STAT 946 - Topics in Probability and Statistics: Mathematical Foundations of Deep Learning
Lecture 12
Professor Mufan Li

Lucas Noritomi-Hartwig
University of Waterloo

Ocotber 20, 2025 from 16h00 to 17h20 in M3 3103

Probability Seminar Mondays 11h30

SOLT November 10: 09h00 - 16h30 Lecture November 10: Starts 16h30

- October 27th: Gustavo Ruihan
- November 17th: Kevin Edward

Project presentations: November 24, Nov 26, Dec 1

Feature learning

- Non-kernel (data independent)
- Non-linear with respect to parameters

Recall: 1-layer network (NTK scaling)

$$f\left(x;\,\theta\right) = \frac{1}{\sqrt{n}} W_1 \varphi\left(W_0 x\right), \qquad W_1 \in \mathbb{R}^{1\times n},\, W_0 \in \mathbb{R}^{n\times n_0},\, x \in \mathbb{R}^{n_0\times 1}$$

Data: $\left(\left(x^\alpha,\, y^\alpha\right)\right)_{\alpha=1}^m$

MSE:

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

Training: $\theta_{k+1} = \theta_k - \eta\nabla\mathcal{L}\left(\theta_k\right)$, $\eta > 0$ is constant.

Hidden: $h^\alpha\left(\theta\right) = W_0 x^\alpha$.

$$\begin{aligned}\Delta h^\alpha &= h^\alpha\left(\theta_1\right) - h^\alpha\left(\theta_0\right)\\ &= -\frac{\eta}{m\sqrt{n}}\sum_{\beta=1}^m \left(f\left(x^\beta;\,\theta_0\right) - y^\beta\right)\left\langle x^\alpha,\, x^\beta\right\rangle W_1^\top \odot \varphi'\left(h^\beta\right)\end{aligned}$$

As $n \to \infty$,

$$\Delta h^\alpha = \Theta\left(\frac{1}{\sqrt{n}}\right) \qquad \text{v.s.} \qquad h^\alpha\left(\theta_0\right) = \Theta\left(1\right)$$
$$\Delta f^\alpha \approx -\frac{\eta}{m} K^\alpha \left(f - y\right) = \Theta\left(1\right)$$

NTK:

$$\begin{aligned} K^{\alpha\beta} &= \left\langle \nabla_\theta f^\alpha,\, \nabla_\theta f^\beta \right\rangle \\ &= \left\langle \nabla_{W_1} f^\alpha,\, \nabla_{W_1} f^\beta \right\rangle + \left\langle \nabla_{W_0} f^\alpha,\, \nabla_{W_0} f^\beta \right\rangle \\ &= \frac{1}{n}\left\langle \varphi\left(h^\alpha\right),\, \varphi\left(\underbrace{h^\beta}_{\Phi_1^{\alpha\beta}}\right)\right\rangle + \underbrace{\left\langle x^\alpha,\, x^\beta \right\rangle}_{\Phi_0^{\alpha\beta}} \underbrace{\frac{1}{n}\sum_{j=1}^{n} \varphi'\left(h_j^\alpha\right)\varphi'\left(h_j^\beta\right)}_{\text{at init. } G_1^{\alpha\beta}} \end{aligned}$$

Need $\Delta h^\alpha = \Theta\left(1\right)$ for NTK to evolve $\to$ feature learning.

Naive method: Increase learning rate $\eta = \eta_0 \sqrt{n}$, η_0 is constant. Non-NTK.

$$\begin{aligned} \implies \Delta h^\alpha &= \Theta\left(1\right) \quad \text{works!} \\ \Delta f^\alpha &= \Theta\left(\sqrt{n}\right) \quad \text{diverges - does not work!} \end{aligned}$$

Compensate: Scale down output.

$$\begin{aligned} f\left(x;\, \theta\right) &= \frac{1}{n} W_1 \varphi\left(W_0 x\right) \\ \implies \Delta h^\alpha &= \Theta\left(\frac{1}{\sqrt{n}}\right) \\ \Delta f^\alpha &= \Theta\left(\frac{1}{\sqrt{n}}\right) \longrightarrow 2 \times \text{effect} \\ K^{\alpha\beta} &= \left\langle \underbrace{\nabla_\theta f^\alpha}_{\frac{1}{\sqrt{n}}},\, \underbrace{\nabla_\theta f^\beta}_{\frac{1}{\sqrt{n}}} \right\rangle \end{aligned}$$

Mean Field Parameterization

Prefactor $\frac{1}{n}$, learning rate: $\eta = \eta_0 n$

$$\begin{aligned} \implies \Delta h^\alpha &= \Theta\left(1\right) \quad \text{works!} \quad \text{accelerated} \\ \Delta f^\alpha &= \Theta\left(1\right) \quad \text{works!} \end{aligned}$$

Effectively three choices:

- NTK:
  - Prefactor: $\dfrac{1}{\sqrt{n}}$ (or 1)
  - Learning rate: 1 (or $\dfrac{1}{\sqrt{n}}$)
  - Initial standard deviation: 1 (or $\dfrac{1}{\sqrt{n}}$)

- Mean field:
  - Prefactor: $\frac{1}{n}$
  - Learning rate: n
  - Initial standard deviation: 1

This kind of equivalence is called ABC-reparameterization (Yang and Hu 2020).

Recall in the NTK: $\Delta f \simeq -\frac{\eta}{m} K\left(f - y\right)$

This is the same in the mean field, however, K is now evolving.

What is the state space of neural network training?

- Full state space: $\theta_{k+1} = \theta_k - \eta \nabla \mathcal{L}\left(\theta_k\right)$
- NTK: f
- Neurons: h, g (in between full and NTK)

We will focus on the full state space.

$\theta = \left(W_1,\, W_0\right) \in \mathbb{R}^{1\times n} \times \mathbb{R}^{n \times n_0}$. As $n \to \infty$, θ becomes infinitely dimensional.

The way to write this is:

$$\begin{aligned} f\left(x;\, \theta\right) &= \frac{1}{n} \sum_{j=1}^{n} W_{1,\,j} \varphi\left(\left\langle W_{0,\,j},\, x\right\rangle\right) \\ &= \int w_1 \varphi\left(\left\langle w_0,\, x\right\rangle\right) d\rho^{(n)}\left(w_1,\, w_0\right) \end{aligned}$$

where

$$\rho^{(n)} = \frac{1}{n} \sum_{j=1}^{n} \delta_{\left(W_{1,\,j},\, W_{0,\,j}\right)}$$

is the empirical measure.

Recall that $\int q\left(w\right) d\delta_{w_0}\left(w_0\right) = q\left(w_0\right)$.

Back to gradient flow:

$$\partial_t \theta\left(t\right) = -\eta \nabla \mathcal{L}\left(\theta\left(t\right)\right)$$

where $\eta = \eta_0 n$.

$$\partial_t W_{1,\,j} = -\frac{\eta_0}{m} \sum_{\beta=1}^{m} \left(f_{(t)}^{\beta} - y^{\beta}\right) \varphi\left(\left\langle W_{0,\,j}^{(t)},\, x^{\beta}\right\rangle\right)$$

where

$$\begin{aligned} f_{(t)}^{\beta} &= \int w_1 \varphi\left(\left\langle w_0,\, x^{\beta}\right\rangle\right) d\rho_t^{(n)} \\ \partial_t W_{0,\,j} &= -\frac{\eta_0}{m} \sum_{\beta=1}^{m} \left(f_{(t)}^{\beta} - y^{\beta}\right) \varphi\left(\left\langle W_{0,\,j}^{(t)},\, x^{\beta}\right\rangle\right) x^{\beta} \end{aligned}$$

$$\theta_j = (W_{1,j}, W_{0,j})$$
$$\partial_t \theta_j(t) = b\left(\theta_j(t), \rho_t^{(n)}\right)$$

Recall: propagation of chaos.

As $n \to \infty$, $\rho_t^{(n)} \to \rho_t \leftarrow \underbrace{\mathcal{L}(\theta_j(t))}_{n \text{ finite}}$ where $\rho_t^{(n)}$ is all particles, and $\theta_j(t)$ is a single particle.

ρ_t is called the mean field measure/distribution, $\rho_t^{(n)}$ is called the mean field ODE (McKean-Vlasov).

Stronger notion of propagation of chaos:

$$\mathcal{L}((\theta_{j_1}, \ldots, \theta_{j_k})) \to \rho_t^{\otimes k}$$

k-particles are independent!

Bound $W_2\left(\mathcal{L}((\theta_{j_1}, \ldots, \theta_{j_k})), \rho_t^{\otimes k}\right) \sim\leq \frac{1}{n}e^{c(t, n_0)}$.

Mean Field PDE

- Cannot differentiate $\rho_t^{(n)}$ or δ_{θ_j}.
- Trick: use test function $q \in C_c^\infty\left(\mathbb{R}^{n_0+1}\right)$.

$$\int q(\theta) \underbrace{\text{``}\nabla_\theta \rho_t^{(n)}(\theta)\text{''}}_{\text{``weak derivative''}} d\theta \overset{\text{IBP}}{=} -\int \nabla_\theta q(\theta)\, d\rho_t^{(n)}(\theta)$$

$$\begin{aligned}
\partial_t \int q(\theta)\, d\rho_t^{(n)}(\theta) &= \partial_t \frac{1}{n}\sum_{j=1}^n q(\theta_j(t)) \\
&= \frac{1}{n}\sum_{j=1}^n \langle \nabla_\theta q(\theta_j(t)), \partial_t \theta_j(t)\rangle \qquad b\left(\theta_j(t), \rho_t^{(n)}\right) \\
&= \int \nabla_\theta q(\theta)\, b\left(\theta, \rho_t^{(n)}\right) d\rho_t^{(n)} \\
\text{``}\int \underbrace{q(\theta)}_{\text{not specified}} \partial_t \rho_t^{(n)} d\theta\text{''} &\overset{\text{IBP}}{=} \int q(\theta)\left(-\text{div}_\theta\left(b\left(\theta, \rho_t^{(n)}\right)\rho_t^{(n)}\right)\right) d\theta\text{''}
\end{aligned}$$

We say that $\rho_t^{(n)}$ is a weak solution of

$$\begin{cases} \partial_t \rho_t = -\text{div}\left(b(\theta, \rho_t)\rho_t\right) & \text{first order non-linear ``Transport equation''} \\ \rho_0 = \rho_0^{(n)} \end{cases}$$

Mean field PDE.

References:

- Mei, Montanari, Nguyen (2018)
- Chizat, Bach (2018)
- Nitanda, Suzuki (2017) Japan
- Rotskoff, Vanden-Eijnden (2018) New York
- Sirigano, Spiliopoulos (2018)