Introduction to Neural Networks

Def: (Fully Connected)

Fix $L \geq 1$, $n_0, \dots, n_{L+1} \geq 1$, $\sigma: \mathbb{R} \to \mathbb{R}$

$$x \in \mathbb{R}^{n_0} \longmapsto z^{(1)}(x) = W^{(1)} x + b^{(1)} \in \mathbb{R}^{n_1}$$

$$\longmapsto z^{(2)}(x) = W^{(2)} \sigma(z^{(1)}(x)) + b^{(2)} \in \mathbb{R}^{n_2}$$

$$\vdots$$

$$\longmapsto z^{(L+1)}(x) = W^{(L+1)} \sigma(z^{(L)}(x)) + b^{(L+1)} \in \mathbb{R}^{n_{L+1}}$$

$W_{ij}^{(\ell)} \sim$ "weights" $\qquad b_i^{(\ell)} \sim$ "biases"

Typical Use:

① **Data Acquistion:** $D = \{ (x^{(i)}, y^{(i)}) \}_{i=1}^{n_{data}}$

$y^{(i)} = f(x^{(i)})$, f "unknown"

② **Model Selection:** Fix L, n_ℓ, σ.

③ **Initialization:** Choose

$\Theta = \{ W^{(\ell)}, b^{(\ell)} \}$ @ random

④ **Optimization:** Do GD:

$$\Theta(t+1) = \Theta(t) - \eta_t \nabla_\Theta \mathcal{L}(\Theta(t))$$

$$\mathcal{L}(\Theta) = \frac{1}{n_{data}} \sum_{k=1}^{n_{data}} \left(y^{(i)} - z(x^{(i)}; \Theta) \right)^2$$

Big Questions

1) Success of non-convex optimization: the loss $\mathcal{L}(\theta)$ is "very" non-convex in θ but GD-type optimization typically finds global opt.

Key: NNs are overparameterized

#params >> #data

"$\Rightarrow$" $\{\nabla_\theta \mathcal{L}(x^{(i)})\}_{i=1}^{n_{data}} \subseteq \mathbb{R}^{\#params}$

linearly indep

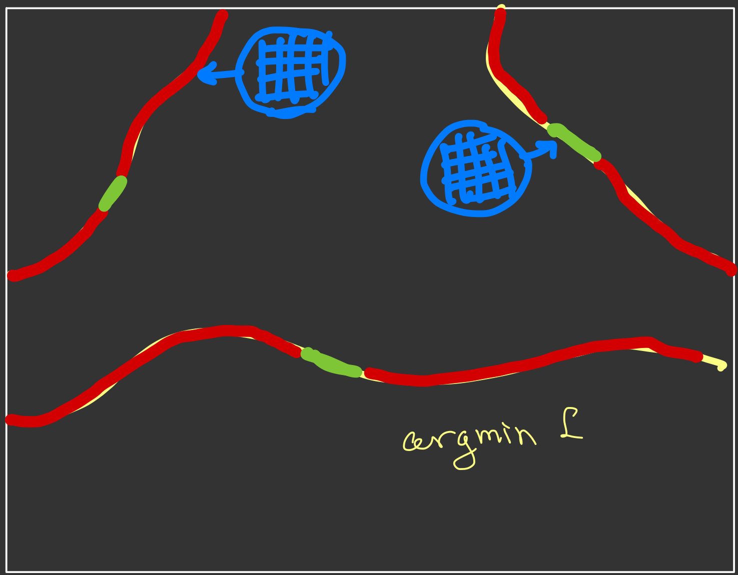
$$\Rightarrow \nabla_\theta \frac{1}{n_{data}} \sum_{i=1}^{n_{data}} \mathcal{L}(x_i;\theta) = 0$$

$$\Leftrightarrow \nabla_\theta \mathcal{L}(x_i;\theta) = 0$$

Neural Tangent Kernel

2) Implicit/Algorithmic Regularization

$\mathbb{R}^{\#params}$



Q: How do diff opt methods influence which global min is found?

Random Neural Nets

Fix $L \geq 1$, $n_0, \dots, n_{L+1} \geq 1$, $\sigma : \mathbb{R} \to \mathbb{R}$

$$x \in \mathbb{R}^{n_0} \mapsto z^{(1)}(x) = W^{(1)} x \in \mathbb{R}^{n_1}$$

$$\mapsto z^{(2)}(x) = W^{(2)} \sigma(z^{(1)}(x)) \in \mathbb{R}^{n_2}$$

$$\vdots$$

$$\mapsto z^{(L+1)}(x) = W^{(L+1)} \sigma(z^{(L)}(x)) \in \mathbb{R}^{n_{L+1}}$$

with

$$W_{ij}^{(\ell)} \sim \mathcal{N}\left(0, \frac{C_W}{n_{\ell-1}}\right)$$

Goal: Fix n_0, n_{L+1}, σ. ~~Str~~ Describe $x \mapsto z^{(L+1)}(x)$ when

$$n_1, \dots, n_L \simeq n \gg 1$$

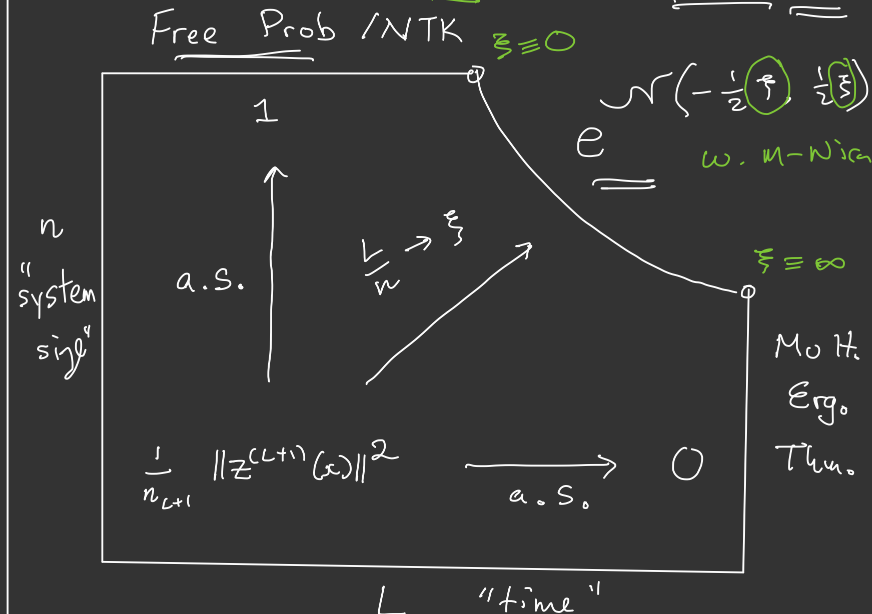
Metaclaim: When n, L large the distribution $z^{(L+1)}(x)$ depends on

$$\frac{L}{n} \left(= \frac{1}{n_1} + \dots + \frac{1}{n_L}\right)$$

and universality class of σ.

Intuition: $\sigma(t) = t$, $C_W = 1$

$$z^{(L+1)} = W^{(L+1)} \cdots W^{(1)} x \qquad \|x\|^2 = n_0$$



$$\frac{1}{n_{L+1}} \| W^{(L+1)} \cdots W^{(1)} x \|^2$$

$$= \frac{1}{n_{L+1}} \left\| W^{(L+1)} \cdots W^{(2)} \frac{W^{(1)} x}{\|W^{(1)} x\|} \right\|^2 \|W^{(1)} x\|^2$$

$$\overset{d}{=} \prod_{\ell=1}^{L+1} \| W^{(\ell)} u \|^2 \approx \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right)$$

$$O\left(\sqrt{\frac{L}{n}}\right)$$

Thm: Fix L, n_0, n_{L+1}. As $n_1, \dots, n_L \to \infty$

$$z^{(L+1)}(x) \in \mathbb{R}^{n_{L+1}} \longrightarrow GP(0, K^{(L+1)})$$

$$\lim_{n_1 \cdots n_L \to \infty} \mathrm{Cov}\left(z_i^{(L+1)}(x), z_j^{(L+1)}(x')\right) = \delta_{ij} K^{(L+1)}(x, x')$$

$$K^{(L+1)}(x, x') = C_W \, \mathbb{E}_{K^{(L)}}\left[\sigma(z_1^{(L)}(x))\, \sigma(z_1^{(L)}(x'))\right]$$

Ex: $\sigma(t) = \mathrm{ReLU}(t) = t\, \mathbb{1}_{t>0}$

$$K^{(L+1)}(x, x) = C_W \, \mathbb{E}_{K^{(L)}}\left[(z_1^{(L)}(x))^2 \, \mathbb{1}_{z_1^{(L)}(x) > 0}\right] = \frac{C_W}{2} K^{(L)}(x, x)$$

$C_W = 2$: $K^{(L)}(x, x) \simeq L^0$

Ex: $\sigma(t) = \tanh(t)$

$C_W = 1$: $K^{(L)}(x, x) \simeq \frac{1}{L^1}$

All This: $\frac{L}{n} \longrightarrow 0$

complexity

$$z^{(L+1)}(x) = W^{(L+1)} \sigma\left(W^{(L)} \sigma\left(\cdots \sigma\left(W^{(2)} \sigma(W^{(1)} x)\right) \cdots\right)\right)$$

Thm: (H) When $n_1, \dots, n_L \simeq n \gg 1$

$$\kappa\left(z_{i_1}^{(L+1)}(x_1), \dots, z_{i_k}^{(L+1)}(x_k)\right) = \begin{cases} 0, & k \text{ odd} \\ O_{\sigma, L}\left(\frac{1}{n^{\frac{k}{2} - 1}}\right), & k \text{ even} \end{cases}$$

$$\Rightarrow \kappa_{2k}^{(L+1)}(x) \equiv \kappa\Big(\underbrace{z_1^{(L+1)}(x), \dots, z_1^{(L+1)}(x)}_{2k \text{ times}}\Big) = O_L\left(\frac{1}{n^{k-1}}\right)$$

Obtain recursions for $\kappa_{2k}^{(L+1)}(x)$ in terms of $\{K_{2j}^{(L)}(x), j \le k\}$ and find: $k = 2, 3, 4$

$$\frac{\kappa_{2k}^{(L+1)}(x)}{\left(\kappa_2^{(L+1)}(x)\right)^k} = C_\sigma \frac{L^{k-1}}{n^{k-1}} + o(n^{-k}) = C_\sigma \left(\frac{L}{n}\right)^{k-1} + o(n^{-k})$$

Cor:

$$\mathrm{Cov}\left(\left(z_1^{(L+1)}(x)\right)^2, \left(z_2^{(L+1)}(x)\right)^2\right) = C_\sigma \cdot \frac{L}{n}$$

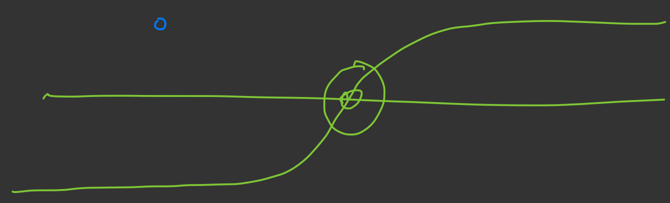
$$\mathrm{Var}\left(\left\|\frac{\partial z_1^{(L+1)}}{\partial x}(x)\right\|^2\right) = C_\sigma \cdot \frac{L}{n}$$

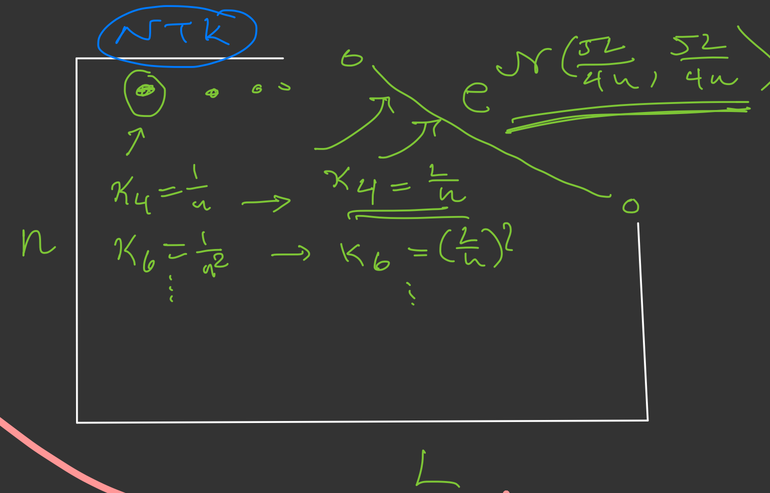
$\iff$ (EVGP happens in FC iff $\frac{L}{n} \gg 1$)

$$\|\mathrm{Hess}_\theta\, z_1^{(L+1)}(x;\theta)\| \simeq C_\sigma \cdot \frac{L}{n}$$

$\iff$ ($\frac{L}{n} > 0$ necessary in FC for feature learning)

⋮





make $\frac{L}{n}$ small

make $\frac{L}{n}$ large

$$\sigma(t) = \left(a_- \mathbb{1}_{t<0} + a_+ \mathbb{1}_{t>0}\right) t$$

$$C_W = \frac{2}{a_-^2 + a_+^2} \qquad (a_- - a_+ = c$$

$$K^{(L)}(x,x) = L^{-\alpha}$$

$$\partial_x \partial_{x'} \Big|_{x=x'} K^{(L)}(x,x') = L^{-\beta}$$

X