STAT 946 - Topics in Probability and Statistics: Mathematical Foundations of Deep Learning
Lecture 4

Lucas Noritomi-Hartwig
University of Waterloo

September 15, 2025 from 16h00 to 17h20 in M3 3103

1 Extension to Deep Networks

Define $f(X^\alpha;\, \theta) = \frac{1}{\sqrt{n}} \underbrace{W_d}_{1\times n} \varphi\left(\underbrace{h_d^\alpha}_{n\times 1}\right)$ where $\alpha = 1, \ldots, m$ is the the data index.

$$h_{l+1}^\alpha = \frac{1}{\sqrt{n}} \underbrace{W_l}_{n\times n} \varphi\left(\underbrace{h_l^\alpha}_{n\times 1}\right)$$
$$h_1^\alpha = \frac{1}{\sqrt{n_0}} \underbrace{W_0}_{n\times n_0} \underbrace{x^\alpha}_{n_0\times 1}$$

$W \in \mathbb{R}^{n\times n}$, $W_{i,\,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,\,1)$, $u,\, v \in \mathbb{R}^n$.

$$\begin{bmatrix} Wu \\ Wv \end{bmatrix} \sim \mathcal{N}\left(0,\, \begin{bmatrix} |u|^2 I_n & \langle u,\, v\rangle\, I_n \\ \langle u,\, v\rangle\, I_n & |v|^2 I_n \end{bmatrix} = \begin{bmatrix} |u|^2 & \langle u,\, v\rangle \\ \langle u,\, v\rangle & |v|^2 \end{bmatrix} \otimes I_n\right)$$

where $A \otimes B = [a_{i,\,j} B]_{i,\,j}$.

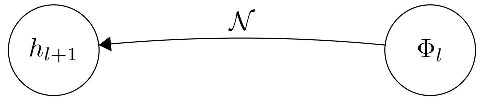
Let $u^\alpha \in \mathbb{R}^n$, $\alpha \in [1:\, m]$

$$[\underbrace{W}_{nm\times 1} u^\alpha]_{\alpha=1}^m \sim \mathcal{N}\left(0,\, [\langle u^\alpha,\, u^\beta\rangle]_{\alpha,\,\beta=1}^m \otimes I_n\right)$$
$$h_{l+1}^\alpha = \frac{1}{\sqrt{n}} W_l \varphi\left(h_l^\alpha\right)$$

Condition on $\mathcal{F}_l = \sigma\left(\{h_k^\alpha\}_{\alpha\in[1:\, m],\, k\le l}\right) \leftarrow \sigma$-algebra.

$$[h_{l+1}^\alpha]_{\alpha=1}^m | \mathcal{F}_l \sim \mathcal{N}\left(0,\, \underbrace{\left[\tfrac{1}{n}\left\langle \varphi\left(h_l\right)^\alpha,\, \varphi\left(h_l^\beta\right)\right\rangle\right]_{\alpha,\,\beta}}_{\Phi_l \in \mathbb{R}^{m\times m}} \otimes I_n\right), \quad \text{where } u^\alpha = \frac{1}{\sqrt{n}} \varphi\left(h_l^\alpha\right)$$

- To characterize neural networks at initialization,



  we only need Φ_l.

- Φ_{l+1} is a deterministic function of $\left[h_{l+1}^{\alpha}\right]_{\alpha=1}^{m}$, i.e.,

$$
\begin{aligned}
\Phi_{l+1} &\overset{\text{det.}}{\longleftarrow} hl+1|\mathcal{F}_l \longleftarrow \Phi_l \\
\Phi_{l+1} &\longleftarrow \Phi_l \\
\Phi_{l+1}|\mathcal{F}_l &\overset{d}{=} \Phi_{l+1}|\sigma\left(\Phi_l\right) && \text{(Weak Markov property)}
\end{aligned}
$$

Define the function $f_n : \Phi_l \mapsto \Phi_{l+1}$ (random map), and $f = \lim_{n\to\infty} f_n$ (deterministic).

$$
\begin{aligned}
\Phi_{l+1}^{\alpha\beta}|\mathcal{F}_l &= \frac{1}{n}\left\langle \varphi\left(h_{l+1}^{\alpha}\right), \varphi\left(h_{l+1}^{\beta}\right)\right\rangle |\mathcal{F}_l \\
&= \frac{1}{n}\sum_{j=1}^{n} \varphi\left(h_{l+1,\,j}^{\alpha}\right)\varphi\left(h_{l+1,\,j}^{\beta}\right)|\mathcal{F}_l \\
&\longrightarrow \mathbb{E}\left[\varphi\left(h_{l+1,\,j}^{\alpha}\right)\varphi\left(h_{l+1,\,j}^{\beta}\right)|\mathcal{F}_l\right]
\end{aligned}
$$

Adding 0...

$$
\begin{aligned}
\Phi_{l+1}^{\alpha\beta}|\mathcal{F}_l &= f\left(\Phi_l\right)^{\alpha\beta} + \frac{1}{n}\sum_{j=1}^{n}\underbrace{\left(\varphi\left(h_{l+1,\,j}^{\alpha}\right)\varphi\left(h_{l+1,\,j}^{\beta}\right) - f\left(\Phi_l\right)^{\alpha\beta}\right)}_{\text{zero mean iid}}|\mathcal{F}_l \\
&= f\left(\Phi_l\right)^{\alpha\beta} + \underbrace{\frac{1}{\sqrt{n}}}_{\text{extra factor}} \underbrace{\frac{1}{\sqrt{n}}\sum_{j=1}^{n} Z_j}_{\Theta(1),\text{ i.e., all moments are }\Theta(1)}, \quad \text{where } \forall j\in[1:\,n],\, Z_j \overset{\text{iid}}{\sim} \mathcal{N}(0,\,1)
\end{aligned}
$$

In context $n\to\infty$, $q_n \in \Theta\left(p(n)\right)$ if $\exists c, C > 0$ such that $cp(n) \le q_n \le Cp(n)$.

This implies that

$$
\Phi_{l+1} = f\left(\Phi_l\right) + \underbrace{\Theta\left(\frac{1}{\sqrt{n}}\right)}_{\to 0}
$$

Thus, in the limit, as $n\to\infty$, Φ_l is deterministic.

$$
\begin{aligned}
\implies \quad \Phi_l &= \underbrace{f\circ\ldots\circ f}_{l\text{ times}}\left(\Phi_0\right) \\
\Phi_0 &= \left[\tfrac{1}{n_0}\left\langle x^{\alpha},\, x^{\beta}\right\rangle\right]_{\alpha,\,\beta=1}^{m}
\end{aligned}
$$

Theorem (NNGP)

- Assume ϕ is "nice" (polynomial tail)

- $n \to \infty$, $\Phi_l \xrightarrow{P} f^{\circ l}(\Phi_0)$

Sequential limits $\neq$ joint limits (joint is a stronger case than sequential).

$$\left(1 + \frac{1}{n}\right)^d = \begin{cases} 1, & n \to \infty \text{ first} \\ \infty, & d \to \infty \text{ first} \\ e^{\frac{d}{n}}, & \frac{d}{n} \to \text{ const.} \end{cases}$$

Order of limits matter.

2 Neural Tangent Kernel (NTK)

(If you come up with a name as good as this, you don't have to do all the theory. The name will stick and people will cite your work.)

$\sim$ Fall 2018, $\sim$ five articles that studied wide neural network training.

- Three of the five articles: Du et al., Allen-Zhu et al., Zou et al. showed that neural network training actually converges.
- Lee et al. (2018) (same group as the NNGP article) showed that the neural network training is linear.
- (Arthur) Jacot et al. (2018) coined the term NTK (wrote this as a PhD student).

Neural network: $f(x^\alpha;\, \theta_k)$, where k is the training time index.

Define the loss function:

$$\mathcal{L}(\theta) = \frac{1}{2m} \sum_{\alpha=1}^{m} \left(f(x^\alpha;\, \theta) - y^\alpha\right)^2$$

$$\theta_{k+1} = \theta_k - \eta \nabla \mathcal{L}(\theta_k) \qquad (\eta > 0)$$

Training:

$$f(x^\alpha;\, \theta_{k+1}) = f(x^\alpha;\, \theta_k) + \langle \nabla_\theta f(x^\alpha;\, \theta_k),\, \theta_{k+1} - \theta_k \rangle + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right),$$

i.e., f is a linear function in terms of θ (not ideal).