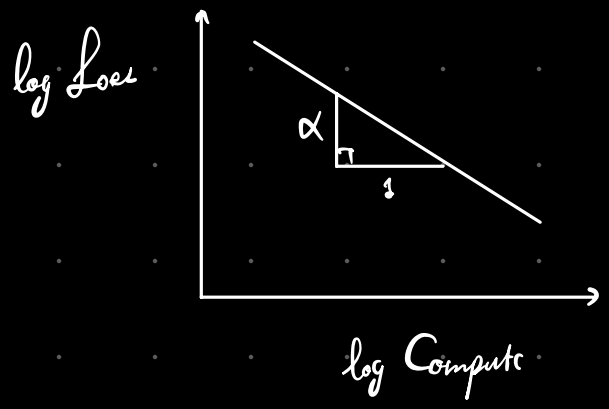
Proportional Scaling Limit

- Scaling law



$$\Rightarrow \text{Loss} = \text{const.} \times \text{Compute}^{-\alpha}$$

- What if we interpret

  Compute $\longrightarrow$ index $n \in \mathbb{N}$

  NN $\longrightarrow$ element $x_n \in \mathcal{X}$ in "space of NNs"

  Loss $\longrightarrow$ test function $L : \mathcal{X} \to \mathbb{R}$

  (morally speaking)

  Then does $x_n \to x^*$ w/ rate $n^{-\alpha}$

  imply $L(x_n) \to L(x^*)$ w/ rate $n^{-\alpha}$?

Example · Let $\alpha > 0$. $X_i \overset{iid}{\sim} \rho$: $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = 1$,

and $\mathbb{E}|X|^{2(1+\alpha)} < \infty$ = the highest moment that exists.

$$\Rightarrow \quad S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(0,1)$$

at rate $\boxed{n^{-\alpha}}$

Remark · Maybe we should study the scaling limit as Compute $\to \infty$.

- Roughly speaking

$$\text{Compute} \propto \underbrace{\text{Width}^2}_{\#\text{ param. per layer}} \times \underbrace{\text{Depth}}_{\#\text{ layers}} \times \underbrace{\text{Batch}}_{\#\text{ data per iter.}} \times \underbrace{\text{Train Iter.}}_{\#\text{ iter.}}$$

$\Rightarrow$ want to study a joint limit (need to specify relationship) of width, depth ("easier"), data, # iter. (hard)

- In this talk, we will study the limit as (w/o skip connections)

$$\begin{cases} \text{width } n, \text{ depth } d \to \infty. \\ \text{ratio } \frac{d}{n} \to \tilde{c} > 0. \end{cases} \longrightarrow \text{"proportional"}$$

Warm Up

- Linear network. one sample $x \in \mathbb{R}^{n_0}$ → input dim.

$$h_1 = \frac{1}{\sqrt{n_0}} W_0 x \qquad W_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$$

$$h_{\ell+1} = \frac{1}{\sqrt{n}} W_\ell h_\ell \qquad \ell \in [d]$$

$$f = \frac{1}{\sqrt{n}} W_d h_d$$

(W_0: $n \times n_0$, h_1: $n \times 1$; W_ℓ: $n \times n$, h_ℓ: $n \times 1$; W_d: $1 \times n$, h_d: $n \times 1$)

Goal · Analyze $\mathcal{L}(h_\ell)$, $\mathcal{L}(f)$ as $d, n \to \infty$.

Trick · Condition on the "right" object

· In this case $\mathcal{F}_\ell := \sigma(\{h_k\}_{k \le \ell})$ → Filtration

"deterministic" ↑ ↳ all the prev. layers

$$h_{\ell+1} \mid \mathcal{F}_\ell = \frac{1}{\sqrt{n}} W_\ell h_\ell \mid \mathcal{F}_\ell$$

W_ℓ: iid $\sim \mathcal{N}(0,1)$

$$\sim \mathcal{N}\left(0, \underbrace{\tfrac{1}{n}\|h_\ell\|^2}_{=:\Phi_\ell} \cdot I_n\right)$$

$\Phi_\ell \longrightarrow$ key quantity

$$\Phi_{\ell+1} = \frac{1}{n}\|h_{\ell+1}\|^2$$

$$= \frac{1}{n}\left\| \frac{1}{\sqrt{n}} W_\ell h_\ell \right\|^2$$

$$= \frac{1}{n}\left\| W_\ell \underbrace{\frac{h_\ell}{\|h_\ell\|}}_{\text{unit vector}} \right\|^2 \underbrace{\frac{1}{n}\|h_\ell\|^2}_{\Phi_\ell}$$

$\frac{h_\ell}{\|h_\ell\|} \sim \mathcal{N}(0, I_n)$

$$= \frac{1}{n}\chi_n^2\, \Phi_\ell \longrightarrow \text{note } \Phi_{\ell+1} \mid \mathcal{F}_\ell = \Phi_{\ell+1} \mid \sigma(\Phi_\ell)$$

$\Rightarrow \Phi_{\ell+1}$ is a Markov chain!

$$= \Phi_\ell + \Phi_\ell\left(\frac{1}{n}\chi_n^2 - 1\right)$$

$\mathbb{E} = 0$, $\mathrm{Var} = 2/n$

$$\Phi_{\ell+1} =: \Phi_\ell + \sqrt{\tfrac{2}{n}}\,\Phi_\ell\,\xi_\ell$$

$\hookrightarrow \mathbb{E}=0$, $\mathrm{Var}=1$

Euler discretization of $d\Phi_\tau = \sqrt{2}\,\Phi_\tau\, dB_\tau$ w/ step size $\frac{1}{n}$.

(Geometric Brownian motion) $\sim \exp(\mathcal{N}(-\tau, 2\tau))$.

- $\ell \mapsto \ell+1$ one step $\Rightarrow$ total d steps (step size $\frac{1}{n}$)

$\Rightarrow$ total depth time $\bar{\tau} := \frac{d}{n}$. current time $\tau = \frac{\ell}{n}$.

hence depth/width should be proportional.

- Furthermore observe that

$$f_\ell \mid \mathcal{F}_\ell \sim \mathcal{N}(0, \Phi_\ell) \Rightarrow \text{Sample } \Phi_\ell \to \Phi_{\ell+1} \to \text{terminal time}$$
$$h_{\ell+1} \mid \mathcal{F}_\ell \sim \mathcal{N}(0, \Phi_\ell) \qquad \Phi_\ell \to \Phi_\infty$$

Remark · Can write down a limiting Fokker-Planck equation for $\mathcal{L}(h_{\ell,i})$ or $\frac{1}{n}\sum_{i=1}^{n} \delta_{h_{\ell,i}}$

Multiple Data Points

- Let $X = [x^1 \cdots x^m] \in \mathbb{R}^{n \times m}$ → # data

$$h_{\ell+1}^\alpha = \frac{1}{\sqrt{n}} W_\ell h_\ell^\alpha \qquad \alpha \in [m]$$
$$h_\ell = [h_\ell^1 \cdots h_\ell^m]$$

- Let $u, v \in \mathbb{R}^n$ be deterministic, then

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

where $A \otimes B = [a_{ij} B]$

$$\Rightarrow \operatorname{vec}(h_{\ell+1}) \mid \mathcal{F}_\ell \overset{d}{=} \mathcal{N}\left(0, \left[\tfrac{1}{n}\langle h_\ell^\alpha, h_\ell^\beta \rangle\right]_{\alpha,\beta=1}^{m} \otimes I_n\right) \quad (*)$$

$\operatorname{vec}: \mathbb{R}^{n \times m} \to \mathbb{R}^{nm}$

$\frac{1}{n} h_\ell^T h_\ell =: \Phi_\ell \in \mathbb{R}^{m \times m}$

Feature covariance matrix.

$$\Rightarrow \Phi_{\ell+1}^{\alpha\beta} \mid \mathcal{F}_\ell = \frac{1}{n} \langle h_{\ell+1}^\alpha, h_{\ell+1}^\beta \rangle \mid \mathcal{F}_\ell$$

note $\mathbb{E}\left[\frac{1}{n}\langle h^\alpha_{\ell+1}, h^\beta_{\ell+1}\rangle \mid \mathcal{F}_\ell\right] = \mathbb{E}\left[h^\alpha_{\ell+1,i}\, h^\beta_{\ell+1,i} \mid \mathcal{F}_\ell\right]$

$= \Phi_\ell^{\alpha\beta}$ ... by (4)

Then $\tilde{\Phi}^{\alpha\beta}_{\ell+1} = \Phi_\ell + \frac{1}{\sqrt{n}} \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(h^\alpha_{\ell+1,i}\, h^\beta_{\ell+1,i} - \Phi_\ell^{\alpha\beta}\right)}$

$\mathbb{E}[\cdot \mid \mathcal{F}_\ell] = 0.$

$\mathrm{Var}[\cdot \mid \mathcal{F}_\ell] = \mathrm{Var}\left(h^\alpha_{\ell+1,i}\, h^\beta_{\ell+1,i} \mid \mathcal{F}_\ell\right)$

$= \Theta(1)$

$\tilde{\Phi}^{\gamma\delta}_{\ell+1} = \cdots\cdots\cdots \left(h^\gamma_{\ell+1,i}\, h^\delta_{\ell+1,i} - \Phi_\ell^{\gamma\delta}\right)$

$\Rightarrow \mathrm{Cov}\left(h^\alpha h^\beta, h^\gamma h^\delta\right)$ where $h \sim \mathcal{N}(0, \Phi_\ell)$

$= \mathbb{E}\, h^\alpha h^\beta h^\gamma h^\delta - \mathbb{E}\, h^\alpha h^\beta\, \mathbb{E}\, h^\gamma h^\delta$

$\mathbb{E}\, h^\alpha h^\beta \to \Phi^{\alpha\beta}$, $\mathbb{E}\, h^\gamma h^\delta \to \Phi^{\gamma\delta}$

Isserlis / Wick

$\mathbb{E}\, h^\alpha h^\beta\, \mathbb{E}\, h^\gamma h^\delta + \Phi^{\alpha\gamma}\Phi^{\beta\delta} + \Phi^{\alpha\delta}\Phi^{\beta\gamma}$

($\mathbb{E}\, h^\alpha h^\beta = \Phi^{\alpha\beta}$, $\mathbb{E}\, h^\gamma h^\delta = \Phi^{\gamma\delta}$)

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

$m(m+1)/2 \times m(m+1)/2$

$\Rightarrow \tilde{\Phi}_{\ell+1} \mid \mathcal{F}_\ell \overset{d}{=} \Phi_\ell + \frac{1}{\sqrt{n}} \Sigma(\Phi_\ell)^{1/2} \xi_\ell$

$\tilde{\Phi}_{\ell+1}$: flatten to $\mathbb{R}^{m(m+1)/2}$

ξ_ℓ: $\mathbb{E} = 0$, $\mathrm{Cov} = I_{m(m+1)/2}$

$\Rightarrow \{\Phi_{\lfloor tn \rfloor}\} \overset{d,n\to\infty}{\xrightarrow{\frac{d}{n} \to \tau}} \begin{cases} d\Phi_\tau = \Sigma(\Phi_\tau)^{1/2}\, dB_\tau, & \tau \in [0, \bar{\tau}] \\ \Phi_0 = \frac{1}{n_0} x^\tau x. \end{cases}$

B_τ: B.M. in $\mathbb{R}^{m(m+1)/2}$

Remark · There is a matrix form

Let $S_\tau := \frac{1}{\sqrt{n}}(W_\tau + W_\tau^T)$ ↳ Brownian motion on $\mathbb{R}^{n\times n}$

B.M on $Sym(n)$

$\Rightarrow d\Phi_\tau = \Phi_\tau^{1/2}\, dS_\tau\, \Phi_\tau^{1/2}$

$\mathbb{R}^{m\times m}$

(Optional)

· $\Sigma(\Phi)$ is the affine-invariant cometric

i.e. let $\mathcal{M} = SPD(m)$, $T_p\mathcal{M} = Sym(m)$

and Riemannian metric

$$g_\Phi(A,B) = \frac{1}{2}Tr(A\Phi^{-1}B\Phi^{-1})$$

$Sym(m)$

$\Rightarrow \Phi_\tau$ is a Brownian motion on $(\mathcal{M}, g)$

(with a specific dual connection ∇^*).

Nonlinear Network

· $h_{\ell+1}^\alpha = \sqrt{\frac{c}{n}}\, W_\ell\, \varphi(h_\ell^\alpha)$, $\varphi: \mathbb{R}\to\mathbb{R}$

$c^{-1} := \mathbb{E}\,\varphi(w)^2$, $w\sim N(0,1)$

(Optional) · Why c? Let $\varphi(x) = \max(x,0)$ (ReLU), $c=1$

then $h_{\ell+1}^\alpha \mid \mathcal{F}_\ell \sim N\left(0, \frac{1}{n}\|\varphi(h_\ell^\alpha)\|^2 \otimes I_n\right)$

$$\mathbb{E}\left[\frac{1}{n}\|\varphi(h_{\ell+1}^\alpha)\|^2 \mid \mathcal{F}_\ell\right] = \mathbb{E}\left[\varphi(h_{\ell+1}^\alpha)^2 \mid \mathcal{F}_\ell\right]$$

$\Phi_{\ell+1}^{\alpha\alpha}$

$$= \frac{1}{2}\mathbb{E}\left[(h_{\ell+1}^\alpha)^2 \mid \mathcal{F}_\ell\right]$$

$$= \frac{1}{2}\,\Phi_\ell^{\alpha\alpha}$$

$$\Rightarrow \mathbb{E}[\Phi_{\ell+1}^{\alpha\alpha} \mid \mathcal{F}_\ell] = \frac{1}{2}\,\Phi_\ell^{\alpha\alpha}$$

$\searrow 0$ as $\ell \to \infty$.

- Let $\varphi_\ell^\alpha := \varphi(h_\ell^\alpha)$

then $\mathrm{vec}(h_{\ell+1}) \mid \mathcal{F}_\ell \sim \mathcal{N}\Big(0, \underbrace{\Big[\tfrac{c}{n}\langle \varphi_\ell^\alpha, \varphi_\ell^\beta \rangle\Big]_{\alpha\beta}}_{=: \tilde{\Phi}_\ell} \otimes I_n\Big)$.

Proposition · Let $\rho_\ell^{\alpha\beta} = \dfrac{\Phi_\ell^{\alpha\beta}}{(\Phi_\ell^{\alpha\alpha}\,\Phi_\ell^{\beta\beta})^{1/2}}$,

then $\mathbb{E}[\rho_{\ell+1}^{\alpha\beta} \mid \mathcal{F}_\ell] = \frac{1}{\pi}\Big[\sqrt{1-(\rho_\ell^{\alpha\beta})^2} + \rho_\ell^{\alpha\beta}(\pi - \arccos(\rho_\ell^{\alpha\beta}))\Big]$

$=: T(\rho_\ell^{\alpha\beta})$ — Cho and Saul kernel

and T has a stable fixed point at 1.

- In other words, $\rho_\ell^{\alpha\beta} \nearrow 1$ as $\ell \to \infty$.

Remark · This is sometimes known as "rank collapse" (since $\mathrm{rank}(h_\ell) \to 1$) and is known to cause vanishing gradients.

Solution · Shape the activation function $\varphi(x) \to x$ as $d \to \infty$.

Intuition: weaken a fixed point iteration $T \circ T \circ T \cdots$ to an Euler discretization $T(\rho) \approx \rho + \cdots$

e.g. for ReLU, let $\varphi_s(x) = s_+ \max(x,0) + s_- \min(x,0)$

where $s_\pm = 1 + \frac{c_\pm}{n^\rho}$ for some $\rho > 0$ TBD.

$$\Rightarrow \Phi_{\ell+1} \mid \mathcal{F}_\ell = \Phi_\ell + \frac{1}{n^\rho}\Big(\overbrace{\tilde{\Phi}_\ell^{\alpha\alpha}\tilde{\Phi}_\ell^{\beta\beta}\, \nu(\rho_\ell^{\alpha\beta})}^{\to b(\Phi_\ell)}\Big)_{\alpha\beta} + \frac{1}{\sqrt{n}} \sum (\Phi_\ell)^{1/2} \xi_\ell + \text{higher order terms}$$

where $\nu(\rho) := \frac{(c_+ - c_-)^2}{2\pi}\left(\sqrt{1-\rho^2} - \rho \arccos(\rho)\right)$

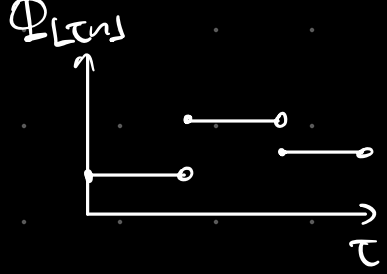
$\Rightarrow \rho = \frac{1}{2}$ matches the step size $\frac{1}{n}$

Theorem (LLNR22)

- As $d, n \to \infty$, $\frac{d}{n} \to \bar{\tau}$, $\rho = \frac{1}{2}$
- $\{\bar{\Phi}_{\lfloor tn \rfloor}\}_{t \in [\bar{\tau}]}$ converges weakly in the Skorohod topology of $\mathbb{R}^{n(n+1)/2}$ to the solution of

$$\begin{cases} d\Phi_t = b(\Phi_t)\,dt + \sum(\Phi_t)^{1/2}\, dB_t, \\ \Phi_0 = \frac{1}{n_0} x^T x. \end{cases}$$

Remark



- Skorohod handles the discontinuities in the process.

- Can extend to other architectures e.g. Shaped Transformers. [NLC+23]

$$h_{\ell+1} = \frac{1}{\sqrt{n}} W_\ell^V h_\ell A_\ell \qquad \text{(next} \times \text{next)}$$

$$A_\ell = \mathrm{Softmax}\left(\frac{1}{s} (W_\ell^Q h_\ell)^\top (W_\ell^K h_\ell) \right) - \frac{1}{n} 1 1^\top + I_n$$

↳ $s \sim n n_k$

$$\Rightarrow h_{\ell+1} \sim \text{linear} + \mathcal{O}\left(\tfrac{1}{n}\right)$$

- Can extend to training (my research talk, arXiv ~ 1–2 months)

$W \in \mathbb{R}^{n \times n}$ iid $\mathcal{N}(0,1)$, $\Psi, g \in \mathbb{R}^{n \times n}$ det.

$$\Rightarrow W \mid \sigma(W\Psi, g^\top W)$$

$$\overset{d}{=} W P_\Psi + P_g W - P_g W P_\Psi + P_g^\perp \tilde{W} P_\Psi^\perp$$

P_Ψ: proj. onto $\mathrm{col}(\Psi)$; $P_g^\perp = I - P_g$; $\tilde{W}$: indep. copy of W.

- Isolates the dependence structure in a clean way.

(arXiv this week)

- Can calculate the spectrum of Φ in the linear case.

specifically let $\rho := \frac{1}{n} \sum_i \delta_{\lambda_i(\Phi_d)}$,

then as $d, n, m \to \infty$, $\frac{d}{n} \to \bar{c}$

$$\rho \to \exp_\# \left(M_{\bar{c}}^{sc} \boxplus \mathrm{Unif}_{[-\bar{c},0]} \right) \boxtimes \rho_0$$

$\exp_\#$: push forward; $\boxplus$: additive free conv.; $\boxtimes$: multiplicative free conv.

$M_{\bar{c}}^{sc} \boxplus \mathrm{Unif}_{[-\bar{c},0]}$: Semicircle law

$\exp_\# (\cdots) \boxtimes \rho_0$: free log-normal