Bayesian inference with finitely wide neural networks

Chi-Ken Lu [*]
Department of Mathematics and Computer Science, Rutgers University, Newark, New Jersey 07102, USA

(Received 7 March 2023; accepted 10 July 2023; published 31 July 2023)

The analytic inference, e.g., predictive distribution being in closed form, may be an appealing benefit for machine learning practitioners when they treat wide neural networks as Gaussian process in a Bayesian setting. The realistic widths, however, are finite and cause weak deviation from the Gaussianity under which partial marginalization of random variables in a model is straightforward. On the basis of multivariate Edgeworth expansion, we propose a non-Gaussian distribution in differential form to model a finite set of outputs from a random neural network, and derive the corresponding marginal and conditional properties. Thus, we are able to derive the non-Gaussian posterior distribution in Bayesian regression task. In addition, in the bottlenecked deep neural networks, a weight space representation of a deep Gaussian process, the non-Gaussianity is investigated through the marginal kernel and the accompanying small parameters.

DOI: 10.1103/PhysRevE.108.014311

I. INTRODUCTION

Neal in his seminal work [1] pointed out that a shallow but infinitely wide random neural network is a Gaussian process (GP) [2] in statistical sense. Subsequent work [3,4] in interpreting neural network with specific nonlinear activation units as kernel machines was also inspired by such idea. More recent reports [5,6] further claimed the equivalence between GP and deep neural networks when each hidden layer in latter is of infinite width. Consequently, machine learning practitioners can perform Bayesian inference by treating a deep and wide neural network as a GP, and exploit the analytic marginal and conditional properties of multivariate Gaussian distribution. Otherwise, one needs to employ gradient-based learning and bootstrap sampling for obtaining predictive distribution [7]. In addition, the equivalence with GP also allows the theoretical methods, such as regression learning curve and PAC-Bayesian analysis [2] to shed light on generalization in neural network learning [8].

In reality, all neural networks have finite width. Therefore, the deviation from Gaussianity requires further quantitative account as practitioners may wonder the corrections to the predictive mean and variance in, for example, a regression task. Yaida [9] and colleagues [10] proposed a perturbative approach for computing the multivariate cumulants by direct application of Wick's contraction theorem. Moreover, the fourth cumulants are shown to be nonzero, scaled by the sum of reciprocal widths, $1/N_1 + 1/N_2 + \cdots + 1/N_{L-1}$, and signaling non-Gaussian aspect of the random processes representing finite-width deep neural networks with L hidden layers [9]. A quartic energy functional for a fixed set of network outputs is formulated field theoretically with which the corrections to the posterior mean and variance due to the *weak* non-Gaussianity are obtained [9].

Although Yaida's approach is appealing from a field theory perspective (see also Refs. [11–13]), the loss of elegant marginal property due to the presence of fourth power term in exponent is critical for analytic inference with finite-width networks. An alternative thinking is to modify the multivariate Gaussian distribution so that the new distribution can match the higher cumulants associated with the networks. In this paper, we wiall use the multivariate Edgeworth series expansion [14] to construct the non-Gaussian distribution for the network's outputs. In particular, we find that the differential representation of Edgeworth expansion greatly facilitates the derivation of marginal and conditional properties of the non-Gaussian distribution. Three main results are reported in this paper. First, the marginal property is intact, and the corrections to conditional mean and variance are derived. Second, with observed data, the non-Gaussian posterior distribution associated with an unobserved output is derived. Third, we derive the marginal covariance [15] of a bottlenecked deep neural network [16], which represents the deep Gaussian process [17] in weight space. It is worthwhile to note that some of the hidden layers in the bottlenecked network are narrow and *strong* non-Gaussianity may be induced.

The paper has the following organization. In Sec. II, we begin by reviewing the computational structure of a shallow cosine network, its equivalence to GP with a Gaussian kernel, and the emergence of non-Gaussianity due to finite width. The shallow network with random parameters in a Bayesian setting is a non-Gaussian prior over function. In Sec. III, the non-Gaussian prior is represented as a differential representation of Edgeworth expansion around Gaussian multivariate distribution, and its marginal and conditional properties are established. The closed-form corrections to the conditional mean and second moment are obtained. Application of the non-Gaussian prior in the Bayesian regression task is discussed in Sec. IV. Finally, Sec. V is devoted to investigating the combined effect of nonlinear activation, depth, and finite width on the non-Gaussian prior for a deep bottlenecked cosine network, which is followed by a discussion in Sec. VI.

[*]CL1178@rutgers.edu

II. WIDE FEED FORWARD NEURAL NETWORK

Let us start with the discussion of a random shallow feed forward network with cosine activation. In the infinite width limit, the network is statistically equivalent to a GP with a Gaussian kernel [18]. The goal is to see the emergence of nonzero fourth cumulant when the width is finite. Consider a single output network with N activation units, the real function value $z_\alpha \in \mathbb{R}$ with the greek subscript is an indexed random variable in association with its input $\mathbf{x}_\alpha \in \mathbb{R}^d$. Explicitly, the output-input of the network has the following relation,

$$z_\alpha = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} w_i \cos\left(\frac{\Omega_i \cdot \mathbf{x}_\alpha}{\sqrt{d}} + \phi_i\right), \tag{1}$$

where the weight variable w's are sampled from the Gaussian distribution $\mathcal{N}(0, 1)$, the scaling variables Ω's are sampled from $\mathcal{N}(0, I_d)$, and the phase variables ϕ's are sampled from uniform distribution $\mathcal{U}([0, 2\pi])$. The normalization factors $\sqrt{2/N}$ and $1/\sqrt{d}$ above follow the parametrization used in Ref. [19]. Because w is the zero mean, the first relevant statistical moment is the covariance, $k(\mathbf{x}_\alpha, \mathbf{x}_\beta)$, namely, the expectation of the product of two function values at two inputs,

$$\begin{aligned} k(\mathbf{x}_\alpha, \mathbf{x}_\beta) :=& \mathbb{E}[z_\alpha z_\beta] \\ =& \frac{1}{N} \sum_i \mathbb{E}\left\{ \cos\left[\frac{\Omega_i \cdot (\mathbf{x}_\alpha - \mathbf{x}_\beta)}{\sqrt{d}}\right] \right. \\ & \left. + \cos\left[\frac{\Omega_i \cdot (\mathbf{x}_\alpha + \mathbf{x}_\beta)}{\sqrt{d}} + 2\phi_i\right]\right\} \\ =& e^{-\frac{1}{2d}|\mathbf{x}_\alpha - \mathbf{x}_\beta|^2}. \end{aligned} \tag{2}$$

In the above, the independence between the random variables w's is used to arrive at the second equality. The third equality is due to the fact that Fourier transformation of Gaussian is Gaussian and the vanishing average of cosine with a uniformly random phase. The fourth moment can be computed in similar manners, but one will note that the product of two cosine terms containing the random phase can generate a nonzero contribution. Here, we focus on the fourth cumulant tensor as the signature of non-Gaussianity,

$$\begin{aligned} V_{\alpha\beta\gamma\delta} :=& \mathbb{E}[z_\alpha z_\beta z_\gamma z_\delta] - K_{\alpha\beta}K_{\gamma\delta} - K_{\alpha\gamma}K_{\beta\delta} - K_{\alpha\delta}K_{\beta\gamma} \\ =& \frac{1}{N^2} \sum_{ij} \mathbb{E}\left\{ \cos\left[\frac{\Omega_i \cdot (\mathbf{x}_\alpha + \mathbf{x}_\beta)}{\sqrt{d}} + 2\phi_i\right] \right. \\ & \left. \times \cos\left[\frac{\Omega_j \cdot (\mathbf{x}_\gamma + \mathbf{x}_\delta)}{\sqrt{d}} + 2\phi_j\right]\right\} + \text{sym. perm.} \\ =& \frac{1}{2N}\left(e^{-\frac{1}{2d}|\mathbf{x}_\alpha + \mathbf{x}_\beta - \mathbf{x}_\gamma - \mathbf{x}_\delta|^2} + e^{-\frac{1}{2d}|\mathbf{x}_\alpha + \mathbf{x}_\gamma - \mathbf{x}_\beta - \mathbf{x}_\delta|^2} \right. \\ & \left. + e^{-\frac{1}{2d}|\mathbf{x}_\alpha + \mathbf{x}_\delta - \mathbf{x}_\beta - \mathbf{x}_\gamma|^2}\right). \end{aligned} \tag{3}$$

Hence, the fourth cumulant tensor receives a contribution proportional to the reciprocal width $1/N$, and it is symmetric with respect to permutation of indices.

III. MULTIVARIATE EDGEWORTH EXPANSION AND NON-GAUSSIAN PRIOR

The fourth cumulant tensor in Eq. (3) signifies that the distribution over the network outputs is non-Gaussian. In other words, the random parameter w's, Ω's, and ϕ's from a prior distribution induce the non-Gaussian prior function distribution due to the finite width. The first interesting consequence of the nonzero fourth cumulant is that the prior distribution over a single output, say z_α, receives a correction term deviating from Gaussian [20],

$$q(z_\alpha) = \mathcal{N}\left(z_\alpha|0, \sigma_0^2\right)\left[1 + \frac{V_{\alpha\alpha\alpha\alpha}}{24}\left(3 - 6\frac{z_\alpha^2}{\sigma_0^2} + \frac{z_\alpha^4}{\sigma_0^4}\right)\right]. \tag{4}$$

The variance $\sigma_0^2 := k(z_\alpha, z_\alpha)$ and the new distribution q has matched moments, i.e., $\mathbb{E}_q[1] = 1$, $\mathbb{E}_q[z_\alpha^2] = \sigma_0^2$, and $\mathbb{E}_q[z_\alpha^4] = 3\sigma_0^4 + V_{\alpha\alpha\alpha\alpha}$. Although the work [20] derived the above non-Gaussian distribution from a renormalization-group perspective [21], such expansion with Hermite polynomials around Gaussian distribution has been known as Edgeworth expansion in statistics literature [20,22,23]. When the width $N \to \infty$, the fourth cumulant V vanishes, and for notational convenience, we denote the limiting distribution by q_∞, which is Gaussian.

A. Joint distribution

As we are mainly interested in the prior distribution over a finite set of function values $\{z_1, z_2, \ldots\}$, the objective here is to construct the non-Gaussian joint distribution as a prior in Bayesian learning. The work [14] suggested the construction of multivariate Edgeworth expansion by replacing the univariate Gaussian $\mathcal{N}(z|0, \sigma_0^2)$ with multivariate one $\mathcal{N}(\mathbf{z}|0, K)$ and the Hermite polynomials with contracted tensor terms containing the inverse covariance matrix $[K^{-1}]$ and vector of function values $\mathbf{z}$ (also see Ref. [24] for an application in cosmology). Please see Appendix A for the expression.

An apparent difficulty with such representation is the demonstration of consistency after marginalization, i.e., $\int dz_1 q(z_1, z_2, z_3, \ldots) = q(z_2, z_3, \ldots)$, which is of critical importance for deriving conditional distribution and subsequent Bayesian learning. Inspired by the fact that the Hermite polynomials are obtained by taking derivative of Gaussian, we can rewrite the multivariate Edgeworth expansion in the following differential form:

$$q(\mathbf{z}) = \left(1 + \frac{V_{ijml}}{24} \partial_{z_i}\partial_{z_j}\partial_{z_m}\partial_{z_l}\right)\mathcal{N}(\mathbf{z}|0, K), \tag{5}$$

where the notation of summation by repeated indices is employed. An advantage of such representation is that the adjoint property of derivative, $\partial_z^\dagger = -\partial_z$, can be exploited in the integration of functions, which vanish at infinity. Hence, it is easy to see the distribution q is normalized, and the second and fourth moments match, for instance,

$$\begin{aligned} \mathbb{E}_q[z_1 z_2] &= K_{12}, \\ \mathbb{E}_q[z_1 z_2 z_3 z_4] &= K_{12}K_{34} + K_{13}K_{24} + K_{14}K_{23} + V_{1234}. \end{aligned} \tag{6}$$

The above non-Gaussian prior can be viewed as a perturbative extension of multivariate Gaussian. In the setting of Bayesian regression, the marginal and conditional properties of the Gaussian are appealing when conjugate likelihood functions are employed. In the following, we will examine the effects of the perturbation in differential form on these properties.

B. Marginal consistency and conditional statistics

We first show that the marginal consistency is intact with the Edgeworth expansion in differential form. Namely, without loss of generality, it suffices to show that marginalization of z_1 of the joint distribution $q(z_1, z_2, z_3, \ldots)$,

$$\begin{aligned}\int dz_1 q(z_1, \tilde{\mathbf{z}}) &= \mathcal{N}(\tilde{\mathbf{z}}|0, \tilde{K}) + v_{\tilde{i}\tilde{j}\tilde{m}\tilde{l}}\partial_{z_{\tilde{i}}}\partial_{z_{\tilde{j}}}\partial_{z_{\tilde{m}}}\partial_{z_{\tilde{l}}} \\ &\quad\times \int dz_1 \mathcal{N}(z_1, \tilde{\mathbf{z}}|0, K) \\ &= \big(1 + v_{\tilde{i}\tilde{j}\tilde{m}\tilde{l}}\partial_{z_{\tilde{i}}}\partial_{z_{\tilde{j}}}\partial_{z_{\tilde{m}}}\partial_{z_{\tilde{l}}}\big)\mathcal{N}(\tilde{\mathbf{z}}|0, \tilde{K}) \\ &= q(\tilde{\mathbf{z}}),\end{aligned} \tag{7}$$

indeed reproduces the joint distribution for $z_2, z_3, \ldots$ in the original form Eq. (5) as the smaller covariance matrix $\tilde{K}$ is the submatrix of K excluding the first row and column. For simplicity, we use v to denote $V/24$. In deriving above, we have employed the fact that $\int dz_1 (\partial_{z_1})^s \mathcal{N}(z_1, \tilde{\mathbf{z}}) = 0$ for any integer power $s \geqslant 1$, meaning that the indices i, j, m, l will exclude any contribution associated with z_1. Thus, the tilded indices $\tilde{i} \in \{2\text{–}4, \ldots\}$, and they can be factored out of the integral.

Here, we wish to stress that the above marginal property of non-Gaussian prior may seems straightforward with the differential representation of Edgeworth expansion, but it does not seem promising to reach the same conclusion if the marginalization is carried out with the explicit representation in Appendix A. Nevertheless, one can easily write the conditional distribution $q(z_1|\tilde{\mathbf{z}}) = q(z_1, \tilde{\mathbf{z}})/q(\tilde{\mathbf{z}})$. More interestingly, the conditional mean is useful for prediction with noiseless observation of data, and shed light on the effect of non-Gaussianity due to the finite width. The details of deriving the following conditional mean:

$$\begin{aligned}\mathbb{E}[z_1|\tilde{\mathbf{z}}] &= \int dz_1 z_1 \frac{q(z_1, \tilde{\mathbf{z}})}{q(\tilde{\mathbf{z}})} \\ &= \mu(\tilde{\mathbf{z}}) + \frac{1}{q(\tilde{\mathbf{z}})} A_{\tilde{i}\tilde{j}\tilde{m}}\partial_{z_{\tilde{i}}}\partial_{z_{\tilde{j}}}\partial_{z_{\tilde{m}}}\mathcal{N}(\tilde{\mathbf{z}}|0, \tilde{K}),\end{aligned} \tag{8}$$

with the finite-width correction term proportional to the third-order tensor,

$$A_{\tilde{i}\tilde{j}\tilde{m}} = \frac{V_{\tilde{i}\tilde{j}\tilde{m}\tilde{l}}}{6}[\tilde{K}^{-1}\mathbf{k}]_{\tilde{l}} - \frac{V_{1\tilde{i}\tilde{j}\tilde{m}}}{6}, \tag{9}$$

can be found in Appendix B. In the GP limit, the conditional mean becomes the well-known result $\mu(\tilde{\mathbf{z}}) = \mathbf{k}^t\tilde{K}^{-1}\tilde{\mathbf{z}}$, as the fourth cumulant V vanishes in the large N limit. We also remind the readers that the tilded symbols are associated with the conditioned outputs $z_{2,3,\ldots}$. In a similar manner, we can also show the conditional second moment,

$$\begin{aligned}\mathbb{E}\big[z_1^2|\tilde{\mathbf{z}}\big] &= \int dz_1 z_1^2 \frac{q(z_1, \tilde{\mathbf{z}})}{q(\tilde{\mathbf{z}})} \\ &= \mu^2(\tilde{\mathbf{z}}) + \sigma^2 + \frac{1}{q(\tilde{\mathbf{z}})}\big[2\mu(\tilde{\mathbf{z}})A_{\tilde{i}\tilde{j}\tilde{m}}\partial_{z_{\tilde{i}}}\partial_{z_{\tilde{j}}}\partial_{z_{\tilde{m}}} + B_{\tilde{i}\tilde{j}}\partial_{z_{\tilde{i}}}\partial_{z_{\tilde{j}}}\big] \\ &\quad\times \mathcal{N}(\tilde{\mathbf{z}}|0, \tilde{K}),\end{aligned} \tag{10}$$

with the second-order tensors denoting

$$B_{\tilde{i}\tilde{j}} = \frac{V_{\tilde{i}\tilde{j}\tilde{m}\tilde{l}}}{2}[\tilde{K}^{-1}\mathbf{k}]_{\tilde{m}}[\tilde{K}^{-1}\mathbf{k}]_{\tilde{l}} - V_{1\tilde{i}\tilde{j}\tilde{m}}[\tilde{K}^{-1}\mathbf{k}]_{\tilde{m}} + \frac{V_{11\tilde{i}\tilde{j}}}{2}. \tag{11}$$

The conditional second moment coincides with that in the GP limit, $\sigma^2 = K_{11} - \mathbf{k}^t\tilde{K}^{-1}\mathbf{k}$. The details of derivation can also be found in Appendix B. It is worthwhile to note that the conditional variance $\mathbb{E}_q[z_1^2|\tilde{\mathbf{z}}] - (\mathbb{E}_q[z_1|\tilde{\mathbf{z}}])^2$ can include contributions related to function values $\tilde{\mathbf{z}}$ through the nonzero A and B terms in above expressions. The lack of such dependence in the GP limit is in fact a shortcoming for modeling, which has motivated the study of a non-Gaussian prior in the machine learning community (for example, the Student-t process in Ref. [25]).

C. An example: bivariate distribution and prediction

Now let us pause to consider a simple example where a shallow network defined in Eq. (1) with width N is used to model two input-output pairs $(\mathbf{x}_1, z_1)$ and $(\mathbf{x}_2, z_2)$. In the end, we will present the predictive mean and variance for z_1 conditioning on z_2. The following notations are used for simplification: the covariance $c := \exp(-|\mathbf{x}_1 - \mathbf{x}_2|^2/2d)$, the 2-×-2 covariance matrix $\Sigma = \begin{pmatrix}1 & c\\ c & 1\end{pmatrix}$ for the bivariate prior, and $z_{21} := (z_2 - cz_1)/(\sqrt{1-c^2})$. Besides, the relations of derivatives of Gaussian in terms of Hermite polynomial will be useful,

$$\begin{aligned}\partial_{z_1}^n\partial_{z_2}^m\mathcal{N}(z_1, z_2|0, \Sigma) &\propto \partial_{z_1}^n\big[\mathcal{N}(z_1|0, 1)\partial_{z_2}^m\mathcal{N}(z_{21}|0, 1)\big] \\ &\propto (-1)^m\frac{\mathcal{N}(z_2|0, 1)}{\sqrt{(1-c^2)^m}} \\ &\quad\times \partial_{z_1}^n[H_m(z_{21})\mathcal{N}(z_{12}|0, 1)],\end{aligned} \tag{12}$$

where we have used $\mathcal{N}(z_1, z_2|0, \Sigma) \propto \mathcal{N}(z_1|0, 1)\mathcal{N}(z_{21}|0, 1)$ up to some irrelevant constant. The probabilist's Hermite polynomials are defined as $H_n(z) = (-1)^n[\mathcal{N}(z|0, 1)]^{-1} d_z^n\mathcal{N}(z|0, 1)$ [22]. Consequently, following the previous discussion, the non-Gaussian bivariate distribution is shown to be

$$\begin{aligned}q(z_1, z_2) = \mathcal{N}(z_1, z_2|0, \Sigma)\bigg\{&1 + \frac{\gamma^4}{16N}[H_4(z_{21}) + H_4(z_{12})] \\ &+ \frac{c\gamma^4}{4N}[3cH_2(z_{21}) + H_3(z_{21})H_1(z_{12}) + z_{12} \leftrightarrow z_{21}] \\ &+ \frac{(c^4+2)\gamma^4}{8N}[2c^2 + 4cH_1(z_{21})H_1(z_{12}) \\ &+ H_2(z_{21})H_2(z_{12})]\bigg\},\end{aligned} \tag{13}$$

and the symbol $\gamma = 1/\sqrt{1-c^2}$ is used to simplify the expression. Besides, the fourth cumulant for the cosine network, $V_{1111} = 3/2N$, $V_{1112} = 3c/2N$, and $V_{1122} = (c^4 + 2)/2N$, have been plugged in above. As for the conditional distribution, e.g., $q(z_1|z_2) = q(z_1, z_2)/q(z_2)$ for the bivariate case, one can simply divide the above with $q(z_2) = \mathcal{N}(z_2|0, 1)[1 + H_4(z_2)/16N]$, which will yield a conditional Gaussian $\mathcal{N}(z_1|z_2)$ multiplied by a factor representing the finite-width effect.

We can directly apply Eq. (8) to obtain the conditional mean in the simple example where the tilded indices there only apply to z_2. Surprisingly, the conditional mean for such noiseless case is

$$\mathbb{E}[z_1|z_2] = cz_2, \tag{14}$$

coinciding with that in the GP limit because the *accidental* cancellation in the third-order tensor $A_{222} \propto (cV_{2222} - V_{1222})$. We will show that the cancellation does not occur in the noisy (the next section) and more general cases. Application of Eq. (10) together with the above conditional mean results in the following conditional variance:

$$\mathrm{Var}[z_1|z_2] = (1-c^2)\left[1 + \frac{4(2-c^2)}{16N + H_4(z_2)} H_2(z_2)\right], \tag{15}$$

which consists of the corresponding variance in GP limit along with a term depending on z_2^2.

IV. BAYESIAN REGRESSION WITH NON-GAUSSIAN PRIOR

Having established the marginal and conditional properties of the non-Gaussian distribution in Eq. (5), we now continue investigating the posterior distribution over the unseen function value z_* associated with input $\mathbf{x}_*$ when the noisy observations $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots\}$ are known. According to Bayes's rule, the objective distribution is the posterior distribution defined as

$$q(z_*|\mathcal{D}) = \frac{q(z_*, \mathcal{D})}{q(\mathcal{D})} = \frac{\int d\mathbf{z}\, q(z_*, \mathbf{z})\mathcal{N}(\mathbf{y}|\mathbf{z}, \Lambda)}{\int d\mathbf{z}\, q(\mathbf{z})\mathcal{N}(\mathbf{y}|\mathbf{z}, \Lambda)}, \tag{16}$$

with Λ denoting a diagonal matrix representing the noise in Gaussian likelihood. When the q's in the numerator and denominator are both Gaussian, which corresponds to the infinitely wide network, the limiting distribution reads

$$\begin{aligned} q_\infty(z_*|\mathcal{D}) &= \int d\mathbf{z}\, \mathcal{N}(z_*|\mathbf{k}_*^t K^{-1}\mathbf{z}, k_{**} - \mathbf{k}_*^t K^{-1}\mathbf{k}_*) \\ &\quad \times \mathcal{N}[\mathbf{z}|K(K+\Lambda)^{-1}\mathbf{y}, \Lambda K(K+\Lambda)^{-1}] \\ &= \mathcal{N}[z_*|\mathbf{k}_*^t(K+\Lambda)^{-1}\mathbf{y}, k_{**} - \mathbf{k}_*^t(K+\Lambda)^{-1}\mathbf{k}_*]. \end{aligned}$$

In the above first equality, the first and second Gaussian distributions on the right hand side represent the conditional density $q_\infty(z_*|\mathbf{z})$ and the posterior $q_\infty(\mathbf{z}|\mathcal{D})$, respectively.

As for the case of finite width, the involved Gaussian likelihood can be dealt with by continuing application of the properties of adjoint differential operator as well as derivative of independent Gaussians. The details of derivation of evidence $q(\mathcal{D})$ can be seen in Appendix C, and the expression in terms of derivatives with respect to the observed y's reads

$$\begin{aligned} q(\mathcal{D}) &= \int d\mathbf{z}\, \mathcal{N}(\mathbf{y}|\mathbf{z}, \Lambda) q(\mathbf{z}) \\ &= \left(1 + v_{ijml}\partial_{y_i}\partial_{y_j}\partial_{y_m}\partial_{y_l}\right)\mathcal{N}(\mathbf{y}|0, K+\Lambda). \end{aligned} \tag{17}$$

Despite its simple form, the evidence term cannot be further manipulated, such as in the infinite-width case. The details of derivation of the posterior distribution for output z_* at the unseen input,

$$\begin{aligned} q(z_*|\mathcal{D}) &= \frac{\int d\mathbf{z}\, \mathcal{N}(\mathbf{y}|\mathbf{z}, \Lambda) q(z_*, \mathbf{z})}{q(\mathcal{D})} = (1 + v_{\hat{i}\hat{j}\hat{m}\hat{l}}\partial_{\hat{i}\hat{j}\hat{m}\hat{l}}) \\ &\quad \times N\left[\begin{pmatrix} z_* \\ \mathbf{y} \end{pmatrix}\middle| 0, \begin{pmatrix} 1 & \mathbf{k}_*^t \\ \mathbf{k}_* & K+\Lambda \end{pmatrix}\right] \Big/ q(\mathcal{D}), \end{aligned} \tag{18}$$

can also be found in Appendix C. Here, the notations can be understood as follows: the hatted indices, $\hat{i}$ for instance, additionally include the symbol $*$ associated with test function value z_*, and the derivative refers to $\partial_* := \partial_{z_*}$ (with respect to function value) and $\partial_i := \partial_{y_i}$ (with respect to the observed values). From the outset, the differential operators are to act on the latent function outputs z's in the prior, but, due to the independent Gaussian likelihood and the adjoint property, it becomes equivalent for them to be pulled out and act on the observed values. Equation (18) has the same form with its noiseless counterpart, i.e., $q(z_*|\mathbf{z})$, and we can apply the expressions in Eqs. (8) and (10) to obtain predictive mean and variance (the replacement of $\tilde{K}$ by $K + \Lambda$ is needed).

We conclude this section with considering the same simple regression example but with noisy observation in y_2 at input $\mathbf{x}_2$, and we wish to predict the unobserved function value z_1. To stress the role of noise parameter σ_n, we only show the predictive mean here,

$$\mathbb{E}[z_1|y_2] = c\alpha_n^2 y_2 - \frac{c\sigma_n^2\alpha_n^5 H_3(\alpha_n y_2)}{4N + \alpha_n^4 H_4(\alpha_n y_2)}, \tag{19}$$

with $\alpha_n = 1/\sqrt{1+\sigma_n^2}$. The correction term vanishes as the noise parameter does and is odd with respect to the sign change in y_2.

V. DEEP BOTTLENECKED NETWORK

Up to this point, we have focused on the emergence of non-Gaussianity in prior function distribution in wide and shallow network, as well as how the non-Gaussian prior affects the inference in a regression task. Our approach can be extended to the deep and wide neural networks as the corresponding prior distributions approach GP [5,6]. However, such wide-width assumption leading to *weak* non-Gaussianity is not valid for the bottlenecked networks in which the hidden layers are alternatively wide and narrow [16,26].

Let us consider a two-layer bottlenecked feed forward network modeling the hierarchical mapping $\mathbf{x} \in \mathbb{R}^d \mapsto \mathbf{z}^{(1)} \in \mathbb{R}^H \mapsto z^{(2)} \in \mathbb{R}$. The hierarchy consists of the following computations:

$$z_{i,\alpha}^{(1)} = \sqrt{\frac{2}{N_1}} \sum_{j=1}^{N_1} w_{ij}^{(1)} \cos\left(\frac{\Omega_j^{(1)} \cdot \mathbf{x}_\alpha}{\sqrt{d}} + \phi_j^{(1)}\right),$$
$$z_\alpha^{(2)} = \sqrt{\frac{2}{N_2}} \sum_{j=1}^{N_2} w_j^{(2)} \cos\left(\frac{\Omega_j^{(2)} \cdot \mathbf{z}_\alpha^{(1)}}{\sqrt{H}} + \phi_j^{(2)}\right), \quad (20)$$

where the parameter w's, Ω's, and ϕ's share the same prior distribution with their counterparts in the shallow network. The widths of hidden layers, N_1 and N_2, are large but the hidden output dimension H is not. In the limit $N_{1,2} \to \infty$, whereas, H remains finite, the two-layer cosine bottlenecked network corresponds to the prior of deep Gaussian process [17,27,28] with the Gaussian kernel, which is a flexible and expressive function prior due to its compositional nature. As its name suggests, the conditional prior $z^{(2)}|\mathbf{z}^{(1)}$ is a GP and so is each component $z_i^{(1)}|\mathbf{x}$ in the hidden output. However, the marginal distribution for $z^{(2)}|\mathbf{x}$ is non-Gaussian as Ref. [15] showed that the fourth cumulant is positive, i.e., a heavy-tailed distribution [29].

Here, we are interested in tracking how the small parameters, $1/N_1$ and $1/N_2$, and the bottleneck parameter $1/H$ enters the second moment. For a deep and finitely wide linear network, the prior is non-Gaussian [9,30], but the second moment does not receive a correction due to the finite width [9,10]. Thus, the effects of nonlinear activation on higher statistical moments are interesting (also see a recent work in Ref. [31]). For the deep cosine network in Eq. (20), we can first consider the wide limit, $N_{1,2} \to \infty$, but the bottleneck width H remains finite. Following the result in Eq. (2), the covariance for the deep model can be computed as the following:

$$k_\infty^{(2)}(\mathbf{x}_\alpha, \mathbf{x}_\beta) = \mathbb{E}_{q_\infty}\left\{\exp\left(-\frac{|\mathbf{z}_\alpha^{(1)} - \mathbf{z}_\beta^{(1)}|^2}{2H}\right)\right\}$$
$$= \left[\int dz_\alpha dz_\beta\, e^{-\frac{(z_\alpha - z_\beta)^2}{2H}} q_\infty(z_\alpha, z_\beta)\right]^H$$
$$= \left[1 + \frac{1 - \exp\left(-\frac{|\mathbf{x}_\alpha - \mathbf{x}_\beta|^2}{2d}\right)}{H/2}\right]^{-H/2}. \quad (21)$$

Note that the decomposition $\mathcal{N}(z_1, z_2|0, K) = \mathcal{N}(z_1 + z_2|0, \sigma_+^2)\mathcal{N}(z_1 - z_2|0, \sigma_-^2)$ [32] has been used in derivation with the variances $\sigma_\pm^2 = 2(K_{11} \pm K_{12})$. We want to stress that the above kernel is exact for any H. For a very narrow case, i.e., $H = 1$, the above result coincides with that in Ref. [15], whereas, in the very wide bottleneck case $H \gg 1$, it can be shown that

$$k_\infty^{(2)}(\mathbf{x}_\alpha, \mathbf{x}_\beta) \approx \exp[k(\mathbf{x}_\alpha, \mathbf{x}_\beta) - 1]\left\{1 + \frac{1}{H}[k(\mathbf{x}_\alpha, \mathbf{x}_\beta) - 1]^2\right\}, \quad (22)$$

where the covariance k in the shallow network is given in (2). In addition, the exponential of shallow kernel in above corresponds to $H \to \infty$, which is the same as the deep kernel in Ref. [33]. Indeed, as suggested in Ref. [10], the appearance of reciprocal width $1/H$ in the correction term is in association with the weak non-Gaussianity even though the other widths $N_{1,2}$ are infinite.

Next, we will compute the covariance for the case where all widths are finite. In fact, the outer width N_2 does not enter the kernel, and we only need a large but finite N_1. The computation is similar, and one only needs to replace the limiting joint distribution $q_\infty(z_\alpha^{(1)}, z_\beta^{(1)})$ with the non-Gaussian q. The details of derivation using the property derivative of Gaussian can be seen in Appendix D. The kernel of the two-layer bottlenecked cosine network with $N_{1,2} < \infty$ reads

$$k^{(2)}(\mathbf{x}_\alpha, \mathbf{x}_\beta) = \left[1 + \frac{1 - k(\mathbf{x}_\alpha, \mathbf{x}_\beta)}{H/2}\right]^{-H/2} (1 + \epsilon)^H, \quad (23)$$

with the correction due to finite N_1,

$$\epsilon = \frac{V_{\alpha\alpha\alpha\alpha} + V_{\beta\beta\beta\beta} - 4V_{\alpha\beta\beta\beta} - 4V_{\beta\alpha\alpha\alpha} + 6V_{\alpha\alpha\beta\beta}}{24} \times \left[\frac{3}{H^2} - \frac{6\sigma_-^2}{H^3(1 + \sigma_-^2/H)} + \frac{3\sigma_-^4}{H^4(1 + \sigma_-^2/H)^2}\right], \quad (24)$$

with $\sigma_-^2 = 2(1 - k)$ is used for easing the notation. Again, the above result is for general H. For the wide bottleneck limit, one can show that these small parameters enter the kernel with the correction of $O(\frac{1}{H})$ followed by $O(\frac{1}{N_1 H})$. Hence, the deep model as a function prior serves as a GP with random kernel whose mean value converges to Eq. (22) when all widths $N_{1,2}$ and H approach infinity. However, the role of H is different from N_1 from our perturbative analysis, and $1/N_1$ does not appear alone in the small parameter. The influence of finite width on the kernel of deep and nonlinear and convolutional models is also studied in Refs. [34–36].

The fourth moment can be computed in a similar manner. The closed form for $N_{1,2} \to \infty$ can be found in Ref. [15]. For the finite width case, the fourth cumulant $V_{\alpha\beta\gamma\delta}^{(2)}$ becomes the Hth power of the permutational symmetrization of $\frac{1}{N_2}\mathbb{E}_q\{\exp[-[z_\alpha^{(1)} + z_\beta^{(1)} - z_\gamma^{(1)} - z_\delta^{(1)}]^2/2H)\}$, the small parameters of which will consist of $1/N_2$, $1/(N_2 H)$, and $1/(N_2 H N_1)$. The analysis of scaling with respect to the reciprocal width is more complex than the deep linear network.

VI. DISCUSSIONS

In essence, the finite width in random neural networks induces nonzero variance of kernel since the fourth cumulant is not zero. From function space perspective, the shallow networks with finitely large width can be regarded as a GP but the kernel itself is a random variable too, so the learned data representation is a distribution over kernels. Therefore, the neural network with very narrow layers may not have enough expressive power for learning, whereas, the network with very wide layers is not flexible enough because it is equivalent to GP learning with one fixed kernel. It was recently suggested in Ref. [37] through studying the partition function [8,38] of finite deep network that Student-t process [25] may suitably represent finite-width network, which offers an alternative description than this paper. The reports in Refs. [28,34] observed the degradation of performance for Bayesian deep neural network when expanding the widths, whereas, the study in Ref. [39] suggested otherwise.

Taking the wide limit with the perturbative approach is appealing because of the analytic and elegant expressions for inference, which may shed light on future investigation of alternative algorithm for classification [40], study of average test error in regression [41,42] and classification [43] tasks using finite-width networks. In this paper, the conditional statistics and perturbed posterior distribution over the unseen output using the non-Gaussian prior are obtained with help of the differential representation of multivariate Edgeworth expansion. In parallel with the equivalence between GP prior and the random wide neural network, the investigation of neural tangent kernel [19,44,45] is important for understanding the learning dynamics during the optimization and the relation with Bayesian neural network learning [46,47].

In a broader context, the techniques for obtaining the non-Gaussian conditional and marginal distributions might be reminiscent of those in computing the transition amplitude between simple harmonic states driven by an external field. Although the non-Gaussianity discussed here originates from the random parameters of the neural network, the non-Gaussian modeling of correlation among cosmic microwave background data might suggest some novel physics beyond the standard model [48].

ACKNOWLEDGMENTS

Correspondences with J. Zavatone-Veth and P. Rotondo are acknowledged.

APPENDIX A: EXPRESSION FOR NON-GAUSSIAN PRIOR DISTRIBUTION

After explicitly taking the derivatives in Eq. (5), it is straightforward to show the expressions for the multivariate non-Gaussian distribution reads

$$q(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^N|K|}} e^{-\frac{1}{2}\mathbf{z}^t K^{-1}\mathbf{z}} \bigg[1 + \frac{V_{\alpha\beta\gamma\delta}}{24} \big(\chi^{(0)}_{\alpha\beta\gamma\delta} - \chi^{(1)}_{\alpha\beta\gamma\delta\mu\nu} z_\mu z_\nu + \chi^{(2)}_{\alpha\beta\gamma\delta\mu\nu\kappa\eta} z_\mu z_\nu z_\kappa z_\eta \big) \bigg], \tag{A1}$$

in which the first perturbation term, a scalar term, is the contraction between the fourth cumulant tensor $V_{\alpha\beta\gamma\delta} \propto 1/N_1$ and the following tensor:

$$\chi^{(0)}_{\alpha\beta\gamma\delta} = (K^{-1})_{\alpha\beta}(K^{-1})_{\gamma\delta} + (K^{-1})_{\alpha\gamma}(K^{-1})_{\beta\delta} + (K^{-1})_{\alpha\delta}(K^{-1})_{\beta\gamma}. \tag{A2}$$

The next contribution results from contraction between the fourth cumulant, the following rank-6 tensor:

$$\begin{aligned}\chi^{(1)}_{\alpha\beta\gamma\delta\mu\nu} &= (K^{-1})_{\alpha\mu}(K^{-1})_{\nu\beta}(K^{-1})_{\gamma\delta} + (K^{-1})_{\alpha\beta}(K^{-1})_{\gamma\mu}(K^{-1})_{\nu\delta} \\ &+ (K^{-1})_{\alpha\mu}(K^{-1})_{\nu\gamma}(K^{-1})_{\beta\delta} + (K^{-1})_{\alpha\gamma}(K^{-1})_{\beta\mu}(K^{-1})_{\nu\delta} \\ &+ (K^{-1})_{\alpha\mu}(K^{-1})_{\nu\delta}(K^{-1})_{\beta\gamma} + (K^{-1})_{\alpha\delta}(K^{-1})_{\beta\mu}(K^{-1})_{\nu\gamma},\end{aligned} \tag{A3}$$

and the rank-2 tensor $z_\mu z_\nu$. The last contribution contains the rank-8 tensor,

$$\chi^{(2)}_{\alpha\beta\gamma\delta\mu\nu\kappa\eta} = (K^{-1})_{\alpha\mu}(K^{-1})_{\beta\nu}(K^{-1})_{\gamma\kappa}(K^{-1})_{\delta\eta}\,, \tag{A4}$$

the rank-4 tensor $z_\mu z_\nu z_\kappa z_\eta$, and the fourth cumulant. It becomes extensively tedious, if possible, proving the marginal and conditional properties with this representation.

APPENDIX B: PROOFS OF CONDITIONAL MEAN EQ. (8) AND SECOND MOMENT EQ. (10)

By plugging the corresponding expression for non-Gaussian joint $q(z_1, \tilde{\mathbf{z}})$, using the conditional property of distribution, and temporarily neglecting the coefficients and indices related to the fourth cumulant V, we can compute the conditional first moment in the following:

$$\begin{aligned}\mathbb{E}[z_1|\tilde{\mathbf{z}}] &\propto \int dz_1 z_1 q(z_1, \tilde{\mathbf{z}}) \\ &= \int dz_1 z_1 \big(1 + V\tilde{\partial}^4 + V\,\tilde{\partial}^3\partial_{z_1} + V\tilde{\partial}^2\partial^2_{z_1} + V\,\tilde{\partial}\partial^3_{z_1} \\ &\quad + V\partial^4_{z_1}\big)[\mathcal{N}(z_1|\tilde{\mathbf{z}})\mathcal{N}(\tilde{\mathbf{z}})] \\ &= (1 + V\tilde{\partial}^4)[\mu(\tilde{\mathbf{z}})\mathcal{N}(\tilde{\mathbf{z}})] - V\tilde{\partial}^3\mathcal{N}(\tilde{\mathbf{z}})\,.\end{aligned} \tag{B1}$$

In the first equality, the derivative terms of equal to or higher order than $\partial^2_{z_1}$ do not contribute as they produce zero when acting on z_1. In the second equality, one can further reduce the first term into $\mu(1 + V\tilde{\partial}^4)\mathcal{N} + 4(\tilde{\partial}\mu)\tilde{\partial}^3\mathcal{N}$ since the μ term is linear in $\tilde{\mathbf{z}}$. Note that the former term is actually $\mu q(\tilde{\mathbf{z}})$. In addition, the minus sign preceding the $\tilde{\partial}^3$ arises from the adjoint property of moving ∂_{z_1} to act on z_1. By restoring the coefficients with some combinatorics and that $\partial_i\mu(\mathbf{z}) = [K^{-1}\mathbf{k}]_i$, it proves Eq. (8).

The conditional second moment can be computed in a similar manner,

$$\begin{aligned}\mathbb{E}\big[z_1^2|\tilde{\mathbf{z}}\big] &\propto \int dz_1 z_1^2 q(z_1, \tilde{\mathbf{z}}) \\ &= \int dz_1 z_1^2 \big(1 + V\tilde{\partial}^4 + V\,\tilde{\partial}^3\partial_{z_1} + V\,\tilde{\partial}^2\partial^2_{z_1} + V\,\tilde{\partial}\partial^3_{z_1} \\ &\quad + V\partial^4_{z_1}\big)[\mathcal{N}(z_1|\tilde{\mathbf{z}})\mathcal{N}(\tilde{\mathbf{z}})] \\ &= (1+V\tilde{\partial}^4)\{[\mu^2(\tilde{\mathbf{z}}) + \sigma^2]\mathcal{N}(\tilde{\mathbf{z}})\} - 2V\tilde{\partial}^3[\mu(\tilde{\mathbf{z}})\mathcal{N}(\tilde{\mathbf{z}})] \\ &\quad + 2V\tilde{\partial}^2[\mathcal{N}(\tilde{\mathbf{z}})].\end{aligned} \tag{B2}$$

The first term in the second equality will contribute a term $\propto q(\tilde{\mathbf{z}})$, the the details are similar with the above.

APPENDIX C: PROOFS OF EVIDENCE IN EQ. (17) AND POSTERIOR IN EQ. (18)

We start with the definition in Eq. (17) and focus on the correction term. The computation involves the marginalization over all z's as well as a complication arising from the derivatives with respect to some four z's. Using the adjoint property of derivative, i.e., $\int f(x)[\partial g(x)/\partial x]dx = \int[-\partial f(x)/\partial x]g(x)dx$, provided that the product fg vanishes at infinity, we can shuffle the operator ∂_z acting on the Gaussian prior on the right most to act on the Gaussian likelihood on the left. Then, use the property, $(\partial_t + \partial_s)\exp[-(s-t)^2] = 0$, we can replace ∂_z with $-\partial_y$ and pull these derivatives out of the integral. The followings summarize these actions,

$$\begin{aligned}\int d\mathbf{z}\,\mathcal{N}(\mathbf{y}|\mathbf{z},\Lambda)\big[v_{ijml}\partial_{z_i}\partial_{z_j}\partial_{z_m}\partial_{z_l}\mathcal{N}(\mathbf{z}|0,K)\big] &= v_{ijml}\int d\mathbf{z}\big[\partial_{z_i}\partial_{z_j}\partial_{z_m}\partial_{z_l}\mathcal{N}(\mathbf{y}|\mathbf{z},\Lambda)\big]\mathcal{N}(\mathbf{z}|0,K)\\ &= v_{ijml}\partial_{y_i}\partial_{y_j}\partial_{y_m}\partial_{y_l}\int d\mathbf{z}\mathcal{N}(\mathbf{y}|\mathbf{z},\Lambda)\mathcal{N}(\mathbf{z}|0,K)\\ &= v_{ijml}\partial_{y_i}\partial_{y_j}\partial_{y_m}\partial_{y_l}\mathcal{N}(\mathbf{y}|0,K+\Lambda). \end{aligned}\tag{C1}$$

As for the proof for Eq. (18), we can proceed with computation of the numerator there, and note that the derivative ∂_{z_*} in $q(z_*, \mathbf{z})$ can be moved out of the integral directly. To accommodate this extra variable, we use the hat on top of the indices to stress that $*$ is included. The following contains the meaning of our notation as well as some details:

$$\begin{aligned}v_{\hat{i}\hat{j}\hat{m}\hat{l}}\int d\mathbf{z}\,\mathcal{N}(\mathbf{y}|\mathbf{z},\Lambda)\big[\partial_{z_{\hat{i}}}\partial_{z_{\hat{j}}}\partial_{z_{\hat{m}}}\partial_{z_{\hat{l}}}\mathcal{N}(z_*,\mathbf{z}|0,\hat{K})\big] &= \big[v_{ijml}\partial_{y_i}\partial_{y_j}\partial_{y_m}\partial_{y_l} + v_{ijm*}\partial_{y_i}\partial_{y_j}\partial_{y_m}\partial_{z_*}\\ &\quad + v_{ij**}\partial_{y_i}\partial_{y_j}\partial^2_{z_*} + v_{i***}\partial_{y_i}\partial^3_{z_*} + v_{****}\partial^4_{z_*}\big]\\ &\quad\times\int d\mathbf{z}\,\mathcal{N}(\mathbf{y}|\mathbf{z},\Lambda)\mathcal{N}(z_*,\mathbf{z}|0,\hat{K})\\ &= v_{\hat{i}\hat{j}\hat{m}\hat{l}}\partial_{\hat{i}\hat{j}\hat{m}\hat{l}}\mathcal{N}\left[\begin{pmatrix} z_*\\ \mathbf{y}\end{pmatrix}\middle|0,\begin{pmatrix}1 & \mathbf{k}_*^t\\ \mathbf{k}_* & K+\Lambda\end{pmatrix}\right],\end{aligned}\tag{C2}$$

where the readers can refer to the text around Eq. (2.21) in Ref. [2] for similar integration leading to the second equality.

APPENDIX D: PROOF OF EQ. (21)

To prove the correction term to the second moment, the previous tricks dealing with Gaussians are still useful. Below, we first shuffle the four derivatives acting on the bivariate Gaussian on the left most to act on the Gaussian factor. One can further use $\partial_{z_{1,2}}\exp[-(z_1-z_2)^2/2] = \pm\partial_{z_-}\exp(-z_-^2/2)$ such that one can rewrite $\partial_{ijlm} = (-1)^{n_1}\partial^4_{z_-}$ with n_1 denoting the number of involved z_1's. The rest of computation is straightforward,

$$\begin{aligned}\int dz_1dz_2\, e^{-\frac{(z_1-z_2)^2}{2H}}(v_{ijml}\partial_i\partial_j\partial_m\partial_l)\mathcal{N}(z_1,z_2|0,K_2) &= \int dz_1dz_2[(v_{ijml}\partial_i\partial_j\partial_m\partial_l)e^{-\frac{(z_1-z_2)^2}{2H}}]\mathcal{N}(z_1,z_2|0,K_2)\\ &= \frac{V_{1111}+V_{2222}-4V_{1222}-4V_{21111}+6V_{1122}}{24}\\ &\quad\times\int\frac{dz_+dz_-}{2}[(\partial_{z_-})^4e^{-\frac{z_-^2}{2H}}]\mathcal{N}(z_+|0,\sigma_+^2)\mathcal{N}(z_-|0,\sigma_-^2)\\ &= V'\int dz\left(\frac{3}{H^2}-\frac{6z^2}{H^3}+\frac{z^4}{H^4}\right)e^{-\frac{z^2}{2H}}\mathcal{N}(z|0,\sigma_-^2)\\ &= \frac{V'}{\sqrt{1+\frac{\sigma_-^2}{H}}}\left[\frac{3}{H^2}-\frac{6\sigma_-^2}{H^3(1+\sigma_-^2/H)}\right.\\ &\quad\left.+\frac{3\sigma_-^4}{H^4(1+\sigma_-^2/H)^2}\right].\end{aligned}\tag{D1}$$

We have used the notation for the variances $\sigma_\pm^2 = 2[1 \pm k(\mathbf{x}_1,\mathbf{x}_2)]$.

[1] R. M. Neal, Monte Carlo implementation of Gaussian process models for Bayesian regression and classification, arXiv:physics/9701026v2.

[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian Process for Machine Learning* (MIT Press, Cambridge, MA, 2006).

[3] C. K. Williams, Computing with infinite networks, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 1997), pp. 295–301.

[4] Y. Cho and L. K. Saul, Kernel methods for deep learning, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2009), pp. 342–350.

[5] A. G. d. G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, Gaussian process behaviour in wide deep neural networks, arXiv:1804.11271.

[6] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, Deep neural networks as Gaussian processes, arXiv:1711.00165.

[7] I. Osband, Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout, in *Workshop on Bayesian Deep Learning* (NIPS, 2016).

[8] H. S. Seung, H. Sompolinsky, and N. Tishby, Statistical mechanics of learning from examples, Phys. Rev. A **45**, 6056 (1992).

[9] S. Yaida, Non-Gaussian processes and neural networks at finite widths, in *Mathematical and Scientific Machine Learning* (PMLR, 2020), pp. 165–192.

[10] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory* (Cambridge University Press, Cambridge, UK, 2021).

[11] E. Dyer and G. Gur-Ari, Asymptotics of wide networks from Feynman diagrams, arXiv:1909.11304.

[12] M. Gabrié, Mean-field inference methods for neural networks, J. Phys. A: Math. Theor. **53**, 223002 (2020).

[13] G. Naveh, O. Ben David, H. Sompolinsky, and Z. Ringel, Predicting the outputs of finite deep neural networks trained with noisy gradients, Phys. Rev. E **104**, 064301 (2021).

[14] Ib. M. Skovgaard, On multivariate Edgeworth expansions, Int. Stat. Rev. **54**, 169 (1986).

[15] C.-K. Lu, S. C.-H. Yang, X. Hao, and P. Shafto, Interpretable deep Gaussian processes with moments, in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2020), pp. 613–623.

[16] D. Agrawal, T. Papamarkou, and J. D. Hinkle, Wide neural networks with bottlenecks are deep Gaussian processes, J. Mach. Learn. Res. **21**, 175 (2020).

[17] A. Damianou and N. Lawrence, Deep Gaussian processes, in *Artificial Intelligence and Statistics* (PMLR, Scottdale, AZ, 2013), pp. 207–215.
[18] A. Rahimi and B. Recht, Random features for large-scale kernel machines, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2008), pp. 1177–1184.
[19] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2018), pp. 8571–8580.
[20] J. M. Antognini, Finite size corrections for neural network Gaussian processes, arXiv:1908.10030.
[21] J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters, and Complexity* (Oxford University Press, New York, 2021), Vol. 14.
[22] A. Papoulis and S. Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes* (McGraw Hill, New York, 2002).
[23] M. Welling, Robust higher order statistics, in *International Workshop on Artificial Intelligence and Statistics* (PMLR, 2005), pp. 405–412.
[24] E. Sellentin, A. H. Jaffe, and A. F. Heavens, On the use of the Edgeworth expansion in cosmology I: How to foresee and evade its pitfalls, arXiv:1709.03452.
[25] A. Shah, A. Wilson, and Z. Ghahramani, Student-t processes as alternatives to Gaussian processes, in *Artificial Intelligence and Statistics* (PMLR, Reykjavik, Iceland, 2014) pp. 877–885.
[26] K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone, Random feature expansions for deep Gaussian processes, in *Proceedings of the 34th International Conference on Machine Learning* (PMLR, Sydney, Australia, 2017), Vol. 70, pp. 884–893.
[27] M. M. Dunlop, M. A. Girolami, A. M. Stuart, and A. L. Teckentrup, How deep are deep Gaussian processes? J Mach. Learn. Res. **19**, 2100 (2018).
[28] G. Pleiss and J. P. Cunningham, The limitations of large width in neural networks: A deep Gaussian process perspective, Adv. Neural Info. Process. Syst. **34**, 3349 (2021).
[29] M. Vladimirova, J. Verbeek, P. Mesejo, and J. Arbel, Understanding priors in Bayesian neural networks at the unit level, in *International Conference on Machine Learning* (PMLR, Long Beach, CA, 2019), pp. 6458–6467.
[30] J. Zavatone-Veth and C. Pehlevan, Exact marginal prior distributions of finite Bayesian neural networks, Adv. Neural Info. Process. Syst. **34**, 3364 (2021).
[31] J. Zavatone-Veth, A. Canatar, B. Ruben, and C. Pehlevan, Asymptotics of representation learning in finite bayesian neural networks, Adv. neural info. Process. Syst. **34**, 24765 (2021).
[32] C.-K. Lu, Scott Cheng-Hsin Yang, and P. Shafto, Standing-wave-decomposition Gaussian process, Phys. Rev. E **98**, 032303 (2018).
[33] D. Duvenaud, O. Rippel, R. Adams, and Z. Ghahramani, Avoiding pathologies in very deep networks, in *Artificial Intelligence and Statistics* (PMLR, Reykjavik, Iceland, 2014), pp. 202–210.
[34] L. Aitchison, Why bigger is not always better: on finite and infinite neural networks, in *International Conference on Machine Learning* (PMLR, 2020), pp. 156–164.
[35] M. B. Li, M. Nica, and D. M. Roy, The neural covariance SDE: Shaped infinite depth-and-width networks at initialization, arXiv:2206.02768.
[36] B. Hanin, Random fully connected neural networks as perturbatively solvable hierarchies, arXiv:2204.01058.
[37] S. Ariosto, R. Pacelli, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo, Statistical mechanics of deep learning beyond the infinite-width limit, arXiv:2209.04882.
[38] Q. Li and H. Sompolinsky, Statistical Mechanics of Deep Linear Neural Networks: The Backpropagating Kernel Renormalization, Phys. Rev. X **11**, 031059 (2021).
[39] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, Finite versus infinite neural networks: an empirical study, Adv. Neural Inf. Process. Syst. **33**, 15156 (2020).
[40] L. Csató, E. Fokoué, M. Opper, B. Schottky, and O. Winther, Efficient approaches to Gaussian process classification, Adv. Neural Inf. Process. Syst. **12**, 251 (1999).
[41] P. Sollich, Learning curves for Gaussian processes, Adv. Neural Inf. Process. Syst. **11**, 344 (1998).
[42] C. K. Williams and F. Vivarelli, Upper and lower bounds on the learning curve for Gaussian processes, Mach. Learn. **40**, 77 (2000).
[43] M. Seeger, PAC-Bayesian generalisation error bounds for Gaussian process classification, J. Mach. Learn. Res. **3**, 233 (2002).
[44] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, On exact computation with an infinitely wide neural net, Adv. Neural Inf. Process. Syst. **32**, 8141 (2019).
[45] B. Hanin and M. Nica, Finite depth and width corrections to the neural tangent kernel, arXiv:1909.05989.
[46] R. Karakida, S. Akaho, and S.-i. Amari, Universal statistics of Fisher information in deep neural networks: Mean field approach, in *The 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, Okinawa, Japan, 2019), pp. 1032–1041.
[47] M. E. E. Khan, A. Immer, E. Abedi, and M. Korzepa, Approximate inference turns deep networks into Gaussian processes, in *Adv. Neural Inf. Process. Syst.* (Curran Associates, Inc., Red Hook, NY, 2019), pp. 3094–3104.
[48] D. Babich, Optimal estimation of non-Gaussianity, Phys. Rev. D **72**, 043003 (2005).