Lecture-10: Rademacher Complexity

1 Rademacher complexity

PAC learning guarantees were for finite hypothesis sets. However typical hypothesis sets in machine learning problems are infinite, e.g. set of all hyperplanes in SVM. We will generalize existing results and derive general learning guarantees for infinite hypothesis sets. We will reduce the infinite hypothesis set to a finite set depending on the notion of complexity. First notion is *Rademacher complexity*, which is difficult to compute empirically for many hypothesis sets. We then study combinatorial notions of complexity, *growth function* and the *VC-dimension*. We relate Rademacher complexity to growth function, and then bound the growth function by the VC-dimension, which are easy to bound or compute in many cases.

Definition 1.1. Consider a hypothesis set $H \subset \mathcal{Y}^{\mathcal{X}}$ and loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Let $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$, then for each hypothesis $h \in H$, we can associate a function $g : \mathcal{Z} \to \mathbb{R}$ defined for all $(x,y) \in \mathcal{Z}$ as $g(x,y) \triangleq L(h(x),y)$, which captures the corresponding loss L. The family of loss function associated to hypothesis set H is defined as $G \triangleq \{(x,y) \mapsto L(h(x),y) : h \in H\}$.

Definition 1.2. For any $g \in \mathbb{R}^{\mathcal{Z}}$ and m-sized sample $z \in \mathcal{Z}^m$, we denote by $g_z \triangleq (g(z_1),\dots,g(z_m)) \in \mathbb{R}^m$.

Definition 1.3 (Rademacher random vector). An *i.i.d.* random vector $X : \Omega \to \{-1,1\}^m$ distributed uniformly is called a *Rademacher random vector*.

Definition 1.4 (Empirical Rademacher complexity). Let $G \subseteq [a,b]^{\mathcal{Z}}$ be a family of functions, a fixed labeled sample $z = (z_1,\dots,z_m) \in \mathcal{Z}^m$ of size m, and $\sigma : \Omega \to \{-1,1\}^m$ an independent m-length Rademacher vector. Then, the *empirical Rademacher complexity* of G with respect to the labeled sample z is defined as

$$\hat{\mathcal{R}}_z(G) \triangleq \mathbb{E}_\sigma \left[\sup_{g \in G} \frac{1}{m} \langle \sigma, g_z \rangle \right] = \mathbb{E} \left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^m \sigma_i g(z_i) \right].$$

Remark 1. The inner product $\langle \sigma, g_z \rangle$ measures the correlation of g_z with random noise σ, and the supremum over all $g \in G$ measures how well the hypothesis class H correlates with σ over the labeled sample z. This is a measure of richness/complexity of class G, since richer families can generate more g_z and better correlate with random noise on average.

Definition 1.5 (Rademacher complexity). Let D be the unknown fixed distribution according to which labeled sample $z \in \mathcal{Z}^m$ is drawn in an *i.i.d.* fashion. For any $m \in \mathbb{N}$, the *Rademacher complexity* of a family of loss functions G is the mean of empirical Rademacher complexity for sample z, and denoted by

$$\mathcal{R}_m(G) \triangleq \mathbb{E}\hat{\mathcal{R}}_z(G).$$

Remark 2. The Rademacher complexity captures the richness of a family of functions by measuring the degree to which a hypothesis set can fit random noise.

Definition 1.6 (Bounded difference property). A function $f : \mathcal{X}^m \to \mathbb{R}$ is said to have the *bounded difference property* with *bounding vector* $c \in \mathbb{R}_+^m$, if for any $x,y \in \mathcal{X}^m$ differing only at location $i \in [m]$,

$$|f(x) - f(y)| \leqslant c_i. \tag{1}$$

Remark 3. Let $G \subseteq [0,1]^{\mathcal{Z}}$ and $a \in [-1,1]^m$, we define a map $k_a : \mathcal{Z}^m \to \mathbb{R}$ for all $z \in \mathcal{Z}^m$ as $k_a(z) \triangleq \sup_{g \in G} \langle a, g_z \rangle$. Fix $i \in [m]$ and choose $w,z \in \mathcal{Z}^m$ such that $w_j = z_j$ for all $j \in [m] \setminus \{i\}$. Then, we have

$$|k_a(z) - k_a(w)| \leqslant \sup_{g \in G} \left| \sum_{j=1}^m a_j g(z_j) - \sum_{j=1}^m a_j g(w_j) \right| = \sup_{g \in G} |a_i| \, |g(z_i) - g(w_i)| \leqslant 1.$$

It follows that map k_a has bounded difference property with bounding vector $\mathbf{1}$.

Lemma 1.7. *Let $G \subseteq [0,1]^{\mathcal{Z}}$ and $\delta > 0$. Then,* $P\left\{\mathcal{R}_m(G) \leqslant \hat{\mathcal{R}}_z(G) + \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\right\} \geqslant 1-\delta$.

Proof. From Remark 3, we observe that $\hat{\mathcal{R}}_z(G)$ satisfies the bounded difference property with bounding vector $\frac{1}{m}\mathbf{1}$. Applying the McDiarmid's inequality to $\hat{\mathcal{R}}_z(G)$ for any $\epsilon > 0$, we obtain

$$P\left\{\hat{\mathcal{R}}_z(G) - \mathbb{E}\hat{\mathcal{R}}_z(G) \leqslant -\epsilon\right\} \leqslant e^{-2m\epsilon^2}.$$

The result follows by setting $2e^{-2m\epsilon^2} = \delta$. □

Definition 1.8. For any labeled sample $z \in \mathcal{Z}^m$ and loss function $g \in G$, we denote the empirical average of g over labeled sample z as $\hat{\mathbb{E}}_z[g] \triangleq \frac{1}{m}\langle \mathbf{1}, g_z\rangle = \frac{1}{m}\sum_{i=1}^m g(z_i)$. The mean of empirical average $\hat{\mathbb{E}}_z[g]$ is denoted by $\mathbb{E}g \triangleq \mathbb{E}\hat{\mathbb{E}}_z[g] = \mathbb{E}g(z_1)$.

Remark 4. When $g(z_i) = L(h(x_i), y_i)$, we have $\hat{\mathbb{E}}_z[g] = \hat{R}_z(h)$ and $\mathbb{E}g = R(h)$.

Remark 5. Consider $k : G \times \mathcal{Z}^m \to \mathbb{R}$ and $\ell : \mathcal{Z}^m \to \mathbb{R}$ defined as $\ell(z) \triangleq \sup_{g\in G} k(g,z)$, then we observe that $k(g,z) \leqslant \ell(z)$ for all $g \in G$. Applying expectation on both sides, we obtain $\mathbb{E}k(g,z) \leqslant \mathbb{E}\ell(z)$ for all $g \in G$ and hence $\sup_{g\in G}\mathbb{E}k(g,z) \leqslant \mathbb{E}\ell(z)$.

Remark 6. Consider a map $g \in G \subseteq [0,1]^{\mathcal{Z}}$ and two *i.i.d.* examples $z_1, z_1' \in \mathcal{Z}$, then $g(z_1) - g(z_1')$ is identical in distribution to $g(z_1') - g(z_1)$. That is, if $\sigma_1 \in \{-1,1\}$ is a zero mean random variable independent of z_1, z_1', then $\sigma_1(g(z_1) - g(z_1'))$ is identically distributed to $g(z_1) - g(z_1')$.

Theorem 1.9. *Consider the following events defined for $\delta > 0, g \in G \subseteq [0,1]^{\mathcal{Z}}$, and* i.i.d. *sample $z \in \mathcal{Z}^m$,*

$$E_g \triangleq \left\{\mathbb{E}g - \hat{\mathbb{E}}_z[g] \leqslant 2\mathcal{R}_m(G) + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\}, \qquad F_g \triangleq \left\{\mathbb{E}g - \hat{\mathbb{E}}_z[g] \leqslant 2\hat{\mathcal{R}}_z(G) + 3\sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\right\}.$$

Then, $P\Big(\cap_{g\in G} E_g\Big) \geqslant 1-\delta$ *and* $P\Big(\cap_{g\in G} F_g\Big) \geqslant 1-\delta$.

Proof. We consider the following function $\Phi : \mathcal{Z}^m \to \mathbb{R}$ defined for all $z \in \mathcal{Z}^m$ as $\Phi(z) \triangleq \sup_{g\in G}(\mathbb{E}g - \hat{\mathbb{E}}_z[g])$. From Remark 3, it follows that Φ has the bounded difference property with bounding vector $\frac{1}{m}\mathbf{1}$. Applying McDiarmid's inequality to Φ, we obtain for any $\delta > 0$

$$P\left\{\Phi(z) \leqslant \mathbb{E}\Phi(z) + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\} = P\Big(\cap_{g\in G}\left\{\mathbb{E}g - \hat{\mathbb{E}}_z[g] \leqslant \mathbb{E}\Phi(z) + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\}\Big) \geqslant 1-\delta.$$

We next bound $\mathbb{E}\Phi(z)$ by the mean of empirical average difference for samples z, z', sampled *i.i.d.* from the fixed unknown distribution D, and applying sub-additivity of supremum function from Remark 5 for conditional expectation $\mathbb{E}_{z'}[\cdot] = \mathbb{E}[\cdot \mid z]$, to obtain

$$\mathbb{E}\Phi(z) = \mathbb{E}\left[\sup_{g\in G}(\mathbb{E}[g] - \hat{\mathbb{E}}_z[g])\right] = \mathbb{E}\left[\sup_{g\in G}\mathbb{E}_{z'}\left[\hat{\mathbb{E}}_{z'}[g] - \hat{\mathbb{E}}_z[g]\right]\right] \leqslant \mathbb{E}\left[\sup_{g\in G}(\hat{\mathbb{E}}_{z'}[g] - \hat{\mathbb{E}}_z[g])\right].$$

Applying Remark 6 to sum of m such differences, we observe that the inner product $\langle\sigma, g_{z'} - g_z\rangle$ for *i.i.d.* Rademacher vector $\sigma \in \{-1,1\}^m$ has an identical distribution to $\langle \mathbf{1}, g_{z'} - g_z\rangle$ for z, z' being *i.i.d.*. Therefore, we have

$$\mathbb{E}\Phi(z) \leqslant \mathbb{E}\left[\sup_{g\in G}\frac{1}{m}\langle\sigma, g_{z'} - g_z\rangle\right] \leqslant \mathbb{E}\left[\sup_{g\in G}\frac{1}{m}\langle\sigma, g_{z'}\rangle\right] + \mathbb{E}\left[\sup_{g\in G}\frac{1}{m}\langle-\sigma, g_z\rangle\right] = 2\mathcal{R}_m(G).$$

It follows that $P(\cap_{g\in G}E_g) \geqslant 1-\delta$. From union bound and Lemma 1.7, we obtain

$$P\left(\cup_{g\in G}\left\{\mathbb{E}g - \hat{\mathbb{E}}_z[g] > 2\mathcal{R}_m(G) + \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\right\} \cup \left\{\mathcal{R}_m(G) > \hat{\mathcal{R}}_z(G) + \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\right\}\right) \leqslant \delta.$$

Using the fact that $E_g \cap \left\{\mathcal{R}_m(G) \leqslant \hat{\mathcal{R}}_z(G) + \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\right\} \subseteq F_g$, we obtain $P(\cap_{g\in G}F_g) \geqslant 1-\delta$. □

Lemma 1.10. *Let $\mathcal{Y} \triangleq \{-1,1\}$ and $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$, the hypothesis set $H \subseteq \mathcal{Y}^{\mathcal{X}}$, and G be the family of loss functions associated to the hypothesis set H for the zero-one loss, i.e. $G \triangleq \left\{(x,y) \mapsto \mathbb{1}_{\{h(x)\neq y\}} : h \in H\right\}$. For any labeled sample $z \in \mathcal{Z}^m$, let $x \in \mathcal{X}^m$ be the unlabeled sample. Then, $\hat{\mathfrak{R}}_z(G) = \frac{1}{2}\hat{\mathfrak{R}}_x(H)$.*

Proof. Since $\sum_{i=1}^m \sigma_i$ remains constant for all $h \in H$ and $\mathbb{1}_{\{h(x_i)\neq y_i\}} = \frac{1-y_ih(x_i)}{2}$, we can write

$$\hat{\mathfrak{R}}_z(G) = \mathbb{E}_\sigma\left[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^m \sigma_i\Big(\frac{1-y_ih(x_i)}{2}\Big)\right] = \mathbb{E}_\sigma\left[\frac{1}{m}\sum_{i=1}^m \sigma_i + \sup_{h\in H}\frac{1}{m}\sum_{i=1}^m -\frac{1}{2}\sigma_i y_i h(x_i)\right].$$

Recall that $\mathbb{E}\sigma_i = 0$ for all $i \in [m]$ and hence from linearity of expectation, we have $\mathbb{E}\frac{1}{m}\sum_{i=1}^m \sigma_i = 0$. Further, $-\sigma \circ y = (-\sigma_i y_i \in \mathcal{Y} : i \in [m])$ has same distribution as $\sigma = (\sigma_i \in \mathcal{Y} : i \in [m])$, and therefore

$$\hat{\mathfrak{R}}_z(G) = \frac{1}{2}\mathbb{E}_\sigma\left[\sup_{h\in H}\frac{1}{m}\langle -\sigma\circ y, h_x\rangle\right] = \frac{1}{2}\mathbb{E}_\sigma\left[\sup_{h\in H}\frac{1}{m}\langle \sigma, h_x\rangle\right] = \frac{1}{2}\hat{\mathfrak{R}}_x(H).$$

□

Theorem 1.11 (Rademacher complexity bounds – binary classification). *For any hypothesis set $H \subseteq \mathcal{Y}^{\mathcal{X}}$, binary labels $\mathcal{Y} = \{-1,+1\}$,* i.i.d. *labeled sample $z \in \mathcal{Z}^m$, and $\delta > 0$, we define events*

$$E_h \triangleq \left\{R(h) \leqslant \hat{R}(h) + \mathfrak{R}_m(H) + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\} \qquad F_h \triangleq \left\{R(h) \leqslant \hat{R}(h) + \hat{\mathfrak{R}}_x(H) + 3\sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\right\}.$$

Then, $P(\cap_{h\in H}E_h) \geqslant 1-\delta$ and $P(\cap_{h\in H}F_h) \geqslant 1-\delta$.

Proof. The result follow from Theorem 1.9 and Lemma 1.10, the fact that $\mathbb{E}g = R(h)$ and $\hat{\mathbb{E}}_z[g] = \hat{R}(h)$, and there is a $g \in G$ for each $h \in H$. □

Remark 7. The second learning bound is data dependent, and very useful if we can efficiently compute the empirical Rademacher complexity $\hat{\mathfrak{R}}_x(H)$. Since σ and $-\sigma$ have the same distribution, we get

$$\hat{\mathfrak{R}}_x(H) \triangleq \mathbb{E}\left[\sup_{h\in H}\frac{1}{m}\langle -\sigma, h\rangle\right] = -\mathbb{E}\left[\inf_{h\in H}\frac{1}{m}\langle \sigma, h\rangle\right].$$

for a fixed value of σ, computing $\inf_{h\in H}\frac{1}{m}\langle\sigma,h\rangle$ is equivalent to an *empirical risk minimization* problem, which is known to be computationally hard for some hypothesis sets.

A McDiarmid's inequality

Definition A.1 (Martingale difference). A random sequence $V : \Omega \to \mathbb{R}^{\mathbb{N}}$ is a *martingale difference sequence* with respect to a random sequence $X : \Omega \to \mathbb{R}^{\mathbb{N}}$ if V_n is a function of $X_1, \ldots, X_n$ for all $n \in \mathbb{N}$, and

$$\mathbb{E}[V_{n+1} \mid X_1, \ldots, X_n] = 0.$$

Lemma A.2. *Let V and Z be random variables satisfying $\mathbb{E}[V \mid Z] = 0$ and $f(Z) \leqslant V \leqslant f(Z) + c$ for some function f and constant $c \geqslant 0$. Then, for all $t > 0$, we have*

$$\mathbb{E}[e^{tV} \mid Z] \leqslant e^{t^2c^2/8}.$$

Proof. The result follows from Hoeffding's Lemma for conditional expectation given Z, where $[a,b] = [f(Z), f(Z)+c]$. □

Theorem A.3 (Azuma's inequality). *Let $V : \Omega \to \mathbb{R}^{\mathbb{N}}$ be a martingale difference sequence with respect to the random sequence $X : \Omega \to \mathbb{R}^{\mathbb{N}}$ and assume that for all $i \in \mathbb{N}$ there is a constant $c_i \geqslant 0$ and random variable Z_i, which is a function of $X_1, \ldots, X_{i-1}$, that satisfies $Z_i \leqslant V_i \leqslant Z_i + c_i$. Defining $\sigma^2 \triangleq \sum_{i=1}^m c_i^2 = \|c\|_2^2$, we have for all $\epsilon > 0$ and $m \in \mathbb{N}$,*

$$P\left\{\sum_{i=1}^m V_i \geqslant \epsilon\right\} \leqslant e^{-2\epsilon^2/\sigma^2}, \qquad P\left\{\sum_{i=1}^m V_i \leqslant -\epsilon\right\} \leqslant e^{-2\epsilon^2/\sigma^2}.$$

Proof. For any $m \in \mathbb{N}$, we can define $S_m \triangleq \sum_{i=1}^m V_i$, then by Chernoff bound and Hoeffding Lemma for martingale difference sequence and conditional expectation, we have

$$P\{S_m \geqslant \epsilon\} \leqslant e^{-t\epsilon}\mathbb{E}[e^{tS_m}] = e^{-t\epsilon}\mathbb{E}[e^{tS_{m-1}}\mathbb{E}[e^{tV_m}|X_1,\dots,X_{m-1}]] \leqslant e^{-t\epsilon}\mathbb{E}[e^{tS_{m-1}}]e^{t^2c_m^2/8} \leqslant \exp\left(-t\epsilon + \frac{t^2\sigma^2}{8}\right).$$

The result for the first part follows by taking $t^* = \frac{4\epsilon}{\sigma^2}$. The second part can be proved similarly. □

Theorem A.4 (McDiarmid's inequality). *Let $f : \mathcal{X}^m \to \mathbb{R}$ be a function with the bounded difference property with bounding vector $c \in \mathbb{R}_+^m$, and $X : \Omega \to \mathcal{X}^m$ be an independent random vector. For all $\epsilon > 0$, we have*

$$P\{f(X) - \mathbb{E}f(X) \geqslant \epsilon\} \leqslant e^{-2\epsilon^2/\|c\|_2^2}, \qquad P\{f(X) - \mathbb{E}f(X) \leqslant -\epsilon\} \leqslant e^{-2\epsilon^2/\|c\|_2^2}.$$

Proof. It suffices to show that $f(X) - \mathbb{E}f(X) = \sum_{i=1}^m V_i$ for some martingale difference sequence $V : \Omega \to \mathbb{R}^m$ with respect to the sequence $X : \Omega \to \mathbb{R}^m$ and for each $i \in [m]$ there exists a constant c_i and a random variable Z_i a function of $X_1, \dots, X_{i-1}$ such that $Z_i \leqslant V_i \leqslant Z_i + c_i$. We define such a random sequence $V : \Omega \to \mathbb{R}^m$ for all $i \in [m]$, as

$$V_i \triangleq \mathbb{E}[f(X) \mid X_1, \dots, X_i] - \mathbb{E}[f(X) \mid X_1, \dots, X_{i-1}].$$

We can verify that $\sum_{k=1}^m V_i = f(X) - \mathbb{E}f(X)$ from the telescopic sum. Next, we observe that for each $i \in [m]$, (a) V_i is a function of $X_1, \dots, X_i$, and $\mathbb{E}[V_i \mid X_1, \dots, X_{i-1}] = 0$ from the tower property of conditional expectation. It follows that V is a martingale difference vector with respect to random vector X. We fix $i \in [m]$. We define upper and lower bounds for V_i as

$$U_i \triangleq \sup_{X_i \in \mathcal{X}} \mathbb{E}[f(X) \mid X_1, \dots, X_{i-1}, X_i] - \mathbb{E}[f(X) \mid X_1, \dots, X_{i-1}],$$
$$L_i \triangleq \inf_{X_i \in \mathcal{X}} \mathbb{E}[f(X) \mid X_1, \dots, X_{i-1}, X_i] - \mathbb{E}[f(X) \mid X_1, \dots, X_{i-1}].$$

It follows that L_i and U_i are functions of $X_1, \dots, X_{i-1}$ and $L_i \leqslant V_i \leqslant U_i$. Consider inputs $X \in \mathcal{X}^m$ and $Y \in \mathcal{X}^m$ such that $X_j = Y_j$ for $j \neq i$ and $X_i \neq Y_i$ where X_i and Y_i are independent and identically distributed. It follows that $\mathbb{E}[f(X) \mid X_1, \dots, X_{i-1}] = \mathbb{E}[f(Y) \mid X_1, \dots, X_{i-1}]$. Further, from bounded difference property (1) of function f, we have $c_i \geqslant f(X) - f(Y)$ for any such X, Y. Taking conditional expectation with respect to last $m - i + 1$ identical elements of vectors X, Y, we get

$$\mathbb{E}[f(X) \mid X_1, \dots, X_{i-1}, X_i] - \mathbb{E}[f(Y) \mid X_1, \dots, X_{i-1}, Y_i] \leqslant c_i.$$

Since this inequality is true for all $X_i, Y_i \in \mathcal{X}$, we can take supremum over all such pairs (X_i, Y_i) on the left hand side of the above equation, to obtain

$$c_i \geqslant \sup_{X_i \in \mathcal{X}} \mathbb{E}[f(X) \mid X_1, \dots, X_{i-1}, X_i] - \inf_{Y_i \in \mathcal{X}} \mathbb{E}[f(Y) \mid X_1, \dots, X_{i-1}, Y_i] = \sup_{X_i \in \mathcal{X}} V_i - \inf_{X_i \in \mathcal{X}} V_i = U_i - L_i.$$

□