LAST TIME :

- Convex Optimization (gradient descent)
- Linear Models (Least-Squares Regression)
- Generalized Linear Models (GLMs)

Rademacher Complexity Bounds :

Let $\mathcal{G}$ is a class of functions from $X \to Y$, where $Y = [a,b] \subseteq \mathbb{R}$.

Empirical Rademacher Complexity:

Let $\mathcal{G}$ be a class of functions from $X \to [a,b]$ and $S = \{x_1, \dots, x_m\} \subseteq X$ a fixed sample of size m. Then the empirical Rademacher complexity of $\mathcal{G}$ with respect to the sample S as

$$\hat{R}_S(\mathcal{G}) = \mathbb{E}_{\sigma}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(x_i)\right]$$

where $\sigma = (\sigma_1, \dots, \sigma_m)$ with σ_i i.i.d random variables taking values uniformly in $\{-1, 1\}$. (Rademacher random variables)

- if $\mathcal{G}$ is a class of boolean f^ns taking values in $\{-1, 1\}$. Suppose S is shattered by $\mathcal{G}$, then $\hat{R}_S(\mathcal{G}) = 1$.

Rademacher Complexity : Let D be a distribution over X. Then for any integer $m \geq 1$, the Rademacher complexity of $\mathcal{G}$ w.r.t. D is the expectation of the empirical R.C over $S \sim D^m$.

$$R_m(\mathcal{G}; D) = \mathbb{E}_{S \sim D^m}\left[\hat{R}_S(\mathcal{G})\right].$$

AIM : When is $\frac{1}{m}\sum_i h(x_i)$ close to $\mathbb{E}_D[h(x)]$ where

- $x_i \sim D$ are drawn i.i.d.
- h may also depend on $x_1, \dots, x_m$.

We will try to bound:

$$\sup_{h \in \mathcal{G}} \left(\frac{1}{m} \sum_i h(x_i) - \mathbb{E}_D[h(x)] \right)$$

<u>McDiarmid's Inequality</u>: Let X be some set and let $f: X^m \to \mathbb{R}$ s.t.

$\forall i, \exists c_i \; \forall x_1, \dots, x_m, x_i' \in X$

$$|f(x_1, \dots, x_m) - f(x_1, \dots, x_i', \dots, x_m)| \le c_i,$$

Let $X_1, \dots, X_m$ be i.i.d. random variables taking values in X, then $\forall \varepsilon > 0$,

$$\mathbb{P}[f(X_1, \dots, X_m) \ge \mathbb{E}[f(X_1, \dots, X_m)] + \varepsilon] \le \exp\left(-\frac{2\varepsilon^2}{\sum_i c_i^2}\right).$$

<u>Theorem</u>: Let $\mathcal{G}$ be a family of functions mapping $X \to [0,1]$. Suppose that a sample S of size m is drawn from some distribution D over X. Then $\forall \delta > 0$, w.p. $\ge 1-\delta$, the following holds $\forall g \in \mathcal{G}$

$$\mathbb{E}_{x \sim D}[g(x)] \le \frac{1}{m} \sum_{i=1}^{m} g(x_i) + 2 \underbrace{R_m(\mathcal{G})} + O\left(\sqrt{\frac{\log 1/\delta}{m}}\right).$$

<u>ASIDE</u>: $g_w: \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}, \quad g_w(x, y) = (w \cdot x - y)^2$

$\mathbb{E}[g_w] = \mathbb{E}_{x,y \sim D}[\underbrace{(w \cdot x - y)^2}]$ (want this to be small)

$\frac{1}{m} \sum_i g_w(x_i, y_i) = \underbrace{\frac{1}{m} \sum_i (w \cdot x_i - y_i)^2}$ (minimize using optimization)

Proof: Define $\Phi(\underbrace{x_1, \ldots, x_m}_{S}) = \sup_{g \in \mathcal{G}} \left(\mathbb{E}_{x \sim D}[g(x)] - \hat{\mathbb{E}}_S(g) \right)$

$$\left(\hat{\mathbb{E}}_S(g) = \frac{1}{m} \sum_i g(x_i) \right)$$

$$\left| \Phi(x_1, \ldots, x_i, \ldots, x_m) - \Phi(x_1, \ldots, x_i', \ldots, x_m) \right| \le \sup_{g \in \mathcal{G}} \frac{1}{m} \left| g(x_i) - g(x_i') \right| \le \frac{1}{m}.$$

We can apply McDiarmid's inequality with $c_i = 1/m \quad \forall i$.

$$\mathbb{P}\left(\Phi(S) \ge \mathbb{E}[\Phi(S)] + \varepsilon \right) \le \exp(-2m\varepsilon^2).$$

w.p. $\ge 1 - \delta$, $\quad \Phi(S) \le \mathbb{E}[\Phi(S)] + \sqrt{\frac{\log 1/\delta}{2m}}$.

Want to bound $\mathbb{E}[\Phi(S)]$.

$S' = \{x_1', \ldots, x_m'\}$ from D.

$$\boxed{\mathbb{E}_D[g] = \mathbb{E}_{S'} \hat{\mathbb{E}}_{S'}[g]}$$

$$\begin{aligned}
\mathbb{E}_S \Phi(S) &= \mathbb{E}_S\left[\sup_{g \in \mathcal{G}} \left(\mathbb{E}_D[g] - \hat{\mathbb{E}}_S g \right) \right] \\
&= \mathbb{E}_S\left[\sup_{g \in \mathcal{G}} \left(\mathbb{E}_{S'} \hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S g \right) \right] \\
&\le \mathbb{E}_{S, S'}\left[\sup_{g \in \mathcal{G}} \left(\hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S[g] \right) \right] \quad \text{(Symmetrization).} \\
&= \mathbb{E}_{S, S'}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_i \left(g(x_i') - g(x_i) \right) \right]. \\
&= \mathbb{E}_{S, S', \sigma}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_i \sigma_i \left(g(x_i') - g(x_i) \right) \right] \quad \left(S, S' \text{ are identically distributed.} \right) \\
&\le \mathbb{E}_{S, \sigma}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_i \sigma_i g(x_i) \right] + \mathbb{E}_{S', \sigma}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_i \sigma_i g(x_i') \right] \\
&= 2 \mathcal{R}_m(\mathcal{G}; D).
\end{aligned}$$

---

w.p. $\ge 1 - \delta$

$$\Phi(S) \le 2 \mathcal{R}_m(\mathcal{G}; D) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

$$\sup_{g \in \mathcal{G}} \left(\mathbb{E}[g] - \hat{\mathbb{E}}_S(g) \right) \le 2 \mathcal{R}_m(\mathcal{G}; D) + \sqrt{\frac{\log 1/\delta}{2m}}.$$

$$\forall g \in \mathcal{G} \quad \mathbb{E}_D[g] \le \frac{1}{m} \sum_i g(x_i) + 2 \mathcal{R}_m(\mathcal{G}; D) + \sqrt{\frac{\log 1/\delta}{2m}}. \quad \left(\text{Slow rates} \right)$$

$\mathcal{Y} = \{+1, -1\}$

$$\mathbb{E}\frac{1}{m}\Big|\sum_i \sigma_i\Big| = \Theta\left(\frac{1}{\sqrt{m}}\right).$$

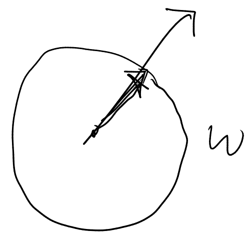
$L = \{ \underline{x} \mapsto \underline{w}\cdot x \mid \underline{w} \in \mathbb{R}^n, \ \|w\| \le W\}$

$\mathcal{X} = \{ \underline{x} \in \mathbb{R}^n, \ \|x\| \le X\}$.

$\forall g \in L, \ g : X \to \mathbb{R}$ $\qquad S = \{x_1, \dots, x_m\}$.

$$\hat{R}_S(L) = \frac{1}{m}\,\mathbb{E}_{\underline{\sigma}}\Big[\sup_{\substack{w \\ \|w\|\le W}} \sum_i \sigma_i(\underline{w}\cdot \underline{x}_i)\Big]$$

$$= \frac{1}{m}\,\mathbb{E}_{\underline{\sigma}}\Big[\sup_{\substack{w \\ \|w\|\le W}} \Big(\underline{w}\cdot\Big(\underbrace{\sum_i \sigma_i \underline{x}_i}\Big)\Big)\Big].$$



$$= \frac{W}{m}\,\mathbb{E}_{\sigma}\Big[\Big\|\sum_i \sigma_i \underline{x}_i\Big\|\Big]$$

$$\le \frac{W}{m}\sqrt{\mathbb{E}_{\sigma}\Big[\Big\|\sum_i \sigma_i \underline{x}_i\Big\|^2\Big]}. \qquad \mathbb{E}[\sigma_i] = 0$$

$$\le \frac{W}{m}\sqrt{\mathbb{E}_{\underline{\sigma}}\Big[\sum_i \|x_i\|^2\Big]}.$$

$$\le \frac{W}{m}\sqrt{m X^2} = \frac{W\cdot X}{\sqrt{m}}. \qquad \|x_i\|^2 \le X^2$$

Talagrand's Lemma:

Let $\mathcal{H}$ be some class of functions and $\phi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz

$$\hat{R}_S(\phi\circ H) \le L\cdot \hat{R}_S(H).$$

$$\phi\circ H = \{\phi\circ h \mid h \in H\}.$$

$\mathcal{H} = \{ (\underline{x}, y) \mapsto (\underline{w} \cdot x - y)^2 \mid \|w\| \le W \}$.

$\underline{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$

In the support of D, $\|\underline{x}\| \le X$, $y \in [-M, M]$.

$\phi(z) = z^2$.

$|w \cdot x - y| \le WX + M$ $\qquad \underline{w} \cdot \underline{x} \le \|\underline{w}\| \cdot \|\underline{x}\| \le WX$

In the interval of relevance, ϕ is indeed $2(WX+M)$ Lipschitz

$$\hat{\mathcal{R}}(\mathcal{H}) \le \frac{2(WX+M) \cdot (WX)}{\sqrt{m}}.$$

[If we do linear regression to find $\hat{w}$, s.t. $\|\hat{w}\| \le W$, then a sample of size $\text{poly}(M, W, X, 1/\varepsilon, 1/\delta)$ is sufficient for learning.]

If $\mathbb{E}[y|x] = \underline{w}^* \cdot x$, we wanted to find $\hat{w}$, st.

? $\boxed{\mathbb{E}_x ((\underline{\hat{w}} \cdot x - \underline{w}^* \cdot x)^2] \le \varepsilon.}$

$$\underbrace{A = \mathbb{E}_D [(\hat{w} \cdot x - y)^2]} \le \hat{\mathbb{E}}_S [(\hat{w} \cdot x - y)^2] + \varepsilon' \qquad (\text{w.h.p})$$

$$\le \hat{\mathbb{E}}_S [(\underline{w}^* x - y)^2] + \varepsilon' \qquad (\text{whp}) \ \hat{w} \in \operatorname{argmin} \hat{\mathbb{E}}_S (wx - y)^2$$

$$\le \mathbb{E}_D [\underline{(w^* \cdot x - y)^2}] + 2\varepsilon'. \qquad \left(\text{apply thm flipping } \mathbb{E}_D \ \& \ \hat{\mathbb{E}}_S\right).$$

$(g - g^* + g^* - y)$

$\mathbb{E}[(g - y)^2] = \underline{\mathbb{E}[(g - g^*)^2] + \mathbb{E}[(g^* - y)^2]}$, $\qquad \mathbb{E}[y|x] = g^*(x)$

$\underbrace{\mathbb{E}[(\hat{w} \cdot x - w^* \cdot x)^2]}_{= A \text{?}} + \cancel{\mathbb{E}[(w^* \cdot x - y)^2]} \le \cancel{\mathbb{E}[(w^* \cdot x - y)^2]} + 2\varepsilon'$

$\mathbb{E}[(g - y)^2] = \mathbb{E}[(g - g^* + g^* - y)^2] = \mathbb{E}[(g - g^*)^2] + \mathbb{E}[(g^* - y)^2] + 2 \underset{x,y}{\mathbb{E}} \underbrace{\cancel{[(g - g^*)(g^* - y)]}}_{0}$