

Neural Networks Gaussian Process and Neural Tangent Kernel Initialization

Richard Xu

April 19, 2021

1 Preamble

In this tutorial, my contribution mainly has been the attempt to summarize the referenced papers and blogs in a unified and (hopefully) more intuitive for Computer Science researchers. I also tried to provide some gentle introduction to people on what is a Gaussian Process and Kernel methods, in order to make this tutorial a bit self contained.

This document contains contents up to NTK for initialization, and the NTK training will be in another document.

In particular, the blogs below are extremely useful, and I encourage you to read the original blog as well.

1. <https://www.uv.es/gonmagar/blog/2019/01/21/DeepNetworksAsGPs>
2. <https://bryn.ai/jekyll/update/2019/04/02/neural-tangent-kernel.html>
3. <http://chenyilan.net/>

1.1 notations

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

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

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 and x' 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

To make this tutorial self-contained, I have included a very quick introduction on the relevant topics include Gaussian Process and Central Limit Theorem

2 Gaussian Process

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

- \mathcal{GP} is a (potentially infinite) collection of RVs, s.t., joint distribution of every finite subset of RVs is multivariate Gaussian:

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

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

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

2.1 marginal and conditional marginal under noisy output

- in a regression setting:

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

- joint $[\mathcal{Y}, y^*]^\top$, after integrate out f :

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

- **predictive distribution** of $y^*|\mathcal{Y}$ using conditional formula of multivariate Gaussian:

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

2.2 marginal and conditional marginal under noiseless output

- **posterior** of f given \mathcal{Y} in regression:

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

for arbitrary variable \mathbf{x}

conditional marginal of $y^*|\mathcal{Y}$ using conditional formula of multivariate Gaussian:

$$p(f|\mathcal{X}, \mathcal{Y}) = \mathcal{GP}\left(K(\mathbf{x}, \mathcal{X})(K(\mathcal{X}, \mathcal{X}) + \sigma_\epsilon^2 \mathbf{I})^{-1} \mathcal{Y},\right. \\ \left. k(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathcal{X})(K(\mathcal{X}, \mathcal{X}) + \sigma_\epsilon^2 \mathbf{I})^{-1} K(\mathcal{X}, \mathbf{x}')\right)$$

- **deterministic function** $y_i = f(x_i)$ is used, e.g., neural network's read-out layer $f(x_i)$, data y_i

$p([\mathcal{Y}, y^*]^\top)$ no longer need to integrate f :

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

predictive distribution $y^*|\mathcal{Y}$ using conditional formula of multivariate Gaussian:

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

3 Kernel methods

consider the equation:

$$\begin{aligned}
 y &= \phi(x)^\top \mathbf{w} \\
 &= \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_m(x) \end{bmatrix}^\top \mathbf{w} \\
 &= [\phi_1(x) \quad \dots \quad \phi_m(x)] \mathbf{w}
 \end{aligned} \tag{2}$$

using definition:

$$\begin{aligned}
 \mathcal{Y} &= [y_1, \dots, y_n]^\top \\
 \Phi &= [\phi(x_1), \dots, \phi(x_n)]^\top \\
 &= \underbrace{\begin{bmatrix} \phi_1(x_1) & \dots & \phi_m(x_1) \\ \vdots & & \vdots \\ \phi_1(x_n) & \dots & \phi_m(x_n) \end{bmatrix}}_{n \times m}
 \end{aligned} \tag{3}$$

Ridge regression can be re-written as:

$$\begin{aligned}
 \mathbf{w}^* &= \arg \min_{\mathbf{w}} \sum_{i=1}^n (y_i - \phi(x_i)^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2 \\
 &= \arg \min_{\mathbf{w}} \|\mathcal{Y} - \Phi \mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2
 \end{aligned} \tag{4}$$

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

$$\mathbf{w}^* = (\underbrace{\Phi^\top \Phi}_{m \times m} + \lambda I)^{-1} \Phi^\top \mathcal{Y} \tag{5}$$

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

$$\begin{aligned}
 y_{\mathbf{w}^*}(x) &= \phi(x)^\top \mathbf{w}^* \\
 &= \phi(x)^\top (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top \mathcal{Y} \\
 &= \underbrace{\phi(x)^\top \Phi^\top}_{1 \times n} (\underbrace{\Phi \Phi^\top}_{n \times n} + \lambda I)^{-1} \mathcal{Y} \\
 &\text{using identity } (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top = \Phi^\top (\Phi \Phi^\top + \lambda I)^{-1}
 \end{aligned} \tag{6}$$

3.1 Kernel trick

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

$$\begin{aligned}
 [\Phi \Phi^\top]_{i,j} &= \phi(x_i)^\top \phi(x_j) = \mathcal{K}(x_i, x_j) \\
 [\phi(x)^\top \Phi^\top]_j &= \phi(x)^\top \phi(x_j) = \mathcal{K}(x, x_j)
 \end{aligned} \tag{7}$$

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

3.2 relationship with Neural Tangent Kernel

Taylor Expansion of $f_{\mathbf{w}}(x)$ around w_0 :

$$f_{\mathbf{w}}(x) \equiv f(\mathbf{w}, x) \approx f(w_0, x) + \underbrace{\nabla_w f(w_0, x)}_{\phi(x)^\top} (w - w_0) + \dots \quad (8)$$

so, in theory, one may solve this using Kernel regression. However, question is **why still using neural networks?**
in here, we have not made any linkage to gradient descend yet.

3.3 relationship with gradient flow

This is a simplified version to Section[??]:

Gradient descend algorithm:

$$\begin{aligned} \theta_{t+1} &= \theta_t - \eta \nabla_{\theta} \mathcal{L}(\theta_{t+1}) \\ \implies \frac{\theta_{t+1} - \theta_t}{\eta} &= -\nabla_{\theta} \mathcal{L}(\theta_{t+1}) \\ \implies \lim_{\eta \rightarrow 0} \frac{\theta_{t+1} - \theta_t}{\eta} &= \frac{d\theta(t)}{dt} = -\nabla_{\theta} \mathcal{L}(\theta) \end{aligned} \quad (9)$$

let's substitute that into **least square** problem:

$$\begin{aligned} \mathcal{L}(\theta) &= \frac{1}{2} \|\tilde{y}(\theta) - y\|_2^2 \\ \implies \nabla_{\theta} \mathcal{L}(\theta) &= \nabla_{\theta} \tilde{y}(\theta) (\tilde{y}(\theta) - y) \\ \implies \frac{d\theta(t)}{dt} &= -\nabla_{\theta} \tilde{y}(\theta) (\tilde{y}(\theta) - y) \end{aligned} \quad (10)$$

so let's look at $\frac{d\tilde{y}(\theta_t)}{dt}$:

$$\begin{aligned} \frac{d\tilde{y}(\theta_t)}{dt} &= \frac{\partial \tilde{y}(\theta(t))}{\partial \theta(t)}^\top \frac{d\theta(t)}{dt} \\ &= \nabla_{\theta} \tilde{y}(\theta) \left(-\nabla_{\theta} \tilde{y}(\theta) (\tilde{y}(\theta) - y) \right) \\ &= -\underbrace{\nabla_{\theta} \tilde{y}(\theta)^\top \nabla_{\theta} \tilde{y}(\theta)}_{K(\theta)} (\tilde{y}(\theta) - y) \\ &\approx -K(\theta_0) (\tilde{y}(\theta) - y) \end{aligned} \quad (11)$$

4 First attempt for modeling Neural Network at initialization

4.1 neural network function

using parameters:

$$\theta \equiv \{W^L, b^L, \dots, W^1, b^1\} \quad (12)$$

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

$$\begin{aligned} f_\theta(X) &= W^L \phi^L(X) + b^L \\ &= W^L (\phi^{L-1}(X) W^{L-1} + b^{L-1}) + b^L \\ &\dots \\ &= W^L \dots (W^1 \phi^1(X) + b^1) + \dots + b^L \end{aligned} \quad (13)$$

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

$$\begin{aligned} \phi^L(X) &\equiv \phi^L(X | \theta^1, \dots, \theta^{L-1}) \\ &\equiv \phi^L(X | W^1, b^1, \dots, W^{L-1}, b^{L-1}) \end{aligned} \quad (14)$$

4.2 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:

$$p\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) = \int \mathcal{N}\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} f_\theta(X) \\ f_\theta(x^*) \end{bmatrix}, \sigma_\epsilon^2 I\right) \mathcal{N}(\theta | 0, \sigma_\theta^2 I) d\theta \quad (15)$$

The integral is **not** analytic!!

4.3 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:

$$p\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) \xrightarrow[N^1, \dots, N^L \rightarrow \infty]{d} ? \quad (16)$$

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

$$p\left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) = ? \quad (17)$$

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

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

4.3.1 look at the mean:

$$\begin{aligned}
& \mathbb{E} \left[\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right] \\
&= \int_y \int_{y^*} \begin{bmatrix} y \\ y^* \end{bmatrix} p \left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right) dy dy^* \\
&= \int_y \int_{y^*} \begin{bmatrix} y \\ y^* \end{bmatrix} \int_{\theta} p \left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \theta, \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right) p(\theta | \sigma_{\theta}^2) d\theta dy dy^* \\
&= \int_{\theta} \int_y \int_{y^*} \underbrace{\begin{bmatrix} y \\ y^* \end{bmatrix} \mathcal{N} \left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} f_{\theta}(X) \\ f_{\theta}(x^*) \end{bmatrix}, \sigma_{\epsilon}^2 I \right)}_{\mathbb{E} \left[\begin{bmatrix} y \\ y^* \end{bmatrix} \right] = \begin{bmatrix} f_{\theta}(X) \\ f_{\theta}(x^*) \end{bmatrix}} dy dy^* \mathcal{N}(\theta | 0, \sigma_{\theta}^2 I) d\theta \\
&= \int \begin{bmatrix} f_{\theta}(X) \\ f_{\theta}(x^*) \end{bmatrix} \mathcal{N}(\theta | 0, \sigma_{\theta}^2 I) d\theta \quad \text{to expand one layer :} \\
&= \int \begin{bmatrix} \phi^L(X) W^L + b^L \\ \phi^L(x^{*\top}) W^L + b^L \end{bmatrix} \mathcal{N}(W^L | 0, \sigma_w^2 I) \mathcal{N}(b^L | 0, \sigma_b^2 I) \mathcal{N}(\theta^1, \dots, \theta^{L-1} | 0, \sigma_{\theta}^2 I) d\theta^1, \dots, \theta^{L-1} dW^L db^L \\
&= \int \left[\underbrace{\phi^L(X) \int W^L \mathcal{N}(W^L | 0, \sigma_w^2 I) dW^L}_{=0} + \underbrace{\int b^L \mathcal{N}(b^L | 0, \sigma_b^2 I) db^L}_{=0} \right] \mathcal{N}(\theta^1, \dots, \theta^{L-1} | 0, \sigma_{\theta}^2 I) d\theta^1, \dots, \theta^{L-1} \\
&\quad \underbrace{\phi^L(x^{*\top}) \int W^L \mathcal{N}(W^L | 0, \sigma_w^2 I) dW^L}_{=0} + \underbrace{\int b^L \mathcal{N}(b^L | 0, \sigma_b^2 I) db^L}_{=0} \\
&= \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{aligned} \tag{18}$$

note we are not dealing with infinity at the moment

4.3.2 look at co-variance

$$\mathbb{E} \left[\begin{bmatrix} y \\ y^* \end{bmatrix} \begin{bmatrix} y^{\top} & y^* \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right] \tag{19}$$

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

$$\begin{aligned}
&= \int_y \int_{y^*} \int_{\theta} p \left(\begin{bmatrix} y \\ y^* \end{bmatrix} \begin{bmatrix} y^{\top} & y^* \end{bmatrix} \middle| \theta, \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right) p(\theta | \sigma_{\theta}^2) d\theta dy dy^* \\
&= \int_{\theta} \int_y \int_{y^*} \underbrace{\begin{bmatrix} y \\ y^* \end{bmatrix} \begin{bmatrix} y^{\top} & y^* \end{bmatrix} \mathcal{N} \left(\begin{bmatrix} y \\ y^* \end{bmatrix} \middle| \begin{bmatrix} f_{\theta}(X) \\ f_{\theta}(x^*) \end{bmatrix}, \sigma_{\epsilon}^2 I \right)}_{\mathbb{E}[Z^2] \quad Z \text{ is not mean-subtracted}} dy dy^* \mathcal{N}(\theta | 0, \sigma_{\theta}^2 I) d\theta
\end{aligned} \tag{20}$$

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

$$\begin{aligned}
&\text{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \implies \mathbb{E}[Z^2] = \text{Var}[Z] + (\mathbb{E}[Z])^2 \\
&= \int_{\theta} \underbrace{\sigma_{\epsilon}^2 I}_{\text{Var}[Z]} + \underbrace{\begin{bmatrix} f_{\theta}(X) \\ f_{\theta}(x^*) \end{bmatrix} \begin{bmatrix} f_{\theta}(X)^{\top} & f_{\theta}(x^*) \end{bmatrix}}_{(\mathbb{E}[Z])^2} \mathcal{N}(\theta | 0, \sigma_{\theta}^2 I) d\theta \\
&= \sigma_{\epsilon}^2 I + \int_{\theta} \left[\begin{pmatrix} \phi^L(X) W^L + b^L \\ \phi^L(x^{*\top}) W^L + b^L \end{pmatrix} \begin{pmatrix} W^{L\top} x^L(X)^{\top} + b^{L\top} \\ W^{L\top} \phi^L(x^*)^{\top} + b^{L\top} \end{pmatrix} \right. \\
&\quad \left. \begin{pmatrix} \phi^L(X) W^L + b^L \\ \phi^L(x^{*\top}) W^L + b^L \end{pmatrix} \begin{pmatrix} W^{L\top} \phi^L(x^{*\top})^{\top} + b^{L\top} \\ W^{L\top} \phi^L(x^*)^{\top} + b^{L\top} \end{pmatrix} \right] \mathcal{N}(\theta | 0, \sigma_{\theta}^2 I) d\theta
\end{aligned} \tag{21}$$

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

$$= \sigma_\epsilon^2 I + \int_\theta \begin{bmatrix} \phi^L(X)W^L W^{L\top} x^L(X)^\top + b^L b^{L\top} & \phi^L(X)W^L W^{L\top} x^L(x^{\star\top})^\top + b^L b^{L\top} \\ \phi^L(x^{\star\top})W^L W^{L\top} \phi^L(X)^\top + b^L b^{L\top} & \phi^L(x^{\star\top})W^L W^{L\top} \phi^L(x^{\star\top})^\top + b^L b^{L\top} \end{bmatrix} \mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) d\theta \quad (22)$$

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

$$\mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) d\theta = \mathcal{N}(\theta^L \mid 0, \sigma_\theta^2 I) \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} \quad (23)$$

$$= \int \begin{bmatrix} \sigma_w^2 \phi^L(X) x^L(X)^\top + \sigma_b^2 & \sigma_w^2 \phi^L(X) \phi^L(x^{\star\top})^\top + \sigma_b^2 \\ \sigma_w^2 \phi^L(x^{\star\top}) \phi^L(X)^\top + \sigma_b^2 & \sigma_w^2 \phi^L(x^{\star\top}) \phi^L(x^{\star\top})^\top + \sigma_b^2 \end{bmatrix} \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} \quad (24)$$

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

$$\begin{aligned} & \int \sigma_w^2 \phi^L(X) \phi^L(X)^\top \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} + \int \sigma_b^2 \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} \\ &= \sigma_w^2 \int \phi^L(X) \phi^L(X)^\top \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} + \sigma_b^2 \end{aligned} \quad (25)$$

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

$$= \sigma_b^2 + \sigma_w^2 \int \left[\phi(W^{L-1} \phi^{L-1}(X) + b^{L-1}) \phi(W^{L-1} \phi^{L-1}(X) + b^{L-1})^\top \right] \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} \quad (26)$$

it's difficult to see what is this distribution is, we need **a trick** to kept it going!

5 Single layer neural network

this section is to describe the paper [1]:

5.1 in summary

in summary, by definition of Gaussian process, a finite collection of “function of data” will be distributed according to Gaussian.

Central Limit Theorem makes them Gaussian distributed under an infinite-width case, despite the fact that the form of ϕ makes distribution calculation difficult.

5.2 notations

$$\begin{aligned} f_k(x) &= b_k + \sum_{j=1}^H v_{jk} h_j(x) \\ h_j(x) &= \tanh \left(a_j + \sum_{i=1}^I u_{ij} x_i \right) \end{aligned} \quad (27)$$

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

$$\begin{aligned} f_k(x) &= \underbrace{b_k}_{z_k^l} + \sum_{j=1}^{\overbrace{H}^{N_l}} \underbrace{v_{jk}}_{W_{k,j}^l} \times \underbrace{\tanh}_{\phi} \left(\underbrace{a_j}_{b_j^{l-1}} + \underbrace{u_{:,j}^\top}_{W_{:,j}^{l-1}{}^\top} x \right) \\ &\quad \underbrace{\hspace{10em}}_{z_j^{l-1}(x)} \\ \implies z_k^l(x) &= b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \times \phi(z_j^{l-1}(x)) \quad \text{modern notation} \\ f_k(\mathbf{x}) &= b_k^{(2)} + \sum_{j=1}^{N_2} W_{k,j}^{(2)} \times \phi \left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{j,i}^{(1)} x_i \right) \quad \text{modern notation} \end{aligned} \quad (28)$$

5.3 $p(z_k^l(x))$ for single input x

We need CLT for computing this probability.

5.3.1 Central Limit Theorem:

$$X^{(1)}, X^{(2)}, \dots, X^{(n)} \quad \text{are i.i.d samples} \quad (29)$$

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:

$$\begin{aligned} \sqrt{n}(\bar{X} - \mathbb{E}[X^{(1)}]) &\xrightarrow{d} \mathcal{N}(0, \sigma^2) \\ (\bar{X} - \mathbb{E}[X^{(1)}]) &\xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \\ \frac{1}{\sigma} \sqrt{n}(\bar{X} - \mathbb{E}[X^{(1)}]) &\xrightarrow{d} \mathcal{N}(0, 1) \end{aligned} \quad (30)$$

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

$$\begin{aligned}
& \sqrt{n}(\bar{X} - \mathbb{E}[X^{(1)}]) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \\
\Rightarrow & \sqrt{n}\sqrt{n}(\bar{X} - \mathbb{E}[X^{(1)}]) \xrightarrow{d} \mathcal{N}(0, \sqrt{n}^2 \sigma^2) = \mathcal{N}(0, n\sigma^2) \\
\Rightarrow & n(\bar{X} - \mathbb{E}[X^{(1)}]) \xrightarrow{d} \mathcal{N}(0, n\sigma^2) \\
\Rightarrow & \left(\sum_{i=1}^n X_i - n\mathbb{E}[X^{(1)}] \right) \xrightarrow{d} \mathcal{N}(0, n\sigma^2)
\end{aligned} \tag{31}$$

choose one of these conditions to suit the situation

5.3.2 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 $b_k^l + \text{sum of i.i.d elements using CLT notations:}$

$$z_k^l(x) = b_k^l + \underbrace{\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x))}_{\sum_{j=1}^{N_l} X_j}, \quad \text{note we are not taking average} \tag{32}$$

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

$$X_j \equiv W_{k,j}^l \phi(z_j^{l-1}(x)) \tag{33}$$

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

Expectation

$$\begin{aligned}
\mathbb{E}[W_{k,j}^l \phi(z_j^{l-1}(x))] &= \mathbb{E}[W_{k,j}^l] \mathbb{E}[\phi(z_j^{l-1}(x))] \quad \text{since } W_{k,j}^l \text{ and } \phi(z_j^{l-1}(x)) \text{ are independent} \\
&\quad \text{as } z_j^{l-1}(x) \text{ depends on } (W^{l-1}, b^{l-1}) \\
&= 0 \times \mathbb{E}[\phi(z_j^{l-1}(x))] \quad \text{because we choose } W_{k,j}^l \sim \mathcal{N}(0, \sigma_w) \\
&= 0
\end{aligned} \tag{34}$$

Variance

$$\begin{aligned}
& \text{Var}[W_{k,j}^l \phi(z_j^{l-1}(x))] \\
&= \mathbb{E}\left[\left(W_{k,j}^l \phi(z_j^{l-1}(x))\right)^2\right] \\
&= \mathbb{E}[(W_{k,j}^l)^2] \mathbb{E}[\phi(z_j^{l-1}(x))^2] \quad \text{since } W_{k,j}^l \text{ and } \phi(z_j^{l-1}(x)) \text{ are independent} \\
&= \sigma_w^2 \underbrace{\mathbb{E}[\phi(z_j^{l-1}(x))^2]}_{\text{bounded}} \Rightarrow \text{Var}[W_{k,j}^l \phi(z_j^{l-1}(x))] \text{ to be bounded} \\
&= \sigma_w^2 \mathbb{E}[\phi(z_j^{l-1}(x))^2]
\end{aligned} \tag{35}$$

we leave in this form, as

$$\mathbb{E}[\phi(z_j^{l-1}(x))^2] \equiv \mathbb{E}_{W^{l-1}, \dots, b^{l-1}, \dots} [\phi(z_j^{l-1}(x))^2] \quad (36)$$

5.3.4 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.

$$\begin{aligned} & \text{using } \left(\sum_{i=1}^n X_i - n\mathbb{E}[X_1] \right) \xrightarrow{d} \mathcal{N}(0, n\sigma^2) \\ \Rightarrow & \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) - 0 \right) \sim \mathcal{N}\left(0, N_l \sigma_w^2 \mathbb{E}[\phi(z_1^{l-1}(x))^2]\right) \quad N_l \rightarrow \infty \end{aligned} \quad (37)$$

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

$$\sigma(W_{k,j}^l) = \sigma_w = \frac{1}{\sqrt{N_l}} \quad \Rightarrow \quad \sigma_w^2 = \frac{1}{N_l} \quad (38)$$

the above is the key, implication is:

$$\begin{aligned} \Rightarrow & \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) - 0 \right) \sim \mathcal{N}\left(0, N_l \frac{1}{N_l} \mathbb{E}[\phi(z_1^{l-1}(x))^2]\right) \\ & = \mathcal{N}\left(0, \underbrace{\mathbb{E}[\phi(z_1^{l-1}(x))^2]}_{\text{bounded}}\right) \end{aligned} \quad (39)$$

finally adding the bias b_k^l :

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

$$\begin{aligned} X & \sim \mathcal{N}(\mu_X, \sigma_X^2) \\ Y & \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \\ Z & = X + Y \quad Z = X + Y \\ \Rightarrow & Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \end{aligned} \quad (40)$$

Therefore:

$$\left(z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) \right) \xrightarrow{d} \mathcal{N}\left(0, \underbrace{\sigma_b^2}_{\sigma_X^2} + \underbrace{\mathbb{E}[\phi(z_1^{l-1}(x))^2]}_{\sigma_Y^2}\right) \quad \text{as } N_l \rightarrow \infty \quad (41)$$

appreciate the recursion here

5.4 given two inputs \mathbf{x}, \mathbf{x}' : compute $\text{Cov}[z_k^l(\mathbf{x}), z_k^l(\mathbf{x}')]]$

5.4.1 Independence after only one layer

$$f_k(\mathbf{x}) = b_k^{(2)} + \sum_{j=1}^{N_2} W_{k,j}^{(2)} \times \phi\left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{j,i}^{(1)} x_i\right) \quad (42)$$

Therefore, individual terms of the outer sum:

$$\text{for } j \neq j' : \begin{cases} W_{k,j}^{(2)} \times \phi\left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{j,i}^{(1)} x_i\right) & \text{random variables } W_{k,j}^{(2)}, W_{j,1}^{(1)}, \dots, W_{j,N_1}^{(1)} \\ W_{k,j'}^{(2)} \times \phi\left(b_{j'}^{(1)} + \sum_{i=1}^{N_1} W_{j',i}^{(1)} x_i\right) & \text{random variables } W_{k,j'}^{(2)}, W_{j',1}^{(1)}, \dots, W_{j',N_1}^{(1)} \end{cases} \quad (43)$$

They involve a different set of random variables. This is **not** the case if one performs another layer

5.4.2 changing from product of two sums into one sum

in general, if we have X_j to be independent of $X_{j'}$ when $j \neq j'$ and dependent when $j = j'$ and for simplicity we let $\mathbb{E}[X_j^{(p)}] = 0 \quad \forall j, p$:

$$\begin{aligned} & \text{Cov}\left[\sum_{j=1}^N X_j^{(p)}, \sum_{j'=1}^N X_{j'}^{(q)}\right] \\ &= \mathbb{E}\left[\sum_{j=1}^N X_j^{(p)} \times \sum_{j'=1}^N X_{j'}^{(q)}\right] \\ &= \mathbb{E}\left[\sum_{j=1}^N X_j^{(p)} X_{j'}^{(q)}\right] \end{aligned} \quad (44)$$

then, by CLT (as we see later in the multi-dimensional CLT)

$$\begin{aligned} & \text{Cov}\left[\sum_{j=1}^N X_j^{(p)}, \sum_{j=1}^N X_j^{(q)}\right] = N \text{Cov}[X_1^{(p)}, X_1^{(q)}] \quad N \rightarrow \infty \\ \implies & \mathbb{E}\left[\sum_{j=1}^N X_j^{(p)} X_j^{(q)}\right] = N \text{Cov}[X_1^{(p)}, X_1^{(q)}] \quad N \rightarrow \infty \quad \text{as well!} \end{aligned} \quad (45)$$

Therefore:

$$\begin{aligned}
& \mathbf{Cov} \left[\sum_{j=1}^{N_2} W_{k,j}^{(2)} \times \phi \left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{j,i}^{(1)} x_i \right), \sum_{j'=1}^{N_2} W_{k,j'}^{(2)} \times \phi \left(b_{j'}^{(1)} + \sum_{i=1}^{N_1} W_{j',i}^{(1)} x'_i \right) \right] \\
&= \mathbb{E} \left[\sum_{j=1}^{N_2} W_{k,j}^{(2)} \times \phi \left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{j,i}^{(1)} x_i \right) \times \sum_{j'=1}^{N_2} W_{k,j'}^{(2)} \times \phi \left(b_{j'}^{(1)} + \sum_{i=1}^{N_1} W_{j',i}^{(1)} x'_i \right) \right] \\
&= \mathbb{E} \left[\sum_{j=1}^{N_2} (W_{k,j}^{(2)})^2 \phi \left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{j,i}^{(1)} x_i \right) \phi \left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{j,i}^{(1)} x'_i \right) \right] \\
&= \mathbb{E} \left[(W_{k,j}^{(2)})^2 \right] \mathbb{E} \left[\sum_{j=1}^{N_2} \phi \left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{j,i}^{(1)} x_i \right) \phi \left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{j,i}^{(1)} x'_i \right) \right] \\
&= \mathbb{E} \left[(W_{k,j}^{(2)})^2 \right] N_2 \mathbb{E} \left[\phi \left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{1,i}^{(1)} x_i \right) \phi \left(b_j^{(1)} + \sum_{i=1}^{N_1} W_{1,i}^{(1)} x'_i \right) \right] \quad N_2 \rightarrow \infty \quad \text{using Eq.(45)} \\
& \tag{46}
\end{aligned}$$

5.4.3 using Multidimensional CLT

Multidimensional CLT only works if \mathbf{X}_i is independent of $\mathbf{X}_j \forall i \neq j$, in the case of:

let $\mathbf{X}_i \in \mathcal{R}^d$:

$$\begin{aligned}
\sum_{i=1}^n \mathbf{X}_i &= \underbrace{\begin{bmatrix} X_1^{(1)} \\ \vdots \\ X_1^{(p)} \\ \vdots \\ X_1^{(q)} \\ \vdots \\ X_1^{(k)} \end{bmatrix}}_{\mathbf{X}_1} + \underbrace{\begin{bmatrix} X_2^{(1)} \\ \vdots \\ X_2^{(p)} \\ \vdots \\ X_2^{(q)} \\ \vdots \\ X_2^{(k)} \end{bmatrix}}_{\mathbf{X}_2} + \cdots + \underbrace{\begin{bmatrix} X_n^{(1)} \\ \vdots \\ X_n^{(p)} \\ \vdots \\ X_n^{(q)} \\ \vdots \\ X_n^{(k)} \end{bmatrix}}_{\mathbf{X}_n} = \underbrace{\begin{bmatrix} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \sum_{i=1}^n X_i^{(k)} \end{bmatrix}}_{\sum_{i=1}^n \mathbf{X}_i} \\
& \tag{47} \\
\Rightarrow \bar{\mathbf{X}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(k)} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{X}}^{(1)} \\ \vdots \\ \bar{\mathbf{X}}^{(p)} \\ \vdots \\ \bar{\mathbf{X}}^{(q)} \\ \vdots \\ \bar{\mathbf{X}}^{(k)} \end{bmatrix}
\end{aligned}$$

1. Sample mean version

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^n [\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i]] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{X}_i - \mathbb{E}[\mathbf{X}_1]) \quad \text{since } p(X_i) = p(X_1) \\
&= \frac{\sqrt{n}}{\sqrt{n}} \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n \mathbf{X}_i \right) - \frac{n}{\sqrt{n}} \mathbb{E}[\mathbf{X}_1] \\
&= \sqrt{n} (\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}_1])
\end{aligned} \tag{48}$$

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^n [\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i]] = \sqrt{n} (\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}_1]) \xrightarrow{d} \mathcal{N}_d(0, \mathbf{Cov}(\mathbf{X}_1)) \\
\Rightarrow & \sqrt{n} \mathbb{E} \left[\underbrace{(\bar{\mathbf{X}}^{(p)} - \mathbb{E}[\bar{\mathbf{X}}_1^{(p)}])}_{\text{scalar}} \underbrace{(\bar{\mathbf{X}}^{(q)} - \mathbb{E}[\bar{\mathbf{X}}_1^{(q)}])}_{\text{scalar}} \right] = \mathbf{Cov}(\mathbf{X}_1)_{(p),(q)}
\end{aligned} \tag{49}$$

for every elements $(p, q) \in \{1, \dots, k\}$:

2. **Sample sum version:**

$$\begin{aligned}
& \left(\left[\sum_i^n \mathbf{X}_i \right] - n \mathbb{E}[\mathbf{X}_1] \right) \xrightarrow{d} \mathcal{N}_k(0, n \boldsymbol{\Sigma}) \\
\Rightarrow & \mathbb{E} \left[\left(\left[\sum_i^n \mathbf{X}_i \right]^{(p)} - n \mathbb{E}[\mathbf{X}_1]^{(p)} \right) \left(\left[\sum_i^n \mathbf{X}_i \right]^{(q)} - n \mathbb{E}[\mathbf{X}_1]^{(q)} \right) \right] = n \boldsymbol{\Sigma}_{(p),(q)} \\
& \Rightarrow \mathbb{E} \left[\left(n \bar{\mathbf{X}}^{(p)} - n \mathbb{E}[X_1^{(p)}] \right) \left(n \bar{\mathbf{X}}^{(q)} - n \mathbb{E}[X_1^{(q)}] \right) \right] = n \boldsymbol{\Sigma}_{(p),(q)} \\
\Rightarrow & \mathbb{E} \left[\left(\left[\sum_i^n \mathbf{X}_i \right]^{(p)} - n \mathbb{E}[X_1^{(p)}] \right) \left(\left[\sum_i^n \mathbf{X}_i \right]^{(q)} - n \mathbb{E}[X_1^{(q)}] \right) \right] = n \boldsymbol{\Sigma}_{(p),(q)}
\end{aligned} \tag{50}$$

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

5.4.4 Put into multidimensional CLT structure:

now, let's look at k^{th} dimension of z^l , i.e., z_k^l , and to see in this dimension, how correlation between pair of data input x and x' is. note that what happen to k^{th} dimension, applies to the rest

$$\begin{aligned}
& \begin{bmatrix} \vdots \\ W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x})) \\ \vdots \\ W_{k,1}^l \phi(z_j^{l-1}(\mathbf{x}')) \\ \vdots \end{bmatrix} + \cdots + \begin{bmatrix} \vdots \\ W_{k,N_l}^l \phi(z_j^{l-1}(\mathbf{x})) \\ \vdots \\ W_{k,N_l}^l \phi(z_j^{l-1}(\mathbf{x}')) \\ \vdots \end{bmatrix} = \underbrace{\begin{bmatrix} \vdots \\ \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x})) \\ \vdots \\ \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x}')) \\ \vdots \end{bmatrix}}_{\begin{bmatrix} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \sum_{i=1}^n X_i^{(k)} \end{bmatrix}} = \underbrace{\begin{bmatrix} \vdots \\ z_k^l(\mathbf{x}) \\ \vdots \\ z_k^l(\mathbf{x}') \\ \vdots \end{bmatrix}}_{\sum_{i=1}^n \mathbf{X}_i}
\end{aligned} \tag{51}$$

compare with the standard notation of Multi-dimensional CLT, and use “sample sum version” of CLT, Eq.[47], and remember $z_k^l(\mathbf{x}) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x}))$, to simplify derivation, let's deliberately not looking at b_k^l for now

$$\begin{aligned}
\sum_i^n X_i^{(p)} &= \left[\sum_i^n X_i \right]^{(p)} \longrightarrow \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x})) \\
\sum_i^n X_i^{(q)} &= \left[\sum_i^n X_i \right]^{(q)} \longrightarrow \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x}')) \\
X_1^{(p)} &\longrightarrow W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x})) \quad \text{be the single term in the sum} \\
X_1^{(q)} &\longrightarrow W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x}'))
\end{aligned} \tag{52}$$

using above identities in Eq.[52]

$$\begin{aligned}
& \mathbb{E} \left[\left(\left[\sum_i^n \mathbf{X}_i \right]^{(p)} - \mathbf{n} \mathbb{E}[X_1^{(p)}] \right) \left(\left[\sum_i^n \mathbf{X}_i \right]^{(q)} - \mathbf{n} \mathbb{E}[X_1^{(q)}] \right) \right] = \mathbf{n} \mathbf{Cov}(\mathbf{X}_1)_{(p),(q)} \\
\Rightarrow & \mathbb{E} \left[\left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x})) - \underbrace{N_l \mathbb{E}[W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}))]}_{=0} \right) \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x}')) - \underbrace{N_l \mathbb{E}[W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}'))]}_{=0} \right) \right] \\
&= N_l \mathbf{Cov} \left(W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x})), W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}')) \right) \\
&= N_l \mathbb{E} \left[W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x})) \times W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}')) \right]
\end{aligned} \tag{53}$$

look at $z_k^l(\mathbf{x})$ with b_k^l too:

$$\begin{aligned}
\mathbf{Cov}(z_k^l(x), z_k^l(x')) &= \sigma_b^2 + \mathbb{E} \left[\left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) \right) \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x')) \right) \right] \\
&= \sigma_b^2 + \textcolor{red}{N_l} \text{Cov}(W_{k,1}^l \phi(z_1^{l-1}(x)), W_{k,1}^l \phi(z_1^{l-1}(x'))) \quad \text{use CLT result above from Eq.[54]} \\
&= \sigma_b^2 + \textcolor{red}{N_l} \sigma_w^2 \text{Cov}(\phi(z_1^{l-1}(x)), \phi(z_1^{l-1}(x'))) \\
&= \sigma_b^2 + \textcolor{red}{N_l} \frac{1}{\textcolor{blue}{N_l}} \text{Cov}(\phi(z_1^{l-1}(x)), \phi(z_1^{l-1}(x'))) \\
&= \sigma_b^2 + \text{Cov}(\phi(z_1^{l-1}(x)), \phi(z_1^{l-1}(x'))) \\
&= \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(x)) \times \phi(z_1^{l-1}(x'))]
\end{aligned} \tag{54}$$

there are many notes this:

1. **note 1** under usual \mathcal{GP} , $f(\cdot)$ has just a one-dimension output, however, in here, function $z^l(\cdot)$ has N_{l+1} outputs, i.e., a vector function. so the “entire” $\mathbf{Cov}(z^l, z^l)$ is of size:

$$\underbrace{N_{l+1}}_{\forall k} \underbrace{|\mathcal{X}|}_{\forall x} \times \underbrace{N_{l+1}}_{\forall k'} \underbrace{|\mathcal{X}|}_{\forall x'} \tag{55}$$

exactly how one may arrange this “gigantic” matrix, either N_{l+1} sub-blocks of $\mathbf{Cov}(x, x')$, or $|\mathcal{X}|$ blocks of $\mathbf{Cov}(k, k')$ has the same effect

2. **note 2**: 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:

$$\begin{aligned}
z_k^l(x) &= b_k + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) \\
z_{k'}^l(x) &= b_{k'} + \sum_{j=1}^{N_l} W_{k',j}^l \phi(z_j^{l-1}(x)) \\
\implies \mathbb{E} \left[W_{k,j}^l \phi(z_j^{l-1}(x)) \times W_{k',j'}^l \phi(z_{j'}^{l-1}(x)) \right] &= 0 \quad \forall \{k, k', j, j'\}
\end{aligned} \tag{56}$$

note 3: in literature, it is written:

$$\begin{aligned}
\mathbb{E}[z_k^l(\textcolor{red}{x}) z_k^l(\textcolor{blue}{x}')] &= \sigma_b^2 + \sigma_w^2 \mathbb{E} \left[\sum_{j=1}^{N_l} \phi(z_j^{l-1}(\textcolor{red}{x})) \phi(z_j^{l-1}(\textcolor{blue}{x}')) \right] \\
\text{instead of } &= \sigma_b^2 + \mathbb{E} \left[\left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\textcolor{red}{x})) \right) \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\textcolor{blue}{x}')) \right) \right]
\end{aligned} \tag{57}$$

see section[5.4.2] for detail

5.4.5 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:

$$\begin{aligned}\mathbb{E}[f(x)] &= 0 \\ \mathbb{E}[f(x, x')] &= \mathbf{Cov}(x, x') = \mathbf{\Sigma}_{x, x'} \\ \implies f &\sim \mathcal{GP}(0, \mathbf{\Sigma})\end{aligned}\tag{58}$$

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

$$\begin{aligned}\text{Cov}[z_k^l(x), z_k^l(x')] &= \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(x)) \times \phi(z_1^{l-1}(x'))] \quad \text{as } N_l \rightarrow \infty \\ z_k^l(x) &\xrightarrow{d} \mathcal{N}\left(0, \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(x))^2]\right) \quad \text{as } N_l \rightarrow \infty\end{aligned}\tag{59}$$

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

$$\begin{aligned}\implies z_k^l(\mathcal{X}) &\sim \mathcal{GP}(0, \mathbf{\Sigma}^l) \\ \text{where specific co-variance } \mathbf{\Sigma}_{x, x'}^l &= \sigma_b^2 + \mathbb{E}[\phi(z_1^{l-1}(x)) \times \phi(z_1^{l-1}(x'))]\end{aligned}\tag{60}$$

5.5 looking at GP systematically

First let's change for the rest of the tutorial:

$$\mathbf{\Sigma}^l \rightarrow K^l\tag{61}$$

$K^l(x, x')$ 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:

$$\sigma(W_{k,j}^l) = \frac{\sigma_w}{\sqrt{N_l}}\tag{62}$$

we look at all GP kernel K^l relate to K^{l-1} :

$$\begin{aligned}K^l(x, x') &= \mathbb{E}[z_k^l(x) z_k^l(x') | z_1^{l-1}] \\ &= \sigma_b^2 + \sigma_w^2 \mathbb{E}[\phi(z_1^{l-1}(x)) \times \phi(z_1^{l-1}(x'))] \quad \text{apply CLT } N_l \rightarrow \infty \\ &= \sigma_b^2 + \sigma_w^2 \underbrace{\mathbb{E}_{z_1^{l-1}(\mathcal{X}) \sim \mathcal{GP}(0, K^{l-1})} [\phi(z_1^{l-1}(x)) \phi(z_1^{l-1}(x'))]}_{F_\phi(K^{l-1})}\end{aligned}\tag{63}$$

since $\mathbb{E}[\phi(z)] = \mathbb{E}_{z \sim p(z)}[\phi(z)]$ and $p(z_1^{l-1}(\mathcal{X})) = \mathcal{GP}(0, K^{l-1})$ just as Eq.[60], and $\phi(z_1^{l-1}(x))$ is function on a specific point x , keep on going:

$$\begin{aligned}&= \sigma_b^2 + \sigma_w^2 \underbrace{F_\phi(K^{l-1}(x, x'), K^{l-1}(x, x), K^{l-1}(x', x'))}_{F_\phi(K^{l-1})} \\ &= \sigma_b^2 + \sigma_w^2 F_\phi(K^{l-1}(x, x'))\end{aligned}\tag{64}$$

5.5.1 using properties of point Marginals of Gaussian Process:

$$\begin{aligned}
F_\phi(K^{l-1}(x, x')) &= \mathbb{E}_{z_j^{l-1} \sim \mathcal{GP}(0, K^{l-1})} \left[\phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x')) \right] \\
&= \mathbb{E}_{\underbrace{(z_j^{l-1}(x), z_j^{l-1}(x'))}_{\text{2 points on function } z_j^{l-1}} \sim \underbrace{\mathcal{N}(0, K^{l-1}(x, x'))}_{\text{2D Gaussian}}} \left[\phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x')) \right] \quad (65)
\end{aligned}$$

$$\begin{bmatrix} z_j^{l-1}(x) \\ z_j^{l-1}(x') \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K^{l-1}(x, x) & K^{l-1}(x, x') \\ K^{l-1}(x, x') & K^{l-1}(x', x') \end{bmatrix}\right) \quad (66)$$

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

$$= F_\phi(K^{l-1}(x, x'), K^{l-1}(x, x), K^{l-1}(x', x')) \quad (67)$$

5.6 in summary

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

$$K^l(x, x') = \sigma_b^2 + \sigma_w^2 \mathbb{E}_{(z_j^{l-1}(x), z_j^{l-1}(x')) \sim \mathcal{N}(0, K^{l-1}(x, x'))} \left[\phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x')) \right] \quad (68)$$

1. some confusion on the dimension of $K^l(x, x')$: the expression above is a scalar
2. however, since each of $z^l(x) \forall x \in \mathcal{X}$ has N_{l+1} values; Therefore, co-variance matrix corresponding to $\{z_1^l(x), \dots, z_{N_{l+1}}^l(x)\}_{x \in \mathcal{X}}$ should be made up of $(N_l \times |\mathcal{X}|) \times (N_l \times |\mathcal{X}|)$ elements
3. the interesting thing is we never need to sample other $z_{j>1}^l(x)$, as CLT made sure only one $z_1^l(x)$ needs to be sampled, at next layer $l+1$
4. we will see the same recursion also applies in NTK, except $\phi \rightarrow \phi'$

6 Expand GP across all layers

this section describe [2]

From the previous section, we know that that “independence” property on $W_{k,j}^{(2)} \times \phi(b_j^{(1)} + \sum_{i=1}^{N_1} W_{j,i}^{(1)} x_i)$ and $W_{k,j'}^{(2)} \times \phi(b_{j'}^{(1)} + \sum_{i=1}^{N_1} W_{j',i}^{(1)} x_i)$ is lost if Neural Network involve more than two layers starting from data \mathbf{x} . However, we need that “independence” property to apply CLT.

However, after bringing each of the layers to infinity, we can make it almost as if it’s just a data. Hence the “independence” property.

6.1 Overall objective

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

$$\begin{aligned} p(z^L|x) &= \int p(z^L, K^0, K^1, \dots, K^L|x) dK^{0,\dots,L} \\ &= \int p(z^L|K^L) \left(\prod_{l=1}^L p(K^l|K^{l-1}) \right) p(K^0|x) dK^{0,\dots,L} \end{aligned} \quad (69)$$

6.2 $p(z^L|K^L)$: conditions on $K^l \equiv \{\phi(z^{l-1})(x))\phi(z^{l-1})(x')\}_{p,q}$

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

$$K^l(x, x') = \mathbb{E}[z_k^l(x) z_k^l(x') | z^{l-1}] \quad (70)$$

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:

$$z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) \quad (71)$$

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

Conditional Marginal

$$\begin{aligned} z_k^l(x) | \{\phi(z_j^{l-1}(x))\}_j^{N_l} &= b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \underbrace{\phi(z_j^{l-1}(x))}_{\text{constant}} \\ \Rightarrow z_k^l(x) | \{\phi(z_j^{l-1}(x))\}_j^{N_l} &\sim \mathcal{N}\left(0, \sigma_b^2 + \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2 \text{Var}[W_{k,j}^l]\right) \\ &= \mathcal{N}\left(0, \sigma_b^2 + \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2\right) \end{aligned} \quad (72)$$

using property of weighted sum of Gaussian:

$$\begin{aligned} X_i &\sim \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \dots, \\ \Rightarrow \sum_{i=1}^n a_i X_i &\sim \mathcal{N}\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \text{Var}[X_i]\right) \end{aligned} \quad (73)$$

Conditional Co-variance

$$\begin{aligned}
& \text{Cov} \left[z_k^l(x), z_k^l(x') \mid \left\{ \phi(z_j^{l-1}(x)), \phi(z_j^{l-1}(x')) \right\}_{j=1}^{N_l} \right] \\
&= \mathbb{E} \left[z_k^l(x) z_k^l(x') \mid \left\{ \phi(z_j^{l-1}(x)), \phi(z_j^{l-1}(x')) \right\}_{j=1}^{N_l} \right] \\
&= \sigma_b^2 + \mathbb{E}_{W_{k,j}^l} \left[\underbrace{\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x'))}_{\text{constant, used as condition}} \right] \\
&= \sigma_b^2 + \sum_{j=1}^{N_l} \text{Var}[W_{k,j}^l] \phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x')) \\
&= \sigma_b^2 + \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x'))
\end{aligned} \tag{74}$$

not using property of weighted sum of Gaussian:
Combine all together

$$\begin{aligned}
& \text{Cov} \left[z_k^l(x), z_k^l(x') \mid \left\{ \phi(z_j^{l-1}(x)), \phi(z_j^{l-1}(x')) \right\}_{j=1}^{N_l} \right] = \sigma_b^2 + \sigma_w^2 \frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x')) \\
& z_k^l(x) \mid \left\{ \phi(z_j^{l-1}(x)) \right\}_j^{N_l} \sim \mathcal{N} \left(0, \sigma_b^2 + \sigma_w^2 \frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2 \right) \\
& \Rightarrow \begin{bmatrix} z^l(x) \\ z^l(x') \end{bmatrix} \mid \left\{ \phi(z_j^{l-1}(x)), \phi(z_j^{l-1}(x')) \right\}_j^{N_l} \sim \mathcal{N} \left(\mathbf{0}, G \left(\begin{bmatrix} K^l(x, x) & K^l(x, x') \\ K^l(x, x') & K^l(x', x') \end{bmatrix} \right) \right)
\end{aligned} \tag{75}$$

in GP paradigm:

$$z^l(x) \mid K^l \sim \mathcal{GP}(z^l; \mathbf{0}, G(K^l)) \tag{76}$$

where

$$\begin{aligned}
K^l(x, x') &= \frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x')) \\
G(K^l(x, x')) &= \sigma_b^2 + \sigma_w^2 K^l(x, x')
\end{aligned} \tag{77}$$

Conveniently, we use K^l as a short-notation collection of $\phi(z_j^{l-1}(x)), \phi(z_j^{l-1}(x')) \quad \forall x, x', j$
also taking care of the layer one, which is just input x :

$$K_{p,q}^l \equiv K^l(x, x') = \begin{cases} \frac{1}{d_{\text{in}}} \sum_{j=1}^{d_{\text{in}}} x_j x'_j = \frac{1}{d_{\text{in}}} x^\top x' & l = 0 \\ \frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x')) & l > 0 \end{cases} \tag{78}$$

to reflect:

$$\text{Cov}(z_k^l, z_{k'}^l) = 0 \quad \forall k, k' \in \{1, \dots, N_{l+1}\} \tag{79}$$

note that

$$K^0(x, x') = \frac{1}{d_{\text{in}}} x^\top x' \quad \text{appears again in NTK} \quad (80)$$

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

$$\begin{aligned} \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 & & \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 & & \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)) \end{aligned} \quad (81)$$

6.3 $p(K^l | K^{l-1})$

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

$$\begin{aligned}
p(K^l | K^{l-1}) &= \int_{z^{l-1}} p(K^l | z^{l-1}) p(z^{l-1} | K^{l-1}) \\
&= \int_{z^{l-1}} p(K^l | z^{l-1}) \mathcal{GP}(z^{l-1}; 0, G(K^{l-1}))
\end{aligned} \tag{82}$$

using GP property, and just look at two points x, x' :

$$\begin{aligned}
p(K_{p,q}^l | K_{p,q}^{l-1}) &= \int_{z^{l-1}(x), z^{l-1}(x')} p\left(\frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^l(x)) \phi(z_j^l(x'))\right) \\
&\quad \mathcal{N}\left(\begin{bmatrix} z^{l-1}(x) \\ z^{l-1}(x') \end{bmatrix}; 0, G\left(\begin{bmatrix} K^{l-1}(x, x) & K^{l-1}(x, x') \\ K^{l-1}(x', x) & K^{l-1}(x', x') \end{bmatrix}\right)\right)
\end{aligned} \tag{83}$$

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

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

$$\begin{aligned}
\underbrace{K_{p,q}^l}_{\bar{X}} &= \frac{1}{N_l} \sum_{j=1}^{N_l} \underbrace{\phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x'))}_{X_j \equiv K_{p,q}^{l,j}} \\
\Rightarrow p(K_{p,q}^{l,1} | K_{p,q}^{l-1}) &= \int_{z^{l-1}(x), z^{l-1}(x')} p(\phi(z_j^l(x)) \phi(z_j^l(x'))) \\
&\quad \mathcal{N}\left(\begin{bmatrix} z^{l-1}(x) \\ z^{l-1}(x') \end{bmatrix}; 0, G\left(\begin{bmatrix} K^{l-1}(x, x) & K^{l-1}(x, x') \\ K^{l-1}(x', x) & K^{l-1}(x', x') \end{bmatrix}\right)\right) \\
&= (F \circ G)(K_{p,q}^{l-1})
\end{aligned} \tag{84}$$

using CLT, pick the most appropriate definition:

$$(\bar{X} - \mathbb{E}[X_1]) \xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}[X_1]}{n}\right) \tag{85}$$

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

$$\begin{aligned}
&(\bar{X} - \mathbb{E}[X_1]) \xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}[X_1]}{n}\right) \\
\Rightarrow (K_{p,q}^l - \mathbb{E}[K_{p,q}^{l,1}]) &\xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}[K_{p,q}^{l,1}]}{N_l}\right) \\
\Rightarrow (K_{p,q}^l - (F \circ G)(K_{p,q}^{l-1})) &\xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}[K_{p,q}^{l,1}]}{N_l}\right) \\
\Rightarrow (K_{p,q}^l | K_{p,q}^{l-1}) &\xrightarrow{d} \mathcal{N}\left((F \circ G)(K^{l-1}), \frac{\text{Var}[K_{p,q}^{l,1}]}{N_l}\right) \\
\Rightarrow \lim_{N_l \rightarrow \infty} p(K^l | K^{l-1}) &= \delta(K^l - (F \circ G)(K^{l-1})) \quad \text{entire matrix}
\end{aligned} \tag{86}$$

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

6.4 putting in the overall objective function

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

$$\begin{aligned}
p(z^L|x) &= \int p(z^L, K^0, K^1, \dots, K^L|x) \, dK^{0,\dots,L} \\
&= \int p(z^L|K^L) \left(\prod_{l=1}^L p(K^l|K^{l-1}) \right) p(K^0|x) \, dK^{0,\dots,L} \\
\lim_{N_L \rightarrow \infty, \dots, N_1 \rightarrow \infty} p(z^L|x) &= \int p(z^L|K^L) \left(\prod_{l=1}^L \delta(K^l - (F \circ G)(K^{l-1})) \right) p(K^0|x) \, dK^{0,\dots,L} \\
&= \int \mathcal{GP}(z^L; 0, G(K^L)) \underbrace{\left(\prod_{l=1}^L \delta(K^l - (F \circ G)(K^{l-1})) \right) \delta\left(K^0 - \frac{1}{d_{\text{in}}} x^\top x\right)}_{= \begin{cases} 1 & \text{if } K^L = (F \circ G)(K^{L-1}) \\ & = (F \circ G)^2(K^{L-2}) \dots \\ & = (F \circ G)^L\left(\frac{1}{d_{\text{in}}} x^\top x\right) \\ 0 & \text{otherwise} \end{cases}} \, dK^{0,\dots,L} \\
&= \mathcal{GP}\left(z^L; 0, G \circ (F \circ G)^L\left(\frac{1}{d_{\text{in}}} x^\top x\right)\right)
\end{aligned} \tag{87}$$

7 NTK at initialization

this section describe [3]

7.1 expression

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

$$\begin{aligned}
 & \begin{bmatrix} \frac{1}{\sqrt{N_l}} W_{1,1}^l \phi(z_1^{l-1}(x)) + \sigma_b b_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} W_{k,1}^l \phi(z_1^{l-1}(x)) + \sigma_b b_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} W_{N_{l+1},1}^l \phi(z_j^{l-1}(x)) + \sigma_b b_{N_{l+1}} \end{bmatrix} + \dots + \begin{bmatrix} \frac{1}{\sqrt{N_l}} W_{1,N_l}^l \phi(z_1^{l-1}(x)) + \sigma_b b_1^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} W_{k,N_l}^l \phi(z_1^{l-1}(x)) + \sigma_b b_k^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} W_{N_{l+1},N_l}^l \phi(z_j^{l-1}(x)) + \sigma_b b_{N_{l+1}}^l \end{bmatrix} \\
 &= \begin{bmatrix} \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{1,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_1^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{N_{l+1},j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_{N_{l+1}}^l \end{bmatrix} = \begin{bmatrix} z_1^l(x) \\ \vdots \\ z_k^l(x) \\ \vdots \\ z_{N_{l+1}}^l(x) \end{bmatrix} \quad (88)
 \end{aligned}$$

7.2 re-parameterized formulation

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

$$\begin{aligned}
 \text{NNGP} \quad z_k^l(x) &= \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l & W_{k,j}^l &\sim \mathcal{N}\left(0, \frac{1}{\sqrt{N_l}}\right) \\
 \text{this section} \quad z_k^l(x) &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l & W_{k,j}^l &\sim \mathcal{N}(0, 1)
 \end{aligned} \quad (89)$$

they are the same, but using re-parameterization!

7.3 Prove by Induction

7.3.1 how does prove by induction work?

by induction works by:

1. proving value at $l = 1$ (or at some initial condition)
2. Then show relationship between l and $l - 1$ in general
3. Finally, it shows what value is at an arbitrary index L

7.3.2 For NTK

we need to show by induction:

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

$$\Theta_{k,k'}^1(x, x') = \underbrace{\left(\frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2 \right)}_{K^1} \delta_{k,k'} \quad (90)$$

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

2. then by assuming:

$$\Theta_{k,k'}^{l-1}(x, x') = \frac{\partial z_k^{l-1}(x, \theta)}{\partial \theta^l}^\top \frac{\partial z_k^{l-1}(x', \theta)}{\partial \theta^l} \xrightarrow{N_l \rightarrow \infty} \Theta_{\infty}^{l-1}(x, x') \delta_{k,k'} \quad (91)$$

we prove:

$$\Theta_{k,k'}^l(x, x') = \frac{\partial z_k^l(x, \theta)}{\partial \theta^l}^\top \frac{\partial z_k^l(x', \theta)}{\partial \theta^l} \xrightarrow{N_{l+1} \rightarrow \infty} \Theta_{\infty}^l(x, x') \delta_{k,k'} \quad (92)$$

7.4 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.(88), we have:

$$\begin{bmatrix} \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W_{1,j}^1 \phi(x_1) + \sigma_b b_1^1 \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W_{k,j}^1 \phi(x_k) + \sigma_b b_k^1 \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W_{N_2,j}^1 \phi(x_{N_1}) + \sigma_b b_{N_2}^1 \end{bmatrix} = \begin{bmatrix} z_1^1(x) \\ \vdots \\ z_k^1(x) \\ \vdots \\ z_{N_2}^1(x) \end{bmatrix} \quad (93)$$

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

$$\begin{aligned} \frac{\partial z_k^l(x)}{\partial W_{i,j}} &= \begin{cases} \frac{1}{\sqrt{d_{\text{in}}}} x_i & \text{if } i = k \text{ i.e., row } k \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{1}{\sqrt{d_{\text{in}}}} \delta_{i,k} x_i \\ \implies \frac{\partial z_{k'}^l(x)}{\partial W_{i,j}} &= \frac{1}{\sqrt{d_{\text{in}}}} \delta_{i,k'} x_i \end{aligned} \quad (94)$$

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

$$\begin{aligned}
\Theta_{k,k'}^1(x, x') &= \sum_{d=1}^{|\theta^1|} \frac{\partial F_k^1(x)}{\partial \theta_d^1} \frac{\partial F_{k'}^1(x')}{\partial \theta_d^1} \quad \theta^1 = \{W^1, b^1\} \\
&= \sum_{d=1}^{|W^1|} \frac{\partial F_k^1(x)}{\partial W_d^1} \frac{\partial F_{k'}^1(x')}{\partial W_d^1} + \sum_{d=1}^{|b^1|} \frac{\partial F_k^1(x)}{\partial b_d^1} \frac{\partial F_{k'}^1(x')}{\partial b_d^1} \\
&= \sum_{i=1}^{N_2} \sum_{j=1}^{d_{\text{in}}} \frac{\partial z_k^1(x)}{\partial W_{i,j}} \frac{\partial z_{k'}^1(x')}{\partial W_{i,j}} + \sum_{i=1}^{N_2} \frac{\partial z_k^1(x)}{\partial b_i} \frac{\partial z_{k'}^1(x')}{\partial b_i} \\
&= \sum_{i=1}^{N_2} \sum_{j=1}^{d_{\text{in}}} \frac{1}{\sqrt{d_{\text{in}}}} x_i \delta_{i,k'} \frac{1}{\sqrt{d_{\text{in}}}} x'_i \delta_{i,k} + \sum_{i=1}^{N_2} \sigma_b \delta_{i,k} \sigma_b \delta_{i,k'} \quad \text{only one } i \in \{1, \dots, N_2\} \text{ in outer sum remain} \\
&= \sum_{j=1}^{d_{\text{in}}} \frac{1}{d_{\text{in}}} x_i x'_i \delta_{k,k'}^2 + \sigma_b^2 \delta_{k,k'} \quad \delta_{i,k'} \delta_{i,k} = \delta_{k,k'} \\
&= \frac{1}{d_{\text{in}}} x^\top x' \delta_{k,k'} + \sigma_b^2 \delta_{k,k'} \\
&= \underbrace{\left(\frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2 \right)}_{K^1} \delta_{k,k'} \\
&\equiv K^1(x, x') \delta_{k,k'}
\end{aligned} \tag{95}$$

the above is just notation for NNGP

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

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

$$\begin{aligned}
\Rightarrow \Theta^1(x, x') &= \left[\underbrace{\begin{bmatrix} K^1(x, x') & \dots & 0 & \dots & 0 \\ 0 & K^1(x, x') & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & K^1(x, x') & 0 \\ 0 & 0 & 0 & 0 & K^1(x, x') \end{bmatrix}}_{k \in \{1, \dots, N_2\}} \right]_{k' \in \{1, \dots, N_2\}} \\
&= \text{repeating diagonal with } K^1(x, x') \delta_{k,k'} \\
&= \underbrace{K^1(x, x')}_{\text{scalar}} \otimes_{\text{outer}} \mathbf{I}_{N_1 \times N_2}
\end{aligned} \tag{96}$$

Θ^1 matrix of square the size of input $|z^1|$, importantly, there is no limit to take for Θ^1

7.5 when $l > 1$

$$\begin{bmatrix} \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{1,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_1^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{N_{l+1},j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_{N_{l+1}}^l \end{bmatrix} = \begin{bmatrix} z_1^l(x) \\ \vdots \\ z_k^l(x) \\ \vdots \\ z_{N_{l+1}}^l(x) \end{bmatrix} \quad (97)$$

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

$$\begin{aligned} \Theta_{k,k'}^l(x, x') &= \sum_{d=1}^{|\theta^l|} \frac{\partial z_k^1(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}} \\ &= \underbrace{\sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^1(x)}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x')}{\partial \{W^l, b^l\}}}_{\textcircled{1}} + \underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^1(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}}}_{\textcircled{2}} \end{aligned} \quad (98)$$

7.5.1 discussion on the term $\Theta_{k,k'}^l(x, x')$

Unlike the NNGP where we assume independence between k, k' and correlation only occur between x, x' . In NTK, we do not assume even independence between k, k' , therefore we must compute the entire correlations between k, k' and x, x' pairs.

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

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

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

$$\begin{aligned} z_k^l &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi\left(\frac{1}{\sqrt{N_{l-1}}} \sum_{i=1}^{N_{l-1}} W_{j,i}^{l-1} \phi(z_i^{l-1}(x)) + \sigma_b b_j^{l-1}\right) + \sigma_b b_k^l \end{aligned} \quad (99)$$

$$\begin{aligned} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} &= \frac{\partial z_k^l(x)}{\partial \phi(z^{l-1}(x))} \frac{\partial \phi(z^{l-1}(x))}{\partial z^{l-1}(x)} \frac{\partial z^{l-1}(x)}{\partial \theta_d^{l-1}} \quad \text{drop index for the last two terms} \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \frac{\partial \phi(z_j^{l-1}(x))}{\partial z_j^{l-1}(x)} \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi'(z_j^{l-1}(x)) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \quad \text{leave last derivative as is, in "recursion"} \end{aligned} \quad (100)$$

substitute it back to (2)

$$\begin{aligned}
& \underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}}}_{(2)} \\
&= \sum_{d=1}^{|\theta^{l-1}|} \left(\frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi'(z_j^{l-1}(x)) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \right) \times \left(\frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k',j}^l \phi'(z_j^{l-1}(x')) \frac{\partial z_j^{l-1}(x')}{\partial \theta_d^{l-1}} \right) \quad \text{by substitution} \\
& \quad (101)
\end{aligned}$$

although it looks like it is in the form of Section[5.4.2], however, $\underbrace{W_{k,j}^l \phi'(z_j^{l-1}(x)) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}}}_{\text{is not}}$

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

$$\begin{aligned}
&= \sum_{d=1}^{|\theta^{l-1}|} \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} \left(W_{k,j}^l \phi'(z_j^{l-1}(x)) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \right) \times \underbrace{\left(W_{k',j'}^l \phi'(z_{j'}^{l-1}(x')) \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}} \right)}_{j \rightarrow j' \text{ in second term}} \quad \text{re-arrange} \\
&= \sum_{d=1}^{|\theta^{l-1}|} \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \phi'(z_j^{l-1}(x)) \phi'(z_{j'}^{l-1}(x')) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}} \quad \text{re-arrange} \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \phi'(z_j^{l-1}(x)) \phi'(z_{j'}^{l-1}(x')) \underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}}}_{\text{definition } \Theta_{j,j'}^{l-1}(x,x')} \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \phi'(z_j^{l-1}(x)) \phi'(z_{j'}^{l-1}(x)) \Theta_{j,j'}^{l-1}(x,x') \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l W_{k',j'}^l \phi'(z_j^{l-1}(x)) \phi'(z_{j'}^{l-1}(x)) \Theta_{\infty}^{l-1}(x,x') \delta_{j,j'} \\
& \quad \text{use induction assumption: } \Theta_{j,j'}^{l-1}(x,x') \rightarrow \underbrace{\Theta_{\infty}^{l-1}(x,x') \delta_{j,j'}}_{\text{deterministic and diagonal limit}} \\
&= \Theta_{\infty}^{l-1}(x,x') \frac{1}{N_l} \sum_{j=1}^{N_l} W_{k,j}^l W_{k',j}^l \phi'(z_j^{l-1}(x)) \phi'(z_j^{l-1}(x')) \\
& \quad (102)
\end{aligned}$$

instead of using CLT, we shall apply LoLN here:

$$\begin{aligned}
& \Theta_{\infty}^{l-1}(x, x') \frac{1}{N_l} \sum_{j=1}^{N_l} W_{k,j}^l W_{k',j}^l \phi'(z_j^{l-1}(x)) \phi'(z_j^{l-1}(x')) \\
&= \underbrace{\Theta_{\infty}^{l-1}(x, x') \mathbb{E}_{W_{k,1}^l, W_{k',1}^l, z_1^{l-1}(x), z_1^{l-1}(x')} \left[W_{k,1}^l, W_{k',1}^l \phi'(z_1^{l-1}(x)) \phi'(z_1^{l-1}(x')) \right]}_{\text{very similar to NNGP}} \\
&= \Theta_{\infty}^{l-1}(x, x') \mathbb{E}_{(z_1^{l-1}(x), z_1^{l-1}(x'))} \left[\phi'(z_1^{l-1}(x)) \phi'(z_1^{l-1}(x')) \right] \mathbb{E}_{W_{k,1}^l, W_{k',1}^l} [W_{k,1}^l W_{k',1}^l] \\
&= \Theta_{\infty}^{l-1}(x, x') \mathbb{E}_{z^{l-1} \sim \mathcal{GP}(0, K^{l-1})} \left[\phi'(z_1^{l-1}(x)) \phi'(z_1^{l-1}(x')) \right] \delta_{k,k'} \\
&= \delta_{k,k'} \dot{K}^l(x, x') \Theta_{\infty}^{l-1}(x, x')
\end{aligned} \tag{103}$$

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

$$\begin{aligned}
\mathbb{E}_{W_{k,1}^l, W_{k',1}^l} [W_{k,1}^l W_{k',1}^l] &= \begin{cases} \mathbb{E}[W_{k,1}^l W_{k',1}^l] & k \neq k' \\ \mathbb{E}[(W_{k,1}^l)^2] & k = k' \end{cases} \\
&= \begin{cases} 0 & k \neq k' \\ 1 & k = k' \end{cases} \quad \text{re-parameterized expression} \quad W_{k,1}^l \sim \mathcal{N}(0, 1) \\
&= \delta_{k,k'}
\end{aligned} \tag{104}$$

2. notice the expression here:

$$\frac{1}{N_l} \sum_{j=1}^{N_l} W_{k,j}^l W_{k',j}^l \phi'(z_j^{l-1}(x)) \phi'(z_j^{l-1}(x')) \tag{105}$$

is the very similar of NNGP formulation, except:

$$\phi(z_j^{l-1}(x)) \rightarrow \phi'(z_j^{l-1}(x)) \tag{106}$$

so expect same CLT/LoLN treatment applies here

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

$$\begin{aligned}
\dot{K}^l(x, x') &= \sigma_w^2 \mathbb{E}_{(z_1^{l-1}(x), z_1^{l-1}(x')) \sim \mathcal{N}(0, K^{l-1}(x, x'))} \left[\phi'(z_1^{l-1}(x)) \phi'(z_1^{l-1}(x')) \right] \\
&= \mathbb{E}_{(z_1^{l-1}(x), z_1^{l-1}(x')) \sim \mathcal{N}(0, K^{l-1}(x, x'))} \left[\phi'(z_1^{l-1}(x)) \phi'(z_1^{l-1}(x')) \right] \quad \text{assume } \sigma_w = 1
\end{aligned} \tag{107}$$

compare with Eq. (68) the recursion in NNGP:

$$K^l(x, x') = \sigma_b^2 + \sigma_w^2 \mathbb{E}_{(z_1^{l-1}(x), z_1^{l-1}(x')) \sim \mathcal{N}(0, K^{l-1}(x, x'))} \left[\phi(z_1^{l-1}(x)) \phi(z_1^{l-1}(x')) \right] \tag{108}$$

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 $\dot{K}^l(x, x')$

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

$$\sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^l(x)}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x')}{\partial \{W^l, b^l\}} \quad (109)$$

and compare that with for $l = 1$:

$$\begin{aligned} \sum_{d=1}^{|\theta^1|} \frac{\partial z_k^1(x)}{\partial \theta_d^1} \frac{\partial z_{k'}^1(x')}{\partial \theta_d^1} \quad \theta^1 = \{W^1, b^1\} \\ = \left(K^1(x, x') \equiv \frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2 \right) \delta_{k, k'} \end{aligned} \quad (110)$$

then, we do know:

$$\begin{aligned} \sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^l(x)}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x')}{\partial \{W^l, b^l\}} \\ = \left(K^l(x, x') \equiv \frac{1}{N_l} \phi(z^l(x))^\top \phi(z^l(x')) + \sigma_b^2 \right) \delta_{k, k'} \end{aligned} \quad (111)$$

7.5.4 putting all together

$$\begin{aligned} \Theta_{k, k'}^l(x, x') &= \sum_{d=1}^{|W^l, b^l|} \frac{\partial z_k^l(x)}{\partial \{W^l, b^l\}} \frac{\partial z_{k'}^l(x')}{\partial \{W^l, b^l\}} + \sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^1(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}} \\ &= K^l(x, x') \delta_{k, k'} + \delta_{k, k'} \dot{K}^l(x, x') \Theta_{\infty}^{l-1}(x, x') \quad N_{l+1} \rightarrow \infty \\ &= \left(K^l(x, x') + \dot{K}^l(x, x') \Theta_{\infty}^{l-1}(x, x') \right) \delta_{k, k'} \\ &= \Theta_{\infty}^l(x, x') \delta_{k, k'} \end{aligned} \quad (112)$$

this does what we want to achieve in Eq.[91], 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:

$$\Theta^l(x, x') = \underbrace{\left(K^l(x, x') + \dot{K}^l(x, x') \Theta_{\infty}^{l-1}(x, x') \right)}_{\text{scalar}} \otimes_{\text{outer}} \underbrace{\mathbf{I}_{N_{l+1} \times N_{l+1}}}_{\text{same value for all } k, k' \text{ pairs}} \quad (113)$$

7.5.5 apply the above to $l = 1$

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

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.