Stat 946 - Topics in Probability and Statistics: Mathematical Foundations of Deep Learning *Lecture 6*

Instructor: *Mufan Li*, Scribe: *Xiaoxi Luo*

September 22, 2025

1 Review of NTK

$$K^{\alpha\beta} = K(x^\alpha, x^\beta) = \left\langle \nabla_\theta f(x^\alpha;\theta), \nabla_\theta f(x^\beta;\theta) \right\rangle.$$

We can interpret it as a kernel. As $n \to \infty$, there are two properties:

1. k^{θ^t} becomes a deterministic function of x^α, x^β.
2. k^{θ^t} is constant in time (in gradient flow training).

These two things guarantees that (1) Neural networks converge exponentially, and (2) We can interpret neural networks as linear method (see below).

Linearization

Discrete time update:

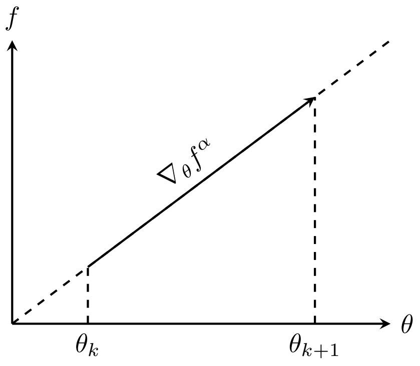
$$\theta_{k+1} = \theta_k - \eta \nabla_\theta \mathcal{L}(\theta_k).$$

By Taylor's theorem:

$$\begin{aligned} f(x^\alpha;\theta_{t+1}) &= f(x^\alpha;\theta_t) + \langle \nabla_\theta f(x^\alpha;\theta_t), \theta_{t+1} - \theta_t \rangle + \overbrace{\nabla_\theta^2 f(x^\alpha;\theta^*)}^{\text{a function of 2 tensors}} \quad [\theta_{k+1} - \theta_k]^{\otimes 2} \\ &= f(x^\alpha;\theta_t) - \eta \frac{1}{m} \sum_{\beta=1}^m K^{\alpha\beta}(f^\beta - y^\beta) \\ &\quad + \frac{\beta^2}{2m^2} \sum_{\gamma,\beta=1}^m (f^\alpha - y^\alpha)(f^\beta - y^\beta) \underbrace{((\nabla f^\beta)^\top \cdot \nabla^2 f^\alpha \cdot \nabla f^\gamma)}_{\mathcal{O}(n^{-\frac{1}{2}}) \to 0} \end{aligned}$$

- That is to say, Taylor's remaining term goes to zero when $n \to \infty$. Only linear term remains, and we characterize it with tangent kernel.
- $f(x^\alpha;\theta)$ is linear in θ. See the figure below.
- We are actually doing linear regression with features $\nabla_\theta f$. The gradient of f has two terms: The first set of features are the hidden layers at initialization (random), same as GP features.

$$\begin{aligned} \nabla_{W_1} f^\alpha &= \frac{1}{\sqrt{n}} \varphi(W_0 x^\alpha)^\top \\ \nabla_{W_0} f^\alpha &= \frac{1}{n} \operatorname{diag}\big(\varphi'(W_0 x^\alpha)\big)(x^\alpha W_1)^\top \end{aligned}$$



2 Extension to MLPs

Depth d (finite):

$$f_\theta^d(x,\Theta) = \frac{1}{\sqrt{n}} \overbrace{W_d}^{1\times n} \varphi(\overbrace{h^{d-1}}^{n\times 1})$$

$$h_{\ell+1}^\alpha = \frac{1}{\sqrt{n}} \overbrace{W_\ell}^{n\times n} \varphi(\overbrace{h_{\ell-1}^\alpha}^{n\times 1}), \quad h_1^\alpha = \frac{1}{\sqrt{n_0}} W_0 x^\alpha$$

with

$$\Theta = \{W_\ell\}_{\ell=1}^d \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad L(\theta) = \frac{1}{2m}\sum_\beta \left(f^\beta - y^\beta\right)^2, \quad \partial_t \theta_t = -\eta \nabla_\theta \mathcal{L}(\theta_t)$$

Since η is a constant at present, we can safely ignore it.

$$k^{\alpha\beta}(x^\alpha, x^\beta) = \left\langle \nabla_\theta f(x^\alpha;\theta), \nabla_\theta f(x^\beta;\theta) \right\rangle = \sum_{\ell=1}^d \left\langle \nabla_{W_\ell} f(x^\alpha;\theta), \nabla_{W_\ell} f(x^\beta;\theta)\right\rangle \tag{1}$$

How do we calculate the gradient of W_ℓ? Consider an entry first:

$$\frac{\partial f^\alpha}{\partial W_{\ell,ij}} = \sum_{k=1}^n \frac{\partial f^\alpha}{\partial h_{\ell+1,k}} \frac{\partial h_{\ell+1,k}}{\partial W_{\ell,ij}}.$$

Since

$$\frac{\partial h_{\ell+1,k}}{\partial W_{\ell,ij}} = \frac{\partial}{\partial W_{ij}}\left(\frac{1}{\sqrt{n}} \sum_{k'} W_{l,kk'} \varphi(h_{lk'}^\alpha)\right)$$

we get

$$\frac{\partial f^\alpha}{\partial W_{\ell,ij}} = \sum_{k=1}^n \sum_{k'=1}^n \frac{\partial f^\alpha}{\partial h_{\ell+1,k}} \frac{1}{\sqrt{n}} \delta_{ik}\delta_{jk'} \varphi(h_{\ell,k'}^\alpha) = \frac{1}{\sqrt{n}} \frac{\partial f^\alpha}{\partial h_{\ell+1,i}} \varphi(h_{\ell,j}^\alpha).$$

Back to Eq. (1),

$$\left\langle \nabla_{W_\ell} f(x^\alpha;\theta), \nabla_{W_\ell} f(x^\beta;\theta)\right\rangle = \sum_{i,j=1}^n \frac{1}{n} \frac{\partial f^\alpha}{\partial h_{\ell+1,i}^\alpha} \frac{\partial f^\beta}{\partial h_{\ell+1,i}^\beta} \varphi(h_{\ell,j}^\alpha)\varphi(h_{\ell,j}^\beta)$$

$$= \underbrace{\sum_i \frac{\partial f^\alpha}{\partial h_{\ell+1,i}} \frac{\partial f^\beta}{\partial h_{\ell+1,i}}}_{\text{Nice if it is also a kernel!}} \overbrace{\frac{1}{n}\left\langle \varphi(h_\ell^\alpha), \varphi(h_\ell^\beta)\right\rangle}^{\Phi_l,\ \text{GP kernel}}$$

$$\frac{\partial f^\alpha}{\partial h_\ell^\alpha} = \frac{\partial f^\alpha}{\partial h_{\ell+1}^\alpha}\frac{\partial h_{\ell+1}^\alpha}{\partial h_\ell^\alpha} = \frac{\partial f^\alpha}{\partial h_{\ell+1}^\alpha}\left(\frac{1}{\sqrt{n}}W_l\frac{\partial}{\partial h_\ell^\alpha}\varphi(h_\ell^\alpha)\right)$$
$$= \frac{1}{\sqrt{n}}\mathrm{diag}\big(\varphi'(h_\ell^\alpha)\big)W_\ell^\top\frac{\partial f^\alpha}{\partial h_{\ell+1}}.$$

Note: $\mathrm{diag}(\varphi'(h_\ell^\alpha))$ only has elements if the indices match. We make the convention that "neurons $\in \Theta(1)$".

Then we define "backward" (post-activation) neurons:

$$g_\ell^\alpha = \sqrt{n}\frac{\partial f^\alpha}{\partial h_\ell}$$

Then the NTK can be written as

$$K^{\alpha\beta} = \sum_{\ell=0}^{d}\frac{1}{n}\langle g_{\ell+1}^\alpha, g_{\ell+1}^\beta\rangle\Phi_\ell^{\alpha\beta}$$

If we define $\langle g_{\ell+1}^\alpha, g_{\ell+1}^\beta\rangle$ as $G_{\ell+1}^{\alpha\beta}$, a backward covariance/kernel, then we could write NTK as

$$K^{\alpha\beta} = \sum_{i=0}^{d} G_{\ell+1}^{\alpha\beta}\Phi_\ell^{\alpha\beta}$$

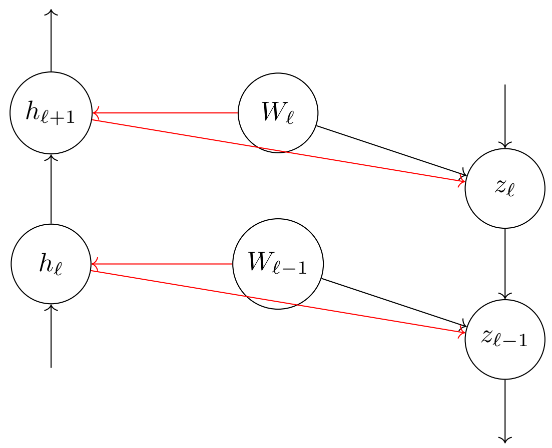
The recursion of g_l^α is given by

$$g_\ell^\alpha = \frac{1}{\sqrt{n}}\,\mathrm{diag}\big(\phi'(h_\ell^\alpha)\big)W_\ell^\top g_{\ell+1}^\alpha.$$

Define the backward "pre-activation"

$$z_\ell^\alpha = \frac{1}{\sqrt{n}}W_\ell^\top g_{\ell+1}^\alpha = \frac{1}{\sqrt{n}}W_\ell^\top\cdot\frac{1}{\sqrt{n}}\underbrace{\mathrm{diag}(\varphi'(h_{\ell+1}^\alpha))}_{D_{\ell+1}^\alpha}W_{\ell+1}^\top\ g_{\ell+2}^\alpha = \frac{1}{\sqrt{n}}W_\ell^\top D_{\ell+1}^\alpha z_{\ell+1}^\alpha \tag{2}$$

The whole structure looks like



In the left way, the $\{h_\ell\}$ have the Markov property. It would be great if $\{z_\ell\}$ is also a Markov chain! If we condition on $h_\ell, h_{\ell+1}, \cdots$, we can remove all the red edges.

However, after conditioning on $\{h_k^\alpha\}_{k,\alpha}$, the random variables W_ℓ and $W_{\ell-1}$ are no longer i.i.d. $N(0,1)$! To handle the new problem, we refer to the following lemma:

Lemma 1 (**Gaussian Condition**)**.** *Let $W \in \mathbb{R}^{n\times m}$ with entries $W_{ij} \overset{iid}{\sim} N(0,1)$. For a deterministic $\phi \in \mathbb{R}^{n\times m}$, we have*

$$W \mid \sigma(W\varphi) \overset{d}{=} WP_\varphi + \widetilde{W}P_\varphi^\perp,$$

where

- *$P_\varphi = \varphi(\varphi^\top\varphi)^\dagger\varphi^\top$ is the projection onto the column space of φ (with $\dagger$ denoting the pseudo-inverse),*
- *$\widetilde{W}$ is an independent copy of W, also with $\widetilde{W}_{ij} \overset{iid}{\sim} N(0,1)$ entries.*

Example Let

$$W = [g_1, g_2], \quad \phi = \begin{bmatrix}1\\0\end{bmatrix}, \quad W\phi = g_1.$$

Then

$$W \mid \sigma(W\phi) \overset{d}{=} \begin{bmatrix}g_1\\0\end{bmatrix} + \begin{bmatrix}0\\\widetilde{g}_2\end{bmatrix},$$

where $\tilde{g}_2$ is an independent copy of g_2.

Define the filter as

$$\mathcal{F}^z_{\ell+1} = \sigma(\{h^\alpha_k\}_{k,\alpha}, \{z^\alpha_k\}_{k\geq\ell+1,\alpha})$$

We apply Lemma 1 on z^α_ℓ, and have

$$\begin{aligned} z^\alpha_\ell \mid F^z_{\ell+1} &= \frac{1}{\sqrt{n}} W_\ell^\top D^\alpha_{\ell+1} z^\alpha_{\ell+1} \mid F^z_{\ell+1} = \frac{1}{\sqrt{n}} (P_{\varphi_\ell} W_\ell^\top + P^\perp_{\varphi_\ell} \widetilde{W}_\ell^\top) D^\alpha_{\ell+1} z^\alpha_{\ell+1} \mid F^z_{\ell+1} \\ &= P_{\varphi_\ell} z^\alpha_\ell + P^\perp_{\varphi_\ell} \cdot \underbrace{\frac{1}{\sqrt{n}} \widetilde{W}_\ell^\top \overbrace{D^\alpha_{\ell+1} z^\alpha_{\ell+1}}^{g^\alpha_{\ell+1}}}_{:=\widetilde{z}^\alpha_\ell \mid \mathcal{F}^z_{\ell+1}} \mid \mathcal{F}^z_{\ell+1} \end{aligned}$$

The first term is due to recursion of $z^\alpha_\ell \mid \mathcal{F}^z_{\ell+1}$ (Eq. 2). In the second term, $\widetilde{z}^\alpha_\ell \mid \mathcal{F}^z_{\ell+1} \sim \mathcal{N}(0, G_{\ell+1} \otimes I_n)$.