability at least $1-\delta$, $D(\ell_{\hat{\theta}}) - r^*$ is bounded by*

$$\inf_{\varepsilon > 0} \left\{ \varepsilon + \tilde{\mathcal{O}}\left(\frac{c_\varepsilon}{n} + \sqrt{\frac{c_\varepsilon(r_\varepsilon + \sqrt{r_\varepsilon/n})}{n}} \right) \right\},$$

where the infimum runs over $\varepsilon > 0$ such that $(L^*(\varepsilon), \mathbf{d}^*(\varepsilon)) \leq (L, \mathbf{d})$.

This theorem captures how the data complexity influences the excess risk of a sample from the Gibbs posterior: the infimum balances the level of excess risk with the complexity of achieving of the desired excess risk.

7 NEURAL SCALING LAWS

So-called "neural scaling laws" are empirical power law relationships between key quantities describing neural networks, including accuracy, model size, number data, compute, etc. (Hoffmann et al., 2022; Kaplan et al., 2020; Sengupta, Goel, and Chakraborty, 2025). With the cost of training runs reaching new heights, empirical "compute-optimal" neural scaling laws are known to have dictated the size of extremely large models and the number of data used to train them.

In this section, we study the scaling laws that arise from our theoretical bounds. We begin with numerical computations of the bounds, to understand the

[2] We will assume the minimum is achieved. If the excess risk condition is met by any (Q-quantized) network, then clearly the minimum is achieved (by finiteness). Quantization, conceivably, could undermine universal approximation guarantees, which might otherwise guarantee us that a network of sufficient size can achieve the excess risk condition (Yarotsky, 2018; Tabuada and Gharesifard, 2023; Hornik, Stinchcombe, and White, 1989; Hornik, 1993; Stinchcombe and White, 1989; Scarselli and Chung Tsoi, 1998). We discuss quantization at the end of the paper.

landscape. We then derive analytical scaling laws in several regimes. Finally, we compare our scaling laws to one's estimated empirically. Up to some coarse approximations, discussed in Section 7.3, we argue that this comparison provides us lower bounds on the complexity of the real world data underlying known scaling laws.

7.1 Synthetic Neural Scaling Laws

In this section, we present numerical scaling laws that we compute from our theoretical bounds. In order to do so, we rely on two ingredients: The first ingredient is the data complexity function $\mathcal{C}(\varepsilon)$, which plays the same role as data does in the empirical estimation of neural scaling laws. The second ingredient is, effectively, a choice for how to scale the number of data, n, with the size of the model, which we will represent by the number of (student) parameters and denote by m. In this section, we adopt so-called compute optimal scalings that have appeared in the literature (Kaplan et al., 2020; Hoffmann et al., 2022).

Fix a data:parameter scaling $m = (n/n_0)^\alpha$ and complexity function $\mathcal{C}(\varepsilon)$. In the following simulations, we perform the following steps: For $n = 10^k$ data, we compute $m = (n/n_0)^\alpha$ and then compute the least width d that has at least m parameters, assuming we have d_0 input dimensions, $L-1$ fully connected layers of width d, and d_L output dimensions. We then optimize ε in Theorem 6.2 to determine the bound, recording the value of ε achieving the minimum, the size of the teacher $\mathcal{C}(\varepsilon)$, and the excess size of the student (the "B" term in c_ε).[3]

We visualize the consequences of $n \propto m$ data:parameter scaling and $\mathcal{C}(\varepsilon) = (1/\varepsilon)^p$ for $p = 9$ in Figure 2. We see that optimizing the balance between the data complexity and teacher's excess risk produces a power law relationship between loss and sample size, similarly as in neural scaling laws. We also observe a power law relationship between other quantities: number of parameters of the optimal teacher and n, teacher's excess risk and n, and number of parameters of the student and n.

7.2 The Growth of Data Complexity

Given Theorem 6.2, we ultimately want to know the rate at which the infimum term vanishes. In the previous section, we used numerical computation to explore the scaling laws induced by our theoretical bounds. Here, we attempt to derive closed form solutions for the rates.

One observation is that contribution of the "B" term to c_ε is of smaller order than that of $\mathcal{C}(\varepsilon)$. Motivated by this, we consider the terms that may dominate in the infimum. If the Bayes error is zero, the term $\sqrt{\mathcal{C}(\varepsilon)\varepsilon/n}$ would appear to dominate. If the Bayes error is strictly positive, this same term behaves like $\sqrt{\mathcal{C}(\varepsilon)/n}$. The following lemma determines the rates when each of these terms dominates. (See Appendix F for the proof.)

Lemma 7.1. *The infimum*

$$\inf_{\varepsilon>0}\left\{\varepsilon + c\sqrt{\frac{\mathcal{C}(\varepsilon)}{n}}\right\} \tag{15}$$

is in the following equivalence classes:

$$\begin{cases} \mathcal{O}\left(n^{-1/2}\right), & \text{if } \mathcal{C}(\varepsilon) \in \mathcal{O}\left(\mathrm{polylog}(\varepsilon^{-1})\right), \\ \mathcal{O}\left(n^{-1/(p+2)}\right), & \text{if } \mathcal{C}(\varepsilon) \in \mathcal{O}\left(\varepsilon^{-p}\right),\ p>0, \\ \mathcal{O}\left(\log^{-1} n\right), & \text{if } \mathcal{C}(\varepsilon) \in \mathcal{O}\left(\exp(\mathrm{poly}(\varepsilon^{-1}))\right). \end{cases} \tag{16}$$

Further, when $\mathcal{C}(\varepsilon) \in \mathcal{O}(\varepsilon^{-p})$, $p>0$, $\mathcal{O}\left(n^{-1/(p+1)}\right)$ contains

$$\inf_{\varepsilon>0}\left\{\varepsilon + c\sqrt{\frac{\mathcal{C}(\varepsilon)\varepsilon}{n}}\right\}, \inf_{\varepsilon>0}\left\{\varepsilon + c\,\frac{\mathcal{C}(\varepsilon)}{n}\right\}. \tag{17}$$

Inspecting Figure 2, we see that the risk bound rate for $p = 9$ is -0.1, which coheres with our analysis above which predicts the rate to be $-1/(p+1)$.

Theorem 6.2 and Lemma 7.1 shows that, if the data complexity stops growing or grows very slowly, then provided the student network is big enough, we get a parametric $n^{-1/2}$ rate. We get slower (nonparametric) rates when the teacher must grow more quickly in size to drive down excess risk.

7.3 Comparison with Empirical Neural Scaling Laws

In this section, we compare the empirical scaling law provided by Hoffmann et al. (2022) to those predicted by our theory, in order to draw conclusions about the complexity of real world data, as captured by our notion of data complexity, $\mathcal{C}(\varepsilon)$. (See Appendix G for mathematical expressions of the scaling laws in these publications.)

This comparison requires several coarse approximations. Four severe ones are: Our approximation of training by Gibbs posteriors, which we motivate but which is undoubtedly imperfect; the fact that the architectures

[3] To optimize over ε, we reparametrize, optimizing teacher width $d^* = \mathcal{C}(\varepsilon)$ using binary search over $\{1, \dots, d\}$, where d is the student width. We optimize Equation (29), presented in Appendix E, taking $q^* = \varepsilon = \mathcal{C}^{-1}(d^*)$, $d^*_{\delta/2} = c_\varepsilon$, and $\delta = 0.05$.

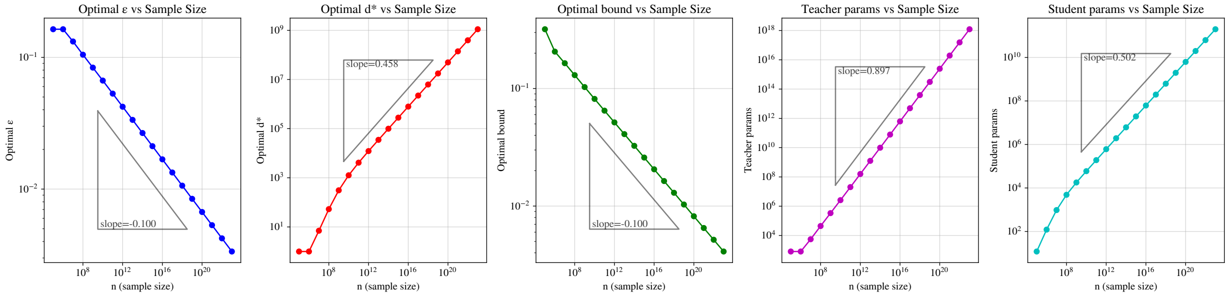

Figure 2: Scaling laws for $\mathcal{C}(\varepsilon) = (1/\varepsilon)^p$ for $p = 9$, data:param scaling $n \propto p$, and Bayes error rate $r^* = 0$. Values are computed numerically based on Theorem 6.2.

we study are distinct from Transformer architectures; that tokens are not i.i.d., and finally our reliance on quantization. In defense of this comparison, we note that the bulk of a Transformer's parameters are in the fully connected layers and that quantization is widely used.

In order to make this comparison, we once again require two key ingredients, namely a choice for data to parameter scaling, as well as a functional form for the data complexity $\mathcal{C}(\varepsilon)$.

While any scaling of data-to-parameter can conceivably shed light on $\mathcal{C}(\varepsilon)$ using our bounds, we adopt their compute-optimal scaling, which scales data and parameters equally, i.e., $m \propto n$.

Using reported values of the estimated exponents for the scaling laws, we computed the rate of decrease in excess risk with the number of data n. In (Hoffmann et al., 2022), the rate is approximately -0.1.

Motivated by the analysis in the prior section, we take $\mathcal{C}(\varepsilon) \in O(\varepsilon^{-p})$. We then tune p to obtain the above rate. We find that $p = 9$ results in a rate of excess risk of -0.1, if we assume the Bayes error rate is zero. (See Figure 2, yet again.) For $p < 9$ (i.e., for an easier data distribution), our theory predicts a faster rate. (For a positive Bayes error rate (0.05), $p = 8$ matched the rate for excess risk, in agreement with Lemma 7.1.)

What can we conclude? Putting our coarse approximations aside, the fact that our upper bound predicts a rate of excess risk of 0.1 for $p = 9$ and faster rates for $p < 9$ implies that the data underlying the empirical scaling law is *at least as hard* as $\mathcal{C}(\varepsilon) = \varepsilon^{-9}$.

There are by now numerous empirical scaling laws we might have compared to. We highlight two cases where comparisons fail to offer any insight. In (Kaplan et al., 2020), they scale parameters to data as $m \propto n^{2.7}$. At this rate, even the number of neurons in the student architecture outpaces the number of data, n. As such, our risk bounds grow without bound. One work around is to adopt a different scaling but to rely on their scaling law for risk (in terms of n and m).

For a suite of standard imagine classification tasks, Rosenfeld et al. (2020) reported empirical fits to scaling laws with faster rates, around -0.7, under data-to-parameters scalings that were, at least, overparameterized (i.e., at least $m \propto n$). However, in overparametrized regimes, our risk bound can achieve a rate no faster than -0.5, since the number of student neurons is a lower bound on our complexity term. This gap points to a limitation in our analysis, which we share with Buzaglo et al. (2024).

In conclusion, by comparing our theoretical upper bounds to known empirical scaling laws, we can provide lower bounds on the complexity of the data underlying the empirical scaling laws. This conclusion, however, rests on coarse approximations, which renders our conclusions more suggestive, rather than definitive.

8 DISCUSSION AND LIMITATIONS

One of our key observations is that the sample complexity in the *agnostic case* scales with the size of the teacher in the same way as it does under realizability. In particular, Corollary 5.3 implies sample complexity bounds for achieving a guaranteed excess risk with respect to any fixed (but potentially distribution-dependent) small teacher.

We then demonstrate that by optimizing over teachers, our excess risk bounds can be expressed as a scaling law, which we connect to empirically studied neural scaling laws. This allows us to make an estimate of the complexity of the data. Our work employs Gibbs posteriors as a model for SGD. SGD may achieve superior rates, which would affect the estimation of the data complexity in these vision and language tasks.

Note that the current approach of lower bounding the probability of interpolation is quite naive, as it only accounts for the probability of sampling the exact teacher network as a student subnetwork. The current approach does not account for permutation or scaling symmetries, nor the possibility of different parameters leading still to the same predictor under D. A large body of work shows redundancy of neurons per given layer (Humayun, Balestriero, and Baraniuk, 2024), supported by the fact that networks are highly prunable (e.g., (Frankle et al., 2019; Frantar and Alistarh, 2023) and many others) and sparsely activated (Li et al., 2023). In addition, recent work shows redundancy and prunability of *entire layers* (Gromov et al., 2025), and competitive performance under a type of depth sparsity (Raposo et al., 2024). All of these observations could serve to produce improved bounds, and get rid of the need for a quantized teacher. The latter is one of the key limitations in the current analysis.

Another drawback of the excess risk bounds presented in our work and (Buzaglo et al., 2024) is that the bounds are tightest when the student size does not exceed the teacher size. While the penalty from having a larger student is only linear in terms of extra neurons per layer and extra layers (Theorem 4.2), the bound still deteriorates with this additional size. One reason for having the student larger than the teacher may be to ease optimization (Jacot, Gabriel, and Hongler, 2018; Du et al., 2019; Mei, Montanari, and Nguyen, 2018). Further, some empirical work on neural scaling laws suggests an aggressive parameter to data scaling to be compute-optimal (Kaplan et al., 2020). We note that our bounds become vacuous in a setting where student's parameters grow faster than n^2.

Nagarajan and Kolter (2019) identify obstacles to proving tight generalization bounds for SGD in deep learning through PAC-Bayes and other uniform convergence bounds. Our analysis side steps this in part because we take Gibbs posteriors to serve as an idealized model and we choose a moderate temperature $\beta = \sqrt{n}$, which, in practice, overfits much less than SGD. The analysis of the Gibbs posterior reveals that we get to make a comparison of our risk to the empirical risk of a fixed teacher, and so we can swap the empirical risk and risk here without invoking uniform convergence. Previous work has suggested to relate the risk of a surrogate and the predictor of interest, in order to sidestep issues coming from the failure of uniform convergence (Negrea, Dziugaite, and Roy, 2020).

We further discuss related work on viewing SGD as sampling, the strong lottery ticket hypothesis, and neural scaling laws in Appendix B.

ACKNOWLEDGEMENT

We thank Jeffrey Negrea, Idan Attias, and Ioannis Mitliagkas for providing early feedback on this work. Daniel M. Roy is supported in part by an NSERC Discovery Grant and a Canada CIFAR AI Chair at the Vector Institute.

References

S. Ahmad and G. Tesauro (1988). “Scaling and Generalization in Neural Networks: A Case Study”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky. Vol. 1. Morgan-Kaufmann. URL.

P. Alquier (2024). “User-friendly Introduction to PAC-Bayes Bounds”. *Foundations and Trends in Machine Learning* 17.2, pp. 174–303. DOI: 10.1561/2200000100. arXiv: 2110.11216. URL.

Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma (2024). “Explaining neural scaling laws”. *Proceedings of the National Academy of Sciences* 121.27, e2311878121. DOI: 10.1073/pnas.2311878121. arXix: 2102.06701. URL.

G. Buzaglo, I. Harel, M. S. Nacson, A. Brutzkus, N. Srebro, and D. Soudry (2024). “How Uniform Random Weights Induce Non-uniform Bias: Typical Interpolating Neural Networks Generalize with Narrow Teachers”. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 5035–5081. arXiv: 2402.06323. URL.

O. Catoni (2007). “PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning”. *IMS Lecture Notes Monograph Series* 56, pp. 1–163. DOI: 10.1214/074921707000000391.

P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina (2019). “Entropy-SGD: biasing gradient descent into wide valleys*”. *Journal of Statistical Mechanics: Theory and Experiment* 2019.12, p. 124018. DOI: 10.1088/1742-5468/ab39d9. URL.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai (2019). “Gradient Descent Finds Global Minima of Deep Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 1675–1685. URL.

G. K. Dziugaite and D. Roy (2018). “Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1377–1386. URL.

J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin (2019). *Stabilizing the Lottery Ticket Hypothesis*. arXiv: 1903.01611.

E. Frantar and D. Alistarh (2023). "SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 10323–10337. URL.

A. Gromov, K. Tirumala, H. Shapourian, P. Glorioso, and D. Roberts (2025). "The Unreasonable Ineffectiveness of the Deeper Layers". In: *The Thirteenth International Conference on Learning Representations*. arXiv: 2403.17887. URL.

P. Grünwald (2011). "Safe Learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity". In: *Proceedings of the 24th Annual Conference on Learning Theory*. Ed. by S. M. Kakade and U. von Luxburg. Vol. 19. Proceedings of Machine Learning Research. Budapest, Hungary: PMLR, pp. 397–420. URL.

K. He, X. Zhang, S. Ren, and J. Sun (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. Rae, and L. Sifre (2022). "Training compute-optimal large language models". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 30016–30030. arXiv: 2203.15556. URL.

K. Hornik (1993). "Some new results on neural network approximation". *Neural Networks* 6.8, pp. 1069–1072. DOI: 10.1016/S0893-6080(09)80018-X. URL.

K. Hornik, M. Stinchcombe, and H. White (1989). "Multilayer feedforward networks are universal approximators". *Neural Networks* 2.5, pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8. URL.

I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio (2018). "Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations". *Journal of Machine Learning Research* 18.187, pp. 1–30. URL.

A. I. Humayun, R. Balestriero, and R. Baraniuk (2024). "Deep Networks Always Grok and Here is Why". In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 20722–20745. arXiv: 2402.15555. URL.

A. Jacot, F. Gabriel, and C. Hongler (2018). "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. URL.

O. Kallenberg (2002). *Foundations of modern probability*. Second. Probability and its Applications (New York). Springer-Verlag, New York, pp. xx+638. DOI: 10.1007/978-1-4757-4015-8. URL.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). *Scaling laws for neural language models*. arXiv: 2001.08361.

J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington (2019). "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. URL.

Z. Li, C. You, S. Bhojanapalli, D. Li, A. S. Rawat, S. J. Reddi, K. Ye, F. Chern, F. Yu, R. Guo, and S. Kumar (2023). "The Lazy Neuron Phenomenon: On Emergence of Activation Sparsity in Transformers". In: *The Eleventh International Conference on Learning Representations*. arXiv: 2210.06313. URL.

E. Malach, G. Yehudai, S. Shalev-Schwartz, and O. Shamir (2020). "Proving the Lottery Ticket Hypothesis: Pruning is All You Need". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 6682–6691. URL.

S. Mandt, M. D. Hoffman, and D. M. Blei (2017). "Stochastic Gradient Descent as Approximate Bayesian Inference". *Journal of Machine Learning Research* 18.134, pp. 1–35. URL.

S. Mei, A. Montanari, and P.-M. Nguyen (2018). "A mean field view of the landscape of two-layer neural networks". *Proceedings of the National Academy of Sciences* 115.33, E7665–E7671. DOI: 10.1073/pnas.1806579115. URL.

C. Mingard, G. Valle-Pérez, J. Skalse, and A. A. Louis (2021). "Is SGD a Bayesian sampler? Well, almost". *Journal of Machine Learning Research* 22.79, pp. 1–64. URL.

V. Nagarajan and J. Z. Kolter (2019). "Uniform convergence may be unable to explain generalization in deep learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. URL.

J. Negrea, G. K. Dziugaite, and D. Roy (2020). "In Defense of Uniform Convergence: Generalization via Derandomization with an Application to Interpolating Predictors". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 7263–7272. URL.

J. Negrea, J. Yang, H. Feng, D. M. Roy, and J. H. Huggins (2023). *Tuning Stochastic Gradient Algorithms for Statistical Inference via Large-Sample Asymptotics*. arXiv: 2207.12395.

V. Ramanujan, M. Wortsman, A. Kembhavi, A. Farhadi, and M. Rastegari (2020). "What's Hidden in a Randomly Weighted Neural Network?" In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11890–11899. DOI: 10.1109/CVPR42600.2020.01191.

D. Raposo, S. Ritter, B. Richards, T. Lillicrap, P. C. Humphreys, and A. Santoro (2024). *Mixture-of-Depths: Dynamically allocating compute in transformer-based language models*. arXiv: 2404.02258.

J. S. Rosenfeld, A. Rosenfeld, Y. Belinkov, and N. Shavit (2020). "A Constructive Prediction of the Generalization Error Across Scales". In: *International Conference on Learning Representations*. arXiv: 1909.12673. URL.

F. Scarselli and A. Chung Tsoi (1998). "Universal Approximation Using Feedforward Neural Networks: A Survey of Some Existing Methods, and Some New Results". *Neural Networks* 11.1, pp. 15–37. DOI: https://doi.org/10.1016/S0893-6080(97)00097-X. URL.

A. Sengupta, Y. Goel, and T. Chakraborty (2025). "How to Upscale Neural Networks with Scaling Law? A Survey and Practical Guidelines". *arXiv preprint arXiv:2502.12051*.

M. Stinchcombe and H. White (1989). "Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions". In: *Proceedings of the International 1989 Joint Conference on Neural Networks*. Vol. 1, pp. 613–617. DOI: 10.1109/IJCNN.1989.118640.

P. Tabuada and B. Gharesifard (2023). "Universal Approximation Power of Deep Residual Neural Networks Through the Lens of Control". *IEEE Transactions on Automatic Control* 68.5, pp. 2715–2728. DOI: 10.1109/TAC.2022.3190051.

M. Tan and Q. Le (2019). "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 6105–6114. URL.

G. Valle-Pérez, C. Q. Camargo, and A. A. Louis (2019). "Deep learning generalizes because the parameter-function map is biased towards simple functions". In: *International Conference on Learning Representations*. arXiv: 1805.08522. URL.

N. Vyas, Y. Bansal, and P. Nakkiran (2022). *Limitations of the NTK for understanding generalization in deep learning*. arXiv: 2206.10012.

M. Welling and Y. W. Teh (2011). "Bayesian learning via stochastic gradient Langevin dynamics". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, pp. 681–688. URL.

D. Yarotsky (2018). "Optimal approximation of continuous functions by very deep ReLU networks". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, pp. 639–649. URL.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. YES
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. N/A
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. N/A

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. YES
   (b) Complete proofs of all theoretical results. YES
   (c) Clear explanations of any assumptions. YES

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). N/A
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). N/A
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). N/A
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). N/A

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. N/A
   (b) The license information of the assets, if applicable. N/A
   (c) New assets either in the supplemental material or as a URL, if applicable. N/A
   (d) Information about consent from data providers/curators. N/A
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. N/A

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. N/A
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. N/A
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. N/A

A FREQUENTLY ASKED QUESTIONS

Q: What's the novelty of this work?

The fundamental goal in this line of work is to explain the empirical success of deep learning. Buzaglo et al. (2024) made important progress by showing that if we measure data complexity appropriately, we can explain why large networks can still generalize well.

Our paper's contribution is not to tell a completely different story—indeed, that would be counterproductive when the target phenomenon (generalization in deep learning) remains fixed. Rather, our contribution is to go beyond realizability and recognize that we can measure data complexity via the function $\mathcal{C}(\varepsilon)$ specifying the size of a network needed to achieve ε excess risk. We then go another step to connect our bounds to empirical scaling laws.

While our analysis builds on established (if niche) theoretical tools, one of the key innovations was identifying that the Gibbs posterior preserves the benefits of having a small teacher without requiring perfect interpolation. This connection was not obvious a priori and required careful technical development.

Q: Why use a posterior sampling setup to model the learning process in modern ML?

We use posterior sampling, specifically Gibbs posteriors, as a theoretical model of SGD. This choice is motivated by both theoretical results (e.g., (Mandt, Hoffman, and Blei, 2017; Negrea, Yang, et al., 2023)) and empirical observations (Valle-Pérez, Camargo, and Louis, 2019; Mingard et al., 2021) that suggest the output distribution of SGD can be well-approximated by samples from a posterior or Gibbs posterior under certain conditions. While no model is perfect, this approach allows us to leverage powerful statistical techniques to analyze the generalization behavior of neural networks trained with SGD. It provides a bridge between the practical training process and rigorous theoretical analysis, offering insights at a level of generality that is not possible to obtain by directly analyzing SGD with available tools.

Q: How does this work relate to or explain existing scaling laws in deep learning?

Our work provides a novel theoretical perspective on scaling laws in deep learning. We derive scaling laws that depend on the complexity of the decision boundary, measured in terms of the size of teacher networks (the $C(\varepsilon)$ function). This leads to an expression for excess risk that aligns with empirically observed scaling laws under certain assumptions about $C(\varepsilon)$. Specifically, if we assume $C(\varepsilon)$ scales polynomially with $1/\varepsilon$, we obtain scaling laws that match those observed empirically by Hoffmann et al. (2022). This highlights that empirical scaling laws reflect the complexity of the underlying decision boundaries in modern machine learning tasks.

Q: Does the bound contradict empirical observations that larger networks often generalize better?

No, our bounds do not contradict these empirical observations. Note that, logically, upper bounds provide information *only when they are small*, not when they are large. (One would need a lower bound in that case.) A large upper bound doesn't predict poor generalization; it is simply uninformative. It is tempting, but misguided, to interpret a large upper bound as predicting poor performance. Our bounds suggest that good generalization can occur when the effective complexity (measured by teacher size) is small relative to the dataset size, even if the student network is much larger. This aligns with observations in practice where large networks can learn simple functions and generalize well. On the other hand, our theory is silent in certain regimes, such as the neural tangent kernel limit. We know that training fundamentally changes in this limit, which is an effect we do not model with our study of Gibbs posteriors.

For example, in NTK scaling limits, one obtains a kernel method in the limit, for which one can apply benign overfitting style analyses of risk under interpolation. However, the limiting kernel method cannot perform feature learning and is strictly worse than finite neural networks (see, e.g., (Vyas, Bansal, and Nakkiran, 2022) and theoretical work by Lee et al. (2019)). More standard configurations of networks do overfit at large size (see Fig. 9 of (Kaplan et al., 2020)). This conflicting behavior might possibly block general bounds from guaranteeing vanishing error under massive overparametrization. It is an interesting direction for further inquiry.

In our opinion, the fact that our bounds become uninformative in this regime is reflective of the current scope of our theory (and (Buzaglo et al., 2024)): both our theories can only explain how moderate levels of overparametrization do not hurt rates. Neither theory can explain the sometimes-beneficial aspects of overparametrization. It must be emphasized, however, that *no current theory* explains generalization *in anything close to full generality* – the NTK regime is too restrictive to capture important deep learning phenomena like feature learning, and mean field (feature learning) limits have only been analyzed under toy data assumptions. One of the key missing ingredients is ways of talking about the easiness of data and this is perhaps the most important contribution of our work.

Our argument connects to the broader question of how to interpret generalization bounds. Mathematical theories of generalization must be evaluated with reference to some clear "scope" – the range of situations and interventions one aims to understand. We believe a productive approach towards explaining deep learning phenomena is to develop theories that offer meaningful insight within relevant scopes. Our framework offers a general, abstract argument for understanding generalization in an important regime (moderate generalization), even if it doesn't explain learnability in the full range of typical parameter settings.

Q: Can the analysis be extended to unbounded loss functions or unquantized networks?

For unbounded loss functions, there are no general results without additional assumptions. While this is an interesting direction for future work, we focused on the bounded loss case as there's still much to understand even in this setting. Our use of quantization simplifies the analysis and allows us to handle deeper networks, though we acknowledge this as a limitation of our current approach. At the same time, it is worth noting that in practice trained neural networks are highly-compressible using quantization and pruning, justifying the use of quantization for the theoretical analysis (Hubara et al., 2018).

Q: How does quantization affect the bounds, and is there a trade-off with the value of Q?

Quantization appears in our bounds through a $\log Q$ term, where Q is the number of quantization levels. Larger Q makes our bounds worse, suggesting we should use small Q. However, there's a trade-off: if we quantize too aggressively (small Q), it becomes harder to find a good teacher network that approximates the true decision boundary well. Using `float32` numbers ($Q = 2^{32}$) introduces a factor of about 22 in the bound, which can be offset by an increase in the number of data. Modern quantization techniques often use much fewer bits (e.g., 4-bit quantization), which would result in even smaller penalties in our bounds.

Q: Are the bounds tight enough to characterize the relationship between student network size and risk empirically?

Our bounds are primarily intended to provide theoretical insights rather than precise empirical predictions. They capture important qualitative relationships, such as how generalization depends on the ratio of effective model complexity to dataset size. However, they may not be tight enough to precisely characterize the quantitative relationship between student size and risk in all regimes. That said, we have demonstrated scenarios where our bounds are non-vacuous for realistic network sizes on the MNIST dataset. While our bounds clearly do not explain many empirical phenomena about network size and generalization, they do provide meaningful and testable predictions in certain regimes.

Q: Can the analysis be extended to multi-class classification or regression tasks?

The key requirement for our bounds is that the loss function is bounded, which is true for standard multi-class classification losses and can be ensured for regression tasks by clipping the predictions. The main ideas of our approach - using Gibbs posteriors, relating student performance to teacher network size, and measuring data complexity through the $C(\varepsilon)$ function – all generalize naturally to these settings. The specific constants in the bounds might change, but the overall form and scaling behaviors would remain similar.

Q: What are the implications of measuring data complexity using teacher network size?

Measuring data complexity using teacher network size provides a novel and intuitive way to characterize the difficulty of learning problems, and one that immediately yields rates on risk, via our theorems. It suggests that the complexity of a task is bounded by the size of the smallest network that can approximate the true decision boundary well, irrespective of other properties of the data distribution. This perspective provides a theoretical justification for why overparameterized models can generalize well when learning "simple" functions.

Q: How does this work connect to concepts like the minimum description length principle?

Our work has connections to the minimum description length (MDL) principle, although we don't explicitly use the MDL formalism. The idea of measuring data complexity through the size of teacher networks is analogous to measuring the complexity of data through the length of its description (in this case, the description is the teacher network). The connection between PAC-Bayes and MDL has been explored in previous work, notably by Grünwald (2011). Our use of PAC-Bayes techniques thus inherits some of these connections. In essence, our approach can be viewed as finding a trade-off between the complexity of the model (teacher size) and its fit to the data, which is a core principle of MDL.

B RELATED WORK

SGD as sampling. The distribution over neural networks induced by training with stochastic gradient descent (SGD) has been characterized in certain regimes. Based on a Gaussian assumption, Mandt, Hoffman, and Blei (2017) show that constant step size SGD corresponds to the Bayesian posterior. Negrea, Yang, et al. (2023) rigorously establish that this Gaussian assumption holds in a large data limit. Further, they verify a number of the consequences derived in (Mandt, Hoffman, and Blei, 2017), but also correct some mistaken deductions and derive yet additional consequences. More is known for variants of SGD. Welling and Teh (2011) proposed stochastic gradient Langevin dynamics algorithm (SGLD), a modification of SGD that adds Gaussian noise to the gradients. Under a decreasing step size, they establish that one obtains samples from the Bayesian posterior distribution.

Sparse subnetworks and the strong lottery ticket hypothesis. Ramanujan et al. (2020) conjectured that within a sufficiently large randomly initialized network, there exists a subnetwork that, without any training, achieves competitive performance to the original large trained network. The authors present an empirical study of finding such subnetworks. Malach et al. (2020) did a theoretical analysis of this phenomena, naming it the strong lottery ticket hypothesis. This line of work can be seen as using a uniform prior to sample over subnetworks of a randomly initialized network. The main results were focused on identifying the needed size of the student network such that the “target” network exists as a subnetwork in the student with high probability.

Scaling Laws. Neural scaling laws describe the empirical phenomenon that the performance of neural networks tends to improve as the size of the model (number of parameters) or the amount of training data increases (Kaplan et al., 2020; Hoffmann et al., 2022; Ahmad and Tesauro, 1988; Rosenfeld et al., 2020; Tan and Le, 2019). The scaling laws often exhibit power law relationships, where performance improves proportionally to a power of the model size or data size.

Recent work has focused on understanding the theoretical underpinnings of these scaling laws, with connections to statistical mechanics and information theory. Bahri et al. (2024) explored the role of model complexity and data distribution on the scaling laws. The authors empirically found that varying the data distribution by switching datasets or adding noise has strong effects on the exponent in the scaling law fit, while superclassing the labels in a classification task did not.

C EMPIRICAL STUDY

In this section, we use empirical methods to evaluate whether Corollary 5.3 is strong enough to guarantee generalization for Gibbs posteriors on real data sets. We will show that Corollary 5.3 can produce nonvacuous risk bounds for MNIST based on a teacher network we have found by training with a small architecture.

C.1 Methods

Corollary 5.3 bounds the risk of a network sampled from the Gibbs posterior by the risk and complexity of other network. And so, to obtain a nonvacuous bound, we simply need to establish the existence of a low-risk, small network.

Our approach will be to use neural network training itself to find small networks. Note that we are allowed to find these low-risk, small networks any way we like. Indeed, due to our oracle inequality hold for all networks simultaneously, *we can go in search of a teacher network using the same data we intend to use in our bound*, provided that the risk of the teacher network is estimated on a separate (validation) set (or we combine the failure probabilities appropriately. Of course, our bounds also tell us about performance on a hypothetical independent training run with (more or less) data from the same distribution.

We directly attempt to train a small network, which is feasible on MNIST. For more complex networks, it would be interesting to combine distillation, pruning, and a host of other approaches to finding small, low-risk teacher networks. These techniques were, evidently, not necessary to obtain a nonvacuous bound for MNIST.

To simplify the reproducibility of our methods, we start from publicly available JAX code for training a MLP on the MNIST dataset.[4] To meet our needs, we need only change the model (lines 45–50) to

```
mlp = hk.Sequential([
    hk.Flatten(),
    hk.Linear(3), jax.nn.relu,
    hk.Linear(NUM_CLASSES),
])
```

which defines a 2-layer MLP with 3 neurons in its single hidden layer and ReLU nonlinearities. (We also increase the number of iterations to 10^5.) On MNIST's 28x28 images, this network has 2388 parameters. The rest of the code trains the model on MNIST, starting from a random initialization.

C.2 Results

After 10^5 iterations, we obtain a network $\theta^\dagger$ with approximately 0.145 risk (i.e., 14.5% probability of misclassification), based on a validation-set estimate. The key observation is that every network—*including* $\theta^\dagger$—is a valid teacher network for Corollary 5.3. Given $\theta^\dagger$'s size and risk, we can compute a nonvacuous bound on the risk of the Gibbs posterior for datasets of sufficient size.

Having a bound for a non-realizable case is critical. An realizable bound does not apply to MNIST using $\theta^\dagger$ as a teacher because $\theta^\dagger$ does not achieve zero risk on MNIST. However, we can construct a distribution, close to MNIST, to which we *can* apply a realizable bound: Define a distribution—call it ALMOST-MNIST—on labelled images (X, y), where the images X have the same distribution as in MNIST, but the labels y are generated by $\theta^\dagger$. (In fact, we must quantize $\theta^\dagger$. If we take $Q = 2^{32}$ then we can exploit the fact that our `float32` training code was in fact quantized. We will assume that using `float8` numbers introduces negligible change to risk.) By construction, ALMOST-MNIST distribution has a total-variation distance of 0.15 to MNIST. Further, given an independent training sample from ALMOST-MNIST, a realizable bound tells you something nonvacuous: since there exists a teacher network ($\theta^\dagger$) with 2388 parameters and risk 0 on the ALMOST-MNIST distribution, we can pick a student architecture with $k \geq 3$ hidden nodes and obtain a (nonvacuous) better-than-chance risk bound for the hard posterior (a theoretical model of SGD) on $n = 50000$ training data provided $\log(2^8)(2388 + k)/n < 0.9$. For $k = 300$ hidden nodes (a 100x increase in size), we obtain, approximately, a 44% risk bound. Indeed, even at $k \approx 1900$, we improve on random guessing with high probability.

Note that such a realizable case bound does not apply to the actual MNIST. In contrast, our fast rate bound yields nonvacuous bounds for datasets that are not too much larger. To address quantization, we take $Q = 2^{32}$ to exploit the fact that our `float32` training code was in fact quantized. This costs us a factor of about 8 in our bound, over state of the art levels (4 bit quantization). In order to reach a better-than-chance bound for the Gibbs posterior, it suffices to train with 315,000 samples.

D PROOFS

D.1 Transformations of Catoni's Theorem

Proof of Corollary 2.3. Recall that $\Phi_a^{-1}(q) = (1 - \exp(-aq))/(1 - \exp(-a))$. For $\lambda > 0$, let $V_{\lambda^{-1}}$ be the event

$$p \leq \Phi_{\lambda/n}^{-1}\left(\frac{(d - \log\delta)}{\lambda}\right) = (1 - \exp(-(d - \log\delta)/n))/(1 - \exp(-\lambda/n)), \tag{18}$$

where d is the log Radon–Nikodym derivative. Since $\hat{p} = 0$ a.s., for all $\lambda > 0$, the event $V_{\lambda^{-1}}$ occurs with probability at least $1 - \delta$. Clearly, $V_{n+1} \subseteq V_n$ for all $n = 1, 2, \ldots$. And so, by the continuity and boundedness of probability measures (Kallenberg, 2002, Lem. 1.14), the probability of V_n converges monotonically to the probability of $V_\infty = \bigcap_{n\geq 1} V_n$ as $n \to \infty$. For every n, V_n has probability at least $1 - \delta$, and so then the same holds for V_∞. We then note that V_∞ is the event that, for all $\lambda > 0$, $p \leq (1 - \exp(-(d - \log\delta)/n))/(1 - \exp(-\lambda/n))$, hence, under this event, $p \leq (1 - \exp(-(d - \log\delta)/n)) \leq (d - \log\delta)/n$, as was to be shown. □

[4] https://github.com/google-deepmind/dm-haiku/blob/main/examples/mnist.py.

D.2 Bounding the probability of interpolation for gated activation and residual connnection neural networks

The following proof closely follows that of the analogous result by Buzaglo et al. (2024).

Proof of Theorem 4.2. Write $W^{(l)} = \begin{bmatrix} W_{11}^{(l)} W_{12}^{(l)} \\ W_{21}^{(l)} W_{22}^{(l)} \end{bmatrix}$, where $W_{11}^{(l)} \in \mathbb{R}^{d_l^* \times d_{l-1}^*}$. Similarly, for vectors $v \in \mathbb{R}^d$, write $v = [v_1 \ v_2]$, where $v_1 \in \mathbb{R}^{d^*}$. We first describe how we embed the teacher network into the wider and deeper student network, as illustrated in Figure 1. Define

$$\begin{aligned} E = \Big\{ h_\theta \in H_{LSFC} \,\Big|\, & \forall l \in [L^* - 1],\ W_{11}^{(l)} = W^{*(l)}, b_1^{(l)} = b^{*(l)}, b_2^{(l)} = 0_{d_l - d_l^*}, \\ & \qquad\qquad\qquad \alpha_1^{(l)} = \alpha^{*(l)}, \alpha_2^{(l)} = 0_{d_l - d_l^*}, \beta^{(l)} = \beta^{*(l)}; \\ & \forall l \in [L^*, L-1],\ \beta^l = 0; \\ & W_{11}^{(L)} = W^{*(L^*)}, b_1^{(L)} = b^{*(L^*)}, b_2^{(L)} = 0_{d_L - d_{L^*}^*}, \\ & \qquad\qquad\qquad \alpha_1^{(L)} = \alpha^{*(L^*)}, \alpha_2^{(L)} = 0_{d_L - d_{L^*}^*} \Big\}. \end{aligned} \tag{19}$$

A straightforward inductive argument (Lemma D.1 below) establishes that, for all $\theta \in E$ and all inputs x, we have $h_\theta(x) = h_{\theta^*}(x)$. Thus, $E \subseteq E_{\theta^*}$, and so it suffices to control $\pi(E)$.

In order to calculate $\pi(E)$, we need only account for the number of parameters in E with values constrained by θ^*. Based on the dimensionalities of the parameterization, it is easy to see that the total number of parameters is $M = L + \sum_{l=1}^{L^*} (d_l^* d_{l-1}^* + 2d_l)$. Since we are working with quantized networks, the total number of ways to set these parameters is Q^M. Thus $\pi(E) = Q^{-M}$. The result then follows. □

For completeness, we provide the inductive argument.

Lemma D.1. *For $\theta \in E$, $h_\theta = h_{\theta^*}$.*

Proof. We will proceed by induction over layers l, establishing that $f^{(l-1)}(x) = f^{*(l-1)}(x)$ for all inputs x.

We begin with the base case of $l = 1$: The first layer of the student network is active since $\beta^{(1)} \neq 0$. Since $\theta \in E$, the output of the first layer of the student is

$$f^{(1)}(x) = \beta^{(1)} \sigma(\alpha^{(1)} \odot W^{(1)} x + b^{(1)}) + x = \beta^{*(1)} \sigma(\alpha^{*(1)} \odot W^{*(1)} x + b^{*(1)}) + x = f^{*(1)}(x). \tag{20}$$

This matches the output of the first layer of the teacher network.

Assume now, for some layer $l-1$, where $1 < l < L$, we have $f^{(l-1)}(x) = f^{*(l-1)}(x)$. We need to show this holds for layer l as well. If $l < L^*$, the layer is active, and therefore

$$\begin{aligned} f^{(l)}(x) &= \beta^{(l)} \sigma(\alpha^{(l)} \odot W^{(l)} f^{(l-1)}(x) + b^{(l)}) + f^{(l-1)}(x) \\ &= \beta^{*(l)} \sigma(\alpha^{*(l)} \odot W^{*(l)} f^{*(l-1)}(x) + b^{*(l)}) + f^{*(l-1)}(x) \\ &= f^{*(l)}(x). \end{aligned} \tag{21}$$

If $l \geq L^*$ and $l < L$, then $\beta^{(l)} = 0$, and this layer is skipped. Thus

$$f^{(l)}(x) = f^{(l-1)}(x) = f^{*(l-1)}(x) = f^{*(L^*-1)}(x). \tag{22}$$

To finish, we note that the output is also the same

$$h_\theta(x) = \operatorname{sign}(W^{(L)} f^{(L-1)}(x) + b^{(L)}) = \operatorname{sign}(W^{*(L^*)} f^{*(L^*-1)}(x) + b^{*(L^*)}) = h_{\theta^*}(x). \tag{23}$$

□

E PROOF OF THEOREM 5.2

Let $d_\delta(\theta) = \log \frac{1}{\delta\,\pi(E_\theta)}$, $\hat{q}(\theta) = \hat{D}_n(\ell_\theta)$, and $q(\theta) = D(\ell_\theta)$. Using the fact that losses are in $[0,1]$, Bernstein's inequality implies that, for every $\theta^* \in \Theta$, writing $q^* = q(\theta^*)$, $\hat{q}^* = \hat{q}(\theta^*)$, and $d^*_\delta = d_\delta(\theta^*)$, we have, with probability at least $1-\delta$,

$$\hat{q}^* \le q^* + \frac{\log(1/\delta)}{3n} + \sqrt{\frac{2q^* \log(1/\delta)}{n}}. \tag{24}$$

Combining this with Equation (12), for every $\theta^* \in \Theta$, we have, with probability at least $1-\delta$,

$$\begin{aligned} D(\ell_{\hat{\theta}}) &\le \Phi^{-1}_{\lambda/n}\left[\hat{q}^* + d^*_{\delta/2}/\lambda\right] & (25)\\ &\le \Phi^{-1}_{\lambda/n}\left[q^* + \frac{\log(2/\delta)}{3n} + \sqrt{\frac{2q^* \log(2/\delta)}{n}} + d^*_{\delta/2}/\lambda\right]. & (26) \end{aligned}$$

We now proceed to invert Φ^{-1}. By Lemma 2.2, and a bit of simplification, we have, for all $v, d \ge 0$,

$$\inf_{\lambda \ge 0} \Phi^{-1}_{\lambda/n}\left[v + d/\lambda\right] \le v + \frac{2d}{n} + \sqrt{\frac{2dv}{n}}. \tag{27}$$

We will assume that there exists a value for λ satisfying this inequality. This assumption may not be valid when $v = 0$ or $d = 0$, but these situations will not be our focus.

We note that the r.h.s. of Equation (26) is nonrandom, and so, for every $\theta^* \in \Theta$, there exists a nonrandom value $\lambda = \lambda^*$ satisfying the above inequality due to Catoni. Thus, for every $\theta^* \in \Theta$, a sample $\hat{\theta}$ from the Gibbs posterior with inverse temperature $\beta = \lambda^*$ satisfies, with probability at least $1-\delta$,

$$\begin{aligned} D(\ell_{\hat{\theta}}) &\le \Phi^{-1}_{\lambda^*/n}\left[\underbrace{q^* + \frac{\log(2/\delta)}{3n} + \sqrt{\frac{2q^* \log(2/\delta)}{n}}}_{v} + d^*_{\delta/2}/\lambda^*\right] & (28)\\ &\le q^* + \frac{\log(2/\delta)}{3n} + \sqrt{\frac{2q^* \log(2/\delta)}{n}} + 2\frac{d^*_{\delta/2}}{n} + \sqrt{\frac{2d^*_{\delta/2}\left(q^* + \frac{\log(2/\delta)}{3n} + \sqrt{\frac{2q^* \log(2/\delta)}{n}}\right)}{n}}. & (29) \end{aligned}$$

Since θ^* was arbitrary, we may choose θ^* to minimize the (nonrandom) bound on the r.h.s. of Equation (29), meaning that $D(\ell_{\hat{\theta}})$ is upper bounded by the infimum of Equation (29) over θ^*, yield an oracle inequality. Summarizing that oracle inequality by dropping lower order terms and log terms, we have

$$D(\ell_{\hat{\theta}}) \le \inf_{\theta^*}\left\{q^* + \tilde{\mathcal{O}}\left(\frac{d^*_\delta}{n} + \sqrt{\frac{d^*_\delta(q^* + \sqrt{q^*/n})}{n}}\right)\right\}. \tag{30}$$

One final remark is that our Gibbs posterior depends on λ, which we have chosen in a nonconstructive way based on distribution-dependent quantities. One way to think about λ is as a hyperparameter that could be tuned.

F PROOF OF LEMMA 7.1

These are standard manipulations in, e.g., nonparametric statistics. We consider the case $\mathcal{C}(\varepsilon) \in \mathcal{O}\left(\mathrm{poly}(1/\varepsilon)\right)$. The other cases are similar. Then, for some $c > 0$ and $p > 0$, the infimum is bounded by

$$\inf_{\varepsilon > 0}\left\{\varepsilon + c\sqrt{\frac{\mathcal{C}(\varepsilon)}{n}}\right\} = \inf_{\varepsilon > 0}\left\{\varepsilon + c\,\frac{\varepsilon^{-\frac{p}{2}}}{\sqrt{n}}\right\}. \tag{31}$$

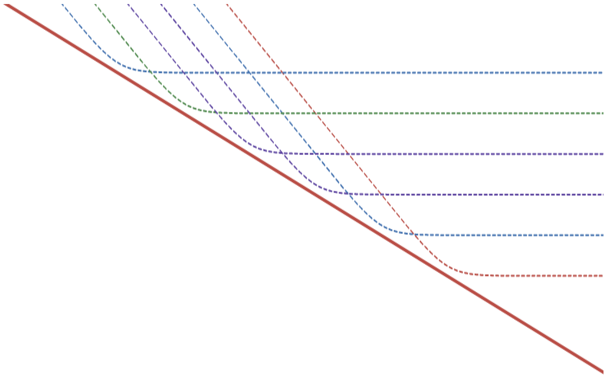

Figure 3: Excess risk (y-axis) as a function of sample size (x-axis), both in log scale. Dashed lines represent standard $n^{-1/2}$ rate excess risk bounds for a fixed network size, which saturate at the best achievable excess risk for that sized architecture. The oracle inequality chooses the right model based on the data set size and the complexity $\mathcal{C}(\varepsilon)$ of the data, as measured by the size of the smallest teacher that can achieve ε excess risk. The solid line represents the final rate achieved by the oracle inequality. In this case, it is a $n^{-1/4}$ rate, for $p = 2$.

From first order optimality conditions, we see the infimum is achieved when $\varepsilon^{\frac{p}{2}+1} = \frac{c\,p}{2\sqrt{n}}$. Since the two terms will have the same rate, the infimum has order $\varepsilon = \mathcal{O}\left(n^{-1/(p+2)}\right)$. See Figure 3 for a visualization in the case $\mathcal{C}(\varepsilon) \in \mathcal{O}\left(\text{poly}(1/\varepsilon)\right)$.

A related infimum arises from a Bernstein bound. Consider

$$\inf_{\varepsilon>0}\left\{\varepsilon + c\sqrt{\frac{\mathcal{C}(\varepsilon)\varepsilon}{n}}\right\} = \inf_{\varepsilon>0}\left\{\varepsilon + c\frac{\varepsilon^{\frac{-(p-1)}{2}}}{\sqrt{n}}\right\}. \tag{32}$$

We see that this is an instance of the previous problem, and so the rate is

$$\mathcal{O}\left(n^{-1/(p+1)}\right). \tag{33}$$

The same argumentation reveals that the rate of

$$\inf_{\varepsilon>0}\left\{\varepsilon + c\,\frac{\mathcal{C}(\varepsilon)}{n}\right\}. \tag{34}$$

is $\mathcal{O}\left(n^{-1/(p+1)}\right)$, which demonstrates that these last two quantities have the same rate.

G EXPRESSIONS OF EMPIRICAL SCALING LAWS IN PRIOR WORK

We examine a scaling law by Rosenfeld et al. (2020), for image classification with ResNet architectures, taking the form

$$D(\ell_{\hat{\theta}}) - r^* = A_n n^{-\alpha_n} + A_m m^{-\alpha_m}, \tag{35}$$

where m represents the total number of parameters in the network (i.e., the number of parameters in the student network), A_n, A_m are constants, and α_n, α_m are the exponents. This law shares a similar form with others in the literature, including those for large language models (Hoffmann et al., 2022).

In our work we also study scaling laws due to Kaplan et al. (2020) for large language models. This work expressed the test loss as a power law of the model size m and the dataset size n,

$$\left(\left(\frac{m_c}{m}\right)^{\alpha_m/\alpha_n} + \frac{n_c}{n}\right)^{\alpha_n}, \tag{36}$$

where m_c and n_c are constants. They investigated the scaling of the model size and the number of data points, and found that the optimal scaling between the two follows a power law: $m \propto n^{\alpha}$, with $\alpha \approx 2.7$.

More recent work by Hoffmann et al. (2022) refined the understanding of compute-optimal scaling. For a fixed compute budget, they have shown that the optimal ratio between model size and the number of training tokens is approximately 1:1 (corresponding to $\alpha = 1$). This result, referred to as "Chinchilla scaling", demonstrated that previous models were significantly undertrained, leading to the development of more efficient and performant LLMs. The functional form of the scaling law studied in (Hoffmann et al., 2022) is the same one as Equation (35).