ORF 543

Lecture 9/14 - NNs at Initialization

Consider a FCNN:

$$\text{input } \vec{x}_\alpha \in \mathbb{R}^{n_0} \to \vec{z}^{(1)}_\alpha = W^{(1)} x_\alpha + \vec{b}^{(1)} \in \mathbb{R}^{n_1}$$

$$\to \vec{z}^{(2)}_\alpha = W^{(2)} \sigma(\vec{z}^{(1)}_\alpha) + \vec{b}^{(2)} \in \mathbb{R}^{n_2}$$

$$\vdots$$

$$\to \vec{z}^{(L+1)}_\alpha \in \mathbb{R}^{n_{L+1}}$$

So, $z^{(\ell+1)}_{i\alpha} = b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma(z_j^{(\ell)})$

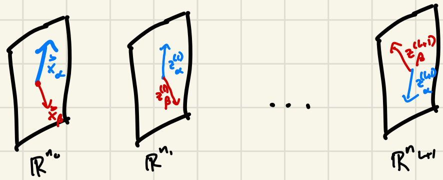
and $\vec{z}^{(\ell)}_\alpha = \langle z^{(\ell)}_{1\alpha}, \dots, z^{(\ell)}_{n_\ell \alpha} \rangle$

We ask how to initialize $W_{ij}^{(\ell)}, b_i^{(\ell)}$ and learning rates for GD?

Gaussian Initialization

Consider $W_{ij}^{(\ell)} \sim N(0, V_W^{(\ell)})$, $\vec{b}_i^{(\ell)} \sim N(0, V_b^{(\ell)})$

We use the information propagation framework, where we want feature dot products to be conserved across layers.



In math, we want to select $V_W^{(\ell)}(n_\ell, \sigma, L)$ and $V_b^{(\ell)}(n_\ell, \sigma, L)$ s.t.

$$\forall \ell \in \{0, \dots, L\} \quad \frac{1}{n_\ell} \langle \vec{z}^{(\ell)}_\alpha, \vec{z}^{(\ell)}_\beta \rangle \approx \frac{1}{n_{\ell+1}} \langle \vec{z}^{(\ell+1)}_\alpha, \vec{z}^{(\ell+1)}_\beta \rangle$$

preserved dot product averaged over dims

There are two useful consequences of conservation of dot product

① We approximately preserve across $\ell \in \{0, \dots, L+1\}$

$$\frac{1}{n_\ell} \|\vec{z}_\alpha^{(\ell)}\|^2 , \quad \left\langle \frac{\vec{z}_\alpha^{(\ell)}}{\|\vec{z}_\alpha^{(\ell)}\|}, \frac{\vec{z}_\beta^{(\ell)}}{\|\vec{z}_\beta^{(\ell)}\|} \right\rangle$$

② The Law of Large Numbers suggests $\forall \ell$,

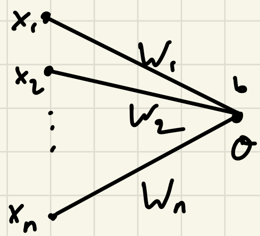
$$\frac{1}{n_\ell} \langle \vec{z}_\alpha^{(\ell)}, \vec{z}_\beta^{(\ell)} \rangle = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} z_{ji\alpha}^{(\ell)} z_{ji\beta}^{(\ell)} \approx \mathbb{E}\{ z_{ij\alpha}^{(\ell)} \cdot z_{ij\beta}^{(\ell)} \}$$

So, information propagation says we preserve the following across ℓ:

$$\mathbb{E}\{ z_{ij\alpha}^{(\ell)} \} , \quad \mathrm{Cov}\{ z_{ij\alpha}^{(\ell)} , z_{ij\beta}^{(\ell)} \}$$

Both conditions basically say mean and variance stay constant.

We can develop the following heuristic for $V_w^{(\ell)}, V_b^{(\ell)}$ with respect to n_ℓ.



Suppose $|x_i| = O(1)$ ← low order, not unreasonably big or small

and $W_j \sim N(0, V_w)$, $b \sim N(0, V_b)$

Since $\vec{z} = \langle \vec{w}, \vec{x} \rangle + b$, we see $\underbrace{\vec{z} \sim N(0, V_b + \sum_{i=1}^{n} V_w \|x_i\|^2)}_{\text{sum of gaussians}}$

$$\Rightarrow \vec{z} \sim N(0, V_b + n V_w \cdot O(1))$$

We arrive at fan-in scaling, where

$$\boxed{\begin{aligned} V_b^{(\ell)} &= C_b, \quad C_b = O(1) \\ V_w^{(\ell)} &= \frac{C_w}{n_{\ell-1}}, \quad C_w = O(1) \end{aligned}}$$

weight variance scales with width

Def: Let T be a set. Then a Gaussian process by T is $\{X_t\}_{t\in T}$ such that $\langle X_{t_1},\dots,X_{t_k}\rangle \in \mathbb{R}^k$ is Gaussian $\forall \{t_1,\dots,t_k\} \subseteq T$

ex/ Let $T=\{1,\dots,n\}$. Let $X=\langle X_1,\dots,X_n\rangle$ be jointly Gaussian.
ex/ Let $T=\mathbb{R}$. $X=X_t$ is gaussian process iff X is a random function on $\mathbb{R}$ with finite-dim distribution (fdd) $\langle X_{t_1},\dots,X_{t_n}\rangle \in \mathbb{R}^n$ Gaussian.

Theorem: (Neal, Lee, ..., Hanin)
Fix n_0, n_{L+1}, σ. Then as $n_1,\dots,n_L \to \infty$ (input/output dims fixed, but else $\to\infty$)

$$\vec{z}^{(L+1)}_{\alpha} \to GP(0, K^{(L+1)}) \quad \text{(Gaussian Process)}$$

i.e. $\mathbb{E}\{z^{(L+1)}_{i\alpha}\}=0$ and $\mathrm{Cov}(z^{(L+1)}_{i\alpha}, z^{(L+1)}_{j\beta}) = \delta_{ij} K^{(L+1)}_{\alpha\beta}$ (cov of two params when passed in $\vec{x}_\alpha, \vec{x}_\beta$)

This describes what happens when we send previous layers to infinite width. We can then recursively define

$$K^{(1)}_{\alpha\beta} = C^{(1)}_b + \frac{C^{(1)}_W}{n_0}\langle \vec{x}_\alpha, \vec{x}_\beta\rangle$$

$$K^{(\ell+1)}_{\alpha\beta} = C^{(\ell)}_b + C^{(\ell)}_W \underbrace{\mathbb{E}_{K^{(\ell)}_{\alpha\beta}}\{\sigma(\vec{z}^{(\ell)}_\alpha)\,\sigma(\vec{z}^{(\ell)}_\beta)\}}$$

$$K^{(\ell)}_{\alpha\beta} = \lim_{n_0,\dots,n_{\ell-1}\to\infty} \mathrm{Cov}(z^{(\ell)}_{i\alpha}, z^{(\ell)}_{i\beta}) \approx \frac{1}{n_\ell}\langle \vec{z}^{(\ell)}_\alpha, \vec{z}^{(\ell)}_\beta\rangle$$

$$\begin{pmatrix} \vec{z}^{(\ell)}_\alpha \\ \vec{z}^{(\ell)}_\beta \end{pmatrix} \sim N\Big(0, \underbrace{\begin{pmatrix} K^{(\ell)}_{\alpha\alpha} & K^{(\ell)}_{\alpha\beta} \\ K^{(\ell)}_{\beta\alpha} & K^{(\ell)}_{\beta\beta} \end{pmatrix}}_{\text{call this } K^{(\ell)}}\Big)$$

$$\int_{\mathbb{R}^2} \frac{\sigma(z^{(\ell)}_\alpha)\,\sigma(z^{(\ell)}_\beta)\, e^{-\frac{1}{2}\langle (K^{(\ell)})^{-1}\left[\begin{smallmatrix} z^{(\ell)}_\alpha \\ z^{(\ell)}_\beta\end{smallmatrix}\right], \left[\begin{smallmatrix} z^{(\ell)}_\alpha \\ z^{(\ell)}_\beta\end{smallmatrix}\right]\rangle}}{\det(2\pi K^{(\ell)})^{\frac{1}{2}}}\, dz^{(\ell)}_\alpha\, dz^{(\ell)}_\beta$$

Info prop $\iff$ $C^{(\ell)}_b, C^{(\ell)}_W$ are set such that $K^{(\ell)}_{\alpha\beta}$ is well-behaved at large ℓ.

Lecture 9/19 Tuning to Criticality

Note: at a particular layer $(\ell+1)$, we are given $\vec{z}_\alpha^{(\ell)}$ which is a random variable every neuron in the layer shares

$\Rightarrow$ $z_{i\alpha}^{(\ell+1)}$ are i.i.d. Gaussian with variance

$$C_b + \frac{C_w}{n_\ell} \|\sigma(\vec{z}_\alpha^{(\ell)})\|^2$$

$\leftarrow$ Gaussian with a random variance

Recall that the goal of info. prop. is to conserve

$$\underbrace{\frac{1}{n_\ell}\langle z_\alpha^{(\ell)}, z_\beta^{(\ell)}\rangle}_{K_{\alpha\beta}^{(\ell)}} \approx \underbrace{\frac{1}{n_{\ell+1}}\langle z_\alpha^{(\ell+1)}, z_\beta^{(\ell+1)}\rangle}_{K_{\alpha\beta}^{(\ell+1)}}$$

So, in the infinite limit, the goal is to find C_b, C_w s.t. $K_{\alpha\beta}^{(\ell)}$ is as constant as possible across ℓ.

Ex/ $\sigma(t) = t$ (Deep Linear Networks)

$\alpha=\beta$: $K_{\alpha\alpha}^{(\ell+1)} = C_b + C_w \mathbb{E}_{K^{(\ell)}}\{\sigma(z_\alpha)^2\}$

This is like supposing the previous layer already went to infinite width $\Rightarrow$ $z_{i\alpha}^{(\ell)} \sim N(0, K_{\alpha\alpha}^{(\ell)})$

$$= C_b + C_w \int_{-\infty}^{\infty} z_\alpha^2 e^{-\frac{z_\alpha^2}{2K_{\alpha\alpha}^{(\ell)}}} \frac{dz_\alpha}{\sqrt{2\pi K_{\alpha\alpha}^{(\ell)}}} = C_b + C_w K_{\alpha\alpha}^{(\ell)}$$

$\alpha \neq \beta$: $K_{\alpha\beta}^{(\ell+1)} = C_b + C_w \mathbb{E}_{K^{(\ell)}}\{\langle z_\alpha, z_\beta\rangle\} = C_b + C_w K_{\alpha\beta}^{(\ell)}$

So, if $\sigma(t) = t$ we want to choose $C_b = 0$, $C_w = 1$.

Remark: If $C_b = 0$ but $C_w \neq 1$, we have an initialization

$$K_{\alpha\beta}^{(\ell+1)} = (C_w)^\ell K_{\alpha\beta}^{(0)}$$

$\leftarrow$ vanishes or explodes if $C_w \neq 1$

Ex/ $\sigma(t) = \mathrm{ReLU}(t) = \max\{0, t\}$

We have $K^{(\ell+1)}_{\alpha\alpha} = C_b + C_W \int_0^\infty \frac{z_\alpha^2 \, e^{-\frac{z_\alpha^2}{2K^{(\ell)}_{\alpha\alpha}}}}{\sqrt{2\pi K^{(\ell)}_{\alpha\alpha}}} dz_\alpha = C_b + \frac{C_W K^{(\ell)}_{\alpha\alpha}}{2}$

With $C_b = 0$, we require $1 = \frac{C_W^{(0)} \cdots C_W^{(\ell)}}{2^\ell}$ $\forall \ell$ $\Rightarrow$ $C_W = 2$ "He-initialization"

However, when $\sigma(t) \neq t$, $\mathbb{E}_{K^{(\ell)}}\{\sigma(z_\alpha)\sigma(z_\beta)\}$ is hard.

We can claim that the recursion

$$(\#) \quad K^{(\ell+1)}_{\alpha\beta} = C_b + C_W \mathbb{E}_{K^{(\ell)}}\{\sigma(z_\alpha), \sigma(z_\beta)\}$$

is a 3d dynamical system with variables $(K^{(\ell)}_{\alpha\alpha}, K^{(\ell)}_{\beta\beta}, K^{(\ell)}_{\alpha\beta})$ with time parameter ℓ.

To solve such a system, we find fixed points, linearize about the fixed points, and ensure the points are stable & critical.

Fixed points at $(*)$ $K_* = C_b + C_W \mathbb{E}_{K_*}\{\sigma^2(z)\}$

$$\left(K^{(\ell)}_{\alpha\alpha} = K_* \Rightarrow K^{(\ell+1)}_{\alpha\alpha} = K_*\right)$$

This condition will have that at deep layers, if $\vec{x}_\alpha \sim N(0, K_*)$, then at large ℓ, $\frac{1}{n_\ell}\|z^{(\ell)}_\alpha\|^2 \approx \frac{1}{n_0}\|x_\alpha\|^2 = K_*$

The second condition is

parallel perturbation $\perp$ of x_α in direction of x_α

$$(\parallel) \quad \left.\frac{\partial K^{(\ell+1)}_{\alpha\alpha}}{\partial K^{(\ell)}_{\alpha\alpha}}\right|_{K^{(\ell)}_{\alpha\alpha} = K_*} = 1 \qquad \left(\underbrace{K^{(\ell)}_{\alpha\alpha} = K_* + \delta K \Rightarrow K^{(\ell+1)}_{\alpha\alpha} = K_* + \delta K + O(\delta K)}_{\text{linearized}}\right)$$

Thirdly,

$$(\perp) \quad \left.\frac{\partial K^{(\ell+1)}_{\alpha\beta}}{\partial K^{(\ell)}_{\alpha\beta}}\right|_{K^\ell_{\alpha\alpha} = K^\ell_{\beta\beta} = K^\ell_{\alpha\beta} = K_*} \qquad \left(\underbrace{K^{(\ell)}_{\alpha\beta} = K_* + \delta K \Rightarrow K^{(\ell+1)}_{\alpha\beta} = K_* + \delta K + O(\delta K)}_{\text{linearized}}\right)$$

These are the dynamical systems constraints for a fixed, stable, critical fixed points. Note that we treat this as small perturbation from a point to generate x_α, x_β, which is why we use linear approx.

Now, $$\frac{\partial K^{(\ell+1)}_{\alpha\alpha}}{\partial k^{(\ell)}_{\alpha\alpha}} = \frac{\partial}{\partial k^{(\ell)}_{\alpha\alpha}}\left(C_b + C_w \mathbb{E}_{K^{(\ell)}}\left\{\sigma(z_\alpha)^2\right\}\right)$$

$$= C_w \frac{\partial}{\partial k^{(\ell)}_{\alpha\alpha}} \int \sigma(z_\alpha)^2 \frac{e^{-\frac{z_\alpha^2}{2k^{(\ell)}_{\alpha\alpha}}}}{\sqrt{2\pi k^{(\ell)}_{\alpha\alpha}}}\, dz_\alpha$$

(Gaussians are alter in Fourier space)

$$\underset{\text{Fourier Transform}}{=} C_w \frac{\partial}{\partial k^{(\ell)}_{\alpha\alpha}} \int \hat{\sigma}^2(\xi)\, e^{-\frac{k^{(\ell)}_{\alpha\alpha}}{2}\xi^2}\, d\xi$$

$$= C_w \int \hat{\sigma}^2(\xi)\left(-\tfrac{1}{2}\xi^2\right) e^{-\frac{k^{(\ell)}_{\alpha\alpha}}{2}\xi^2}\, d\xi$$

F.T. diagonalizes the derivative

$$= C_w \int \frac{1}{2}\partial^2_{z_\alpha}\left(\sigma(z_\alpha)\right) \frac{e^{-\frac{z_\alpha^2}{2k^{(\ell)}_{\alpha\alpha}}}}{\sqrt{2\pi k^{(\ell)}_{\alpha\alpha}}}\, dz_\alpha$$

$$\chi_{\parallel}(K_*) = \frac{C_w}{2}\mathbb{E}_{K_*}\left\{\partial^2\left(\sigma^2(z)\right)\right\} = 1$$

We can do the same thing to find

$$\chi_{\perp}(K_*) = C_w \mathbb{E}_{K_*}\left\{\left(\partial\sigma(z)\right)^2\right\} = 1$$

So, the constraints of "tuning to criticality" result with

$$\boxed{\begin{aligned} &(*)\quad K_* = C_b + C_w \mathbb{E}_{K_*}\left\{\sigma^2(z)\right\} \\ &(\parallel)\quad \chi_{\parallel}(K_*) \equiv \frac{C_w}{2}\mathbb{E}_{K_*}\left\{\partial^2\left(\sigma^2(z)\right)\right\} = 1 \\ &(\perp)\quad \chi_{\perp}(K_*) \equiv C_w \mathbb{E}_{K_*}\left\{\left(\partial\sigma(z)\right)^2\right\} = 1 \end{aligned}}$$

These conditions confirm that if you have two inputs x_α, x_β "close" with $\mathrm{Cov}(x_\alpha, x_\beta) = 1-\varepsilon$, things don't exponentially explode or vanish (K_* is fixed point).

We can return to $\sigma(t) = \text{ReLU}(t)$

$$(*)\ K_* = C_b + C_w \frac{K_*}{2}$$

$\Longrightarrow$

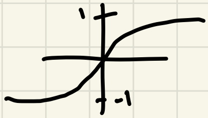
Gaussian integral

$$(\parallel)\ 1 = \frac{C_w}{2}\mathbb{E}_{K_*}\left\{\frac{\partial^2}{\partial z^2}\left(z^2 \mathbb{1}_{z>0}\right)\right\} = \frac{C_w}{2}\mathbb{E}_{K_*}\{2\,\mathbb{1}_{z>0}\} = \frac{C_w}{2}$$

$$(\perp)\ 1 = C_w\,\mathbb{E}_{K_*}\left\{(\partial z \mathbb{1}_{z>0})^2\right\} = C_w\,\mathbb{E}_{K_*}\{\mathbb{1}_{z>0}\} = \frac{C_w}{2}$$

So, $C_w = 2$, $C_b = 0$, $K_* \geq 0$ arbitrary

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



Note: the only fixed point is $K_* = 0$.

$$\chi_{\parallel}(K_*) = \frac{C_w}{2}\mathbb{E}_{K_*}\left\{\partial^2(\sigma^2(z))\right\} = C_w\,\mathbb{E}\left\{\partial\left(\sigma(z)\sigma'(z)\right)\right\}$$

$$= C_w\,\mathbb{E}_{K_*}\{\sigma(z)\sigma''(z)\} + \chi_{\perp}(K_*)$$

So, if you want $\chi_{\parallel}(K_*) = \chi_{\perp}(K_*) = 1$, we require

$$C_w\,\mathbb{E}_{K_*}\{\sigma(z)\sigma''(z)\} = 0 \iff K_* = 0$$

$\sigma\sigma''$ is even and 0 at origin

So what happens is that, at criticality,

$$C_b = 0,\ C_w = 1,\ K^{(\ell)}_{\alpha\alpha} = \frac{(C_w)^{\ell}}{2\ell} = \frac{1}{2\ell} \text{ at large } \ell.$$

Covariances approach fixed point, don't do exponential stuff.

Lecture 9/21 - NNGP

Theorem: Fix $L \geq 1$, $n_0, n_L \geq 1$, $\sigma: \mathbb{R} \to \mathbb{R}$. Define

$$z_{i\alpha}^{(l+1)} = \begin{cases} b_i^{(l+1)} + \sum_{j=1}^{n_l} W_{ij}^{(l+1)} \sigma(z_{j\alpha}^{(l)}) & l \geq 1 \\ b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_{j\alpha} & l = 0 \end{cases}$$

with $W_{ij}^{(l+1)} \sim N(0, \frac{C_W}{n_l})$, $b_i^{(l+1)} \sim N(0, C_b)$

If σ is poly bounded (i.e. $\exists n \geq 1, C > 0$ s.t. $\sup_{x \in \mathbb{R}} \frac{|\sigma(x)|}{1 + x^{2n}} \leq C$)

then for any $\vec{x}_A = (x_{\alpha_1}, \dots, x_{\alpha_k})$, $x_{\alpha_j} \in \mathbb{R}^{n_0}$, the output vector $\vec{z}_A^{(L+1)} = \langle z_{\alpha_1}^{(L+1)}, \dots, z_{\alpha_k}^{(L+1)} \rangle \in \mathbb{R}^{k \times n_{L+1}}$ converge in distribution as $n_1, \dots, n_L \to \infty$ to a mean 0 Gaussian with

$$\lim_{n_1, \dots, n_L \to \infty} \mathrm{Cov}(z_{i\alpha}^{(L+1)}, z_{j\beta}^{(L+1)}) = \delta_{ij} K_{\alpha\beta}^{(L+1)}$$

where

$$\begin{cases} K_{\alpha\beta}^{(l+1)} = C_b + C_W \mathbb{E}_{K^{(l)}} \{\sigma(z_\alpha) \sigma(z_\beta)\} & l \geq 1 \\ K_{\alpha\beta}^{(l+1)} = C_b + \frac{C_W}{n_0} \vec{x}_\alpha \cdot \vec{x}_\beta & l = 0 \end{cases}$$

Recall:

(1) Suppose $\vec{X}_n \in \mathbb{R}^k$ is a random variable with

$$\mathbb{E}\{e^{-i \vec{X}_n \cdot \vec{\xi}}\} \xrightarrow{n \to \infty} \mathbb{E}\{e^{-i \vec{X} \cdot \vec{\xi}}\} \quad \forall \vec{\xi} \in \mathbb{R}^k.$$

Then, $X_n \xrightarrow{d} X$ ← in distribution

(2) Suppose $X \sim N(\vec{\mu}, \Sigma) \in \mathbb{R}^k$. Then,

$$\mathbb{E}\{e^{-i \vec{X}_n \cdot \vec{\xi}}\} = e^{-i \vec{\mu} \cdot \vec{\xi} - \frac{1}{2} \vec{\xi}^T \Sigma \vec{\xi}}$$

Proof: We WTS that for any $\gamma = \langle \xi_1, \dots, \xi_{n_{L+1}} \rangle$, $\xi_j \in \mathbb{R}^k$,

$$\lim_{n_1, \dots, n_L \to \infty} \mathbb{E}\left\{ e^{-i \vec{z}_A^{(L+1)} \cdot \vec{\gamma}} \right\} = e^{-\frac{1}{2} \sum_{j=1}^{n_{L+1}} \xi_j^T K_A^{(L+1)} \xi_j} \quad (*)$$

where

$$K_A^{(L+1)} = \begin{pmatrix} K_{\alpha_1\alpha_1}^{(L+1)} & \cdots & \vdots \\ \vdots & \ddots & \\ \cdots & & K_{\alpha_k\alpha_k}^{(L+1)} \end{pmatrix}$$

Step 1: Vibes: we can think of the layers moving through the network as a Markov chain.

Given $z^{(L)}_{A}$, we find $z^{(L+1)}_{jA} \equiv \underbrace{\langle z^{(L+1)}_{j\alpha_1}, \dots, z^{(L+1)}_{j\alpha_k} \rangle}^{\text{mean 0 independent Gaussians}}$ and

$$\mathrm{Cov}\left(z^{(L+1)}_{j\alpha}, z^{(L+1)}_{j\beta} \,\middle|\, z^{(L)}_{A}\right)$$

Recall: If $\vec{X} \sim N(0, \Sigma) \in \mathbb{R}^k$ and $\vec{u}, \vec{v} \in \mathbb{R}^k$, $\langle \vec{X}\cdot\vec{u}, \vec{X}\cdot\vec{v} \rangle$ is Gaussian with mean 0 and $\mathrm{Cov}(\vec{X}\cdot\vec{u}, \vec{X}\cdot\vec{v}) = \vec{u}^T \Sigma \vec{v}$

Note that $z^{(L+1)}_{i\alpha} = \langle b^{(L+1)}_i, W^{(L+1)}_{i_1}, \dots, W^{(L+1)}_{i n_L} \rangle \cdot \langle 1, \sigma(z^{(L)}_{1\alpha}), \dots, \sigma(z^{(L)}_{n_L \alpha}) \rangle$

$$\Rightarrow \mathrm{Cov}\left(z^{(L+1)}_{j\alpha}, z^{(L+1)}_{j\beta}\right) = \begin{bmatrix} 1 \\ \sigma(\vec{z}^{(L)}_{\alpha}) \end{bmatrix}^T \begin{bmatrix} C_b & & 0 \\ & \frac{C_w}{n_1} & \\ 0 & \ddots & \frac{C_w}{n_L} \end{bmatrix} \begin{bmatrix} 1 \\ \sigma(\vec{z}^{(L)}_{\beta}) \end{bmatrix}$$

$$= C_b + \frac{C_w}{n_L} \sum_{j=1}^{n_L} \sigma(z^{(L)}_{j\alpha}) \sigma(z^{(L)}_{j\beta}) = \hat{K}^{(L+1)}_{\alpha\beta}$$

Thus,

$$\mathbb{E}\left\{ e^{-i \vec{z}^{(L+1)}_{A} \cdot \vec{\gamma}} \right\} = \mathbb{E}\left\{ \mathbb{E}\left\{ e^{-i \vec{z}^{(L+1)}_{A} \cdot \vec{\gamma}} \,\middle|\, \vec{z}^{(L)}_{A} \right\} \right\}$$

$$= \mathbb{E}\left\{ e^{-\frac{1}{2} \sum_{j=1}^{n_{L+1}} \vec{\gamma}_j^T \hat{K}^{(L+1)}_{A} \vec{\gamma}_j} \right\} \quad (\#)$$

we want this to approach constant Σ

" $z^{(L+1)}_{iA}$ are i.i.d. Gaussian with mean 0 but with some covariance $\hat{K}^{(L)}_{A}$ "

Step 2: Vibes: Each transition between layers is symmetric to permutation of the neurons. So, only averages can matter

Each entry of $\hat{K}^{(L+1)}_{A}$ has form

$$O^{(L)}_{f} = \frac{1}{n_L} \sum_{j=1}^{n_L} f(z^{(L)}_{jA})$$

$$= \frac{1}{n_L} \sum_{j=1}^{n_L} \left(b + C_w \sigma(z^{(L)}_{j\alpha}) \sigma(z^{(L)}_{j\beta})\right)$$

We can use the following Proposition:

Prop: If f is poly bounded, $\sup_{n_1, \dots, n_L \geq 1} \left|\mathbb{E}\{O^{(L)}_{f}\}\right| < \infty$ (always bounded)

and $\lim_{n_1, \dots, n_L \to \infty} \mathrm{Var}(O^{(L)}_{f}) = 0$ (goes to constant)

Corollary: If we define $K_{\alpha\beta}^{(L+1)} = \lim_{n_1,\dots,n_L \to \infty} \mathbb{E}\left\{\hat{K}_{\alpha\beta}^{(L+1)}\right\}$, then (#) $\Rightarrow$ (*).

Proof of corollary: The proposition gives $\hat{K}_{\alpha\beta}^{(L+1)} \xrightarrow{d} K_{\alpha\beta}^{(L+1)}$.
Also, the map $K \mapsto e^{-\frac{1}{2}\vec{\xi}^T K \vec{\xi}}$ is bounded & $\overset{\text{continuous}}{C^0}$.

So, all the network outputs' variances converge to the same shared deterministic covariance $K_{\alpha\beta}^{(L+1)}$. □

We now know that the output vectors converge in distributions to mean 0 Gaussians with $\lim_{n_1,\dots,n_L\to\infty} \mathrm{Cov}\left(z_{i\alpha}^{(L+1)}, z_{j\beta}^{(L+1)}\right) = \delta_{ij} K_{\alpha\beta}^{(L+1)}$

We complete the proof by deriving a recurrence relation for $K_{\alpha\beta}^{(L+1)}$. We know

$$K_{\alpha\beta}^{(L+1)} = \lim_{n_1,\dots,n_L\to\infty} \mathrm{Cov}\left(z_{i\alpha}^{(L+1)}, z_{i\beta}^{(L+1)}\right)$$

$\mathrm{Cov}(X,Y) = \mathbb{E}\{\mathrm{Cov}(X,Y|Z)\} + \mathrm{Cov}(\mathbb{E}\{X|Z\}, \mathbb{E}\{Y|Z\})$ = Law of Total Covariance

$$= \lim_{n_1,\dots,n_L\to\infty} \mathbb{E}\left\{\mathrm{Cov}\left(z_{i\alpha}^{(L+1)}, z_{i\beta}^{(L+1)} \,\middle|\, z_A^{(L)}\right)\right\} + \mathrm{Cov}\Big(\underbrace{\cancel{\mathbb{E}\{z_{i\alpha}^{(L+1)}|z_A^{(L)}\}}, \cancel{\mathbb{E}\{z_{i\beta}^{(L+1)}|z_A^{(L)}\}}}_{\text{mean 0 Gaussians}}\Big)$$

$$= \lim_{n_1,\dots,n_L\to\infty} \mathbb{E}\Big\{C_b + \frac{C_w}{n_L}\sum_{j=1}^{n_L} \underbrace{\sigma(z_{j\alpha}^{(L)})\sigma(z_{j\beta}^{(L)})}_{\text{all these have same expectation because of symmetry}}\Big\}$$

$$= C_b + C_w\, \mathbb{E}_{K^{(L)}}\left\{\sigma(z_\alpha)\sigma(z_\beta)\right\}$$

limiting outputs. So limit of expectation is expectation of limits via Continuous Mapping Theorem

We can repeat this logic via induction to get the recurrence relation.

We finish by going back and proving the proposition:

Prop: If f is poly bounded, $\sup_{n_1,\dots,n_L\geq 1} \left|\mathbb{E}\{O_f^{(L)}\}\right| < \infty$ (always bounded)

and $\lim_{n_1,\dots,n_L\to\infty} \mathrm{Var}\left(O_f^{(L)}\right) = 0$ (goes to constant)

Proof: We induct on L. When $L=1$,

$z_{iA}^{(1)} = \langle z_{i\alpha_1}^{(1)}, \dots, z_{i\alpha_k}^{(1)}\rangle$ are i.i.d Gaussian with mean 0 and $\mathrm{Cov}\left(z_{i\alpha}^{(1)}, z_{i\beta}^{(1)}\right) = C_b + \frac{C_w}{n_0}\vec{x}_\alpha \cdot \vec{x}_\beta$

$\leftarrow$ all have same variance

Thus, $\mathbb{E}\{O_f^{(1)}\} = \mathbb{E}\{f(z_{jA}^{(1)})\}$ is finite because f is poly bounded independently of n_1. Furthermore,

$$\mathrm{Var}(O_f^{(1)}) = \mathrm{Var}\left(\frac{1}{n_1}\sum_{j=1}^{n_1} f(z_{jA}^{(1)})\right) = \frac{1}{n_1}\mathrm{Var}\left(f(z_{jA}^{(1)})\right)$$

$$\leq \frac{1}{n_1}\mathbb{E}\{f(z_{jA}^{(1)})^2\}$$

$$\to 0 \text{ as } L \to \infty.$$

The inductive step happens because f is poly bounded.

$\square$

Lecture ? - LR in NTK/GP Regime

Last time: We saw

① How to set C_B, C_W in a random FCNN at large width of the form $z_{i\alpha}^{(\ell+1)} = b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\left(z_{j\alpha}^{(\ell)}\right)$

with $W_{ij}^{(\ell+1)} \sim \mathcal{N}\left(0, \frac{C_W}{n_\ell}\right)$ and $b_i^{(\ell+1)} \sim \mathcal{N}(0, C_B)$

② That as $n_1, \dots, n_L \to \infty$ $\vec{z}_\alpha^{(L+1)} \to GP(0, K^{(L+1)})$ with

$$\lim_{n_1,\dots,n_L \to \infty} \mathrm{Cov}\left(z_{i\alpha}^{(L+1)}, z_{j\beta}^{(L+1)}\right) = \delta_{ij} K_{\alpha\beta}^{(L+1)}$$

and the relation

$$K_{\alpha\beta}^{(\ell+1)} = C_b + C_W \mathbb{E}_{K^{(\ell)}}\left[\sigma(\vec{z}_\alpha)\sigma(\vec{z}_\beta)\right]$$

with $\chi_{\parallel} = \frac{C_W}{2}\mathbb{E}_{K_*}\left[\partial^2 \sigma^2(z)\right] = 1$, $\chi_\perp = C_W \mathbb{E}_{K_*}\left[\sigma'(z)^2\right] = 1$

Today: We ask how to set LR for GD to be "well-behaved"?

$$\theta(t+1) = \theta(t) - \eta_t \vec{\nabla}_\theta \mathcal{L}(\theta(t))$$

Intuition 1: Denote that $\vec{z}(\vec{x}, \theta) = \theta\vec{x}$, $Y = \theta_* X$, $\mathcal{L}(\theta) = \frac{1}{2}\|\theta X - Y\|_2^2$

This yields

$$\vec{\nabla}_\theta \mathcal{L}(\theta) = (\theta X - Y)X^T = (\theta - \theta_*)XX^T$$

So, the GD update step becomes

$$\begin{aligned}\theta(t+1) - \theta_* &= \theta(t) - \theta_* - \eta(\theta(t) - \theta_*)XX^T \\ &= (\theta(t) - \theta_*)(I - \eta XX^T)\end{aligned}$$

$$\Rightarrow \boxed{\eta < \frac{2}{\lambda_{max}(XX^T)} = \frac{2}{\|\mathrm{Hess}(\mathcal{L})\|_{op}} = \frac{2}{\lambda_{max}(\vec{\nabla}_\theta z (\vec{\nabla}_\theta z)^T)}}$$

Under this condition,

$$\begin{aligned}\|\theta(t+1) - \theta_*\|_2 &\le \|\theta(t) - \theta_*\|_2 (1 - \eta\lambda_{min}(XX^T)) \\ &\le \|\theta(t) - \theta_*\|_2 \, e^{-\eta t \lambda_{min}(XX^T)}\end{aligned}$$

So, the best convergence rate is $e^{\frac{-2t}{\kappa(XX^T)}}$, $\kappa(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$

Intuition 2: Suppose we have noisy gradients

$$\theta(t+1) = \theta(t) - \eta_t\left(\vec{\nabla}_\theta \mathcal{L}(\theta(t)) + \xi_t\right), \quad \xi_t \sim \mathcal{N}(0, \sigma^2)$$

$$\Rightarrow \theta(t+1) - \theta_* = (\theta(t) - \theta_*)(I - \eta_t X X^T) + \eta_t \xi_t$$

$$\Rightarrow \|\theta(t+1) - \theta_*\|_2 \le \|\theta(t) - \theta_*\|_2\, e^{-\eta_t \lambda_{min}} + \eta_t \|\xi_t\|$$

$$\le \|\theta(0) - \theta_*\|_2\, e^{-\sum_{s=0}^{t} \eta_t \lambda_{min}} + \sum_{s=0}^{t} \eta_s \|\xi_s\|\, e^{-\sum_{s'=s+1}^{t} \eta_{s'} \lambda_{min}}$$

So, we need $\boxed{\sum_{s=0}^{\infty} \eta_s = \infty \quad \text{and} \quad \eta_s \to 0}$ (also $\sum_{s=0}^{\infty} \eta_s^2 < \infty$)

Now, returning to wide NNs with scalar output ($n_{L+1} = 1$), the effective Jacobian is

$$\vec{\eta} \odot \vec{\nabla}_\theta z_{1\alpha}^{(L+1)} = \left(\eta_{\theta_j} \partial_{\theta_j} z_{1\alpha}^{(L+1)}, \quad j \in \{1, \dots, \#\text{params}\}\right)$$

elementwise

LR for each param

$$\Rightarrow \lambda_{max} = \|\vec{\eta} \odot \vec{\nabla}_\theta z_{1\alpha}^{(L+1)}\|^2$$

$$\Rightarrow \dots$$

$$\Rightarrow \boxed{\begin{array}{ll} \eta_b^{(\ell)} = O(1) & \left(\text{or } O\left(\frac{1}{L}\right)\right) \\ \eta_w^{(\ell)} = O\left(\frac{1}{\sqrt{n_{\ell-1}}}\right) & \left(\text{or } O\left(\frac{1}{L\sqrt{n_{\ell-1}}}\right)\right) \end{array}}$$

Lecture - Pathologies of NTK/GP Regime

Pathologies: ① As $n_1, \dots, n_L \to \infty$, GD on MSE equivalent to linear/Kernel method

$$\boxed{z_\alpha^{(L+1)}(\theta) \to \tilde{z}_\alpha^{(L+1)}(\theta) = z_\alpha^{(L+1)}(\theta(0)) + \eta \odot \vec{\nabla}_\theta z_\alpha^{(L+1)}(\theta(0))(\theta - \theta(0))}$$

learning happens in last layer and to first order in hidden layers ② No feature learning!

Fix: Mean-field init $W_{ij}^{(\ell)} \sim \begin{cases} \mathcal{N}\left(0, \frac{C_w}{n_{\ell-1}}\right) & \ell \le L \\ \mathcal{N}\left(0, \frac{1}{n_L^2}\right) & \ell = L+1 \end{cases}$ and $\eta_w^{(\ell)} = \eta_b^{(\ell)} = O(1)$

Lecture 10/3 - Loss Hessian

We can summarize the optimization of our network via

- Loss Hessian $\quad \mathrm{Hess}_\theta \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial \theta_1^2} & \frac{\partial^2 \mathcal{L}}{\partial \theta_1 \partial \theta_2} & \cdots \\ \vdots & \ddots & \\ \vdots & & \end{bmatrix} \approx$ negative inverse of Fisher Information

- NNGP (NN Gaussian Process) $\quad \vec{z}^{(L)}(D)^T \vec{z}^{(L)}(D) \in \mathbb{R}^{n_L \times n_L}$
  - infinite-width limit of Bayesian network, such as a randomly initialized NN like Lecture 9/21

- NTK (Neural Tangent Kernel) $\quad \vec{\nabla}_\theta z(D;\vec{\theta})\left(\vec{\nabla}_\theta z(D;\vec{\theta})\right)^T$
  - Kernel methods replace learning feature vectorizations with weighting the training inputs and inferring a Kernel $K(\vec{x},\vec{x}'): \mathcal{X}\times\mathcal{X}\to\mathbb{R}$
  - Kernels are great when $K(\vec{x},\vec{x}') = \langle \varphi(\vec{x}), \varphi(\vec{x}')\rangle_V$ for some vector space V and some $\varphi: \mathcal{X}\to V$
  - The NTK is a kernel $\varphi: \mathbb{R}^{n_{in}} \times \mathbb{R}^{n_{in}} \to \mathbb{R}^{n_{out}\times n_{out}}$ with

$$\varphi_{jk}(\vec{x},\vec{y};\vec{\theta}) = \sum_{\theta_i} \partial\theta_i z_j(\vec{x};\vec{\theta})\,\partial\theta_i z_k(\vec{y};\vec{\theta})$$

  - The NTK represents the influence of the loss gradient $\partial_w \mathcal{L}(w,\vec{y}_i)\big|_{w=\vec{z}(\vec{x}_i;\theta)}$ w.r.t. example $(\vec{x}_i,\vec{y}_i)$ on the evolution of the NN $\vec{z}(\cdot,\theta)$ through GD step.
  - In large width (large parameter) limit, NTK is constant & deterministic!

Hessian Eigenvalues (Sagun et. al.)

Figure: Bulk + Outliers Dynamics in Hessian spectrum with CE loss on 2 class task with clusters at (−1, −1) and (1, 1) with increasing variance.

Spectrum of $\quad \mathrm{Hess}_\theta \vec{z}(x, \underbrace{\theta(\infty)}_{\text{converged params}})$

- decomposes into bulk + outliers
- bulk has small eigenvalues (some negative)
- # outliers ≈ # of classes
- outlier size depends on batch size
- left edge of spectrum gets negative!

Properties in the Wild

- Hessian has rank at most $\min\{\#\text{data}, \#\text{params}\}$
- Larger eigenvalue $\Rightarrow$ sharper loss surface, faster optimization

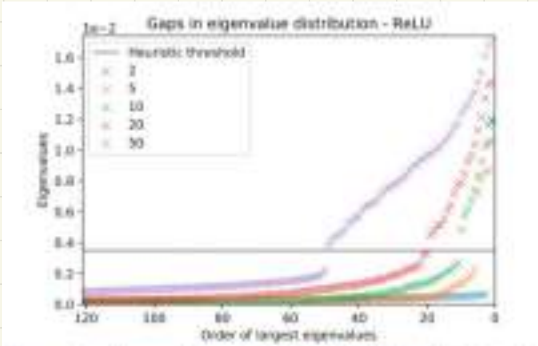


Figure: # outliers ≈ # classes in 100 - 30 - 30 - k - SM network.

Outlier eigenvalues correspond to class means??

Hessian Eigenvectors (Gur-Ari, Roberts, Dyer, "Gradient Descent Happens ...")

2 results that are robust!

① Top eigenvectors <u>stabilize</u> as training converges.

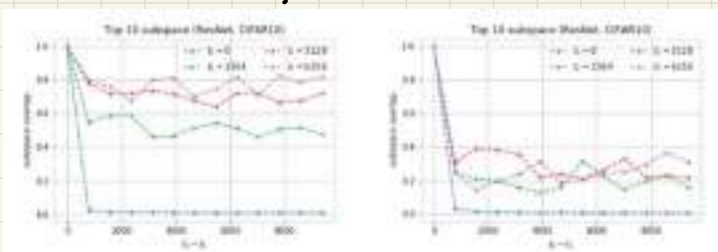


Figure: Stabilization of top eigenspace of Loss Hessian.

② Loss gradients are in span of top eigenvectors.

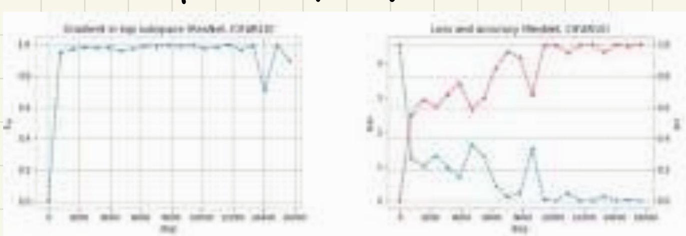


Figure: Loss gradients concentrate in top eigenspace of Loss Hessian.

Lecture 10/15 · Classifiers (Papyan, "Traces of Class")

Let x_{ic} be input, $c \in \{1, \dots, C\}$ is class and $i \in \{1, \dots, n\}$ is index.
Model output is $f(x_i, c) \in \mathbb{R}^C$ with softmax as $p(x_i, c) = \begin{bmatrix} e^{f(x_i,c)_1} / \sum_{c'=1}^{C} e^{f(x_i,c)_{c'}} \\ \vdots \end{bmatrix}$
Let $g_{icc'} = \vec{\nabla}_\theta \ell$ for an example x_{ic} if assigned label was c'.

Structure of loss Hessian

Gauss-Newton decomposition

where interesting eigenvalues happen

$$H_\theta[\ell(f(\theta))] = \vec{\nabla}_\theta f \, H_f \ell \, \vec{\nabla}_\theta f^T + \vec{\nabla}_f \ell \, H_\theta f := G + E$$

≈ 0 near convergence

For cross-entropy, G is 2^{nd} moment matrix:

$$g_{icc'} = \partial_\theta \ell(f(x_{ic}; \theta), y_{c'}) \qquad G = \operatorname*{Avg}_{i,c,c'} \{ g_{icc'} g_{icc'}^T \}$$

We decompose G into $\quad G = G_{class} + G_{cross} + G_{within} + G_{c=c'}$

covariance in a class $\quad G_{class} = \sum_c w_c g_c g_c^T$

covariance within class group $\quad G_{within} = \sum_{i,c,c'} w_{icc'} (g_{icc'} - g_{cc'})(g_{icc'} - g_{cc'})^T$

covariance between class groups and global incorrect average $\quad G_{cross} = \sum_{c,c'} w_{cc'} (g_{cc'} - g_c)(g_{cc'} - g_c)^T$

where

$$g_{cc'} = \operatorname*{Avg}_i \{ g_{icc'} \} \qquad g_c = \operatorname*{Avg}_{c \neq c'} \{ g_{cc'} \}$$

avg. gradient for class c if label were c' $\qquad$ Avg. incorrect gradient

3 level structure in unsupervised learning? what is C, and is the structure there
cool research question

We see the contributions of different parts to the 3-level structure.
(Bulk, C^2 outliers with higher eigenvalue of H, C outliers with even higher).

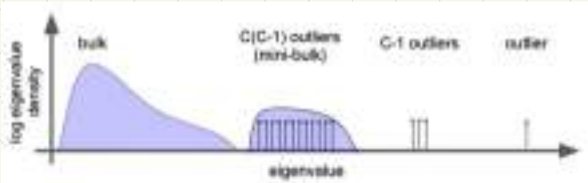


Figure: Cartoon of bulk + outlier structure in Hessian spectrum.

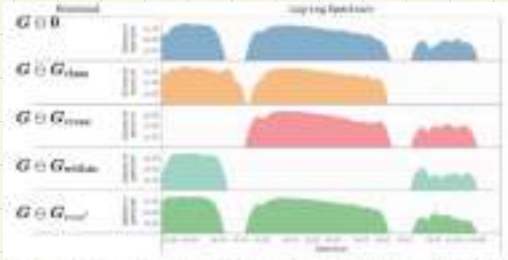


Figure: Bulk + outliers in FIM spectrum of VGG11 on CIFAR10

Structure of activations

Consider $\vec{h}^{\ell}_{ic} = \sigma\left(W^{(\ell)} \vec{h}^{\ell-1}_{ic}\right)$. Let $\underbrace{H^\ell \equiv \operatorname*{Avg}_{i,c}\left(h^\ell_{ic}\,(h^\ell_{ic})^T\right)}_{\text{feature covariance}}$

post-activations

We decompose $H^\ell = H^\ell_{class} + H^\ell_{within}$

between-class moment $\quad H^\ell_{class} = \operatorname*{Avg}_{c}\left\{\vec{h}^\ell_c \vec{h}^{\ell\,T}_c\right\}$ (mean)

within class 2nd moment $\quad H^\ell_{within} = \operatorname*{Avg}_{i,c}\left\{(\vec{h}^\ell_{ic} - \vec{h}^\ell_c)(\vec{h}^\ell_{ic} - \vec{h}^\ell_c)^T\right\}$ (variance)

where $\quad \underbrace{\vec{h}^\ell_c \equiv \operatorname*{Avg}_{i}\left\{\vec{h}^\ell_{ic}\right\}}_{\text{feature class means}} \qquad \vec{h}^\ell_G \equiv \operatorname*{Avg}_{c}\left\{\vec{h}^\ell_c\right\}$

We find larger eigenvalues and interesting outlier stuff happening for H^ℓ_{class}. The largest eigenvalue is class-agnostic.

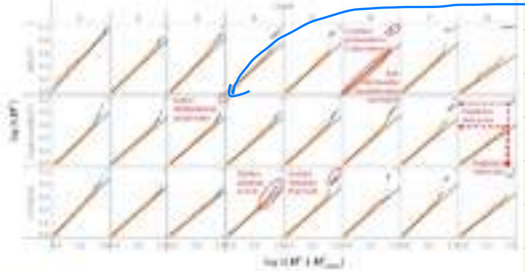


Figure: Eigenvalues of $H^\ell \perp H^\ell_{class}$ (x axis) vs H^ℓ_{class} (y axis). Outliers come from H^ℓ_{class}, especially in later layers.

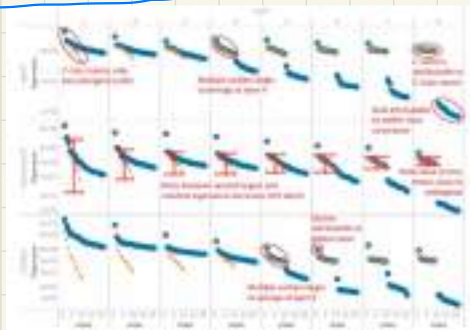


Figure: Eigenvalues of H^ℓ (blue) vs H^ℓ_{class} (orange).

Structure of backprop. grads

Let $\delta^\ell_{icc'}$ = layer ℓ grads, $\qquad \Delta^\ell = \operatorname*{Avg}_{i,c}\left\{\delta^\ell_{icc'}\,(\delta^\ell_{icc'})^T\right\}$

We decompose $\Delta^\ell = \Delta^\ell_{class} + \Delta_{cross} + \Delta^\ell_{within} + \Delta_{c=c'}$

where

$$\Delta^\ell_{class} = \operatorname*{Avg}_{c}\left\{\delta^\ell_c (\delta^\ell_c)^T\right\}$$

$$\Delta^\ell_{within} = \operatorname*{Avg}_{i,c}\left\{(\delta^\ell_{icc'} - \delta^\ell_{cc'})(\delta^\ell_{icc'} - \delta^\ell_{cc'})^T\right\}$$

$$\Delta^\ell_{cross} = \operatorname*{Avg}_{c\neq c'}\left\{(\delta^\ell_{cc'} - \delta^\ell_{c'})(\delta^\ell_{cc'} - \delta^\ell_{c'})^T\right\}$$

with

$\delta^\ell_{cc'} = \operatorname*{Avg}_{i}\left\{\delta^\ell_{icc'}\right\}$ grad cross-class means

$\delta^\ell_c = \operatorname*{Avg}_{c\neq c'}\left\{\delta^\ell_{cc'}\right\}$ grad class means

Neural collapse (Papyan, "Prevalence of Neural Collapse")

Call the layer ℓ output $\vec{h}^{\,\ell}_{ic}$.

We want to understand the late-time dynamics of $\vec{h}^{\,\ell}_{ic}$ via means

$$\vec{\mu}_G = \operatorname*{Avg}_{ic} \vec{h}_{ic} \qquad \vec{\mu}_c = \operatorname*{Avg}_{i} \vec{h}_{ic}$$

global mean — class mean

and covariances

$$\Sigma_B = \operatorname*{Avg}_{c} \left\{ (\vec{\mu}_c - \vec{\mu}_G)(\vec{\mu}_c - \vec{\mu}_G)^T \right\}$$

$$\Sigma_W = \operatorname*{Avg}_{ic} \left\{ (\vec{h}_{ic} - \vec{\mu}_c)(\vec{h}_{ic} - \vec{\mu}_c)^T \right\}$$

Phenomena of Neural Collapse

(1) Variability collapse $\Sigma_W \to 0$ (predictions approach class mean)

(2) $\{\mu_c \mid c \in 1, \dots, C\}$ approaches simplex vertices (class means are maximally orthogonal and same magnitude)

(3) Classification becomes nearest neighbors

(4) $W_c \approx \vec{\mu}_c - \vec{\mu}_G$

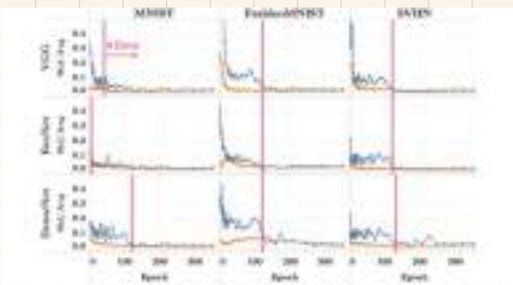


Figure: Coefficients of variation for $\{\|\mu_c\|\}$ (orange) and $\{\|\mu_c - \mu_G\|\}$ (blue). Simple datasets

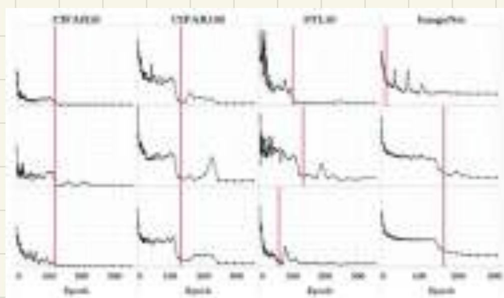


Figure: Mismatch between nearest-neighbor and NN classifiers. Complex datasets

"$\mathrm{Var}(\|\vec{\mu}_c\|)$ and $\mathrm{Var}(\|\vec{\mu}_c - \vec{\mu}_G\|) \to 0$"
Sorta holds for complex datasets
(1) and (2)

"Nearest-neighbor and NN behave similarly"
Doesn't hold for complex datasets
(3)

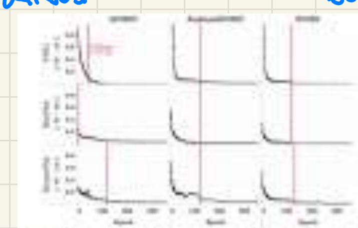


Figure: $\|M - W\|_F$ ... Simple datasets

∅ Pic of $\|W - \{\vec{\mu}_c - \vec{\mu}_G \mid c \in 1, \dots, C\}\|_F$ for (4)

"final layer approaches class means" **(4)**

Sharpness (Cohen et al, "Edge of Stability")

Train a NN with a fixed learning rate η.
We track "sharpness", or $\lambda_{max} = \lambda_{max}(\text{Hessian}_\theta)$ over time
← Hessian of loss

The theoretical expectation is that η should not be much larger than $2/\lambda_{max}$.

The empirical observation is that λ_{max} grows until $\lambda_{max} \approx 2/\eta$

They interpret this that the model finds "sharpest" parts during training so that steps are most meaningful.

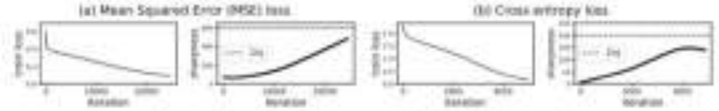


Figure: Dynamics of $2/\lambda_{max}$ over time for different losses.

Sharpness λ_{max} approaches $2/\eta$

Large Learning Rates - (Lewkowycz et al, "Catapult Phase")

We ask about fixing the NN and varying large η.
The finding is three phases

- lazy phase $0 < \eta < 2/\lambda_{max}(\text{NTK})$ ← mystery threshold
- catapult phase $2/\lambda_{max}(\text{NTK}) < \eta < C_*/\lambda_{max}(\text{NTK})$ ← here things look like losses grow divergently, but then they work out
- divergent phase $C_*/\lambda_{max}(\text{NTK}) < \eta$

output sensitivity to parameters

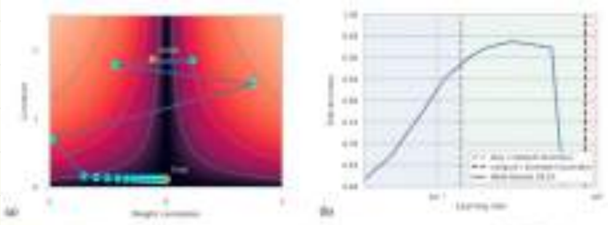


Figure: Weight correlation over training and test accuracy on CIFAR10 with fixed number of training steps.

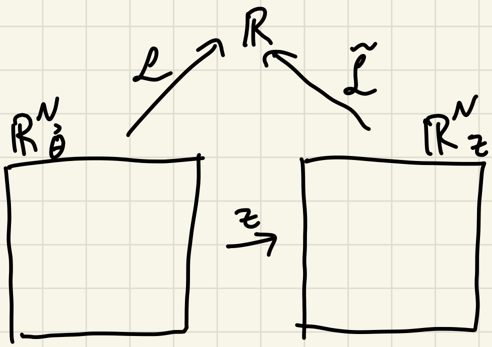
Best η is in catapult region.

Lecture 10/10 - Intro to NTK

Consider GD with

$$\vec{\theta}(t+1) = \vec{\theta}(t) - \eta \vec{\nabla}_{\theta} \mathcal{L}(\vec{\theta}(t))$$

Suppose that $\mathcal{L}(\vec{\theta}) = \tilde{\mathcal{L}}(z(\vec{\theta}))$ for some z! ← change of coordinates



In $z(\vec{\theta})$ variables,

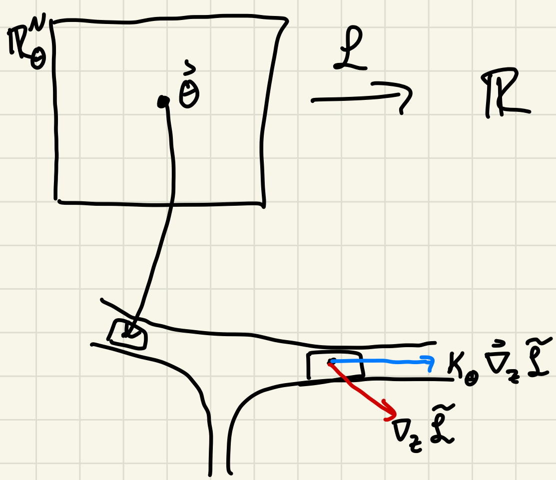
$$z(t+1) = z(\vec{\theta}(t+1))$$
$$= z(\vec{\theta}(t) - \eta \vec{\nabla}_{\theta} \mathcal{L}(\vec{\theta}(t)))$$
$$= z(\vec{\theta}(t) - \eta \vec{\nabla}_{\theta} \tilde{\mathcal{L}}(z(\vec{\theta}(t))))$$
$$= z(\vec{\theta}(t)) - \eta \vec{\nabla}_{\theta} z\big|_{\theta(t)} \cdot \vec{\nabla}_z \tilde{\mathcal{L}}\big|_{z=z(\vec{\theta}(t))} + O(\eta^2)$$
$$= z(t) - \eta \underbrace{(\vec{\nabla}_{\theta} z)^T}_{\substack{\text{Jacobian} \\ \in \mathbb{R}^{N \times n}}}(\vec{\nabla}_{\theta} z)\big|_{\theta(t)} \cdot \vec{\nabla}_z \tilde{\mathcal{L}}(z(t))$$

So,

$$\boxed{z(t+1) = z(t) - \eta K_{\vec{\theta}(t)} \vec{\nabla}_z \tilde{\mathcal{L}}(z(t))}$$

where $K_{\vec{\theta}(t)} \equiv \underbrace{(\nabla_{\vec{\theta}} z)^T (\nabla_{\vec{\theta}} z)}_{\text{Neural Tangent Kernel}} \in \mathbb{R}^{n \times n}$

The picture looks like



K_{θ} makes update steps move along the manifold allowed by $\mathrm{im}(z)$

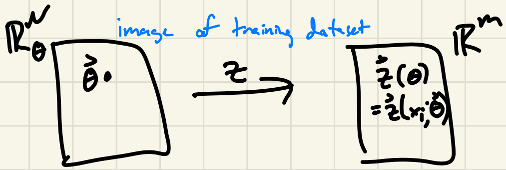
Ex/ MSE

Consider a NN $z(\vec{x};\vec{\theta})$ and m training data points $D=\{(\vec{x}_i,y_i), i=1,\dots,m\}$ and loss

$$\mathcal{L}(\vec{\theta})=\tilde{\mathcal{L}}(z(\vec{\theta}))=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(y_i-z(\vec{x}_i,\vec{\theta})\right)^2$$

We can use the change of coordinates induced by z to get that

$$\vec{z}(t)=\{z(\vec{x}_i;\vec{\theta}(t))\}_{i=1}^{m} \quad \text{and} \quad \vec{Y}=\{y_i\}_{i=1}^{m}$$



We have $\mathcal{L}(\vec{\theta}(t))=\frac{1}{2}\|\vec{Y}-\vec{z}(t)\|^2$ and $\vec{z}(t+1)=\vec{z}(t)-\eta K_{\theta(t)}(\vec{z}(t)-\vec{Y})$
where

$$\boxed{(K_{\theta(t)})_{ij}=\left(\vec{\nabla}_\theta z(\vec{x}_i;\vec{\theta}(t))\right)^T\vec{\nabla}_\theta z(\vec{x}_j;\vec{\theta}(t)) \qquad i,j\in\{1,\dots,m\}}$$

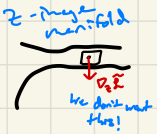
Gram Matrix

Key points!

- If $K_{\theta(t)}=K$ is independent of $\vec{\theta}$, this is "Kernel methods" on $\tilde{\mathcal{L}}(z)=\frac{1}{2}\|\vec{Y}-\vec{z}\|^2$. This is a time-varying Kernel
- Suppose $\exists\lambda_0>0$ s.t. $\forall t\geq 0$, $\lambda_{min}(K_{\theta(t)})\geq\lambda_0 \iff \lambda_\infty > K_{\theta(t)}\geq\lambda_0 I$ (#)

  K is PD and bounded

  This condition promises successful optimization.



if K_θ PSD
$\Rightarrow\langle K_\theta\vec{\nabla}_z\tilde{\mathcal{L}},\vec{\nabla}_z\tilde{\mathcal{L}}\rangle>0$,
we can always move in the direction given by $\vec{\nabla}_z\tilde{\mathcal{L}}$

Proof: $\mathcal{L}(\vec{\theta}(t+1))=\frac{1}{2}\|\vec{Y}-\vec{z}(t+1)\|^2$ and $\vec{z}(t+1)-\vec{Y}=(I-\eta K_{\theta(t)})(\vec{z}(t)-\vec{Y})$

Thus, if $\eta<\frac{1}{\lambda_{max}}$, $\mathcal{L}(\vec{\theta}(t+1))=\frac{1}{2}\|(I-\eta K_{\theta(t)})(\vec{z}(t)-\vec{Y})\|^2$

$$\leq\frac{1}{2}\|\vec{z}(t)-\vec{Y}\|^2(1-\eta\lambda_0)^2\leq\mathcal{L}(\theta(t))e^{-2\eta\lambda_0}$$

$$\Rightarrow\mathcal{L}(\vec{\theta}(t+1))\leq e^{-2\eta\lambda_0}\mathcal{L}(\theta(0))$$

$\square$

The goal is as follows:

For wide NNs w/ NTK init, MSE loss, small η, and $L<\infty$ fixed, the "Meta Theorem" is that $K_{\theta(t)}$ satisfies (#) (bounded PD). $\Rightarrow \mathcal{L} \to 0$.

The intuition is that if $K_{\theta(t)} \geq \lambda_0 I$, we can move $z(x_i;\vec{\theta})$ at will and always make progress. So, the data points cannot fight each other.

> To show (#) typically,
> (i) Show $K_{\theta(0)} \geq \lambda I$ and (ii) Show $\sup_{t\geq 0} \|K_{\theta(t)} - K_{\theta(0)}\| \leq \frac{\lambda_0}{2}$

Ex/ Simple NN $z_\alpha^{(2)} = \sum_{j=1}^{n} \frac{1}{\sqrt{n}} W_j^{(2)} \sigma(W_j^{(1)} \cdot \vec{x}_\alpha)$ with 1D output and 1 hidden layer.

Suppose that $|\sigma|, |\sigma'|, |\sigma''| \leq 1$ and $\|\vec{x}_\alpha\| = 1$ and we have NTK init

$$\left.\begin{array}{l} W_j^{(2)} \sim N(0,1) \\ W_j^{(1)} \sim N(0,I) \end{array}\right\} \vec{\theta}(0)$$

(i) We have

$$\left(K_{\theta(0)}\right)_{\alpha\beta} = \sum_{k=1}^{n} \partial_{W_k^{(2)}} z_\alpha^{(2)} \partial_{W_k^{(2)}} z_\beta^{(2)} + \langle \partial_{W_k^{(1)}} z_\alpha^{(2)}, \partial_{W_k^{(1)}} z_\beta^{(2)} \rangle$$

$$= \frac{1}{n} \sum_{k=1}^{n} \underbrace{\sigma(W_k^{(1)} \vec{x}_\alpha)\, \sigma(W_k^{(1)} \vec{x}_\beta)}_{K^{(2)}_{\alpha\beta}} + \underbrace{(W_k^{(2)})^2 \sigma'(W_k^{(1)} \vec{x}_\alpha)\, \sigma'(W_k^{(1)} \vec{x}_\beta)\, \vec{x}_\alpha \cdot \vec{x}_\beta}_{K^{(1)}_{\alpha\beta}}$$

$\Rightarrow K_{\theta(0)} \geq K^{(2)}_{\theta(0)}$ We only need $K^{(2)}_{\theta(0)} \geq \lambda_0 I$!

Now,

$$\left(K^{(2)}_{\theta(0)}\right)_{\alpha\beta} = \frac{1}{n} \sum_{k=1}^{n} \sigma(z^{(1)}_{k;\alpha})\, \sigma(z^{(1)}_{k;\beta}) = \frac{1}{n} \sum_{k=1}^{n} K^{(2)}_{k;\theta(0)}$$

avg. of iid matrices

<u>Idea:</u> we will write $K^{(2)}_{\theta(0)} = \mathbb{E}\{K^{(2)}_{1;\theta(0)}\} + \frac{1}{n} \sum_{k=1}^{n} K^{(2)}_{k;\theta(0)} - \mathbb{E}\{K^{(2)}_{k;\theta(0)}\}$

where $\left(K^{(2)}_{k;\theta(0)}\right)_{\alpha\beta} = \sigma(z^{(1)}_{k\alpha})\, \sigma(z^{(1)}_{k\beta})$ to get concentration bound on $K^{(2)}_{\theta(0)} - \mathbb{E}\{K^{(2)}_{\theta(0)}\}$

Theorem: Matrix Bernstein Inequality

Let $Z = \sum_{j=1}^{n} S_j$, where $S_j \sim$ iid with $\mathbb{E}\{S_j\} = 0 \ \forall j$

and $\|S_j\|_{op}$ ← largest eigenvalue of S_j $\le L$.

Let $V = \max\left\{ \left\|\sum_{j=1}^{n} \mathbb{E}\{S_j S_j^T\}\right\|_{op}, \left\|\sum_{j=1}^{n} \mathbb{E}\{S_j^T S_j\}\right\|_{op} \right\}$.

Then, $\mathbb{P}\{\|Z\|_{op} > t\} \le e^{\frac{-t^2/2}{V + Lt/3}}$

For us, $S_j = \frac{1}{n} K^{(2)}_{j;\theta(0)} - \mathbb{E}\{K^{(2)}_{j;\theta(0)}\} \Rightarrow \|S_j\|_{op} \le \left\|\frac{1}{n} K^{(2)}_{j,\theta(0)}\right\|_\infty \cdot m \le 2\frac{m}{n}$

Similarly, $V \le C\frac{m}{n} \overset{M.B.}{\Rightarrow} \mathbb{P}\{\|K^{(2)}_{\theta(0)} - \mathbb{E}\{K^{(2)}_{\theta(0)}\}\|_{op} > t\} \le e^{\frac{-ct^2}{(1+t)\frac{m}{n}}}$ set $t = \sqrt{\frac{m}{n}}$

$\Rightarrow \|K^{(2)}_{\theta(0)} - \mathbb{E}\{K^{(2)}_{\theta(0)}\}\| \le c\sqrt{\frac{m}{n}}$ with high probability.

So, $K^{(2)}_{\theta(0)}$ concentrates well about the mean. We now want to show the result for the expectation.

We WTS that if
- σ is not poly
- $\vec{x}_\alpha \ne \vec{x}_\beta$ if $\alpha \ne \beta \quad \forall \alpha, \beta$
- $\|x_\alpha\| = 1 \quad \forall \alpha$

$\Rightarrow \mathbb{E}\{K^{(2)}_{\theta(0)}\} \ge \lambda_0 I \iff \left(\mathbb{E}\{\sigma(W^{(1)} \cdot \vec{x}_\alpha)\sigma(W^{(1)} \cdot \vec{x}_\beta)\}\right)_{\alpha\beta} \ge \lambda_0 I$

Note that we can move from expectations in $\{\vec{x}_\alpha\}$ space to an infinite dimensional Hilbert space $\mathcal{H} = \{\Phi_\alpha\}$ s.t. $\mathcal{H} = L^2(\mathbb{R}^{n_0}, e^{-\frac{1}{2}\|W^{(1)}\|^2})$ ← inner products in $\mathcal{H}$ are expectations over $W^{(1)}$

$$\left(\text{Also, } \mathcal{H} = \{f : \mathbb{R}^{n_0} \to \mathbb{R} \mid \mathbb{E}\{f(W)^2\} < \infty\}\right)$$

So, $\mathcal{H}$ gives $\mathbb{E}\{K^{(2)}_{\theta(0)}\}_{\alpha\beta} = \langle \Phi_\alpha, \Phi_\beta \rangle_{\mathcal{H}}$ where $\Phi_\alpha^{(w)} = \sigma(W \cdot \vec{x}_\alpha)$

$$\Rightarrow \mathbb{E}\{K^{(2)}_{\theta(0)}\} = \begin{bmatrix} \langle \Phi_1, \Phi_1 \rangle & \langle \Phi_2, \Phi_1 \rangle & \cdots \\ \vdots & \ddots & \\ \vdots & & \end{bmatrix} \text{ is a Gram Matrix}$$

Theorem: (Gram)

For a Gram matrix $A = B^T B$, the following are equivalent:

(1) $A > 0$ — A is PD (2) $\det A > 0$ (3) $\mathrm{vol}(\text{Parallelepiped}(\{\vec{B}_i\}))^2 > 0$ — rows that generate A (4) All rows $\{B_i\}$ linearly independent

We want to show that $\{\Phi_\alpha\}_{\alpha=1}^m$ are linearly independent in $\mathcal{H}$, as this will give us that $\mathbb{E}\{K^{(L)}_{\theta(0)}\} > 0$ by the above theorem.

As usual, suppose that

$$\sum_{\alpha=1}^m c_\alpha \Phi_\alpha = 0 \text{ in } \mathcal{H} \text{ for some } c_\alpha\text{'s.}$$

We want to show that this implies $c_\alpha = 0 \ \forall \alpha$. Now,

$$\sum_{\alpha=1}^m c_\alpha \Phi_\alpha = 0 \text{ in } \mathcal{H} \iff \forall f \in \mathcal{H}, \ \sum_{\alpha=1}^m c_\alpha \langle \Phi_\alpha, f\rangle_{\mathcal{H}} = 0$$

$$\iff \forall f \in \mathcal{H} \quad \sum_{\alpha=1}^m c_\alpha \mathbb{E}\{\sigma(W_{x_\alpha}) f(W)\} = 0$$

Since the Hermite polynomials are orthogonal w.r.t. weight measure e^{-x^2}, we can use them as an orthonormal basis for $\mathcal{H}$ to decompose σ:

$$\sigma(t) = \sum_{j=0}^\infty \frac{\sigma_j}{\sqrt{j!}} H_j(t) \qquad (\sigma \text{ non-poly} \Rightarrow \sigma_k \neq 0 \ \forall k)$$

Let β be arbitrary. Since our assumption holds $\forall f \in \mathcal{H}$, clearly it holds for $\{f_k(W)\}_{k=1}^\infty$, where $f_k(W) = \frac{\sigma_k}{\sqrt{k!}} H_k(W \cdot \vec{x}_\beta)$

The assumption gives

$$\forall k \in \mathbb{N}, \quad 0 = \sum_{\alpha=1}^m c_\alpha \mathbb{E}\left\{\sum_{j=0}^\infty \frac{\sigma_j}{\sqrt{j!}} H_j(W\cdot\vec{x}_\alpha) \frac{\sigma_k}{\sqrt{k!}} H_k(W\cdot\vec{x}_\beta)\right\}$$

$$\overset{\text{Hermite polynomials orthonormal}}{=} \sum_{\alpha=1}^m c_\alpha (\vec{x}_\alpha \cdot \vec{x}_\beta)^k$$

As $k \to \infty$, we find that $(\vec{x}_\alpha \cdot \vec{x}_\beta) \to \delta_{\alpha\beta} \Rightarrow c_\beta = 0$.

This line of reasoning holds for all β, and so all the c_α's are 0.

This means that the $\{\Phi_\alpha\}_{\alpha=1}^m$ are linearly independent in $\mathcal{H}$.

So, by the Gram Theorem,

$$\mathbb{E}\{K^{(L)}_{\theta(0)}\} = \mathrm{Gram}\left(\{\Phi_\alpha\}_{\alpha=1}^m\right) > 0.$$

Since $K^{(L)}_{\theta(0)}$ concentrates well about its expectation and $K_{\theta(0)} \geq K^{(L)}_{\theta(0)}$, we achieve the result that the NTK $K_{\theta(0)}$ is PD at $t=0$.

$\square$

Lecture 10/12 - NTK Sends $\mathcal{L} \to 0$

Recall that we consider the small example

$$z_\alpha^{(2)}(t) = z_\alpha^{(2)}(\Theta(t)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i^{(2)}(t)\, \sigma(w_i^{(1)}(t)\, \vec{x}_\alpha)$$

with $w_i^{(2)} \sim \mathrm{Unif}([-1,1])$, $w_i^{(1)} \sim \mathcal{N}(0, I)$, $\|\sigma\|_\infty, \|\sigma'\|_\infty, \|\sigma''\|_\infty \le 1$.
We inspect gradient descent on MSE

$$\mathcal{L}(\theta) = \frac{1}{2m} \sum_{j=1}^{m} \left(z_{\alpha_j}^{(2)}(\theta) - y_{\alpha_j}\right)^2 \qquad \theta(t+1) = \theta(t) - \eta \vec{\nabla}_\theta \mathcal{L}(\theta(t))$$

Assume the following:

$$\frac{d}{dt}\theta(t) = -\eta \vec{\nabla}_\theta \mathcal{L}(\theta(t)), \qquad W_i^{(2)}(t) = W_i^{(2)}(0)$$

freeze 2nd layer

We still have the NTK

$$\boxed{\underbrace{\left(K_{\theta(t)}\right)_{ij}}_{m \times m} = \underbrace{\left(\vec{\nabla}_\theta z_{\alpha_i}^{(2)}(\theta(t))\right)^T \left(\vec{\nabla}_\theta z_{\alpha_j}^{(2)}(\theta(t))\right)}_{\text{Gram matrix}}}$$

<u>The overall goal</u>: Show that w.h.p. $\mathcal{L}(\theta(t)) \xrightarrow{t \to \infty} 0$

Last time we split this into two subproblems:

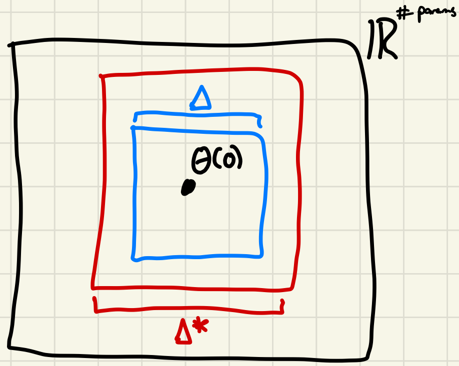
(i) $\exists \lambda_0 > 0$ s.t. $K_{\theta(0)} \succeq \lambda_0 I$ w.h.p. ($K_{\theta(0)}$ is PD, showed this last time)

(ii) $K_{\theta(t)} \succeq \frac{\lambda_0}{2} I \quad \forall t \ge 0$ ($K_{\theta(t)}$ stays PD, show this this time)

In other words, today we want to show

$$\boxed{\forall t \ge 0, \quad \|K_{\theta(t)} - K_{\theta(0)}\|_{op} \le \frac{\lambda_0}{2}}$$

The idea is as follows: (Du et. al.)



(1) $\|\theta - \theta(0)\| \le \Delta(n, m, \lambda_0, \dots)$ ← stay within blue box

$\Rightarrow \|K_\theta - K_{\theta(0)}\|_{op} \le \frac{\lambda_0}{4}$ ← K_θ stays positive

We WTS the implication and that the premises hold.

(2) While $K_{\theta(t)} \ge \frac{\lambda_0}{2} I$, $\mathcal{L}(\theta(t))$ decays exponentially, $\|\frac{d}{dt}\theta(t)\|^2 \approx \mathcal{L}(t)$

We want $\Delta^* \le \Delta$ so that we never leave the box of size Δ so that (1) will give us $K_{\theta(t)} \ge \frac{\lambda_0}{4} I$.

We first show the implication in (1).

Lemma 1: Let $\Delta \in (0, 1]$. If $\forall i$, $\|w_i^{(1)} - w_i^{(1)}(0)\| \le \Delta$, then $\|K_\theta - K_{\theta(0)}\|_{op} \le 2m\Delta$

If params don't change too much, $K_{\theta(t)}$ stays PD.

Proof: We have $(K_\theta)_{ij} = \frac{1}{n}\sum_{k=1}^{n} w_k^{(2)}(0)^2 \underbrace{(\vec{x}_{\alpha_i} \cdot \vec{x}_{\alpha_j})}_{|\cdot| \le 1} \sigma'(w_k^{(1)} \cdot \vec{x}_{\alpha_i}) \sigma'(w_k^{(1)} \vec{x}_{\alpha_j})$

Note that because σ' is bounded, we see that $W \in \mathbb{R}^{n_0} \mapsto \sigma'(W\vec{x}_{\alpha_i})\sigma'(W\vec{x}_{\alpha_j})$ is 2-Lipschitz.

To see this,

bounded differences

$$\begin{aligned}
&\left|\sigma'(W\vec{x}_{\alpha_i})\sigma'(W\vec{x}_{\alpha_j}) - \sigma'(\bar{W}\vec{x}_{\alpha_i})\sigma'(\bar{W}\vec{x}_{\alpha_j})\right| \\
&= \left|(\sigma'(W\vec{x}_{\alpha_i}) - \sigma'(\bar{W}\vec{x}_{\alpha_i}))\sigma'(W\vec{x}_{\alpha_j}) + \sigma'(\bar{W}\vec{x}_{\alpha_i})(\sigma'(W\vec{x}_{\alpha_j}) - \sigma'(\bar{W}\vec{x}_{\alpha_j}))\right| \\
&\le 2\|W - \bar{W}\|
\end{aligned}$$

largest diff.

Thus, $\|K_\theta - K_{\theta(0)}\|_\infty \le 2\Delta$. Lastly, since $A \in \mathbb{R}^{m \times m}$, $\|A\|_{op} \le m\|A\|_\infty$.

$\Rightarrow \|K_\theta - K_{\theta(0)}\|_{op} \le 2m\Delta$. $\square$

Corollary: If $\|w_i^{(1)}(t) - w_i^{(1)}(0)\| \le \frac{\lambda_0}{8m}$ $\forall t \ge 0$, $\Rightarrow K_{\theta(t)} \ge \frac{\lambda_0}{4} I$ $\forall t \ge 0$.

This tells us that we wish to set $\boxed{\Delta \equiv \frac{\lambda_0}{8m}}$

This proves the implication of (1): if $\|W_i^{(1)}(t) - W_i^{(1)}(0)\| \leq \frac{\lambda_0}{8m}$ $\forall t \geq 0$

$$\Rightarrow K_{\theta(t)} \geq \frac{\lambda_0}{4} I \quad \forall t \geq 0.$$

Now, all that is left to show is that

$$\|\theta(t) - \theta(0)\| \leq \int^\infty \|\tfrac{d}{ds}\theta(s)\| ds \leq \Delta^* \quad \text{for some } \Delta^* \leq \Delta = \frac{\lambda_0}{8m}$$

With this, we can use the Corollary to show that $K_{\theta(t)}$ stays P.D.
We now show the premises.

Lemma 2: Fix $t \geq 0$ and suppose that $\forall s < t$, $K_{\theta(s)} \geq \frac{\lambda_0}{2} I$ (*)

Then $\forall s < t$, $\|W_i^{(1)}(s) - W_i^{(1)}(0)\| \leq \Delta^* = \dfrac{2m\, \mathcal{L}(0)^{\frac{1}{2}}}{\lambda_0 n^{\frac{1}{2}}}$

If K_θ stays PD, then params don't change too much

Proof: We have $\|W_i^{(1)}(s) - W_i^{(1)}(0)\| = \left\| \int_0^s \frac{d}{d\tau} W_i^{(1)}(\tau) d\tau \right\| \leq \int_0^\infty \left\| \frac{d}{d\tau} W_i^{(1)}(\tau) d\tau \right\|$

For fixed i, τ, we compute

$$\frac{d}{d\tau} W_i^{(1)}(\tau) = -\eta\, \partial_{W_i^{(1)}} \mathcal{L}(\theta(\tau)) = -\eta\, \partial_{W_i^{(1)}} \left[\frac{1}{2m} \sum_{j=1}^m \left(z_{\alpha_j}^{(2)}(\tau) - y_{\alpha_j} \right)^2 \right]$$

$$= \frac{-\eta}{m} \sum_{j=1}^m \left(z_{\alpha_j}^{(2)}(\tau) - y_{\alpha_j} \right) \left(\partial_{W_i^{(1)}} z_{\alpha_j}^{(2)}(\tau) \cdot \vec{x}_{\alpha_j} \right)$$

$$= \frac{-\eta}{\sqrt{n}\, m} \sum_{j=1}^m \left(z_{\alpha_j}^{(2)}(\tau) - y_{\alpha_j} \right) \underbrace{W_i^{(2)}(0)\, \sigma'\left(W_i^{(1)}(\tau) \vec{x}_{\alpha_j} \right)}_{|\cdot| \text{ bounded by } 1}$$

$$\Rightarrow \left\| \frac{d}{d\tau} W_i^{(1)}(\tau) \right\| \leq \frac{\eta}{m} \sum_{j=1}^m \left| z_{\alpha_j}^{(2)}(\tau) - y_{\alpha_j} \right|$$

We can use the Power-Mean inequality:

$$\boxed{\forall p < p', \quad \left(\frac{1}{m} \sum_{j=1}^m a_j^{p} \right)^{\frac{1}{p}} \leq \left(\frac{1}{m} \sum_{j=1}^m a_j^{p'} \right)^{\frac{1}{p'}}}$$

$\Rightarrow \| \partial_{W_i^{(1)}} \mathcal{L}(\theta(\tau)) \| \leq \frac{\eta}{\sqrt{n}} \mathcal{L}(\theta(\tau))^{\frac{1}{2}}$. Therefore,

apply part (2) because (*)

$$\|W_i^{(1)}(s) - W_i^{(1)}(0)\| \leq \frac{\eta}{m} \int_0^\infty e^{-\frac{\eta \lambda_0 \tau}{2m}} d\tau\, \mathcal{L}(0)^{\frac{1}{2}} = \frac{\eta}{\sqrt{n}} \cdot \frac{2m}{\lambda_0 \eta} \mathcal{L}(0)^{\frac{1}{2}}$$

$$= \frac{2m\, \mathcal{L}(0)^{\frac{1}{2}}}{\lambda_0 n^{\frac{1}{2}}}.$$

$\square$

This tells us to set $\boxed{\Delta^* = \dfrac{2m\, \mathcal{L}(0)^{\frac{1}{2}}}{\lambda_0 n^{\frac{1}{2}}}}$

Here is a quick proof of (2), which we used above.

$$\boxed{\text{Recall: } \frac{d}{dt}\left(z^{(2)}(t) - y\right) = -\frac{\eta}{m} K_{\theta(t)}\left(z^{(2)}(t) - y\right)}$$

To see (2) (K₀ PD ⇒ $\mathcal{L}(t)$ exponential decay), note that

MSE

$$\mathcal{L}(t) = \frac{1}{2m}\left(z^{(2)}(t) - y\right)^T\left(z^{(2)}(t) - y\right) \Rightarrow \frac{d}{dt}\mathcal{L}(t) = -\frac{\eta}{m}\left(z(t) - y\right)^T K_{\theta(t)}\left(z(t) - y\right)$$

$$\underset{(K_\theta \geq \frac{\lambda_0}{2} I)}{\leq} -\frac{\eta}{m}\lambda_0 \mathcal{L}(t)$$

$$\Rightarrow \mathcal{L}(t) \leq e^{-\frac{\eta\lambda_0}{m}t}\mathcal{L}(0)$$

At this point, we proved that

Lemma 1 $\|w_i^{(1)}(t) - w_i^{(1)}(0)\| \leq \Delta \Rightarrow K_{\theta(t)} \geq \frac{\lambda_0}{2} I$ &

Lemma 2 $\forall s < t,\ K_{\theta(s)} \geq \frac{\lambda_0}{2} I \Rightarrow \|w_i^{(1)}(s) - w_i^{(1)}(0)\| \leq \Delta^*$

Suppose that $\Delta^* < \Delta \iff \frac{2m\mathcal{L}(0)^{\frac{1}{2}}}{\lambda_0 n^{\frac{1}{2}}} < \frac{\lambda_0}{2m} \iff n > \frac{16m^4\mathcal{L}(0)}{\lambda_0^4}$

Define $t_K = \inf\{t > 0 \text{ s.t. } K_{\theta(t)} \leq \frac{\lambda_0}{2} I\}$ ← first t that NTK isn't PD enough

$t_\Delta = \inf\{t > 0 \text{ s.t. } \exists j \in \{1, \dots, n\} \text{ s.t. } \|w_j^{(1)}(t) - w_j^{(1)}(0)\| > \Delta^*\}$

first t s.t. weights grow a lot

$t^* = \min\{t_K, t_\Delta\}$

We claim that <u>t^* must be ∞.</u>

?

<u>Proof:</u> Suppose BWOC that $t^* < \infty$.

<u>Case 1:</u> $t^* = t_\Delta \leq t_K$

Then, $\forall t < t^*$, we have $\|w_i^{(1)}(t) - w_i^{(1)}(0)\| \leq \Delta^* < \Delta \overset{\text{Lemma 1}}{\Longrightarrow} K_{\theta(t)} \geq \frac{\lambda_0}{2} I$

$\rightarrow\leftarrow$ by definition of t_K.

<u>Case 2:</u> $t^* = t_K \leq t_\Delta$

Then, $\forall t < t^*$, we have $K_{\theta(t)} \geq \frac{\lambda_0}{2} I \overset{\text{Lemma 2}}{\Longrightarrow} \|w_i^{(1)}(t) - w_i^{(1)}(0)\| \leq \Delta^* < \Delta$

$\rightarrow\leftarrow$ by our definitions.

Thus, $t^* = \infty$.

□

So, we showed that the weights always stay within Δ and therefore that the NTK is always PD. Applying part (2) as $t \to \infty$, we have shown that $\mathcal{L}(t) \to 0$!

Lecture 10/31 - Kernels

Def: Let $\Omega \subset \mathbb{R}^d$. A Kernel on Ω is $K: \Omega \times \Omega \to \mathbb{R}$ s.t. $\forall \vec{x}_1, \ldots, \vec{x}_N, \vec{y} \in \Omega$, $\forall a_1, \ldots, a_N \in \mathbb{R}$

- $K(\vec{x}, \vec{y}) = K(\vec{y}, \vec{x})$
- K is "positive" $\Leftrightarrow$ $\sum_{i,j=1}^{N} a_i a_j K(\vec{x}_i, \vec{x}_j) > 0$ if $\|\vec{a}\| \neq 0$

We can think of K as a infinite analog of positive definite matrices.

Ex/ 1) If Ω finite $(\vec{x}_1, \ldots, \vec{x}_k \in \mathbb{R}^d)$, $K \in \mathbb{R}^{k \times k}$, $K(\vec{x}_i, \vec{x}_j) = K_{ij}$

$\Rightarrow$
- K is symmetric
- $\sum_{i,j=1}^{N} a_i a_j K(\vec{x}_i, \vec{x}_j) = \vec{a}^T K \vec{a} > 0$ if $\vec{a} \neq 0$

2) $\Omega = \mathbb{R}^d$, $K(\vec{x}, \vec{y}) = \langle \vec{x}, \vec{y} \rangle$
- dot product is commutative
- $\sum_{i,j=1}^{N} a_i a_j K(\vec{x}_i, \vec{x}_j) = \|\sum_i a_i \vec{x}_i\|^2$

3) $\Omega \in \mathbb{R}^d$, $K(\vec{x}, \vec{y}) = e^{-\|x-y\|^2 / 2\sigma^2}$

The general case is defined via feature maps!

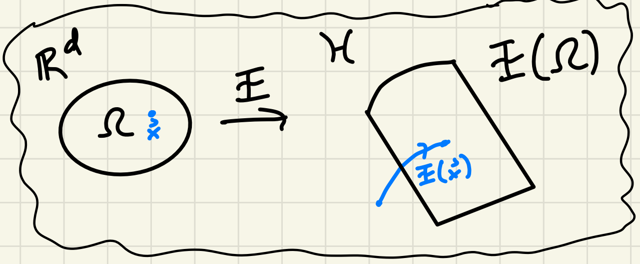
Def: The feature map $\Phi: \Omega \to \mathcal{H}$ (arbitrary Hilbert space) is given by

$$K(\vec{x}, \vec{y}) = \langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle_{\mathcal{H}}$$

where

$$\underset{\Omega}{\underset{\Uparrow}{\vec{x}}} \overset{\Phi}{\longmapsto} \langle \psi_1(\vec{x}), \psi_2(\vec{x}), \ldots \rangle$$

(vector in $\mathcal{H}$ of coefficients in the ONB)



where $\{\psi_j\}$ is an ONB of $\mathcal{H}$.

Theorem: Every kernel comes from a feature map.

Proof: (Ω compact, K is C^0)

Fix $\mu \in P(\Omega)$ as a subset, and let $T_K : L^2(\Omega, \mu) \circlearrowleft$ be defined s.t.

any measure on Ω

$$(T_K f)(\vec{x}) = \int_\Omega K(\vec{x}, \vec{y}) f(\vec{y}) \, d\mu(\vec{y})$$

Note that T_K is compact. We can apply the spectral theorem:

$$T_K = \sum_{j=0}^{\infty} \lambda_j \varphi_j \varphi_j^T \quad \text{for an orthonormal basis } \{\varphi_j\}$$

Moreover, $K(\vec{x}, \cdot) \in L^2(\Omega, \mu)$

$$\Rightarrow \forall \vec{x} \in \Omega, \quad K(\vec{x}, \vec{y}) = \sum_{j=0}^{\infty} a_j(\vec{x}) \varphi_j(\vec{y}) \qquad (\{\varphi_j\} \text{ ONB})$$

Further,

$$\lambda_j \varphi_j(\vec{x}) = (T_K \varphi_j)(\vec{x}) = \int_\Omega K(\vec{x}, \vec{y}) \varphi_j(\vec{y}) \, d\mu(\vec{y})$$

$$= \sum_{k=0}^{\infty} a_k(\vec{x}) \underbrace{\int_\Omega \varphi_k(\vec{y}) \varphi_j(\vec{y}) \, d\mu(\vec{y})}_{\delta_{kj}}$$

$$= a_j(\vec{x})$$

$$\Rightarrow K(\vec{x}, \vec{y}) = \sum_{j=0}^{\infty} \lambda_j \varphi_j(\vec{x}) \varphi_j(\vec{y}) = \langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle_{\mathcal{H}},$$

where $\Phi(\vec{x}) = \langle \underbrace{\sqrt{\lambda_j} \varphi_j(\vec{x})}_{\psi_j(\vec{x})}, \; j = 0, 1, 2, \dots \rangle$

So, $\mathcal{H} = \ell_2$ with the ONB $\{\psi_j\}$.

$\square$

Def Given kernel K, the reproducing kernel Hilbert space (RKHS) is

$$\mathcal{H}_K = T_K^{\frac{1}{2}} L^2(\Omega, \mu) = \left\{ \sum_{j=0}^{\infty} a_j \sqrt{\lambda_j} \varphi_j \;\middle|\; \vec{a} \in \ell_2 \right\}$$

$$\Rightarrow \langle f, g \rangle_{\mathcal{H}_K} = \langle T^{-1} f, g \rangle_{L^2} = \langle T^{-\frac{1}{2}} f, T^{-\frac{1}{2}} g \rangle_{L^2}$$

Properties of RKHS:

① $K(\vec{x}, \cdot) \in \mathcal{H}_K$ s.t.

K in RKHS

$$\|K(\vec{x}, \cdot)\|^2_{\mathcal{H}_K} = \left\langle \sum_{j=0}^{\infty} \lambda_j \varphi_j(\vec{x}) \varphi_j(\cdot), \sum_{k=0}^{\infty} \lambda_k \varphi_k(\vec{x}) \varphi_k(\cdot) \right\rangle_{\mathcal{H}_K}$$

$$= \sum_{j,k=0}^{\infty} \varphi_j(\vec{x}) \varphi_k(\vec{x}) \lambda_j \lambda_k \langle \varphi_j, \varphi_k \rangle_{\mathcal{H}_K}$$

$$= \sum_{j,k=0}^{\infty} \varphi_j(\vec{x})\varphi_k(\vec{x})\,\lambda_j\lambda_k\,\lambda_j^{-\frac{1}{2}}\lambda_k^{-\frac{1}{2}}\langle \varphi_j,\varphi_k\rangle_{L^2}$$

$$= \sum_{j=0}^{\infty} \lambda_j \varphi_j(\vec{x})^2 = K(\vec{x},\vec{x})$$

"reproducing" property
the equates point evaluation to $\langle \cdot,\cdot \rangle_{\mathcal{H}_K}$

② $\forall f \in \mathcal{H}_K$, $\quad \langle f(\cdot), K(\vec{x},\cdot)\rangle_{\mathcal{H}_K} = \langle f(\cdot), \Phi(\vec{x})\rangle_{\mathcal{H}_K} \equiv f(\vec{x})$

So, $f \mapsto f(\vec{x})$ is <u>bounded</u> (linear functionals in Hilbert spaces are bounded)

Note: you are an RKHS if and only if point evaluation is bounded.

③ $\langle K(\vec{x},\cdot), K(\vec{y},\cdot)\rangle_{\mathcal{H}_K} = K(\vec{x},\vec{y})$

④ $\mathcal{H}_K$ is the closure of $\left\{ \sum_{j=1}^{N} a_j K(\vec{x}_j,\cdot) \right\}$
with respect to

$$\langle K(\vec{x},\cdot), K(\vec{x}',\cdot)\rangle_{\mathcal{H}_K} = K(\vec{x},\vec{x}')$$

This means that we can describe $\mathcal{H}_K$ via a dataset and function evaluations $\{f(\vec{x}_j)\}$.

To recap, we saw an equivalence

$$\begin{tikzcd}
\overset{\text{kernel}}{K} \arrow[r] \arrow[d] & \Phi_K : \Omega \to \mathcal{H}_K \arrow[dl, Leftrightarrow] \\
\mathcal{H}_K &
\end{tikzcd}$$

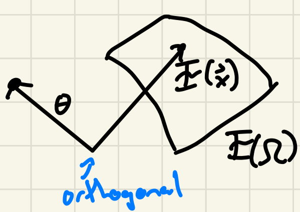
$$K(\vec{x},\vec{y}) = \sum_{j=0}^{\infty} \psi_j(\vec{x})\,\psi_j(\vec{y}) \quad \text{where } \{\psi_j\} \text{ ONB for } \mathcal{H}_K.$$

<u>ML Applications</u>

Given $\Phi = (\psi_0, \psi_1, \dots)$ with $\psi_j : \Omega \to \mathbb{R}$, we wish to find the function

$$f(\vec{x};\theta) = \sum_{j=0}^{\infty} \theta_j \psi_j(\vec{x}) = \langle \theta, \Phi(\vec{x})\rangle \quad \text{that minimizes}$$

$$\sum_{i=1}^{m} \ell(f(\vec{x}_i,\theta), y_i) + \frac{\lambda}{2}\|\theta\|_2^2$$



Option 1: $\ell(a,b) = \frac{1}{2}(a-b)^2 \Rightarrow \mathcal{L}_\lambda(\theta) = \frac{1}{2}\|Y - \Phi^T\theta\|^2 + \frac{\lambda}{2}\|\theta\|^2$

Yongwei method

We have
$$\vec{\nabla}_\theta \mathcal{L}_\lambda = -\Phi(Y - \Phi^T\theta) + \lambda\theta$$
and so
$$\vec{\nabla}_\theta \mathcal{L}_\lambda = 0 \iff \theta = \underbrace{(\Phi\Phi^T + \lambda I)^{-1}}_{\text{shitty to invert, } \in \mathbb{R}^{n\text{features} \times n\text{features}}}\Phi Y$$

Option 2: Let's write $K(\vec{x},\vec{y}) = \langle \Phi(\vec{x}), \Phi(y)\rangle_{\ell_2}$ and deal with things in $\mathcal{H}_K = \text{span}\{\Phi\}$. (So, Φ is ONB for $\mathcal{H}_K$).

Kernel method

We have
$$f(\vec{x};\theta) = \sum_{j=0}^{\infty} \theta_j \varphi_j(\vec{x}) \in \mathcal{H}_K, \quad \|\theta\|_2^2 = \|f\|_{\mathcal{H}_K}^2$$
$$\Rightarrow f_* = \underset{f\in\mathcal{H}_K}{\text{argmin}} \underbrace{\sum_{i=1}^{m} \ell(f(\vec{x}_i), y_i)}_{\substack{\text{only depends on} \\ \{f(\vec{x}_i)\} = \{\langle K(\vec{x}_i,\cdot), f\rangle_{\mathcal{H}_K}\}_{i=1}^m}} + \frac{\lambda}{2}\|f\|_{\mathcal{H}_K}^2$$

Let us consider the (finite-dim) subspace of $\mathcal{H}_K$ along the training dataset given by $\Pi_X : \mathcal{H}_K \to \text{Span}\{K(\vec{x}_i,\cdot)\}_{i=1}^m$

and so $\sum_{i=1}^{m} \ell(f(\vec{x}_i), y_i)$ depends only on $\Pi_X f$.

The minimization problem is:
$$f_* = \underset{f\in\mathcal{H}_K}{\text{argmin}}\ \mathcal{L}(\Pi_X f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}_K}^2$$

However, $\|f\|_{\mathcal{H}_K}^2 = \|\Pi_X f\|_{\mathcal{H}_K}^2 + \|\Pi_X^\perp f\|_{\mathcal{H}_K}^2$.

Since $\mathcal{L}$ doesn't see $\Pi_X^\perp f$ (it only sees function eval. at data points),
$$f_* = \underset{f\in \underbrace{\text{Span}\{K(\vec{x}_i,\cdot)\}_{i=1}^m}_{\Pi_X(\mathcal{H}_K)}}{\text{argmin}} \sum_{i=1}^{m} \ell(f(\vec{x}_i), y_i) + \frac{\lambda}{2}\|f\|_{\mathcal{H}_K}^2$$

We parameterize $f(x) = \sum_{j=1}^{m} \underline{\underline{a_j}}\, K(\vec{x}_j,\cdot)$ and solve for $\ell(a,b) = \frac{1}{2}(a-b)^2$

$$\Rightarrow f(\vec{x}_i) = \left\langle K(\vec{x}_i, \cdot), \sum_{j=1}^{m} a_j K(\vec{x}_j, \cdot) \right\rangle_{\mathcal{H}_K} = \sum_{j=1}^{m} a_j K(\vec{x}_i, \vec{x}_j) = K\vec{a}$$

$$K = \begin{bmatrix} K(\vec{x}_1, \vec{x}_1) & K(\vec{x}_2, \vec{x}_1) & \cdots \\ \vdots & \ddots & \end{bmatrix}$$

$$\Rightarrow \|f\|^2_{\mathcal{H}_K} = \left\langle \sum_{j=1}^{m} a_j K(\vec{x}_j, \cdot), \sum_{i=1}^{m} a_i K(\vec{x}_i, \cdot) \right\rangle_{\mathcal{H}_K}$$

$$= \sum_{i,j=1}^{m} a_i a_j K(\vec{x}_i, \vec{x}_j) = \vec{a}^T K \vec{a}$$

$$\Rightarrow \vec{a}_* = \underset{\vec{a}}{\operatorname{argmin}} \ \frac{1}{2}\|Y - K\vec{a}\|_2^2 + \frac{\lambda}{2} \underbrace{\vec{a}^T K \vec{a}}_{\|\vec{a}\|^2_{K^{-1}}}$$

So, $\vec{\nabla}_a = -K(Y - K\vec{a}) + \lambda K \vec{a} = 0 \iff \vec{a}_* = (K + \lambda I)^{-1} Y$

$$\iff f_* = K\vec{a}_* = K(\underbrace{K + \lambda I}_{\in \mathbb{R}^{\#\text{data} \times \#\text{data}}})^{-1} Y$$

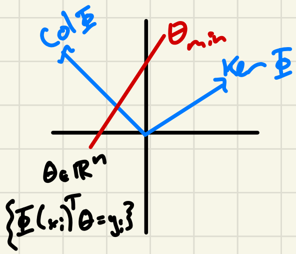
To sum, kernel methods for a given kernel K yield:

* Φ_K - feature map
* $\mathcal{H}_K$ - RKHS (Representer Theorem)
* f_K - Gaussian Process on Ω with
  $\mathbb{E}\{f_K(\vec{x})\} = 0$,
  $\mathrm{Cov}(f_K(\vec{x}), f_K(\vec{y})) = K(\vec{x}, \vec{y})$
* DPP X_K on Ω

Lecture 11/2 - Quadratic Models

Last time - We considered linear models

$$z(x;\theta) = \Phi^T(x)\theta = \sum_{j=0}^{n} \theta_j \psi_j(x)$$



All solutions to $\mathcal{L}(\theta) = \frac{1}{2}\sum_{i=1}^{n} (z(x_i;\theta) - y_i)^2 = 0$

are solutions of $\Phi\Phi^T\theta = \Phi Y$

Today we study quadratic models

$\Psi(x)$ symmetric

$$\boxed{z(x;\theta) = \Phi(x)^T\theta + \frac{\varepsilon}{2}\theta^T \Psi(x)\theta} = \sum_{j=0}^{n} \theta_j \psi_j(x) + \frac{\varepsilon}{2}\sum_{j_1,j_2=0}^{n} \theta_{j_1}\theta_{j_2}\psi_{j_1,j_2}(x)$$

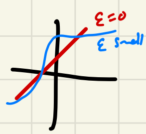
We motivate this via Taylor expansion

$$f(x;\theta) = f(x;0) + \nabla_\theta f(x;0)^T\theta + \frac{1}{2}\theta^T H_\theta f(x;0)\,\theta + \dots$$

With the same loss $\mathcal{L}_A(\theta) = \sum_{\alpha\in A} \frac{1}{2}(y_\alpha - z(x_\alpha;\theta))^2$

and the goal to find minima of $\mathcal{L}_A(\theta)$ to 1^{st} order in ε.

Notation: We define $\nabla_\theta z(x;\theta) \equiv \Phi^E(x;\theta) = \Phi(x) + \varepsilon\Psi(\theta)$



To solve $\nabla_\theta \mathcal{L}_A(\theta) = 0$ to first order in ε, we have

$$\nabla_\theta \mathcal{L}_A(\theta) = \sum_{\alpha\in A} \Phi^E(x_\alpha;\theta)\,(z(x_\alpha;\theta) - y_\alpha)$$

$$(1)\quad = \sum_{\alpha\in A} (\Phi(x_\alpha) + \varepsilon\Psi(x_\alpha)\theta) \times (\Phi(x_\alpha)^T\theta + \frac{\varepsilon}{2}\theta^T\Psi(x_\alpha)\theta - y_\alpha)$$

Let's write $\theta_* = \overset{\text{free}}{\theta^F} + \varepsilon \overset{\text{interacting}}{\theta^I} + O(\varepsilon^2)$, where $\Phi\Phi^T\theta^F = \Phi Y$

So, $0 = \nabla_\theta \mathcal{L}_A(\theta)$ gives

$$0 = \sum_{\alpha\in A}\left(\Phi(x_\alpha) + \varepsilon\Phi(x_\alpha)(\theta^F + \varepsilon\theta^I)\right) * \left(\Phi(x_\alpha)^T(\theta^F+\varepsilon\theta^I) + \frac{\varepsilon}{2}(\theta^F)^T\Phi(x_\alpha)\theta^F - y_\alpha\right)$$

zeroth order terms ε^0 $\Rightarrow$ $0 = \varepsilon^0\left[\sum_{\alpha\in A}\overbrace{\Phi(x_\alpha)\Phi(x_\alpha)^T}^{\text{rank-1 matrices}}\theta^F - \Phi(x_\alpha)y_\alpha\right]$ ← this is 0 because $\Phi\Phi^T\theta^F = \Phi Y$

first order terms ε^1 $\Rightarrow$ $0 = \varepsilon^1\left[\sum_{\alpha\in A}\Phi(x_\alpha)\theta^F(\Phi(x_\alpha)^T\theta^F - y_\alpha) + \sum_{\alpha\in A}\frac{1}{2}\Phi(x_\alpha)(\theta^F)^T\Phi(x_\alpha)\theta^F + \sum_{\alpha\in A}\Phi(x_\alpha)\Phi(x_\alpha)^T\theta^I\right]$

equal

$$\Rightarrow \sum_{\alpha\in A}\Phi(x_\alpha)\theta^F y_\alpha = \sum_{\alpha\in A}\Phi(x_\alpha)\theta^F\underbrace{\Phi(x_\alpha)^T\theta^F}_{y_\alpha\text{ prediction linear model made}} + \sum_{\alpha\in A}\frac{1}{2}\Phi(x_\alpha)(\theta^F)^T\Phi(x_\alpha)\theta^F + \underbrace{\sum_{\alpha\in A}\Phi(x_\alpha)\Phi(x_\alpha)^T}_{\Phi\Phi^T}\theta^I$$

$$\Rightarrow \boxed{\theta^I = -\underbrace{(\Phi\Phi^T)^+}_{\text{pseudo-inverse}}\sum_{\alpha\in A}\frac{1}{2}\Phi(x_\alpha) * (\theta^F)^T\Phi(x_\alpha)\theta^F}$$

elementwise mult.

Interpretation:

(1) We can write $(\theta^F)^T\Phi(x_\alpha)\theta^F = \langle\theta^F, \Phi(x_\alpha)\theta^F\rangle$ which is $\underset{\Phi\text{ not pos. def.}}{\underline{\text{almost}}}$ $\|\theta^F\|^2_{\Phi(x_\alpha)}$

(2) Also, $\theta^I = (\Phi\Phi^T)^+(\Phi Y^I)$ ← transformed Y via these coefficients

So, if we had changed $\underbrace{Y \mapsto Y + \Phi Y^I}_{\text{deform } Y \text{ with useful features – feature learning!}}$

and solved least-squares with a linear model, we would get the same predictions

$\Longleftrightarrow$

GD on nonlinear models learns label features to an linear model on

(3) Note that $(\Phi\Phi^T)\theta^I = -\sum_{\alpha\in A} \frac{1}{2}\Psi(x_\alpha)(\theta^F)^T\Psi(x_\alpha)\theta^F$

only determines θ^I on $\mathrm{span}\{\Phi(x_\alpha)\}$; so, it is unclear what happens to $\theta^I_\perp$ ← this may depend on optimization method and allow weird things to happen

When we do gradient flow (continuous GD),

$$\frac{d}{dt}\theta_t = -\eta \nabla_\theta \mathcal{L}_A(\theta_t)$$

Recall the effective features $\Phi^E(x;\theta) = \Phi(x) + \varepsilon \underbrace{\Psi(x)\theta}_{\nabla_\theta z(x;\theta)}$

We can write

$$\frac{d}{dt}\Phi^E(x;\theta_t) = \varepsilon\Psi(x)\underbrace{\frac{d}{dt}\theta^F_t}_{\in \mathrm{col}(\Phi)} + O(\varepsilon^2)$$

Interpretations:

- Φ^E changes!
- $\frac{d}{dt}\theta^F_t \in \mathrm{col}(\Phi) \Rightarrow \frac{d}{dt}\Phi(x;\theta) \in \mathrm{span}\{\Psi(x)\Phi(x_\alpha), \alpha\in A\}$

Moreover, since θ^F_t solves the linear model,

$$\frac{d}{dt}\theta^F_t = \frac{d}{dt}(\theta^F_t - \theta_*) = -\eta\Phi\Phi^T\underbrace{(\theta^F_t - \theta_*)}_{\text{vector of residuals}}$$

$$\Rightarrow \theta^F_t - \theta_* = e^{-\eta t\Phi\Phi^T}(\theta^F_0 - \theta_*)$$

Therefore,

$$\frac{d}{dt}\Phi^E(x;\theta_t) = \varepsilon\Psi(x)(-\eta\Phi\Phi^T)(\theta^F_t - \theta_*)$$
$$= \varepsilon\Psi(x)(-\eta\Phi\Phi^T)e^{-\eta t\Phi\Phi^T}(\theta^F_0 - \theta_*)$$

$$\Rightarrow \Phi^E(x;\theta_t) - \Phi^E(x;\theta_0) = \varepsilon\Psi(x)(I - e^{-\eta t\Phi\Phi^T})_*(\theta^F_0 - \theta_*)$$

At $t\to\infty$,

$$\boxed{\Phi^E(x;\theta_\infty) = \Phi^E(x;\theta_0) + \varepsilon\Psi(x)(\theta^F_0 - \theta_*)}$$

To recap:

① We got the NTK $\Phi^{\varepsilon}(\Phi^{\varepsilon})^T$ @ all times

② Formula for what happens to θ_* to leading order in ε on $\mathrm{span}\{\Phi\}$

what happens to $\theta_\perp$?

Next,
$$\frac{d}{dt}\theta_t = \underbrace{-\eta\nabla_\theta \mathcal{L}_A(\theta_t)}_{\text{grad flow}} = -\eta\sum_{\alpha\in A}\Phi^{\varepsilon}(x_\alpha;\theta_t)\times\left(z(x_\alpha;\theta_t)-y_\alpha\right)$$

Therefore,
$$\frac{d}{dt}\theta_t = -\eta\sum_{\alpha\in A}\left(\Phi(x_\alpha)+\varepsilon\Psi(x_\alpha)\left(I-e^{-\eta t\Phi\Phi^T}\right)(\theta_0^F-\theta_*)\right)$$
$$\times\left(\Phi(x_\alpha)^T\theta_t+\frac{\varepsilon}{2}\theta_t^T\Psi(x_\alpha)\theta_t-y_\alpha\right)$$

$\frac{d}{dt}\theta_*^F$ cancels w/ $\Phi\Phi^T\theta-\Phi y$

$$\Rightarrow \frac{d}{dt}\theta_t^I = -\eta\sum_{\alpha\in A}\Phi(x_\alpha)\Phi(x_\alpha)^T\theta_t^I + \Psi(x_\alpha)\left(I-e^{-\eta t\Phi\Phi^T}\right)(\theta_0^F-\theta_*)\left(\Phi(x_\alpha)^T\theta_t^F-y_\alpha\right)$$
$$+\frac{1}{2}\Phi(x_\alpha)(\theta_t^F)^T\Psi(x_\alpha)\theta_t^F$$

Projecting onto the orthogonal complement of $\mathrm{span}\{\Phi(x_\alpha)\}$

$$\Rightarrow \frac{d}{dt}\theta_{t\perp}^I = -\eta\sum_{\alpha\in A}\Psi^{\perp}(x_\alpha)\left(I-e^{-\eta t\Phi\Phi^T}\right)(\theta_0^F-\theta_*)\left(\Phi(x_\alpha)^T\theta_t^F-y_\alpha\right)$$

Lecture 11/7

Recall last time: gradient flow on quadratic models

$$(\#)\quad \frac{d}{dt}\theta(t) = -\nabla_\theta \mathcal{L}(\theta(t)) \qquad\qquad z(x;\theta) = \Phi^T(x)\,\theta + \frac{\epsilon}{2}\theta^T \Psi(x)\,\theta$$

where $\mathcal{L}_A(\theta) = \sum_{\alpha\in A} \frac{1}{2}\left(z(x_\alpha;\theta) - y_\alpha\right)^2$.

Note that gradient flow $(\#)$ is the limit $\eta \to 0$ of gradient descent

$$(\#\#)\quad \theta(t+1) = \theta(t) - \eta\nabla_\theta \mathcal{L}(\theta(t))$$

We have seen that η small vs. η large can make qualitative differences.

Today: We consider "large" η in quadratic approximations to 1-layer ReLU nets:

$$z(x;\theta) = \sum_{i=1}^m \frac{v_i}{\sqrt{m}}\,\sigma\!\left(\frac{u_i^T}{\sqrt{d}}x\right),$$

σ: ReLU
$\vec{u} \in \mathbb{R}^d$ input
$\vec{v} \in \mathbb{R}^{d\times m}$ weights

Writing the quadratic approx. (evaluated at $t=0$)

$$z(x;\theta) \approx z^0 + \sum_{i=1}^m \nabla^0_{u_i}\underbrace{\left(u_i - u_i(0)\right)}_{\in\mathbb{R}^d} + \partial^0_{v_i}\underbrace{\left(v_i - v_i(0)\right)}_{\in\mathbb{R}} + \underbrace{\left(u_i - u_i(0)\right)}_{\in\mathbb{R}^d}^T \overbrace{H_i^0}^{\in\mathbb{R}^{d\times 1}}\underbrace{\left(v_i - v_i(0)\right)}_{\in\mathbb{R}}$$

where $\quad z^0 = z(x;\theta(0)), \quad \nabla^0_{u_i} = \nabla_{u_i} z(x;\theta(0)), \quad \partial^0_{v_i} = \partial_{v_i} z(x;\theta(0))$

$$H_i^0 = \partial_{v_i}\nabla_{u_i} z(x;\theta(0))$$

no second derivatives $\partial^2_{u_i}$ or $\partial^2_{v_i}$ because ReLU assumption!

Goal: Following Zhu et. al, we consider one training datapoint (x,y) and show that the "catapult phase" occurs.

Explicitly,

$$\lambda(u,v) = \|\nabla_\theta z(x;\theta)\|^2 = \sum_{i=1}^m \|\nabla_{u_i} z(x;\theta)\|^2 + \left(\partial_{v_i} z(x;\theta)\right)^2$$

NTK $\cong$ is 1×1 and $\lambda(t) = \lambda(u(t),v(t)), \quad \mathcal{L}(t) = \mathcal{L}(u(t),v(t))$

We have the following "phase diagram" for optimization:

"Thm:" [Zhu] when $m \gg 1$

$\underline{0 < \eta < \frac{2}{\lambda(0)}}$: optimization "looks linear" in the sense that
$$\mathcal{L}(t) \approx c(1-\varepsilon)^t, \quad \lambda(t) \approx \lambda(0)$$

$\underline{\frac{2}{\lambda(0)} < \eta < \frac{4}{\lambda(0)}}$: "catapult phase"

loss grows exponentially — if $t \in [0, \underset{\sim \log m}{T_1})$: $\mathcal{L}(t) \approx (1+\varepsilon)^t$, $\lambda(t) \approx \lambda(0)$

loss settles — if $t \in [T_1, \underset{\sim \log m}{T_2})$: $\mathcal{L}(t) = \Theta(m)$ plateaus, $\underbrace{\lambda(t+1) < \lambda(t)}_{\text{this looks like flatter NTK}}$

loss shrinks exponentially — if $t \in [T_2, \infty)$: $\mathcal{L}(t) \approx (1-\varepsilon)^t$, $\lambda(t) \to \lambda(\infty)$ small

$\underline{\frac{4}{\lambda(0)} < \eta}$: optimization diverges $\mathcal{L}(t) \approx (1+\varepsilon)^t \quad \forall t$

<u>Interpretation of catapult phase:</u>

* $\mathcal{L}(t) = (1+\varepsilon)^t \Rightarrow \theta(t)$ leaves the region around $\theta(0)$
* $\lambda(t+1) < \lambda(t) \Rightarrow$ find a "flat part" of parameter space. Since $\mathcal{H}_\theta \mathcal{L}$ and the NTK (λ) are isospectral, $\mathcal{H}_\theta \mathcal{L} = \nabla_\theta z (\nabla_\theta z)^T$ has the same nonzero eigenvalues as $\lambda = (\nabla_\theta z)^T \nabla_\theta z$
  $\Rightarrow$ max eigenvalue keeps decreasing

The <u>key step</u> is to derive a closed set of equations for two "order parameters", which are

$$\left.\begin{cases} \text{residual} \rightarrow z(t+1) - y = f\big(z(t) - y, \lambda(t)\big) \\ \text{NTK} \rightarrow \lambda(t+1) = f\big(z(t) - y, \lambda(t)\big) \end{cases}\right\} \text{coupled recursion of two parameters}$$

Prop. ⋆ (we will prove this at the end)

$$z(t+1) - y = (z(t) - y)\left[\overbrace{1 - \eta\lambda(t)}^{\text{linear part}} + \frac{\|x\|^2}{md}\eta^2 z(t)(z(t) - y)\right] \quad (1)$$

$$\lambda(t+1) = \underbrace{\lambda(t)}_{\text{linear part}} + \eta\frac{\|x\|^2}{md}(z(t) - y)^2\left[\eta\lambda(t) - \frac{4z(t)}{z(t) - y}\right] \quad (2)$$

First, though, we will prove the theorem from this proposition.

(1) Since $m \gg 1$, the two quadratic terms above scale like $\sim \frac{1}{m}$ <u>unless</u> the residuals scale with $z(t) \sim \sqrt{m}$

(note that we can think of ε from the previous lecture to be like $\frac{1}{m}$ (the thing that scales the Hessian))

$\Rightarrow$ early dynamics (before $z(t)$ gets too big or converge) are always $\approx$ linear

(2) So, if $\eta < \frac{2}{\lambda(0)}$, $|z(t)-y| \approx Ce^{-t} \ll \sqrt{m}$ $\forall t$, yielding the first phase

"if you look linear and are driven linearly, you behave linearly"

(3) If $\eta > \frac{2}{\lambda(0)}$, we diverge linearly with $|z(T_1)| \approx Ce^{T_1} \sim \sqrt{m} \Rightarrow T_1 = O(\log m)$
$\lambda(t) \approx \lambda(0)$

Around time $t = T_1$, the recursion in Prop. ☆ yields

$$\lambda(t+1) \approx \lambda(t) + \eta \frac{\|x\|^2}{d}\left[\eta\lambda(0) - 4\right]$$

So, if $\eta < \frac{4}{\lambda(0)}$, $\lambda(t+1) < \lambda(t)$ decreases and $|1 - \eta\lambda(t)|$ gets smaller.

$\Rightarrow z(t) - y$ stops growing until $|1 - \eta\lambda(t)| < 1$, and we re-enter the linear regime with $\mathcal{L}(t) \to 0$ exponentially.

This yields the result! The residuals and NTK fight each other in the quadratic case.

Now, we prove the recursion.

$\boxed{\underline{\text{Proof of Prop. ☆}}}$ -

Recall that

$$z(t) = z^0 + \sum_{i=1}^{m} \nabla_{u_i}^0 (u_i(t) - u_i(0)) + \partial_{v_i}^0 (v_i(t) - v_i(0)) + (u_i(t) - u_i(0))^T H_i^0 (v_i(t) - v_i(0))$$

Taking a gradient,

$$\nabla_{u_i} z(t) = \nabla_{u_i}^0 + H_i^0 (v_i(t) - v_i(0))$$

We can also write out

$$\nabla_{u_i}^0 = \nabla_{u_i}\left[\sum_{j=1}^{m} \frac{1}{\sqrt{m}} v_j(0)\, \sigma\!\left(u_j(0)^T \frac{x}{\sqrt{d}}\right)\right] = \frac{x}{\sqrt{md}} v_i(0)\, \mathbb{1}_{\{u_i(0)^T x \geq 0\}}$$

Also,

$$\partial_{v_i} z(t) = \partial_{v_i}^0 + (u_i(t) - u_i(0))^T H_i^0$$

and

$$\partial_{v_i}^0 = \partial_{v_i}\left[\sum_{j=1}^{m} \frac{1}{\sqrt{m}} v_j(0)\, \sigma\!\left(\frac{u_j(0)^T x}{\sqrt{d}}\right)\right] = \frac{u_i(0)^T x}{\sqrt{md}} \mathbb{1}_{\{u_i(0)^T x \geq 0\}}$$

The mixed derivative is $H_i^0 = \frac{x}{\sqrt{md}} \mathbb{1}_{\{u_i(0)^T x \geq 0\}}$

So, we can compute the residual

$$z(t+1) - y = -y + z^0 + \sum_{i=1}^{m} \nabla_{u_i}^0 (u_i(t+1) - u_i(0)) + \partial_{v_i}^0 (v_i(t+1) - v_i(0))$$

$$+ (u(t+1) - u(0))^T H_i^0 (v_i(t+1) - v_i(0))$$

$$= -y + z^0 + \sum_{i=1}^{m} \nabla_{u_i}^0 (u_i(t) - u_i(0) - \eta \nabla_{u_i} \mathcal{L}(t)) + \partial_{v_i}^0 (v_i(t) - v_i(0) - \eta\, \partial_{v_i} \mathcal{L}(t))$$

$$+ ((u_i(t) - u_i(0) - \eta \nabla_{u_i} \mathcal{L}(t)) H_i^0 (v_i(t) - v_i(0) - \eta\, \partial_{v_i} \mathcal{L}(t))$$

$$= z(t) - y - \eta \left[\begin{array}{l} \sum_{i=1}^{m} \nabla_{u_i}^0 (\nabla_{u_i}[z(t)](z(t) - y)) + \partial_{v_i}^0 (\partial_{v_i}[z(t)](z(t) - y)) \\ + (\nabla_{u_i}[z(t)])^T H_i^0 (v_i(t) - v_i(0))(z(t) - y) + (u_i(t) - u_i(0))^T H_i^0 (\partial_{v_i}[z(t)]) \end{array}\right]$$

$$+ \eta^2 (z(t) - y)^2 \times (\nabla_{u_i}[z(t)])^T H_i^0\, \partial_{v_i}[z(t)]$$

$$= z(t) - y - (z(t) - y)\eta \lambda(t) + \eta^2 (z(t) - y)^2 \times \underbrace{(\nabla_{u_i}[z(t)])^T H_i^0\, \partial_{v_i}[z(t)]}_{\frac{\|x\|^2}{md} z(t)}$$

$$= (z(t) - y)\left[1 - \eta \lambda(t) + \frac{\|x\|^2}{md} \eta^2 z(t)(z(t) - y)\right]$$

$\square$

Open problems for quadratic models!

* $\sigma \neq$ ReLU, one datapoint (perhaps σ s.t. σ' monotone)
* # data ≥ 2, $d \geq 2$ (Zhu et. al do $d=1$, # data $= 2$)
* $Z(T_1) \sim \sqrt{m}$, same as mean field scaling?? ★ relationship between growing function outputs & feature learning/catapult??
* What happens for Φ, Ψ random?
* Do cubic models have another "catapult phase"?

Lecture 11/9 - Implicit Bias

Recall: Last time we considered $z(x;\theta) = \Phi(x)^T\theta + \frac{\varepsilon}{2}\theta^T \Psi(x)\theta$ for small ε expansions for GD and GF.

Today: [Woodworth]

Consider quadratic models of the form

elementwise square

$$\vec{\theta} = \begin{pmatrix}\theta^+ \\ \theta^-\end{pmatrix} \in \mathbb{R}^{2d} \quad \text{where} \quad z(x;\theta) = \langle \beta_\theta ; x\rangle = \langle \theta_+^{\circ 2} - \theta_-^{\circ 2}, x\rangle$$

$$= 0^T \cdot \theta + \frac{2}{2}\theta^T \begin{pmatrix}\mathrm{Diag}(x) & 0 \\ 0 & \mathrm{Diag}(x)\end{pmatrix}\theta$$

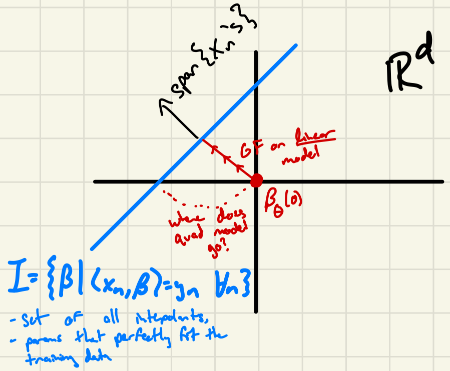
$\Phi = 0$, $\Psi = \begin{pmatrix}\mathrm{Diag}\, x & 0 \\ 0 & \mathrm{Diag}\, x\end{pmatrix}$, $\varepsilon = 2$

The motivation for this is that we are able to express all linear functions with a nonlinear parameterization.

We train by gradient flow (GF)

$$\frac{d}{dt}\theta(t) = -\nabla_\theta \mathcal{L}(\theta(t)), \quad \mathcal{L}(\theta(t)) = \sum_{n=1}^{N} \frac{1}{2}\left(z(x_n;\theta) - y_n\right)^2$$

$$\theta(0) = \alpha\begin{pmatrix}\theta_0 \\ \theta_0\end{pmatrix} \Rightarrow \beta_\theta(0) = \theta_+^{\circ 2}(0) - \theta_-^{\circ 2}(0) = 0$$



The question: As a function of "scale" α, "shape" θ_0, which minimum on I does GF find?

$\left(\begin{array}{l}\text{Mean-field is } \alpha \to 0 \\ \text{NTK parameterization is } \alpha \to \infty\end{array}\right)$ "Implicit bias"

Theorem: If GF with some initialization converges to a minimum loss of $\mathcal{L}$, the minimum is given by

$$\beta^*_{\alpha,\theta_0} = \underset{\beta \in \mathbb{R}^d}{\arg\min}\ \underbrace{Q_{\alpha,\theta_0}(\beta)}_{\text{"implicit" bias}} \text{ subject to } X^T\beta = y.$$

$Q_{\alpha\theta_0}$ is strictly convex:

$$Q_{\alpha\theta_0}(\beta) = \sum_{i=1}^{d} \alpha^2 \theta_{0,i}^2\, q\left(\frac{\beta_i}{\alpha^2\theta_{0,i}^2}\right)$$

where

$$q(z) = 2 - \sqrt{4+z^2} + z\,\text{arcsinh}\left(\frac{z}{2}\right)$$

Reduce nonlinear optimization to linear optimization with explicit penalty!

Interpretation:

(1) This says that we have implicit regularization Q_{α,θ_0} s.t. GF returns

$$\underset{\beta}{\arg\min} \left\{ \sum_{n=1}^{N} (\langle \beta, x_n\rangle - y_n)^2 + \lambda Q_{\alpha,\theta_0}(\beta) \right\} \text{ with } \lambda \to 0$$

minimize loss first then regularizer

(2) $\alpha \to 0$ causes $Q_{\alpha,\theta_0}(\beta) \to \|\beta\|_1$ implicit L_1 regularization "feature selection"

(3) $\alpha \to \infty$ causes $\alpha^2 Q_{\alpha,\theta_0}(\beta) \to \frac{1}{4}\sum_{i=1}^{d} \frac{\beta_i^2}{\theta_{i,0}^2}$ implicit weighted L_2 regularization

(4) For $\alpha \in (0,\infty)$, Q somehow interpolates between the two.

Proof:

write diff eq

Lemma 1: We have $$\frac{d}{dt}\theta(t) = -2\left(\underbrace{(X,-X)^T}_{\in\mathbb{R}^{2d\times N}}\underbrace{\vec{r}(t)}_{\in\mathbb{R}^N}\right) \odot \underbrace{\theta(t)}_{\in\mathbb{R}^{2d}}$$

element-wise product

where $\vec{r}(t) = \begin{pmatrix} r_1(t) \\ \vdots \\ r_N(t) \end{pmatrix}$, $r_n(t) = \langle \beta(t), x_n\rangle - y_n$ residuals

solve diff eq

Lemma 2: We solve

$$\beta_{\alpha,\theta_0}(\infty) = \underbrace{2\alpha^2\theta_0^{\odot 2}}_{\in\mathbb{R}^d} \odot \sinh\left(-4\underbrace{X^T}_{\in\mathbb{R}^{d\times N}}\underbrace{\int_0^\infty \vec{r}(s)\,ds}_{\in\mathbb{R}^N}\right)$$

conclude **Lemma 3:** Show Lemma 2 $\Rightarrow Q_{\alpha,\theta_0}(\beta)$ as stated.

Proof of Lemma 1:

Fix $w=(w_1,\dots,w_d)$. We can see $\partial_{w_k}\langle w^{\odot 2}, x\rangle = 2x_k w_k$

$$\Rightarrow \vec{\nabla}_w \langle \vec{w}^{\odot 2}, \vec{x}\rangle = 2\vec{x}\odot\vec{w}$$

Thus,

$$\frac{d}{dt}\theta_\pm(t) = -\vec{\nabla}_{\theta_\pm}\mathcal{L}(\theta(t)) = -\sum_{n=1}^{N} r_n(t)\vec{\nabla}_{\theta_\pm}\langle\beta,\vec{x}_n\rangle = -\sum_{n=1}^{N} r_n(t)\,\vec{\nabla}_{\theta_\pm}\langle\theta_+^{\odot 2}-\theta_-^{\odot 2},\vec{x}_n\rangle$$

$$= -\sum_{n=1}^{N} r_n(t)\left(\pm 2\vec{x}_n\odot\theta_\pm(t)\right) = \left(\mp\sum_{n=1}^{N} 2\vec{x}_n r_n(t)\right)\odot\theta_\pm(t)$$

$$= \left(-2(\pm X)^T r(t)\right)\odot\theta_\pm(t)$$

So, $$\frac{d}{dt}\theta(t) = \begin{pmatrix}\frac{d}{dt}\theta_+(t)\\ \frac{d}{dt}\theta_-(t)\end{pmatrix} = \begin{pmatrix}-2X^T\vec{r}(t)\odot\theta_+(t)\\ 2X^T\vec{r}(t)\odot\theta_-(t)\end{pmatrix} = -2\left(\underbrace{(X,-X)}_{\text{concatenation}}{}^T\overset{\text{function of }\theta}{\vec{r}(t)}\right)\odot\theta(t)$$

$\square$

Proof of Lemma 2:

We have $\beta(t) = \theta_+^{\odot 2}(t) - \theta_-^{\odot 2}(t)$

By Lemma 1,

$$\theta(t) = \theta(0)\odot e^{-2(X,-X)^T\int_0^t \vec{r}(s)ds}$$

$$\Rightarrow \theta_\pm(t) = \alpha\theta_0\odot e^{\mp 2X^T\int_0^t \vec{r}(s)\,ds}$$

$$\Rightarrow \beta(t) = \alpha^2\theta_0^{\odot 2}\odot\left(e^{-4X^T\int_0^t\vec{r}(s)ds} - e^{4X^T\int_0^t\vec{r}(s)ds}\right)$$

$$= 2\alpha^2\theta_0^{\odot 2}\odot\sinh\left(-4X^T\int_0^t\vec{r}(s)ds\right)$$

$\square$

Note: $-4X^T\int_0^\infty \vec{r}(s)ds \in \mathrm{col}(X^T)$ is some vector in data span

This moves us orthogonally to the interpolant hyperplane $\mathcal{I}$, acting as a Lagrange Multiplier.

Proof of Lemma 3:

Suppose $\beta^*_{\alpha,\theta_0} \equiv \beta_{\alpha;\theta_0}(\infty)$ is a global minimum of $\mathcal{L}$.

$$\Rightarrow \langle \beta^*_{\alpha,\theta_0}, \vec{x}_n \rangle = y_n \quad \forall n$$

Let's write $f_{\alpha,\theta_0}(\beta) \equiv 2\alpha^2 \theta_0^{\odot 2} \odot \sinh(\beta)$ for notation

The KKT conditions (optimality for Lagrange multipliers) for

$$\beta^* = \underset{\beta}{\operatorname{argmin}}\ Q_{\alpha,\theta_0}(\beta) \text{ s.t. } X\beta = Y$$

are $X\beta^* = Y$ and $\exists \nu$ s.t. $\vec{\nabla}_\beta Q_{\alpha,\theta_0}(\beta^*) = X^T \nu$

grad of constraints lies in column space of X (i.e. is orthogonal to interpolant line)

But if we have

$$\vec{\nabla}_\beta Q_{\alpha,\theta_0}\left(f_{\alpha,\theta_0}(X^T\nu)\right) = X^T\nu$$

then KKT constraints are satisfied. So, we want

∇Q evaluated at f

$$\left(\vec{\nabla}_\beta Q_{\alpha,\theta_0}\right) \circ f_{\alpha,\theta_0} = \text{Identity} \Leftrightarrow \vec{\nabla}_\beta Q_{\alpha,\theta_0}(\beta) = f^{-1}_{\alpha,\theta_0}(\beta) \Leftrightarrow Q_{\alpha,\theta_0}(\beta) = \vec{\nabla}^{-1}\left(f^{-1}_{\alpha,\theta_0}(\beta)\right)$$

Since $f_{\alpha,\theta_0} = \alpha^2 \theta_0^{\odot 2} \odot \sinh(\beta)$, we find

$$Q_{\alpha,\theta_0}(\beta) = \sum_{i=1}^{d} \alpha^2 \theta_{0i}^2\, q\left(\frac{\beta_i}{\alpha^2 \theta_{0i}^2}\right) \quad \text{where} \quad q(z) = 2 - \sqrt{4+z^2} + z \operatorname{arcsinh}\left(\frac{z}{2}\right)$$

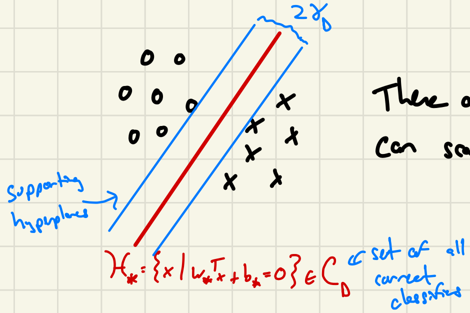
$\square$

Open problems for quadratic models!

* Implicit bias of general quadratic model?
  (how does optimizer & Φ, Ψ interact)
  - $\Phi = 0$, Ψ general or $\Phi = 0$, $\Psi(x)$ simultaneously diagonalizable or Φ expansion in ε
* Catapult phase for general quadratic models
* Convergence of gradient flow

Lecture 11/14 – Implicit Bias II

Consider a dataset $D = \{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^d$, $y_i \in \{\pm 1\}$ that is "linearly separable". i.e. $\exists b_* \in \mathbb{R}$, $w_* \in \mathbb{R}^d$ s.t. $y_i(\vec{w}_*^T \vec{x}_i + b_*) > 0$ ← same sign



There are ∞ many classifiers, since we can scale w_* and b_* to get the same classifier.

Goal: Find implicit bias of GF

$$\begin{cases} \frac{d}{dt} w(t) = -\vec{\nabla}\mathcal{L}(w(t)) \\ \mathcal{L}(w(t)) = \sum_{i=1}^{N} \ell(y_i w^T x_i) \end{cases}$$

← turn off the bias

where $\ell(u) = e^{-u}, \log(1 + e^{-u}), \ldots$

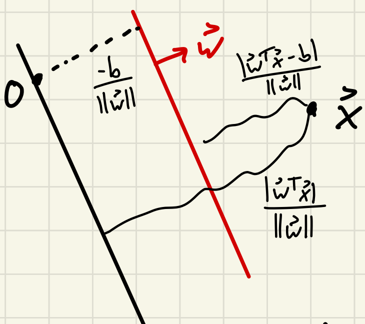
Margins, Support Vectors

Given a classifier $x \to y(x) = \mathrm{sgn}(w^T x + b)$ with $(w, b) \in C_D$, the **margin** is

$$\gamma_{(x_i, y_i)}(w, b) = \text{"margin on } (x_i, y_i)\text{"}$$

$$= \frac{y_i(w^T x_i + b)}{\|w\|} \underset{\text{because } (w,b)\in C_D}{=} \frac{|w^T x_i + b|}{\|w\|}$$

$$= \mathrm{dist}(x_i, \text{decision boundary})$$



We define the **margin on the dataset** by $\gamma_D(w, b) = \min_{(x,y)\in D} \gamma_{(x_i, y_i)}(w, b)$

We define the **max-margin classifier** $\hat{w}, \hat{b}$ as a classifier that maximizes $\max_{(w,b)} \gamma_D(w, b)$ (#)

Note that $\forall (w,b) \in C_D$, $\dfrac{y_i(w^T x_i + b)}{\|w\|}$ is invariant to the transformation $(w,b) \to C(w,b)$ for some $C>0$. So, $\forall (w,b) \in C_D$ we can find $(\tilde{w},\tilde{b}) \in C_D$ s.t.

$$\gamma_D(w,b) = \gamma_D(\tilde{w},\tilde{b}) \quad \text{and} \quad \min_i y_i(\tilde{w}^T x_i + \tilde{b}) = 1 \iff y_i(\tilde{w}^T x_i + \tilde{b}) \geq 1 \ \forall i$$

So, (#) can be viewed with the numerator of γ_D as a constraint in the form $\max_w \dfrac{1}{\|w\|}$ s.t. $y_i(w^T x_i + b) \geq 1 \ \forall i$

$$\iff \boxed{\min_w \tfrac{1}{2}\|w\|^2 \ \text{s.t.} \ y_i(w^T x_i + b) \geq 1 \ \forall i}$$

max-margin classifier objective (##)

Since this is convex objective over convex region, we can find a dual problem

dual variable in $\mathbb{R}^{\#\text{ constraints}} = \mathbb{R}^d$

$$\mathcal{L}(\vec{w}, b, \vec{a}) = \tfrac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^n a_i\left(y_i(w^T x_i + b) - 1\right)$$

So, solutions to (##) must have

$$\vec{\nabla}\mathcal{L} = 0, \qquad y_i(w^T x_i + b) - 1 \geq 0 \qquad a_i \geq 0 \qquad a_i\left(y_i(w^T x_i + b) - 1\right) = 0$$

stationary point — primal feasibility — dual feasibility — boundary constraints (we are tight and on the boundary of either primal or dual)

$$\Rightarrow 0 = \vec{\nabla}_w \mathcal{L} = w - \sum_{i=1}^n a_i y_i x_i, \qquad 0 = \vec{\nabla}_b \mathcal{L} = \sum_{i=1}^n a_i y_i$$

The boundary constraint gives $\forall i$, $a_i = 0$ or $\underbrace{\gamma_{(x_i,y_i)}(w,b) = 1}_{\text{we are on one of the supporting, blue hyperplanes}}$

So, define $S = \{i \mid a_i \neq 0\} \Rightarrow \underbrace{\{\vec{x}_i \mid i \in S\}}_{\text{support vectors}}$

The gradient constraint gives $w = \sum_{i=1}^n a_i y_i x_i \in \text{span}\{x_i, i \in S\}$.
So, the max-margin classifier $\hat{w}$ is defined by the support vectors! If we get new, easier data, we don't change anything. Particularly, for any new point $\vec{x}$,

$$y(\vec{x}; \hat{w}, \hat{b}) = \text{sgn}(\hat{w}^T \vec{x} + \hat{b}) = \text{sgn}\Big(\sum_{i \in S} a_i y_i \underbrace{x_i^T x}_{\text{dot product kernel!}} + \hat{b}\Big)$$

hopefully small # of support vectors

If we replace with feature representations, we get Kernel SVM.

Theorem: Given any init, as $t \to \infty$

- $\|w(t)\| \to \infty$
- $\mathcal{L}(w(t)) \to 0$
- $\frac{w(t)}{\|w(t)\|} = \frac{\hat{w}}{\|\hat{w}\|} + O\left(\frac{1}{\log t}\right)$ where $\hat{w}$ is the max-margin classifier

Proof:

First, suppose WOLOG that all points $(x_i, y_i) \to (y_i x_i, 1)$. ↙ find a half-space all the points are in

This works since we set the bias to 0.

Note that GF and the definition of $\mathcal{L}$ gives

$$\frac{d}{dt} w_*^T w(t) = w_*^T\left(-\vec{\nabla}_w \mathcal{L}(w(t))\right) = -w_*^T \sum_{i=1}^n y_i x_i \ell'(w(t)^T x_i y_i)$$

$$= -\sum_{i=1}^n \underbrace{\overbrace{(y_i w_*^T x_i)}^{\text{independent of } t}}_{>0 \text{ because } w_* \in C_0} \underbrace{\overbrace{\ell'(w(t)^T x_i y_i)}^{\text{uniform bound}}}_{<0 \text{ because } \ell(\cdot) \sim e^{-u}} > 0$$

So, if the data is linearly separable, $\frac{d}{dt} w_*^T w(t) > 0$

Suppose BWOC that $\|w(t)\| \le R \ \forall t \ge 0$ (bounded). Then, $\exists \delta > 0$ s.t.

$$\ell'(y_i w(t)^T x_i) \le -\delta < 0 \Rightarrow \frac{d}{dt} w_*^T w(t) \ge \delta n \min_i (y_i w_*^T x_i)$$

This is a contradiction, since the derivative is uniformly bounded from below, and so must diverge. Therefore, $\boxed{\|w(t)\| \to \infty}$

Now, GF grants that

$$\frac{d}{dt}\mathcal{L}(w(t)) = -\|\vec{\nabla}_w \mathcal{L}(w(t))\|^2$$
$$\le -\left(w_*^T \vec{\nabla}_w \mathcal{L}(w(t))\right)^2$$

Similar logic gives $\boxed{\mathcal{L}(w(t)) \to 0}$.

Lastly, let $r(t) = w(t) - \hat{w}\log(t) - \tilde{w}$, where

$$\forall i \in S, \ e^{-x_i^T \tilde{w}} \equiv a_i \Rightarrow \hat{w} = \sum_{i=1}^{\hat{n}} e^{-x_i^T \tilde{w}} x_i$$

set $y_i = 1$ cutely

↙ we can do this by considering linearly independent subset of S

Let $\theta \equiv \min_{i \notin S} x_i^T \hat{w} > 0$ } min non-supported margin

We want to show that $\|r(t)\|$ is bounded.

We can do this by showing the derivative is integrable.

Now, $\frac{d}{dt}\frac{1}{2}\|r(t)\|^2 = \left(\frac{d}{dt}r(t)\right)^T r(t) = \left(-\vec{\nabla}\mathcal{L}(w(t)) - \frac{1}{t}\hat{\omega}\right)^T r(t)$

$$= \left(\sum_{i=1}^{\hat{n}} x_i e^{-w(t)^T x_i} - \frac{1}{t}\hat{\omega}\right) r(t)$$

$$= \left(\sum_{i=1}^{\hat{n}} x_i e^{(-r(t) - \hat{\omega}\log(t) - \tilde{\omega})^T x_i} - \frac{1}{t}\sum_{i\in S} e^{-x_i^T\tilde{\omega}} x_i\right)^T r(t)$$

$$= \left(\sum_{i=1}^{\hat{n}} x_i \left(\frac{1}{t}\right)^{\overset{\text{margin!}}{\hat{\omega}^T x_i}} e^{-r(t)^T x_i} e^{-\tilde{\omega}^T x_i} - \frac{1}{t}\sum_{i\in S} e^{-x_i^T\tilde{\omega}} x_i\right)^T r(t)$$

Collecting terms with $i \in S$,

$$\frac{1}{t}\sum_{i\in S} x_i e^{-x_i^T\tilde{\omega}}\left(e^{-r(t)^T x_i} - 1\right)^T r(t)$$

$$= \frac{1}{t}\sum_{i\in S} \underbrace{x_i^T r(t)\, e^{-x_i^T\tilde{\omega}}\left(e^{-x_i^T r(t)} - 1\right)}_{c\cdot z(e^{-z}-1) \leq 0!}$$

For $i \notin S$,

$$\left\|\sum_{i\notin S} x_i \left(\frac{1}{t}\right)^{\gamma_{(x_i,y_i)}(\hat{\omega})} e^{-\text{const.}}\right\| \leq n\cdot c\cdot \frac{1}{t^{\theta}} \Rightarrow \int_0^\infty \frac{d}{dt}\|r(t)\|^2 < \infty.$$

$\theta > 1 \Rightarrow$ bounded integral

So, $\|r(t)\|$ is bounded $\Rightarrow$ $\boxed{\frac{w(t)}{\|w(t)\|} = \frac{\hat{\omega}}{\|\hat{\omega}\|} + O\left(\frac{1}{\log t}\right)}$

<u>Things we see:</u>

- Larger gradient signal for small margins
- optimization moves in support vector directions
  - for these directions, the optimization has unique solution

$\square$

Remarks

(Conv, ReLU, avg) are pos. homogenous

1) This thing works for <u>any</u> homogenous classifier (scaling changes score by a power of the scalar)
2) Convergence is slow $\frac{1}{\log(t)}$

Open problems

* Include a bias?
* new losses?
* quadratic models

Paper: Implicit bias of GD on separable data

Lecture 11/16 - SGD Implicit Bias

Suppose that we are given a model $z = (x;\theta)$, $\theta \in \mathbb{R}^n$ which we train by SGD:

$$\mathcal{L}(\theta) = \frac{1}{m}\sum_{k=1}^{m} \ell(x_k;\theta) \qquad \theta(t+1) = \theta(t) - \eta \vec{\nabla}\mathcal{L}^B(\theta)$$

$$\mathcal{L}^B(\theta) = \frac{1}{|B|}\sum_{k=1}^{m} \mathbb{1}_{\{x_k \in B\}}\, \ell(x_k;\theta) \qquad x_k \in B \text{ w.p. } \frac{|B|}{m} \text{ independently}$$

<u>The goal:</u> We wish to understand the implicit bias of SGD.

<u>tl;dr</u>: SGD prefers "wider minima" or "flatter parts" of $\mathbb{R}^n$

<u>Yaida</u>

Yaida uses the dynamical systems perspective that $\theta \sim \mathbb{P}_{ss}$ "steady state" and for any observable $O: \mathbb{R}^n \to \mathbb{R}$,

$$\boxed{\langle O(\theta)\rangle = \langle [[O(\theta - \eta\vec{\nabla}\mathcal{L}^B(\theta))]]\rangle}$$

"fluctuation/dissipation relationship"

where $\langle\cdot\rangle$ is an average w.r.t. the $\mathbb{P}_{ss}$ distribution, and $[[\cdot]]$ is an average w.r.t. batching B.

The philosophy about this is to use FDR to
(i) Taylor expand the RHS
(ii) Collect powers of η
(iii) Compute properties of the steady-state $\mathbb{P}_{ss}$.

<u>Def:</u> Let $\tilde{C}_{ij}(\theta) = (2^{\text{nd}} \text{ moments of } \vec{\nabla}\mathcal{L}^B \text{ w.r.t. } B)$ ← Second moment

$$= [[\partial_{\theta_i}\mathcal{L}^B(\theta)\, \partial_{\theta_j}\mathcal{L}^B(\theta)]]$$

Lemma: In the steady state distribution,

(i) $\langle \vec{\nabla}\mathcal{L}(\theta)\rangle = 0$ $\qquad$ (ii) $\langle \theta \cdot \vec{\nabla}\mathcal{L}(\theta)\rangle = \langle \tfrac{1}{2}\eta\, \mathrm{tr}(\tilde{C})\rangle \geq 0$ $\qquad$ diagonal terms are squares

no net gradient in steady state $\qquad$ alignment is related to batch-averaged loss covariances

Proof: Consider the identity observable $O(\theta)=\theta$. Then, FDR gives

$$\langle\theta\rangle = \langle [[\theta - \eta\vec{\nabla}\mathcal{L}^B(\theta)]]\rangle = \langle\theta\rangle - \eta\langle[[\vec{\nabla}\mathcal{L}^B(\theta)]]\rangle$$

However, $\forall\theta$ we have

$$[[\vec{\nabla}\mathcal{L}^B(\theta)]] = [[\frac{1}{|B|}\sum_{i=1}^{m} \underbrace{\mathbb{1}_{x_k\in B}}_{\frac{|B|}{m}\text{ in expectation}} \vec{\nabla}\ell(x_k;\theta)]]$$

$$= \frac{1}{m}\sum_{i=1}^{m}\vec{\nabla}\ell(x_k;\theta) = \vec{\nabla}\mathcal{L}(\theta) \Rightarrow (i).$$

So, $\langle\theta\rangle = \langle\theta\rangle - \eta\langle\vec{\nabla}\mathcal{L}(\theta)\rangle \Rightarrow \langle\vec{\nabla}\mathcal{L}(\theta)\rangle = 0.$

Next, if $O(\theta) = \frac{1}{2}\theta_i^2$, FDR gives

$$\langle\tfrac{1}{2}\theta_i^2\rangle = \langle[[\tfrac{1}{2}(\theta_i - \eta\partial_{\theta_i}\mathcal{L}^B(\theta))^2]]\rangle$$

$$= \langle[[\tfrac{1}{2}\theta_i^2 - \eta\theta_i\partial_{\theta_i}\mathcal{L}^B(\theta) + \tfrac{1}{2}\eta^2(\partial_{\theta_i}\mathcal{L}^B(\theta))^2]]\rangle$$

$$= \langle\tfrac{1}{2}\theta_i^2\rangle - \eta\langle\theta_i\partial_{\theta_i}\mathcal{L}^B(\theta)\rangle + \frac{\eta^2}{2}\langle[[(\partial_{\theta_i}\mathcal{L}^B(\theta))^2]]\rangle$$

$$\Rightarrow \eta\langle\theta_i\partial_{\theta_i}\mathcal{L}(\theta)\rangle = \frac{\eta^2}{2}\langle\tilde{C}_{ii}\rangle.$$

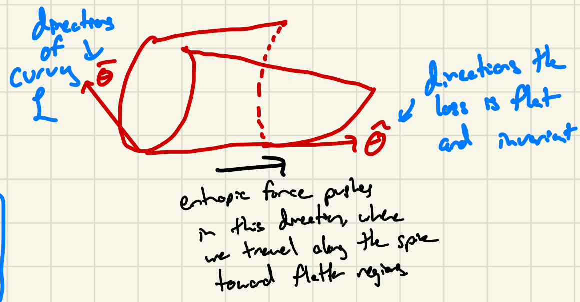
Summing over i's, we get (ii). $\square$

Wei-Schwab-

Typically $m \ll n$, and so there $\exists$ $n-m$ directions in which $\mathcal{L}$ is flat.

Consider $\mathcal{L}(\theta) = \sum_{i=1}^{n}\bar{\theta}_i^2\underbrace{\lambda_i(\hat{\theta}_{n+1},\dots,\hat{\theta}_N)}_{\geq 0}$. The landscape looks like

$\theta = (\bar{\theta},\hat{\theta}) \in \mathbb{R}^N$ $\qquad$ N ← # params



At fixed $\hat{\theta}$ directions, the loss function as a function of $\bar{\theta}$ looks like a curved landscape with

$$\mathrm{Hess}_{\bar{\theta}}(\mathcal{L}) = 2\begin{bmatrix}\lambda_{n+1}(\hat{\theta}) & & 0\\ & \ddots & \\ 0 & & \lambda_N(\hat{\theta})\end{bmatrix}$$

We have the intuition: $\hat{\theta}$ directions evolve "slowly" w.r.t. $\bar{\theta}$ directions since $\bar{\theta}$ directions drive the loss down.

So, we assume $\bar{\theta} \sim \mathbb{P}_{ss}$ and ask then about one step in $\hat{\theta}$ directions.

This assumes that $\bar{\theta}$ already equilibrates before substantial movement in $\hat{\theta}$ directions.

We sketch below some helpful claims and lemmas:

"Claim": Write $C_{ij} = \mathrm{Cov}_{mb}\left(\partial_{\theta_i} \mathcal{L}^B(\theta), \partial_{\theta_j} \mathcal{L}^B(\theta)\right)$. ← variance

At late times, $C = \alpha H_{\bar{\theta}}(\mathcal{L}) \qquad \alpha > 0$

Proof: Lol we don't prove this. Use it as a tool for later thm. □

Note: For each $i = 1, \dots, n$, $\tilde{C}_{ii} = C_{ii} + \left(\partial_{\theta_i} \mathcal{L}^B(\theta)\right)^2 \geq C_{ii}$.

This is the relation between second moments $\tilde{C}$ and covariances C.

Corollary to note: $\langle \bar{\theta}_i^2 \rangle \geq \frac{\alpha \eta}{4} + O(\eta^2)$

↑ how high up the walls we walk in $\bar{\theta}$ $\mathbb{P}_{ss}$ directions

Proof: We apply FDR with the observable $O(\theta) = \langle \theta_i^2 \rangle$. Up to $O(\eta^2)$, we get

$$\langle \bar{\theta}_i \partial_{\bar{\theta}_i} \mathcal{L} \rangle = \eta/2 \langle \tilde{C}_{ii} \rangle$$

by Lemma (ii) evaluated component-wise.

take a derivative

$$= \langle 2 \bar{\theta}_i^2 \lambda_i(\hat{\theta}) \rangle = \eta/2 \langle \tilde{C}_{ii} \rangle$$

$$\| \quad 2\lambda_i(\hat{\theta}) \langle \bar{\theta}_i^2 \rangle \implies \langle \bar{\theta}_i^2 \rangle = \frac{\eta}{4} \frac{\langle \tilde{C}_{ii} \rangle}{\lambda_i(\hat{\theta})} \geq \frac{\eta}{4} \frac{\alpha \cdot H_{ii}(\hat{\theta})}{\lambda_i(\hat{\theta})}$$

"claim"

$$\implies \langle \bar{\theta}_i^2 \rangle \geq \frac{\eta}{4} \alpha .$$

□

Note that in what we do above, we think of $\hat{\theta}$ as fixed as we determine the overall effect of $\mathbb{P}_{ss}$. Only after this do we consider a step in $\hat{\theta}$'s.

$\mathbb{E}$ w.r.t. both $\mathbb{P}_{ss}$ and minibatches

Prop: $\mathbb{E}\left\{ tr\left(H_{\bar{\theta}}(\mathcal{L}(\theta(t+1)))\right) - tr\left(H_{\bar{\theta}}(\mathcal{L}(\theta(t)))\right)\right\} \le 0$

So, we go to places with smaller λ's over time (flatter region in loss). We call this the entropic force.

Proof: Fix $i \in 1,\ldots,n$, $\hat{\theta}(t)$. We have

after $\bar{\theta}$'s equilibrate ($\mathbb{P}_{ss}$ avg.), take a step in $\hat{\theta}$

$$\langle [[\lambda_i(\theta(t+1)) - \lambda_i(\theta(t))]] \rangle$$

$$= \langle [[\lambda_i(\hat{\theta}(t) - \eta \vec{\nabla}_{\hat{\theta}} \mathcal{L}^{B}(\theta(t))) - \lambda_i(\theta(t))]] \rangle$$

Taylor expand

$$= \langle [[-\eta \vec{\nabla}_{\hat{\theta}} \lambda_i(\hat{\theta}(t)) \cdot \vec{\nabla}_{\hat{\theta}} \mathcal{L}^{B}(\theta(t)) + O(\eta^2)]] \rangle$$

$$= -\eta \langle \vec{\nabla}_{\hat{\theta}} \lambda_i(\hat{\theta}(t)) \cdot \vec{\nabla}_{\hat{\theta}} \mathcal{L}(\theta(t)) \rangle + O(\eta^2)$$

take derivative

$$= -\eta \, \vec{\nabla}_{\hat{\theta}} \lambda_i(\hat{\theta}(t)) \sum_{j=1}^{\hat{n}} \langle \bar{\theta}_j^2 \rangle \vec{\nabla}_{\hat{\theta}} \lambda_j(\hat{\theta}(t)) + O(\eta^2)$$

Corollary

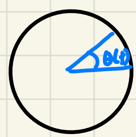
$$\le -\frac{\alpha}{4}\eta \, \vec{\nabla}_{\hat{\theta}} \lambda_i(\hat{\theta}(t)) \sum_{j=1}^{\hat{n}} \vec{\nabla}_{\theta} \lambda_j(\hat{\theta}(t)) + O(\eta^2)$$

Summing this over all i, we get our result.

Since we are riding up the walls (see corollary), we are moving in a direction to give us more room to ride up the walls.

$\square$

Challenge Problem

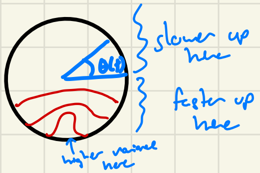


Suppose $\theta(t) \in [0, 2\pi)$.

We have the dynamics $\theta(t+dt) = \theta(t) \pm dt$ with probability $\frac{1}{2}$.

The steady state is given by $d\mathbb{P}_{ss}(\theta) = \frac{1}{2\pi} d\theta$ uniform distribution.

Suppose now that we have the dynamics



$$\theta(t+dt)-\theta(t)=\begin{cases}\pm\, dt & \theta<\pi\\ \pm 2\,dt & \theta\geq\pi\end{cases}$$

We expect more probability mass up top, since we bounce around the bottom 2x faster.

The answer is $d\mathbb{P}_{ss}(\theta)=\frac{\sqrt{2}}{\sqrt{2}+1}\cdot\frac{1}{\pi}\,d\theta\,\mathbb{1}_{\theta<\pi}+\frac{1}{\sqrt{2}+1}\cdot\frac{1}{\pi}\,d\theta\,\mathbb{1}_{\theta\geq\pi}$

far away from the boundary.

The result we see is that we spend less time where the variance is larger. Coming back to SGD, we have

$$\theta(t+1)-\theta(t)=-\eta\vec{\nabla}\mathcal{L}^{B}(\theta)=\underbrace{-\eta\vec{\nabla}\mathcal{L}(\theta)}_{\substack{\text{mean,}\\ \text{GF drift}}}+\underbrace{\eta\left(\vec{\nabla}\mathcal{L}(\theta)-\vec{\nabla}\mathcal{L}^{B}(\theta)\right)}_{\substack{\sim\mathcal{N}(0,C(\theta))\\ \text{state-dependent diffusion term}}}$$

$\Rightarrow$ an implicit bias of SGD is that, in addition to minimizing loss (which the mean takes care of), to also $\boxed{\text{minimize } \operatorname{tr}(C(\theta))}$

We look for areas where the between-batch loss variance is low. This can be thought of as finding flat/isotropic/nice regions of the space of the loss.

Fill in comment

note here

Lecture 11/28 - Entropy + Widths

First, observe that generalization only makes sense given a priori complexity information about the function f we want to learn.

To see this precisely, note that $\forall\, \Omega \subseteq \mathbb{R}^n$, $m \geq 1$, there exists $f: \Omega \to \mathbb{R}$ s.t. we cannot learn f from any dataset of size m.

Proof: Discretize $\Omega = \coprod_{j=1}^{n} \Omega_j$ with $n \gg m$ and $f|_{\Omega_j}$ i.i.d. random. $\square$

We need better notions to talk about how complex a function is to encode/learn.

Def: A model class (class of functions) K is a compact subset of a Banach space $(X, \|\cdot\|_X)$

Some examples:

① $K = \{ f: \Omega \to \mathbb{R} \mid \int_\Omega |f(x)|^2 + \|\nabla f\|^2 \, dx \leq 1 \} \subset L^2(\Omega)$

② $K = \{ f: \Omega \to \mathbb{R} \mid \|f\|_{Lip} \leq 1 \} \subset C^0(\Omega)$

Lipschitz constant

The question is: Given any method for "learning" $f \in K$, how do you measure "how well you did"?

Entropy (Kolmogorov '30s)

Def: Let $\varepsilon_n(K) = n^{th}$ entropy # of K (model class to learn) $= \inf \left\{ \varepsilon > 0 \;\middle|\; \exists \text{ covering of } K \text{ by } 2^n \text{ balls of radius } \varepsilon \right\}$

Intuitions

① K compact $\Rightarrow$ finite cover $\Rightarrow$ $\varepsilon_n(K) < \infty$ $\forall n$

② $\varepsilon_n(K)$ = error in $\|\cdot\|_X$ of best n-bit compression of K

$\underbrace{K \subset \bigcup_{i=1}^{2^n} N_\varepsilon(f_i)}_{\text{def. of } \varepsilon_n(K)}$ yields a bijection $\{f_i\} \longleftrightarrow \{0,1\}^n$ where $f \in K \mapsto$ nearest ball center in K

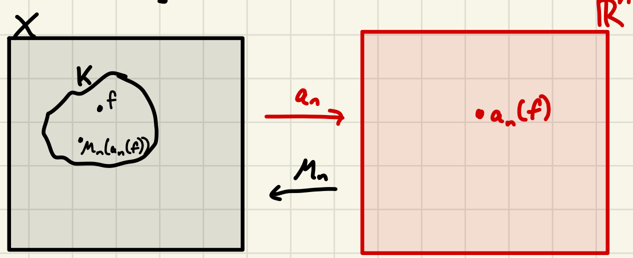
③ $\varepsilon_n(K)$ typically can be computed as $n \to \infty$, but this only tells us how hard a function is to learn, not how well a learning procedure does (not yet).

Stable Width

Def: An n-param approximation scheme for learning K is a pair of functions

"param extraction" $\quad a_n : K \to \mathbb{R}^n$

"reconstruction" $\quad M_n : \mathbb{R}^n \to K$



Def: The error of (a_n, M_n) is

$$E_{a_n,M_n}(K) = \sup_{f \in K} \|f - M_n(a_n(f))\|_X$$

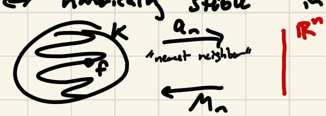
worst reconstruction error over $f \in K$

Def: The stable n-width of K is

$$\delta_n(K) = \inf_{a_n, M_n \in \text{Lip-}2} E_{a_n,M_n}(K)$$

best error we can do

Note that intuitively, Lipschitz $\Leftrightarrow$ "numerically stable" in the sense that it excludes space-filling curves.



The amazing result is that $\varepsilon_n(K)$ and $\delta_n(K)$ are equivalent! We prove this below. First, recall the following results:

Theorem: (Johnson-Lindenstrauss Lemma)
Let $\varepsilon \in (0,1)$. For any $x_1, \dots, x_k \in X$, $\exists$ a 1-Lipschitz (and linear!) function $A: X \to \mathbb{R}^m$ s.t. $\forall i,j$ $\quad (1-\varepsilon)\|x_i - x_j\|_X \leq \|Ax_i - Ax_j\|_{\mathbb{R}^m} \leq \|x_i - x_j\|_X$
as long as $\quad m \geq \frac{8}{\varepsilon^2}\log(k)$.

Theorem: (Kirzbraun Extension Theorem)
If $f: U_1 \to \mathcal{H}_2$, $U_1 \subseteq \mathcal{H}_1$ is Lipschitz, then $\exists F: \mathcal{H}_1 \to \mathcal{H}_2$ s.t.

$$F|_{U_1} = f \quad \text{and} \quad \|F\|_{Lip} = \|f\|_{Lip}$$

same Lipschitz constant

With this machinery, we can prove both directions.

($\Rightarrow$)

Theorem: $\forall n \quad \delta_{32n}(K) \leq 3\varepsilon_n(K)$

definition of $\varepsilon_n(K)$

Proof: Fix n. Choose $\{f_i, i \in [2^n]\} \subseteq K$ s.t. $K \subseteq \bigcup_{i=1}^{2^n} N_{\varepsilon_n(K)}(f_i)$
Applying JL on these ball centers with $\varepsilon = \frac{1}{2}$, $k = 2^n$, $x_i = f_i$, we get
$a: K \to \mathbb{R}^{32n}$ s.t. $\forall i,j$, $\quad \frac{1}{2}\|f_i - f_j\|_X \leq \|a(f_i) - a(f_j)\|_{\mathbb{R}^{32n}} \leq \|f_i - f_j\|_X$

Note that over $U_1 = \{a(f_i)\} \subseteq \mathbb{R}^{32n}$, a function $M_1: U_1 \to X$ that inverts a on the ball centers (i.e. $M_1(a(f_i)) = f_i$) is 2-Lipschitz by the JL inequality. So, by the extension theorem, there exists $M: \mathbb{R}^{32n} \to X$ that is 2-Lipschitz with $M(a(f_i)) = f_i \ \forall i$. So, $\forall f \in K$,

$$\|f - M(a(f))\|_X \leq \underbrace{\|f - f_i\|_X}_{\leq \varepsilon_n(K)} + \underbrace{\|f_i - M(a(f_i))\|_X}_{=0} + \underbrace{\|M(a(f_i)) - M(a(f))\|_X}_{\leq 2\varepsilon_n(K) \text{ because } \|M\|_{Lip}=2,\ \|a\|_{Lip}=1 \Rightarrow \|M \circ a\|_{Lip}=2} \quad \text{Triangle ineq.}$$

$$\leq \varepsilon_n(K) + 0 + 2\varepsilon_n(K) = 3\varepsilon_n(K)$$

Since this holds for all $f \in K$,

$$\delta_{32n}(K) \leq E_{a,M}(K) \leq 3\varepsilon_n(K)$$

$\square$

($\Leftarrow$)

Theorem: Fix $r>0$. Then, $\delta_n(K) \le n^{-r} \Rightarrow \varepsilon_n(K) \le (n/\log n)^{-r}$
(ε, δ go to 0 together)

Proof: Fix n and consider a near-optimal (a_n, M_n) s.t. $\delta = E_{a_n, M_n}(K)$ and $\delta_n(K) \le \delta \le 2\delta_n(K)$. Suppose $\underbrace{a_n(K)}_{\text{image of } K} \subset N_R(?) \subset \mathbb{R}^n$.

Let $\{N_{2\delta}(f_i)\}_{i=1}^{P_\delta(K)}$ be a maximal 2δ-packing of K.
($P_\delta(K)$ is max # of disjoint balls of radius 2δ fitting in K)

Note that $\{N_{4\delta}(f_i)\}_{i=1}^{P_\delta(K)}$ is a covering of K (if not, we could have fit another 2δ ball in the packing). We analyze the functions of a_n, M_n at each ball center f_i. Note that $\forall i,j \in [P_\delta(K)]$,

$$\|M_n(a_n(f_i)) - M_n(a_n(f_j))\|_X \ge 2\delta \Rightarrow \|a_n(f_i) - a_n(f_j)\|_{\mathbb{R}^n} \ge \delta \quad \text{by } M_n \text{ 2-Lipschitz.}$$

Thus, $\{N_\delta(a_n(f_i)),\ i \in [P_\delta(K)]\}$ is a δ-packing of $\underbrace{N_R(?)}_{\text{we know the volume!}}$ in $\mathbb{R}^n$.

Hence, $P_\delta(K) \le \left(\frac{6R}{\delta}\right)^n = 2^{n \log\left(\frac{c}{\delta}\right)}$ for some c

So, $\varepsilon_{n\log\left(\frac{c}{\delta}\right)}(K) \le 4\delta \le 8\delta_n(K)$

Then, if $\delta_n(K) \le n^{-r}$, $\underbrace{\varepsilon_{n\log n}}_{N}(K) \le \underbrace{n}_{\frac{N}{\log N}}^{-r}$.

still shaky about these 3 lines ...

$\square$

We combine these as follows:

★ **Theorem:** (Carl, Cohen, DeVore, ...)

When $K \subset\subset X$ and X is a Hilbert space, $\varepsilon_n(K) \asymp \delta_n(K)$ as $n \to \infty$.

differ by a universal constant, & grow the same

Proof: Results of the two above theorems as $n \to \infty$. $\square$

Open problems!

★ Add a dataset of size m (restrict a to something factorable over an evaluation map at m points)

★ How regular (Lipschitz?) are NN functions?

★ Solidify relationship between above and statistical learning ($\varepsilon_n(K)$ is basically VC dim.)

Lecture 11/30 - Path Counting for ReLU

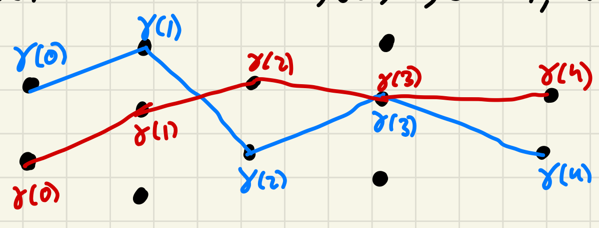
Consider a ReLU FCNN
$$z_{i;\alpha}^{(\ell+1)} = \begin{cases} \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma(z_{j;\alpha}^{(\ell)}) & \ell \geq 1 \\ \sum_{j=1}^{n_\ell} W_{ij}^{(\ell)} x_{j;\alpha} & \ell = 0 \end{cases}$$

with widths $n_1, \ldots, n_L$ and
$$\sigma(t) = t \mathbb{1}_{t>0}, \qquad W_{ij}^{(\ell+1)} = \sqrt{\frac{2}{n_\ell}}\, \hat{W}_{ij}^{(\ell+1)}, \qquad \hat{W}_{ij}^{(\ell+1)} \sim \mu \text{ i.i.d}$$

where the distribution μ is symmetric about the mean with variance 1 and finite moments.

The goal is to explain a $\overbrace{\underline{\text{combinatorial approach}}}^{\text{path counting}}$ to study __any__ statistic of random ReLU at a single input $x_\alpha \neq 0$.

__Def:__ For each $n \geq 1$ write $[n] = \{1, \ldots, n\}$.
The space of paths in a FCNN with widths $n_0, \ldots, n_{L+1}$ is
$$\Gamma = [n_0] \times \cdots \times [n_{L+1}]$$
i.e. $\gamma \in \Gamma$ is $\gamma = (\gamma(0), \gamma(1), \ldots, \gamma(L+1))$, $\gamma(\ell) \in [n_\ell]$ $\forall \ell$.



__Notation:__ For each $\ell = 1, \ldots, L+1$, let
$$W_\gamma^{(\ell)} = W_{\gamma(\ell), \gamma(\ell-1)}^{(\ell)}, \qquad z_\gamma^{(\ell)} = z_{\gamma(\ell);\alpha}^{(\ell)}, \qquad \#\Gamma = \prod_{\ell=0}^{L+1} n_\ell$$
$$\Gamma_{p,q} = \{\gamma \in \Gamma : \gamma(0) = p,\ \gamma(L+1) = q\}$$

Prop. We have

$$z_{q;\alpha}^{(L+1)} = \sum_{p=1}^{n_0} x_{p;\alpha} \sum_{\gamma \in \Gamma_{p,q}} W_\gamma^{(L+1)} \prod_{\ell=1}^{L} W_\gamma^{(\ell)} \xi_\gamma^{(\ell)} \qquad (\star)$$

where $\xi_\gamma^{(\ell)} = \mathbb{1}_{\{z_\gamma^{(\ell)} \geq 0\}}$ indicates if the neuron $\gamma(\ell)$ is on.

Proof: Given $\vec{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$, $\sigma(\vec{v}) = D_v \vec{v}$, where $D_v = \mathrm{Diag}(\mathbb{1}_{\{v_1 \geq 0\}}, \dots, \mathbb{1}_{\{v_n \geq 0\}})$. Thus,

$$\vec{z}_\alpha^{(L+1)} = W^{(L+1)} \sigma(W^{(L)} \sigma(\dots \sigma(W^{(1)} \vec{x}_\alpha) \dots)) = W^{(L+1)} D^{(L)} W^{(L)} \dots D^{(1)} W^{(1)} \vec{x}_\alpha$$

where $D^{(\ell)} = \mathrm{Diag}(\mathbb{1}_{\{z_{i;\alpha}^{(\ell)} \geq 0\}}, i = 1, \dots, n_\ell)$. Hence,

$$\begin{aligned} z_{q;\alpha}^{(L+1)} &= \sum_{p=1}^{n_0} x_{p;\alpha} \left(W^{(L+1)} D^{(L)} W^{(L)} \dots D^{(1)} W^{(1)} \right)_{pq} \\ &= \sum_{p=1}^{n_0} x_{p;\alpha} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \dots \sum_{i_L=1}^{n_L} D_{i_1 i_1}^{(1)} W_{i_1 p}^{(1)} \cdot \ldots \cdot D_{i_L i_L}^{(L)} W_{i_L i_{L-1}}^{(L)} \cdot W_{q i_L}^{(L+1)} \\ &= \sum_{p=1}^{n_0} x_{p;\alpha} \sum_{\gamma \in \Gamma_{p,q}} W_\gamma^{(L+1)} \prod_{\ell=1}^{L} W_\gamma^{(\ell)} \xi_\gamma^{(\ell)} \end{aligned}$$

$\square$

Prop: At init,

$$\mathbb{1}_{\{z_{i;\alpha}^{(\ell)} \geq 0\}} \overset{d}{=} \mathrm{Bernoulli}(\tfrac{1}{2}) \text{ i.i.d}$$

and independent of any <u>even</u> function of the weights $W^{(\ell)}$'s.

i.e. symmetric under $W^{(\ell)} \mapsto -W^{(\ell)}$

Proof:

Idea: given $\vec{z}_\alpha^{(\ell)}$, i.i.d weights symmetric about 0 means

$$\mathbb{1}_{\{z_{i;\alpha}^{(\ell+1)} \geq 0\}} \overset{d}{=} \mathrm{Bernoulli}(\tfrac{1}{2}) \text{ given } \vec{z}_\alpha^{(\ell)}.$$

Since this distribution is the same regardless of $\vec{z}_\alpha^{(\ell)}$, we are done.

For the weight-wise independence, check the paper :-

$\square$

Corollary: We have $$z_\alpha^{(L+1)} \overset{d}{=} W^{(L+1)} \hat{D}^{(L)} W^{(L)} \cdots \hat{D}^{(1)} W^{(1)} \vec{x}_\alpha$$
where $\hat{D}_{ii}^{(\ell)} \sim \text{Bernoulli}(\tfrac{1}{2})$ i.i.d.

Proof: Duh. $\square$

Corollary: We have $$\frac{\partial z_{q;\alpha}^{(L+1)}}{\partial x_{p;\alpha}} = \sum_{\gamma \in \Gamma_{p,q}} W_\gamma^{(L+1)} \prod_{\ell=1}^{L} W_\gamma^{(\ell)} \xi_\gamma^{(\ell)}$$

Proof: Duh. $\square$

Lemma: For any $n_0, \dots, n_{L+1}$, $$\mathbb{E}\left\{\left(\frac{\partial z_{q;\alpha}^{(L+1)}}{\partial x_{p;\alpha}}\right)^2\right\} = \frac{2}{n_0}.$$

Proof: Let $A = \mathbb{E}\left\{\left(\frac{\partial z_{q;\alpha}^{(L+1)}}{\partial x_{p;\alpha}}\right)^2\right\}$. Then
$$A = \mathbb{E}\Bigg\{\underbrace{\sum_{\gamma_1,\gamma_2 \in \Gamma_{p,q}} \prod_{k=1}^{2}\left(W_{\gamma_k}^{(L+1)} \prod_{\ell=1}^{L} W_{\gamma_k}^{(\ell)} \xi_{\gamma_k}^{(\ell)}\right)}_{\text{even function in } W^{(\ell)}\text{, so } \xi\text{'s independent}}\Bigg\}$$
$$\Rightarrow A = \sum_{\gamma_1,\gamma_2 \in \Gamma_{p,q}} \mathbb{E}\left\{\prod_{k=1}^{2} W_{\gamma_k}^{(L+1)}\right\} \cdot \prod_{\ell=1}^{L} \mathbb{E}\left\{\prod_{k=1}^{2} W_{\gamma_k}^{(\ell)}\right\} \mathbb{E}\left\{\prod_{k=1}^{2} \xi_{\gamma_k}^{(\ell)}\right\}$$
But note that $$\mathbb{E}\left\{\prod_{k=1}^{2} W_{\gamma_k}^{(\ell)}\right\} = \frac{2}{n_{\ell-1}} \underbrace{\delta_{\gamma_1(\ell)\gamma_2(\ell)}\, \delta_{\gamma_1(\ell-1)\gamma_2(\ell-1)}}_{\substack{\text{if not the same, this} \\ \text{is a product of i.i.d.} \\ \text{mean 0 weights}}}$$
variance

So, we sum over $\gamma = \gamma_1 = \gamma_2$. For this, $\mathbb{E}\left\{\prod_{k=1}^{2} \xi_{\gamma_k}^{(\ell)}\right\} = \mathbb{E}\left\{\xi_\gamma^{(\ell)2}\right\} = \frac{1}{2}$.

All together,
$$A = \sum_{\gamma \in \Gamma_{p,q}} \frac{2}{n_L} \prod_{\ell=1}^{L} \frac{\cancel{2}}{n_{\ell-1}} \cdot \frac{1}{\cancel{2}} = \frac{2}{n_0} \frac{1}{\prod_{\ell=1}^{L} n_\ell} \sum_{\gamma \in \Gamma_{p,q}} 1$$

$$= \frac{2}{n_0} \; \mathcal{E}\{1\} \leftarrow \text{expectation over uniform measure in path space}$$

where $\mathcal{E}$ is an average over choices of random $\gamma \in \Gamma_{p,q}$ with $\gamma(0) = p$, $\gamma(L+1) = q$, $\gamma(\ell) \sim \mathrm{Unif}([n_\ell])$ independently.

Clearly, $A = \frac{2}{n_0}$

$\square$

Theorem: (Boris spittin)

When $n_1, \dots, n_L$ are large, let $\beta = 5 \sum_{\ell=1}^{L} \frac{1}{n_\ell}$ ($5 \cdot$ aspect ratio ($r = \frac{L}{n}$))

Then,
$$\left(\frac{\partial z_{q;\alpha}^{(L+1)}}{\partial x_{p;\alpha}} \right)^2 = \exp\left(\mathcal{N}\left(-\tfrac{\beta}{2}, \beta\right) + O\left(\tfrac{\beta}{n}\right) \right)$$
$\leftarrow O\left(\frac{L}{n^2}\right)$

Exponentially sensitive in the aspect ratio!

Proof: Nope :)

$\square$

Exercise: Show that
$$\mathbb{E}\left\{ \left(\frac{\partial z_{q;\alpha}^{(L+1)}}{\partial x_{p;\alpha}} \right)^4 \right\} \simeq \frac{\text{const}}{n_0^2} \exp\left(5 \sum_{\ell=1}^{L} \frac{1}{n_\ell} \right)$$

Lemma: Consider the on-diagonal NTK
$$\Theta_{\alpha\alpha}^{(L+1)} = \| \vec{\nabla}_\theta z_\alpha^{(L+1)} \|^2 \quad \text{where } n_{L+1} = 1$$
$$= \sum_{\ell=1}^{L+1} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_{\ell-1}} \underbrace{\left(\frac{\partial z_{1;\alpha}^{(L+1)}}{\partial \hat{W}_{ij}^{(\ell)}} \right)^2}_{\text{call this } B}$$
$$\Rightarrow \mathbb{E}\{ \Theta_{\alpha\alpha}^{(L+1)} \} = 2L \frac{\|\vec{x}_\alpha\|^2}{n_0}$$

Proof: Note that
$$\frac{\partial z_{1;\alpha}^{(L+1)}}{\partial \hat{W}_{ij}^{(\ell)}} = \sum_{p=1}^{n_0} x_{p;\alpha} \cdot \sum_{\substack{\gamma \in \Gamma_{p,1} \\ \gamma(\ell) = i \\ \gamma(\ell-1) = j}} \frac{\sqrt{\frac{2}{n_L}}\, \hat{W}_\gamma^{(L+1)} \prod_{\ell=1}^{L} \sqrt{\frac{2}{n_{\ell-1}}}\, \hat{W}_\gamma^{(\ell)} \xi_\gamma^{(k)}}{\hat{W}_{ij}^{(\ell)}} \quad \text{by } \bigstar$$

Then,
$$\mathbb{E}\{B\} = \mathbb{E}\left\{ \sum_{p_1, p_2 = 1}^{n_0} x_{p_1;\alpha} x_{p_2;\alpha} \cdot \sum_{\substack{\gamma_1, \gamma_2 \in \Gamma_{p_1} \\ \gamma_k \ni W_{ij}^{(\ell)}}} \frac{\frac{2}{n_L} \prod_{k=1}^{2} \hat{W}_{\gamma_k}^{(L+1)} \prod_{\ell=1}^{L} \frac{2}{n_{\ell-1}} \prod_{k=1}^{2} \hat{W}_{\gamma_k}^{(\ell)} \xi_{\gamma_k}^{(\ell)}}{(\hat{W}_{ij}^{(\ell)})^2} \right\}$$

$$= \sum_{p_1,p_2}^{n_0} x_{p_1\alpha} x_{p_2\alpha} \sum_{\substack{\gamma_1,\gamma_2 \in \Gamma_{p,1} \\ \gamma_k \ni W_{ij}^{(\ell)}}} \prod_{\ell' \neq \ell} \frac{2}{n_{\ell'-1}} \mathbb{E}\left\{\prod_{k=1}^{2} \hat{W}_{\gamma_k}^{(\ell)}\right\} \mathbb{E}\left\{\xi_{\gamma_k}^{(\ell)}\right\} \mathbb{E}\left\{\frac{2}{n_L} \prod_{k=1}^{2} \hat{W}_{\gamma_k}^{(L+1)}\right\}$$

$$\mathbb{E}\left\{\frac{2}{n_{\ell-1}} \prod_{k=1}^{2} \xi_{\gamma_k}^{(\ell)}\right\}$$

$$= \sum_{p=1}^{n_0} x_{p;\alpha}^2 \cdot 2 \prod_{\ell=0}^{L} \frac{1}{n_\ell} \sum_{\substack{\gamma \in \Gamma_{p,1} \\ \hat{W}_{ij}^{(\ell)} \in \gamma}} 1 = \frac{2\|\vec{x}_\alpha\|^2}{n_0} \cdot \frac{1}{\prod_{\ell=1}^{L} n_\ell} \#\left\{\gamma \in \Gamma_{p,1} : W_{ij}^{(\ell)} \in \gamma\right\}$$

$$= \frac{2\|\vec{x}_\alpha\|^2}{n_0} (n_\ell n_{\ell-1})^{-1}$$

So,
$$\mathbb{E}\left\{\Theta_{\alpha\alpha}^{(L+1)}\right\} = \sum_{\ell=1}^{L} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_{\ell-1}} \frac{2\|\vec{x}_\alpha\|^2}{n_0} (n_\ell n_{\ell-1})^{-1} = \frac{2L\|\vec{x}_\alpha\|^2}{n_0}$$

$\square$

Lemma: Consider the off-diagonal NTK

$$\Theta^{(L+1)}_{\alpha\beta} = \left(\vec{\nabla}_\theta z_\alpha^{(L+1)}\right)^T \left(\vec{\nabla}_\theta z_\beta^{(L+1)}\right) \quad \text{where} \quad n_{L+1}=1$$

$$= \sum_{\ell=1}^{L+1} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_{\ell-1}} \underbrace{\frac{\partial z_{1;\alpha}^{(L+1)}}{\partial \hat{W}_{ij}^{(\ell)}} \frac{\partial z_{1;\beta}^{(L+1)}}{\partial \hat{W}_{ij}^{(\ell)}}}_{B} \quad \Big\} = 0 \text{ when } \ell = L+1\text{??}$$

Proof: Note that

$$\frac{\partial z_{1;\alpha}^{(L+1)}}{\partial \hat{W}_{ij}^{(\ell)}} = \sum_{p=1}^{n_0} x_{p;\alpha} \cdot \sum_{\substack{\gamma\in\Gamma_{p,1}\\ \gamma(\ell)=i\\ \gamma(\ell-1)=j}} \frac{\sqrt{\frac{2}{n_L}}\, \hat{W}_\gamma^{(L+1)} \prod_{\ell'=1}^{L} \sqrt{\frac{2}{n_{\ell'-1}}}\, \hat{W}_\gamma^{(\ell')} \xi_{\gamma,\alpha}^{(\ell')}}{\hat{W}_{ij}^{(\ell)}} \quad \text{by } \bigstar$$

and similarly

$$\frac{\partial z_{1;\beta}^{(L+1)}}{\partial \hat{W}_{ij}^{(\ell)}} = \sum_{p=1}^{n_0} x_{p;\beta} \cdot \sum_{\substack{\gamma\in\Gamma_{p,1}\\ \gamma(\ell)=i\\ \gamma(\ell-1)=j}} \frac{\sqrt{\frac{2}{n_L}}\, \hat{W}_\gamma^{(L+1)} \prod_{\ell'=1}^{L} \sqrt{\frac{2}{n_{\ell'-1}}}\, \hat{W}_\gamma^{(\ell')} \xi_{\gamma,\beta}^{(\ell')}}{\hat{W}_{ij}^{(\ell)}}$$

Then,

$$\mathbb{E}\{B\} = \mathbb{E}\left\{ \sum_{p_\alpha,p_\beta=1}^{n_0} x_{p_\alpha;\alpha} x_{p_\beta;\alpha} \cdot \sum_{\substack{\gamma_\alpha\in\Gamma_{p_\alpha,1}\\ \gamma_\beta\in\Gamma_{p_\beta,1}\\ \gamma_\alpha(\ell)=\gamma_\beta(\ell)=i\\ \gamma_\alpha(\ell-1)=\gamma_\beta(\ell-1)=j}} \frac{\frac{2}{n_L} \hat{W}_{\gamma_\alpha}^{(L+1)} \hat{W}_{\gamma_\beta}^{(L+1)} \prod_{\ell'=1}^{L} \frac{2}{n_{\ell'-1}} \hat{W}_{\gamma_\alpha}^{(\ell')} \xi_{\gamma_\alpha\alpha}^{(\ell')} \hat{W}_{\gamma_\beta}^{(\ell')} \xi_{\gamma_\beta\beta}^{(\ell')}}{\left(\hat{W}_{ij}^{(\ell)}\right)^2} \right\}$$

⊥ b.c. ξ ind. from each fn. of weights

$$= \sum_{p_\alpha,p_\beta=1}^{n_0} x_{p_\alpha,\alpha} x_{p_\beta,\beta} \sum_{\substack{\gamma_\alpha\in\Gamma_{p_\alpha,1}\\ \gamma_\beta\in\Gamma_{p_\beta,1}\\ \gamma_\alpha(\ell)=\gamma_\beta(\ell)=i\\ \gamma_\alpha(\ell-1)=\gamma_\beta(\ell-1)=j}} \frac{2}{n_L} \mathbb{E}\{\hat{W}_{\gamma_\alpha}^{(L+1)} \hat{W}_{\gamma_\beta}^{(L+1)}\} \prod_{\ell'\neq\ell} \frac{2}{n_{\ell'-1}} \mathbb{E}\{\hat{W}_{\gamma_\alpha}^{(\ell')} \hat{W}_{\gamma_\beta}^{(\ell')}\} \mathbb{E}\{\xi_{\gamma_\alpha\alpha}^{(\ell')} \xi_{\gamma_\beta\beta}^{(\ell')}\}$$
$$\frac{2}{n_{\ell-1}} \mathbb{E}\{\xi_{\gamma_\alpha\alpha}^{(\ell)} \xi_{\gamma_\beta\beta}^{(\ell)}\}$$

⊥ between layers

For any paths $\gamma_\alpha \in \Gamma_{p_\alpha,1}$, $\gamma_\beta \in \Gamma_{p_\beta,1}$ that are distinct, their contributions disappear. Precisely, suppose that $\gamma_\alpha(\mathcal{L}) \neq \gamma_\beta(\mathcal{L}) \in [n_\mathcal{L}]$ for some $\mathcal{L}\neq\ell$. Then, $\hat{W}^{(\mathcal{L})}_{\gamma_\alpha(\mathcal{L}),\gamma_\alpha(\mathcal{L}-1)} \perp \hat{W}^{(\mathcal{L})}_{\gamma_\beta(\mathcal{L}),\gamma_\beta(\mathcal{L}-1)}$ and so $\mathbb{E}\{\hat{W}_{\gamma_\alpha}^{(\mathcal{L})} \hat{W}_{\gamma_\beta}^{(\mathcal{L})}\} = \mathbb{E}\{\hat{W}_{\gamma_\alpha}^{(\mathcal{L})}\}\mathbb{E}\{\hat{W}_{\gamma_\beta}^{(\mathcal{L})}\} = 0$.

γ_α and γ_β cannot disagree at ℓ either, since these do not contribute to the derivative w.r.t. $\hat{W}_{ij}^{(\ell)}$ (we required $\gamma_\alpha(\ell)=\gamma_\beta(\ell)=i$, $\gamma_\alpha(\ell-1)=\gamma_\beta(\ell-1)=j$). This logic at $\mathcal{L}=1$ requires that $p_\alpha = p_\beta$ as well. So, we sum over identical paths to get

$$= \sum_{p=1}^{n_0} x_{p;\alpha} x_{p;\beta} \sum_{\substack{\gamma\in\Gamma_{p,1}\\ \gamma(\ell)=i\\ \gamma(\ell-1)=j}} \frac{2}{n_L} \underbrace{\mathbb{E}\{\hat{W}_\gamma^{(L+1)^2}\}}_{=\mathrm{Var}\,\hat{W}=1} \prod_{\ell'\neq\ell} \underbrace{\mathbb{E}\{\hat{W}_\gamma^{(\ell')^2}\}}_{=1} \prod_{\ell'=1}^{L} \frac{2}{n_{\ell'-1}} \mathbb{E}\{\xi_{\gamma;\alpha}^{(\ell')} \xi_{\gamma;\beta}^{(\ell')}\}$$

$$= \sum_{p=1}^{n_0} x_{p\alpha} x_{p\beta} \cdot \frac{2^{L+1}}{\prod_{\ell'=0}^{L} n_{\ell'}} \cdot \sum_{\substack{\gamma\in\Gamma_{p,1}\\ \gamma(\ell)=i\\ \gamma(\ell-1)=j}} \prod_{\ell'=1}^{L} \mathbb{E}\{\xi_{\gamma;\alpha}^{(\ell')} \xi_{\gamma;\beta}^{(\ell')}\}$$

Due to symmetry over paths (same $W_{ij}^{(\ell)} \sim \mu$), $\prod_{\ell'=1}^{L} \mathbb{E}\{\xi_{\gamma;\alpha}^{(\ell')} \xi_{\gamma;\beta}^{(\ell')}\}$ is the same $\forall \gamma$.

$$= \sum_{p=1}^{n_0} x_{p\alpha} x_{p\beta} \cdot \frac{2^{L+1}}{\prod_{\ell'=0}^{L} n_{\ell'}} \cdot \prod_{\ell'=1}^{L} \mathbb{E}\left\{ z^{(\ell')}_{\gamma;\alpha} z^{(\ell')}_{\gamma;\beta} \right\} \cdot \Big| \underbrace{\left\{ \gamma \in \Gamma_{p,1} \;\middle|\; \gamma(\ell)=i,\ \gamma(\ell-1)=j \right\}}_{\frac{\prod_{\ell'=1}^{L} n_{\ell'}}{n_\ell \cdot n_{\ell-1}}} \Big|$$

$$= \sum_{p=1}^{n_0} x_{p\alpha} x_{p\beta} \cdot \frac{2^{L+1}}{n_0 n_{\ell-1} n_\ell} \cdot \prod_{\ell'=1}^{L} \mathbb{E}\left\{ z^{(\ell')}_{\gamma;\alpha} z^{(\ell')}_{\gamma;\beta} \right\} = \vec{x}_\alpha \cdot \vec{x}_\beta \cdot \frac{2^{L+1}}{n_0 n_{\ell-1} n_\ell} \cdot \prod_{\ell'=1}^{L} \mathbb{E}\left\{ z^{(\ell')}_{\gamma;\alpha} z^{(\ell')}_{\gamma;\beta} \right\}$$

This gives

$$\mathbb{E}\left\{ \Theta^{(L+1)}_{\alpha\beta} \right\} = \sum_{\ell=1}^{L+1} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_{\ell-1}} \vec{x}_\alpha \cdot \vec{x}_\beta \cdot \frac{2^{L+1}}{n_0 n_{\ell-1} n_\ell} \cdot \prod_{\ell'=1}^{L} \mathbb{E}\left\{ z^{(\ell')}_{\gamma;\alpha} z^{(\ell')}_{\gamma;\beta} \right\}$$

$$= \vec{x}_\alpha \cdot \vec{x}_\beta \cdot \frac{2^{L+1}}{n_0} \cdot \sum_{\ell=1}^{L} \prod_{\ell'=1}^{L} \mathbb{E}\left\{ z^{(\ell')}_{\gamma;\alpha} z^{(\ell')}_{\gamma;\beta} \right\}$$

this sum goes to L instead of L+1. why??

$$= \underbrace{\boxed{\vec{x}_\alpha \cdot \vec{x}_\beta \cdot \frac{L}{n_0} \cdot 2^{L+1} \prod_{\ell=1}^{L} \mathbb{E}\left\{ z^{(\ell)}_{\gamma;\alpha} z^{(\ell)}_{\gamma;\beta} \right\}}}_{\text{$\mathbb{P}$ both inputs turn on ALL neurons in a path}}$$

We have $K \in \mathbb{R}^{m \times m}$ as the expected NTK at init. Then define $\mu_{max} = \lambda_{max}(K)$. Since the top eigenvalue is $\leq$ any matrix norm, we get

$$\mu_{max} \leq \max_\beta \sum_{\alpha=1}^{m} |K_{\alpha\beta}| = \max_\beta \sum_{\alpha=1}^{m} |\vec{x}_\alpha \cdot \vec{x}_\beta| \, \frac{L}{n_0} 2^{L+1} \cdot \prod_{\ell=1}^{L} \mathbb{E}\left\{ z^{(\ell)}_{\gamma\alpha} z^{(\ell)}_{\gamma\beta} \right\}$$

We have $\mathbb{E}\left\{ z^{(1)}_{\gamma\alpha} z^{(1)}_{\gamma\beta} \right\} = \frac{1}{2} - \frac{1}{2\pi} \arccos\left(\frac{\vec{x}_\alpha \cdot \vec{x}_\beta}{\|\vec{x}_\alpha\| \|\vec{x}_\beta\|_2} \right)$

and $\mathbb{E}\left\{ z^{(\ell)}_{\gamma\alpha} z^{(\ell)}_{\gamma\beta} \right\} \leq \frac{1}{2}$

row/col. norm

$$\Rightarrow \mu_{max} \leq \max_\beta \sum_{\alpha=1}^{m} |\vec{x}_\alpha \cdot \vec{x}_\beta| \, \frac{2L}{n_0} \left(1 - \frac{1}{\pi} \arccos\left(\frac{\vec{x}_\alpha \cdot \vec{x}_\beta}{\|\vec{x}_\alpha\| \|\vec{x}_\beta\|} \right) \right)$$

Also, $\mu_{max}^2 \le \sum_{\alpha=1}^{m} \sum_{\beta=1}^{m} (\vec{x}_\alpha \cdot \vec{x}_\beta)^2 \frac{4L^2}{n_0^2} \left(1 - \frac{1}{\pi} \arccos\left(\frac{\vec{x}_\alpha \cdot \vec{x}_\beta}{\|\vec{x}_\alpha\| \|\vec{x}_\beta\|}\right)\right)^2$

Frobenius

Let
$$C_D = \min \left\{ \max_\beta \sum_{\alpha=1}^{m} |\vec{x}_\alpha \cdot \vec{x}_\beta| \frac{2L}{n_0} \left(1 - \frac{1}{\pi} \arccos\left(\frac{\vec{x}_\alpha \cdot \vec{x}_\beta}{\|\vec{x}_\alpha\| \|\vec{x}_\beta\|}\right)\right), \sqrt{\sum_{\alpha=1}^{m} \sum_{\beta=1}^{m} (\vec{x}_\alpha \cdot \vec{x}_\beta)^2 \frac{4L^2}{n_0^2} \left(1 - \frac{1}{\pi} \arccos\left(\frac{\vec{x}_\alpha \cdot \vec{x}_\beta}{\|\vec{x}_\alpha\| \|\vec{x}_\beta\|}\right)\right)^2} \right\}$$

be dataset dependent. Then, $\mu_{max} \le C_D$.

the following

Proposition 3 (Pure weight moments for $K_N, \Delta K_N$). *We have*
$$\mathbb{E}[K_w] = \frac{d}{n_0} \|x\|_2^2.$$
Moreover,
$$\mathbb{E}[K_w^2] \simeq \frac{d^2}{n_0^2} \|x\|_2^4 \exp(5\beta) \left(1 + O\left(\sum_{i=1}^{d} \frac{1}{n_i^2}\right)\right), \qquad \beta := \sum_{i=1}^{d} \frac{1}{n_i}.$$
Finally,

Suppose $\mathbb{E}\{K_{\alpha\alpha}^2\} \le C_1 \frac{4d^2}{n_0^2} \|\vec{x}_\alpha\|^4 e^{5\beta}$

$$\Rightarrow \mathrm{Var}(K_{\alpha\alpha}) \le (C_1 e^{5\beta} - 1) \frac{4d^2}{n_0^2} \|\vec{x}_\alpha\|^4$$

$$\Rightarrow \sigma \le \sqrt{C_1 e^{5\beta} - 1} \cdot \frac{2d}{n_0} \|\vec{x}_\alpha\|^2$$

Chebyshev

$$\Rightarrow \mathbb{P}\left\{ K_{\alpha\alpha} \ge \left(1 + \sqrt{\frac{m}{\delta}} \sqrt{C_1 e^{5\beta} - 1}\right) \cdot \frac{2d}{n_0} \|x_\alpha\|^2 \right\} \le \frac{\delta}{m}$$

union bound

$$\Rightarrow \mathbb{P}\left\{ \mathrm{Tr}(K) \le \left(1 + \sqrt{\frac{m}{\delta}} \sqrt{C_1 e^{5\beta} - 1}\right) \cdot \frac{2d}{n_0} \sum \|\vec{x}_\alpha\|^2 \right\} \ge 1 - \delta$$

$$\Rightarrow R = \left(1 + \sqrt{\frac{m}{\delta}} \sqrt{C_1 e^{5\beta} - 1}\right) \cdot \frac{2d}{n_0} \sum \|\vec{x}_\alpha\|^2$$

Then, Matrix Chernoff gives

$$\mathbb{P}\left\{\lambda_{max}(K) \geq (1+\varepsilon)\mu_{max}\right\} \leq m\left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\mu_{max}/R}$$

$$\Rightarrow \mathbb{P}\left\{\lambda_{max}(k) \geq (1+\varepsilon)\, C_D\right\} \leq m\left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{C_m/R}$$

where

$$C_D = \min\left\{\frac{\max_{\beta} \sum_{\alpha=1}^{m} |\vec{x}_\alpha \cdot \vec{x}_\beta| \frac{2L}{n_0}\left(1 - \frac{1}{\pi}\arccos\left(\frac{\vec{x}_\alpha \cdot \vec{x}_\beta}{\|\vec{x}_\alpha\|\|\vec{x}_\beta\|}\right)\right)}{\sqrt{\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m} (\vec{x}_\alpha \cdot \vec{x}_\beta)^2 \frac{nL^2}{n_0^2}\left(1 - \frac{1}{\pi}\arccos\left(\frac{\vec{x}_\alpha \cdot \vec{x}_\beta}{\|\vec{x}_\alpha\|\|\vec{x}_\beta\|}\right)\right)^2}}\right\}$$

$$C_m = \frac{2L}{n_0} \max_{\alpha}\left\{\|\vec{x}_\alpha\|^2\right\}$$

$$R = \left(1 + \sqrt{\frac{m}{8}}\sqrt{c_1 e^{5\beta} - 1}\right) \cdot \frac{2d}{n_0} \sum \|\vec{x}_\alpha\|^2$$

Lecture 12/5 - Linear Regions

Consider a FC ReLU net:

$$\vec{z}_i^{(\ell+1)}(\vec{x}) = \vec{b}_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma(\vec{z}_j^{(\ell)})$$

with $n_{L+1} = 1$. Note that $\vec{x} \in \mathbb{R}^{n_0} \mapsto z^{(L+1)}(x) \in \mathbb{R}$ is continuous, piecewise linear.

One question we can ask is <u>how many pieces we get</u> in best/worst/avg cases?

We can use this result as a very rough measure of the complexity of ReLU nets.

Examples

Example 1

$n_0 = L = 1$

$$z^{(2)}(x) = b^{(2)} + \sum_{j=1}^{n_1} W_j^{(2)} \sigma(W_j^{(1)} x + b_j^{(1)})$$

As on the pset, we define breakpoints $\xi_j = -\frac{b_j^{(1)}}{W_j^{(1)}}$

Since $\frac{dz^{(2)}}{dx}$ is constant between break points,

$\#$ pieces $\le n_1 + 1$

Example 2

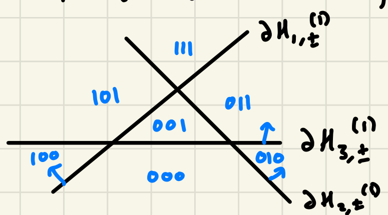
$n_0 \ge 2$, $L = 1$

$$z^{(2)}(\vec{x}) = b^{(2)} + \sum_{j=1}^{n_1} \sigma(\vec{W}_j^{(1)} \cdot \vec{x} + \vec{b}_j^{(1)})$$

For each $j = 1, \dots, n_1$, define

$$H_{j,\pm}^{(1)} = \{x \in \mathbb{R}^{n_0} \mid \operatorname{sgn}(\vec{w}_j^{(1)} \cdot \vec{x} + \vec{b}_j^{(1)}) = \pm 1\}$$

In $\mathbb{R}^2$, this makes a planar subdivision:



↑ = direction of on

Note that in each component of $\mathbb{R}^{n_0} \setminus \bigcup_{j=1}^{n_1} \partial H_{j,1}^{(1)}$ (the cells of the hyperplane arrangement)
each neuron is either on or off. Thus,
$\vec{\nabla}_x z^{(2)}(x)$ is constant on each cell of the hyperplane arrangement, and so

$$\# \text{ pieces} \le \underbrace{\# \text{ cells in arrangement of } n_1 \text{ hyperplanes in } \mathbb{R}^{n_0} \le \sum_{i=0}^{n_0} \binom{n_1}{i}}_{\substack{\text{Zaslavsky's Theorem,} \\ \text{equality when in general position}}} = \begin{cases} n_1^{n_0}/n_0! & n_1 \gg n_0 \\ 2^{n_1} & n_1 \le n_0 \end{cases}$$

General Setting

Def: A linear region is a <u>maximal</u> n_0-dimensional connected set on which $\vec{\nabla}_x z^{L+1}(\vec{x})$ is constant.

(Worst Case)

Lemma: For any n_0, L, any $n_1, \dots n_L$, $\quad \# \text{ linear regions} \le 3^{\# \text{ neurons}}$ ← each neuron partitions space in 3 parts

Proof: For each assignment of neuron on/offs

$$\vec{\varepsilon} = (\varepsilon_i^{(\ell)} \dots) \in \{-1, 0, 1\}^{[n_1] \times \dots \times [n_L]}$$

define $\quad P(\vec{\varepsilon}) = \{ x \in \mathbb{R}^{n_0} \mid \operatorname{sgn}(z_i^{(\ell)}(\vec{x})) = \varepsilon_i^{(\ell)} \}$.
Each $P(\vec{\varepsilon})$ is a region of input space with the same signs of preactivations, and so $\vec{\nabla}_x z^{(L+1)}(\vec{x})$ is constant on each $P(\vec{\varepsilon})$.
They also partition $\mathbb{R}^{n_0}$ (disjoint union), i.e. $\mathbb{R}^{n_0} = \coprod_{\vec{\varepsilon}} P(\vec{\varepsilon})$.

We want to show that each $P(\vec{\varepsilon})$ is a connected set. In fact, we will show that each $P(\vec{\varepsilon})$ is a <u>convex polytope!</u>

Write $\quad P(\vec{\varepsilon}) = \bigcap_{\ell=1}^{L} P^{(\ell)}(\vec{\varepsilon}) \quad$ and $\quad P^{\ell}(\vec{\varepsilon}) = \bigcap_{i=1}^{n_\ell} P_i^{(\ell)}(\vec{\varepsilon})$, where

$$P_i^{(\ell)}(\vec{\varepsilon}) = \{ x \in \mathbb{R}^{n_0} \mid \operatorname{sgn}(z_i^{(\ell)}(\vec{x})) = \varepsilon_i^{(\ell)} \}$$

Note that $P^{(1)}(\vec{\varepsilon})$ is a convex polytope because each

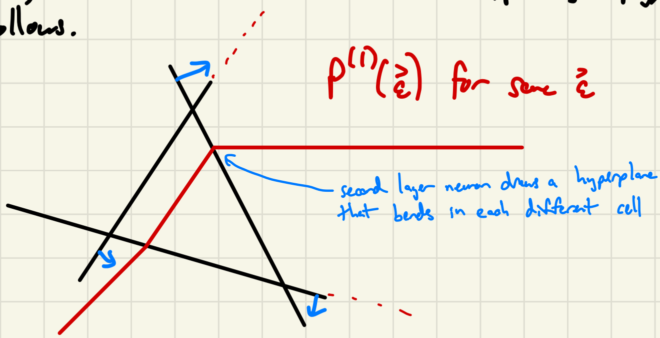
$$P_i^{(1)} = \{ x \in \mathbb{R}^{n_0} \mid \operatorname{sgn}(\vec{w}_i^{(1)} \cdot \vec{x} + \vec{b}^{(1)}) = \varepsilon_i^{(1)} \}$$

is either a half-space or a hyperplane. So, $P^{(1)}(\vec{\varepsilon}) = \bigcap_{i=1}^{n_1} P_i^{(1)}(\vec{\varepsilon})$ is a convex polytope.

Next, note that on $P^{(1)}(\vec{\varepsilon})$, if $\dim(P^{(1)}(\vec{\varepsilon})) = n_0$, $\vec{\nabla}_x z^{(2)}(\vec{x})$ is constant. So, $P^{(1)}(\vec{\varepsilon}) \cap P_i^{(2)}(\vec{\varepsilon})$ is the intersection of $P^{(1)}(\vec{\varepsilon})$ with a hyperplane or a half-space. Thus,

$$P^{(2)}(\vec{\varepsilon}) = \bigcap_{i=1}^{n_2} P^{(1)}(\vec{\varepsilon}) \cap P_i^{(2)}(\vec{\varepsilon})$$

is a convex polytope. Repeat inductively to see that $P(\vec{\varepsilon})$ is a convex polytope, and is therefore connected. Since there are $3^{\#\text{neurons}}$ of possible $\vec{\varepsilon}$'s, each of which makes a new (possibly empty) region, the result follows.



∎

<u>Upshot:</u> $L \geq 2 \Rightarrow$ # pieces grows quickly because "bent hyperplanes wiggle"

<u>Open Problems:</u>

* $\mathbb{P}\{\exists$ bounded bent hyperplane$\} = ?$

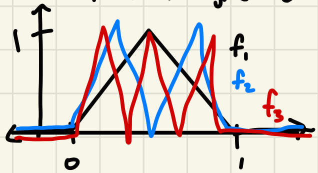
* for $n_0 = 2, L = 1$ $\mathbb{E}\{\#$ sides of polygon containing the origin$\}$?

(Worst case - exponential in # neurons - can exist)

<u>Theorem A:</u> (Telgarsky)

Suppose $n_0 = 1$. Then, $\exists$ a ReLU net with large enough L s.t.

- depth $= 2L$
- # neurons $= 3L-1$
- # linear regions $= 2^L$



<u>Proof:</u> Define $f(x) = \sigma(2\sigma(x) - 4\sigma(x - \frac{1}{2}))$

Let $f_\ell(x) = \underbrace{f \circ f \circ \cdots \circ f}_{\ell \text{ times}}$

So, f_L is a ReLU net with L spikes and so $2 + 2^L$ regions. ∎

(Avg. Case)

Theorem B: (Hanin-Rolnick)

Suppose $W_{ij}^{(\ell+1)} \sim \mathcal{N}(0, 2/n_\ell)$, $b_i^{(\ell+1)} \sim \mathcal{N}(0, C_b)$, $n_0 = 1$,

Then, $\mathbb{E}\{\#\text{ linear regions in } [a,b]\} \le C \cdot |a-b| \cdot \#\text{ neurons}$

Proof idea: look up co-area formula!

Coarea formula

From Wikipedia, the free encyclopedia

Open Problems:

* Count # of global regions (set $[a,b]$ to $\mathbb{R}$).

Lecture 12/7 - Bayesian Interpolation w/ Linear Nets

NNs have many large parameters:
- depth L
- width n_ℓ
- input dim. n_0
- # train datapoints P

We want to ask how do L, n_ℓ, n_0, P influence "model quality", i.e. feature learning, robustness, generalization, etc.

There are some challenges with any analysis:

① model is nonlinear in its parameters

② limits as $P, L, n_0, n_\ell \to \infty$ in different orders don't commute

Examples of non-commuting limits

Ex 1/ (Marchenko-Pastur)

Suppose $X \sim \mathbb{R}^{P \times n_0}$ with $X_{ij} \sim \mathcal{N}(0,1)$ and

sample covariance determines what linear regression do →

$$\Sigma_{n_0,P} = \frac{1}{n_0} X X^T \in \mathbb{R}^{P \times P}$$

Since $\Sigma_{n_0,P}$ is PSD, write $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_P \geq 0$ as eigenvalues of $\Sigma_{n_0,P}$ and

$$\mu_{n_0,P} \equiv \frac{1}{P} \sum_{j=1}^{P} \delta_{\lambda_j}$$ ← counting measure on eigenvalues

Theorem: If $n_0, P \to \infty$ with $P/n_0 \to \alpha \in (0,1)$, then

$$\mu_{n_0,P} \xrightarrow{w} \mu_{MP;\alpha}$$, where

converges in distribution weakly almost surely where $\quad d\mu_{MP;\alpha}(x) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{\alpha x}\mathbb{1}_{[\lambda_-,\lambda_+]}(x), \qquad \lambda_\pm = (1\pm\sqrt{\alpha})^2$

Ex 2/ Deep Linear Network

Consider $\quad z(\vec{x};\theta) = \underbrace{W^{(L+1)}\cdots W^{(1)}}_{\in\mathbb{R}^{n_0}}\vec{x} = \vec{\theta}^T\vec{x} \quad$ where $\quad W^{(\ell)}_{ij}\sim\mathcal{N}\left(0,\frac{1}{n_{\ell-1}}\right)$.

We have $\quad \vec{\theta} = \frac{\vec{\theta}}{\|\vec{\theta}\|}\|\vec{\theta}\|, \quad$ but $\quad \frac{\vec{\theta}}{\|\vec{\theta}\|}\sim \mathrm{Unif}(S^{n_0-1}) \perp \|\vec{\theta}\|$

uniform on sphere

Recall the following fact: if $W\in\mathbb{R}^{n\times m}$ has $W_{ij}\sim\mathcal{N}(0,\sigma^2)$, then $W = UWO$ for $U\in O(n)$, $O\in O(m)$

(rotation/reflection invariant!)

So, $\quad \|W^{(L+1)}\cdots W^{(1)}\| = \underbrace{\|W^{(L+1)}\|}_{\text{distributed as chi-squared!}} \left\|\frac{W^{(L+1)}}{\|W^{(L+1)}\|}W^{(L)}\cdots W^{(1)}\right\|$

$$\overset{d}{=}\left(\frac{1}{n_L}\chi^2_{n_L}\right)^{\frac{1}{2}}\underbrace{\left\|\frac{W^{(L+1)}}{\|W^{(L+1)}\|}W^{(L)}\cdots W^{(1)}\right\|}_{W^{(L)}\text{ rotationally invariant, can replace }\frac{W^{(L+1)}}{\|W^{(L+1)}\|}\text{ with }e_1}$$

$$\overset{d}{=}\ \cdots\ \overset{d}{=}\left(\prod_{\ell=0}^{L}\underbrace{\frac{1}{n_\ell}\chi^2_{n_\ell}}_{\text{indep. w/ mean 1 and var. }\frac{1}{n_\ell}}\right)^{\frac{1}{2}}$$

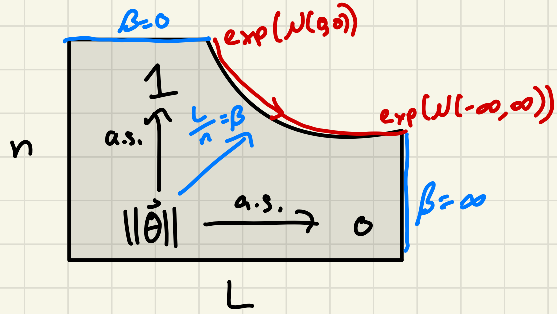
So, as $n\to\infty$, $\|\vec{\theta}\|\to 1$ almost surely.

However, we can also do

$$\|\vec{\theta}\| = \exp\left(\frac{1}{2}\sum_{\ell=1}^{L}\log\left(\frac{1}{n_\ell}\chi^2_{n_\ell}\right)\right)$$

$$\overset{n,L\gg1}{\approx}\exp\left(\mathcal{N}\left(-\frac{L}{4n},\frac{L}{4n}\right)\right)$$

So, as $L \to \infty$, $\|\vec{\theta}\| \to \exp(\mathcal{N}(-\infty,\infty)) = 0$.
This looks like the picture



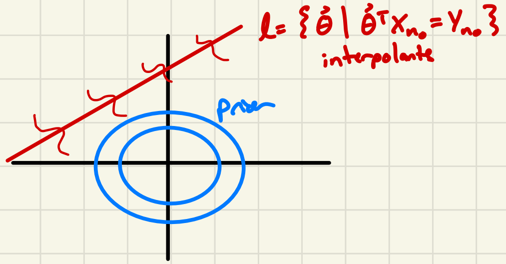
Bayesian Interpolation (Hanin + Alex Zlokapa)

Model: $z(\vec{x};\vec{\theta}) = W^{(L+1)} \cdots W^{(1)} \vec{x} = \vec{\theta}^T \vec{x}$

Data: $X_{n_0} = (\vec{x}_{i,n_0}) \in \mathbb{R}^{n_0 \times P}$, $\quad Y_{n_0} = (y_{j,n_0}) \in \mathbb{R}^{1 \times P}$

Prior: $W_{ij}^{(\ell)} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n_{\ell-1}}\right)$

NLL: $\mathcal{L}_D(\vec{\theta}) = \frac{1}{2} \|\vec{\theta}^T X_{n_0} - Y_{n_0}\|_2^2$ $\quad$ likelihood $\propto \exp\left(-\frac{\beta}{2}\mathcal{L}_D(\vec{\theta})\right)$



The Bayesian inference on each model is

$$d\mathbb{P}_{post}(\theta \mid X_{n_0}, Y_{n_0}, L, n_\ell, \sigma^2) = \lim_{\beta \to \infty} \frac{d\mathbb{P}_{prior}(\theta \mid L, n_\ell, \sigma^2) \times \exp\left(-\frac{\beta}{2}\mathcal{L}_D(\vec{\theta})\right)}{\underbrace{Z_\beta(X_{n_0}, Y_{n_0} \mid L, n_\ell, \sigma^2)}_{\text{partition function, normalizes distribution of posterior, } = \mathbb{P}\{\text{data} \mid \text{model}\}}}$$

Bayesian evidence

$$\propto \delta_\ell(\vec{\theta})\, \mathbb{P}_{prior}(\vec{\theta})$$

We can then perform Bayesian model selection to

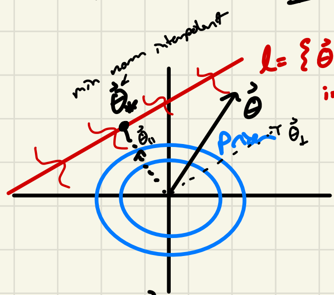
maximize $Z_\infty(X_{n_0}, Y_{n_0} \mid L, n_\ell, \sigma^2)$ $\iff$ MLE on space of NNs

maximize volume of interpolating models $\vec{\theta}$ over L, n_ℓ, σ^2 architecture

The results thus

* $\mathbb{P}_{post}(\vec{\theta})$, Z_∞ are exactly computable (not asymptotically!)
* Effective depth $P \cdot \frac{L}{n} \left(= P \sum_{\ell=1}^{L} \frac{1}{n_\ell}\right) = \lambda_{post}$ determines posterior!
* $\lambda_{post} \to \infty \Rightarrow$ optimal feature learning from data-agnostic priors ($\sigma^2 = 1$)

Claim: Any $\vec{\theta}$ can be decomposed into $\vec{\theta}_\parallel + \vec{\theta}_\perp$, where $\vec{\theta}_\parallel \in Col(X_{n_0})$ and $\vec{\theta}_\perp \in Col(X_{n_0})^\perp$



We claim if $\vec{\theta} \sim \mathbb{P}_{post}$, then

$\vec{\theta}_\parallel = \vec{\theta}_*$

$$\vec{\theta} = \vec{\theta}_* + u\|\vec{\theta}_\perp\|,$$

where $u \sim Unif(S\, col(X_{n_0})^\perp)$ independently of $\|\vec{\theta}_\perp\|$

Interpolation For a test point $\vec{x}$, $\vec{x} = \vec{x}_\parallel + \vec{x}_\perp$ by projection onto X_{n_0}. Then,

$$f(\vec{x}) = \vec{\theta}^T\vec{x} = \vec{\theta}_*^T\vec{x}_\parallel + (u\vec{x}_\perp)\|\vec{\theta}_\perp\|$$
$$\approx \mathcal{N}\left(\vec{\theta}_*^T\vec{x}_\parallel, \frac{\|x_\perp\|^2}{n_0 - P}\|\theta_\perp\|^2\right)$$

this is how Bayesian inference can learn features!

So, $\|\vec{\theta}_\perp\|$ controls overall prediction scale in unseen directions!

Theorem: Suppose $n_0, P \to \infty$ with $P/n_0 \to \alpha \in (0,1)$ s.t.

$$\|\vec{\theta}_{*,n_0}\| \xrightarrow{d} \|\vec{\theta}_*\|$$

Then,

$$\theta_*^2 = \underset{\theta^2}{\operatorname{argmax}} \lim_{\substack{n_0, P \to \infty \\ P/n_0 \to \alpha}} Z_\infty(X_{n_0}, Y_{n_0} \mid L, n_\ell, \theta^2)$$

gives

$$\lim_{P, n_0 \to \infty} \mathbb{P}_{post}(\|\vec{\theta}_*\| \mid \theta^2 = \theta_*^2, L, n_\ell) = \delta_{\frac{\|\vec{\theta}_*\|^2}{\alpha}}$$

best data-dependent prior doesn't depend on architecture

Furthermore,

$$\lim_{\substack{P, n_0 \to \infty \\ P/n_0 \to \alpha}} \mathbb{P}_{post}(\|\vec{\theta}_\perp\| \mid \theta^2 = 1, L, n_\ell) = \delta_{\frac{\|\vec{\theta}_*^{\,2}\|}{\alpha} z(\lambda)}$$

As $\lambda \to \infty$, $z(\lambda) \to 1$.

in ∞ effective depth, we match best prior in data-agnostic way