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, 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 $g(..)$ needs to be fine-tuned for exact computations, a change of variable reveals a robust linear scaling with P for any $g(..)$.

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 & 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 approaches Aitchison (2019); Li & Sompolinsky (2021); Aitchison (2021); Seroussi et al. (2023a); Ariosto et al. (2022); Bordelon & Pehlevan (2022); Rubin et al. (2024; 2025); Ringel et al. (2025) and Saad and

Solla type approaches Saad & 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-principle predictions on sample complexity and feature learning effects in networks trained to equilibrium. Our approach consists of three main parts: (1) a taxonomy of feature learning patterns per layer along with concrete variational probabilities representing these patterns (Sec. 4.3), (2) a tractable proxy ($\tilde{E}$) to the Kullback-Leibler divergence between each variational probability and the true distribution (Secs. 4.1-4.2), and (3) a set of claims establishing how feature/kernel-eigenmode enhancement in one layer propagates to the next layer and affects $\tilde{E}$ (Sec. 4.4). Combining these three, together with a bound on sample complexity, 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 it for convolutional neural networks (CNNs) as needed. Our FCNs are defined by

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

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 & Izmailov (2020); Wilson (2020); Naveh et al. (2021). Alternatively, they represent an exact solution to Langevin dynamics with weight decay Welling & Teh (2011b). We denote the target function by y and the training sample size by P, and assume training with MSE loss. The quadratic weight decay for each layer l is set to $2\kappa N_{l-1}/\sigma_l^2$, where κ is the ridge parameter and $N_0 = d$ is the input dimension. 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_ν 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, $p\left(\Theta\right)$ 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 of order 1 under the prior He et al. (2015).

A main focus of this work is characterizing 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). 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. This alignment represents the extent to which the network learns a function that is proportional to the target, and bounds the test MSE via the Cauchy–Schwarz inequality $\int d\mu_x (f(x) - y(x))^2 \geq \langle y, y \rangle (A_f - 1)^2$. We thus consider the condition $A_f \approx 1$ to be a certificate of 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 $\alpha \sim \mathcal{O}(1)$. 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 $\Pr_{p_0}[A_f \geq \alpha]$ and $\Pr_{\pi}[A_f \geq \alpha]$ respectively. Following simple arguments (see App. A), we obtain the following bound on the log posterior

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

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 $\Pr_{\pi}[A_f \geq \alpha] \approx O(1)$. Since a random network is unlikely to achieve strong alignment, $\log \Pr_{p_0}[A_f \geq \alpha]$ would typically be highly negative for large α. Therefore, a sufficiently large data term is required to cancel this effect. Explicitly,

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

Thus, up to the ridge parameter and an $\mathcal{O}(1)$ 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 $k/\kappa = \mathcal{O}(1)$. 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.

In App. B, we provide an asymptotically exact solution for estimating the bound for the prior alignment $\Pr_{p_0}[A_f \geq \alpha]$ for a two-layer non-linear FCN using Large Deviation Theory (LDT). 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. 1). 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.

4 Heuristics for Computing Prior Alignment

In this section, we adopt a variational approach to estimating P_* by comparing different modes of feature learning [1] under a certain loss (see (8) below). While many approaches predict different feature learning mechanisms Pacelli et al. (2023); Fischer et al. (2024); Meegen & Sompolinsky (2024); Buzaglo et al. (2025); Li & 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.

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

4.1 VARIATIONAL ANALYSIS

Our next objective is to make the sample complexity bound, P_*, tractable. This requires estimating the prior probability term, $\Pr_{p_0}[A_f \geq \alpha]$, for alignments $\alpha \sim \mathcal{O}(1)$. As a first step, we simplify this by relating the cumulative probability to the probability density denoted by $p_{A_f}(\alpha)$. As shown in App. A.2, for large alignments, we have $p_{A_f}(\alpha) \gtrsim \Pr_{p_0}[A_f \geq \alpha]$. 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 in terms of pre-activations h (1). Accordingly, in App. C.2, we follow standard statistical mechanics techniques to express this density as

$$p_{A_f}(\alpha) = \int Dh \, \exp[-H_{p,\alpha}(h)]/Z_p, \quad \text{where} \quad Z_p = \int_{-\infty}^{\infty} d\alpha \int Dh \, \exp[-H_{p,\alpha}(h)]. \tag{6}$$

Per α, the above implies a probability over h's namely $\exp[-H_{p,\alpha}(h)]/\int Dh \, \exp[-H_{p,\alpha}(h)]$. We next approximate this probability by an analytically tractable variational estimate $\hat{q}(h) = e^{-H_q(h)}/Z_q$. The variational computation follows by looking for $\hat{q}(h)$ which minimizes the KL divergence. 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 & Jr (2009); Huber (1968). Here we provide a brief description – for the full derivation see App. C.1. By applying the Feynman–Bogoliubov inequality, we obtain an upper bound on $E(\alpha)$

$$E(\alpha) \leq \log(Z_p/Z_q) + \tilde{E}_q, \quad \tilde{E}_q = \mathbb{E}_{h\sim\hat{q}}[H_{p,\alpha}(h) - H_q(h)] \tag{7}$$

We argue in App. C.3 that for $\alpha \sim \mathcal{O}(1)$, the $\log(Z/Z')$ terms are subleading w.r.t. $\tilde{E}_q$. Thus, for $\alpha \sim \mathcal{O}(1)$, we can estimate $E(\alpha) \approx \tilde{E}_{q_*}$ where $\hat{q}_*$ minimizes $\tilde{E}_q$.

4.2 AN ESTIMATE FOR THE VARIATIONAL ENERGY

Next, we turn to estimating the variational energy $\tilde{E}_q$ (7). In App. C.2 we provide general expressions for $H_{p,\alpha}(h)$ and show that it decouples into independent distributions for each neuron and layer. Computing $\tilde{E}_q$ is difficult for all but deep linear networks due to $\sigma(h)$-dependent terms appearing in $H_{p,\alpha}(h)$. The argument put forward in App. C.3 is that non-linear terms scale the same way as the simpler quadratic terms appearing in $H_{p,\alpha}(h)$. They are therefore irrelevant to the scale analysis to follow.[2]

Since $H_{p,\alpha}(h)$ decouples, we consider candidate $\hat{q}$ patterns that do so in the same manner. This choice of variational ansatz 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). Concretely, we define $\hat{q}$ as a layer- and neuron-independent Gaussian $\hat{q}(h) = \prod_{l=1}^{L-1}\prod_{i=1}^{N_l}\mathcal{N}\left(h_i^l \mid \mu_{l,i}, Q_{l,i}\right)$. As shown in App. C.3, the variational energy estimate of a pattern $\hat{q}$, assuming an alignment of $\alpha \sim \mathcal{O}(1)$, is given by

$$\tilde{E}_q \propto \sum_{l=1}^{L-1}\sum_{i=1}^{N_l} \underbrace{\left(\mathbb{E}_{h\sim\mathcal{N}(\mu_{l,i},Q_{l,i})}\left[\left\langle h, K_{l-1}^{-1} - Q_{l,i}^{-1}, h\right\rangle\right] + \left\langle \mu_{l,i}, Q_{l,i}^{-1}, \mu_{l,i}\right\rangle\right)}_{=:\Delta_{l,i}} + \underbrace{\langle y, K_{L-1}, y\rangle^{-1}}_{=:a_y}, \tag{8}$$

where we define

$$K_{l>0}(x,x') = \frac{\sigma_{l+1}^2}{N_l}\sum_{i=1}^{N_l}\mathbb{E}_{h_i^l\sim\hat{q}_{l,i}}\left[\sigma\left(h_i^l(x)\right)\sigma\left(h_i^l(x')\right)\right], \quad K_0(x,x') = \frac{\sigma_1^2}{d}x\cdot x'. \tag{9}$$

[2] $\sigma(..)$ still enters our analysis via feature propagation effects which determine K_l^{-1} (see (9)).

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 $\alpha \sim \mathcal{O}(1)$. Requiring that $\hat{q}$ minimize $\tilde{E}_q$ and $\alpha \sim \mathcal{O}(1)$, we estimate $\tilde{E}_q \propto E(\alpha)$, which, in turn, enables us to estimate P_*.

4.3 Feature Learning Patterns

While the above variational approach allows a variety of candidate $\hat{q}$'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. 4.2, 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.2. Concretely, per layer and neuron pre-activation ($h_i^l(x)$), we allow one of the following choices:

(1) Gaussian Process (GP). Here, $\hat{q}_{l,i}$ is a Gaussian process (GP) so that $h_{l,i} \sim \mathcal{N}(0, K_{l-1})$ with K_{l-1} defined in (9). This choice defines the "base model" of feature learning. For FCNs,[3] it implies that the network propagates feature structure forward without altering latent features (see Sec. 4.4). 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.[4] Thus, here too, the distribution is a GP but with a different covariance $Q_{l,i}$ given by

$$Q_{l,i}(x, x') = K_{l-1}(x, x') + D\langle \Phi_*^l, K_{l-1}, \Phi_*^l \rangle \, \Phi_*^l \Phi_*^l. \tag{10}$$

(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 $\mu_{l,i}\Phi_*^l$. Explicitly, we define the distribution of the specialized neuron as

$$\hat{q}_{l,i}(h_i^l) = \delta[h_i^l - \mu_{l,i}\Phi_*^l], \tag{11}$$

4.4 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. 4.3), 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 $\hat{q}_{l,i}$ 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 & Ringel (2025)) for measuring function overlaps. Depending on the input, other choices can also be considered

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

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

(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.4.

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 as $\mathcal{O}(N_l/M)$. **Justification:** When M neurons specialize, the next layer's kernel is approximately $K_l(x,x') = A + \frac{M}{N_l}\sigma(\Phi_*^l(x))\sigma(\Phi_*^l(x'))$, 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. 3.

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 $D = 1$.

5 Concrete Examples

We now apply the heuristic principles of Sec. 4.2, 4.3, 4.4 to derive sample complexity bounds in a few examples. We first benchmark this method on a two-layer network, a setting that is well-studied in the literature and for which the prior's upper bound can be computed directly using LDT. We then extend the analysis to deep networks, going beyond the current state of the art.

5.1 The Two-Layer Network

In the two-layer setting, an exact solution can be obtained, so we begin by comparing our heuristic approach to the exact solution. In this case, we consider both two-layer FCNs as well as CNNs with non-overlapping convolution windows. Together, these are given by

$$f(x) = \sum_{i=1}^{N_w} \sum_{j=1}^{N} w_{ij}^{(2)}\, \sigma(w_j^{(1)} \cdot x_i), \tag{12}$$

where $x \in \mathbb{R}^d$ is drawn from $\mathcal{N}(0, I_d)$. We take $d = N_w S$ so that $w_j, x_i \in \mathbb{R}^S$. The vector x_i is given by the $((i-1)S+1)$-th to iS-th coordinates of x. We train these networks on a polynomial target of degree m given by $y(x) = \sum_{i=1}^{N_w} He_m(w_* \cdot x_i)$ 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 Lengevin dynamics Welling & Teh (2011b), with ridge parameter κ, quadratic weight decay, and standard scaling. For an extension to mean-field scaling, see App. D.1.1.

Fully Connected Networks. We turn to compute $\tilde{E}_q$ for a range of feature learning patterns. For this shallow FCN (and $N_w = 1$), our choice of feature learning patterns amounts to considering distributions in a single layer. We consider the following three scenarios (though more combinations are possible): (1) all neurons are GP distributed, (2) all are GFL distributed with amplification D, and (3) M neurons specialize on a the same feature, while those remaining are GP distributed. As the kernel of the first layer can only express linear features, the only relevant feature to be considered for the GFL and M-specialization patterns is $\Phi_*(x) = w_* \cdot x$. We compute the scale of the optimal variational energy for each pattern:

(1) **GP**: Here, $\Delta_{1,i} = 0$ since $K_{l-1} = Q_l$. In this baseline setting, learning $m > 1$ is hard since $\langle He_m | K_2 | He_m \rangle = \mathcal{O}(d^{-m})$ (see Sec. 4.4). Thus, in total, we have $\tilde{E}_{\hat{q}\sim\text{GP}} \propto d^m$.

(2) **GFL**: Following (8), this pattern incurs a cost of $\Delta_{1,i} = D$ per neuron i, resulting in a total cost of ND. The a_y term can be calculated utilizing Claim (ii). This leads to a D^m-factor decrease in the RKHS norm relative to the GP, so that $a_y \propto (d/D)^m$. In total, we find that $\tilde{E}_{\hat{q}\sim\text{GFL}} \propto ND + (d/D)^m$. Minimizing w.r.t. D, we obtain $D_{\min} = (d^m/N)^{1/(m+1)}$, and, substituting back, we obtain $\tilde{E}_{\hat{q}\sim\text{GFL}} \propto (Nd)^{\frac{m}{m+1}}$.

(3) **M-Specialization**: Following 8, this pattern incurs a cost of $\Delta_1 = M\langle \Phi_*, K_0^{-1}\Phi_* \rangle = Md$, where we denote $\Delta_l = \sum_i \Delta_{l,i}$. Utilizing Claim (i), this results in adding a spike with an M/N coefficient along $\sigma(w_* \cdot x)$ in K_1 appearing in a_y. Before this spike, $He_m(x)$ only had overlaps with the m-th order Taylor expansion of the kernel, leading to a d^{-m} scaling. However, since $\sigma(w_* \cdot x)$ has an $\mathcal{O}(1)$ overlap with $He_m(x)$, so that $a_y \propto N/M$, we obtain $\tilde{E}_{\hat{q}\sim\text{M-Sp}} \propto dM + N/M$. Minimizing further over M, the number of specializing neurons leads to $M_{\min} = \sqrt{N/d}$ and therefore $\tilde{E}_{\hat{q}\sim\text{M-Sp}} \propto \sqrt{dN}$.

Now, we can compare the different feature learning patterns. Taking the most common linear scaling where $N \propto d$, the specialization scenario has the lowest variational energy. Our scaling theory then predicts an $\mathcal{O}(d)$ sample complexity as well as multimodal distribution of w along w_* with $\mathcal{O}(1)$ specializing neurons. Taking $m > 1$ and $N \gg d^5$, lazy learning wins and leads to $\mathcal{O}(d^m)$ complexity. When $m = 1$, GFL and M-specialization are on par for $N \propto d$. These calculations coincide with both experimental and direct LDT results, as demonstrated in Fig. 1 for networks trained on He_3. In terms of sample complexity, both predictions agree with experiment, with a scaling of $P_* \propto d$, as seen in Fig. 1 (b). Our heuristic approach correctly predicts the scaling of the number of specializing neurons with N, as seen in Fig. 1 (c). Finally, as shown in panel (a), the analytical LDT method recovers the correct pre-activation distribution, which corresponds to $\hat{q}(h)$ for $\hat{q} \sim$ M-Sp.

CNN with Non-overlapping Patches. Our approach can be extended to the CNN in (12) with $N_w > 1$. In this case, it is better to focus on the covariance of w_i, namely, $\Sigma = N^{-1}\sum_{i=1}^{N} w_i w_i^T$, than on the covariance of pre-activations on each path. One can then show that the cost becomes $\Delta_1 = \mathrm{E}_{\hat{q}_w} w^T \left[\Sigma^{-1} - I_S/S\right] w$.

(1) **GP**: In this scenario, the output kernel is given by $K_{2,\text{CNN}}(x, x') = N_w^{-1}\sum_{i=1}^{N_w} K_{2,\text{FCN}}(x_i, x_i')$, where $K_{2,\text{FCN}}$ is the FCN kernel ($N_w = 1$). Focusing on a linear target for simplicity, we can work out the scaling of the relevant (linear) kernel feature by Taylor expanding $K_{2,\text{FCN}} = a_1 x_i \cdot x_i'/S$ to get $K_{2,\text{CNN}} = \frac{a_1}{N_w S} x \cdot x'$, with a_1 being some $\mathcal{O}(1)$ constant. Lazy learning then yields $\tilde{E}_{\hat{q}\sim\text{GP}} = N_w S = d$, as in a FCN with no weight sharing.

(2) **GFL**: Here, we take $\Sigma = I_S/S + D w_* w_*^T$, resulting in $\Delta_{l=1} = ND$. The leading term of the Taylor expansion now equals $K_{2,\text{CNN}} = \frac{1}{N_w S}\sum_{i=1}^{N_w} x_i^T \Sigma x_i'$ leading to a $D/(N_w S)$ scaling of the target. All in all, we find that $\tilde{E}_{\hat{q}\sim\text{GFL}} = ND + (N_w S)/D$, leading to $\tilde{E}_{q_*} = \sqrt{NM_w S}$ for the optimal $\hat{q}_* \sim$ GFL.

(3) **M-Specialization**: In this case, we obtain a $\Delta_{l=1} = MS$ cost. The contribution of the M specializing neurons to the kernel goes as $\frac{a_1 M}{N_w N}\sum_{i=1}(w_* \cdot x)(w_* \cdot x')$, leading to $a_y = \langle y, K_L, y \rangle = \mathcal{O}(M/N_w N)$. Thus, $\tilde{E}_{\hat{q}\sim\text{M-Sp}} = MS + N_w N/M$ resulting in the optimal variational energy $\sqrt{SN_w N}$.

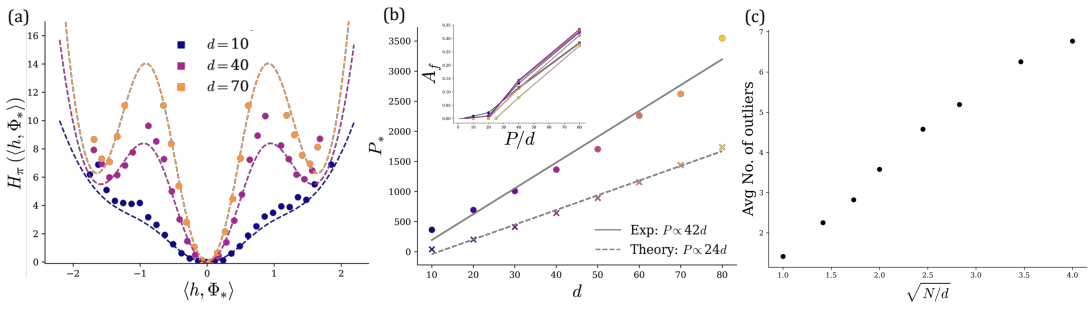

Figure 1: Numerical and experimental results for a two-layer erf network trained on the normalized third Hermite polynomial ($m = 3$). On the leftmost panel (a), the experimental negative log density of the hidden layer pre-activation alignment with the linear feature is shown, along with theoretical predictions computed utilizing LDT (see App. B). Indeed, the pre-activation distribution corresponds to $\hat{q}(h)$ for $\hat{q} \sim$ M-Sp, as predicted by our heuristic approach. The center 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 $P_* \propto d$. Finally, on the rightmost panel (c), we keep P and d fixed, and plot the number of specialized neurons in the hidden layer as a function of our predicted $\sqrt{N/d}$ scaling.

Both GFL and M-Specialization patterns, in the proportionate limit $N \propto N_w \propto S \to \infty$, lead to $P_* \propto S^{3/2} = d^{3/4}$. This recovers results reported in Ringel et al. (2025) computed via a mean-field approach.

5.2 The Three-Layer Network

Extending our analysis to the analytically complex setting of a three-layer network allows us to address a previously intractable challenge: predicting where and how feature learning emerges within the network. For brevity, we present only a representative subset of choices to illustrate how varying the width ratio influences the most likely pattern to emerge. As before, we consider three scenarios: (1) both layers are GP distributed, (2) the first layer is GP distributed, the second has M_2 specialized neurons that learn the linear feature, while those remaining are GP distributed, and (3) M_1 neurons in the first layer specialize on the linear feature, while all neurons in the second layer specialize on the cubic feature, with a small proportionality constant $\mu_{2,i} = \pm\sqrt{\beta/N}$. We refer to the latter pattern as magnetization. We next turn to compute the variational energy for each one of these cases.

(1) **GP–GP:** As in the two-layer FCN, both layers satisfy $\Delta_{l=1,2,i} = 0$. Since no features are amplified in this setting, we obtain $\tilde{E}_{\hat{q}\sim\text{GP}-\text{GP}} \propto d^3$.

(2) **GP–Specialization:** In the first layer we have $\Delta_{1,i} = 0$, while the second layer contributes $M_2 d$ through Δ_2. The contribution from a_y is N_2/M_2 (see App. C.4). Altogether, we have $\tilde{E}_{\hat{q}\sim\text{GP}-\text{SP}} \propto M_2 d + N_2/M_2$. Minimizing with respect to M_2 yields $\tilde{E}_{\hat{q}\sim\text{GP}-\text{SP}} \propto \sqrt{N_2 d}$ with $M_2 = \sqrt{N_2/d}$.

(3) **Specialization–Magnetization:** The cost of the first layer is $\Delta_1 = dM_1$, as in the two-layer FCN. Following Claim (i), K_1 has a spike of size M_1/N_1 in the linear direction, so that the contribution from the second layer is $\beta N_1/M_1$. The cubic feature is spiked in the second layer, contributing an alignment term of N_2/β. In total, we have $\tilde{E}_{\hat{q}\sim\text{SP-MAG}} \propto dM_1 + \frac{N_1}{M_1}\beta + \frac{N_2}{\beta}$. Minimizing w.r.t. β and M_1, we obtain $\tilde{E}_{\hat{q}\sim\text{SP-MAG}} \propto (N_2 N_1 d)^{1/3}$ with $M_1 = (N_2 N_1/d^2)^{1/3}$.

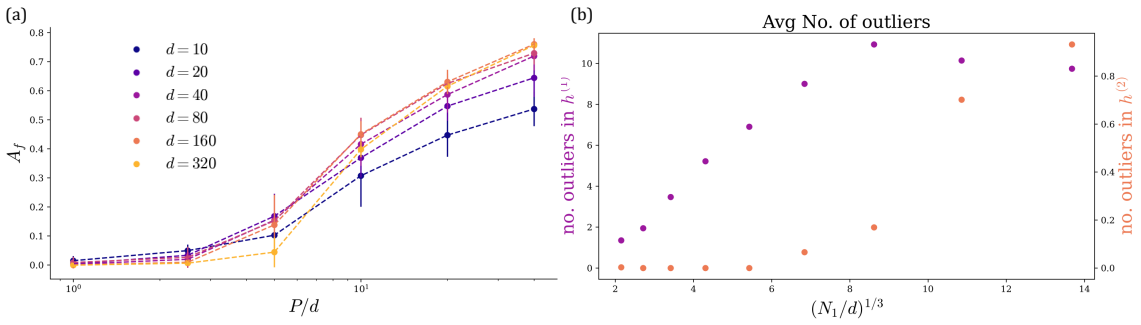

Figure 2: Heuristic predictions accurately capture sample complexity and neuron specialization in three-layer erf FCNs with the 3-rd Hermite polynomial as the target. All experiments use a three-layer network with $N_2 = d$. Panel (a) shows network alignment as a function of sample to input dimension ratio P/d. The data for different input dimensions (d) collapse onto a single curve, confirming the predicted $\mathcal{O}(d)$ sample complexity, where good alignment is achieved. Here, $N_1 = d$ as well. Panel (b) shows the number of linearly specialized neurons in the first (purple) and second (orange) layers as the first-layer width N_1 varies (with fixed P, d and N_2). The number of first-layer specialists initially follows the predicted $(N_1/d)^{(1/3)}$ scaling before the predicted transition occurs, where second-layer neurons begin to specialize.

Across all choices of $\hat{q}$ 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. 2(a), 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 Fig. 4(b)) 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 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). Finally, it would 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)

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=q5hfPgAACAAJ.

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.

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].

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 (eds.), *Mathematical Methods in Solid State and Superfluid Theory*, pp. 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.

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].

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, pp. 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, pp. 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.