# STAT 946 - Topics in Probability and Statistics: Mathematical Foundations of Deep Learning
# Lecture 15
# Professor Mufan Li

Lucas Noritomi-Hartwig
University of Waterloo

November 5, 2025 from 16h00 to 17h20 in M3 3103

Recall: From Hanin and Nica (2019) "Products of Many Random Matrices"

Result:

$$h_{l+1} = \frac{1}{\sqrt{n}} W_l h_l \qquad W_{l,\,jk} \overset{\text{iid}}{\sim} \mathcal{N}(0,\,1)$$
$$h_1 = \frac{1}{\sqrt{n_0}} W_0 x$$
$$\log\left(\frac{1}{n}|h_d|^2\right) - \log\left(\frac{1}{n_0}|x|^2\right) \overset{d,\,n\to\infty,\,\frac{d}{n}\to\bar{\tau}}{\longrightarrow} \mathcal{N}(-\bar{\tau},\,2\bar{\tau})$$

This is the law of geometric Brownian motion:

$$dX_\tau = \sqrt{2} X_\tau dB_\tau$$

Li (this Professor Mufan Li), Nica, Roy (2022) "Neural Covariance SDEs ..."

Recall: $h_{l+1} \mid \mathcal{F}_l \left(= \sigma\left((h_k)_{k\le l}\right)\right) \overset{\text{d}}{=} \mathcal{N}\left(0,\, \underbrace{\frac{1}{n}|h_l|^2}_{\Phi_l} \otimes I_n\right)$

$$\begin{aligned}
\Phi_{l+1} &= \Phi_l \frac{1}{n}|\xi_l|^2, \qquad \xi_l \sim \mathcal{N}(0,\,I) \\
&= \Phi_l\left(1 + \sqrt{\frac{2}{n}}\xi_l\varepsilon_l\right), \qquad \varepsilon_l \sim \mathcal{D}\left(\mathbb{E}[\varepsilon_l] = 0,\, \mathbb{E}\left[\varepsilon_l^2\right] = 1\right) \\
\implies \quad \Phi_{l+1} &= \Phi_l + \underbrace{\frac{1}{\sqrt{n}}}_{\text{step } \frac{1}{n}} \underbrace{\sqrt{2}\Phi_l}_{\sigma(\Phi_l)} \underbrace{\varepsilon_l}_{\text{zero mean}} \\
\Phi_\tau^{(n)} &:= \Phi_{\lfloor \tau n \rfloor} \\
&\overset{d,\,n\to\infty,\,\frac{d}{n}\to\bar{\tau}}{\longrightarrow} d\Phi_\tau \\
&= \sqrt{2}\Phi_\tau dB_\tau
\end{aligned}$$

Geometric Brownian motion!

$$\text{Itô} \quad d\log(\Phi_\tau) = (-1)\, d\tau + \sqrt{2} dB_\tau$$

Thus,

$$\begin{aligned}\log(\Phi_\tau) - \log(\Phi_0) &= \int_0^\tau (-1)\, d\tau' + \int_0^\tau \sqrt{2} dB_{\tau'} \\ &= -\tau + \sqrt{2} B_\tau \\ &\sim \mathcal{N}(-\tau,\, 2\tau)\end{aligned}$$

**Mutliple data points**: $m \geq 2$.

$$\begin{aligned}h_{l+1}^\alpha &= \frac{1}{\sqrt{n}} W_l h_l^\alpha \\ h_1^\alpha &= \frac{1}{\sqrt{n_0}} W_0 x^\alpha\end{aligned}$$

Recall:

$$\left[h_{l+1}^\alpha\right]_{\alpha=1}^m \big|\, \mathcal{F}_l \left(= \sigma\left((h_k)_{k\leq l}\right)\right) \overset{\mathrm{d}}{=} \mathcal{N}\left(0,\, \underbrace{\left[\tfrac{1}{n}\left\langle h_l^\alpha,\, h_l^\beta \right\rangle\right]_{\alpha,\,\beta=1}^m}_{\Phi_l = \left[\Phi_l^{\alpha\beta}\right]_{\alpha\beta}} \otimes I_n\right)$$

$$\begin{aligned}\mathbb{E}\left[\Phi_{l+1}^{\alpha\beta} \mid \mathcal{F}_l\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n h_{l+1,\,j}^\alpha h_{l+1,\,j}^\beta \mid \mathcal{F}_l\right] \\ &= \mathbb{E}\left[h_{l+1,\,j}^\alpha h_{l+1,\,j}^\beta \mid \mathcal{F}_l\right], \quad \forall j \in [1:\, n] \\ &= \Phi_l^{\alpha\beta}\end{aligned}$$

Subtract the mean:

$$\begin{aligned}\Phi_{l+1}^{\alpha\beta} - \Phi_l \mid \mathcal{F}_l &= \frac{1}{n}\sum_{j=1}^n h_{l+1,\,j}^\alpha h_{l+1,\,j}^\beta - \Phi_l^{\alpha\beta} \mid \mathcal{F}_l \\ &= \underbrace{\frac{1}{\sqrt{n}}}_{\text{step size } \frac{1}{n}} \underbrace{\frac{1}{\sqrt{n}}\sum_{j=1}^n \left(\underbrace{h_{l+1,\,j}^\alpha h_{l+1,\,j}^\beta - \Phi_l^{\alpha\beta}}_{\text{zero mean}}\right)}_{=:\xi_l^{\alpha\beta}\ \text{CLT term} = \Theta(1)} \mid \mathcal{F}_l\end{aligned}$$

Thus,

$$\Phi_{l+1}^{\alpha\beta} = \Phi_l^{\alpha\beta} + \frac{1}{\sqrt{n}} \xi_l^{\alpha\beta}, \quad \alpha,\, \beta \in [1:\, m]$$

We will flatten the Markov chain:

- $\Phi$ is symmetric (PSD)
- $\mathrm{vec} : \mathrm{Sym}(m) \to \mathbb{R}^{\bar{m} = m(m+1)/2}$, $\mathrm{vec} : \Phi \mapsto \left[\Phi^{\alpha\beta}\right]_{\alpha\leq\beta} \to$ upper-triangular entries (unique entries)

Markov chain is always in the space $\mathbb{R}^{\bar{m}}$. So,

$$\operatorname{vec}(\Phi_{l+1}) = \operatorname{vec}(\Phi_l) + \frac{\operatorname{vec}(\xi_l)}{\sqrt{n}}, \qquad \xi_l \to \left[\xi_l^{\alpha\beta}\right]_{\alpha\beta}$$

Abusing notation:

$$\Phi_{l+1} = \Phi_l + \frac{1}{\sqrt{n}}\xi_l$$

We nned the covariance matrix: $\operatorname{Cov}(\xi_l)$:

$$\begin{aligned}
\mathbb{E}\left[\xi_l^{\alpha\beta}\xi_l^{\gamma\delta}\right] &= \mathbb{E}\left[\left(h_{l+1,\,j}^{\alpha}h_{l+1,\,j}^{\beta} - \Phi_l^{\alpha\beta}\right)\left(h_{l+1,\,j}^{\gamma}h_{l+1,\,j}^{\delta} - \Phi_l^{\gamma\delta}\right)\right] \\
&= \mathbb{E}\left[h^{\alpha}h^{\beta}h^{\gamma}h^{\delta}\right] - \Phi^{\alpha\beta}\Phi^{\gamma\delta} \\
&= \left(\mathbb{E}\left[h^{\alpha}h^{\beta}\right]\mathbb{E}\left[h^{\gamma}h^{\delta}\right] + \mathbb{E}\left[h^{\alpha}h^{\gamma}\right]\mathbb{E}\left[h^{\beta}h^{\delta}\right] + \mathbb{E}\left[h^{\alpha}h^{\delta}\right]\mathbb{E}\left[h^{\gamma}h^{\beta}\right]\right) - \Phi^{\alpha\beta}\Phi^{\gamma\delta} \qquad \text{(Wick's Isserlis)} \\
&= \left(\Phi^{\alpha\beta}\Phi^{\gamma\delta} + \Phi^{\alpha\delta}\Phi^{\beta\delta} + \Phi^{\alpha\delta}\Phi^{\gamma\beta}\right) - \Phi^{\alpha\beta}\Phi^{\gamma\delta} \\
&= \Phi^{\alpha\delta}\Phi^{\beta\delta} + \Phi^{\alpha\delta}\Phi^{\gamma\beta}
\end{aligned}$$

where $\xi_l^{\alpha\beta} \in \mathbb{R}^{\bar{m}}$ and $\alpha \le \beta$, $\gamma \le \delta$.

Thus,

$$\operatorname{vec}(\Phi_{l+1}) = \operatorname{vec}(\Phi_l) + \frac{1}{\sqrt{n}}\underbrace{\Sigma(\Phi_l)^{1/2}}_{\mathbb{R}^{\bar{m}\times\bar{m}}}\varepsilon_l, \quad \varepsilon_l \in \mathbb{R}^{\bar{m}},\ \varepsilon_l \sim \mathcal{D}\left(\mathbb{E}[\varepsilon_l] = 0,\ \mathbb{E}\left[\varepsilon_l^2\right] = 1\right)$$

$$\Sigma(\Phi)^{\alpha\beta,\,\gamma\delta} = \Phi^{\alpha\delta}\Phi^{\beta\delta} + \Phi^{\alpha\delta}\Phi^{\gamma\beta}$$

$$\xrightarrow{d,\,n\to\infty,\,\frac{d}{n}\to\bar{\tau}} \begin{cases} d\underbrace{\Phi_\tau}_{\mathbb{R}^{\bar{m}}} = \Sigma(\Phi_\tau)^{1/2}\,d\underbrace{B_\tau}_{BM \text{ on } \mathbb{R}^{\bar{m}}} \\ \Phi_0 = \left[\frac{1}{n_0}\left\langle x^{\alpha},\, x^{\beta}\right\rangle\right]_{\alpha,\,\beta=1}^{m} \end{cases}$$

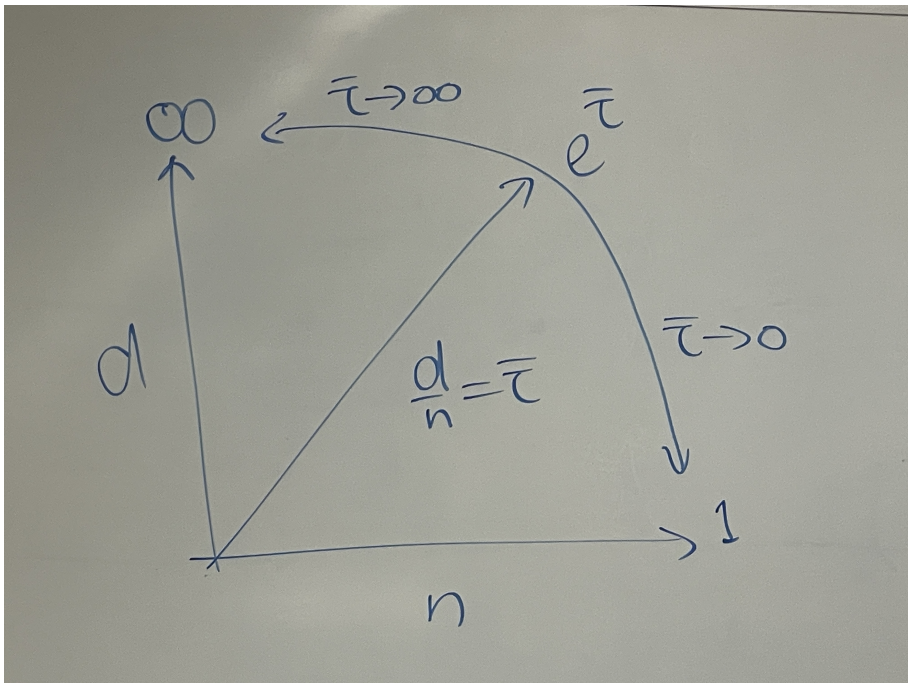
$$\Phi_{\bar{\tau}} \leftarrow \left[\frac{1}{n_0}\left\langle h_d^{\alpha},\, h_d^{\beta}\right\rangle\right]_{\alpha,\,\beta=1}^{m}$$

$$d\Phi_\tau = \underbrace{\Sigma(\Phi_\tau)^{1/2}}_{\text{inverse Riemannian metric (later)}} dB_\tau$$

Intuition:

$$\left(\underbrace{1}_{\text{each layer} \approx 1} + \frac{1}{n}\right)^{d} \to \begin{cases} 1, & n \to \infty \\ \infty, & d \to \infty \\ e^{\bar{\tau}}, & n \to \infty,\ d \to \infty,\ \frac{d}{n} \to \bar{\tau} \end{cases}$$

$$\prod_{l=1}^{d-1}\left(1 + \sqrt{\frac{2}{n}}\varepsilon_l\right)^{d}$$




## Nonlinear Networks

$(m = 1)$

Most naïve thing one cane do:

$$h_{l+1} = \frac{1}{\sqrt{n}} W_l \varphi(h_l)$$

$$h_{l+1} \mid \mathcal{F}_l \sim \mathcal{N}\left(0, \underbrace{\frac{1}{n}|\varphi(h_l)|^2}_{\Phi_l} \otimes I_n\right)$$

$$\mathbb{E}[\Phi_{l+1} \mid \mathcal{F}_l] = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \varphi(h_{l+1,j})^2 \mid \mathcal{F}_l\right]$$

$$= \mathbb{E}\left[\underbrace{\varphi(h_{l+1,j})^2}_{h_{l+1,j}^2 \mathbb{1}_{\{h_{l+1,j}>0\}}}\right]$$

$$\begin{aligned}
\mathbb{E}\left[w^2 \mathbb{1}_{\{w>0\}}\right] &= \int_0^\infty x^2 d\rho(x), \quad \rho(x) \sim \mathcal{N}(0, \Phi) \\
&= \frac{1}{2} \underbrace{\int_{-\infty}^{\infty} x^2 d\rho(x)}_{\mathbb{E}[w^2]=\Phi} \\
\implies \mathbb{E}[\Phi_{l+1} \mid \mathcal{F}_l] &= \frac{1}{2}\Phi_l \\
\Phi_{l+1} &= \underbrace{\frac{\Phi}{2}}_{\text{unstable! } \Phi_l \to 0 \text{ as } l\to\infty} + \frac{1}{\sqrt{n}}\xi_l
\end{aligned}$$

(Kaiming)He - Init.

$$\begin{aligned}
\frac{1}{c} &= \mathbb{E}\left[\varphi(w)^2\right], \qquad w \sim \mathcal{N}(0, \infty) \\
\varphi(|c|x) &= |c|\varphi(x) \\
\implies h_{l+1} &= \sqrt{\frac{c}{n}} W_l \varphi(h_l) \\
h_{l+1} \mid \mathcal{F}_l &\overset{\mathrm{d}}{=} \mathcal{N}\left(0, \underbrace{\frac{c}{n}|\varphi(h_l)|^2}_{\Phi_l \text{ Def'n changed!}} \otimes I_n\right)
\end{aligned}$$

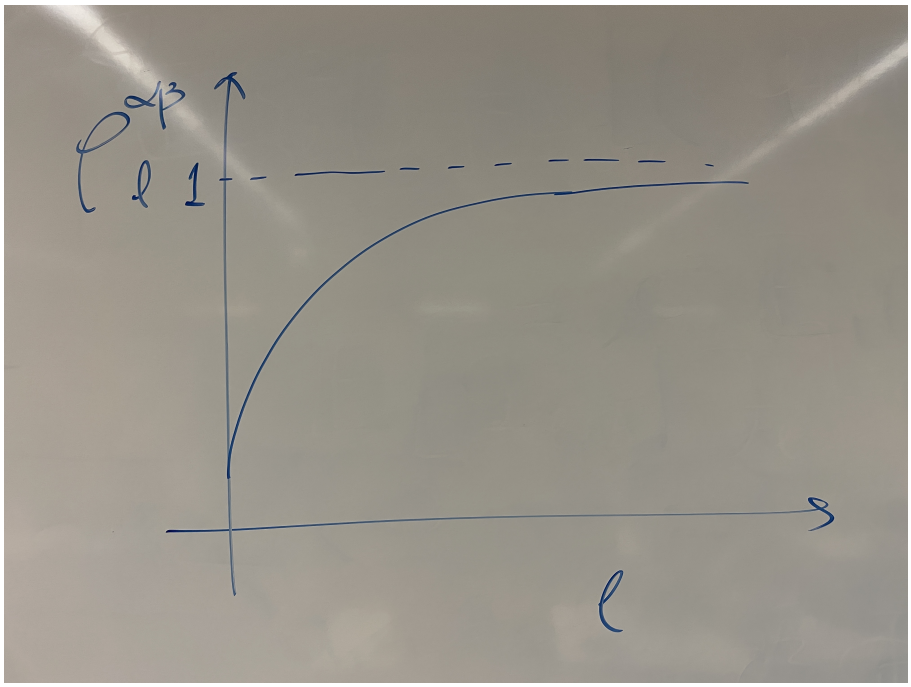
$$\begin{aligned}
\mathbb{E}[\Phi_{l+1} \mid \mathcal{F}_l] &= \mathbb{E}\left[\frac{c}{n}\sum_{j=1}^{n} \varphi(h_{l+1,j})^2 \mid \mathcal{F}_l\right] \\
&= \mathbb{E}\left[c\varphi(h_{l+1,j})^2 \mid \mathcal{F}_l\right] \\
&= 2\mathbb{E}\left[\varphi(w)^2\right], \quad w \sim \mathcal{N}(0, \Phi_l) \\
&= 2\frac{1}{2}\Phi_l \\
\implies \Phi_{l+1} &= \Phi_l + \frac{1}{\sqrt{n}} \underbrace{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(c\varphi(h_{l+1,j})^2 - \Phi_l\right)}_{\xi_l}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\xi_l^2 \mid \mathcal{F}_l\right] &= \mathbb{E}\left[\left(c\varphi(h_{l+1,j})^2 - \Phi_l\right)^2 \mid \mathcal{F}_l\right] \\
&= c^2\mathbb{E}\left[\underbrace{\varphi(h_{l+1,j})^4}_{h_{l+1,j}^4 \mathbb{1}_{\{h_{l+1,j}>0\}}} \mid \mathcal{F}_l\right] - \Phi_l^2 \\
&= \underbrace{\frac{c^2}{2}}_{=2} \mathbb{E}\left[\underbrace{h_{l+1,j}^4}_{3\Phi_l^2} \mid \mathcal{F}_l\right] - \Phi_l^2 \\
&= 5\Phi_l^2
\end{aligned}$$

$$\Phi_{l+1} = \Phi_l + \sqrt{5}\Phi_l \varepsilon_l, \quad \varepsilon_l \sim \mathcal{D}\left(\mathbb{E}\left[\varepsilon_l\right] = 0, \mathbb{E}\left[\varepsilon_l^2\right] = 1\right)$$
$$\to d\Phi_l = \sqrt{5}\Phi_l dB_\tau, \quad \frac{d}{n} \to \bar{\tau}$$
$$\implies \frac{\Phi_l}{\Phi_0} \sim \exp\left(\mathcal{N}\left(-\frac{5}{2}\bar{\tau}, 5\bar{\tau}\right)\right)$$

In the case of $m \geq 2$,

$$h_{l+1}^{\alpha} = \sqrt{\frac{c}{n}} W_l \varphi\left(h_l^{\alpha}\right)$$
$$\rho_l^{\alpha\beta} = \frac{\Phi_l^{\alpha\beta}}{\sqrt{\Phi_l^{\alpha\alpha}\Phi_l{}^{\beta\beta}}} \in [-1;\ 1]$$
$$\rho_{l+1}^{\alpha\beta} = f\left(\rho_l^{\alpha\beta}\right) + \dots$$
$$\rho_l^{\alpha\beta} \to 1 \text{ as } l \to \infty$$




- $h_l^{\alpha} \approx h_l^{\beta}$ constant
- Transformers with $\rho \approx 1$
- $\implies$ vanishing gradients (Noci et al., 2022)