

Infinite-width Neural Networks: Relationship with Gaussian Process and Neural Tangent Kernel

Richard Xu

August 22, 2020

1 Preamble

In this tutorial, my contribution mainly has been the attempt to summarize the following papers and blogs in a unified and (hopefully) more intuitive ways - (particularly, the three blogs below are extremely useful):

- Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In Advances in neural information processing systems, pages 8571–8580, 2018
- J. H. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. ICLR, 2018
- Radford M. Neal. Priors for infinite networks (tech. rep. no. crg-tr-94-1). University of Toronto, 1994
- <https://www.uv.es/gonmagar/blog/2019/01/21/DeepNetworksAsGPs>
- <https://bryn.ai/jekyll/update/2019/04/02/neural-tangent-kernel.html>
- http://chenyilan.net/files/ntk_derivation.pdf

1.1 notations

- I attempted to unify notations, where I used the following definition for Neural Network functions:

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

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

1. $k \in \{1, \dots, N_{l+1}\}$ indexes elements of z^l
2. $i \in \{1, \dots, N_{l+1}\}$ also indexes elements of z^l , and it is used when k is reserved to a specific index
3. $j \in \{1, \dots, N_k\}$ indexes elements of z^{l-1}
4. $W^l \in \mathcal{R}^{N_{l+1} \times N_l}$
5. $x^{(p)}$ and $x^{(q)}$ are used to indicate two data points
6. k and k' indexes two functional output of z^l
7. size of data input is $|d_{\text{in}}|$

1.2 Others minor contributions

- I made the derivations a bit more verbose for people to follow
- a very quick introduction on Gaussian Process, and Central Limit Theorem

2 Gaussian Process

This tutorial makes frequent references to GP, so we talk about it briefly:

- if one is to perform a predictive distribution $p(y^*|y, X, x^*)$ through GP:

$$\begin{aligned} p\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) &= \int p\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}, \textcolor{red}{f}\right) p(\textcolor{red}{f}|X) \mathrm{d}f \\ &= \int \mathcal{N}\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} \textcolor{red}{f}(X) \\ \textcolor{red}{f}(x^{*\top}) \end{bmatrix}, \sigma_\epsilon^2 I\right) p(\textcolor{red}{f}|X, x^*) \mathrm{d}f \end{aligned}$$

- This is the **key**: prior $p(f|X, x^*)$ is defined over function $f(X)$ instead of X
- Imagine, if instead, prior is defined over X , i.e., $p(X)$ is the prior:

$$\int \mathcal{N}\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} f(X) \\ f(x^{*\top}) \end{bmatrix}, \sigma_\epsilon^2 I\right) p(X) \mathrm{d}X$$

Then, non-linear f is **not** making integral tractable!

3 GP for Neural Network: Direct computation

3.1 neural network function

using parameters:

$$\theta \equiv \{W^L, b^L, \dots, W^1, b^1\}$$

Deep neural network function $f_\theta(X)$ is defined as:

$$\begin{aligned} f_\theta(X) &= W^L \phi^L(X) + b^L \\ &= W^L (\phi^{L-1}(X) W^{L-1} + b^{L-1}) + b^L \\ &\dots \\ &= W^L \dots (W^1 \phi^1(X) + b^1) + \dots + b^L \end{aligned}$$

it should be noted that non-linear output $\phi^l(\cdot)$:

$$\begin{aligned} \phi^L(X) &\equiv \phi^L(X | \theta^1, \dots, \theta^{L-1}) \\ &\equiv \phi^L(X | W^1, b^1, \dots, W^{L-1}, b^{L-1}) \end{aligned}$$

3.2 Apply NN function in predictive distribution

- However, applying NN function in predictive distribution: prior is defined over θ instead of over f . i.e., i.i.d noises are injected to each element of θ . The predictive distribution:

$$p\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) = \int \mathcal{N}\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} f_\theta(X) \\ f_\theta(x^*) \end{bmatrix}, \sigma_\epsilon^2 I\right) \mathcal{N}(\theta | 0, \sigma_\theta^2 I) d\theta$$

- The integral is **not** analytic!!

3.3 what is the predictive distribution

- eventually, we will need to ask an even harder question on, i.e., suppose we let $N^l \equiv |W^l|$, i.e., the “width” of the neural network at each layer l , and we would like to study the effect of:

$$p\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) \xrightarrow[N^1, \dots, N^L \rightarrow \infty]{d} ?$$

- however, firstly, we ask the question on, what is:

$$p\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) = ?$$

- attempt to compute it **directly**, by looking the **mean** and **variance**:

$$\begin{aligned} &- \mathbb{E} \left[\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right] \\ &- \mathbb{E} \left[\begin{bmatrix} y \\ y^* \end{bmatrix} \begin{bmatrix} y^\top & y^{*\top} \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right] \end{aligned}$$

3.3.1 look at the mean:

$$\begin{aligned}
& \mathbb{E} \left[\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right] \\
&= \int_y \int_{y^*} \begin{bmatrix} y \\ y^* \end{bmatrix} p \left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right) dy dy^* \\
&= \int_y \int_{y^*} \begin{bmatrix} y \\ y^* \end{bmatrix} \int_{\theta} p \left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \theta, \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right) p(\theta | \sigma_{\theta}^2) d\theta dy dy^* \\
&= \underbrace{\int_{\theta} \int_y \int_{y^*} \begin{bmatrix} y \\ y^* \end{bmatrix} \mathcal{N} \left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} f_{\theta}(X) \\ f_{\theta}(x^*) \end{bmatrix}, \sigma_{\epsilon}^2 I \right) dy dy^* \mathcal{N}(\theta | 0, \sigma_{\theta}^2 I) d\theta}_{\mathbb{E} \left[\begin{bmatrix} y \\ y^* \end{bmatrix} \right] = \begin{bmatrix} f_{\theta}(X) \\ f_{\theta}(x^*) \end{bmatrix}} \\
&= \int \begin{bmatrix} f_{\theta}(X) \\ f_{\theta}(x^*) \end{bmatrix} \mathcal{N}(\theta | 0, \sigma_{\theta}^2 I) d\theta \quad \text{to expand one layer :} \\
&= \int \begin{bmatrix} \phi^L(X)W^L + b^L \\ \phi^L(x^{*\top})W^L + b^L \end{bmatrix} \mathcal{N}(W^L | 0, \sigma_w^2 I) \mathcal{N}(b^L | 0, \sigma_b^2 I) \mathcal{N}(\theta^{1, \dots, L-1} | 0, \sigma_{\theta}^2 I) d\theta^{1, \dots, L-1} dW^L db^L \\
&= \int \left[\underbrace{\phi^L(X) \int W^L \mathcal{N}(W^L | 0, \sigma_w^2 I) dW^L}_{=0} + \underbrace{\int b^L \mathcal{N}(b^L | 0, \sigma_b^2 I) db^L}_{=0} \right. \\
&\quad \left. \underbrace{\phi^L(x^{*\top}) \int W^L \mathcal{N}(W^L | 0, \sigma_w^2 I) dW^L}_{=0} + \underbrace{\int b^L \mathcal{N}(b^L | 0, \sigma_b^2 I) db^L}_{=0} \right] \mathcal{N}(\theta^{1, \dots, L-1} | 0, \sigma_{\theta}^2 I) d\theta^{1, \dots, L-1} \\
&= \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{aligned}$$

note we are not dealing with infinity at the moment

3.3.2 look at co-variance

$$\mathbb{E} \left[\begin{bmatrix} y \\ y^* \end{bmatrix} \begin{bmatrix} y^{\top} & y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right]$$

Apply same trick as calculating mean, i.e., introducing θ and then integrate it out:

$$\begin{aligned}
&= \int_y \int_{y^*} \int_{\theta} p \left(\begin{bmatrix} y \\ y^* \end{bmatrix} \begin{bmatrix} y^{\top} & y^* \end{bmatrix} \middle| \theta, \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right) p(\theta | \sigma_{\theta}^2) d\theta dy dy^* \\
&= \underbrace{\int_{\theta} \int_y \int_{y^*} \begin{bmatrix} y \\ y^* \end{bmatrix} \begin{bmatrix} y^{\top} & y^* \end{bmatrix} \mathcal{N} \left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} f_{\theta}(X) \\ f_{\theta}(x^*) \end{bmatrix}, \sigma_{\epsilon}^2 I \right) dy dy^* \mathcal{N}(\theta | 0, \sigma_{\theta}^2 I) d\theta}_{\mathbb{E}[Z^2] \quad Z \text{ is not mean-subtracted}}
\end{aligned}$$

$$\text{Let } Z = \begin{bmatrix} y \\ y^* \end{bmatrix}:$$

$$\begin{aligned}
&\text{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \implies \mathbb{E}[Z^2] = \text{Var}[Z] + (\mathbb{E}[Z])^2 \\
&= \int_{\theta} \underbrace{\sigma_{\epsilon}^2 I}_{\text{Var}[Z]} + \underbrace{\begin{bmatrix} f_{\theta}(X) \\ f_{\theta}(x^*) \end{bmatrix} \begin{bmatrix} f_{\theta}(X)^{\top} & f_{\theta}(x^*) \end{bmatrix}}_{(\mathbb{E}[Z])^2} \mathcal{N}(\theta | 0, \sigma_{\theta}^2 I) d\theta \\
&= \sigma_{\epsilon}^2 I + \int_{\theta} \left[\begin{pmatrix} \phi^L(X)W^L + b^L \\ \phi^L(x^{*\top})W^L + b^L \end{pmatrix} \begin{pmatrix} W^{L\top} \phi^L(X)^{\top} + b^{L\top} \\ W^{L\top} \phi^L(x^{*\top})^{\top} + b^{L\top} \end{pmatrix} \right. \\
&\quad \left. \begin{pmatrix} \phi^L(X)W^L + b^L \\ \phi^L(x^{*\top})W^L + b^L \end{pmatrix} \begin{pmatrix} W^{L\top} \phi^L(x^{*\top})^{\top} + b^{L\top} \\ W^{L\top} \phi^L(X)^{\top} + b^{L\top} \end{pmatrix} \right] \mathcal{N}(\theta | 0, \sigma_{\theta}^2 I) d\theta
\end{aligned}$$

realize $\mathbf{Cov}(x^L(X)W^L, b^L) = 0$:

$$= \sigma_\epsilon^2 I + \int_\theta \begin{bmatrix} \phi^L(X)W^L W^{L\top} x^L(X)^\top + b^L b^{L\top} & \phi^L(X)W^L W^{L\top} x^L(x^{\star\top})^\top + b^L b^{L\top} \\ \phi^L(x^{\star\top})W^L W^{L\top} \phi^L(X)^\top + b^L b^{L\top} & \phi^L(x^{\star\top})W^L W^{L\top} \phi^L(x^{\star\top})^\top + b^L b^{L\top} \end{bmatrix} \mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) d\theta$$

factorize $\mathcal{N}(\theta)$ as each element of θ is independent:

$$\mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) d\theta = \mathcal{N}(\theta^L \mid 0, \sigma_\theta^2 I) \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1}$$

$$= \int \begin{bmatrix} \sigma_w^2 \phi^L(X) x^L(X)^\top + \sigma_b^2 & \sigma_w^2 \phi^L(X) \phi^L(x^{\star\top})^\top + \sigma_b^2 \\ \sigma_w^2 \phi^L(x^{\star\top}) \phi^L(X)^\top + \sigma_b^2 & \sigma_w^2 \phi^L(x^{\star\top}) \phi^L(x^{\star\top})^\top + \sigma_b^2 \end{bmatrix} \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1}$$

let's taking the **left corner** element, and expand θ by one:

$$\begin{aligned} & \int \sigma_w^2 \phi^L(X) \phi^L(X)^\top \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} + \int \sigma_b^2 \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} \\ &= \sigma_w^2 \int \phi^L(X) \phi^L(X)^\top \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} + \sigma_b^2 \end{aligned}$$

as we know $\phi^L(X) \phi^L(X)^\top \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} + \sigma_b^2$:

$$= \sigma_b^2 + \sigma_w^2 \int \left[\phi(W^{L-1} \phi^{L-1}(X) + b^{L-1}) \phi(W^{L-1} \phi^{L-1}(X) + b^{L-1})^\top \right] \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1}$$

it's difficult to see what is this distribution is.

4 Single layer neural network

$$f_k(x) = b_k + \sum_{j=1}^H v_{jk} h_j(x)$$

$$h_j(x) = \tanh \left(a_j + \sum_{i=1}^I u_{ij} x_i \right)$$

this is very strange way to define neural network, and it defines it to part of the second layer:

$$f_k(x) = \underbrace{b_k}_{z_k^l} + \sum_{j=1}^H \underbrace{v_{jk}}_{W_{k,j}^l} \times \underbrace{\tanh}_{\phi} \left(\underbrace{a_j}_{b_j^{l-1}} + \underbrace{u_{:,j}^\top}_{W_{:,j}^{l-1}{}^\top} x \right)$$

$$\implies z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \times \phi(z_j^{l-1}(x)) \quad \text{modern notation}$$

4.1 $p(z_k^l(x))$ for single input x

We need CLT for computing this probability.

4.1.1 Central Limit Theorem:

$X^{(1)}, X^{(2)}, \dots, X^{(n)}$ are i.i.d samples

- note any **arbitrary** distribution with *bounded variance* for $X^{(i)}$ will do
- let \bar{X} be sample mean, and let: $\sigma^2 = \text{Var}[X^{(1)}]$
- Limiting form of the distribution:

$$\sqrt{n}(\bar{X} - \mathbb{E}[X^{(1)}]) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

$$(\bar{X} - \mathbb{E}[X^{(1)}]) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$

$$\frac{1}{\sigma} \sqrt{n}(\bar{X} - \mathbb{E}[X^{(1)}]) \xrightarrow{d} \mathcal{N}(0, 1)$$

- Similarly, instead of “**sample mean**”, it can be also be applied to “**sample sum**” of i.i.d random variables:

$$\sqrt{n}(\bar{X} - \mathbb{E}[X^{(1)}]) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

$$\implies \sqrt{n}\sqrt{n}(\bar{X} - \mathbb{E}[X^{(1)}]) \xrightarrow{d} \mathcal{N}(0, \sqrt{n}^2 \sigma^2) = \mathcal{N}(0, n\sigma^2)$$

$$\implies n(\bar{X} - \mathbb{E}[X^{(1)}]) \xrightarrow{d} \mathcal{N}(0, n\sigma^2)$$

$$\implies \left(\sum_{i=1}^n X_i - n\mathbb{E}[X^{(1)}] \right) \xrightarrow{d} \mathcal{N}(0, n\sigma^2)$$

choose one of these conditions to suit the situation

4.1.2 Apply CLT to compute $p(z_k^l(x))$

- let's pick any arbitrary x , since we only pick a single x , so the index is **not** important, there is no need to use $x^{(1)}$ like in the literature:
- computing $p(z_k^l(x))$ directly is hard!
- however, $z_k^l(x)$ is $b_k^l +$ sum of i.i.d elements using CLT notations:

$$z_k^l(x) = b_k^l + \underbrace{\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x))}_{\underbrace{\sum_{j=1}^{N_l} X_j}} \quad \text{note we are not taking average}$$

- therefore, we can just compute mean and variance of its individual element, i.e., an arbitrary $j = 1$ and then apply CLT!

$$X_j \equiv W_{k,j}^l \phi(z_j^{l-1}(x))$$

4.1.3 mean and variance of $W_{k,j}^l \phi(z_j^{l-1}(x))$

- **Expectation**

$$\begin{aligned} \mathbb{E}[W_{k,j}^l \phi(z_j^{l-1}(x))] &= \mathbb{E}[W_{k,j}^l] \mathbb{E}[\phi(z_j^{l-1}(x))] \quad \text{since } W_{k,j}^l \text{ and } \phi(z_j^{l-1}(x)) \text{ are independent} \\ &\quad \text{as } z_j^{l-1}(x) \text{ depends on } (W^{l-1}, b^{l-1}) \\ &= 0 \times \mathbb{E}[\phi(z_j^{l-1}(x))] \quad \text{because we choose } W_{k,j}^l \sim \mathcal{N}(0, \sigma_w) \\ &= 0 \end{aligned}$$

- **Variance**

$$\begin{aligned} \text{Var}[W_{k,j}^l \phi(z_j^{l-1}(x))] &= \mathbb{E}\left[\left(W_{k,j}^l \phi(z_j^{l-1}(x))\right)^2\right] \\ &= \mathbb{E}[(W_{k,j}^l)^2] \mathbb{E}[\phi(z_j^{l-1}(x))^2] \quad \text{since } W_{k,j}^l \text{ and } \phi(z_j^{l-1}(x)) \text{ are independent} \\ &= \sigma_w^2 \underbrace{\mathbb{E}[\phi(z_j^{l-1}(x))^2]}_{\text{bounded}} \implies \text{Var}[W_{k,j}^l \phi(z_j^{l-1}(x))] \text{ to be bounded} \\ &= \sigma_w^2 \mathbb{E}[\phi(z_j^{l-1}(x))^2] \end{aligned}$$

we leave in this form, as

$$\mathbb{E}[\phi(z_j^{l-1}(x))^2] \equiv \mathbb{E}_{W^{l-1}, \dots, b^{l-1}, \dots}[\phi(z_j^{l-1}(x))^2]$$

4.1.4 apply CLT:

However, we can apply CLT: making $p(z^l(x))$ distributed as Gaussian where its variance is dependent on variance of previous layer, a recursion.

$$\begin{aligned} \text{using } \left(\sum_{i=1}^n X_i - n\mathbb{E}[X_1] \right) &\xrightarrow{d} \mathcal{N}(0, n\sigma^2) \\ \Rightarrow \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) - 0 \right) &\sim \mathcal{N}\left(0, N_l \sigma_w^2 \mathbb{E}[\phi(z_1^{l-1}(x))^2]\right) \quad N_l \rightarrow \infty \end{aligned}$$

- However, variance under this expression $N_l \sigma_w^2 [\phi(z_1^{l-1}(x))^2]$ is divergent because of N_l !
- luckily, we can take control the choice of σ_w^2 , if we let:

$$\sigma(W_{k,j}^l) = \sigma_w = \frac{1}{\sqrt{N_l}} \quad \Rightarrow \quad \sigma_w^2 = \frac{1}{N_l}$$

- the above is the key, implication is:

$$\begin{aligned} \Rightarrow \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) - 0 \right) &\sim \mathcal{N}\left(0, N_l \frac{1}{N_l} \mathbb{E}[\phi(z_1^{l-1}(x))^2]\right) \\ &= \mathcal{N}\left(0, \underbrace{\mathbb{E}[\phi(z_1^{l-1}(x))^2]}_{\text{bounded}}\right) \end{aligned}$$

- finally adding the bias b_k^l :

Note that sum of two **independent** Gaussian random variables is also Gaussian: (not to confuse with GMM!)

$$\begin{aligned} X &\sim \mathcal{N}(\mu_X, \sigma_X^2) \\ Y &\sim \mathcal{N}(\mu_Y, \sigma_Y^2) \\ Z = X + Y &\quad Z = X + Y \\ \Rightarrow Z &\sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \end{aligned}$$

Therefore:

$$\left(z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) \right) \xrightarrow{d} \mathcal{N}\left(0, \underbrace{\sigma_b^2}_{\sigma_X^2} + \underbrace{\mathbb{E}[\phi(z_1^{l-1}(x))^2]}_{\sigma_Y^2}\right) \quad \text{as } N_l \rightarrow \infty$$

- appreciate the recursion here

4.2 given two inputs $x^{(p)}$, $x^{(q)}$: compute $\text{Cov}[z_k^l(x^{(p)}) z_k^l(x^{(q)})]$

To do so, we need to use **Multidimensional CLT**

4.2.1 Multidimensional CLT:

$$\begin{aligned}
\sum_{i=1}^n \mathbf{X}_i &= \underbrace{\begin{bmatrix} X_1^{(1)} \\ \vdots \\ X_1^{(p)} \\ \vdots \\ X_1^{(q)} \\ \vdots \\ X_1^{(k)} \end{bmatrix}}_{\mathbf{X}_1} + \underbrace{\begin{bmatrix} X_2^{(1)} \\ \vdots \\ X_2^{(p)} \\ \vdots \\ X_2^{(q)} \\ \vdots \\ X_2^{(k)} \end{bmatrix}}_{\mathbf{X}_2} + \cdots + \underbrace{\begin{bmatrix} X_n^{(1)} \\ \vdots \\ X_n^{(p)} \\ \vdots \\ X_n^{(q)} \\ \vdots \\ X_n^{(k)} \end{bmatrix}}_{\mathbf{X}_n} = \underbrace{\begin{bmatrix} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \sum_{i=1}^n X_i^{(k)} \end{bmatrix}}_{\sum_{i=1}^n \mathbf{X}_i} \\
\Rightarrow \bar{\mathbf{X}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(k)} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{X}}^{(1)} \\ \vdots \\ \bar{\mathbf{X}}^{(p)} \\ \vdots \\ \bar{\mathbf{X}}^{(q)} \\ \vdots \\ \bar{\mathbf{X}}^{(k)} \end{bmatrix}
\end{aligned}$$

Therefore:

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^n [\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i]] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{X}_i - \mathbb{E}[\mathbf{X}_1]) = \frac{\sqrt{n}}{\sqrt{n}} \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n \mathbf{X}_i \right) - n \mathbb{E}[\mathbf{X}_1] \\
&= \sqrt{n} (\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}_1])
\end{aligned}$$

- Sample mean version:

$$\Rightarrow \sqrt{n} \mathbb{E} \left[\underbrace{\left(\bar{\mathbf{X}}^{(p)} - \mathbb{E}[\bar{\mathbf{X}}_1^{(p)}] \right)}_{\text{scalar}} \underbrace{\left(\bar{\mathbf{X}}^{(q)} - \mathbb{E}[\bar{\mathbf{X}}_1^{(q)}] \right)}_{\text{scalar}} \right] = \Sigma_{(p),(q)}$$

for each co-variance/non-diagonal elements $(p, q) \in \{1, \dots, k\}$:

- Sample sum version:

$$\begin{aligned}
& \left(\left[\sum_i^n \mathbf{X}_i \right] - n\mathbb{E}[\mathbf{X}_1] \right) \xrightarrow{d} \mathcal{N}_k(0, n\boldsymbol{\Sigma}) \\
\Rightarrow & \mathbb{E} \left[\left(\left[\sum_i^n \mathbf{X}_i \right]^{(p)} - n\mathbb{E}[\mathbf{X}_1]^{(p)} \right) \left(\left[\sum_i^n \mathbf{X}_i \right]^{(q)} - n\mathbb{E}[\mathbf{X}_1]^{(q)} \right) \right] = n\boldsymbol{\Sigma}_{(p),(q)} \\
& \Rightarrow \mathbb{E} \left[\left(n\bar{\mathbf{X}}^{(p)} - n\mathbb{E}[X_1^{(p)}] \right) \left(n\bar{\mathbf{X}}^{(q)} - n\mathbb{E}[X_1^{(q)}] \right) \right] = n\boldsymbol{\Sigma}_{(p),(q)} \\
\Rightarrow & \mathbb{E} \left[\left(\left[\sum_i^n \mathbf{X}_i \right]^{(p)} - n\mathbb{E}[X_1^{(p)}] \right) \left(\left[\sum_i^n \mathbf{X}_i \right]^{(q)} - n\mathbb{E}[X_1^{(q)}] \right) \right] = n\boldsymbol{\Sigma}_{(p),(q)}
\end{aligned}$$

where $\boldsymbol{\Sigma}_{(p),(q)} = \text{Cov}(X_1^{(p)}, X_1^{(q)})$

4.2.2 put in Multidimensional CLT structure:

$$\begin{aligned}
& \begin{bmatrix} \vdots \\ W_{k,1}^l \phi(z_1^{l-1}(x^{(p)})) \\ \vdots \\ W_{k,1}^l \phi(z_j^{l-1}(x^{(q)})) \\ \vdots \end{bmatrix} + \dots + \begin{bmatrix} \vdots \\ W_{k,N_l}^l \phi(z_j^{l-1}(x^{(p)})) \\ \vdots \\ W_{k,N_l}^l \phi(z_j^{l-1}(x^{(q)})) \\ \vdots \end{bmatrix} = \underbrace{\begin{bmatrix} \vdots \\ \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(p)})) \\ \vdots \\ \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(q)})) \\ \vdots \end{bmatrix}}_{\begin{bmatrix} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \sum_{i=1}^n X_i^{(k)} \end{bmatrix}} = \underbrace{\begin{bmatrix} \vdots \\ z_k^l(x^{(p)}) \\ \vdots \\ z_k^l(x^{(q)}) \\ \vdots \end{bmatrix}}_{\sum_{i=1}^n \mathbf{X}_i}
\end{aligned}$$

Since we already know that:

$$\begin{aligned}
& \mathbb{E} \left[\left(\left[\sum_i^n \mathbf{X}_i \right]^{(p)} - n\mathbb{E}[X_1^{(p)}] \right) \left(\left[\sum_i^n \mathbf{X}_i \right]^{(q)} - n\mathbb{E}[X_1^{(q)}] \right) \right] = n\boldsymbol{\Sigma}_{(p),(q)} \\
\Rightarrow & \mathbb{E} \left[\left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(p)})) - \underbrace{N_l \mathbb{E}[W_{k,1}^l \phi(z_1^{l-1}(x^{(p)}))]}_{=0} \right) \times \right. \\
& \left. \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(q)})) - \underbrace{N_l \mathbb{E}[W_{k,1}^l \phi(z_1^{l-1}(x^{(q)}))]}_{=0} \right) \right] = N_l \boldsymbol{\Sigma}_{(p),(q)}
\end{aligned}$$

for any arbitrary $j = 1$, and then:

$$\begin{aligned}
& \mathbb{E} \left[\left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(p)})) \right) \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(q)})) \right) \right] \\
&= N_l \Sigma_{(p),(q)} \\
&= N_l \text{Cov} \left(W_{k,1}^l \phi(z_1^{l-1}(x^{(p)})), W_{k,1}^l \phi(z_1^{l-1}(x^{(q)})) \right) \\
&= N_l \mathbb{E} \left[W_{k,1}^l \phi(z_1^{l-1}(x^{(p)})) \times W_{k,1}^l \phi(z_1^{l-1}(x^{(q)})) \right]
\end{aligned}$$

add b_k^l into, and look at $z_k^l(x)$:

$$\begin{aligned}
\mathbb{E} [z_k^l(x^{(p)}) z_k^l(x^{(q)})] &= \sigma_b^2 + \mathbb{E} \left[\left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(p)})) \right) \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(q)})) \right) \right] \\
&= \sigma_b^2 + N_l \text{Cov} (W_{k,1}^l \phi(z_1^{l-1}(x^{(p)})), W_{k,1}^l \phi(z_1^{l-1}(x^{(q)}))) \quad \text{use CLT result above} \\
&= \sigma_b^2 + N_l \sigma_w^2 \text{Cov} (\phi(z_1^{l-1}(x^{(p)})), \phi(z_1^{l-1}(x^{(q)}))) \\
&= \sigma_b^2 + N_l \frac{1}{N_l} \text{Cov} (\phi(z_1^{l-1}(x^{(p)})), \phi(z_1^{l-1}(x^{(q)}))) \\
&= \sigma_b^2 + \text{Cov} (\phi(z_1^{l-1}(x^{(p)})), \phi(z_1^{l-1}(x^{(q)}))) \\
&= \sigma_b^2 + \mathbb{E} [\phi(z_1^{l-1}(x^{(p)})) \times \phi(z_1^{l-1}(x^{(q)}))]
\end{aligned}$$

- **note 1:** this co-variance is same $\forall k$ in $z_k^l(x)$, so right hand side does not need to keep k index because in this particular setting, since b_k , $b_{k'}$, $W_{k,j}$ and $W_{k',j'}$ are independent variables, co-variance between any of them are zero:

$$\begin{aligned}
z_{\mathbf{k}}^l(x) &= b_{\mathbf{k}} + \sum_{j=1}^{N_l} W_{\mathbf{k},j}^l \phi(z_j^{l-1}(x)) \\
z_{\mathbf{k}'}^l(x) &= b_{\mathbf{k}'} + \sum_{j=1}^{N_l} W_{\mathbf{k}',j}^l \phi(z_j^{l-1}(x)) \\
\implies \mathbb{E} [W_{\mathbf{k},j}^l \phi(z_j^{l-1}(x)) \times W_{\mathbf{k}',j'}^l \phi(z_{j'}^{l-1}(x))] &= 0 \quad \forall \{k, k', j, j'\}
\end{aligned}$$

- **note 2:** in literature, it is written:

$$\begin{aligned}
\mathbb{E} [z_k^l(x^{(\mathbf{p})}) z_k^l(x^{(\mathbf{q})})] &= \sigma_b^2 + \sigma_w^2 \mathbb{E} \left[\sum_{j=1}^{N_l} \phi(z_j^{l-1}(x^{(\mathbf{p})})) \phi(z_j^{l-1}(x^{(\mathbf{q})})) \right] \\
\text{instead of } &= \sigma_b^2 + \mathbb{E} \left[\left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(\mathbf{p})})) \right) \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(\mathbf{q})})) \right) \right]
\end{aligned}$$

This is because of **note1** above

- regardless of this special property CLT still apply.

4.2.3 Relationship with Gaussian Process (GP):

let $f_k(x) \equiv z_k^l(x)$ be some function, and since for every arbitrary point pair, $x^{(p)}$ and $x^{(q)}$, we have:

$$\begin{aligned}\mathbb{E}[f(x)] &= 0 \\ \mathbb{E}[f(x^{(p)}), f(x^{(q)})] &= \Sigma_{(p),(q)} \\ \implies f &\sim \mathcal{GP}(0, \Sigma)\end{aligned}$$

- looking at mean and co-variance as $N_l \rightarrow \infty$

$$\begin{aligned}\text{Cov}[z_k^l(x^{(p)}), z_k^l(x^{(q)})] &= \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(x^{(p)})) \times \phi(z_1^{l-1}(x^{(q)}))] \quad \text{as } N_l \rightarrow \infty \\ z_k^l(x) &\xrightarrow{d} \mathcal{N}\left(0, \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(x))^2]\right) \quad \text{as } N_l \rightarrow \infty\end{aligned}$$

- putting it in layer specific GP:

$$\begin{aligned}\implies z_k^l(x) &\sim \mathcal{GP}(0, \Sigma) \\ \text{where } \Sigma_{p,q} &= \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(x^{(p)})) \times \phi(z_1^{l-1}(x^{(q)}))] \quad \text{as } N_l \rightarrow \infty\end{aligned}$$

4.3 more on GP

- First define $K^l(x^{(p)}, x^{(q)})$ in terms of pre-activation $z_k^l(x)$ in this section, it will be changed later to post-activation
- instead of letting $\sigma(W_{k,j}^l) = \frac{1}{\sqrt{N_l}}$ in previous section, we let it be more generically:

$$\sigma(W_{k,j}^l) = \frac{\sigma_w}{\sqrt{N_l}}$$

$$\begin{aligned}K^l(x^{(p)}, x^{(q)}) &= \mathbb{E}[z_k^l(x^{(p)})z_k^l(x^{(q)}) | z^{l-1}] \\ &= \mathbb{E}\left[\left(b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(p)}))\right) \times \left(b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(q)}))\right)\right] \\ &= \sigma_b^2 + \frac{\sigma_w^2}{N_l} \mathbb{E}\left[\sum_{j=1}^{N_l} \phi(z_j^{l-1}(x^{(p)})) \times \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x^{(q)}))\right] \\ &= \sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(z_1^{l-1}(x^{(p)})) \times \phi(z_1^{l-1}(x^{(q)}))] \quad \text{apply CLT } N_l \rightarrow \infty \\ &= \sigma_b^2 + \sigma_w^2 \underbrace{\mathbb{E}_{z_1^{l-1} \sim \mathcal{GP}(0, K^{l-1})} [\phi(z_1^{l-1}(x^{(p)})) \phi(z_1^{l-1}(x^{(q)}))]}_{\text{since } \mathbb{E}[\phi(z)] = \mathbb{E}_{z \sim p(z)} [\phi(z)]} \\ &= \sigma_b^2 + \sigma_w^2 \underbrace{F_\phi(K^{l-1}(x^{(p)}, x^{(q)}), K^{l-1}(x^{(p)}, x^{(p)}), K^{l-1}(x^{(q)}, x^{(q)}))}_{F_\phi(K^{l-1})} \\ &= \sigma_b^2 + \sigma_w^2 F_\phi(K^{l-1}(x^{(p)}, x^{(q)}))\end{aligned}$$

using properties of point Marginals of Gaussian Process:

$$\begin{aligned}
F_\phi(K^{l-1}(x^{(p)}, x^{(q)})) &= \mathbb{E}_{z_j^{l-1} \sim \mathcal{GP}(0, K^{l-1})} \left[\phi(z_j^{l-1}(x^{(p)})) \phi(z_j^{l-1}(x^{(q)})) \right] \\
&= \underbrace{\mathbb{E}_{\left(z_j^{l-1}(x^{(p)}), z_j^{l-1}(x^{(q)}) \right)}}_{\text{2 points on function } z_j^{l-1}} \underbrace{\sim \mathcal{N}(0, K^{l-1}(x^{(p)}, x^{(q)}))}_{\text{2D Gaussian}} \left[\phi(z_j^{l-1}(x^{(p)})) \phi(z_j^{l-1}(x^{(q)})) \right]
\end{aligned}$$

$$\begin{bmatrix} z_j^{l-1}(x^{(p)}) \\ z_j^{l-1}(x^{(q)}) \end{bmatrix} \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} K^{l-1}(x^{(p)}, x^{(p)}) & K^{l-1}(x^{(p)}, x^{(q)}) \\ K^{l-1}(x^{(p)}, x^{(q)}) & K^{l-1}(x^{(q)}, x^{(q)}) \end{bmatrix} \right)$$

assume z^{l-1} can be integrated out:

$$= F_\phi(K^{l-1}(x^{(p)}, x^{(q)}), K^{l-1}(x^{(p)}, x^{(p)}), K^{l-1}(x^{(q)}, x^{(q)}))$$

4.4 in summary

this is how K^l relates to K^{l-1} :

$$\textcolor{red}{K^l}(x^{(p)}, x^{(q)}) = \sigma_b^2 + \sigma_w^2 \mathbb{E}_{(z_j^{l-1}(x^{(p)}), z_j^{l-1}(x^{(q)}))} \sim \mathcal{N}(0, \textcolor{red}{K^{l-1}}(x^{(p)}, x^{(q)})) \left[\phi(z_j^{l-1}(x^{(p)})) \phi(z_j^{l-1}(x^{(q)})) \right] \quad (1)$$

we will see the same recursion also applies in NTK, except $\phi \rightarrow \phi'$

5 Expand GP across all layers

5.1 Overall objective

Looking the probability of the final layer output z^L depending on input x :

$$\begin{aligned} p(z^L|x) &= \int p(z^L, K^0, K^1, \dots, K^L|x) \, dK^{0,\dots,L} \\ &= \int p(z^L|K^L) \left(\prod_{l=1}^L p(K^l|K^{l-1}) \right) p(K^0|x) \, dK^{0,\dots,L} \end{aligned}$$

5.2 $p(z^L|K^L)$: conditions on $K^l \equiv \{\phi(z^{l-1})(x^{(p)})\}\phi(z^{l-1})(x^{(q)})\}_{p,q}$

(J. H. Lee et. al 2018) presents an **alternative** definition of K^l , where no longer define K from pre-activation:

$$K^l(x^{(p)}, x^{(q)}) = \mathbb{E}[z_k^l(x^{(p)})z_k^l(x^{(q)}) | z^{l-1}]$$

instead it define K^l in terms of post-activation of previous later $\phi(z^{l-1})$ for reason illustrated later

- look at Neural Network function:

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

let's make it dependent on $\{\phi(z_j^{l-1}(x))\}_j^{N_l}$, i.e.:

- Conditional Marginal

$$\begin{aligned} z_k^l(x) | \{\phi(z_j^{l-1}(x))\}_j^{N_l} &= b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \underbrace{\phi(z_j^{l-1}(x))}_{\text{constant}} \\ \implies z_k^l(x) | \{\phi(z_j^{l-1}(x))\}_j^{N_l} &\sim \mathcal{N}\left(0, \sigma_b^2 + \sum_{j=1}^{N_l} \underbrace{\phi(z_j^{l-1}(x))^2}_{\text{constant}} \text{Var}[W_{k,j}^l]\right) \\ &= \mathcal{N}\left(0, \sigma_b^2 + \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2\right) \end{aligned}$$

using property of weighted sum of Gaussian:

$$\begin{aligned} X_i &\sim \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \dots, \\ \implies \sum_{i=1}^n a_i X_i &\sim \mathcal{N}\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \text{Var}[X_i]\right) \end{aligned}$$

- Conditional Co-variance

$$\begin{aligned}
& \text{Cov} \left[z_k^l(x^{(p)}), z_k^l(x^{(q)}) \mid \left\{ \phi(z_j^{l-1}(x^{(p)})), \phi(z_j^{l-1}(x^{(q)})) \right\}_{j=1}^{N_l} \right] \\
&= \mathbb{E} \left[z_k^l(x^{(p)}) z_k^l(x^{(q)}) \mid \left\{ \phi(z_j^{l-1}(x^{(p)})), \phi(z_j^{l-1}(x^{(q)})) \right\}_{j=1}^{N_l} \right] \\
&= \sigma_b^2 + \mathbb{E}_{W_{k,j}^l} \left[\sum_{j=1}^{N_l} W_{k,j}^l \underbrace{\phi(z_j^{l-1}(x^{(p)})) \phi(z_j^{l-1}(x^{(q)}))}_{\text{constant, used as condition}} \right] \\
&= \sigma_b^2 + \sum_{j=1}^{N_l} \text{Var}[W_{k,j}^l] \phi(z_j^{l-1}(x^{(p)})) \phi(z_j^{l-1}(x^{(q)})) \\
&= \sigma_b^2 + \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x^{(p)})) \phi(z_j^{l-1}(x^{(q)}))
\end{aligned}$$

not using property of weighted sum of Gaussian:

- Combine all together

$$\begin{aligned}
& \text{Cov} \left[z_k^l(x^{(p)}), z_k^l(x^{(q)}) \mid \left\{ \phi(z_j^{l-1}(x^{(p)})), \phi(z_j^{l-1}(x^{(q)})) \right\}_{j=1}^{N_l} \right] = \sigma_b^2 + \sigma_w^2 \frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x^{(p)})) \phi(z_j^{l-1}(x^{(q)})) \\
& z_k^l(x) \mid \left\{ \phi(z_j^{l-1}(x)) \right\}_{j=1}^{N_l} \sim \mathcal{N} \left(0, \sigma_b^2 + \sigma_w^2 \frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2 \right) \\
& \Rightarrow \begin{bmatrix} z^l(x^{(p)}) \\ z^l(x^{(q)}) \end{bmatrix} \mid \begin{bmatrix} \phi(z_j^{l-1}(x^{(p)})) \\ \phi(z_j^{l-1}(x^{(q)})) \end{bmatrix} \sim \mathcal{N} \left(\mathbf{0}, G \left(\begin{bmatrix} K^l(x^{(p)}, x^{(p)}) & K^l(x^{(p)}, x^{(q)}) \\ K^l(x^{(p)}, x^{(q)}) & K^l(x^{(q)}, x^{(q)}) \end{bmatrix} \right) \right)
\end{aligned}$$

- in GP paradigm:

$$z^l(x) \mid K^l \sim \mathcal{GP}(z^l; \mathbf{0}, G(K^l))$$

where

$$\begin{aligned}
K^l(x^{(p)}, x^{(q)}) &= \frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x^{(p)})) \phi(z_j^{l-1}(x^{(q)})) \\
G(K^l(x^{(p)}, x^{(q)})) &= \sigma_b^2 + \sigma_w^2 K^l(x^{(p)}, x^{(q)})
\end{aligned}$$

Conveniently, we use K^l as a short-notation collection of $\phi(z_j^{l-1}(x^{(p)}))$, $\phi(z_j^{l-1}(x^{(q)})) \quad \forall p, q, j$

- also taking care of the layer one, which is just input x :

$$K_{p,q}^l \equiv K^l(x^{(p)}, x^{(q)}) = \begin{cases} \frac{1}{d_{\text{in}}} \sum_{j=1}^{d_{\text{in}}} x_j^{(p)} x_j^{(q)} & l = 0 \\ \frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x^{(p)})) \phi(z_j^{l-1}(x^{(q)})) & l > 0 \end{cases}$$

- to reflect:

$$\text{Cov}(z_k^l, z_{k'}^l) = 0 \quad \forall k, k' \in \{1, \dots, N_{l+1}\}$$

one may construct giant co-variance matrix with $N_{l+1} \times N_{l+1}$ diagonal blocks:

$$\mathbf{z}^l = \underbrace{\begin{bmatrix} \color{red}{z_1^l(x^{(1)})} & \color{red}{z_1^l(x^{(2)})} & \dots & z_1^l(x^{(|\mathcal{D}|)}) \\ \vdots & \vdots & \dots & \vdots \\ z_j^l(x^{(1)}) & z_j^l(x^{(2)}) & \dots & z_j^l(x^{(|\mathcal{D}|)}) \\ \vdots & \vdots & \ddots & \vdots \\ z_{N_{l+1}}^l(x^{(1)}) & z_{N_{l+1}}^l(x^{(2)}) & \dots & z_{N_{l+1}}^l(x^{(|\mathcal{D}|)}) \end{bmatrix}}_{|\mathcal{D}|} \left. \vphantom{\begin{bmatrix} \color{red}{z_1^l(x^{(1)})} & \color{red}{z_1^l(x^{(2)})} & \dots & z_1^l(x^{(|\mathcal{D}|)}) \\ \vdots & \vdots & \dots & \vdots \\ z_j^l(x^{(1)}) & z_j^l(x^{(2)}) & \dots & z_j^l(x^{(|\mathcal{D}|)}) \\ \vdots & \vdots & \ddots & \vdots \\ z_{N_{l+1}}^l(x^{(1)}) & z_{N_{l+1}}^l(x^{(2)}) & \dots & z_{N_{l+1}}^l(x^{(|\mathcal{D}|)}) \end{bmatrix}} \right\} \text{width} \Rightarrow \text{vec}(\mathbf{z}^l) = \begin{bmatrix} \color{red}{z_1^l(x^{(1)})} \\ z_2^l(x^{(1)}) \\ \vdots \\ z_{N_{l+1}}^l(x^{(1)}) \\ \color{red}{z_1^l(x^{(2)})} \\ z_2^l(x^{(2)}) \\ \vdots \\ z_{N_{l+1}}^l(x^{(2)}) \\ \vdots \\ z_1^l(x^{(|\mathcal{D}|)}) \\ z_2^l(x^{(|\mathcal{D}|)}) \\ \vdots \\ z_{N_{l+1}}^l(x^{(|\mathcal{D}|)}) \end{bmatrix}$$

$$\sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} G(K_{1,1}^l) & \dots & 0 & \dots & G(K_{1,|\mathcal{D}|}^l) & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & G(K_{1,1}^l) & \dots & 0 & 0 & G(K_{1,|\mathcal{D}|}^l) \\ \color{red}{G(K_{2,1}^l)} & \dots & 0 & \dots & G(K_{2,|\mathcal{D}|}^l) & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & G(K_{2,1}^l) & \dots & 0 & 0 & G(K_{2,|\mathcal{D}|}^l) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ G(K_{|\mathcal{D}|,1}^l) & \dots & 0 & \dots & G(K_{|\mathcal{D}|,|\mathcal{D}|}^l) & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & G(K_{|\mathcal{D}|,1}^l) & \dots & 0 & 0 & G(K_{|\mathcal{D}|,|\mathcal{D}|}^l) \end{bmatrix} \right)$$

$$\Rightarrow p(\mathbf{z}^l | K^l) = \mathcal{N}(\mathbf{0}, G(K^l) \otimes \mathbf{I}_{N_{l+1} \times N_{l+1}})$$

$$= \mathcal{GP}(\mathbf{z}^l; \mathbf{0}, G(K^l))$$

5.3 $p(K^l | K^{l-1})$

Use marginal property of GP and look at: $p(K^l | K^{l-1})$:

$$\begin{aligned}
p(K^l|K^{l-1}) &= \int_{z^{l-1}} p(K^l|z^{l-1})p(z^{l-1}|K^{l-1}) \\
&= \int_{z^{l-1}} p(K^l|z^{l-1})\mathcal{GP}(z^{l-1}; 0, G(K^{l-1}))
\end{aligned}$$

- using GP property, and just look at two points $x^{(p)}, x^{(q)}$:

$$\begin{aligned}
p(K_{p,q}^l|K_{p,q}^{l-1}) &= \int_{z^{l-1}(x^{(p)}), z^{l-1}(x^{(q)})} p\left(\frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^l(x^{(p)}))\phi(z_j^l(x^{(q)}))\right) \\
&\quad \mathcal{N}\left(\begin{bmatrix} z^{l-1}(x^{(p)}) \\ z^{l-1}(x^{(q)}) \end{bmatrix}; 0, G\left(\begin{bmatrix} K^{l-1}(x^{(p)}, x^{(p)}) & K^{l-1}(x^{(p)}, x^{(q)}) \\ K^{l-1}(x^{(q)}, x^{(p)}) & K^{l-1}(x^{(q)}, x^{(q)}) \end{bmatrix}\right)\right)
\end{aligned}$$

5.3.1 what happen to sum $\sum_{j=1}^{N_l} \phi(z_j^{l-1}(x^{(p)}))\phi(z_j^{l-1}(x^{(q)}))$ as $N_l \rightarrow \infty$ using CLT:

- look at $K_{p,q}^l$ and notice it's sum of iid random variable $K_{p,q}^{l,j}$:

$$\begin{aligned}
\underbrace{K_{p,q}^l}_{\bar{X}} &= \frac{1}{N_l} \sum_{j=1}^{N_l} \underbrace{\phi(z_j^{l-1}(x^{(p)}))\phi(z_j^{l-1}(x^{(q)}))}_{X_j \equiv K_{p,q}^{l,j}} \\
\Rightarrow p(K_{p,q}^{l,1}|K_{p,q}^{l-1}) &= \int_{z^{l-1}(x^{(p)}), z^{l-1}(x^{(q)})} p(\phi(z_j^l(x^{(p)}))\phi(z_j^l(x^{(q)}))) \\
&\quad \mathcal{N}\left(\begin{bmatrix} z^{l-1}(x^{(p)}) \\ z^{l-1}(x^{(q)}) \end{bmatrix}; 0, G\left(\begin{bmatrix} K^{l-1}(x^{(p)}, x^{(p)}) & K^{l-1}(x^{(p)}, x^{(q)}) \\ K^{l-1}(x^{(q)}, x^{(p)}) & K^{l-1}(x^{(q)}, x^{(q)}) \end{bmatrix}\right)\right) \\
&= (F \circ G)(K_{p,q}^{l-1})
\end{aligned}$$

- using CLT, pick the most appropriate definition:

$$(\bar{X} - \mathbb{E}[X_1]) \xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}[X_1]}{n}\right)$$

- let's see what is $\lim_{N_l \rightarrow \infty} p(K^l|K^{l-1})$:

$$\begin{aligned}
&(\bar{X} - \mathbb{E}[X_1]) \xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}[X_1]}{n}\right) \\
\Rightarrow (K_{p,q}^l - \mathbb{E}[K_{p,q}^{l,1}]) &\xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}[K_{p,q}^{l,1}]}{N_l}\right) \\
\Rightarrow (K_{p,q}^l - (F \circ G)(K_{p,q}^{l-1})) &\xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}[K_{p,q}^{l,1}]}{N_l}\right) \\
\Rightarrow (K_{p,q}^l|K_{p,q}^{l-1}) &\xrightarrow{d} \mathcal{N}\left((F \circ G)(K^{l-1}), \frac{\text{Var}[K_{p,q}^{l,1}]}{N_l}\right) \\
\Rightarrow \lim_{N_l \rightarrow \infty} p(K^l|K^{l-1}) &= \delta(K^l - (F \circ G)(K^{l-1})) \quad \text{entire matrix}
\end{aligned}$$

- **note** using CLT, sample mean converge to δ_μ , can be exploited for other application
- note that this single step conditional is quite easy

5.4 putting in the overall objective function

let width of all layers to $\rightarrow \infty$:

$$\begin{aligned}
p(z^L|x) &= \int p(z^L, K^0, K^1, \dots, K^L|x) \, dK^{0,\dots,L} \\
&= \int p(z^L|K^L) \left(\prod_{l=1}^L p(K^l|K^{l-1}) \right) p(K^0|x) \, dK^{0,\dots,L} \\
\lim_{N_L \rightarrow \infty, \dots, N_1 \rightarrow \infty} p(z^L|x) &= \int p(z^L|K^L) \left(\prod_{l=1}^L \delta(K^l - (F \circ G)(K^{l-1})) \right) p(K^0|x) \, dK^{0,\dots,L} \\
&= \int \mathcal{GP}(z^L; 0, G(K^L)) \underbrace{\left(\prod_{l=1}^L \delta(K^l - (F \circ G)(K^{l-1})) \right) \delta\left(K^0 - \frac{1}{d_{\text{in}}} x^\top x\right)}_{= \begin{cases} 1 & \text{if } K^L = (F \circ G)(K^{L-1}) \\ & = (F \circ G)^2(K^{L-2}) \dots \\ & = (F \circ G)^L\left(\frac{1}{d_{\text{in}}} x^\top x\right) \\ 0 & \text{otherwise} \end{cases}} \, dK^{0,\dots,L} \\
&= \mathcal{GP}\left(z^L; 0, G \circ (F \circ G)^L\left(\frac{1}{d_{\text{in}}} x^\top x\right)\right)
\end{aligned}$$

6 Neural Tangent Kernel

6.1 what do you hope for functional gradient $\frac{\delta C}{\delta f(\theta)}$

- Under some training regime, e.g., **gradient descend**, there will be parameter dynamics $\frac{d\theta}{dt}$:
- what you hope: under $\frac{d\theta}{dt}$, functional gradient $\frac{\delta C}{\delta f(\theta)}$ is **negative all the time!**

6.1.1 Change Cost C with respect to parameter θ

More concretely, **question** is if we vary the parameters of our network:

$$\theta \rightarrow \theta + \epsilon \eta$$

in a **direction vector** η , then, how does the cost change. Formally, we want the following limit:

$$\lim_{\epsilon \rightarrow 0} \frac{C[f(\theta + \epsilon \eta)] - C[f(\theta)]}{\epsilon}$$

it turns out functional $C[f]$ is in square bracket

- Since f is a function, we need functional solution **Riesz-Markov-Kakutani Representation Theorem** says:

$$\int_{\mathbf{X}} \frac{\delta J}{\delta g}(x)^\top \phi(x) dx = \lim_{\epsilon \rightarrow 0} \frac{J[g + \epsilon \phi] - J[g]}{\epsilon}$$

- if g was a variable instead of a function, then, the above is analogous to:

$$\phi^\top \nabla_g J$$

i.e., directional derivative of J in the direction of ϕ , and there is no integral $\int_{\mathbf{X}} dx$!

- we can **not** substitute into RMK Representation directly, because our changes $\epsilon \eta$ occur in function's argument:

$$\lim_{\epsilon \rightarrow 0} \frac{C[f(\theta + \epsilon \eta)] - C[f^\theta \eta]}{\epsilon}$$

- But we must get it in the form of $C[f(\theta) + \epsilon \eta]$. Therefore, we need to use Taylor Expansion:

$$\begin{aligned}
& C\left(\underbrace{f^\theta}_g + \epsilon \underbrace{\eta \cdot \frac{\partial f^\theta}{\partial \theta}}_\phi + O(\epsilon^2)\right) - C(f^\theta) \\
\Rightarrow \lim_{\epsilon \rightarrow 0} & \frac{\epsilon}{\epsilon} \quad \text{matching with RMK representation} \\
& = \int_{\mathbf{x}} \left(\frac{\delta C}{\delta f(\theta)}(x) \right)^\top \left(\eta \cdot \frac{\partial f(\theta)}{\partial \theta} \right) dx \\
& = \sum_{d=1}^{|\theta|} \int_{\mathbf{x}} \left(\frac{\delta C}{\delta f(\theta)}(x) \right)^\top \left(\eta \cdot \frac{\partial f(\theta)}{\partial \theta_d} \right) dx \\
& = \sum_{d=1}^{|\theta|} \int_{\mathbf{x}} \sum_{i=1}^N \left(\frac{\delta C}{\delta f(\theta)}(x) \right)_i \left(\eta \cdot \frac{\partial f(\theta)}{\partial \theta_d} \right)_i dx \\
& = \sum_{d=1}^{|\theta|} \sum_{i=1}^N \eta \int_{\mathbf{x}} \left(\frac{\delta C}{\delta f(\theta)}(x) \right)_i \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_i dx \quad \text{change order of integral and sum}
\end{aligned}$$

6.1.2 Change Cost C with respect to parameter dynamics $\frac{\partial \theta}{\partial t}$

- above tells how much does C change if $\theta \rightarrow \theta + \epsilon \eta$
- since we can choose any direction η , we equally choose a direction to be **parameter dynamics**, i.e.:

$$\eta \equiv \frac{\partial \theta}{\partial t}$$

- by substitution:

$$\lim_{\epsilon \rightarrow 0} \frac{C[f(\theta + \epsilon \eta)] - C[f^\theta]}{\epsilon} = \sum_{d=1}^{|\theta|} \sum_{i=1}^N \left(\frac{\partial \theta}{\partial t} \right) \int_{\mathbf{x}} \left(\frac{\delta C}{\delta f(\theta)}(x) \right)_i \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_i dx$$

- if **gradient descent** regime is used:

$$\begin{aligned}
\frac{\partial \theta}{\partial t} &= - \frac{\partial C[f(\theta)]}{\partial \theta} \\
&= - \lim_{\epsilon \rightarrow 0} \frac{C[f(\theta + \epsilon \mathbf{I})] - C[f^\theta]}{\epsilon} \\
&= - \sum_{d=1}^{|\theta|} \sum_{k=1}^N \int_{\mathbf{x}'} \left(\frac{\delta C}{\delta f(\theta)}(x') \right)_k \left(\frac{\partial f_k(\theta)}{\partial \theta_d} \right)_k dx' \quad \text{change index to } k \text{ and } x \rightarrow x'
\end{aligned}$$

- substitution:

$$\begin{aligned}
& \lim_{\epsilon \rightarrow 0} \frac{C[f(\theta + \epsilon \eta)] - C[f^\theta]}{\epsilon} \\
&= \sum_{d=1}^{|\theta|} \sum_{i=1}^N \left(\frac{\partial \theta}{\partial t} \right) \int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x) \right)_i \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_i dx \\
&= - \sum_{d=1}^{|\theta|} \sum_{i=1}^N \left[\sum_{d=1}^{|\theta|} \sum_{k=1}^N \int_{\mathbf{X}'} \left(\frac{\delta C}{\delta f(\theta)}(x') \right)_k \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_k dx' \right] \left[\int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x) \right)_i \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_i dx \right] \\
&= - \sum_{d=1}^{|\theta|} \sum_{i=1}^N \left[\sum_{k=1}^N \int_{\mathbf{X}'} \left(\frac{\delta C}{\delta f(\theta)}(x') \right)_k \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_k dx' \right] \int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x) \right)_i \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_i dx \\
&= - \sum_{d=1}^{|\theta|} \sum_{i=1}^N \sum_{k=1}^N \left(\int_{\mathbf{X}'} \left(\frac{\delta C}{\delta f(\theta)}(x') \right)_k \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_k dx' \right) \left(\int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x) \right)_i \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_i dx \right) \\
&\quad \text{can take sum } \sum_k \text{ out of bracket because second term has no } k \\
&= - \sum_{d=1}^{|\theta|} \sum_{i=1}^N \sum_{k=1}^N \int_{\mathbf{X}'} \int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x') \right)_k \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_k \left(\frac{\delta C}{\delta f(\theta)}(x) \right)_i \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_i dx dx' \\
&= - \int_{\mathbf{X}'} \int_{\mathbf{X}} \underbrace{\sum_{i=1}^N \sum_{k=1}^N \left(\frac{\delta C}{\delta f(\theta)}(x') \right)_k \left[\sum_{d=1}^{|\theta|} \left(\frac{\partial f_k(\theta)}{\partial \theta_d} \right)_k \left(\frac{\partial f(\theta)}{\partial \theta_d} \right)_i \right] \left(\frac{\delta C}{\delta f(\theta)}(x) \right)_i}_{\sum_i \sum_j x_i x_j M_{i,j} = \bar{x}^\top M \bar{x}} dx dx' \\
&= - \int_{\mathbf{X}'} \int_{\mathbf{X}} \underbrace{\left(\frac{\delta C}{\delta f(\theta)}(x') \right)^\top}_{\Theta(x, x')} \Theta(x, x') \left(\frac{\delta C}{\delta f(\theta)}(x) \right) dx dx'
\end{aligned}$$

- note $\Theta(x, x')$ above has nothing to do Neural Networks, i.e., the above is true under gradient descent regardless of $f(\theta)$ used

6.1.3 What happens $\Theta(x^{(p)}, x^{(q)})$ is positive definite

- the above implies that if NTK is positive definite (which is the NTK paper is all about):

$$\lim_{\epsilon \rightarrow 0} \frac{C[f(\theta + \epsilon \frac{\partial \theta}{\partial t})] - C[f^\theta]}{\epsilon} = \text{negative value}$$

cost will converge to a global optima.

- it's not immediately apparent why the term inside the integral is actually positive. to make it clearer, we rewrite the following using simple notations:

$$\int_{x^{(p)}} \int_{x^{(q)}} \underbrace{\tilde{f}(x^{(p)})^\top \Theta(x^{(p)}, x^{(q)}) \tilde{f}(x^{(q)})}_{\text{positive}}$$

- for a specific term

$$\bar{f}(x^{(p)})^\top K(x^{(p)}, x^{(q)}) \bar{f}(x^{(q)})$$

it may not be positive as left vector $\bar{f}(x^{(p)})$ and right vector $\bar{f}(x^{(q)})$ may not equate. However, by summing all **four** elements concerning the co-efficient of $\Theta(i, j) \equiv \Theta_{i,j}(x^{(p)}, x^{(q)})$:

$$\begin{aligned} A &\equiv \textcolor{red}{f_i(x^{(p)})} \Theta(i, j) f_j(x^{(p)}) + \textcolor{red}{f_i(x^{(p)})} \Theta(i, j) f_j(x^{(q)}) = \textcolor{red}{f_i(x^{(p)})} \Theta(i, j) (f_j(x^{(p)}) + f_j(x^{(q)})) \\ B &\equiv \textcolor{blue}{f_i(x^{(q)})} \Theta(i, j) f_j(x^{(p)}) + \textcolor{blue}{f_i(x^{(q)})} \Theta(i, j) f_j(x^{(q)}) = \textcolor{blue}{f_i(x^{(q)})} \Theta(i, j) (f_j(x^{(p)}) + f_j(x^{(q)})) \\ A + B &= \underbrace{(f_i(x^{(p)}) + f_i(x^{(q)}))}_{g_i(x^{(p)}, x^{(q)})} \Theta(i, j) \underbrace{(f_j(x^{(p)}) + f_j(x^{(q)}))}_{g_j(x^{(p)}, x^{(q)})} \end{aligned}$$

it's invariant to value in $x^{(p)}$ and $x^{(q)}$, as both are used

6.1.4 What does NTK paper aims to prove

- NTK paper is all about proving in addition to gradient descend regime:

1. $f(\theta)$ is neural network
2. Gaussian initialization is applied
3. having $N_1, \dots, N_L \rightarrow \infty$:

then,

1. NTK is indeed positive definite, in a Scalar matrix form: “some positive scalar” $\times \mathbf{I}_{N_{l+1}}$
2. remains approximately constant throughout training

Consequently, leading $\lim_{\epsilon \rightarrow 0} \frac{C[f(\theta + \epsilon \frac{\partial \theta}{\partial t})] - C[f(\theta)]}{\epsilon}$ to stay negative, i.e., cost always going down

6.2 NTK in Neural Networks

- we use the re-parameterization version of NN function:

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

where $W_{k,j}^l, b_k^l \sim \mathcal{N}(0, 1)$

- **Neural Tangent Kernel** at each Layer l :

$$\begin{aligned}
\Theta^l(x^{(p)}, x^{(q)}) &= \sum_{d=1}^{|\theta|} \frac{\partial z^l(x^{(p)})}{\partial \theta_d} \otimes_{\text{outer}} \frac{\partial z^l(x^{(q)})}{\partial \theta_d} \\
&= \sum_{d=1}^{|\theta|} \begin{bmatrix} \frac{\partial z_1^l(x^{(p)})}{\partial \theta_d} & \dots & \frac{\partial z_{N_{l+1}}^l(x^{(p)})}{\partial \theta_d} \end{bmatrix} \begin{bmatrix} \frac{\partial z_1^l(x^{(q)})}{\partial \theta_d} & \dots & \frac{\partial z_{N_{l+1}}^l(x^{(q)})}{\partial \theta_d} \end{bmatrix}^\top \\
&= \sum_{d=1}^{|\theta|} \begin{bmatrix} \frac{\partial z_1^l(x^{(p)})}{\partial \theta_d} \frac{\partial z_1^l(x^{(q)})}{\partial \theta_d} & \dots & \frac{\partial z_{N_{l+1}}^l(x^{(p)})}{\partial \theta_d} \frac{\partial z_{N_{l+1}}^l(x^{(q)})}{\partial \theta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_1^l(x^{(p)})}{\partial \theta_d} \frac{\partial z_1^l(x^{(q)})}{\partial \theta_d} & \dots & \frac{\partial z_{N_{l+1}}^l(x^{(p)})}{\partial \theta_d} \frac{\partial z_{N_{l+1}}^l(x^{(q)})}{\partial \theta_d} \end{bmatrix}
\end{aligned}$$

- note that size of Θ^l is $N_{l+1} \times N_{l+1}$, it has nothing to do with $|\theta|$ (it is used in the sum)
- loosely speaking:
 1. NTK studies “pseudo-correlations” between a pair of output (k, k') of a vector function z^l by summing over their derivatives over all parameters.
(derivative correlations between **function’s output**)
 2. which is opposite to fisher information matrix:

$$\mathbf{I}_{i,j} = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta_i} \log f(X; \theta) \right) \left(\frac{\partial}{\partial \theta_j} \log f(X; \theta) \right) \right]$$

where FIM studies correlation between log derivative of pair of parameters (θ_i, θ_j) from a scalar function f .
(derivative correlations between **function’s parameters**)

- note that symbol here \otimes above is outer product as oppose to kronecker product everywhere else in this tutorial. But the two are related:

$$\mathbf{u} \otimes_{\text{Kron}} \mathbf{v}^\top = \mathbf{u} \mathbf{v}^\top = \mathbf{u} \otimes_{\text{outer}} \mathbf{v}$$

6.3 NTK at initialization

in GP, we have the following matrix to showcase correlation between $x^{(p)}$ and $x^{(q)}$, but the same k for $z_k^l(x)$:

$$\begin{bmatrix} \vdots \\ W_{k,1}^l \phi(z_1^{l-1}(x^{(p)})) \\ \vdots \\ W_{k,1}^l \phi(z_j^{l-1}(x^{(q)})) \\ \vdots \end{bmatrix} + \dots + \begin{bmatrix} \vdots \\ W_{k,N_l}^l \phi(z_j^{l-1}(x^{(p)})) \\ \vdots \\ W_{k,N_l}^l \phi(z_j^{l-1}(x^{(q)})) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(p)})) \\ \vdots \\ \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(q)})) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ z_k^l(x^{(p)}) \\ \vdots \\ z_k^l(x^{(q)}) \\ \vdots \end{bmatrix}$$

in NTK, we change this to showcase correlations in z_k^l and $z_{k'}^l$, and having the same x :

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

6.3.1 when $l = 1$

$$\begin{bmatrix} \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W_{1,j}^1 \phi(x_1) + \sigma_b b_1^1 \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W_{k,j}^1 \phi(x_k) + \sigma_b b_k^1 \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W_{N_1,j}^1 \phi(x_{N_1}) + \sigma_b b_{N_1}^1 \end{bmatrix} = \begin{bmatrix} z_1^1(x) \\ \vdots \\ z_k^1(x) \\ \vdots \\ z_{N_1}^1(x) \end{bmatrix}$$

- when computing: $\frac{\partial z_k^1(x)}{\partial W_{i,j}}$, we use i here, because k is fixed by $z_k^1(x)$:
- note when computing $\frac{\partial z_k^1(x)}{\partial W_{i,j}}$ only k^{th} row going to return a gradient!

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

- now, each element of the outer product matrix $\Theta^l(x^{(p)}, x^{(q)}) = \sum_{d=1}^{|\theta|} \frac{\partial F_k^l(x^{(p)})}{\partial \theta_d} \otimes \frac{\partial F_{k'}^l(x^{(q)})}{\partial \theta_d}$ at k, k' is:

$$\begin{aligned}
\Theta_{k,k'}^1(x^{(p)}, x^{(q)}) &= \sum_{d=1}^{|\theta^1|} \frac{\partial F_k^1(x^{(p)})}{\partial \theta_d^1} \frac{\partial F_{k'}^1(x^{(q)})}{\partial \theta_d^1} \quad \theta^1 = \{W^1, b^1\} \\
&= \sum_{d=1}^{|W^1|} \frac{\partial F_k^1(x^{(p)})}{\partial W_d^1} \frac{\partial F_{k'}^1(x^{(q)})}{\partial W_d^1} + \sum_{d=1}^{|b^1|} \frac{\partial F_k^1(x^{(p)})}{\partial b_d^1} \frac{\partial F_{k'}^1(x^{(q)})}{\partial b_d^1} \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{d_{\text{in}}} \frac{\partial z_k^1(x^{(p)})}{\partial W_{i,j}} \frac{\partial z_{k'}^1(x^{(q)})}{\partial W_{i,j}} + \sum_{i=1}^{N_1} \frac{\partial z_k^1(x^{(p)})}{\partial b_i} \frac{\partial z_{k'}^1(x^{(q)})}{\partial b_i} \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{d_{\text{in}}} \frac{1}{\sqrt{d_{\text{in}}}} x_i^{(p)} \delta_{i,k'} \frac{1}{\sqrt{d_{\text{in}}}} x_i^{(q)} \delta_{i,k} + \sum_{i=1}^{N_1} \sigma_b \delta_{i,k} \sigma_b \delta_{i,k'} \quad \text{only one } i \in \{1, \dots, N_1\} \text{ in outer sum remain} \\
&= \sum_{j=1}^{d_{\text{in}}} \frac{1}{d_{\text{in}}} x_i^{(p)} x_i^{(q)} \delta_{k,k'}^2 + \sigma_b^2 \delta_{k,k'} \quad \delta_{i,k'} \delta_{i,k} = \delta_{k,k'} \\
&= \frac{1}{d_{\text{in}}} x^{(p)\top} x^{(q)} \delta_{k,k'} + \sigma_b^2 \delta_{k,k'} \\
&= \underbrace{\left(\frac{1}{d_{\text{in}}} x^{(p)\top} x^{(q)} + \sigma_b^2 \right)}_{K^1} \delta_{k,k'} \\
&\equiv K^1(x^{(p)}, x^{(q)}) \delta_{k,k'} \quad \text{conform to notation in GP for NN section}
\end{aligned}$$

- now we have each element $\Theta_{k,k'}^l$, the final Θ^l is:

$$\begin{aligned}
\Rightarrow \Theta^1(x^{(p)}, x^{(q)}) &= \begin{bmatrix} G(K^1)(x^{(p)}, x^{(q)}) & 0 & \dots & 0 \\ 0 & K^1(x^{(p)}, x^{(q)}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & K^1(x^{(p)}, x^{(q)}) \\ 0 & 0 & 0 & 0 & K^1(x^{(p)}, x^{(q)}) \end{bmatrix} \\
&\quad \text{repeating diagonal with } K^1(x^{(p)}, x^{(q)}) \\
&= \underbrace{K^1(x^{(p)}, x^{(q)})}_{\text{scalar}} \mathbf{I}_{N_{l+1} \times N_{l+1}}
\end{aligned}$$

Θ^1 matrix of square the size of input $|z^1|$

- importantly, there is no limit to take for Θ^1

6.3.2 when $l > 1$

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

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

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

- looking at this specific term: $\frac{\partial z_k^1(x^{(p)})}{\partial \theta_d^{l-1}}$, write $x^{(p)} \equiv x$, and definition again:

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

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

- substitution:

$$\begin{aligned}
& \sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^1(x^{(p)})}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x^{(q)})}{\partial \theta_d^{l-1}} \\
&= \sum_{d=1}^{|\theta^{l-1}|} \left(\frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi'(z_j^{l-1}(x^{(p)})) \frac{\partial z_j^{l-1}(x^{(p)})}{\partial \theta_d^{l-1}} \right) \times \left(\frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k',j}^l \phi'(z_j^{l-1}(x^{(q)})) \frac{\partial z_j^{l-1}(x^{(q)})}{\partial \theta_d^{l-1}} \right) \\
&= \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 \phi'(z_j^{l-1}(x^{(p)})) \frac{\partial z_j^{l-1}(x^{(p)})}{\partial \theta_d^{l-1}} \right) \times \underbrace{\left(W_{k',j'}^l \phi'(z_{j'}^{l-1}(x^{(q)})) \frac{\partial z_{j'}^{l-1}(x^{(q)})}{\partial \theta_d^{l-1}} \right)}_{j \rightarrow j'} \\
&= \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 \phi'(z_j^{l-1}(x^{(p)})) \phi'(z_{j'}^{l-1}(x^{(q)})) \frac{\partial z_j^{l-1}(x^{(p)})}{\partial \theta_d^{l-1}} \frac{\partial z_{j'}^{l-1}(x^{(q)})}{\partial \theta_d^{l-1}} \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \phi'(z_j^{l-1}(x^{(p)})) \phi'(z_{j'}^{l-1}(x^{(q)})) \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^{(p)}, x^{(q)})} \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \phi'(z_j^{l-1}(x)) \phi'(z_{j'}^{l-1}(x)) \Theta_{j,j'}^{l-1}(x^{(p)}, x^{(q)}) \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \phi'(z_j^{l-1}(x^{(p)})) \phi'(z_{j'}^{l-1}(x^{(q)})) \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)}) \delta_{j,j'} \\
&\quad \text{assume } \Theta_{j,j'}^{l-1}(x^{(p)}, x^{(q)}) \rightarrow \text{deterministic and diagonal limit } \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)}) \delta_{j,j'} \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} W_{k,j}^l W_{k',j}^l \phi'(z_j^{l-1}(x^{(p)})) \phi'(z_j^{l-1}(x^{(q)})) \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)}) \quad \text{change } j' \rightarrow j \text{ and remove } \sum_{j'=1}^{N_l} W_{k,j}^l
\end{aligned}$$

- apply CLT, we know that:

$$\begin{aligned}
& \mathbb{E} \left[\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^1(x^{(p)})}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x^{(q)})}{\partial \theta_d^{l-1}} \right] \\
&= \mathbb{E} \left[W_{k,j}^l W_{k',j}^l \phi'(z_j^{l-1}(x^{(p)})) \phi'(z_j^{l-1}(x^{(q)})) \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)}) \right] \quad N_l \rightarrow \infty \\
&= \mathbb{E} \left[W_{k,j}^l W_{k',j}^l \right] \mathbb{E} \left[\phi'(z_j^{l-1}(x^{(p)})) \phi'(z_j^{l-1}(x^{(q)})) \right] \underbrace{\Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)})}_{\text{constant}} \\
&= \delta_{k,k'} \mathbb{E} \left[\phi'(z_j^{l-1}(x^{(p)})) \phi'(z_j^{l-1}(x^{(q)})) \right] \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)})
\end{aligned}$$

- we have seen previously Eq. (1):

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

- however, this time we need to define a similar auxiliary variable \tilde{K}^l , notice it **has no** σ_b^2 term, describing expectation of $\phi'()$

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

- also notice the above equation is **not a recursion**, i.e., $\dot{K}^l(x^{(p)}, x^{(q)})$ and $K^{l-1}(x^{(p)}, x^{(q)})$ are not the same thing.

$$\begin{aligned}
&= \delta_{k,k'} \mathbb{E}_{(z_j^{l-1}(x^{(p)}), z_j^{l-1}(x^{(q)}))} \sim \mathcal{N}(0, K^{l-1}(x^{(p)}, x^{(q)})) \left[\phi'(z_j^{l-1}(x^{(p)})) \phi'(z_j^{l-1}(x^{(q)})) \right] \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)}) \\
&= \delta_{k,k'} \dot{K}^l(x^{(p)}, x^{(q)}) \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)})
\end{aligned}$$

- look at $\{W^l, b^l\}$ part:

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

and compare that with for $l = 1$:

$$\sum_{d=1}^{|\theta^1|} \frac{\partial F_k^1(x^{(p)})}{\partial \theta_d^1} \frac{\partial F_{k'}^1(x^{(q)})}{\partial \theta_d^1} \quad \theta^1 = \{W^1, b^1\}$$

it's the same if we replace

$$\left(K^1(x^{(p)}, x^{(q)}) \equiv \frac{1}{d_{\text{in}}} x^{(p)\top} x^{(q)} + \sigma_b^2 \right) \delta_{k,k'} \rightarrow \left(K^l(x^{(p)}, x^{(q)}) \equiv \frac{1}{N_l} \phi(z^l(x^{(p)}))^{\top} \phi(z^l(x^{(q)})) + \sigma_b^2 \right) \delta_{k,k'}$$

$$\begin{aligned}
\Theta_{k,k'}^l(x^{(p)}, x^{(q)}) &= \sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^l(x^{(p)})}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x^{(q)})}{\partial \{W^l, b^l\}} + \sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^1(x^{(p)})}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^1(x^{(q)})}{\partial \theta_d^{l-1}} \\
&= K^l(x^{(p)}, x^{(q)}) \delta_{k,k'} + \delta_{k,k'} \dot{K}^l(x^{(p)}, x^{(q)}) \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)}) \\
&= \left(K^l(x^{(p)}, x^{(q)}) + \dot{K}^l(x^{(p)}, x^{(q)}) \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)}) \right) \delta_{k,k'} \\
&\text{repeating diagonal with } K^l(x^{(p)}, x^{(q)}) + \dot{K}^l(x^{(p)}, x^{(q)}) \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)}) \\
&= \underbrace{\left(K^l(x^{(p)}, x^{(q)}) + \dot{K}^l(x^{(p)}, x^{(q)}) \Theta_{\infty}^{l-1}(x^{(p)}, x^{(q)}) \right)}_{\text{scalar}} \mathbf{I}_{N_{l+1} \times N_{l+1}}
\end{aligned}$$