A brief note on understanding neural networks as Gaussian processes

Mengwu Guo*

Applied Analysis, Department of Applied Mathematics, University of Twente

Abstract: As a generalization of the work in [5] (Lee et al., 2017), this note briefly discusses when the prior of a neural network output follows a Gaussian process, and how a neural-network-induced Gaussian process is formulated. The posterior mean functions of such a Gaussian process regression lie in the reproducing kernel Hilbert space defined by the neural-network-induced kernel. In the case of two-layer neural networks, the induced Gaussian processes provide an interpretation of the reproducing kernel Hilbert spaces whose union forms a Barron space.

Keywords: neural network, Gaussian process, neural-network-induced kernel, reproducing kernel Hilbert space, Barron space

1 Neural network

We consider an \(L\)-hidden-layer fully-connected neural network [8] with width \(N_l\) and nonlinear activation function \(\phi\) for the \(l\)-th layer, \(1 \leq l \leq L\). At the \(j\)-th neuron in the \(l\)-th layer of the network, the pre-bias and post-activation are denoted by \(z_{ij}^{[l]}\) and \(x_{ij}^{[l]}\), respectively, \(1 \leq i \leq N_l\). Let \(x = x^{[0]} \in \mathbb{R}^{d_{in}}\) denote the inputs of the network and \(y = z^{[L+1]} \in \mathbb{R}^{d_{out}}\) denote the outputs. Note that we also let \(N_0 = d_{in}\) and \(N_{L+1} = d_{out}\). Weight and bias parameters between the \((l-1)\)-th and \(l\)-th layers are represented by \(W_{ij}^{[l]}\) and \(b_{i}^{[l]}\), respectively, \(1 \leq l \leq (L+1), 1 \leq i \leq N_l, 1 \leq j \leq N_{l-1}\). Then one has

\[
\begin{align*}
  z_{ij}^{[l]}(x) &= \sum_{j=1}^{N_{l-1}} W_{ij}^{[l]} x_{j}^{[l-1]}(x), \quad x_{ij}^{[l]}(x) = \phi(z_{ij}^{[l]}(x) + b_{i}^{[l]}), \quad 1 \leq i \leq N_l, 1 \leq l \leq L, \quad \text{and} \\
  y_{i}(x) &= z_{i}^{[L+1]}(x) = \sum_{j=1}^{N_{L}} W_{ij}^{[L+1]} x_{j}^{[L]}(x), \quad 1 \leq i \leq d_{out}.
\end{align*}
\]

A multivariate function \(f : \Omega \rightarrow \mathbb{R}^{d_{out}}\) (\(\Omega \subset \mathbb{R}^{d_{in}}\)) is approximated by a vector-valued surrogate of neural network, denoted by \(f^{\text{NN}}(\cdot ; W, b) : \Omega \rightarrow \mathbb{R}^{d_{out}}\). Here, vectors \(W\) and \(b\) collect all the weight and bias parameters, respectively. Based on the data of \(M\) input-output pairs \(\{(x^{(m)}, y^{(m)})\}_{m=1}^{M}\), the training of such a neural network is often performed by minimizing a loss function:

\[
(W, b) = \arg\min_{W, b} \left\{ \frac{1}{M} \sum_{m=1}^{M} \left\| y^{(m)} - f^{\text{NN}}(x^{(m)}; W, b) \right\|_2^2 + \lambda \|W\|_2^2 \right\},
\]

in which the first term is the mean square error and the second is an \(L_2\)-regularization term with \(\lambda \geq 0\) being the penalty coefficient.

*E-mail: m.guo@utwente.nl
2 Neural network prior as Gaussian processes

The prior of a neural network output is a Gaussian process under the following assumptions [5, 6].

- For the first layer ($l = 1$), $(w_{i1}^{[1]}, b_{i}^{[1]}) \sim \pi$ are independent and identically distributed (i.i.d) with respect to $1 \leq i \leq N_1$, $w_{i1}^{[1]} := \{W_{i1}^{[1]}, W_{i2}^{[1]}, \ldots, W_{id_{out}}^{[1]}\}^T$, and $\pi$ can be any distribution;
- All the other weights and biases are independently drawn;
- For the $l$-th layer, $2 \leq l \leq L + 1$, $W_{ij}^{[l]}$’s are i.i.d. with mean value $\mu_{ij}^{[l]}/N_{l-1}$ and variance $\sigma_{ij}^{[l]}^2/N_{l-1}$, $1 \leq i \leq N_l$, $1 \leq j \leq N_{l-1}$;
- For the $l$-th layer, $2 \leq l \leq L$, $b_i^{[l]}$’s are i.i.d. with mean value $\mu_i^{[l]}$ and variance $\sigma_i^{[l]}$, $1 \leq i \leq N_l$; $b_i^{[L+1]} = 0$, $1 \leq i \leq d_{out}$; and
- $N_l \to \infty$, $1 \leq l \leq L$.

First we show by induction that $\{z_i^{[l]}(x) | i \in \mathbb{N}^+\}$ are independent, identical random fields over $\Omega$, $1 \leq l \leq L + 1$. When $l = 1$, $z_i^{[1]}(x) = w_i^{[1]T} x$, and the proposition apparently holds true. Assume it holds true for $l$, $1 \leq l \leq L$. We consider two arbitrary locations $x, x' \in \Omega$ and hence have

$$E[z_i^{[l+1]}(x)] = \mu_i^{[l+1]}E_{z_i^{[l]}(x)}[\phi(z_i^{[l]}(x)) + b_i^{[l]}], \quad (3)$$

and

$$\text{Cov}[z_i^{[l+1]}(x), z_j^{[l+1]}(x')]
= \sum_{p,q} \text{Cov}[W_{ip}^{[l+1]} x_p^{[l]}(x), W_{jq}^{[l+1]} x_q^{[l]}(x')]
= \sum_{p,q} \left\{ E[W_{ip}^{[l+1]} W_{jq}^{[l+1]}] \cdot E[x_p^{[l]}(x) x_q^{[l]}(x')] - E[W_{ip}^{[l+1]}] \cdot E[W_{jq}^{[l+1]}] \cdot E[x_p^{[l]}(x)] \cdot E[x_q^{[l]}(x')] \right\}
= \sum_p \left\{ E[W_{ip}^{[l+1]} W_{jp}^{[l+1]}] \cdot E[x_p^{[l]}(x) x_p^{[l]}(x')] - \frac{(\mu_{ij}^{[l+1]})^2}{N_l^2} \cdot E[x_p^{[l]}(x)] \cdot E[x_p^{[l]}(x')] \right\}
= \sum_{p \neq q} \frac{(\mu_{ij}^{[l+1]})^2}{N_l^2} \cdot \text{Cov}[x_p^{[l]}(x), x_p^{[l]}(x')]
= 0,$$

since $z_p^{[l]}$ and $z_q^{[l]}$ are independent when $p \neq q$.

$$= \frac{(\mu_{ij}^{[l+1]})^2}{N_l} \cdot \text{Cov}[x_p^{[l]}(x), x_p^{[l]}(x')] + \sigma_{ij}^2 \delta_{ij} E[x_p^{[l]}(x), x_p^{[l]}(x')]
\to \sigma_{ij}^2 \delta_{ij} E_{z_i^{[l]}(x)}[\phi(z_i^{[l]}(x)) + b_i^{[l]}] \cdot \phi(z_i^{[l]}(x')) + b_i^{[l]}], \quad \text{as } N_l \to \infty.$$  

Here $\delta_{ij}$ denotes the Kronecker delta. When $i \neq j$, the covariance value equals zero, implying that $\{z_i^{[l+1]}(x) | i \in \mathbb{N}^+\}$ are all uncorrelated. In fact, they are independent as each $z_i^{[l+1]}(x)$ is independently defined as a combination of $\{x_i^{[l]} | i \in \mathbb{N}^+\}$. It is also trivial to see that these random fields are identical, meaning that the proposition holds true for $l + 1$ as well, which completes its proof.

As defined in [1], furthermore, $z_i^{[l]}$ equals a sum of $N_{l-1}$ i.i.d. random variables whose mean and variance are proportional to $1/N_{l-1}$. As $N_{l-1} \to \infty$, the central limit theorem [2] can be applied, which gives that $\{z_i^{[l]}(x) | i \in \mathbb{N}^+\}$

1 The discussion in this section generalizes the work in [5].
follow identical, independent Gaussian processes, $2 \leq l \leq L + 1$, i.e., $z^{[l]} \sim \mathcal{GP}(h^{[l]}, k^{[l]})$ whose mean and covariance functions are

$$
    h^{[l]}(x) = \mu^{[l]} + \mathbb{E}_{z^{[l-1]}, b^{[l-1]}}[\phi(z^{[l-1]}(x) + b^{[l-1]})], \quad \text{and}
    \quad k^{[l]}(x, x') = \sigma^2_w^{[l]} + \mathbb{E}_{z^{[l-1]}, b^{[l-1]}}[\phi(z^{[l-1]}(x) + b^{[l-1])} \cdot \phi(z^{[l-1]}(x') + b^{[l-1]})].
$$

(5)

Therefore, the neural network outputs, collected in the $(L + 1)$-th layer, also follow independent, identical Gaussian process priors, written as $y \sim \mathcal{GP}(h_{NN}, k_{NN}) = \mathcal{GP}(h^{[L+1]}, k^{[L+1]})$. Often referred to as neural-network-induced Gaussian processes, such priors can be explicitly formulated through the recursive relation (5).

### 3 Regression using neural-network-induced Gaussian processes

- **Gaussian process regression**

  As discussed in Section 2, each output $y(x)$ follows a prior of neural-network-induced Gaussian process. Here we assume that the output is corrupted by an independent Gaussian noise, written as

$$
    y(x) \sim \mathcal{GP}(h_{NN}(x), k_{NN}(x, x')) + \mathcal{N}(0, \sigma^2) .
$$

(6)

Conditioned on the training data $(X, y) = \{(x^{(m)}, y^{(m)})\}_{m=1}^{M} \in \mathbb{R}^{d_x \times M} \times \mathbb{R}^{M}$, the noise-free posterior output $y^*(x)|X, y$ follows a new Gaussian process $\mathcal{GP}(h^*_{NN}(x), k^*_{NN}(x, x'))$ whose mean and covariance functions are given as

$$
    h^*_{NN}(x) = h_{NN}(x) + k_{NN}(X, x)^T[K + \sigma^2_0 I_M]^{-1}(y - h_{NN}(X)) , \quad \text{and}
    \quad k^*_{NN}(x, x') = k_{NN}(x, x') - k_{NN}(X, x)^T[K + \sigma^2_0 I_M]^{-1}k_{NN}(X, x') ,
$$

(7)

in which $h_{NN}(X) := [h_{NN}(x^{(1)}), \ldots, h_{NN}(x^{(M)})]^T \in \mathbb{R}^{M}$, $k_{NN}(X, x) := [k_{NN}(x^{(i)}, x^{(j)})]_{i,j=1}^{M} \in \mathbb{R}^{M \times M}$. In addition, the hyperparameters, including the noise $\sigma^2_0$ and the kernel parameters $\theta = \{\mu^w[1], \sigma^2_w[1] , \ldots, \mu^w[L+1], \sigma^2_w[L+1], \mu^b[1], \sigma^2_b[1], \ldots, \mu^b[L], \sigma^2_b[L]\}$, can be determined by maximizing the log marginal likelihood as follows

$$
    (\theta, \sigma^2_0) = \arg \max_{\theta, \sigma^2_0} \log p(y|X, \theta, \sigma^2) = \arg \max_{\theta, \sigma^2_0} \left\{ \frac{1}{2}(y - h_{NN}(X; \theta))^T(K(\theta) + \sigma^2_0 I)^{-1}(y - h_{NN}(X; \theta)) - \frac{1}{2} \log |K(\theta) + \sigma^2_0 I| - \frac{M}{2} \log (2\pi) \right\} .
$$

(8)

- **Viewpoint of kernel ridge regression**

  A reproducing kernel Hilbert space $\mathcal{H}_{k_{NN}}(\Omega)$ is induced by the kernel function $k_{NN}$ and provides a completion of the following function space of reproducing kernel map reconstruction $\mathcal{G}$:

$$
    \mathcal{G}_{k_{NN}} = \left\{ f(x) = \sum_{m=1}^{M} \beta_m k_{NN}(x^{(m)}, x) \bigg| M \in \mathbb{N}^+, x^{(m)} \in \Omega, \beta_m \in \mathbb{R} \right\} .
$$

(9)

Note that $\mathcal{G}_{k_{NN}}$ is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{G}_{k_{NN}}}$ defined as follows: if $f(x) = \sum_{m=1}^{M} \beta_m k_{NN}(x^{(m)}, x)$, and $g(x) = \sum_{m'=1}^{M'} \beta_m' k_{NN}(x^{(m')}, x)$, then

$$
    \langle f, g \rangle_{\mathcal{G}_{k_{NN}}} = \sum_{m=1}^{M} \sum_{m'=1}^{M'} \beta_m \beta_m' k_{NN}(x^{(m)}, x^{(m')}). \quad \text{(10)}
$$

(10)
Comparing the posterior mean \( h^*_{\text{NN}}(x) \) with the prior mean \( h_{\text{NN}}(x) \), the correction term, \( \Delta_{\text{NN}}(x) = h^*_{\text{NN}}(x) - h_{\text{NN}}(x) = k_{\text{NN}}(x, x)^T \beta = \sum_{m=1}^{M} \beta_m k_{\text{NN}}(x^{(m)}, x) \), is evidently a reproducing kernel map reconstruction, and thus \( \Delta_{\text{NN}} \in H_{k_{\text{NN}}} \). Here

\[
\beta = [K + \sigma^2_k I_M]^{-1} (y - h_{\text{NN}}(X)) = \arg \min_{\beta \in \mathbb{R}^M} \left\{ \| y - h_{\text{NN}}(X) - \Delta_{\text{NN}}(X) \|_2^2 + \sigma^2_k \| \Delta_{\text{NN}} \|_{H_{k_{\text{NN}}}}^2 \right\},
\]

in which \( \| \Delta_{\text{NN}} \|_{H_{k_{\text{NN}}}}^2 = \beta^T K \beta \), i.e., the combination coefficients of the kernel map reconstruction are determined through a least squares problem regularized by \( \| \cdot \|_{H_{k_{\text{NN}}}}^2 \), often referred to as a kernel ridge regression \([1]\).

### 4 Two-layer neural network

In this section we consider the case when \( L = 1 \). For the sake of conciseness, the notation of first-layer weights and biases is simplified as \( w := w^{[1]} \) and \( b := b^{[1]} \), respectively, and we take \( \mu_w^{[2]} = 0 \) and \( \sigma^2_w^{[2]} = 1 \) without loss of generality.

- **Reproducing kernel Hilbert space**

  In such a case we have \( h_{\text{NN}}(x) = 0 \) and

  \[
k_{\text{NN}}(x, x') = E_{(w,b) \sim \pi} [\phi(w^T x + b) \cdot \phi(w^T x' + b)],
\]

  in which \( k_{\pi} \) denotes the kernel function induced by a two-layer neural network when \((w, b) \sim \pi\). The corresponding reproducing kernel Hilbert space by \( k_{\pi} \) is hence denoted by \( H_{k_{\pi}}(\Omega) \).

  We define for a fixed \( \pi \in P(S_{\Delta_{\pi}}) \) that

  \[
  H_{\pi}(\Omega) := \left\{ f(x) = \int_{S_{\Delta_{\pi}}} \alpha(w, b) \phi(w^T x + b) \, d\pi(w, b) \parallel f \parallel_{H_{\pi}} < \infty \right\},
  \]

  with \( \parallel f \parallel_{H_{\pi}}^2 := E_{(w,b) \sim \pi} [\alpha(w, b)^2] \),

  where \( S_{\Delta_{\pi}} := \{(w, b) \parallel \{w^T, b\}^T \parallel_1 = 1\} \), and \( P(S_{\Delta_{\pi}}) \) denotes the collection of all probability measures on \((S_{\Delta_{\pi}},\mathcal{F})\), \( \mathcal{F} \) being the Borel \( \sigma \)-algebra on \( S_{\Delta_{\pi}} \). It has been shown that \( H_{\pi} = H_{k_{\pi}} \), and \([3, 7]\) are referred to for more details.

- **Barron space**

  Naturally connected with the aforementioned reproducing kernel Hilbert spaces is the Barron space \([3, 4]\) defined as

  \[
  B_2(\Omega) := \left\{ f(x) = \int_{S_{\Delta_{\pi}}} \alpha(w, b) \phi(w^T x + b) \, d\pi(w, b) \parallel f \parallel_{B_2} < \infty \right\},
  \]

  with \( \parallel f \parallel_{B_2}^2 := \inf_{\pi \in P} E_{(w,b) \sim \pi} [\alpha(w, b)^2] \).

  Thus we have

  \[
  B_2(\Omega) = \bigcup_{\pi \in P(S_{\Delta_{\pi}})} H_{\pi}(\Omega) = \bigcup_{\pi \in P(S_{\Delta_{\pi}})} H_{k_{\pi}}(\Omega),
  \]

  i.e., the Barron space \( B_2 \) is the union of a class of reproducing kernel Hilbert spaces \( H_{k_{\pi}} \) that are defined by the neural-network-induced kernels \( k_{\pi} \) through two-layer neural networks.

  Upon \((w, b) \sim \pi \in S_{\Delta_{\pi}}\), we consider an additional random variable \( a \) such that \((a, w, b) \sim \rho, \rho \in P(\mathbb{R} \times S_{\Delta_{\pi}})\), and \( E_{(a,w,b) \sim \rho} [a | w, b] = \alpha(w, b) \). Hence a function \( f \in B_2(\Omega) \) admits the form

  \[
  f(x) = \int_{S_{\Delta_{\pi}}} E_{\rho}[a | w, b] \cdot \phi(w^T x + b) \, d\pi(w, b) = \int_{\mathbb{R} \times S_{\Delta_{\pi}}} \alpha \phi(w^T x + b) \, d\rho(a, w, b).
  \]
A Monte Carlo estimate with $N$ samples of $(a_i, w_i, b_i)$ drawn from $\rho$ is then given as

$$f(x) \approx \frac{1}{N} \sum_{i=1}^{N} a_i \phi(w_i^T x + b_i),$$

which coincides with the expression of a two-layer neural network’s output. Considering that the weights of first-layer neurons are $(a_i/N)$’s whose mean and variance are both proportional to $1/N$, the distribution from which $(a_i/N, w_i, b_i)$’s are drawn almost aligns with the assumptions made in the beginning of Section 2, except that $a_i$’s are not independent of $(w_i, b_i)$’s. In fact such a dependency does not impact the outputs’ being Gaussian processes, as it will only impact the second layer’s infinite sum which does not exist in a two-layer network.

As $N = N_1 \to \infty$, the Monte Carlo (17) can approximate all the functions in $B_2(\Omega)$ as $\rho$ ranging over all possible options, whereas the assumption of infinite width in Section 2 guarantees the applicability of the central limit theorem. The kernel map constructions in all $H_{k_r}$’s that are induced by the network, as given in (15), and the Monte Carlo estimates sampled from all $\rho$’s, as given in (17), will eventually recover the same function space $B_2(\Omega)$.

References

[1] P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. arXiv: 2103.09177, 2021.

[2] R. Durrett. Probability: Theory and Examples. Cambridge university press, 2019.

[3] W. E, C. Ma, and L. Wu. A priori estimates of the population risk for two-layer neural networks. arXiv: 1810.06397, 2018.

[4] W. E, C. Ma, and L. Wu. The Barron space and the flow-induced function spaces for neural network models. Constructive Approximation, pages 1–38, 2021.

[5] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. arXiv: 1711.00165, 2017.

[6] R. M. Neal. Bayesian Learning for Neural Networks. Springer Science & Business Media, 2012.

[7] A. Rahimi and B. Recht. Uniform approximation of functions with random bases. In 2008 46th Annual Allerton Conference on Communication, Control, and Computing, pages 555–561. IEEE, 2008.

[8] G. Strang. Linear Algebra and Learning from Data. Wellesley-Cambridge Press, 2019.

[9] C. K. Williams and C. E. Rasmussen. Gaussian Processes for Machine Learning. MIT press Cambridge, MA, 2006.

A preliminary version of this note was used for the internal discussions within the Dynamics & Deep Learning Reading Club at the University of Twente.