STAT 946 Scribed Note for Lecture 16
Professor Mufan Li

Scribed by Ruihan Lin

Nov. 10

Set-up:

$$h_{l+1}^{\alpha} = \sqrt{\frac{c}{n}} W_l \varphi(h_l^{\alpha})$$
$$h_l^{\alpha} = \frac{1}{\sqrt{n}} x^{\alpha};\ W_{l,ij} \overset{iid}{\sim} N(0,1);\ \varphi(x) = max(x,0);\ c^{-1} = E\varphi^2(\omega),\ \omega \sim N(0,1)$$

$$\left[h_{l+1}^{\alpha}\right]_{\alpha=1}^{m} \Big|\, \mathcal{F}_l \sim N\left(0,\ \left[\frac{c}{n}\left\langle \varphi_l^{\alpha},\, \varphi_l^{\beta}\right\rangle\right]_{\alpha,\beta=1}^{m} \otimes I_n\right)$$

where $\mathcal{F}_l$ is the sigma field generated from the first to the l-th layer, and let $\Phi_l := \left[\frac{c}{n}\left\langle \varphi_l^{\alpha},\, \varphi_l^{\beta}\right\rangle\right]_{\alpha,\beta=1}^{m} \otimes I_n$.

$$\begin{aligned}
\mathbb{E}\Big[\Phi_{\ell+1}^{\alpha\beta} \,\Big|\, \mathcal{F}_\ell\Big] &= \mathbb{E}\left[\frac{c}{n}\sum_{i=1}^{n} \varphi(h_{\ell,i}^{\alpha})\,\varphi(h_{\ell,i}^{\beta}) \,\middle|\, \mathcal{F}_\ell\right] \\
&= \mathbb{E}\Big[c\,\varphi(h_{\ell,i}^{\alpha})\,\varphi(h_{\ell,i}^{\beta}) \,\Big|\, \mathcal{F}_\ell\Big] \qquad \text{(Joint Gaussian)} \\
&= \left(\Phi_\ell^{\alpha\alpha}\,\Phi_\ell^{\beta\beta}\right)^{1/2} \times \mathbb{E}\left[c\,\varphi\left(\frac{h_{\ell,i}^{\alpha}}{(\Phi_\ell^{\alpha\alpha})^{1/2}}\right)\,\varphi\left(\frac{h_{\ell,i}^{\beta}}{\left(\Phi_\ell^{\beta\beta}\right)^{1/2}}\right) \,\middle|\, \mathcal{F}_\ell\right].
\end{aligned}$$

As

$$c > 0, \qquad \varphi(cx) = c\,\varphi(x).$$

Let $\frac{h_{\ell,i}^{\alpha}}{\left(\Phi_\ell^{\alpha\alpha}\right)^{1/2}}$, $\frac{h_{\ell,i}^{\beta}}{\left(\Phi_\ell^{\beta\beta}\right)^{1/2}}$ be w^{α}, w^{β},

$$\begin{pmatrix} w^{\alpha} \\ w^{\beta} \end{pmatrix} \sim \mathcal{N}\left(0,\ \begin{bmatrix} 1 & \rho_\ell^{\alpha\beta} \\ \rho_\ell^{\beta\alpha} & 1 \end{bmatrix}\right), \qquad \text{where}$$

$$\rho_\ell^{\alpha\beta} = \frac{\Phi_\ell^{\alpha\beta}}{\left(\Phi_\ell^{\alpha\alpha}\,\Phi_\ell^{\beta\beta}\right)^{1/2}}.$$

By Cho and Saul (2009),

$$\mathbb{E}\big[\varphi(w^\alpha)\,\varphi(w^\beta)\big] = \frac{\sqrt{1-\rho^2} \;+\; \rho\big(\pi - \arccos(-\rho)\big)}{2\pi}$$
$$=:\; \bar{J}_1(\rho_l^{\alpha\beta})$$

$$\Rightarrow \quad \mathbb{E}\Big[\Phi_{\ell+1}^{\alpha\beta}\,\Big|\,\mathcal{F}_\ell\Big] = c\,\bar{J}_1\Big(\rho_\ell^{\alpha\beta}\Big)\,\Big(\Phi_\ell^{\alpha\alpha}\Phi_\ell^{\beta\beta}\Big)^{1/2}$$

Aside $n \to \infty$,

$$\Phi_{\ell+1}^{\alpha\beta} = c\,\bar{J}_1\Big(\rho_\ell^{\alpha\beta}\Big) = \;\bar{f}\circ\bar{f}\circ\cdots\circ\bar{f}\,\big(c\,\rho_0^{\alpha\beta}\big)$$

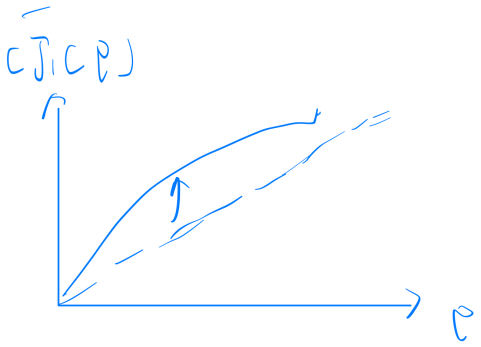

Figure 1: $\bar{J}_1(\rho) > \rho,\ \rho \in [0,1]$, $\rho = 1$ is a fixed point $p_{\ell+1} = c\,\bar{J}_1(p_\ell),\ \rho_0 > 0$

$\rho_l \to 1$ as $l \to \infty$.

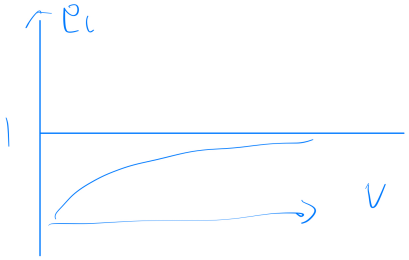

Figure 2: Remark: Noci et al.(2022), $\rho = 1$ implies vanishing gradient

Aside: $\rho_\ell^{\alpha\beta} = 1 - \Theta\big(\ell^{-2}\big)\,,$

$q_\ell^{\alpha\beta} = \ell^2(1-\rho_\ell^{\alpha\beta}) \longrightarrow$ SDE (Li and Nica, 2024).

Solutions:

1. **Skip connection:**
$$h_{\ell+1} = h_\ell + \frac{1}{\sqrt{n}}\sqrt{\frac{c}{n}}\, W_\ell\, \varphi(h_\ell)$$
2. **Modify activations:** Morten *et al.* (2021), "*Deep Kernel Shaping*". Zhang et al. (2022): "Tailored Activation Transform"

shaped activation function:
$$\varphi_s(x) = c\,\varphi(ax+b) + d$$
For relu:
$$\varphi_s(x) = \max\{x,0\} \;-\; a\,\min\{x,0\}$$
Treat $\rho_d^{\alpha\beta}$ as a function of a, b, c, d; optimize
$$\rho_d^{\alpha\beta} = \rho_d^{\alpha\beta}(a,b,c,d) \qquad (n\to\infty)$$
optimize $\rho_d^{\alpha\beta} = 0.3$

Result: can train deep networks w/o heuristics (No adam. skip normalization)

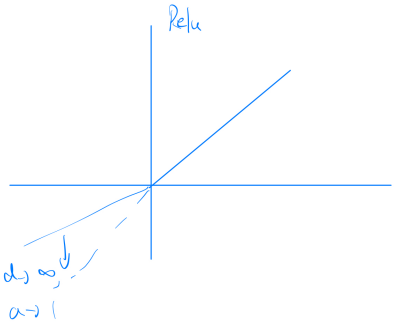

Figure 3: Relu

Zhang et al. show: as $h\to\infty,\ d\to\infty$, ρ_l converges to an ode,
$$\partial_\tau \rho_\tau^{\alpha\beta} = U\big(\rho_\tau^{\alpha\beta}\big)\,.$$
$$U(\rho) = \frac{\sqrt{1-\rho^2} \;+\; \rho\,(\pi - \arccos(\rho))}{2\pi}$$

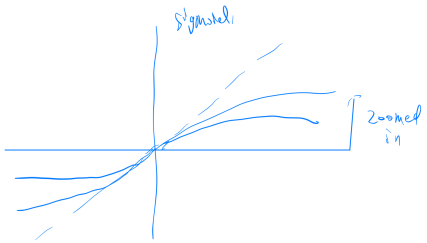

Figure 4: Sigmoid

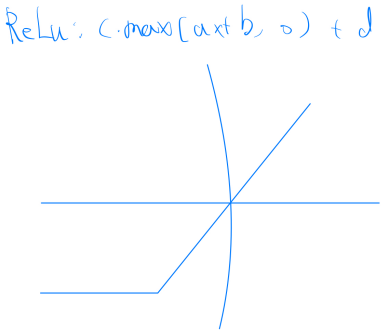

Figure 5: Shaped Relu

Shaping recipe:

For Relu:

$$\varphi_s(x) = s_+ \max(x, 0) \; + \; s_- \min(x, 0)$$

$$s_\pm = 1 + \frac{C_\pm}{n^p}, \qquad p \text{ to be determined}$$

Smoothing activation:

$$\varphi \in C^4(\mathbb{R}), \qquad \varphi(0) = 0, \qquad \varphi'(0) = 1$$

$$\varphi_s(x) = s\,\varphi\Big(\frac{x}{s}\Big), \qquad s = a\,n^p; \quad \big|\varphi^{(4)}(x)\big| \;\le\; C_1\,\big(1 + |x|\big)^q$$