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

Richard Xu

September 3, 2020

1 Preamble

In this tutorial, my contribution mainly has been the attempt to summarize the following papers and blogs in a unified and (hopefully) more intuitive for Computer Science researchers. In particular, the blogs below are extremely useful, and I encourage you to read the orginal blog as well.

- Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. arXiv preprint arXiv:1902.06720, 2019
- Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In Advances in neural information processing systems, pages 8571–8580, 2018
- J. H. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. ICLR, 2018
- Radford M. Neal. Priors for infinite networks (tech. rep. no. crg-tr-94-1). University of Toronto, 1994
- `https://www.uv.es/gonmagar/blog/2019/01/21/DeepNetworksAsGPs`
- `https://bryn.ai/jekyll/update/2019/04/02/neural-tangent-kernel.html`
- `http://chenyilan.net/files/ntk_derivation.pdf`
- `http://chenyilan.net/files/linearization.pdf`

1.1 notations

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

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

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

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

1.2 Others minor contributions

- I made the derivations a bit more verbose for people to follow
- 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:

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

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

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

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

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

3 GP for Neural Network: Direct computation

3.1 neural network function

using parameters:

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

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

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

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

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

3.2 Apply NN function in predictive distribution

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

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

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

3.3 what is the predictive distribution

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

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

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

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

- attempt to compute it **directly**, by looking the **mean** and **variance**:
  - $\mathbb{E}\left[\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix} \right]$
  - $\mathbb{E}\left[\begin{bmatrix} y \\ y^\star \end{bmatrix} \begin{bmatrix} y^\top & y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{\star\top} \end{bmatrix} \right]$

3.3.1 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}
$$

note we are not dealing with infinity at the moment

3.3.2 look at co-variance

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4 Single layer neural network

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

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

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

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

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

We need CLT for computing this probability.

4.1.1 Central Limit Theorem:

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

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

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

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

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

choose one of these conditions to suit the situation

4.1.2 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:

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

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

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

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

- **Expectation**

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

- **Variance**

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

we leave in this form, as

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

4.1.4 apply CLT:

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

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

- 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:

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

- the above is the key, implication is:

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

- finally adding the bias b_k^l:

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

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

  Therefore:

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

- appreciate the recursion here

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

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

4.2.1 Multidimensional CLT:

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

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

Therefore:

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

- **Sample mean version**:

$$\sqrt{n}\,\left(\overline{\mathbf{X}} - \mathbb{E}[\mathbf{X}_1]\right) \xrightarrow{d} \mathcal{N}_k(0, \boldsymbol{\Sigma})$$

$$\implies \sqrt{n}\,\mathbb{E}\Big[\Big(\underbrace{\overline{\mathbf{X}}^{(p)} - \mathbb{E}[\overline{\mathbf{X}}_1^{(p)}]}_{\text{scalar}}\Big)\Big(\underbrace{\overline{\mathbf{X}}^{(q)} - \mathbb{E}[\,\overline{\mathbf{X}}_1^{(q)}]}_{\text{scalar}}\Big)\Big] = \boldsymbol{\Sigma}_{(p),(q)}$$

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

- **Sample sum version**:

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

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

4.2.2 put in Multidimensional CLT structure:

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

Since we already know that:

$$
\begin{aligned}
&\mathbb{E}\Bigg[\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(p)} - n\mathbb{E}\big[X_1^{(p)}\big]\bigg)\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(q)} - n\mathbb{E}\big[X_1^{(q)}\big]\bigg)\Bigg] = n\mathbf{\Sigma}_{(p),(q)} \\
\implies &\mathbb{E}\Bigg[\bigg(\sum_{j=1}^{N_l} W_{k,j}^l\phi(z_j^{l-1}(x^{(p)})) - N_l\underbrace{\mathbb{E}\big[W_{k,1}^l\phi(z_1^{l-1}(x^{(p)}))\big]}_{=0}\bigg)\times \\
&\qquad\qquad\bigg(\sum_{j=1}^{N_l} W_{k,j}^l\phi(z_j^{l-1}(x^{(q)})) - N_l\underbrace{\mathbb{E}\big[W_{k,1}^l\phi(z_1^{l-1}(x^{(q)}))\big]\bigg)}_{=0}\Bigg] = N_l\mathbf{\Sigma}_{(p),(q)}
\end{aligned}
$$

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

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(p)}))\Big)\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(q)}))\Big)\Big] \\
&= N_l \, \boldsymbol{\Sigma}_{(p),(q)} \\
&= N_l \, \mathrm{Cov}\Big(W_{k,1}^l \phi(z_1^{l-1}(x^{(p)})), W_{k,1}^l \phi(z_1^{l-1}(x^{(q)}))\Big) \\
&= N_l \, \mathbb{E}\Big[W_{k,1}^l \phi(z_1^{l-1}(x^{(p)})) \times W_{k,1}^l \phi(z_1^{l-1}(x^{(q)}))\Big]
\end{aligned}$$

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

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

- **note 1**: this co-variance is same $\forall k$ in $z_k^l(x)$, so right hand side does not need to keep k index

  because in this particular setting, since b_k, $b_{k'}$, $W_{k,j}$ and $W_{k',j'}$ are independent variables, co-variance between any of them are zero:

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

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

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

  This is because of `note1` above

- regardless of this special property CLT still apply.

4.2.3 Relationship with Gaussian Process (GP):

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

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

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

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

- putting it in layer specific GP:

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

4.3 more on GP

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

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

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

using properties of point Marginals of Gaussian Process:

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

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

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

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

4.4 in summary

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

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

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

5 Expand GP across all layers

5.1 Overall objective

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

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

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

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

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

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

- look at Neural Network function:

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

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

- Conditional Marginal

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

using property of weighted sum of Gaussian:

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

- Conditional Co-variance

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

not using property of weighted sum of Gaussian:

- Combine all together

$$
\mathrm{Cov}\Big[z_k^l(x^{(p)}), z_k^l(x^{(q)}) \;\Big|\; \Big\{\phi\big(z_j^{l-1}(x^{(p)})\big), \phi\big(z_j^{l-1}(x^{(q)})\big)\Big\}_{j=1}^{N_l}\Big] = \sigma_b^2 + \sigma_w^2\,\frac{1}{N_l}\sum_{j=1}^{N_l} \phi\big(z_j^{l-1}(x^{(p)})\big)\,\phi\big(z_j^{l-1}(x^{(q)})\big)
$$

$$
z_k^l(x)\big|\,\{\phi(z_j^{l-1}(x))\}_j^{N_l} \sim \mathcal{N}\Big(0, \sigma_b^2 + \sigma_w^2\,\frac{1}{N_l}\sum_{j=1}^{N_l}\phi(z_j^{l-1}(x))^2\Big)
$$

$$
\implies \begin{bmatrix} z^l(x^{(p)}) \mid \phi\big(z_j^{l-1}(x^{(p)})\big) \\ z^l(x^{(q)}) \mid \phi\big(z_j^{l-1}(x^{(q)})\big) \end{bmatrix} \sim \mathcal{N}\Big(\mathbf{0}, G\Big(\begin{bmatrix} K^l(x^{(p)}, x^{(p)}) & K^l(x^{(p)}, x^{(q)}) \\ K^l(x^{(p)}, x^{(q)}) & K^l(x^{(q)}, x^{(q)}) \end{bmatrix}\Big)\Big)
$$

- in GP paradigm:

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

where

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

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

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

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

- to reflect:

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

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

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

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

$$\begin{aligned} \implies p(\mathbf{z}^l|K^l) &= \mathcal{N}\Big(\mathbf{0}, G\big(K^l\big) \otimes \mathbf{I}_{N_{l+1} \times N_{l+1}}\Big) \\ &= \mathcal{GP}\big(\mathbf{z}^l; 0, G(K^l)\big) \end{aligned}$$

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

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

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

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

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

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

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

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

- using CLT, pick the most appropriate definition:

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

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

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

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

5.4 putting in the overall objective function

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

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

6 Neural Tangent Kernel

6.1 The problem

- since Cost (or output layer) can be defined in convex function terms of post-activation last layer neurons $\phi(z^L(x))$, for example:

$$C = \|y - \phi(z^L(x))\|^2$$

  there must be a global minimal if we were to optimize it in term of $\phi(z^L(x))$

- however, current training regime:

  1. gradient descend are **not** optimized using $\frac{\delta C}{\delta \phi(z^L(x))}$
  2. but, it is computed through $\frac{\partial \delta C}{\partial \theta}$

- so it's unclear if a tiny step taken when $\theta(t) \to \theta(t+\epsilon)$ is to lead towards a negative gradient value in $\frac{\delta C}{\delta \phi(z^L(x))}$

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

- Under any training regime, there will be parameter dynamics (gradient flow) $\frac{\mathrm{d}\theta}{\mathrm{d}t}$
- what you hope: under **gradient descend** with infinitesimal step size (a.k.a. gradient flow)

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = -\frac{\partial C}{\partial \theta}$$

  functional gradient $\frac{\delta C}{\delta f(\theta)}$ is **negative all the time**!

- because C is convex functional of $f(\theta)$, and if gradient is negative all the time, it will eventually reach the global minima
- and **no!** this doesn't work all the time, it only occur under specific conditions listed below:

6.2.1 $\frac{\delta C}{\delta f(\theta)}$ under arbitrary infinitesimal step change $\theta \to \theta + \epsilon\eta$

So the **question** is, when θ undertakes infinitesimal step change in a **direction vector** η, i.e.,:

$$\theta \to \theta + \epsilon\eta$$

how does $\frac{\delta C}{\delta f(\theta)}$ change. Formally, we want to compute the following limit:

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

it is a mathematical traditional to write functional $C[f]$ is in square bracket

- Since $C[f]$ is a functional, we need to use **Riesz-Markov-Kakutani Representation Theorem**:

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

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

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

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

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

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

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

$$\begin{aligned}
&\implies \lim_{\epsilon \to 0} \frac{C\Big[\underbrace{f(\theta)}_{g} + \epsilon \underbrace{\eta \cdot \frac{\partial f(\theta)}{\partial \theta}}_{\phi} + O(\epsilon^2)\Big] - C\big[f(\theta)\big]}{\epsilon} \qquad \text{matching with RMK representation} \\
&= \int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x)\right)^\top \left(\eta \cdot \frac{\partial f(\theta)}{\partial \theta}\right) \mathrm{d}x \\
&= \sum_{d=1}^{|\theta|} \int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x)\right)^\top \left(\eta \cdot \frac{\partial f(\theta)}{\partial \theta_d}\right) \mathrm{d}x \\
&= \sum_{d=1}^{|\theta|} \int_{\mathbf{X}} \sum_{i=1}^{N} \left(\frac{\delta C}{\delta f(\theta)}(x)\right)_i \left(\eta \cdot \frac{\partial f(\theta)}{\partial \theta_d}\right)_i \mathrm{d}x \\
&= \sum_{d=1}^{|\theta|} \sum_{i=1}^{N} \eta \int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x)\right)_i \left(\frac{\partial f(\theta)}{\partial \theta_d}\right)_i \mathrm{d}x \qquad \text{change order of integral and sum}
\end{aligned}$$

6.2.2 $\frac{\delta C}{\delta f(\theta)}$ under gradient flow in gradient descend training

- above tells how much does C change if $\theta \to \theta + \epsilon\eta$
- since we can choose any direction η, we can equally (and meaningfully) choose a direction to be **gradient flow**, i.e.:

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

which correspond to the training regime used

- by substitution:

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

- if **gradient descent** training regime is used, then:

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

- substitution:

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

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

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

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

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

  cost will converge to a global optima.

- it is important to know the term **inside** the integral is actually **not** guaranteed to be positive.
- It is only become positive when the integrals are taken. To make it clear, we rewrite the following using simple notations:

$$\int_{x^{(p)}} \int_{x^{(q)}} \underbrace{\bar{f}(x^{(p)})^{\top} \Theta(x^{(p)}, x^{(q)}) \bar{f}(x^{(q)})}$$

- for a specific term

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

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

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

  since g is non-specific to value in $x^{(p)}$ and $x^{(q)}$, as both are used.

- therefore, $K(x^{(p)}, x^{(q)})$ is positive definitely **condition** on the fact that $x^{(p)}$ and $x^{(q)}$ are distributed from the same distribution, e.g., p^{in}.
- formally, we can write it as:

$$\begin{aligned}
& K \text{ is positive definite with respect to } \|\cdot\|_{p^{\text{in}}} \quad \text{if} \\
& \mathbb{E}_{x,x' \sim p^{\text{in}}}\big[f(x)^{\top} f(x')\big] > 0 \implies \mathbb{E}_{x,x' \sim p^{\text{in}}}\big[f(x)^{\top} K f(x')\big] > 0
\end{aligned}$$

6.2.4 What does NTK paper aims to prove

- NTK paper is all about, Proving under gradient descend training regime/gradient field

  and with the following conditions:

  1. $f(\theta)$ is a neural network

2. θ has appropriate Gaussian initialization is applied
3. having $N_1, \dots N_L \to \infty$:

Then,

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

Consequently, leading $\lim_{\epsilon \to 0} \frac{C[f(\theta + \epsilon \frac{\partial \theta}{\partial t})] - C[f(\theta)]}{\epsilon}$ to stay negative, i.e., cost always going down in a convex function, so it will eventually reach global minimum.

6.3 NTK in Neural Networks

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

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

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

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

- note that size of Θ^l is $N_{l+1} \times N_{l+1}$, it has nothing to do with $|\theta|$ (it is used in the sum)
- loosely speaking:
  1. NTK studies "pseudo-correlations" between a pair of output (k, k') of a vector function z^l by summing over their derivatives over all parameters from two data $x^{(p)}$ and $x^{(q)}$
     (derivative correlations between **function's output**)
  2. which is different to fisher information matrix:
     $$\mathbf{I}_{i,j} = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \log f(X;\theta) \right) \left(\frac{\partial}{\partial \theta_j} \log f(X;\theta) \right) \right]$$
     where FIM studies correlation between log derivative of pair of parameters (θ_i, θ_j) from a scalar function f.
     (derivative correlations between **function's parameters**)
- note that symbol here $\otimes$ above is outer product as oppose to kronecker product everywhere else in this tutorial. But the two are related:

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

7 NTK at initialization

given an input x, we show the following matrix form for correlations of their output:

$$\begin{bmatrix} \frac{1}{\sqrt{N_l}} W^l_{1,1}\phi\big(z_1^{l-1}(x)\big) + \sigma_b b_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{k,1}\phi\big(z_1^{l-1}(x)\big) + \sigma_b b_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{N_{l+1},1}\phi\big(z_j^{l-1}(x)\big) + \sigma_b b_{N_{l+1}} \end{bmatrix} + \cdots + \begin{bmatrix} \frac{1}{\sqrt{N_l}} W^l_{1,N_l}\phi\big(z_1^{l-1}(x)\big) + \sigma_b b^l_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{k,N_l}\phi\big(z_1^{l-1}(x)\big) + \sigma_b b^l_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{N_{l+1},N_l}\phi\big(z_j^{l-1}(x)\big) + \sigma_b b^l_{N_{l+1}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{1,j}\phi\big(z_j^{l-1}(x)\big) + \sigma_b b^l_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{k,j}\phi\big(z_j^{l-1}(x)\big) + \sigma_b b^l_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{N_{l+1},j}\phi\big(z_j^{l-1}(x)\big) + \sigma_b b^l_{N_{l+1}} \end{bmatrix} = \begin{bmatrix} z_1^l(x) \\ \vdots \\ z_k^l(x) \\ \vdots \\ z^l_{N_{l+1}}(x) \end{bmatrix}$$

7.1 when $l = 1$

From the above:

$$\begin{bmatrix} \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W^1_{1,j}\phi\big(x_1\big) + \sigma_b b^1_1 \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W^1_{k,j}\phi\big(x_k\big) + \sigma_b b^1_k \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W^1_{N_1,j}\phi\big(x_{N_1}\big) + \sigma_b b^1_{N_1} \end{bmatrix} = \begin{bmatrix} z_1^1(x) \\ \vdots \\ z_k^1(x) \\ \vdots \\ z^1_{N_1}(x) \end{bmatrix}$$

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

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

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

$$
\begin{aligned}
\Theta^1_{k,k'}(x^{(p)},x^{(q)}) &= \sum_{d=1}^{|\theta^1|} \frac{\partial F^1_k(x^{(p)})}{\partial \theta^1_d}\frac{\partial F^1_{k'}(x^{(q)})}{\partial \theta^1_d} \qquad \theta^1 = \{W^1, b^1\} \\
&= \sum_{d=1}^{|W^1|} \frac{\partial F^1_k(x^{(p)})}{\partial W^1_d}\frac{\partial F^1_{k'}(x^{(q)})}{\partial W^1_d} + \sum_{d=1}^{|b^1|} \frac{\partial F^1_k(x^{(p)})}{\partial b^1_d}\frac{\partial F^1_{k'}(x^{(q)})}{\partial b^1_d} \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{d_{\text{in}}} \frac{\partial z^1_k(x^{(p)})}{\partial W_{i,j}}\frac{\partial z^1_{k'}(x^{(q)})}{\partial W_{i,j}} + \sum_{i=1}^{N_1} \frac{\partial z^1_k(x^{(p)})}{\partial b_i}\frac{\partial z^1_{k'}(x^{(q)})}{\partial b_i} \\
&= \sum_{i=1}^{N_1}\sum_{j=1}^{d_{\text{in}}} \frac{1}{\sqrt{d_{\text{in}}}}x_i^{(p)}\delta_{i,k'}\ \frac{1}{\sqrt{d_{\text{in}}}}x_i^{(q)}\delta_{i,k} + \sum_{i=1}^{N_1}\sigma_b\delta_{i,k}\ \sigma_b\delta_{i,k'} \qquad \text{only one } i\in\{1,\dots N_1\} \text{ in outer sum remain} \\
&= \sum_{j=1}^{d_{\text{in}}} \frac{1}{d_{\text{in}}}x_i^{(p)}x_i^{(q)}\delta^2_{k,k'} + \sigma_b^2\delta_{k,k'} \qquad \delta_{i,k'}\delta_{i,k} = \delta_{k,k'} \\
&= \frac{1}{d_{\text{in}}}x^{(p)\top}x^{(q)}\delta_{k,k'} + \sigma_b^2\delta_{k,k'} \\
&= \underbrace{\left(\frac{1}{d_{\text{in}}}x^{(p)\top}x^{(q)} + \sigma_b^2\right)}_{K^1}\delta_{k,k'} \\
&\equiv K^1(x^{(p)},x^{(q)})\delta_{k,k'} \qquad \text{conform to notation in GP for NN section}
\end{aligned}
$$

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

$$
\implies \Theta^1(x^{(p)},x^{(q)}) = \begin{bmatrix}
G(K^1)(x^{(p)},x^{(q)}) & \dots & 0 & \dots & 0 \\
0 & K^1(x^{(p)},x^{(q)}) & 0 & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & \dots & \dots & K^1(x^{(p)},x^{(q)}) & 0 \\
0 & 0 & 0 & 0 & K^1(x^{(p)},x^{(q)})
\end{bmatrix}
$$

repeating diagonal with $K^1(x^{(p)},x^{(q)})$

$$
= \underbrace{K^1(x^{(p)},x^{(q)})}_{\text{scalar}}\mathbf{I}_{N_{l+1}\times N_{l+1}}
$$

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

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

7.2 when $l > 1$

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

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

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

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

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

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

- substitution:

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

- apply CLT, we know that:

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

- we have seen previously Eq. (1):

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

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

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

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

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

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

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

and compare that with for $l = 1$:

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

it's the same if we replace

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

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

8 NTK during training

Looking at training the Last-layer:

8.1 single data x under mean-square error

- for a single data x in $\mathcal{R}^{d_{\text{in}}}$, and its associated label y, imagine last layer parameter is:

$$\theta^{L+1} = (W^{L+1}, b^{L+1})$$

then, objective is:

$$\begin{aligned} C &= \frac{1}{2}\|f(x) - y\|_2^2 \\ &= \frac{1}{2}\left\| \left(\frac{\sigma_w}{\sqrt{N_l}} W^{L+1} \phi\big(z^L(x)\big) + \sigma_b b^{L+1} \right) - y \right\|_2^2 \end{aligned}$$

- the above defines last layer as if it is the linear layers in NN, non-standard part is to re-parameterization $\frac{\sigma_w}{\sqrt{N_l}}$ and σ_b are added to allow:

$$W^{L+1} \sim \mathcal{N}(0,1) \quad \text{and } b^{L+1} \sim \mathcal{N}(0,1)$$

then, the above is written as:

$$\begin{aligned} C &= \frac{1}{2}\left\| \left(\underbrace{\begin{bmatrix} W^{L+1} & b^{L+1} \end{bmatrix}}_{\theta^{L+1}} \underbrace{\begin{bmatrix} \frac{\sigma_w}{\sqrt{N_l}} \phi\big(z^L(x)\big) & \sigma_b \end{bmatrix}^\top}_{\bar{a}(x)} \right) - y \right\|_2^2 \\ &= \frac{1}{2}\left\| \big(\bar{a}(x)^\top \theta^{L+1}\big) - y \right\|_2^2 \\ &= \frac{1}{2}\left\| \hat{y}_t(x) - y \right\|_2^2 \\ \implies \frac{\partial C}{\partial \theta^{L+1}} &= \bar{a}(x)^\top \big(\bar{a}(x)\theta^{L+1} - y\big) \end{aligned}$$

- reason to write this way is to express derivative in θ^{L+1} jointly, instead of writing W^{L+1} and b^{L+1} separately

8.2 entire dataset $\mathcal{X}$:

8.2.1 softmax:

$$\hat{y}_t(\mathcal{X}) = \begin{bmatrix} \hat{y}_t^1(x_1) \\ \vdots \\ \hat{y}_t^{N^{L+1}}(x_1) \\ \vdots \\ \hat{y}_t^1(x_k) \\ \vdots \\ \hat{y}_t^{N^{L+1}}(x_k) \\ \vdots \\ \hat{y}_t^1(x_{|D|}) \\ \vdots \\ \hat{y}_t^{N^{L+1}}(x_{|D|}) \end{bmatrix} = \mathrm{vec}\big([\hat{y}_t^i(x)]_{x\in\mathcal{X}}\big) \in \mathcal{R}^{N^{L+1}\times|D|\times 1}$$

8.2.2 mean-square error:

we focus on MSE:

$$\begin{aligned} C &= \frac{1}{2}\left\|\big(\bar{a}(\mathcal{X})^\top \theta^{L+1}\big) - \mathcal{Y}\right\|_2^2 \\ &= \frac{1}{2}\left\|\hat{y}_t(\mathcal{X}) - \mathcal{Y}\right\|_2^2 \\ \implies \frac{\partial C}{\partial \theta^{L+1}} &= \bar{a}(\mathcal{X})^\top\big(\bar{a}(\mathcal{X})\theta^{L+1} - \mathcal{Y}\big) \end{aligned}$$

$$\hat{y}_t(\mathcal{X}) = \begin{bmatrix} \hat{y}_t(x_1) \\ \vdots \\ \hat{y}_t(x_k) \\ \vdots \\ \hat{y}_t(x_{|D|}) \end{bmatrix} = \mathrm{vec}\big([\hat{y}_t(x)]_{x\in\mathcal{X}}\big) \in \mathcal{R}^{|D|} \times 1$$

$$\implies \frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta} = \begin{bmatrix} \frac{\mathrm{d}\hat{y}_t(x_1)}{\mathrm{d}\theta_1} & \cdots & \frac{\mathrm{d}\hat{y}_t(x_1)}{\mathrm{d}\theta_{|\theta|}} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}\hat{y}_t(x_k)}{\mathrm{d}\theta_1} & \cdots & \frac{\mathrm{d}\hat{y}_t(x_k)}{\mathrm{d}\theta_{|\theta|}} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}\hat{y}_t(x_{|D})}{\mathrm{d}\theta_1} & \cdots & \frac{\mathrm{d}\hat{y}_t(x_{|D})}{\mathrm{d}\theta_{|\theta|}} \end{bmatrix}$$

$$\implies \hat{\Theta}(\mathcal{X}, \mathcal{X}) = \sum_{i=1}^{|\theta|} \frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta_i} \frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta_i}^{\top} \qquad \text{empirical Tangent Kernel}$$

$$= \begin{bmatrix} \sum_{i=1}^{|\theta|} \frac{\partial \hat{y}_t(x_1)}{\partial \theta_i} \frac{\partial \hat{y}_t(x_1)}{\partial \theta_i} & \cdots & \sum_{i=1}^{|\theta|} \frac{\partial \hat{y}_t(x_1)}{\partial \theta_i} \frac{\partial \hat{y}_t(x_{|D|})}{\partial \theta_i} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{|\theta|} \frac{\partial \hat{y}_t(x_1)}{\partial \theta_i} \frac{\partial \hat{y}_t(x_k)}{\partial \theta_k} & \cdots & \sum_{i=1}^{|\theta|} \frac{\partial \hat{y}_t(x_k)}{\partial \theta_i} \frac{\partial \hat{y}_t(x_{|D|})}{\partial \theta_i} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{|\theta|} \frac{\partial \hat{y}_t(x_1)}{\partial \theta_i} \frac{\partial \hat{y}_t(x_{|D|})}{\partial \theta_k} & \cdots & \sum_{i=1}^{|\theta|} \frac{\partial \hat{y}_t(x_{|D|})}{\partial \theta_i} \frac{\partial \hat{y}_t(x_{|D|})}{\partial \theta_i} \end{bmatrix}$$

8.3 Sketch of Proof

- We are interested to study the behavior of $\hat{y}_t(x)$ for a singular data x as θ_t evolves
- the expression $\hat{y}_t(x)$ can be misleading: it should be written instead as $\hat{y}_t(x|\mathcal{X})$ as it depends on training dataset $(\mathcal{X}, \mathcal{Y})$, and θ as well.
- Note that we can interchangeably write:

$$\hat{y}_t(\mathcal{X}, \theta) \equiv \hat{y}_t(\mathcal{X}) \equiv \hat{y}(\mathcal{X}, \theta_t)$$

Also, as we do not have expression for θ_t, we must start from $\frac{\mathrm{d}\theta}{\mathrm{d}t}$ using gradient descent:

1. find expression for $\hat{y}_t(\mathcal{X}, \theta)$: it has two versions:
   using gradient descend:

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = -\eta \left(\frac{\partial \hat{y}_t(\mathcal{X}, \theta)}{\partial \theta} \right)^{\top} \frac{\partial C}{\partial \hat{y}_t(\mathcal{X}, \theta)}$$

   then,

   version 1: assume $\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}$ is time-invariant

   (a) from $\frac{\mathrm{d}\theta}{\mathrm{d}t}$, use ODE to obtain θ_t

   (b) then, obtain $\hat{y}_t(\mathcal{X}, \theta) = \big(\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\big)^{\top} \theta_t$

version 2:

(a) obtain expression for:

$$\frac{\mathrm{d}\hat{y}}{\mathrm{d}t} = \left(\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\right)^{\top} \frac{\mathrm{d}\theta}{\mathrm{d}t}$$

(b) then, use ODE to obtain $\hat{y}_t(\mathcal{X}, \theta)$

2. now we have expression of $\hat{y}_t(\mathcal{X}, \theta)$,

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \frac{\mathrm{d}\theta}{\mathrm{d}t} = -\eta \left(\frac{\partial \hat{y}_t(\mathcal{X}, \theta)}{\partial \theta}\right)^{\top} \frac{\partial C}{\partial \hat{y}_t(\mathcal{X}, \theta)}$$

3. from $\frac{\mathrm{d}\theta}{\mathrm{d}t}$, use ODE or straight integration to obtain:

$$\theta_t, \quad \text{or} \quad \omega_t = \theta_t - \theta_0$$

4. Finally obtain how change of parameter contribute to last layer of single data in linear (in terms of θ_t) model:

$$\hat{y}(x, \theta_t) = \hat{y}(x, \theta_0) + \left.\frac{\mathrm{d}\hat{y}(x, \theta_t)}{\mathrm{d}t}\right|_{\theta \to \theta_0} \omega_t$$

think above as: instead of taking Euclidean step, i.e., $\theta_t = \theta_0 + h$, it now follows gradient descent path:

$$\theta_t = \theta_0 + \omega_t$$

- dependency chain is:

$$(\mathcal{X}, \mathcal{Y}) \quad \to \quad \frac{\mathrm{d}\theta}{\mathrm{d}t} \quad \to \quad (\theta_t, \omega_t) \quad \to \quad f_t^{\mathrm{lin}}(x, \theta)$$

- note the following:
  - We need all training data pairs $(\mathcal{X}, \mathcal{Y})$ to determine the change in θ
  - so $\frac{\mathrm{d}\theta}{\mathrm{d}t}$ is a function of $(\mathcal{X}, \mathcal{Y})$, more specifically, $\hat{y}(\mathcal{X}, \theta)$ and $\mathcal{Y}$, this is in Section (8.5)
  - then we can work out how $\frac{\mathrm{d}\theta}{\mathrm{d}t}$ may impact $\hat{y}_t(x)$
  - simply obtained expression of $\hat{y}_t(\mathcal{X})$ won't give you expression for $\hat{y}_t(x)$

8.4 General linear ODE solution

We need basic tools on ODE solution:

- looking at the equation, treating everything in 1-d:

$$
\begin{aligned}
\dot{x} &= Ax + b \\
\implies \frac{\dot{x}}{Ax+b} &= 1 \\
\implies \frac{\dot{x}}{x+\frac{b}{A}} &= A \\
\implies \frac{\mathrm{d}}{\mathrm{d}t}\log\left(x+\frac{b}{A}\right) &= A \qquad \text{easy to see: } \frac{\mathrm{d}}{\mathrm{d}t}\log\left(x+\frac{b}{A}\right) = \frac{\frac{\mathrm{d}x}{\mathrm{d}t}}{x+\frac{b}{A}} \\
\implies \int_t \frac{\mathrm{d}}{\mathrm{d}t}\log\left(x+\frac{b}{A}\right)\mathrm{d}t &= \int_t A\mathrm{d}t \\
\log\left(x+\frac{b}{A}\right) &= At + h \\
x + \frac{b}{A} &= \exp\left(At + {\color{red}h}\right) \\
\implies x(t) &= -\frac{b}{A} + {\color{red}C}\exp\left(At\right)
\end{aligned}
$$

- when things are in multi-dimensions:

$$
x(t) = -A^{-1}b + C\exp\left(At\right) \quad b \text{ is column vector}
$$

let $t = 0$:

$$
\begin{aligned}
x(0) &= -A^{-1}b + C \\
\implies C &= \left(x(0) + A^{-1}b\right)
\end{aligned}
$$

substitute in C:

$$
x(t) = -A^{-1}b + \left(x(0) + A^{-1}b\right)\exp\left(At\right) \tag{2}
$$

8.5 find expression for ${\color{red}\hat{y}_t(\mathcal{X}, \theta)}$:

- we are using **version 1** (link 1) to get expression for $\theta^{L+1}(t)$ from $\frac{\mathrm{d}\theta^{L+1}(t)}{\mathrm{d}t}$:
- in MSE context:

$$
\begin{aligned}
C &= \frac{1}{2}\left\|\left(\bar{a}(\mathcal{X})^\top\theta^{L+1}\right) - \mathcal{Y}\right\|_2^2 \\
&= \frac{1}{2}\left\|\hat{y}_t(\mathcal{X}) - \mathcal{Y}\right\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}\theta^{L+1}(t)}{\mathrm{d}t} &= -\eta\left(\frac{\partial\hat{y}_t(\mathcal{X}, \theta^{L+1})}{\partial\theta^{L+1}}\right)^\top \frac{\partial C}{\partial\hat{y}_t(\mathcal{X}, \theta^{L+1})} \\
&= -\eta\bar{a}(\mathcal{X})^\top\left(\bar{a}(\mathcal{X})\theta^{L-1} - \mathcal{Y}\right) \\
&= \underbrace{-\eta\bar{a}(\mathcal{X})^\top\bar{a}(\mathcal{X})}_{A}\theta^{L-1}(t) + \underbrace{\eta\bar{a}(\mathcal{X})^\top\mathcal{Y}}_{b}
\end{aligned}
$$

- so by substitution, using Eq. (2), we have:

$$
\begin{aligned}
A &= -\eta \bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X}) \\
\implies A^{-1} &= \frac{-1}{\eta} \left(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\right)^{-1} \\
b &= \eta \bar{a}(\mathcal{X})^\top \mathcal{Y}
\end{aligned}
$$

$$
\begin{aligned}
\theta^{L+1}(t) &= -\frac{-1}{\eta}\Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\eta \bar{a}(\mathcal{X})^\top \mathcal{Y}\Big) \\
&\quad + \Big[\theta^{L+1}(0) + \frac{-1}{\eta}\Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\eta \bar{a}(\mathcal{X})^\top \mathcal{Y}\Big)\Big] \ \exp\Big(-\eta \bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X}) t\Big) \\
&= \Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\bar{a}(\mathcal{X})^\top \mathcal{Y}\Big) \\
&\quad + \Big[\theta^{L+1}(0) - \Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\bar{a}(\mathcal{X})^\top \mathcal{Y}\Big)\Big] \ \exp\Big(-\eta \bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X}) t\Big) \\
\theta^{L+1}(t)\bar{a}(\mathcal{X}) &= \Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\bar{a}(\mathcal{X})^\top \mathcal{Y}\, \bar{a}(\mathcal{X})\Big) \\
&\quad + \Big[\theta^{L+1}(0)\bar{a}(\mathcal{X}) - \Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\bar{a}(\mathcal{X})^\top \mathcal{Y}\bar{a}(\mathcal{X})\Big)\Big] \ \exp\Big(-\eta \bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X}) t\Big) \\
\hat{y}_t(\mathcal{X}) &= \mathcal{Y} + \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big) \ \exp\big(-\eta \bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X}) t\big) \\
&= \underbrace{\mathcal{Y}} + \hat{y}_0(\mathcal{X}) \exp\big(-\eta \bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X}) t\big) - \underbrace{\mathcal{Y} \exp\big(-\eta \bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X}) t\big)} \\
&= \big(\mathbf{I} - \exp\big(-\eta \bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X}) t\big)\big)\mathcal{Y} + \hat{y}_0(\mathcal{X}) \exp\big(-\eta \bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X}) t\big) \\
&= \big(\mathbf{I} - \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X}) t\big)\big)\mathcal{Y} + \hat{y}_0(\mathcal{X}) \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X}) t\big)
\end{aligned}
$$

- no close-form solution when $\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})$ is also a function of time, but we can approximate it using linearlized model:
- however since we are dealing with last layer, $\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})$ is constant, therefore:

$$
\begin{aligned}
\hat{y}_t(\mathcal{X}) &= \mathcal{Y} + \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big) \ \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X}) t\big) \\
\frac{\mathrm{d}\hat{y}_t(\mathcal{X})}{\mathrm{d}t} &= -\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big) \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X}) t\big) \\
&= -\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\big(\hat{y}_t(\mathcal{X}) - \mathcal{Y}\big)
\end{aligned} \tag{3}
$$

8.6 computing ω_t

- using gradient descend, but substituting $\hat{y}(\mathcal{X})$:

$$
\begin{aligned}
\frac{\mathrm{d}\theta}{\mathrm{d}t} &= -\eta \frac{\partial C}{\partial \theta} \\
&= -\eta \Big(\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\Big)^\top \frac{\partial C}{\partial \hat{y}_t(\mathcal{X})} \\
&= -\eta \Big(\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\Big)^\top \big(\hat{y}_t(\mathcal{X}) - \mathcal{Y}\big) \\
&= -\eta \Big(\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\Big)^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big) \quad \text{using equation (3)}
\end{aligned}
$$

- to work out $\theta(t) \equiv \theta_t \equiv \omega_t$:

$$
\begin{aligned}
\theta(t) &= \int_{\tau=0}^{t} \frac{\mathrm{d}\theta}{\mathrm{d}\tau}\mathrm{d}\tau \\
&= \int_{\tau=0}^{t} -\eta \Big(\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\Big)^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\tau\big)\mathrm{d}\tau \\
&= -\eta \Big(\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\Big)^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \underbrace{\int_{\tau=0}^{t} \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\tau\big)\mathrm{d}\tau}
\end{aligned}
$$

- looking at:

$$
\begin{aligned}
&\int_{\tau=0}^{t} \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\tau\big)\mathrm{d}\tau \\
&= \Big[-\frac{1}{\eta}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\tau\big)\Big]_{\tau=0}^{t} \\
&= -\frac{1}{\eta}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big) + \frac{1}{\eta}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1} \\
&= \underbrace{\frac{1}{\eta}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big)}
\end{aligned}
$$

- Finally $\theta(t) \equiv w_t$:

$$
\begin{aligned}
\theta(t) &= -\eta \Big(\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\Big)^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \frac{1}{\eta}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\Big) \\
&= -\Big(\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\Big)^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big)
\end{aligned}
$$

8.7 Compute $f^{\text{lin}}(x, \theta_t)$

$$
\begin{aligned}
f_t^{\text{lin}}(x) &\equiv f_0(x) + \underbrace{\frac{\partial f(x,\theta)}{\partial \theta}\Big|_{\theta\to\theta_0}}_{\text{constant in } t} \underbrace{\big(\theta_t - \theta_0\big)}_{\omega_t} \\
&= f_0(x) + \frac{\partial f(x,\theta)}{\partial \theta}\Big|_{\theta\to\theta_0}\ \omega_t \qquad \text{note } \omega \text{ refer to change, irrespective of starting position } \theta_0
\end{aligned}
$$

substitute into $f_t^{\text{lin}}(x) \equiv \hat{y}_t(x)$, we have:

$$
\begin{aligned}
\hat{y}(x,\theta_t) &= \hat{y}(x,\theta_0) + \frac{\partial \hat{y}(x,\theta)}{\partial \theta}\Big|_{\theta\to\theta_0} \omega_t \\
&= \hat{y}(x,\theta_0) + \bar{a}(x)\Big(-\frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\Big)^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big) \\
&= \hat{y}(x,\theta_0) - \bar{a}(x)\bar{a}(\mathcal{X})^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big) \\
&= \hat{y}(x,\theta_0) - \bar{a}(x)\bar{a}(\mathcal{X})^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big) \\
&= \hat{y}(x,\theta_0) - \hat{\mathcal{K}}(x,\mathcal{X})\big(\hat{y}(\mathcal{X},\theta_0) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big)
\end{aligned}
$$

8.8 expectation and variance

- **mean**

$$
\begin{aligned}
\mathbb{E}\big[\hat{y}(x,\theta_t)\big] &= \mathbb{E}\Big[\hat{y}(x,\theta_0) - \hat{\mathcal{K}}(x,\mathcal{X})\big(\hat{y}(\mathcal{X},\theta_0) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big)\Big] \\
&= \underbrace{\mathbb{E}\big[\hat{y}(x,\theta_0)\big]}_{=0} - \underbrace{\mathbb{E}\big[\hat{y}(\mathcal{X},\theta_0)\big]}_{=0}\ \mathbb{E}\Big[\hat{\mathcal{K}}(x,\mathcal{X})\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big)\Big] \\
&\qquad + \mathbb{E}\Big[\hat{\mathcal{K}}(x,\mathcal{X})\mathcal{Y}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big)\Big] \\
&= \mathbb{E}\Big[\hat{\mathcal{K}}(x,\mathcal{X})\mathcal{Y}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big)\Big] \\
&= \mathcal{K}(x,\mathcal{X})\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{Y} \quad \text{deterministic in infinite width}
\end{aligned}
$$

- **variance**

$$
\begin{aligned}
&\hat{y}(x,\theta_t) - \mathbb{E}\big[\hat{y}(x,\theta_t)\big] \qquad \text{let infinite width} \\
=&\hat{y}(x,\theta_0) - \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\big(\hat{y}(\mathcal{X},\theta_0) - \mathcal{Y}\big) \\
&\qquad - \mathcal{K}(x,\mathcal{X})\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{Y} \\
=&\hat{y}(x,\theta_0) - \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\ \hat{y}(\mathcal{X},\theta_0) \\
&\qquad + \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{Y} \\
&\qquad - \mathcal{K}(x,\mathcal{X})\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{Y} \\
=&\hat{y}(x,\theta_0) - \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\ \hat{y}(\mathcal{X},\theta_0)
\end{aligned}
$$

then:

$$
\begin{aligned}
&\mathrm{Var}\big[\hat{y}(x,\theta_t)\big] \\
&= \mathbb{E}\Big[\Big(\hat{y}(x,\theta_t) - \mathbb{E}\big[\hat{y}(x,\theta_t)\big]\Big)^{\top}\ \Big(\hat{y}(x,\theta_t) - \mathbb{E}\big[\hat{y}(x,\theta_t)\big]\Big)\Big] \\
&= \mathbb{E}\Big[\Big(\hat{y}(x,\theta_0) - \underbrace{\mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\ \hat{y}(\mathcal{X},\theta_0)}\Big) \\
&\qquad\qquad \Big(\hat{y}(x,\theta_0) - \underbrace{\mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\hat{y}(\mathcal{X},\theta_0)}\Big)^{\top}\Big]
\end{aligned}
$$

knowing that when $t = 0$:

$$
\begin{aligned}
\mathrm{Cov}\Big[\hat{y}(x,\theta_0)\ ,\ \hat{y}(\mathcal{X},\theta_0)\Big] &= \mathbb{E}\Big[\hat{y}(x,\theta_0)\hat{y}(\mathcal{X},\theta_0)^{\top}\Big] = \mathcal{K}(x,\mathcal{X}) \\
\mathrm{Cov}\Big[\hat{y}(\mathcal{X},\theta_0)\ ,\ \hat{y}(x,\theta_0)\Big] &= \mathbb{E}\Big[\hat{y}(\mathcal{X},\theta_0)\hat{y}(x,\theta_0)^{\top}\Big] = \mathcal{K}(\mathcal{X},x) \\
\mathrm{Var}\Big[\hat{y}(x,\theta_0)\ ,\ \hat{y}(x,\theta_0)\Big] &= \mathbb{E}\Big[\hat{y}(x,\theta_0)\hat{y}(x,\theta_0)^{\top}\Big] = \mathcal{K}(x,x) \\
&\Big(\mathcal{K}(x,\mathcal{X})\ \underbrace{\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)}_{\text{symmetric}}\ \hat{y}(\mathcal{X},\theta_0)\Big)\Big)^{\top} \\
&= \Big(\hat{y}(\mathcal{X},\theta_0)^{\top}\ \underbrace{\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)}_{\text{symmetric}}\mathcal{K}(\mathcal{X},x)\ \Big)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}\big[\hat{y}(x,\theta_t)\big] =& \mathbb{E}\big[\hat{y}(x,\theta_0)\ \hat{y}(x,\theta_0)^{\top}\big] \\
&- \mathbb{E}\Big[\hat{y}(x,\theta_0)\ \hat{y}(\mathcal{X},\theta_0)^{\top}\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\ \mathcal{K}(\mathcal{X},x)\Big] \\
&- \mathbb{E}\Big[\mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\hat{y}(\mathcal{X},\theta_0)\ \hat{y}(x,\theta_0)^{\top}\Big] \\
&+ \mathbb{E}\Big[\mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\ \hat{y}(\mathcal{X},\theta_0)\hat{y}(\mathcal{X},\theta_0)^{\top}\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{K}(\mathcal{X},x \\
=& \mathcal{K}(x,x) \\
&- \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\ \mathcal{K}(\mathcal{X},x) \\
&- \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\ \mathcal{K}(\mathcal{X},x) \\
&+ \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{K}(\mathcal{X},\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{K}(\mathcal{X},x) \\
=& \mathcal{K}(x,x) - 2\mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\ \mathcal{K}(\mathcal{X},x) \\
&+ \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{K}(\mathcal{X},x) \\
=& \mathcal{K}(x,x) - \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(2\mathbf{I} - \underbrace{2\exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)}\Big)\ \mathcal{K}(\mathcal{X},x) \\
&+ \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \underbrace{2\exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)} + \exp\big(-2\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{K}(\mathcal{X},x)
\end{aligned}
$$

terms outside of the red bits are the same:

$$
\mathrm{Var}\big[\hat{y}(x,\theta_t)\big] = \mathcal{K}(x,x) - \mathcal{K}(x,\mathcal{X})\ \mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-2\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{K}(\mathcal{X},x)
$$

9 Infinite width networks are linearized networks

for every $x \in \mathcal{R}^{N_0}$ with $\|x\|_2 \leq 1$ with probablity close to 1 over random initialization:

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

check relevant publications for the proof