LECTURE 2: GENERALIZATION AND UNIFORM CONVERGENCE THEORY

SUPERVISED LEARNING PROBLEM:

- Data: $\{(y_i, x_i)\}_{i \leq n}$ $y_i \in \mathbb{R}$ $x_i \in \mathbb{R}^d$ iid $\sim \mathbb{P}$
- Loss fct: $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$
- Goal: Fit a predictor $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ that has small

<u>Expected/population/test risk</u>: $R(\hat{f}) := \mathbb{E}_{(y,x) \sim \mathbb{P}}[\ell(y, \hat{f}(x))]$

<u>Empirical risk</u>: $\hat{R}_n(f) := \frac{1}{n} \sum_{i \leq n} \ell(y_i, f(x_i))$

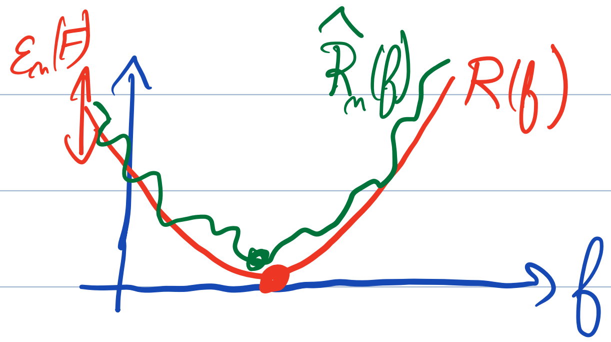
<u>ERM</u>: $\hat{f}_n := \underset{f}{\operatorname{argmin}} \{ \hat{R}_n(f) \text{ s.t. } f \in \mathcal{F} \}$ class of models

<u>Generalizat° question</u>: → what is $R(\hat{f}_n)$?
→ how to guarantee that we do well on new unseen data?

<u>"Classical" idea</u>: Uniform convergence:

If $\mathcal{E}_n(F) := \sup_{f\in F} |\hat{R}_n(f) - R(f)|$ is small

then $R(\hat{f}_n) \approx \inf_{f\in F} R(f)$



"Proof":

$$f_* = \operatorname*{argmin}_{f\in F} R(f)$$

$$R(\hat{f}_n) = R(f_*) + [R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)] + \underbrace{[\hat{R}_n(\hat{f}_n) - \hat{R}_n(f_*)]}_{\leq 0}$$

$$+ [\hat{R}_n(f_*) - R(f_*)]$$

$$\leq R(f_*) + 2 \underbrace{\sup_{f\in F} |R(f) - \hat{R}_n(f)|}_{\equiv: \mathcal{E}_n(F)} \quad \square$$

Rmk 1: $\hat{R}_n(f) - R(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)) - \mathbb{E}[\ell(y, f(x))]$

For fixed f $\quad = O(1/\sqrt{n})$ CLT

$\rightarrow$ here we want a bound that holds uniformly over all $f\in F$

Rmk 2: In classical statistical learning theory,

"a model learnt from data is effective at predicting on new data" is a consequence of U.C.

→ It is the mechanism that guarantees generalizat°

→ Highly influential idea: research then consists in carefully crafting fct classes such that $\varepsilon_n(F)$ small

↳ constrain number of parameters OR regularize weights

↳ in classical SLT: models are underparametrized or over-constrained

→ In Deep Learning, it seems good generalization happens thanks to a completely different principle

⇒ More in 2 next lectures

→ UC Theory still highly informative

Also help to compare and contrast what happens in DL

→ dependency on different parameters

GOAL rest of lecture

* Basic definitions + tools to bound $\varepsilon_n(F)$
* Ex 1: F = 2-layer NNs
* Ex 2: F = Multilayer NNs

Def [Rademacher complexity] • $(\mathcal{Z}, \mathbb{P})$ proba space
• $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ class of fcts $g : \mathcal{Z} \to \mathbb{R}$

$$Rd_m(\mathcal{G}) := \mathbb{E}\left[\sup_{g \in \mathcal{G}} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \, g(z_i) \right| \right]$$

where $z_1, \dots, z_m \underset{iid}{\sim} \mathbb{P}$
$\varepsilon_1, \dots, \varepsilon_m \underset{iid}{\sim} \mathrm{Unif}(\{\pm 1\})$

Rmk: In our case, $z_i = (y_i, x_i)$ $\quad g(z_i) = \ell(y_i, f(x_i))$
$\mathcal{G} = \ell \circ F$

Lemma 1:
$$\frac{1}{2} Rd_m(\mathcal{G}) \leq \mathbb{E}\left[\sup_{g \in \mathcal{G}} \left| \frac{1}{m} \sum_{i=1}^{m} g(z_i) - \mathbb{E} g \right| \right] \leq 2 \, Rd_m(\mathcal{G})$$

→ will only prove this upper bound

Proof: [In this lecture, we are only interested in upper-bound]
Decoupling argument:

$$\mathbb{E}_z \sup_g \left| \frac{1}{m} \sum_{i=1}^{m} g(z_i) - \mathbb{E} g \right|$$

$$= \mathbb{E}_z \sup_g \left| \mathbb{E}_{z'} \left[\frac{1}{m} \sum_{i=1}^{m} g(z_i) - g(z_i') \right] \right|$$

$$\leq_{\text{(Jensen's)}} \mathbb{E}_{\substack{Z, Z' \\ \sigma}} \sup_g \left| \frac{1}{n} \sum_{i=1}^n [g(Z_i) - g(Z_i')] \sigma_i \right|$$ (5)

$$= \mathbb{E}_{\substack{Z, Z' \\ \sigma}} \sup_g \left| \frac{1}{n} \sum_{i=1}^n g(Z_i) \sigma_i - \frac{1}{n} \sum_{i=1}^n g(Z_i') \sigma_i \right|$$

$$\leq_{\text{(triangular ineq)}} 2 \, Rd_n(\mathcal{G}) \qquad \square$$

Rmk: • This implies $\mathbb{E}[R(\hat{f}_n)] \leq \inf_{f \in F} R(f) + 2 Rd_n(\ell \circ F)$

• To show without expectation, need to show $\sup_{f \in F} |\hat{R}_n(f) - R(f)|$ concentrates around its $\mathbb{E}[\cdot]$

E.g. Markov's inequality: w.p. $\geq 1-\delta$

$$R(\hat{f}_n) \leq \inf_{f \in F} R(F) + \frac{2}{\delta} Rd_n(\ell \circ F)$$

Can improve using concentrat°: e.g. $\sqrt{\log(\frac{1}{\delta})}$ if ℓ is bounded by McDiarmid's bounded difference inequality

Right now: $\mathrm{Rad}_m(\ell \circ \mathcal{F})$ $\longrightarrow$ want $\mathrm{Rad}_m(\mathcal{F})$

Lemma 2: [Contraction inequality]

$\phi_1, \dots, \phi_m : \mathbb{R} \to \mathbb{R}$ λ-Lipschitz $\phi_i(0) = 0$

$$\mathbb{E}_{\substack{Z\\ \sigma}}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \phi_i(f(Z_i))\, \sigma_i \right| \right] \leq \lambda\, \mathbb{E}_{\substack{Z\\ \sigma}}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} f(Z_i)\, \sigma_i \right| \right]$$

Proof: (skipped during class) Add in lecture notes

you can try to do it by yourself

$\Rightarrow$ Can reduce to ourselves to $\lambda = 1$

and $\phi_2 = \dots = \phi_m(t) = t$

$\Rightarrow$ Can do $\phi_1(f(Z_1)) \longrightarrow f(Z_1)$ "by hand"

(Add it in the lecture notes)

Bound on $\mathcal{E}_n(F)$: $\quad \ell(y,\cdot)$ L-lipschitz

$$\mathbb{E}[\mathcal{E}_n(F)] \leq 2\,Rd_n(\ell\circ F)$$

$$Rd_n(\ell\circ F) \leq \mathbb{E}\ \left|\frac{1}{n}\sum_{i=1}^{n}\ell(y_i,0)\varepsilon_i\right|$$

$$+\mathbb{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left[\ell(y_i,f(x_i)) - \ell(y_i,0)\right]\right|$$

$\leq$ Jensen's

$$\leq \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j}\varepsilon_i\varepsilon_j\,\ell(y_i,0)\,\ell(y_j,0)\right]^{1/2}$$

$$+ L\,\mathbb{E}\sup_{f\in F}\left|\frac{1}{n}\sum\varepsilon_i f(x_i)\right|$$

$$\leq \frac{1}{\sqrt{n}}\,\mathbb{E}\left[\ell(y,0)^2\right]^{\frac{1}{2}} + L\cdot Rd_n(F)$$

Summary: $\mathbb{E}[R(\hat{f}_n)] \leq \inf_{f\in F} R(f) + \frac{C}{\sqrt{n}}\,\mathbb{E}\left[\ell(y,0)^2\right]^{\frac{1}{2}}$

$$+ C\,L\cdot Rd_n(F)$$

$\boxed{\text{Example 1:}}$ $\mathcal{F} :=$ 2-layer neural nets

$$f(x;\theta) = \frac{1}{M}\sum_{j=1}^{M} a_j \,\sigma(\langle w_j, x\rangle) \qquad x \in \mathbb{R}^d$$

$$\theta = \{(a_j, w_j)\}_{j\le m}$$

Assume: $x \in B_2^d(C_x\sqrt{d})$ = ball of radius $C_x\sqrt{d}$

$$\mathcal{F} = \begin{cases} w_j \in B_2^d(C_w) \\ a = (a_1, \dots, a_M) \in S_{p,M} = \left\{ \frac{1}{M}\sum_{i=1}^{M} |a_i|^p \le r_0^p \right\} \end{cases}$$

$$= \left\{ \|a\|_p \le r_0 M^{1/p} \right\}$$

$$= B_p^M(r_0 M^{1/p})$$

By Jensen's $S_{p,M} \subseteq S_{p',M} \quad p' < p$

$$Rd_m(\mathcal{F}) = \mathbb{E} \sup_{a,W} \left| \frac{1}{m}\sum_{i=1}^{m} z_i \frac{1}{M}\sum_{\ell=1}^{M} a_\ell \,\sigma(\langle w_\ell, x_i\rangle)\right|$$

$$= \frac{1}{M}\mathbb{E} \sup_{W} \sup_{a\in S_{p,M}} \left\langle a, \frac{1}{m}\sum_{i=1}^{m} z_i \,\sigma(W x_i)\right\rangle$$

$$\sup_{\|a\|_p\le C} \langle a, v\rangle = C\|v\|_q \qquad q = \frac{p}{p-1}$$

$$= n_0 M^{\frac{1}{p}-1} \mathbb{E} \sup_W \left\| \frac{1}{m} \sum_{i=1}^m z_i \sigma(W x_i) \right\|_q$$

$$= n_0 M^{\frac{1}{p}-1} \mathbb{E} \sup_{w_1 \dots w_M} \left(\sum_{\ell=1}^M \left| \frac{1}{m} \sum_{i=1}^m z_i \sigma(\langle w_\ell, x_i \rangle) \right|^q \right)^{\frac{1}{q}}$$

$$M \sup_{w_\ell}$$

$$= n_0 M^{\frac{1}{p}+\frac{1}{q}-1} \mathbb{E} \sup_w \left| \frac{1}{m} \sum_{i=1}^m z_i \sigma(\langle w, x_i \rangle) \right|$$

$= n_0$ Rd of one neuron $\rightarrow$ indep of M and p

Contraction

$$\leq n_0 L_\sigma \mathbb{E} \sup_w \left| \langle w, \frac{1}{m} \sum_i z_i x_i \rangle \right|$$

$$= n_0 L_\sigma C_w \mathbb{E} \left\| \frac{1}{m} \sum_i z_i x_i \right\|_2$$

$$\leq n_0 L_\sigma C_w \mathbb{E} \left[\left\| \frac{1}{m} \sum_i z_i x_i \right\|_2^2 \right]^{1/2} \qquad \leq \frac{1}{m^2} \sum_i \mathbb{E} \|x_i\|_2^2$$

$$\leq n_0 L_\sigma C_w \frac{1}{\sqrt{m}} \mathbb{E} \left[\|X_1\|^2 \right]^{1/2}$$

$$\leq n_0 L_\sigma C_w C_x \sqrt{\frac{d}{m}}$$

Conclusion: $\boxed{\mathrm{Rd}_m(\ell\circ\mathcal{F}) \leq \frac{C}{\sqrt{m}} + L\cdot L_\sigma \cdot r_0 C_\omega C_x \sqrt{\frac{d}{m}}}$

Rmk: ① Bound independent of M

↳ NNs can generalize even with $M=\infty$

as long as $\|a\|_p \ll \sqrt{\frac{m}{d}}$ [Bartlett, 1996]

1^{st} realized : # param not the right "complexity" measure (VC dim) → norm of the parameters.

Natural $M=\infty$: μ Prob meas on $B_2^d(C_\omega)$

$$\mathcal{F}_p(r_0) := \left\{ f(x) = \int \sigma(\langle \omega, x\rangle)\, a(\omega)\, \mu(d\omega) \quad \|a\|_{L^p(\mu)} = \left(\int |a(\omega)|^p \mu(d\omega)\right)^{1/p} \leq r_0 \right\}$$

$$\mathrm{Rd}_m(\mathcal{F}_p(r_0)) \leq L_\sigma\, r_0\, C_x C_\omega \sqrt{\frac{d}{m}}$$

② Independent of p: $\mathcal{F}_p(r_0) \subset \mathcal{F}_{p'}(r_0) \subset \mathcal{F}_1(r_0)$

$1 < p' < p$

↳ strongest case for $\underline{p=1}$ $\quad \int |a(\omega)|\, \mu(d\omega) \leq r_0$

"Convex Neural networks"

$$\overline{\mathcal{F}_1(r_0)} = \left\{ f(x) = \int \sigma(\langle \omega, x \rangle)\, \nu(d\omega) : \|\nu\|_{TV} \le r_0 \right\}$$

Total variation

$\nu = \nu_+ - \nu_-$ $\quad \|\nu\|_{TV} = \nu_+(\Omega) + \nu_-(\Omega)$ ← signed measure

→ Will go back to this model later this semester

<u>Rmk:</u> $p = 2$ $\quad \mathcal{F}_2$ is a RKHS

(also talk about it later this semester)

EX 2: Multilayer NNs

[Bartlett, Foster, Telgarsky, 2017]
("spectral-based" generalization bound for multi-layer)
(Lipschitz "norm of the NN")

Here slightly simplified setting

$$f(x; W) = \sigma \circ W_L \circ \sigma \circ W_{L-1} \circ \dots \circ \sigma \circ W_1 x$$

* $\sigma : \mathbb{R} \to \mathbb{R}$ 1- Lipschitz $\sigma(0) = 0$
* $W = (W_L, W_{L-1}, \dots, W_1)$

$W_L \in \mathbb{R}^{1\times d}$ $W_{L-1}, \dots, W_2 \in \mathbb{R}^{M\times M}$ $W_1 \in \mathbb{R}^{M\times d}$

$$\Theta(b_1 \dots b_L, s_1, \dots, s_L) = \{ \|W_\ell\|_{op} \leq s_\ell \quad \|W_\ell\|_{1,2} \leq b_\ell \}$$

$$\|W\|_{1,2} = \| (\underbrace{\|W_{\cdot,1}\|_1}_{\ell_1 \text{ norm of columns.}}, \dots, \|W_{\cdot,M}\|_1) \|_2$$

$$F := \{ f(\cdot, W) : W \in \Theta(b_1, \dots, b_L, s_1, \dots s_L) \}$$

Goal: bound $Rd_n(F)$

$$X = [x_1, \dots, x_d] \in \mathbb{R}^{d \times m} \qquad \text{data}$$

$$F_m(X) := \{ f_m(W) : W \in \Theta \} \subseteq \mathbb{R}^m$$

$$f_m(W) = \left(f(x_1; W), \dots, f(x_m; W) \right)$$

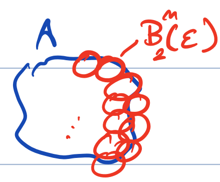
$\boxed{\text{Approach:}}$

Def [Conditional Rademacher complexity]

$$Rd_m(F_m(X)) := \mathbb{E}_z \sup_{W \in \Theta} \frac{1}{m} \sum_{i=1}^{m} z_i f(x_i; W)$$

$$= \mathbb{E}_z \sup_{W \in \Theta} \frac{1}{m} \langle f_m(W), z \rangle$$

Rmk: $Rd_m(F) = \mathbb{E}_X Rd_m(F_m(X))$



Def: [Covering number] $A \subseteq \mathbb{R}^m$

$\mathcal{N}(A; \varepsilon) :=$ min # of ℓ_2 balls of radius ε to cover A

Lemma 3: Assume $0 \in \mathcal{F}_n(X) \subseteq [0,1]^n$

$$Rd_n(\mathcal{F}_n(X)) \leq \inf_{\alpha \geq 0} \left\{ \frac{2\alpha}{\sqrt{n}} + \frac{24}{n} \int_\alpha^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}_n(X), \varepsilon)}\, d\varepsilon \right\}$$

"Dudley integral" "metric entropy"

Proof: Classical "chaining" argument [Vershynin]
$\longrightarrow$ will add in the notes

Goal: Bound $\mathcal{N}(\mathcal{F}_n(X), \varepsilon)$, i.e, construct ε-covering

$$\mathcal{F}_n(X) := \{ f_n(W) : W \in \Theta = \Theta_1 \times \dots \times \Theta_L \}$$

$$\Theta_\ell = \{ W_\ell : \|W_\ell\|_{op} \leq s_\ell,\ \|W_\ell\|_{1,2} \leq b_\ell \}$$

$$f_n(W) = \sigma \circ W_L \circ \dots \circ \sigma \circ A_1 X \in \mathbb{R}^n$$

$$f_n^{(\ell)}(W^\ell) = \sigma \circ W_\ell \circ \dots \circ \sigma \circ A_1 X \in \mathbb{R}^{M \times n}$$

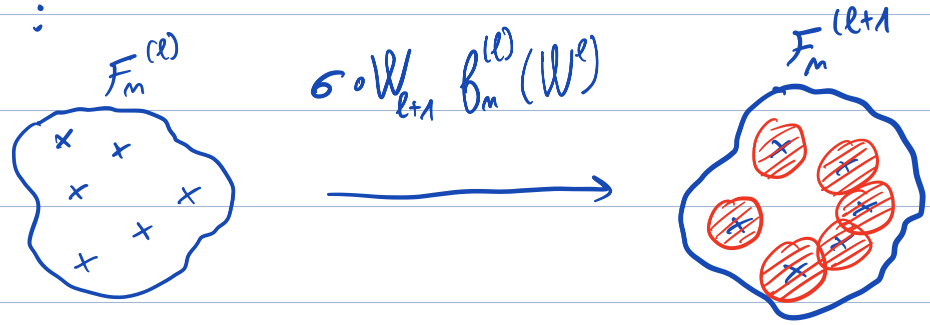
output of intermediary layers

$$f_n^{(\ell+1)}(W^{\ell+1}) = \sigma \circ W_{\ell+1}\, f_n^{(\ell)}(W^\ell)$$

Idea: Construct covering $\mathcal{C}_L$ by constructing sequentially coverings $\mathcal{C}_\ell$ δ_ℓ covering of

$$F_m^{(\ell)}(X) := \left\{ f_m^{(\ell)}(W^\ell) : W^\ell \in \Theta^\ell \right\} \subseteq \mathbb{R}^{M \times m}$$

Given $\mathcal{C}_\ell$ δ_ℓ-covering of $F_m^{(\ell)}$, we construct $\delta_{\ell+1}$-covering $\mathcal{C}_{\ell+1}$ as follows:



Rmk: ε-covering will be ε-cov. of $\sigma(W_{\ell+1} z)$ by Lipschitzness

$$z \in \mathcal{C}_\ell \longrightarrow \left\{ W_{\ell+1} z : W_{\ell+1} \in \Theta_{\ell+1} \right\}$$

$B_{\ell+1}(z; \varepsilon_{\ell+1})$ $\varepsilon_{\ell+1}$ covering for each of these sets $\to$ union gives $\mathcal{C}_{\ell+1}$

$$|\mathcal{C}_{\ell+1}| \leq \sum_{z \in \mathcal{C}_\ell} |B_{\ell+1}(z; \varepsilon_{\ell+1})| \leq |\mathcal{C}_\ell| \max_{z \in \mathcal{C}_\ell} |B_{\ell+1}(z; \varepsilon_{\ell+1})|$$

$$\delta_{\ell+1} \leq \sup_{W_{\ell+1} \in \Theta_{\ell+1}} \|W_{\ell+1}\|_{op} \, \delta_\ell + \varepsilon_{\ell+1}$$

$$\left[\begin{aligned} f &= W_{\ell+1} \tilde{f} \quad \to \delta_\ell \text{ close to } z \in \mathcal{C}_\ell \\ &= \underline{W_{\ell+1}(\tilde{f} - z)} + \underbrace{W_{\ell+1} z}_{\varepsilon_{\ell+1}} \end{aligned} \right]$$

Hence $\delta_{\ell+1} \leq s_{\ell+1} \delta_\ell + \varepsilon_{\ell+1}$ $\qquad \delta_L \leq \sum_{\ell=1}^{L} \prod_{j=\ell+1}^{L} s_j \cdot \varepsilon_\ell$

$$\sup_{z \in \mathcal{C}_\ell} |B_{\ell+1}(z; \varepsilon_{\ell+1})| \leq \sup_{f_m \in F_m^{(\ell)}} \mathcal{N}(T_{\ell+1}(f_m) ; \varepsilon_{\ell+1})$$

where $T_\ell(f_m) := \{ W_\ell f_m : W_\ell \in \Theta_\ell \}$

$$\|W_\ell\|_{op} \leq s_\ell$$

$$\|W_\ell\|_{1,2} \leq b_\ell$$

$$\boxed{\log N(F_m; \varepsilon) \leq \sum_{\ell=1}^{L} \sup_{f_m \in F_m^{(\ell)}} \log N(T_\ell(f_m) ; \varepsilon_\ell)}$$

$$\boxed{\varepsilon := \sum_{\ell=1}^{L} \prod_{j=\ell+1}^{L} s_j \cdot \varepsilon_\ell}$$

Lemma 4: $T(Z; b) := \{ A Z : \|A\|_{1,2} \leq b \}$ $\qquad A \in \mathbb{R}^{M \times M}$, $Z \in \mathbb{R}^{M \times N}$

$$\log N(T(Z, b); \varepsilon) \leq 2 \left(\frac{b \|Z\|_F}{\varepsilon} \right)^2 \log(2M)$$

Will prove later if I have the time.

(this lemma is the reason for $\|W_\ell\|_{1,2} \leq b_\ell$ contraint)

$$R_\ell := \sup_{f_m \in \mathcal{F}_m^{(\ell-1)}} \log \mathcal{N}\left(T_\ell(f_m); \varepsilon_\ell\right)$$

$$\leq 2 \left(\frac{b_\ell}{\varepsilon_\ell} \sup_{f_m \in \mathcal{F}_m^{(\ell-1)}} \|f_m\|_F \right)^2 \log(2M)$$

$$f_m^{(\ell-1)}(W^{\ell-1}) = \sigma \circ W_{\ell-1} \circ \sigma \dots \circ \sigma \circ W_1 X \in \mathbb{R}^{M \times m}$$

$$\|f_m^{(\ell-1)}\|_F \leq \|W_{\ell-1}\|_{op} \dots \|W_1\|_{op} \|X\|_F$$

$$\leq \|X\|_F \cdot \prod_{i=1}^{\ell-1} s_i$$

$$X_\ell \leq 2 \left(\frac{b_\ell}{\varepsilon_\ell} \prod_{i=1}^{\ell-1} s_i \right)^2 \|X\|_F^2 \log(2M)$$

$$\log \mathcal{N}\left(\mathcal{F}_m(X); \varepsilon\right) \leq 2 \sum_{\ell=1}^{L} \left(\frac{b_\ell}{\varepsilon_\ell} \prod_{i=1}^{\ell-1} s_i \right)^2 \|X\|_F^2 \log(2M)$$

$$\varepsilon = \sum_{\ell=1}^{L} \underbrace{\prod_{j=\ell+1}^{L} s_j \, \varepsilon_\ell}_{=: \alpha_\ell \cdot \varepsilon}$$

$$\sum_{\ell=1}^{L} \alpha_\ell = 1$$

optimize over them!

$$\log \mathcal{N}(\mathcal{F}_m, \varepsilon) \leq \frac{2\|X\|_F^2 \log 2M}{\varepsilon^2} \sum_{\ell=1}^{L} \left(\frac{b_\ell}{\alpha_\ell} \prod_{j \neq \ell} s_j \right)^2$$

$$= \left(\prod_{j=1}^{L} s_j \right)^2 \sum_{\ell=1}^{L} \left(\frac{b_\ell}{\alpha_\ell s_\ell} \right)^2$$

$$c_\ell := \frac{b_\ell}{s_\ell} \quad \Rightarrow \quad \min \sum_{\ell=1}^{L} \left(\frac{c_\ell}{\alpha_\ell} \right)^2 \quad \text{s.t.} \quad \sum_{\ell=1}^{L} \alpha_\ell = 1, \quad \alpha_\ell \geq 0$$

Convince yourself: $\alpha_\ell = \dfrac{c_\ell^{2/3}}{\sum_{\ell'} c_{\ell'}^{2/3}}$ [Hölder's inequality]

$$Q := \left(\sum_{\ell=1}^{L} \left(\frac{b_\ell}{s_\ell} \right)^{2/3} \right)^{3/2} \left(\prod_{i=1}^{L} s_i \right) \|X\|_F \sqrt{\log 2M}$$

$$\Rightarrow \boxed{\log \mathcal{N}(\mathcal{F}_m ; \varepsilon) \leq 2 \frac{Q^2}{\varepsilon^2}}$$

$$Rd_m(\mathcal{F}_m) \leq \frac{2\alpha}{\sqrt{m}} + \frac{24}{m} \int_\alpha^{\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}_m ; \varepsilon)} \, d\varepsilon$$

$$\leq \frac{2\alpha}{\sqrt{m}} + \frac{24}{m} Q \sqrt{2} \int_\alpha^{\sqrt{m}} \frac{1}{\varepsilon} \, d\varepsilon \quad \to \log \frac{\sqrt{m}}{\varepsilon}$$

$$\overset{\alpha = \frac{1}{m}}{\leq} \frac{100}{m} Q \log(m)$$

Conclusion: $\quad Rd_n(\mathcal{F}_n(X)) \leq C \cdot R_0 \cdot \|X\|_F \dfrac{\log n}{n} \cdot \sqrt{\log(2M)}$

where "spectral complexity" of the network is given by

$$R_0 := \left(\sum_{\ell=1}^{L} \left(\frac{b_\ell}{s_\ell} \right)^{2/3} \right)^{\frac{3}{2}} \prod_{i=1}^{L} s_i$$

$$\|W_\ell\|_{op} \leq s_\ell \qquad \|W_\ell\|_{1,2} \leq b_\ell$$

Rmk:
- Mild dependency on $\log(M)$
- scales with "Lipschitz cste" $\prod_{i=1}^{L} s_i$ of network
- dependency on # layers $\propto L^{\frac{3}{2}}$ (outside Lipschitz constant)

EX: $\|X_i\|_2 \asymp C\sqrt{d} \qquad W_\ell$ iid $\frac{1}{M}$ sub-Gaussian

$\|X\|_F \asymp \sqrt{nd}$

$\|W_\ell\|_{op} \asymp c \qquad \|W_\ell\|_{1,2} \asymp M$

$$Rd_n(\mathcal{F}_n) \asymp \sqrt{\frac{d M^2}{n}}\, L^{3/2} c^L$$

same issue I talked last week

For $L = O(1)$: requires $n \gg M^2 d \gg \underbrace{M^2 \vee Md}_{\#\text{parameters}}$

Can these UCs explain why NNs generalize?
(very different opinions, especially about right complexity measure)

- IF NNs verify these bounds → explain why it generalizes
- Some correlat° between real & gen bounds
  ↳ lots of empirical studies: overall none seem highly predictive of performance
- "Requires rethinking gene" → when interpolate break down

↳ above bound "cons"
- exponential dependency on L
- no assumption on the distribution of the data
- need # samples >> # parameters

Proof of Lemma 4

$$T(Z; b) := \{ AZ : \|A\|_{1,2} \le b \}$$

$$AZ = \underbrace{(A \odot C)}_{B} \hat{Z} \qquad Z = \begin{pmatrix} c_1 & & 0 \\ & \ddots & \\ 0 & & c_M \end{pmatrix} \hat{Z}$$

has normalized rows

$$c_i = \|Z_{i,\cdot}\|_2$$

$$C = \begin{pmatrix} c_1 & c_2 & \cdots & c_M \\ \vdots & \vdots & & \vdots \\ c_1 & c_2 & & c_M \end{pmatrix} \Big\updownarrow M$$

$$\overset{\longleftrightarrow}{M}$$

$$A \operatorname{diag}\begin{pmatrix} c_1 & & \\ & \ddots & \\ & & c_m \end{pmatrix} = A \odot C$$

$$\|B\|_1 = \sum_{ij} |B_{ij}| = \langle C, |A| \rangle \le \|C\|_{\infty,2} \|A\|_{1,2}$$

$$\left(\|M\|_{p,q} = \left\| \left(\|M_{\cdot,1}\|_p, \dots, \|M_{\cdot,m}\|_p \right) \right\|_q \right)$$

$$\langle X, Y \rangle \le \|X\|_{p,s} \|Y\|_{q,r} \qquad \frac{1}{p} + \frac{1}{q} = 1$$

$$\frac{1}{s} + \frac{1}{r} = 1$$

$$\|C\|_{\infty,2} = \|(|c_1|, \dots, |c_M|)\|_2 = \left(\sum_i \|Z_{i,\cdot}\|_2^2\right)^{\frac{1}{2}}$$
$$= \|Z\|_F .$$

Hence $\|B\|_1 \leq \|Z\|_F \cdot \|A\|_{1,2} \leq \|Z\|_F \cdot b =: \tilde{b}$

$$T(z_i\, b) \subseteq \{ B\hat{Z} : \|B\|_1 \leq \tilde{b} \}$$

Lemma 5: Let $u = \sum_{i=1}^{p} \lambda_i v_i$ vectors v_i, u, $\lambda_i \geq 0$

$\Rightarrow \exists\, k_1, \dots k_p \geq 0$ integers $\sum_{i=1}^{p} k_i = k$

s.t. $\left\| u - \frac{\|\lambda\|_1}{k} \sum_{i=1}^{p} k_i v_i \right\|_2^2 \leq \frac{\|\lambda\|_1^2}{k} \max_{i \leq p} \|v_i\|_2^2$

Rmk: Result does not depend on the dimension of v_i's
(hold for $\dim(v_i) = \infty$)

Proof: [Add a proof in the notes.]

Corollary: $S(v_1, \dots, v_p ; r) := \{ u = \sum \lambda_j v_j : \|\lambda\|_1 \leq r \}$

$$N(S ; \varepsilon) \leq (2p)^k$$

where $\varepsilon^2 = \frac{r^2}{k} \max \|v_i\|_2^2$

We have $T(Z, b) \subseteq \{ B\hat{Z} : \|B\|_1 \leq b \}$

$$B\hat{Z} = \sum_{i,j=1}^{M} B_{ij} \left[e_i e_j^T \hat{Z}\right] = \sum_{ij=1}^{M} B_{ij} \quad V_{ij} \in \mathbb{R}^{M \times m}$$

Above corollary gives $\mathcal{N}(\{B\hat{Z} : \|B\|_1 \leq b\}; \varepsilon) \leq (2M^2)^k$

$$\varepsilon^2 = \frac{b^2}{k} \max_{ij} \|e_i e_j^T \hat{Z}\|_F^2 = \frac{b^2}{k}$$

$$= \|\hat{Z}_{j\cdot}\|_2^2 = 1$$

$$\Rightarrow \log \mathcal{N}(T(Z, b), \varepsilon) \leq k \log (2M)^2$$

$$\leq 2 \left(\frac{b}{\varepsilon}\right)^2 \log(2M) \qquad \square$$