Asymptotics of representation learning in finite Bayesian neural networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent works have suggested that finite Bayesian neural networks may outperform their infinite cousins because finite networks can flexibly adapt their internal representations. However, our theoretical understanding of how the learned hidden layer representations of finite networks differ from the fixed representations of infinite networks remains incomplete. Perturbative finite-width corrections to the network prior and posterior have been studied, but the asymptotics of learned features have not been fully characterized. Here, we argue that the leading finite-width corrections to the average feature kernels for any Bayesian network with linear readout and quadratic cost have a largely universal form. We illustrate this explicitly for two classes of fully connected networks: deep linear networks and networks with a single nonlinear hidden layer. Our results begin to elucidate which features of data wide Bayesian neural networks learn to represent.

1 Introduction

The expressive power of deep neural networks critically depends on their ability to learn to represent the features of data [1–22]. However, the structure of their hidden layer representations is only theoretically well-understood in certain infinite-width limits, in which these representations cannot flexibly adapt to learn data dependent features [3–10, 22]. In the Bayesian setting, these representations are described by fixed, deterministic kernels [3–10]. As a result of this inflexibility, recent works have suggested that finite Bayesian neural networks may generalize better than their infinite counterparts because of their ability to learn representations [9].

Theoretical exploration of how finite and infinite Bayesian neural networks differ has largely focused on the properties of the prior and posterior distributions over network outputs [11–16]. In particular, several works have studied the leading perturbative finite-width corrections to these distributions [11–15]. Yet, the corresponding asymptotic corrections to the feature kernels have only been studied in a few special cases [15]. Therefore, the structure of these corrections, as well as their dependence on network architecture, remain poorly understood. In this paper, we make the following contributions towards the goal of a complete understanding of feature learning at asymptotically large but finite widths:

- We argue that the leading finite-width corrections to the average hidden layer kernels of any Bayesian neural network with linear readout and least-squares loss have a largely prescribed form (Conjecture 1). In particular, we argue that these assumptions fix the dependency of the correction on the target outputs, as well as the fact that the size of the correction should scale with the ratio of the number of outputs to the typical hidden layer width. We show analytically and numerically that these constraints should hold for fully-connected feedforward networks, and conjecture that they should extend to other architectures.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

- We explicitly compute the leading finite-width corrections for deep linear networks (§4.1) and networks with a single nonlinear hidden layer (§4.2), showing that they are of the general form claimed. These tractable examples show how certain directions in the output kernel matrix can be enhanced or suppressed in the hidden layer representations depending on the structure of the input similarities. Moreover, even in this simple setting, different choices of nonlinearity can affect the particular combinations of input and output similarities on which each element of the resulting feature kernel depends.

Our results begin to elucidate the structure of learned representations in wide Bayesian neural networks. The assumptions of our general argument are satisfied in many regression settings, hence our qualitative conclusions should be broadly applicable.

2 Preliminaries

We begin by defining our notation, setup, and assumptions. We will index training and test examples by Greek subscripts $\mu, \nu, \ldots$, and layer dimensions (that is, neurons) by Latin subscripts $j, l, \ldots$. Layers will be indexed by the script Latin letter ℓ. Matrix- or vector-valued quantities corresponding to a given layer will be indexed with a parenthesized superscript, while scalar quantities that depend only on the layer will be indexed with a subscript. Depending on context, $\| \cdot \|$ will denote the ℓ_2 norm on vectors or the Frobenius norm on matrices. We denote the standard Euclidean inner product of two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ by $\mathbf{a} \cdot \mathbf{b}$.

2.1 Bayesian neural networks

Our primary results concern deep Bayesian neural networks with fully connected linear readout. Such a network $\mathbf{f} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_d}$ with d layers can be written as

$$\mathbf{f}(\mathbf{x}; W^d, \mathcal{W}) = \frac{1}{\sqrt{n_{d-1}}} W^{(d)} \boldsymbol{\psi}(\mathbf{x}; \mathcal{W}), \tag{1}$$

where the feature map $\boldsymbol{\psi}(\cdot; \mathcal{W}) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{d-1}}$ includes all $d-1$ hidden layers, collectively parameterized by $\mathcal{W}$. Here, $\boldsymbol{\psi}$ can be some combination of fully-connected feedforward networks, convolutional networks, recurrent networks, et cetera, such that it is expressible within the Tensor Programs framework of Yang [7]. We assume isotropic Gaussian priors over the trainable parameters [1–21], with

$$W_{ij}^{(d)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_d^2) \tag{2}$$

in particular.

In our analysis, we fix an arbitrary training dataset $\mathcal{D} = \{(\mathbf{x}_\mu, \mathbf{y}_\mu)\}_{\mu=1}^p$ of p examples. For analytical tractability, we consider a least-squares cost

$$E(\Theta; \mathcal{D}) = \frac{1}{2} \sum_{\mu=1}^{p} \|\mathbf{f}(\mathbf{x}_\mu) - \mathbf{y}_\mu\|^2, \tag{3}$$

and take the un-normalized likelihood to be $p(\mathcal{D} \,|\, \Theta) \propto \exp(-\beta E)$, where $\beta \geq 0$ is an inverse temperature parameter that tempers the likelihood and $\Theta = \{W^{(d)}, \mathcal{W}\}$. We then introduce the Bayes posterior over parameters given these data:

$$p(\Theta \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \Theta) p(\Theta)}{p(\mathcal{D})}; \tag{4}$$

we denote averages with respect to this distribution by $\langle \cdot \rangle$. This is equivalent to considering the equilibrium distribution of Langevin sampling of the parameters at inverse temperature β with temperature-dependent weight decay:

$$d\Theta^{(\ell)}(t) = -(\beta^{-1} \Sigma \Theta + \nabla_\Theta E) dt + \sqrt{2\beta^{-1}} dB^{(\ell)}(t) \tag{5}$$

where $B^{(\ell)}(t)$ is a standard Wiener process and Σ is the diagonal matrix of prior variances. Though this choice of temperature dependence may *a priori* seem to be one of convenience, we show in Appendices F and G that alternative choices yield pathological low-temperature behavior in deep linear networks. Moreover, by tuning β, one can then adjust whether the posterior (4) is dominated by the prior ($\beta \ll 1$) or the likelihood ($\beta \gg 1$).

2.2 Feature kernels and the Gaussian process limit

We quantify representations using the hidden layer feature kernels, which measure how the similarities between inputs evolve as they are propagated through the network. Schematically, these kernels are defined as

$$K^{(\ell)}_{\mu\nu} \equiv \frac{1}{n_\ell}\phi_\ell(\mathbf{h}^{(\ell)}_\mu)\cdot\phi_\ell(\mathbf{h}^{(\ell)}_\nu) \qquad (\ell = 1,\ldots,d-1), \tag{6}$$

where $\mathbf{h}^{(\ell)}_\mu$ is the n_ℓ-dimensional vector of preactivations at the ℓ-th hidden layer for the μ-th training input, ϕ_ℓ is the activation function, and the inner product is taken over internal indices of the layer. These kernels are the central objects of our study; our concrete goal is to study their averages $\langle K^{(\ell)}\rangle$ with respect to the Bayes posterior.

We consider the limit of large hidden layer widths $n_1, n_2, \ldots, n_{d-1} \to \infty$ with n_0, n_d, p, and d fixed. More precisely, we consider a limit in which $n_\ell = \alpha_\ell n$ for $\ell = 1,\ldots,d-1$, where $\alpha_\ell \in (0,\infty)$ and $n \to \infty$, as studied by [3–14, 16–18, 22] and others. Importantly, we note that size of n_0 relative to n is unimportant for our results, whereas n_d/n and d/n must be small [9, 11, 16]. As mentioned above, we assume that $\boldsymbol{\psi}$ is such that the network can be expressed in the Tensor Programs framework of Yang [7]. Then, in this infinite-width limit, the kernels converge in probability to the deterministic Neural Network Gaussian Process (GP) kernels [3–8], defined recursively as

$$[K^{(\ell)}_{\text{GP}}]_{\mu\nu} = \mathbb{E}\left[\phi_\ell(h^{(\ell)}_\mu)\phi_\ell(h^{(\ell)}_\nu)\ :\ h^{(\ell)}_\mu \sim \mathcal{N}(0, \sigma^2_\ell K^{(\ell-1)}_{\text{GP}} + \zeta^2_\ell \mathbf{1}\mathbf{1}^\top)\right] \tag{7}$$

with base case $K^{(0)}_{\text{GP}} = G_{xx}$ given by the Gram matrix $[G_{xx}]_{\mu\nu} \equiv n_0^{-1}\mathbf{x}_\mu\cdot\mathbf{x}_\nu$ of the inputs. Thus, in the GP limit, the hidden layer kernels are independent of the target outputs.

One heuristically expects the kernels to provide a reasonably complete description of feature learning at large widths. As a possible complication, one could imagine a situation in which the activation distributions are 'rotated' during inference without changing the kernels. However, Yang and Hu [8] have shown that this cannot happen in the infinite-width limit for a restricted class of nonlinearities, and conjecture that it is forbidden more broadly. Therefore, our goal is to compute the leading asymptotic finite-width corrections to the NNGP kernels for $n \gg 1$ as a step towards understanding feature learning in finite width networks.

3 Universal aspects of the asymptotics of learned representations

We first present our main result, which shows that the form of the leading perturbative correction to the average hidden layer kernels of a Bayesian neural network is tightly constrained by the assumptions that the readout is linear, that the cost is quadratic, and that the NNGP limit is well-defined.

3.1 Pertubative corrections to the average feature kernels

Our main result is as follows:

Conjecture 1 *Consider a Bayesian neural network with linear readout of the form* (1)*, with posterior* (4)*. Assume that this network admits a well-defined GP limit, i.e., that its architecture is covered in the convergence proofs of Yang [7]. Then, the averages of the hidden layer kernels with respect to the posterior have well-behaved asymptotic series expansions at large width, with leading terms*

$$\langle K^{(\ell)}_{\mu\nu}\rangle = \mathbb{E}_{\mathcal{W}} K^{(\ell)}_{\mu\nu} + \frac{1}{2}n_d \sum_{\rho,\lambda=1}^{p} [\beta^2\Gamma^{-1}G_{yy}\Gamma^{-1} - \beta\Gamma^{-1}]_{\rho\lambda} \operatorname*{cov}_{\mathcal{W}}(K^{(\ell)}_{\mu\nu}, K^{(d-1)}_{\rho\lambda}) + \ldots, \tag{8}$$

where $[G_{yy}]_{\mu\nu} \equiv n_d^{-1}\mathbf{y}_\mu\cdot\mathbf{y}_\nu$ *is the Gram matrix of the outputs and* $\Gamma \equiv I_p + \beta\sigma_d^2 K^{(d-1)}_{GP}$*. Here, the cumulants of the kernels are computed with respect to the* prior*, and are themselves given by asymptotic series at large widths. The ellipsis denotes terms that are of subleading order in the inverse hidden layer widths.*

In Appendix B, we derive this result perturbatively by expanding the posterior cumulant generating function of the kernels in powers of deviations from their GP values. However, the resulting

perturbation series may not rigorously be an asymptotic series for the average kernels, and this method does not yield quantitative bounds for the width-dependence of the terms. We therefore frame it as a conjecture. For fully-connected networks, also known as multilayer perceptrons (MLPs), a quantitative version of this result follows from the work of Yaida [11]. For MLPs, he showed that $\mathbb{E}_{\mathcal{W}} K^{(\ell)} = K_{\text{GP}}^{(\ell)} + \mathcal{O}(n^{-1})$, $\text{cov}_{\mathcal{W}}(K_{\mu\nu}^{(\ell)}, K_{\rho\lambda}^{(d-1)}) = \mathcal{O}(n^{-1})$, and that higher cumulants are of $\mathcal{O}(n^{-2})$.

Conjecture 1 posits that there are two possible types of leading finite-width corrections to the GP kernels. The first class of corrections are deviations of $\mathbb{E}_{\mathcal{W}} K^{(\ell)}$ from $K_{\text{GP}}^{(\ell)}$. These terms reflect corrections to the prior, and do not reflect non-trivial representation learning as they are independent of the outputs. The second class of corrections are those from the output-dependent $\text{cov}_{\mathcal{W}}(K_{\mu\nu}^{(\ell)}, K_{\rho\lambda}^{(d-1)})$ term. For deep linear MLPs or MLPs with a single hidden layer, $\mathbb{E}_{\mathcal{W}} K^{(\ell)}$ is exactly equal to $K_{\text{GP}}^{(\ell)}$ at any width (see Appendix B) [11], and only the covariance term contributes. More broadly, the results of Yaida [11] show that these two terms should be of the same order for any MLP; $\mathbb{E}_{\mathcal{W}} K^{(\ell)} - K_{\text{GP}}^{(\ell)}$ should not dominate the covariance term.

The leading output-dependent correction has several interesting features. First, it includes a factor of n_d, hence, at least for MLPs, it scales with n_d/n and not with $1/n$ alone. This mirrors the findings of Aitchison [9]; feature learning in wide Bayesian networks with many outputs is qualitatively different from that in networks with few outputs relative to their hidden layer width. If n_d/n does not tend to zero with increasing n, the infinite-width behavior is not described by a standard GP [7, 9]. Moreover, we note that the matrix Γ is invertible at any finite temperature, even when $K_{\text{GP}}^{(d-1)}$ is singular. Therefore, provided that one can extend the GP kernel by continuity to non-invertible G_{xx}, this result can be applied in the data-dense regime $n_0 < p$ as well as the data-sparse regime $n_0 > p$. Finally, we observe that the correction depends on the outputs only through their Gram matrix G_{yy}. This result is intuitively sensible, since with the prior (2) and cost (3) the induced posterior over the output activations is invariant under simultaneous rotation of the output activations and targets. Furthermore, G_{yy} is transformed by factors of the matrix Γ^{-1}, hence the correction depends on certain interactions between the output similarities and the GP kernel $K_{\text{GP}}^{(d-1)}$.

3.2 High- and low-temperature limits of the leading correction

To gain some intuition for the properties of the leading finite-width corrections, we consider their high- and low-temperature limits. These limits correspond to tuning the posterior (4) to be dominated by the prior and the likelihood, respectively. At high temperatures ($\beta \ll 1$), expanding Γ^{-1} as a Neumann series (see Appendix A and [23]) yields

$$\beta^2 \Gamma^{-1} G_{yy} \Gamma^{-1} - \beta \Gamma^{-1} = -\beta I_p + \beta^2 (G_{yy} + \sigma_d^2 K_{\text{GP}}^{(d-1)}) + \mathcal{O}(\beta^3). \tag{9}$$

Thus, at high temperatures, the outputs only influence the average kernels of Conjecture 1 to subleading order in both width and β, which is intuitively sensible given that the likelihood is discounted in this regime. Moreover, the leading output-dependent contribution averages together G_{yy} and $K_{\text{GP}}^{(d-1)}$, hence, intuitively, there is no way to 'cancel' the GP contributions to the average kernels. We note that, at infinite temperature ($\beta = 0$), the posterior reduces to the prior, and all finite-width corrections to the average kernels arise from the discrepancy between $\mathbb{E}_{\mathcal{W}} K^{(\ell)}$ and $K_{\text{GP}}^{(\ell)}$.

At low temperatures ($\beta \gg 1$), the behavior of Γ^{-1} differs depending on whether or not $K_{\text{GP}}^{(d-1)}$ is of full rank. In particular, let P be an orthogonal projection matrix from $\mathbb{R}^p$ to the $\text{rk}\, K_{\text{GP}}^{(d-1)}$-dimensional image of $K_{\text{GP}}^{(d-1)}$. As $K_{\text{GP}}^{(d-1)}$ is a positive-semidefinite real matrix, it has a positive-semidefinite Moore-Penrose pseudoinverse $[K_{\text{GP}}^{(d-1)}]^+$ satisfying $K_{\text{GP}}^{(d-1)} [K_{\text{GP}}^{(d-1)}]^+ = [K_{\text{GP}}^{(d-1)}]^+ K_{\text{GP}}^{(d-1)} = P$. If $K_{\text{GP}}^{(d-1)}$ is full-rank, $P = I_p$ and $[K_{\text{GP}}^{(d-1)}]^+ = [K_{\text{GP}}^{(d-1)}]^{-1}$. Then, we have

$$\beta^2 \Gamma^{-1} G_{yy} \Gamma^{-1} - \beta \Gamma^{-1} = \sigma_d^{-4} [K_{\text{GP}}^{(d-1)}]^+ (G_{yy} - \sigma_d^2 K_{\text{GP}}^{(d-1)}) [K_{\text{GP}}^{(d-1)}]^+ + \mathcal{O}(\beta^{-1}). \tag{10}$$

Therefore, the leading-order low temperature correction depends on the difference between the target and NNGP kernels, while the leading non-trivial high temperature correction depends on their sum. Furthermore, the low-temperature correction projects the target outputs into the image of $K_{\text{GP}}^{(d-1)}$.

4 Learned representations in tractable network architectures

Having derived the general form of the leading perturbative finite-width correction to the average feature kernels, we now consider several example network architectures. For these tractable examples, we explicitly show that Conjecture 1 holds perturbatively, and provide concrete formulas for the corrections.

4.1 Deep linear networks

We first consider deep linear networks with no bias terms, for which the required computations are particularly simple applications of Isserlis' theorem (see Appendix A and [24, 25]). For such a network, we have $K_{\text{GP}}^{(\ell)} = m_\ell^2 G_{xx}$, where $m_\ell^2 \equiv \sigma_\ell^2 \sigma_{\ell-1}^2 \cdots \sigma_1^2$ is the product of prior variances. Deferring the details of our calculation to Appendix D, we find that the assumptions of our general argument are satisfied, and that (8) simplifies to

$$\frac{1}{m_\ell^2}\langle K^{(\ell)}\rangle = G_{xx} + \left(\sum_{\ell'=1}^{\ell} \frac{n_d}{n_{\ell'}}\right) m_d^2 \beta^2 G_{xx}\Gamma^{-1}\left(G_{yy} - m_d^2 G_{xx} - \frac{1}{\beta} I_p\right)\Gamma^{-1} G_{xx} + \mathcal{O}(n^{-2}), \tag{11}$$

where $\Gamma \equiv I_p + \beta m_d^2 G_{xx}$. Thus, the leading corrections to the normalized average kernels $\langle K^{(\ell)}\rangle / m_\ell^2$ are identical up to a scalar factor that encodes the width-dependence of the correction. This sum-of-inverse-widths dependence was previously noted by Yaida [11] in his study of the corrections to the prior of a deep linear network. For a network with hidden layers of equal width n, we have the simple linear dependence $\sum_{\ell'=1}^{\ell}(n_d/n_{\ell'}) = n_d \ell / n$. If one instead includes a narrow bottleneck in an otherwise wide network, this dependence predicts that the kernels before the bottleneck should be close to their GP values, while those after the bottleneck should deviate strongly. In Appendix H, we show that the correction remains of the same form even if one allows arbitrary forward skip connections, though the dependence on width and depth is given by a more complex recurrence relation.

This result simplifies further at low temperatures, where, by the result of §3.2, we have

$$\frac{1}{m_\ell^2}\langle K^{(\ell)}\rangle = G_{xx} + \left(\sum_{\ell'=1}^{\ell} \frac{n_d}{n_{\ell'}}\right)\left(m_d^{-2} G_{yy} - G_{xx}\right) + \mathcal{O}(n^{-2}, \beta^{-1}), \tag{12}$$

in the regime in which G_{xx} is invertible. We thus obtain the simple qualitative picture that the low-temperature average kernels linearly interpolate between the input and output Gram matrices. In Appendix E, we show that this limiting result can be recovered from the recurrence relation derived through other methods by Aitchison [9], who did not use it to compute finite-width corrections.

We can gain some additional understanding of the structure of the correction by using the eigendecomposition of G_{xx}. As G_{xx} is by definition a real positive semidefinite matrix, it admits a unitary eigendecomposition $G_{xx} = U\Lambda U^\dagger$ with non-negative eigenvalues $\Lambda_{\mu\mu}$. In this basis, the average kernel is

$$\frac{1}{m_\ell^2} U^\dagger \langle K^{(\ell)}\rangle U = \Lambda + \left(\sum_{\ell'=1}^{\ell} \frac{n_d}{n_{\ell'}}\right)\left(m_d^{-2}\tilde{\Lambda} U^\dagger G_{yy} U \tilde{\Lambda} - \tilde{\Lambda}\Lambda\right) + \mathcal{O}(n^{-2}), \tag{13}$$

where we have defined the diagonal matrix

$$\tilde{\Lambda} \equiv \frac{\beta m_d^2 \Lambda}{I_p + \beta m_d^2 \Lambda}. \tag{14}$$

As $\beta m_d^2 \Lambda \geq 0$, the diagonal elements of $\tilde{\Lambda}$ are bounded as $0 \leq \tilde{\Lambda}_{\mu\mu} \leq 1$. Thus, the factors of $\Gamma^{-1} G_{xx}$ by which G_{yy} is conjugated have the effect of suppressing directions in the projection of G_{yy} onto the eigenspace of G_{xx} with small eigenvalues. We can see that this effect will be enhanced at high temperatures ($\beta \ll 1$) and small scalings ($m_d^2 \ll 1$), and suppressed at low temperatures and large scalings. For this linear network, similarities are not enhanced, only suppressed. Moreover, if G_{xx} is diagonal, then a given element of the average kernel will depend only on the corresponding element of G_{yy}.

4.2 Networks with a single nonlinear hidden layer

We now would like to characterize how including nonlinearity affects the structure of learned representations. For a general deep MLP, it is generally not possible to analytically compute $\mathrm{cov}_{\mathcal{W}}(K^{(\ell)}_{\mu\nu}, K^{(d-1)}_{\rho\lambda})$ to the required perturbative order [8, 11, 17, 18]. However, we can make some analytical progress in a network with a single nonlinear hidden layer and no bias terms. As detailed in Appendix I, for such a network we have the exact expressions

$$[K_{\text{GP}}]_{\mu\nu} = \mathbb{E}\left[\phi(h^{(1)}_{\mu})\phi(h^{(1)}_{\nu})\right] \tag{15}$$

and

$$\underset{\mathcal{W}}{\mathrm{cov}}(K_{\mu\nu}, K_{\rho\lambda}) = \frac{1}{n_1}\left(\mathbb{E}\left[\phi(h^{(1)}_{\mu})\phi(h^{(1)}_{\nu})\phi(h^{(1)}_{\rho})\phi(h^{(1)}_{\lambda})\right] - [K_{\text{GP}}]_{\mu\nu}[K_{\text{GP}}]_{\rho\lambda}\right), \tag{16}$$

where we drop the layer index on the kernel for brevity as there is only one hidden layer, and expectations are taken over the p-dimensional Gaussian random vector $h^{(1)}_{\mu}$, which has mean zero and covariance $\mathrm{cov}(h^{(1)}_{\mu}, h^{(1)}_{\nu}) = \sigma_1^2(G_{xx})_{\mu\nu}$.

Though these expressions are easy to define, it is not possible to evaluate the four-point expectation in closed form for general Gram matrices G_{xx} and nonlinearities ϕ. This obstacle has been noted in previous studies [8, 11, 14], and makes it challenging to extend approaches similar to those used here to deeper nonlinear networks. For polynomial activation functions, the required expectations can be evaluated using Isserlis' theorem (see Appendix A). However, even for a quadratic activation function $\phi(x) = x^2$, the resulting formula for the kernel will involve many elementwise matrix products, and cannot be simplified into an intuitively comprehensible form.

To gain some intuition for how different choices of nonlinear activation function affect the learned representations, we consider the case in which G_{xx} is diagonal. In this special case, the four-point term simplifies dramatically. In particular, we have

$$(K_{\text{GP}})_{\mu\nu} = \mathrm{var}[\phi(h^{(1)}_{\mu})]\delta_{\mu\nu} + \mathbb{E}[\phi(h^{(1)}_{\mu})]\mathbb{E}[\phi(h^{(1)}_{\nu})] \tag{17}$$

and

$$\begin{aligned}\underset{\mathcal{W}}{\mathrm{cov}}(K_{\mu\nu}, K_{\rho\lambda}) = \frac{1}{n_1}\Big(&\mathrm{var}[\phi(h^{(1)}_{\mu})^2]\delta_{\mu\nu}\delta_{\mu\rho}\delta_{\mu\lambda} \\ &+ \mathrm{var}[\phi(h^{(1)}_{\mu})]\,\mathrm{var}[\phi(h^{(1)}_{\nu})](1-\delta_{\mu\nu})(\delta_{\mu\rho}\delta_{\nu\lambda} + \delta_{\mu\lambda}\delta_{\nu\rho})\Big),\end{aligned} \tag{18}$$

which yields

$$\begin{aligned}\langle K_{\mu\nu}\rangle = (K_{\text{GP}})_{\mu\nu} &+ \frac{1}{2}\frac{n_2}{n_1}\sigma_2^2(\beta^2\Gamma^{-1}G_{yy}\Gamma^{-1} - \beta\Gamma^{-1})_{\mu\nu} \\ &\times\left[\mathrm{var}[\phi(h^{(1)}_{\mu})^2]\delta_{\mu\nu} + 2\,\mathrm{var}[\phi(h^{(1)}_{\mu})]\,\mathrm{var}[\phi(h^{(1)}_{\nu})](1-\delta_{\mu\nu})\right] + \mathcal{O}(n_1^{-2}).\end{aligned} \tag{19}$$

Moreover, applying the Sherman-Morrison formula [23], we have

$$\Gamma^{-1}_{\mu\nu} = \frac{\delta_{\mu\nu}}{\gamma_\mu} - \frac{1}{1 + \sum_{\rho=1}^{p}\mathbb{E}[\phi(h^{(1)}_{\rho})]^2/\gamma_\rho}\frac{\mathbb{E}[\phi(h^{(1)}_{\mu})]}{\gamma_\mu}\frac{\mathbb{E}[\phi(h^{(1)}_{\nu})]}{\gamma_\nu}, \tag{20}$$

where we have defined the vector $\gamma_\mu \equiv 1 + \beta\sigma_2^2\,\mathrm{var}[\phi(h^{(1)}_{\mu})]$ for brevity.

These results reveal an interesting distinction between the behavior of activation functions that yield $\mathbb{E}\phi(h) = 0$ and those that yield $\mathbb{E}\phi(h) \neq 0$. In particular, for those that have $\mathbb{E}\phi(h) = 0$, the NNGP kernel is diagonal, and a given element of the NLO correction depends only on the corresponding element of G_{yy}. However, if $\mathbb{E}\phi(h) \neq 0$, then the NNGP kernel includes a rank-1 component, and each element of the NLO correction depends on all elements of G_{yy}. Moreover, this means that the case in which G_{xx} is diagonal is qualitatively distinct from the case in which there is only a single training input.

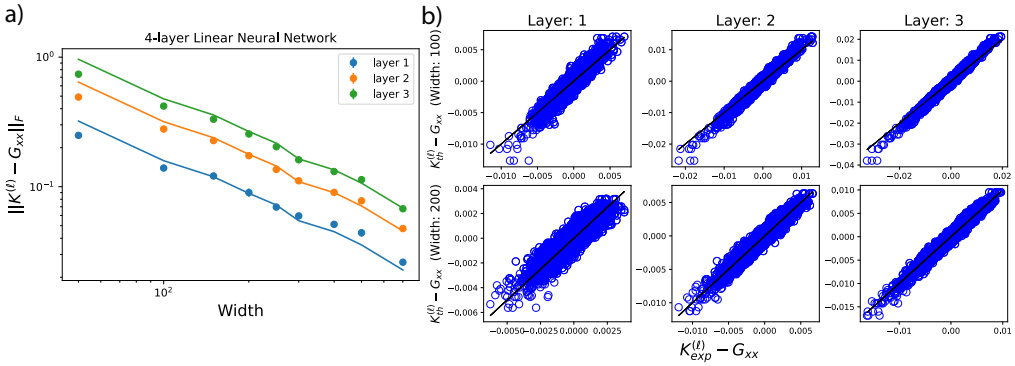

Figure 1: 3-hidden-layer neural network with linear activations trained via Langevin sampling on an artificial dataset (see Appendix J). (a) The Frobenius norm of the deviation of the empirical average kernel of each layer from its GP value (in this case, simply G_{xx}) for varying widths. We see perfect match with theoretical predictions, which are shown as solid lines. We obtain the predicted $1/n$ decay with increasing width and the linear scaling with the depth. Error bars represent the standard deviation of the mean over three realizations but are smaller than the markers. (b) Scatter plot of individual elements of the experimental (ordinate) and theoretical (abscissa) kernels. For low widths a slight deviation is visible between experiment and theory, while for larger widths the agreement is better.

5 Numerical experiments

In this section, we numerically probe the predictions of Conjecture 1, starting with simulations of deep linear networks. We devise an artificial task in which $n_0 = 10$-dimensional inputs $\mathbf{x}_\mu$ are chosen randomly from a standard Gaussian distribution and are randomly assigned to $n_d = 10$ dimensional one-hot labels $\mathbf{x}_\mu$. This generative model yields structured output correlations. Then, the weights of a depth-d linear network are trained using a discretization of the Langevin dynamics in (5). We provide a detailed discussion of our numerical methods in Appendix J [26]. As illustrated in Figure 1, we find excellent agreement with our theory for widths as low as 150.

Finally, we study how narrow bottlenecks affect representation learning in a more realistic nonlinear network. We train a network with three hidden layers and error function (erf) activations on a subset of the MNIST dataset [27]. Despite its analytical simplicity, erf is among the activation functions for which the covariance term in Conjecture 1 cannot be evaluated in closed form (see §4.2). However, it is easy to simulate numerically. Consistent with the predictions of our theory for linear networks, we find that introducing a narrow bottleneck leads to more representation learning in subsequent hidden layers, even if those layers are quite wide (Figure 2). Quantitatively, if one increases the width of the hidden layers between which the fixed-width bottleneck is sandwiched, the deviation of the first layer's kernel from its GP value decays roughly as $1/n$ with increasing width, while the deviations for the bottleneck and subsequent layers remain roughly constant. In contrast, the kernel deviations throughout a network with equal-width hidden layers decay roughly as $1/n$ (Figure 2). These observations are qualitatively consistent with the width-dependence of the linear network kernel (11). Precise characterization of nonlinear networks will be an interesting objective for future work.

6 Related work

Our work is closely related to several recent analytical studies of finite-width neural networks. First, Aitchison [9] argued that the kernel flexibility afforded by finite-width Bayesian networks can be advantageous. He derived a recurrence relation for the learned feature kernels in deep linear networks, which he solved in the limits of infinite width and few outputs, narrow width and many outputs, and infinite width and many outputs. As discussed in §4.1 and in Appendix E, our results on deep linear networks extend those of his work. Moreover, our numerical results support his suggestion that networks with narrow bottlenecks may learn interesting features.

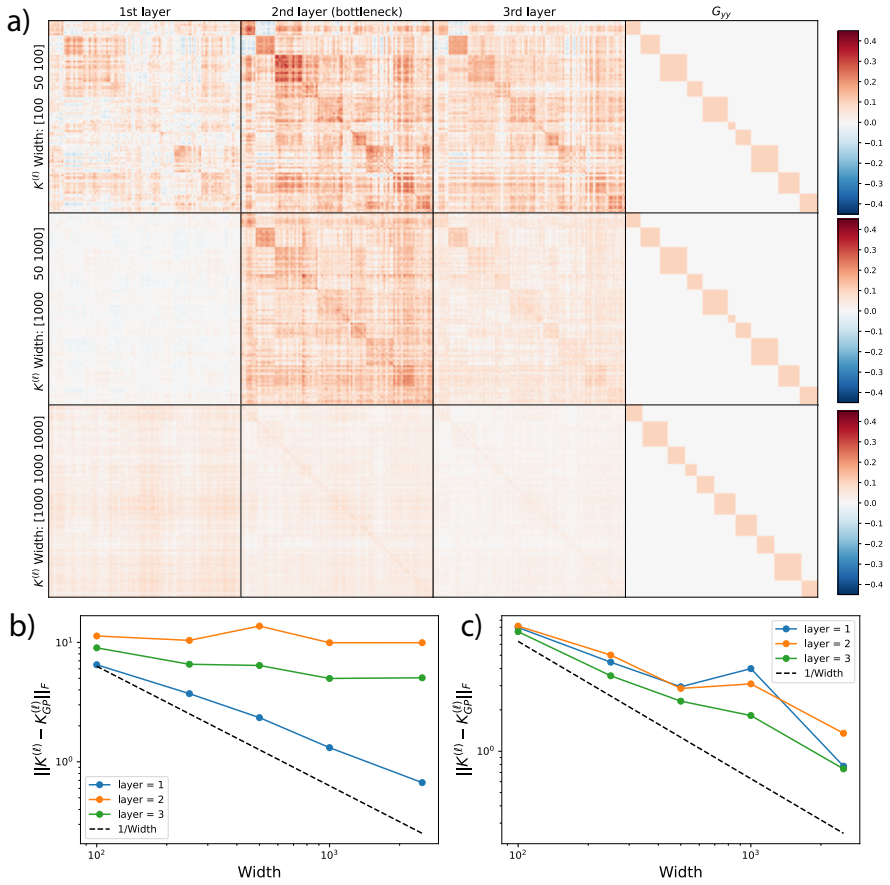

Figure 2: 3-hidden layer neural network with erf activations trained via Langevin sampling on MNIST (see Appendix J). (a) The empirical average kernels compared to the GP kernels for all layers with varying widths. Labels on the y-axes indicate the widths of each layer and the last column shows the output kernel. We observe that for networks with bottleneck layers, the deviation from $K_{\text{GP}}^{(\ell)}$ is largest at the bottleneck indicating representation learning; without a bottleneck deviations are considerably less (the last row). The off-diagonal elements in the 3rd layer for bottleneck networks acquire smaller corrections than the diagonal ones, indicating learning of output correlations G_{yy}. (b) Hidden layer kernel deviation from GP kernels as a function of width for bottleneck networks. While the first layer shows $1/n$ scaling, the bottleneck layer and the 3rd layer deviations stay almost constant. This behavior is predicted analytically for linear networks. (c) As in (b) for networks without a bottleneck. Consistent with our theory, all layers display $1/n$ decay.

Our approach and the asymptotic regime we consider mirror recent perturbative studies of finite-width Bayesian networks. As noted in §3 and Appendix B, we make use of the results of Yaida [11], who derived recurrence relations for the pertubative corrections to the cumulants of the finite-width prior for an MLP. However, Yaida did not attempt to study the statistics of learned features; the goal of his work was to establish a general framework for the study of finite-width corrections. Perturbative corrections to the network prior and posterior have also been studied by Halverson et al. [12] and Naveh et al. [14], respectively. Our work builds upon these studies by perturbatively characterizing the internal representations that are learned upon inference.

Li and Sompolinsky [15] considered deep linear networks in an alternative asymptotic regime in which the input dimensionality, hidden layer width, and number of training examples were all taken to infinity with fixed ratios, i.e., $n_1, n, p \to \infty$ with $n_1 = n_2 = \cdots = n_{d-1} = n$, $n_0/n \in (0, \infty)$, and $p/n \in (0, \infty)$. This asymptotic regime is standard in studies of neural networks from the standpoint of statistical physics [28], though most studies of wide Bayesian neural networks consider the fixed-p regime investigated here [3–14, 16–18, 22]. In Appendix E, we show that our result (12) for the low-temperature kernel can be recovered as the $p/n \downarrow 0$ limit of their result. As the dataset size p appears only implicitly in our approach, we leave the incorporation of corrections at large dataset size as an interesting objective for future work.

Finally, Zavatone-Veth and Pehlevan [16] have recently derived exact formulas for the prior densities over the outputs of finite Bayesian networks with linear or ReLU activation functions. They showed that the priors for a single training input have heavy tails that are not captured by leading perturbative corrections to the density. This leaves open the possibility that there could be non-perturbative corrections to the average kernels that are not captured by Conjecture 1. However, if the target outputs are in some sense sufficiently small and the likelihood sufficiently concentrated near the origin, these effects may be strongly suppressed in the posterior provided that the width is large. It will be interesting to investigate the possibility of non-perturbative effects on feature learning in future work.

7 Conclusions

In this paper, we have shown that the leading perturbative feature learning corrections to the NNGP kernels of wide Bayesian neural networks with linear readout and least-squares cost should be of a tightly constrained form. We demonstrate analytically and with numerical experiments that these results hold for MLPs, and conjecture that they should extend to more general network architectures that admit a well-defined NNGP limit.

Limitations. We emphasize that our perturbative argument for Conjecture 1 is not rigorous, and, as discussed above, does not account for possible non-perturbative effects. Moreover, it is non-quantitative: we have not obtained rigorous asymptotic bounds on the remainder for general network architectures. We leave rigorous proofs of the applicability of our results to more general architectures and of the smallness of the remainder as an objective for future work. Furthermore, we have considered only one possible asymptotic regime: that in which the width is taken to infinity with a finite training dataset and small output dimensionality. As discussed above in reference to the work of Aitchison [9] and Li and Sompolinsky [15], investigation of alternative limits in which output dimension, dataset size, depth, and hidden layer width are all taken to infinity with fixed ratios may be an interesting subject for future work.

References

[1] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT Press, Cambridge, MA, USA, 2016.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[3] Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

[4] Christopher KI Williams. Computing with infinite networks. *Advances in neural information processing systems*, pages 295–301, 1997.

[5] Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

[6] Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are Gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1g30j0qF7.

[7] Greg 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.

[8] Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

[9] Laurence Aitchison. Why bigger is not always better: on finite and infinite neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 156–164. PMLR, July 2020. URL http://proceedings.mlr.press/v119/aitchison20a.html.

[10] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020.

[11] Sho Yaida. Non-Gaussian processes and neural networks at finite widths. In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 165–192, Princeton University, Princeton, NJ, USA, July 2020. PMLR. URL http://proceedings.mlr.press/v107/yaida20a.html.

[12] James Halverson, Anindita Maiti, and Keegan Stoner. Neural networks and quantum field theory. *Machine Learning: Science and Technology*, 2021.

[13] Joseph M Antognini. Finite size corrections for neural network Gaussian processes. *arXiv preprint arXiv:1908.10030*, 2019.

[14] Gadi Naveh, Oded Ben-David, Haim Sompolinsky, and Zohar Ringel. Predicting the outputs of finite networks trained with noisy gradients. *arXiv preprint arXiv:2004.01190*, 2020.

[15] Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The back-propagating renormalization group. *arXiv preprint arXiv:2012.04030*, 2020.

[16] Jacob A Zavatone-Veth and Cengiz Pehlevan. Exact priors of finite neural networks. *arXiv preprint arXiv:2104.11734*, 2021.

[17] Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from Feynman diagrams. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gFvANKDS.

[18] Kyle Aitken and Guy Gur-Ari. On the asymptotics of wide networks with polynomial activations. *arXiv preprint arXiv:2006.06687*, 2020.

[19] Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10248–10259. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/wenzel20a.html.

[20] Vincent Fortuin, Adrià Garriga-Alonso, Florian Wenzel, Gunnar Rätsch, Richard Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. *arXiv preprint arXiv:2102.06571*, 2021.

[21] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are Bayesian neural network posteriors really like? *arXiv preprint arXiv:2104.14421*, 2021.

[22] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[23] Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

[24] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

[25] Gian-Carlo Wick. The evaluation of the collision matrix. *Physical Review*, 80(2):268, 1950.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[27] Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[28] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] As discussed in §3 and the Conclusion, the primary limitation of our work is that our main result is obtained through a non-rigorous perturbative argument. We therefore frame its most general form as a conjecture.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not anticipate that our work will have negative societal impacts as outlined in the ethics guidelines, as it is purely theoretical.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See §2 and the statement of Conjecture 1.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendices B, D, F, H, and I. As noted above, we frame our main result as a conjecture due to the fact that some of our methods are not fully rigorous.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code will be included as a Python notebook with the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All experimental details have been detailed in supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See details in Appendix J of the supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use the MNIST dataset [27], which we cite appropriately. We also use open-source libraries such as PyTorch [26], which we also cite.
   (b) Did you mention the license of the assets? [Yes] We mention licenses in Appendix J of the supplemental material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] Our work involves no new assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]