Neural Net Classifiers and their Rademacher Complexities

$x \in \mathcal{X}$ $\qquad g_f(x) = \operatorname{sgn} f(x)$

$\qquad\qquad f : \mathcal{X} \to \mathbb{R}$ in $\mathcal{F}$

Review:

- linear classifiers: $\mathcal{X} \subseteq \mathbb{R}^d$

  $f(x) = \langle w, x \rangle \qquad w \in \mathbb{R}^d$

  (can cover affine $f(x) = \langle w, x \rangle + b$ by adding an all-1 coordinate to x:

  $x \mapsto \binom{x}{1} \in \mathbb{R}^{d+1}$)

$$\mathcal{F} := \{ x \mapsto \langle w, x \rangle : \|w\| \le B \}$$

$$R_n(\mathcal{F}(x^n)) = \frac{1}{n} \mathbb{E}_\varepsilon \left[\sup_{\|w\| \le B} \left| \sum_{i=1}^n \varepsilon_i \langle w, x_i \rangle \right| \right]$$

$$\le \frac{B}{n} \sqrt{\sum_{i=1}^n \|x_i\|^2}$$

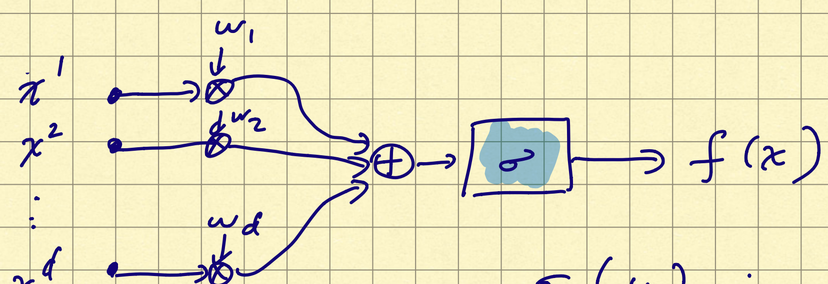
$$\mathcal{X} = \{ x \in \mathbb{R}^d : \|x\| \le R \} \implies R_n(\mathcal{F}(x^n)) \le \frac{BR}{\sqrt{n}}.$$

- simple nonlinearity: single neuron

$$f(x) = \sigma\big(\langle w, x \rangle\big) \qquad x, w \in \mathbb{R}^d$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a continuous fcn,
$\sigma(0) = 0$, Lipschitz continuous: $|\sigma(u) - \sigma(v)| \le L|u - v|$

$x = (x^1, x^2, \dots, x^d)^T$



: simple nonlinear unit

$\sigma(u)$: $\tanh(u) = \dfrac{e^u - e^{-u}}{e^u + e^{-u}}$

$\qquad (u)_+ = \max(0, u)$, ReLU

$$\mathcal{F}_\sigma := \left\{ x \mapsto \sigma\left(\sum_{j=1}^{d} w_j x^j \right) : \|w\| \le B \right\}$$

$$\mathcal{F}_\sigma = \sigma \circ \mathcal{F} \quad (\mathcal{F} : \text{lin. classifiers})$$

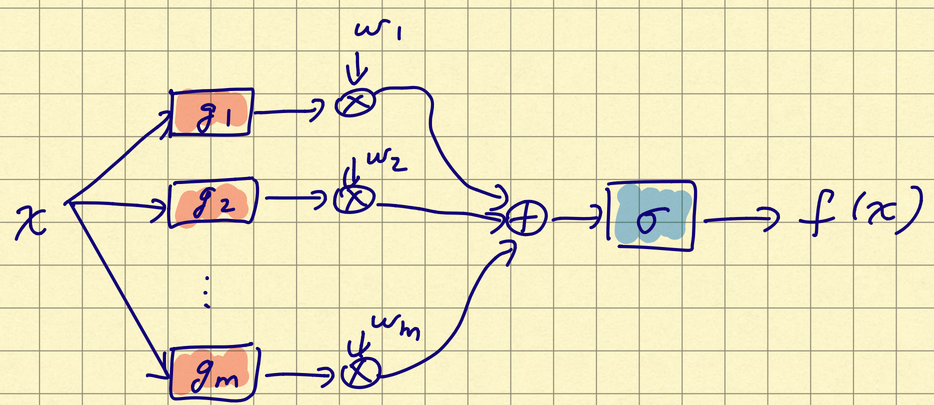
Contraction principle:

$$R_n(\mathcal{F}_\sigma) \le 2L\, R_n(\mathcal{F}) \le \frac{2LBR}{\sqrt{n}}$$

$$\text{if} \quad \mathcal{X} = \{ x \in \mathbb{R}^d : \|x\| \le R \}$$

- neural nets

$\mathcal{G}$: base classifiers $g : \mathcal{X} \to \mathbb{R}$



$$\mathcal{F}_1 := \left\{ x \mapsto \sigma\left(\sum_{j=1}^{m} w_j g_j(x) \right) : \begin{array}{l} m \in \mathbb{N} \\ \|w\|_1 = \sum_{j=1}^{m} |w_j| \le B \\ g_1, \dots, g_m \in \mathcal{G} \end{array} \right\}$$

$$\mathcal{F}_1 = \sigma \circ (B \cdot \operatorname{absconv}(\mathcal{G}))$$

Contraction principle again:

$\sigma(0) = 0$

σ L-Lip.

$$\begin{aligned} R_n(\mathcal{F}_1) &\le 2L \cdot R_n(B \cdot \operatorname{absconv}(\mathcal{G})) \\ &= 2LB \cdot R_n(\operatorname{absconv}(\mathcal{G})) \\ &= 2LB \cdot R_n(\mathcal{G}). \end{aligned}$$

- adding layers (making neural nets deeper)

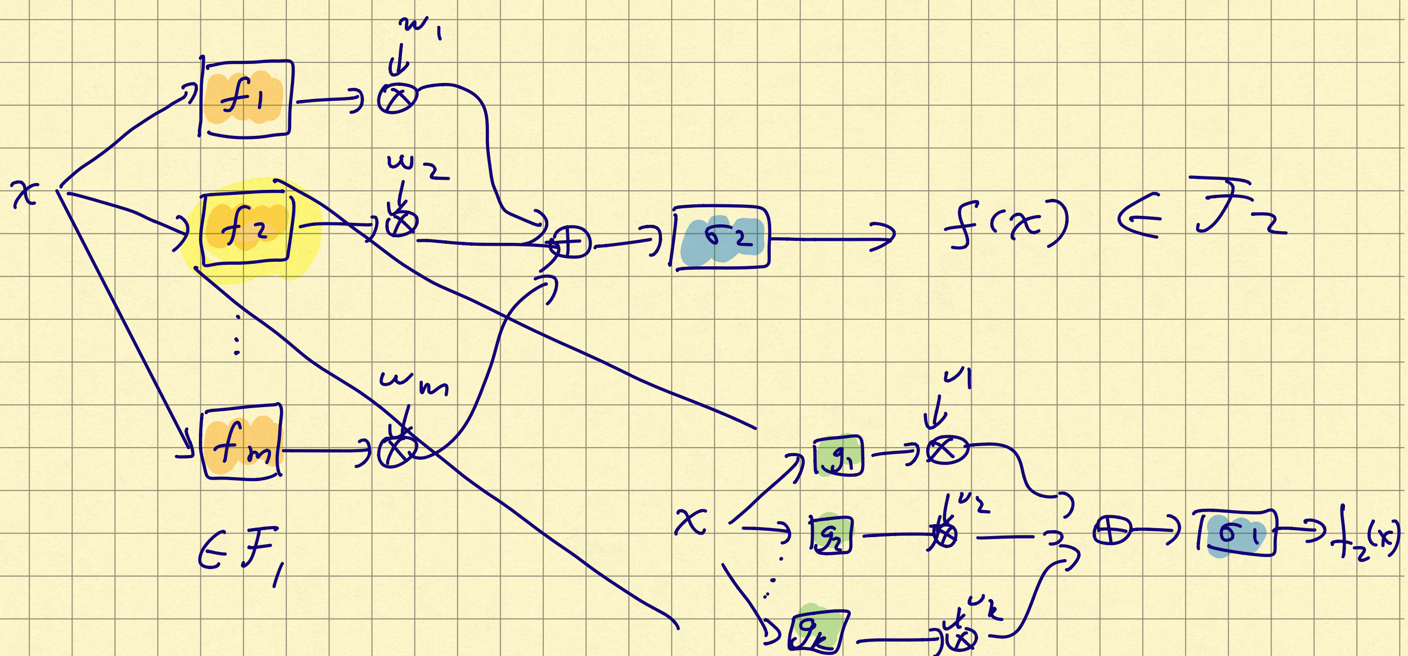
$f_1, \dots, f_m \in \mathcal{F}_1$ above

$$f_j(x) = \sigma_1\left(\sum_{k=1}^{m_j} w_k^j \, g_k^j(x)\right)$$

where $m_j \in \mathbb{N}$; $\|w^j\|_1 = \sum_{k=1}^{m_j} |w^j_k| \le B_1$

$g_1^j, \dots, g_{m_j}^j \in \mathcal{G}$

$$f(x) = \sigma_2\left(\sum_{j=1}^{m} w_j f_j(x)\right) \qquad \|w\|_1 \le B_2$$



$$\mathcal{F}_2 = \sigma_2 \circ B_2 \operatorname{absconv}(\mathcal{F}_1)$$

$$\begin{aligned} R_n(\mathcal{F}_2) &\le 2L_2 B_2 \cdot R_n(\mathcal{F}_1) \\ &\le 2L_2 B_2 \cdot 2L_1 B_1 \cdot R_n(\mathcal{G}) \\ &= 2^2 L_2 L_1 B_2 B_1 \cdot R_n(\mathcal{G}) \end{aligned}$$

– add more layers: $\mathcal{F}_j = \sigma_j \circ B_j \operatorname{absconv}(\mathcal{F}_{j-1})$

$\dots$ $\qquad \mathcal{F}_0 = \mathcal{G}$

$j = 1, \dots, \ell$

ℓ – # layers

$$R_n(\mathcal{F}_\ell) \leq \underbrace{2^\ell \cdot \prod_{j=1}^{\ell}(L_j B_j)}_{\text{grows as } \exp(\ell)} \cdot R_n(\mathcal{G})$$

(recursive use of contraction principle: K–P. 2002)

Can we do better? $\quad 2^\ell \longrightarrow \sqrt{\ell}$

$\prod_{j=1}^{\ell}(L_j B_j)$ still present

- Bartlett–Foster–Telgarsky (2017)
- Golowich–Rakhlin–Shamir (2017)

GRS : peeling layer-by-layer + log exp trick

$$\mathcal{F}_j = \sigma_j \circ B_j \operatorname{absconv}(\mathcal{F}_{j-1}) \qquad j=1,\dots,\ell$$

$$\mathcal{F}_0 = \mathcal{G}$$

Step 1

$$R_n(\mathcal{F}_\ell) = R_n(\mathcal{F}_\ell(X^n)), \quad X^n \text{ fixed}$$

$$R_n(\mathcal{F}_\ell) = \frac{1}{n} \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_\ell} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right]$$

$$= \frac{1}{\lambda n} \mathbb{E}_\varepsilon \left[\log \exp \left(\lambda \sup_{f \in \mathcal{F}_\ell} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right) \right]$$

$\lambda > 0$: to be tuned

$$\leq \frac{1}{\lambda n} \log \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_\ell} \exp \left(\lambda \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right) \right]$$

Note: $G(u) := e^{\lambda u}$ $(\lambda > 0)$ is convex, nondecreasing

Step 2

Lemma 1 (GRS) Let G be a convex, nondecreasing fcn $\mathbb{R} \to \mathbb{R}$; let $\mathcal{F}$ be of the form

$$\mathcal{F} = \sigma \circ B\,\mathrm{absconv}(\mathcal{F}')$$

for $\sigma : \mathbb{R} \to \mathbb{R}$, $\sigma(0) = 0$, L-Lip.. Then

$$\mathbb{E}_\varepsilon \left[G\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right) \right]$$
$$\leq 2\, \mathbb{E}_\varepsilon \left[G\left(LB \cdot \sup_{f' \in \mathcal{F}'} \left| \sum_{i=1}^n \varepsilon_i f'(x_i) \right| \right) \right]$$

Let's apply lemma 1 to $\mathcal{F}_\ell = \sigma_\ell \circ B_\ell\, \mathrm{absconv}(\mathcal{F}_{\ell-1})$

$$G(u) = e^{\lambda u}$$

$$\mathbb{E}_\varepsilon \left[\exp\left(\lambda \sup_{f \in \mathcal{F}_\ell} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right) \right]$$
$$\leq 2\, \mathbb{E}_\varepsilon \left[\exp\left(\lambda L_\ell B_\ell \cdot \sup_{f \in \mathcal{F}_{\ell-1}} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right) \right]$$
$$\vdots$$
$$\leq 2^\ell\, \mathbb{E}_\varepsilon \left[\exp\left(\lambda \prod_{j=1}^{\ell} (L_j B_j) \cdot \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n \varepsilon_i g(x_i) \right| \right) \right]$$

Step 3

$$M := \prod_{j=1}^{\ell} (L_j B_j)$$

$$\mathbb{E}_\varepsilon \left[\exp\left(\lambda M \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n \varepsilon_i g(x_i) \right| \right) \right] \leq \;?$$

$$\mathcal{A} := \{ (g(x_1), \dots, g(x_n)) : g \in \mathcal{G} \}$$

Lemma 2 (GRS) Let $A \subset \mathbb{R}^n$ be a bdd set. Then, $\forall \lambda > 0$,

$$\mathbb{E}_\varepsilon \left[\exp\left(\lambda \sup_{a \in A} \left| \sum_{i=1}^n \varepsilon_i a_i \right| \right) \right]$$

$$\leq \exp\left(\frac{\lambda^2}{2} \sum_{i=1}^n \sup_{a \in A} |a_i|^2 \right) \exp\left(\lambda n R_n(A) \right)$$

Proof idea $U(\varepsilon_1, \dots, \varepsilon_n) := \sup_{a \in A} \left| \sum_{i=1}^n \varepsilon_i a_i \right|$

$U(\varepsilon_1, \dots, \varepsilon_i, \dots, \varepsilon_n) - U(\varepsilon_1, \dots, -\varepsilon_i, \dots, \varepsilon_n) \leq \sup_{a \in A} |a_i|$

— mimic proof of McDiarmid, to bound $\mathbb{E}[e^{\lambda U}]$. ∎

Let $A := \mathcal{G}(x^n)$ $\qquad M := \prod_{j=1}^{\ell} (L_j B_j)$

$$\mathbb{E}_\varepsilon \left[\exp\left(\lambda M \cdot \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n \varepsilon_i g(x_i) \right| \right) \right]$$

$$\leq \exp\left(\frac{\lambda^2 M^2}{2} \cdot \sum_{i=1}^n \sup_{g \in \mathcal{G}} |g(x_i)|^2 \right) \cdot \exp\left(\lambda M n R_n(\mathcal{G}(x^n)) \right)$$

Step 4

$$R_n(\mathcal{F}_\ell) \leq \frac{1}{\lambda n} \log 2^\ell \, \mathbb{E}_\varepsilon \left[\exp\left(\lambda M \cdot \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n \varepsilon_i g(x_i) \right| \right) \right]$$

$$\leq \frac{1}{\lambda n} \left(\ell \log 2 + \frac{\lambda^2 M^2}{2} \sum_{i=1}^n \sup_{g \in \mathcal{G}} |g(x_i)|^2 + \lambda M n R_n(\mathcal{G}) \right)$$

$$= M R_n(\mathcal{G}) + \lambda \cdot \frac{M^2}{2n} \sum_{i=1}^n \sup_{g \in \mathcal{G}} |g(x_i)|^2 + \frac{1}{\lambda} \cdot \frac{\ell \log 2}{n}$$

$$\inf_{\lambda \geq 0} \left\{ \frac{a}{\lambda} + b\lambda \right\} = 2\sqrt{ab} \qquad (a, b \geq 0)$$

min. over $\lambda \geq 0$:

$$\min_{\lambda \geq 0} = \frac{1}{n} \cdot 2 \sqrt{\ell \log 2 \cdot \frac{M^2}{2} \left(\sum_{i=1}^{n} \sup_{g \in \mathcal{G}} |g(x_i)|^2 \right)}$$

$$= \frac{M}{n} \sqrt{2 \ell \cdot \log 2 \cdot \sum_{i=1}^{n} \sup_{g \in \mathcal{G}} |g(x_i)|^2}$$

$$\therefore \boxed{\begin{aligned} R_n(\mathcal{F}_\ell) &\leq \prod_{j=1}^{\ell} (L_j B_j) \cdot R_n(\mathcal{G}(X^n)) \\ &+ \frac{1}{n} \prod_{j=1}^{\ell} (L_j B_j) \underbrace{\sqrt{\ell \log 4 \cdot \sum_{i=1}^{n} \sup_{g \in \mathcal{G}} |g(x_i)|^2}}_{O(\sqrt{\ell n})} \end{aligned}}$$

(check constant in front in $\sqrt{\ell \ldots}$ term).

- take $L_j, B_j = 1 \ \forall j$ —

$$R_n(\mathcal{F}_\ell) \leq R_n(\mathcal{G}) + \frac{c}{n} \sqrt{\ell \cdot \sum_{i=1}^{n} \sup_{g \in \mathcal{G}} |g(x_i)|^2}$$

$$\lesssim R_n(\mathcal{G}) + C' \sqrt{\ell/n} .$$

- take $L_1 = \cdots = L_\ell = 1$
- B_j : largest ℓ_1 norm of weights in layer j