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

Richard Xu

January 8, 2021

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

Alexander G de G Matthews, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Sample-then-optimize posterior sampling for Bayesian linear models. In NeurIPS Workshop on Advances in Approximate Bayesian Inference, 2017.

`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:

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

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 $\mathcal{X} \equiv \{x_1, \dots x_k\}$

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

2.1 marginal and conditional marginal under noisy output

- in a regression setting:

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

- 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:

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

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 sample and optimize

in words can be summarized as follows:

given a linear equation with non-empty null-space, and some prior distribution $w \sim \mathcal{N}()$ defined over $w \in \mathcal{W}$. Since null-space is non-empty, there exist a "subset" of $\mathcal{W}$, namely $w^\star$ satisfy linear equation, and forming a new posterior distribution:

$$w^\star \sim p\big(w \mid \|Aw = y\|_2^2\big) \tag{2}$$

We intend to draw a sample $w^\star$ from this posterior.

sample-then-optimize allows us to sample $w^0 \sim \mathcal{N}()$, and optimize using gradient descend to obtain $w^\star$. The following proves they indeed have the same probability as sampling from posterior directly

3.0.1 proof

$$\begin{aligned} w &\sim \mathcal{N}(0, \lambda I) \\ y &= Aw \quad A \in \mathcal{R}^{N \times M} \end{aligned} \tag{3}$$

when $N < M$, there is a family of solutions and this can be understood as to see the following:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,M-1} & A_{1,M} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_{N,1} & A_{N,2} & \dots & A_{N,M-1} & A_{N,M} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_N \\ w_{N+1} \\ \vdots \\ w_M \end{bmatrix} \tag{4}$$

general solution is:

$$\begin{aligned} \hat{w} &= A^+ y + (I - A^+ A)\, t \qquad \text{for arbitrary } t \in \mathcal{R}^M \\ &= \hat{w}_{||} + \hat{w}_{\text{null}} \end{aligned} \tag{5}$$

simple verification would see that,

$$\begin{aligned} & A(\hat{w}_{||}) = y \\ \implies\ & A(\hat{w}_{||} + \hat{w}_{\text{null}}) = y \\ \text{because}\quad & A(\hat{w}_{||}) + \underbrace{A(\hat{w}_{\text{null}})}_{0} = y \end{aligned} \tag{6}$$

informally $\hat{w}_{||} = A^+ y$ gives the solution $\hat{w}$ for:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,M-1} & A_{1,M} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_{N,1} & A_{N,2} & \dots & A_{N,M-1} & A_{N,M} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_N \\ w_{N+1} \\ \vdots \\ w_M \end{bmatrix} \tag{7}$$

and $\hat{w}_{\text{null}} = (I - A^+ A)\, t$ to ensure $\hat{w}$ also simultaneously satisfy:

$$\begin{bmatrix}0\\ \vdots\\ 0\end{bmatrix} = \begin{bmatrix}0 & 0 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & 0 & 0\end{bmatrix} \begin{bmatrix}w_1\\ \vdots\\ w_N\\ w_{N+1}\\ \vdots\\ w_M\end{bmatrix} \tag{8}$$

pseudo-inverse:

$$\begin{aligned} AA^+A &= A \\ \implies A^+ &= A^\top(AA^\top)^{-1} \qquad A^+ \in \mathcal{R}^{M\times N} \quad (AA^\top) \in \mathcal{R}^{N\times N} \quad \text{full rank} \\ &= (A^\top A)^+A^\top \qquad A^\top A \in \mathcal{R}^{M\times M} \quad \text{not full rank} \\ \implies A^+ &= \lim_{\delta^2\to 0} A^\top\left(AA^\top + \delta^2 I\right)^{-1} \quad \text{limit definition of pseudo inverse} \end{aligned} \tag{9}$$

it is easy to show $(I - A^+A)$ is projection into null(A) because:

$$\begin{aligned} A\ (I - A^+\ A) &= A\ (I - A^\top(AA^\top)^{-1}\ A) \\ &= A - (AA^\top)(AA^\top)^{-1}A \\ &= \mathbf{0} \end{aligned} \tag{10}$$

since each of the columns of the matrix $(I - A^+A)$ span null space of A, then:
$\hat{w}_{\text{null}} = (I - A^+A)\ t$ is some linear combination of columns of $(I - A^+A)$ and hence:

$$\hat{w}_{\text{null}} = (I - A^+A)\ t\ \in\ \text{null}(A) \tag{11}$$

3.0.2 what happen to w during gradient descend?

suppose we have the objective function:

$$\begin{aligned} f(w) &= \|Aw - y\|_2^2 \\ \implies \frac{\partial f(w)}{\partial w} &= \frac{\partial (Aw-y)^\top(Aw-y)}{\partial w} \\ &= \frac{\partial (w^\top A^\top - y^\top)(Aw-y)}{\partial w} \\ &= \frac{\partial (w^\top A^\top Aw - y^\top Aw - w^\top A^\top y + y^\top y)}{\partial w} \\ &= 2A^\top Aw - 2A^\top y \\ \implies \triangle w &= -\eta\left(A^\top Aw - A^\top y\right) \end{aligned} \tag{12}$$

therefore, say at time $t = 0$ we let the initial point w^0:

$$w^0 \sim \mathcal{N}(0, \lambda I) \tag{13}$$

and break up w^0 into two components with axes in different sub-spaces:

$$w^0 = w^0_{\text{null}} + w^0_{||} \tag{14}$$

and apply the following transform in gradient descend:

$$\begin{aligned}\triangle w(0) &\propto A^\top A w^0 - A^\top y \\ &= A^\top A\big(w^0_{\text{null}} + w^0_{||}\big) - A^\top y \\ &= A^\top A w^0_{\text{null}} + A^\top A w^0_{||} - A^\top y\end{aligned} \tag{15}$$

because: $g^\top A^\top f = f^\top A g = 0 \qquad g \in \mathcal{R}^M, \quad f \in \mathcal{R}^N \quad g \in \text{null}(A)$

$$\begin{aligned}\triangle w(0) &\propto \underbrace{A}_{f}{}^\top A \underbrace{w^0_{\text{null}}}_{g} + A^\top_f A w^0_{||} - A^\top y \\ &= 0 + A^\top A w^0_{||} - A^\top y\end{aligned} \tag{16}$$

the above means that there is no "movement" for w_{null} component, $\forall t$ in null(A). only movement occur in the $w_{||}$ subspace, assume gradient descend converge eventually. So at the end of gradient descend at time $t = \tau$:

$$\begin{aligned}&\triangle w(\tau) = 0 \\ \implies\ & A^\top A w^\tau_{||} - A^\top y = 0 \\ \implies\ & A^\top A w^\tau_{||} = A^\top y \\ \implies\ & w^\tau_{||} = (A^\top A)^+ A^\top y \\ & = A^+ y\end{aligned} \tag{17}$$

now we look at the general solution in Eq.(5) and replacing arbitary t with starting point w^0 as a repeating applications of gradient descend:

$$\begin{aligned}w^\star &= w^\tau_{||} + w^\tau_{\text{null}} \\ &= \underbrace{A^+ y}_{\mu} + \underbrace{(I - A^+ A)}_{\sigma} w^0\end{aligned} \tag{18}$$

so having $w^0 \sim \mathcal{N}(0, \lambda I)$, w^* is just apply an Affine transform with the following Gaussian distribution:

$$w^\star \sim \mathcal{N}\big(A^+ y,\ \lambda(I - A^+ A)\big) \tag{19}$$

so moral of the story, instead of sample from $p(w|y)$ given the above, one can just sample w^0 from prior $p(w)$, and then optimize using gradient:

$$w^\star = \arg\min_w \|Aw - y\|_2^2 \tag{20}$$

3.0.3 apply to Gaussian Process

so if you want to sample $w^\star$ from the posterior:

$$\begin{aligned}\mu_\star(y^\star|x^\star, \mathcal{X}, \mathcal{Y}) &= \mathcal{K}(x^\star, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}\mathcal{Y} \\ \Sigma_\star(y^\star|x^\star, \mathcal{X}, \mathcal{Y}) &= \mathcal{K}(x^\star, x^\star) \ - \ \mathcal{K}(x^\star, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}\,\mathcal{K}(\mathcal{X}, x^\star)\end{aligned} \tag{21}$$

one may sample $\theta_0 \sim \mathcal{N}(\mathbf{0}, \mathcal{K})$, and then because of the assumption:

in MSE context of read-out only layer:

$$
\begin{aligned}
C &= \frac{1}{2}\big\|(\underbrace{\bar{a}(\mathcal{X})^\top}_{A}\underbrace{\theta^{L+1}}_{w}) - \mathcal{Y}\big\|_2^2 \\
&= \frac{1}{2}\big\|\hat{y}_t(\mathcal{X}) - \mathcal{Y}\big\|_2^2
\end{aligned}
\tag{22}
$$

the only difference, is instead of representing a prior for θ_0^{L+1}, we have a prior for $\hat{y}(x, \theta_0)$, which we know how to express (NNGP) and sample

and, instead of express posterior of $w^\star \equiv \theta^{L+1^\star} \equiv \theta_\infty$, we let:

$$
w^\star \equiv \hat{y}(x, \theta_\infty) \quad \text{since } \hat{y}(x, \theta_\infty) \text{ is linear in } \theta_{t=\infty} \tag{23}
$$

We can derive the expression of posterior for $\hat{y}(x, \theta_\infty)$, but we do not sample from it, and instead we use **sample-then-optimize**

4 Kernel methods

consider the equation:

$$\begin{aligned} y &= \phi(x)^\top \boldsymbol{w} \\ &= \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_m(x) \end{bmatrix}^\top \boldsymbol{w} \\ &= \begin{bmatrix} \phi_1(x) & \dots & \phi_m(x) \end{bmatrix} \boldsymbol{w} \end{aligned} \tag{24}$$

using definition:

$$\begin{aligned} \mathcal{Y} &= [y_1, \dots, y_n]^\top \\ \Phi &= [\phi(x_1), \dots, \phi(x_n)]^\top \\ &= \underbrace{\begin{bmatrix} \phi_1(x_1) & \dots & \phi_m(x_1) \\ \vdots & \vdots & \vdots \\ \phi_1(x_n) & \dots & \phi_m(x_n) \end{bmatrix}}_{n \times m} \end{aligned} \tag{25}$$

Ridge regression can be re-written as:

$$\begin{aligned} \boldsymbol{w}^\star &= \arg\min_{\boldsymbol{w}} \sum_{i=1}^{n} \left(y_i - \phi(x_i)^\top \boldsymbol{w}\right)^2 + \lambda \|\boldsymbol{w}\|_2^2 \\ &= \arg\min_{\boldsymbol{w}} \|\mathcal{Y} - \Phi \boldsymbol{w}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2 \end{aligned} \tag{26}$$

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

$$\boldsymbol{w}^\star = \big(\underbrace{\Phi^\top \Phi}_{m \times m} + \lambda I\big)^{-1} \Phi^\top \mathcal{Y} \tag{27}$$

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

$$\begin{aligned} y_{\boldsymbol{w}^\star}(x) &= \phi(x)^\top \boldsymbol{w}^\star \\ &= \phi(x)^\top \left(\Phi^\top \Phi + \lambda I\right)^{-1} \Phi^\top \mathcal{Y} \\ &= \underbrace{\phi(x)^\top \Phi^\top}_{1 \times n} \big(\underbrace{\Phi \Phi^\top}_{n \times n} + \lambda I\big)^{-1} \mathcal{Y} \\ &\quad \text{using identity } \left(\Phi^\top \Phi + \lambda I\right)^{-1} \Phi^\top = \Phi^\top \left(\Phi \Phi^\top + \lambda I\right)^{-1} \end{aligned} \tag{28}$$

4.1 Kernel trick

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

$$\begin{aligned} \left[\Phi \Phi^\top\right]_{i,j} &= \phi(x_i)^\top \phi(x_j) = \mathcal{K}(x_i, x_j) \\ \left[\phi(x)^\top \Phi^\top\right]_j &= \phi(x)^\top \phi(x_j) = \mathcal{K}(x, x_j) \end{aligned} \tag{29}$$

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

4.2 relationship with Neural Tangent Kernel

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

$$f_{\boldsymbol{w}}(x) \equiv f(\boldsymbol{w}, x) \approx f(w_0, x) + \underbrace{\nabla_w f(w_0, x)^\top}_{\phi(x)^\top}(w - w_0) + \dots \tag{30}$$

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.

4.3 relationship with gradient flow

This is a simplified version to Section[8]:

Gradient descend algorithm:

$$\begin{aligned} \theta_{t+1} &= \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_{t+1}) \\ \implies \frac{\theta_{t+1} - \theta_t}{\eta} &= -\nabla_\theta \mathcal{L}(\theta_{t+1}) \\ \implies \lim_{\eta \to 0} \frac{\theta_{t+1} - \theta_t}{\eta} &= \frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\nabla_\theta \mathcal{L}(\theta) \end{aligned} \tag{31}$$

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

$$\begin{aligned} \mathcal{L}(\theta) &= \frac{1}{2}\|\tilde{y}(\theta) - y\|_2^2 \\ \implies \nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \tilde{y}(\theta)\big(\tilde{y}(\theta) - y\big) \\ \implies \frac{\mathrm{d}\theta(t)}{\mathrm{d}t} &= -\nabla_\theta \tilde{y}(\theta)\big(\tilde{y}(\theta) - y\big) \end{aligned} \tag{32}$$

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

$$\begin{aligned} \frac{\mathrm{d}\tilde{y}(\theta_t)}{\mathrm{d}t} &= \frac{\partial \tilde{y}(\theta(t))}{\partial \theta(t)}^\top \frac{\mathrm{d}\theta(t)}{\mathrm{d}t} \\ &= \nabla_\theta \tilde{y}(\theta)\Big(- \nabla_\theta \tilde{y}(\theta)\big(\tilde{y}(\theta) - y\big)\Big) \\ &= -\underbrace{\nabla_\theta \tilde{y}(\theta)^\top \nabla_\theta \tilde{y}(\theta)}\big(\tilde{y}(\theta) - y\big) \\ &\approx -K(\theta_0)\big(\tilde{y}(\theta) - y\big) \end{aligned} \tag{33}$$

5 GP for Neural Network: Direct computation

5.1 neural network function

using parameters:

$$\theta \equiv \{W^L, b^L, \dots W^1, b^1\} \tag{34}$$

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

$$\begin{aligned} f_\theta(X) &= W^L \phi^L(X) + b^L \\ &= W^L \left(\phi^{L-1}(X) W^{L-1} + b^{L-1}\right) + b^L \\ &\dots \\ &= W^L \cdots (W^1 \phi^1(X) + b^1) + \dots) + b^L \end{aligned} \tag{35}$$

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

$$\begin{aligned} \phi^L(X) &\equiv \phi^L(X \,|\theta^1, \dots, \theta^{L-1}) \\ &\equiv \phi^L(X \,|W^1, b^1, \dots, W^{L-1}, b^{L-1}) \end{aligned} \tag{36}$$

5.2 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:

$$p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{\star\top} \end{bmatrix}\right) = \int \mathcal{N}\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} f_\theta(X) \\ f_\theta(x^\star) \end{bmatrix}, \sigma_\epsilon^2 I\right) \mathcal{N}(\theta|0, \sigma_\theta^2 I) \mathrm{d}\theta \tag{37}$$

The integral is **not** analytic!!

5.3 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:

$$p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) \xrightarrow[N^1, \dots, N^L \to \infty]{d} ? \tag{38}$$

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

$$p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) = ? \tag{39}$$

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

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

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

5.3.1 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|\, \theta, \underline{\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,\dots,L-1} \mid 0, \sigma_\theta^2 I) \mathrm{d}\theta^{1,\dots,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,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} \\
&= \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{aligned} \tag{40}
$$

note we are not dealing with infinity at the moment

5.3.2 look at co-variance

$$
\mathbb{E}\left[\begin{bmatrix} y \\ y^\star \end{bmatrix} \begin{bmatrix} y^\top & y^\star \end{bmatrix} \,\middle|\, \begin{bmatrix} X \\ x^{\star\top} \end{bmatrix}\right] \tag{41}
$$

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

$$
\begin{aligned}
&= \int_y \int_{y^\star} \int_\theta p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \begin{bmatrix} y^\top & y^\star \end{bmatrix} \,\middle|\, \theta, \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) p(\theta|\sigma_\theta^2) \, \mathrm{d}\theta \, \mathrm{d}y \, \mathrm{d}y^\star \\
&= \int_\theta \underbrace{\int_y \int_{y^\star} \begin{bmatrix} y \\ y^\star \end{bmatrix} \begin{bmatrix} y^\top & y^\star \end{bmatrix} \mathcal{N}\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \,\middle|\, \begin{bmatrix} f_\theta(X) \\ f_\theta(x^\star) \end{bmatrix}, \sigma_\epsilon^2 I\right) \mathrm{d}y \, \mathrm{d}y^\star}_{\mathbb{E}[Z^2] \quad Z \text{ is } \textbf{not} \text{ mean-subtracted}} \mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) \mathrm{d}\theta
\end{aligned} \tag{42}
$$

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

$$
\begin{aligned}
&\mathrm{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \quad \Longrightarrow \quad \mathbb{E}[Z^2] = \mathrm{Var}[Z] + (\mathbb{E}[Z])^2 \\
&= \int_\theta \underbrace{\sigma_\epsilon^2 I}_{\mathrm{Var}[Z]} + \underbrace{\begin{bmatrix} f_\theta(X) \\ f_\theta(x^\star) \end{bmatrix} \begin{bmatrix} f_\theta(X)^\top & f_\theta(x^\star) \end{bmatrix}}_{(\mathbb{E}[Z])^2} \mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) \mathrm{d}\theta \\
&= \sigma_\epsilon^2 I + \int_\theta \begin{bmatrix} \left(\phi^L(X) W^L + b^L\right) \left(W^{L\top} x^L(X)^\top + b^{L\top}\right) & \left(\phi^L(X) W^L + b^L\right) \left(W^{L\top} \phi^L(x^{\star\top})^\top + b^{L\top}\right) \\ \left(\phi^L(\phi^{\star\top}) W^L + b^L\right) \left(W^{L\top} \phi^L(X)^\top + b^{L\top}\right) & \left(\phi^L(x^{\star\top}) W^L + b^L\right) \left(W^{L\top} \phi^L(x^{\star\top})^\top + b^{L\top}\right) \end{bmatrix} \mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) d\theta
\end{aligned} \tag{43}
$$

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

$$= \sigma_e^2 I + \int_\theta \begin{bmatrix} \phi^L(X)W^L W^{L\top} x^L(X)^\top + b^L b^{L\top} & \phi^L(X)W^L W^{L\top} x^L(x^{\star\top})^\top + b^L b^{L\top} \\ \phi^L(x^{\star\top})W^L W^{L\top} \phi^L(X)^\top + b^L b^{L\top} & \phi^L(x^{\star\top})W^L W^{L\top} \phi^L(x^{\star\top})^\top + b^L b^{L\top} \end{bmatrix} \mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) d\theta \tag{44}$$

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

$$\mathcal{N}(\theta \mid 0, \sigma_\theta^2 I) d\theta = \mathcal{N}(\theta^L \mid 0, \sigma_\theta^2 I)\, \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} \tag{45}$$

$$= \int \begin{bmatrix} \sigma_w^2 \phi^L(X) x^L(X)^\top + \sigma_b^2 & \sigma_w^2 \phi^L(X)\phi^L(x^{\star\top})^\top + \sigma_b^2 \\ \sigma_w^2 \phi^L(x^{\star\top})\phi^L(X)^\top + \sigma_b^2 & \sigma_w^2 \phi^L(x^{\star\top})\phi^L(x^{\star\top})^\top + \sigma_b^2 \end{bmatrix} \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I) d\theta^{1,\dots,L-1} \tag{46}$$

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

$$\begin{aligned} &\int \sigma_w^2 \phi^L(X)\phi^L(X)^\top \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I)\, \mathrm{d}\theta^{1,\dots,L-1} + \int \sigma_b^2 \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I)\, \mathrm{d}\theta^{1,\dots,L-1} \\ =&\sigma_w^2 \int \phi^L(X)\phi^L(X)^\top \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I)\, \mathrm{d}\theta^{1,\dots,L-1} + \sigma_b^2 \end{aligned} \tag{47}$$

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

$$= \sigma_b^2 + \sigma_w^2 \int \left[\phi(W^{L-1}\phi^{L-1}(X) + b^{L-1})\phi(W^{L-1}\phi^{L-1}(X) + b^{L-1})^\top\right] \mathcal{N}(\theta^{1,\dots,L-1} \mid 0, \sigma_\theta^2 I)\, \mathrm{d}\theta^{1,\dots,L-1} \tag{48}$$

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

6 Single layer neural network

$$
\begin{aligned}
f_k(x) &= b_k + \sum_{j=1}^{H} v_{jk} h_j(x) \\
h_j(x) &= \tanh\left(a_j + \sum_{i=1}^{I} u_{ij} x_i\right)
\end{aligned}
\tag{49}
$$

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

$$
\begin{aligned}
\underbrace{f_k(x)}_{z_k^l} &= \underbrace{b_k}_{b_k^l} + \overbrace{\sum_{j=1}^{H}}^{N_l} \underbrace{v_{jk}}_{W_{k,j}^l} \times \underbrace{\tanh}_{\phi}\Big(\underbrace{\underbrace{a_j}_{b_j^{l-1}} + \underbrace{u_{:,j}^{\top}}_{W_{:,j}^{l-1\top}} x}_{z_j^{l-1}(x)}\Big) \\
\implies z_k^l(x) &= b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \times \phi\big(z_j^{l-1}(x)\big) \quad \text{modern notation}
\end{aligned}
\tag{50}
$$

6.1 $p\big(z_k^l(x)\big)$ for single input x

We need CLT for computing this probability.

6.1.1 Central Limit Theorem:

$$
X^{(1)}, X^{(2)}, \ldots, X^{(n)} \quad \text{are i.i.d samples} \tag{51}
$$

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:

$$
\begin{aligned}
\sqrt{n}\big(\overline{X} - \mathbb{E}[X^{(1)}]\big) &\xrightarrow{d} \mathcal{N}(0, \sigma^2) \\
\big(\overline{X} - \mathbb{E}[X^{(1)}]\big) &\xrightarrow{d} \mathcal{N}\big(0, \frac{\sigma^2}{n}\big) \\
\frac{1}{\sigma}\sqrt{n}\big(\overline{X} - \mathbb{E}[X^{(1)}]\big) &\xrightarrow{d} \mathcal{N}(0, 1)
\end{aligned}
\tag{52}
$$

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

$$
\begin{aligned}
&\sqrt{n}\big(\overline{X} - \mathbb{E}[X^{(1)}]\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \\
\implies &\sqrt{n}\sqrt{n}\big(\overline{X} - \mathbb{E}[X^{(1)}]\big) \xrightarrow{d} \mathcal{N}(0, \sqrt{n}^2\sigma^2) = \mathcal{N}(0, n\sigma^2) \\
\implies &n\big(\overline{X} - \mathbb{E}[X^{(1)}] \xrightarrow{d} \mathcal{N}(0, n\sigma^2) \\
\implies &\Big(\sum_{i=1}^{n} X_i - n\mathbb{E}[X^{(1)}]\Big) \xrightarrow{d} \mathcal{N}(0, n\sigma^2)
\end{aligned}
\tag{53}
$$

choose one of these conditions to suit the situation

6.1.2 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:

$$z_k^l(x) = b_k^l + \underbrace{\sum_{j=1}^{N_l} \underbrace{W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)}_{X_j}}_{\sum_{j=1}^{N_l} X_j}, \quad \text{note we are not taking average} \tag{54}$$

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

$$X_j \equiv W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) \tag{55}$$

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

Expectation

$$\begin{aligned}
\mathbb{E}\big[W_{k,j}^l\ \phi\big(z_j^{l-1}(x)\big)\big] &= \mathbb{E}[W_{k,j}^l]\ \mathbb{E}\big[\phi\big(z_j^{l-1}(x)\big)\big] \qquad \text{since } W_{k,j}^l \text{ and } \phi\big(z_j^{l-1}(x)\big) \text{ are independent} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{as } z_j^{l-1}(x) \text{ depends on } (W^{l-1}, b^{l-1}) \\
&= 0 \times \mathbb{E}[\phi\big(z_j^{l-1}(x)\big)] \qquad \text{because we choose} \quad W_{k,j}^l \sim \mathcal{N}(0, \sigma_w) \\
&= 0
\end{aligned} \tag{56}$$

Variance

$$\begin{aligned}
&\mathrm{Var}\big[W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)\big] \\
&= \mathbb{E}\Big[\Big(W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)\Big)^2\Big] \\
&= \mathbb{E}\big[\big(W_{k,j}^l\big)^2\big]\ \mathbb{E}\big[\phi\big(z_j^{l-1}(x)\big)^2\big] \quad \text{since } W_{k,j}^l \text{ and } \phi\big(z_j^{l-1}(x)\big) \text{ are independent} \\
&= \sigma_w^2 \mathbb{E}\big[\underbrace{\phi\big(z_j^{l-1}(x)\big)^2}_{\text{bounded}}\big] \quad \Longrightarrow \mathrm{Var}\big[W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)\big] \text{ to be bounded} \\
&= \sigma_w^2\ \mathbb{E}\big[\phi\big(z_j^{l-1}(x)\big)^2\big]
\end{aligned} \tag{57}$$

we leave in this form, as

$$\mathbb{E}\big[\phi\big(z_j^{l-1}(x)\big)^2\big] \equiv \mathbb{E}_{W^{l-1},\ldots,b^{l-1},\ldots}\big[\phi\big(z_j^{l-1}(x)\big)^2\big] \tag{58}$$

6.1.4 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.

$$\text{using} \quad \left(\sum_{i=1}^{n} X_i - n\mathbb{E}[X_1]\right) \xrightarrow{d} \mathcal{N}(0, n\sigma^2)$$
$$\implies \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) - 0\right) \sim \mathcal{N}\Big(0, N_l\ \sigma_w^2\ \mathbb{E}\big[\phi(z_1^{l-1}(x))^2\big]\Big) \quad N_l \to \infty \tag{59}$$

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

$$\sigma(W_{k,j}^l) = \sigma_w = \frac{1}{\sqrt{N_l}} \quad \implies \sigma_w^2 = \frac{1}{N_l} \tag{60}$$

the above is the key, implication is:

$$\begin{aligned} \implies \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) - 0\right) &\sim \mathcal{N}\Big(0, N_l \frac{1}{N_l}\ \mathbb{E}\big[\phi(z_1^{l-1}(x))^2\big]\Big) \\ &= \mathcal{N}\Big(0, \underbrace{\mathbb{E}\big[\phi(z_1^{l-1}(x))^2\big]}_{\text{bounded}}\Big) \end{aligned} \tag{61}$$

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

$$\begin{aligned} X &\sim \mathcal{N}(\mu_X, \sigma_X^2) \\ Y &\sim \mathcal{N}(\mu_Y, \sigma_Y^2) \\ Z = X + Y \quad Z &= X + Y \\ \implies Z &\sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \end{aligned} \tag{62}$$

Therefore:

$$\left(z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x))\right) \xrightarrow{d} \mathcal{N}\Big(0, \underbrace{\sigma_b^2}_{\sigma_X^2} + \underbrace{\mathbb{E}\big[\phi(z_1^{l-1}(x))^2\big]}_{\sigma_Y^2}\Big) \quad \text{as } N_l \to \infty \tag{63}$$

appreciate the recursion here

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

To do so, we need to used results from **Multidimensional CLT**:

6.2.1 Multidimensional CLT:

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

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

$$\implies \overline{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(k)} \end{bmatrix} = \begin{bmatrix} \overline{\mathbf{X}}^{(1)} \\ \vdots \\ \overline{\mathbf{X}}^{(p)} \\ \vdots \\ \overline{\mathbf{X}}^{(q)} \\ \vdots \\ \overline{\mathbf{X}}^{(k)} \end{bmatrix} \tag{64}$$

1. **Sample mean version**

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big[\mathbf{X}_i - \mathbb{E}\big[\mathbf{X}_i\big]\big] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{X}_i - \mathbb{E}\big[\mathbf{X}_1\big]) \qquad \text{since } p(X_i) = p(X_1) \\ &= \frac{\sqrt{n}}{\sqrt{n}} \frac{1}{\sqrt{n}} \Big(\sum_{i=1}^{n} \mathbf{X}_i \Big) - \frac{n}{\sqrt{n}} \mathbb{E}\big[\mathbf{X}_1\big] \\ &= \sqrt{n} \left(\overline{\mathbf{X}} - \mathbb{E}\big[\mathbf{X}_1\big] \right) \end{aligned} \tag{65}$$

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big[\mathbf{X}_i - \mathbb{E}\big[\mathbf{X}_i\big]\big] = \sqrt{n} \left(\overline{\mathbf{X}} - \mathbb{E}\big[\mathbf{X}_1\big] \right) \xrightarrow{d} \mathcal{N}_d\big(0, \boldsymbol{\Sigma}(\mathbf{X}_1)\big) \\ \implies &\sqrt{n}\, \mathbb{E}\Big[\Big(\underbrace{\overline{\mathbf{X}}^{(p)} - \mathbb{E}\big[\overline{\mathbf{X}}_1^{(p)}\big]}_{\text{scalar}} \Big) \Big(\underbrace{\overline{\mathbf{X}}^{(q)} - \mathbb{E}\big[\,\overline{\mathbf{X}}_1^{(q)}\big]}_{\text{scalar}} \Big) \Big] = \boldsymbol{\Sigma}(\mathbf{X}_1)_{(p),(q)} \end{aligned} \tag{66}$$

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

2. **Sample sum version**:

$$
\begin{aligned}
\left(\Big[\sum_i^n \mathbf{X}_i\Big] - n\mathbb{E}\big[\mathbf{X}_1\big]\right) &\xrightarrow{d} \mathcal{N}_k(0, n\boldsymbol{\Sigma}) \\
\implies \mathbb{E}\bigg[\Big(\Big[\sum_i^n \mathbf{X}_i\Big]^{(p)} - n\mathbb{E}\big[\mathbf{X}_1\big]^{(p)}\Big)\Big(\Big[\sum_i^n \mathbf{X}_i\Big]^{(q)} - n\mathbb{E}\big[\mathbf{X}_1\big]^{(q)}\Big)\bigg] &= n\boldsymbol{\Sigma}_{(p),(q)} \\
\implies \mathbb{E}\Big[\big(n\overline{\mathbf{X}}^{(p)} - n\mathbb{E}\big[X_1^{(p)}\big]\big)\big(n\overline{\mathbf{X}}^{(q)} - n\mathbb{E}\big[X_1^{(q)}\big]\big)\Big] &= n\boldsymbol{\Sigma}_{(p),(q)} \\
\implies \mathbb{E}\Big[\Big(\Big[\sum_i^n \mathbf{X}_i\Big]^{(p)} - n\mathbb{E}\big[X_1^{(p)}\big]\Big)\Big(\Big[\sum_i^n \mathbf{X}_i\Big]^{(q)} - n\mathbb{E}\big[X_1^{(q)}\big]\Big)\Big] &= n\boldsymbol{\Sigma}_{(p),(q)}
\end{aligned}
\tag{67}
$$

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

6.2.2 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

$$
\begin{bmatrix} \vdots \\ W_{k,1}^l \phi\big(z_1^{l-1}(x)\big) \\ \vdots \\ W_{k,1}^l \phi\big(z_j^{l-1}(x')\big) \\ \vdots \end{bmatrix} + \cdots + \begin{bmatrix} \vdots \\ W_{k,N_l}^l \phi\big(z_j^{l-1}(x)\big) \\ \vdots \\ W_{k,N_l}^l \phi\big(z_j^{l-1}(x')\big) \\ \vdots \end{bmatrix} = \underbrace{\begin{bmatrix} \vdots \\ \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) \\ \vdots \\ \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x')\big) \\ \vdots \end{bmatrix}}_{\begin{bmatrix} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \sum_{i=1}^n X_i^{(k)} \end{bmatrix}} = \underbrace{\begin{bmatrix} \vdots \\ z_k^l(x) \\ \vdots \\ z_k^l(x') \\ \vdots \end{bmatrix}}_{\sum_{i=1}^n \mathbf{X}_i}
\tag{68}
$$

compare with the standard notation of Multi-dimensional CLT, and use "sample sum version" of CLT, Eq.[64], and remember $z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)$, to simplify derivation, let's deliberately not looking at b_k^l for now

$$
\begin{aligned}
\sum_i^n X_i^{(p)} = \Big[\sum_i^n X_i\Big]^{(p)} &\triangleq \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) \\
\sum_i^n X_i^{(q)} = \Big[\sum_i^n X_i\Big]^{(p)} &\triangleq \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x')\big) \\
X_1^{(p)} &\triangleq W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) \\
X_1^{(q)} &\triangleq W_{k,j}^l \phi\big(z_j^{l-1}(x')\big) \\
\implies \mathbf{\Sigma}(\mathbf{X}_1)_{(p),(q)} &\triangleq \mathbf{Cov}\Big(W_{k,1}^l \phi(z_1^{l-1}(x)), W_{k,1}^l \phi(z_1^{l-1}(x'))\Big)
\end{aligned}
\tag{69}
$$

using above identities in Eq.[69]

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\Big[\sum_i^n \mathbf{X}_i\Big]^{(p)} - n\mathbb{E}\big[X_1^{(p)}\big]\Big)\Big(\Big[\sum_i^n \mathbf{X}_i\Big]^{(q)} - n\mathbb{E}\big[X_1^{(q)}\big]\Big)\Big] = n\mathbf{\Sigma}(\mathbf{X}_1)_{(p),(q)} \\
\implies &\mathbb{E}\Big[\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) - N_l \underbrace{\mathbb{E}\big[W_{k,1}^l \phi(z_1^{l-1}(x))\big]}_{=0}\Big)\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x')) - N_l \underbrace{\mathbb{E}\big[W_{k,1}^l \phi(z_1^{l-1}(x'))\big]}_{=0}\Big)\Big] \\
&= N_l \mathbf{Cov}\Big(W_{k,1}^l \phi(z_1^{l-1}(x)), W_{k,1}^l \phi(z_1^{l-1}(x'))\Big) \\
&= N_l\, \mathbb{E}\Big[W_{k,1}^l \phi(z_1^{l-1}(x)) \times W_{k,1}^l \phi(z_1^{l-1}(x'))\Big]
\end{aligned}
\tag{70}
$$

look at $z_k^l(x)$ with b_k^l too:

$$
\begin{aligned}
\mathbf{Cov}\big(z_k^l(x), z_k^l(x')\big) &= \sigma_b^2 + \mathbb{E}\Big[\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)\Big)\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x')\big)\Big)\Big] \\
&= \sigma_b^2 + N_l\, \mathrm{Cov}\big(W_{k,1}^l \phi\big(z_1^{l-1}(x)\big), W_{k,1}^l \phi\big(z_1^{l-1}(x')\big)\big) \quad \text{use CLT result above from Eq.[71]} \\
&= \sigma_b^2 + N_l \sigma_w^2 \mathrm{Cov}\big(\phi\big(z_1^{l-1}(x)\big), \phi\big(z_1^{l-1}(x')\big)\big) \\
&= \sigma_b^2 + N_l \frac{1}{N_l} \mathrm{Cov}\big(\phi\big(z_1^{l-1}(x)\big), \phi\big(z_1^{l-1}(x')\big)\big) \\
&= \sigma_b^2 + \mathrm{Cov}\big(\phi\big(z_1^{l-1}(x)\big), \phi\big(z_1^{l-1}(x')\big)\big) \\
&= \sigma_b^2 + \mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big) \times \phi\big(z_1^{l-1}(x')\big)\big]
\end{aligned}
\tag{71}
$$

there are many notes this:

1. **note 1** by definition, $\dim(z^l) = N_{l+1}$, so the "entire" $\mathbf{Cov}\big(z^l, z^l\big)$ is of size:

$$
\underbrace{N_{l+1}}_{\forall k}\ \underbrace{|\mathcal{X}|}_{\forall x} \quad \times \quad \underbrace{N_{l+1}}_{\forall k'}\ \underbrace{|\mathcal{X}|}_{\forall x'}
\tag{72}
$$

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:

$$\begin{aligned} z_k^l(x) &= b_k + \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) \\ z_{k'}^l(x) &= b_{k'} + \sum_{j=1}^{N_l} W_{k',j}^l \phi\big(z_j^{l-1}(x)\big) \\ \implies \mathbb{E}\Big[W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) \times\ & W_{k',j'}^l \phi\big(z_{j'}^{l-1}(x)\big)\Big] = 0 \quad \forall\{k,k',j,j'\} \end{aligned} \tag{73}$$

note 3: in literature, it is written:

$$\begin{aligned} \mathbb{E}\big[z_k^l(x) z_k^l(x')\big] &= \sigma_b^2 + \sigma_w^2\, \mathbb{E}\Big[\sum_{j=1}^{N_l} \phi\big(z_j^{l-1}(x)\big)\phi\big(z_j^{l-1}(x')\big)\Big] \\ \text{instead of } &= \sigma_b^2 + \mathbb{E}\Big[\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)\Big)\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x')\big)\Big)\Big] \end{aligned} \tag{74}$$

This is because of **note1** above
regardless of this special property CLT still apply.

6.2.3 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:

$$\begin{aligned} \mathbb{E}\big[f(x)\big] &= 0 \\ \mathbb{E}\big[f(x,x')\big] &= \mathbf{\Sigma}_{x,x'} \\ \implies f &\sim \mathcal{GP}(0, \mathbf{\Sigma}) \end{aligned} \tag{75}$$

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

$$\begin{aligned} \text{Cov}\Big[z_k^l(x), z_k^l(x')\Big] &= \sigma_b^2 + \mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big) \times \phi\big(z_1^{l-1}(x')\big)\big] \quad \text{as } N_l \to \infty \\ z_k^l(x) &\xrightarrow{d} \mathcal{N}\Big(0, \sigma_b^2 + \mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big)^2\big]\Big) \quad \text{as } N_l \to \infty \end{aligned} \tag{76}$$

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

$$\begin{aligned} &\implies z_k^l(\mathcal{X}) \sim \mathcal{GP}(0, \mathbf{\Sigma}^l) \\ \text{where specific co-variance} \quad &\mathbf{\Sigma}_{x,x'}^l = \sigma_b^2 + \mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big) \times \phi\big(z_1^{l-1}(x')\big)\big] \end{aligned} \tag{77}$$

6.3 looking at GP systematically

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

$$\mathbf{\Sigma}^l \to K^l \tag{78}$$

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

$$\sigma(W_{k,j}^l) = \frac{\sigma_w}{\sqrt{N_l}} \tag{79}$$

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

$$\begin{aligned} K^l(x, x') &= \mathbb{E}\big[z_k^l(x) z_k^l(x') \big|\, z^{l-1}\big] \\ &= \sigma_b^2 + \sigma_w^2 \mathbb{E}\big[\phi(z_1^{l-1}(x)) \;\times\; \phi(z_1^{l-1}(x'))\big] \quad \text{apply CLT} \quad N_l \to \infty \\ &= \sigma_b^2 + \sigma_w^2 \underbrace{\mathbb{E}_{z_1^{l-1}(\mathcal{X}) \sim \mathcal{GP}(0, K^{l-1})} \Big[\phi(z_1^{l-1}(x)) \phi(z_1^{l-1}(x'))\Big]}_{} \end{aligned} \tag{80}$$

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.[77], and $\phi(z_1^{l-1}(x))$ is function on a specific point x, keep on going:

$$\begin{aligned} &= \sigma_b^2 + \sigma_w^2 \underbrace{F_\phi\big(K^{l-1}(x, x'), K^{l-1}(x, x), K^{l-1}(x', x')\big)}_{F_\phi(K^{l-1})} \\ &= \sigma_b^2 + \sigma_w^2 \; F_\phi(K^{l-1}(x, x')) \end{aligned} \tag{81}$$

using properties of point Marginals of Gaussian Process:

$$\begin{aligned} F_\phi(K^{l-1}(x, x')) &= \mathbb{E}_{z_j^{l-1} \sim \mathcal{GP}(0, K^{l-1})} \Big[\phi(z_j^{l-1}(x)) \phi(z_j^{l-1}(x'))\Big] \\ &= \mathbb{E}_{\underbrace{\left(z_j^{l-1}(x), z_j^{l-1}(x')\right)}_{\text{2 points on function } z_j^{l-1}} \sim \underbrace{\mathcal{N}\big(0, K^{l-1}(x, x')\big)}_{\text{2D Gaussian}}} \Big[\phi\big(z_j^{l-1}(x)\big) \phi\big(z_j^{l-1}(x')\big)\Big] \end{aligned} \tag{82}$$

$$\begin{bmatrix} z_j^{l-1}(x) \\ z_j^{l-1}(x') \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}\,, \begin{bmatrix} K^{l-1}(x, x) & K^{l-1}(x, x') \\ K^{l-1}(x, x') & K^{l-1}(x', x') \end{bmatrix}\right) \tag{83}$$

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

$$= F_\phi\big(K^{l-1}(x, x'), K^{l-1}(x, x), K^{l-1}(x', x')\big) \tag{84}$$

6.4 in summary

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

$$K^l(x, x') = \sigma_b^2 + \sigma_w^2 \, \mathbb{E}_{\left(z_j^{l-1}(x), z_j^{l-1}(x')\right) \sim \mathcal{N}\left(0, K^{l-1}(x, x')\right)} \Big[\phi\big(z_j^{l-1}(x)\big) \phi\big(z_j^{l-1}(x')\big)\Big] \tag{85}$$

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

7 Expand GP across all layers

7.1 Overall objective

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

$$\begin{aligned} p(z^L|x) &= \int p(z^L, K^0, K^1, \dots, K^L|x)\, \mathrm{d}K^{0,\dots,L} \\ &= \int p(z^L|K^L)\Big(\prod_{l=1}^{L} p(K^l|K^{l-1})\Big)p(K^0|x)\, \mathrm{d}K^{0,\dots,L} \end{aligned} \tag{86}$$

7.2 $p(z^L|K^L)$: conditions on $K^l \equiv \left\{\phi\big(z^{l-1}\big)(x)\big)\phi\big(z^{l-1}\big)(x')\big)\right\}_{p,q}$

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

$$K^l(x, x') = \mathbb{E}\big[z_k^l(x) z_k^l(x')\big|\, z^{l-1}\big] \tag{87}$$

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:

$$z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) \tag{88}$$

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

$$\begin{aligned} z_k^l(x)\big|\, \left\{\phi(z_j^{l-1}(x))\right\}_j^{N_l} &= b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \underbrace{\phi(z_j^{l-1}(x))}_{\text{constant}} \\ \implies z_k^l(x)\big|\, \left\{\phi(z_j^{l-1}(x))\right\}_j^{N_l} &\sim \mathcal{N}\Big(0, \sigma_b^2 + \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2 \mathrm{Var}\big[W_{k,j}^l\big]\Big) \\ &= \mathcal{N}\Big(0, \sigma_b^2 + \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2\Big) \end{aligned} \tag{89}$$

using property of weighted sum of Gaussian:

$$\begin{aligned} X_i &\sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad i = 1, \dots, \\ \implies \sum_{i=1}^{n} a_i X_i &\sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2\, \mathrm{Var}[X_i]\right) \end{aligned} \tag{90}$$

Conditional Co-variance

$$
\begin{aligned}
&\mathrm{Cov}\Big[z_k^l(x), z_k^l(x') \;\Big|\; \Big\{\phi\big(z_j^{l-1}(x)\big), \phi\big(z_j^{l-1}(x')\big)\Big\}_{j=1}^{N_l}\Big] \\
&= \mathbb{E}\Big[z_k^l(x) z_k^l(x') \;\Big|\; \Big\{\phi\big(z_j^{l-1}(x)\big), \phi\big(z_j^{l-1}(x')\big)\Big\}_{j=1}^{N_l}\Big] \\
&= \sigma_b^2 + \mathbb{E}_{W_{k,j}^l}\Big[\sum_{j=1}^{N_l} {W_{k,j}^l}^2 \underbrace{\phi\big(z_j^{l-1}(x)\big)\, \phi\big(z_j^{l-1}(x')\big)}_{\text{constant, used as condition}}\Big] \\
&= \sigma_b^2 + \sum_{j=1}^{N_l} \mathrm{Var}\big[W_{k,j}^l\big]\, \phi\big(z_j^{l-1}(x)\big)\, \phi\big(z_j^{l-1}(x')\big) \\
&= \sigma_b^2 + \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi\big(z_j^{l-1}(x)\big)\, \phi\big(z_j^{l-1}(x')\big)
\end{aligned}
\tag{91}
$$

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

$$
\begin{aligned}
\mathrm{Cov}\Big[z_k^l(x), z_k^l(x') \;\Big|\; \Big\{\phi\big(z_j^{l-1}(x)\big), \phi\big(z_j^{l-1}(x')\big)\Big\}_{j=1}^{N_l}\Big] &= \sigma_b^2 + \sigma_w^2\, \frac{1}{N_l} \sum_{j=1}^{N_l} \phi\big(z_j^{l-1}(x)\big)\, \phi\big(z_j^{l-1}(x')\big) \\
z_k^l(x) \big|\, \{\phi(z_j^{l-1}(x))\}_j^{N_l} &\sim \mathcal{N}\Big(0, \sigma_b^2 + \sigma_w^2\, \frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2\Big) \\
\implies \begin{bmatrix} z^l(x) \mid \phi\big(z_j^{l-1}(x)\big) \\ z^l(x') \mid \phi\big(z_j^{l-1}(x')\big) \end{bmatrix} &\sim \mathcal{N}\Big(\mathbf{0}, G\Big(\begin{bmatrix} K^l(x,x) & K^l(x,x') \\ K^l(x,x') & K^l(x',x') \end{bmatrix}\Big)\Big)
\end{aligned}
\tag{92}
$$

in GP paradigm:

$$
z^l(x) | K^l \sim \mathcal{GP}\big(z^l; \mathbf{0}, G(K^l)\big) \tag{93}
$$

where

$$
\begin{aligned}
K^l(x,x') &= \frac{1}{N_l} \sum_{j=1}^{N_l} \phi\big(z_j^{l-1}(x)\big)\, \phi\big(z_j^{l-1}(x')\big) \\
G\big(K^l(x,x')\big) &= \sigma_b^2 + \sigma_w^2 K^l(x,x')
\end{aligned}
\tag{94}
$$

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:

$$
K_{p,q}^l \equiv K^l(x,x') = \begin{cases} \frac{1}{d_{\text{in}}} \sum_{j=1}^{d_{\text{in}}} x_j x'_j = \frac{1}{d_{\text{in}}} x^\top x' & l = 0 \\ \frac{1}{N_l} \sum_{j=1}^{N_l} \phi\big(z_j^{l-1}(x)\big) \phi\big(z_j^{l-1}(x')\big) & l > 0 \end{cases} \tag{95}
$$

to reflect:

$$
\mathrm{Cov}(z_k^l, z_{k'}^l) = 0 \;\forall\, k, k' \in \{1, \dots N_{l+1}\} \tag{96}
$$

note that

$$K^0(x,x') = \frac{1}{d_{\text{in}}} x^\top x' \quad \text{appears again in NTK} \tag{97}$$

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

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

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

$$\implies p(\mathbf{z}^l|K^l) = \mathcal{N}\Big(\mathbf{0}, G\big(K^l\big) \otimes \mathbf{I}_{N_{l+1}\times N_{l+1}}\Big) = \mathcal{GP}\big(\mathbf{z}^l; 0, G(K^l)\big) \tag{98}$$

7.3 $p(K^l|K^{l-1})$

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

$$
\begin{aligned}
p(K^l|K^{l-1}) &= \int_{z^{l-1}} p(K^l|z^{l-1}) p(z^{l-1}|K^{l-1}) \\
&= \int_{z^{l-1}} p(K^l|z^{l-1}) \mathcal{GP}\big(z^{l-1}; 0, G(K^{l-1})\big)
\end{aligned}
\tag{99}
$$

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

$$
p(K^l_{p,q}|K^{l-1}_{p,q}) = \int_{z^{l-1}(x)\,, z^{l-1}(x')} p\Big(\frac{1}{N_l}\sum_{j=1}^{N_l} \phi\big(z^l_j(x)\big)\phi\big(z^l_j(x')\big)\Big) \mathcal{N}\Big(\begin{bmatrix} z^{l-1}(x) \\ z^{l-1}(x') \end{bmatrix}; 0, G\Big(\begin{bmatrix} K^{l-1}(x,x) & K^{l-1}(x,x') \\ K^{l-1}(x,x') & K^{l-1}(x',x') \end{bmatrix}\Big)\Big)
\tag{100}
$$

7.3.1 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 $K^{l,j}_{p,q}$:

$$
\begin{aligned}
\underbrace{K^l_{p,q}}_{\overline{X}} &= \frac{1}{N_l}\sum_{j=1}^{N_l} \underbrace{\phi\big(z^{l-1}_j(x)\big)\phi\big(z^{l-1}_j(x')\big)}_{X_j \equiv K^{l,j}_{p,q}} \\
\implies p(K^{l,1}_{p,q}|K^{l-1}_{p,q}) &= \int_{z^{l-1}(x)\,, z^{l-1}(x')} p\big(\phi(z^l_j(x))\phi(z^l_j(x'))\big) \mathcal{N}\Big(\begin{bmatrix} z^{l-1}(x) \\ z^{l-1}(x') \end{bmatrix}; 0, G\Big(\begin{bmatrix} K^{l-1}(x,x) & K^{l-1}(x,x') \\ K^{l-1}(x,x') & K^{l-1}(x',x') \end{bmatrix}\Big)\Big) \\
&= (F \circ G)(K^{l-1}_{p,q})
\end{aligned}
\tag{101}
$$

using CLT, pick the most appropriate definition:

$$
(\overline{X} - \mathbb{E}[X_1]) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\text{Var}[X_1]}{n}\Big)
\tag{102}
$$

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

$$
\begin{aligned}
(\overline{X} - \mathbb{E}[X_1]) &\xrightarrow{d} \mathcal{N}\Big(0, \frac{\text{Var}[X_1]}{n}\Big) \\
\implies (K^l_{p,q} - \mathbb{E}[K^{l,1}_{p,q}]) &\xrightarrow{d} \mathcal{N}\Big(0, \frac{\text{Var}[K^{l,1}_{p,q}]}{N_l}\Big) \\
\implies (K^l_{p,q} - (F \circ G)(K^{l-1}_{p,q})) &\xrightarrow{d} \mathcal{N}\Big(0, \frac{\text{Var}[K^{l,1}_{p,q}]}{N_l}\Big) \\
\implies (K^l_{p,q}|K^{l-1}_{p,q}) &\xrightarrow{d} \mathcal{N}\Big((F \circ G)(K^{l-1}), \frac{\text{Var}[K^{l,1}_{p,q}]}{N_l}\Big) \\
\implies \lim_{N_l \to \infty} p(K^l|K^{l-1}) &= \delta\big(K^l - (F \circ G)(K^{l-1})\big) \quad \text{entire matrix}
\end{aligned}
\tag{103}
$$

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

7.4 putting in the overall objective function

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

$$
\begin{aligned}
p(z^L|x) &= \int p(z^L, K^0, K^1, \dots, K^L|x)\, \mathrm{d}K^{0,\dots,L} \\
&= \int p(z^L|K^L)\Big(\prod_{l=1}^{L} p(K^l|K^{l-1})\Big) p(K^0|x)\, \mathrm{d}K^{0,\dots,L} \\
\lim_{N_L\to\infty,\dots,N_1\to\infty} p(z^L|x) &= \int p(z^L|K^L)\Big(\prod_{l=1}^{L} \delta\big(K^l - (F\circ G)(K^{l-1})\big)\Big) p(K^0|x)\, \mathrm{d}K^{0,\dots,L} \\
&= \int \mathcal{GP}\big(z^L; 0, G(K^L)\big) \underbrace{\Big(\prod_{l=1}^{L} \delta\big(K^l - (F\circ G)(K^{l-1})\big)\Big) \delta\Big(K^0 - \frac{1}{d_{\text{in}}} x^\top x\Big)}_{= \begin{cases} = 1 & \text{if } K^L = (F\circ G)(K^{L-1}) \\ & \quad = (F\circ G)^2(K^{L-2}) \dots \\ & \quad = (F\circ G)^L\Big(\frac{1}{d_{\text{in}}} x^\top x\Big) \\ = 0 & \text{otherwise} \end{cases}} \mathrm{d}K^{0,\dots,L} \\
&= \mathcal{GP}\Big(z^L; 0,\ G\circ (F\circ G)^L\big(\frac{1}{d_{\text{in}}} x^\top x\big)\Big)
\end{aligned}
\tag{104}
$$

8 Gradient Flow in terms of Neural Tangent Kernel

8.1 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:

$$C = \|y - \phi(z^L(x))\|^2 \tag{105}$$

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 $\frac{\delta C}{\delta \phi(z^L(x))}$

8.2 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)

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = -\frac{\partial C}{\partial \theta} \tag{106}$$

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:

8.2.1 $\frac{\delta C}{\delta f(\theta)}$ 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.,:

$$\theta \to \theta + \epsilon\eta \tag{107}$$

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

$$\lim_{\epsilon \to 0} \frac{C\big[f(\theta + \epsilon\eta\big] - C\big[f(\theta]\big)}{\epsilon} \tag{108}$$

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**:

$$\int_{\mathbf{X}} \frac{\delta J}{\delta g}(x)^\top \phi(x) \mathrm{d}x = \lim_{\epsilon \to 0} \frac{J[g + \epsilon\phi] - J[g]}{\epsilon} \tag{109}$$

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

$$\phi^\top \nabla_g J \tag{110}$$

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:

$$\lim_{\epsilon \to 0} \frac{C\big[f(\theta + \epsilon\eta)\big] - C\big[f(\theta)\big]}{\epsilon} \tag{111}$$

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

$$\begin{aligned}
&\implies \lim_{\epsilon \to 0} \frac{C\Big[\underbrace{f(\theta)}_{g} + \epsilon \underbrace{\eta \cdot \frac{\partial f(\theta)}{\partial \theta}}_{\phi} + O(\epsilon^2)\Big] - C\big[f(\theta)\big]}{\epsilon} && \text{matching with RMK representation} \\
&= \int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x)\right)^{\top} \left(\eta \cdot \frac{\partial f(\theta)}{\partial \theta}\right) \mathrm{d}x \\
&= \sum_{d=1}^{|\theta|} \int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x)\right)^{\top} \left(\eta \cdot \frac{\partial f(\theta)}{\partial \theta_d}\right) \mathrm{d}x \\
&= \sum_{d=1}^{|\theta|} \int_{\mathbf{X}} \sum_{i=1}^{N} \left(\frac{\delta C}{\delta f(\theta)}(x)\right)_i \left(\eta \cdot \frac{\partial f(\theta)}{\partial \theta_d}\right)_i \mathrm{d}x \\
&= \sum_{d=1}^{|\theta|} \sum_{i=1}^{N} \eta \int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x)\right)_i \left(\frac{\partial f(\theta)}{\partial \theta_d}\right)_i \mathrm{d}x && \text{change order of integral and sum}
\end{aligned} \tag{112}$$

8.2.2 $\frac{\delta C}{\delta f(\theta)}$ 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.:

$$\eta \equiv \frac{\partial \theta}{\partial t} \tag{113}$$

which correspond to the training regime used
by substitution:

$$\lim_{\epsilon \to 0} \frac{C[f(\theta + \epsilon\eta)] - C[f^{\theta}]}{\epsilon} = \sum_{d=1}^{|\theta|} \sum_{i=1}^{N} \left(\frac{\partial \theta}{\partial t}\right) \int_{\mathbf{X}} \left(\frac{\delta C}{\delta f(\theta)}(x)\right)_i \left(\frac{\partial f(\theta)}{\partial \theta_d}\right)_i \mathrm{d}x \tag{114}$$

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

$$\begin{aligned}
\frac{\partial \theta}{\partial t} &= -\frac{\partial C[f(\theta)]}{\partial \theta} \\
&= -\lim_{\epsilon \to 0} \frac{C[f(\theta + \epsilon\, \mathbf{I})] - C[f^{\theta}]}{\epsilon} \\
&= -\sum_{d'=1}^{|\theta|} \sum_{k=1}^{N} \int_{\mathbf{X}'} \left(\frac{\delta C}{\delta f(\theta)}(x')\right)_k \left(\frac{\partial f_k(\theta)}{\partial \theta_{d'}}\right)_k \mathrm{d}x' && \text{change index to } k \text{ and } x \to x'
\end{aligned} \tag{115}$$

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}
\tag{116}
$$

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

8.2.3 What happens $\Theta(x, x')$ is positive definite

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

$$
\lim_{\epsilon \to 0} \frac{C[f(\theta + \epsilon \frac{\partial \theta}{\partial t})] - C[f^\theta]}{\epsilon} = \text{negative value} \tag{117}
$$

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:

$$
\int_x \int_{x'} \underbrace{\bar{f}(x)^\top \Theta(x, x') \bar{f}(x')}_{} \tag{118}
$$

for a specific term

$$
\bar{f}(x)^\top K(x, x') \bar{f}(x') \tag{119}
$$

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

$$
\begin{aligned}
A \equiv \quad & f_i(x)\Theta(i,j)f_j(x) \;+\; f_i(x)\Theta(i,j)f_j(x') = f_i(x)\Theta(i,j)\big(f_j(x)+f_j(x')\big) \\
B \equiv \quad & f_i(x')\Theta(i,j)f_j(x) \;+\; f_i(x')\Theta(i,j)f_j(x') = f_i(x')\Theta(i,j)\big(f_j(x)+f_j(x')\big) \\
& A + B = \underbrace{\big(f_i(x)+f_i(x')\big)}_{g_i(x,x')}\,\Theta(i,j)\,\underbrace{\big(f_j(x)+f_j(x')\big)}_{g_j(x,x')}
\end{aligned}
\tag{120}
$$

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

$$
\begin{aligned}
& K \text{ is positive definite with respect to } \|\cdot\|_{p^{\text{in}}} \quad \text{if} \\
& \mathbb{E}_{x,x'\sim p^{\text{in}}}\big[f(x)^\top f(x')\big] > 0 \implies \mathbb{E}_{x,x'\sim p^{\text{in}}}\big[f(x)^\top K f(x')\big] > 0
\end{aligned}
\tag{121}
$$

8.2.4 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.

8.3 NTK in Neural Networks

we use the re-parameterization version of NN function:

$$
\begin{aligned}
z_k^{(l)} &= \frac{1}{\sqrt{N_l}}\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l \\
\text{where} \quad & W_{k,j}^l, b_k^l \sim \mathcal{N}(0,1)
\end{aligned}
\tag{122}
$$

Neural Tangent Kernel at each Layer l:

$$
\begin{aligned}
\Theta^l(x,x') &= \sum_{d=1}^{|\theta|} \frac{\partial z^l(x)}{\partial \theta_d} \otimes_{\text{outer}} \frac{\partial z^l(x')}{\partial \theta_d} \\
&= \sum_{d=1}^{|\theta|} \begin{bmatrix} \frac{\partial z_1^l(x)}{\partial \theta_d} & \cdots & \frac{\partial z_{N_{l+1}}^l(x)}{\partial \theta_d} \end{bmatrix} \begin{bmatrix} \frac{\partial z_1^l(x')}{\partial \theta_d} & \cdots & \frac{\partial z_{N_{l+1}}^l(x')}{\partial \theta_d} \end{bmatrix}^\top \\
&= \sum_{d=1}^{|\theta|} \begin{bmatrix} \frac{\partial z_1^l(x)}{\partial \theta_d}\frac{\partial z_1^l(x')}{\partial \theta_d} & \cdots & \frac{\partial z_{N_{l+1}}^l(x)}{\partial \theta_d}\frac{\partial z_{N_{l+1}}^l(x')}{\partial \theta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_1^l(x)}{\partial \theta_d}\frac{\partial z_1^l(x')}{\partial \theta_d} & \cdots & \frac{\partial z_{N_{l+1}}^l(x)}{\partial \theta_d}\frac{\partial z_{N_{l+1}}^l(x')}{\partial \theta_d} \end{bmatrix}
\end{aligned}
\tag{123}
$$

note that size of $\Theta^l(x, x')$ is $N_{l+1} \times N_{l+1}$, it has nothing to do with $|\theta|$ (it is used in the sum)

however, the **entire NTK** Θ^l will be the size $\big(N_{l+1} \times |\mathcal{D}|\big) \times \big(N_{l+1} \times |\mathcal{D}|\big)$
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 and x'
   (derivative correlations between **function's output**)

   which is different to fisher information matrix:

$$\mathbf{I}_{i,j} = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \log f(X;\theta)\right)\left(\frac{\partial}{\partial \theta_j} \log f(X;\theta)\right)\right] \tag{124}$$

2. 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:

$$\mathbf{u} \otimes_{\text{Kron}} \mathbf{v}^{\mathbf{T}} = \mathbf{u}\mathbf{v}^{\mathbf{T}} = \mathbf{u} \otimes_{\text{outer}} \mathbf{v} \tag{125}$$

9 NTK at initialization

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

$$\begin{bmatrix} \frac{1}{\sqrt{N_l}} W^l_{1,1}\phi\big(z^{l-1}_1(x)\big) + \sigma_b b_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{k,1}\phi\big(z^{l-1}_1(x)\big) + \sigma_b b_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{N_{l+1},1}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b_{N_{l+1}} \end{bmatrix} + \cdots + \begin{bmatrix} \frac{1}{\sqrt{N_l}} W^l_{1,N_l}\phi\big(z^{l-1}_1(x)\big) + \sigma_b b^l_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{k,N_l}\phi\big(z^{l-1}_1(x)\big) + \sigma_b b^l_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} W^l_{N_{l+1},N_l}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b^l_{N_{l+1}} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{1,j}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b^l_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{k,j}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b^l_k \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W^l_{N_{l+1},j}\phi\big(z^{l-1}_j(x)\big) + \sigma_b b^l_{N_{l+1}} \end{bmatrix} = \begin{bmatrix} z^l_1(x) \\ \vdots \\ z^l_k(x) \\ \vdots \\ z^l_{N_{l+1}}(x) \end{bmatrix} \tag{126}$$

9.1 prove by induction

9.1.1 prove by induction

induction works by proving value at $l = 1$, then show relationship between l and $l-1$ in general. Finally, it shows what value is at an arbitrary index L

9.1.2 for NTK

we need to show by induction,

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

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

even better, no need to show: $\Theta^1_{k,k'}(x, x') \to K^1 \delta_{k,k'}$, it is actually equal! (127)

2. then by assuming:

$$\Theta^{l-1}_{k,k'}(x, x') = \frac{\partial z^{l-1}_k(x, \theta)}{\partial \theta^l}^\top \frac{\partial z^{l-1}_k(x', \theta)}{\partial \theta^l} \xrightarrow{N_l \to \infty} \Theta^{l-1}_\infty(x, x') \delta_{k,k'} \tag{128}$$

3. we prove:

$$\Theta^l_{k,k'}(x, x') = \frac{\partial z^l_k(x, \theta)}{\partial \theta^l}^\top \frac{\partial z^l_k(x', \theta)}{\partial \theta^l} \xrightarrow{N_{l+1} \to \infty} \Theta^l_\infty(x, x') \delta_{k,k'} \tag{129}$$

9.2 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.(126), we have:

$$\begin{bmatrix} \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W^1_{1,j}\phi(x_1) + \sigma_b b^1_1 \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W^1_{k,j}\phi(x_k) + \sigma_b b^1_k \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W^1_{N_1,j}\phi(x_{N_1}) + \sigma_b b^1_{N_1} \end{bmatrix} = \begin{bmatrix} z^1_1(x) \\ \vdots \\ z^1_k(x) \\ \vdots \\ z^1_{N_1}(x) \end{bmatrix} \tag{130}$$

when computing: $\frac{\partial z^1_k(x)}{\partial W_{i,j}}$, here, we use i to index entries of W, because k is used 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, i.e., $\frac{\partial z^1_k(x)}{\partial W_{i,j}} = 0$ if $i \neq k$

$$\begin{aligned} \frac{\partial z^l_k(x)}{\partial W_{i,j}} &= \begin{cases} \frac{1}{\sqrt{d_{\text{in}}}} x_i & \text{if } i = k \text{ i.e., row } k \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{1}{\sqrt{d_{\text{in}}}} \delta_{i,k} x_i \\ \implies \frac{\partial z^l_{k'}(x)}{\partial W_{i,j}} &= \frac{1}{\sqrt{d_{\text{in}}}} \delta_{i,k'} x_i \end{aligned} \tag{131}$$

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

$$\begin{aligned} \Theta^1_{k,k'}(x, x') &= \sum_{d=1}^{|\theta^1|} \frac{\partial F^1_k(x)}{\partial \theta^1_d} \frac{\partial F^1_{k'}(x')}{\partial \theta^1_d} \qquad \theta^1 = \{W^1, b^1\} \\ &= \sum_{d=1}^{|W^1|} \frac{\partial F^1_k(x)}{\partial W^1_d} \frac{\partial F^1_{k'}(x')}{\partial W^1_d} + \sum_{d=1}^{|b^1|} \frac{\partial F^1_k(x)}{\partial b^1_d} \frac{\partial F^1_{k'}(x')}{\partial b^1_d} \\ &= \sum_{i=1}^{N_1} \sum_{j=1}^{d_{\text{in}}} \frac{\partial z^1_k(x)}{\partial W_{i,j}} \frac{\partial z^1_{k'}(x')}{\partial W_{i,j}} + \sum_{i=1}^{N_1} \frac{\partial z^1_k(x)}{\partial b_i} \frac{\partial z^1_{k'}(x')}{\partial b_i} \\ &= \sum_{i=1}^{N_1} \sum_{j=1}^{d_{\text{in}}} \frac{1}{\sqrt{d_{\text{in}}}} x_i \delta_{i,k'} \frac{1}{\sqrt{d_{\text{in}}}} x'_i \delta_{i,k} + \sum_{i=1}^{N_1} \sigma_b \delta_{i,k} \, \sigma_b \delta_{i,k'} \quad \text{only one } i \in \{1, \dots N_1\} \text{ in outer sum remain} \\ &= \sum_{j=1}^{d_{\text{in}}} \frac{1}{d_{\text{in}}} x_i x'_i \delta^2_{k,k'} + \sigma_b^2 \delta_{k,k'} \qquad \delta_{i,k'} \delta_{i,k} = \delta_{k,k'} \\ &= \frac{1}{d_{\text{in}}} x^\top x' \delta_{k,k'} + \sigma_b^2 \delta_{k,k'} \\ &= \Big(\underbrace{\frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2}_{K^1} \Big) \delta_{k,k'} \\ &\equiv K^1(x, x') \delta_{k,k'} \quad \text{conform to notation in NNGP section} \end{aligned} \tag{132}$$

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

$$\implies \Theta^1(x,x') = \underbrace{\begin{bmatrix} K^1(x,x') & \dots & 0 & \dots & 0 \\ 0 & K^1(x,x') & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & K^1(x,x') & 0 \\ 0 & 0 & 0 & 0 & K^1(x,x') \end{bmatrix}}_{k\in\{1,\dots N_l\}} \Bigg\} k' \in \{1,\dots N_l\}$$
$$= \text{repeating diagonal with } K^1(x,x')\delta_{k,k'}$$
$$= \underbrace{K^1(x,x')}_{\text{scalar}} \otimes_{\text{outer}} \mathbf{I}_{N_{l+1}\times N_{l+1}} \tag{133}$$

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

9.3 when $l > 1$

$$\begin{bmatrix} \frac{1}{\sqrt{N_l}}\sum_{j=1}^{N_l} W^l_{1,j}\phi(z^{l-1}_j(x)) + \sigma_b b^l_1 \\ \vdots \\ \frac{1}{\sqrt{N_l}}\sum_{j=1}^{N_l} W^l_{k,j}\phi(z^{l-1}_j(x)) + \sigma_b b^l_k \\ \vdots \\ \frac{1}{\sqrt{N_l}}\sum_{j=1}^{N_l} W^l_{N_{l+1},j}\phi(z^{l-1}_j(x)) + \sigma_b b^l_{N_{l+1}} \end{bmatrix} = \begin{bmatrix} z^l_1(x) \\ \vdots \\ z^l_k(x) \\ \vdots \\ z^l_{N_{l+1}}(x) \end{bmatrix} \tag{134}$$

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

$$\begin{aligned} \Theta^l_{k,k'}(x,x') &= \sum_{d=1}^{|\theta^l|} \frac{\partial z^1_k(x)}{\partial \theta^{l-1}_d}\frac{\partial z^l_{k'}(x')}{\partial \theta^{l-1}_d} \\ &= \sum_{d=1}^{|W^l,b^l|} \frac{\partial z^1_k(x)}{\partial\{W^l,b^l\}}\frac{\partial z^l_{k'}(x')}{\partial\{W^l,b^l\}} + \sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z^1_k(x)}{\partial \theta^{l-1}_d}\frac{\partial z^l_{k'}(x')}{\partial \theta^{l-1}_d} \end{aligned} \tag{135}$$

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

$$\begin{aligned} z^l_k &= \frac{1}{\sqrt{N_l}}\sum_{j=1}^{N_l} W^l_{k,j}\phi(z^{l-1}_j(x)) + \sigma_b b^l_k \\ &= \frac{1}{\sqrt{N_l}}\sum_{j=1}^{N_l} W^l_{k,j}\phi\Big(\frac{1}{\sqrt{N_{l-1}}}\sum_{j=1}^{N_{l-1}} W^{l-1}_{j,i}\phi(z^{l-1}_i(x)) + \sigma_b b^{l-1}_j\Big) + \sigma_b b^l_j \end{aligned} \tag{136}$$

$$
\begin{aligned}
\frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} &= \frac{\partial z_k^1(x)}{\partial \phi(z^{l-1}(x))} \frac{\partial \phi(z^{l-1}(x))}{\partial z^{l-1}(x)} \frac{\partial z^{l-1}(x)}{\partial \theta_d^{l-1}} && \text{drop index for the last two terms} \\
&= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \frac{\partial \phi(z_j^{l-1}(x))}{\partial z_j^{l-1}(x)} \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \\
&= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \, \phi'(z_j^{l-1}(x)) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} && \text{leave last derivative as is, in "recursion"}
\end{aligned} \tag{137}
$$

substitution:

$$
\begin{aligned}
&\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}} \\
&= \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) \\
&= \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'} \\
&= \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}} \\
&= \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'} \\
&\qquad \text{subsitute induction criteria: } \Theta_{j,j'}^{l-1}(x,x') \to \underbrace{\Theta_{\infty}^{l-1}(x,x') \delta_{j,j'}}_{\text{deterministic and diagonal limit}} \\
&= \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') \quad \text{change } j' \to j \text{ and remove } \sum_{j'=1}^{N_l} W_{k,j}^l
\end{aligned} \tag{138}
$$

apply CLT, we know that:

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{d=1}^{|\theta^{l-1}|}\frac{\partial z_k^1(x)}{\partial \theta_d^{l-1}}\frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}}\Big] \\
&= \mathbb{E}\Big[W_{k,j}^l\, W_{k',j}^l \phi'(z_j^{l-1}(x))\;\phi'(z_j^{l-1}(x'))\;\Theta_\infty^{l-1}(x,x')\Big] \qquad N_l\to\infty \\
&= \mathbb{E}\Big[W_{k,j}^l\, W_{k',j}^l\Big]\mathbb{E}\Big[\phi'(z_j^{l-1}(x))\;\phi'(z_j^{l-1}(x'))\Big]\;\underbrace{\Theta_\infty^{l-1}(x,x')}_{\text{constant}} \\
&= \delta_{k,k'}\mathbb{E}\Big[\phi'(z_j^{l-1}(x))\;\phi'(z_j^{l-1}(x'))\Big]\;\Theta_\infty^{l-1}(x,x')
\end{aligned}
\tag{139}
$$

we have seen previously Eq. (85):

$$
K^l(x,x') = \sigma_b^2 + \sigma_w^2\,\mathbb{E}_{\left(z_j^{l-1}(x),z_j^{l-1}(x')\right)\sim\mathcal{N}\left(0,K^{l-1}(x,x')\right)}\Big[\phi\big(z_j^{l-1}(x)\big)\phi\big(z_j^{l-1}(x')\big)\Big] \tag{140}
$$

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 $\phi'()$

$$
\begin{aligned}
\dot{K}^l(x,x') &= \sigma_w^2\,\mathbb{E}_{\left(z_j^{l-1}(x),z_j^{l-1}(x')\right)\sim\mathcal{N}\left(0,K^{l-1}(x,x')\right)}\Big[\phi'\big(z_j^{l-1}(x)\big)\phi'\big(z_j^{l-1}(x')\big)\Big] \\
&= \mathbb{E}_{\left(z_j^{l-1}(x),z_j^{l-1}(x')\right)\sim\mathcal{N}\left(0,K^{l-1}(x,x')\right)}\Big[\phi'\big(z_j^{l-1}(x)\big)\phi'\big(z_j^{l-1}(x')\big)\Big] \qquad \text{assume } \sigma_w=1
\end{aligned}
\tag{141}
$$

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

$$
\begin{aligned}
&= \delta_{k,k'}\mathbb{E}_{\left(z_j^{l-1}(x),z_j^{l-1}(x')\right)\sim\mathcal{N}\left(0,K^{l-1}(x,x')\right)}\Big[\phi'\big(z_j^{l-1}(x)\big)\phi'\big(z_j^{l-1}(x')\big)\Big]\;\Theta_\infty^{l-1}(x,x') \\
&= \delta_{k,k'}\,\dot{K}^l(x,x')\;\Theta_\infty^{l-1}(x,x')
\end{aligned}
\tag{142}
$$

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

$$
\sum_{d=1}^{|W^l,b^l|}\frac{\partial z_k^l(x)}{\partial\{W^l,b^l\}}\;\frac{\partial z_{k'}^l(x')}{\partial\{W^l,b^l\}} \tag{143}
$$

and compare that with for $l=1$:

$$
\sum_{d=1}^{|\theta^1|}\frac{\partial F_k^1(x)}{\partial\theta_d^1}\frac{\partial F_{k'}^1(x')}{\partial\theta_d^1} \qquad \theta^1=\{W^1,b^1\} \tag{144}
$$

it's the same if we replace

$$
\Big(K^1(x,x')\equiv\frac{1}{d_{\text{in}}}x^\top x'+\sigma_b^2\Big)\delta_{k,k'} \to \Big(K^l(x,x')\equiv\frac{1}{N_l}\phi\big(z^l(x)\big)^\top\phi\big(z^l(x)\big)+\sigma_b^2\Big)\delta_{k,k'} \tag{145}
$$

$$
\begin{aligned}
\Theta^l_{k,k'}(x,x') &= \sum_{d=1}^{|W^l,b^l|} \frac{\partial z^l_k(x)}{\partial\{W^l,b^l\}} \frac{\partial z^l_{k'}(x')}{\partial\{W^l,b^l\}} + \sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z^1_k(x)}{\partial \theta^{l-1}_d} \frac{\partial z^l_{k'}(x')}{\partial \theta^{l-1}_d} \\
&= K^l(x,x')\ \delta_{k,k'} + \delta_{k,k'} \dot{K}^l(x,x')\ \Theta^{l-1}_\infty(x,x') \\
&= \Big(K^l(x,x')\ + \dot{K}^l(x,x') \Theta^{l-1}_\infty(x,x') \Big) \delta_{k,k'} \\
&\text{repeating diagonal with } K^l(x,x')\ + \dot{K}^l(x,x') \Theta^{l-1}_\infty(x,x') \\
&= \underbrace{\Big(K^l(x,x')\ + \dot{K}^l(x,x') \Theta^{l-1}_\infty(x,x') \Big)}_{\text{scalar}} \otimes_{\text{outer}} \mathbf{I}_{N_{l+1} \times N_{l+1}}
\end{aligned}
\tag{146}
$$

10 NTK during training

Looking at training the Last-layer:

10.1 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:

$$\theta^{L+1} = (W^{L+1}, b^{L+1}) \tag{147}$$

then, objective is:

$$\begin{aligned} C &= \frac{1}{2}\|f(x) - y\|_2^2 \\ &= \frac{1}{2}\left\| \left(\frac{\sigma_w}{\sqrt{N_l}} W^{L+1} \phi\big(z^L(x)\big) + \sigma_b b^{L+1} \right) - y \right\|_2^2 \end{aligned} \tag{148}$$

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:

$$W^{L+1} \sim \mathcal{N}(0,1) \quad \text{ and } b^{L+1} \sim \mathcal{N}(0,1) \tag{149}$$

then, the above is written as:

$$\begin{aligned} C &= \frac{1}{2}\left\| \left(\underbrace{\begin{bmatrix} W^{L+1} & b^{L+1} \end{bmatrix}}_{\theta^{L+1}} \underbrace{\begin{bmatrix} \frac{\sigma_w}{\sqrt{N_l}} \phi\big(z^L(x)\big) & \sigma_b \end{bmatrix}^\top}_{\bar{a}(x)} \right) - y \right\|_2^2 \\ &= \frac{1}{2}\left\| \big(\bar{a}(x)^\top \theta^{L+1}\big) - y \right\|_2^2 \\ &= \frac{1}{2}\left\| \hat{y}_t(x) - y \right\|_2^2 \\ \implies \frac{\partial C}{\partial \theta^{L+1}} &= \bar{a}(x)^\top \big(\bar{a}(x)\theta^{L+1} - y\big) \end{aligned} \tag{150}$$

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

also note that if only read-out layer is trained then $\bar{a}(x)$ is constant in t:

$$\begin{aligned} \frac{\partial \hat{y}_t(x)}{\partial \theta_t} &= \bar{a}(x) \\ \implies \frac{\partial \hat{y}_t(x)}{\partial \theta_t}^\top \frac{\partial \hat{y}_t(x)}{\partial \theta_t} &= \bar{a}(x)^\top \bar{a}(x) = \Theta_t(x,x) = \Theta_0(x,x) \end{aligned} \tag{151}$$

10.2 entire dataset $\mathcal{X}$:

10.2.1 softmax:

$$\hat{y}_t(\mathcal{X}) = \begin{bmatrix} \hat{y}_t^1(x_1) \\ \vdots \\ \hat{y}_t^{N^{L+1}}(x_1) \\ \vdots \\ \hat{y}_t^1(x_k) \\ \vdots \\ \hat{y}_t^{N^{L+1}}(x_k) \\ \vdots \\ \hat{y}_t^1(x_{|D|}) \\ \vdots \\ \hat{y}_t^{N^{L+1}}(x_{|D|}) \end{bmatrix} = \text{vec}\big([\hat{y}_t^i(x)]_{x\in\mathcal{X}}\big) \in \mathcal{R}^{N^{L+1}\times|D|\times 1} \tag{152}$$

10.2.2 mean-square error:

we focus on MSE:

$$\begin{aligned} C &= \frac{1}{2}\big\|\big(\bar{a}(\mathcal{X})^\top\theta^{L+1}\big) - \mathcal{Y}\big\|_2^2 \\ &= \frac{1}{2}\big\|\hat{y}_t(\mathcal{X}) - \mathcal{Y}\big\|_2^2 \\ \implies \frac{\partial C}{\partial\theta^{L+1}} &= \bar{a}(\mathcal{X})^\top\big(\bar{a}(\mathcal{X})\theta^{L+1} - \mathcal{Y}\big) \end{aligned} \tag{153}$$

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

$$\implies \frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta} = \begin{bmatrix} \frac{\mathrm{d}\hat{y}_t(x_1)}{\mathrm{d}\theta_1} & \cdots & \frac{\mathrm{d}\hat{y}_t(x_1)}{\mathrm{d}\theta_{|\theta|}} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}\hat{y}_t(x_k)}{\mathrm{d}\theta_1} & \cdots & \frac{\mathrm{d}\hat{y}_t(x_k)}{\mathrm{d}\theta_{|\theta|}} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}\hat{y}_t(x_{|D})}{\mathrm{d}\theta_1} & \cdots & \frac{\mathrm{d}\hat{y}_t(x_{|D})}{\mathrm{d}\theta_{|\theta|}} \end{bmatrix} \tag{154}$$

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

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

10.3 what is linearized model: $f(x,\theta_t) \to f_t^{\text{lin}}(x,\theta) \equiv f^{\text{lin}}(x,\theta_t)$

$$f_t^{\text{lin}}(x) \equiv f_0(x) + \underbrace{\frac{\partial f(x,\theta)}{\partial \theta}\Big|_{\theta\to\theta_0}}_{\text{constant in } t} \underbrace{(\theta_t - \theta_0)}_{\triangle\theta(t)}$$

$$= f_0(x) + \frac{\partial f(x,\theta)}{\partial \theta}\Big|_{\theta\to\theta_0} \triangle\theta(t) \qquad \text{note } \triangle\theta(t) \text{ refer to change, irrespective of starting position } \theta_0 \tag{155}$$

10.4 Sketch of Proof to compute $f_t^{\text{lin}}(x,\theta)$

We are interested to study the behavior of $f_t^{\text{lin}}(x)$ for a singular data x as θ_t evolves

the expression $f_t^{\text{lin}}(x)$ can be misleading: it should be written instead as $f_t^{\text{lin}}(x|\mathcal{X})$ as it depends on training dataset $(\mathcal{X},\mathcal{Y})$, and θ as well.

Note that we can interchangeably write:

$$f_t^{\text{lin}}(\mathcal{X},\theta) \equiv f_t^{\text{lin}}(\mathcal{X}) \equiv f^{\text{lin}}(\mathcal{X},\theta_t) \tag{156}$$

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

1. find expression for $f_t^{\text{lin}}(\mathcal{X},\theta)$: it has two versions:
   using gradient descend:

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = -\eta\left(\frac{\partial f_t^{\text{lin}}(\mathcal{X},\theta)}{\partial\theta}\right)^\top \frac{\partial C}{\partial f_t^{\text{lin}}(\mathcal{X},\theta)} \tag{157}$$

then,

version 1: assume $f_t^{\text{lin}}(\mathcal{X})$ is a linear model

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

(b) then, obtain $f_t^{\text{lin}}(\mathcal{X},\theta) = \left(\frac{\partial f_t^{\text{lin}}(\mathcal{X})}{\partial\theta}\right)^\top \triangle\theta(t)$

version 2:

(a) obtain expression for:

$$\frac{\mathrm{d}\hat{y}}{\mathrm{d}t} = \left(\frac{\partial f_t^{\text{lin}}(\mathcal{X})}{\partial\theta}\right)^\top \frac{\mathrm{d}\theta}{\mathrm{d}t} \tag{158}$$

(b) then, use ODE to obtain $f_t^{\text{lin}}(\mathcal{X},\theta)$

2. now we have expression of $f_t^{\text{lin}}(\mathcal{X},\theta)$, substitute it back to gradient descend:

$$\frac{\mathrm{d}\triangle\theta(t)}{\mathrm{d}t} = \frac{\mathrm{d}\theta}{\mathrm{d}t} = -\eta\left(\frac{\partial f_t^{\text{lin}}(\mathcal{X},\theta)}{\partial\theta}\right)^\top \frac{\partial C}{\partial f_t^{\text{lin}}(\mathcal{X},\theta)} \tag{159}$$

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

$$\triangle\theta(t) = \theta_t - \theta_0 \tag{160}$$

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

$$f_t^{\text{lin}}(x,\theta_t) = f(x,\theta_0) + \left.\frac{\mathrm{d}f(x,\theta_t)}{\mathrm{d}t}\right|_{\theta\to\theta_0} \triangle\theta(t) \tag{161}$$

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

$$\theta_t = \theta_0 + \triangle\theta(t) \tag{162}$$

dependency chain is:

$$(\mathcal{X},\mathcal{Y}) \quad\to\quad \frac{\mathrm{d}\triangle\theta(t)}{\mathrm{d}t} \quad\to\quad \triangle\theta(t) \quad\to\quad \hat{y}_t^{\text{lin}}(x,\theta) \tag{163}$$

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, $f(\mathcal{X},\theta)$ and $\mathcal{Y}$, this is in Section (10.6)
- then we can work out how $\frac{\mathrm{d}\theta}{\mathrm{d}t}$ may impact $f_t^{\text{lin}}(x)$
- simply obtained expression of $f_t^{\text{lin}}(\mathcal{X})$ won't give you expression for $f_t^{\text{lin}}(x)$

10.5 General linear ODE solution

We need basic tools on ODE solution:
looking at the equation, treating everything in 1-d:

$$\begin{aligned}
\dot{x} &= Ax + b \\
\implies \frac{\dot{x}}{Ax+b} &= 1 \\
\implies \frac{\dot{x}}{x+\frac{b}{A}} &= A \\
\implies \frac{\mathrm{d}}{\mathrm{d}t}\log\left(x+\frac{b}{A}\right) &= A \qquad \text{easy to see:} \quad \frac{\mathrm{d}}{\mathrm{d}t}\log\left(x+\frac{b}{A}\right) = \frac{\frac{\mathrm{d}x}{\mathrm{d}t}}{x+\frac{b}{A}} \\
\implies \int_t \frac{\mathrm{d}}{\mathrm{d}t}\log\left(x+\frac{b}{A}\right)\mathrm{d}t &= \int_t A\mathrm{d}t \\
\log\left(x+\frac{b}{A}\right) &= At + h \\
x+\frac{b}{A} &= \exp\left(At+h\right) \\
\implies x(t) &= -\frac{b}{A} + C\exp\left(At\right)
\end{aligned} \tag{164}$$

when things are in multi-dimensions:

$$x(t) = -A^{-1}b + C\exp\left(At\right) \quad b \text{ is column vector} \tag{165}$$

let $t = 0$:

$$\begin{aligned}
x(0) &= -A^{-1}b + C \\
\implies C &= \left(x(0) + A^{-1}b\right)
\end{aligned} \tag{166}$$

substitute in C:

$$x(t) = -A^{-1}b + \left(x(0) + A^{-1}b\right)\exp\left(At\right) \tag{167}$$

10.6 find expression for $f_t^{\text{lin}}(\mathcal{X}, \theta)$:

10.6.1 generic MSE loss for f_t^{lin}

we use **version 2**:

$$\begin{aligned}
\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} &= -\eta\left(\frac{\partial f_t^{\text{lin}}(\mathcal{X},\theta)}{\partial\theta}\right)^\top \frac{\partial C}{\partial f_t^{\text{lin}}(\mathcal{X},\theta)} \\
&= -\eta\left(\frac{\partial f_0(\mathcal{X},\theta)}{\partial\theta}\right)^\top \frac{\partial C}{\partial f_t^{\text{lin}}(\mathcal{X},\theta)} \qquad \text{using property of linearlised model Eq.(155)} \\
\implies \frac{\mathrm{d}f_t^{\text{lin}}(\mathcal{X},\theta)}{\mathrm{d}t} &= \frac{\partial f_t^{\text{lin}}(\mathcal{X},\theta)}{\partial\theta_t}\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} \\
&= -\eta\frac{\partial f_t^{\text{lin}}(\mathcal{X},\theta)}{\partial\theta}\left(\frac{\partial f_0(\mathcal{X},\theta)}{\partial\theta}\right)^\top \frac{\partial C}{\partial f_t^{\text{lin}}(\mathcal{X},\theta)} \\
&= -\eta\Theta_0(\mathcal{X},\mathcal{X})\frac{\partial C}{\partial f_t^{\text{lin}}(\mathcal{X},\theta)}
\end{aligned} \tag{168}$$

assume C is MSE:

$$\begin{aligned}\frac{\mathrm{d}f_t^{\text{lin}}(\mathcal{X},\theta)}{\mathrm{d}t} &= -\eta\Theta_0(\mathcal{X},\mathcal{X})\big(f_t^{\text{lin}}(\mathcal{X})-\mathcal{Y}\big)\\ &= \underbrace{-\eta\Theta_0(\mathcal{X},\mathcal{X})}_{A} f_t^{\text{lin}}(\mathcal{X}) + \underbrace{\eta\Theta_0(\mathcal{X},\mathcal{X})\mathcal{Y}}_{b}\end{aligned} \tag{169}$$

so by substitution, using Eq. (167), we have:

$$\begin{aligned}&A = -\eta\Theta_0(\mathcal{X},\mathcal{X})\\ \implies\ &A^{-1} = \frac{-1}{\eta}\big(\Theta_0(\mathcal{X},\mathcal{X})\big)^{-1}\\ &b = \eta\Theta_0(\mathcal{X},\mathcal{X})\mathcal{Y}\end{aligned} \tag{170}$$

we have:

$$\begin{aligned}&x(t) = -A^{-1}b + \big(x(0)+A^{-1}b\big)\exp\Big(At\Big)\\ \implies\ &\theta^{L+1}(t) = -\frac{-1}{\eta}\Big(\Theta_0(\mathcal{X},\mathcal{X})\Big)^{-1}\Big(\eta\Theta_0(\mathcal{X},\mathcal{X})\mathcal{Y}\Big)\\ &\qquad + \Big[f_0(\mathcal{X}) + \frac{-1}{\eta}\Big(\Theta_0(\mathcal{X},\mathcal{X})\Big)^{-1}\Big(\eta\Theta_0(\mathcal{X},\mathcal{X})\mathcal{Y}\Big)\Big]\ \exp\Big(-\eta\Theta_0(\mathcal{X},\mathcal{X})t\Big)\\ &\qquad = \mathcal{Y} + \big(f_0(\mathcal{X})-\mathcal{Y}\big)\ \exp\big(-\eta\Theta_0(\mathcal{X},\mathcal{X})t\big)\end{aligned} \tag{171}$$

solution is:

$$f_t^{\text{lin}}(\mathcal{X},\theta) = \mathcal{Y} + \exp\big(-\eta\Theta_0(\mathcal{X},\mathcal{X})\,t\big)\big(f_0(\mathcal{X})-\mathcal{Y}\big) \tag{172}$$

10.6.2 read-out layer only under MSE

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:

$$\begin{aligned}C &= \frac{1}{2}\big\|\big(\bar{a}(\mathcal{X})^\top\theta^{L+1}\big)-\mathcal{Y}\big\|_2^2\\ &= \frac{1}{2}\big\|\hat{y}_t(\mathcal{X})-\mathcal{Y}\big\|_2^2\end{aligned} \tag{173}$$

$$\begin{aligned}\frac{\mathrm{d}\theta^{L+1}(t)}{\mathrm{d}t} &= -\eta\bigg(\frac{\partial\hat{y}_t(\mathcal{X},\theta^{L+1})}{\partial\theta^{L+1}}\bigg)^\top\frac{\partial C}{\partial\hat{y}_t(\mathcal{X},\theta^{L+1})}\\ &= -\eta\bar{a}(\mathcal{X})^\top\big(\bar{a}(\mathcal{X})\theta^{L-1}-\mathcal{Y}\big)\\ &= \underbrace{-\eta\bar{a}(\mathcal{X})^\top\bar{a}(\mathcal{X})}_{A}\theta^{L-1}(t) + \underbrace{\eta\bar{a}(\mathcal{X})^\top\mathcal{Y}}_{b}\end{aligned} \tag{174}$$

so by substitution, using Eq. (167), we have:

$$\begin{aligned}&A = -\eta\bar{a}(\mathcal{X})^\top\bar{a}(\mathcal{X})\\ \implies\ &A^{-1} = \frac{-1}{\eta}\big(\bar{a}(\mathcal{X})^\top\bar{a}(\mathcal{X})\big)^{-1}\\ &b = \eta\bar{a}(\mathcal{X})^\top\mathcal{Y}\end{aligned} \tag{175}$$

$$
\begin{aligned}
x(t) &= -A^{-1}b + \big(x(0) + A^{-1}b\big)\exp\big(At\big) \\
\implies \theta^{L+1}(t) &= -\frac{-1}{\eta}\Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\eta\bar{a}(\mathcal{X})^\top \mathcal{Y}\Big) \\
&\quad + \left[\theta^{L+1}(0) + \frac{-1}{\eta}\Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\eta\bar{a}(\mathcal{X})^\top \mathcal{Y}\Big)\right]\exp\Big(-\eta\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})t\Big) \\
&= \Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\bar{a}(\mathcal{X})^\top \mathcal{Y}\Big) \\
&\quad + \left[\theta^{L+1}(0) - \Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\bar{a}(\mathcal{X})^\top \mathcal{Y}\Big)\right]\exp\Big(-\eta\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})t\Big)
\end{aligned}
\tag{176}
$$

look at MSE again: $\hat{y}_t(x) = \bar{a}(x)^\top \theta^{L+1}$:

$$
\begin{aligned}
\hat{y}_t(\mathcal{X}) = \theta^{L+1}(t)\bar{a}(\mathcal{X}) &= \Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\bar{a}(\mathcal{X})^\top \mathcal{Y}\,\bar{a}(\mathcal{X})\Big) \\
&\quad + \left[\theta^{L+1}(0)\bar{a}(\mathcal{X}) - \Big(\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})\Big)^{-1}\Big(\bar{a}(\mathcal{X})^\top \mathcal{Y}\bar{a}(\mathcal{X})\Big)\right]\exp\Big(-\eta\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})t\Big) \\
&= \mathcal{Y} + \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\exp\big(-\eta\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})t\big) \\
&= \mathcal{Y} + \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})\big)
\end{aligned}
\tag{177}
$$

$\hat{y}_t(\mathcal{X})$ can be written in various forms:

$$
\begin{aligned}
\hat{y}_t(\mathcal{X}) &= \underbrace{\mathcal{Y}}\ + \hat{y}_0(\mathcal{X})\exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big) - \underbrace{\mathcal{Y}\exp\big(-\eta\bar{a}(\mathcal{X})^\top \bar{a}(\mathcal{X})t\big)} \\
&= \big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})\,t\big)\big)\mathcal{Y} + \hat{y}_0(\mathcal{X})\exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})\,t\big)
\end{aligned}
\tag{178}
$$

relationship between $\hat{y}_t(\mathcal{X})$ and $\frac{\mathrm{d}\hat{y}_t(\mathcal{X})}{\mathrm{d}t}$:

$$
\begin{aligned}
\hat{y}_t(\mathcal{X}) &= \mathcal{Y} + \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big) \\
\frac{\mathrm{d}\hat{y}_t(\mathcal{X})}{\mathrm{d}t} &= -\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big) \\
&= -\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\big(\hat{y}_t(\mathcal{X}) - \mathcal{Y}\big)
\end{aligned}
\tag{179}
$$

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

10.6.3 Two are equal!

looking at Eq. (177) and compare it with Eq. (172) side by side:

$$
\begin{aligned}
\hat{y}_t(\mathcal{X}) &= \mathcal{Y} + \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})\big) \\
f_t^{\text{lin}}(\mathcal{X},\theta) &= \mathcal{Y} + \exp\big(-\eta\Theta_0(\mathcal{X},\mathcal{X})\,t\big)\big(f_0(\mathcal{X}) - \mathcal{Y}\big)
\end{aligned}
\tag{180}
$$

it looks the same as long as $\mathcal{K}(\mathcal{X},\mathcal{X}) \equiv \Theta_0(\mathcal{X},\mathcal{X})$, but wait, if one look at Eq. (132)

$$
\Theta^1_{k,k'}(x,x') \equiv K^1(x,x')\delta_{k,k'}
$$

since we fixed all the preceeding layers, and only train readout layers, so it's like just K^1 using last-layer post-activation as oppose to input x and x'

this means all thereafter sections we can using either one in Eq.(180)

10.7 computing $\triangle\theta(t)$

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

$$\begin{aligned}\frac{\mathrm{d}\theta}{\mathrm{d}t} &= -\eta\frac{\partial C}{\partial\theta}\\ &= -\eta\left(\frac{\partial\hat{y}_t(\mathcal{X})}{\partial\theta}\right)^\top\frac{\partial C}{\partial\hat{y}_t(\mathcal{X})}\\ &= -\eta\left(\frac{\partial\hat{y}_t(\mathcal{X})}{\partial\theta}\right)^\top(\hat{y}_t(\mathcal{X})-\mathcal{Y})\\ &= -\eta\left(\frac{\partial\hat{y}_t(\mathcal{X})}{\partial\theta}\right)^\top(\hat{y}_0(\mathcal{X})-\mathcal{Y})\ \exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\quad\text{using equation (179)}\end{aligned}\tag{181}$$

to work out $\triangle\theta(t)$:

$$\begin{aligned}\triangle\theta(t) &= \int_{\tau=0}^{t}\frac{\mathrm{d}\theta}{\mathrm{d}\tau}\mathrm{d}\tau\\ &= \int_{\tau=0}^{t}-\eta\left(\frac{\partial\hat{y}_t(\mathcal{X})}{\partial\theta}\right)^\top(\hat{y}_0(\mathcal{X})-\mathcal{Y})\ \exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\tau\big)\mathrm{d}\tau\\ &= -\eta\left(\frac{\partial\hat{y}_t(\mathcal{X})}{\partial\theta}\right)^\top(\hat{y}_0(\mathcal{X})-\mathcal{Y})\ \underbrace{\int_{\tau=0}^{t}\exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\tau\big)\mathrm{d}\tau}\end{aligned}\tag{182}$$

looking at:

$$\begin{aligned}&\int_{\tau=0}^{t}\exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\tau\big)\mathrm{d}\tau\\ &= \left[-\frac{1}{\eta}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})\tau\big)\right]_{\tau=0}^{t}\\ &= -\frac{1}{\eta}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)+\frac{1}{\eta}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\\ &= \underbrace{\frac{1}{\eta}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\big(\mathbf{I}-\exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\big)}\end{aligned}\tag{183}$$

Finally $\triangle\theta(t)$:

$$\begin{aligned}\triangle\theta(t) &= -\eta\left(\frac{\partial\hat{y}_t(\mathcal{X})}{\partial\theta}\right)^\top(\hat{y}_0(\mathcal{X})-\mathcal{Y})\ \frac{1}{\eta}\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I}-\exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\Big)\\ &= -\left(\frac{\partial\hat{y}_t(\mathcal{X})}{\partial\theta}\right)^\top(\hat{y}_0(\mathcal{X})-\mathcal{Y})\ \hat{\mathcal{K}}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I}-\exp\big(-\eta\hat{\mathcal{K}}(\mathcal{X},\mathcal{X})t\big)\Big)\end{aligned}\tag{184}$$

10.8 putting togehter for $\hat{y}(x, \theta_t)$

$$
\begin{aligned}
\hat{y}(x, \theta_t) &= \hat{y}(x, \theta_0) + \frac{\partial \hat{y}(x, \theta)}{\partial \theta}\Big|_{\theta \to \theta_0} \triangle\theta(t) \\
&= \hat{y}(x, \theta_0) + \bar{a}(x)\Big(- \frac{\partial \hat{y}_t(\mathcal{X})}{\partial \theta}\Big)^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})t\big)\Big) \\
&= \hat{y}(x, \theta_0) - \bar{a}(x)\bar{a}(\mathcal{X})^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})t\big)\Big) \\
&= \hat{y}(x, \theta_0) - \bar{a}(x)\bar{a}(\mathcal{X})^\top \big(\hat{y}_0(\mathcal{X}) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})t\big)\Big) \\
&= \hat{y}(x, \theta_0) - \hat{\mathcal{K}}(x, \mathcal{X})\big(\hat{y}(\mathcal{X}, \theta_0) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})t\big)\Big)
\end{aligned}
\tag{185}
$$

10.8.1 simillarly for f_t^{lin}

looking at pattern, you will see that:

$$
f_t^{\text{lin}}(x, \theta) = f(x, \theta_0) - \Theta_0(x, \mathcal{X})\big(f(\mathcal{X}, \theta_0) - \mathcal{Y}\big)\ \Theta_0(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\Theta_0(\mathcal{X}, \mathcal{X})t\big)\Big) \tag{186}
$$

10.9 expectation and variance

mean

$$
\begin{aligned}
\mathbb{E}\big[\hat{y}(x, \theta_t)\big] &= \mathbb{E}\Big[\hat{y}(x, \theta_0) - \hat{\mathcal{K}}(x, \mathcal{X})\big(\hat{y}(\mathcal{X}, \theta_0) - \mathcal{Y}\big)\ \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})t\big)\Big)\Big] \\
&= \underbrace{\mathbb{E}\big[\hat{y}(x, \theta_0)\big]}_{=0} - \underbrace{\mathbb{E}\big[\hat{y}(\mathcal{X}, \theta_0)\big]}_{=0}\ \mathbb{E}\Big[\hat{\mathcal{K}}(x, \mathcal{X})\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})t\big)\Big)\Big] \\
&\qquad + \mathbb{E}\Big[\hat{\mathcal{K}}(x, \mathcal{X})\mathcal{Y}\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})t\big)\Big)\Big] \\
&= \mathbb{E}\Big[\hat{\mathcal{K}}(x, \mathcal{X})\mathcal{Y}\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})t\big)\Big)\Big] \\
&= \mathcal{K}(x, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\mathcal{K}(\mathcal{X}, \mathcal{X})t\big)\Big)\mathcal{Y} \quad \text{deterministic in infinite width}
\end{aligned}
\tag{187}
$$

variance

$$
\begin{aligned}
&\hat{y}(x, \theta_t) - \mathbb{E}\big[\hat{y}(x, \theta_t)\big] \qquad \text{let infinite width} \\
=&\hat{y}(x, \theta_0) - \mathcal{K}(x, \mathcal{X})\ \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\mathcal{K}(\mathcal{X}, \mathcal{X})t\big)\Big)\big(\hat{y}(\mathcal{X}, \theta_0) - \mathcal{Y}\big) \\
&\qquad - \mathcal{K}(x, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\mathcal{K}(\mathcal{X}, \mathcal{X})t\big)\Big)\mathcal{Y} \\
=&\hat{y}(x, \theta_0) - \mathcal{K}(x, \mathcal{X})\ \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\mathcal{K}(\mathcal{X}, \mathcal{X})t\big)\Big)\ \hat{y}(\mathcal{X}, \theta_0) \\
&\qquad + \mathcal{K}(x, \mathcal{X})\ \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\mathcal{K}(\mathcal{X}, \mathcal{X})t\big)\Big)\mathcal{Y} \\
&\qquad - \mathcal{K}(x, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\mathcal{K}(\mathcal{X}, \mathcal{X})t\big)\Big)\mathcal{Y} \\
=&\hat{y}(x, \theta_0) - \mathcal{K}(x, \mathcal{X})\ \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(- \eta\mathcal{K}(\mathcal{X}, \mathcal{X})t\big)\Big)\ \hat{y}(\mathcal{X}, \theta_0)
\end{aligned}
\tag{188}
$$

then:

$$
\begin{aligned}
&\mathrm{Var}\big[\hat{y}(x,\theta_t)\big] \\
&= \mathbb{E}\Big[\Big(\hat{y}(x,\theta_t) - \mathbb{E}[\hat{y}(x,\theta_t)]\Big)^\top \Big(\hat{y}(x,\theta_t) - \mathbb{E}[\hat{y}(x,\theta_t)]\Big)\Big] \\
&= \mathbb{E}\Big[\Big(\hat{y}(x,\theta_0) - \underbrace{\mathcal{K}(x,\mathcal{X})\,\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\,\hat{y}(\mathcal{X},\theta_0)}\Big) \\
&\qquad\qquad \Big(\hat{y}(x,\theta_0) - \underbrace{\mathcal{K}(x,\mathcal{X})\,\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\hat{y}(\mathcal{X},\theta_0)}\Big)^\top\Big]
\end{aligned}
\tag{189}
$$

knowing that when $t = 0$:

$$
\begin{aligned}
\mathrm{Cov}\Big[\hat{y}(x,\theta_0)\,,\,\hat{y}(\mathcal{X},\theta_0)\Big] &= \mathbb{E}\Big[\hat{y}(x,\theta_0)\hat{y}(\mathcal{X},\theta_0)^\top\Big] = \mathcal{K}(x,\mathcal{X}) \\
\mathrm{Cov}\Big[\hat{y}(\mathcal{X},\theta_0)\,,\,\hat{y}(x,\theta_0)\Big] &= \mathbb{E}\Big[\hat{y}(\mathcal{X},\theta_0)\hat{y}(x,\theta_0)^\top\Big] = \mathcal{K}(\mathcal{X},x) \\
\mathrm{Var}\Big[\hat{y}(x,\theta_0)\,,\,\hat{y}(x,\theta_0)\Big] &= \mathbb{E}\Big[\hat{y}(x,\theta_0)\hat{y}(x,\theta_0)^\top\Big] = \mathcal{K}(x,x) \\
&\qquad \Big(\mathcal{K}(x,\mathcal{X})\ \underbrace{\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)}_{\text{symmetric}}\ \hat{y}(\mathcal{X},\theta_0)\Big)\Big)^\top \\
&\qquad = \Big(\hat{y}(\mathcal{X},\theta_0)^\top\ \underbrace{\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)}_{\text{symmetric}}\ \mathcal{K}(\mathcal{X},x)\ \Big)
\end{aligned}
\tag{190}
$$

$$
\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\ldots \\
=& \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}
\tag{191}
$$

terms outside of the red bits are the same:

$$
\mathrm{Var}\big[\hat{y}(x,\theta_t)\big] = \mathcal{K}(x,x) - \mathcal{K}(x,\mathcal{X})\,\mathcal{K}(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-2\eta\mathcal{K}(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{K}(\mathcal{X},x) \tag{192}
$$

10.9.1 similar in f_t^{lin}

- mean:

$$\mathbb{E}\big[f_t^{\text{lin}}(x,\theta_t)\big] = \Theta(x,\mathcal{X})\Theta_0(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-\eta\Theta_0(\mathcal{X},\mathcal{X})t\big)\Big)\mathcal{Y} \tag{193}$$

- variance:

$$\text{Var}\big[f_t^{\text{lin}}(x,\theta_t)\big] = \Theta_0(x,x) - \Theta_0(x,\mathcal{X})\ \Theta_0(\mathcal{X},\mathcal{X})^{-1}\Big(\mathbf{I} - \exp\big(-2\eta\Theta_0(\mathcal{X},\mathcal{X})t\big)\Big)\Theta_0(\mathcal{X},x) \tag{194}$$

11 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:

$$\left.\begin{array}{l} \sup\limits_{t\geq 0} \|f_t(x) - f_t^{\text{lin}}\|_2 \\ \sup\limits_{t\geq 0} \dfrac{\|\theta_t - \theta_0\|_2}{\sqrt{n}} \\ \sup\limits_{t\geq 0} \|\hat{\Theta}_t - \hat{\Theta}_0\|_F \end{array}\right\} = \mathcal{O}(n^{-\frac{1}{2}}) \quad \text{as} \quad n \to \infty \tag{195}$$

check relevant publications for the proof