Lecture 7

* You can signup for class presentations.
* Scribe lecture notes for lectures 1-5 are online.
* Last time
  - → VC Theorem
  - → Rademacher complexity

Today

Finish Rademacher complexity

Let $Z = X \times Y$

$\mathcal{F}$: function class $Z \to \mathbb{R}$

D : distribution over Z.

Definition (Rademacher Complexity).

Let $\mathcal{F}$ be a family of real-valued functions $f: Z \to \mathbb{R}$ where $Z = X \times Y$. Then the Rademacher complexity $R(\mathcal{F})$ is defined as

$$R(\mathcal{F}) = \frac{1}{n} \mathbb{E}_{\sigma \sim \{\pm 1\}^n} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(z_i) \right].$$

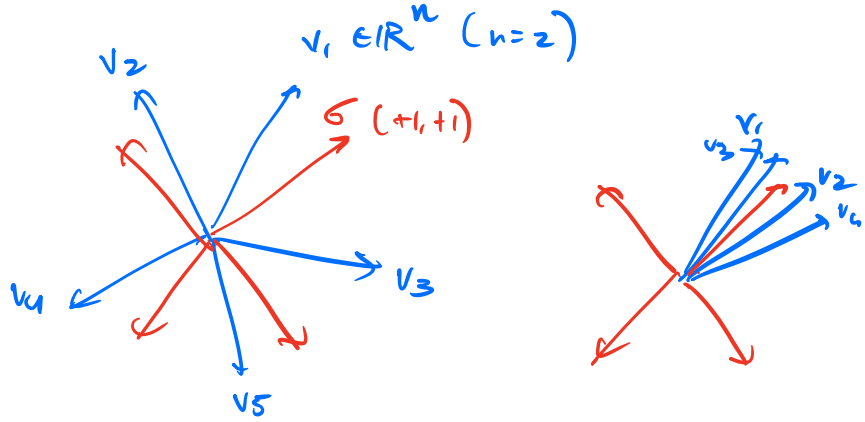
More generally, given a set of vectors $A \subset \mathbb{R}^n$, the Rademacher complexity $R(A)$ is defined as

$$R(A) = \frac{1}{n} \mathbb{E}_{\sigma \sim \{\pm 1\}^n} \sup_{a \in A} \sum_{i=1}^{n} \sigma_i (a)_i .$$

Intuition: $R(\mathcal{F})$ captures how well can $\mathcal{F}$ fit random noise.

If $\mathcal{F}$ can fit random noise, $\mathcal{F}$ will probably overfit.

Geometric picture



In expectation over $\sigma \sim \{\pm 1\}^n$, what is the max. inner product we can get with σ?

How do we use Rademacher?

$S = \{(x_i, y_i),\ i \in [n]\}$

$\mathcal{H}$: function from $\mathcal{X} \to \mathcal{Y}$.

$\mathcal{H} \circ S = \{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}$

$\ell(h(x), y)$: instead of writing $\ell(h(x), y)$ we can write $\ell(h, z) = \ell(h(x), y)$ where $z = (x, y)$

$\ell \circ \mathcal{H} \circ S = \{(\ell(h, z_i),\ i \in [n]) : h \in \mathcal{H}\}$

For e.g. $\mathcal{H} = \{h_1, h_2, h_3\}$

$$\ell \circ \mathcal{H} \circ S = \{(\ell(h_1, z_1), \ldots, \ell(h_1, z_n)),$$
$$(\ell(h_2, z_1), \ldots, \ell(h_2, z_n)),$$
$$(\ell(h_3, z_1), \ldots, \ell(h_3, z_n))\}$$

Lemma (Symmetrization with Rademacher)

$$\mathbb{E}_{S\sim D^n} \sup_{h\in\mathcal{H}} \left(R(h) - \hat{R}_S(h)\right) \le 2\,\mathbb{E}_{S\sim D^n} R(\ell\circ\mathcal{H}\circ S).$$

Proof:

$$\mathbb{E}_{S\sim D^n} \sup_{h\in\mathcal{H}} \left(R(h) - \hat{R}_S(h)\right)$$

$$\le \mathbb{E}_{S,S'} \sup_{h\in\mathcal{H}} \frac{1}{n}\left(\sum_{i=1}^{n} \left(\ell(h,z_i) - \ell(h,z_i')\right)\right)$$

$$= \mathbb{E}_{S,S',\sigma_{1:n}} \sup_{h\in\mathcal{H}} \frac{1}{n}\left(\sum_{i=1}^{n} \sigma_i\left(\ell(h,z_i) - \ell(h,z_i')\right)\right)$$

$$\le \mathbb{E}_S \mathbb{E}_{\sigma_{1:n}} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\,\ell(h,z_i)$$

$$+ \mathbb{E}_{S'} \mathbb{E}_{\sigma_{1:n}} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} (-\sigma_i)\,\ell(h,z_i')$$

$$\mathbb{E}_S \sup_{h\in\mathcal{H}} \left(R(h) - \hat{R}_S(h)\right) \le 2\,\mathbb{E}_{S,\sigma_{1:n}} \sup_{h\in\mathcal{H}} \left(\frac{1}{n}\sum_{i=1}^{n} \sigma_i\,\ell(h,z_i)\right)$$

$$= 2\,\mathbb{E}_S R(\ell\circ\mathcal{H}\circ S)$$

■

Theorem (Excess risk bounds using Rademacher)

Assume that for all z and $h \in \mathcal{H}$ we have that $|\ell(h,z)| \leq c$. Then we probability at least $(1-\delta)$ over $S \sim D^n$,

(1) $$\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \leq 2 \mathbb{E}_{S'} R(\ell \circ \mathcal{H} \circ S') + c \sqrt{\frac{2 \log(1/\delta)}{n}}.$$

(2) $$\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \leq 2 R(\ell \circ \mathcal{H} \circ S) + 3c \sqrt{\frac{2 \log(2/\delta)}{n}}$$

(3) for any $h^* \in \mathcal{H}$,

$$R(ERM_{\mathcal{H}}(S)) - R(h^*) \leq 2 R(\ell \circ \mathcal{H} \circ S) + 4c \sqrt{\frac{2 \log(4/\delta)}{n}}.$$

(this holds for $h^* = \operatorname{arg\,min}_{h \in \mathcal{H}} R(h)$)

Proof

We will keep using McDiarmid's.

Part (i)

Note that $\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right)$ satisfies the bounded differences property with constant $\frac{2c}{n}$.

(changing any (x_i, y_i) changes the loss by at most $\frac{2c}{n}$).

$\therefore$ Using McDiarmid's,

$$\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \leq \mathbb{E}\left(\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \right) + \varepsilon$$

$$\text{w.p. } 1 - \exp\left(- \frac{2\varepsilon^2}{n(2c/n)^2} \right)$$

$$= 1 - \underbrace{\exp\left(- \frac{n\varepsilon^2}{2c^2} \right)}_{\delta}$$

$$\text{Choose } \varepsilon = c\sqrt{\frac{2\log(1/\delta)}{n}}$$

We get, w.p. $(1 - \delta)$,

$$\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \leq \mathbb{E}\left(\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \right) + c\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Now use symmetrization lemma, and result follows.

∎

Part (2)

$$R(\ell \circ \mathcal{H} \circ S) = \mathbb{E}_{\sigma_{i:n}} \left(\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(h, z_i) \right)$$

this also satisfies bounded differences with constant $\frac{2c}{n}$ (swapping σ_i by σ_i' changes by $\leq \frac{2c}{n}$) w.p. $1-\delta$,

$$\therefore \quad R(\ell \circ \mathcal{H} \circ S) \geq \mathbb{E}_{S'} \left(R(\ell \circ \mathcal{H} \circ S') \right) - c\sqrt{\frac{2\log(1/\delta)}{n}}.$$

$$\Rightarrow \mathbb{E}_{S'} \left(R(\ell \circ \mathcal{H} \circ S) \right) \leq R(\ell \circ \mathcal{H} \circ S) + c\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Now set $\delta = \delta'/2$, w.p. $1 - \frac{\delta'}{2}$,

$$\mathbb{E}_{S'} \left(R(\ell \circ \mathcal{H} \circ S) \right) \leq R(\ell \circ \mathcal{H} \circ S) + c\sqrt{\frac{2\log(2/\delta')}{n}}$$

$$\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \leq 2\mathbb{E}_{S'} \left(R(\ell \circ \mathcal{H} \circ S) \right) + c\sqrt{\frac{2\log(2/\delta)}{n}}$$

(By part (1)).

Result follows by doing a union bound.

w.p. $\geq 1-\delta$,

$$\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \leq 2\, R(\ell \circ \mathcal{H} \circ S) + 3c \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Part 3

Let $h_S = ERM_{\mathcal{H}}(S)$.

Part (2) bounds $R(h) - \hat{R}(h) \leq (\cdot)$
For ② we need bound on $\hat{R}(h) - R(h) \leq (\cdot)$
∴ Part (2) doesn't work.

$$R(h_S) - R(h^*) = \underbrace{R(h_S) - \hat{R}_S(h_S)}_{①} + \underbrace{\hat{R}_S(h_S) - \hat{R}_S(h^*)}_{\leq 0} + \underbrace{\hat{R}_S(h^*) - R(h^*)}_{②}$$

① is bounded by Part (2)

For ② we use Hoeffding's (McDiarmid's with constant $\frac{2c}{n}$)

By using this, we get that w.p. $1 - \frac{\delta}{2}$,

$$\hat{R}_S(h^*) - R(h^*) \leq c \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Combining ① & ②,

$$R(h_S) - R(h^*) \leq 2\, R(\ell \circ \mathcal{H} \circ S) + 4c \sqrt{\frac{2\log(4/\delta)}{n}}.$$

Take aways.

→ Could be much better than VC bound: Rademacher complexity takes the data distribution into account, and could give tighter bounds than worst case VC bounds.

→ Data-dependent bound. (3) in our Theorem is a data-dependent bound, we use a training set S both for learning a hypothesis from $\mathcal{H}$, and for estimating its generalization error (we can check if we are overfitting).

Rademacher calculus.

Claim (Translation & Scaling)

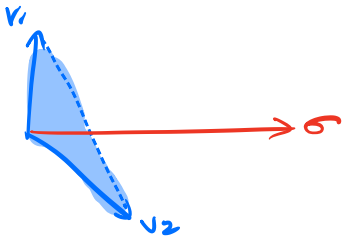
Let $A' = \{\varphi a + v, a \in A\}$

Then $R(A') = \varphi R(A)$.

Proof [Exercise]

Can also show that

$R(\{\text{convex hull of } A\}) = R(A)$.



Lemma (Massart Lemma) Let $A = \{v_1, \dots, v_m\}$ be a finite set of vectors in $\mathbb{R}^n$. Let $\bar{v} = \frac{1}{m}\sum_{i=1}^{m} v_i$. Then

$$R(A) \leq \max_i \|v_i - \bar{v}\|_2 \cdot \frac{\sqrt{2\log(m)}}{n}$$

Proof [Exercise]

Hint: First, by translation invariance, we can take $\bar{v} = 0$ without loss of generality. Then use the mat. of sub-Gaussians result from last time. ∎

Note: This gives a bound for finite hypothesis classes.

Lemma (Contraction Lemma)

For each $i \in [m]$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function i.e. $|\phi_i(x) - \phi_i(y)| \leq \rho |x-y| \quad \forall x, y \in \mathbb{R}$.

For any $a \in \mathbb{R}^n$ define $\phi(a) \in \mathbb{R}^n$ as $\phi(a) = (\phi((a)_1), \dots, \phi((a)_n))$. For a set A, let $\phi \circ A = \{\phi(a) : a \in A\}$. Then

$$\mathcal{R}(\phi \circ A) \leq \rho \, R(A).$$

Proof. Refer to book. ∎

Rademacher complexity of linear classes.

$$\mathcal{H}_1 = \{ h_w(x) = \langle w, x \rangle : \|w\|_1 \le B_1 \}$$

$$\mathcal{H}_2 = \{ h_w(x) = \langle w, x \rangle : \|w\|_2 \le B_2 \}$$

Lemma (ℓ2 bounded linear predictor)

Let $S = (x_1, \ldots, x_n)$. Define $\mathcal{H}_2 \circ S = \{ (\langle w, x_1 \rangle, \ldots \langle w, x_n \rangle) : \|w\|_2 \le B_2 \}$.

Then $R(\mathcal{H}_2 \circ S) \le B_2 \frac{\max_i \|x_i\|_2}{\sqrt{n}}$

Proof By Cauchy-Schwartz : $\langle w, x \rangle \le \|w\|_2 \|x\|_2$

$$\therefore n R(\mathcal{H}_2 \circ S) = \mathbb{E}_\sigma \left[\sup_{a \in \mathcal{H}_2 \circ S} \sum_{i=1}^n \sigma_i a_i \right]$$

$$= \mathbb{E}_\sigma \left[\sup_{w : \|w\|_2 \le B_2} \sum_{i=1}^n \sigma_i \langle w, x_i \rangle \right]$$

$$= \mathbb{E}_\sigma \left[\sup_{w : \|w\|_2 \le B_2} \left\langle w, \sum_{i=1}^n \sigma_i x_i \right\rangle \right]$$

$$\le B_2 \cdot \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2 \right] - (1)$$

Using Jensen's,

$$\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2 \right] = \mathbb{E}_\sigma \left[\left(\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2 \right)^{1/2} \right]$$

$$\leq \left(\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2 \right)^{1/2} \quad -(2)$$

$$\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2 \right] = \mathbb{E}_\sigma \left[\sum_{i,j} \sigma_i \sigma_j \langle x_i, x_j \rangle \right]$$

Since σ_i are independent $\mathbb{E}_\sigma [\sigma_i \sigma_j] = 0 \quad (\forall\, i \neq j)$

$$\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2 \right] = \sum_{i=1}^n \|x_i\|_2^2$$

$$\leq n \max_i \|x_i\|_2^2 \quad -(3)$$

Combining (1), (2) & (3) gives bound. ∎

Lemma (ℓ_1 bounded linear model)

Let $S = (x_1, \ldots, x_n)$ where $x_i \in \mathbb{R}^d \ \forall\, i \in [n]$.

Then $$R(\mathcal{H}_1 \circ S) \leq B_1 \max_i \|x_i\|_\infty \sqrt{\frac{2\log(2d)}{n}}.$$

Proof By Holder's inequality $\langle w, v \rangle \leq \|w\|_1 \|v\|_\infty$

$$n \quad R(\mathcal{H}_1 \circ S) = \mathbb{E}_\sigma \left[\sup_{a \in \mathcal{H}_1 \circ S} \sum_{i=1}^n \sigma_i a_i \right]$$

$$= \mathbb{E}_\sigma \left[\sup_{w : \|w\|_1 \leq B_1} \sum_{i=1}^n \sigma_i \langle w, x_i \rangle \right]$$

$$= \mathbb{E}_\sigma \left[\sup_{w : \|w\|_1 \leq B_1} \left\langle w, \sum_{i=1}^n \sigma_i x_i \right\rangle \right]$$

$$\leq \mathbb{E}_\sigma \left[B_1 \left\| \sum_{i=1}^n \sigma_i x_i \right\|_\infty \right]$$

$$= B_1 \, \mathbb{E}_\sigma \left[\max_{j \in [d]} \left| \sum_{i=1}^{n} \sigma_i (x_i)_j \right| \right]$$

each term $\sigma_i (x_i)_j$ is $|(x_i)_j|$ sub-Gaussian

since $|(x_i)_j| \le \max_i \|x_i\|_\infty$,

$\therefore$ each term $\sigma_i (x_i)_j$ is $\max_i \|x_i\|_\infty$ sub-Gaussian

the sum $\sum_{i=1}^{n} \sigma_i (x_i)_j$ is sub-Gaussian

$$\le \left(\sum_{i=1}^{n} \left(\max_i \|x_i\|_\infty \right)^2 \right)^{1/2}$$

$$\le \sqrt{n} \cdot \max_i \|x_i\|_\infty$$

By bound for max. of sub-Gaussian, including negations to take care of $|\cdot|$,

$$n \, R(H_1 \circ S) \le B_1 \sqrt{n} \max_i \|x_i\|_\infty \sqrt{2 \log(2d)}$$

■