STAT 946 - Topics in Probability and Statistics: Mathematical Foundations of Deep Learning
Lecture 18
Professor Mufan Li

Lucas Noritomi-Hartwig
University of Waterloo

November 19, 2025 from 16h00 to 17h20 in M3 3103

Shaped Transformer (Noci et al. 2023) - co-authored by Professor Mufan Li

The "usual" self-attention is defined as follows:

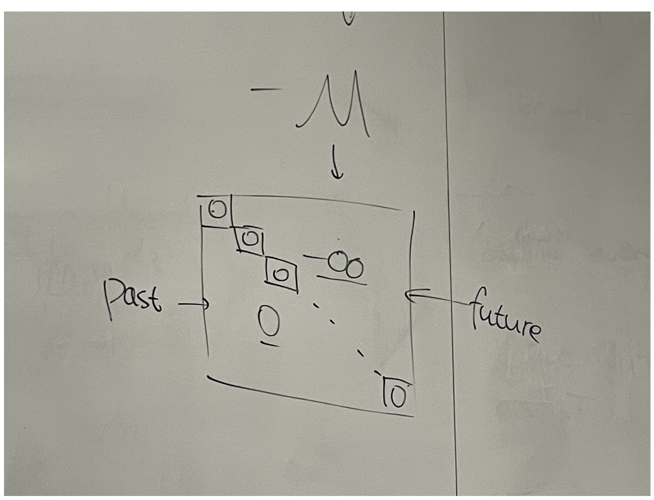
$$h_l = \left[h_l^1 \dots h_l^m\right] \in \mathbb{R}^{n\times m} \qquad \text{(no longer vertically stacking)}$$
$$h_{l+1} = \frac{1}{\sqrt{n}} W_l^V h_l A_l$$

where $W_l^V \in \mathbb{R}^{n\times n}$, $h_l \in \mathbb{R}^{n\times m}$, and $A_l \in \mathbb{R}^{m\times m}$, and A_l is given by:

$$A_l = \underbrace{\text{Softmax}}_{\text{column-wise}} \left(\frac{1}{\tau} \left(\frac{1}{\sqrt{n}} W_l^Q h_l \right)^\top \left(\frac{1}{\sqrt{n}} W_l^K h_l \right) - \mathcal{M} \right)$$

where $\tau \propto \sqrt{n_k}$, and W_l^Q, $W_l^K \in \mathbb{R}^{n_k \times n}$, think $n_k \sim n$. Each token "talks" to each other.

The matrix $\mathcal{M}$ is a mask where the upper-triangular section is the "future" of the token sequence. We set these values to $-\infty$ so that future tokens do not contribute towards prediction in softmax.



$$\text{Softmax}\left(y\right) = \frac{e^{y}}{\sum_{\alpha=1}^{m} e^{y^{\alpha}}} \text{(entrywise)}, \quad y \in \mathbb{R}^{m \times 1}$$

τ is referred to as tempurature (notion comes from statistical physics):

Gibbs distribution: $\propto e^{\frac{H}{\tau}}$ where H is a Hamiltonian.

For attention:

- as $\tau \to 0$, A_l concentrates (on the largest y^{α} entry)
- as $\tau \to \infty$, A_l is uniform $\frac{1}{m}\mathbb{1}$ where $\mathbb{1} \in \mathbb{R}^m$ is a vector of all "1"s.

$$\text{Softmax}\left(\frac{1}{\tau} y\right) = \frac{1}{m}\mathbb{1} + \frac{1}{\tau m}\left(y - \bar{y}\right) + \frac{1}{2\tau^2 m}\left(\left(y - \bar{y}\right)^2 - \left(\bar{y}^2 - \bar{y^2}\right)\right) + \mathcal{O}\left(\tau^{-3}\right)$$

where $\bar{y} = \frac{1}{m}\sum_{\alpha=1}^{m} y^{\alpha}$. This is a taylor expansion around $\frac{1}{\tau}$. After one layer of the first term, $\frac{1}{m}\mathbb{1}$, we end up with $\rho = 1$ or "rank collapse" which leads to gradients vanishing.

We choose center I_m:

$$\begin{aligned} \tau \to \infty &\implies A_l \to I_m \\ &\implies h_{l+1} = \frac{1}{\sqrt{n}} W_l h_l \quad \text{(Stable!)} \end{aligned}$$

Open question: Centered at linear attention?

$$\tau \to \infty \implies h_{l+1} = \frac{1}{\sqrt{n}} W_l^V h_l \frac{1}{\sqrt{n_k}}\left(\frac{1}{\sqrt{n}} W_l^Q h_l\right)^{\top}\left(\frac{1}{\sqrt{n}} W_l^K h_l\right)$$

Recall that

$$\varphi_S\left(x\right) = x + frac1\sqrt{n}\psi_1\left(x\right) + \frac{1}{n}\psi_2\left(x\right) + \mathcal{O}\left(n^{-3/2}\right)$$

The recipe is:

$$\begin{aligned} A_l &= I_m + \text{Softmax}\left(\ldots\right) - \frac{1}{m}\mathbb{1}\mathbb{1}^{\top} \\ h_{l+1} &= \frac{1}{\sqrt{n}} W_l^V h_l\left(I_m + \frac{1}{\tau}\left(\ldots\right) + \frac{1}{\tau^2}\left(\ldots\right) + \mathcal{O}\left(\tau^{-3}\right)\right) \end{aligned}$$

From this, we get a SDE limit for Φ. What this implies is that $\rho \neq 1$, thus we do not have vanishing gradients. Note that if we only have linear networks, we want a learning rate for $\frac{1}{\sqrt{n}} W_l^V h_l$ that is not MUP, however, we need a learning rate for $I_m + \frac{1}{\tau}\left(\ldots\right) + \frac{1}{\tau^2}\left(\ldots\right) + \mathcal{O}\left(\tau^{-3}\right)$ that is MUP.

The remaining results covered are not yet published.

Spectrum of Φ (In-progress, Li, de Dios Pont, Nica, Roy)

Consider a linear network:

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

where $W_{l,\,jk} \overset{\text{iid}}{\sim} \mathcal{N}(0,\,1)$, $n,\, d \to \infty$, $\frac{d}{n} \to \bar{\tau}$ Thus,

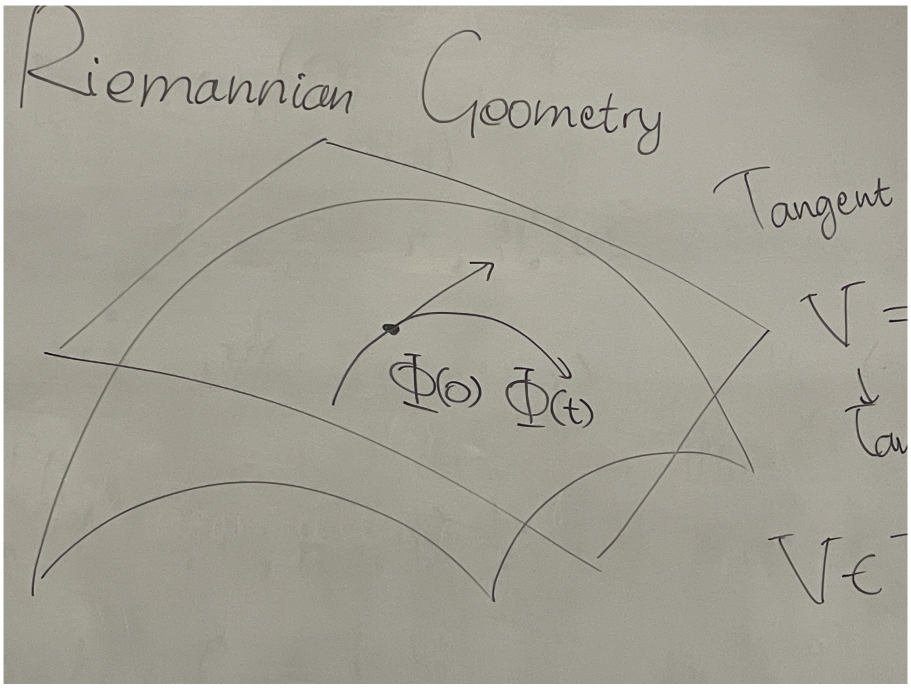
$$d\Phi_\tau = \Sigma(\Phi)^{1/2}\, dB_\tau$$

where $\Phi_\tau \in \mathbb{R}^{\bar{m}}$, $\bar{m} = \frac{1}{2}m(m+1)$ which is the number of upper triangular entries of covariance matrix, $\Sigma \in \mathbb{R}^{\bar{m}\times\bar{m}}$, and $B_\tau \in \mathbb{R}^{\bar{m}\times 1}$.

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

Detour into Riemannian Geometry:

Manifold: $\Phi \in \mathrm{SPD}(m) =: \mathcal{M}$. Tangent space:



$V = \frac{d}{dt}\Phi(t)\big|_{t=0}$ is a tangent vector. It turns out that $V \in T_\Phi\mathcal{M} = \mathrm{Sym}(m)$.

Object: Coordinate

- vec:$\mathcal{M} \to \mathbb{R}^{\bar{m}}$
- vec:$T_\Phi\mathcal{M} \to \mathbb{R}^{\bar{m}}$

Object: Riemannian Metric. For each Φ, there is a map g_Φ given by

$$g_\Phi : T_\Phi\mathcal{M} \times T_\Phi\mathcal{M} \to \mathbb{R}$$

and is an inner product.

In coordinates, we can interpret g_Φ as an $m \times m$ matrix: $\langle v,\, v\rangle_{g_\Phi} = u^\top g(\Phi)\, v$, where $u,\, v \in \mathbb{R}^{\bar{m}\times 1}$.

If we have a Brownian motion (in coordinates - not the same thing as Euclidean Brownian motion),

$$dX_\tau = \frac{1}{2}\text{gradient} \log\det\left(g_\Phi(X_\tau)\right) d\tau + g_\Phi(X_\tau)^{-1/2}\, dB_\tau$$

where $B_\tau \in \mathbb{R}^m$ is a Brownian motion, and X_τ is a Brownian motion in coordinates.

Question: Is $\Sigma(\Phi)^{-1}$ a Riemannian metric? Short answer: Yes.

Theorem 1. *If A, $B \in Sym(m)$ and*

$$\begin{aligned} vec(A)^\top &= \Sigma(\Phi)^{-1} vec(B) \\ &= \frac{1}{2} Tr\left(A\Phi^{-1}B\Phi^{-1}\right) \to \textit{the affine-invariant metric} \end{aligned}$$

- *Intuition: $\Sigma(\Phi)$ is a degree 2 polynomial in Φ,*
- *Intuition: The affine-invariant metrix is "degree -2" polynomial in Φ*
- *Verified symbolicly $m = 2$ (correct)*

Lemma. (Affine-invariance)
Let $P : \Phi_l \to \Phi_{l+1}$ (random Markov chain map). Then, we can equivalently define $P_\tau : \Phi_0 \to \Phi_\tau$ (stochastic flow). If $A \in \mathbb{R}^{m\times m}$ is full rank, then,

$$AP(\Phi)A^\top \overset{\mathrm{d}}{=} P\left(A\Phi A^\top\right)$$

and equivalently,

$$AP_\tau(\Phi)A^\top \overset{\mathrm{d}}{=} P_\tau\left(A\Phi A^\top\right)$$

Remarks:

- Symmetry $\implies$ geometry. The way to think about this is that the neural network randomness is symmetric, which causes the SDE to also be symmetric.
- $P_\tau(\Phi_0) = \Phi_0^{1/2} P_\tau(I_m)\Phi_0^{1/2}$.
- If $A \in \mathcal{O}(m)$, i.e., $AA^\top = I_m$, the $AP_\tau(I_m)A^\tau \overset{\mathrm{d}}{=} P_\tau(I_m)$. Thus, if $\Phi_0 = I_m$ then $A\Phi_\tau A^\top \overset{\mathrm{d}}{==} \Phi_\tau$, where $A \in \mathcal{O}(m)$. This is free diagonalization (GOE).

Theorem 2. *If $\lambda_j = \lambda_j(\Phi_\tau)$ where $\lambda_1 < \lambda_2 < \ldots < \lambda_m$, then*

$$d\lambda_j = \sqrt{2}\lambda_j dB_\tau^{(j)} + \underbrace{\sum_{k=1,\, k\neq j}^{m} \frac{\lambda_j\lambda_k}{\lambda_j - \lambda_k}}_{\Theta(m-1)} d\tau$$

If we replace the above with

$$d\lambda_j = \sqrt{2}\cdot 1\cdot dB_\tau^{(j)} + \sum_{k=1,\, k\neq j}^{m} \frac{1}{\lambda_j - \lambda_k} d\tau$$

then we get the Dyson Brownian motion. Thus, we have the Geometric Dyson Brownian motion.

Remark:

- The proof is easy with orthogonal invariance (we can diagonalize for free, and thus study the eigenvalues directly).
- However, it is still possible without using the invariance.
- This is a very tractable process due to the decoupled Brownian motions.

We want to take $m \to \infty$. Time change $\tau \to \frac{\tau}{m}$

$$\begin{aligned} &\implies d\lambda_j = \sqrt{\frac{2}{m}}\lambda_j dB_\tau^{(j)} + \frac{1}{m}\sum_{k\neq j}\frac{\lambda_j\lambda_k}{\lambda_j-\lambda_k} d\tau \\ &\overset{m\to\infty}{\longrightarrow} 0 + \int \frac{\lambda_j y}{\lambda_j - y}\rho_\tau(y) \end{aligned}$$

where

$$\rho_\tau(y) = \lim_{m\to\infty} \frac{1}{m} \sum_j \delta_{\lambda_j(\tau)}$$

We introduce the T-Transform:

$$\mathcal{G}_\tau(z) = \int \frac{x}{z-x} \rho_\tau(dx), \quad z \in \mathbb{C}$$

if we replace this with

$$\mathcal{G}_\tau(z) = \int \frac{1}{z-x} \rho_\tau(dx), \quad z \in \mathbb{C}$$

we get the Stieltjes transform.

Theorem 3. *As $m \to \infty$,*

$$\partial_\tau \mathcal{G}_\tau(z) = -z\mathcal{G}_\tau(z)\,\partial_z \mathcal{G}_\tau(z)$$

if we replace this with

$$\partial_\tau \mathcal{G}_\tau(z) = -1 \cdot \mathcal{G}_\tau(z)\,\partial_z \mathcal{G}_\tau(z)$$

we get the complex Burgers Equation. If $\Phi_0 = I_m$ $(\rho_0 = \delta_1)$, then

$$\mathcal{G}_\tau(z) = \frac{1}{ze^{\tau G_\tau(z)} - 1}$$

We can solve the above equation by fixed-point iterations.

<u>Remarks</u>:

- This equation is similar to the Lambert-W function, and so there is likely no closed-form solution.
- This is sometimes called the "free log-normal":
  - Semi-circle = free normal
  - Marchenko-Pastur = free Poisson
- if $\tau << 1$, $e^{-\tau\mathcal{G}} = 1 - \tau\mathcal{G} + \mathcal{O}(\tau^2)$ which is quadratic and thus solvable:

  $$\rho_\tau(x) = \frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{2\pi x^2 \tau} + \mathcal{O}(\tau^2)$$

  $\sigma^2 = 1+\tau$, $\lambda = \frac{\tau}{1+\tau}$, and $\lambda_\pm = \sigma^2\left(1 \pm \sqrt{\lambda}\right)^2$. if we replace this with

  $$\rho_\tau(x) = \frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{2\pi x \tau} + \mathcal{O}(\tau^2)$$

  we get MUP.

This is Professor Mufan Li's favourite result.