2021 55th Asilomar Conference on Signals, Systems, and Computers | 978-1-6654-5828-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEEECONF53345.2021.9723137

Depth induces scale-averaging in overparameterized linear Bayesian neural networks

Jacob A. Zavatone-Veth
Department of Physics
Harvard University
Cambridge, MA, United States
jzavatoneveth@g.harvard.edu

Cengiz Pehlevan
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA, United States
cpehlevan@seas.harvard.edu

***Abstract*—Inference in deep Bayesian neural networks is only fully understood in the infinite-width limit, where the posterior flexibility afforded by increased depth washes out and the posterior predictive collapses to a shallow Gaussian process. Here, we interpret finite deep linear Bayesian neural networks as data-dependent scale mixtures of Gaussian process predictors across output channels. We leverage this observation to study representation learning in these networks, allowing us to connect limiting results obtained in previous studies within a unified framework. In total, these results advance our analytical understanding of how depth affects inference in a simple class of Bayesian neural networks.**

***Index Terms*—Bayesian inference, neural networks, representation learning**

I. Introduction

Understanding the effect of depth and width on inference is among the central goals of the modern theory of neural networks. Recent theoretical advances have elucidated the behavior of networks in the infinite-width limit, in which the complexity introduced by depth washes out and inference is described by Gaussian process regression [1]–[7]. However, inference at finite widths, where hidden layers retain the flexibility to learn task-relevant representations, remains incompletely understood [7]–[11]. In the setting of gradient-based maximum likelihood optimization, some insights have been gained through the study of finite overparameterized deep linear neural networks [12]–[15]. In the fully Bayesian setting, the behavior of this simple class of models has been characterized in several limiting cases, including asymptotically at large but finite width [1], [8], [10], [11], [16]. However, a unifying perspective on these results is lacking, and our understanding of inference in deep linear Bayesian neural networks (henceforth ℓBNNs) at finite width remains incomplete.

Here, we make the following contributions toward a more comprehensive understanding of ℓBNN inference:

1) We express the moment generating function of the posterior predictive of a finite, overparameterized deep ℓBNN as a data-dependent continuous scale mixture of Gaussian process (GP) generating functions. This scale average induces coupling across output channels, and compliments previous interpretations of deep ℓBNNs in terms of mixing over an adaptive kernel distribution [17], [18]. This observation is mathematically straightforward, but yields some useful insights into inference in finite ℓBNNs. We extend this argument to compute the posterior mean feature kernel of the network's first layer, allowing us to study the representations learned by finite ℓBNNs.
2) We study the asymptotic behavior of these scale mixtures in several limits, allowing us to connect our results to previous work on the asymptotics of ℓBNNs [8], [10], [11], [16]. We identify several interesting areas for future investigation, and point to challenges for precise characterization of how ℓBNNs behave in certain asymptotic regimes.

II. Setup

We begin by defining our setup and our notation, which is mostly standard [19]–[22]. Depending on context, $\| \cdot \|$ will denote the ℓ_2 norm on vectors or the Frobenius norm on matrices. We will use the shorthand that integrals without specified domains are taken over all real matrices of the implied dimension. We use the standard Loewner order on real symmetric matrices, such that $A \succeq 0$ (respectively $A \succ 0$) means that the matrix A is positive semi-definite, or PSD (respectively positive-definite, or PD). For a matrix $A \in \mathbb{R}^{p \times n}$, we let $\mathrm{v}(A) \in \mathbb{R}^{pn}$ be its row-major vectorization. Then, denoting the Kronecker product by $\otimes$, we have $\mathrm{v}(ABC) = (A \otimes C^\top)\,\mathrm{v}(B)$ for conformable matrices A, B, and C. For brevity, we define the shorthand $\operatorname{etr}(X) = \exp \operatorname{tr}(X)$.

For a set of compatibly-sized matrices $W_1 \in \mathbb{R}^{n_1 \times n_0}$, $W_2 \in \mathbb{R}^{n_2 \times n_1}$, $\dots$, $W_d \in \mathbb{R}^{n_d \times n_{d-1}}$, we define a depth-$d$ ℓBNN as the linear map

$$\begin{aligned} f : \mathbb{R}^{n_0} &\to \mathbb{R}^{n_d} \\ x &\mapsto W_d \cdots W_1 x. \end{aligned} \tag{1}$$

We will assume that the "hidden layer widths" $n_1, n_2, \dots, n_{d-1}$ are all greater than or equal to the output dimension n_d, such that the rank of the end-to-end weight matrix $W_d \cdots W_1$ is not constrained by an intermediate bottleneck. We make the standard choice of isotropic Gaussian priors over the weight matrices:

$$[W_\ell]_{ij} \sim_{\text{i.i.d.}} \mathcal{N}\left(0, n_{\ell-1}^{-1}\right), \tag{2}$$

This work was supported by the Harvard Data Science Initiative Competitive Research Fund, the Harvard Dean's Competitive Fund for Promising Scholarship, and a Google Faculty Research Award.

with variances chosen such that the prior variances of the activations at any layer do not diverge with increasing width [1]–[6], [23]. One could allow general layer-dependent variances σ_ℓ^2/n_ℓ, but for ℓBNNs the additional factors can always be absorbed into the definition of the input so long as they are finite and non-zero. Thus, for the sake of notational clarity, we make the simplest choice of prior variances.

For a training dataset $\mathcal{D} = \{(x_\mu, y_\mu)\}_{\mu=1}^{p}$ of p examples, we choose an isotropic Gaussian likelihood

$$p(\mathcal{D} \,|\, W_1, \dots, W_d) \propto \exp\left(-\frac{\beta}{2} \sum_{\mu=1}^{p} \|f(x_\mu) - y_\mu\|^2 \right); \quad (3)$$

we will refer to the inverse variance $\beta \geq 0$ as the *inverse temperature* in analogy with statistical mechanics. The Bayes posterior over the weight matrices is then given up to normalization as $p(W_1, \dots, W_d \,|\, \mathcal{D}) \propto p(\mathcal{D} \,|\, W_1, \dots, W_d) p(W_1) \cdots p(W_d)$.

We collect the training inputs and targets into data matrices $X \in \mathbb{R}^{p \times n_0}$ and $Y \in \mathbb{R}^{p \times n_d}$ with elements $X_{\mu j} = x_{\mu,j}$ and $Y_{\mu j} = y_{\mu,j}$, respectively. We will sometimes find it useful to consider a differentiated test dataset $\hat{\mathcal{D}} = \{(\hat{x}_{\hat{\mu}}, \hat{y}_{\hat{\mu}})\}_{\hat{\mu}=1}^{\hat{p}}$ with corresponding data matrices $\hat{X} \in \mathbb{R}^{\hat{p} \times n_0}$ and $\hat{Y} \in \mathbb{R}^{\hat{p} \times n_d}$. For these data, we define the associated normalized Gram matrices $G_{xx} \equiv n_0^{-1} X X^\top$, $G_{x\hat{x}} \equiv n_0^{-1} X \hat{X}^\top$, $G_{\hat{x}\hat{x}} \equiv n_0^{-1} \hat{X} \hat{X}^\top$, $G_{yy} \equiv n_d^{-1} Y Y^\top$, and $G_{\hat{y}\hat{y}} \equiv n_d^{-1} \hat{Y} \hat{Y}^\top$. Our assumptions on the data will be given purely in terms of conditions on these Gram matrices. In particular, we will assume that the training input Gram matrix G_{xx} is invertible; other conditions will be introduced as needed. We note that this invertibility condition, combined with our assumption that the hidden layer widths are wide enough such that the end-to-end weight matrix is not rank-constrained, means that the ℓBNNs we consider can linearly interpolate their training data, and are thus overparameterized.

III. Scale-averaging in deep ℓBNNs

A. The function-space prior as a scale mixture

We begin with the nearly trivial observation that, for some input data matrix X, the induced prior over network outputs $F = X W_1^\top \cdots W_d^\top$ can be expressed as a continuous scale mixture of matrix Gaussians. This expression will prove useful in our subsequent study of the posterior predictive by allowing us to compute integrals over network outputs rather than over network weights. This simplification is allowed thanks to the fact that the likelihood (3) models the targets Y as being independent of the parameters given the network outputs.

Recall from §II that the prior distribution of the first layer's weight matrix is a matrix Gaussian:

$$W_1^\top \sim \mathcal{MN}_{n_1 \times n_0}(0, n_0^{-1} I_{n_0}, I_{n_1}). \quad (4)$$

Then, for $W_2, \dots, W_d$ fixed, the distribution of F induced by the prior over W_1 can be read off using the properties of the matrix Gaussian under linear transformations [20]:

$$F = X W_1^\top (W_2^\top \cdots W_d^\top) \sim \mathcal{MN}_{p \times n_d}(0, G_{xx}, L), \quad (5)$$

where we have recognized the normalized Gram matrix G_{xx} and defined the $n_d \times n_d$ matrix

$$L \equiv W_d \cdots W_2 W_2^\top \cdots W_d^\top. \quad (6)$$

For this to make sense, both G_{xx} and L must be of full rank. As stated in §II, we assume the dataset to be such that G_{xx} is invertible. Moreover, denoting the prior distribution over L induced by the priors over $W_2, \dots, W_d$ by ϖ, the stated assumption that $n_1, \dots, n_{d-1} \geq n_d$, implies that L is invertible ϖ-almost surely [19], [20].

Using the law of total expectation, we conclude that the prior over outputs for any fixed set of inputs with invertible Gram matrix is a continuous scale mixture of matrix Gaussians, with the prior density explicitly given as

$$\begin{aligned} p_d(F \,|\, X) = \mathbb{E}_{L \sim \varpi} \Bigg[&(2\pi)^{-n_d p/2} \det(L \otimes G_{xx})^{-1/2} \\ &\times \operatorname{etr}\left(-\frac{1}{2} L^{-1} F^\top G_{xx}^{-1} F \right) \Bigg]. \end{aligned} \quad (7)$$

For a depth-two network, ϖ is a Wishart distribution $L \sim \mathcal{W}_{n_2}(n_1^{-1} I_{n_2}, n_1)$, which simplifies to a scalar Gamma-distributed random variable $\lambda \sim \mathrm{Gamma}(n_1/2, 2/n_1)$ when $n_2 = 1$ [20]. These results allow one to easily write down the density of ϖ with respect to Lebesgue measure in the two-layer case. We note that the density of ϖ is expressible for deeper networks in terms of the Meijer G-function [24]; we will not further pursue this line of analysis in the present work. One could also integrate out weight matrices beyond W_1, but this would yield more complicated formulas for the prior density, which do not permit easy analysis of the posterior predictive [9], [10]. In particular, it is unclear how one might obtain an exact expression for the joint function-space prior density for all p examples [9].

B. The cumulant generating function of the posterior predictive

We now exploit the observations of the previous section to study the posterior predictives of finite ℓBNNs. To do so, we will consider the moment generating function

$$Z(\beta, J) = \mathbb{E}_{W_1, \cdots, W_d \,|\, X, Y} \operatorname{etr}(J^\top W_d \cdots W_1 \hat{X}^\top) \quad (8)$$

of the posterior predictive for some test data $\hat{X}$. To leverage the mixture-of-Gaussians interpretation of the prior, we express the generating function as an integral over function outputs $F = X W_1^\top \cdots W_d^\top$ and $\hat{F} = \hat{X} W_1^\top \cdots W_d^\top$, yielding

$$\begin{aligned} Z(\beta, J) \propto \int dF \, d\hat{F} \, \exp\left(\operatorname{tr}(\hat{F}^\top J) - \frac{\beta}{2} \|F - Y\|^2 \right) \\ \times p_d(F, \hat{F} \,|\, X, \hat{X}) \end{aligned} \quad (9)$$

in terms of the joint prior $p_d(F, \hat{F} \,|\, X, \hat{X})$, where the implied constant of proportionality ensures that $Z(\beta, 0) = 1$. Here, the joint prior $p_d(F, \hat{F} \,|\, X, \hat{X})$ is given by substituting the combined dataset $\begin{bmatrix} X \\ \hat{X} \end{bmatrix}$ into (7) under the temporary assumption that the Gram matrix of the combined dataset is invertible.

We now exchange integration over F and $\hat{F}$ with expectation over the scale matrix L, which allows us to evaluate the

Gaussian integrals over F and $\hat{F}$ exactly. This calculation is easily performed using row-major vectorization [21]; we defer a detailed sketch to the Appendix and merely summarize the result here. We define the $pn_d \times pn_d$ symmetric matrix

$$\Gamma_L \equiv I_{pn_d} + \beta G_{xx} \otimes L, \tag{10}$$

the $\hat{p}n_d \times \hat{p}n_d$ symmetric matrix

$$\Sigma_L \equiv G_{\hat{x}\hat{x}} \otimes L - \beta(G_{x\hat{x}}^\top \otimes L)\Gamma_L^{-1}(G_{x\hat{x}} \otimes L), \tag{11}$$

and the $\hat{p}n_d$-dimensional vector

$$\mu_L \equiv \beta(G_{x\hat{x}}^\top \otimes L)\Gamma_L^{-1}\,\mathrm{v}(Y). \tag{12}$$

We let ρ be a probability measure over $n_d \times n_d$ positive semidefinite matrices, defined by its density

$$\frac{d\rho}{d\varpi} \propto \det\left(\Gamma_L\right)^{-1/2} \exp\left(-\frac{1}{2}\beta\,\mathrm{v}(Y)^\top\Gamma_L^{-1}\,\mathrm{v}(Y)\right) \tag{13}$$

with respect to ϖ; the implied constant of proportionality ensures that $\int_{L\succeq 0} d\rho(L) = 1$. Then,

$$Z(\beta, J) = \mathbb{E}_{L\sim\rho} \exp\left(\mu_L^\top\,\mathrm{v}(J) + \frac{1}{2}\,\mathrm{v}(J)^\top\Sigma_L\,\mathrm{v}(J)\right). \tag{14}$$

From this moment generating function, we can immediately read off that the mean and covariance of the posterior predictive are

$$\langle \hat{F}_{\hat{\mu}j}\rangle = \mathbb{E}_{L\sim\rho}\mu_L^\top\,\mathrm{v}(\chi_{\hat{\mu}j}) \tag{15}$$

and

$$\begin{aligned}\mathrm{cov}(\hat{F}_{\hat{\mu}j}, \hat{F}_{\hat{\nu}k}) = \mathbb{E}_{L\sim\rho}\,&\mathrm{v}(\chi_{\hat{\mu}j})^\top\Sigma_L\,\mathrm{v}(\chi_{\hat{\nu}k}) \\ &+ \mathrm{cov}_{L\sim\rho}\left(\mu_L^\top\,\mathrm{v}(\chi_{\hat{\mu}j}), \mu_L^\top\,\mathrm{v}(\chi_{\hat{\nu}k})\right),\end{aligned} \tag{16}$$

respectively, where we define the $\hat{p} \times n_2$ matrix $[\chi_{\hat{\mu}j}]_{\hat{\rho}l} = \delta_{\hat{\mu}\hat{\rho}}\delta_{jl}$. We remark that all of these results extend to the training set predictor with the replacement $G_{x\hat{x}} \leftarrow G_{xx}$.

We recognize (14) as a scale-average of the GP generating function of a single-layer ℓBNN, for which $L = I_{n_d}$ [1]–[6], [22]. Similarly, the mean predictor is a scale-average of GP mean predictors, while the predictor covariance includes an additional term beyond the average of the GP covariance, as per the law of total covariance [20]. We emphasize that the scale distribution ρ is data-dependent: depth allows the ℓBNN to adaptively couple its output channels in a way that a single-layer network cannot. We finally remark that, unlike in studies of gradient-based maximum likelihood estimation in deep linear networks [12]–[15], no exceptional assumptions on the weight distribution or data are required to obtain this intuitive picture.

The above results are rendered somewhat complicated by the need to average over $n_d \times n_d$ PSD matrices. If $n_d = 1$, the situation simplifies substantially, as the scale variable is now a scalar λ, and the Kronecker products can be eliminated. Concretely, for $\lambda \geq 0$, we define

$$\Gamma_\lambda \equiv I_p + \beta\lambda G_{xx} \in \mathbb{R}^{p\times p}, \tag{17}$$

$$\Sigma_\lambda \equiv \lambda G_{\hat{x}\hat{x}} - \beta\lambda^2 G_{x\hat{x}}^\top\Gamma_\lambda^{-1}G_{x\hat{x}} \in \mathbb{R}^{\hat{p}\times\hat{p}}, \quad \text{and} \tag{18}$$

$$\mu_\lambda \equiv \lambda\beta G_{x\hat{x}}^\top\Gamma_\lambda^{-1}y \in \mathbb{R}^{\hat{p}}, \tag{19}$$

and let ρ be a probability measure on $[0, \infty)$, defined by its density

$$\frac{d\rho}{d\varpi} \propto \det(\Gamma_\lambda)^{-1/2} \exp\left(-\frac{1}{2}\beta y^\top\Gamma_\lambda^{-1}y\right) \tag{20}$$

with respect to ϖ; the implied constant of proportionality ensures that $\int_0^\infty d\rho(\lambda) = 1$. Then, the cumulant generating function of the posterior predictive of a deep ℓBNN with scalar output can be expressed as

$$Z(\beta, j) = \mathbb{E}_{\lambda\sim\rho} \exp\left(\mu_\lambda^\top j + \frac{1}{2}j^\top\Sigma_\lambda j\right). \tag{21}$$

From this, we obtain correspondingly simplified expressions for the predictor mean and covariance, which reduce to

$$\langle\hat{f}\rangle = \mathbb{E}_{\lambda\sim\rho}\mu_\lambda \tag{22}$$

and

$$\mathrm{cov}(\hat{f}) = \mathbb{E}_{\lambda\sim\rho}\Sigma_\lambda + \mathrm{cov}_{\lambda\sim\rho}(\mu_\lambda), \tag{23}$$

respectively. Again, these results represent scale-averages of shallow GP predictors, but they are of a simpler form thanks to the lack of mixing between outputs. Even in this simplified setting, and even if one makes a further restriction to the case in which there is only a single training example, the averages defy exact analysis for general values of the hyperparameters due to the terms of the form $1/(1 + \beta G_{xx}\lambda)$ in the exponent [24].

C. The zero-temperature limit

Though analysis of the scale distribution is challenging for general values of the likelihood variance, the situation simplifies somewhat in the zero-temperature limit $\beta \to \infty$ of vanishing likelihood variance. In this limit, the likelihood tends to a collection of Dirac masses that enforce the constraint that the ℓBNN interpolates its training set. For this interpretation to be sensible at the level of the posterior predictive, the training dataset must be linearly interpolatable, i.e., there must exist some matrix $W \in \mathbb{R}^{n_0\times n_d}$ such that $XW = Y$. We will focus on this case, and operate under the assumption that the training dataset Gram matrix G_{xx} is invertible. Then, we expect all expectations over L to be sufficiently regular such that we can interchange the limit in β with the integrals, which should allow us to compute them using the pointwise limit of the density $d\rho/d\varpi$. We note that, though this limit is convenient for theoretical analysis, it is somewhat unnatural from a Bayesian perspective, as it models the targets as a deterministic function of the outputs [23], [25].

Under these regularity assumptions, we expect to have the almost-sure low-temperature limit $\beta\Gamma_L^{-1} \to G_{xx}^{-1} \otimes L^{-1}$, which yields the almost-sure limiting behavior

$$\mu_L \to (G_{x\hat{x}}^\top G_{xx}^{-1} \otimes I_{n_2})\,\mathrm{v}(Y), \tag{24}$$

$$\Sigma_L \to (G_{\hat{x}\hat{x}} - G_{x\hat{x}}^\top G_{xx}^{-1}G_{x\hat{x}}) \otimes L. \tag{25}$$

Then, the limiting mean predictor simplifies to

$$\lim_{\beta\to\infty}\langle\hat{F}\rangle = G_{x\hat{x}}^\top G_{xx}^{-1}Y. \tag{26}$$

This precisely corresponds to the least-norm pseudoinverse solution to the system $XW = Y$, which is intuitively sensible. Moreover, we have the limiting covariance

$$\begin{aligned}\lim_{\beta\to\infty} \operatorname{cov}(\hat{F}_{\hat{\mu}j}, \hat{F}_{\hat{\nu}k}) = (G_{\hat{x}\hat{x}} - G_{x\hat{x}}^{\top} G_{xx}^{-1} G_{x\hat{x}})_{\hat{\mu}\hat{\nu}} \\ \times \lim_{\beta\to\infty} \mathbb{E}_{L\sim\rho} L_{jk},\end{aligned} \tag{27}$$

which is precisely the GP posterior sample-sample covariance, multiplied by a coupling between output channels.

This argument also yields an approximate density

$$\frac{d\rho}{d\varpi} \propto \frac{1}{\det(L)^{p/2}} \operatorname{etr}\left(-\frac{1}{2} Y^{\top} G_{xx}^{-1} Y L^{-1}\right). \tag{28}$$

In the two-layer case, where L follows a Wishart distribution, the limiting density of ρ with respect to Lebesgue measure on PSD matrices should then be given as

$$\begin{aligned}\frac{d\rho}{dL} &\propto \det(L)^{(n_1-p)/2-(n_2+1)/2} \\ &\times \operatorname{etr}\left(-\frac{1}{2}(n_1 L + Y^{\top} G_{xx}^{-1} Y L^{-1})\right).\end{aligned} \tag{29}$$

This implies that L follows a matrix generalized inverse Gaussian (MGIG) distribution at low temperatures [26], [27]:

$$L \sim \mathcal{MGIG}_{n_2}\left(Y^{\top} G_{xx}^{-1} Y, n_1 I_{n_2}, \frac{n_1 - p}{2}\right). \tag{30}$$

This observation yields several insights. First, it implies that the moment generating functions of L and L^{-1} are given in terms of Bessel functions of matrix argument of the second kind $B_{\nu}(Z)$ [24], [26]–[29]. Second, neither the mean $\mathbb{E}L$ nor the reciprocal mean $\mathbb{E}L^{-1}$ of the MGIG are known in closed form for general values of the parameters [26], [27]. We will therefore resort to studying the behavior of these expectations in various asymptotic limits in §V. However, reasonably efficient algorithms for sampling from the MGIG are available; the situation is of course particularly simple when $n_2 = 1$ [27]. Therefore, this formulation could allow faster numerical studies of two-layer ℓBNNs at low temperatures than is possible through naïve sampling of the weights, as the dimensionality of the search space is reduced from $n_0 n_1 + n_1 n_2$ to n_2^2.

IV. Average first-layer feature kernels in ℓBNNs

We now use the methods of §III to study the average feature kernels of deep ℓBNNs. For technical convenience, we restrict our attention to the kernel of the first hidden layer evaluated on the training set:

$$K \equiv \frac{1}{n_1} X W_1 W_1^{\top} X^{\top}. \tag{31}$$

Then, we can proceed as before to integrate W_1 out of the posterior moment generating function of K:

$$Z(\beta, J) = \mathbb{E}_{W_1,\ldots,W_d \,|\, X,Y} \operatorname{etr}\left(\frac{1}{n_1} J X W_1^{\top} W_1 X^{\top}\right). \tag{32}$$

As discussed in the Appendix, the required computation is straightforward as all integrals are Gaussian. Whereas we considered the full generating function of the posterior predictive, we focus only on the posterior mean of the kernel. Defining the $p \times p$ scale-dependent matrix

$$\begin{aligned}[\Delta_L]_{\mu\nu} \equiv \beta^2 \operatorname{v}(Y)^{\top} \Gamma_L^{-1} (G_{xx} \chi_{\mu\nu} G_{xx} \otimes L) \Gamma_L^{-1} \operatorname{v}(Y) \\ - \beta \operatorname{tr}[\Gamma_L^{-1} (G_{xx} \chi_{\mu\nu} G_{xx} \otimes L)]\end{aligned} \tag{33}$$

for $\chi_{\mu\nu}$ is the $p \times p$ matrix $[\chi_{\mu\nu}]_{\rho\lambda} = \delta_{\mu\rho}\delta_{\nu\lambda}$, the posterior-averaged feature kernel can be expressed as

$$\langle K \rangle = G_{xx} + \frac{1}{n_1} \mathbb{E}_{L\sim\rho} \Delta_L. \tag{34}$$

As the shallow GP result is simply $\langle K \rangle = G_{xx}$, this yields a natural interpretation of the mean kernel of a finite-width deep ℓBNN as the GP kernel plus some correction. Though the complexity of the matrix Δ_L renders this result somewhat less than fully transparent, the situation again simplifies for the case of scalar output, for which we have

$$\Delta_{\lambda} = \lambda \beta^2 G_{xx} \Gamma_{\lambda}^{-1} y y^{\top} \Gamma_{\lambda}^{-1} G_{xx} - \lambda \beta G_{xx} \Gamma_{\lambda}^{-1} G_{xx}. \tag{35}$$

We observe that the first term in this result is the outer product of the non-scale-averaged mean training set predictor $\beta G_{xx} \Gamma_{\lambda}^{-1} y$ with itself. Strikingly, the matrix Δ_{λ}—when evaluated at $\lambda = 1$—is precisely the matrix that appears as the asymptotic correction to the average kernel computed in our previous work [10].

Following the discussion of §III-C, we have the almost-sure pointwise low-temperature limit

$$\lim_{\beta\to\infty} \Delta_L = Y L^{-1} Y^{\top} - n_d G_{xx}. \tag{36}$$

Thus, to compute the average kernel at low temperatures, we must compute the limiting reciprocal mean of L. As noted in §III-C, this is not known in closed form.

V. Asymptotic behavior of ℓBNNs

We now consider the asymptotic behavior of ℓBNNs in various limits, allowing us to connect our results to those of previous works. So as to make contact with as many previous works as possible [1]–[6], [8], [10], [11], [16], we will largely focus on the behavior of the average kernel $\langle K \rangle$. In all cases, we will assume that the hidden layer widths $n_1, \ldots, n_{d-1}$ are of a comparable scale n, such that the ratios n_{ℓ}/n remain fixed as n is taken to be large. As in the rest of the paper, we will assume that G_{xx} is invertible, which requires that $n_0 \geq p$. Thus, in limits in which p is taken to be large, we implicitly also take n_0 to be large. We will only consider limits in which the depth is held fixed and finite, or, at least, tends to infinity far more slowly than the hidden layer width, such that d/n is perturbatively small [10], [11], [30]. For the sake of analytical tractability, will often restrict our attention to two-layer networks ($d = 2$) in the zero-temperature limit ($\beta \to \infty$). For notational brevity, we define the ratios $\alpha \equiv p/n$ and $\gamma \equiv n_d/n$, which, under our assumptions, are bounded as $0 \leq \alpha, \gamma \leq 1$.

A. $n \to \infty$, p and n_d fixed

We first consider the regime in which the hidden layer widths tend to infinity with fixed input dimension, output dimension, training dataset size, and depth. This is the most commonly considered asymptotic regime for BNNs [1]–[6], [10], [11]. In this limit, a simple saddle-point argument shows that the data-dependence in ρ can be neglected, and that the expectations over L should be dominated by the mode of ϖ, which is $L_* = \mathbb{E}_{L\sim\varpi} L = I_{n_d}$ [31]. Applying this result to evaluate the expectations in the posterior predictive (14), we recover the expected correspondence between infinitely-wide BNNs and Gaussian processes [1]–[6], [10], [11].

Moreover, we can use this simple argument to recover the leading asymptotic correction to the average hidden layer kernel computed in our previous work [10]. As the expectation in (34) carries an overall factor of $1/n_1$, the leading correction is simply given by evaluating Δ_L at the saddle-point value of L; corrections to the saddle point at large but finite widths will lead subleading corrections to the kernel [31]. After some algebraic simplification, this yields

$$\langle K\rangle = G_{xx} + \gamma G_{xx}\Gamma_\infty^{-1}(G_{yy} - \Gamma_\infty)\Gamma_\infty^{-1}G_{xx} + \mathcal{O}(\gamma^{-2}), \quad (37)$$

where $\Gamma_\infty \equiv G_{xx} + I_p/\beta$. This matches the result of [10], and is consistent with the observation in §IV that the finite-width kernel in the scalar output setting is simply the average of the asymptotic correction over scales. More generally, one could treat L as a small perturbation of the identity, and use perturbative methods similar to those of our previous work to recover the results on corrections to predictor statistics given there [10].

B. $p \to \infty$, n and n_d fixed

We next consider the regime in which the dataset size is taken to be large relative to the hidden layer width and output dimension. This regime is of interest because one expects posterior concentration to occur in the large-dataset regime [23], [25]. Focusing on the zero-temperature limit, we make the simple approximation of neglecting all terms in the density (29) that do not scale with p, leaving $d\rho/dL \propto \exp\{-p[\operatorname{tr}(Y^\top G_{xx}^{-1} Y L^{-1})/p + \log\det(L)]/2\}$. We then evaluate the integral over L by a saddle-point approximation, yielding $L = Y^\top G_{xx}^{-1} Y/p$. This yields an average kernel of

$$\langle K\rangle \approx (1-\gamma)G_{xx} + \frac{1}{n_1} Y\left(\frac{1}{p}Y^\top G_{xx}^{-1}Y\right)^{-1} Y^\top \quad (38)$$

under the reasonable assumption that $Y^\top G_{xx}^{-1} Y$ is of full rank in this regime. Notably, the correction to the GP kernel need not be vanishingly small.

C. $n, p \to \infty$ and n_d fixed

We now consider the limit in which the hidden layer width and training dataset size tend to infinity for fixed depth and output dimension, as previously studied by Li & Sompolinsky [16]. We focus—as those authors did—on the zero-temperature limit, and restrict our attention to the two-layer case for the sake of analytical tractability. Then, exploiting the results of §III-C, we expect the expectations over L to be dominated by the mode of the MGIG. Concretely, we neglect terms of order n_d/n, while keeping the term $Y^\top G_{xx}^{-1} Y$ as we expect it to be of order p. Then, the mode of ρ is determined by a continuous algebraic Ricatti equation (CARE) [27], [28]:

$$I_{n_2} - L^{-1}\left(\frac{Y^\top G_{xx}^{-1} Y}{n_1}\right)L^{-1} - (1-\alpha)L^{-1} = 0. \quad (39)$$

Using the fact that the solutions of this equation commute with the matrix $Y^\top G_{xx}^{-1} Y$ [28], this is identical to the defining equation for the "renormalization matrix" of [16] in the depth-two case. In particular, using the result of §III-C and §IV, this immediately implies that we recover their results for the predictor statistics and zero-temperature kernel.

After some simplification, the solution to this CARE yields

$$\langle K\rangle \approx \frac{1}{2}G_{xx}\left[(1+\alpha)I_p + \left((1-\alpha)^2 I_p + 4\gamma G_{xx}^{-1}G_{yy}\right)^{1/2}\right], \quad (40)$$

but there may be corrections to the saddle point at non-vanishing γ (in particular, if n_d^2 grows faster than roughly $n^{1/3}$ [31]). To leading order in γ, we have

$$\langle K\rangle \approx (1-\gamma)G_{xx} + \frac{\gamma}{1-\alpha}G_{yy} + \mathcal{O}(\gamma^2), \quad (41)$$

which is easily seen to agree with the result in the finite-p regime upon expanding when $\alpha \ll 1$.

D. $n, n_d \to \infty$ and p fixed

Our analysis in the preceding sections was facilitated by the fact that the dimensionality of the scale integral remained finite. However, regimes in which the number of outputs tends to infinity with the hidden layer width can also be of interest. In particular, this limit is relevant to the study of autoencoding in high dimensions, and potentially also to classification tasks with many groups (e.g., ImageNet [32]). Though the same techniques that permit easy asymptotic analysis of other limits cannot be directly applied [31], the problem of computing kernel statistics can be reformulated as an integral over the $p \times p$ kernel matrices themselves, as noted by Aitchison [8]. Then, provided that p is held fixed, the kernel can be computed using a saddle-point approximation. As in the case above, it is easiest to make analytical progress in two-layer networks at zero temperature. There, one finds that the limiting kernel is determined by the solution to the CARE [8], [33]

$$G_{xx}^{-1} - \gamma K^{-1}G_{yy}K^{-1} + (\gamma-1)K^{-1} = 0. \quad (42)$$

The solution to this CARE yields

$$\langle K\rangle \approx \frac{1}{2}G_{xx}\left[(1-\gamma)I_p + \left((1-\gamma)^2 I_p + 4\gamma G_{xx}^{-1}G_{yy}\right)^{1/2}\right]. \quad (43)$$

In particular, when $\gamma = 1$, we recover Aichison [8]'s result that $K = G_{xx}(G_{xx}^{-1}G_{yy})^{1/2} = (G_{yy}G_{xx}^{-1})^{1/2}G_{xx}$. More generally, we observe that this result is suggestively similar to the kernel in the case of large p and finite n_d. In particular, this result can be recovered by making what is in principle an unjustified naïve Laplace approximation to the integral over L as in the preceding section while keeping terms of order γ and ignoring possible corrections to the saddle point from the high-dimensional measure. Further exploration of this will be an interesting subject for future investigation.

E. $n, n_d, p \to \infty$

Finally, one might consider the regime in which the hidden layer width, output dimension, and dataset size tend jointly to infinity. This regime is more challenging to study than those discussed previously, as there is not a clear way to reduce the problem to a finite-dimensional integral. The natural setup for this joint asymptotic limit is a random design teacher-student setting, in which the input examples are independent and identically distributed samples from some distribution and the targets are generated by a linear model with random coefficient matrix. Then, the ℓBNN problem is closely related to the random-design linear-rank matrix inference task, which is known to be challenging to analyze [34]–[36]. We direct the interested reader to recent works by Barbier and Macris [35] and by Maillard *et al.* [36] on this problem, and defer more detailed analysis to future work.

VI. Discussion and conclusions

In this short paper, we have studied some aspects of inference in finite overparameterized ℓBNNs. We presented a simple argument that leads to a clear conceptual picture of the effect of depth, and exploited those methods to connect the results of previous studies. However, we note that our approach is specialized to linear networks, and would not extend easily to nonlinear BNNs. Taken together, our results provide some insight into finite-width effects in a model where depth does not affect the hypothesis class, but does affect inference.

The output-mixing scale average interpretation studied in this work compliments previous interpretations of deep BNNs as mixtures of GPs across a data-adaptive distribution of the kernel that measures similarities between input examples. This interpretation has been pursued in a series of recent works by Aitchison and colleagues [8], [18], [33], starting with the abovementioned work on kernel statistics in deep ℓBNNs [8]. Those authors have also studied a class of models that generalizes this interpretation of deep BNNs by explicitly fixing prior distributions over data-adaptive kernels, resulting in model predictions [8], [33]. This adaptive-kernel description has also recently been considered by Pleiss and Cunningham [17], who showed that the mean predictor of a two-layer deep GP can be interpreted as a data-dependent mixture of function bases. Their result covers a much broader model class than just ℓBNNs—the class of all possible BNNs, linear or nonlinear, is a degenerate subclass of the set of deep GPs—but does not capture higher moments of the posterior predictive.

For a deep ℓBNNs, the adaptive-kernel interpretation arises naturally if one integrates the readout weight matrix W_d out of the prior, rather than integrating out the first layer weight matrix W_1 as we did here [8], [18]. Other than in the limit of large output dimension, integrating out W_1 rather than W_d affords some advantages—some merely aesthetic, others technical—if one has the specific objective of analytically characterizing inference in ℓBNNs. Both approaches allow for the study of the limit of large width and fixed dataset size, but the need to average over the dataset-size-dimensional kernel matrix makes the large-dataset limit harder to study in the adaptive-kernel interpretation. If one studies the posterior predictive generating function (8) using the adaptive kernel interpretation, one must contend with the need to average over the kernel matrix for the combined train-test set. The blocks of this combined kernel matrix do not appear on equal footing in the generating function because the likelihood only involves the training set; this results in conceptually more complex expressions. Finally, the approach taken here has the advantage of simplifying dramatically for single-output networks; such a simplification is not as obvious in the adaptive-kernel interpretation.

In concurrent work, Lee *et al.* [37] have proposed to manually introduce scale mixing to wide BNNs by fixing priors over the prior variances of the last layer's weights. With this setup, taking the limit of infinite hidden layer width results in a scale mixture of GP predictors. Here, we observe that such scale-averaging arises naturally as an effect of depth in finite-width ℓBNNs, hence their setup could be interpreted as manually compensating for the effective loss of depth in an infinite BNN. Based on numerical experiments, they claim that this method can in some cases improve generalization performance relative to that of the fixed-scale GP predictor corresponding to the infinite-width limit of a BNN with fixed prior weight variances. However, importantly, their setup does not consider coupling across multiple output channels. Comprehensive investigation of when data-adaptive scale mixing yields better generalization performance than a fixed-scale GP will be an interesting subject for future investigation.

To conclude, the results of this work illustrate several important conceptual points. Notably, the behavior of networks with many outputs is qualitatively distinct from those with scalar outputs, as there are interactions between output channels which are apparent neither in the scalar output case nor in the limit of infinite width and fixed output dimension. These interactions render both finite-size and asymptotic analyses more challenging. This issue is not merely one of abstract theoretical interest. Rather, it is potentially relevant to attempts to explain empirical results in deep learning. As modern image recognition tasks often include thousands of classes, the ratio of depth to output dimension of realistic networks may non-negligible [32]. Thus, we believe that careful analysis of representation learning in joint limits of infinite hidden layer width, output dimension, depth, and dataset size will be an important subject for future work.

Acknowledgments

We thank A. Atanasov and B. Bordelon for useful conversations and helpful comments on our manuscript.

References

[1] R. M. Neal, "Priors for infinite networks," in *Bayesian Learning for Neural Networks*. Springer, 1996, pp. 29–53.

[2] C. K. Williams, “Computing with infinite networks,” *Advances in Neural Information Processing Systems*, pp. 295–301, 1997.

[3] J. Lee, J. Sohl-Dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, “Deep neural networks as Gaussian processes,” in *International Conference on Learning Representations*, 2018.

[4] A. G. d. G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, “Gaussian process behaviour in wide deep neural networks,” in *International Conference on Learning Representations*, 2018.

[5] G. Yang, “Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation,” *arXiv preprint arXiv:1902.04760*, 2019.

[6] J. Hron, Y. Bahri, R. Novak, J. Pennington, and J. Sohl-Dickstein, “Exact posterior distributions of wide Bayesian neural networks,” *arXiv preprint arXiv:2006.10541*, 2020.

[7] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, “Finite versus infinite neural networks: an empirical study,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 15 156–15 172.

[8] L. Aitchison, “Why bigger is not always better: on finite and infinite neural networks,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. Daumé III and A. Singh, Eds., vol. 119. PMLR, 07 2020, pp. 156–164.

[9] J. A. Zavatone-Veth and C. Pehlevan, “Exact marginal prior distributions of finite Bayesian neural networks,” in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[10] J. A. Zavatone-Veth, A. Canatar, B. S. Ruben, and C. Pehlevan, “Asymptotics of representation learning in finite Bayesian neural networks,” in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[11] D. A. Roberts, S. Yaida, and B. Hanin, “The principles of deep learning theory,” *arXiv preprint arXiv:2106.10165*, 2021.

[12] K. Fukumizu, “Effect of batch learning in multilayer neural networks,” in *Proceedings of the 5th International Conference on Neural Information Processing*, 1998, pp. 67–70.

[13] A. M. Saxe, J. L. McClelland, and S. Ganguli, “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks,” *arXiv preprint arXiv:1312.6120*, 2013.

[14] C. Yun, S. Krishnan, and H. Mobahi, “A unifying view on implicit bias in training linear neural networks,” in *International Conference on Learning Representations*, 2021.

[15] A. Atanasov, B. Bordelon, and C. Pehlevan, “Neural networks as kernel learners: The silent alignment effect,” *arXiv preprint arXiv:2111.00034*, 2021.

[16] Q. Li and H. Sompolinsky, “Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization,” *Phys. Rev. X*, vol. 11, p. 031059, 09 2021.

[17] G. Pleiss and J. P. Cunningham, “The limitations of large width in neural networks: A deep Gaussian process perspective,” in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[18] L. Aitchison, A. Yang, and S. W. Ober, “Deep kernel processes,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 130–140.

[19] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2012.

[20] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018, vol. 47.

[21] J. R. Magnus and H. Neudecker, *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

[22] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3.

[23] D. J. MacKay, “A practical Bayesian framework for backpropagation networks,” *Neural Computation*, vol. 4, no. 3, pp. 448–472, 1992.

[24] “*NIST Digital Library of Mathematical Functions*,” http://dlmf.nist.gov/, Release 1.1.1 of 2021-03-15, 2021, f. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[25] A. G. Wilson and P. Izmailov, “Bayesian deep learning and a probabilistic perspective of generalization,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 4697–4708.

[26] R. W. Butler, “Generalized inverse Gaussian distributions and their Wishart connections,” *Scandinavian Journal of Statistics*, vol. 25, no. 1, pp. 69–75, 1998.

[27] F. Fazayeli and A. Banerjee, “The matrix generalized inverse Gaussian distribution: Properties and applications,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 648–664.

[28] R. W. Butler and A. T. Wood, “Laplace approximation for Bessel functions of matrix argument,” *Journal of Computational and Applied Mathematics*, vol. 155, no. 2, pp. 359–382, 2003.

[29] C. S. Herz, “Bessel functions of matrix argument,” *Annals of Mathematics*, pp. 474–523, 1955.

[30] B. Hanin, “Random neural networks in the infinite width limit as Gaussian processes,” *arXiv preprint arXiv:2107.01562*, 2021.

[31] Z. Shun and P. McCullagh, “Laplace approximation of high dimensional integrals,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 4, pp. 749–760, 1995.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[33] L. Aitchison, “Deep kernel machines and fast solvers for deep kernel machines,” *arXiv preprint arXiv:2108.13097*, 2021.

[34] J. Bun, R. Allez, J.-P. Bouchaud, and M. Potters, “Rotational invariant estimator for general noisy matrices,” *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7475–7490, 2016.

[35] J. Barbier and N. Macris, “Statistical limits of dictionary learning: random matrix theory and the spectral replica method,” *arXiv preprint arXiv:2109.06610*, 2021.

[36] A. Maillard, F. Krzakala, M. Mézard, and L. Zdeborová, “Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising,” *arXiv preprint arXiv:2110.08775*, 2021.

[37] H. Lee, E. Yun, H. Yang, and J. Lee, “Scale mixtures of neural network Gaussian processes,” *arXiv preprint arXiv:2107.01408*, 2021.

Appendix

In this short appendix, we sketch the derivations of our results for the moment generating function of the posterior predictive (reported in §III) and the posterior average kernel (reported in §IV). Following the setup in §III, computation of the moment generating function of the posterior predictive requires only the evaluation of a single Gaussian integral, hence we will omit many intermediate steps for brevity. We proceed under the assumption that the combined Gram matrix

$$\tilde{G}_{xx} = \begin{bmatrix} G_{xx} & G_{x\hat{x}} \\ G_{x\hat{x}}^{\top} & G_{\hat{x}\hat{x}} \end{bmatrix} \tag{44}$$

is invertible; the result extends to the general case by continuity [19], [20].

We start by using the representation of the generating function as an integral over predictions (9) and the expression for the function-space prior density as a continuous scale mixture (7). Then, assuming that that we can apply Fubini's theorem to interchange the integrals over F and $\hat{F}$ with the integral over L, our first task is to evaluate the matrix Gaussian integral

$$\begin{aligned}&(2\pi)^{-n_d\tilde{p}/2}\det(L)^{-\tilde{p}/2}\det(\tilde{G}_{xx})^{-n_d/2}\\ &\times\int d\tilde{F}\ \mathrm{etr}\bigg(-\frac{1}{2}\beta(F-Y)(F-Y)^{\top}+\hat{F}^{\top}J\\ &\qquad -\frac{1}{2}L^{-1}\tilde{F}^{\top}\tilde{G}_{xx}^{-1}\tilde{F}\bigg),\end{aligned} \tag{45}$$

where $\tilde{F} \equiv [F^\top, \hat{F}^\top]^\top$. This integral is easiest to evaluate using row-major vectorization, for which $\mathrm{v}(\tilde{F}) = [\mathrm{v}(F)^\top, \mathrm{v}(\hat{F})^\top]^\top$. Then, defining the matrix

$$A = \begin{bmatrix} \beta I_{pn_d} & 0 \\ 0 & 0 \end{bmatrix} + \tilde{G}_{xx}^{-1} \otimes L^{-1} \tag{46}$$

and the vector

$$b = \begin{bmatrix} \beta\, \mathrm{v}(Y) \\ \mathrm{v}(J) \end{bmatrix}, \tag{47}$$

the integral of interest can be expressed as

$$\begin{aligned} &(2\pi)^{-n_d \tilde{p}/2} \det(\tilde{G}_{xx} \otimes L)^{-1/2} \\ &\times \int d\,\mathrm{v}(\tilde{F}) \exp\left(-\frac{1}{2} \mathrm{v}(\tilde{F})^\top A\, \mathrm{v}(\tilde{F}) + b^\top \mathrm{v}(\tilde{F}) \right) \\ &= \det(\tilde{G}_{xx} \otimes L)^{-1/2} \det(A)^{-1/2} \exp\left(\frac{1}{2} b^\top A^{-1} b \right) \end{aligned} \tag{48}$$

up to a normalizing factor of $\mathrm{etr}(-\beta Y Y^\top/2)$. Using properties of the Kroenecker product, we find after a bit of algebra that [19]

$$\det(\tilde{G}_{xx} \otimes L) \det(A) = \det[(\tilde{G}_{xx} \otimes L) A] = \det(\Gamma_L) \tag{49}$$

and

$$\begin{aligned} \frac{1}{2} b^\top A^{-1} b - \frac{1}{2} \beta \operatorname{tr}(Y Y^\top) = &-\frac{1}{2} \beta\, \mathrm{v}(Y)^\top \Gamma_L^{-1} \mathrm{v}(Y) \\ &+ \mu_L^\top \mathrm{v}(J) \\ &+ \frac{1}{2} \mathrm{v}(J)^\top \Sigma_L \mathrm{v}(J), \end{aligned}$$

where we have defined the matrices Γ_L and Σ_L and the vector μ_L as in (10), (11), and (12) of the main text, respectively. We then conclude the desired result upon grouping L-dependent terms that do not depend on the source J into the density $d\rho/d\varpi$.

The kernel statistics may be derived through an analogous procedure. We start with the posterior moment generating function

$$Z(\beta, J) = \mathbb{E}_{W_1, \ldots, W_d \,|\, X, Y} \,\mathrm{etr}\left(-\frac{1}{2} \frac{1}{n_1} J X W_1^\top W_1 X^\top \right), \tag{50}$$

where the source term is defined with a factor of $-1/2$ for convenience. As before, the first layer weight matrix can be integrated out, yielding

$$\begin{aligned} &Z(\beta, J) \\ &\quad \propto \mathbb{E}_{L \sim \varpi} (2\pi)^{-n_d p/2} \det(L)^{-p/2} \det(G_{xx})^{-n_d/2} \\ &\qquad \times \int dF \exp\left(-\frac{\beta}{2} \| F - Y \|_F^2 \right) \\ &\qquad\quad \times \mathrm{etr}\left(-\frac{1}{2} L^{-1} F^\top G_{xx}^{-1} \left(I_p + \frac{1}{n_1} G_{xx} J \right) F \right). \end{aligned} \tag{51}$$

This is again a Gaussian integral, hence it can be evaluated by direct computation using the vectorization method discussed above. After varying the result with respect to J, one obtains the formula (34) reported in the main text.