

NTK Derivation

Yilan Chen

August 3, 2020

1 Problem Setup

Consider a fully connected neural network.

$$\alpha^{(0)}(x; \theta) = x \quad (1)$$

$$\tilde{\alpha}^{(l+1)}(x; \theta) = \frac{1}{\sqrt{n_l}} W^{(l)} \alpha^{(l)}(x; \theta) + \beta b^{(l)}, \quad \text{for } l = 0, \dots, L-1 \quad (2)$$

$$\alpha^{(l)}(x; \theta) = \sigma(\tilde{\alpha}^{(l)}(x; \theta)), \quad \text{for } l = 0, \dots, L-1 \quad (3)$$

$$f_\theta(x) := \tilde{\alpha}^{(L)}(x; \theta) \quad (4)$$

where $W^{(l)} \in \mathbb{R}^{n_{l+1} \times n_l}$, $b^{(l)} \in \mathbb{R}^{n_{l+1}}$, whose elements $w_{i,j}^{(l)}, b_i^{(l)} \sim \mathcal{N}(0, 1)$

Neural Tangent Kernel(NTK):

$$\Theta^{(L)}(\theta) = \sum_{p=1}^P \partial_{\theta_p} F^{(L)}(\theta) \otimes \partial_{\theta_p} F^{(L)}(\theta) \quad (5)$$

$$\begin{aligned} \Theta^{(L)}(x, x') &= \sum_{p=1}^P \frac{\partial F^{(L)}(\theta, x)}{\partial \theta_p} \otimes \frac{\partial F^{(L)}(\theta, x')}{\partial \theta_p} \\ &= \sum_{p=1}^P \left[\frac{\partial F_1^{(L)}(\theta, x)}{\partial \theta_p}, \dots, \frac{\partial F_{n_L}^{(L)}(\theta, x)}{\partial \theta_p} \right]^T \otimes \left[\frac{\partial F_1^{(L)}(\theta, x')}{\partial \theta_p}, \dots, \frac{\partial F_{n_L}^{(L)}(\theta, x')}{\partial \theta_p} \right]^T \\ &= \sum_{p=1}^P \begin{bmatrix} \frac{\partial F_1^{(L)}(\theta, x)}{\partial \theta_p} & \frac{\partial F_1^{(L)}(\theta, x')}{\partial \theta_p} & \dots & \frac{\partial F_1^{(L)}(\theta, x)}{\partial \theta_p} & \frac{\partial F_{n_L}^{(L)}(\theta, x')}{\partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial F_{n_L}^{(L)}(\theta, x)}{\partial \theta_p} & \frac{\partial F_{n_L}^{(L)}(\theta, x')}{\partial \theta_p} & \dots & \frac{\partial F_{n_L}^{(L)}(\theta, x)}{\partial \theta_p} & \frac{\partial F_{n_L}^{(L)}(\theta, x')}{\partial \theta_p} \end{bmatrix} \in \mathbb{R}^{n_L \times n_L} \end{aligned} \quad (6)$$

is the sum of P (number of parameters) matrices.

2 Gaussian Process

Let $\tilde{\alpha}_k^{(l+1)}(x; \theta)$ be the k_{th} entry of $\tilde{\alpha}^{(l+1)}(x; \theta)$, $k = 1, \dots, n_{l+1}$,

$$\tilde{\alpha}_k^{(l+1)}(x; \theta) = \frac{1}{\sqrt{n_l}} w_k^{(l)} \cdot \alpha^{(l)}(x; \theta) + \beta b_k^{(l)} = \frac{1}{\sqrt{n_l}} \sum_{j=1}^{n_l} w_{kj}^{(l)} \alpha_j^{(l)}(x; \theta) + \beta b_k^{(l)} \quad (7)$$

where $w_k^{(l)} \in \mathbb{R}^{n_l}$, is the k_{th} row of $W^{(l)}$, $b_k^{(l)} \in \mathbb{R}$,

For any output of any layer, we have

$$\begin{aligned} \mathbb{E}[\tilde{\alpha}_k^{(l+1)}(x; \theta)] &= \mathbb{E}\left[\frac{1}{\sqrt{n_l}} w_k^{(l)} \cdot \alpha^{(l)}(x; \theta) + \beta b_k^{(l)}\right] \\ &= \frac{1}{\sqrt{n_l}} \mathbb{E}[w_k^{(l)}] \cdot \mathbb{E}[\alpha^{(l)}(x; \theta)] + \beta \mathbb{E}[b_k^{(l)}] \\ &= 0, \quad \text{for } l = 0, \dots, L-1 \end{aligned} \quad (8)$$

The covariance for k_{th} and k'_{th} ($k \neq k'$) entry of outputs for any layer is

$$\begin{aligned}\mathbb{E}[\tilde{\alpha}_k^{(l+1)}(x; \theta) \tilde{\alpha}_{k'}^{(l+1)}(x'; \theta)] &= \mathbb{E}\left[\left[\frac{1}{\sqrt{n_l}} w_k^{(l)} \cdot \alpha^{(l)}(x; \theta) + \beta b_k^{(l)}\right] \left[\frac{1}{\sqrt{n_l}} w_{k'}^{(l)} \cdot \alpha^{(l)}(x'; \theta) + \beta b_{k'}^{(l)}\right]\right] \\ &= 0, \quad \text{for } k \neq k', l = 0, \dots, L-1\end{aligned}\quad (9)$$

So different elements of outputs for any layer is independent.

The covariance for the same entry of outputs is as follows.

When $L = 1$,

$$\begin{aligned}\Sigma^{(1)}(x, x') &= \mathbb{E}[\tilde{\alpha}_k^{(1)}(x; \theta) \tilde{\alpha}_k^{(1)}(x'; \theta)] = \mathbb{E}\left[\left[\frac{1}{\sqrt{n_0}} w_k^{(0)} \cdot \alpha^{(0)}(x; \theta) + \beta b_k^{(0)}\right] \cdot \left[\frac{1}{\sqrt{n_0}} w_k^{(0)} \cdot \alpha^{(0)}(x'; \theta) + \beta b_k^{(0)}\right]\right] \\ &= \frac{1}{n_0} \alpha^{(0)}(x; \theta) \cdot \alpha^{(0)}(x'; \theta) + \beta^2 = \frac{1}{n_0} x^T x' + \beta^2\end{aligned}\quad (10)$$

Recursively, for $l = 0, \dots, L-1$,

$$\begin{aligned}\tilde{\Sigma}^{(l+1)}(x, x') &= \mathbb{E}[\tilde{\alpha}_k^{(l)}(x; \theta) \tilde{\alpha}_k^{(l)}(x'; \theta)] = \mathbb{E}\left[\left[\frac{1}{\sqrt{n_l}} w_k^{(l)} \cdot \alpha^{(l)}(x; \theta) + \beta b_k^{(l)}\right] \cdot \left[\frac{1}{\sqrt{n_l}} w_k^{(l)} \cdot \alpha^{(l)}(x'; \theta) + \beta b_k^{(l)}\right]\right] \\ &= \frac{1}{n_l} \alpha^{(l)}(x; \theta) \cdot \alpha^{(l)}(x'; \theta) + \beta^2 = \frac{1}{n_l} \sigma(\tilde{\alpha}^{(l)}(x; \theta)) \cdot \sigma(\tilde{\alpha}^{(l)}(x'; \theta)) + \beta^2 \\ &= \frac{1}{n_l} \sum_{i=1}^{n_l} \sigma(\tilde{\alpha}_i^{(l)}(x; \theta)) \sigma(\tilde{\alpha}_i^{(l)}(x'; \theta)) + \beta^2\end{aligned}\quad (11)$$

when $n_l \rightarrow \infty$,

$$\begin{aligned}\tilde{\Sigma}^{(l+1)}(x, x') &\rightarrow \mathbb{E}[\sigma(\tilde{\alpha}_i^{(l)}(x; \theta)) \sigma(\tilde{\alpha}_i^{(l)}(x'; \theta))] + \beta^2 \\ &\rightarrow \Sigma^{(l+1)}(x, x') = \mathbb{E}_{g \sim \mathcal{N}(0, \Sigma^{(l)})}[\sigma(g(x)) \sigma(g(x'))] + \beta^2\end{aligned}\quad (12)$$

taking the expectation with respect to a centered Gaussian process g of covariance $\Sigma^{(l)}$, which is equivalent to integrating against the joint distribution of only $g(x)$ and $g(x')$ (a zero mean, two-dimensional Gaussian whose covariance matrix has distinct entries $\Sigma^{(l)}(x, x')$, $\Sigma^{(l)}(x, x)$ and $\Sigma^{(l)}(x', x')$).

From the above, we can see the output functions $f_{\theta, k}(x)$, for $k = 1, \dots, n_L$ tend (in law) to iid centered Gaussian processes of covariance $\Sigma^{(L)}$ in the limit as $n_1, \dots, n_{L-1} \rightarrow \infty$ sequentially. (Proposition 1 in [1])

3 NTK at Initialization

Using induction. When $L = 1$,

$$f_{\theta}(x) = \tilde{\alpha}^{(1)}(x; \theta) = \frac{1}{\sqrt{n_0}} W^{(0)} x + \beta b^{(0)} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_k(x) \\ \vdots \\ f_{n_1}(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} w_{0i}^{(0)} x_i + \beta b_0^{(0)} \\ \vdots \\ \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} w_{ki}^{(0)} x_i + \beta b_k^{(0)} \\ \vdots \\ \frac{1}{\sqrt{n_0}} \sum_{i=1}^{n_0} w_{n_1 i}^{(0)} x_i + \beta b_{n_1}^{(0)} \end{bmatrix} \in \mathbb{R}^{n_1} \quad (13)$$

The partial derivative of the k_{th} entry of $f_{\theta}(x)$ with respect to w_{ji} is

$$\frac{\partial f_k(x)}{\partial w_{ji}} = \frac{1}{\sqrt{n_0}} x_i \delta_{jk} \quad (14)$$

$$\frac{\partial f_k(x)}{\partial w_{ji}} \frac{\partial f_{k'}(x')}{\partial w_{ji}} = \frac{1}{\sqrt{n_0}} x_i \frac{1}{\sqrt{n_0}} x'_i \delta_{jk} \delta_{jk'} = \frac{1}{n_0} x_i x'_i \delta_{jk} \delta_{jk'} \quad (15)$$

$$\frac{\partial f_k(x)}{\partial b_j^{(0)}} \frac{\partial f_{k'}(x')}{\partial b_j^{(0)}} = \beta \delta_{jk} \delta_{jk'} = \beta^2 \delta_{jk} \delta_{jk'} \quad (16)$$

See(6), the k, k' entry of NTK $\Theta^{(L)}(x, x')$ is the sum for all parameters.

$$\begin{aligned}
\Theta_{kk'}^{(1)}(x, x') &= \sum_{p=1}^P \frac{\partial F_k^{(1)}(\theta, x)}{\partial \theta_p} \frac{\partial F_{k'}^{(1)}(\theta, x')}{\partial \theta_p} = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \frac{\partial f_k(x)}{\partial w_{ji}} \frac{\partial f_{k'}(x')}{\partial w_{ji}} + \sum_{j=1}^{n_1} \frac{\partial f_k(x)}{\partial b_j^{(0)}} \frac{\partial f_{k'}(x')}{\partial b_j^{(0)}} \\
&= \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \frac{1}{n_0} x_i x'_i \delta_{jk} \delta_{jk'} + \sum_{j=1}^{n_1} \beta^2 \delta_{jk} \delta_{jk'} = \frac{1}{n_0} x^T x' \delta_{kk'} + \beta^2 \delta_{kk'} \\
&= \Sigma^{(1)}(x, x') \delta_{kk'}
\end{aligned} \tag{17}$$

$$\Theta^{(1)}(x, x') = \Sigma^{(1)}(x, x') \otimes I_{n_0} \in \mathbb{R}^{n_0 \times n_0} \tag{18}$$

which is a deterministic and diagonal matrix.

When $L \geq 1$, assume the neural tangent kernel $\Theta^{(L)}(x, x')$ of the smaller network converges to a deterministic limit:

$$\Theta_{ii'}^{(L)}(x, x') = (\partial_{\tilde{\theta}} \tilde{\alpha}_i^{(L)}(x; \theta))^T (\partial_{\tilde{\theta}} \tilde{\alpha}_{i'}^{(L)}(x'; \theta)) \rightarrow \Theta_{\infty}^{(L)}(x, x') \delta_{ii'} \tag{19}$$

where we split the parameters into the parameters of the first L layers $\tilde{\theta} = (W^{(0)}, b^{(0)}, \dots, W^{(L-1)}, b^{(L-1)})$ and those of the last layer $(W^{(L)}, b^{(L)})$. Now we need to prove the Convergence for $\Theta^{(L+1)}(x, x')$.

For $L + 1$,

$$f_{\theta}(x) = \tilde{\alpha}^{(L+1)}(x; \theta) = \frac{1}{\sqrt{n_L}} W^{(L)} \alpha^{(L)}(x; \theta) + \beta b^{(L)} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_k(x) \\ \vdots \\ f_{n_{L+1}}(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} w_{0i}^{(L)} \alpha_i^{(L)}(x; \theta) + \beta b_0^{(0)} \\ \vdots \\ \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} w_{ki}^{(L)} \alpha_i^{(L)}(x; \theta) + \beta b_k^{(0)} \\ \vdots \\ \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} w_{n_{L+1},i}^{(L)} \alpha_i^{(L)}(x; \theta) + \beta b_{n_{L+1}}^{(0)} \end{bmatrix} \in \mathbb{R}^{n_L} \tag{20}$$

The partial derivative of the k_{th} entry of $f_{\theta}(x)$ with respect to one of the $\tilde{\theta}$ is

$$\begin{aligned}
\partial_{\tilde{\theta}_p} f_{\theta,k}(x) &= \frac{\partial f_{\theta,k}(x)}{\partial \tilde{\theta}_p} = \frac{\partial f_{\theta,k}(x)}{\partial \alpha_i^{(L)}(x; \theta)} \frac{\partial \alpha_i^{(L)}(x; \theta)}{\partial \tilde{\alpha}_i^{(L)}(x; \theta)} \frac{\partial \tilde{\alpha}_i^{(L)}(x; \theta)}{\partial \tilde{\theta}_p} = \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} w_{ki}^{(L)} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \frac{\partial \tilde{\alpha}_i^{(L)}(x; \theta)}{\partial \tilde{\theta}_p} \\
&= \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} w_{ki}^{(L)} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \partial_{\tilde{\theta}_p} \tilde{\alpha}_i^{(L)}(x; \theta)
\end{aligned} \tag{21}$$

$$\begin{aligned}
\frac{\partial f_{\theta,k}(x)}{\partial \tilde{\theta}_p} \frac{\partial f_{\theta,k'}(x')}{\partial \tilde{\theta}_p} &= \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} w_{ki}^{(L)} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \frac{\partial \tilde{\alpha}_i^{(L)}(x; \theta)}{\partial \tilde{\theta}_p} \frac{1}{\sqrt{n_L}} \sum_{i'=1}^{n_L} w_{k'i'}^{(L)} \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x'; \theta)) \frac{\partial \tilde{\alpha}_{i'}^{(L)}(x'; \theta)}{\partial \tilde{\theta}_p} \\
&= \frac{1}{n_L} \sum_{i=1}^{n_L} \sum_{i'=1}^{n_L} w_{ki}^{(L)} w_{k'i'}^{(L)} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x'; \theta)) \frac{\partial \tilde{\alpha}_i^{(L)}(x; \theta)}{\partial \tilde{\theta}_p} \frac{\partial \tilde{\alpha}_{i'}^{(L)}(x'; \theta)}{\partial \tilde{\theta}_p}
\end{aligned} \tag{22}$$

Take the sum for all parameters of $\tilde{\theta}$,

$$\begin{aligned}
\sum_{p=1}^{|\tilde{\theta}|} \frac{\partial f_{\theta,k}(x)}{\partial \tilde{\theta}_p} \frac{\partial f_{\theta,k'}(x')}{\partial \tilde{\theta}_p} &= \frac{1}{n_L} \sum_{i=1}^{n_L} \sum_{i'=1}^{n_L} w_{ki}^{(L)} w_{k'i'}^{(L)} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x'; \theta)) \sum_{p=1}^{|\tilde{\theta}|} \frac{\partial \tilde{\alpha}_i^{(L)}(x; \theta)}{\partial \tilde{\theta}_p} \frac{\partial \tilde{\alpha}_{i'}^{(L)}(x'; \theta)}{\partial \tilde{\theta}_p} \\
&= \frac{1}{n_L} \sum_{i=1}^{n_L} \sum_{i'=1}^{n_L} w_{ki}^{(L)} w_{k'i'}^{(L)} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x'; \theta)) \frac{\partial \tilde{\alpha}_i^{(L)}(x; \theta)}{\partial \tilde{\theta}} \cdot \frac{\partial \tilde{\alpha}_{i'}^{(L)}(x'; \theta)}{\partial \tilde{\theta}} \\
&= \frac{1}{n_L} \sum_{i=1}^{n_L} \sum_{i'=1}^{n_L} w_{ki}^{(L)} w_{k'i'}^{(L)} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x'; \theta)) \Theta_{ii'}^{(L)}(x, x') \\
&= \frac{1}{n_L} \sum_{i=1}^{n_L} \sum_{i'=1}^{n_L} w_{ki}^{(L)} w_{k'i'}^{(L)} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x'; \theta)) \Theta_{\infty}^{(L)}(x, x') \delta_{ii'} \\
&= \frac{1}{n_L} \sum_{i=1}^{n_L} w_{ki}^{(L)} w_{k'i}^{(L)} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x'; \theta)) \Theta_{\infty}^{(L)}(x, x')
\end{aligned} \tag{23}$$

when $n_L \rightarrow \infty$, this tends to its expectation

$$\begin{aligned}
\mathbb{E}[w_{ki}^{(L)} w_{k'i}^{(L)} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x'; \theta)) \Theta_\infty^{(L)}(x, x')] &= \mathbb{E}[w_{ki}^{(L)} w_{k'i}^{(L)}] \mathbb{E}[\dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x'; \theta))] \Theta_\infty^{(L)}(x, x') \\
&= \delta_{kk'} \mathbb{E}[\dot{\sigma}(\tilde{\alpha}_i^{(L)}(x; \theta)) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x'; \theta))] \Theta_\infty^{(L)}(x, x') \\
&= \delta_{kk'} \mathbb{E}_{g \sim \mathcal{N}(0, \Sigma^{(L)})}[\dot{\sigma}(g(x)) \dot{\sigma}(g(x'))] \Theta_\infty^{(L)}(x, x') \\
&= \delta_{kk'} \dot{\Sigma}^{(L+1)}(x, x') \Theta_\infty^{(L)}(x, x')
\end{aligned} \tag{24}$$

For the parameters of last layer $(W^{(L)}, b^{(L)})$, the derivation is similar as $L=1$ but just replace the x_i with $\alpha_i^{(L)}(x; \theta)$. We can get the result as

$$\frac{1}{n_L} \alpha_i^{(L)}(x; \theta)^T \alpha_i^{(L)}(x'; \theta) \delta_{kk'} + \beta^2 \delta_{kk'} \rightarrow \Sigma^{(L+1)}(x, x') \delta_{kk'} \tag{25}$$

Take the sum of these two parts, we can get

$$\Theta_{kk'}^{(L+1)}(x, x') = \Theta_\infty^{(L)}(x, x') \dot{\Sigma}^{(L+1)} \delta_{kk'} + \Sigma^{(L+1)}(x, x') \delta_{kk'} \tag{26}$$

$$\Theta^{(L+1)}(x, x') = [\Theta_\infty^{(L)}(x, x') \dot{\Sigma}^{(L+1)}(x, x') + \Sigma^{(L+1)}(x, x')] \otimes I_{n_{L+1}} \in \mathbb{R}^{n_{L+1} \times n_{L+1}} \tag{27}$$

which is a deterministic and diagonal matrix.

4 NTK During Training

Please reference the original papers.

References

- [1] Jacot A, Gabriel F, Hongler C. Neural tangent kernel: Convergence and generalization in neural networks[C]//Advances in neural information processing systems. 2018: 8571-8580.
- [2] Arora S, Du S S, Hu W, et al. On exact computation with an infinitely wide neural net[C]//Advances in Neural Information Processing Systems. 2019: 8141-8150.
- [3] Lee J, Bahri Y, Novak R, et al. Deep neural networks as gaussian processes[J]. arXiv preprint arXiv:1711.00165, 2017.
- [4] Lee J, Xiao L, Schoenholz S S, et al. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent[J]. arXiv preprint arXiv:1902.06720, 2019.
- [5] Yang, Greg. "Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation." arXiv preprint arXiv:1902.04760 (2019).
- [6] Huang W, Du W, Da Xu R Y. On the Neural Tangent Kernel of Deep Networks with Orthogonal Initialization[J]. arXiv preprint arXiv:2004.05867, 2020.