7 Rademacher Complexity: Properties & Applications

From now on, we will rely on the following (informal) inequality to establish generalization. With probability at least $1-\delta$,

$$R(\hat{f}) - R(f) \leq 4\mathfrak{R}_n(\mathcal{G}) + B\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{7.1}$$

where $\mathcal{G} = \{(y,x) \to \ell((y,x),f) \text{ where } f \in \mathcal{F}\}$. Formal statement is given in Theorem 23. Key observation is that, in order to achieve generalization, we only need to find an upper bound to Rademacher complexity $\mathfrak{R}_n(\mathcal{G})$ that decays with n.

7.1 Properties of Rademacher complexity

Below, we state some properties of Rademacher complexity.

1. Monotonicity: if $\mathcal{F}_1 \subseteq \mathcal{F}_2$ then $\mathfrak{R}_n(\mathcal{F}_1) \leq \mathfrak{R}_n(\mathcal{F}_2)$
2. Linear combination: if $\mathcal{F}_1 + \mathcal{F}_2 = \{f_1 + f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ then $\mathfrak{R}_n(\mathcal{F}_1 + \mathcal{F}_2) = \mathfrak{R}_n(\mathcal{F}_1) + \mathfrak{R}_n(\mathcal{F}_2)$
3. Scaling: if $c \in \mathbb{R}$ and $c\mathcal{F} = \{cf : f \in \mathcal{F}\}$ then $\mathfrak{R}_n(c\mathcal{F}) = |c|\mathfrak{R}_n(\mathcal{F})$
4. Convex Hull of $\mathcal{F}$: if $|\mathcal{F}| < \infty$ then $\mathfrak{R}_n(\text{convex-hull}(\mathcal{F})) = \mathfrak{R}_n(\mathcal{F})$

The above properties follow from the definition of Rademacher complexity, and their proof is left to reader as an exercise.

We notice that in the generalization bound 7.1, the Rademacher complexity of the function class $\mathcal{G}$ plays a key role. Our objective is to connect this bound to the complexity of the hypothesis functions $\mathcal{F}$. The following strong result serves to that purpose.

Lemma 28 (**Talagrand's contraction principal**). *Let g be an L-Lipschitz continuous function, and $\mathcal{F}$ is a function class. Then,*

$$\mathfrak{R}_n(g \circ \mathcal{F}) \leq L \cdot \mathfrak{R}_n(\mathcal{F}).$$

Proof of the above lemma is involved and skipped in class. We emphasize that Talagrand's lemma can be used to map the Rademacher complexity of $\mathcal{G}$, to that of $\mathcal{F}$ which is known for certain function classes. We first go over an example to demonstrate how to use the above result.

Example. [Support Vector Machines] In our first example, we visit a classical learning algorithm. As before, we denote our data pairs with $z = (x,y)$ and $y \in \{\pm 1\}$, $x \in \mathbb{R}^d$ and loss $\ell(z,f) = \max\{0, 1 - y \cdot f(x)\}$ which is often called as the hinge-loss. Let's define the function $\phi(s) = \max\{0, 1-s\}$, and notice that $\ell(z,f) = \phi(yf(x))$ and ϕ is 1-Lipschitz continuous.

Recall that the generalization bound we obtained in (7.1) relies on the Rademacher complexity of the loss class $\mathfrak{R}_n(\mathcal{G})$, where $\mathcal{G} = \{z = (y,x) \to \phi(yf(x)), f \in \mathcal{F}\}$. In order to connect this to the complexity of $\mathcal{F}$, we define $\mathcal{H} = \{z = (y,x) \to yf(x), f \in \mathcal{F}\}$, and we notice that $\mathcal{G} = \phi \circ \mathcal{H}$. By the Talagrand's contraction principal, we can bound the Rademacher complexity of $\mathcal{H}$ as

$$\mathfrak{R}_n(\mathcal{G}) \leq 1 \cdot \mathfrak{R}_n(\mathcal{H}),$$

since ϕ is 1-Lipschitz. But

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{H}) &= \mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h(z_i)\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i y_i f(x_i)\right], \quad \sigma_i y_i \overset{d}{=} \sigma_i \ \ \{**\} \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} n^{-1}\sum_{i}\sigma_i f(x_i)\right] \\
&= \mathfrak{R}_n(\mathcal{F}).
\end{aligned}$$

Note that in the second line of equality, $y_i \overset{d}{=} \sigma_i y_i$ comes from the fact that $\sigma_i y_i \perp\!\!\!\perp x_i$ even though $y_i \not\perp\!\!\!\perp x_i$ (verify this).

Therefore we can conclude that if we characterize $\mathfrak{R}_n(\mathcal{F})$, then we can characterize $\mathfrak{R}_n(\mathcal{G})$. This still doesn't complete the whole picture, but we are making progress.

Example. [Smooth relaxations to 0-1 loss] Smooth surrogate relaxations to 0-1 loss are commonly employed in machine learning. The basic idea is that we would like to minimize the misclassification error which is based on the 0-1 loss, but in practice we cannot minimize this ill-behaved loss function due to its discontinuous behavior. For this reason, we consider surrogate losses to 0-1 loss which are its smoothed versions. In this example, we will see how using a surrogate loss may lead to a worsened generalization error.

We consider a binary classification problem where we denote the data as $z = (y, x)$, with class labels $y \in \{\pm 1\}$, and $f \in \mathcal{F}$, then the 0-1 loss function is given as $\mathbb{1}_{\{yf(x)\leq 0\}}$ and can be equivalently written as follows.

$$\ell_0(z, f) \triangleq \ell_0(yf(x)) \ \text{ where } \ \ell_0(s) = \begin{cases} 1 & \text{if } s < 0, \\ 0 & \text{if } s \geq 0. \end{cases}$$

We will assume that the product of response and prediction satisfies the following property.

Assumption 1. *We assume the following holds*

$$\exists C > 0, \forall f \in \mathcal{F} \ \ \mathbb{P}(0 \leq yf(x) \leq \tau) \leq C\tau.$$

for small τ. The assumption is simply stating that the probability of misclassifying a sample gets smaller with smaller margin. This assumption is not very transparent as is and it can unpacked for certain data distributions, but for now, let's work with this assumption to obtain our result.

Let's introduce a surrogate loss function which will serve as a relaxation to 0-1 loss.

$$\ell_\tau(s) = \begin{cases} 1 & \text{if } s < 0 \\ 1 - \frac{s}{\tau} & \text{if } 0 \leq s < \tau \\ 0 & \text{if } s \geq \tau \end{cases}$$

Another motivation for using the above loss is that the 0-1 loss function assignes the same penalty for low and high confidence predictions. Instead we would like to encourage higher confidence predictions with τ-margin sensitivity.

The loss function $\ell_\tau(s)$ is Lipschitz continuous with constant $L = 1/\tau$. Denote

$$\hat{f}_\tau = \operatorname*{argmin}_{f \in \mathcal{F}} \ \hat{R}_\tau(f)$$
$$f^*_\tau = \operatorname*{argmin}_{f \in \mathcal{F}} \ R_\tau(f)$$

With this notation, we have f^*_0 as the minimizer of the population risk $R_0(f)$. We make the following observations:

1. By Theorem 7.1, with probability at least $1 - \delta$, we have

$$R_\tau(\hat{f}_\tau) \leq R_\tau(f^*_\tau) + 4\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{2\log(2/\delta)}{n}}$$

   where $\mathcal{G} = \{z \to \ell_\tau(z, f) \ : \ f \in \mathcal{F}\}$.

2. By an argument similar to the one in previous example, we have $\mathfrak{R}_n(\mathcal{G}) \leq (1/\tau)\mathfrak{R}_n(\mathcal{F})$ (by Talagrand's contraction principal).

3. Since $\ell_\tau \geq \ell_0$, we have $R_\tau \geq R_0$ and $\hat{R}_\tau \geq \hat{R}_0$.

4. Also, by the Assumption 1, we have

$$\sup_{f \in \mathcal{F}} R_\tau(f) - R_0(f) \leq \sup_{f \in \mathcal{F}} \mathbb{P}(0 \leq yf(x) \leq \tau) \leq C\tau$$

   for small τ. This allows us to write

$$R_\tau(f^*_\tau) \leq R_\tau(f^*_0) \leq R_0(f^*_0) + C\tau.$$

Combining the above observations, we can write with probability $1 - \delta$

$$R_0(\hat{f}_\tau) - R_0(f^*_0) \leq C\tau + \frac{4}{\tau}\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Notice the trade-off on τ in the above bound. As we will see in the next result, we typically have $\mathfrak{R}_n(\mathcal{F}) = \mathcal{O}(1/\sqrt{n})$. Therefore the above is of order

$$R_0(\hat{f}_\tau) - R_0(f^*_0) \lesssim \tau + \frac{1}{\tau\sqrt{n}},$$

which yields a rate of $1/n^{1/4}$ after optimizing the bound over τ. Notice the sharp drop in the convergence rate, from $1/n^{1/2}$ to $1/n^{1/4}$, which is due to using a surrogate loss.

7.2 Rademacher complexity of constrained linear models

So far, we have shown that the generalization bounds can be written in terms of $\mathfrak{R}_n(\mathcal{F})$. In the following, we will show that $\mathfrak{R}_n(\mathcal{F})$ decays with n which completes the picture in terms of achieving a generalization bound.

Theorem 29 (**Rademacher Complexity of linear models**). *Define the function class of ball constrained linear models as* $\mathcal{F} = \{f(x) = \langle x, \theta\rangle, \ \|\theta\| \leq r\}$. *We have*

1. $\hat{\mathfrak{R}}_n(\mathcal{F}) \le \frac{r}{n}\sqrt{\sum_{i=1}^n \|x_i\|^2}$
2. *If* $\mathbb{E}[\|x_i\|^2] \le \kappa^2$, *then* $\mathfrak{R}_n(\mathcal{F}) \le \frac{r\kappa}{\sqrt{n}}$.

Remark. The above bound tells us that the Rademacher complexity of decays with a rate $1/\sqrt{n}$. Plugging this back in the bound (7.1), we can achieve generalization. For example, using this for linear SVMs, we obtain a generalization bound of $\mathcal{O}(1/\sqrt{n})$.

Proof. We first prove the first result. We write

$$\begin{aligned}
\widehat{\mathfrak{R}}_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i \sigma_i f(x_i)\Big| x_{1:n}\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i \sigma_i \langle\theta, x_i\rangle\Big| x_{1:n}\right] \\
&= \mathbb{E}\left[\sup_{\|\theta\|\le r} \langle\theta, \frac{1}{n}\sum_i \sigma_i x_i\rangle\Big| x_{1:n}\right], \\
&\underset{(i)}{=} r\cdot\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_i \sigma_i x_i\Big\|\Big| x_{1:n}\Big] \\
&\underset{(ii)}{\le} r\cdot\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_i \sigma_i x_i\Big\|^2\Big| x_{1:n}\Big]^{1/2} \\
&= \frac{r}{n}\mathbb{E}\left[\sum_i \sigma_i^2\|x_i\|^2 + \sum_{i\ne j}\sigma_i\sigma_j\langle x_i, x_j\rangle | x_{1:n}\right]^{1/2}, \\
&= \frac{r}{n}\left(\sum_i \|x_i\|^2\right)^{1/2}
\end{aligned}$$

where step (i) follows from the dual formulation of ℓ_2-norm, i.e., $\sup_{\|\theta\|=1} \langle\theta, u\rangle = \|u\|$, step (ii) follows from Jensen's inequality.

For the second part, we write

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}\big[\widehat{\mathfrak{R}}_n(\mathcal{F})\big] &\le \mathbb{E}\left[\frac{r}{n}\sqrt{\sum_i \|x_i\|^2}\right] \\
&\le \frac{r}{n}\sqrt{\sum_i \mathbb{E}[\|x_i\|^2]} \\
&\le \frac{r}{\sqrt{n}}\kappa,
\end{aligned}$$

where the second inequality follows from Jensen's inequality, and the last one follows from the assumption $\mathbb{E}[\|x_i\|^2] \le \kappa^2$. □

We should remark that κ is typically of order $\sqrt{d}$, so the generalization bound we get is like $\sqrt{d/n}$ as expected.

7.3 Massart's Finite Lemma

We have already worked out the generalization performance of finite function classes in Section 4.2. But in this section, we would like to use our new tool, the Rademacher complexity for the same

purpose. This will allow us to compare bounds obtained through different techniques. The following result is very useful in that respect.

We introduce the Massart's Lemma that will be used throughout next few lectures.

Lemma 30 (Massart's Finite Lemma). *Suppose that $\mathcal{F}$ satisfies $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(z_i)^2 \le \kappa^2$, then the empirical Rademacher complexity of the function class is bounded, i.e.*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \le \kappa\sqrt{\frac{2\log|\mathcal{F}|}{n}}$$

Remark.

1. The above bound is only useful (for now) when $|\mathcal{F}| < \infty$.
2. When the loss is bounded by B, the condition above is immediately satisfied for $\kappa = B$.
3. Plugging this into (7.1), we get a generalization bound with probability at least $1-\delta$
   - by Rademacher Complexity: $4B\sqrt{\frac{2\log(|\mathcal{F}|)}{n}} + B\sqrt{\frac{2\log(2/\delta)}{n}}$
   - by union bound: $B\sqrt{\frac{2\log(|\mathcal{F}|)}{n} + \frac{2\log(1/\delta)}{n}}$.

   Although these two bounds have the same rate of convergence, we notice that the latter bound is slightly tighter.
4. Perhaps the most important observation we can make is that, the function class $\mathcal{F}$ enters the above bound only through function evaluations over the data points $z_{1:n}$. This observation will be crucial in the next section.

Proof. Note that throughout this proof, we denote data with $z_{1:n} = \{z_1, \dots, z_n\}$. We will first obtain a bound for $\exp\{t\cdot\hat{\mathfrak{R}}_n(\mathcal{F})\}$, and convert this to a bound on $\hat{\mathfrak{R}}_n(\mathcal{F})$.

$$\begin{aligned}
\exp\{t\cdot\hat{\mathfrak{R}}_n(\mathcal{F})\} &= \exp\left\{t\cdot\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(z_i)\Big|z_{1:n}\right]\right\} \qquad (7.2)\\
&\le \mathbb{E}\left[\exp\left\{t\cdot\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(z_i)\right\}\Big|z_{1:n}\right] \quad \text{(by Jensens's inequality)}\\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\exp\left\{t\cdot\frac{1}{n}\sum_{i=1}^n\sigma_i f(z_i)\right\}\Big|z_{1:n}\right] \quad \text{(sup on a monotone transformation)}\\
&\underset{(*)}{\le} \sum_{f\in\mathcal{F}}\mathbb{E}\left[\exp\left\{t\cdot\frac{1}{n}\sum_{i=1}^n\sigma_i f(z_i)\right\}\Big|z_{1:n}\right] \quad (\mathcal{F}\text{ is finite and exp() is positive})\\
&= \sum_{f\in\mathcal{F}}\prod_{i=1}^n M\Big(\tfrac{t}{n}f(z_i)\Big) \qquad M_\sigma(t)\text{ is the MGF of }\sigma\text{, i.e.,}\\
M_\sigma(t) &= \mathbb{E}\big[\exp\{t\sigma\}\big|z_{1:n}\big] = \cosh(t).
\end{aligned}$$

In the above derivation, in step $(*)$, we replaced sup over a set with a summation over that set. It is important to pay attention to this step as in the next section, we will obtain a general bound by simply tightening this inequality.

We proceed by noticing that $x^2/2 \geq \log\cosh(x)$ (check this!) which implies $\exp\{x^2/2\} \geq \cosh(x)$. Therefore, we can write

$$\begin{aligned}
\sum_{f\in\mathcal{F}}\prod_{i=1}^{n} M\Big(\tfrac{t}{n}f(z_i)\Big) &\leq \sum_{f\in\mathcal{F}}\prod_{i=1}^{n}\exp\left\{\frac{t^2}{2n^2}f(z_i)^2\right\} \\
&= \sum_{f\in\mathcal{F}}\exp\left\{\frac{t^2}{2n}\underbrace{\frac{1}{n}\sum_{i=1}^{n}f(z_i)^2}_{\leq\kappa^2}\right\} \\
&\leq |\mathcal{F}|\exp\left\{\frac{t^2}{2n}\kappa^2\right\}.
\end{aligned}$$

The final bound we obtained can be written as

$$\begin{aligned}
\exp\left\{t\cdot\hat{\mathfrak{R}}_n(\mathcal{F})\right\} &\leq |\mathcal{F}|\exp\left\{\frac{t^2\kappa^2}{2n}\right\} \\
\implies \hat{\mathfrak{R}}_n(\mathcal{F}) &\leq \frac{\log|\mathcal{F}|}{t} + \frac{t\kappa^2}{2n}
\end{aligned}$$

which holds for all $t \geq 0$. By optimizing over t, we will obtain the final result. That is, differentiating the RHS above with respect to t and solving for the optimal value gives $2\kappa\sqrt{\log|\mathcal{F}|/2n}$. $\square$