Machine Learning Theory Lecture 4: Neural Network Gaussian Process and NTK

Richard Xu

September 29, 2021

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

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

1.1 marginal and conditional marginal under noisy output

- in a regression setting:

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

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

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

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

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

1.2 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:

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

- **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:

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

2 Kernel methods

consider the equation, where $\phi(\cdot) \in \mathbb{R}^m$:

$$
\begin{aligned}
y &= \phi(x)^\top \boldsymbol{w} \\
&= \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_m(x) \end{bmatrix}^\top \boldsymbol{w} \\
&= \begin{bmatrix} \phi_1(x) & \dots & \phi_m(x) \end{bmatrix} \boldsymbol{w}
\end{aligned} \tag{1}
$$

using definition:

$$
\begin{aligned}
\mathcal{Y} &= [y_1, \dots, y_n]^\top \\
\Phi &= [\phi(x_1), \dots, \phi(x_n)]^\top \\
&= \underbrace{\begin{bmatrix} \phi_1(x_1) & \dots & \phi_m(x_1) \\ \vdots & \vdots & \vdots \\ \phi_1(x_n) & \dots & \phi_m(x_n) \end{bmatrix}}_{n \times m}
\end{aligned} \tag{2}
$$

Ridge regression can be re-written as:

$$
\begin{aligned}
\boldsymbol{w}^\star &= \arg\min_{\boldsymbol{w}} \sum_{i=1}^{n} \left(y_i - \phi(x_i)^\top \boldsymbol{w}\right)^2 + \lambda \|\boldsymbol{w}\|_2^2 \\
&= \arg\min_{\boldsymbol{w}} \|\mathcal{Y} - \Phi \boldsymbol{w}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2
\end{aligned} \tag{3}
$$

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

$$
\boldsymbol{w}^\star = \big(\underbrace{\Phi^\top \Phi}_{m \times m} + \lambda I\big)^{-1} \Phi^\top \mathcal{Y} \tag{4}
$$

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

$$
\begin{aligned}
y_{\boldsymbol{w}^\star}(x) &= \phi(x)^\top \boldsymbol{w}^\star \\
&= \phi(x)^\top \left(\Phi^\top \Phi + \lambda I\right)^{-1} \Phi^\top \mathcal{Y} \\
&= \underbrace{\phi(x)^\top \Phi^\top}_{1 \times n} \big(\underbrace{\Phi \Phi^\top}_{n \times n} + \lambda I\big)^{-1} \mathcal{Y} \\
&\text{using identity } \left(\Phi^\top \Phi + \lambda I\right)^{-1} \Phi^\top = \Phi^\top \left(\Phi \Phi^\top + \lambda I\right)^{-1}
\end{aligned} \tag{5}
$$

2.1 Kernel trick

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

$$
\begin{aligned}
\left[\Phi \Phi^\top\right]_{i,j} &= \phi(x_i)^\top \phi(x_j) = \mathcal{K}(x_i, x_j) \\
\left[\phi(x)^\top \Phi^\top\right]_j &= \phi(x)^\top \phi(x_j) = \mathcal{K}(x, x_j)
\end{aligned} \tag{6}
$$

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

3 Neutral network with Gaussian initialization

$$
\begin{aligned}
z_k^l(x) &= b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \times \phi\Big(z_j^{l-1}(x)\Big) \qquad W_{k,j}^l \sim \mathcal{N}\Big(0, \frac{1}{\sqrt{N_l}}\Big) \quad b_k^l \sim \mathcal{N}\big(0, \sigma_b\big) \quad \text{or :} \\
z_k^l(x) &= \sigma_b b_k^l + \sum_{j=1}^{N_l} \frac{1}{\sqrt{N_l}} W_{k,j}^l \times \phi\Big(z_j^{l-1}(x)\Big) \qquad W_{k,j}^l \sim \mathcal{N}\big(0,1\big) \quad b_k^l \sim \mathcal{N}\big(0,1\big)
\end{aligned} \tag{7}
$$

4 Neural Network Expressivity in Gaussian Process

this is to paraphrase [1] [2]

4.1 pre-activation layer 1

putting in data $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$, we have:

$$
z^1(\mathbf{x}) = \begin{bmatrix} z_1^1 \\ \vdots \\ z_{N_2}^1 \end{bmatrix} = \begin{bmatrix} W_{1,1}^1 & \dots & W_{1,d_{\text{in}}}^1 \\ \vdots & \ddots & \vdots \\ W_{N_2,1}^1 & \dots & W_{N_2,d_{\text{in}}}^1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_{d_{\text{in}}} \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix} \tag{8}
$$

similarly, we can have another expression for $\mathbf{x}' \in \mathbb{R}^d$

$$
z^1(\mathbf{x}') = \begin{bmatrix} z_1^1 \\ \vdots \\ z_{N_2}^1 \end{bmatrix} = \begin{bmatrix} W_{1,1}^1 & \dots & W_{1,d_{\text{in}}}^1 \\ \vdots & \ddots & \vdots \\ W_{N_2,1}^1 & \dots & W_{N_2,d_{\text{in}}}^1 \end{bmatrix} \begin{bmatrix} x'_1 \\ \vdots \\ x'_{d_{\text{in}}} \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_{N_2} \end{bmatrix} \tag{9}
$$

Obviously, regardless if we use $(\mathbf{x}, \mathbf{x})$ or $(\mathbf{x}, \mathbf{x}')$, when $k \neq k'$:

$$
\begin{cases} \text{Cov}\big(z_k^1(\mathbf{x}), z_{k'}^1(\mathbf{x})\big) & = 0 \\ \text{Cov}\big(z_k^1(\mathbf{x}), z_{k'}^1(\mathbf{x}')\big) & = 0 \end{cases} \quad \forall k \neq k' \tag{10}
$$

4.1.1 $p(z_k^1(\mathbf{x}))$

This should really be written as $p(z_k^1(\mathbf{x})\ |\mathbf{x})$ to be inline with $p(z_k^l(\mathbf{x})\ |z^{l-1})$, using the definition at:

$$
\begin{aligned}
z_k^1(\mathbf{x}) &= \sum_{j=1}^{d_{\text{in}}} W_{k,j}^1 x_j + b_k = \sum_{j=1}^{N_1} W_{k,j}^1 x_j + b_k \\
\implies z_k^1(\mathbf{x}) &\sim \mathcal{N}\Big(0, \sigma_b^2 + \sum_{j=1}^{N_1} \Big(\frac{1}{\sqrt{N_1}} x_j\Big)^2\Big) \\
&= \mathcal{N}\Big(0,\ \sigma_b^2 + \frac{1}{N_1} \sum_{j=1}^{N_1} x_j^2\Big) = \mathcal{N}\Big(0,\ \sigma_b^2 + \frac{1}{N_1} \mathbf{x}^\top \mathbf{x}\Big)
\end{aligned} \tag{11}
$$

similarly,

$$
z_k^1(\mathbf{x}') \sim \mathcal{N}\Big(0, \sigma_b^2 + \frac{1}{N_1} \mathbf{x}'^\top \mathbf{x}'\Big) \tag{12}
$$

and co-variance would be:

$$
\begin{aligned}
\mathrm{Cov}\big(z_k^1(\mathbf{x}), z_k^1(\mathbf{x})\big) &= \mathbb{E}\Big[\Big(\sum_{j=1}^{N_1} W_{k,j}^1 x_j\Big)\ \Big(\sum_{j=1}^{N_1} W_{k,j}^1 x_j\Big)\Big] \\
&= \sum_{j=1}^{N_1} \mathbb{E}\big[(W_{k,j}^1)^2\big] x_j x_j + \sum_{j=1}\sum_{i=1} \mathbb{E}[W_{k,j}^1]\mathbb{E}[W_{k,i}^1] x_j x_i' \\
&= \frac{1}{N_1}\mathbf{x}^\top \mathbf{x}'
\end{aligned} \tag{13}
$$

for any pairs of data $\mathbf{x}$ and $\mathbf{x}'$, we have, $\forall k$:

$$
z_k^1(\mathbf{x}) \sim \mathcal{GP}(K^1) \quad \text{where} \quad K^1(\mathbf{x}, \mathbf{x}') = \sigma_b^2 + \frac{1}{N_1}\mathbf{x}^\top \mathbf{x}' \tag{14}
$$

4.1.2 adding activation ϕ:

$$
\phi(z^1(\mathbf{x})) = \begin{bmatrix} \phi(z_1^1) \\ \vdots \\ \phi(z_k^1) \end{bmatrix} \tag{15}
$$

It's difficult to tell what distribution this is

4.2 pre-activation layer l

$$
z^l(\mathbf{x}) = \begin{bmatrix} z_1^l \\ \vdots \\ z_{N_{l+1}}^l \end{bmatrix} = \begin{bmatrix} W_{1,1}^l & \cdots & W_{1,N_l}^l \\ \vdots & \ddots & \vdots \\ W_{N_{l+1},1}^l & \cdots & W_{k,N_l}^l \end{bmatrix} \begin{bmatrix} \phi(z_1^{l-1}(\mathbf{x})) \\ \vdots \\ \phi(z_{N_l}^{l-1}(\mathbf{x})) \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix} \tag{16}
$$

similarly, we can have:

$$
z^l(\mathbf{x}') = \begin{bmatrix} z_1^l \\ \vdots \\ z_{N_{l+1}}^l \end{bmatrix} = \begin{bmatrix} W_{1,1}^l & \cdots & W_{1,N_l}^l \\ \vdots & \ddots & \vdots \\ W_{N_{l+1},1}^l & \cdots & W_{N_{l+1},N_l}^l \end{bmatrix} \begin{bmatrix} \phi\big(z_1^{l-1}(\mathbf{x}')\big) \\ \vdots \\ \phi\big(z_{N_l}^{l-1}(\mathbf{x}')\big) \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_{N_{l+1}} \end{bmatrix} \tag{17}
$$

for a specific k^{th} row:

$$
z_k^l(\mathbf{x}) = \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x})) + b_k^l \tag{18}
$$

4.2.1 marginal $p(z_k^l(\mathbf{x})|z^{l-1})$

problem is due to non-linearity of $\phi(z_j^{l-1}(\mathbf{x}))$, we do not know what distribution $z_k^l(\mathbf{x})$ is!

However, let's look at an individual term inside the sum: $\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(\mathbf{x}))$

$$
\begin{aligned}
\mathbb{E}\big[W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}))\big] &= 0 \\
\mathrm{Var}\big[W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}))\big] &= \mathbb{E}\big[\big(W_{k,1}^l \phi(z_1^{l-1}(\mathbf{x}))\big)^2\big] \\
&= \mathbb{E}\big[\big(W_{k,1}^l\big)^2\big]\mathbb{E}\big[\phi(z_1^{l-1}(\mathbf{x}))\big)^2\big] \\
&= \frac{1}{N_l}\mathrm{Var}\big[\phi(z_1^{l-1}(\mathbf{x}))\big]
\end{aligned} \tag{19}
$$

Since Var$[\cdot]$ can be chosen to be bounded, and each $W^l_{k,j}\phi(z^{l-1}_j(\mathbf{x}))$ to be i.i.d, so we can apply CLT, as $N_l \to \infty$, and also $\{z^l_j(\mathbf{x})\}$ are i.i.d given z^{l-1}:

$$\begin{aligned}\sum_{j=1}^{N_l} W^l_{k,j}\phi(z^{l-1}_j(\mathbf{x})) &\xrightarrow{d} \mathcal{N}\Big(0,\ \text{Var}\big[W^l_{k,j}\phi(z^{l-1}_j(\mathbf{x}))\big]N_l\Big) \\ &\xrightarrow{d} \mathcal{N}\Big(0,\ \frac{1}{N_l}\text{Var}\big[\phi(z^{l-1}_1(\mathbf{x}))\big]N_l\Big) \quad \text{substitute Eq.(19)} \\ &\xrightarrow{d} \mathcal{N}\Big(0,\ \text{Var}\big[\phi(z^{l-1}_1(\mathbf{x}))\big]\Big)\end{aligned} \tag{20}$$

4.2.2 joint density $p\big(z^l_k(\mathbf{x}), z^l_k(\mathbf{x}')\ |z^l(\mathbf{x}), z^l(\mathbf{x}')\big)$

Here we use a multivariate version of CLT where each i.i.d team inside the sum is a vector: $\begin{bmatrix} z^l_k(\mathbf{x}) \\ z^l_k(\mathbf{x}') \end{bmatrix}$:

$$\begin{bmatrix} z^l_k(\mathbf{x}) \\ z^l_k(\mathbf{x}') \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{N_l} W^l_{k,1}\phi(z^{l-1}_j(\mathbf{x})) \\ \sum_{j=1}^{N_l} W^l_{k,1}\phi(z^{l-1}_j(\mathbf{x}')) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}\,, \Sigma\left(\begin{bmatrix} W^l_{k,1}\phi(z^{l-1}_1(\mathbf{x})) \\ W^l_{k,1}\phi(z^{l-1}_1(\mathbf{x}')) \end{bmatrix}\right) N_l\right) \tag{21}$$

use the notation for zero-meaned R.V:

$$\Sigma\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) = \mathbb{E}\left[\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\begin{bmatrix} y_1 & y_2 \end{bmatrix}\right] = \begin{bmatrix} \text{Var}(y_1) & \text{Cov}(y_1, y_2) \\ \text{Cov}(y_1, y_2) & \text{Var}(y_2) \end{bmatrix} \tag{22}$$

We already know the variance (diagonal) from Eq.(20). How about the co-variance (off-diagonal) term: $\text{Cov}(y_1, y_2) \equiv \text{Cov}\big[W^{l-1}_{k,1}\phi(z^{l-1}_1(\mathbf{x}))\ ,\ W^{l-1}_{k,1}\phi(z^{l-1}_1(\mathbf{x}'))\big]$:

$$\begin{aligned}&\text{Cov}\big[W^l_{k,1}\phi(z^{l-1}_1(\mathbf{x}))\ ,\ W^l_{k,1}\phi(z^{l-1}_1(\mathbf{x}'))\big] \\ &= \mathbb{E}\big[W^l_{k,1}\phi(z^{l-1}_1(\mathbf{x}))\ W^l_{k,1}\phi(z^{l-1}_1(\mathbf{x}'))\big] \\ &= \mathbb{E}\big[(W^l_{k,1})^2\big]\ \mathbb{E}[\phi(z^{l-1}_1(\mathbf{x}))\phi(z^{l-1}_1(\mathbf{x}'))] \quad \text{we didn't need } \mathbb{E} \text{ for } \mathbf{x}^\top\mathbf{x}' \text{ as in Eq.(13)} \\ &= \frac{1}{N_l}\mathbb{E}\big[\phi(z^{l-1}_1(\mathbf{x}))\phi(z^{l-1}_1(\mathbf{x}'))\big]\end{aligned} \tag{23}$$

therefore, canceling out $\frac{1}{N_l} \times N_l$ and add σ^2_b to each of the entries.

It is important to note that σ^2_b also appears in the off-diagonal entries as well as the diagonal entry.

$$\begin{aligned}&\begin{bmatrix} z^l_k(\mathbf{x}) = b_k + \sum_{j=1}^{N_l} W^l_{k,1}\phi(z^{l-1}_j(\mathbf{x})) \\ z^l_k(\mathbf{x}') = b_k + \sum_{j=1}^{N_l} W^l_{k,1}\phi(z^{l-1}_j(\mathbf{x}')) \end{bmatrix} \xrightarrow{d} \\ &\mathcal{N}\left(\mathbf{0}\,,\ \begin{bmatrix} \sigma^2_b + \mathbb{E}\big[\phi(z^{l-1}_1(\mathbf{x}))\phi(z^{l-1}_1(\mathbf{x}))\big] & \sigma^2_b + \mathbb{E}\big[\phi(z^{l-1}_1(\mathbf{x}))\phi(z^{l-1}_1(\mathbf{x}'))\big] \\ \sigma^2_b + \mathbb{E}\big[\phi(z^{l-1}_1(\mathbf{x}))\phi(z^{l-1}_1(\mathbf{x}'))\big] & \sigma^2_b + \mathbb{E}\big[\phi(z^{l-1}_1(\mathbf{x}'))\phi(z^{l-1}_1(\mathbf{x}'))\big] \end{bmatrix}\right)\end{aligned} \tag{24}$$

4.2.3 Relationship with Gaussian Process (GP):

let $f(x) \equiv z^l_k(x)$ be some function, and since for every arbitrary point pair, x and x', we have:

$$\begin{aligned}&\begin{bmatrix} f(x) \\ f(x') \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}\,, \begin{bmatrix} K(x,x) & K(x,x') \\ K(x,x') & K(x',x') \end{bmatrix}\right) \\ &\implies f \sim \mathcal{GP}(0, \mathbf{K})\end{aligned} \tag{25}$$

looking at mean and co-variance as $N_l \to \infty$, for each x, x' pair:

$$\begin{aligned} \begin{bmatrix} z_k^l(x) \\ z_k^l(x') \end{bmatrix} &\sim \mathcal{N}\left(\mathbf{0}\,,\begin{bmatrix} K^l(x,x) & K^l(x,x') \\ K^l(x,x') & K^l(x',x') \end{bmatrix}\right) \\ \text{marginal} \quad z_k^l(x) &\xrightarrow{d} \mathcal{N}\left(0, \sigma_b^2 + \mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big)^2\big]\right) \quad \text{as } N_l \to \infty \\ \text{where:} \quad \mathrm{Cov}\left[z_k^l(x), z_k^l(x')\right] &= K^l(x,x') = \sigma_b^2 + \mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big) \times \phi\big(z_1^{l-1}(x')\big)\big] \end{aligned} \tag{26}$$

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

$$\implies z_k^l(\mathcal{X}) \sim \mathcal{GP}(0, \mathbf{K}^l) \tag{27}$$

The recursion tells us $z_1^{l-1}(\mathcal{X}) \sim \mathcal{GP}(0, \mathbf{K}^{l-1})$. Remove the suffix $z_1 \to z$:

$$\begin{aligned} &\implies z_k^l(\mathcal{X}) \sim \mathcal{GP}(0, \mathbf{K}^l) \quad \forall k \\ &\text{where} \quad \mathbf{K}^l = \sigma_b^2 + \mathbb{E}_{z^{l-1}(\mathcal{X}) \sim \mathcal{GP}(0, \mathbf{K}^{l-1})}\big[\phi\big(z^{l-1}(\mathcal{X})\big)\phi\big(z^{l-1}(\mathcal{X})\big)^\top\big] \end{aligned} \tag{28}$$

5 NTK at initialization

this section describe [3]

5.1 expression

5.2 re-parameterized formulation

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

$$\begin{aligned} &\text{NNGP} \quad z_k^l(x) = \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) + \sigma_b b_k^l \qquad W_{k,j}^l \sim \mathcal{N}\Big(0, \frac{1}{\sqrt{N_l}}\Big) \\ &\text{in NTK we use re-parameterization} \quad z_k^l(x) = \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) + \sigma_b b_k^l \qquad W_{k,j}^l \sim \mathcal{N}(0,1) \quad \sigma_b \sim \mathcal{N}(0,1) \end{aligned} \tag{29}$$

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

$$\begin{bmatrix} z_1^l(x) \\ \vdots \\ z_k^l(x) \\ \vdots \\ z_{N_{l+1}}^l(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{1,j}^l \phi\big(z_j^{l-1}(x)\big) + \sigma_b b_1^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) + \sigma_b b_k^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{N_{l+1},j}^l \phi\big(z_j^{l-1}(x)\big) + \sigma_b b_{N_{l+1}}^l \end{bmatrix} \tag{30}$$

5.3 Prove by Induction

5.3.2 For NTK

we need to show by induction:

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

$$\Theta_{k,k'}^1(x, x') = \Bigg(\underbrace{\frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2}_{K^1}\Bigg)\delta_{k,k'} \tag{31}$$

even better, no need to show: $\quad \Theta_{k,k'}^1(x,x') \to K^1 \delta_{k,k'}$. it is actually equal! Besides there is no N_1 to take limit to ∞

2. then by assuming:

$$\Theta^{l-1}_{k,k'}(x,x') = \frac{\partial z^{l-1}_k(x,\theta)}{\partial \theta^l}^\top \frac{\partial z^{l-1}_k(x',\theta)}{\partial \theta^l} \xrightarrow{N_l \to \infty} \Theta^{l-1}_\infty(x,x')\delta_{k,k'} \tag{32}$$

we can prove:

$$\Theta^{l}_{k,k'}(x,x') = \frac{\partial z^{l}_k(x,\theta)}{\partial \theta^l}^\top \frac{\partial z^{l}_k(x',\theta)}{\partial \theta^l} \xrightarrow{N_{l+1} \to \infty} \Theta^{l}_\infty(x,x')\delta_{k,k'} \tag{33}$$

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

From the Eq.(30), we have:

$$\begin{bmatrix} \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W^1_{1,j} x_1 + \sigma_b b^1_1 \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W^1_{k,j} x_2 + \sigma_b b^1_k \\ \vdots \\ \frac{1}{\sqrt{d_{\text{in}}}} \sum_{j=1}^{d_{\text{in}}} W^1_{N_2,j} x_{d_{\text{in}}} + \sigma_b b^1_{N_2} \end{bmatrix} = \begin{bmatrix} z^1_1(x) \\ \vdots \\ z^1_k(x) \\ \vdots \\ z^1_{N_2}(x) \end{bmatrix} \tag{34}$$

note when computing $\frac{\partial z^1_k(x)}{\partial W^1_{i,j}}$ only k^{th} row going to return a gradient, i.e., $\frac{\partial z^1_k(x)}{\partial W^1_{i,j}} = 0$ if $i \neq k$, and the gradient correspond to $\frac{\cdot}{\partial W^1_{i,j}}$ is x_j:

$$\begin{aligned} \frac{\partial z^1_k(x)}{\partial W^1_{i,j}} &= \begin{cases} \frac{1}{\sqrt{d_{\text{in}}}} x_j & \text{if } i = k \text{ i.e., row } k \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{1}{\sqrt{d_{\text{in}}}} \delta_{i,k} x_j \\ \implies \frac{\partial z^1_{k'}(x)}{\partial W^1_{i,j}} &= \frac{1}{\sqrt{d_{\text{in}}}} \delta_{i,k'} x_j \end{aligned} \tag{35}$$

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

The individual element of $\Theta^l(x,x')$ at k, k' is:

$$
\begin{aligned}
\Theta^1_{k,k'}(x,x') &= \sum_{d=1}^{|\theta^1|} \frac{\partial F^1_k(x)}{\partial \theta^1_d} \frac{\partial F^1_{k'}(x')}{\partial \theta^1_d} \qquad \theta^1 = \{W^1, b^1\} \\
&= \sum_{d=1}^{|W^1|} \frac{\partial F^1_k(x)}{\partial W^1_d} \frac{\partial F^1_{k'}(x')}{\partial W^1_d} + \sum_{d=1}^{|b^1|} \frac{\partial F^1_k(x)}{\partial b^1_d} \frac{\partial F^1_{k'}(x')}{\partial b^1_d} \\
&= \sum_{i=1}^{N_2} \sum_{j=1}^{d_{\text{in}}} \frac{\partial z^1_k(x)}{\partial W_{i,j}} \frac{\partial z^1_{k'}(x')}{\partial W_{i,j}} + \sum_{i=1}^{N_2} \frac{\partial z^1_k(x)}{\partial b_i} \frac{\partial z^1_{k'}(x')}{\partial b_i} \\
&= \sum_{i=1}^{N_2} \sum_{j=1}^{d_{\text{in}}} \frac{1}{\sqrt{d_{\text{in}}}} x_j \delta_{i,k'} \; \frac{1}{\sqrt{d_{\text{in}}}} x'_j \delta_{i,k} + \sum_{i=1}^{N_2} \sigma_b \delta_{i,k} \; \sigma_b \delta_{i,k'} \quad \text{only one } i \in \{1, \ldots N_2\} \text{ in outer sum remain} \\
&= \sum_{j=1}^{d_{\text{in}}} \frac{1}{d_{\text{in}}} x_j x'_j \delta^2_{k,k'} + \sigma^2_b \delta_{k,k'} \qquad \delta_{i,k'} \delta_{i,k} = \delta_{k,k'} \\
&= \frac{1}{d_{\text{in}}} x^\top x' \delta_{k,k'} + \sigma^2_b \delta_{k,k'} \\
&= \underbrace{\left(\frac{1}{d_{\text{in}}} x^\top x' + \sigma^2_b \right)}_{K^1} \delta_{k,k'} \\
&\equiv K^1(x,x') \delta_{k,k'}
\end{aligned} \tag{36}
$$

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

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

$$
\begin{aligned}
\implies \Theta^1(x,x') &= \left.\underbrace{\begin{bmatrix} K^1(x,x') & \ldots & 0 & \ldots & 0 \\ 0 & K^1(x,x') & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & \ldots & K^1(x,x') & 0 \\ 0 & 0 & 0 & 0 & K^1(x,x') \end{bmatrix}}_{k \in \{1,\ldots,N_2\}} \right\} k' \in \{1, \ldots, N_2\} \\
&= \text{repeating diagonal with } K^1(x,x')\delta_{k,k'} \\
&= \underbrace{K^1(x,x')}_{\text{scalar}} \otimes_{\text{outer}} \mathbf{I}_{N_1 \times N_2}
\end{aligned} \tag{37}
$$

Θ^1 matrix of square the size of input $(N_2 \times |\mathcal{X}|) \times (N_2 \times |\mathcal{X}|)$, importantly, there is no limit to take for Θ^1, and so $\Theta^1 = K^1(\mathcal{X}, \mathcal{X}) \otimes_{\text{outer}} \mathbf{I}_{N_1 \times N_2}$

5.5 when $l > 1$

$$\begin{bmatrix} z_1^l(x) \\ \vdots \\ z_k^l(x) \\ \vdots \\ z_{N_{l+1}}^l(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{1,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_1^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l \\ \vdots \\ \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{N_{l+1},j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_{N_{l+1}}^l \end{bmatrix} \tag{38}$$

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

$$\begin{aligned} \Theta_{k,k'}^l(x,x') &= \sum_{d=1}^{|\theta^l|} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}} \\ &= \underbrace{\sum_{d=1}^{|W^l,b^l|} \frac{\partial z_k^l(x)}{\partial\{W^l,b^l\}} \frac{\partial z_{k'}^l(x')}{\partial\{W^l,b^l\}}}_{1} + \underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial \theta_d^{l-1}}}_{2} \end{aligned} \tag{39}$$

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

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

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

$$\begin{aligned} z_k^l &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x)) + \sigma_b b_k^l \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \phi\Big(\frac{1}{\sqrt{N_{l-1}}} \sum_{j=1}^{N_{l-1}} W_{j,i}^{l-1} \phi(z_i^{l-1}(x)) + \sigma_b b_j^{l-1} \Big) + \sigma_b b_j^l \end{aligned} \tag{40}$$

$$\begin{aligned} \frac{\partial z_k^l(x)}{\partial \theta_d^{l-1}} &= \frac{\partial z_k^l(x)}{\partial \phi(z^{l-1}(x))} \frac{\partial \phi(z^{l-1}(x))}{\partial z^{l-1}(x)} \frac{\partial z^{l-1}(x)}{\partial \theta_d^{l-1}} \quad \text{drop index for the last two terms} \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \frac{\partial \phi(z_j^{l-1}(x))}{\partial z_j^{l-1}(x)} \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \\ &= \frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \dot{\phi}(z_j^{l-1}(x)) \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \quad \text{leave last derivative as is, in “recursion”} \end{aligned} \tag{41}$$

substitute it back to 2

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

$$
= \sum_{d=1}^{|\theta^{l-1}|} \left(\frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k,j}^l \, \dot{\phi}(z_j^{l-1}(x)) \, \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \right) \times \left(\frac{1}{\sqrt{N_l}} \sum_{j=1}^{N_l} W_{k',j}^l \, \dot{\phi}(z_j^{l-1}(x')) \, \frac{\partial z_j^{l-1}(x')}{\partial \theta_d^{l-1}} \right) \quad \text{by substitution} \tag{42}
$$

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

$$
\begin{aligned}
&= \sum_{d=1}^{|\theta^{l-1}|} \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} \left(W_{k,j}^l \, \dot{\phi}(z_j^{l-1}(x)) \, \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \right) \times \underbrace{\left(W_{k',j'}^l \, \dot{\phi}(z_{j'}^{l-1}(x')) \, \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}} \right)}_{j \to j' \;\; \text{in second term}} \quad \text{re-arrange} \\
&= \sum_{d=1}^{|\theta^{l-1}|} \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l \; W_{k',j'}^l \dot{\phi}(z_j^{l-1}(x)) \; \dot{\phi}(z_{j'}^{l-1}(x')) \; \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \; \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}} \quad \text{re-arrange} \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l \; W_{k',j'}^l \dot{\phi}(z_j^{l-1}(x)) \; \dot{\phi}(z_{j'}^{l-1}(x')) \; \underbrace{\sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_j^{l-1}(x)}{\partial \theta_d^{l-1}} \; \frac{\partial z_{j'}^{l-1}(x')}{\partial \theta_d^{l-1}}}_{\text{definition } \Theta_{j,j'}^{l-1}(x,x')} \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l \; W_{k',j'}^l \dot{\phi}(z_j^{l-1}(x)) \; \dot{\phi}(z_{j'}^{l-1}(x)) \; \Theta_{j,j'}^{l-1}(x,x') \\
&= \frac{1}{N_l} \sum_{j=1}^{N_l} \sum_{j'=1}^{N_l} W_{k,j}^l \; W_{k',j'}^l \dot{\phi}(z_j^{l-1}(x)) \; \dot{\phi}(z_{j'}^{l-1}(x')) \; \Theta_{\infty}^{l-1}(x,x') \delta_{j,j'} \\
&\qquad \text{use induction assumption: } \Theta_{j,j'}^{l-1}(x,x') \to \underbrace{\Theta_{\infty}^{l-1}(x,x')\delta_{j,j'}}_{\text{deterministic and diagonal limit}} \\
&= \Theta_{\infty}^{l-1}(x,x') \frac{1}{N_l} \sum_{j=1}^{N_l} W_{k,j}^l \; W_{k',j}^l \dot{\phi}(z_j^{l-1}(x)) \; \dot{\phi}(z_j^{l-1}(x')) \quad \text{only terms remain are } j = j'
\end{aligned} \tag{43}
$$

instead of using CLT, we shall apply LoLN here:

$$
\begin{aligned}
&\Theta_\infty^{l-1}(x,x')\underbrace{\frac{1}{N_l}\sum_{j=1}^{N_l} W_{k,j}^l\ W_{k',j}^l \dot\phi(z_j^{l-1}(x))\ \dot\phi(z_j^{l-1}(x'))} \\
&= \Theta_\infty^{l-1}(x,x')\underbrace{\mathbb{E}_{W_{k,1}^l, W_{k',1}^l, z_1^{l-1}(x), z_1^{l-1}(x')}\Big[W_{k,1}^l, W_{k',1}^l \dot\phi(z_1^{l-1}(x))\ \dot\phi(z_1^{l-1}(x'))\Big]} \quad \text{very similar to NNGP} \\
&= \Theta_\infty^{l-1}(x,x')\mathbb{E}_{\left(z_1^{l-1}(x), z_1^{l-1}(x')\right)}\Big[\dot\phi\big(z_1^{l-1}\dot\phi\big(z_1^{l-1}(x')\big)\Big]\mathbb{E}_{W_{k,1}^l, W_{k',1}^l}\big[W_{k,1}^l\ W_{k',1}^l\big] \\
&= \Theta_\infty^{l-1}(x,x')\mathbb{E}_{z^{l-1}\sim\mathcal{GP}\left(0,K^{l-1}\right)}\Big[\dot\phi\big(z_1^{l-1}(x)\big)\dot\phi\big(z_1^{l-1}(x')\big)\Big]\delta_{k,k'} \\
&= \delta_{k,k'}\dot K^l(x,x')\ \Theta_\infty^{l-1}(x,x')
\end{aligned} \tag{44}
$$

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

$$
\begin{aligned}
\mathbb{E}_{W_{k,1}^l, W_{k',1}^l}\big[W_{k,1}^l\ W_{k',1}^l\big] &= \begin{cases}\mathbb{E}\big[W_{k,1}^l\ W_{k',1}^l\big] & k\neq k' \\ \mathbb{E}\big[(W_{k,1}^l)^2\big] & k=k'\end{cases} \\
&= \begin{cases}0 & k\neq k' \\ 1 & k=k'\end{cases} \qquad \text{re-parameterized expression} \quad W_{k,1}^l\sim\mathcal{N}(0,1) \\
&= \delta_{k,k'}
\end{aligned} \tag{45}
$$

2. notice the expression here:

$$
\frac{1}{N_l}\sum_{j=1}^{N_l} W_{k,j}^l\ W_{k',j}^l \dot\phi(z_j^{l-1}(x))\ \dot\phi(z_j^{l-1}(x')) \tag{46}
$$

   is the very similar of NNGP formulation, except:

$$
\phi(z_j^{l-1}(x)) \to \dot\phi(z_j^{l-1}(x)) \tag{47}
$$

   so expect same CLT/LoLN treatment applies here

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

$$
\begin{aligned}
\dot K^l(x,x') &= \sigma_w^2\ \mathbb{E}_{\left(z_1^{l-1}(x), z_1^{l-1}(x')\right)\sim\mathcal{N}\left(0,K^{l-1}(x,x')\right)}\Big[\dot\phi\big(z_1^{l-1}(x)\big)\dot\phi\big(z_1^{l-1}(x')\big)\Big] \\
&= \ \mathbb{E}_{\left(z_1^{l-1}(x), z_1^{l-1}(x')\right)\sim\mathcal{N}\left(0,K^{l-1}(x,x')\right)}\Big[\dot\phi\big(z_1^{l-1}(x)\big)\dot\phi\big(z_1^{l-1}(x')\big)\Big] \qquad \text{assume } \sigma_w=1
\end{aligned} \tag{48}
$$

   compare with Eq. (**??**) the recursion in NNGP:

$$
K^l(x,x') = \sigma_b^2 + \sigma_w^2\ \mathbb{E}_{\left(z_1^{l-1}(x), z_1^{l-1}(x')\right)\sim\mathcal{N}\left(0,K^{l-1}(x,x')\right)}\Big[\phi\big(z_1^{l-1}(x)\big)\phi\big(z_1^{l-1}(x')\big)\Big] \tag{49}
$$

   note $\dot K^l(x,x')$ is **not** a recursion, and $K^l(x,x')$ is expressed in recursion

4. note $\delta_{k,k'}\dot K^l(x,x')\ \Theta_\infty^{l-1}(x,x')$ is a scalar, in particular $\dot K^l(x,x')$ is a scalar. However, $\Theta(x,x')$ is the constructed matrix, where elements are of $\dot K^l(x,x')$

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

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

$$\sum_{d=1}^{|W^l,b^l|} \frac{\partial z_k^l(x)}{\partial\{W^l,b^l\}} \frac{\partial z_{k'}^l(x')}{\partial\{W^l,b^l\}} \tag{50}$$

and compare that with for $l = 1$:

$$\begin{aligned} &\sum_{d=1}^{|\theta^1|} \frac{\partial z_k^1(x)}{\partial\theta_d^1} \frac{\partial z_{k'}^1(x')}{\partial\theta_d^1} \qquad \theta^1 = \{W^1, b^1\} \\ &= \left(K^1(x,x') \equiv \frac{1}{d_{\text{in}}} x^\top x' + \sigma_b^2\right)\delta_{k,k'} \end{aligned} \tag{51}$$

then, we do know:

$$\begin{aligned} &\sum_{d=1}^{|W^l,b^l|} \frac{\partial z_k^l(x)}{\partial\{W^l,b^l\}} \frac{\partial z_{k'}^l(x')}{\partial\{W^l,b^l\}} \\ &= \left(K^l(x,x') \equiv \frac{1}{N_l}\phi\big(z^l(x)\big)^\top \phi\big(z^l(x)\big) + \sigma_b^2\right)\delta_{k,k'} \end{aligned} \tag{52}$$

5.5.4 putting all together

$$\begin{aligned} \Theta_{k,k'}^l(x,x') &= \sum_{d=1}^{|W^l,b^l|} \frac{\partial z_k^l(x)}{\partial\{W^l,b^l\}} \frac{\partial z_{k'}^l(x')}{\partial\{W^l,b^l\}} + \sum_{d=1}^{|\theta^{l-1}|} \frac{\partial z_k^1(x)}{\partial\theta_d^{l-1}} \frac{\partial z_{k'}^l(x')}{\partial\theta_d^{l-1}} \\ &= K^l(x,x')\ \delta_{k,k'} + \delta_{k,k'} \dot{K}^l(x,x')\ \Theta_\infty^{l-1}(x,x') \qquad N_{l+1}\to\infty \\ &= \left(K^l(x,x')\ + \dot{K}^l(x,x')\Theta_\infty^{l-1}(x,x')\right)\delta_{k,k'} \\ &= \Theta_\infty^l(x,x')\delta_{k,k'} \end{aligned} \tag{53}$$

this does what we want to achieve in Eq.[32], by assuming $\Theta_{k,k'}^{l-1}(x,x') \xrightarrow{N_l\to\infty} \Theta_\infty^{l-1}(x,x')\delta_{k,k'}$, we prove: $\Theta_{k,k'}^l(x,x') \xrightarrow{N_{l+1}\to\infty} \Theta_\infty^l(x,x')\delta_{k,k'}$

then finally:

$$\Theta^l(x,x') = \underbrace{\left(K^l(x,x')\ + \dot{K}^l(x,x')\Theta_\infty^{l-1}(x,x')\right)}_{\text{scalar}} \otimes_{\text{outer}} \underbrace{\mathbf{I}_{N_{l+1}\times N_{l+1}}}_{\text{same value for all } k,k' \text{ pairs}} \tag{54}$$

5.5.5 apply the above to $l = 1$

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

6 linearized model

See NTK in action: Linearized model [4]:

6.1 $f_t^{\text{lin}}(x, \theta_t)$ and $\dot{\omega}$

linearized model is:

$$\begin{aligned} f_t^{\text{lin}}(x, \theta_t) &= f_0(x, \theta_0) + \nabla_\theta f(x, \theta_t)\Big|_{\theta_t \to \theta_0} \triangle\theta(t) \\ &= f_0(x, \theta_0) + \nabla_\theta f_0(x, \theta_0) \ \big(\theta(t) - \theta(0)\big) \\ &= f_0(x, \theta_0) + \nabla_\theta f_0(x, \theta_0) \ \omega_t \end{aligned} \tag{55}$$

both $f_0(x, \theta_0)$ and $\nabla_\theta f_0(x, \theta_0)$ are constants

6.1.1 dynamics of $\dot{\omega}$

looking at the dynamics of linearized gradient flow of **linearized model**, it is obvious that $\dot{\omega}$ only depends on $\mathcal{X}$ instead, as parameter dynamics only depends on training data $\mathcal{X}$:

$$\begin{aligned} \theta_{t+1} &= \theta_t - \eta \nabla_{\theta_t} \mathcal{L}(\cdot) \\ \theta_{t+1} - \theta_t &= -\eta \nabla_{\theta_t} \mathcal{L}(\cdot) \\ \dot{\omega} = \theta_{t+1} - \theta_t &= -\eta \nabla_{\theta_t} \mathcal{L}(\cdot) \qquad \text{why not } \dot{\omega} = \lim_{h \to 0} \frac{\theta_{t+1} - \theta_t}{\eta} = -\eta \nabla_{\theta_t} \mathcal{L}(\cdot)? \\ &= -\eta \Big(\nabla_{\theta_t} f_t^{\text{lin}} \nabla_{f_t^{\text{lin}}(\mathcal{X})} \mathcal{L}(\cdot) \Big) \\ &= -\eta \Big(\nabla_{\theta_t} \big[f_0(x, \theta_0) + \nabla_\theta f_0(x, \theta_0) \ \big(\theta(t) - \theta(0)\big) \big] \nabla_{f_t^{\text{lin}}(\mathcal{X})} \mathcal{L}(\cdot) \Big) \\ &= -\eta \nabla_\theta f(\mathcal{X}, \theta_0)^\top \nabla_{f_t^{\text{lin}}(\mathcal{X})} \mathcal{L}(\cdot) \end{aligned} \tag{56}$$

$$\dot{\omega} = -\eta \nabla_\theta f(\mathcal{X}, \theta_0)^\top \nabla_{f_t^{\text{lin}}(\mathcal{X})} \mathcal{L}(\cdot)$$

6.1.2 dimensionality

$\nabla_\theta f_0(\mathcal{X}, \theta_0) \in \mathbb{R}^{|\mathcal{X}| \times |\theta|}$:

$$\nabla_\theta f(\mathcal{X}, \theta) \nabla_\theta f(\mathcal{X}, \theta)^\top = \sum_{i=1}^{|\theta|} \big(\nabla_\theta f(\mathcal{X}, \theta_i)\big) \big(\nabla_\theta f(\mathcal{X}, \theta_i)\big)^\top = \hat{\Theta}(\mathcal{X}, \mathcal{X}) \tag{57}$$

one of the important NTK is when $t = 0$, i.e., at initialization:

$$\nabla_\theta f(\mathcal{X}, \theta_0) \nabla_\theta f(\mathcal{X}, \theta_0)^\top = \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) \tag{58}$$

6.1.3 dynamics of $\dot{f}_t^{\text{lin}}$

$$\begin{aligned}\dot{f}_t^{\text{lin}}(x,\theta_t) &= \nabla_\theta f_0(x,\theta_0)\ \dot{\omega}(t)\\ &= \nabla_\theta f_0(x,\theta_0)\ \big[-\eta\nabla_\theta f(\mathcal{X},\theta_0)^\top \nabla_{f_t^{\text{lin}}(\mathcal{X})}\mathcal{L}(\cdot)\big]\\ &= -\eta\hat{\Theta}_0(x,\mathcal{X})\nabla_{f_t^{\text{lin}}(\mathcal{X})}\mathcal{L}(\cdot)\end{aligned} \tag{60}$$

6.1.4 ODE solution using $\mathcal{L} = \frac{1}{2}\|f(\mathcal{X}) - \mathcal{Y}\|_2^2$

$$\begin{aligned}\dot{f}_t^{\text{lin}}(\mathcal{X},\theta_t) &= -\eta\hat{\Theta}_0(x,\mathcal{X})\nabla_{f_t^{\text{lin}}(\mathcal{X})}\mathcal{L}(\cdot)\\ &= -\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\big(f_t^{\text{lin}}(\mathcal{X}) - \mathcal{Y}\big)\end{aligned} \tag{61}$$

then ODE has the close-form solution:

1. note the following has terms in $\mathcal{X}$:

$$\begin{aligned}f_t^{\text{lin}}(\mathcal{X},\theta) &= \mathcal{Y} + \big(f_0(\mathcal{X},\theta_0) - \mathcal{Y}\big)\exp^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\ t}\\ &= \mathcal{Y} + f_0(\mathcal{X},\theta_0)\exp^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\ t} - \mathcal{Y}\exp^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\ t} +\\ &= \big(\mathbf{I} - \exp^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\ t}\big)\mathcal{Y} + \exp^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\ t} f_0(\mathcal{X},\theta_0)\end{aligned} \tag{62}$$

   (a) $t=0$: $\quad f_t^{\text{lin}}(\mathcal{X},\theta)|_{t=0} = f_0(\mathcal{X},\theta_0)$
   (b) $t=\infty$: $\quad f_t^{\text{lin}}(\mathcal{X},\theta)|_{t=\infty} = \mathcal{Y}$
   (c) it makes sense as f_t^{lin} is an interpolation between $f_0(\mathcal{X},\theta_0)$ and $\mathcal{Y}$

2. ODE solution for parameter ω_t is:

$$\omega_t = -\nabla_\theta f(\mathcal{X},\theta_0)^\top \hat{\Theta}_0(\mathcal{X},\mathcal{X})^{-1}\big(\mathbf{I} - \exp^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\ t}\big)\big(f_0(\mathcal{X},\theta_0) - \mathcal{Y}\big) \tag{63}$$

3. prediction of x is:

$$\begin{aligned}f_t^{\text{lin}}(x,\theta_t) &= f_0(x,\theta_0) + \nabla_\theta f_0(x,\theta_0)\ \omega_t \qquad \text{subtitute Eq.(63)}\\ &= f_0(x,\theta_0) - \nabla_\theta f_0(x,\theta_0)\ \nabla_\theta f(\mathcal{X},\theta_0)^\top \hat{\Theta}_0(\mathcal{X},\mathcal{X})^{-1}\big(\mathbf{I} - \exp^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\ t}\big)\big(f_0(\mathcal{X},\theta_0) - \mathcal{Y}\big)\\ &= f_0(x,\theta_0) - \hat{\Theta}_0(x,\mathcal{X})\hat{\Theta}_0(\mathcal{X},\mathcal{X})^{-1}\big(\mathbf{I} - \exp^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\ t}\big)\big(f_0(\mathcal{X},\theta_0) - \mathcal{Y}\big)\end{aligned} \tag{64}$$

   alternatively it is written as:

$$\begin{aligned}f_t^{\text{lin}}(x,\theta) = {}& \hat{\Theta}_0(x,\mathcal{X})\hat{\Theta}_0(\mathcal{X},\mathcal{X})^{-1}\big(\mathbf{I} - \exp^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\ t}\big)\mathcal{Y}\\ &+ f_0(x,\theta_0) + \hat{\Theta}_0(x,\mathcal{X})\hat{\Theta}_0(\mathcal{X},\mathcal{X})^{-1}\big(\mathbf{I} - \exp^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\ t}\big)f_0(\mathcal{X},\theta_0)\end{aligned} \tag{65}$$

(a) $t = 0$: $f_t^{\text{lin}}(x,\theta)|_{t=0} = f_0(x,\theta)$

(b) $t = \infty$:

$$f_t^{\text{lin}}(x,\theta)|_{t=\infty} = \hat{\Theta}_0(x,\mathcal{X})\hat{\Theta}_0(\mathcal{X},\mathcal{X})^{-1}\mathcal{Y} \quad + \quad f_0(x,\theta_0) - \hat{\Theta}_0(x,\mathcal{X})\hat{\Theta}_0(\mathcal{X},\mathcal{X})^{-1}f_0(\mathcal{X},\theta_0) \tag{66}$$

6.2 lazy training

finally one need to prove these:

$$\left.\begin{array}{l} \sup\limits_{t\geq 0} \|f_t(x) - f_t^{\text{lin}}\|_2 \\ \sup\limits_{t\geq 0} \dfrac{\|\theta_t - \theta_0\|_2}{\sqrt{n}} \\ \sup\limits_{t\geq 0} \|\hat{\Theta}_t - \hat{\Theta}_0\|_F \end{array}\right\} = \mathcal{O}(n^{-\frac{1}{2}}) \quad \text{as} \quad n \to \infty \tag{67}$$

References

[1] Radford M Neal, "Priors for infinite networks (tech. rep. no. crg-tr-94-1)," *University of Toronto*, 1994.

[2] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein, "Deep neural networks as gaussian processes," *arXiv preprint arXiv:1711.00165*, 2017.

[3] Arthur Jacot, Franck Gabriel, and Clément Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *arXiv preprint arXiv:1806.07572*, 2018.

[4] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, 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.