"Rademacher complexity and generalization bounds"

Michael Arbel, Julien Mairal, Pierre Gaillard

Exercice 1. Rademacher complexity
Let $\mathbb{P}$ be some unknown distribution on $\mathcal{X}\times\mathcal{Y}$, with $\mathcal{Y}=\{-1,1\}$. Assume we are given a dataset $S_n := (X_i, Y_i)_{1\leq i\leq n}$ of i.i.d. points in $\mathcal{X}\times\mathcal{Y}$ distributed according to some probability $\mathbb{P}$. Let $\mathcal{F}$ be a set of real-valued functions defined on $\mathcal{X}$. Let $\sigma_1,\dots,\sigma_n$ be n i.i.d. Rademacher variables, i.e. $\sigma_i \in \{-1,1\}$ with $\mathbb{P}(\sigma_i = 1) = \frac{1}{2}$. We define the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ of $\mathcal{F}$ to be:

$$\mathcal{R}_n(\mathcal{F}) = \frac{2}{n}\mathbb{E}_{X,\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(X_i)\right]. \tag{1}$$

Remark: The above definition is slightly different than the one in the lecture slides as it does not take the absolute value. Both definitions are considered in the literature, but this one is easier to deal with.

Introduce the set $\mathcal{G} := \{g(x,y) := \varphi(yf(x)) | f\in\mathcal{G}\}$ for some real-valued function φ that is L-Lipschitz, i.e. $\varphi(t)-\varphi(s)\leq L|t-s|$. For simplicity we write $z=(x,y)$ for any $\mathcal{X}\times\mathcal{Y}$ and define $Z_i = (X_i, Y_i)$.

1. Define $\mathcal{R}_n(\mathcal{G}) = \frac{2}{n}\mathbb{E}[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i\varphi(Y_i f(X_i))]$. Prove that:

$$\mathcal{R}_n(\mathcal{G}) \leq L\mathcal{R}_n(\mathcal{F}).$$

2. Define the population risk $R_\varphi(f)$ and empirical risk $R^n_\varphi(f,\mathcal{S}_n)$ of a function f to be:

$$R_\varphi(f) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\varphi(yf(x))\right], \qquad R^n_\varphi(f,S_n) = \frac{1}{n}\sum_{i=1}^{n}\varphi(Y_i f(X_i)). \tag{2}$$

Prove that:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} R_\varphi(f) - R^n_\varphi(f, S_n)\right] \leq 2L\mathcal{R}_n(\mathcal{F}).$$

and that:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} R^n_\varphi(f, S_n) - R_\varphi(f)\right] \leq 2L\mathcal{R}_n(\mathcal{F}).$$

3. Consider $\hat{f}_n$ to be a minimizer of the empirical risk $R^n_\varphi(\hat{f}_n, \mathcal{S}_n) = \min_{f\in\mathcal{F}} R^n_\varphi(f, \mathcal{S}_n)$ and denote by $R^\star_\varphi$ the optimal population risk over the class of measurable functions. Show that:

$$\mathbb{E}_{S_n}[R_\varphi(\hat{f}_n)] - R^\star_\varphi \leq 4L\mathcal{R}_n(\mathcal{F}) + \inf_{f\in\mathcal{F}} R_\varphi(f) - R^\star_\varphi.$$

Proof. • Proof of (1).

Consider a set of maps $\alpha_i(f)$ and $\beta_i(f)$ indexed by $1 \leq i \leq n$ defined as:

$$\alpha_i(f) = \varphi(Y_i f(X_i)), \qquad \beta_i(f) = Lf(X_i). \tag{3}$$

Introduce the vectors maps $\Psi_j(f)$ for $0 \leq j \leq n$ with: $\Psi_0(f) := (\alpha_1(f), ..., \alpha_n(f))$ and $\Psi_n(f) := (\beta_1(f), ..., \beta_n(f))$, and for $0 < j < n$:

$$\Psi_j(f) := (\beta_1(f), ..., \beta_j(f), \alpha_{j+1}(f), ..., \alpha_n(f)).$$

Finally, for some vector map $\Psi = (\psi_1, ..., \psi_n)$, where φ_i can be either α_i or β_i, we introduce the notation $\mathcal{R}(\Psi_j(\mathcal{F}))$:

$$\mathcal{R}(\Psi(\mathcal{F})) := \frac{2}{n}\mathbb{E}[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \psi_i(f)].$$

With this notation and if, we get

$$\mathcal{R}(\Psi_0(\mathcal{F})) = \mathcal{R}_n(\mathcal{G}), \qquad \mathcal{R}(\Psi_n(\mathcal{F})) = L\mathcal{R}_n(\mathcal{F}) \tag{4}$$

We will prove that for any $0 \leq j < n$:

$$\mathcal{R}(\Psi_j(\mathcal{F})) \leq \mathcal{R}(\Psi_{j+1}(\mathcal{F})).$$

The above inequality means that we can always "flip" a component $\alpha_j(f)$ to $\beta_j(f)$ without decreasing the Rademacher complexity. It allows to directly conclude that $\mathcal{R}(\Psi_0(\mathcal{F})) \leq \mathcal{R}(\Psi_n(\mathcal{F}))$ which is the desired result. Without loss of generality, we only need to prove the inequality for $j = 0$, as the proof can be applied similarly to $j > 0$.

$$
\begin{aligned}
\mathcal{R}(\Psi_0(\mathcal{F})) &= \frac{2}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i\alpha_i(f)\right] \\
&= \frac{2}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sigma_1\alpha_1(f) + \sum_{i=2}^{n}\sigma_i\alpha_i(f)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\alpha_1(f) + \sum_{i=2}^{n}\sigma_i\alpha_i(f)\right) + \sup_{f\in\mathcal{F}}\left(-\alpha_1(f) + \sum_{i=2}^{n}\sigma_i\alpha_i(f)\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{f,f'\in\mathcal{F}}\left(\varphi(Y_1f(X_1)) - \varphi(Y_1f'(X_1)) + \sum_{i=2}^{n}\sigma_i(\alpha_i(f) + \alpha_i(f'))\right)\right] \\
&\leq \frac{1}{n}\mathbb{E}\left[\sup_{f,f'\in\mathcal{F}}\left(L|f(X_1) - f'(X_1)| + \sum_{i=2}^{n}\sigma_i(\alpha_i(f) + \alpha_i(f'))\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{f,f'\in\mathcal{F}}\left(Lf(X_1) - Lf'(X_1)| + \sum_{i=2}^{n}\sigma_i(\alpha_i(f) + \alpha_i(f'))\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{f,f'\in\mathcal{F}}\left(\beta_1(f) + \sum_{i=2}^{n}\sigma_i\alpha_i(f)\right) + \left(-\beta_1(f') + \sum_{i=2}^{n}\sigma_i\alpha_i(f')\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\beta_1(f) + \sum_{i=2}^{n}\sigma_i\alpha_i(f)\right) + \sup_{f'\in\mathcal{F}}\left(-\beta_1(f') + \sum_{i=2}^{n}\sigma_i\alpha_i(f')\right)\right] \\
&= \frac{2}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\sigma_1\beta_1(f) + \sum_{i=2}^{n}\sigma_i\alpha_i(f)\right)\right] = \mathcal{R}(\Psi_1(\mathcal{F}))
\end{aligned}
$$

We are able to drop the absolute value (in the step after the inequality), since the roles of f and f' are symmetric and the supremum is achieved when $f(X_1) - f'(X_1)$ is positive. This completes the proof.

- Proof of (2).

Consider an copy $S'_n = (X'_i, Y'_i)_{1\leq i\leq n}$ of the data S_n that is independent of it. It is easy to see that $R_\varphi(f) = \mathbb{E}_{S'_n}[R^n_\varphi(f, S'_n)]$

$$\begin{aligned}
\mathbb{E}_{S_n}\left[\sup_{f\in\mathcal{F}} R_\varphi(f) - R^n_\varphi(f,S_n)\right] &= \mathbb{E}_{S_n}\left[\sup_{f\in\mathcal{F}} \mathbb{E}_{S'_n} R^n_\varphi(f,S'_n) - R^n_\varphi(f,S_n)\right] \\
&\leq \mathbb{E}_{S_n,S'_n}\left[\sup_{f\in\mathcal{F}} R^n_\varphi(f,S'_n) - R^n_\varphi(f,S_n)\right] \\
&\leq \frac{1}{n}\mathbb{E}_{S_n,S'_n}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \varphi(Y_i^{'} f(X_i^{'})) - \varphi(Y_i f(X_i))\right] \\
&= \frac{1}{n}\mathbb{E}_{S_n,S'_n,\sigma}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \sigma_i\left(\varphi(Y_i^{'} f(X_i^{'})) - \varphi(Y_i f(X_i))\right)\right] \\
&\leq 2\mathcal{R}_n(\mathcal{G}) \leq 2L\mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

The same proof holds for the second inequality.

- Proof of (3).

Assume for simplicity that $f^\star$ is a minimizer of the population risk $R_\varphi(f)$.

$$\begin{aligned}
R_\varphi(\hat{f}_n) - R_\varphi(f^\star) &= \left(R_\varphi(\hat{f}_n) - R^n_\varphi(\hat{f}_n,\mathcal{S}_n)\right) + \underbrace{\left(R^n_\varphi(\hat{f}_n,\mathcal{S}_n) - R^n_\varphi(f^\star,\mathcal{S}_n)\right)}_{\leq 0} + \left(R^n_\varphi(f^\star,\mathcal{S}_n) - R_\varphi(f^\star)\right) \\
&\leq \left(R_\varphi(\hat{f}_n) - R^n_\varphi(\hat{f}_n,\mathcal{S}_n)\right) + \left(R^n_\varphi(f^\star,\mathcal{S}_n) - R_\varphi(f^\star)\right) \\
&\leq \sup_{f\in\mathcal{F}}\left(R_\varphi(f) - R^n_\varphi(f,\mathcal{S}_n)\right) + \sup_{f\in\mathcal{F}}\left(R^n_\varphi(f,\mathcal{S}_n) - R_\varphi(f)\right)
\end{aligned}$$

Taking the expectation w.r.t. data we get:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_n}[R_\varphi(\hat{f}_n)] - R_\varphi(f^\star) &\leq \mathbb{E}_{\mathcal{S}_n}\left[\sup_{f\in\mathcal{F}}\left(R_\varphi(f) - R^n_\varphi(f,\mathcal{S}_n)\right)\right] + \mathbb{E}_{\mathcal{S}_n}\left[\sup_{f\in\mathcal{F}}\left(R^n_\varphi(f,\mathcal{S}_n) - R_\varphi(f)\right)\right] \\
&\leq 2L\mathcal{R}_n(\mathcal{F}) + 2L\mathcal{R}_n(\mathcal{F}) = 4L\mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

Recalling that $R_\varphi(f^\star) = \inf_{f\in\mathcal{F}} R_\varphi(f)$, we get the desired result. □