# STAT 946 - Deep Learning Theory Lecture 3

Mufan Li

September 10, 2025

## 1 Neural Network Gaussian Process (NNGP)

- Neal (1995)
- One hidden layer at initialization.

For $\theta = \{W_1, W_0\}$ and $W_{l,ij} \sim N(0,1)$.

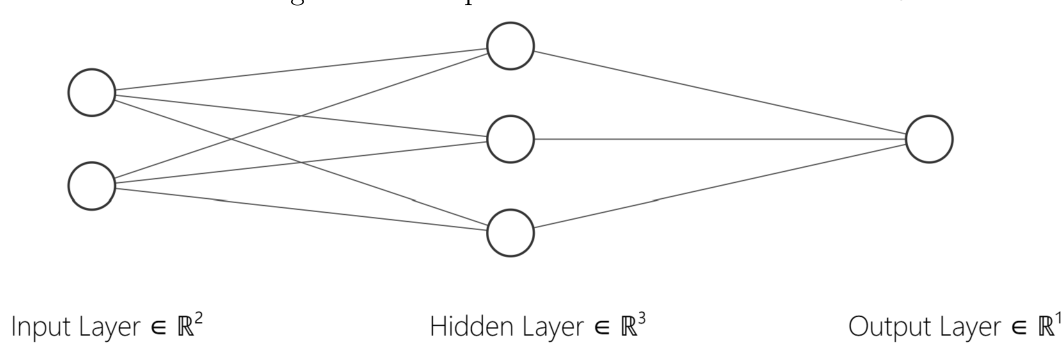
$$\underset{\substack{\downarrow \\ \text{input}}}{f(x;\theta)} = \frac{1}{\sqrt{n}} \underset{\substack{\downarrow \\ 1\times n}}{W_1} \phi\left( \underset{\substack{\downarrow \\ n\times n_0}}{W_0} \underset{\substack{\downarrow \\ n_0\times 1}}{x} \right)$$

Alternatively, we can write

$$f(x;\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{W_{1,i} \phi\left(\langle W_{0,i\cdot}, x\rangle\right)}_{\text{iid with mean } 0, \triangleq Z_i}$$

Where $W_{0,i\cdot}$ is the $i$th row of $W_0$.

Figure 1: Example of a neural network with $n = 3$




Recall that for $\{X\}_n$ with mean 0 and variance $\sigma^2$, $\frac{1}{\sqrt{n}} \sum X_i \xrightarrow{d} N(0, \sigma^2)$ as $n \to \infty$. Hence we have

$$\begin{aligned}\sigma^2 =& \mathbb{E}\left[Z_i^2\right] \\ =& \mathbb{E}\left[\left(W_{1,i}\phi\left(\langle W_{0,i\cdot}, x\rangle\right)\right)^2\right] \\ =& \mathbb{E}\left[W_{1,i}^2\right]\mathbb{E}\left[\phi\left(\langle W_{0,i\cdot}, x\rangle\right)^2\right] \\ =& 1\cdot\mathbb{E}\left[\phi\left(\langle W_{0,i\cdot}, x\rangle\right)^2\right] \\ =& \mathbb{E}\left[\phi\left(\langle W_{0,i\cdot}, x\rangle\right)^2\right]\end{aligned}$$

Note that if $\phi(x) = \max(x, 0)$, then there is closed form formulae (Cho and Saul (2007)). Therefore,

$$\begin{bmatrix} f(x^1;\theta) \\ f(x^2;\theta) \end{bmatrix} \xrightarrow{d} N\left(0, \begin{bmatrix} \sigma(x_1)^2 & ? \\ ? & \sigma(x^2)^2 \end{bmatrix}\right)$$

*Exercise* 1.1. Show all linear combinations $a_1 f(x^1;\theta) + a_2 f(x^2;\theta) \xrightarrow{d} N(\cdots)$.

To find the covariance,

$$\begin{aligned}\mathbb{E}\left[Z_i(x^1)Z_i(x^2)\right] =& 1\times\mathbb{E}\left[\phi\left(\langle W_{0,i\cdot}, x^1\rangle\right)\phi\left(\langle W_{0,i\cdot}, x^2\rangle\right)\right] \\ \triangleq& \Phi(x^1, x^2)\end{aligned}$$

**Definition 1.1** (Gaussian Process)**.** *A random funtion $f : \mathbb{R}^{n_0} \to \mathbb{R}$ is a Gaussian Process (GP) with mean $m : \mathbb{R}^{n_0} \to \mathbb{R}$, covariate kernel $\Phi : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}$ if*

1. *$\Phi$ is symmetric positive semi-definite.*
2. $$[f(x^\alpha)]_{\alpha=1}^m \sim N\left([m(x^\alpha)]_{\alpha=1}^m, \left[\Phi(x^\alpha, x^\beta)\right]_{\alpha,\beta=1}^m\right)$$

**Theorem 1.1** (Neal, 1995)**.** *Assume "$\phi$ is nice". As $n \to \infty$, the nerual network defined above $f(\cdot;\theta) : \mathbb{R}^{n_0} \to \mathbb{R}$, we have*

$$f(\cdot;\theta) \xrightarrow{d} GP(0, \Phi)$$

*For $f \in C(\mathbb{R}^{n_0})$.*

*Remark* 1.1. There exists sequence $\{f_n\}_{n=1}^\infty \in C^{\mathbb{N}}$ and $f_n \to f^*$ but $f^* \notin C$. For example, let

$$f_n(x) = \max(0, \min(1, nx))$$

Which

$$\lim_{n\to\infty} f_n = \begin{cases} 1, & x > 0 \\ 0, & x \le 0 \end{cases}$$

**Definition 1.2** (Stochastic equicontinuity (equivalent to tightness in $C$))**.** *A function $f$ is stochastic equicontinuous if $\exists \alpha, \beta, C > 0 : \forall x^1, x^2 \in \mathbb{R}^{n_0}$*

$$\mathbb{E}\left|f_n(x^1;\theta) - f_n(x^2;\theta)\right|^\alpha \le C|x^1, x^2|^{1+\beta}, \forall n \ge 1$$

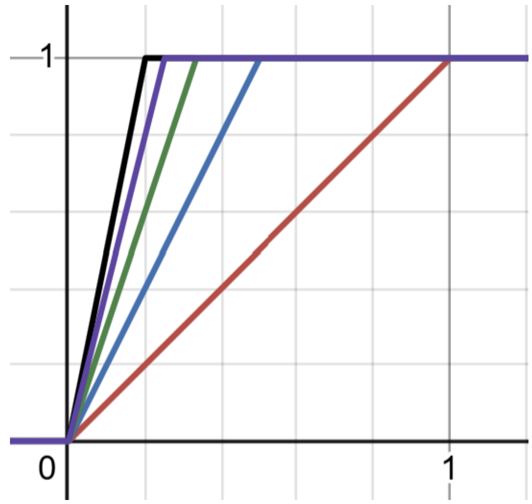



Figure 2: Plot of $f_n$ for $n = 1, 2, 3, 4, 5$

*Exercise* 1.2. Show if $|\phi(u) - \phi(v)| \le C|u - v|^\gamma$ for $\gamma > \frac{1}{2}$ then stochastic equicontinuity of $\phi$ with $\beta = 2\gamma - 1$.

*Remark* 1.2.

- $GP(0, \Phi)$ is a "prior".
- Infinite width network is characterized by one function $\Phi$.
- Neal: "... it may be possible to implement Bayesian inference ... without any for an actual network."

Depth $d > 1$ (finite). (Lee et al. 2017)

$$f(x^\alpha; \theta) = \frac{1}{\sqrt{n}} \underset{\substack{\downarrow \\ 1\times n}}{W_d} \phi(\underset{\substack{\downarrow \\ n\times 1}}{h_d^\alpha})) \quad \text{(MLP)}$$

$$h_{l+1}^\alpha = \frac{1}{\sqrt{n}} \underset{\substack{\downarrow \\ n\times n}}{W_l} \phi(\underset{\substack{\downarrow \\ n\times 1}}{h_l^\alpha}))$$

$$h_1^\alpha = \frac{1}{\sqrt{n_0}} \underset{\substack{\downarrow \\ n\times n_0}}{W_0} \phi(\underset{\substack{\downarrow \\ n_0\times 1}}{x^\alpha}))$$

Alternatively,

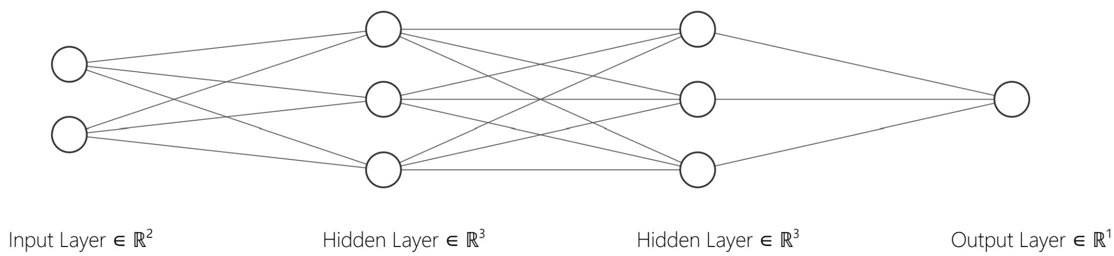
$$f(x^\alpha; \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underset{\substack{\downarrow \\ \text{mean 0, iid}}}{W_{d,i}} \phi(\underset{\substack{\downarrow \\ \text{distribution changes with } n}}{h_{d,i}^\alpha})$$

We would like to study

$$\frac{1}{\sqrt{n}} \sum_i \frac{x_i - \mu_n}{\sigma_n} \to ?$$

However, this requires $x_i \overset{\text{iid}}{\sim} P_n$. One approach is call sequential limit by setting $n_1 \to \infty, n_2 \to \infty$ etc.

Figure 3: Example of a neural network with two hidden layers




## 2 Properties fo Gaussians

- Let

$$\underset{n\times 1}{\underset{\downarrow}{g}} \sim N(\underset{1\times n}{\underset{\downarrow}{\mu}}, \underset{n\times n}{\underset{\downarrow}{\Sigma}})$$

  And $A \in \mathbb{R}^{m\times n}$. Then

$$Ag \sim N(A\mu, A\Sigma A^T)$$

- Let $g \sim N(0, I_n)$ and $u, v \in \mathbb{R}^n$. Then

$$\begin{bmatrix} \langle g, u\rangle \\ \langle g, v\rangle \end{bmatrix} \sim N\left(0, \begin{bmatrix} |u|^2 & \langle u, v\rangle \\ \langle u, v\rangle & |v|^2 \end{bmatrix}\right)$$

- Let $W \in \mathbb{R}^{n\times n}, W_{ij} \overset{\text{iid}}{\sim} N(0,1)$ and $u, v \in \mathbb{R}^n$. Then

$$\begin{bmatrix} Wu \\ Wv \end{bmatrix} \sim N(0, \Phi)$$

  Where

$$\begin{aligned} \Phi &= \left[\begin{array}{ccc:ccc} |u|^2 & & & \langle u, v\rangle & & \\ & \ddots & & & \ddots & \\ & & |u|^2 & & & \langle u, v\rangle \\ \hdashline \langle u, v\rangle & & & |v|^2 & & \\ & \ddots & & & \ddots & \\ & & \langle u, v\rangle & & & |v|^2 \end{array}\right] \\ &= \begin{bmatrix} |u|^2 I_n & \langle u, v\rangle I_n \\ \langle u, v\rangle I_n & |v|^2 I_n \end{bmatrix} \\ &= \begin{bmatrix} |u|^2 & \langle u, v\rangle \\ \langle u, v\rangle & |v|^2 \end{bmatrix} \otimes I_n \end{aligned}$$

  The entries of the covariance matrix can be obtained by

$$\mathbb{E}\langle W_i, u\rangle\langle W_j, u\rangle = |u|^2\delta_{ij}$$

$$\mathbb{E}\langle W_i, u\rangle\langle W_j, v\rangle = \langle u, v\rangle\delta_{ij}$$

**Definition 2.1** (Kronecker Product)**.** *Let* $A = [a_{ij}]_{ij}$*, then*

$$A \otimes B = [a_{ij}B]_{ij} = \begin{bmatrix} a_{11}B & a_{12}B & \cdots \\ a_{21} & a_{22}B & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$