NTK deduce

221900180 田永铭　tianym2022@smail.nju.edu.cn

2024 年 9 月 24 日

section: The network

Denote the network function by $f(\cdot;\theta) : R^{n_0} \to R^{n_L}$, $f_\theta(x) = \tilde{\alpha}^{(L)}(x;\theta)$, where $\tilde{\alpha}$ means pre-activation. Let σ denote the non-linearity, which is Lipschitz and twice differentiable, with bounded second derivative.

And the network can be represented as:

$$\begin{cases} \alpha^{(0)}(x;\theta) = x \\ \tilde{\alpha}^{(\ell+1)}(x;\theta) = \dfrac{1}{\sqrt{n_\ell}} W^{(\ell)} \alpha^{(\ell)}(x;\theta) + \beta b^{(\ell)} \\ \alpha^{(\ell)}(x;\theta) = \sigma(\tilde{\alpha}^{(\ell)}(x;\theta)) \end{cases}$$

Here, $\frac{1}{\sqrt{n_\ell}}$ scales $W^{(\ell)}$ so that $W^{(\ell)} \sim \mathcal{N}(0,1)$, while $\frac{1}{\sqrt{n_\ell}} W^{(\ell)} \sim \mathcal{N}(0, \frac{1}{n_\ell})$, which is the LeCun initialization. β is used to control the influence of the term with $W^{(\ell)}$.

section: NTK from the loss function

Denote the loss function by $\mathcal{L}(\theta) = \frac{1}{N}\Sigma_{i=1}^N l(f(x^{(i)};\theta), y^{(i)})$.

So $\nabla_\theta \mathcal{L}(\theta) = \frac{1}{N}\Sigma_{i=1}^N \nabla_\theta f(x^{(i)};\theta) \nabla_f l(f, y^{(i)})$ (chain rule).

Let us neglect the learning rate, and the learning direction can be expressed as:

$$\frac{d\theta}{dt} = -\nabla_\theta \mathcal{L}(\theta).$$

Again, by the chain rule, we have:

$$\begin{aligned}\frac{df(x;\theta)}{dt} &= \frac{df(x;\theta)}{d\theta}\cdot\frac{d\theta}{dt}\\ &= -\frac{1}{N}\Sigma_{i=1}^{N}\underbrace{\boxed{\nabla_\theta f(x;\theta)^T\nabla_\theta f(x;\theta)}}_{NTK}\nabla_f l(f,y^{(i)})\end{aligned}$$

section: A glimpse of NTK

NTK can be expressed as:

$K(x,x';\theta)=\nabla_\theta f(x;\theta)^T\nabla_\theta f(x';\theta)$,

where $K_{m,n}(x,x';\theta)=\Sigma_{p=1}^{P}\frac{\partial f_m(x;\theta)}{\partial\theta_p}\cdot\frac{\partial f_n(x';\theta)}{\partial\theta_p}$.
P is the number of the arguments of the network, which is easy to compute.

Let's take a look at the change of dimension:

The input x,x' is $\mathrm{n}_0\times 1$, $f:n_0\mapsto n_L$, $\nabla_\theta f(x;\theta)^T$ is $n_L\times P$, and by the effect of $K:R^{n_0}\times R^{n_0}\mapsto R^{n_L}\times R^{n_L}$, the K finally maps the input to a $n_L\times n_L$ matrix.

Denote $\nabla_\theta f(x;\theta)$ by $\varphi(x)$, we get a beautiful form of NTK:

$NTK(x,x')=\langle\varphi(x),\varphi(x')\rangle$, where the$\langle\cdot,\cdot\rangle$ is the inner product.

section: Proof of Proposition 1

Proposition 1: As $n_1,n_2\cdots n_{L-1}\to\infty$, the output functions $f_{\theta,k}$(for $k=1,\cdots,n_L$), tend to i.i.d centred Gaussian processes of covariance $\Sigma^{(L)}$, defined recursively as follows:

$$\begin{cases}\Sigma^{(1)}(x,x')=\dfrac{1}{n_0}x^Tx'+\beta^2\\ \Sigma^{(l+1)}(x,x')=E_{f\sim\mathcal{N}(0,\Sigma^{(l)})}[\sigma(f(x))\sigma(f(x'))]+\beta^2.\end{cases}$$

Proof:

We prove by induction.

① The case when L = 1:

The output is $f(x;\theta) = \tilde{\alpha}^{(1)}(x) = \frac{1}{\sqrt{n_0}} w^{(0)^T} x + \beta b^{(0)}$, where $\tilde{\alpha}^{(1)}_m = \frac{1}{\sqrt{n_0}} \Sigma_{i=1}^{n_0} W^{(0)}_{im} x_i + \beta b^{(0)}_m, 1 \leq m \leq n_1$.

Note the assumption that all the parameters are initialized i.i.d as $\mathcal{N}(0,1)$, we will use that in the deduction. Let's calculate $\Sigma^{(1)}(x,x')$:

$$\begin{aligned}
\Sigma^{(1)}(x,x') &= E[\tilde{\alpha}^{(1)}(x) \cdot \tilde{\alpha}^{(1)}(x')] \qquad \text{(inner product)} \\
&= E\langle \frac{1}{\sqrt{n_0}} w^{(0)^T} x + \beta b^{(0)}, \frac{1}{\sqrt{n_0}} w^{(0)^T} x' + \beta b^{(0)} \rangle \\
&= E\langle \frac{1}{\sqrt{n_0}} w^{(0)^T} x, \frac{1}{\sqrt{n_0}} w^{(0)^T} x' \rangle + E\langle \frac{1}{\sqrt{n_0}} w^{(0)^T} x, \beta b^{(0)} \rangle \\
&\quad + E\langle \frac{1}{\sqrt{n_0}} w^{(0)^T} x', \beta b^{(0)} \rangle + E\langle \beta b^{(0)}, \beta b^{(0)} \rangle.
\end{aligned}$$

Note that we are calculating the expectation of the parameters θ.

There are four terms in the equation above, let's talk about each:

- The fourth term: each component of the term is $\beta^2 b_i^{(0)^2}$, and $b_i^{(0)} \sim \mathcal{N}(0,1)$, it's easy to get that the result is β^2.
- The second and the third terms: They are with a $b^{(0)}$ whose expectation is zero, so the result is zero too.
- The first term:

$$\begin{aligned}
& E\langle \frac{1}{\sqrt{n_0}} w^{(0)^T} x, \frac{1}{\sqrt{n_0}} w^{(0)^T} x' \rangle \\
&= \frac{1}{n_0} E\langle w^{(0)^T} x, w^{(0)^T} x' \rangle \quad (w^{(0)} : n_0 \times n_1, x : n_0 \times 1) \\
&= \frac{1}{n_0} E\langle \Sigma_{i=1}^{n_0} w_i^{(0)^T} x_i, \Sigma_{i=1}^{n_0} w_i^{(0)^T} x'_i \rangle \quad (w_i^{(0)^T} \text{is the column vector of } w^{(0)^T}) \\
&= \frac{1}{n_0} \Sigma_{i=1}^{n_0} E\langle w_i^{(0)^T} x_i, w_i^{(0)^T} x'_i \rangle \quad \text{(by independency)} \\
&= \frac{1}{n_0} (\Sigma_{i=1}^{n_0} E\langle w_i^{(0)^T}, w_i^{(0)^T} \rangle x_i \cdot x'_i) \quad (x_i, x'_i : 1 \times 1) \\
&= \frac{1}{n_0} \Sigma_{i=1}^{n_0} x_i x'_i \\
&= \frac{1}{n_0} x^T x'.
\end{aligned}$$

□

② We assume the conclusion is valid for $L = l$.

So $\tilde{\alpha}_m^{(l)}$ is a GP with covariance $\Sigma^{(l)}$ and $\tilde{\alpha}_i^{(l)}(i = 1, \cdots, n_l)$ are i.i.d.

③ Then we talk about the case when $L = l + 1$:

The output here can be expressed as:

$f(x;\theta) = \tilde{\alpha}^{(l+1)}(x) = \frac{1}{\sqrt{n_l}} {w^{(l)}}^T \sigma(\tilde{\alpha}^{(l)}(x)) + \beta b^{(l)}$,

where $\tilde{\alpha}_m^{(l+1)}(x) = \frac{1}{\sqrt{n_l}} \Sigma_{i=1}^{n_l} w_i {m^{(l)}}^T \sigma(\tilde{\alpha}_i^{(l)}(x)) + \beta b_m^{(l)}, 1 \leq m \leq n_{l+1}$.

Similarly, we have:

$\Sigma^{(\tilde{l+1})}(x, x') = \frac{1}{n_l} \sigma(\tilde{\alpha}^{(l)}(x))^T \sigma(\tilde{\alpha}^{(l)}(x')) + \beta^2$.

Here, by the central limit theorem and the assumption of the induction:

$\Sigma^{(l+1)}(x, x') \to E_{f \sim \mathcal{N}(0, \Sigma^{(l)})}[\sigma(f(x))\sigma(f(x'))] + \beta^2$.

□

section: Proof of Theorem 1

Theorem 1: For a network of depth L at initialization, with a Lipschitz nonlinearity σ, and in the limit as the layers width $n_1, ..., n_{L-1} \to \infty$, the NTK $\Theta^{(L)}$ converges in probability to a deterministic limiting kernel:

$$\Theta^{(L)} \to \Theta_\infty^{(L)} \otimes Id_{n_L}.$$

The scalar kernel $\Theta_\infty^{(L)} : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}$ is defined recursively by

$$\begin{aligned} \Theta_\infty^{(1)}(x, x') &= \Sigma^{(1)}(x, x') \\ \Theta_\infty^{(L+1)}(x, x') &= \Theta_\infty^{(L)}(x, x') \dot{\Sigma}^{(L+1)}(x, x') + \Sigma^{(L+1)}(x, x'), \end{aligned}$$

where

$$\dot{\Sigma}^{(L+1)}(x, x') = \mathbb{E}_{f \sim \mathcal{N}\left(0, \Sigma^{(L)}\right)} \left[\dot{\sigma}(f(x)) \dot{\sigma}(f(x'))\right],$$

taking the expectation with respect to a centred Gaussian process f of covariance $\Sigma^{(L)}$, and where $\dot{\sigma}$ denotes the derivative of σ.

Here,$\otimes$ denotes the inner product, not the Kronecker product! Id_{n_L} denotes the identity matrix with the dimension of n_L.

Proof:

We prove by induction.
① The case when L = 1:

Firstly, we give a proof of the base case which is not that solid, but can serve as a sketch:

The output is $f(x;\theta) = \tilde{\alpha}^{(1)}(x) = \frac{1}{n_0} w^{(0)^T} x + \beta b^{(0)}$.

Split the parameters θ into W and b, we can get:

$$\begin{aligned}
K^{(1)}(x, x';\theta) &= (\frac{\partial f(x';\theta)}{\partial w^{(0)}})^T (\frac{\partial f(x;\theta)}{\partial w^{(0)}}) + (\frac{\partial f(x';\theta)}{\partial b^{(0)}})^T (\frac{\partial f(x;\theta)}{\partial b^{(0)}}) \\
&= \frac{1}{\sqrt{n_0}} \frac{1}{\sqrt{n_0}} x^T x' + \beta \cdot \beta \\
&= \frac{1}{n_0} x^T x' + \beta^2 \\
&= \Sigma^{(1)}(x, x')
\end{aligned}$$

However, the $\delta_{kk'}$ is necessary, in detail:

Take the term $(\frac{\partial f(x';\theta)}{\partial w^{(0)}})^T (\frac{\partial f(x;\theta)}{\partial w^{(0)}})$ for example:

$$f_k(x';\theta) = \frac{1}{\sqrt{n_0}} \begin{bmatrix} w_1^{(0)^T} \\ w_2^{(0)^T} \\ \cdots \\ w_{n_1}^{(0)^T} \end{bmatrix} x' + \beta b^{(0)}.$$

$$f_k'(x';\theta) = \frac{1}{\sqrt{n_0}} \begin{bmatrix} w_1^{(0)^T} \\ w_2^{(0)^T} \\ \cdots \\ w_{n_1}^{(0)^T} \end{bmatrix} x + \beta b^{(0)}.$$

Denote $w_i^{(0)^T}$ as the row vector of $w^{(0)^T}$, and by the independency of $w_i^{(0)^T}$ $(i = 1, \cdots, n_1)$, we have:

$$
\begin{aligned}
\frac{\partial f_k(x';\theta)}{\partial w^{(0)}} &= \left[\frac{\partial f_k(x';\theta)}{\partial w_1^{(0)}}, \frac{\partial f_k(x';\theta)}{\partial w_2^{(0)}}, \cdots, \frac{\partial f_k(x';\theta)}{\partial w_{n_1}^{(0)}}\right] \\
&= \frac{1}{\sqrt{n_0}}\left[\begin{bmatrix} x' \\ 0 \\ \cdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ x' \\ \cdots \\ 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ \\ \cdots \\ x' \end{bmatrix}\right]. \\
\frac{\partial f_k'(x';\theta)}{\partial w^{(0)}} &= \left[\frac{\partial f_k'(x;\theta)}{\partial w_1^{(0)}}, \frac{\partial f_k'(x;\theta)}{\partial w_2^{(0)}}, \cdots, \frac{\partial f_k'(x;\theta)}{\partial w_{n_1}^{(0)}}\right] \\
&= \frac{1}{\sqrt{n_0}}\left[\begin{bmatrix} x \\ 0 \\ \cdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ x \\ \cdots \\ 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ \\ \cdots \\ x \end{bmatrix}\right].
\end{aligned}
$$

Thus, we have:
$(\frac{\partial f(x';\theta)}{\partial w^{(0)}})^T(\frac{\partial f(x;\theta)}{\partial w^{(0)}}) = \frac{1}{n_0}x^T x' Id_{n_1}$, in another word: The first part of $\Theta_{kk'}(x,x') =$

$$
\begin{cases} \dfrac{1}{n_0}x^T x' & , \quad \text{if} \quad k = k' \\ \quad 0 & , \quad \text{if} \quad k \neq k' \end{cases}
$$

That is to say:

The first part of $\Theta_{kk'}(x,x') = \frac{1}{n_0}x^T x' \delta_{kk'}$.

Similarly, we can get the second part of $\Theta_{kk'}(x,x')$, which is $\beta^2\delta_{kk'}$.

$\square$

② We assume the conclusion is valid for $L = l$.

③ Then we talk about the case when $L = l + 1$:

We split the parameters θ into two part, the first l layers($\tilde{\theta}$) and the layer $l+1$ $(w^{(l)}, b^{(l)})$.

By Proposition 1 and the induction hypothesis, as $n_1, ..., n_{l-1} \to \infty$ the pre-activations $\tilde{\alpha}_i^{(l)}$ are i.i.d centered Gaussian with covariance $\Sigma^{(l)}$ and the neural tangent kernel $\Theta_{ii'}^{(l)}(x,x')$

of the smaller network converges to a deterministic limit:

$$\left(\partial_{\tilde{\theta}}\tilde{\alpha}_i^{(l)}(x;\theta)\right)^T \partial_{\tilde{\theta}}\tilde{\alpha}_{i'}^{(l)}(x';\theta) \to \Theta_\infty^{(l)}(x,x')\delta_{ii'}.$$

And we have:

$\partial_{\tilde{\theta}_p} f_{\theta,k}(x) = \frac{1}{\sqrt{n_l}}\Sigma_{i=1}^{n_1}\partial_{\tilde{\theta}_p}\tilde{\alpha}_i^{(l)}(x;\theta)\dot{\sigma}(\tilde{\alpha}_i^l(x;\theta)w_{ik}^l.$

We can prove that by the chain rule:

$$\begin{aligned}
&\frac{\partial f_k(x,\theta)}{\partial\tilde{\theta}_p}\\
&=\langle\frac{\partial f_k(x;\theta)}{\partial\tilde{\alpha}^{(l)}(x;\theta)},\frac{\partial\tilde{\alpha}^{(l)}(x;\theta)}{\partial\tilde{\theta}_p}\rangle\\
&=\Sigma_{i=1}^{n_l}\boxed{\frac{\partial f_k(x;\theta)}{\partial\tilde{\alpha}_i^{(l)}(x;\theta)}}\cdot\frac{\partial\tilde{\alpha}_i^{(l)}(x;\theta)}{\partial\tilde{\theta}_p}\\
&=\Sigma_{i=1}^{n_l}\boxed{\frac{\partial f_k(x;\theta)}{\partial\alpha_i^{(l)}(x;\theta)}\cdot\frac{\partial\alpha_i^{(l)}(x;\theta)}{\partial\tilde{\alpha}_i^{(l)}(x;\theta)}}\cdot\frac{\partial\tilde{\alpha}_i^{(l)}(x;\theta)}{\partial\tilde{\theta}_p}\\
&=\Sigma_{i=1}^{n_l}\frac{1}{\sqrt{n_l}}w_{ik}^{(l)}\cdot\dot{\sigma}\cdot\frac{\partial\tilde{\alpha}_i^{(l)}(x;\theta)}{\partial\tilde{\theta}_p}.
\end{aligned}$$

□

Now let's get the first part of the kernel $\Theta_{kk'}^{(l+1)}(x;x')$(with the parameters of the first l layers), which is the inner product of $\frac{\partial f_k(x,\theta)}{\partial\tilde{\theta}_p}$ and $\frac{\partial f_k(x',\theta)}{\partial\tilde{\theta}_p}$:

$$\begin{aligned}
&\Theta_{kk'}^{(l+1)}(x;x') \quad \text{(the first part)}\\
&=\frac{1}{n_l}\Sigma_{i=1,i'=1}^{n_l}\theta_{ii'}^{(l)}(x,x')\dot{\sigma}(\tilde{\alpha}_i^{(l)}(x;\theta))\dot{\sigma}(\tilde{\alpha}_{i'}^{(l)}(x';\theta))w_{ik}^{(l)}w_{i'k'}^{(l)}\\
&\to\frac{1}{n_l}\Sigma_{i=1}^{n_l}\theta_\infty^{(l)}(x,x')\dot{\sigma}(\tilde{\alpha}_i^{(l)}(x;\theta))\dot{\sigma}(\tilde{\alpha}_i^{(l)}(x';\theta))w_{ik}^{(l)}w_{ik'}^{(l)} \qquad \text{(by the assumption of the induction)}\\
&\to\Theta_\infty^{(l)}(x,x')\dot{\Sigma}^{(l+1)}(x,x')\delta_{kk'} \qquad \text{(by the central limit theorem).}
\end{aligned}$$

As for the second part of $\Theta_{kk'}^{(l+1)}(x;x')$ (with the parameters of the lth layer):

$$
\begin{aligned}
&\Theta^{(l+1)}_{kk'}(x;x') \quad \text{(the second part)}\\
&= (\frac{\partial f(x';\theta)}{\partial w^{(l)}})^T(\frac{\partial f(x;\theta)}{\partial w^{(l)}}) + (\frac{\partial f(x';\theta)}{\partial b^{(l)}})^T(\frac{\partial f(x;\theta)}{\partial b^{(l)}})\\
&\to \Sigma^{(l+1)}\delta_{kk'} \quad \text{(similarly as we prove the base case)}
\end{aligned}
$$

And the result is now obvious.

□

section: Train

pass

section: The positiveness of the $\Theta^{(L)}_\infty$

pass

section: references

[1] Some Math behind Neural Tangent Kernel

[2] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks, February 2020. arXiv:1806.07572 [cs, math, stat].