

Machine Learning Theory Lecture 4: Neural Network Gaussian Process and NTK

Richard Xu

October 12, 2021

1 Gaussian Process

We make frequent references to GP, so we talk about it briefly:

1.1 definition

1. \mathcal{GP} is a (potentially infinite) collection of RVs, such that the joint distribution of every finite subset of RVs is multivariate Gaussian:

$$f \sim \mathcal{GP}(\mu(x), \mathcal{K}(x, x')) \quad \text{for any arbitrary } x, x'$$

2. **prior** defined over $p(f|\mathcal{X})$, instead of $p(x)$ over $\mathcal{X} \equiv \{x_1, \dots, x_k\}$

$$p(f|\mathcal{X}) \equiv p\left(\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{bmatrix}\right) = \mathcal{N}(0, K) = \mathcal{N}\left(0, \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_k) \\ \vdots & \ddots & \vdots \\ k(x_k, x_1) & \dots & k(x_k, x_k) \end{bmatrix}\right)$$

1.2 Noisy output setting

in a regression with *noisy output* setting:

$$y_i = f(x_i) + \epsilon_i \quad \epsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$$

1. **joint distribution** $[\mathcal{Y}, y^\star]^\top$, after integrate out f :

$$\begin{aligned} p\left(\begin{bmatrix} \mathcal{Y} \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} \mathcal{X} \\ x^{\star\top} \end{bmatrix}, \sigma_\epsilon^2\right) &= \int p\left(\begin{bmatrix} \mathcal{Y} \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} \mathcal{X} \\ x^{\star\top} \end{bmatrix}, \mathbf{f}\right) p(\mathbf{f}|\mathcal{X}, x^\star) d\mathbf{f} \\ &= \int \mathcal{N}\left(\begin{bmatrix} \mathcal{Y} \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} \mathbf{f}(\mathcal{X}) \\ \mathbf{f}(x^{\star\top}) \end{bmatrix}, \sigma_\epsilon^2 I\right) p(\mathbf{f}|\mathcal{X}, x^\star) d\mathbf{f} \\ &= \mathcal{N}\left(0, \begin{bmatrix} \underbrace{K(\mathcal{X}, \mathcal{X}) + \sigma_\epsilon^2 I}_{\Sigma_{1,1}} & \underbrace{K(\mathcal{X}, x^\star)}_{\Sigma_{1,2}} \\ \underbrace{K(x^\star, \mathcal{X})}_{\Sigma_{2,1}} & \underbrace{K(x^\star, x^\star) + \sigma_\epsilon^2}_{\Sigma_{2,2}} \end{bmatrix}\right) \end{aligned}$$

2. **predictive distribution** of $y^*|\mathcal{Y}$ using conditional formula from the above *joint* multivariate Gaussian:

$$\begin{aligned}
p(y^*|\mathcal{Y}, \mathcal{X}, x^*) &= \mathcal{N}\left(\underbrace{\mathbf{0}}_{\mu_2} + \underbrace{K(x^*, \mathcal{X})}_{\Sigma_{2,1}} \underbrace{(K(\mathcal{X}, \mathcal{X}) + \sigma_\epsilon^2 I)^{-1}(\mathcal{Y} - \underbrace{\mathbf{0}}_{\mu_1})}_{\Sigma_{1,1}^{-1}}, \right. \\
&\quad \left. \underbrace{k(x^*, x^*) + \sigma_\epsilon^2}_{\Sigma_{2,2}} - \underbrace{K(x^*, \mathcal{X})}_{\Sigma_{2,1}} \underbrace{(K(\mathcal{X}, \mathcal{X}) + \sigma_\epsilon^2 I)^{-1}}_{\Sigma_{1,1}^{-1}} \underbrace{K(\mathcal{X}, x^*)}_{\Sigma_{1,2}}\right)
\end{aligned}$$

1.3 noiseless output setting

in a *noiseless output* setting, for example, neural network's read-out layer $f(x_i)$:

$$y_i = f(x_i) \quad (1)$$

1. **joint distribution** of y^* and \mathcal{Y} since *deterministic function* is used, $p([\mathcal{Y}, y^*]^\top)$ no longer need to integrate f :

$$p\left(\begin{bmatrix} \mathcal{Y} \\ y^* \end{bmatrix} \middle| \begin{bmatrix} \mathcal{X} \\ x^{*\top} \end{bmatrix}\right) = p\left(\begin{bmatrix} f(\mathcal{X}) \\ f(x^*) \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} K(\mathcal{X}, \mathcal{X}) & K(\mathcal{X}, x^*) \\ K(x^*, \mathcal{X}) & K(x^*, x^*) \end{bmatrix}\right)$$

replace symbols $x^* \rightarrow x, y^* \rightarrow f(x)$, we have:

$$\begin{aligned}
p\left(\begin{bmatrix} \mathcal{Y} \\ f \end{bmatrix} \middle| \begin{bmatrix} \mathcal{X} \\ \textcolor{red}{x}^\top \end{bmatrix}\right) &= p\left(\begin{bmatrix} f(\mathcal{X}) \\ f(\textcolor{red}{x}) \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} K(\mathcal{X}, \mathcal{X}) + \sigma_\epsilon^2 \mathbf{I} & K(\mathcal{X}, \textcolor{red}{x}) \\ K(\textcolor{red}{x}, \mathcal{X}) & K(\textcolor{red}{x}, \textcolor{red}{x}) \end{bmatrix}\right) \\
&\text{for arbitrary variable } \textcolor{red}{x}
\end{aligned}$$

2. **predictive distribution** of $y^*|\mathcal{Y}$ using conditional formula from the above *joint* multivariate Gaussian:

$$\begin{aligned}
p(y^*|\mathcal{Y}, \mathcal{X}, x^*) &= \mathcal{N}\left(K(x^*, \mathcal{X})K(\mathcal{X}, \mathcal{X})^{-1}\mathcal{Y}, \right. \\
&\quad \left. k(x^*, x^*) - K(x^*, \mathcal{X})K(\mathcal{X}, \mathcal{X})^{-1}K(\mathcal{X}, x^*)\right)
\end{aligned} \quad (2)$$

replace symbols $x^* \rightarrow x, y^* \rightarrow f$, we have:

$$\begin{aligned}
p(f|\mathcal{X}, \mathcal{Y}) &= \mathcal{GP}\left(K(\textcolor{red}{x}, \mathcal{X})K(\mathcal{X}, \mathcal{X})^{-1}\mathcal{Y}, \right. \\
&\quad \left. k(\textcolor{red}{x}, \textcolor{red}{x}) - K(\textcolor{red}{x}, \mathcal{X})K(\mathcal{X}, \mathcal{X})^{-1}K(\mathcal{X}, \textcolor{red}{x})\right)
\end{aligned} \quad (3)$$

2 Kernel methods

consider the equation, where $\phi(\cdot) \in \mathbb{R}^m$:

$$\begin{aligned} y &= \phi(x)^\top \mathbf{w} \\ &= \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_m(x) \end{bmatrix}^\top \mathbf{w} \\ &= [\phi_1(x) \quad \dots \quad \phi_m(x)] \mathbf{w} \end{aligned} \tag{4}$$

using definition:

$$\begin{aligned} \mathcal{Y} &= [y_1, \dots, y_n]^\top \\ \Phi &= [\phi(x_1), \dots, \phi(x_n)]^\top \\ &= \underbrace{\begin{bmatrix} \phi_1(x_1) & \dots & \phi_m(x_1) \\ \vdots & & \vdots \\ \phi_1(x_n) & \dots & \phi_m(x_n) \end{bmatrix}}_{n \times m} \end{aligned} \tag{5}$$

Ridge regression can be re-written as:

$$\begin{aligned} \mathbf{w}^\star &= \arg \min_{\mathbf{w}} \sum_{i=1}^n (y_i - \phi(x_i)^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2 \\ &= \arg \min_{\mathbf{w}} \|\mathcal{Y} - \Phi \mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2 \end{aligned} \tag{6}$$

just like the normal ridge regression, the least-square solution is:

$$\mathbf{w}^\star = (\underbrace{\Phi^\top \Phi}_{m \times m} + \lambda I)^{-1} \Phi^\top \mathcal{Y} \tag{7}$$

substitute \mathbf{w}^\star back to $y = \phi(x)^\top \mathbf{w}$ for a single pair of data,output (x, y) :

$$\begin{aligned} y_{\mathbf{w}^\star}(x) &= \phi(x)^\top \mathbf{w}^\star \\ &= \phi(x)^\top (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top \mathcal{Y} \\ &= \underbrace{\phi(x)^\top \Phi^\top}_{1 \times n} (\underbrace{\Phi \Phi^\top}_{n \times n} + \lambda I)^{-1} \mathcal{Y} \\ &\text{using identity } (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top = \Phi^\top (\Phi \Phi^\top + \lambda I)^{-1} \end{aligned} \tag{8}$$

2.1 Kernel trick

the above looks all good, except we want to avoid computing $\phi(x)$ explicitly, especially when m is large! However, knowing

$$\begin{aligned} [\Phi \Phi^\top]_{i,j} &= \phi(x_i)^\top \phi(x_j) = \mathcal{K}(x_i, x_j) \\ [\phi(x)^\top \Phi^\top]_j &= \phi(x)^\top \phi(x_j) = \mathcal{K}(x, x_j) \end{aligned} \tag{9}$$

we dodged the bullet of computing $\phi(x)$ explicitly!

3 Neural Network Expressivity in Gaussian Process, [1] [2]

3.1 Key takeaway

Elements of pre-activation layer z_k^l of a neural network is i.i.d GP when width tends to infinity:

$$z_k^l(\mathcal{X}) \sim \mathcal{GP}(0, K^l) \quad \forall k$$

$$\text{where } K^l = \sigma_b^2 + \mathbb{E}_{z_1^{l-1}(\mathcal{X}) \sim \mathcal{GP}(0, K^{l-1})} [\phi(z_1^{l-1}(\mathcal{X})) \phi(z_1^{l-1}(\mathcal{X}))^\top] \quad N_l \rightarrow \infty \quad (10)$$

3.2 Neutral network with Gaussian initialization

$$z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \times \phi(z_j^{l-1}(x)) \quad W_{k,j}^l \sim \mathcal{N}\left(0, \frac{1}{\sqrt{N_l}}\right) \quad b_k^l \sim \mathcal{N}(0, \sigma_b) \quad \text{or :} \quad (11)$$

$$z_k^l(x) = \sigma_b b_k^l + \sum_{j=1}^{N_l} \frac{1}{\sqrt{N_l}} W_{k,j}^l \times \phi(z_j^{l-1}(x)) \quad W_{k,j}^l \sim \mathcal{N}(0, 1) \quad b_k^l \sim \mathcal{N}(0, 1)$$

3.3 pre-activation layer 1

putting in data $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$, we have:

$$z^1(\mathbf{x}) = \begin{bmatrix} z_1^1 \\ \vdots \\ z_{N_2}^1 \end{bmatrix} = \begin{bmatrix} W_{1,1}^1 & \cdots & W_{1,d_{\text{in}}}^1 \\ \vdots & \ddots & \vdots \\ W_{N_2,1}^1 & \cdots & W_{N_2,d_{\text{in}}}^1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_{d_{\text{in}}} \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_{N_2} \end{bmatrix} \quad (12)$$

similarly, we can have another expression for $\mathbf{x}' \in \mathbb{R}^d$

$$z^1(\mathbf{x}') = \begin{bmatrix} z_1^1 \\ \vdots \\ z_{N_2}^1 \end{bmatrix} = \begin{bmatrix} W_{1,1}^1 & \cdots & W_{1,d_{\text{in}}}^1 \\ \vdots & \ddots & \vdots \\ W_{N_2,1}^1 & \cdots & W_{N_2,d_{\text{in}}}^1 \end{bmatrix} \begin{bmatrix} x'_1 \\ \vdots \\ x'_{d_{\text{in}}} \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_{N_2} \end{bmatrix} \quad (13)$$

Obviously, regardless if we use (\mathbf{x}, \mathbf{x}) or $(\mathbf{x}, \mathbf{x}')$, when $k \neq k'$:

$$\begin{cases} \text{Cov}(z_k^1(\mathbf{x}), z_{k'}^1(\mathbf{x})) &= 0 \\ \text{Cov}(z_k^1(\mathbf{x}), z_{k'}^1(\mathbf{x}')) &= 0 \end{cases} \quad \forall k \neq k' \quad (14)$$

3.3.1 $p(z_k^1(\mathbf{x}))$

$$z_k^1(\mathbf{x}) = \sum_{j=1}^{d_{\text{in}}} W_{k,j}^1 x_j + b_k = \sum_{j=1}^{N_1} W_{k,j}^1 x_j + b_k$$

$$\implies z_k^1(\mathbf{x}) \sim \mathcal{N}\left(0, \sigma_b^2 + \sum_{j=1}^{N_1} \left(\frac{1}{\sqrt{N_1}} x_j\right)^2\right) \quad (15)$$

$$= \mathcal{N}\left(0, \sigma_b^2 + \frac{1}{N_1} \sum_{j=1}^{N_1} x_j^2\right) = \mathcal{N}\left(0, \sigma_b^2 + \frac{1}{N_1} \mathbf{x}^\top \mathbf{x}\right)$$

similarly,

$$z_k^1(\mathbf{x}') \sim \mathcal{N}\left(0, \sigma_b^2 + \frac{1}{N_1} \mathbf{x}'^\top \mathbf{x}'\right) \quad (16)$$

and **co-variance** would be:

$$\begin{aligned} \text{Cov}(z_k^1(\mathbf{x}), z_k^1(\mathbf{x}')) &= \mathbb{E}\left[\left(\sum_{j=1}^{N_1} W_{k,j}^1 x_j + b_k\right) \left(\sum_{j=1}^{N_1} W_{k,j}^1 x'_j + b_k\right)\right] \\ &= \sum_{j=1}^{N_1} \mathbb{E}[(W_{k,j}^1)^2] x_j x_j + \sum_{j=1}^{N_1} \sum_{i \neq j} \mathbb{E}[W_{k,j}^1] \mathbb{E}[W_{k,i}^1] x_j x'_i + b_k^2 \\ &= \sigma_b^2 + \frac{1}{N_1} \mathbf{x}^\top \mathbf{x}' \end{aligned} \quad (17)$$

for any pairs of data \mathbf{x} and \mathbf{x}' , we have, $\forall k$:

$$\begin{bmatrix} z_k^1(\mathbf{x}) \\ z_k^1(\mathbf{x}') \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_b^2 + \frac{1}{N_1} \mathbf{x}^\top \mathbf{x} & \sigma_b^2 + \frac{1}{N_1} \mathbf{x}^\top \mathbf{x}' \\ \sigma_b^2 + \frac{1}{N_1} \mathbf{x}^\top \mathbf{x}' & \sigma_b^2 + \frac{1}{N_1} \mathbf{x}'^\top \mathbf{x}' \end{bmatrix}\right) \quad (18)$$

$$z_k^1(\mathcal{X}) \sim \mathcal{GP}(0, K^1) \quad \text{where} \quad K^1(\mathbf{x}, \mathbf{x}') = \sigma_b^2 + \frac{1}{N_1} \mathbf{x}^\top \mathbf{x}' \quad (19)$$

3.3.2 adding activation ϕ :

$$\phi(z^1(\mathbf{x})) = \begin{bmatrix} \phi(z_1^1) \\ \vdots \\ \phi(z_k^1) \end{bmatrix} \quad (20)$$

It's difficult to tell what distribution this is

3.4 pre-activation layer l

$$z^l(\mathbf{x}) = \begin{bmatrix} z_1^l \\ \vdots \\ z_{N_{l+1}}^l \end{bmatrix} = \begin{bmatrix} W_{1,1}^l & \cdots & W_{1,N_l}^l \\ \vdots & \ddots & \vdots \\ W_{N_{l+1},1}^l & \cdots & W_{N_{l+1},N_l}^l \end{bmatrix} \begin{bmatrix} \phi(z_1^{l-1}(\mathbf{x})) \\ \vdots \\ \phi(z_{N_l}^{l-1}(\mathbf{x})) \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_{N_{l+1}} \end{bmatrix} \quad (21)$$

similarly, we can have:

$$z^l(\mathbf{x}') = \begin{bmatrix} z_1^l \\ \vdots \\ z_{N_{l+1}}^l \end{bmatrix} = \begin{bmatrix} W_{1,1}^l & \cdots & W_{1,N_l}^l \\ \vdots & \ddots & \vdots \\ W_{N_{l+1},1}^l & \cdots & W_{N_{l+1},N_l}^l \end{bmatrix} \begin{bmatrix} \phi(z_1^{l-1}(\mathbf{x}')) \\ \vdots \\ \phi(z_{N_l}^{l-1}(\mathbf{x}')) \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_{N_{l+1}} \end{bmatrix} \quad (22)$$

for a specific k^{th} row:

$$z_k^l(\mathbf{x}) = \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x})) + b_k^l \quad (23)$$

3.4.1 Prove $\text{Cov}(z_k^l(\mathbf{x}), z_{k'}^l(\mathbf{x})) = 0$ by induction

This fact is surprising as both $z_k^l(\mathbf{x})$ and $z_{k'}^l(\mathbf{x})$ share the same $z^{l-1}(\mathbf{x})$ or $\phi(z^{l-1}(\mathbf{x}))$! However, we can prove by induction. Firstly, we see that $z_k^1(\mathbf{x})$ and $z_{k'}^1(\mathbf{x})$ are independent.

Assume z_i^{l-1} and z_j^{l-1} are i.i.d Gaussian Processes, i.e., $z_j^{l-1} \stackrel{\text{i.i.d}}{\sim} \mathcal{GP}(0, K^{l-1})$ and hence $z_i^{l-1}(\mathbf{x})$ and $z_{j, j \neq i}^{l-1}(\mathbf{x})$ are independent too. Then, forward looking in Eq.(26), if we can prove that:

$$z_k^l = \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x})) + b_k \xrightarrow{d} \mathcal{N}\left(0, \text{Var}[\phi(z_1^{l-1}(\mathbf{x}))] + \sigma_b^2\right) \quad (24)$$

You see that the RHS $z_1^{l-1}(\mathbf{x})$ is an arbitrary random variable, CLT made it independent of the actual values of $\{z_j^{l-1}(\mathbf{x})\}$. Therefore $z_k^l(\mathbf{x})$ and $z_{k'}^l(\mathbf{x})$ are independent. It also implies that $z_k^{l-1}(\mathbf{x})$ and $z_{k'}^{l-1}(\mathbf{x})$ are independent.

3.4.2 marginal $p(z_k^l(\mathbf{x}))$

problem is due to non-linearity of $\phi(z_j^{l-1}(\mathbf{x}))$, we do not know what distribution $z_k^l(\mathbf{x})$ is!

However, let's look at an individual term inside the sum: $\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x}))$

$$\begin{aligned} \mathbb{E}[W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}))] &= 0 \\ \text{Var}[W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}))] &= \mathbb{E}[(W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x})))^2] \\ &= \mathbb{E}[(W_{k,1}^l)^2] \mathbb{E}[\phi(z_1^{l-1}(\mathbf{x}))^2] \\ &= \frac{1}{N_l} \text{Var}[\phi(z_1^{l-1}(\mathbf{x}))] \end{aligned} \quad (25)$$

Since $\text{Var}[\phi(z_j^{l-1}(\mathbf{x}))]$ can be chosen to be bounded, and each $W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x}))$ to be i.i.d, (from section 3.4.1), so we can apply CLT and let $N_l \rightarrow \infty$:

$$\begin{aligned} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x})) &\xrightarrow{d} \mathcal{N}\left(0, \text{Var}[W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}))] N_l\right) \\ &\xrightarrow{d} \mathcal{N}\left(0, \frac{1}{N_l} \text{Var}[\phi(z_1^{l-1}(\mathbf{x}))] N_l\right) \quad \text{substitute Eq.(25)} \\ &\xrightarrow{d} \mathcal{N}\left(0, \text{Var}[\phi(z_1^{l-1}(\mathbf{x}))]\right) \end{aligned} \quad (26)$$

3.4.3 joint density $p(z_k^l(\mathbf{x}), z_{k'}^l(\mathbf{x}'))$

Here we use a multivariate version of CLT where each i.i.d team inside the sum is a vector: $\begin{bmatrix} z_k^l(\mathbf{x}) \\ z_{k'}^l(\mathbf{x}') \end{bmatrix}$:

looking at the part without bias term b_k :

$$\left[\sum_{j=1}^{N_l} W_{k,1}^l \phi(z_j^{l-1}(\mathbf{x})) \right] \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Sigma \left(\begin{bmatrix} W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x})) \\ W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}')) \end{bmatrix} \right) N_l\right) \quad (27)$$

use the notation for zero-meaned R.V:

$$\Sigma \left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right) = \mathbb{E} \left[\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \begin{bmatrix} y_1 & y_2 \end{bmatrix} \right] = \begin{bmatrix} \text{Var}(y_1) & \text{Cov}(y_1, y_2) \\ \text{Cov}(y_1, y_2) & \text{Var}(y_2) \end{bmatrix} \quad (28)$$

We already know the variance (diagonal) from Eq.(26). How about the co-variance (off-diagonal) term:
 $\text{Cov}(y_1, y_2) \equiv \text{Cov}[W_{k,1}^{l-1}\phi(z_1^{l-1}(\mathbf{x})), W_{k,1}^{l-1}\phi(z_1^{l-1}(\mathbf{x}'))]:$

$$\begin{aligned} & \text{Cov}[W_{k,1}^l\phi(z_1^{l-1}(\mathbf{x})), W_{k,1}^l\phi(z_1^{l-1}(\mathbf{x}'))] \\ &= \mathbb{E}[W_{k,1}^l\phi(z_1^{l-1}(\mathbf{x})) W_{k,1}^l\phi(z_1^{l-1}(\mathbf{x}'))] \\ &= \mathbb{E}[(W_{k,1}^l)^2] \mathbb{E}[\phi(z_1^{l-1}(\mathbf{x}))\phi(z_1^{l-1}(\mathbf{x}'))] \quad \text{we didn't need } \mathbb{E} \text{ for } \mathbf{x}^\top \mathbf{x}' \text{ as in Eq.(17)} \\ &= \frac{1}{N_l} \mathbb{E}[\phi(z_1^{l-1}(\mathbf{x}))\phi(z_1^{l-1}(\mathbf{x}'))] \end{aligned} \quad (29)$$

therefore, canceling out $\frac{1}{N_l} \times N_l$ and add σ_b^2 to each of the entries.

It is important to note that σ_b^2 also appears in the off-diagonal entries as well as the diagonal entry.

$$\begin{aligned} & \begin{bmatrix} z_k^l(\mathbf{x}) = b_k + \sum_{j=1}^{N_l} W_{k,1}^l\phi(z_j^{l-1}(\mathbf{x})) \\ z_k^l(\mathbf{x}') = b_k + \sum_{j=1}^{N_l} W_{k,1}^l\phi(z_j^{l-1}(\mathbf{x}')) \end{bmatrix} \xrightarrow{d} \\ & \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(\mathbf{x}))\phi(z_1^{l-1}(\mathbf{x}))] & \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(\mathbf{x}))\phi(z_1^{l-1}(\mathbf{x}'))] \\ \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(\mathbf{x}))\phi(z_1^{l-1}(\mathbf{x}'))] & \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(\mathbf{x}'))\phi(z_1^{l-1}(\mathbf{x}'))] \end{bmatrix}\right) \end{aligned} \quad (30)$$

3.4.4 Relationship with Gaussian Process (GP):

let $f(x) \equiv z_k^l(x)$ be some function, and since for every arbitrary point pair, x and x' , we have:

$$\begin{aligned} & \begin{bmatrix} f(x) \\ f(x') \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(x, x) & K(x, x') \\ K(x, x') & K(x', x') \end{bmatrix}\right) \\ & \implies f \sim \mathcal{GP}(0, \mathbf{K}) \end{aligned} \quad (31)$$

looking at mean and co-variance as $N_l \rightarrow \infty$, for each x, x' pair:

$$\begin{aligned} & \begin{bmatrix} z_k^l(x) \\ z_k^l(x') \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K^l(x, x) & K^l(x, x') \\ K^l(x, x') & K^l(x', x') \end{bmatrix}\right) \\ & \text{marginal } z_k^l(x) \xrightarrow{d} \mathcal{N}(0, \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(x))^2]) \quad \text{as } N_l \rightarrow \infty \\ & \text{where: } \text{Cov}[z_k^l(x), z_k^l(x')] = K^l(x, x') = \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(x)) \times \phi(z_1^{l-1}(x'))] \end{aligned} \quad (32)$$

putting it in layer specific GP define over some domain \mathcal{X} as $N_l \rightarrow \infty$:

$$\implies z_k^l(\mathcal{X}) \sim \mathcal{GP}(0, K^l) \quad (33)$$

The recursion tells us $z_1^{l-1}(\mathcal{X}) \sim \mathcal{GP}(0, K^{l-1})$. Remove the suffix $z_1 \rightarrow z$:

$$\begin{aligned} & \implies z_k^l(\mathcal{X}) \sim \mathcal{GP}(0, K^l) \quad \forall k \\ & \text{where } K^l = \sigma_b^2 + \mathbb{E}_{z_1^{l-1}(\mathcal{X}) \sim \mathcal{GP}(0, K^{l-1})}[\phi(z_1^{l-1}(\mathcal{X}))\phi(z_1^{l-1}(\mathcal{X}))^\top] \end{aligned} \quad (34)$$

4 NTK at initialization [3]

4.1 Key takeaway

$$\Theta_{k,k'}^l(x, x') \xrightarrow{N_{l+1} \rightarrow \infty} \Theta_{\infty}^l(x, x') \delta_{k,k'} \quad (35)$$

$$\Theta^l(x, x') = \underbrace{\left(K^l(x, x') + \dot{K}^l(x, x') \Theta_{\infty}^{l-1}(x, x') \right)}_{\text{scalar}} \otimes_{\text{outer}} \underbrace{\mathbf{I}_{N_{l+1} \times N_{l+1}}}_{\text{same value for all } k, k' \text{ pairs}} \quad (36)$$

4.2 expression of NTK

at layer l :

$$\begin{aligned} \Theta_{k,k'}^l(x, x') &= \frac{\partial z_k^l(x, \theta)}{\partial \theta^l} \frac{\partial z_{k'}^l(x', \theta)}{\partial \theta^l} \\ &= \sum_i^{| \theta |} \frac{\partial z_k^l(x, \theta)}{\partial \theta_i^l} \frac{\partial z_{k'}^l(x', \theta)}{\partial \theta_i^l} \end{aligned} \quad (37)$$

why do you think it's called neural tangent kernel?

4.3 re-parameterized formulation

different to NNGP, we now write neural network expression as:

$$\begin{aligned} \text{NNGP} \quad z_k^l(x) &= \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l & W_{k,j}^l &\sim \mathcal{N}\left(0, \frac{1}{\sqrt{N_l}}\right) \\ \text{in NTK we use re-parameterization} \quad z_k^l(x) &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l & W_{k,j}^l &\sim \mathcal{N}(0, 1) \quad \sigma_b \sim \mathcal{N}(0, 1) \end{aligned} \quad (38)$$

Given a single input x , we show the following is the relationship between two adjacent layers $z^{l-1}(x) \rightarrow z^l(x)$:

$$\begin{bmatrix} z_1^l(x) \\ \vdots \\ z_k^l(x) \\ \vdots \\ z_{N_{l+1}}^l(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{1,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_1^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{N_{l+1},j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_{N_{l+1}}^l \end{bmatrix} \quad (39)$$

4.4 Prove by Induction

4.4.2 For NTK

we need to show by induction:

1. assume for a small network, at $l = 1$ we prove:

$$\Theta_{k,k'}^1(x, x') = \underbrace{\left(\frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2 \right)}_{K^1} \delta_{k,k'} \quad (40)$$

even better, no need to show: $\Theta_{k,k'}^1(x, x') \rightarrow K^1 \delta_{k,k'}$. it is actually equal! Besides there is no N_1 to take limit to ∞

2. then by assuming:

$$\Theta_{k,k'}^{l-1}(x, x') = \frac{\partial z_k^{l-1}(x, \theta)}{\partial \theta^l}^\top \frac{\partial z_k^{l-1}(x', \theta)}{\partial \theta^l} \xrightarrow{N_l \rightarrow \infty} \Theta_\infty^{l-1}(x, x') \delta_{k,k'} \quad (41)$$

we can prove:

$$\Theta_{k,k'}^l(x, x') = \frac{\partial z_k^l(x, \theta)}{\partial \theta^l}^\top \frac{\partial z_k^l(x', \theta)}{\partial \theta^l} \xrightarrow{N_{l+1} \rightarrow \infty} \Theta_\infty^l(x, x') \delta_{k,k'} \quad (42)$$

4.5 when $l = 1$: $\Theta_{k,k'}^1(x, x') = \left(\frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2 \right) \delta_{k,k'}$

From the Eq.(39), we have:

$$\begin{bmatrix} \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W_{1,j}^1 x_1 + \sigma_b b_1^1 \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W_{k,j}^1 x_2 + \sigma_b b_k^1 \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W_{N_2,j}^1 x_{d_{\text{in}}} + \sigma_b b_{N_2}^1 \end{bmatrix} = \begin{bmatrix} z_1^1(x) \\ \vdots \\ z_k^1(x) \\ \vdots \\ z_{N_2}^1(x) \end{bmatrix} \quad (43)$$

note when computing $\frac{\partial z_k^1(x)}{\partial W_{i,j}^1}$ only k^{th} row going to return a gradient, i.e., $\frac{\partial z_k^1(x)}{\partial W_{i,j}^1} = 0$ if $i \neq k$, and the gradient correspond to $\frac{\partial z_k^1(x)}{\partial W_{i,j}^1}$ is x_j :

$$\begin{aligned} \frac{\partial z_k^1(x)}{\partial W_{i,j}^1} &= \begin{cases} \frac{1}{\sqrt{d_{\text{in}}}} x_j & \text{if } i = k \text{ i.e., row } k \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{1}{\sqrt{d_{\text{in}}}} \delta_{i,k} x_j \\ \Rightarrow \frac{\partial z_{k'}^1(x)}{\partial W_{i,j}^1} &= \frac{1}{\sqrt{d_{\text{in}}}} \delta_{i,k'} x_j \end{aligned} \quad (44)$$

now, taking pair of data x and x' , each element of the outer product matrix $\Theta^l(x, x') = \sum_{d=1}^{|\theta|} \frac{\partial F_k^l(x)}{\partial \theta_d} \otimes \frac{\partial F_{k'}^l(x')}{\partial \theta_d}$.

The individual element of $\Theta^l(x, x')$ at k, k' is:

$$\begin{aligned}
\Theta_{k,k'}^1(x, x') &= \sum_{d=1}^{|\theta^1|} \frac{\partial F_k^1(x)}{\partial \theta_d^1} \frac{\partial F_{k'}^1(x')}{\partial \theta_d^1} \quad \theta^1 = \{W^1, b^1\} \\
&= \sum_{d=1}^{|W^1|} \frac{\partial F_k^1(x)}{\partial W_d^1} \frac{\partial F_{k'}^1(x')}{\partial W_d^1} + \sum_{d=1}^{|b^1|} \frac{\partial F_k^1(x)}{\partial b_d^1} \frac{\partial F_{k'}^1(x')}{\partial b_d^1} \\
&= \sum_{i=1}^{N_2} \sum_{j=1}^{d_{\text{in}}} \frac{\partial z_k^1(x)}{\partial W_{i,j}} \frac{\partial z_{k'}^1(x')}{\partial W_{i,j}} + \sum_{i=1}^{N_2} \frac{\partial z_k^1(x)}{\partial b_i} \frac{\partial z_{k'}^1(x')}{\partial b_i} \\
&= \sum_{i=1}^{N_2} \sum_{j=1}^{d_{\text{in}}} \frac{1}{\sqrt{d_{\text{in}}}} x_j \delta_{i,k'} \frac{1}{\sqrt{d_{\text{in}}}} x'_j \delta_{i,k} + \sum_{i=1}^{N_2} \sigma_b \delta_{i,k} \sigma_b \delta_{i,k'} \quad \text{only one } i \in \{1, \dots, N_2\} \text{ in outer sum remain} \\
&= \sum_{j=1}^{d_{\text{in}}} \frac{1}{d_{\text{in}}} x_j x'_j \delta_{k,k'}^2 + \sigma_b^2 \delta_{k,k'} \quad \delta_{i,k'} \delta_{i,k} = \delta_{k,k'} \\
&= \frac{1}{d_{\text{in}}} x^\top x' \delta_{k,k'} + \sigma_b^2 \delta_{k,k'} \\
&= \underbrace{\left(\frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2 \right)}_{K^1} \delta_{k,k'} \\
&\equiv K^1(x, x') \delta_{k,k'}
\end{aligned} \tag{45}$$

4.5.1 structure of $\Theta^1(x, x')$

now we have each element $\Theta_{k,k'}^1(x, x')$, the final $\Theta^1(x, x')$ is:

$$\begin{aligned}
\Rightarrow \Theta^1(x, x') &= \left[\underbrace{\begin{bmatrix} K^1(x, x') & \dots & 0 & \dots & 0 \\ 0 & K^1(x, x') & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & K^1(x, x') & 0 \\ 0 & 0 & 0 & 0 & K^1(x, x') \end{bmatrix}}_{k \in \{1, \dots, N_2\}} \right]_{k' \in \{1, \dots, N_2\}} \\
&= \text{repeating diagonal with } K^1(x, x') \delta_{k,k'} \\
&= \underbrace{K^1(x, x')}_{\text{scalar}} \otimes_{\text{outer}} \mathbf{I}_{N_1 \times N_2}
\end{aligned} \tag{46}$$

4.6 when $l > 1$

$$\begin{bmatrix} z_1^l(x) \\ \vdots \\ z_k^l(x) \\ \vdots \\ z_{N_l+1}^l(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{1,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_1^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{N_l+1,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_{N_l+1}^l \end{bmatrix} \quad (47)$$

split sum into two parts: $\{W^l, b^l\}$ and θ^{l-1}

$$\begin{aligned} \Theta_{k,k'}^l(x, x') &= \sum_{d=1}^{|\theta^l|} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}} \\ &= \underbrace{\sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^l(x)}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x')}{\partial \{W^l, b^l\}}}_{\textcircled{1}} + \underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}}}_{\textcircled{2}} \end{aligned} \quad (48)$$

4.6.2 Expression for $\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}}$

in expression $\underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}}}_{\textcircled{2}}:$

derivatives with respect to the single terms: $\frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}}$

$$\begin{aligned} z_k^l &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi\left(\frac{1}{\sqrt{N_{l-1}}} \sum_{i=1}^{N_{l-1}} W_{j,i}^{l-1} \phi(z_i^{l-2}(x)) + \sigma_b b_j^{l-1}\right) + \sigma_b b_k^l \end{aligned} \quad (49)$$

$$\begin{aligned} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} &= \frac{\partial z_k^l(x)}{\partial \phi(z^{l-1}(x))} \frac{\partial \phi(z^{l-1}(x))}{\partial z^{l-1}(x)} \frac{\partial z^{l-1}(x)}{\partial \theta_d^{l-1}} \quad \text{drop index for the last two terms} \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \frac{\partial \phi(z_j^{l-1}(x))}{\partial z_j^{l-1}(x)} \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \dot{\phi}(z_j^{l-1}(x)) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \quad \text{leave last derivative as is, in "recursion"} \end{aligned} \quad (50)$$

substitute it back to (2)

$$\begin{aligned}
& \underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}}}_{(2)} \\
&= \sum_{d=1}^{|\theta^{l-1}|} \left(\frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \dot{\phi}(z_j^{l-1}(x)) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \right) \times \left(\frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k',j}^l \dot{\phi}(z_j^{l-1}(x')) \frac{\partial z_j^{l-1}(x')}{\partial \theta_d^{l-1}} \right) \quad \text{by substitution}
\end{aligned} \tag{51}$$

although it looks like it is in the form of Section[??], however, $\underbrace{W_{k,j}^l \dot{\phi}(z_j^{l-1}(x)) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}}}_{\text{is not}}$

independent of $\underbrace{W_{k',j'}^l \dot{\phi}(z_{j'}^{l-1}(x')) \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}}}_{\text{for } j \neq j', \text{ therefore:}}$

$$\begin{aligned}
&= \sum_{d=1}^{|\theta^{l-1}|} \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} \left(W_{k,j}^l \dot{\phi}(z_j^{l-1}(x)) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \right) \times \underbrace{\left(W_{k',j'}^l \dot{\phi}(z_{j'}^{l-1}(x')) \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}} \right)}_{j \rightarrow j' \text{ in second term}} \quad \text{re-arrange} \\
&= \sum_{d=1}^{|\theta^{l-1}|} \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \dot{\phi}(z_j^{l-1}(x)) \dot{\phi}(z_{j'}^{l-1}(x')) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}} \quad \text{re-arrange} \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \dot{\phi}(z_j^{l-1}(x)) \dot{\phi}(z_{j'}^{l-1}(x')) \underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}}}_{\text{definition } \Theta_{j,j'}^{l-1}(x,x')} \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \dot{\phi}(z_j^{l-1}(x)) \dot{\phi}(z_{j'}^{l-1}(x')) \Theta_{j,j'}^{l-1}(x,x') \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \dot{\phi}(z_j^{l-1}(x)) \dot{\phi}(z_{j'}^{l-1}(x')) \Theta_{\infty}^{l-1}(x,x') \delta_{j,j'} \\
&\quad \text{use induction assumption: } \Theta_{j,j'}^{l-1}(x,x') \rightarrow \underbrace{\Theta_{\infty}^{l-1}(x,x') \delta_{j,j'}}_{\text{deterministic and diagonal limit}} \\
&= \Theta_{\infty}^{l-1}(x,x') \frac{1}{N_l} \sum_{j=1}^{N_l} W_{k,j}^l W_{k',j}^l \dot{\phi}(z_j^{l-1}(x)) \dot{\phi}(z_j^{l-1}(x')) \quad \text{only terms remain are } j = j'
\end{aligned} \tag{52}$$

instead of using CLT, we shall apply LoLN here:

$$\begin{aligned}
& \Theta_{\infty}^{l-1}(x, x') \frac{1}{N_l} \sum_{j=1}^{N_l} W_{k,j}^l W_{k',j}^l \dot{\phi}(z_j^{l-1}(x)) \dot{\phi}(z_j^{l-1}(x')) \\
&= \underbrace{\Theta_{\infty}^{l-1}(x, x') \mathbb{E}_{W_{k,1}^l, W_{k',1}^l, z_1^{l-1}(x), z_1^{l-1}(x')} \left[W_{k,1}^l, W_{k',1}^l \dot{\phi}(z_1^{l-1}(x)) \dot{\phi}(z_1^{l-1}(x')) \right]}_{\text{very similar to NNGP}} \\
&= \Theta_{\infty}^{l-1}(x, x') \mathbb{E}_{(z_1^{l-1}(x), z_1^{l-1}(x'))} \left[\dot{\phi}(z_1^{l-1}(x)) \dot{\phi}(z_1^{l-1}(x')) \right] \mathbb{E}_{W_{k,1}^l, W_{k',1}^l} [W_{k,1}^l W_{k',1}^l] \\
&= \Theta_{\infty}^{l-1}(x, x') \mathbb{E}_{z^{l-1} \sim \mathcal{GP}(0, K^{l-1})} \left[\dot{\phi}(z_1^{l-1}(x)) \dot{\phi}(z_1^{l-1}(x')) \right] \delta_{k,k'} \\
&= \delta_{k,k'} \dot{K}^l(x, x') \Theta_{\infty}^{l-1}(x, x')
\end{aligned} \tag{53}$$

1. Derivation of $\delta_{k,k'}$ part:

$$\begin{aligned}
\mathbb{E}_{W_{k,1}^l, W_{k',1}^l} [W_{k,1}^l W_{k',1}^l] &= \begin{cases} \mathbb{E}[W_{k,1}^l W_{k',1}^l] & k \neq k' \\ \mathbb{E}[(W_{k,1}^l)^2] & k = k' \end{cases} \\
&= \begin{cases} 0 & k \neq k' \\ 1 & k = k' \end{cases} \quad \text{re-parameterized expression} \quad W_{k,1}^l \sim \mathcal{N}(0, 1) \\
&= \delta_{k,k'}
\end{aligned} \tag{54}$$

2. notice the expression here:

$$\frac{1}{N_l} \sum_{j=1}^{N_l} W_{k,j}^l W_{k',j}^l \dot{\phi}(z_j^{l-1}(x)) \dot{\phi}(z_j^{l-1}(x')) \tag{55}$$

is the very similar of NNGP formulation, except:

$$\phi(z_j^{l-1}(x)) \rightarrow \dot{\phi}(z_j^{l-1}(x)) \tag{56}$$

so expect same CLT/LoLN treatment applies here

3. looking at abbreviation symbol $\dot{K}^l(x, x')$:

$$\begin{aligned}
\dot{K}^l(x, x') &= \sigma_w^2 \mathbb{E}_{(z_1^{l-1}(x), z_1^{l-1}(x')) \sim \mathcal{N}(0, K^{l-1}(x, x'))} \left[\dot{\phi}(z_1^{l-1}(x)) \dot{\phi}(z_1^{l-1}(x')) \right] \\
&= \mathbb{E}_{(z_1^{l-1}(x), z_1^{l-1}(x')) \sim \mathcal{N}(0, K^{l-1}(x, x'))} \left[\dot{\phi}(z_1^{l-1}(x)) \dot{\phi}(z_1^{l-1}(x')) \right] \quad \text{assume } \sigma_w = 1
\end{aligned} \tag{57}$$

compare with Eq. (??) the recursion in NNGP:

$$K^l(x, x') = \sigma_b^2 + \sigma_w^2 \mathbb{E}_{(z_1^{l-1}(x), z_1^{l-1}(x')) \sim \mathcal{N}(0, K^{l-1}(x, x'))} \left[\phi(z_1^{l-1}(x)) \phi(z_1^{l-1}(x')) \right] \tag{58}$$

note $\dot{K}^l(x, x')$ is **not** a recursion, and $K^l(x, x')$ is expressed in recursion

4. note $\delta_{k,k'} \dot{K}^l(x, x') \Theta_{\infty}^{l-1}(x, x')$ is a scalar, in particular $\dot{K}^l(x, x')$ is a scalar. However, $\Theta(x, x')$ is the constructed matrix, where elements are of $\dot{K}^l(x, x')$

4.6.3 Expression for $\sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^1(x)}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x')}{\partial \{W^l, b^l\}}$

in expression $\underbrace{\sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^1(x)}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x')}{\partial \{W^l, b^l\}}}_{\textcircled{1}}:$

$$\sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^l(x)}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x')}{\partial \{W^l, b^l\}} \quad (59)$$

and compare that with for $l = 1$:

$$\begin{aligned} \sum_{d=1}^{|\theta^1|} \frac{\partial z_k^1(x)}{\partial \theta_d^1} \frac{\partial z_{k'}^1(x')}{\partial \theta_d^1} \quad \theta^1 = \{W^1, b^1\} \\ = \left(K^1(x, x') \equiv \frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2 \right) \delta_{k, k'} \end{aligned} \quad (60)$$

then, we do know:

$$\begin{aligned} \sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^l(x)}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x')}{\partial \{W^l, b^l\}} \\ = \left(K^l(x, x') \equiv \frac{1}{N_l} \phi(z^l(x))^\top \phi(z^l(x')) + \sigma_b^2 \right) \delta_{k, k'} \end{aligned} \quad (61)$$

4.6.4 putting all together

$$\begin{aligned} \Theta_{k, k'}^l(x, x') &= \sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^l(x)}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x')}{\partial \{W^l, b^l\}} + \sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^1(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}} \\ &= K^l(x, x') \delta_{k, k'} + \delta_{k, k'} \dot{K}^l(x, x') \Theta_{\infty}^{l-1}(x, x') \quad N_{l+1} \rightarrow \infty \\ &= \left(K^l(x, x') + \dot{K}^l(x, x') \Theta_{\infty}^{l-1}(x, x') \right) \delta_{k, k'} \\ &= \Theta_{\infty}^l(x, x') \delta_{k, k'} \end{aligned} \quad (62)$$

this does what we want to achieve in Eq.[41], by assuming $\Theta_{k, k'}^{l-1}(x, x') \xrightarrow{N_l \rightarrow \infty} \Theta_{\infty}^{l-1}(x, x') \delta_{k, k'}$,

we prove: $\Theta_{k, k'}^l(x, x') \xrightarrow{N_{l+1} \rightarrow \infty} \Theta_{\infty}^l(x, x') \delta_{k, k'}$

then finally:

$$\Theta^l(x, x') = \underbrace{\left(K^l(x, x') + \dot{K}^l(x, x') \Theta_{\infty}^{l-1}(x, x') \right)}_{\text{scalar}} \otimes_{\text{outer}} \underbrace{\mathbf{I}_{N_{l+1} \times N_{l+1}}}_{\text{same value for all } k, k' \text{ pairs}} \quad (63)$$

4.6.5 apply the above to $l = 1$

apply the above to $l = 1$, when $l = 1$, $\dot{\phi}(\cdot) = 0 \implies \dot{K}$ just a zero matrix. This is as expected just data x , i.e., constant.

5 linearized model [4]

NTK property during training illustrated using a linearized regime:

5.1 $f_t^{\text{lin}}(x, \theta_t)$ and $\dot{\omega}$

linearized model is:

$$\begin{aligned} f_t^{\text{lin}}(x, \theta_t) &= f_0(x, \theta_0) + \nabla_{\theta} f(x, \theta_t) \Big|_{\theta_t \rightarrow \theta_0} \Delta\theta(t) \\ &= f_0(x, \theta_0) + \nabla_{\theta} f_0(x, \theta_0) (\theta(t) - \theta(0)) \\ &= f_0(x, \theta_0) + \nabla_{\theta} f_0(x, \theta_0) \omega_t \end{aligned} \quad (64)$$

both $f_0(x, \theta_0)$ and $\nabla_{\theta} f_0(x, \theta_0)$ are constants

5.1.1 dynamics of $\dot{\omega}$

looking at the dynamics of linearized gradient flow of **linearized model**, it is obvious that $\dot{\omega}$ only depends on \mathcal{X} instead, as parameter dynamics only depends on training data \mathcal{X} :

$$\begin{aligned} \theta_{t+1} &= \theta_t - \eta \nabla_{\theta_t} \mathcal{L}(\cdot) \\ \theta_{t+1} - \theta_t &= -\eta \nabla_{\theta_t} \mathcal{L}(\cdot) \\ \dot{\omega} = \theta_{t+1} - \theta_t &= -\eta \nabla_{\theta_t} \mathcal{L}(\cdot) \\ &= -\eta \nabla_{\theta} f(\mathcal{X}, \theta_0)^{\top} \nabla_{f_t^{\text{lin}}(\mathcal{X})} \mathcal{L}(\cdot) \end{aligned} \quad (65)$$

$$\dot{\omega} = -\eta \nabla_{\theta} f(\mathcal{X}, \theta_0)^{\top} \nabla_{f_t^{\text{lin}}(\mathcal{X})} \mathcal{L}(\cdot)$$

5.1.2 Dimensionality

$\nabla_{\theta} f_0(\mathcal{X}, \theta_0) \in \mathbb{R}^{|\mathcal{X}| \times |\theta|}$.

$$\nabla_{\theta} f(\mathcal{X}, \theta) \nabla_{\theta} f(\mathcal{X}, \theta)^{\top} = \sum_{i=1}^{|\theta|} (\nabla_{\theta_i} f(\mathcal{X}, \theta)) (\nabla_{\theta_i} f(\mathcal{X}, \theta))^{\top} = \hat{\Theta}(\mathcal{X}, \mathcal{X}) \quad (66)$$

One of the important NTK is when $t = 0$, i.e., at initialization:

$$\nabla_{\theta} f(\mathcal{X}, \theta_0) \nabla_{\theta} f(\mathcal{X}, \theta_0)^{\top} = \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) \quad (67)$$

5.1.3 dynamics of \dot{f}_t^{lin}

$$\begin{aligned}
\dot{f}_t^{\text{lin}}(x, \theta_t) &= \nabla_{\theta} f_0(x, \theta_0) \dot{\omega}(t) \\
&= \nabla_{\theta} f_0(x, \theta_0) \left[-\eta \nabla_{\theta} f(\mathcal{X}, \theta_0)^{\top} \nabla_{f_t^{\text{lin}}(\mathcal{X})} \mathcal{L}(\cdot) \right] \\
&= -\eta \hat{\Theta}_0(x, \mathcal{X}) \nabla_{f_t^{\text{lin}}(\mathcal{X})} \mathcal{L}(\cdot)
\end{aligned} \tag{69}$$

5.1.4 ODE solution using $\mathcal{L} = \frac{1}{2} \|f(\mathcal{X}) - \mathcal{Y}\|_2^2$

$$\begin{aligned}
\dot{f}_t^{\text{lin}}(\mathcal{X}, \theta_t) &= -\eta \hat{\Theta}_0(x, \mathcal{X}) \nabla_{f_t^{\text{lin}}(\mathcal{X})} \mathcal{L}(\cdot) \\
&= -\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) (f_t^{\text{lin}}(\mathcal{X}) - \mathcal{Y})
\end{aligned} \tag{70}$$

then ODE has the close-form solution:

1. note the following has terms in \mathcal{X} :

$$f_t^{\text{lin}}(\mathcal{X}, \theta) = \mathcal{Y} + (f_0(\mathcal{X}, \theta_0) - \mathcal{Y}) \exp^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t} \tag{71}$$

- (a) $t = 0$: $f_t^{\text{lin}}(\mathcal{X}, \theta)|_{t=0} = f_0(\mathcal{X}, \theta_0)$
- (b) $t = \infty$: $f_t^{\text{lin}}(\mathcal{X}, \theta)|_{t=\infty} = \mathcal{Y}$
- (c) it makes sense as f_t^{lin} is an interpolation between $f_0(\mathcal{X}, \theta_0)$ and \mathcal{Y}

2. ODE solution for parameter ω_t is:

$$\omega_t = \nabla_{\theta} f(\mathcal{X}, \theta_0)^{\top} \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} (\mathbf{I} - \exp^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t}) (\mathcal{Y} - f_0(\mathcal{X}, \theta_0)) \tag{72}$$

3. prediction of x is:

$$\begin{aligned}
f_t^{\text{lin}}(x, \theta_t) &= f_0(x, \theta_0) + \nabla_{\theta} f_0(x, \theta_0) \omega_t \quad \text{substitute Eq.(72)} \\
&= f_0(x, \theta_0) + \hat{\Theta}_0(x, \mathcal{X}) \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} (\mathbf{I} - \exp^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t}) (\mathcal{Y} - f_0(\mathcal{X}, \theta_0))
\end{aligned} \tag{73}$$

- (a) $t = 0$: $f_t^{\text{lin}}(x, \theta)|_{t=0} = f_0(x, \theta)$
- (b) $t = \infty$:

$$f_t^{\text{lin}}(x, \theta)|_{t=\infty} = f_0(x, \theta_0) + \hat{\Theta}_0(x, \mathcal{X}) \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} (\mathcal{Y} - f_0(\mathcal{X}, \theta_0)) \tag{74}$$

compare with:

$$\begin{aligned}
y_{\mathbf{w}^*}(x) &= \phi(x)^{\top} \Phi^{\top} (\Phi \Phi^{\top} + \lambda I)^{-1} \mathcal{Y} \\
\text{or without ridge part: } y_{\mathbf{w}^*}(x) &= \phi(x)^{\top} \Phi^{\top} (\Phi \Phi^{\top})^{-1} \mathcal{Y}
\end{aligned} \tag{75}$$

5.2 mean and variance of $f_t^{\text{lin}}(x, \theta_t)$

look at $f_t^{\text{lin}}(x, \theta_t) = f_0(x, \theta_0) - \hat{\Theta}_0(x, \mathcal{X})\hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1}(\mathbf{I} - \exp^{-\eta\hat{\Theta}_0(\mathcal{X}, \mathcal{X})t})(f_0(\mathcal{X}, \theta_0) - \mathcal{Y})$:

$$\mathbb{E}[f_t^{\text{lin}}(x, \theta_t)] = \hat{\Theta}_0(x, \mathcal{X})\hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1}(\mathbf{I} - \exp^{-\eta\hat{\Theta}_0(\mathcal{X}, \mathcal{X})t})\mathcal{Y}$$

when $n \rightarrow \infty$, we have $\hat{\Theta}_0 \rightarrow \Theta_\infty \equiv \Theta$, and $\hat{\mathcal{K}} \rightarrow \mathcal{K}$, by letting:

$$\begin{aligned}\mathbb{E}[f_0(x, \theta_0)f_0(x, \theta_0)^\top] &= \mathcal{K}(x, x) \\ \mathbb{E}[f_0(\mathcal{X}, \theta_0)f_0(\mathcal{X}, \theta_0)^\top] &= \mathcal{K}(\mathcal{X}, \mathcal{X}) \\ \mathbb{E}[f_0(x, \theta_0)f_0(\mathcal{X}, \theta_0)^\top] &= \mathcal{K}(x, \mathcal{X}) \\ \mathbb{E}[f_0(\mathcal{X}, \theta_0)f_0(x, \theta_0)^\top] &= \mathcal{K}(\mathcal{X}, x)\end{aligned}$$

$$\begin{aligned}\text{Var}[f_t^{\text{lin}}(x, \theta_t)] &= \mathcal{K}(x, x) \\ &+ \Theta(x, \mathcal{X})\Theta^{-1}(\mathcal{X}, \mathcal{X})(\mathbf{I} - \exp^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\mathcal{K}(\mathcal{X}, \mathcal{X})(\mathbf{I} - \exp^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\Theta^{-1}(\mathcal{X}, \mathcal{X})\Theta(\mathcal{X}, x) \\ &- \mathcal{K}(x, \mathcal{X})(\mathbf{I} - \exp^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\Theta^{-1}(\mathcal{X}, \mathcal{X})\Theta(\mathcal{X}, x) \\ &- \Theta(x, \mathcal{X})\Theta^{-1}(\mathcal{X}, \mathcal{X})(\mathbf{I} - \exp^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\mathcal{K}(\mathcal{X}, x)\end{aligned}$$

5.2.1 special case when $\hat{y}_t(\mathcal{X}, \theta^{L+1}) = \bar{a}(x)^\top \theta_t^{L+1}$

$$\hat{y}(x, \theta_t^{L+1}) = \hat{y}(x, \theta_0^{L+1}) - \hat{\mathcal{K}}(x, \mathcal{X})\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1}(\mathbf{I} - \exp^{-\eta\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})t})(\hat{y}(\mathcal{X}, \theta_0^{L+1}) - \mathcal{Y}) \quad (76)$$

when $n \rightarrow \infty$, we have $\hat{\mathcal{K}} \rightarrow \mathcal{K}$:

$$\mathbb{E}[\hat{y}_t(x, \theta^{L+1})] = \mathcal{K}(x, \mathcal{X})\mathcal{K}^{-1}(\mathcal{X}, \mathcal{X})(\mathbf{I} - \exp^{-\eta\mathcal{K}t})\mathcal{Y} \quad (77)$$

think about the case when $t \rightarrow \infty$

$$\mathbb{E}[\hat{y}_t(x, \theta^{L+1})] = \mathcal{K}(x, \mathcal{X})\mathcal{K}^{-1}(\mathcal{X}, \mathcal{X})\mathcal{Y} \quad (78)$$

now expand every term:

$$\text{Var}[\hat{y}_t(x, \theta^{L+1})] = \mathcal{K}(x, x) - \mathcal{K}(x, \mathcal{X})\mathcal{K}^{-1}(\mathcal{X}, \mathcal{X})(\mathbf{I} - \exp^{-2\eta\mathcal{K}(\mathcal{X}, \mathcal{X})t})\mathcal{K}(\mathcal{X}, x) \quad (79)$$

think about the case when $t \rightarrow \infty$

$$\text{Var}[\hat{y}_t(x, \theta^{L+1})] = \mathcal{K}(x, x) - \mathcal{K}(x, \mathcal{X})\mathcal{K}^{-1}(\mathcal{X}, \mathcal{X})\mathcal{K}(\mathcal{X}, x) \quad (80)$$

compare that with:

$$\begin{aligned}p(f|\mathcal{X}, \mathcal{Y}) &= \mathcal{GP}\left(K(\mathbf{x}, \mathcal{X})K(\mathcal{X}, \mathcal{X})^{-1}\mathcal{Y}, \right. \\ &\quad \left. k(\mathbf{x}, \mathbf{x}) - K(\mathbf{x}, \mathcal{X})K(\mathcal{X}, \mathcal{X})^{-1}K(\mathcal{X}, \mathbf{x})\right)\end{aligned} \quad (81)$$

5.3 lazy training

finally one need to prove these:

$$\left. \begin{aligned} & \sup_{t \geq 0} \|f_t(x) - f_t^{\text{lin}}\|_2 \\ & \sup_{t \geq 0} \frac{\|\theta_t - \theta_0\|_2}{\sqrt{n}} \\ & \sup_{t \geq 0} \|\hat{\Theta}_t - \hat{\Theta}_0\|_F \end{aligned} \right\} = \mathcal{O}(n^{-\frac{1}{2}}) \quad \text{as } n \rightarrow \infty \quad (82)$$

References

- [1] Radford M Neal, “Priors for infinite networks (tech. rep. no. crg-tr-94-1),” *University of Toronto*, 1994.
- [2] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein, “Deep neural networks as gaussian processes,” *arXiv preprint arXiv:1711.00165*, 2017.
- [3] Arthur Jacot, Franck Gabriel, and Clément Hongler, “Neural tangent kernel: Convergence and generalization in neural networks,” *arXiv preprint arXiv:1806.07572*, 2018.
- [4] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington, “Wide neural networks of any depth evolve as linear models under gradient descent,” *arXiv preprint arXiv:1902.06720*, 2019.