Bayesian Inference with Shaped Deep Non-linear MLPs

Boris Hanin*, Tianze Jiang†

May, 2026

Abstract

Studying the limiting behavior when scaling up model dimensions provides extensive insight into deep learning theory, but mathematically characterizing how the limit of different dimensions jointly control non-linear network fitting and prediction remain largely open. In this work, we study Bayesian inference in deep non-linear MLPs in the regime where the number of training samples (P), the input dimension (N_0), the hidden layer width (N), and the number of hidden layers (L) are simultaneously large. Building on the SDE scaling limit of the layer-wise conjugate kernels, we justify a mathematically rigorous framework to analyze Bayesian inference of neural networks in the proportional $LP \propto N$ limit. We show that Bayesian partition functions of the MLP scaling limit can be equivalently expressed as statistics of the solution of the Neural Covariance SDE [LNR22] and perturbatively solve the Bayesian inference problem for small LP/N. Our framework covers both smooth and ReLU activation functions and applies to arbitrary temperature. With our unifying framework, we rigorously recover several prior results and generate novel technical messages in architecture selection (Bayesian evidence) and feature learning (Bayesian predictive posterior).

*Princeton University. Email: bhanin@princeton.edu.
†Princeton University. Email: tzjiang@princeton.edu

1 Introduction

Rooted in many deep results from approximation to generalization [Zha+17; AS17; Bel+19; Han19; Dau+19], deep Multi-layer Perceptrons (MLPs) are quintessential architectural substrates for analyzing the dynamics of neural networks with a scaling size. As a result, it becomes critical to theoretically understand and rigorously characterize how scaling up different dimensions, such as the number of hidden layers (L), the width of each hidden layer (N), and the size of the dataset (P) jointly shape the behavior of large MLPs at initialization or during training.

Many works studying scaling up neural networks focus on the infinite width limit under a fixed depth and data, giving rise to many classical results such as the Neural Tangent Kernel (NTK) and the Neural Network Gaussian Process (NNGP) [Lee+17; COB19; Du+19; COB19; JGH20; Yan21]. However, the infinite-width limit fundamentally suppresses phenomena in very deep architectures [HN19b; HDR19; HDR22; LNR22] when $L \propto N$ or with a large dataset [LS21; Pac+23] when $P \propto N$. More recent studies [HZ24; Han24; Cam+25; Li+26] also focuses on proportional limits, in which depth, width, and data diverge jointly, revealing new stochastic and geometric structures.

Despite recent progress, how jointly scaling architecture dimensions (L, N, P and input dimension N_0) in non-linear networks impact training-time fitting and test-time prediction is still mysterious. When and why do deep neural networks behave like kernel methods, and what enables feature-learning? When feature learning is possible, what features do MLPs learn? What combinations of data and architecture gets benefits from depth?

In this work, we present the first mathematically rigorous way to answer these questions on non-linear deep MLPs from the Bayesian inference perspective [Mac92; Nea12; HZ23; HZ24; Bas+25]. Bayesian inference analyzes a neural network architecture $x \to f(x;\Theta)$ by choosing a specific prior on the weights $\mathbb{P}(\Theta)$ and computing the posterior $\mathbb{P}(\Theta|\mathcal{D})$ (conditional on the training dataset $\mathcal{D}$) via tilting the prior with the training loss $\mathbb{P}(\mathcal{D}|\Theta) \propto \exp(-\beta\mathcal{L}(\mathcal{D};\Theta))$. The posterior network prediction on test data x_0 is therefore the push-forward of $f(x_0;\Theta)$ on the weights $\Theta \sim \mathbb{P}(\Theta|\mathcal{D})$. The marginal of training data $\mathbb{P}(\mathcal{D})$, equivalently the likelihood of the architecture and prior, can be used for model selection.

Our main contribution in this paper is characterizing and justifying a rigorous unifying framework of analyzing Bayesian inference in non-linear MLPs when $LP \propto N$, building on top of the Neural Covariance SDE (NSDE) limit [LNR22; LN24; Noc+23]. Our method of computing Bayesian statistics (outlined in Section 2 and elaborated in Appendix A) covers a wide range of non-linear activation functions (leaky ReLU in Appendix A.2 and smooth functions in Appendix A.3) and can naturally be applied at any temperature. As examples of applying our framework to answer questions about learning and inference, we compute the perturbative expansion at small effective-depth (Section 4.2 and Section 4.3). Our calculations in the perturbative expansion recover and generalize many results in prior literature for linear networks [HZ23] and weakly non-linear networks with cubic activation function [HZ24]. We further generalize these results and provide novel insights on the perturbative prediction feature map as well as the Bayesian evidence analysis with different activation functions.

1.1 Preliminaries and technical setup

Model We consider MLPs with L fully connected layers $f(x;\Theta): x \in \mathbb{R}^{N_0} \to y \in \mathbb{R}$[1]:

$$z_1 = \frac{1}{\sqrt{N_0}} W_0 x, \quad \phi_\ell = \phi_s(z_\ell), \quad z_{\ell+1} = \sqrt{\frac{c}{N}} W_\ell \phi_\ell, \quad y = z_{\text{out}} = \sqrt{\frac{c}{N}} W_{\text{out}} \phi_L \in \mathbb{R} \tag{1}$$

where $\Theta = (W_0, W_1, \ldots, W_{L-1}, W_{\text{out}})$ is the set of weights, $z_i \in \mathbb{R}^N, i = 1, 2, \ldots, L$ are hidden pre-activations, c is a normalizing constant such that $c^{-1} \triangleq \mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\phi_s(g)^2\right]$ so each hidden neuron is $\mathcal{N}(0,1)$ at init, ϕ_s is the shaped activation function with parameter s (see below), and the interior recursion runs from $\ell = 1, 2, \ldots, L-1$. At initialization (also as the network prior), we assume that all W_i's are initialized with i.i.d. $\mathcal{N}(0,1)$ entries. We wish to characterize the network posterior when $L, P, N \to \infty$ jointly under the squared loss $\mathcal{L}(\mathcal{D};\Theta) \triangleq \frac{1}{2}\sum_{\mu=1}^{P}\left(f\left(x_\mu;\Theta\right) - y_\mu\right)^2$ over a training dataset $\mathcal{D} = \{(x_\mu, y_\mu), \mu = 1, \ldots, P\}$, where $x_\mu \in \mathbb{R}^{N_0}, y_\mu \in \mathbb{R}$ are training data.

Proportional width and depth limit Our results will be described asymptotically in L, N but **not** asymptotically in P, N_0. In other words, we first consider *fixed finite* P, N_0 in the $L, N \to \infty$ limit; then we take P, N_0 to be large. For a fixed P, prior studies focusing on the product of L random $N \times N$ matrices [Han19; HZ23; BLR24; Li+26] indicate that $t = L/N$ controls the initialization behavior in deep linear networks when $L, N \to \infty$, and that both $t = 0$ and $t = \infty$ yield degenerate scaling regimes. While such conclusion does not immediately extend to networks with a non-linear activation function, one can apply a *shaping* mechanism to the non-linear activation function ϕ_s such that the ratio $t = L/N$ still uniquely controls the initialization distribution in non-linear MLPs, as we see next.

Shaped activation function For a growing $L \to \infty$, to ensure that the output distribution at initialization is non-degenerate, various prior works [Mar+21; LNR22; ZBM22; HZ24] have established the necessity of *shaping* the activation function ϕ_s. Specifically, for a smooth base $\phi_{\text{base}}(x)$ normalized to $\phi_{\text{base}}(0) = \phi'_{\text{base}}(0) - 1 = 0$, we consider

$$\phi_s(x) = s\phi_{\text{base}}(x/s) = x + \frac{1}{2s}\phi''_{\text{base}}(0)\cdot x^2 + \frac{1}{6s^2}\phi'''_{\text{base}}(0)\cdot x^3 + \ldots \tag{2}$$

for a shaping factor $s = s(N, L, P)$ that is thought to scale to infinity with N, L, P. Intuitively, the shaping $s > 1$ effectively diminishes the effect of non-linear activation at each layer ($\phi_s(x) \approx x$), such that the cumulative effect becomes controllable over a large number of $L \to \infty$ layers. Similarly, for leaky-ReLU $\phi_{\text{base}}(x) = c_+ \max(x, 0) + c_- \min(x, 0)$ we consider

$$\phi_s(x) = x + \frac{1}{s}\left(c_+ \max(x, 0) + c_- \min(x, 0)\right). \tag{3}$$

We also refer to [LN24] for a rigorous discussion on why shaping is necessary and the different types of degenerate limits with mis-specified s. In short, there usually exists a critical $s^\star(N, L, P)$ such that shaping with $s \in \omega(s^\star)$ results in effective equivalence to a

[1] While our analysis easily extends to $y \in \mathbb{R}^d$ output for $d \in \Theta(1)$, we present scalar output results for simplicity.

deep linear network ($\phi_s = \mathrm{id}$), and shaping with $s \in o(s^\star)$ results in a L/N-independent degenerate limit. Intuitively, the shaping mechanism ensures that the (cumulative) effect of applying a non-linear activation function at each layer is on the same scale as the effect of linear layers.

Bayesian inference The Bayesian inference [Mac92; Nea94] point of view for studying (1) is based on a prior distribution over the trainable weights and studying the posterior measure over the weights defined by:

$$\mathbb{P}_{\text{post}}(\Theta|\mathcal{D}) \propto \mathbb{P}_{\text{prior}}(\Theta) \exp(-\beta \mathcal{L}(\mathcal{D}; \Theta))$$

where β is the inverse temperature. The characteristic function of the predictive posterior on a test data $x_0 \in \mathbb{R}^{N_0}$ is therefore:[2]

$$\mathbb{E}_{\text{post},\beta}[\exp[-i\kappa f(x_0; \Theta)]] = \frac{Z_\beta(x_0, \kappa)}{Z_\beta(0)} \tag{4}$$

where Z_β is the *partition function*:

$$Z_\beta(x_0, \kappa) \triangleq (2\pi\beta)^{P/2} \cdot \mathbb{E}_{\text{prior}}\left[\exp\left[-\beta \mathcal{L}(\mathcal{D}; \Theta) - i\kappa f(x_0; \Theta)\right]\right], \quad Z_\beta(0) = Z_\beta(x_0; 0) \tag{5}$$

and the marginal on the data, equivalently the likelihood of the model and prior, is

$$\mathbb{P}_{\text{prior}}(\mathcal{D}) = \mathbb{E}_{\text{prior}}\left[\exp\left[-\beta \mathcal{L}(\mathcal{D}; \Theta)\right]\right] \propto Z_\beta(0)$$

and also known as the *Bayesian evidence*, maximizing which has been the objective to many results on architecture selection. We refer to [Mac92] for a more detailed preliminary on Bayesian inference.

Conjugate kernel under the prior Because the partition function Z in (5) concerns only the prior distribution of f, it is sufficient to study the distribution of $f(\mathcal{D} \cup \{x_0\}; \Theta) \in \mathbb{R}^{P+1}$ in (1) at initialization. An important observation of (1) at initialization is that, for each ℓ, conditioned on all prior weights $W_{[0:\ell-1]}$, the next layer $z_{\ell+1} \in \mathbb{R}^N$ is N i.i.d. Gaussian random variables with mean zero. Over a finite set of inputs, the covariance matrix between individual neurons from different $z_{\ell+1}(x)$'s follows the *Conjugate Kernel* $\Phi^{(\ell)}$ [CS09; El 10; PW17; FW20; LNR22; BP22; Cho23; WWF24] defined (on a pair of inputs x_α, x_β) as:

$$\Phi^{(\ell)}_{\alpha\beta} \triangleq \mathrm{cov}\left([z_{\ell+1}(x_\alpha)]_1, [z_{\ell+1}(x_\beta)]_1\right) = \frac{c}{N}\langle \phi_\ell(x_\alpha), \phi_\ell(x_\beta)\rangle. \tag{6}$$

Furthermore, because weights at each layer are sampled independently, the following is a Markov chain for the forward pass:

$$X \to \Phi^{(0)} \to z_1 \to \phi_1 \to \Phi^{(1)} \to z_2 \to \phi_2 \to \Phi^{(2)} \dots \to \Phi^{(L)} \to y$$

[2] While our computations seamlessly carry to characterizing the predictive posterior of multiple test data points jointly, we only derive the case of predicting one posterior for presentation here.

where the transition is only random at the $\Phi^{(\ell)} \to z_{\ell+1}$ steps. In other words, it suffices to study the discrete Markov chain

$$\frac{1}{N_0} X^\top X = \Phi^{(0)} \to \Phi^{(1)} \to \cdots \to \Phi^{(L)} \tag{7}$$

with the final output of the model $f(\mathcal{D} \cup \{x_0\}; \Theta) | W_{[0:L-1]} \sim \mathcal{N}(0, \Phi^{(L)})$. As a result, studying Bayesian inference reduces to studying the distributional properties (expectation over test functions) of conjugate kernel $\Phi^{(L)}$ under the prior with random hidden weights.

2 Main idea: Bayesian inference via Neural Covariance SDEs

To formalize the connection between the conjugate kernel and Bayesian inference, a crucial argument we use in this work (see Lemma 4) is as follows: when $f(x) = W^\top h(x)$ for some output projection weight $W \in \mathbb{R}^{N \times 1}$ whose prior is i.i.d. Gaussian $\mathcal{N}(0, \mathbb{I}_N)$ and $\mathcal{L}$ is the squared loss, the partition function $Z_\beta(x_0, \kappa)$ in (5) has another equivalent form as an integral:

$$Z_\beta(x_0, \kappa) = \int_{\mathbb{R}^P} \exp\left[-\frac{1}{2\beta}\|p\|^2 + ip^\top Y\right] \cdot \mathbb{E}_{\text{prior}}\left[\exp\left(-\frac{1}{2} v^\top \Phi v\right)\right] \, \mathrm{d}p \tag{8}$$

where $v = [p^\top, \kappa]^\top \in \mathbb{R}^{P+1}$ and $\Phi = [h(X), h(x_0)]^\top [h(X), h(x_0)] \in \mathbb{R}^{(P+1)\times(P+1)}$ is the matrix version of our conjugate kernel (6) at the penultimate layer. The formulation (8) also allows us to take $\lim_{\beta\to\infty} Z_\beta$. For other loss functions, one can also write the respective expressions of Z with only $\mathbb{E}_{\text{prior}}$ of statistics of Φ. As an example, the Binary Cross-Entropy (BCE) loss gives equivalent partition function

$$Z_\beta^{\text{BCE}}(x_0, \kappa) = 2^{-\beta P} \int_{\mathbb{R}_+^P} \nu_\beta^{\text{PG}}(\mathrm{d}\omega) \int_{\mathbb{R}^P} \frac{\exp(-\frac{1}{2} p^\top \Omega^{-1} p)}{(2\pi)^{P/2}\sqrt{\det \Omega}} \cdot \mathbb{E}_{\text{prior}}\left[\exp\left(-\frac{1}{2} v^\top \Phi v\right)\right] \, \mathrm{d}p$$

where $\Omega = \operatorname{diag}(\omega_1, \ldots, \omega_P)$, $v = \left[p + i\beta\left(Y - \frac{1}{2}\mathbb{1}\right), \kappa\right]^\top$, and ν_β^{PG} is the product Polya-Gamma law with marginals $\omega_\mu \sim \text{PG}(\beta, 0)$ [PSW13]. In this paper, we focus specifically on the squared error because of the simplicity of (8) (specifically the integration form), but in principle our techniques to compute $\mathbb{E}_{\text{prior}}[e^{-v^\top \Phi v/2}]$ provide solutions to cross-entropy and other loss functions as well.

Having established that (8) only depends on the prior through $\mathbb{E}_{\text{prior}}\left[\exp\left(-\frac{1}{2} v^\top \Phi v\right)\right]$, it thus suffices to compute this expectation for each p. In the simplest case where the network prior to Φ is deterministically (or at least approximately) equal to a kernel $\bar{\Phi}$ with an associated *feature map* $f : x_\alpha \to \bar{x}_\alpha$ such that $\bar{\Phi}_{\alpha\beta} = \langle \bar{x}_\alpha, \bar{x}_\beta \rangle$, the expectation in (8) is moot and Bayesian inference reduces to what we refer to as the *kernel method* where the partition function $Z_\beta^{\text{kernel}}(x_0, \kappa; f) \triangleq \int_{\mathbb{R}^P} \exp\left[-\frac{1}{2\beta}\|p\|^2 + ip^\top Y - \frac{1}{2} v^\top \bar{\Phi} v\right] \mathrm{d}p$ can be computed directly:

$$Z_\beta^{\text{kernel}} = (2\pi)^{\frac{P}{2}} \exp\left[-\frac{1}{2}(Y^\top A^{-1} Y + \log\det A) - i\kappa \bar{\Phi}_{\kappa p} A^{-1} Y - \frac{1}{2}\kappa^2(\bar{\Phi}_{\kappa\kappa} - \bar{\Phi}_{\kappa p} A^{-1} \bar{\Phi}_{p\kappa})\right] \tag{9}$$

where $\bar{\Phi} = \begin{bmatrix} \bar{\Phi}_{pp} & \bar{\Phi}_{p\kappa} \\ \bar{\Phi}_{\kappa p} & \bar{\Phi}_{\kappa\kappa} \end{bmatrix}$ and $A = \frac{1}{\beta}\mathbb{I}_P + \bar{\Phi}_{pp}$. This induces the predictive posterior (4) to be an exact *Gaussian* random variable with $\mathbb{E}_{\text{post}}[f(x_0)] = \bar{\Phi}_{\kappa p}A^{-1}Y$ and $\text{var}_{\text{post}}(f(x_0)) = \bar{\Phi}_{\kappa\kappa} - \bar{\Phi}_{\kappa p}A^{-1}\bar{\Phi}_{p\kappa}$. Intuitively, the predictive posterior mean weighs the training labels according to alignment between test and train in the feature space, and the variance measures the distance between test towards the orthogonal projection into train in the feature space.

In general, it is often not the case that the prior distribution of Φ concentrates on a single $\bar{\Phi}$. Without concentration, it is also not always true that the posterior prediction is a Gaussian (equivalently, $\log Z$ is not a quadratic polynomial of κ), hence the non-triviality of dealing with the expectation in (8). For a finite P, we are interested in the proportional $L/N = t^\star > 0$ limit, where there exist an important tool for analyzing the evolution (7): the Neural Covariance SDE (NSDE) [LNR22; LN24], which (informally) states the following:

Proposition 1 (Neural Covariance SDE, informally)**.** *Fixing a terminal $t^\star > 0$ and scale $N \to \infty, L = \lfloor t^\star N \rfloor$. For shaped smooth activations (2) and leaky ReLU (3), when taken $s = \alpha_{\text{act}}^{-1/2}\sqrt{N}$ for some $\alpha_{\text{act}} > 0$, the interpolated covariance process $t \to \Phi^{(\lfloor tN \rfloor)}$ converges to a diffusion process $\Phi_t \in \mathbb{R}^{P\times P}$ solving*[3]

$$\mathrm{d}\Phi_t = \alpha_{\text{act}} \cdot b(\Phi_t)\ \mathrm{d}t + \Phi_t^{1/2}\ \mathrm{d}B_t\Phi^{1/2}, t \in [0, t^\star], \quad \Phi_0 = \frac{1}{N_0}X^\top X. \tag{10}$$

Here $B_t = \frac{1}{\sqrt{2}}\left(\widetilde{B}_t + \widetilde{B}^\top\right) \in \mathbb{R}^{P\times P}$ where $\widetilde{B}_t \in \mathbb{R}^{P\times P}$ has i.i.d. Brownian motion entries. The drift components $b(\Phi)$ depends only on the base non-linearity ϕ_{base}.

We formalize Proposition 1 in Appendix A. Recall that our goal is to evaluate the expectation under prior $\mathbb{E}\left[\exp\left(-\frac{1}{2}v^\top\Phi^{(L)}v\right)\right]$. To do that, simply plugging in (10) for $s_t(v) \triangleq v^\top\Phi_t v \in \mathbb{R}$ yields

$$\mathrm{d}s_t(v) = v^\top\Phi_t^{1/2}\ \mathrm{d}B_t\Phi_t^{1/2}v + \alpha_{\text{act}}v^\top b(\Phi_t)v\ \mathrm{d}t = \sqrt{2}s_t(v)\ \mathrm{d}W_t + \alpha_{\text{act}}v^\top b(\Phi_t)v\ \mathrm{d}t$$

where $(W_t)_{t\geq 0}$ is a standard Brownian Motion on $\mathbb{R}$. The SDE on $s_t(v)$ thus (in theory) contains all the necessary information to solve $\mathbb{E}_{\text{prior}}[\exp(-s_t(v)/2)]$, integrating which in (8) will then lead to a closed form solution to the Bayesian partition function.

Although our treatments for P is non-asymptotic in nature, it is still convenient to balance scale preliminarily. Consider a deep linear network, in which (10) has $\alpha_{\text{act}} = 0$ or equivalently $s = \infty$, we have the linear SDE $\mathrm{d}\Phi_t = \Phi_t^{1/2}\ \mathrm{d}B_t\Phi^{1/2}$, the RHS of which has operator norm (in fact the entire spectrum) $\left\|\Phi_t^{1/2}\ \mathrm{d}B_t\Phi^{1/2}\right\|_{op} \asymp \sqrt{P\ \mathrm{d}t}\|\Phi_t\|_{op}$. As a result, for large P, we will make a time change $\mathrm{d}\tau \triangleq P\ \mathrm{d}t$ such that (10) becomes (denote $c_{\text{act}} \triangleq \alpha_{\text{act}}P^{-1}$):

$$\mathrm{d}\Phi_\tau = c_{\text{act}} \cdot b(\Phi_\tau)\ \mathrm{d}\tau + \sqrt{\frac{1}{P}}\Phi_\tau^{1/2}\ \mathrm{d}B_\tau\Phi_\tau^{1/2}, \quad \tau \in [0, \tau^\star \triangleq Pt^\star = LP/N]. \tag{11}$$

and we have

$$\mathrm{d}s_\tau(v) = \sqrt{2/P} \cdot s_\tau(v)\ \mathrm{d}W_\tau + c_{\text{act}} \cdot v^\top b(\Phi_\tau)v\ \mathrm{d}\tau, \quad s_0(v) = v^\top\Phi_0 v. \tag{12}$$

[3]The exact form in [LNR22] is different but algebraically equivalent to ours, see Lemma 3.

We will refer to $\tau \in [0, LP/N]$ as the *depth-time*, which solely controls the Φ dynamics (normalized for large P), and $\tau^\star$ to be the terminal time. The resulting $\frac{LP}{N} \in \Theta(1)$ regime has also been the central pursuit of characterizing the proportional $N, L, P \to \infty$ limits in many related studies [HZ23; HZ24; Li+26].

Our results will be parameterized by a $\tau^\star > 0$ and in the limit of

$$\lim_{P\to\infty} \left(N, L \to \infty \text{ with } L/N = \tau^\star P^{-1}\right). \tag{13}$$

Methodologically, we first fix a terminal time $\tau^\star > 0$. For any finite P, Bayesian statistics in the asymptotic limit with $L, N \to \infty$ proportionally such that $L/N = t^\star = \tau^\star P^{-1}$ can be computed as a function of P and $\tau^\star$. Finally, we normalize the result so that the limit $P \to \infty$ is well-defined for fixed $\tau^\star$. Our calculation of the Bayesian partition function $Z^{(N,L,\mathcal{D})}$ with (1) will be based on the following result to $Z^{(\tau)}$ from solving the NSDE, which we justify in Appendix C (see also Section 4.4).

Theorem 1 (Pointwise convergence of MLP to SDE partition function)**.** *Fix any $t^\star > 0, x_0 \in \mathbb{R}^{N_0}, \kappa \in \mathbb{R}$ and finite $\beta > 0$. In the width and depth limit $N, L \to \infty, L/N = t^\star$, $\tau^\star \triangleq Pt^\star$ with a fixed P, Bayesian inference with MLP (1) from $\mathbb{E}_{\text{post},\beta}\left[e^{-i\kappa f(x_0;\Theta)}\right] \propto Z_\beta^{(N,L,\mathcal{D})}(x_0, \kappa)$ has that:*

$$\begin{aligned}\lim_{N,L\to\infty} Z_\beta^{(N,L,\mathcal{D})}(x_0,\kappa) &= Z_\beta^{(\tau^\star)}(x_0,\kappa) \\ &= \int_{\mathbb{R}^P} \exp\left[-\frac{1}{2\beta}\|p\|^2 + ip^\top Y\right] \mathbb{E}_{(12)}\left[\exp\left(-\frac{1}{2}s_{\tau^\star}(v)\right)\right] \mathrm{d}p \end{aligned} \tag{14}$$

where $v = [p^\top, \kappa]^\top$ and $s_{Pt^\star}$ is the solution to the SDEs (11) and (12).

From Theorem 1, one can also easily proceed to take $P \to \infty$ at (14), which justifies dropping $O(P^{-1})$-terms at the RHS via the Lévy's continuity theorem. Due to the analytic form of $Z^{(\tau)}$, in the majority of the following results when not explicitly stated otherwise, we will analyze $Z = Z^{(\tau)}$ the SDE partition function, as opposed to the actual MLP partition function $\lim Z^{(N,L,\mathcal{D})}$ (see Section 4.4).

In general, while solving (12) may be intractable for general $\tau^\star > 0$, the small τ expansion (leading partial derivatives at $\tau = 0$) of $\mathbb{E}[e^{-s/2}]$ is still tractable in closed form. Furthermore, the system can be explicitly closed in special cases, such as when $\alpha_{\text{act}} = 0$ (diffusion-only) and when $\tau^\star = c_{\text{act}}^{-1} \to 0$ (equivalently drift-only), corresponding to different MLP scaling regimes. Our main technical contribution in this paper is thus to extract insights from partially solving (11) and (12) and derive corollaries towards Bayesian inference in MLPs.

Assumptions Our most important assumption is that $P < N_0$ and that the initial $\Phi_0 = \frac{1}{N_0} X^\top X \in \mathbb{R}^{P\times P} \succ 0$ is full rank. Although our calculation of $Z^{(Pt)}$ does not require assumptions on the base activation function beyond those in the NSDE (see Appendix A), the final step of taking $P \to \infty$ and extracting qualitative results requires more assumptions in the dataset $\mathcal{D}$. Specifically, we assume that $\max_\alpha \|X_\alpha\|^2, \min_\alpha \|X_\alpha\|^2, \|\Phi_0\|_{op}, \|\Phi_0^{-1}\|_{op} \in \Theta(1)$ which is true in Marchenko-Pastur-type Gram matrices (when $\lim P/N_0 = \lambda < 1$)

and that $Y \in \mathbb{R}^P$ has $\Theta(1)$-entries. When these large-P assumptions are not met (e.g. if $\|\Phi_0\|_{op} \asymp P$), our finite P partition function calculations are not affected, but the $P \to \infty$ limit will render different results (i.e. the set of dominating vs negligible terms may be different).

3 In relation to prior literature

Bayesian inference A central theme in modern studies of large networks is that neural networks at initialization can be studied in *function space* rather than *parameter space*. In the Bayesian view, randomness in the weights induces a prior over functions. Earlier work [Nea94; Wil96; Nea12] showed that for a broad class of priors, the single-hidden-layer network prior converges (as the width scales) to a Gaussian process, and that the resulting Gaussian process perspective can be leveraged for principled uncertainty quantification and inference. These ideas were later extended to deep fully connected networks by taking $L \to \infty$ after $N \to \infty$ first: iterating the layer-wise covariance map yields a deterministic kernel recursion at fixed depth, giving rise to the Neural Network Gaussian Process (NNGP) formalism [Lee+17; Mat+18; Tre23]. In this regime, hidden feature Gram matrix (i.e. our conjugate kernel Φ) concentrate in the large width limit, enabling a clean characterization of "typical" network behavior at initialization.

Outside of the strictly infinite-width setting, however, exact Bayesian inference is generally intractable, motivating a large body of approximation methods that aim to preserve calibrated uncertainty while remaining computationally feasible [Blu+15; GG16]. Parallel to these algorithmic developments, recent theory has begun to characterize Bayesian posteriors in *scaling limits* where computations become analyzable. For deep linear networks, [HZ23] first derives exact non-asymptotic expressions for the Bayesian evidence and predictive posterior and highlights the resulting *effective posterior depth* LP/N. This is extended by [Bas+25] which studies deep linear networks with convolution layers. In the resulting $LP/N \in \Theta(1)$ joint scaling limit with shaped nonlinear networks, [HZ24] develops perturbative first-order LP/N-expansions around the infinite-width $N \gg LP$ baseline. The results in [HZ24] (many of which we will recover and extend, see Section 4) are closest in spirit to ours. However, their calculations are conducted at a physics level of rigor and rely on many simplifications of the data and the asymptotic limit. Our framework, on the other hand, is both mathematically rigorous, conceptually simple, and easier to generalize.

Infinite width and depth scaling Taking depth large introduces new phenomena even before considering training: stability of signal propagation, gradient explosion/vanishing, and emergent transitions between ordered and chaotic behavior. Early dynamical mean-field analyses of large random networks identified sharp transitions in typical dynamics and provided a conceptual template for later "edge-of-chaos" perspectives [SCS88]. In deep *feed-forward* settings, mean-field methods quantify how correlations and norms propagate across layers, and connect expressivity to transient chaotic amplification mechanisms [Poo+16]. Closely related work on "deep information propagation" formalizes how correlation maps and their fixed points govern trainability and motivates critical initialization schemes that keep signal propagation non-degenerate over many layers [Sch+17; PW17].

Complementing the mean-field approaches, various studies on product of many large random matrices at network initialization also give further insights to deep network initialization and conditioning [HN19a; HN19b; HP21; Li+26; HJ25]. Within these, an important direction is to study the *joint* limits in which width and depth grow together ($N, L \to \infty$ concurrently), a regime where deterministic infinite-width recursions are no longer adequate because finite-width fluctuations can accumulate over a large number of layers. For the forward pass of a single input token in the proportional $N \propto L$ limit, [HN19b] demonstrated that in the linear networks, accumulated feature fluctuations lead to log-normal distributions in hidden layers and gradients, revealing a rich stochastic structure absent in the deterministic kernel limit. [LNR22] furthers this discovery to the flexible framework of NSDEs that incorporates non-linear activation functions, which are later extended by [Noc+23; LN24] to ResNets. This joint-scaling perspective is complementary to fixed-depth infinite-width neural matrix law frameworks that characterize broad classes of architectures at initialization [Yan21]. In the present work, these simultaneous width and depth scalings are particularly natural because they yield stochastic limiting objects for Gram matrices [LNR22], allowing spectral observables and Bayesian quantities to be expressed as functionals of the terminal covariance state.

4 Overview of technical takeaways

In this section, we go into details of our exact technical results following expressing the partition function with Theorem 1. Each `Takeaway` is meant to be an informal statement which we expand in the appendix. They include both the recovery and extension of prior results derived with different methods as special cases of our framework, as well as novel results demonstrating our capability of solving previously open regimes. For example, while the positive temperature case $\beta < \infty$ was somewhat difficult to handle in the earlier frameworks of [HZ23; HZ24], a non-trivial β fits easily into our calculations of (8) with (12). Finally, we provide insights extracted from Bayesian inference computations in different cases to answer questions regarding model architecture selection and feature learning.

4.1 Recovering and extending prior results in Bayesian inference

Our first set of results will come from special cases in (11). Specifically, we consider when either the diffusion term dominates the SDE (equivalently $\alpha = 0, s \to \infty$), which yields the deep linear network, or when the drift term dominates the SDE (equivalently $c_{\text{act}} = \tau^{-1} \to \infty, \tau \to 0$), which yields the infinite-width $N \gg LP$ limit.

Linear networks and diffusion-only The simplest starting point in analyzing the network (1) is when $\phi_s = \text{id}$ the identity map. In this case f is simply a deep linear network one only needs the limiting product of many large random matrices. Here, the Bayesian partition function $Z_\beta(x_0, \kappa)$ has a mathematically closed solution in the form of Meijer-G integrals [HZ23] at the $N, L, P \to \infty, LP/N \in \Theta(1)$ limit. Below in `Takeaway 1.1`, we recover two main results in [HZ23].

- `Takeaway 1.1`: **Deep linear network**: The Bayesian partition function in the linear

network $Z_\beta^{\text{linear}}(x_0, \kappa)$ has the explicit integration form (for any τ):

$$Z_\beta^{\text{linear}(\tau)} \propto \int_{\mathbb{R}^P} \mathbb{E}\left[\exp\left(-\frac{1}{2\beta}\|p\|^2 + ip^\top Y - \frac{1}{2}\exp\left(-\frac{\tau}{P} + \sqrt{\frac{2\tau}{P}}G\right) v^\top \Phi_0 v\right)\right] \, \mathrm{d}p$$

where expectation is over $G \sim \mathcal{N}(0,1)$ and $v = [p^\top, \kappa]^\top$. As an exemplary corollary, the first-order expansion at $\tau = 0$ with $\beta = \infty$ is:

$$\left.\frac{\partial \log Z_\infty^{\text{linear}(\tau)}(\cdot)}{\partial \tau}\right|_{\tau=0} = \frac{1 + O(P^{-1})}{4P}\left[\left(P(1-\nu_0) + \kappa^2\|x_0^\perp\|^2\right)^2 + 2P(1-2\nu_0)\right]$$

where $\nu_0 = P^{-1}Y^\top \Phi_0^{-1} Y$ is the linear interpolator norm of $\mathcal{D}$, and $x_0^\perp \perp \text{span}(X_\mathcal{D})$ is the orthogonal component projecting test x_0 to the span of the training inputs $X_\mathcal{D}$.

We will show `Takeaway 1.1` with our NSDE framework in Appendix A.4.

Infinite-width kernel limit While non-linear networks are remarkably more complicated than their linear counterparts, the infinite-width $\tau_{\text{terminal}} = LP/N \to 0$ limit, which can be shown to effectively kills the diffusion terms by a time-change and leaves an equivalent **Neural Covariance ODE**, is solvable. In this limit, without the activation function, the conjugate kernel of linear networks is trivially $\Phi_L = \Phi_0$ with the identity feature map. This trivial reduction is also true if $s \in \Theta(N^{1/2})$ by Proposition 1. Alternatively, we show that a stronger shaping $s \in \Theta(L^{1/2})$ (which is equivalent to $s \in \Theta(N^{1/2})$ with positive $\lim L/N > 0$) induces a non-trivial deterministic kernel method when $\lim L/N = 0$. This fact was established in [HZ24] with cubic activation $\phi_{\text{base}} = x + \psi x^3, \phi_s(x) = x + \frac{\psi}{L}x^3$, which we will extend to general activation functions. Since an ODE guaranties a deterministic answer, Bayesian inference reduces to a kernel method, which gives a deterministic yet training-data agnostic feature map (that explicitly only depends on c, ϕ_{base}).

- `Takeaway 1.2`: **Data agnostic feature map in the infinite width limit**: In the large width $N \gg LP$ limit, Bayesian inference (at any temperature) for MLP (1) with shaping (3) and (2) where $s = (c_{\text{act}}/L)^{-1/2}$, reduces to a data-agnostic kernel method (9) with a feature map $f_0(x; c_{\text{act}}, \phi_{\text{base}}) \in \mathbb{R}^\infty$. In the zero temperature case, the predictive posterior at test x_0 is a Gaussian with mean and variance

$$\mu_0(x_0) = \sum_{\mu \in \mathcal{D}} a_\mu Y_\mu, \quad \sigma_0^2(x_0) = \|f_0(x_0)^\perp\|^2 \tag{15}$$

  where $f_0(x_0) = f_0(x_0)^\perp + \sum_{\mu\in\mathcal{D}} a_\mu f_0(x_\mu)$ and $f_0(x_0)^\perp$ is orthogonal to $\text{span}_\mu(f_0(x_\mu))$.

We justify this statement precisely in Appendix A.5. It will also be apparent that $c_{\text{act}} \in o(1)$ implies a trivial feature map (equivalent to the linear network), and $c_{\text{act}} = \infty$ admits a unique (although possibly divergent) limit for the kernel, thus the criticality $s = \Theta(L^{1/2})$ at the infinite-width limit. Since we mostly assume a fixed ϕ_{base}, we will also use $f_0(x; c_{\text{act}})$ to short-hand the feature map and discuss the role of c_{act} later.

4.2 Small depth-time perturbation

Now we turn to the more general case where neither the infinite width non-linearity induced kernel nor the linear network effect dominates the partition function. Firstly, we need to decide the shaping factor of ϕ_s, such that non-linearity effects contribute at the same order as linear networks. Motivated by the infinite-width limit as well as (11), we consider the following shaping:

$$s_{\text{shape}} = \sqrt{c_{\text{act}}^{-1} N/P} = \sqrt{c_{\text{act}}^{-1} L/\tau} \tag{16}$$

with $c_{\text{act}} \in \Theta(1)$. Our main technical result concerns perturbatively solving (5) under (16) in the first order of $\tau = LP/N$. Noticing that $Z^{(0)}$ is the kernel method (9) with input kernel $\bar{\Phi} = \Phi_0$ regardless of model, we can thus compare $\frac{1}{Z^{(0)}}\,\partial_\tau Z|_{\tau=0} = \partial_\tau \log Z|_{\tau=0}$.

- `Takeaway 2`: **Perturbative expansion of $Z^{(\tau)}$.** The 1st order partial derivative of small $\tau = \frac{LP}{N}$ in the joint limit (13) with shaping according to (16):

$$\partial_\tau \log Z_\beta^{(\tau)}(\cdot)\Big|_{\tau=0} = \partial_\tau \log Z_\beta^{\text{linear}}(\cdot)\Big|_{\tau=0} + c_{\text{act}}\, \partial_c \log Z_\beta^{\text{kernel}}(\cdot; f_0(\cdot; c))\Big|_{c=0} \tag{17}$$

  is the *sum* of first order derivative in the *linear network component* in `Takeaway 1.1` and *infinite-width kernel component* with partition function (9) in `Takeaway 1.2`.

Balancing the two terms in `Takeaway 2` immediately implies the critical $c_{\text{act}} \in \Theta(1)$ in (16) is indeed the appropriate scaling. We show (17), as well as the following set of corollary takeaways in Appendix B.

Perturbative Bayesian evidence Eqn. (17) is already a strong statement that allows us to draw interesting corollaries. For instance, this implies that first order Bayes evidence $\mathbb{P}_{\text{prior}}(\mathcal{D}) \propto Z_\beta(0)$ in (5) is just the sum of the linear network evidence and the infinite-width kernel evidence. See Appendix B.1 for an elaborate account.

- `Takeaway 2.1`: **Bayesian evidence at small τ.** In the first order of small τ expansion in the joint limit (13), the Bayesian evidence is given by (at $\tau = 0$ and up to $O(P^{-1})$ remaining terms):

$$\frac{\partial \log Z_\beta^{(\tau)}(0)}{P\partial\tau} = \frac{1}{4}\left(\nu_0^\beta - \frac{1}{P}\operatorname{Tr}\left(\Phi_0(\Phi_0 + \beta^{-1}\mathbb{I})^{-1}\right)\right)^2 + \frac{c_{\text{act}}}{P}\frac{\partial \log Z_\beta^{\text{kernel}}(0; f_0)}{\partial c} \tag{18}$$

  where $\nu_0^\beta = P^{-1}Y^\top(\Phi_0 + \beta^{-1}\mathbb{I})^{-1}\Phi_0(\Phi_0 + \beta^{-1}\mathbb{I})^{-1}Y$ is the (normalized) RKHS norm of the Kernel Ridge Regression estimator. We normalize both sides by $1/P$ because the log marginal of data is $\log \mathbb{P}(\mathcal{D}) \propto P$. When $c_{\text{act}} = 0$, the remaining first term is equivalent to that of the linear network in `Takeaway 1.1`.

It is notable that while the linear network evidence can be shown to always increase with depth, such is not true with the infinite-width kernel in `Takeaway 1.2`. As a result, depending on the exact scale of c_{act}, we may encounter phase transitions in the form of "depth is only beneficial if non-linearity is not too strong". We list some specific corollaries of this flavor with concrete $\mathcal{D}$ in Section 4.3.

Adaptive prediction kernel at $\tau > 0$ Another corollary of (17) is that we can derive an equivalent feature map for the predictive posterior in the zero temperature regime. We show this in Appendix B.2.

- `Takeaway 2.2`: **Equivalent data-dependent feature map at small τ.** In the first order of small τ expansion in the joint limit (13) at $\beta = \infty$, the Bayesian predictive posterior of at test data x_0 is (approximately) equivalently given by kernel method with the training-data dependent feature map:

$$x_\alpha \to \sum_{x_\mu \in \mathcal{D}} a_\mu f_0(x_\mu; c_{\text{act}}\tau) + \left(1 + \frac{1}{2}(\nu_0 - 1)\tau\right) f_0^\perp(x_\alpha; c_{\text{act}}\tau) \tag{19}$$

  where a_μ follows the definition in (15). The predictive mean remain un-changed from (15), whereas the variance is reshaped as a function of the training dataset.

By approximately given, we mean that we drop $O(P^{-1}) + O(\tau^2)$ terms in the partition function Z. An important remark is that while the predictive posterior is equivalently given by a kernel method with an explicit feature map, the full Bayesian partition function is *not* given by the same kernel method (in contrast to `Takeaway 1.2`). Such equivalence is only possible by the fact that the leading order derivative of the partition function $\partial_\tau \mathbb{E}_{\text{post}}[e^{-i\kappa f(x_0)}] = \partial_\tau \left[\log Z(x_0, \kappa) - \log Z(x_0, 0)\right]$ is a *quadratic* polynomial of κ (after dropping $o(1)$ terms) at $\tau = 0$ (Proposition 9).

4.3 Interpretations of perturbative results

Given the above results, it is natural to ask: what do our technical takeaways imply in terms of model selection and prediction? In particular, what data distribution does deeper network or non-linearity prefer or penalize, as evaluated by the Bayesian evidence $\partial_\tau \mathbb{P}(\mathcal{D}) \geq 0$?

We make some concrete interpretations of our findings when $\Phi_0 = \rho\mathbb{I} + o(1)$, i.e., when the input data have a fixed norm $\rho > 0$ and are (approximately) orthogonal to each other. Such conditions typically arise when input data are from isotropic Gaussians, or when batch norm is applied. In the proportional joint limit with a small terminal LP/N, we justify two novel findings from our Bayes calculations in the form of when data favors depth or non-linear activation strength in the model likelihood. We show these results in Appendix B.1.

- `Takeaway 3.1`: **Smooth activation.** Under normalized inputs, for deep MLPs with c_{act}-shaped smooth activation (2) and (16), depth is beneficial (at first order LP/N) to fitting if and only if:

$$\begin{aligned} 0 \leq & \frac{1}{4}\left(\frac{\rho\|y\|^2}{(\rho+\beta^{-1})^2} - \frac{\rho}{\rho+\beta^{-1}}\right)^2 + c_{\text{act}}\left\{\frac{\rho\left(2\left(c_1+c_2\right)\rho - \left(3c_1+2c_2\right)\right)}{2(\rho+\beta^{-1})^2}\|y\|^2\right. \\ & \left. + \frac{c_1\rho^2}{2(\rho+\beta^{-1})^2}\|\widetilde{y}\|^2 - \frac{\left(3c_1+2c_2\right)\rho(\rho-1)}{2(\rho+\beta^{-1})}\right\} \end{aligned}$$

  where $c_1 = \frac{1}{4}(\phi''_{\text{base}}(0))^2, c_2 = \frac{1}{2}\phi'''_{\text{base}}(0)$, and $\|y\|^2 = \frac{1}{P}\sum_{\mathcal{D}} Y_\alpha^2, \|\widetilde{y}\|^2 = \frac{1}{P}(\sum_{\mathcal{D}} Y_\alpha)^2$.

- `Takeaway 3.2:` **ReLU activation.** For deep MLPs with c_{act}-shaped ReLU activation (3) and (16), depth is beneficial if and only if:

$$0 \leq \frac{1}{4}\left(\frac{\rho\|y\|^2}{(\rho+\beta^{-1})^2} - \frac{\rho}{\rho+\beta^{-1}}\right)^2 + c_{\text{act}} \cdot \frac{(c_+ - c_-)^2\rho}{4\pi(\rho+\beta^{-1})^2}\left(\|\widetilde{y}\|^2 - \|y\|^2\right)$$

where $\|y\|^2 = \frac{1}{P}\sum_{\mathcal{D}} Y_\alpha^2$, $\|\widetilde{y}\|^2 = \frac{1}{P}(\sum_{\mathcal{D}} Y_\alpha)^2$.

4.4 Perturbative convergence from MLP to SDE

Because our above takeaways use the SDE partition functions, not statistics from the actual MLP, it is thus necessary to explain how they are mathematically rigorous and what types of limits are necessary to translate SDE takeaways to MLPs. Theorem 1 gives us point-wise convergence (for each τ) along the forward passage, but perturbative takeaways do not immediately follow. While taking the perturbative limit in discrete Markov chains is not well-defined, we can show a stronger convergence from MLP layer-wise to SDE, which we will justify (alongside with the proof of Theorem 1) in Appendix C.

- `Takeaway 4:` **Uniform convergence of generator.** Fix $v \in \mathbb{R}^{P+1}$ and time $T > 0$ and let $F(\Phi) \triangleq e^{-\frac{1}{2}v^\top \Phi v}$. Define the rescaled discrete-time difference per layer in (1)

$$D_N F(t, \Phi_0) \triangleq N\left(\mathbb{E}_{\text{prior}}\left[F\left(\Phi^{(\lfloor tN\rfloor+1)}\right)\right] - \mathbb{E}_{\text{prior}}\left[F\left(\Phi^{(\lfloor tN\rfloor)}\right)\right]\right),$$

and the limit (taken from applying Ito's Lemma on $F(\Phi_t)$)

$$DF(t, \Phi_0) \triangleq \mathbb{E}_{(10)}\left[\mathcal{L}F(\Phi_t)\right] = \mathbb{E}_{(10)}\left[F(\Phi_t)\left(-\frac{1}{2}\alpha_{\text{act}} v^\top b(\Phi_t) v + \frac{1}{4}\left(v^\top \Phi_t v\right)^2\right)\right].$$

Then for any compact K,

$$\sup_{t\in[0,T],\Phi_0\in K} |D_N F(t, \Phi_0) - DF(t, \Phi_0)| \to 0.$$

As a corollary with dominated convergence theorem, for any x_0, κ and $\ell, \beta > 0$, we have

$$\frac{N}{P}\left(Z_\beta^{(N,\ell+1,\mathcal{D})} - Z_\beta^{(N,\ell,\mathcal{D})}\right) \to \partial_\tau Z_\beta^{(\tau)}(x_0, \kappa)\Big|_{\tau=P\ell/N}. \tag{20}$$

We present this convergence formally in Theorem 5.

References

[AS17] Madhu S. Advani and Andrew M. Saxe. *High-dimensional dynamics of generalization error in neural networks*. 2017. arXiv: 1710.03667 [stat.ML]. URL: https://arxiv.org/abs/1710.03667.

[Bas+25] Federico Bassetti et al. *Feature learning in finite-width Bayesian deep linear networks with multiple outputs and convolutional layers*. 2025. arXiv: 2406.03260 [stat.ML]. URL: https://arxiv.org/abs/2406.03260.

[Bel+19] Mikhail Belkin et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (July 2019), pp. 15849–15854. ISSN: 1091-6490. DOI: 10.1073/pnas.1903070116. URL: http://dx.doi.org/10.1073/pnas.1903070116.

[BLR24] Federico Bassetti, Lucia Ladelli, and Pietro Rotondo. *Proportional infinite-width infinite-depth limit for deep linear neural networks*. 2024. arXiv: 2411.15267 [stat.ML]. URL: https://arxiv.org/abs/2411.15267.

[Blu+15] Charles Blundell et al. *Weight Uncertainty in Neural Networks*. 2015. arXiv: 1505.05424 [stat.ML]. URL: https://arxiv.org/abs/1505.05424.

[BP22] Lucas Benigni and Sandrine Péché. *Largest Eigenvalues of the Conjugate Kernel of Single-Layered Neural Networks*. 2022. arXiv: 2201.04753 [math.PR]. URL: https://arxiv.org/abs/2201.04753.

[Cam+25] Francesco Camilli et al. *Information-theoretic reduction of deep neural networks to linear models in the overparametrized proportional regime*. 2025. arXiv: 2505.03577 [math.ST]. URL: https://arxiv.org/abs/2505.03577.

[Cho23] Clément Chouard. "Deterministic equivalent of the conjugate kernel matrix associated to artificial neural networks". In: *arXiv preprint arXiv:2306.05850* (2023).

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. "On lazy training in differentiable programming". In: *Advances in neural information processing systems* 32 (2019).

[CS09] Youngmin Cho and Lawrence Saul. "Kernel methods for deep learning". In: *Advances in neural information processing systems* 22 (2009).

[Dau+19] I. Daubechies et al. *Nonlinear Approximation and (Deep) ReLU Networks*. 2019. arXiv: 1905.02199 [cs.LG]. URL: https://arxiv.org/abs/1905.02199.

[Du+19] Simon S. Du et al. *Gradient Descent Provably Optimizes Over-parameterized Neural Networks*. 2019. arXiv: 1810.02054 [cs.LG]. URL: https://arxiv.org/abs/1810.02054.

[El 10] Noureddine El Karoui. "The spectrum of kernel random matrices". In: (2010).

[FW20] Zhou Fan and Zhichao Wang. "Spectra of the Conjugate Kernel and Neural Tangent Kernel for linear-width neural networks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 7710–7721. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/572201a4497b0b9f02d4f279b09ec30d-Paper.pdf.

[GG16] Yarin Gal and Zoubin Ghahramani. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. 2016. arXiv: 1506.02142 [stat.ML]. URL: https://arxiv.org/abs/1506.02142.

[Han19] Boris Hanin. "Universal function approximation by deep neural nets with bounded width and relu activations". In: *Mathematics* 7.10 (2019), p. 992.

[Han24] Boris Hanin. "Random Fully Connected Neural Networks as Perturbatively Solvable Hierarchies". In: *Journal of Machine Learning Research* 25.267 (2024), pp. 1–58. URL: http://jmlr.org/papers/v25/23-0643.html.

[HDR19] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. "Training dynamics of deep networks using stochastic gradient descent via neural tangent kernel". In: *CoRR* (2019).

[HDR22] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. "The curse of depth in kernel regime". In: *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021*. PMLR. 2022, pp. 41–47.

[HJ25] Boris Hanin and Tianze Jiang. *Global Universality of Singular Values in Products of Many Large Random Matrices*. 2025. arXiv: 2503.07872 [math.PR]. URL: https://arxiv.org/abs/2503.07872.

[HN19a] Boris Hanin and Mihai Nica. "Finite depth and width corrections to the neural tangent kernel". In: *arXiv preprint arXiv:1909.05989* (2019).

[HN19b] Boris Hanin and Mihai Nica. "Products of Many Large Random Matrices and Gradients in Deep Neural Networks". In: *Communications in Mathematical Physics* 376.1 (Dec. 2019), pp. 287–322. ISSN: 1432-0916. DOI: 10.1007/s00220-019-03624-z. URL: http://dx.doi.org/10.1007/s00220-019-03624-z.

[HP21] Boris Hanin and Grigoris Paouris. "Non-asymptotic results for singular values of Gaussian matrix products". In: *Geometric and Functional Analysis* 31.2 (2021), pp. 268–324.

[HZ23] Boris Hanin and Alexander Zlokapa. "Bayesian interpolation with deep linear networks". In: *Proceedings of the National Academy of Sciences* 120.23 (2023), e2301345120.

[HZ24] Boris Hanin and Alexander Zlokapa. "Bayesian inference with deep weakly nonlinear networks". In: *arXiv preprint arXiv:2405.16630* (2024).

[JGH20] Arthur Jacot, Franck Gabriel, and Clément Hongler. *Neural Tangent Kernel: Convergence and Generalization in Neural Networks*. 2020. arXiv: 1806.07572 [cs.LG]. URL: https://arxiv.org/abs/1806.07572.

[Lee+17] Jaehoon Lee et al. "Deep neural networks as gaussian processes". In: *arXiv preprint arXiv:1711.00165* (2017).

[Li+26] Mufan Li et al. *Geometric Dyson Brownian Motions and the Free Log-Normal Limit for a Non-Square Product of Random Matrices*. Unpublished Manuscript. 2026.

[LN24] Mufan Bill Li and Mihai Nica. *Differential Equation Scaling Limits of Shaped and Unshaped Neural Networks*. 2024. arXiv: 2310.12079 [stat.ML]. URL: https://arxiv.org/abs/2310.12079.

[LNR22] Mufan Li, Mihai Nica, and Dan Roy. "The neural covariance SDE: Shaped infinite depth-and-width networks at initialization". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 10795–10808.

[LS21] Qianyi Li and Haim Sompolinsky. "Statistical Mechanics of Deep Linear Neural Networks: The Backpropagating Kernel Renormalization". In: *Physical Review X* 11.3 (Sept. 2021). ISSN: 2160-3308. DOI: 10.1103/physrevx.11.031059. URL: http://dx.doi.org/10.1103/PhysRevX.11.031059.

[Mac92] David J. C. MacKay. "A Practical Bayesian Framework for Backpropagation Networks". In: *Neural Computation* 4.3 (1992), pp. 448–472. DOI: 10.1162/neco.1992.4.3.448.

[Mar+21] James Martens et al. "Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping". In: *arXiv preprint arXiv:2110.01765* (2021).

[Mat+18] Alexander G de G Matthews et al. "Gaussian process behaviour in wide deep neural networks". In: *arXiv preprint arXiv:1804.11271* (2018).

[Nea12] Radford M Neal. *Bayesian learning for neural networks*. Vol. 118. Springer Science & Business Media, 2012.

[Nea94] Radford M. Neal. "Bayesian Learning for Neural Networks". Ph.D. thesis. PhD thesis. Department of Computer Science, University of Toronto, 1994. URL: https://glizen.com/radfordneal/ftp/thesis.pdf.

[Noc+23] Lorenzo Noci et al. *The Shaped Transformer: Attention Models in the Infinite Depth-and-Width Limit*. 2023. arXiv: 2306.17759 [stat.ML]. URL: https://arxiv.org/abs/2306.17759.

[Pac+23] R. Pacelli et al. "A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-width limit". In: *Nature Machine Intelligence* 5.12 (Dec. 2023), pp. 1497–1507. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00767-6. URL: http://dx.doi.org/10.1038/s42256-023-00767-6.

[Poo+16] Ben Poole et al. *Exponential expressivity in deep neural networks through transient chaos*. 2016. arXiv: 1606.05340 [stat.ML]. URL: https://arxiv.org/abs/1606.05340.

[PSW13] Nicholas G Polson, James G Scott, and Jesse Windle. "Bayesian inference for logistic models using Pólya–Gamma latent variables". In: *Journal of the American statistical Association* 108.504 (2013), pp. 1339–1349.

[PW17] Jeffrey Pennington and Pratik Worah. "Nonlinear random matrix theory for deep learning". In: *Advances in neural information processing systems* 30 (2017).

[Sch+17] Samuel S. Schoenholz et al. *Deep Information Propagation*. 2017. arXiv: 1611.01232 [stat.ML]. URL: https://arxiv.org/abs/1611.01232.

[SCS88] H. Sompolinsky, A. Crisanti, and H. J. Sommers. "Chaos in Random Neural Networks". In: *Phys. Rev. Lett.* 61 (3 July 1988), pp. 259–262. DOI: 10.1103/PhysRevLett.61.259. URL: https://link.aps.org/doi/10.1103/PhysRevLett.61.259.

[Tre23] Dario Trevisan. *Wide Deep Neural Networks with Gaussian Weights are Very Close to Gaussian Processes*. 2023. arXiv: 2312.11737 [math.ST]. URL: https://arxiv.org/abs/2312.11737.

[Wil96] Christopher Williams. "Computing with Infinite Networks". In: *Advances in Neural Information Processing Systems*. Ed. by M.C. Mozer, M. Jordan, and T. Petsche. Vol. 9. MIT Press, 1996. URL: https://proceedings.neurips.cc/paper_files/paper/1996/file/ae5e3ce40e0404a45ecacaaf05e5f735-Paper.pdf.

[WWF24] Zhichao Wang, Denny Wu, and Zhou Fan. *Nonlinear spiked covariance matrices and signal propagation in deep neural networks*. 2024. arXiv: 2402.10127 [stat.ML]. URL: https://arxiv.org/abs/2402.10127.

[Yan21] Greg Yang. *Tensor Programs III: Neural Matrix Laws*. 2021. arXiv: 2009.10685 [cs.NE]. URL: https://arxiv.org/abs/2009.10685.

[ZBM22] Guodong Zhang, Aleksandar Botev, and James Martens. "Deep learning without shortcuts: Shaping the kernel with tailored rectifiers". In: *arXiv preprint arXiv:2203.08120* (2022).

[Zha+17] Chiyuan Zhang et al. *Understanding deep learning requires rethinking generalization*. 2017. arXiv: 1611.03530 [cs.LG]. URL: https://arxiv.org/abs/1611.03530.

A Computing the prior: Neural Covariance SDE

Let us first lay out the foundations of NSDE as we need. Recall again the forward pass (1):

$$z_1 = \frac{1}{\sqrt{N_0}} W_0 x, \quad \phi_\ell = \phi_s(z_\ell), \quad z_{\ell+1} = \sqrt{\frac{c}{N}} W_\ell \phi_\ell, \quad y = z_{\text{out}} = \sqrt{\frac{c}{N}} W_{\text{out}} \phi_L \in \mathbb{R}$$

Consider network weights at initialization with shaped activation ϕ_s in (1) on a finite dataset $\mathcal{D}$. The key is to analyze how $\Phi^{(\ell)} = \left[\frac{c}{N}\langle \phi_\ell^\alpha, \phi_\ell^\beta \rangle\right]_{\alpha,\beta\in[P]}$ changes from layer ℓ to $\ell+1$. Using the conditional Gaussian structure of $(g_\ell^\alpha)_\alpha$ and a Taylor expansion of ϕ_s at zero, [LNR22] shows that for large n the update:

$$\Phi^{(\ell+1)} = \Phi^{(\ell)} + \frac{1}{N} b(\Phi^{(\ell)}) + \frac{1}{\sqrt{N}} (\Phi^{(\ell)})^{1/2} \xi_\ell (\Phi^{(\ell)})^{1/2} + \text{higher ordered terms}$$

where $\xi_\ell \in \mathbb{R}^{P\times P}$ is sampled from a Gaussian Orthogonal Ensemble , and b is a deterministic bias functions that connects with the activation. Intuitively, the $1/\sqrt{N}$ zero-mean fluctuations come from the variance averaging over N neurons; the drift $b(\Phi)$ comes from the first non-vanishing nonlinear terms in the Taylor series of the activation function. We then let $N \to \infty$ and $L \to \infty$ with $\frac{L}{N} \to T \in (0, \infty)$ and define the continuous "depth-time" index $t \in [0, T]$ via $\ell \approx tN$. Interpolating $\Phi^{(\ell)}$ as a càdlàg process $\Phi^{(\lfloor tN \rfloor)} \triangleq \Phi_t^{(N)}$, Ethier-Kurtz-type results on convergence of Markov chains to diffusions imply our convergence to the solution of a stochastic differential equation, as we see below.

A.1 Local convergence

Let us first set up the proper notion of local convergence needed from the MLP to the Neural Covariance SDE.

Definition 1 (Local convergence in the Skorokhod topology)**.** *We say a sequence of processes X^n converge locally to X in the Skorokhod topology under a continuous test function $f : \cdot \to [0, \infty]$ if for any $r > 0$, the following stopping times*

$$\tau_r^{(n)} \triangleq \inf\{t \geq 0 : f(X_t^n) \geq r\}, \quad \tau_r \triangleq \inf\{t \geq 0 : f(X_t) \geq r\}$$

has that $X^n_{t\wedge\tau^{(n)}} \Rightarrow X_{t\wedge\tau}$ in the Skorokhod topology. Furthermore, we say that the process X_t does not have finite-time explosion if $\mathbb{P}(\lim_{r\to\infty} \tau_r = \infty) = 1$.

Building on the definition, we can immediately show the following (proof deferred to Appendix D).

Lemma 1 (Non-explosion of pre-limit)**.** *Suppose a sequence of processes X^n converge locally to X in the Skorokhod topology under a test function $f : \cdot \to [0, \infty]$, and that the limit process X_t has continuous paths and does not have finite-time explosion. Then for any $r, t > 0$:*

$$\limsup_n \mathbb{P}\left(\tau_r^{(n)} \leq t\right) \leq \mathbb{P}(\tau_r \leq t)$$

and as a result $\lim_{r\to\infty} \limsup_n \mathbb{P}\left(\tau_r^{(n)} \leq t\right) = 0$.

The main reason for the necessity of local convergence is that the drift terms in Equation (10) are not always globally Lipschitz. To that end, we had to introduce the stopping times τ_r such that under F_r, the drift is bounded and Lipschitz. Let us now specify the exact statement for convergence from MLP to Neural Covariance SDE below.

A.2 ReLU networks

Consider the network defined in (1) where $\phi_s(x) = s_+ \max(x, 0) + s_- \min(x, 0)$ is (leaky) ReLU-like with shaping from (3). The NSDE is formally defined as follows.

Proposition 2 (NSDE in ReLU networks, see Theorem 3.2 in [LNR22])**.** *In the notations of (1), let $\Phi^{(\ell)} \in \mathbb{R}^{P\times P}$ where $\Phi^{(\ell)}_{\alpha\beta} \triangleq \frac{c}{N}\left\langle \phi^\alpha_\ell, \phi^\beta_\ell \right\rangle$ and shaped activation*

$$\phi_s(x) = x + \frac{1}{s}\left(c_+ \max(x, 0) + c_- \min(x, 0)\right), \quad s = (N/\alpha_{\text{act}})^{1/2}.$$

In the limit $N \to \infty, \frac{L}{N} \to T < \infty$, the interpolated process $\Phi^{(\lfloor tn \rfloor)}$ converges locally in distribution in the Skorokhod topology of $D_{\mathbb{R}_+,\mathbb{R}^{P\times P}}$ under the test function $f(\Phi) = \max\{(\min_\alpha \Phi_{\alpha\alpha})^{-1}, \max_\alpha \Phi_{\alpha\alpha}\}$ without finite-time explosion to the solution of the SDE

$$\mathrm{d}\Phi_t = \alpha_{\text{act}} \cdot b(\Phi_t)\ \mathrm{d}t + \Phi_t^{1/2}\ \mathrm{d}B_t \Phi_t^{1/2}, \quad \Phi_0 = \left[\frac{1}{N_0}\left\langle x^\alpha, x^\beta \right\rangle\right]_{1\le\alpha\le\beta\le P},$$

where if we let $\rho(x) \triangleq \frac{(c_+-c_-)^2}{2\pi}\left(\sqrt{1-x^2} - x\arccos x\right)$, $D_t \triangleq \operatorname{diag}(\Phi_t)$, then

$$b(\Phi_t) = D_t^{1/2}\rho\left(D_t^{-1/2}\Phi_t D_t^{-1/2}\right) D_t^{1/2} \in \mathbb{R}^{P\times P}$$

in which ρ is applied entry-wise to the normalized $P \times P$ matrix.

In particular, $\rho(1) = 0$. As a result, the diagonal entries follow a geometric Brownian Motion without drift, which explains the finite-time non-explosion. We will also use the fact that the drift-only ODE evolution (Proposition 2 dropping the diffusion term) admits a deterministic feature map independent of $\mathcal{D}$.

Proposition 3 (ODE feature map, ReLU)**.** *There exists a feature map $f : x \in \mathbb{R}^{N_0}, c \in \mathbb{R} \to \mathbb{R}^\infty$ such that the $P \times P$ kernel matrix*

$$(\Phi_c)_{\alpha\beta} = \langle f(x_\alpha, c), f(x_\beta, c)\rangle$$

satisfies the matrix ODE (following notations of Proposition 2),

$$\mathrm{d}\Phi_c = b(\Phi_c)\ \mathrm{d}c, \quad b(\Phi_t) = D_t^{1/2}\rho\left(D_t^{-1/2}\Phi_t D_t^{-1/2}\right) D_t^{1/2} \in \mathbb{R}^{P\times P}$$

Proof of Proposition 3. From Proposition 6 and Lemma 5, we know that the ODE path is always PSD. For any finite set $S \subset \mathbb{R}^{N_0}$, solve the ODE for the restriction Φ_c^S. Since the vector field is entry-wise and compatible with restriction to subsets, these finite-dimensional solutions are consistent under restriction. Hence they define a positive-definite kernel $K_c(x, x')$ on the whole input space, and Moore-Aronszajn applies. Thus, a feature representation induced by the respective RKHS exists. Because this kernel operates on the entire space $\mathbb{R}^{N_0}$ of possible inputs, it cannot be dependent on the specific $\mathcal{D}$. □

A.3 Smooth networks

Consider smooth base activation function $\phi_{\text{base}} \in C^4(\mathbb{R})$ before shaping via (2) that satisfies:

`A1` First two orders of derivatives $0 = \phi(0) = \phi'(0) - 1$.

`A2` There exists some $C > 0$ such that $|\phi^{(4)}(x)| \le C\left(1 + |x|^C\right)$.

`A3` Negative third order derivative $\frac{3}{4}\phi''(0)^2 + \phi'''(0) \le 0$.

The second and third assumptions are technical ones arising from the derivation of NSDE, we refer to [LNR22] for a careful narrative of their necessity. These assumptions were satisfied by common activation functions such as Sigmoid, Tanh, and Soft-Plus.

Proposition 4 (See Theorem 3.9 and Proposition 3.7 in [LNR22])**.** *Assuming that ϕ satisfies assumptions* `A1-3`*. Consider the forward pass (1) with shaping $\phi_s(x) = s\phi(x/s), s = (\alpha_{\text{act}}/N)^{-1/2}$. In the limit $N \to \infty, \frac{L}{N} \to T < \infty$, the interpolated process $\Phi^{(\lfloor tn \rfloor)}$ converges locally in distribution in the Skorokhod topology of $D_{\mathbb{R}_+,\mathbb{R}^{P\times P}}$ under the $f(\Phi) = \max\{(\min_\alpha \Phi_{\alpha\alpha})^{-1}, \max_\alpha \Phi_{\alpha\alpha}\}$ without finite-time explosion to the solution of the SDE*

$$\mathrm{d}\Phi_t = \alpha_{\text{act}} \cdot b(\Phi_t)\ \mathrm{d}t + \Phi_t^{1/2}\ \mathrm{d}B_t \Phi_t^{1/2}, \quad \Phi_0 = \left[\frac{1}{N_0}\left\langle x^\alpha, x^\beta \right\rangle\right]_{1\le\alpha\le\beta\le P},$$

Here $B_t = \frac{1}{\sqrt{2}}\left(\widetilde{B}_t + \widetilde{B}^\top\right) \in \mathbb{R}^{P\times P}$ where $\widetilde{B}_t \in \mathbb{R}^{P\times P}$ has i.i.d. Brownian motion entries. The drift components $b(\Phi)$ is given by:

$$b_{\alpha\beta}(\Phi) = \frac{1}{4}\phi''(0)^2\left[\Phi_{\alpha\alpha}\Phi_{\beta\beta} + \Phi_{\alpha\beta}\left(2\Phi_{\alpha\beta} - 3\right)\right] + \frac{1}{2}\phi'''(0)\Phi_{\alpha\beta}\left(\Phi_{\alpha\alpha} + \Phi_{\beta\beta} - 2\right).$$

for all $\alpha, \beta \in [P]$.

Following Proposition 3, we also have the ODE feature map statement.

Proposition 5 (ODE feature map, Smooth)**.** *There exists a feature map $f : x \in \mathbb{R}^{N_0}, c \in \mathbb{R} \to \mathbb{R}^\infty$ such that the $P \times P$ kernel matrix $(\Phi_c)_{\alpha\beta} = \langle f(x_\alpha, c), f(x_\beta, c)\rangle$ satisfies the matrix ODE $\mathrm{d}\Phi_c = b(\Phi_c)\ \mathrm{d}c$ where*

$$b_{\alpha\beta}(\Phi) = \frac{1}{4}\phi''(0)^2\left[\Phi_{\alpha\alpha}\Phi_{\beta\beta} + \Phi_{\alpha\beta}\left(2\Phi_{\alpha\beta} - 3\right)\right] + \frac{1}{2}\phi'''(0)\Phi_{\alpha\beta}\left(\Phi_{\alpha\alpha} + \Phi_{\beta\beta} - 2\right).$$

Proof. The proof follows immediately from the proof of Proposition 3 as well as combining Proposition 6 and Lemma 5. □

PSD-ness of the SDE solution As a final remark, because the MLP conjugate kernel is PSD by definition, and that the SDE kernel (10) has a continuous path, (local) Skorokhod convergence implies weak convergence pointwise (up to each local stopping time). As a result, both convergences in Proposition 2 and Proposition 4 guarantee positive semidefiniteness on the Neural SDE solution $\Phi_t \succeq 0$ for any $t \ge 0$ in the limit. See Lemma 5.

A.4 Special case 1: deep linear networks

Let us justify `Takeaway 1.1`, the main result in [HZ23], as the first special case. An important aspect of Proposition 1 is that the NSDE separates two effects in the Gram matrix Φ_t dynamic: a multiplicative stochastic component already present in deep linear networks and a drift term induced by shaped nonlinearity. Indeed, with a deep linear network (identity activation function), taking $\phi_s(x) = x$ yields equivalently $c_\pm = 0$ in Proposition 2. In this case, the NSDE becomes

$$\mathrm{d}\Phi_t = \Phi_t^{1/2}\,\mathrm{d}B_t\Phi^{1/2}, \quad t_{\text{depth}} = \ell/N, \quad t_{\text{terminal}} = L/N$$

which is diffusion-only. The terminal time t_{terminal} parameterizes the action at fixed P. As in the main text, for a large $P \to \infty$, a analysis of the operator norm yields

$$\|\Phi_t^{1/2} G \Phi_t^{1/2}\|_{op} \asymp \sqrt{P}\|\Phi_t\|_{op},$$

for a symmetric iid Gaussian G. To balance the sides we will naturally make a time-change $\mathrm{d}\tau = P\,\mathrm{d}t$ such that

$$\mathrm{d}\Phi_\tau = \frac{1}{\sqrt{P}}\Phi_\tau^{1/2}\,\mathrm{d}B_\tau\Phi_\tau^{1/2}, \quad \tau_{\text{depth}} = P\ell/N, \quad \tau \in [0, \tau_{\text{terminal}} = LP/N] \tag{21}$$

which recovers the LP/N rate independently derived in different ways from [HZ23; HZ24; Li+26]. Let us now solve (5) in the case of deep linear networks by (21). Simple algebra reveals that $s_\tau(v) \triangleq v^\top\Phi_\tau v \in \mathbb{R}$ follows the SDE (for a fixed $p \in \mathbb{R}^P$)

$$\mathrm{d}s_\tau(v) = \frac{1}{\sqrt{P}}v^\top\Phi_\tau^{1/2}\,\mathrm{d}B_\tau\Phi_\tau^{1/2}v = \sqrt{\frac{2}{P}}s_\tau(v)\,\mathrm{d}W_\tau, \quad s_0 = v^\top\Phi_0 v$$

where (W_τ) is a standard Brownian motion in $\mathbb{R}$. This means that $s_t(p)$ is a Geometric Brownian motion with a closed terminal expression

$$s_\tau(v) =_d s_0(v)\exp\left(-\frac{\tau}{P} + \sqrt{\frac{2\tau}{P}}G\right), \quad G \sim \mathcal{N}(0,1) \tag{22}$$

and thus the partition function in the propositional limit reads (in this subsection we will use $Z^{(\tau)}$ to denote linear networks for convenience)

$$Z_\beta^{(\tau)}(x_0,\kappa) \propto \int_{\mathbb{R}^P} \mathbb{E}\left[\exp\left(-\frac{1}{2\beta}\|p\|^2 + ip^\top Y - \frac{1}{2}\exp\left(-\frac{\tau}{P} + \sqrt{\frac{2\tau}{P}}G\right)v^\top\Phi_0 v\right)\right]\,\mathrm{d}p$$

where the expectation is taken over $G \sim \mathcal{N}(0,1)$. This simple derivation justifies the first part of `Takeaway 1.1`. To take the perturbative limit $\partial_\tau Z^{(\tau)}$ without differentiating under the integral, consider applying Ito's lemma on $s_\tau = v^\top\Phi_\tau v \geq 0$:[4]

$$\mathrm{d}s_\tau = \sqrt{\frac{2}{P}}s_\tau dB_\tau, \qquad \mathrm{d}e^{-s_\tau/2} = e^{-s_\tau/2}\cdot\frac{1}{4P}s_\tau^2\,\mathrm{d}\tau + (\cdot)\,\mathrm{d}W_\tau$$

[4]We drop the distinction between $\frac{1}{P}$ versus $\frac{1}{P+1}$ here because technically the NSDE is now applied on the $(P+1)\times(P+1)$ Gram matrix on $\mathcal{D}\cup\{x_0\}$. The P vs. $P+1$ distinction is absorbed in $O(P^{-1})$ factors that will be present in the partition function in any case.

and as a result (see Lemma 2 for justification)

$$\partial_\tau \mathbb{E}[e^{-s_\tau/2}]\Big|_{\tau=0} = \frac{1}{4P}\exp\left(-\frac{s_0}{2}\right)s_0^2$$

and since $\exp(-x/2), x \geq 0$ is bounded below and above

$$\partial_\tau Z_\beta^{(\tau)}(x_0,\kappa)\Big|_{\tau=0} = \int_{\mathbb{R}^P} \exp\left(-\frac{1}{2\beta}\|p\|^2 + iY^\top p - \frac{s_0}{2}\right)\cdot\frac{1}{4P}s_0^2\,\mathrm{d}p$$

This is a Gaussian integral because s_0 is a quadratic function of p, so $\partial_\tau Z$ can be seen as the expectation of $\frac{1}{4P}s_0^2$ up to a normalizing constant. In fact, the normalizing constant is exactly

$$\int_{\mathbb{R}^P} \exp\left(-\frac{1}{2\beta}\|p\|^2 + iY^\top p - \frac{1}{2}s_0\right) = Z_\beta^{(\tau=0)}(x_0,\kappa)$$

by (8). As a result, we have that:

$$\partial_\tau \log Z_\beta^{(\tau)}(x_0,\kappa)\Big|_{\tau=0} = \frac{1}{Z_\beta^{(0)}(x_0,\kappa)}\partial_\tau Z_\beta^{(\tau)}(x_0,\kappa)\Big|_{\tau=0} = \mathbb{E}_{p\sim\mathcal{N}(\cdot,\cdot)}\left[\frac{1}{4P}s_0^2\right]$$

where p follows the (complex) Gaussian distribution

$$p \sim \mathcal{N}\left((\Phi_{pp}^\beta)^{-1}(iY - \Phi_{p\kappa}\kappa),\quad (\Phi_{pp}^\beta)^{-1} \triangleq \left(\Phi_{pp} + \frac{1}{\beta}\mathbb{I}\right)^{-1}\right) \tag{23}$$

where $\Phi_0 = \begin{bmatrix} \Phi_{pp} \in \mathbb{R}^{P\times P} & \Phi_{p\kappa} \\ \Phi_{\kappa p} & \Phi_{\kappa\kappa} \in \mathbb{R} \end{bmatrix}$ is the decomposition of input Gram matrix. This gives us a closed-form formula for the first-order perturbation of the linear network partition function without integrating.

Let us now focus on the $\beta = \infty$ case and evaluate the expectation directly

$$s_0 = p^\top \Phi_{pp} p + 2\kappa p^\top \Phi_{p\kappa} + \kappa^2 \Phi_{\kappa\kappa} \tag{24}$$

is a quadratic polynomial of a Gaussian. When $\beta = \infty$, simple calculation yields that:

$$\mathbb{E}_{p\sim(23)}\left[s_0^2\right] = \left[\left(-P + Y^\top \Phi_{pp}^{-1} Y - \kappa^2\left(\Phi_{\kappa\kappa} - \Phi_{\kappa p}\Phi_{pp}^{-1}\Phi_{p\kappa}\right)\right)^2 + 2P - 4Y^\top \Phi_{pp}^{-1} Y\right] \tag{25}$$

which directly proves `Takeaway 1.1`.

Dropping $O(P^{-1})$ terms. Expanding (25) and denoting $\nu_0 = P^{-1}Y^\top \Phi_{pp}^{-1} Y$, $\|x_0^\perp\|^2 = \Phi_{\kappa\kappa} - \Phi_{\kappa p}\Phi_{pp}^{-1}\Phi_{p\kappa}$ gives

$$\partial_\tau \log Z_\infty^{(\tau)}(x_0,\kappa)\Big|_{\tau=0} = \frac{1}{4}\left[P(1-\nu_0)^2 + 2(1-2\nu_0) + 2(1-\nu_0)\|x_0^\perp\|^2\kappa^2 + \frac{1}{P}\|x_0^\perp\|^4\kappa^4\right]$$

is a polynomial of κ. Keeping only the top order (in terms of P) in each coefficient yields

$$\partial_\tau \log Z_\infty^{(\tau)}(x_0,\kappa)\Big|_{\tau=0} = \frac{1+O(P^{-1})}{4}\left[P(1-\nu_0)^2 + 2(1-\nu_0)\|x_0^\perp\|^2\kappa^2\right].$$

As a sanity check, this scale is correct because $\partial_\tau \log Z_\infty(0) = \partial_\tau \log \mathbb{P}(\mathcal{D}) \in \Theta(P)$ and that $\partial_\tau \mathrm{var}(f_{\mathrm{post}}(x_0)) \in \Theta(1)$. Therefore, taking $P \to \infty$ is well-defined.

Positive temperature $\beta^{-1} > 0$**.** Finally, let us address the effect of temperature. Plugging in the completed Gaussian $p \sim \mathcal{N}\left((\Phi_{pp}^{\beta})^{-1}(iY - \Phi_{p\kappa}\kappa), (\Phi_{pp}^{\beta})^{-1} \triangleq \left(\Phi_{pp} + \frac{1}{\beta}\mathbb{I}\right)^{-1}\right)$ into the expectation $\frac{1}{4P}\mathbb{E}\left[(p^\top \Phi_{pp} p + 2\kappa p^\top \Phi_{p\kappa} + \kappa^2 \Phi_{\kappa\kappa})^2\right]$ yields the polynomial expansion

$$\frac{1}{4P}\mathbb{E}\left[s_0^2\right] = \frac{C_0}{4P} + \frac{C_1}{4P}\kappa + \frac{C_2}{4P}\kappa^2 + \frac{C_3}{P}\kappa^3 + \frac{C_4}{4P}\kappa^4,$$

where denoting $\mathbb{I}_\beta = \Phi_{pp}(\Phi_{pp}^{\beta})^{-1} \prec \mathbb{I}$, we specifically need the following coefficients

$$\begin{aligned}
C_0 &= \left[\left(\mathrm{Tr}(\mathbb{I}_\beta) - Y^T(\Phi_{pp}^{\beta})^{-1}\mathbb{I}_\beta Y\right)\right]^2 + 2\,\mathrm{Tr}\left((\mathbb{I}_\beta)^2\right) - 4Y^\top(\Phi_{pp}^{\beta})^{-1}(\mathbb{I}_\beta)^2 Y, \\
C_3 &= i\beta^{-1}\left[\Phi_{\kappa\kappa} - \Phi_{\kappa p}(\Phi_{pp}^{\beta})^{-1}\Phi_{p\kappa} - \beta^{-1}\Phi_{\kappa p}(\Phi_{pp}^{\beta})^{-2}\Phi_{p\kappa}\right]\Phi_{\kappa p}(\Phi_{pp}^{\beta})^{-2}Y \\
C_4 &= \left[\Phi_{\kappa\kappa} - \Phi_{\kappa p}(\Phi_{pp}^{\beta})^{-1}\Phi_{p\kappa} - \beta^{-1}\Phi_{\kappa p}(\Phi_{pp}^{\beta})^{-2}\Phi_{p\kappa}\right]^2.
\end{aligned}$$

Under our standard assumptions on Φ_0, one has that $\Phi_{\kappa\kappa} = \|x_0\|^2 \in \Theta(1), \|\Phi_{\kappa p}\|^2 \in \Theta(1)$ so $C_4 \in O(1)$. Furthermore, $\|Y\| \in O(\sqrt{P})$ so $C_3 \in O(\sqrt{P})$. In other words,

$$\frac{1}{4P}\mathbb{E}\left[s_0^2\right] = \frac{C_0}{4P} + \frac{C_1}{4P}\kappa + \frac{C_2}{4P}\kappa^2 + O(P^{-1/2})$$

is *still* (approximately) a quadratic polynomial of κ.

Another identity we will use later on Bayesian evidence ($\kappa = 0$) at positive temperature $\beta < \infty$ by plugging in $s_0 = p^\top \Phi_{pp} p$ and (23) into $\mathbb{E}[s_0^2]$ is that

$$\frac{1}{4P}\mathbb{E}[s_0^2] = \frac{1}{4P}\left[\left(\mathrm{Tr}(\mathbb{I}_\beta) - Y^T(\Phi_{pp}^{\beta})^{-1}\mathbb{I}_\beta Y\right)\right]^2 + \frac{1}{2P}\,\mathrm{Tr}\left((\mathbb{I}_\beta)^2\right) - \frac{1}{P}Y^\top(\Phi_{pp}^{\beta})^{-1}(\mathbb{I}_\beta)^2 Y$$

Keeping only the leading orders of large P, we are left with

$$\partial_\tau \log Z_\beta^{(\tau)}(0)\Big|_{\tau=0} = \frac{1}{4P}\left[\mathrm{Tr}(\Phi_{pp}(\Phi_{pp}^{\beta})^{-1}) - Y^T(\Phi_{pp}^{\beta})^{-1}\Phi_{pp}(\Phi_{pp}^{\beta})^{-1}Y\right]^2 + O(1) \tag{26}$$

in which $\beta = \infty$ recovers Corollary 3.9 in [HZ23].

A.5 Special case 2: Neural ODE in the infinite-width limit

Let us now study the second special case, which involves the infinite-width limit and $t_{\text{terminal}} = L/N \to 0$. At a first glance, (10) when $t_{\text{terminal}} = 0$ will only return $\Phi_L = \Phi_0$ which leads to kernel method with the trivial identity feature map. Let us copy the NSDE with $s = \sqrt{N/\alpha_{\text{act}}}$ and a fixed P here as

$$\mathrm{d}\Phi_t = \alpha_{\text{act}} \cdot b(\Phi_t)\,\mathrm{d}t + \Phi_t^{1/2}\,\mathrm{d}B_t\Phi_t^{1/2}, \quad t \in [0, L/N].$$

With a direct time-change $\mathrm{d}c = \alpha_{\text{act}}\,\mathrm{d}t$, we have that

$$\mathrm{d}\Phi_c = b(\Phi_c)\,\mathrm{d}c + \frac{1}{\sqrt{\alpha_{\text{act}}}}\Phi_c^{1/2}\,\mathrm{d}B_c\Phi_c^{1/2}, \quad c \in [0, c_{\text{terminal}} = \alpha_{\text{act}}t = L/s^2]. \tag{27}$$

Equation (27) post-time change now allows sending $t \to 0$ so long as $c_{\text{terminal}} > 0$, and we end up with the ODE $\mathrm{d}\Phi_c = b(\Phi_c)\,\mathrm{d}c$. In other words, when we take the shaping $s = (c_{\text{terminal}}/L)^{-1/2}$, the infinite-width limit admits a unique ODE parameterized by c.

Let us now be more precise in the exact limit taken and MLP convergence to the ODE by the following theorem in [LN24] which covers the limit $L = N^{1-\epsilon}$. The strict $N \to \infty$ first, then $L \to \infty$ (effectively $p = 0$) limiting case has been derived using independent methods in [ZBM22; HZ24]. The resulting ODE remains unchanged.

Proposition 6 (Neural Covariance ODE, Proposition 3.4 in [LN24])**.** *Fix any $p \in (0, 1/2)$ and $P > 0$ and consider the limit where $L = N^{2p} \to \infty, s \in \Theta(L^{1/2})$. Conjugate kernel evolution in the shaped ReLU MLP (resp. shaped smooth MLP) converges to the drift-only ODE $\mathrm{d}\Phi_c = b(\Phi_c)\,\mathrm{d}c, c \in [0, L/s^2]$ weakly with respect to the Skorokhod topology of $D_{\mathbb{R}_+, \mathcal{S}^P}$.*

Because the ODE removes stochasticity in the diagonals, the convergence is thus global and not just local. As justified in Proposition 3 and Proposition 5, in both smooth and ReLU cases, the Neural Covariance ODE yields a deterministic, $\mathcal{D}$-independent feature map. Combining the above arguments and Proposition 6 into, we immediately get the following.

Theorem 2 (`Takeaway 1.2`: Kernel method in the Neural Covariance ODE)**.** *Fix any $p \in (0, 1/2)$ and $P > 0$ and consider the limit where $L = N^{2p} \to \infty, s \in \Theta(L^{1/2})$. Bayesian inference is equivalent to a kernel method with feature map $f_0(x; c = L/s^2)$ defined by the Neural ODE that only depends on s and ϕ_{base}.*

Finally, when $L/s^2 = c \in o(1)$, the neural ODE gives the identity feature map, and when $c \to \infty$, the kernel becomes the solution to the ODE at $c = \infty$. It is not hard to check that in both cases of Proposition 2 and Proposition 4, the drift-only ODE at infinite-time converges to a single fixed stationary solution (depending on Φ_0). Therefore, $s \in \Theta(L^{1/2})$ is the unique non-degenerate scaling.

B Perturbative expansion with small depth-time

In this section, we justify the takeaways in Section 4.2 for the NSDE partition function when perturbing around small $\tau = LP/N$. Let us first present a modified Dynkin's formula (the proof deferred to Appendix D).

Lemma 2 (Differentiation of expectation)**.** *Assume that the coefficients $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ are locally Lipschitz and continuous. Let $x_0 \in \mathbb{R}^d$, and let $(X_t)_{0\le t<\tau}$ be the unique local strong solution to the stochastic differential equation:*

$$\mathrm{d}X_t = b\left(X_t\right)\mathrm{d}t + \sigma\left(X_t\right)\mathrm{d}W_t, \quad X_0 = x_0$$

where $\tau = \lim_R \tau_R$ is the explosion stopping time such that $b(\Phi_t), \sigma(\Phi_t)$ are bounded and Lipschitz within $[0, \tau_R]$. Then, for any function $f \in C_b^2\left(\mathbb{R}^d\right)$, we have:

$$\lim_{t\to 0} \frac{\mathbb{E}\left[f\left(X_t\right)\mathbb{1}_{\{t<\tau\}}\right] - f\left(x_0\right)}{t} = b\left(x_0\right)^\top \nabla f\left(x_0\right) + \frac{1}{2}\operatorname{Tr}\left(\sigma\left(x_0\right)\sigma\left(x_0\right)^\top \nabla^2 f\left(x_0\right)\right)$$

In particular, if $\mathbb{P}(\tau = \infty) = 1$, then

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0^+} \mathbb{E}\left[f\left(X_t\right)\right] = b(x_0)^\top \nabla f(x_0) + \frac{1}{2}\operatorname{Tr}\left(\sigma(x_0)\sigma(x_0)^\top \nabla^2 f(x_0)\right).$$

Lemma 2 immediately allow us to show `Takeaway 2` by a simple Ito's Lemma for a finite P via the following arguments. Recall that $s_t(v) = v^\top \Phi_t v$, the above lemma applied on $F(\Phi) = \exp\left(-v^\top \Phi v/2\right)$ suggests that

$$\partial_\tau \mathbb{E}\left[\exp\left(-\frac{s_\tau}{2}\right)\right] = \exp\left(-\frac{s_0}{2}\right)\left(-\frac{1}{2}c_{\text{act}} \cdot v^\top b(\Phi_0)v + \frac{1}{4P}s_0^2\right) \tag{28}$$

applied on (12). Very similar to the linear case (Appendix A.4), one can proceed to integrate over p via the equivalent Gaussian integral (23). To justify `Takeaway 2`, let us first explain why differentiation and integral can be swapped, or in other words, why

$$\partial_\tau \int_{\mathbb{R}^P} e^{iY^\top p - \frac{1}{2\beta}\|p\|^2}\mathbb{E}\left[\exp\left(-\frac{s_\tau}{2}\right)\right]\,\mathrm{d}p = \int_{\mathbb{R}^P} e^{iY^\top p - \frac{1}{2\beta}\|p\|^2}\partial_\tau\mathbb{E}\left[\exp\left(-\frac{s_\tau}{2}\right)\right]\,\mathrm{d}p. \tag{29}$$

Writing both sides as expectations over $p \sim$ (23) as in the linear case, this is equivalent with swapping differentiation with expectation (over p) in

$$\int_{\mathbb{R}^P} \exp\left(iY^\top p - \frac{1}{2\beta}\|p\|^2 - \frac{1}{2}s_0\right)\mathbb{E}\left[\exp\left(-\frac{s_\tau - s_0}{2}\right)\right]\,\mathrm{d}p \propto \mathbb{E}_{p\sim(23)}\mathbb{E}\left[\exp\left(-\frac{s_\tau}{2}\right)\right].$$

By dominated convergence, we only need to show that:

$$\mathbb{E}_{p\sim(23)}\left[\left|\partial_\tau \mathbb{E}\exp\left(-\frac{s_\tau}{2}\right)\right|\right] = \mathbb{E}_{p\sim(23)}\left[\left|-\frac{1}{2}c_{\text{act}} \cdot v^\top b(\Phi_0)v + \frac{1}{4P}s_0^2\right|\right] < \infty$$

This is true because for any architecture, $c_{\text{act}}b(\Phi_0)$ is a constant matrix so $-\frac{1}{2}c_{\text{act}}\cdot v^\top b(\Phi_0)v + \frac{1}{4P}s_0^2$ is simply a polynomial of p, which has bounded first moment. Furthermore, since $s_0 \geq 0$ one has $e^{-s/2} \in (0,1)$, so the condition for dominated convergence is valid. Hence (29) holds for any $\Phi_0 \succ 0$.

Combining (28) and (29), we get the corollary statement which justifies `Takeaway 2`.

Theorem 3 (First-order expansion of log partition function)**.** *The first-order small τ partition function of the Neural Covariance SDE* (11) *with shaping* (16) *can be written as:*

$$\partial_\tau \log Z_\beta^{(\tau)}\Big|_{\tau=0} = c_{\text{act}} \cdot \mathbb{E}_{p\sim(23)}\left[-\frac{1}{2}v^\top b(\Phi_0)v\right] + \mathbb{E}_{p\sim(23)}\left[\frac{1}{4P}s_0^2\right] \tag{30}$$

at any $\beta > 0$.

The second term in (30) is the linear term from Appendix A.4 based on the fact that the Gaussian distribution integrated over p does not change at $\tau = 0$. To see what the first term represents more clearly, consider again the Neural ODE kernel from Appendix A.5. From the ODE $\mathrm{d}\Phi_c = b(\Phi_c)\,\mathrm{d}c$ we have that:

$$\mathrm{d}\exp\left(-\frac{1}{2}v^\top \Phi_c v\right)\Big|_{c=0} = -\frac{1}{2}\exp\left(-\frac{1}{2}s_0\right)v^\top b(\Phi_0)v$$

Now we can swap in differentiation placement for the exact reason as above to get:

$$\frac{1}{Z_\beta^{\text{kernel}}(\cdot; f_0(\cdot;0) = \text{id})}\partial_c Z_\beta^{\text{kernel}}(\cdot; f_0(\cdot;c))\Big|_{c=0} = \mathbb{E}_{(23)}\left[-\frac{1}{2}v^\top b(\Phi_0)v\right]$$

which is exactly the first term in (30), in turn justifying Equation (17).

B.1 First-order evidence and interpretation

From Equation (17), the perturbative evidence (18) in `Takeaway 2.1` follows easily by taking $\kappa = 0$ (so that $Z(x_0, 0) = Z(0)$). The non-linear component is copied verbatim, and the linear evidence follows from (26) directly. Having now the exact expression for

$$\partial_\tau Z(0) = \partial_\tau \mathbb{P}_{\text{prior}}(\mathcal{D})$$

we can interpret the expression as architecture preference for the data. Specifically, we analyze the question of

$$\text{when does a larger } \frac{LP}{N} \text{ increase Bayesian evidence (deeper is better) ?}$$

We do so by evaluating (recall (26))

$$\partial_\tau \log Z(0) = \frac{1}{4P}\left[\text{Tr}(\Phi(\Phi^\beta)^{-1}) - Y^T(\Phi^\beta)^{-1}\Phi(\Phi^\beta)^{-1}Y\right]^2 + c_{\text{act}}\partial_c \log Z_\beta^{\text{kernel}}(0; f_0(\cdot, c))$$

at $\tau = 0$. For the data-generating process, we assume that there exists a norm $\rho > 0$ such that

$$\|\Phi_0 - \rho \mathbb{I}_P\|_{op} \in o(1)$$

that input data are close to orthogonal to each other and all have norm close to ρ. Furthermore, we assume that $|\mathbb{1}^\top Y| = |\sum_{\mathcal{D}} Y_\alpha| \in \Theta(\sqrt{P})$. In this case, the linear evidence evaluates to the simplified expression

$$\frac{1}{4P}\left[\text{Tr}(\Phi(\Phi^\beta)^{-1}) - Y^T(\Phi^\beta)^{-1}\Phi(\Phi^\beta)^{-1}Y\right]^2 = \frac{P}{4}\left(\frac{\rho\|Y\|^2}{P(\rho+\beta^{-1})^2} - \frac{\rho}{\rho+\beta^{-1}}\right)^2$$

Now for the non-linear part we need

$$\partial_c \log Z_\beta^{\text{kernel}}(0; f_0(\cdot, c))\Big|_{c=0} = \mathbb{E}_{p\sim(23)}\left[-\frac{1}{2}p^\top b(\Phi_0)p\right].$$

Below we evaluate the drift-induced evidence for each different b.

Smooth activation function In the smooth case we have:

$$b_{\alpha\beta}(\Phi) = c_1\left[\Phi_{\alpha\alpha}\Phi_{\beta\beta} + \Phi_{\alpha\beta}(2\Phi_{\alpha\beta} - 3)\right] + c_2\Phi_{\alpha\beta}\left(\Phi_{\alpha\alpha} + \Phi_{\beta\beta} - 2\right), \quad \forall \alpha, \beta \in [P]$$

where $c_1 = \frac{1}{4}(\phi''(0))^2, c_2 = \frac{1}{2}\phi'''(0)$. The Gaussian expectation evaluates to:

$$\begin{aligned}\mathbb{E}\left[-\frac{1}{2}p^\top b(\Phi)p\right] =& \frac{\rho\left(2\left(c_1+c_2\right)\rho - \left(3c_1+2c_2\right)\right)}{2(\rho+\beta^{-1})^2}\|Y\|^2 \\ &+ \frac{c_1\rho^2}{2(\rho+\beta^{-1})^2}\left(\mathbb{1}^\top Y\right)^2 - P\frac{\left(3c_1+2c_2\right)\rho(\rho-1)}{2(\rho+\beta^{-1})}\end{aligned}$$

As a result, we have the following statement.

Proposition 7 (Architectural takeaway for smooth activation). *Under the* $\|\Phi_0 - \rho\mathbb{I}\|_{op} = o(1)$ *training data* $X_{\mathcal{D}}$ *assumption, deeper is better, equivalently* $\partial_\tau \mathbb{P}(\mathcal{D}) \geq 0$*, if and only if:*

$$0 \leq \frac{1}{4}\left(\frac{\rho\|Y\|^2}{P(\rho+\beta^{-1})^2} - \frac{\rho}{\rho+\beta^{-1}}\right)^2 + c_{\text{act}} \cdot \left\{\frac{\rho\left(2\left(c_1+c_2\right)\rho - \left(3c_1+2c_2\right)\right)}{2P(\rho+\beta^{-1})^2}\|Y\|^2 + \frac{c_1\rho^2}{2P(\rho+\beta^{-1})^2}\left(\mathbb{1}^\top Y\right)^2 - \frac{\left(3c_1+2c_2\right)\rho(\rho-1)}{2(\rho+\beta^{-1})}\right\}$$

where $c_1 = \frac{1}{4}(\phi''(0))^2, c_2 = \frac{1}{2}\phi'''(0)$.

ReLU activation function Similar to the smooth case, in ReLU networks recall that

$$b\left(\Phi_t\right) = \left(c_+ - c_-\right)^2 D_t^{1/2}\rho\left(D_t^{-1/2}\Phi_t D_t^{-1/2}\right)D_t^{1/2}$$

for $\rho(x) \triangleq \frac{1}{2\pi}\left(\sqrt{1-x^2} - x\arccos x\right)$. The Gaussian expectation evaluates to:

$$\mathbb{E}\left[-\frac{1}{2}p^\top b(\Phi_0)p\right] = \frac{\left(c_+ - c_-\right)^2\rho}{4\pi(\rho+\beta^{-1})^2}\left[\left(\mathbb{1}^\top Y\right)^2 - \|Y\|^2\right].$$

As a result, we have the following statement.

Proposition 8 (Architectural takeaway for ReLU activation). *Under the* $\|\Phi_0 - \rho\mathbb{I}\|_{op} = o(1)$ *training data* $X_{\mathcal{D}}$ *assumption, deeper is better, equivalently* $\partial_\tau \mathbb{P}(\mathcal{D}) \geq 0$*, if and only if:*

$$0 \leq \frac{1}{4}\left(\frac{\rho\|Y\|^2}{P(\rho+\beta^{-1})^2} - \frac{\rho}{\rho+\beta^{-1}}\right)^2 + c_{\text{act}} \cdot \frac{\left(c_+ - c_-\right)^2\rho}{4P\pi(\rho+\beta^{-1})^2}\left[\left(\mathbb{1}^\top Y\right)^2 - \|Y\|^2\right]$$

B.2 First-order predictive posterior

Let us apply (17) to the predictive posterior at test x_0 and justify `Takeaway 2.2`. The predictive posterior can be read off from the partition function directly as the characteristic function

$$\log \mathbb{E}_{\text{post}}[\exp\{-i\kappa f(x_0;\Theta)\}] = \log Z_\beta(x_0,\kappa) - \log Z_\beta(x_0,0).$$

Because the constant ($\kappa = 0$) terms are subtracted from the posterior, we will only work with κ-dependence and use a constant C to placehold any quantity irrelevant to κ. Taking differentiation on both sides, we have thus:

$$\partial_\tau \log \mathbb{E}_{\text{post},\tau}[\exp\{-i\kappa f(x_0;\Theta)\}] = \partial_\tau\left(\log Z_\beta^{(\tau)}(x_0,\kappa) - \log Z_\beta^{(\tau)}(x_0,0)\right).$$

At $\tau = 0$ the init condition of the forward pass, $\log \mathbb{E}_{\text{post}}$ is a quadratic polynomial of κ (as one can tell from Equation (9)). Furthermore, the posterior is a Gaussian random variable if and only if $\log \mathbb{E}_{\text{post}}$ is a quadratic polynomial of κ.

Let us examine the κ-dependence of $\log Z_\beta(x_0,\kappa)$ in light of (30). Note that

$$\mathbb{E}_{p\sim(23)}\left[-\frac{1}{2}v^\top b(\Phi_0)v\right], \quad v = [p^\top,\kappa]^\top$$

is *by definition* a quadratic of κ. From Appendix A.4, we also know that $\mathbb{E}_{p\sim(23)}\left[\frac{1}{4P}s_0^2\right]$ is (approximately) a quadratic polynomial of κ for any β. This gives us the following result.

Proposition 9 (First order predictive posterior Gaussianity)**.** *For fixed κ, the first order perturbation $\partial_\tau \log \mathbb{E}_{\text{post}}[e^{-i\kappa f(x_0)}]$ is a quadratic polynomial of κ up to $O(P^{-1/2})$ terms at any $\beta > 0$. As a result, at the first order of small LP/N the predictive posterior remains Gaussian in the $P \to \infty$ limit.*

Let us carefully compute the exact Gaussian for the simplified case $\beta \to \infty$, in which the higher order terms κ^3, κ^4 from linear networks are smaller at $O(P^{-1})$. Recall that (writing $s_0 = v^\top \Phi_0 v$ as a function of κ and following notations in Appendix A.4):

$$\frac{1}{4P}\mathbb{E}_{p\sim(23)}\left[s_0(\kappa)^2\right] = \frac{1+O(P^{-1})}{2}(1-\nu_0)\|x_0^\perp\|^2\kappa^2 + C$$

where C is independent with κ. Hence

$$\partial_\tau \log Z_\infty^{(\tau)} = c_{\text{act}} \cdot \mathbb{E}_{p\sim(23)}\left[-\frac{1}{2}v^\top b(\Phi_0)v\right] + \frac{1+O(P^{-1})}{2}(1-\nu_0)\|x_0^\perp\|^2\kappa^2 + C$$

Let us now find the respective feature map $f^{(\tau)}(\cdot)$ such that the induced kernel method (9) has the same predictive posterior equivalent to the above. Recall that the Neural ODE (Appendix A.5) gives that for

$$\partial_\tau \log Z_\infty^{\text{kernel}}(x_0, \kappa; f_0(\cdot;\tau)) = \mathbb{E}_{p\sim(23)}\left[-\frac{1}{2}v^\top b(\Phi_0)v\right] + C$$

so

$$\partial_\tau \log Z_\infty^{\text{kernel}}(x_0, \kappa; f_0(\cdot;c_{\text{act}}\tau)) = c_{\text{act}} \cdot \mathbb{E}_{(23)}\left[-\frac{1}{2}v^\top b(\Phi_0)v\right] + C. \tag{31}$$

Denote the quadratic polynomial expansion of $i\kappa$:

$$\mathbb{E}_{(23)}\left[-\frac{1}{2}v^\top b(\Phi_0)v\right] = C - i\kappa c_1 - \frac{1}{2}\kappa^2 c_2, \quad c_1, c_2 \in \mathbb{R}.$$

so

$$\log \frac{Z_\infty^{\text{kernel}}(\cdot; f_0)}{Z_\infty^{\text{kernel}}(0; f_0)} = -i\kappa\left(\Phi_{\kappa p}\Phi_{pp}^{-1}Y + \tau \cdot c_{\text{act}}c_1\right) - \frac{1}{2}\kappa^2\left(\|x_0^\perp\|^2 + \tau \cdot c_{\text{act}}c_2\right) + O(\tau^2 + P^{-1}).$$

Note that (9) asserts (when $\beta = \infty$) that for kernel method the κ^2 coefficient in $\log Z$ only depends on the orthogonal component of test $f^{(\tau)}(x_0)$ to $f^{(\tau)}(\mathcal{D})$, whereas the mean is determined from the linear combination of training labels. Here we see that the mean of first-order predictive posterior remains unchanged from the neural ODE, and the variance gets changed. As a result, it is natural to define the feature map:

$$f^{(\tau)} : x_0 \to \sum_{x_\mu \in \mathcal{D}} a_\mu f_0(x_\mu; c_{\text{act}}\tau) + \lambda f_0^\perp(x_0; c_{\text{act}}\tau)$$

where $f_0(x_\alpha) = f_0(x_\alpha)^\perp + \sum_{\mu\in\mathcal{D}} a_\mu f_0(x_\mu)$ and $f_0(x_\alpha)^\perp$ is orthogonal to $\text{span}_{\mathcal{D}}(f_0(\cdot))$ (see also (15)). This yields the effective

$$\log \frac{Z^{\text{kernel}}(\cdot; f^{(\tau)})}{Z^{\text{kernel}}(0; f^{(\tau)})} = -i\kappa\left(\Phi_{\kappa p}\Phi_{pp}^{-1}Y + \tau c_{\text{act}}c_1\right) - \frac{1}{2}\kappa^2\lambda^2\left(\|x_0^\perp\|^2 + \tau \cdot c_{\text{act}}c_2\right) + O(\tau^2 + P^{-1})$$

Note that (17) evaluates $\log \frac{Z_\infty^{(\tau)}(x_0,\kappa)}{Z_\infty^{(\tau)}(0)}$ to

$$-i\kappa\left(\Phi_{\kappa p}\Phi_{pp}^{-1}Y+\tau c_{\text{act}}c_1\right)-\frac{1}{2}\kappa^2\left(\|x_0^\perp\|^2+\tau\left[c_{\text{act}}c_2+(\nu_0-1)\|x_0^\perp\|^2\right]\right)+O(\tau^2+P^{-1})$$

we only need (*independent* of c_1 and c_2)

$$\lambda^2=\frac{\|x_0^\perp\|^2+\tau\left[c_{\text{act}}c_2+(\nu_0-1)\|x_0^\perp\|^2\right]}{\|x_0^\perp\|^2+\tau\cdot c_{\text{act}}c_2}=1+(\nu_0-1)\tau+O(\tau^2)$$

for

$$\log\frac{Z^{\text{kernel}}(x_0,\kappa;f^{(\tau)})}{Z^{\text{kernel}}(0;f^{(\tau)})}=\log\frac{Z_\infty^{(\tau)}(x_0,\kappa)}{Z_\infty^{(\tau)}(0)}+O(\tau^2+P^{-1}).$$

To summarize the above, we have the following result.

Theorem 4 (Equivalent feature map)**.** *At zero temperature $\beta\to\infty$, kernel method induced by the feature map*

$$f^{(\tau)}:x_0\to\sum_{x_\mu\in\mathcal{D}}a_\mu f_0(x_\mu;c_{\text{act}}\tau)+\lambda f_0^\perp(x_0;c_{\text{act}}\tau),\quad \lambda=1+\frac{1}{2}(\nu_0-1)\tau,$$

where $f_0(x_0;c_{\text{act}}\tau=L/s^2)$ is the ODE feature map from Theorem 2 and coefficients a_μ follow (15), gives an equivalent predictive posterior characteristic function $\log Z(x_0,\kappa)$ (at any test point x_0) for Bayesian inference with NSDE up to $O(\tau^2+P^{-1})$ terms.

C MLP convergence to SDE

Finally, let us take care of the exact convergence from MLP to NSDE regarding the partition function. We will show two statements: a pointwise convergence statement (Theorem 1) as well as a (stronger) statement on the perturbative limit (`Takeaway 4`, see Theorem 5 below). Because one cannot define the partial differentiation with discrete Markov Chains in the MLP (as opposed to the NSDE in Appendix B), our perturbative analysis for the MLP partition function will be translated through the *uniform* convergence of layer-wise generators. Denote:

$$F_v(\Phi)\triangleq\exp\left(-\frac{1}{2}v^\top\Phi v\right)\qquad\text{for some fixed } v.$$

We will use $\mathbb{E}^{\Phi_0}$ to denote the prior since both the MLP and SDE conjugate kernel dynamics are Markovian and depend only on Φ_0. Note that F is bounded by $(0,1]$ by PSD-ness of both MLP and SDE kernel trajectories (see Appendix A). In this section we will also use $\Phi_N^{(\ell)}$ to denote the conjugate kernel at layer ℓ for a width-N MLP and use Φ_t to (only) denote the NSDE from (10) kernel without the $t\to\tau$ time change. Our layer-wise convergence result (formalizing `Takeaway 4`) can be stated as follows.

Theorem 5 (Layer-wise uniform convergence)**.** *Fix a radius $r > 0$ and a terminal time $T > 0$. Let $K_r = \{\Phi : f_{\text{test}}(\Phi) \le r\} \cap \mathcal{S}_+^P$ for the f_{test} in Proposition 2 and Proposition 4. Define the stopping times*

$$\kappa_r^{(N)} \triangleq \frac{1}{N} \inf\left\{\ell \ge 0 : \Phi_N^{(\ell)} \notin K_r\right\}, \quad \kappa_r \triangleq \inf\{t \ge 0 : \Phi_t \notin K_r\},$$

and the stopped difference quotient

$$D_N^r F(t, \Phi_0) \triangleq N\left(\mathbb{E}^{\Phi_0}\left[F\left(\Phi_N^{(\lfloor tN\rfloor+1)\wedge N\kappa_r^{(N)}}\right)\right] - \mathbb{E}^{\Phi_0}\left[F\left(\Phi_N^{(\lfloor tN\rfloor\wedge N\kappa_r^{(N)})}\right)\right]\right),$$

with limit

$$D^r F(t, \Phi_0) \triangleq \mathbb{E}^{\Phi_0}\left[h\left(\Phi^{t\wedge\kappa_r}\right)\mathbb{1}_{\{t<\kappa_r\}}\right].$$

where $h \triangleq \mathcal{L}F$ is the generator. Then in both cases of ReLU and smooth activations,

$$\lim_{N\to\infty} \sup_{t\in[0,T], \Phi_0\in K_{r/2}} |D_N^r F(t, \Phi_0) - D^r F(t, \Phi_0)| = 0.$$

Proof of Theorem 5. Because the MLP is a time-homogeneous Markov chain, let

$$A_N F(\Phi) \triangleq N\mathbb{E}^{\Phi}\left[F(\Phi_N^{(1)}) - F(\Phi)\right], \quad D_N^r F(t, \Phi_0) = \mathbb{E}^{\Phi_0}\left[A_N F\left(\Phi_N^{(\lfloor tN\rfloor)}\right)\mathbb{1}_{\left\{\lfloor tN\rfloor<\kappa_r^{(N)}\right\}}\right]$$

be the one-step generator of the unstopped discrete chain. By definition we have $F \in C_b^\infty$ (with constants depending only on v), and $h \triangleq \mathcal{L}F$ is continuous and bounded on K_r. [LNR22] shows convergence of the generator

$$\Phi_N^{(\ell+1)} - \Phi_N^{(\ell)} = \frac{1}{N} b_N\left(\Phi_N^{(\ell)}\right) + \frac{1}{\sqrt{N}}\sigma_N(\Phi_N^{(\ell)})\xi_{\ell+1} + O\left(N^{-3/2}\right)$$

with $b_N \to b, \sigma_N\sigma_N^\top \to \Sigma$ uniformly on compact sets and Lipschitz on K_r. From [LNR22, Lemma A.5 and Proposition A.6], we have the local uniform generator convergence

$$\varepsilon_N \triangleq \sup_{\Phi\in K_r} |A_N F(\Phi) - \mathcal{L}F(\Phi)| \to 0.$$

Define the killed semigroups

$$Q_{N,t}^r h(\Phi_0) \triangleq \mathbb{E}^{\Phi_0}\left[h\left(\Phi_N^{(\lfloor tN\rfloor)}\right)\mathbb{1}_{\left\{t<\kappa_r^{(N)}\right\}}\right], \quad Q_t^r h(\Phi_0) \triangleq \mathbb{E}^{\Phi_0}\left[h(\Phi_t)\mathbb{1}_{\{t<\kappa_r\}}\right].$$

Then

$$D_N^r F(t, \Phi_0) = Q_{N,t}^r(A_N F)(\Phi_0), \quad D^r F(t, \Phi_0) = Q_t^r(\mathcal{L}F)(\Phi_0).$$

Applying the Feller semigroup convergence criterion [LNR22, Theorem A.3], which gives convergence uniformly for bounded times once the generators converge on a core. Combined with [LNR22, Proposition A.6] we get

$$\eta_N \triangleq \sup_{t\in[0,T],\ \Phi_0\in K_{r/2}} \left|Q_{N,t}^r h(\Phi_0) - Q_t^r h(\Phi_0)\right| \to 0.$$

Although the core in [LNR22] is C_0^∞, our application is local (equivalently, one may multiply F by a smooth cutoff equal to 1 on a neighborhood of K_r). Combining the above and using that $Q^r_{N,t}$ is a contraction in the sup norm, we obtain

$$\begin{aligned}
&\sup_{t\in[0,T],\ \Phi_0\in K_{r/2}} |D^r_N F(t,\Phi_0) - D^r F(t,\Phi_0)| \\
\leq &\sup_{\Phi\in K_r} |A_N F(\Phi) - \mathcal{L}F(\Phi)| + \sup_{t\in[0,T],\ \Phi_0\in K_{r/2}} \left|Q^r_{N,t}h(\Phi_0) - Q^r_t h(\Phi_0)\right| \\
=&\varepsilon_N + \eta_N \to 0
\end{aligned}$$

which proves the claim in both the ReLU and smooth activation regimes. □

Let us conclude with the pointwise convergence in Theorem 1, in which we wanted to show that

$$\begin{aligned}
\lim_{N\to\infty, L=t^\star N} Z_\beta^{(N,L,\mathcal{D})}(x_0,\kappa) \triangleq &\lim \int_{\mathbb{R}^P} \exp\left[-\frac{1}{2\beta}\|p\|^2 + ip^\top Y\right] \mathbb{E}\left[F(\Phi_N^{(L)})\right]\ \mathrm{d}p \\
&= \int_{\mathbb{R}^P} \exp\left[-\frac{1}{2\beta}\|p\|^2 + ip^\top Y\right] \mathbb{E}_{(10)}\left[F(\Phi_{t^\star})\right]\ \mathrm{d}p \triangleq Z_\beta^{(\tau)}(x_0,\kappa)
\end{aligned}$$

Proof of Theorem 1. We start by showing that for any fixed $t > 0$ and any $\Phi_0 \in \mathcal{S}_+^P$, one has that:

$$\lim_{N\to\infty} \mathbb{E}\left[F\left(\Phi_N^{(\lfloor tN\rfloor)}\right)\right] = \mathbb{E}\left[F\left(\Phi_t\right)\right]. \tag{32}$$

for both ReLU and smooth activations (satisfying the assumptions `A1-3`). Once this is shown, then Theorem 1 is concluded automatically by dominated convergence because F is bounded.

For any large stopping radius R, again define $\kappa_R^{(N)}$ and κ_R are the local stopping times without finite-time explosion. The limiting process has continuous sample paths, and hence by the continuous mapping theorem for any fixed t:

$$\Phi_N^{(\lfloor (t\wedge\kappa_R^{(N)})N\rfloor)} \Rightarrow \Phi_{t\wedge\kappa_R}, \quad \forall t$$

by further the boundedness of F:

$$\mathbb{E}^{\Phi_0} F\left(\Phi_N^{(\lfloor (t\wedge\kappa_R^{(N)})N\rfloor)}\right) \to \mathbb{E}^{\Phi_0} F\left(\Phi_{t\wedge\kappa_R}\right), \quad \forall t.$$

We compare stopped and unstopped expectations. By boundedness

$$\left|\mathbb{E}^{\Phi_0} F\left(\Phi_N^{(\lfloor tN\rfloor)}\right) - \mathbb{E}^{\Phi_0} F\left(\Phi_N^{(\lfloor (t\wedge\kappa_R^{(N)})N\rfloor)}\right)\right| \leq \mathbb{P}\left(\kappa_R^{(N)} \leq t\right).$$

Similarly,

$$\left|\mathbb{E}^{\Phi_0} F\left(\Phi_t\right) - \mathbb{E}^{\Phi_0} F\left(\Phi_{t\wedge\kappa_R}\right)\right| \leq \mathbb{P}(\kappa_R \leq t).$$

So we have

$$\left|\mathbb{E}^{\Phi_0} F\left(\Phi_N^{(\lfloor tN\rfloor)}\right) - \mathbb{E}^{\Phi_0} F\left(\Phi_t\right)\right| \leq \mathbb{P}(\kappa_R \leq t) + \mathbb{P}\left(\kappa_R^{(N)} \leq t\right).$$

By Lemma 1, sending $R \to \infty$ allows $\mathbb{P}(\kappa_R \leq t) + \mathbb{P}\left(\kappa_R^{(N)} \leq t\right) \to 0$ and we are done. □

D Technical Lemmas and their proofs

Proof of Lemma 1. Denote $F_r \triangleq \{x : f(x) \geq r\}$. Let $Y_s^{n,r} \triangleq X^n_{s\wedge\tau_r^{(n)}}$ and $Y_s^r \triangleq X_{s\wedge\tau_r}$. By local convergence, $Y^{n,r} \Rightarrow Y^r$ in the Skorokhod topology. Fix $r, t > 0$. Choose $a > t$ such that the evaluation map $y \mapsto y(a)$ is a.s. continuous under Y^r. Such a 's are dense because Càdlàg paths have at most countably many fixed-time discontinuities with positive probability. Since F_r is closed and the paths are right-continuous, whenever $\tau_r^{(n)} < \infty$, the hitting point satisfies

$$X^n_{\tau_r^{(n)}} \in F_r$$

Hence, if $\tau_r^{(n)} \leq t < a$, then the stopped process has already hit F_r, so

$$Y_a^{n,r} = X^n_{\tau_r^{(n)}} \in F_r, \quad \left\{\tau_r^{(n)} \leq t\right\} \subseteq \{Y_a^{n,r} \in F_r\}.$$

Thus

$$\limsup_n \mathbb{P}\left(\tau_r^{(n)} \leq t\right) \leq \limsup_n \mathbb{P}\left(Y_a^{n,r} \in F_r\right)$$

By the continuous mapping theorem, $Y_a^{n,r} \Rightarrow Y_a^r$. Since F_r is closed, Portmanteau gives

$$\limsup_n \mathbb{P}\left(Y_a^{n,r} \in F_r\right) \leq \mathbb{P}\left(Y_a^r \in F_r\right)$$

But $Y_a^r \in F_r$ if and only if $\tau_r \leq a$. Hence

$$\limsup_n \mathbb{P}\left(\tau_r^{(n)} \leq t\right) \leq \mathbb{P}\left(\tau_r \leq a\right) \quad \forall a > t.$$

Now take $a \downarrow t$ and by continuity of probability from above, we get the desired claim. □

Proof of Lemma 2. Denote $Lf(x) \triangleq b(x)^\top \nabla f(x) + \frac{1}{2}\operatorname{Tr}\big(\sigma(x)\sigma(x)^\top \nabla^2 f(x)\big)$. By Itô's formula applied to $X_{t\wedge\tau_R}$,

$$f(X_{t\wedge\tau_R}) - f(x_0) = \int_0^t \mathbf{1}_{\{s<\tau_R\}} Lf(X_s)\, ds + \int_0^t \mathbf{1}_{\{s<\tau_R\}} \nabla f(X_s)^\top \sigma(X_s)\, dW_s.$$

The stochastic integral is a martingale, hence

$$\frac{\mathbb{E}[f(X_{t\wedge\tau_R})] - f(x_0)}{t} = \frac{1}{t}\int_0^t \mathbb{E}\big[\mathbf{1}_{\{s<\tau_R\}} Lf(X_s)\big]\, ds.$$

As $s \to 0$, $X_s \to x_0$ a.s. and $\mathbf{1}_{\{s<\tau_R\}} \to 1$ a.s. Therefore, by dominated convergence,

$$\mathbb{E}\big[\mathbf{1}_{\{s<\tau_R\}} Lf(X_s)\big] \longrightarrow Lf(x_0),$$

and consequently

$$\lim_{t\to 0} \frac{\mathbb{E}[f(X_{t\wedge\tau_R})] - f(x_0)}{t} = Lf(x_0).$$

It remains to compare this with the killed process. Since f is bounded,

$$\left|\mathbb{E}[f(X_t)\mathbf{1}_{\{t<\tau\}}] - \mathbb{E}[f(X_{t\wedge\tau_R})]\right| \leq 2\|f\|_\infty \mathbb{P}(\tau_R \leq t).$$

We show that $\mathbb{P}(\tau_R \le t) = o(t)$ for small t for any R such that $\tau_R > 0$ a.s. Let $K_R \triangleq \{\Phi : \max\{(\min_\alpha \Phi_{\alpha\alpha})^{-1}, \max_\alpha \Phi_{\alpha\alpha} \le R\} \cap \mathcal{S}_+$ be a compact domain, then $K_b \triangleq \sup_{K_R} |b(x)|$ and $K_\sigma \triangleq \sup_{K_R} \|\sigma(x)\|_F$ are both finite. For $u \le t$,

$$X_{u\wedge\tau_R} - x_0 = \int_0^{u\wedge\tau_R} b(X_s)\,ds + \int_0^{u\wedge\tau_R} \sigma(X_s)\,dW_s.$$

If t is small enough that $K_b t \le R/2$, then

$$\{\tau_R \le t\} \subseteq \left\{ \sup_{u\in[0,t]} \left| \int_0^{u\wedge\tau_R} \sigma(X_s)\,dW_s \right| \ge R/2 \right\}.$$

By the Burkholder–Davis–Gundy inequality,

$$\mathbb{P}(\tau_R \le t) \le \frac{C}{R^4}\mathbb{E}\left[\sup_{u\in[0,t]} \left| \int_0^{u\wedge\tau_R} \sigma(X_s)\,dW_s \right|^4 \right] \le \frac{CK_\sigma^4}{R^4} t^2 = o(t).$$

Hence

$$\frac{\mathbb{E}[f(X_t)\mathbf{1}_{\{t<\tau\}}] - \mathbb{E}[f(X_{t\wedge\tau_R})]}{t} \to 0.$$

Combining the preceding limits yields exactly $\lim_{t\downarrow 0} \frac{\mathbb{E}[f(X_t)\mathbf{1}_{\{t<\tau\}}] - f(x_0)}{t} = Lf(x_0)$. □

Lemma 3 (Equivalent Neural Covariance SDE in matrix form)**.** *Given a symmetric positive-definite $\Phi \in \mathbb{R}^{m\times m}$, let $M = m(m+1)/2$ and define $\Sigma \triangleq [\Phi_{\alpha\gamma}\Phi_{\beta\delta} + \Phi_{\alpha\delta}\Phi_{\beta\gamma}]_{\alpha\le\beta,\gamma\le\delta} \in \mathbb{R}^{M\times M}$. Suppose $B \in \mathbb{R}^M$ has i.i.d. $\mathcal{N}(0,1)$ entries and $G \in \mathbb{R}^{m\times m}$ is symmetric and has $\mathcal{N}(0,1)$ off-diagonal and $\mathcal{N}(0,2)$ diagonal entries, then*

$$\left(\Sigma^{1/2}B\right)_{\alpha\le\beta} =_d \left(\Phi^{1/2}G\Phi^{1/2}\right)_{\alpha\le\beta}$$

equal in distribution for all $\alpha, \beta \in [m]$.

Proof. Let $A \triangleq \Phi^{1/2} = A^\top$, then we only need to show that the covariance matrix of $(AGA)_{\alpha\le\beta} \in \mathbb{R}^M$ is exactly Σ. For any $\alpha, \beta, \gamma, \delta$,

$$(AGA)_{\alpha\beta} = \sum_{i,j=1}^m A_{\alpha i} G_{ij} A_{\beta j}.$$

Since $\mathbb{E}\left[G_{ij}G_{k\ell}\right] = \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}$ for all i, j, k, ℓ:

$$\begin{aligned}
\mathbb{E}\left[(AGA)_{\alpha\beta}(AGA)_{\gamma\delta}\right] &= \sum_{i,j,k,\ell} A_{\alpha i}A_{\beta j}A_{\gamma k}A_{\delta\ell}\mathbb{E}\left[G_{ij}G_{k\ell}\right] \\
&= \sum_{i,j,k,\ell} A_{\alpha i}A_{\beta j}A_{\gamma k}A_{\delta\ell}\left(\delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}\right) \\
&= \sum_{i,j} A_{\alpha i}A_{\beta j}A_{\gamma i}A_{\delta j} + \sum_{i,j} A_{\alpha i}A_{\beta j}A_{\gamma j}A_{\delta i} \\
&= \left(\sum_i A_{\alpha i}A_{\gamma i}\right)\left(\sum_j A_{\beta j}A_{\delta j}\right) + \left(\sum_i A_{\alpha i}A_{\delta i}\right)\left(\sum_j A_{\beta j}A_{\gamma j}\right) \\
&= \left(AA^\top\right)_{\alpha\gamma}\left(AA^\top\right)_{\beta\delta} + \left(AA^\top\right)_{\alpha\delta}\left(AA^\top\right)_{\beta\gamma}
\end{aligned}$$

Because $A = \Phi^{1/2}$ is symmetric, $AA^\top = \Phi$. Therefore

$$\mathbb{E}\left[(AGA)_{\alpha\beta}(AGA)_{\gamma\delta}\right] = \Phi_{\alpha\gamma}\Phi_{\beta\delta} + \Phi_{\alpha\delta}\Phi_{\beta\gamma}.$$

Since the mean of G is zero, this concludes the proof. This proof also justifies the difference in form between our (10) versus the vector version in [LNR22]. □

Lemma 4 (Bayesian partition function to conjugate kernel)**.** *For any model $f(x) = W_{\text{out}}^\top h(x)$ with a linear layer as its last where $W_{\text{out}} \sim \mathcal{N}(0, \mathbb{I}_N)$, the Bayes partition function of the predictive posterior (5) under MSE loss at test point x_0 (and training data $X \in \mathbb{R}^{\cdot\times P}, Y \in \mathbb{R}^P$):*

$$Z_\beta(x_0, \kappa) = (2\pi\beta)^{P/2}\mathbb{E}_{\text{prior}}\left[\exp\left[-\frac{\beta}{2}\left\|Y - W_{\text{out}}^\top h(X)\right\|_2^2 - i\kappa f(x_0)\right]\right]$$

is equal to

$$Z_\beta(x_0, \kappa) = \int_{\mathbb{R}^P} \exp\left[-\frac{1}{2\beta}\|p\|^2 + ip^\top Y\right] \cdot \mathbb{E}_{\text{prior}}\left[\exp\left(-\frac{v^\top \Phi v}{2}\right)\right] \, \mathrm{d}p$$

where $v = [p^\top, \kappa]^\top \in \mathbb{R}^{P+1}$ and $\Phi = [h(X), h(x_0)]^\top [h(X), h(x_0)] \in \mathbb{R}^{(P+1)\times(P+1)}$.

Proof. From the Gaussian integral

$$1 = \int_{\mathbb{R}^P} \frac{\mathrm{d}t}{(2\pi\beta)^{P/2}} \exp\left[-\frac{1}{2\beta}\left\|t^\top - i\beta\left(Y - W_{\text{out}}^\top h(X)\right)\right\|^2\right]$$

we have

$$\exp\left[-\frac{\beta}{2}\left\|Y - W_{\text{out}}^\top h(X)\right\|^2\right] = \int_{\mathbb{R}^P} \frac{\mathrm{d}t}{(2\pi\beta)^{P/2}} \exp\left[-\frac{1}{2\beta}\|t\|^2 + it^\top\left(Y - W_{\text{out}}^\top h(X)\right)\right]$$

so

$$\begin{aligned}
Z_\beta(x_0, \kappa) &= (2\pi\beta)^{P/2}\mathbb{E}_{\text{prior}}\left[\exp\left[-\frac{\beta}{2}\left\|Y - W_{\text{out}}^\top h(X)\right\|_2^2 - i\kappa f(x_0)\right]\right] \\
&= \mathbb{E}_{\text{prior}}\left[\int_{\mathbb{R}^P} \exp\left[-\frac{1}{2\beta}\|t\|^2 + it^\top Y - it^\top W_{\text{out}}^\top h(X) - i\kappa W_{\text{out}}^\top h(x_0)\right] \, \mathrm{d}t\right] \\
&= \int_{\mathbb{R}^P} \exp\left[-\frac{1}{2\beta}\|t\|^2 + it^\top Y\right] \cdot \mathbb{E}_{\text{prior}}\left[\exp\left(-\frac{[t^\top, \kappa]\Phi[t^\top, \kappa]^\top}{2}\right)\right] \, \mathrm{d}t
\end{aligned}$$

where in the last line, we integrate over W_{out} to get the desired result. □

Lemma 5 (PSD of the limit)**.** *Suppose $X^n \Rightarrow X$ locally in $D_{\mathbb{R}_+, \mathbb{R}^{m\times m}}$ under test function f without finite-time explosion. If the pre-limit is PSD*

$$\mathbb{P}\left(X_t^n \in \mathcal{S}_+, \forall t \geq 0\right) = 1 \quad \text{for every } n,$$

then the limit is also PSD: $\mathbb{P}\left(X_t \in \mathcal{S}_+, \forall t \geq 0\right) = 1$.

Proof. Let $D(\mathcal{S}_+) \triangleq \left\{x \in D_{\mathbb{R}_+,\mathbb{R}^{m\times m}} : x(t) \in \mathcal{S}_+ \forall t \geq 0\right\}$. First note that $D(\mathcal{S}_+)$is closed in the Skorokhod topology. Indeed, $\mathcal{S}_+ \subseteq \mathbb{R}^{m\times m}$ is closed. If $x^k \to x$ in Skorokhod topology and $x^k(t) \in \mathcal{S}_+$for all t, then, on every compact interval, there exist time changes λ_k such that

$$\sup_{s\leq T} \left\| x^k(\lambda_k(s)) - x(s) \right\| \to 0.$$

Since λ_k is onto, $x^k(\lambda_k(s)) \in \mathcal{S}_+$. Thus $x(s)$ is a limit of points in $\mathcal{S}_+$, so $x(s) \in \mathcal{S}_+$. Hence $x \in D(\mathcal{S}_+)$. Fix $r > 0$, and define the stopped processes

$$Y_t^{n,r} := X^n_{t\wedge\tau_r^{(n)}}, \quad Y_t^r := X_{t\wedge\tau_r}.$$

By local convergence, $Y^{n,r} \Rightarrow Y^r$ in the Skorokhod topology. Since X^n lives in $\mathcal{S}_+$a.s., so does $Y^{n,r}$; hence

$$\mathbb{P}(Y^{n,r} \in D(\mathcal{S}_+)) = 1.$$

Because $D(\mathcal{S}_+)$ is closed, Portmanteau gives

$$1 = \limsup \mathbb{P}(Y^{n,r} \in D(\mathcal{S}_+)) \leq \mathbb{P}(Y^r \in D(\mathcal{S}_+)).$$

so $\mathbb{P}(Y^r \in D(\mathcal{S}_+)) = 1$. Now use non-explosion. On the event $\{\tau_r = \infty\}$, we have

$$Y_t^r = X_{t\wedge\tau_r} = X_t, \quad \{\tau_r = \infty\} \cap \{Y^r \in D(\mathcal{S}_+)\} \subseteq \{X \in D(\mathcal{S}_+)\}.$$

Since $\mathbb{P}(Y^r \in D(\mathcal{S}_+)) = 1$, it follows that

$$\mathbb{P}(X \in D(\mathcal{S}_+)) \geq \mathbb{P}(\tau_r = \infty).$$

Using non-explosion, $\lim_{r\to\infty} \mathbb{P}(\tau_r = \infty) = 1$, we obtain $\mathbb{P}(X_t \in \mathcal{S}_+ \forall t \geq 0) = 1$. □