Lecture 24

Last time

- Controlling the VC dimension
- Generalization for linear models.

Today

- Linear Models continued
- Convexity & stability

Linear models continued

Lemma: Consider $X = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ and

$$\mathcal{H} = \{a \mapsto \langle w, a \rangle \mid \|w\|_2 \leq 1\}.$$

Then,

$$\bar{\mathcal{R}}(\mathcal{H}(X)) \leq \sqrt{\sum_{i=1}^{n} \|x_i\|_2^2},$$

thus,

$$\bar{\mathcal{R}}_n(\mathcal{H}) \leq \frac{\mathbb{E} \max \|x_i\|_2}{\sqrt{n}}.$$

Dimension free.

Proof: Expanding

$$\bar{\mathcal{R}}(\mathcal{H}(X)) = \mathbb{E}_\varepsilon \sup_{\|w\|_2 \leq 1} \sum_{i=1}^{n} \varepsilon_i \langle w, x_i \rangle$$

$$= \mathbb{E}_\varepsilon \sup_{\|w\|_2 \leq 1} \langle w, \sum \varepsilon_i x_i \rangle$$

$$= \mathbb{E}_\varepsilon \left\| \sum \varepsilon_i x_i \right\|_2$$

Jensen's ↘

$$\leq \sqrt{\mathbb{E}_{\varepsilon} \left\| \sum \varepsilon_i x_i \right\|_2^2}$$

ε_i iid ↘

$$= \sqrt{\sum \|x_i\|_2^2}.$$

Recall

$$\bar{R}_n(\mathcal{H}) = \mathbb{E}_X \bar{R}\left(\frac{1}{n} \mathcal{H}(X)\right)$$

$$\leq \frac{1}{n} \mathbb{E} \sqrt{\sum \|x_i\|_2^2}$$

Jensen's ↘

$$\leq \frac{1}{\sqrt{n}} \mathbb{E}_X \max \|x_i\|_2. \qquad \square$$

Lastly, lets do the same with the ℓ_1 norm.

Lemma: Consider $X = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ and

$$\mathcal{H} = \{a \mapsto \langle w, a \rangle \mid \|w\|_1 \leq 1\}.$$

Then,

$$\bar{R}(\mathcal{H}(X)) \leq \sqrt{2n \log(2d)} \max \|x_i\|_\infty,$$

thus,

$$\bar{R}_n(\mathcal{H}) \le \sqrt{\frac{2\log 2d}{n}} \max \|x_i\|_\infty.$$

Almost dim free.

Proof: Same strategy as before

$$\bar{R}(\mathcal{H}(X)) = \mathbb{E}_\varepsilon \sup_{\|w\|_1 \le 1} \langle w, \sum_i \varepsilon_i x_i \rangle$$

(why?)

$$= \mathbb{E}_\varepsilon \left\| \sum_i \varepsilon_i x_i \right\|_\infty$$

$$= \mathbb{E}_\varepsilon \sup_{j \in [d]} \left| \sum_i \varepsilon_i x_{ij} \right|$$

Massart's lemma.

$$\le \sqrt{2 \log(2d)} \sup_{x_i} \sqrt{\sum_j x_{ij}^2}$$

$$\le \sqrt{2n \log(2d)} \max \|x_i\|_\infty.$$

□

Generalization via convexity and stability

Suppose we wanted to minimize

(*) $\min_{\theta \in \Theta} F(\theta)$ with $F(\theta) = \mathbb{E}_{x \sim p} f(\theta, x)$.

and we had an algorithm A s.t. given an iid sample $X = (x_1, \dots x_n)$ it outputs $\hat{\theta} = A(X)$ an estimate of a minimizer of (*). Our goal next is to establish generalization guarantees for $A(X)$ (not necessarily uniformly).

The main idea is to use stability. Define

$$X^i = (x_1, \dots, x_{i-1}, x', x_{i+1}, \dots, x_n)$$

where $x' \sim P$ and independent of X.

Theorem: We have that

$$\mathop{\mathbb{E}}_{X}\left[F(A(X)) - \frac{1}{n}\sum_{i=1}^{n} f(A(X), x_i)\right]$$

$$= \mathop{\mathbb{E}}_{\substack{(X, x') \sim P \\ i \sim \mathrm{Unif}([n])}} \left[f(A(X^i), x_i) - f(A(X), x_i)\right].$$

Proof: The proof is simple, for every i we have

$$\mathop{\mathbb{E}}_{X} F(A(X)) = \mathop{\mathbb{E}}_{X, x'}\left[f(A(X), x'\right) = \mathop{\mathbb{E}}_{X, x'}\left[f(A(X^i, x_i)\right].$$

Also

$$\mathop{\mathbb{E}}_{X}\left[\frac{1}{n}\sum f(A(X), x_i)\right] = \mathop{\mathbb{E}}_{X,i} f(A(X), x_i).$$

Substituting these two gives (♡). □

Intuition: If the value of the loss is stable to changing one element of the training set, then the method generalize well.

Def: We say that an algorithm $A(\cdot)$ is replace-one-stable with rate $\varepsilon(n)$ if

$$\mathop{\mathbb{E}}_{\substack{X, X' \overset{iid}{\sim} P \\ i \sim \text{Unif}[n]}} \left[f(A(X^i), x_i) - f(A(X), x_i) \right] \leq \varepsilon(n).$$

Assumption (:): We will focus on the case $f(\cdot, x)$ ρ-Lipschitz convex for all x and $\mathcal{H}$ is a convex closet set.

Just as we did for ridge regression we consider a regularized estima

for

(A) $A(X) := \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} f(\theta, x_i) + \frac{\lambda}{2} \|\theta\|_2^2$.

In turn, this objective exhibits nice properties.

Def: A function $g: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is α-strongly convex if $g - \frac{\alpha}{2}\|\cdot\|^2$ is convex.

The objective in (A) is λ-strongly convex by construction.

Lemma (V): (Proven in Nonlinear I) Suppose g is α-strongly convex. Then it has a unique minimizer θ^* and

$$\frac{\alpha}{2}\|\theta - \theta^*\|^2 \leq g(\theta) - g(\theta^*) \quad \forall \theta.$$

Thus, $A(\cdot)$ is well defined. Next, we establish a generalization guarantee.

Theorem: Suppose Assumption (:-) holds. Then, algorithm (A) is

replace-one-stable with rate $\frac{2\rho^2}{\lambda n}$.

Thus,

$$\mathbb{E}_X\left[F(A(X)) - \frac{1}{n}\sum_{i=1}^{n} f(A(X), x_i)\right] \le \frac{2\rho^2}{\lambda n}.$$

Proof: Notice that by ρ-Lipschitzness

$$f(A(X^i), x_i) - f(A(X), x_i) \le \rho \|A(X^i) - A(X)\|_2.$$

So it suffices to show that

$$\|A(X^i) - A(X)\| \le \frac{2\rho}{\lambda n}.$$

Define the function

$$f_X(\theta) := \frac{1}{n}\sum_{i=1}^{n} f(\theta, x_i) + \frac{\lambda}{2}\|\theta\|_2^2.$$

By Lemma (V) we have that for each i:

$$\frac{\lambda}{2}\|A(X^i) - A(X)\|_2^2$$

$$\le f_X(A(X^i)) - f_X(A(X))$$

$$= \frac{1}{n}\sum_{x\in X} f(A(X^i), x) + \frac{\lambda}{2}\|A(X^i)\|_2^2.$$

$$- \frac{1}{n}\sum_{x\in X} f(A(X), x) - \frac{\lambda}{2}\|A(X)\|_2^2.$$

$$= \frac{1}{n}\sum_{x\in X^i} f(A(X^i), x) + \frac{\lambda}{2}\|A(X^i)\|_2^2.$$

↖ Notice the index set

$$- \frac{1}{n} \sum_{x \in X^i} f(A(X), x) - \frac{\lambda}{2} \|A(X)\|_2^2.$$

$$+ \frac{f(A(X^i), x_i) - f(A(X), x_i)}{n}$$

$$+ \frac{f(A(X), x') - f(A(X^i), x')}{n}$$

$$\leq f_{X^i}(A(X^i)) - f_{X^i}(A(X^i))$$

$$+ \frac{2\rho}{n} \|A(X) - A(X^i)\|_2$$

$$\leq -\frac{\lambda}{2} \|A(X^i) - A(X)\|_2^2 + \frac{2\rho}{n} \|A(X) - A(X^i)\|_2$$

$A(X^i)$ minimizes f_{X^i} so this uses Lemma (V).

Rearranging gives

$$\|A(X^i) - A(X)\|_2^2 \leq \frac{2\rho}{\lambda n},$$

as we wanted. □

We can use this to prove an excess risk bound.

Corollary: Let θ^* be any minimizer of

(★). Then, we have

$$\mathbb{E}_X F(A(X)) \le \min_{\theta \in \Theta} F(\theta) + \frac{\lambda}{2}\|\theta^*\|_2^2 + \frac{2\rho}{\lambda n}.$$

Thus, if we set $\lambda = \sqrt{\frac{4\rho^2}{n\|\theta^*\|^2}}$,

$$\mathbb{E}_X F(A(X)) \le \min_{\theta \in \Theta} F(\theta) + \frac{2\rho\|\theta^*\|}{\sqrt{n}}.$$

Proof: Adding and subtracting

$$\mathbb{E}_X F(A(X)) = \mathbb{E}_X\left[\frac{1}{n}\sum_{x \in X} f(A(X), x)\right] + \mathbb{E}_X\left[F(A(S)) - \frac{1}{n}\sum_{x \in X} f(A(X), x)\right]$$

$$\le \mathbb{E}_X\left[\frac{1}{n}\sum_{x \in X} f(A(X), x)\right] + \frac{2\rho^2}{\lambda n}.$$

Notice that by definition

$$\mathbb{E}_X\left[\frac{1}{n}\sum_{x \in X} f(A(X), x)\right] \le \mathbb{E}_X\left[f_X(A(X))\right]$$

$$\leq \mathbb{E}_X [f_X(\theta^*)]$$

$$= F(\theta^*) + \frac{\lambda}{2} \|\theta^*\|_2^2.$$

Combining these two bounds yield the first stated bound. The second one follows by optimizing λ. $\square$