An brief introduction to the neural network Gaussian process from the perspective of mean field theory

Jacob A. Zavatone-Veth*

May 6, 2023

Abstract

In these lecture notes, we provide a brief, heuristic derivation of the neural network Gaussian process (NNGP) correspondence using the framework of self-consistent mean field theory.

These notes were originally prepared to accompany a tutorial given at the 2022 HHMI Janelia Junior Scientist Workshop on Theoretical Neuroscience.

1 Introduction

Consider a deep feedforward neural network with L layers,

$$\mathbf{h}^{(0)}(\mathbf{x}) = \mathbf{x} \in \mathbb{R}^{n_0} \tag{1}$$

$$\mathbf{h}^{(\ell)}(\mathbf{x}) = \frac{1}{\sqrt{n_{\ell-1}}}\mathbf{W}^{(\ell)}\phi_\ell(\mathbf{h}^{(\ell-1)}(\mathbf{x})) \in \mathbb{R}^{n_\ell} \qquad \ell = 1, \ldots, L. \tag{2}$$

For simplicity, we focus on the case of networks with no bias terms, as their introduction does not qualitatively affect our subsequent results.

As is common practice at the level of the prior in Bayesian neural networks, or at initialization in the context of gradient-descent-based maximum likelihood estimation, assume that the weight distribution is isotropic and Gaussian

$$W_{ij}^{(\ell)} \sim_{\text{i.i.d.}} \mathcal{N}(0,1). \tag{3}$$

We will consider the infinite-width limit

$$n_1, \cdots, n_L \to \infty \tag{4}$$

for fixed input dimension n_0, depth L, and dataset size P. To be more precise, we let

$$n_\ell = N\alpha_\ell, \tag{5}$$

and take the infinite-width limit by taking $N \to \infty$ for fixed ratios $\alpha_\ell \in (0, \infty)$. In a practical application, $\mathbf{h}^{(L)}$ would be followed by a fixed-dimensional linear readout [10].

*Department of Physics and Center for Brain Science,
Harvard University,
Cambridge MA 02138,
jzavatoneveth@g.harvard.edu

Our goal is to characterize the resulting statistics of the network activities for some set of P inputs $\mathbf{x}_\mu$. We allow these inputs, and the input dimension, to be arbitrary up to the condition that the $P \times P$ input Gram matrix

$$G_{\mu\nu} = \mathbf{x}_\mu \cdot \mathbf{x}_\nu \tag{6}$$

is invertible. In particular, we may also take the input dimension to infinity with the layer widths.

Our overall approach follows recent work by Segadlo et al. [6], though our notation and some steps of the computation hew more closely to our own work in Zavatone-Veth et al. [10]. To study activity statistics, we consider the moment generating function

$$Z = \mathbb{E}_{\mathbf{W}^{(1)},\cdots,\mathbf{W}^{(L)}} \exp\left(i \sum_{\ell=1}^{L} \sum_{\mu=1}^{P} \mathbf{b}_\mu^{(\ell)} \cdot \mathbf{h}_\mu^{(\ell)} \right), \tag{7}$$

where we write

$$\mathbf{h}_\mu^{(\ell)} = \mathbf{h}^{(\ell)}(\mathbf{x}_\mu) \tag{8}$$

for brevity. The MGF has the important property that it is equal to unity at zero source, i.e.,

$$Z\Big|_{\{\mathbf{b}_\mu^{(\ell)}=\mathbf{0}\}} = 1. \tag{9}$$

2 Integrating out the weights

The expectation over the weights in (7) is challenging to evaluate because they appear in the iterative linear-nonlinear function composition by which the network is defined. To unroll these expectations, we multiply by one. Less glibly, we unpack the definitions of the hidden layer activities by multiplying by integrals of δ-distributions that enforce their definitions:

$$1 = \int d\mathbf{h}_\mu^{(\ell)} \, \delta\left(\mathbf{h}_\mu^{(\ell)} - \frac{1}{\sqrt{n_{\ell-1}}} \mathbf{W}^{(\ell)} \phi_\ell(\mathbf{h}_\mu^{(\ell-1)}) \right). \tag{10}$$

Then, we have

$$Z = \int \prod_{\ell=1}^{L} \prod_{\mu=1}^{P} d\mathbf{h}_\mu^{(\ell)} \exp\left(i \sum_{\ell=1}^{L} \sum_{\mu=1}^{P} \mathbf{b}_\mu^{(\ell)} \cdot \mathbf{h}_\mu^{(\ell)} \right) \mathbb{E}_{\mathbf{W}^{(1)},\cdots,\mathbf{W}^{(L)}} \prod_{\ell=1}^{L} \prod_{\mu=1}^{P} \delta\left(\mathbf{h}_\mu^{(\ell)} - \frac{1}{\sqrt{n_{\ell-1}}} \mathbf{W}^{(\ell)} \phi_\ell(\mathbf{h}_\mu^{(\ell-1)}) \right). \tag{11}$$

We now replace the δ-distributions by their Fourier representations

$$\delta\left(\mathbf{h}_\mu^{(\ell)} - \frac{1}{\sqrt{n_{\ell-1}}} \mathbf{W}^{(\ell)} \phi_\ell(\mathbf{h}_\mu^{(\ell-1)}) \right) = \frac{1}{(2\pi)^{n_\ell}} \int d\hat{\mathbf{h}}_\mu^{(\ell)} \exp\left[i \hat{\mathbf{h}}_\mu^{(\ell)} \cdot \left(\mathbf{h}_\mu^{(\ell)} - \frac{1}{\sqrt{n_{\ell-1}}} \mathbf{W}^{(\ell)} \phi_\ell(\mathbf{h}_\mu^{(\ell-1)}) \right) \right], \tag{12}$$

which yields

$$\begin{aligned} Z = \int \prod_{\ell=1}^{L} \prod_{\mu=1}^{P} \frac{d\mathbf{h}_\mu^{(\ell)} \, d\hat{\mathbf{h}}_\mu^{(\ell)}}{(2\pi)^{n_\ell}} \exp\left(i \sum_{\ell=1}^{L} \sum_{\mu=1}^{P} [\mathbf{b}_\mu^{(\ell)} + \hat{\mathbf{h}}_\mu^{(\ell)}] \cdot \mathbf{h}_\mu^{(\ell)} \right) \\ \times \mathbb{E}_{\mathbf{W}^{(1)},\cdots,\mathbf{W}^{(L)}} \exp\left[i \sum_{\ell=1}^{L} \frac{1}{\sqrt{n_{\ell-1}}} \sum_{\mu=1}^{P} \hat{\mathbf{h}}_\mu^{(\ell)} \cdot \mathbf{W}^{(\ell)} \phi_\ell(\mathbf{h}_\mu^{(\ell-1)}) \right]. \end{aligned} \tag{13}$$

The Gaussian expectations over the weights now factor, and can be easily evaluated

$$\mathbb{E}_{\mathbf{W}^{(1)},\cdots,\mathbf{W}^{(L)}} \exp\left[i \sum_{\ell=1}^{L} \frac{1}{\sqrt{n_{\ell-1}}} \sum_{\mu=1}^{P} \hat{\mathbf{h}}_\mu^{(\ell)} \cdot \mathbf{W}^{(\ell)} \phi_\ell(\mathbf{h}_\mu^{(\ell-1)}) \right]$$

$$= \prod_{\ell=1}^{L} \mathbb{E}_{\mathbf{W}^{(\ell)}} \exp\left[i \frac{1}{\sqrt{n_{\ell-1}}} \sum_{\mu=1}^{P} \hat{\mathbf{h}}_\mu^{(\ell)} \cdot \mathbf{W}^{(\ell)} \phi_\ell(\mathbf{h}_\mu^{(\ell-1)}) \right] \tag{14}$$

$$= \prod_{\ell=1}^{L} \exp\left[-\frac{1}{2} \sum_{\mu,\nu=1}^{P} \hat{\mathbf{h}}_\mu^{(\ell)} \cdot \hat{\mathbf{h}}_\nu^{(\ell)} \left(\frac{1}{n_{\ell-1}} \phi_\ell(\mathbf{h}_\mu^{(\ell-1)}) \cdot \phi_\ell(\mathbf{h}_\nu^{(\ell-1)}) \right) \right]. \tag{15}$$

At this point, we have an expression for the moment generating function in terms of only the activities $\mathbf{h}_\mu^{(\ell)}$ and the corresponding Lagrange multipliers $\hat{\mathbf{h}}_\mu^{(\ell)}$:

$$Z = \int \prod_{\ell=1}^{L} \prod_{\mu=1}^{P} \frac{d\mathbf{h}_\mu^{(\ell)} \, d\hat{\mathbf{h}}_\mu^{(\ell)}}{(2\pi)^{n_\ell}} \exp\left(i \sum_{\ell=1}^{L} \sum_{\mu=1}^{P} [\mathbf{b}_\mu^{(\ell)} + \hat{\mathbf{h}}_\mu^{(\ell)}] \cdot \mathbf{h}_\mu^{(\ell)} \right)$$
$$\times \prod_{\ell=1}^{L} \exp\left[-\frac{1}{2} \sum_{\mu,\nu=1}^{P} \hat{\mathbf{h}}_\mu^{(\ell)} \cdot \hat{\mathbf{h}}_\nu^{(\ell)} \left(\frac{1}{n_{\ell-1}} \phi_\ell(\mathbf{h}_\mu^{(\ell-1)}) \cdot \phi_\ell(\mathbf{h}_\nu^{(\ell-1)}) \right) \right]. \tag{16}$$

However, this expression is not particularly tractable because different layers and neurons are coupled together.

3 Introducing the kernels

To make progress, we observe that the layers are coupled only through the kernels

$$K_{\mu\nu}^{(\ell)} = \frac{1}{n_\ell} \phi_\ell(\mathbf{h}_\mu^{(\ell)}) \cdot \phi_\ell(\mathbf{h}_\nu^{(\ell)}). \tag{17}$$

Therefore, we can decouple layers by introducing these kernels as integration variables. To do so, we multiply by one:

$$1 = \prod_{\ell=1}^{L} \prod_{\mu,\nu=1}^{P} \int dK_{\mu\nu}^{(\ell)} \, \delta\left(K_{\mu\nu}^{(\ell)} - \frac{1}{n_\ell} \phi_\ell(\mathbf{h}_\mu^{(\ell)}) \cdot \phi_\ell(\mathbf{h}_\nu^{(\ell)}) \right) \tag{18}$$

$$= \prod_{\ell=1}^{L} \prod_{\mu,\nu=1}^{P} \int \frac{dK_{\mu\nu}^{(\ell)} \, d\hat{K}_{\mu\nu}^{(\ell)}}{4\pi/n_\ell} \exp\left[\frac{1}{2} i \hat{K}_{\mu\nu}^{(\ell)} \left(n_\ell K_{\mu\nu}^{(\ell)} - \phi_\ell(\mathbf{h}_\mu^{(\ell)}) \cdot \phi_\ell(\mathbf{h}_\nu^{(\ell)}) \right) \right] \tag{19}$$

$$= \int \prod_{\ell=1}^{L} \prod_{\mu,\nu=1}^{P} \frac{dK_{\mu\nu}^{(\ell)} \, d\hat{K}_{\mu\nu}^{(\ell)}}{4\pi/n_\ell} \exp\left(\frac{1}{2} i \sum_{\ell=1}^{L} n_\ell \sum_{\mu,\nu=1}^{P} K_{\mu\nu}^{(\ell)} \hat{K}_{\mu\nu}^{(\ell)} \right) \prod_{\ell=1}^{L} \prod_{j=1}^{n_\ell} \exp\left(-\frac{1}{2} i \sum_{\mu,\nu=1}^{P} \hat{K}_{\mu\nu}^{(\ell)} \phi_\ell(h_{\mu j}^{(\ell)}) \phi_\ell(h_{\nu j}^{(\ell)}) \right). \tag{20}$$

Here, we also make use of the fact that

$$K_{\mu\nu}^{(0)} = \frac{1}{n_0} \mathbf{x}_\mu \cdot \mathbf{x}_\nu \tag{21}$$

is a fixed object depending only on the dataset, which by our assumptions on the data is an invertible matrix.

Multiplying by one and interchanging the order of integration, we can now factor the integrals over the hidden layer activities:

$$Z = \int \prod_{\ell=1}^{L} \prod_{\mu,\nu=1}^{P} \frac{dK_{\mu\nu}^{(\ell)} d\hat{K}_{\mu\nu}^{(\ell)}}{4\pi/n_\ell} \exp[NS(\mathbf{K}^{(1)}, \cdots, \mathbf{K}^{(L)}, \hat{\mathbf{K}}^{(1)}, \ldots, \hat{\mathbf{K}}^{(L)})], \tag{22}$$

where we have defined the action

$$S(\mathbf{K}^{(1)}, \cdots, \mathbf{K}^{(L)}, \hat{\mathbf{K}}^{(1)}, \ldots, \hat{\mathbf{K}}^{(L)}) = \frac{1}{2} i \sum_{\ell=1}^{L} \alpha_\ell \sum_{\mu,\nu=1}^{P} K_{\mu\nu}^{(\ell)} \hat{K}_{\mu\nu}^{(\ell)} + \frac{1}{N} \sum_{\ell=1}^{L} \sum_{j=1}^{n_\ell} \log z_j^{(\ell)}(b_{\mu j}^{(\ell)}, \mathbf{K}^{(\ell-1)}, \hat{\mathbf{K}}^{(\ell)}) \tag{23}$$

for single-site generating functions

$$z_j^{(\ell)}(b_{\mu j}^{(\ell)}, \mathbf{K}^{(\ell-1)}, \hat{\mathbf{K}}^{(\ell)}) = \int \prod_{\mu=1}^{p} \frac{dh_{\mu j}^{(\ell)} d\hat{h}_{\mu j}^{(\ell)}}{2\pi} \exp\left(i \sum_{\mu=1}^{p} (b_{\mu j}^{(\ell)} + \hat{h}_{\mu j}^{(\ell)}) h_{\mu j}^{(\ell)} - \frac{1}{2} \sum_{\mu,\nu=1}^{p} K_{\mu\nu}^{(\ell-1)} \hat{h}_{\mu j}^{(\ell)} \hat{h}_{\nu j}^{(\ell)} \right)$$
$$\times \exp\left(-\frac{1}{2} i \sum_{\mu,\nu=1}^{p} \hat{K}_{\mu\nu}^{(\ell)} \phi_\ell(h_{\mu j}^{(\ell)}) \phi_\ell(h_{\nu j}^{(\ell)}) \right). \tag{24}$$

4 Saddle-point evaluation

We now observe that the functions $z_j^{(\ell)}(b_{\mu j}^{(\ell)}, \mathbf{K}^{(\ell-1)}, \hat{\mathbf{K}}^{(\ell)})$ are identical across $j = 1, \ldots, n_\ell$ up to the choice of source $b_{\mu j}^{(\ell)}$. Thus, we expect to have

$$\frac{1}{N} \sum_{\ell=1}^{L} \sum_{j=1}^{n_\ell} \log z_j^{(\ell)}(b_{\mu j}^{(\ell)}, \mathbf{K}^{(\ell-1)}, \hat{\mathbf{K}}^{(\ell)}) \sim \mathcal{O}(1). \tag{25}$$

Therefore, the integral over the kernels and the corresponding Lagrange multipliers is a finite $(2LP^2)$ dimensional integral of an integrand of the form $\exp[NS]$ for $S \sim \mathcal{O}(1)$. In the limit $N \to \infty$, we therefore expect to be able to asymptotically approximate the integral using the method of steepest descent. Roughly speaking, this allows us to approximate an integral with a sharply peaked integrand simply by evaluating the integrand at its maximal value. This gives

$$\lim_{N\to\infty} \frac{1}{N} \log Z \sim \operatorname*{extr}_{\mathbf{K}^{(1)}, \cdots, \mathbf{K}^{(L)}, \hat{\mathbf{K}}^{(1)}, \ldots, \hat{\mathbf{K}}^{(L)}} S(\mathbf{K}^{(1)}, \cdots, \mathbf{K}^{(L)}, \hat{\mathbf{K}}^{(1)}, \ldots, \hat{\mathbf{K}}^{(L)}), \tag{26}$$

where the notation $\operatorname{extr}_{\mathbf{K}^{(1)}, \cdots, \mathbf{K}^{(L)}, \hat{\mathbf{K}}^{(1)}, \ldots, \hat{\mathbf{K}}^{(L)}}$ means that we should evaluate the kernels at a stationary point, where

$$\frac{\delta S}{\delta \mathbf{K}^{(\ell)}} = \mathbf{0}, \qquad \frac{\delta S}{\delta \hat{\mathbf{K}}^{(\ell)}} = \mathbf{0} \qquad \ell = 1, \ldots, L. \tag{27}$$

From $\delta S / \delta \mathbf{K}^{(\ell)} = \mathbf{0}$, we have

$$i\hat{K}_{\mu\nu}^{(\ell)} = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \langle \hat{h}_{\mu j}^{(\ell)} \hat{h}_{\nu j}^{(\ell)} \rangle_j, \tag{28}$$

where we introduce the notation

$$\langle\cdot\rangle_j = \frac{1}{z_j^{(\ell)}(b_{\mu j}^{(\ell)}, \mathbf{K}^{(\ell-1)}, \hat{\mathbf{K}}^{(\ell)})} \int \prod_{\mu=1}^{p} \frac{dh_{\mu j}^{(\ell)}\, d\hat{h}_{\mu j}^{(\ell)}}{2\pi} (\cdot) \exp\left(i \sum_{\mu=1}^{p} (b_{\mu j}^{(\ell)} + \hat{h}_{\mu j}^{(\ell)}) h_{\mu j}^{(\ell)} - \frac{1}{2} \sum_{\mu,\nu=1}^{p} K_{\mu\nu}^{(\ell-1)} \hat{h}_{\mu j}^{(\ell)} \hat{h}_{\nu j}^{(\ell)} \right) \times \exp\left(-\frac{1}{2} i \sum_{\mu,\nu=1}^{p} \hat{K}_{\mu\nu}^{(\ell)} \phi_\ell(h_{\mu j}^{(\ell)}) \phi_\ell(h_{\nu j}^{(\ell)}) \right). \tag{29}$$

Similarly, from $\delta S / \delta \hat{\mathbf{K}}^{(\ell)} = \mathbf{0}$, we have

$$K_{\mu\nu}^{(\ell)} = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \langle \phi_\ell(h_{\mu j}^{(\ell)}) \phi_\ell(h_{\nu j}^{(\ell)}) \rangle_j. \tag{30}$$

We claim that $\hat{\mathbf{K}}^{(1)} = \cdots = \hat{\mathbf{K}}^{(L)} = \mathbf{0}$ is a self-consistent solution. This follows from the requirement that $Z = 1$ when $b_{\mu j}^{(\ell)} = 0$. Then, we can see that $z_j^{(\ell)}(b_{\mu j}^{(\ell)}, \mathbf{K}^{(\ell-1)}, \mathbf{0})$ is simply the moment generating function of a P-dimensional Gaussian random vector with mean zero and covariance $\mathbf{K}^{(\ell)}$.

5 Conclusion: the NNGP

At this point, we have found that

$$h_{\mu j}^{(\ell)} \sim \mathcal{N}(0, K_{\mu\nu}^{(\ell-1)} \delta_{ij}), \tag{31}$$

where the kernels are determined through the recurrence

$$K_{\mu\nu}^{(\ell)} = \mathbb{E}_{h_\mu^{(\ell)} \sim \mathcal{N}(0, K_{\mu\nu}^{(\ell-1)})} [\phi_\ell(h_\mu^{(\ell)}) \phi_\ell(h_\nu^{(\ell)})] \tag{32}$$

with initial condition

$$K_{\mu\nu}^{(0)} = \frac{1}{n_0} \mathbf{x}_\mu \cdot \mathbf{x}_\nu. \tag{33}$$

In particular, different neurons in a given layer are statistically independent.

Though our approach here has not been fully mathematically rigorous [1, 6, 10], these results can be established rigorously [2–4, 7]. They can also be extended to other network architectures [5, 9]. Moreover, one can compute finite-size corrections to the NNGP prior and posterior [6, 8, 10], and in some cases compute the prior exactly at finite size [11].

References

[1] A. Crisanti and H. Sompolinsky, "Path integral approach to random neural networks", Phys. Rev. E **98**, 062120 (2018), https://link.aps.org/doi/10.1103/PhysRevE.98.062120.

[2] J. Lee, J. Sohl-Dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, "Deep neural networks as Gaussian processes", in International conference on learning representations (2018), https://openreview.net/forum?id=B1EA-M-0Z.

[3] A. G. d. G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, "Gaussian process behaviour in wide deep neural networks", in International conference on learning representations (2018), https://openreview.net/forum?id=H1-nGgWC-.

[4]R. M. Neal, "Priors for infinite networks", in *Bayesian learning for neural networks* (Springer, 1996), pp. 29–53.

[5]R. Novak, L. Xiao, Y. Bahri, J. Lee, G. Yang, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, "Bayesian deep convolutional networks with many channels are Gaussian processes", in International conference on learning representations (2019), https://openreview.net/forum?id=B1g30j0qF7.

[6]K. Segadlo, B. Epping, A. van Meegen, D. Dahmen, M. Krämer, and M. Helias, "Unified field theoretical approach to deep and recurrent neuronal networks", Journal of Statistical Mechanics: Theory and Experiment **2022**, 103401 (2022), https://dx.doi.org/10.1088/1742-5468/ac8e57.

[7]C. K. Williams, "Computing with infinite networks", Advances in Neural Information Processing Systems, 295–301 (1997), https://papers.nips.cc/paper/1996/hash/ae5e3ce40e0404a45ecacaaf05e5f735-Abstract.html.

[8]S. Yaida, "Non-Gaussian processes and neural networks at finite widths", in Proceedings of the first mathematical and scientific machine learning conference, Vol. 107, edited by J. Lu and R. Ward, Proceedings of Machine Learning Research (July 2020), pp. 165–192, http://proceedings.mlr.press/v107/yaida20a.html.

[9]G. Yang, "Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation", arXiv preprint arXiv:1902.04760 (2019).

[10]J. A. Zavatone-Veth, A. Canatar, B. S. Ruben, and C. Pehlevan, "Asymptotics of representation learning in finite Bayesian neural networks", in Advances in neural information processing systems, Vol. 34, edited by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (2021), https://proceedings.neurips.cc/paper/2021/hash/cf9dc5e4e194fc21f397b4cac9cc3ae9-Abstract.html.

[11]J. A. Zavatone-Veth and C. Pehlevan, "Exact marginal prior distributions of finite Bayesian neural networks", in Advances in neural information processing systems, Vol. 34, edited by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (2021), https://proceedings.neurips.cc/paper/2021/hash/1baff70e2669e8376347efd3a874a341-Abstract.html.