Mitigating the Curse of Detail: Scaling Arguments for Feature Learning and Sample Complexity

Anonymous authors
Paper under double-blind review

Abstract

Two pressing topics in the theory of deep learning are the interpretation of feature learning mechanisms and the determination of implicit bias of networks in the rich regime. Current theories of rich feature learning effects revolve around networks with one or two trainable layers or deep linear networks. Furthermore, even under such limiting settings, predictions often appear in the form of high-dimensional non-linear equations, which require computationally intensive numerical solutions. Given the many details that go into defining a deep learning problem, this analytical complexity is a significant and often unavoidable challenge. Here, we propose a powerful heuristic route for predicting the data and width scales at which various patterns of feature learning emerge. This form of scale analysis is considerably simpler than such exact theories and reproduces the scaling exponents of various known results. In addition, we make novel predictions on complex toy architectures, such as three-layer non-linear networks and attention heads, thus extending the scope of first-principle theories of deep learning.

1 Introduction

There is a clear need for a better theoretical understanding of deep learning. However, efforts to construct such theories inevitably suffer from a "curse of details". Indeed, since any choice of architecture, activation, data measure, and training protocol affects performance, finding a theory with true predictive power that accurately accounts for all those details is unlikely. One workaround is to focus on analytically tractable toy models, an approach that can often uncover interesting fundamental aspects. However, analytical tractability is a fragile, fine-tuned property; thus, a large explainability gap remains between such toy models and more complex data/architecture settings.

An alternative approach focuses on scaling properties of neural networks, which appear more robust. Two well-established examples are empirically predicting network performance by extrapolating learning curves using power laws Kaplan et al. (2020); Hestness et al. (2017), and providing theory-inspired suggestions for hyperparameter transfer techniques Yang et al. (2022); Bordelon et al. (2023). Indeed, it is often the case Cardy (1996) that predicting scaling exponents is easier than predicting exact or approximate behaviors. As a simple toy model of this, consider the integral $\int_{-\infty}^{\infty} dx g(x/P)$. While $g(\cdot)$ needs to be fine-tuned for exact computations, a change of variable reveals a robust linear scaling with P for any $g(\cdot)$.

This work focuses on scaling properties of feature learning. Feature learning, or, more generally, interpretability, have been studied extensively both from the practical and theoretical side. On the practical side, mechanistic interpretability Bereska and Gavves (2024) has provided us with statistical explanations for why some predictions are made and the underlying decision mechanisms. On the theory side, kernel-based ap-

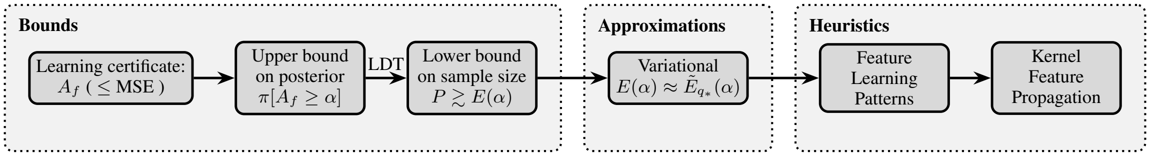
Bounds
Learning certificate: A_f ($\leq$ MSE)
Upper bound on posterior $\pi[A_f \geq \alpha]$
LDT
Lower bound on sample size $P \gtrsim E(\alpha)$
Approximations
Variational $E(\alpha) \approx \tilde{E}_{q_*}(\alpha)$
Heuristics
Feature Learning Patterns
Kernel Feature Propagation

Figure 1: Logical flow of sample complexity derivation. **Bounds**: **(i)-(iii)** deriving lower bounds on sample complexity using LDT (Sec. 3). **Approximations**: **(iv)** approximating the intractable lower bound (Sec. 4), **Heuristics**: **(v)-(vi)** providing heuristic methods for manually computing the approximated bound (Sec. 5). Each section is composed of intermediate steps as detailed in the diagram.

proaches Aitchison (2019); Li and Sompolinsky (2021); Aitchison (2021); Seroussi et al. (2023a); Ariosto et al. (2022); Bordelon and Pehlevan (2022); Rubin et al. (2024; 2025); Ringel et al. (2025) and Saad and Solla type approaches Saad and Solla (1995); Arnaboldi et al. (2023); Bietti et al. (2022) (and their Bayesian counterparts Cui (2025)) allow us to solve simple non-linear teacher-student networks in the rich regime. However, our ability to capture more elaborate and compositional feature learning effects, such as those involving depth and emergence, is hampered by said analytical difficulties.

In this work, we introduce a novel framework addressing the challenging task of making first-principles predictions on sample complexity and feature learning effects in networks trained to equilibrium The undeniable success of various deep learning models across diverse domains underscores the importance of theoretically predicting the conditions under which these models exhibit failure modes. Accordingly, a main focus of this work is lower-bounding sample complexity. Specifically, we aim to determine the scaling behavior of P_*, the threshold sample size at which learning becomes possible, as a function of input dimension, layer width, regularization, and parametrization choices (e.g., mean-field versus standard scaling).

Our Bayesian approach, capturing networks fully trained using SGLD [1] Mandt et al. (2017), is described schematically in Fig. 1. It consists of the following steps: **(i)** we lower bound the test MSE by $(1 - A_f)^2$, where A_f is the alignment of output and target; **(ii)** we establish an upper bound on the probability to observe good learning (ie, strong alignment $A_f \geq \alpha \approx 1$) in the posterior using the negative-log-probability of the rare event of good learning in the prior; **(iii)** we leverage an *upper* bound on the prior to obtain a *lower* bound $E(\alpha)$ on the minimal sample size necessary for alignment of at least α; **(iv)** we derive a variational approximation, $\tilde{E}_q(\alpha)$, for $E(\alpha)$ with an explicit formula using kernel-adaptation type approximations; **(v)** we propose *feature learning patterns* as heuristic variational probabilities q and choose such q that minimizes $\tilde{E}_q(\alpha)$; finally **(vi)** our fully analytic computation of $\tilde{E}_q(\alpha)$ further relies on heuristic scaling relations for how a feature amplified in one layer *propagates* to downstream layers.

Using the above, we re-derive, in a relatively straightforward manner, known results on two-layer networks: sample complexity benefits of rich learning and Grokking transitions. We demonstrate the power of this approach by expanding the scope of tractable models. Specifically, we study non-linear 3-layer networks in the rich regime and predict sample complexity, layer-wise location of learning, and scaling of the number of specializing neurons.

2 Setup

We consider here several types of feedforward networks, but, for the sake of clarity, we illustrate the main derivation on deep fully connected networks (FCNs) and, later, when we analyze specific problems, augment

[1] which can be thought of as a proxy for SGD Mingard et al. (2020)

it for convolutional neural networks (CNNs) as needed. Our FCNs are defined by

$$f(x) = \sum_{i=1}^{N_{L-1}} w_i^L \sigma(h_i^{L-1}(x)), \quad \text{where } h_i^{l>1}(x) = \sum_{j=1}^{N_{l-1}} W_{ij}^l \sigma(h_j^{l-1}(x)), \;\; h_j^1(x) = \sum_{k=1}^{d} W_{jk}^1 x_k, \tag{1}$$

where σ can be any activation function, and we refer to h_i^l's as *pre-acitvations*. We consider Bayesian neural networks, as Bayesian descriptions are a commonly used proxy for network behavior after long-time stochastic training Wilson and Izmailov (2020); Wilson (2020); Naveh et al. (2021). Alternatively, they represent an exact solution to Langevin dynamics with weight decay Welling and Teh (2011b). We denote the target function by y and the training sample size by P, and assume training with Mean Squared Error (MSE) loss. The quadratic weight decay for each layer l is set to $\kappa N_{l-1}/\sigma_l^2$, where κ is the ridge parameter and $N_0 = d$ is the input dimension. This choice of weight decay results in a Gaussian prior distribution for the weights given by $W_{i,j}^l \sim \mathcal{N}(0, \sigma_l^2/N_{l-1})$ for $i = 1, .., N_l$, $j = 1, ..., N_{l-1}$. The possible outputs f of such a network, given y and P, are then distributed according to the posterior:

$$\pi\left(f \mid y, P\right) = \frac{1}{Z} \exp\left(-\frac{1}{2\kappa} \sum_{\nu=1}^{P} \left[f(x_\nu) - y(x_\nu)\right]^2\right) p_0(f), \tag{2}$$

where Z is the normalization constant, $\{x_\nu\}_{\nu=1}^{P}$ is the training dataset of size P, and $p_0(f)$ is the prior defined as $p_0(f) = \int d\Theta p\left(\Theta\right) \delta\left[f - f_\Theta\right]$, determined by the weight decay. Here, Θ is the collection of all network weights, and $p\left(\Theta\right)$ corresponds to the density of the prior weight distribution, which we take to be Gaussian with a diagonal covariance, representing quadratic weight decay, and f_Θ is the network architecture with weights Θ. We further set $p(\Theta)$ such that pre-activations are all $\mathcal{O}(1)$ under the prior He et al. (2015). For classification tasks see App. A.2. As a measure for learning, we consider an observable, which we refer to as *alignment*, given by

$$A_f := \langle f, y\rangle / \langle y, y\rangle, \tag{3}$$

where $\langle g, h\rangle = \int d\mu_x g(x)h(x)$ is the functional inner product, and $d\mu_x$ is some test measure, which, conveniently, does not need to be the measure from which the training set was drawn. We similarly define $\langle g, K, h\rangle = \int d\mu_x d\mu_{x'} g(x)K(x, x')h(x')$ for any kernel K. Alignment represents the extent to which the network learns a function that is proportional to the target. It bounds the test MSE via the Cauchy–Schwarz inequality $\int d\mu_x(f(x) - y(x))^2 \geq \langle y, y\rangle(A_f - 1)^2$. Having $A_f \approx 1$ is thus a necessary condition for successful learning.

3 Alignment and Sample Complexity

We turn to analyze sample complexity via an upper bound on the probability of finding $A_f \geq \alpha$ for $\alpha \approx 1$. We begin with a theoretical bound on the posterior that mainly depends on the chance that a random network, chosen from the prior, produces an alignment of at least α. We denote the prior and posterior alignment probabilities by $\Pr_{p_0}\left[A_f \geq \alpha\right]$ and $\Pr_\pi\left[A_f \geq \alpha\right]$ respectively. Following simple arguments (see App. A and App. A.2 for generalization to classification.), we obtain the following bound on the log posterior [2]

$$\log\left(\Pr_\pi\left[A_f \geq \alpha\right]\right) < Pk/\left(2\kappa\right) + \log\left(\Pr_{p_0}\left[A_f \geq \alpha\right]\right), \tag{4}$$

where $k = P^{-1} \sum_{\nu=1}^{P} \mathbb{E}_{p_0}[(f(x_\nu) - y(x_\nu))^2]$ is the only training-set dependent quantity and is generally of order one. The Bayesian interpretation of successful learning is having $\Pr_\pi\left[A_f \geq \alpha\right] \approx \mathcal{O}(1)$. Since a random network is unlikely to achieve strong alignment, $\log \Pr_{p_0}\left[A_f \geq \alpha\right]$ would typically be highly negative for large α. Therefore, a sufficiently large data term is required to cancel this effect. Explicitly,

$$P \gtrsim -2\kappa \log \Pr_{p_0}\left[A_f \geq \alpha\right]/k. \tag{5}$$

[2] See also Lavie and Ringel (2025), for a similar data-agnostic bound in the context of lazy learning.

Thus, up to the ridge parameter and the $\mathcal{O}(1)$ factor k, depending on the training set, the log probability of prior alignment with the target lower bounds the sample complexity. Here, it is worth noting that the bound becomes tight when overfitting effects are small, which is typically the case for $k/\kappa \sim \mathcal{O}(1)$. Taking $\kappa \to 0$ encourages overfitting (though often benignly Bartlett et al. (2020)) and trivializes this bound. We conjecture that, in this case, κ should be kept $\mathcal{O}(1)$ based on the effective ridge treatment Canatar et al. (2021); Cohen et al. (2021); Bartlett et al. (2020). Establishing this conjecture is outside the scope of this work. From a PAC-Bayesian perspective, an analogous bound would require P to be much larger than the KL-divergence between the prior and posterior (e.g. McAllester (1999)).[3]. More recently, prior-posterior relations have been studied in the context of complex Boolean functions Mingard et al. (2025).

Following the Chernoff inequality, we can find an upper bound for the probability (and a lower bound for P) via

$$P \geq -2\kappa/k\ \log \mathrm{Pr}_{p_0}\left[A_f \geq \alpha\right] \geq 2\kappa/k\ E(\alpha), \quad E(\alpha) = -\log \inf_{t>0} e^{-t\alpha}\mathbb{E}_{p_0}\left[e^{tA_f}\right]. \tag{6}$$

Where we refer to $E(\alpha)$ as the *energy*. We can thus express the minimal sample size necessary for learning, P_*, through the energy as $P_* \propto E(\alpha)$. In App. B, we provide an asymptotically exact solution for $E(\alpha)$, and compute it explicitly for a two layer network . We also argue and demonstrate that our bound is inherently tied to feature learning. Indeed, a network sampled from the prior that achieves such alignment is a statistical outlier, driven by the emergence of an internal structure which mimics feature learning (see also Fig. 2). Nevertheless, such a direct LDT approach is computationally prohibitive in most cases of interest. We therefore introduce a heuristic LDT-based method for evaluating P_*. This method not only enables predicting the scaling of P_* but also the feature learning effects that lead to successful learning.

In this section, we adopt a variational approach to estimating P_* by comparing different modes of feature learning [4] under a certain loss (see (9) below). While many approaches predict different feature learning mechanisms Pacelli et al. (2023); Fischer et al. (2024); Meegen and Sompolinsky (2024); Buzaglo et al. (2025); Li and Sompolinsky (2021); Seroussi et al. (2023b); Rubin et al. (2025; 2024), they are often case-dependent, highly detailed, and complex. Thus, we propose a method that abstracts key feature learning mechanisms from these frameworks into distinct, comparable patterns.

4 Variational Analysis

Our next objective is to make the sample complexity bound tractable. This requires estimating the prior probability term, $\mathrm{Pr}_{p_0}\left[A_f \geq \alpha\right]$, for alignments $\alpha \approx 1$. As a first step, we simplify this by relating the cumulative distribution function to the probability density denoted by $p_{A_f}(\alpha)$. As shown in App. A.3, for large alignments, we have $E(\alpha) \approx -\log p_{A_f}(\alpha)$. This allows us to re-express P_* in terms of the density: $P_* = -2\kappa \log p_{A_f}(\alpha)/k$. However, computing $p_{A_f}(\alpha)$ directly remains intractable. We therefore turn to a variational approach to estimate it. As explained in the next section, we wish to express the variational probability density in terms of pre-activations h (1). Accordingly, in App. C.1, we follow standard statistical mechanics techniques to express this density as

$$p_{A_f}(\alpha) = \int Dh\ \exp[-H_{p,\alpha}(h)]/Z_p, \quad \text{where} \quad Z_p = \int_{-\infty}^{\infty} d\alpha \int Dh\ \exp[-H_{p,\alpha}(h)]. \tag{7}$$

We comment that the above equation explicitly constitutes a definition of $H_{p,\alpha}$, namely, for every α one can find a function $H_{p,\alpha}(h)$ and normalizing constant Z_p such that (7) holds. Further requiring that

[3] Following the data-processing-inequality, one can lower-bound the KL-divergence between the full prior and posterior probabilities by the KL-divergence of a coarser probability of an $A_f \geq \alpha$ event in the prior and posterior. The latter KL divergence is given by $-\log \mathrm{Pr}_{p_0}\left[A_f \geq \alpha\right]$

[4] Viewed here formally as emergent weight/pre-activation structures enabling the outlier.

$\min_h H_{p,\alpha}(h) = 0$ ensures the uniqueness of this definition. Similarly, the above implies a measure over h's given by $\exp[-H_{p,\alpha}(h)]/\int \mathcal{D}h \, \exp[-H_{p,\alpha}(h)]$. We next approximate this measure per α by an analytically tractable variational estimate, q. Here too we follow the same notation as in (7), so that for any q, α we define $q_\alpha(h) := e^{-H_{q,\alpha}(h)}/Z_{q,\alpha}$ where here $Z_{q,\alpha} = \int \mathcal{D}h e^{-H_{q,\alpha}(h)}$. The variational computation follows by looking for $q_\alpha(h)$ which minimizes the KL divergence between the measure on h which defines p_{A_f}, and q. The KL divergence can also be used in the estimation of $E(\alpha) := -\log(p_{A_f}(\alpha))$, following the Feynman–Bogoliubov inequality Kuzemsky (2015); Bogolubov and Jr (2009); Huber (1968). Here we provide a brief description – for the full derivation see App. C.2. By applying the Feynman–Bogoliubov inequality, we obtain an upper bound on $E(\alpha)$

$$E(\alpha) \approx \min_{q_\alpha}(\log(Z_p/Z_{q,\alpha}) + \tilde{E}_q(\alpha)), \qquad \tilde{E}_q(\alpha) = \mathbb{E}_{h\sim q_\alpha}[H_{p,\alpha}(h) - H_{q,\alpha}(h)] \tag{8}$$

We argue in App. C.4 that for $\alpha \approx 1$, the log terms are subleading w.r.t. $\tilde{E}_q(\alpha)$. Defining $q_{*,\alpha}$ to be the measure that minimizes $\tilde{E}_q(\alpha)$, we obtain $E(\alpha) \approx \tilde{E}_{q_*}(\alpha)$.

Next, we turn to estimating the variational energy $\tilde{E}_q(\alpha)$ Eq. (8) for $\alpha \sim 1$, omitting all α indices for brevity. In App. C.1 we simplify p_{A_f}, and show that the distribution in each layer depends only on the previous through a fluctuating non-linear operator. Next, we assume that this kernel is weakly fluctuating, and replace it with its expectation w.r.t. the variational distribution. This choice approximation aligns with various works on deep non-linear networks, where layer-wise kernels are identified as the relevant and sufficient set of order parameters Rubin et al. (2025); Fischer et al. (2024); Seroussi et al. (2023b); Ringel et al. (2025). We further take a decoupled Gaussian variational ansatz so that $q(h) = \prod_{l=1}^{L-1}\prod_{i=1}^{N_l} q_{l,i}(h_i^l)$ where $q_{l,i}$ is Gaussian with mean $\mu_{l,i}$ and variance $Q_{l,i}$. As shown in App. C.3, the variational energy estimate is then given by

$$\tilde{E}_q \propto \sum_{l=1}^{L-1}\sum_{i=1}^{N_l} \underbrace{\left(\mathbb{E}_{h\sim\mathcal{N}(\mu_{l,i},Q_{l,i})}\left[\left\langle h, K_{l-1}^{-1} - Q_{l,i}^{-1}, h\right\rangle\right] + \left\langle \mu_{l,i}, Q_{l,i}^{-1}, \mu_{l,i}\right\rangle\right)}_{=:\Delta_{l,i}} + \underbrace{\langle y, K_{L-1}, y\rangle^{-1}}_{=:a_y}, \tag{9}$$

where we define

$$K_{l>0}(x,x') = \frac{\sigma_{l+1}^2}{N_l}\sum_{i=1}^{N_l}\mathbb{E}_{h_i^l\sim q_{l,i}}\left[\sigma\left(h_i^l(x)\right)\sigma\left(h_i^l(x')\right)\right], \quad K_0(x,x') = \frac{\sigma_1^2}{d}x\cdot x'. \tag{10}$$

Here, the $\Delta_{l,i}$ terms arise from the difference between the approximated kernel and the actual one, and the a_y term results from enforcing an alignment $\alpha \approx 1$. Requiring that q minimize $\tilde{E}_q$ and $\alpha \approx 1$, we estimate $\tilde{E}_q \propto E(\alpha \approx 1) \propto P_*$. Another interpretation of $\Delta_{l,i}$, discussed in App. D, is the excess weight due to feature learning. This viewpoint is useful for feature learning patterns involving circuits, as the latter have a sharp imprint in weight-space. The above kernel viewpoint is, however, more general and can be used both for circuits and for more distributed learning patterns.

5 Heuristics for Manual Computation of Variational approximation

5.1 Feature Learning Patterns

While the above variational approach allows a variety of candidate q's, we focus on the previously mentioned set of feature-learning scenarios that have been extensively studied in the literature. Although this subset may appear restrictive, by varying behaviors among layers and between different neurons of the same layer, it already captures a wide range of phenomena. We then need to compare the variational energy ($\tilde{E}_q$), as detailed in Sec. 4, for such combinations and select the minimizer. The optimal pattern is an indication of the feature learning that emerges in the network to enable strong alignment, as motivated in App. A.3. Concretely, per layer and neuron pre-activation ($h_i^l(x)$), we allow one of the following choices, as illustrated in Fig.

(1) Gaussian Process (GP). Here, $q_{l,i}$ is a Gaussian process (GP) so that $h_{l,i} \sim \mathcal{N}(0, K_{l-1})$ with K_{l-1} defined in (10). This choice defines the "base model" of feature learning. For FCNs,[5] it implies that the network propagates feature structure forward without altering latent features (see Sec. 5.2). When all layers and neurons follow this distribution, the network reduces to the neural network GP (NNGP) Neal (1996), where no feature learning occurs. Introducing any of the patterns below in a subset of neurons enables feature learning to emerge.

(2) Gaussian Feature Learning (GFL). In this scenario, pre-activations remain Gaussian with zero mean, but the covariance is modified relative to the GP scenario (1): the kernel of the previous layer is amplified by a factor D in the direction of a specific feature (e.g. an eigenfunction of K_{l-1}) Φ_*^l.[6] Thus, here too, the distribution is a GP but with a different covariance $Q_{l,i}$ given by

$$Q_{l,i}(x, x') = K_{l-1}(x, x') + D\langle \Phi_*^l, K_{l-1}, \Phi_*^l \rangle \, \Phi_*^l(x)\Phi_*^l(x'). \tag{11}$$

(3) Specialization. In this scenario, a given neuron specializes to a particular feature Φ_*^l with proportionality constant $\mu_{l,i}$. This pattern corresponds to a Gaussian distribution which is sharply peaked around a non-zero mean $\mu_{l,i}\Phi_*^l$ [7]. Explicitly, we define the distribution of the specialized neuron as

$$q_{l,i}(\langle h_i^l, \Phi_*^l \rangle) = \delta[\langle h_i^l, \Phi_*^l \rangle - \mu_{l,i}], \quad q_{l,i}(\langle h_i^l, \Phi_\perp^l \rangle) = \mathcal{N}(0, \langle \phi_\perp^l, K_{l-1}, \Phi_\perp^l \rangle). \tag{12}$$

5.2 Layer-wise Feature Propagation

Since the variational energy of each layer depends on the kernel of the previous layer, an important element in our heuristic is understanding how the choice of pattern in a given layer affects the kernel and its spectrum in the subsequent layer. To this end, we define feature learning as any deviation from the baseline GP pattern (see Sec. 5.1), such as introducing a non-zero mean to the distribution (i.e., specialization) or altering its covariance structure (i.e., GFL). In our framework, a "feature" refers either to the mean $\mu_{l,i}$ of $q_{l,i}$ or to an eigenfunction of its covariance operator $Q_{l,i}(x, x')$.

We now outline several key claims concerning how features typically propagate between layers in FCNs. In this context, we consider a data measure that is i.i.d. Gaussian with zero mean and variance 1, not because it approximates the data well, but rather because it provides an unbiased baseline (see also Lavie and Ringel (2025)) for measuring function overlaps. Depending on the input, other choices can also be considered (e.g. permutation-symmetric measures over discrete tokens Lavie et al. (2024)). The following claims with their justifications should be understood as heuristic principles or rationalizations of empirically observed phenomena. Proving them in general or augmenting for different architectures is left for future work. For further details and empirical results, see App. C.5.

Claim (i): Neuron specialization creates a spectral spike. Assume that M neurons in layer l specialize on a single feature $\Phi_*^l(x)$, the subsequent kernel K_l develops a new, dominant spectral feature corresponding to $\sigma(\Phi_*^l(x))$. The corresponding RKHS norm of this feature is amplified, scaling

[5] For CNNs, even in the lazy regime, deeper kernels have different input scope and hence do generate new structure.

[6] One may also consider generalizations to several features.

[7] Taking equilibrated networks and increasing the amount of data, specialization was shown in Rubin et al. (2024) to emerge as a first-order phase transition where the average of preactivations suddenly shifts to $\mu_{l,i}$. This behavior was further associated with Grokking, suggesting a potential specialization-grokking link.

as $\mathcal{O}(N_l/M)$. **Justification:** When M neurons specialize, the next layer's kernel is approximately $K_l(x,x') = A(x,x') + \frac{M}{N_l}\sigma(\Phi_*^l(x))\sigma(\Phi_*^l(x'))$, where A is the contribution from the non-specialized neurons. Treating the specialization term as a rank-1 update, the Sherman-Morrison formula shows that its RKHS norm becomes $(R_A^{-1} + M/N)^{-1}$, where R_A is the RKHS norm of A, which satisfies $R_A^{-1} \ll M/N$ in typical high-dimensional settings.

Claim (ii): Amplified features in the pre-activation kernel create amplified higher-order features in the post-activation kernel. If a feature $\Phi_*^l(x)$ in kernel K_l has its eigenvalue enhanced by a factor D (i.e., $\lambda_* \to \lambda_* D$), then the corresponding m-th order power of this feature $(\Phi_*^l)^m(x)$ will have the bulk of its spectral decomposition, under the downstream kernel, shifted up by D^m, with similar effect on the inverse RKHS norm. **Justification:** A Taylor expansion of K_{l+1} in terms of the eigenfunctions of K_l shows that the term corresponding to $(\Phi_*^l)^m(x)$ will have a coefficient scaling with $(\lambda_* D)^m$. We argue that this term is difficult to span using other terms in this expansion, allowing us to treat it as a spectral spike and analyze it similarly to Claim (i). A numerical demonstration of this effect is shown in Fig. 5.

Claim (iii): Lazy layers preserve the relative scale of features from the previous layer. In the absence of feature learning, a properly normalized lazy layer approximately preserves the eigenspectrum of the previous kernel. If a feature $\Phi_*^l(x)$ has an eigenvalue λ_* with respect to the pre-activation kernel given by K_{l-1}, its effective eigenvalue with respect to the post-activation kernel K_l will also be proportional to λ_*. **Justification:** Follows from Claim (ii) taking $D = 1$.

Propagation rules for FCNs

(1) **Specialization**: Layer l specialized M neurons on Φ_*^l. For any feature Φ satisfying $\langle \sigma(\Phi_*^l), \Phi \rangle \neq 0$, we approximate $\langle \Phi, K_l^{-1}, \Phi \rangle \propto \left[\sum_{i\text{ sp.}} \frac{\mu_{i,l}^2}{N_l}\right]^{-1}$, where we sum over all specializing neurons.
(2) **GFL**: Layer l amplified fluctuation along Φ_*^l by D so that $\langle \Phi_*^l, K_l^{-1}, \Phi_*^l \rangle = (D\lambda_*)^{-1}$, where λ_* is the GP value of the inner product. Then for any m we have $\langle (\Phi_*^l)^m, K_l^{-1}, (\Phi_*^l)^m \rangle \propto (D\lambda_*)^{-m}$.

6 Concrete Examples

We now apply the heuristic principles of Sec. 4, 5.1, 5.2 to derive sample complexity bounds in a few examples. In App. E.1, we benchmark our method on two-layer FCNs and simple CNNs with non-overlapping patches. There we reproduce both the sample complexity exponent $P_* = d^{3/4}$ identified for CNNs in Ringel et al. (2025) and further predict that $P_* = d$ for two-layer FCNs studying a Hermite-3 target as well as the scaling of the number of specializing neurons with width (see Fig. 2). The latter is also a setting for which the prior's upper bound can be computed directly from $E(\alpha)$ using Large Deviation Theory, leading to a good match with experiment (Fig. 2 panel (a)).

Going to what we believe is beyond the current analytical state of the art, in App. E.3, we predict a $P_* = \sqrt{L}$, where L is the context length, of a softmax attention layer learning a cubic target (see Fig. 3). Another such instance, discussed in detail below, is that of a 3-layer non-linear network learning a non-linear target. In App. A.2 we show that this approach can be extended to classification tasks as well. In App. E.2 we apply our heuristics to a concrete setting, of a two-layer ReLU network trained on a parity task. We find there as well are able to predict the emergent feature pattern, which qualitatively differs from erf networks. Rather than identifying the scaling number of specializing neurons, we find that there is a finite number of neurons and we are able to predict their scale.

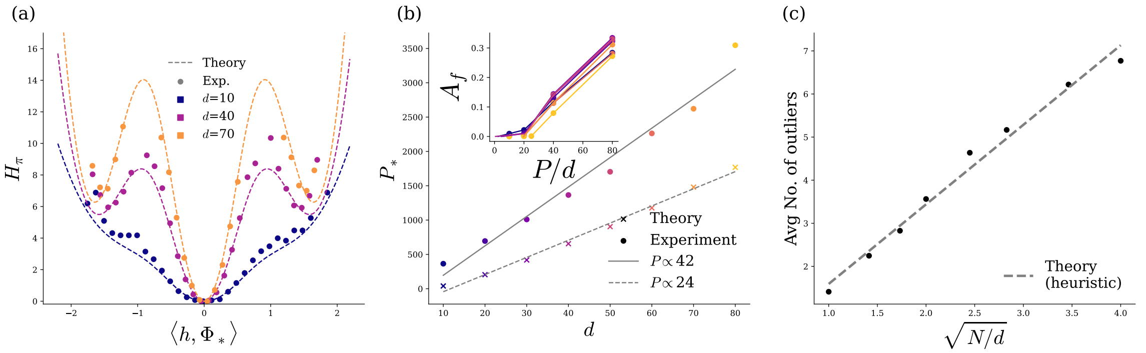

Figure 2: Numerical and experimental results for a two-layer erf network trained on the normalized third Hermite polynomial ($m = 3$). In panel (a) we compare the experimental results and exact theoretical predictions (computed utilizing LDT, see App. B) for the distribution of the alignment of the hidden layer pre-activation with the linear feature. Here we follow the same notation as in (7), so that H_π is the negative log posterior of the preactivations up to an additive constant that enforces zero minimum. We also find the pre-activation distribution corresponds to $q(h)$ for $q \sim$ M-Sp, as predicted by our heuristic approach. Panel (b) compares theoretical and experimental predictions for P_*, defined as alignment $\alpha > 0.1$ (inset shows alignment as a function of sample size). Both theoretical and experimental results agree on $P_* \propto d$. In (c), we increase N and keep P and d fixed, and plot the number of specialized neurons in the hidden layer. In agreement with our heuristic predictions, the number of neurons increases linearly with $\sqrt{N/d}$.

6.1 The Three-layer Network

Here we consider a three-layer FCNs given by

$$f(x) = \sum_{i=1}^{N_2} a_i \sigma \left(\sum_{j=1}^{N_1} w_{ij}^{(2)} \sigma(w_j^{(1)} \cdot x) \right), \tag{13}$$

where $x \in \mathbb{R}^d$ is drawn from $\mathcal{N}(0, I_d)$. We train these networks on a polynomial target of degree m given by $y(x) = He_m(w_* \cdot x)$ where He_m is the m-th probabilist Hermite polynomial, which is the standard polynomial choice under our choice of data measure, and $w_* \in \mathbb{R}^S$ is some normalized vector. The networks are trained via Langevin dynamics Welling and Teh (2011b), with ridge parameter κ, quadratic weight decay, and standard scaling. For an extension to mean-field scaling, see App. E.1.1.

As a starting point for our analysis, consider the simplest pattern (q"="**GP-GP**), having two GP/lazy layers where taking an l'th layer to be lazy means $Q_{l,i} = K_{l-1}$ and $\mu_{l,i} = 0$. Following the choice of pattern, our goal is to estimate the scaling of $\tilde{E}_q$. Examining Eq. (9), we find that the $\Delta_{l=1,2,i}$ contributions all cancel by our above choice of Q's and μ's. The only non-zero contribution thus comes from the final layer and is given by the inverse of $\langle y, K_2, y \rangle$. Because of lazy learning, $K_2(x, x')$ is a standard dot product FCN kernel which can be expanded as $\sum_{n=1}^{\infty} a_n (x \cdot x'/d)^n$, with $a_n = O(1)$ w.r.t. $N_{1,2}, d, P$. It can then be shown that $\langle y, K_2, y \rangle = O(d^{-m})$. Leading to $\tilde{E}_{GP-GP} = a_y \propto d^m$.

Next, we consider a feature learning pattern wherein the first layer is GP distributed but the second has feature learning (q"="**GP-Sp.**). Specifically, for the $l = 1$ layer, we take $Q_{1,i} = K_0; \mu_{1,i} = 0$ thereby nullifying again $\Delta_{1,0}$ in Eq. 9 for $\tilde{E}_q$. For the second layer, we assume M_2 specializing neurons (e.g. $i = 1..M_2$) which fluctuate around the linear feature $(w_* \cdot x)$, while others are lazy, namely $Q_{2,i>M_2} =$

$K_1, \mu_{2,i>M_2} = 0$ and $Q_{2,i=1..M_2} = K_1, \mu_{2,i=1..M_2}(x) = (w_* \cdot x)$. Examining $\sum_{i=1}^{N_2} \Delta_{2,i}$ in Eq. 9, we get zero contributions from $\Delta_{2,i>M_2} = 0$ and $M_2 \langle (w_* \cdot x), K_1^{-1}, (w_* \cdot x) \rangle$ from $i = 1..M_2$. As K_1 is again a simple FCN dot product kernel with no feature learning effects, normalized linear functions such as $w_* \cdot x$ have an $O(d)$ RKHS norm. We thus obtain an overall contribution to $\tilde{E}_q$ from the $l = 2$ layer equal to $M_2 d$. Finally, we need to estimate $a_y = \langle y | K_2 | y \rangle^{-1}$. Note that K_2 is not a standard FCN kernel anymore, since Q_2, which contains target information, is used in its definition (Eq. 10). According to our feature propagation rule (i), with $\Phi_* = (w_* \cdot x)$, we have $a_y = N_2/M_2$. Given M_2, we thus obtain a variational energy of $M_2 d + N_2/M_2$. We next need to minimize over free parameters, namely M_2 leading to $M_2 = \sqrt{N_2/d}$ and finally $\tilde{E}_{\mathbf{GP-Sp.}} = \sqrt{N_2 d}$. Provided N_2 scales less than d^{2m-1} ($N_2 = o(d^{2m-1})$), this pattern is favorable to $\mathbf{GP} - \mathbf{GP}$.

Finally, we consider what turns out to be the favorable pattern consisting of M_1 neurons specializing on $(w_* \cdot x)$ in the first layer and all neurons in the second layer specializing $He_m(w_* \cdot x)$ in the second layer, with a small proportionality constant $\mu_{2,i} = \pm\sqrt{\beta/N_2}$ (q"="**Sp.-Mag.**). We refer to the second-layer pattern as magnetization. Following straightforward adaptation previous argument to this pattern, the variational energy for this pattern as well as others, for $m = 3$, can be found in Table 1.

Feature Pattern	Δ_1	Δ_2	a_y	Minimizing Parameters	$\tilde{E}$
GP-GP	0	0	d^3	$-$	d^3
GP-Sp.	0	$M_2 d$	$\dfrac{N_2}{M_2}$	$M_2 = \sqrt{\dfrac{N_2}{d}}$	$\sqrt{N_2 d}$
Sp.-Mag.	$M_1 d$	$\dfrac{N_1}{M_1}\beta$	$\dfrac{N_2}{\beta}$	$\beta = \left(\dfrac{N_2^2}{N_1 d}\right)^{1/3}, \quad M_1 = \left(\dfrac{N_2 N_1}{d^2}\right)^{1/3}$	$(N_1 N_2 d)^{1/3}$

Table 1: Variational energy $\tilde{E}$ for different choices of feature-learning patterns in a three-layer FCN trained on $y(x) = \mathrm{He}_3(w_* \cdot x)$. The patterns shown here are (first/second layer): GP-GP, GP-Specialization, and Specialization-Magnetization. For each pattern, the components of the variational energy (Δ_1, Δ_2, a_y) together with the corresponding minimizing parameters are shown. We comment that the GP-GP pattern is favorable only for $d > N_2^5$, and otherwise feature learning will emerge.

In the non-GP q patterns, we obtain the same scaling of $\tilde{E}_q$ in the proportional limit ($N_1 \propto N_2 \propto d$), namely, $P_*/\kappa \propto d$. This observation is validated experimentally in Fig. 3, where the transition to non-zero alignment becomes sharper in the thermodynamic limit ($d \to \infty$). However, the mechanism by which this scaling is realized changes. In the specialization-magnetization pattern the sample complexity scales with $N_1^{1/3}$, therefore, it increases with N_1. However, under the GP-specialization pattern, sample complexity does not scale with N_1, making this pattern preferable. This prediction is in line with experimental results (see Figs. 8 and 3 panel (c)) where increasing N_1 causes the described change in feature learning patterns. Our prediction also accurately determines the scaling of the number of specializing neurons with N_1.

6.2 Softmax Attention

Here, we consider an attention block of the form

$$f(X) = \frac{1}{\sqrt{L}} \sum_{h=1}^{H} \sum_{a,b=1}^{L} @_{ab;h}(X)(w_h \cdot x^b) \quad @_{ab;h}(X) = e^{[x^a]^\top A_h x^b} \left(\sum_{c=1}^{L} e^{[x^a]^\top A_h x^c} \right)^{-1}, \tag{14}$$

where $X \in \mathbb{R}^{L \times d}$, $A \in \mathbb{R}^{d \times d}$, and $x^a \in \mathbb{R}^d$ is the a-th row of X, and $w_h \in \mathbb{R}^d$. Our prior on network weights is $\prod_{h=1}^{H} \mathcal{N}(0, I_{d^2}/d^2; A_h) \mathcal{N}(0, I_d/(dH); w_h)$. The only dependence on the context length L arises from the pre-factor $1/\sqrt{L}$, which ensures that for $X_i^a \sim \mathcal{N}(0,1)$, we have $f(X) = \mathcal{O}(1)$. The target function is given by $y(X) = \sum_{a,b} \frac{1}{\sqrt{L(L-1)}} x_1^a x_2^b x_3^b$, also normalized to be $\mathcal{O}(1)$. Following our approach, we propose two patterns for this architecture: GP (or lazy learning) and specialization (where we take $A_h \sim \mathcal{N}(\mu\sigma_x \otimes I_{(d-2)}, I_{d^2})$ for $\sigma_x = [1,0;0,1]$ and optimize over μ). As detailed in E.3, the variational energy scales as Ld^3 for the GP pattern and as $\sqrt{HLd^3}$ for the specialization pattern, the latter thus being favorable for H scaling less than $\sqrt{Ld^3}$. As shown in Fig. 3, this scaling of P with L and d indeed matches the dependence of the sample complexity on L.

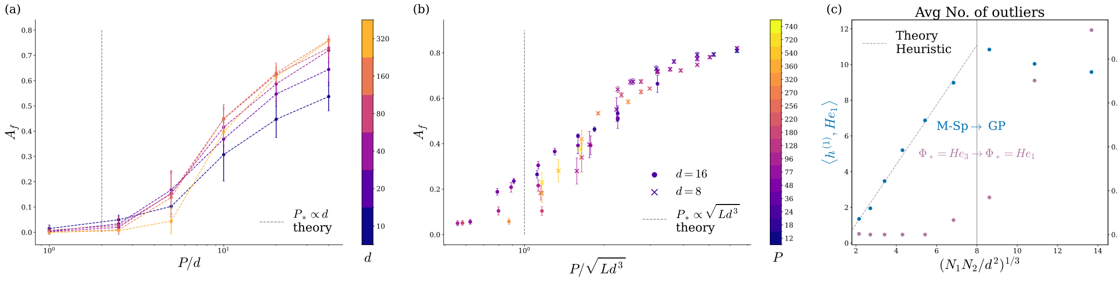

Figure 3: **Sample complexity:** Heuristic predictions accurately capture sample complexity in both three-layer erf FCNs and softmax attention heads, as well as feature learning scaling. Panels (a),(b) both track how the network alignment changes as a function of the ratio between the sample size, P, and the predicted sample complexity- P/d for the FCN in panel (a), and $P/\sqrt{Ld^3}$ for the attention head in panel (b). In both cases, we observe that the alignment collapses onto a single curve, confirming the predicted sample complexity, where good alignment is achieved. See Fig. 9 for comparison to MSE. See E.3 and F.1 for experimental details. **Feature learning patterns:** Panel (c) tracks the number of linearly specialized neurons, in both the first (blue) and second (purple) layers as the first-layer width N_1 varies (with fixed P, d and N_2). The number of first-layer specializing neurons initially follows the predicted $(N_1/d)^{(1/3)}$ scaling before the predicted transition occurs, where second-layer neurons begin to specialize on the linear feature rather than the cubic one, and the first layer neurons approach the GP distribution.

7 Discussion

This paper presents a novel methodology for analyzing the scaling behavior of sample complexity, through which we can also understand how distinct learning mechanisms emerge. Its strength lies in abstracting away from fine-grained details to isolate the core principles at play. By providing a common language for disparate phenomena, our work aims to unify fragmented theoretical perspectives, paving the way for an accessible and cohesive theory of representation learning. We hope such a strategy would remove barriers and expedite connections between mechanistic interpretability and first-principles scientific approaches.

Limitations. Notwithstanding our contributions, several avenues for improvement remain. In particular, extending the approach to overfitting patterns, quantifying feature propagation in more general CNNs and transformers, and addressing multi-feature interaction effects as those appearing in the context of superposition Elhage et al. (2022). It would also be desirable to extend our heuristic to dynamics of learning, potentially drawing insights from previous work relating equilibrium and dynamical phenomena Power et al. (2022); Rubin et al. (2024); Bahri et al. (2024); Nam et al. (2024). Since Bayesian convergence times can be exceptionally slow, correctly predicting the emergence of feature learning in early stages of training may also be highly advantageous. Finally, in some cases, such as under mean-field scaling, overfitting effects can emerge. It would therefore be valuable to extend our approach to account for patterns that lead to overfitting.

REFERENCES

Laurence Aitchison. Why bigger is not always better: on finite and infinite neural networks. *arXiv preprint arXiv:1910.08013*, 2019.

Laurence Aitchison. Deep kernel machines and fast solvers for deep kernel machines. *arXiv preprint arXiv:2108.13097*, 2021.

S Ariosto, R Pacelli, M Pastore, F Ginelli, M Gherardi, and P Rotondo. Statistical mechanics of deep learning beyond the infinite-width limit. *arXiv preprint arXiv:2209.04882*, 2022.

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. *arXiv e-prints*, art. arXiv:2302.05882, February 2023. doi: 10.48550/arXiv.2302.05882.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi: 10.1073/pnas.2311878121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2311878121.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, December 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1907378117. URL https://pnas.org/doi/full/10.1073/pnas.1907378117.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024. URL https://arxiv.org/abs/2404.14082.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in Neural Information Processing Systems*, 35:9768–9783, 2022.

N. N. Bogolubov and Nickolai N. Bogolubov Jr. *Introduction To Quantum Statistical Mechanics (2nd Edition)*. World Scientific Publishing Company, December 2009. ISBN 978-981-310-095-4. Google-Books-ID: t2JIDQAAQBAJ.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35:32240–32256, 2022.

Blake Bordelon, Lorenzo Noci, Mufan Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. URL https://openreview.net/forum?id=6pfCFDPhy6.

Gon Buzaglo, Itamar Harel, Mor Shpigel Nacson, Alon Brutzkus, Nathan Srebro, and Daniel Soudry. How Uniform Random Weights Induce Non-uniform Bias: Typical Interpolating Neural Networks Generalize with Narrow Teachers, February 2025. URL http://arxiv.org/abs/2402.06323. arXiv:2402.06323 [cs].

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12 (1):2914, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23103-1. URL https://www.nature.com/articles/s41467-021-23103-1.

J.L. Cardy. *Scaling and Renormalization in Statistical Physics*. Cambridge lecture notes in physics. Cambridge University Press, 1996. ISBN 9787506238229. URL https://books.google.co.il/books?id=g5hfPgAACAAJ.

Youngmin Cho and Lawrence K Saul. Kernel Methods for Deep Learning. *NIPS*, pages 1–9, 2009.

Omry Cohen, Or Malka, and Zohar Ringel. Learning Curves for Deep Neural Networks: A Gaussian Field Theory Perspective. *Physical Review Research*, 3(2):023034, April 2021. ISSN 2643-1564. doi: 10.1103/PhysRevResearch.3.023034. URL `http://arxiv.org/abs/1906.05301`. arXiv:1906.05301 [cs].

Hugo Cui. High-dimensional learning of narrow neural networks, 2025. URL `https://arxiv.org/abs/2409.13904`.

Bernard Derrida. Random-energy model: Limit of a family of disordered models. *Physical Review Letters*, 45(2):79–82, 1980. doi: 10.1103/PhysRevLett.45.79.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

Kirsten Fischer, Javed Lindner, David Dahmen, Zohar Ringel, Michael Krämer, and Moritz Helias. Critical feature learning in deep neural networks, May 2024. URL `http://arxiv.org/abs/2405.10761`. arXiv:2405.10761 [cond-mat].

Florentin Guth, Brice Ménard, Gaspar Rochette, and Stéphane Mallat. A rainbow in deep network black boxes. *Journal of Machine Learning Research*, 25(350):1–59, 2024. URL `http://jmlr.org/papers/v25/23-1573.html`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv e-prints*, art. arXiv:1502.01852, February 2015. doi: 10.48550/arXiv.1502.01852.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017. URL `https://arxiv.org/abs/1712.00409`.

Albrecht Huber. Variational Principles in Quantum Statistical Mechanics. In R. C. Clark and G. H. Derrick, editors, *Mathematical Methods in Solid State and Superfluid Theory*, pages 364–392. Springer US, Boston, MA, 1968. ISBN 978-1-4899-6214-0 978-1-4899-6435-9. doi: 10.1007/978-1-4899-6435-9_14. URL `http://link.springer.com/10.1007/978-1-4899-6435-9_14`.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. 2020.

A. L. Kuzemsky. Variational Principle of Bogoliubov and Generalized Mean Fields in Many-Particle Interacting Systems. *International Journal of Modern Physics B*, 29(18):1530010, July 2015. ISSN 0217-9792, 1793-6578. doi: 10.1142/S0217979215300108. URL `http://arxiv.org/abs/1507.00563`. arXiv:1507.00563 [cond-mat].

Itay Lavie and Zohar Ringel. Demystifying Spectral Bias on Real-World Data, February 2025. URL `http://arxiv.org/abs/2406.02663`. arXiv:2406.02663 [stat].

Itay Lavie, Guy Gur-Ari, and Zohar Ringel. Towards Understanding Inductive Bias in Transformers: A View From Infinity, May 2024. URL `http://arxiv.org/abs/2402.05173`. arXiv:2402.05173 [cs].

Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Phys. Rev. X*, 11:031059, Sep 2021. doi: 10.1103/PhysRevX.11.031059. URL https://link.aps.org/doi/10.1103/PhysRevX.11.031059.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

David A. McAllester. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, page 164–170, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131674. doi: 10.1145/307400.307435. URL https://doi.org/10.1145/307400.307435.

Alexander van Meegen and Haim Sompolinsky. Coding schemes in neural networks learning classification tasks, June 2024. URL http://arxiv.org/abs/2406.16689. arXiv:2406.16689 [cs].

Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A Louis. Is sgd a bayesian sampler? well, almost. *arXiv preprint arXiv:2006.15191*, 2020.

Chris Mingard, Lukas Seier, Niclas Göring, Andrei-Vlad Badelita, Charles London, and Ard Louis. Characterising the inductive biases of neural networks on boolean data, 2025. URL https://arxiv.org/abs/2505.24060.

Yoonsoo Nam, Nayara Fonseca, Seok Hyeong Lee, Chris Mingard, and Ard A. Louis. An exactly solvable model for emergence and scaling laws in the multitask sparse parity problem, 2024. URL https://arxiv.org/abs/2404.17563.

Gadi Naveh, Oded Ben David, Haim Sompolinsky, and Zohar Ringel. Predicting the outputs of finite deep neural networks trained with noisy gradients. *Physical Review E*, 104(6), Dec 2021. ISSN 2470-0053. doi: 10.1103/physreve.104.064301. URL http://dx.doi.org/10.1103/PhysRevE.104.064301.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 1996. ISBN 978-0-387-94724-2 978-1-4612-0745-0. doi: 10.1007/978-1-4612-0745-0. URL http://link.springer.com/10.1007/978-1-4612-0745-0.

R. Pacelli, S. Ariosto, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo. A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-width limit. *Nature Machine Intelligence*, 5(12): 1497–1507, December 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00767-6. URL https://www.nature.com/articles/s42256-023-00767-6.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias, and Inbar Seroussi. Applications of Statistical Field Theory in Deep Learning, April 2025. URL http://arxiv.org/abs/2502.18553. arXiv:2502.18553 [stat].

Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a First Order Phase Transition in Two Layer Networks, May 2024. URL http://arxiv.org/abs/2310.03789. arXiv:2310.03789 [stat].

Noa Rubin, Kirsten Fischer, Javed Lindner, David Dahmen, Inbar Seroussi, Zohar Ringel, Michael Krämer, and Moritz Helias. From Kernels to Features: A Multi-Scale Adaptive Theory of Feature Learning, 2025. URL https://arxiv.org/abs/2502.03210. Version Number: 2.

David Saad and Sara A. Solla. Exact solution for on-line learning in multilayer neural networks. *Phys. Rev. Lett.*, 74:4337–4340, May 1995. doi: 10.1103/PhysRevLett.74.4337. URL https://link.aps.org/doi/10.1103/PhysRevLett.74.4337.

Inbar Seroussi, Gadi Naveh, and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some cnns. *Nature Communications*, 14(1):908, Feb 2023a. ISSN 2041-1723. doi: 10.1038/s41467-023-36361-y. URL https://doi.org/10.1038/s41467-023-36361-y.

Inbar Seroussi, Gadi Naveh, and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some CNNs. *Nature Communications*, 14(1):908, February 2023b. ISSN 2041-1723. doi: 10.1038/s41467-023-36361-y. URL https://www.nature.com/articles/s41467-023-36361-y.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 681–688, USA, 2011a. Omnipress. ISBN 978-1-4503-0619-5. URL http://dl.acm.org/citation.cfm?id=3104482.3104568.

Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, Madison, WI, USA, 2011b. Omnipress. ISBN 9781450306195.

Andrew Gordon Wilson. The Case for Bayesian Deep Learning, 2020. URL https://arxiv.org/abs/2001.10995. Version Number: 1.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian Deep Learning and a Probabilistic Perspective of Generalization, 2020. URL https://arxiv.org/abs/2002.08791. Version Number: 4.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.