PNAS RESEARCH ARTICLE | STATISTICS

Bayesian interpolation with deep linear networks

Boris Hanin[a,1,2] and Alexander Zlokapa[b,c,1]

Edited by David Donoho, Stanford University, Stanford, CA; received January 24, 2023; accepted April 25, 2023

Characterizing how neural network depth, width, and dataset size jointly impact model quality is a central problem in deep learning theory. We give here a complete solution in the special case of linear networks with output dimension one trained using zero noise Bayesian inference with Gaussian weight priors and mean squared error as a negative log-likelihood. For any training dataset, network depth, and hidden layer widths, we find nonasymptotic expressions for the predictive posterior and Bayesian model evidence in terms of Meijer-G functions, a class of meromorphic special functions of a single complex variable. Through asymptotic expansions of these Meijer-G functions, a rich new picture of the joint role of depth, width, and dataset size emerges. We show that linear networks make provably optimal predictions at infinite depth: the posterior of infinitely deep linear networks with data-agnostic priors is the same as that of shallow networks with evidence-maximizing data-dependent priors. This yields a principled reason to prefer deeper networks when priors are forced to be data-agnostic. Moreover, we show that with data-agnostic priors, Bayesian model evidence in wide linear networks is maximized at infinite depth, elucidating the salutary role of increased depth for model selection. Underpinning our results is an emergent notion of effective depth, given by the number of hidden layers times the number of data points divided by the network width; this determines the structure of the posterior in the large-data limit.

deep learning | Bayesian inference | neural networks | linear networks

Significance

Understanding the interplay between network architecture, dataset statistics, and learning algorithms is a key challenge in deep learning. We overcome this challenge analytically for zero-noise Bayesian inference in neural networks with linear activations and output dimension one but arbitrary depth, width, and training data. Our results provide rigorous solutions to learning with the simplest class of neural networks in which depth, width, and dataset can vary freely. We identify an emergent notion of effective depth, which combines depth, width, and dataset size, that must be large for networks to compute optimal posteriors from universal priors. Our results both present examples of models admitting exact solutions to Bayesian inference and give the first principled Bayesian justification for preferring deeper networks.

A central aim of deep learning theory is to understand the properties of overparameterized networks trained to fit large datasets. Key questions include: How do learned networks use training data to make predictions on test points? Which neural network architectures lead to more parsimonious models? What are the joint scaling laws connecting the quality of learned models to the number of training data points, network depth, and network width (1–4)?

The present article gives the exact answers to such questions for a class of neural networks in which one can simultaneously vary input dimension, number of training data points, network width, and network depth. This is significant because the limits where these four structural parameters tend to infinity do not commute, causing all prior work to miss important aspects of how they jointly influence learning. Our results pertain specifically to deep linear networks

$$f(x) = W^{(L+1)} \cdots W^{(1)} x, \qquad [1]$$

with input dimension N_0, L hidden layers of widths N_ℓ, and output dimension $N_{L+1} = 1$. As a form of learning, we take zero noise Bayesian interpolation starting from a Gaussian prior on network weights $W^{(\ell)} \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and empirical mean squared error over a training dataset of P examples as the negative log-likelihood. Deep linear networks, while linear in x, are nonlinear in their parameters and have been extensively studied as models for learning with neural networks using both gradient-based methods (5–9) and Bayesian inference (10–12).

Since we are considering an output dimension of 1, we may write $f(x) = \theta^T x$ for a vector $\theta \in \mathbb{R}^{N_0}$. What differentiates our work from a classical Bayesian analysis of Gaussian linear regression is that as soon as $L \geq 1$ the components of θ are correlated and non-Gaussian under the prior. Predictions $f(x)$ on inputs x orthogonal to inputs from the training data therefore differ under the prior and posterior. Specifically, as shown in Fig. 1, we may decompose $\theta = \theta_{||} + \theta_\perp$ into its projections onto directions spanned by the training data and their complement. By our Bayesian construction, $\theta_{||}$ is responsible for fitting the training data. Due to the correlations under the prior between $\theta_{||}$ and $\theta_\perp$, however, information from the training data will influence the posterior distribution of $\theta_\perp$. It is precisely this data-dependent extrapolation displayed by deep linear networks

Author affiliations: [a]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540; [b]Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139; and [c]Google Quantum AI, Venice, CA 90291

Author contributions: B.H. and A.Z. designed research; performed research; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]B.H. and A.Z. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: bhanin@princeton.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2301345120/-/DCSupplemental.

Published May 30, 2023.

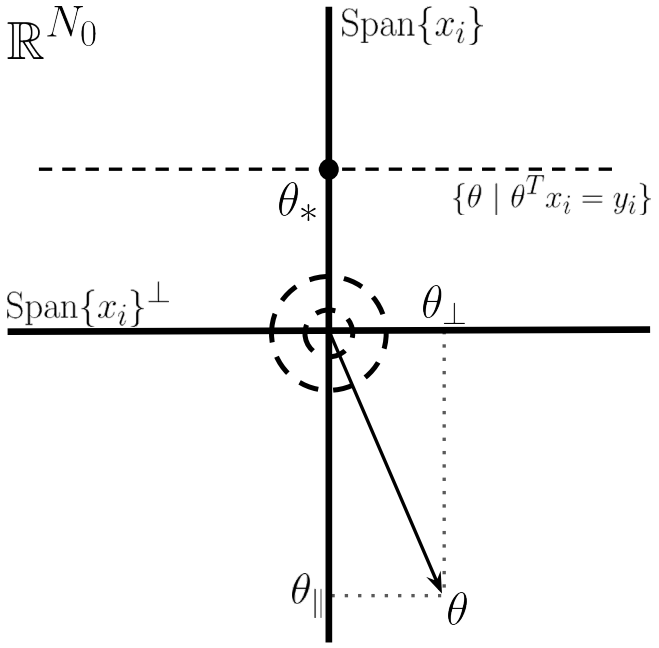

Fig. 1. The input space $\mathbb{R}^{N_0}$ is decomposed into directions $\mathrm{Span}\{x_i\}$ spanned by inputs from the training data and its orthogonal complement. The minimal ℓ_2 norm interpolant is the intersection between the space of interpolants (i.e., the dashed line showing all θ satisfying $\theta^T x_i = y_i$ for all i) and $\mathrm{Span}\{x_i\}$. When generating predictions $f(x) = \theta^T x$, the parameter vector θ can be decomposed into its projections $\theta_{||}$ and $\theta_\perp$ onto $\mathrm{Span}\{x_i\}$ and $\mathrm{Span}\{x_i\}^\perp$, respectively. When fully trained, $\theta_{||}$ will equal θ_*. Circles centered at the origin represent equiprobable lines for the prior over θ, which is always radial but non-Gaussian for any $L \geq 1$.

with $L \geq 1$ that differentiates them from the linear models obtained by taking $L = 0$.

Relation to Prior Work. To put our work in context, we briefly summarize prior approaches to understanding learning with neural networks. The neural tangent kernel (NTK) and other kernel-based models (13–17) reduce neural networks to linear models. Training by gradient descent or Bayesian inference consequently does not affect predictions $f(x_\perp)$ of test inputs orthogonal to the training dataset. Moreover, the NTK regime only considers the limit of infinite width with finite depth and dataset size. More recent work (18–24) shows that feature learning emerges when taking depth and width to infinity simultaneously with the effective prior depth $\lambda_{\text{prior}} = L/N$ Eq. **9** determining both the behavior under both gradient descent and Bayesian inference. Still, such analyses are restricted to finite dataset size P (or more precisely dataset sizes that are much smaller than both network depth and width).

Phenomena such as double descent (25, 26) and benign overfitting (27) appear in linear models when taking width, dataset size, and dataset dimension to infinity simultaneously (28–34). However, as mentioned previously, these linear models do not learn to make data-adaptive predictions in directions not already present in the training data. Moreover, such approaches are restricted to studying fixed depth. Other approaches consider neural networks in the mean-field limit (35–42). In this regime, networks do make data-adaptive predictions $f(x_\perp)$. However, mean-field limits have only been considered at fixed depth and recast optimization in terms of complex nonlinear evolution equations, whose dependence on the training data is typically difficult to access.

The literature most directly related to the present article studies Bayesian inference with either nonlinear (43–48) or linear networks (10–12, 49). These works consider only the regime where depth is either fixed or much smaller than both width and size of the training dataset. We find, in contrast, that the full role of depth in model selection and extrapolation can only be understood in the regime where depth, width, and dataset size are simultaneously large. Finally, our results characterize zero temperature Bayesian inference with deep neural networks at any joint scaling of depth, width, dataset size, and dataset dimension. In this sense, they can be viewed as giving exact expressions for the predictive posterior over deep Gaussian processes (50) with Euclidean covariance in every layer.

Overview of Results. Our first result, Theorem **1** below, gives exact nonasymptotic formulas for both the predictive posterior (i.e., the distribution of $f(x)$ jointly for all inputs x when $W^{(\ell)}$ are drawn from the posterior) and the Bayesian model evidence in terms of a class of meromorphic special functions in one complex variable called Meijer-G functions (51). These results hold for arbitrary training datasets, input dimensions, hidden layer widths, and network depth. They represent an enlargement of the class of priors over θ for which posteriors can be computed in closed form. In particular, they show that zero noise Bayesian inference is exactly solvable for deep Gaussian processes (50) with Euclidean covariances.

To glean insights from the nonasymptotic results in Theorem **1**, we provide in Theorem **2** new asymptotic expansions of Meijer-G functions that allow us to compute expressions for the Bayesian model evidence and the predictive posterior under essentially any joint scalings of P, N_ℓ, L in the large-data limit where $P \to \infty$ with $P/N_0 \to \alpha_0 \in (0, 1)$. To understand the role of depth, we consider regimes in which L either stays finite or grows together with P, N_ℓ.

We focus in this article on zero-noise, or interpolating, posteriors that fit the training data exactly Eq. **6** and the discussion in Optimal Extrapolation. For parametric models, interpolation often causes overfitting at large sample sizes. An important empirical (52) and theoretical (25, 27, 28) observation, however, is that in many nonparameteric overparameterized models, such as the deep linear neural networks we study, this does not occur. Our results therefore give new information about the joint effects of depth, width, and sample size on the nature of interpolating models.

What emerges from our analysis is a rich new picture of the role of depth in linear networks in determining the nature of extrapolation and Bayesian model selection, given by maximizing the Bayesian model evidence, i.e., the likelihood of the training data under the posterior (cf §4, §5 in ref. 53). We present here an informal explanation of our main results, starting with Theorem **3** which implies the following:

> **Takeaway:** Evidence-maximizing priors give the same Gaussian predictive posterior for any architecture in the large-data limit.

This distinguished posterior represents, from a Bayesian point of view, a notion of optimal extrapolation. Indeed, because $f(x)$ is linear in x, it is natural to decompose

$$x = x_{||} + x_\perp, \qquad x \in \mathbb{R}^{N_0},$$

where $x_{||}$ is the projection of x onto directions spanned by inputs from the training data, and $x_\perp$ is the projection of x onto the orthogonal complement (like the decomposition of θ in Fig. 1).

Predictions under the evidence-maximizing posterior at test input x are Gaussian and take the form

$$f(x) \sim \mathcal{N}\left(\theta_*^T x_{||},\ \frac{||\theta_*||^2}{\alpha_0}||x_\perp||^2\right), \qquad [2]$$

where θ_* is the minimal norm interpolant of the training data (cf. Fig. 1 and Eq. **17**). As we explain in Optimal Extrapolation below, the deterministic mean $\theta_*^T x_{||}$ appears because, by construction, our posteriors are concentrated on θ that interpolate the training data (Eq. **6**) and thus we must have $\theta_{||} = \theta_*$. The scalar $||\theta_*||^2/\alpha_0$, in contrast, sets a data-dependent variance for making predictions in directions orthogonal to inputs in the training data. This particular value for the variance is the most likely given that our posteriors are necessarily isotropic in directions orthogonal to the training data. Moreover, the approximate normality of the evidence-maximizing posterior at large sample sizes is due to the fact that the coordinates of a spherically symmetric random vector in high dimensions are approximately Gaussian (see the discussion just below Eq. **12**).

In general, a linear network that maximizes Bayesian evidence, and hence produces the posterior Eq. **2**, may require the prior distribution over network weights to be data-dependent. In machine learning contexts, we hope instead to find optimal but data-agnostic priors. Theorems **6** and **8** (in combination with Theorem **3**) show this is possible in linear networks with large depth. Informally they give:

> **Takeaway:** Wide linear networks (at comparable depth, width, number of data points) with data-agnostic priors give the same predictive posterior as shallow networks with optimal data-dependent priors.

This result highlights the remarkable role of depth in shaping the posterior over predictions in directions of feature space orthogonal to those present in the training data. This can only happen in nonlinear models such as deep linear networks. Quantifying how large network depth must be to ensure optimal extrapolation is explained in Theorem **8**, which provides universal scaling laws for Bayesian posteriors in terms of a single parameter that couples depth, width, and dataset size. Informally, we have the following:

> **Takeaway:** Consider linear networks in the regime $1 \ll$ depth, dataset size $\ll$ width. With data-agnostic priors, the posterior depends only on the effective posterior depth
>
> $$\lambda_{\text{post}} := \frac{(\text{network depth}) \times (\text{dataset size})}{\text{network width}}.$$
>
> As $\lambda_{\text{post}} \to \infty$, evidence grows and the posterior converges to the evidence-maximizing posterior Eq. **2**.

Since λ_{post} determines both the bias and the variance of the posterior, the preceding takeaway can be viewed as a scaling law relating depth, width, and training set size (1–4). In particular, it shows that for large linear networks it is λ_{post}, rather than depth, width, and dataset size separately, that determines the quality of the learned model.

As the preceding statement suggests, at least with data-agnostic priors and wide linear networks, maximizing Bayesian evidence requires large depth, as measured by λ_{post}. Moreover, evidence maximization is not possible at finite λ_{post}. The final result we present (Theorem **6**) concerns maximization of Bayesian evidence—a principled method of comparing different architectures (53)—and is summarized as follows:

> **Takeaway:** With data-agnostic priors and width that is proportional to dataset size, Bayesian evidence is maximal in networks with depth equal to a data-dependent constant times width.

Mis-specification of this constant only results in an order one decrease in evidence and does not affect the posterior. In comparison, a network with smaller depth has exponentially smaller evidence and a suboptimal posterior. The preceding takeaways give perhaps the first principled Bayesian justification for preferring neural networks with large depth, albeit in the restricted setting of linear networks.

Preliminaries

Setup. We fix a training set with P examples

$$X_{N_0} = (x_{1,N_0}, \ldots, x_{P,N_0}) \in \mathbb{R}^{N_0 \times P},\ Y_{N_0} = (y_1, \ldots, y_P) \in \mathbb{R}^P.$$

We will assume X_{N_0} has full rank. Since we study zero noise posteriors supported on the models that minimize the likelihood Eq. **4**, we assume also that $1 \le P \le N_0$. Otherwise, the set of minima of the likelihood consists of a single θ and our posteriors would have zero variance. A key role in our results will be played by the minimal ℓ_2-norm solution to ordinary linear least squares regression of Y_{N_0} onto X_{N_0}:

$$\theta_{*,N_0} := \underset{\theta \in \mathbb{R}^{N_0}}{\arg\min} ||\theta||_2 \quad \text{s.t.} \quad \theta^T X_{N_0} = Y_{N_0}. \qquad [3]$$

Further, we fix $N_1, \ldots, N_L \ge 1$ and consider fitting the training data (X_{N_0}, Y_{N_0}) by a linear model

$$f(x) = \theta^T x \in \mathbb{R}, \quad \theta, x \in \mathbb{R}^{N_0},$$

equipped with quadratic negative log-likelihood

$$\mathcal{L}(\theta \mid X_{N_0}, Y_{N_0}) := \frac{1}{2}||\theta^T X_{N_0} - Y_{N_0}||_2^2, \qquad [4]$$

and a deep linear prior

$$\theta \sim \mathbb{P}_{\text{prior}} \quad \Longleftrightarrow \quad \theta = W^{(L+1)} \cdots W^{(1)}, \qquad [5]$$

in which

$$W^{(\ell)} \in \mathbb{R}^{N_\ell \times N_{\ell-1}}, \quad W_{ij}^{(\ell)} \sim \mathcal{N}\left(0, \frac{\sigma^2}{N_{\ell-1}}\right) \quad \text{independent},$$

where $\sigma > 0$. Our goal is to study the posterior distribution over the set of $\theta \in \mathbb{R}^{N_0}$ that exactly fit the training data. Explicitly, writing $d\mathbb{P}_{\text{prior}}(\theta)$ for the prior density, we study zero noise posteriors

$$d\mathbb{P}_{\text{post}}\left(\theta \mid L, N_\ell, \sigma^2, X_{N_0}, Y_{N_0}\right) := \lim_{\beta \to \infty} \frac{d\mathbb{P}_{\text{prior}}\left(\theta \mid N_0, L, N_\ell, \sigma^2\right) \exp\left[-\beta \mathcal{L}(\theta \mid X_{N_0}, Y_{N_0})\right]}{Z_\beta\left(X_{N_0}, Y_{N_0} \mid L, N_\ell, \sigma^2\right)}. \qquad [6]$$

Writing $\mathbb{E}_{\text{post}}[\cdot]$ for the expectation with respect to the posterior Eq. **6**, we describe the posterior by giving exact formulas for its characteristic function

$$\mathbb{E}_{\text{post}}\left[\exp\left\{-i\mathbf{t} \cdot \theta\right\}\right] = \frac{Z_\infty(\mathbf{t})}{Z_\infty(\mathbf{0})}, \qquad \mathbf{t}, \theta \in \mathbb{R}^{N_0}, \qquad [7]$$

where $Z_\infty(\mathbf{t}) = Z_\infty(\mathbf{t} \mid L, N_\ell, \sigma^2, X_{N_0}, Y_{N_0})$ is the zero-temperature partition function given by taking $\beta \to \infty$ in

$$Z_\beta(\mathbf{t}) := A_\beta \int \exp\left[-\frac{\beta}{2}||Y - \prod_{\ell=1}^{L+1} W^{(\ell)} X||_2^2 - i\theta \cdot \mathbf{t} - \sum_{\ell=1}^{L+1} \frac{N_{\ell-1}}{2\sigma^2}||W^{(\ell)}||_F^2 \right] \prod_{\ell=1}^{L+1} dW^{(\ell)}. \quad [8]$$

The normalizing constant A_β cancels in the ratio Eq. **6** and in any computations involving maximizing ratios of model evidence (*SI Appendix*, B).

The denominator $Z_\infty(\mathbf{0})$ is often called the Bayesian model evidence and represents the probability of the data (X_{N_0}, Y_{N_0}) given the model (i.e., the depth L, layer widths $N_1, \ldots, N_L$ and prior variance σ^2). As detailed in §4, §5 of ref. 53, maximizing the Bayesian model evidence is therefore equivalent to maximum likelihood estimation over the space of models and gives a principled way to select among different models, all of which interpolate the training data. Before stating our technical results, we briefly explain how to compute effective depth and how to reason about optimal extrapolation in linear networks.

Effective Depth. The number of layers L does not provide a useful measure of complexity for the prior distribution over network outputs $f(x) = \theta^T x$ when θ is drawn from the deep linear prior Eq. **5**. This is true in both linear and nonlinear networks at large width (e.g., refs. 18, 21, 54 and 23 for a treatment of deep nonlinear networks). A more useful notion of depth is

$$\lambda_{\text{prior}} = \text{ effective depth of prior } := \sum_{\ell=1}^{L} \frac{1}{N_\ell}, \quad [9]$$

and it is indeed λ_{prior} that plays an important role in our results. Let us provide a brief justification for why λ_{prior} is a natural measure of complexity for the prior (*SI Appendix*, B). With $N_1 = \cdots = N_L = N$, Theorem 1.2 in ref. 55 shows that when $\sigma^2 = 1$, under the prior, the squared singular values of $\left(W^{(L)} \cdots W^{(1)}\right)^{1/L}$ converge to the uniform distribution on $[0, 1]$. Hence, only the squared singular values of $\left(W^{(L)} \cdots W^{(1)}\right)^{1/L}$ lying in intervals of the form $[1 - CL^{-1}, 1]$ correspond to singular values of $W^{(L)} \cdots W^{(1)}$ that remain uniformly bounded away from 0 at large L. At least heuristically, this implies that $W^{(L)} \cdots W^{(1)}$ is supported on matrices of rank approximately $\lambda_{\text{prior}}^{-1}$.*

Viewing $\lambda_{\text{prior}}^{-1}$ as a natural measure of the number of degrees of freedom in the prior motivates the introduction of a posterior effective depth

$$\lambda_{\text{post}} := \frac{P}{\lambda_{\text{prior}}^{-1}} = \sum_{\ell=1}^{L} \frac{P}{N_\ell}. \quad [10]$$

λ_{post} is a ratio between the number of degrees of freedom in the training data (given by the number of training data points) and in the prior. We will see in Theorem **8** that it is precisely λ_{post} that controls the structure of the posterior.

*We do not know this for sure since uniformity for the distribution of singular values at the right edge of the spectrum does not follow from a result only about the global density of singular values.

Optimal Extrapolation. This section explains key structural properties of Bayesian posteriors in linear networks. Consider the model $f(x) = \theta^T x$ and decompose the parameters $\theta = \theta_{||} + \theta_\perp$ as in Fig. 1. Since zero noise posteriors fit the training data, we have

$$\theta \sim \mathbb{P}_{\text{post}} \quad \Longrightarrow \quad \theta_{||} = \theta_{*,N_0},$$

where θ_{*,N_0} is the minimum-norm interpolant Eq. **3**. Moreover, the prior over θ is invariant under all orthogonal transformations and the likelihood is invariant under arbitrary transformations of $\theta_\perp$. Hence, in distribution

$$\theta \sim \mathbb{P}_{\text{post}} \quad \Longrightarrow \quad \theta_\perp \overset{d}{=} u \cdot ||\theta_\perp||, \quad [11]$$

where u is independent of $||\theta_\perp||$ and is uniformly distributed on the unit sphere in $\text{col}(X_{N_0})^\perp \subseteq \mathbb{R}^{N_0}$. The only degree of freedom in the posterior is therefore the distribution of the radial part $||\theta_\perp||$ of the vector $\theta_\perp$. Given a test data point $x = x_{||} + x_\perp$, we find

$$\theta \sim \mathbb{P}_{\text{post}} \quad \Longrightarrow \quad f(x) = (\theta_{*,N_0})^T x_{||} + \|\theta_\perp\| \cdot u^T x_\perp.$$

The distribution of $\|\theta_\perp\|$ controls the scale of predictions for data not spanned by the training set, i.e., for the task of extrapolation. For example if $x = x_{||} \in \text{col}(X_{N_0})$, then $f(x)$ has zero variance since it is determined completely by the training data. More generally, by the Poincare–Borel theorem (refs. 56 and 57) we have for $N_0, P \gg 1$ that

$$f(x) \approx \mathcal{N}\left((\theta_{*,N_0})^T x_{||},\ ||\theta_\perp||^2 \frac{||x_\perp||^2}{N_0 - P} \right). \quad [12]$$

Indeed, since $\widehat{\theta}_\perp = \theta_\perp / ||\theta_\perp|| \in \mathbb{R}^{N_0 - P}$ is rotationally invariant under the posterior, the Poincare–Borel theorem (e.g., Theorem 2 in [21] together with the fact that the mixing measure μ is precisely a point-mass at 1 by (1) in ref. 56) shows that the joint distribution of any fixed (or even slowly growing) number of marginals $\{\widehat{\theta}_\perp^T x_1, \ldots, \widehat{\theta}_\perp^T x_k\}$ is approximately Gaussian when $N_0 - P$ is large. In particular, since predictions under the posterior are of the $f(x; \theta) = \theta_{||}^T x_{||} + ||\theta_\perp|| \widehat{\theta}_\perp^T x_\perp$ and $||\theta_\perp||$ is independent of $\widehat{\theta}_\perp$, we see that in the high-dimensional regime $N_0, P \gg 1$ the finite-dimensional distributions of the posterior are approximately normal.

At $L = 0$, the prior and posterior distributions over $\|\theta_\perp\|^2$ are identical, preventing any feature learning from occurring. For $L \geq 1$, all components of θ are correlated under the prior, allowing information from the training data to be encoded into $\theta_\perp$. We shall see (Theorem **8**) that λ_{post} quantifies how much information the model learns about $\theta_\perp$. In particular, increasing λ_{post} causes the posterior distribution of $\|\theta_\perp\|^2$ to be more and more concentrated around a particular value:

$$\|\theta_\perp\|^2 \approx ||\theta_{*,N_0}||^2 (1 - \alpha_0)/\alpha_0.$$

This special choice of scale maximizes Bayesian evidence, in accordance with the first takeaway described in the Introduction section. It corresponds to the natural estimate for the true signal strength $||v||^2$ under a zero noise generative process $y_i = v^T x_i$ in which x_i are isotropic, given that one observes only the projection $v_{||} = \theta_{*,N_0}$ of v onto directions in the training data.

Main Results

Nonasymptotic Results. We are now ready to formulate our first main result (Theorem **1**), which expresses the partition function $Z_\infty(\mathbf{t})$ defined in Eq. **8** in terms of the Meijer-G function; this allows the Bayesian evidence and predictive posterior to be written in exact closed form for any choice of network depth, hidden layer widths, dataset size, and input dimension. Compared to prior work using either iterative saddle point approximations of integrals encoding the Bayesian posterior (10) or more involved methods such as the replica trick (12), our method provides a direct representation of the network posterior via the partition function in terms of a single contour integral without specializing to limiting cases. Additional quantities, such as the variance of the posterior, are simply expressed as a ratio of Meijer-G functions. We shall later recover known limiting cases and uncover new asymptotic results from expansions of the Meijer-G function (Theorem **2**).

Theorem 1 (Predictive Posterior and Evidence). *Fix $P, L, N_0, \ldots, N_L \geq 1$, $\sigma^2 > 0$ as well as training data X_{N_0}, Y_{N_0}. Fix $\boldsymbol{t} \in \mathbb{R}^{N_0}$ and write*

$$\boldsymbol{t} = \boldsymbol{t}_{||} + \boldsymbol{t}_{\perp}, \qquad \boldsymbol{t}_{||} \in \mathrm{col}(X_{N_0}), \quad \boldsymbol{t}_{\perp} \in \mathrm{col}(X_{N_0})^{\perp}.$$

Define $4M = \prod_{\ell=0}^{L} 2\sigma^2/N_\ell$ and introduce the following shorthand for the Meijer-G functions (SI Appendix, A) with parameters given by layer widths:

$$G^{L+1,0}_{0,L+1}\left(\frac{||\theta_{*,N_0}||^2}{4M} \,\middle|\, \underset{\frac{P}{2}, \frac{\boldsymbol{N}}{2}+k}{-}\right) := G^{L+1,0}_{0,L+1}\left(\frac{||\theta_{*,N_0}||^2}{4M} \,\middle|\, \underset{\frac{P}{2}, \frac{N_1}{2}+k, \ldots, \frac{N_L}{2}+k}{-}\right).$$

*The partition function $Z_\infty(\boldsymbol{t})$ of the predictive posterior defined in Eq. **8** is*

$$Z_\infty(\boldsymbol{t}) = \left(\frac{4\pi}{||\theta_*||^2}\right)^{\frac{P}{2}} \exp\left[-i\langle\theta_{*,N_0}, \boldsymbol{t}\rangle\right] \prod_{\ell=1}^{L} \Gamma\left(\frac{N_\ell}{2}\right)^{-1} \times \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \left|\left|\boldsymbol{t}_{\perp}\right|\right|^{2k} M^k G^{L+1,0}_{0,L+1}\left(\frac{||\theta_{*,N_0}||^2}{4M} \,\middle|\, \underset{\frac{P}{2}, \frac{\boldsymbol{N}}{2}+k}{-}\right). \quad [13]$$

In particular, the Bayesian model evidence $Z_\infty(\mathbf{0})$ equals

$$\frac{(4\pi)^{P/2}}{||\theta_*||^P \prod_{\ell=1}^{L} \Gamma\left(\frac{N_\ell}{2}\right)} G^{L+1,0}_{0,L+1}\left(\frac{||\theta_{*,N_0}||^2}{4M} \,\middle|\, \underset{\frac{P}{2}, \frac{\boldsymbol{N}}{2}}{-}\right). \quad [14]$$

Further, given $x \in \mathbb{R}^{N_0}$, the mean of the predictive posterior is

$$\mathbb{E}_{post}\left[f(x)\right] = (\theta_{*,N_0})^T x, \quad [15]$$

while the posterior variance $\mathrm{Var}_{\mathrm{post}}\left[f(x)\right]$ is

$$2M||x_{\perp}||^2 \frac{G^{L+1,0}_{0,L+1}\left(\frac{||\theta_{*,N_0}||^2}{4M} \,\middle|\, \underset{\frac{P}{2}, \frac{\boldsymbol{N}}{2}+1}{-}\right)}{G^{L+1,0}_{0,L+1}\left(\frac{||\theta_{*,N_0}||^2}{4M} \,\middle|\, \underset{\frac{P}{2}, \frac{\boldsymbol{N}}{2}}{-}\right)}, \quad [16]$$

where $x_\perp$ is the projection of x onto the orthogonal complement of the span $\mathrm{col}(X_{N_0})$ of the training data.

SI Appendix, C for a proof.

Asymptotic Results. To evaluate the predictive posterior and evidence in Theorem **1** in the limits where N_0, P, N_ℓ (and potentially L) tend to infinity, we require expansions of the Meijer-G function obtained by the Laplace method. We are interested in regimes where N_0, P grow and will assume a mild compatibility condition on the training data: for all $\alpha_0 \in (0, 1)$, we assume there exists constant $||\theta_*||$ such that

$$\lim_{\substack{P,N_0\to\infty \\ P/N_0\to\alpha_0\in(0,1)}} ||\theta_{*,N_0}|| = ||\theta_*||, \quad [17]$$

where the convergence is in distribution. This assumption is very generic and is (for example) satisfied for a Gaussian data model where inputs are Gaussian

$$X_{N_0} = (x_{i,N_0},\ i = 1, \ldots, P), \quad x_{i,N_0} \sim \mathcal{N}(0, \Sigma_{N_0}) \quad \text{iid},$$

outputs are linear plus noise

$$Y = V_{N_0} X_{N_0} + \epsilon_{N_0},\ V_{N_0} \sim \mathrm{Unif}\left(S^{N_0-1}\right), \quad \epsilon_{N_0} \sim \mathcal{N}\left(0, \sigma_\epsilon^2 I_{N_0}\right),$$

and the spectral density of the design matrices Σ_{N_0} converges weakly as $N_0 \to \infty$ to a fixed probability measure on $\mathbb{R}_+$ with finite moments.

To minimize notation, we report here the expansions in terms of a single layer width $N = N_1 = \cdots = N_L$, but expansions with distinct N_ℓ (and to higher order) are provided in the proof (*SI Appendix*, D).

Theorem 2 (Asymptotic Expansions of Meijer-*G*). *Set $N_1, \ldots, N_L = N$ and define $\boldsymbol{N} = (N_1, \ldots, N_L)$. Suppose that the training data satisfies* Eq. **17**. *In different limiting cases such that $\{P, N\} \to \infty$ with fixed $P/N_0 = \alpha_0$, we evaluate the quantities*

$$\log G := \log G^{L+1,0}_{0,L+1}\left(\frac{||\theta_{*,N_0}||^2}{4M} \,\middle|\, \underset{\frac{P}{2}, \frac{\boldsymbol{N}}{2}+k}{-}\right)$$

$$\Delta(\log G)[k] := \log G^{L+1,0}_{0,L+1}\left(\frac{||\theta_{*,N_0}||^2}{4M} \,\middle|\, \underset{\frac{P}{2}, \frac{\boldsymbol{N}}{2}+k}{-}\right) - \log G^{L+1,0}_{0,L+1}\left(\frac{||\theta_{*,N_0}||^2}{4M} \,\middle|\, \underset{\frac{P}{2}, \frac{\boldsymbol{N}}{2}}{-}\right).$$

We will see in each case that to leading order $\log G$ does not depend on k while $\Delta(\log G)[k]$ does.

1. *Fix $L < \infty$, $\alpha, \sigma^2 > 0$. Suppose $P, N \to \infty$ with $P/N \to \alpha$. Then,*

$$\log G = \frac{N\alpha}{2}\left[\log\left(\frac{N\alpha}{2}\right) + \log\left(1 + \frac{z_*}{\alpha}\right) - \left(1 + \frac{z_*}{\alpha}\right)\right] + \frac{NL}{2}\left[\log\left(\frac{N}{2}\right) + \log\left(1 + z_*\right) - \left(1 + z_*\right)\right], \quad [18]$$

plus an error of size $\tilde{O}(1)$ and

$$\Delta(\log G)[k] = kL\left[\log\left(\frac{N}{2}\right) + \log(1 + z_*)\right], \quad [19]$$

plus an error of size $\tilde{O}(1/N)$, where $z_* > \min\{-\alpha, -1\}$ is the unique solution to

$$\left(1 + \frac{z_*}{\alpha}\right)(1 + z_*)^L = \frac{||\theta_*||^2}{\sigma^{2(L+1)}\alpha_0}. \quad [20]$$

2. *Fix* $\lambda_{\text{prior}}, \alpha > 0$ *and suppose* $L = \lambda_{\text{prior}} N$, $P, N \to \infty$, $P/N \to \alpha$, $\sigma^2 = 1$. *Then,*

$$\log G = \frac{\lambda_{\text{prior}} N^2}{2}\left[\log\left(\frac{N}{2}\right) - 1\right] + \frac{N\alpha}{2}\left[\log\left(\frac{N\alpha}{2}\right) - 1\right] + \frac{\lambda_{\text{prior}} N}{2}\left[-\log\left(\frac{N}{2}\right) + \log(2\pi)\right] + \tilde{O}(1), \quad [21]$$

and, up to an error of size $\tilde{O}(1/N)$,

$$\Delta(\log G)[k] = k\left[\lambda_{\text{prior}} N \log\left(\frac{N}{2}\right) + \log\left(\frac{||\theta_*||^2}{\alpha_0}\right)\right]. \quad [22]$$

3. *Fix* $\lambda_{\text{post}} > 0$. *Suppose* $N, P, L \to \infty$ *with* $LP/N \to \lambda_{\text{post}}$, $\sigma^2 = 1$ *and* $L/N \to 0$. *Then,*

$$\log G = \frac{P}{2}\left[\log\left(\frac{P}{2}\right) - 1\right] + \frac{\lambda_{\text{post}} N}{2}\frac{N}{P}\left[\log\left(\frac{N}{2}\right) - 1\right] + \frac{P}{2}\left[\log(1 + t_*) - t_*\left(1 + \frac{\lambda_{\text{post}} t_*}{2}\right)\right] + \frac{\lambda_{\text{post}} N}{2}\left[\frac{N}{P}\left(1 + \frac{P}{N}t_*\right)\log\left(1 + \frac{P}{N}t_*\right) - t_*\right] + \tilde{O}(1), \quad [23]$$

and

$$\Delta(\log G)[k] = k\left[L \log\left(\frac{N}{2}\right) + \lambda_{\text{post}} t_*\right] + \tilde{O}\left(\frac{1}{N}\right), \quad [24]$$

where t_* *is the unique solution to*

$$e^{\lambda_{\text{post}} t_*}(1 + t_*) = ||\theta_*||^2/\alpha_0. \quad [25]$$

In all the estimates above, $\tilde{O}(1) = O(\max\{\log P, \log N\})$ *suppresses lower-order terms of order 1, up to logarithmic factors; similarly,* $\tilde{O}(1/N) = O(\max\{\log P, \log N\}/N)$. *Suppressed terms are included in SI Appendix, D.*

Model Selection and Extrapolation. We combine our nonasymptotic formulas for the posterior and evidence from Theorem **1** with the Meijer-G function expansions from Theorem **2** to investigate two fundamental questions about Bayesian interpolation with linear networks. Together they give, at least in the restricted setting of deep linear networks with output dimension 1, a principled reason to prefer deeper networks.

- **Model Selection.** How should we choose the prior weight variance σ^2, model depth L and layer widths N_ℓ? Recall that the Bayesian model evidence $Z_\infty(\mathbf{0})$ from Eq. **7** represents the likelihood of observing the data (X_{N_0}, Y_{N_0}) given the architecture L, N_ℓ and the prior weight variance σ^2. Maximizing the Bayesian evidence is therefore maximum likelihood estimation over the space of models and gives a principled method for model selection. We shall see that data-agnostic priors maximize the evidence for networks with infinite λ_{post}, while shallower networks require data-dependent priors.
- **Extrapolation.** How optimal is the posterior of a linear network? Predictions on inputs from directions orthogonal to the training data are determined by the distribution $\theta_\perp$. At $\lambda_{\text{post}} = 0$, its prior and posterior coincide. As λ_{post} increases, however, we shall find that information from the training set mixes into $\theta_\perp$ and ultimately maximizes evidence, producing optimal extrapolation.

For simplicity, we report our results here in terms of a single hidden layer width $N = N_1 = \cdots = N_L$. The general form with generic N_ℓ, as well as related results not stated here, are available by direction application of the general expansions in *SI Appendix*, D. For convenience, we define

$$\nu := ||\theta_*||^2/\alpha_0. \quad [26]$$

As elsewhere, we emphasize that we focus on the regime

$$P/N_0 \to \alpha_0 \in (0, 1),$$

as $P, N_0 \to \infty$. When $\alpha_0 \geq 1$, Theorem **1** still holds but immediately yields that $\theta = \theta_{*,N_0}$ almost surely since $\theta_\perp = 0$. As described by Eq. **12** and stated rigorously in *SI Appendix*, A, the predictive posterior in the regime $P < N_0$ is Gaussian with variance determined by $||\theta_\perp||^2$, which converges to a constant. We rewrite this posterior in the form

$$f(x) \quad \to \quad \mathcal{N}(\mu_*, \nu c \Sigma_\perp), \qquad x = (x_{i,N_0},\ i = 1, \ldots, k),$$

for free scalar c, scalar μ_*, and $k \times k$ PSD matrix $\Sigma_\perp$ given by

$$\mu_* := \langle \theta_{*,N_0}, x \rangle, \quad \Sigma_\perp = \frac{\left\langle x^\perp_{i,N_0}, x^\perp_{j,N_0} \right\rangle}{N_0 - P},$$

in the limit $P, N_0 \to \infty$ and $P/N_0 \to \alpha_0 \in (0, 1)$. The following results report the Bayesian evidence and value of c under different join scalings of P, N, and L. First, we observe that maximizing Bayesian evidence always results in the posterior corresponding to $c = 1$ (proven in *SI Appendix*, E).

Theorem 3 (Universal Maximal Evidence Posterior). *Fix* L, $N_1 \ldots, N_L \geq 1$, $\sigma^2 > 0$, $\alpha_0 \in (0, 1)$, *and consider sequences of training datasets* $X_{N_0} \in \mathbb{R}^{N_0 \times P}$, $Y_{N_0} \in \mathbb{R}^{1 \times P}$ *such that*

$$P, N_0 \to \infty, \qquad P/N_0 \to \alpha_0,$$

and Eq. **17** *holds. Let* $Z_\infty(\mathbf{0})$ *denote the limiting Bayesian model evidence. The following statements are equivalent:*

(a) σ^2 *maximizes* $Z_\infty(\mathbf{0})$, *and*
(b) the posterior over predictions $f(x)$ *converges weakly to a Gaussian* $\mathcal{N}(\mu_*, \nu\Sigma_\perp)$.

We emphasize that Theorem **3** holds for any choice of depth and hidden layer widths. At finite depth, we shall see that identifying the evidence-maximizing prior σ^2_* requires knowledge of the data-dependent parameter ν. In contrast, infinite-depth networks have evidence maximized by $\sigma^2_* = 1$, allowing them to successfully infer ν from training data despite a data-agnostic prior. We proceed to consider different joint scalings of P, N, and L in Theorems **4**, **6** and **8**, which follow directly from writing the partition function in terms of the Meijer-G function (Theorem **1**) and applying the suitable asymptotic expansion (Theorem **2**).

Finite depth. We present here a characterization of the infinite-width posterior and model evidence for networks with a fixed number of hidden layers.

Theorem 4 (Posterior and Evidence at Finite *L*). *For each $N_0, P \geq 1$, consider training data X_{N_0}, Y_{N_0} satisfying* Eq. **17**. *Fix constants $L \geq 0$, $\sigma^2 > 0$ and suppose that*

$$P, N_0, N \to \infty, \quad P/N_0 \to \alpha_0 \in (0,1), \quad P/N \to \alpha \in (0,\infty).$$

In this limit, the posterior over predictions $f(x)$ converges weakly to a Gaussian

$$\mathcal{N}\left(\mu_*,\ \nu\Sigma_\perp\left(1+\frac{z_*}{\alpha}\right)^{-1}\right),$$

where z_* is the unique solution to

$$\frac{\nu}{\sigma^{2(L+1)}} = \left(1+\frac{z_*}{\alpha}\right)(1+z_*)^L, \quad z_* > \max\{-1,-\alpha\}. \tag{27}$$

Additionally, we have the following large-P expansion of the Bayesian model evidence

$$\log Z_\infty(\mathbf{0}) \to \frac{P}{2}\left[\log\left(\frac{P}{2}\right)+\log\left(1+\frac{z_*}{\alpha}\right)-\left(1+\frac{z_*}{\alpha}\right)-\log\left(\frac{||\theta_*||^2}{4\pi}\right)\right]+\frac{NL}{2}\left[\log(1+z_*)-z_*\right]+O(\max\{\log P,\log N\}).$$

In regimes where z_* tends to zero, the posterior will converge to the evidence-maximizing posterior independent of the architecture and prior. By Eq. **27**, this occurs for instance in the limit of infinite depth with $\sigma = 1$ or when $\nu = \sigma^{2(L+1)}$. This last condition corresponds to maximizing the Bayesian evidence $Z_\infty(\mathbf{0})$ at finite L.

Corollary 5 (Bayesian Model Selection at Finite *L*). *In the setting of Theorem **4** the Bayesian evidence $Z_\infty(\mathbf{0})$ satisfies:*

$$\sigma_*^2 = \arg\max_\sigma Z_\infty(\mathbf{0}) = \nu^{\frac{1}{L+1}} \tag{28}$$

$$L_* = \arg\max_L Z_\infty(\mathbf{0}) = \frac{\log(\nu)}{\log(\sigma^2)} - 1. \tag{29}$$

In particular, given a prior variance with $\operatorname{sgn}(\nu-1) = \operatorname{sgn}(\sigma^2-1)$ satisfying $|\sigma^2-1| \leq \epsilon$, the optimal depth network satisfies $L_* \geq |\log(\nu)|/\epsilon$.

To put this Corollary into context, note that in the large-width limit of Theorem **4** there are only two remaining model parameters, L and σ^2. Model selection can therefore be done in two ways. The first is an empirical Bayes approach in which one uses the training data to determine a data-dependent prior σ_*^2 given by Eq. **28**. The other approach, which more closely follows the use of neural networks in practice, is to seek a universal, data-agnostic value of σ^2 and optimize instead the network architecture. The expression Eq. **28** shows that the only way to choose σ^2, L to (approximately) maximize model evidence for any fixed ν is to take $\sigma^2 \approx 1$ with $\operatorname{sgn}(\sigma^2-1) = \operatorname{sgn}(\nu-1)$ and $L \to \infty$. Hence, restricting to the data-agnostic prior $\sigma^2 = 1$ naturally leads to a Bayesian preference for infinite-depth networks, regardless of the training data. This motivates us to consider large-N limits in which L tends to infinity, which we take up in the next two sections.

Infinite depth. We fix $\sigma^2 = 1$ and investigate extrapolation and model selection in regimes where $N, L, P \to \infty$ simultaneously.

Theorem 6 (Posterior and Evidence at Fixed λ_{prior}). *For each $N_0, P \geq 1$, consider training data X_{N_0}, Y_{N_0} satisfying* Eq. **17**. *Moreover, fix $\lambda_{\text{prior}}, \alpha \in (0,\infty)$, $\alpha_0 \in (0,1)$. Suppose that $N_1 = \cdots = N_L = N$ and that*

$$P, N_\ell, L \to \infty, \quad P/N_0 \to \alpha_0, \quad P/N \to \alpha, \quad L/N \to \lambda_{\text{prior}}.$$

In this limit, the posterior over predictions $f(x)$ converges weakly to a Gaussian

$$\mathcal{N}(\mu_*,\ \nu\Sigma_\perp),$$

which is independent of $\alpha, \lambda_{\text{prior}}$. Additionally, the evidence admits the following asymptotic expansion:

$$\log Z_\infty(\mathbf{0}) = \frac{P}{2}\left[\log\left(\frac{P}{2}\right)-1-\log\left(\frac{||\theta_*||^2}{4\pi}\right)\right]+O(\max\{\log P,\log N\}).$$

This result highlights the remarkable nature of data-driven extrapolation in deep networks.

Corollary 7 (Optimal Learning by Deep Networks). *In the setting of Theorem **6**, the posterior distribution over regression weights θ is the same in the following two settings:*

- *We fix $L \geq 0$, take the data-dependent prior variance σ_* that maximizes the Bayesian model evidence as a function of the training data and network depth as in Eq. **28**, and send the network width N to infinity.*
- *We fix $\lambda_{\text{prior}} > 0$, take a data-agnostic prior $\sigma^2 = 1$, and send both the network depth $L := \lambda_{\text{prior}} \cdot N$ and network width N to infinity together.*

This corollary makes precise the statement that infinitely deep networks with data-agnostic priors performs as optimally finite depth networks with fine-tuned data-dependent priors.

In *SI Appendix*, F, we find that the ratio of model evidence at fixed depth to the model evidence at infinite depth vanishes like $\exp[-O(N)]$. In comparison, mis-specifying the value of constant λ_{prior} only results in an $O(1)$ ratio of model evidences, and it does not affect the posterior to leading order. We conclude that, at $\sigma^2 = 1$, wide networks with depth comparable to width are robustly preferred to shallow networks.

Scaling laws for optimal learning. To emphasize the similarity between dataset size and depth, we take the limit of $L, P, N \to \infty$ while holding LP/N constant. This results in a scaling law for feature learning that only depends on λ_{post}, as seen in the following theorem.

Theorem 8 (Posterior and Evidence at Fixed λ_{post}). *For each $N_0, P \geq 1$, consider training data X_{N_0}, Y_{N_0} satisfying* Eq. **17**. *Moreover, fix constants $\lambda_{\text{post}} > 0$, $\alpha_0 \in (0,1)$, $\sigma^2 = 1$. Suppose that $N_1 = \cdots = N_L = N$ and suppose that*

$$P, N_\ell, L \to \infty, \ \frac{P}{N_0} \to \alpha_0, \ \frac{P}{N} \to 0, \ \frac{L}{N} \to 0, \ \frac{LP}{N} \to \lambda_{\text{post}}.$$

In this limit, the posterior over predictions $f(x)$ converges weakly to a Gaussian $\mathcal{N}\left(\mu_*,\ \nu\Sigma_\perp(1+t_*)^{-1}\right)$, where t_* is the unique solution to

$$\nu = (1+t_*)e^{\lambda_{\text{post}} t_*}, \quad t_* > -1.$$

Additionally, we have the following asymptotic expansion for the Bayesian model evidence

$$\log Z_\infty(\mathbf{0}) = \frac{P}{2}\left(\log\left(\frac{P}{2}\right) - 1\right) - \frac{P}{2}\log\left(\frac{||\theta_*||^2}{4\pi}\right) + \frac{P}{2}\left[\log(1+t_*) - t_* - \frac{1}{2}\lambda_{\text{post}} t_*^2\right] + \tilde{O}(\max\{\log P, \log N, \log L\}).$$

In the limit $\lambda_{\text{post}} \to \infty$, Theorem **8** implies that optimal feature learning is rapidly approached—specifically, like the harmonic mean of 1 and $\lambda_{\text{post}}/\log \nu$. Bayesian model selection also drives λ_{post} to infinity.

Corollary 9 (Bayesian Model Selection at Fixed λ_{post}). *In the setting of Theorem* ***8****, the Bayesian evidence $Z_\infty(\mathbf{0})$ is monotonically increasing in λ_{post}. Specifically,*

$$\frac{\partial \log Z_\infty(\mathbf{0})}{\partial \lambda_{\text{post}}} = \frac{P t_*^2}{4} \geq 0. \qquad [30]$$

These results provide simple scaling laws that apply independently of the choice of dataset, demonstrating the behavior of the predictor's posterior distribution and model evidence in terms of λ_{post}. The coupling of depth and dataset size in λ_{post} provides an interpretation of depth as a mechanism to improve learning in a manner similar to additional data: Larger datasets and larger depths contribute equally toward aligning the prior $\sigma^2 = 1$ toward the correct posterior.

Properties of Deep Linear Networks. We relate our work to prior results in the literature: variance-limited scaling laws (3) and sample-wise double descent (12, 25, 26). To make the discussion concrete, we shall focus on the architecture introduced in Theorem **6**, where depth, width, and dataset size scale linearly with each other to produce a posterior that exhibits optimal feature learning given prior $\sigma^2 = 1$.

Variance-limited scaling laws. We examine the scaling behavior of model error in the infinite-width or infinite-dataset limits. The work of (3) shows that the difference between the finite-size loss and the infinite-size loss scales like $1/x$ for $x = N$ or P while the other parameter is held fixed (P or N, respectively). Here, we demonstrate an analogous scaling law when $N \propto P \propto L$. Similarly to the results of ref. 3, this scaling law is independent of the choice of dataset and consequently provides a universal insight into how performance improves with larger models or more data. The proof is found in *SI Appendix*, G.

Theorem 10 (Variance-Limited Scaling Law). *For each $N_0 \geq 1$ consider training data X_{N_0}, Y_{N_0} satisfying* Eq. **17**. *Fix constants $\lambda_{\text{prior}} > 0, \alpha > 0$, $\sigma^2 = 1$. Suppose that $N_1 = \cdots = N_L = N$, that $L = \lambda_{\text{prior}} N$ and number of data points satisfies $P = N\alpha$. Then, as $N \to \infty$,*

$$\text{Var}_{\text{post}}[f(x)] = \lim_{N\to\infty} \text{Var}_{\text{post}}[f(x)] + \frac{C}{N} + O\left(\frac{\log N}{N^2}\right),$$

where $C \in \mathbb{R}$ is a universal constant.

Double descent. We demonstrate double descent in $\alpha_0 = P/N_0$ consistent with previous literature (12). As a concrete example, we shall consider a Gaussian data model and evaluate double descent for the posterior of optimal feature learning; this posterior is achieved by, for example, deep networks with $\sigma^2 = 1$ (Theorem **6**), or finite-depth networks with data-tuned priors σ_*^2 (Theorem **4**). The proof is found in *SI Appendix*, H.

Theorem 11 (Double Descent in α_0). *Consider generative data model*

$$x_i \in \mathbb{R}^{N_0} \sim \mathcal{N}(0, \mathrm{I}),\ y_i = V_0 x_i + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2),\ ||V_0||^2 = 1,$$

and posterior distribution $\mathcal{N}(\mu_*, \nu\Sigma_\perp)$. We have error

$$\mathbb{E}_{x,X,\epsilon}\left[\langle f(x) - V_0 x\rangle^2\right] = \begin{cases} \frac{1}{\alpha_0} - \alpha_0 + \frac{1}{1-\alpha_0}\sigma_\epsilon^2, & \alpha_0 < 1 \\ \frac{1}{\alpha_0 - 1}\sigma_\epsilon^2, & \alpha_0 \geq 1 \end{cases},$$

which diverges at $\alpha_0 = 1$.

Discussion

Neural networks are nonlinear functions of their parameters, making an analytic understanding of their properties difficult. Here, we adopted the simplification of studying linear neural networks, which remain nonlinear in their parameters but are more tractable. To conclude, we emphasize three limitations of our work. First, we consider only *linear networks*, which are linear as a function of their inputs. However, they are not linear as a function of their parameters, making learning by Bayesian inference nontrivial. Second, we study learning by Bayesian inference and leave extensions to learning by gradient descent to future work. Finally, our results characterize the predictive posterior, i.e., the distribution over model predictions after inference. It would be interesting to derive the form of the posterior over network weights as well.

Data, Materials, and Software Availability. There are no data associated with this manuscript.

ACKNOWLEDGMENTS. This work started at the 2022 Summer School on the Statistical Physics of Machine Learning held at École de Physique des Houches. We are grateful for the wonderful atmosphere at the school and would like to express our appreciation to the session organizers Florent Krzakala and Lenka Zdeborová as well as to Haim Sompolinsky for his series of lectures on Bayesian analysis of deep linear networks. We further thank Edward George for pointing out the connection between our work and the deep Gaussian process literature. Finally, we thank Matias Cattaneo, Isaac Chuang, David Dunson, Jianqing Fan, Aram Harrow, Jason Klusowski, Cengiz Pehlevan, Veronika Rockova, and Jacob Zavatone-Veth for their feedback and suggestions. B.H. is supported by NSF grants DMS-2143754, DMS-1855684, and DMS-2133806. A.Z. is supported by the Hertz Foundation, and by the DoD NDSEG. We also thank two anonymous reviewers for improving aspects of the exposition and for pointing out a range of typos in the original manuscript.

1. M. Geiger *et al.*, Scaling description of generalization with number of parameters in deep learning. *J. Stat. Mech.: Theory Exp.* **2020**, 023401 (2020).
2. J. Kaplan *et al.*, Scaling laws for neural language models. arXiv [Preprint] (2020). http://arxiv.org/abs/2001.08361 (Accessed 14 May 2023).
3. Y. Bahri , E. Dyer , J. Kaplan , J. Lee , U. Sharma , Explaining neural scaling laws. arXiv [Preprint] (2021) http://arxiv.org/abs/2102.06701 (Accessed 14 May 2023).
4. J. W. Rae *et al.*, Scaling language models: Methods, analysis & insights from training gopher. arXiv [Preprint] (2021). http://arxiv.org/abs/2112.11446 (Accessed 14 May 2023).
5. S. Arora, N. Cohen, N. Golowich, W. Hu, "A convergence analysis of gradient descent for deep linear neural networks" in *ICLR* (2019).
6. S. Arora, N. Cohen, E. Hazan, On the optimization of deep networks: Implicit acceleration by overparameterization. ICML (2018).

7. A. M. Saxe, J. L. McClelland, S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks" in *ICLR* (2014).
8. A. Saxe, S. Sodhani, S. J. Lewallen, "The neural race reduction: Dynamics of abstraction in gated networks" in *International Conference on Machine Learning (PMLR)* (2022), pp. 19287–19309.
9. K. Kawaguchi, "Deep learning without poor local minima" in *Advances in Neural Information Processing Systems* (2016), pp. 586–594.
10. Q. Li, H. Sompolinsky, Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Phys. Rev. X* **11**, 031059. (2021).
11. J. Zavatone-Veth, C. Pehlevan, Exact marginal prior distributions of finite Bayesian neural networks. *Adv. Neural Inf. Process. Syst.* **34** (2021).
12. J. A. Zavatone-Veth, W. L. Tong, C. Pehlevan, Contrasting random and learned features in deep Bayesian linear regression. *Phys. Rev. E* **105**, 064118. (2022).
13. S. S. Du, J. D. Lee, Y. Tian, B. Poczos, A. Singh, "Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima" in *ICML* (2018).
14. S. S. Du, X. Zhai, B. Poczos, A Singh, "Gradient descent provably optimizes over-parameterized neural networks" in *International Conference on Learning Representations* (2019).
15. Z. Allen-Zhu, Y. Li, Y. Liang, "Learning and generalization in overparameterized neural networks, going beyond two layers" in *Advances in Neural Information Processing Systems* (2019), pp. 6158–6169.
16. A. Jacot, F. Gabriel, C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks" in *Advances in Neural Information Processing Systems* (2018), pp. 8571–8580.
17. C. Liu, L. Zhu, M. Belkin, Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Appl. Comput. Harm. Anal.* **59**, 85–116 (2022).
18. B. Hanin, "Which neural net architectures give rise to exploding and vanishing gradients?" in *Advances in Neural Information Processing Systems* (2018).
19. B. Hanin, M. Nica, Products of many large random matrices and gradients in deep neural networks. *Commun. Math. Phys.* **376**, 287–322 (2020).
20. B. Hanin, D. Rolnick, "How to start training: The effect of initialization and architecture" in *Advances in Neural Information Processing Systems* (2018), pp. 571–581.
21. B. Hanin, M. Nica, "Finite depth and width corrections to the neural tangent kernel" in *ICLR 2020* (2020).
22. M. B. Li, M. Nica, D. M. Roy, "The neural covariance SDE: Shaped infinite depth-and-width networks at initialization" in *NeurIPS 2022* (2022).
23. D. A. Roberts, S. Yaida, B. Hanin, *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks* (Cambridge University Press, 2022).
24. S. Yaida, "Non-gaussian processes and neural networks at finite widths" in *MSML* (2020).
25. M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15849–15854 (2019).
26. M. Belkin, D. Hsu, J. Xu, Two models of double descent for weak features. *SIAM J. Math. Data Sci.* **2**, 1167–1180 (2020).
27. P. L. Bartlett, P. M. Long, G. Lugosi, A. Tsigler, Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. U.S.A.*, (2020).
28. T. Hastie, A. Montanari, S. Rosset, R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Stat.* **50**, 949–986 (2022).
29. A. Montanari, Y. Zhong, The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *Ann. Stat.* **50**, 2816–2847 (2022).
30. B. Adlam, J. Levinson, J. Pennington, "A random matrix perspective on mixtures of nonlinearities for deep learning" in *AISTATS* (2022).
31. B. Adlam, J. Pennington, "The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization" in *International Conference on Machine Learning (PMLR)* (2020), pp. 74–84.
32. M. S. Advani, A. M. Saxe, H. Sompolinsky, High-dimensional dynamics of generalization error in neural networks. *Neural Netw.* **132**, 428–446 (2020).
33. S. Mei, A. Montanari, The generalization error of random features regression: Precise asymptotics and the double descent curve. *Commun. Pure Appl. Math*, (2019).
34. S. Mei, T. Misiakiewicz, A. Montanari, Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Appl. Comput. Harm. Anal.*, (2021).
35. S. Mei, A. Montanari, P. M. Nguyen, A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E7665–E7671 (2018).
36. G. Rotskoff, E. Vanden-Eijnden, Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. *Adv. Neural Inf. Process. Syst.* **31** (2018).
37. L. Chizat, F. Bach, "On the global convergence of gradient descent for over-parameterized models using optimal transport" in *Advances in Neural Information Processing Systems* (2018), pp. 3036–3046.
38. J. Sirignano, K. Spiliopoulos, Mean field analysis of neural networks: A central limit theorem. *Stochast. Process. Their Appl.* **130**, 1820–1852 (2020).
39. J. Sirignano, K. Spiliopoulos, Mean field analysis of deep neural networks. *Math. Oper. Res.* (2021).
40. J. Yu, K. Spiliopoulos, Normalization effects on shallow neural networks and related asymptotic expansions. *Found. Data Sci.* **3**, 151–200 (2021).
41. H. T. Pham, P. M. Nguyen, "Global convergence of three-layer neural networks in the mean field regime" in *ICLR* (2021).
42. G. Yang, E. J. Hu, "Tensor programs iv: Feature learning in infinite-width neural networks" in *International Conference on Machine Learning (PMLR)* (2021), pp. 11727–11737.
43. J. Lee *et al.*, "Deep neural networks as gaussian processes" in *ICML 2018* and arXiv [Preprint] (2018). http://arxiv.org/abs/1711.00165 (Accessed 14 May 2023).
44. S. Ariosto *et al.*, Statistical mechanics of deep learning beyond the infinite-width limit. arXiv [Preprint] (2022). http://arxiv.org/abs/2209.04882 (Accessed 14 May 2023).
45. J. Hron, R. Novak, J. Pennington, J. Sohl-Dickstein, "Wide Bayesian neural networks have a simple weight posterior: Theory and accelerated sampling" in *International Conference on Machine Learning (PMLR)* (2022), pp. 8926–8945.
46. H. Cui , F. Krzakala , L. Zdeborová , Optimal learning of deep random networks of extensive-width. arXiv [Preprint] (2023). http://arxiv.org/abs/2302.00375 (Accessed 14 May 2023).
47. G. Naveh, Z. Ringel, A self consistent theory of gaussian processes captures feature learning effects in finite CNNs. *Adv. Neural Inf. Process. Syst.* **34** (2021).
48. I. Seroussi , Z. Ringel , Separation of scales and a thermodynamic description of feature learning in some CNNs. arXiv [Preprint] (2021). http://arxiv.org/abs/2112.15383 (Accessed 14 May 2023).
49. L. Noci, G. Bachmann, K. Roth, S. Nowozin, T. Hofmann, Precise characterization of the prior predictive distribution of deep ReLU networks. *Adv. Neural Inf. Process. Syst.* **34**, 20851–20862 (2021).
50. A. Damianou, N. D. Lawrence, "Deep Gaussian processes" in *Artificial Intelligence and Statistics (PMLR)* (2013), pp. 207–215.
51. C. Meijer, Über whittakersche bzw. besselsche funktionen und deren produkte. *Nieuw Arch. Voor Wiskunde* **18**, 10–29 (1936).
52. C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, "Understanding deep learning requires rethinking generalization" in *International Conference on Learning Representations, (ICLR)* (2017).
53. D. J. MacKay, Bayesian interpolation. *Neural Comput.* **4**, 415–447 (1992).
54. B. Hanin , Random fully connected neural networks as perturbatively solvable hierarchies. arXiv [Preprint] (2022). http://arxiv.org/abs/2204.01058 (Accessed 14 May 2023).
55. B. Hanin, G. Paouris, Non-asymptotic results for singular values of Gaussian matrix products. *Geom. Funct. Anal.* **31**, 268–324 (2021).
56. P. Diaconis, D. Freedman, "A dozen de finetti-style results in search of a theory" in *Annales de l'IHP Probabilites et statistiques* (1987), vol. 23, pp. 397–423.
57. E. Borel, *Introduction géométrique à quelques théories physiques* (Gauthier-Villars, 1914).