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:

$$
\begin{aligned}
&z^l_k(x) = b^l_k + \sum_{j=1}^{N_l} W^l_{k,j} \times \phi\Big(z^{l-1}_j(x)\Big) \qquad W^l_{k,j} \sim \mathcal{N}\Big(0, \frac{1}{\sqrt{N_l}}\Big) \quad b^l_k \sim \mathcal{N}\big(0, \sigma_b\big) \quad \text{or :} \\
&z^l_k(x) = \sigma_b b^l_k + \sum_{j=1}^{N_l} \frac{1}{\sqrt{N_l}} W^l_{k,j} \times \phi\Big(z^{l-1}_j(x)\Big) \qquad W^l_{k,j} \sim \mathcal{N}\big(0, 1\big) \quad b^l_k \sim \mathcal{N}\big(0, 1\big)
\end{aligned} \tag{1}
$$

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 turorial self-contained, I have included a very quick introduction on the relavant topics incluce 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:

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

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

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

2.1 marginal and conditional marginal under noisy output

- in a regression setting:

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

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

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

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

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

2.2 marginal and conditional marginal under noiseless output

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

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

for arbitrary variable x

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

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

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

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

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

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

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

3 Kernel methods

consider the equation:

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

using definition:

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

Ridge regression can be re-written as:

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

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

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

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

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

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

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

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

3.2 relationship with Neural Tangent Kernel

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

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

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:

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

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

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

so let's look at $\frac{\mathrm{d}\tilde{y}(\theta_t)}{\mathrm{d}t}$:

$$\begin{aligned} \frac{\mathrm{d}\tilde{y}(\theta_t)}{\mathrm{d}t} &= \frac{\partial \tilde{y}(\theta(t))}{\partial \theta(t)}^\top \frac{\mathrm{d}\theta(t)}{\mathrm{d}t} \\ &= \nabla_\theta \tilde{y}(\theta)\Big(- \nabla_\theta \tilde{y}(\theta)\big(\tilde{y}(\theta) - y\big)\Big) \\ &= - \underbrace{\nabla_\theta \tilde{y}(\theta)^\top \nabla_\theta \tilde{y}(\theta)}\big(\tilde{y}(\theta) - y\big) \\ &\approx -K(\theta_0)\big(\tilde{y}(\theta) - y\big) \end{aligned} \tag{11}$$

4 First attempt for modeling Neural Network at initialization

4.1 neural network function

using parameters:

$$\theta \equiv \{W^L, b^L, \dots W^1, b^1\} \tag{12}$$

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

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

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

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

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:

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

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:

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

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

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

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

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

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

4.3.1 look at the mean:

$$
\begin{aligned}
&\mathbb{E}\left[\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right] \\
&= \int_y \int_{y^\star} \begin{bmatrix} y \\ y^\star \end{bmatrix} p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) \mathrm{d}y \, \mathrm{d}y^\star \\
&= \int_y \int_{y^\star} \begin{bmatrix} y \\ y^\star \end{bmatrix} \int_\theta p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \underline{\theta, \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}}\right) p(\theta|\sigma_\theta^2) \, \mathrm{d}\theta \, \mathrm{d}y \, \mathrm{d}y^\star \\
&= \int_\theta \underbrace{\int_y \int_{y^\star} \begin{bmatrix} y \\ y^\star \end{bmatrix} \mathcal{N}\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \underline{\begin{bmatrix} f_\theta(X) \\ f_\theta(x^\star) \end{bmatrix}}, \sigma_\epsilon^2 I\right) \mathrm{d}y \, \mathrm{d}y^\star}_{\mathbb{E}\left[\begin{bmatrix} y \\ y^\star \end{bmatrix}\right] = \begin{bmatrix} f_\theta(X) \\ f_\theta(x^\star) \end{bmatrix}} \mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) \mathrm{d}\theta \\
&= \int \begin{bmatrix} f_\theta(X) \\ f_\theta(x^\star) \end{bmatrix} \mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) \, \mathrm{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 \mid 0, \sigma_w^2 I) \mathcal{N}(b^L \mid 0, \sigma_b^2 I) \mathcal{N}(\theta^{1,\ldots,L-1} \mid 0, \sigma_\theta^2 I) \mathrm{d}\theta^{1,\ldots,L-1} \mathrm{d}W^L \, \mathrm{d}b^L \\
&= \int \begin{bmatrix} \phi^L(X) \underbrace{\int W^L \mathcal{N}(W^L \mid 0, \sigma_w^2 I) \mathrm{d}W^L}_{=0} + \underbrace{\int b^L \mathcal{N}(b^L \mid 0, \sigma_b^2 I) \mathrm{d}b^L}_{=0} \\ \phi^L(x^{*\top}) \underbrace{\int W^L \mathcal{N}(W^L \mid 0, \sigma_w^2 I) \mathrm{d}W^L}_{=0} + \underbrace{\int b^L \mathcal{N}(b^L \mid 0, \sigma_b^2 I) \mathrm{d}b^L}_{=0} \end{bmatrix} \mathcal{N}(\theta^{1,\ldots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\ldots,L-1} \\
&= \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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We need CLT for computing this probability.

5.3.1 Central Limit Theorem:

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

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

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

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

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

choose one of these conditions to suit the situation

5.3.2 Apply CLT to compute $p\big(z_k^l(x)\big)$

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 $p\big(z_k^l(x)\big)$ directly is hard!

however, $z_k^l(x)$ is $b_k^l +$ sum of i.i.d elements using CLT notations:

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

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

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

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

Expectation

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

Variance

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

we leave in this form, as

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

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.

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

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

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

the above is the key, implication is:

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

finally adding the bias b_k^l:
Note that sum of two **independent** Gaussian random variables is also Gaussian: (not to confuse with GMM!)

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

Therefore:

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

appreciate the recursion here

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

5.4.1 Independence after only one layer

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

Therefore, individual terms of the outer sum:

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

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 $\mathbb{E}\big[X_j^{(p)}\big] = 0 \quad \forall j, p$:

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

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

$$\begin{aligned} \mathbf{Cov}\Big[\sum_{j=1}^{N} X_j^{(p)}, \sum_{j=1}^{N} X_j^{(q)}\Big] &= N\mathbf{Cov}\big[X_1^{(p)}, X_1^{(q)}\big] \qquad N \to \infty \\ \implies \mathbb{E}\Big[\sum_{j=1}^{N} X_j^{(p)} X_j^{(q)}\Big] &= N\mathbf{Cov}\big[X_1^{(p)}, X_1^{(q)}\big] \qquad N \to \infty \quad \text{as well!} \end{aligned} \tag{45}$$

Therefore:

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

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

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

$$
\implies \overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_i^{(k)} \end{bmatrix} = \begin{bmatrix} \overline{\mathbf{X}}^{(1)} \\ \vdots \\ \overline{\mathbf{X}}^{(p)} \\ \vdots \\ \overline{\mathbf{X}}^{(q)} \\ \vdots \\ \overline{\mathbf{X}}^{(k)} \end{bmatrix}
$$

1. **Sample mean version**

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

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

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

2. **Sample sum version**:

$$
\begin{aligned}
\left(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]-n\mathbb{E}[\mathbf{X}_1]\right)&\xrightarrow{d}\mathcal{N}_k(0,n\mathbf{\Sigma}) \\
\implies \mathbb{E}\Bigg[\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(p)}-n\mathbb{E}[\mathbf{X}_1]^{(p)}\bigg)\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(q)}-n\mathbb{E}[\mathbf{X}_1]^{(q)}\bigg)\Bigg]&=n\mathbf{\Sigma}_{(p),(q)} \\
\implies \mathbb{E}\Big[\Big(n\overline{\mathbf{X}}^{(p)}-n\mathbb{E}[X_1^{(p)}]\Big)\Big(n\overline{\mathbf{X}}^{(q)}-n\mathbb{E}[X_1^{(q)}]\Big)\Big]&=n\mathbf{\Sigma}_{(p),(q)} \\
\implies \mathbb{E}\Bigg[\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(p)}-n\mathbb{E}[X_1^{(p)}]\bigg)\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(p)}-n\mathbb{E}[X_1^{(q)}]\bigg)\Bigg]&=n\mathbf{\Sigma}_{(p),(q)}
\end{aligned} \tag{50}
$$

where $\mathbf{\Sigma}_{(p),(q)}=\mathrm{Cov}\big(X_1^{(p)},X_1^{(q)}\big)$

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

$$
\begin{bmatrix} \vdots \\ W^l_{k,1}\phi\big(z^{l-1}_1(x)\big) \\ \vdots \\ W^l_{k,1}\phi\big(z^{l-1}_j(x')\big) \\ \vdots \end{bmatrix} + \cdots + \begin{bmatrix} \vdots \\ W^l_{k,N_l}\phi\big(z^{l-1}_j(x)\big) \\ \vdots \\ W^l_{k,N_l}\phi\big(z^{l-1}_j(x')\big) \\ \vdots \end{bmatrix} = \underbrace{\begin{bmatrix} \vdots \\ \sum_{j=1}^{N_l} W^l_{k,j}\phi\big(z^{l-1}_j(x)\big) \\ \vdots \\ \sum_{j=1}^{N_l} W^l_{k,j}\phi\big(z^{l-1}_j(x')\big) \\ \vdots \end{bmatrix}}_{\begin{bmatrix} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \sum_{i=1}^n X_i^{(k)} \end{bmatrix}} = \underbrace{\begin{bmatrix} \vdots \\ z^l_k(x) \\ \vdots \\ z^l_k(x') \\ \vdots \end{bmatrix}}_{\sum_{i=1}^n \mathbf{X}_i} \tag{51}
$$

compare with the standard notation of Multi-dimensional CLT, and use "sample sum version" of CLT, Eq.[47], and remember $z^l_k(x) = b^l_k + \sum_{j=1}^{N_l} W^l_{k,j}\phi\big(z^{l-1}_j(x)\big)$, to simplify derivation, let's deliberately not looking at b^l_k for now

$$
\begin{aligned}
\sum_i^n X_i^{(p)} = \Big[\sum_i^n X_i\Big]^{(p)} &\longrightarrow \sum_{j=1}^{N_l} W^l_{k,j}\phi\big(z^{l-1}_j(x)\big) \\
\sum_i^n X_i^{(q)} = \Big[\sum_i^n X_i\Big]^{(p)} &\longrightarrow \sum_{j=1}^{N_l} W^l_{k,j}\phi\big(z^{l-1}_j(x')\big) \\
X_1^{(p)} &\longrightarrow W^l_{k,j}\phi\big(z^{l-1}_j(x)\big) \quad \text{be the single term in the sum} \\
X_1^{(q)} &\longrightarrow W^l_{k,j}\phi\big(z^{l-1}_j(x')\big)
\end{aligned} \tag{52}
$$

using above identities in Eq.[52]

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\Big[\sum_i^n \mathbf{X}_i\Big]^{(p)} - n\mathbb{E}\big[X_1^{(p)}\big]\Big)\Big(\Big[\sum_i^n \mathbf{X}_i\Big]^{(q)} - n\mathbb{E}\big[X_1^{(q)}\big]\Big)\Big] = n\mathbf{Cov}(\mathbf{X}_1)_{(p),(q)} \\
\Longrightarrow\ &\mathbb{E}\Big[\Big(\sum_{j=1}^{N_l} W^l_{k,j}\phi(z^{l-1}_j(x)) - N_l\underbrace{\mathbb{E}\big[W^l_{k,1}\phi(z^{l-1}_1(x))\big]}_{=0}\Big)\Big(\sum_{j=1}^{N_l} W^l_{k,j}\phi(z^{l-1}_j(x')) - N_l\underbrace{\mathbb{E}\big[W^l_{k,1}\phi(z^{l-1}_1(x'))\big]}_{=0}\Big)\Big] \\
&= N_l\mathbf{Cov}\Big(W^l_{k,1}\phi(z^{l-1}_1(x)), W^l_{k,1}\phi(z^{l-1}_1(x'))\Big) \\
&= N_l\,\mathbb{E}\Big[W^l_{k,1}\phi(z^{l-1}_1(x)) \times W^l_{k,1}\phi(z^{l-1}_1(x'))\Big]
\end{aligned} \tag{53}
$$

look at $z^l_k(x)$ with b^l_k too:

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

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" $\mathbf{Cov}\big(z^l, z^l\big)$ is of size:

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

   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:

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

   note 3: in literature, it is written:

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

   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:

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

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

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

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

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

5.5 looking at GP systematically

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

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

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

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

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

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

5.5.1 using properties of point Marginals of Gaussian Process:

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

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

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

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

5.6 in summary

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

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

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 $\{z_1^l(x), \ldots, z_{N_{l+1}}^l(x)\}_{x \in \mathcal{X}}$ 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 $\phi \to \phi'$

6 Expand GP across all layers

this section describe [2]

From the previous section, we know that that "independence" property on $W^{(2)}_{k,j} \times \phi\big(b^{(1)}_j + \sum_{i=1}^{N_1} W^{(1)}_{j,i} x_i\big)$ and $W^{(2)}_{k,j'} \times \phi\big(b^{(1)}_{j'} + \sum_{i=1}^{N_1} W^{(1)}_{j',i} x_i\big)$ 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:

$$\begin{aligned} p(z^L|x) &= \int p(z^L, K^0, K^1, \dots, K^L|x)\, \mathrm{d}K^{0,\dots,L} \\ &= \int p(z^L|K^L) \Big(\prod_{l=1}^{L} p(K^l|K^{l-1}) \Big) p(K^0|x)\, \mathrm{d}K^{0,\dots,L} \end{aligned} \tag{69}$$

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

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

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

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:

$$z^l_k(x) = b^l_k + \sum_{j=1}^{N_l} W^l_{k,j} \phi(z^{l-1}_j(x)) \tag{71}$$

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

$$\begin{aligned} z^l_k(x) \big|\, \{\phi(z^{l-1}_j(x))\}^{N_l}_j &= b^l_k + \sum_{j=1}^{N_l} W^l_{k,j} \underbrace{\phi(z^{l-1}_j(x))}_{\text{constant}} \\ \implies\ z^l_k(x) \big|\, \{\phi(z^{l-1}_j(x))\}^{N_l}_j &\sim \mathcal{N}\Big(0, \sigma_b^2 + \sum_{j=1}^{N_l} \phi(z^{l-1}_j(x))^2 \mathrm{Var}\big[W^l_{k,j}\big]\Big) \\ &= \mathcal{N}\Big(0, \sigma_b^2 + \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(z^{l-1}_j(x))^2\Big) \end{aligned} \tag{72}$$

using property of weighted sum of Gaussian:

$$\begin{aligned} X_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad i = 1, \dots, \\ \implies \sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2\ \mathrm{Var}[X_i] \right) \end{aligned} \tag{73}$$

Conditional Co-variance

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

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

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

in GP paradigm:

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

where

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

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

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

to reflect:

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

note that

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

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

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

$$\begin{aligned}
&\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} G(K_{1,1}^l) & \dots & 0 & \dots\dots\dots & G(K_{1,|\mathcal{D}|}^l) & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & G(K_{1,1}^l) & \dots\dots\dots & 0 & 0 & G(K_{1,|\mathcal{D}|}^l) \\ -- & -- & -- & ---- & -- & -- & -- \\ G(K_{2,1}^l) & \dots & 0 & \dots\dots\dots & G(K_{2,|\mathcal{D}|}^l) & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & G(K_{2,1}^l) & \dots\dots\dots & 0 & 0 & G(K_{2,|\mathcal{D}|}^l) \\ -- & -- & -- & ---- & -- & -- & -- \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ -- & -- & -- & -- & -- & -- & -- \\ G(K_{|\mathcal{D}|,1}^l) & \dots & 0 & \dots\dots\dots & G(K_{|\mathcal{D}|,|\mathcal{D}|}^l) & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & G(K_{|\mathcal{D}|,1}^l) & \dots\dots\dots & 0 & 0 & G(K_{|\mathcal{D}|,|\mathcal{D}|}^l) \end{bmatrix}\right) \\
\implies p(\mathbf{z}^l|K^l) &= \mathcal{N}\Big(\mathbf{0}, G(K^l) \otimes \mathbf{I}_{N_{l+1}\times N_{l+1}}\Big) \\
&= \mathcal{GP}\big(\mathbf{z}^l; 0, G(K^l)\big)
\end{aligned} \tag{81}$$

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

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

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

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

$$
\begin{aligned}
p(K^l_{p,q}|K^{l-1}_{p,q}) = \int_{z^{l-1}(x)\ ,z^{l-1}(x')} & p\Big(\frac{1}{N_l}\sum_{j=1}^{N_l}\phi\big(z^l_j(x)\big)\phi\big(z^l_j(x')\big)\Big) \\
& \mathcal{N}\Bigg(\begin{bmatrix} z^{l-1}(x) \\ z^{l-1}(x') \end{bmatrix}; 0, G\Bigg(\begin{bmatrix} K^{l-1}(x,x) & K^{l-1}(x,x') \\ K^{l-1}(x,x') & K^{l-1}(x',x') \end{bmatrix}\Bigg)\Bigg)
\end{aligned} \tag{83}
$$

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

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

$$
\begin{aligned}
\underbrace{K^l_{p,q}}_{\overline{X}} &= \frac{1}{N_l}\sum_{j=1}^{N_l}\underbrace{\phi\big(z^{l-1}_j(x)\big)\phi\big(z^{l-1}_j(x')\big)}_{X_j \equiv K^{l,j}_{p,q}} \\
\implies p(K^{l,1}_{p,q}|K^{l-1}_{p,q}) &= \int_{z^{l-1}(x)\ ,z^{l-1}(x')} p\big(\phi\big(z^l_j(x)\big)\phi\big(z^l_j(x')\big)\big) \\
& \qquad \mathcal{N}\Bigg(\begin{bmatrix} z^{l-1}(x) \\ z^{l-1}(x') \end{bmatrix}; 0, G\Bigg(\begin{bmatrix} K^{l-1}(x,x) & K^{l-1}(x,x') \\ K^{l-1}(x,x') & K^{l-1}(x',x') \end{bmatrix}\Bigg)\Bigg) \\
&= (F \circ G)(K^{l-1}_{p,q})
\end{aligned} \tag{84}
$$

using CLT, pick the most appropriate definition:

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

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

$$
\begin{aligned}
(\overline{X} - \mathbb{E}[X_1]) &\xrightarrow{d} \mathcal{N}\Big(0, \frac{\text{Var}[X_1]}{n}\Big) \\
\implies \big(K^l_{p,q} - \mathbb{E}[K^{l,1}_{p,q}]\big) &\xrightarrow{d} \mathcal{N}\Big(0, \frac{\text{Var}[K^{l,1}_{p,q}]}{N_l}\Big) \\
\implies \big(K^l_{p,q} - (F\circ G)(K^{l-1}_{p,q})\big) &\xrightarrow{d} \mathcal{N}\Big(0, \frac{\text{Var}[K^{l,1}_{p,q}]}{N_l}\Big) \\
\implies \big(K^l_{p,q}|K^{l-1}_{p,q}\big) &\xrightarrow{d} \mathcal{N}\Big((F\circ G)(K^{l-1}), \frac{\text{Var}[K^{l,1}_{p,q}]}{N_l}\Big) \\
\implies \lim_{N_l\to\infty} p(K^l|K^{l-1}) &= \delta\big(K^l - (F\circ G)(K^{l-1})\big) \quad \text{entire matrix}
\end{aligned} \tag{86}
$$

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 $\to \infty$:

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

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 $z^{l-1}(x) \to z^l(x)$:

$$\begin{bmatrix} \frac{1}{\sqrt{N_l}} W^l_{1,1}\phi\big(z^{l-1}_1(x)\big) + \sigma_b b_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{k,1}\phi\big(z^{l-1}_1(x)\big) + \sigma_b b_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{N_{l+1},1}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b_{N_{l+1}} \end{bmatrix} + \cdots + \begin{bmatrix} \frac{1}{\sqrt{N_l}} W^l_{1,N_l}\phi\big(z^{l-1}_1(x)\big) + \sigma_b b^l_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{k,N_l}\phi\big(z^{l-1}_1(x)\big) + \sigma_b b^l_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{N_{l+1},N_l}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b^l_{N_{l+1}} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{1,j}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b^l_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{k,j}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b^l_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{N_{l+1},j}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b^l_{N_{l+1}} \end{bmatrix} = \begin{bmatrix} z^l_1(x) \\ \vdots \\ z^l_k(x) \\ \vdots \\ z^l_{N_{l+1}}(x) \end{bmatrix} \tag{88}$$

7.2 re-parameterized formulation

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

$$\begin{aligned} \text{NNGP} \quad & z^l_k(x) = \sum_{j=1}^{N_l} W^l_{k,j}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b^l_k \qquad W^l_{k,j} \sim \mathcal{N}\Big(0, \frac{1}{\sqrt{N_l}}\Big) \\ \text{this section} \quad & z^l_k(x) = \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{k,j}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b^l_k \qquad W^l_{k,j} \sim \mathcal{N}(0,1) \end{aligned} \tag{89}$$

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:

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

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

2. then by assuming:

$$\Theta^{l-1}_{k,k'}(x,x') = \frac{\partial z_k^{l-1}(x,\theta)}{\partial \theta^l}^\top \frac{\partial z_k^{l-1}(x',\theta)}{\partial \theta^l} \xrightarrow{N_l \to \infty} \Theta^{l-1}_\infty(x,x')\delta_{k,k'} \tag{91}$$

we prove:

$$\Theta^{l}_{k,k'}(x,x') = \frac{\partial z_k^{l}(x,\theta)}{\partial \theta^l}^\top \frac{\partial z_k^{l}(x',\theta)}{\partial \theta^l} \xrightarrow{N_{l+1} \to \infty} \Theta^{l}_\infty(x,x')\delta_{k,k'} \tag{92}$$

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

From the Eq.(88), we have:

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

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 $i \neq k$

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

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:

$$
\begin{aligned}
\Theta^1_{k,k'}(x,x') &= \sum_{d=1}^{|\theta^1|} \frac{\partial F^1_k(x)}{\partial \theta^1_d} \frac{\partial F^1_{k'}(x')}{\partial \theta^1_d} \qquad \theta^1 = \{W^1, b^1\} \\
&= \sum_{d=1}^{|W^1|} \frac{\partial F^1_k(x)}{\partial W^1_d} \frac{\partial F^1_{k'}(x')}{\partial W^1_d} + \sum_{d=1}^{|b^1|} \frac{\partial F^1_k(x)}{\partial b^1_d} \frac{\partial F^1_{k'}(x')}{\partial b^1_d} \\
&= \sum_{i=1}^{N_2} \sum_{j=1}^{d_{\text{in}}} \frac{\partial z^1_k(x)}{\partial W_{i,j}} \frac{\partial z^1_{k'}(x')}{\partial W_{i,j}} + \sum_{i=1}^{N_2} \frac{\partial z^1_k(x)}{\partial b_i} \frac{\partial z^1_{k'}(x')}{\partial b_i} \\
&= \sum_{i=1}^{N_2} \sum_{j=1}^{d_{\text{in}}} \frac{1}{\sqrt{d_{\text{in}}}} x_i \delta_{i,k'} \ \frac{1}{\sqrt{d_{\text{in}}}} x'_i \delta_{i,k} + \sum_{i=1}^{N_2} \sigma_b \delta_{i,k} \ \sigma_b \delta_{i,k'} \quad \text{only one } i \in \{1, \dots N_2\} \text{ in outer sum remain} \\
&= \sum_{j=1}^{d_{\text{in}}} \frac{1}{d_{\text{in}}} x_i x'_i \delta^2_{k,k'} + \sigma_b^2 \delta_{k,k'} \qquad \delta_{i,k'}\delta_{i,k} = \delta_{k,k'} \\
&= \frac{1}{d_{\text{in}}} x^\top x' \delta_{k,k'} + \sigma_b^2 \delta_{k,k'} \\
&= \underbrace{\left(\frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2 \right)}_{K^1} \delta_{k,k'} \\
&\equiv K^1(x,x') \delta_{k,k'}
\end{aligned}
\tag{95}
$$

the above is just notation for NNGP

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

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

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

Θ^1 matrix of square the size of input $(N_2 \times |\mathcal{X}|) \times (N_2 \times |\mathcal{X}|)$, importantly, there is no limit to take for Θ^1

7.5 when $l > 1$

$$\begin{bmatrix} \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{1,j} \phi(z^{l-1}_j(x)) + \sigma_b b^l_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{k,j} \phi(z^{l-1}_j(x)) + \sigma_b b^l_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{N_{l+1},j} \phi(z^{l-1}_j(x)) + \sigma_b b^l_{N_{l+1}} \end{bmatrix} = \begin{bmatrix} z^l_1(x) \\ \vdots \\ z^l_k(x) \\ \vdots \\ z^l_{N_{l+1}}(x) \end{bmatrix} \tag{97}$$

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

$$\begin{aligned} \Theta^l_{k,k'}(x,x') &= \sum_{d=1}^{|\theta^l|} \frac{\partial z^1_k(x)}{\partial \theta^{l-1}_d} \frac{\partial z^l_{k'}(x')}{\partial \theta^{l-1}_d} \\ &= \underbrace{\sum_{d=1}^{|W^l,b^l|} \frac{\partial z^1_k(x)}{\partial \{W^l,b^l\}} \frac{\partial z^l_{k'}(x')}{\partial \{W^l,b^l\}}}_{1} + \underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z^1_k(x)}{\partial \theta^{l-1}_d} \frac{\partial z^l_{k'}(x')}{\partial \theta^{l-1}_d}}_{2} \end{aligned} \tag{98}$$

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

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 $\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z^1_k(x)}{\partial \theta^{l-1}_d} \frac{\partial z^l_{k'}(x')}{\partial \theta^{l-1}_d}$

in expression $\underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z^1_k(x)}{\partial \theta^{l-1}_d} \frac{\partial z^l_{k'}(x')}{\partial \theta^{l-1}_d}}_{2}$:

derivatives with respect to the single terms: $\frac{\partial z^1_k(x)}{\partial \theta^{l-1}_d}$

$$\begin{aligned} z^l_k &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{k,j} \phi(z^{l-1}_j(x)) + \sigma_b b^l_k \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{k,j} \phi\left(\frac{1}{\sqrt{N_{l-1}}} \sum_{j=1}^{N_{l-1}} W^{l-1}_{j,i} \phi(z^{l-1}_i(x)) + \sigma_b b^{l-1}_j \right) + \sigma_b b^l_j \end{aligned} \tag{99}$$

$$\begin{aligned} \frac{\partial z^l_k(x)}{\partial \theta^{l-1}_d} &= \frac{\partial z^l_k(x)}{\partial \phi(z^{l-1}(x))} \frac{\partial \phi(z^{l-1}(x))}{\partial z^{l-1}(x)} \frac{\partial z^{l-1}(x)}{\partial \theta^{l-1}_d} \quad \text{drop index for the last two terms} \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{k,j} \frac{\partial \phi(z^{l-1}_j(x))}{\partial z^{l-1}_j(x)} \frac{\partial z^{l-1}_j(x)}{\partial \theta^{l-1}_d} \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{k,j} \, \phi'(z^{l-1}_j(x)) \frac{\partial z^{l-1}_j(x)}{\partial \theta^{l-1}_d} \qquad \text{leave last derivative as is, in “recursion”} \end{aligned} \tag{100}$$

substitute it back to 2

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

$$
= \sum_{d=1}^{|\theta^{l-1}|} \left(\frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \ \phi'(z_j^{l-1}(x)) \ \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \right) \times \left(\frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k',j}^l \ \phi'(z_j^{l-1}(x')) \ \frac{\partial z_j^{l-1}(x')}{\partial \theta_d^{l-1}} \right) \quad \text{by substitution} \tag{101}
$$

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

$$
= \sum_{d=1}^{|\theta^{l-1}|} \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} \left(W_{k,j}^l \ \phi'(z_j^{l-1}(x)) \ \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \right) \times \underbrace{\left(W_{k',j'}^l \ \phi'(z_{j'}^{l-1}(x')) \ \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}} \right)}_{j \to j' \ \text{ in second term}} \quad \text{re-arrange}
$$

$$
= \sum_{d=1}^{|\theta^{l-1}|} \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l \ W_{k',j'}^l \phi'(z_j^{l-1}(x)) \ \phi'(z_{j'}^{l-1}(x')) \ \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \ \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}} \quad \text{re-arrange}
$$

$$
= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l \ W_{k',j'}^l \phi'(z_j^{l-1}(x)) \ \phi'(z_{j'}^{l-1}(x')) \ \underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \ \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}}}_{\text{definition } \Theta_{j,j'}^{l-1}(x,x')}
$$

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

$$
= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l \ W_{k',j'}^l \phi'(z_j^{l-1}(x)) \ \phi'(z_{j'}^{l-1}(x')) \ \Theta_{\infty}^{l-1}(x,x')\delta_{j,j'}
$$

use induction assumption: $\Theta_{j,j'}^{l-1}(x,x') \to \underbrace{\Theta_{\infty}^{l-1}(x,x')\delta_{j,j'}}_{\text{deterministic and diagonal limit}}$

$$
= \Theta_{\infty}^{l-1}(x,x') \frac{1}{N_l} \sum_{j=1}^{N_l} W_{k,j}^l \ W_{k',j}^l \phi'(z_j^{l-1}(x)) \ \phi'(z_j^{l-1}(x')) \tag{102}
$$

instead of using CLT, we shall apply LoLN here:

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

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

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

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:

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

   so expect same CLT/LoLN treatment applies here

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

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

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

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

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

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

and compare that with for $l=1$:

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

then, we do know:

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

7.5.4 putting all together

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

this does what we want to achieve in Eq.[91], by assuming $\Theta_{k,k'}^{l-1}(x,x') \xrightarrow{N_l\to\infty} \Theta_\infty^{l-1}(x,x')\delta_{k,k'}$, we prove: $\Theta_{k,k'}^l(x,x') \xrightarrow{N_{l+1}\to\infty} \Theta_\infty^l(x,x')\delta_{k,k'}$

then finally:

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

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.