Machine Learning Theory (CSC 482A/581A) - Lectures 13–15

Nishant Mehta

1 The bounded differences inequality

We previously have seen how to derive high probability upper bounds on the excess risk for VC classes, in either the realizable case and the more general agnostic setting. As we will see in this lecture and the next, the bound in the agnostic setting can be recovered using a different notion of the complexity called *Rademacher complexity*. In addition, bounds based on Rademacher complexity readily generalize other loss functions such as squared loss, thereby providing risk bounds for regression as well.

Before deriving a risk bound, we first need a powerful concentration inequality known as the *bounded differences inequality* and often called McDiarmid's inequality (McDiarmid, 1989). This inequality holds for functions which satisfy a bounded differences condition.

Definition 1. We say that a function $g \colon \mathcal{Z}^n \to \mathbb{R}$ satisfies the *bounded differences condition* with constants $c_1, \ldots, c_n$ if, for all $i \in [n]$,

$$\sup_{z_1, \ldots, z_n, z_i' \in \mathcal{Z}} |g(z_1, \ldots, z_n) - g(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c_i.$$

Throughout this lecture, $\mathbf{Z}_1^n$ will denote $(Z_1, \ldots, Z_n)$.

Theorem 1. *Let $Z_1, \ldots, Z_n$ be independent random variables taking values in $\mathcal{Z}$. If a function $g \colon \mathcal{Z}^n \to \mathbb{R}$ satisfies the bounded differences condition with constants $c_1, \ldots, c_n$, then for any $t > 0$,*

$$\Pr\big(g(\mathbf{Z}_1^n) - \mathsf{E}[g(\mathbf{Z}_1^n)] \geq t\big) \leq \exp\left(-\frac{2t^2}{\sum_{j=1}^n c_j^2}\right).$$

Remarks. By applying the above result with $-g$ instead of g, it also holds that

$$\Pr\big(g(\mathbf{Z}_1^n) - \mathsf{E}[g(\mathbf{Z}_1^n)] \leq -t\big) \leq \exp\left(-\frac{2t^2}{\sum_{j=1}^n c_j^2}\right).$$

Therefore, we also have the two-sided inequality

$$\Pr\Big(\left|g(\mathbf{Z}_1^n) - \mathsf{E}\left[g(\mathbf{Z}_1^n)\right]\right| \geq t\Big) \leq 2\exp\left(-\frac{2t^2}{\sum_{j=1}^n c_j^2}\right).$$

2 Excess risk bound based on Rademacher complexity

Since uniform convergence is sufficient for learning (and, in particular, sufficient for learning using ERM), we begin with the familiar random quantity

$$\sup_{f \in \mathcal{F}} \left\{ \mathsf{E}\left[\ell_f(Z)\right] - \frac{1}{n}\sum_{j=1}^n \ell_f(Z_j) \right\}, \tag{1}$$

where $Z = (X, Y)$ (likewise $Z_j = (X_j, Y_j)$) and $\ell_f(Z)$ is the loss suffered by hypothesis f under outcome Z. In the case of classification, the loss function is zero-one loss, and so $\ell_f(Z) = \mathbf{1}\left[f(X) \neq Y\right]$. Our goal is to obtain a high probability upper bound on (1), and we do this in 2 steps.

Step 1: Relating (1) to its expectation

Let us assume that the loss function is bounded, so that we always have $\ell_f(z) \in [0, b]$ for some constant $b > 0$. Then, letting $g(\mathbf{Z}_1^n) = \sup_{f\in\mathcal{F}} \left\{ \mathsf{E}\left[\ell_f(Z)\right] - \frac{1}{n}\sum_{j=1}^n \ell_f(Z_j) \right\}$, it is simple to verify that g satisfies the conditions of Theorem 1 with $c_j = b$ for all $j \in [n]$. Indeed, since for any $i \in [n]$,

$$\begin{aligned} g(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n) &= \sup_{f\in\mathcal{F}} \left\{ \mathsf{E}\left[\ell_f(Z)\right] - \frac{1}{n}\sum_{j=1}^n \ell_f(z_j) + \frac{1}{n}\left(\ell_f(z_i) - \ell_f(z_i')\right) \right\} \\ &\leq \sup_{f\in\mathcal{F}} \left\{ \mathsf{E}\left[\ell_f(Z)\right] - \frac{1}{n}\sum_{j=1}^n \ell_f(z_j) \right\} + \frac{1}{n}\sup_{f\in\mathcal{F}}\{\ell_f(z_i) - \ell_f(z_i')\} \\ &\leq g(z_1, \ldots, z_n) + b. \end{aligned}$$

Applying Theorem 1 (with inversion), it holds that with probability at least $1 - \delta$,

$$\sup_{f\in\mathcal{F}} \left\{ \mathsf{E}\left[\ell_f(Z)\right] - \frac{1}{n}\sum_{j=1}^n \ell_f(Z_j) \right\} \leq \mathsf{E}\left[\sup_{f\in\mathcal{F}} \left\{ \mathsf{E}\left[\ell_f(Z)\right] - \frac{1}{n}\sum_{j=1}^n \ell_f(Z_j) \right\} \right] + b\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Step 2: Symmetrization

We now perform symmetrization by ghost sample and symmetrization by random signs in succession. To this end, let $Z_1', \ldots, Z_n'$ be an independent copy of $Z_1, \ldots, Z_n$. Then

$$\mathsf{E}\left[\sup_{f\in\mathcal{F}} \left\{ \mathsf{E}\left[\ell_f(Z)\right] - \frac{1}{n}\sum_{j=1}^n \ell_f(Z_j) \right\} \right] = \mathsf{E}\left[\sup_{f\in\mathcal{F}} \left\{ \mathsf{E}\left[\frac{1}{n}\sum_{j=1}^n \ell_f(Z_j') \right] - \frac{1}{n}\sum_{j=1}^n \ell_f(Z_j) \right\} \right].$$

Next, since the supremum of an expectation is at most the expectation of the supremum, the above is at most

$$\mathsf{E}\left[\sup_{f\in\mathcal{F}} \left\{ \frac{1}{n}\sum_{j=1}^n \Big(\ell_f(Z_j') - \ell_f(Z_j) \Big) \right\} \right].$$

Let $\sigma_1, \ldots, \sigma_1$ be independent Rademacher random variables (i.e. taking values -1 and $+1$ with equal probability $\frac{1}{2}$). Since $Z_1, \ldots, Z_n, Z_1', \ldots, Z_n'$ are i.i.d., the above quantity is equal to

$$\mathsf{E}\left[\sup_{f\in\mathcal{F}} \left\{ \frac{1}{n}\sum_{j=1}^n \sigma_j \Big(\ell_f(Z_j') - \ell_f(Z_j) \Big) \right\} \right],$$

which is at most

$$\mathsf{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{j=1}^n \sigma_j \ell_f(Z_j') \right] + \mathsf{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{j=1}^n (-\sigma_j) \ell_f(Z_j) \right] = 2\,\mathsf{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{j=1}^n \sigma_j \ell_f(Z_j) \right].$$

Defining the class $\mathcal{L}_\mathcal{F} := \{\ell_f : f \in \mathcal{F}\}$, we have just obtained an upper bound in term of the Rademacher complexity of $\mathcal{L}_\mathcal{F}$.

Definition 2. Let $\mathcal{G}$ be a class of functions mapping from $\mathcal{Z}$ to $\mathbb{R}$. The *empirical Rademacher complexity* $\mathcal{G}$ (with respect to sample $\mathbf{Z}_1^n$) is defined as

$$\widehat{\mathcal{R}}_n(\mathcal{G}) = \mathsf{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{j=1}^{n} \sigma_j g(Z_j) \,\middle|\, \mathbf{Z}_1^n\right].$$

The *Rademacher complexity of* $\mathcal{G}$ (with respect to probability distribution P) is defined as

$$\mathcal{R}_n(\mathcal{G}) = \mathsf{E}\left[\widehat{\mathcal{R}}_n(\mathcal{G})\right].$$

We will make a few remarks about Rademacher complexity shortly; but first, let's formulate the conclusion of Steps 1 and 2.

Theorem 2. *With probability at least* $1-\delta$*,*

$$\sup_{f\in\mathcal{F}}\left\{\mathsf{E}\left[\ell_f(Z)\right] - \frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j)\right\} \le 2\mathcal{R}_n(\mathcal{L}_{\mathcal{F}}) + b\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

The following excess risk bound is almost immediate.

Corollary 1. *Let* $\hat{f}$ *be any ERM estimator and let* f^* *be the risk minimizer over class* $\mathcal{F}$*. If* $\ell_f(Z)$ *takes values in* $[0,b]$ *for all* $f\in\mathcal{F}$*, then with probability at least* $1-\delta$*,*

$$R(\hat{f}) \le R(f^*) + 2\mathcal{R}_n(\mathcal{L}_{\mathcal{F}}) + b\sqrt{\frac{2\log\frac{2}{\delta}}{n}}.$$

Proof. First,

$$\begin{aligned} R(\hat{f}) - R(f^*) &= \left(\hat{R}(\hat{f}) - \hat{R}(f^*)\right) + \left(R(\hat{f}) - \hat{R}(\hat{f})\right) + \left(\hat{R}(f^*) - R(f^*)\right) \\ &\le \left(R(\hat{f}) - \hat{R}(\hat{f})\right) + \left(\hat{R}(f^*) - R(f^*)\right). \end{aligned}$$

From Theorem 2, with probability at least $1-\delta/2$,

$$R(\hat{f}) - \hat{R}(\hat{f}) \le 2\mathcal{R}_n(\mathcal{L}_{\mathcal{F}}) + b\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Also, from either the bounded differences inequality (Theorem 1) or Hoeffding's inequality, with probability at least $1-\delta/2$,

$$\hat{R}(f^*) - R(f^*) \le b\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

□

If $\mathcal{F}$ is a set of classifiers and the loss function is the zero-one loss, then the Rademacher complexity of $\mathcal{L}_{\mathcal{F}}$ can be expressed in terms of the Rademacher complexity of the class $\mathcal{F}$ itself.

Let $\mathcal{Y} = \{-1, 1\}$. Then

$$\begin{aligned}
\widehat{\mathcal{R}}_n(\mathcal{L}_\mathcal{F}) &= \mathsf{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j \, \mathbf{1}\left[f(X_j) \neq Y_j \right] \right] \\
&= \mathsf{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j \frac{1 - Y_j f(X_j)}{2} \right] \\
&= \frac{1}{2} \mathsf{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j Y_j f(X_j) \right] \\
&= \frac{1}{2} \mathsf{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j f(X_j) \right] \\
&= \frac{1}{2} \widehat{\mathcal{R}}_n(\mathcal{F}),
\end{aligned}$$

where the third equality follows since each Rademacher random variable has zero mean, and the fourth equality follows because, conditional on $\mathbf{Z}_1^n$, the distributions of $\sigma_j Y_j$ and σ_j are identical. It therefore also holds that $\mathcal{R}_n(\mathcal{L}_\mathcal{F}) = \frac{1}{2} \mathcal{R}_n(\mathcal{F})$.

Using this relation, we now have an excess risk bound for sets of classifiers that depends only on the Rademacher complexity of $\mathcal{F}$:

Theorem 3. *Let $\mathcal{F}$ be a set of classifiers, let $\hat{f}$ be any ERM estimator over $\mathcal{F}$, and let f^* be the risk minimizer over class $\mathcal{F}$. Then with probability at least $1 - \delta$,*

$$R(\hat{f}) \leq R(f^*) + \mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}.$$

Remarks. As above, let $\mathcal{Y} = \{-1, 1\}$. Consider $\widehat{R}_n(\mathcal{F})$, the empirical Rademacher complexity of a set of classifiers $\mathcal{F}$ with respect to a fixed sequence of inputs $\mathbf{X}_1^n = (X_1, \ldots, X_n)$. Letting $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ and $f(\mathbf{X}_1^n) = (f(X_1), \ldots f(X_n))$, this may be rewritten as

$$\frac{1}{n} \mathsf{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \langle f(\mathbf{X}_1^n), \boldsymbol{\sigma} \rangle \right].$$

Thus, the empirical Rademacher complexity of $\mathcal{F}$ measures the ability of functions from $\mathcal{F}$ to fit random noise (specifically, random sign noise, or random labels). Now, if for every sign vector $\boldsymbol{\sigma} \in \{-1, 1\}^n$ there is a function $f \in \mathcal{F}$ which labels as 1 precisely those examples X_j for which $\sigma_j = 1$ (and labels the other examples -1), then $\mathcal{F}$ shatters $\mathbf{X}_1^n$ and it holds that $\widehat{\mathcal{R}}_n(\mathcal{F}) = 1$. The class $\mathcal{F}$ clearly is a rich class compared to the sample size, and this exactly corresponds to it being a class of maximum Rademacher complexity (at this sample size); as a result, Theorem 3 provides only a trivial excess risk bound, as we should expect.

3 Bounding Rademacher complexity

Massart's finite class lemma (to leverage the growth function).

Lemma 1. *Let A be a finite subset of $\mathbb{R}^n$, with $r = \max_{a \in A} \|a\|_2$. Let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher random variables. Then*

$$\mathsf{E}\left[\max_{a \in A} \sum_{j=1}^{n} \sigma_j a_j\right] \le r\sqrt{2 \log |A|}.$$

Proof. First, from Jensen's inequality and the convexity of $x \mapsto \exp(x)$, it holds that

$$\exp\left(\lambda\, \mathsf{E}\left[\max_{a \in A} \sum_{j=1}^{n} \sigma_j a_j\right]\right) \le \mathsf{E}\left[\exp\left(\lambda \max_{a \in A} \sum_{j=1}^{n} \sigma_j a_j\right)\right] = \mathsf{E}\left[\max_{a \in A} \exp\left(\lambda \sum_{j=1}^{n} \sigma_j a_j\right)\right].$$

Since the max is bounded by the sum, this is at most

$$\sum_{a \in A} \mathsf{E}\left[\exp\left(\lambda \sum_{j=1}^{n} \sigma_j a_j\right)\right] = \sum_{a \in A} \prod_{j=1}^{n} \mathsf{E}\left[e^{\lambda \sigma_j a_j}\right] = \sum_{a \in A} \prod_{j=1}^{n} \left(\frac{e^{-\lambda a_j} + e^{\lambda a_j}}{2}\right). \tag{2}$$

Next, we use the inequality[1] $\frac{e^{-x}+e^{x}}{2} \le e^{x^2/2}$, which follows by taking a Taylor expansion:

$$\begin{aligned}
\frac{1}{2}\left(e^{-x} + e^{x}\right) &= \frac{1}{2}\left(1 - x + \frac{x^2}{2} - \frac{x^3}{3!} + \frac{x^4}{4!} \ldots\right) + \frac{1}{2}\left(1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots\right) \\
&= 1 + \frac{x^2}{2} + \frac{x^4}{4!} + \frac{x^6}{6!} \ldots \\
&\le 1 + \frac{x^2}{2} + \frac{(x^2)^2}{2^2 \cdot 2} + \frac{(x^2)^3}{2^3 \cdot 3!} \ldots \\
&= e^{x^2/2}.
\end{aligned}$$

Therefore, (2) is at most

$$\sum_{a \in A} \exp\left(\sum_{j=1}^{n} \lambda^2 a_j^2 / 2\right) \le \sum_{a \in A} e^{\lambda^2 r^2/2} \le |A| e^{\lambda^2 r^2/2}.$$

Putting together the sequence of inequalities, taking the log, and dividing by λ yields

$$\mathsf{E}\left[\max_{a \in A} \sum_{j=1}^{n} \sigma_j a_j\right] \le \frac{\log |A|}{\lambda} + \frac{\lambda r^2}{2}.$$

It remains to tune λ, which can be done by finding a value of λ such that the derivative is equal to zero. Setting $\lambda = \sqrt{\frac{2 \log |A|}{r^2}}$ yields the result. □

[1] This could be referred to as the hyperbolic cosine trick, since $\frac{1}{2}\left(e^{-x} + e^{x}\right) = \cosh(x)$.

Recovering excess risk bounds for VC classes. With Massart's finite class lemma in hand, it is now simple to obtain an excess risk bound for VC classes. Let $\mathcal{F}$ be a class whose VC dimension is V. Starting from Theorem 3, it remains to bound the Rademacher complexity of $\mathcal{F}$. Observe that

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) = \mathsf{E}\left[\mathsf{E}\left[\widehat{\mathcal{R}}_n(\mathcal{F}) \mid \mathbf{X}_1^n\right]\right] &= \mathsf{E}\left[\mathsf{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{j=1}^n \sigma_j f(X_j) \,\middle|\, \mathbf{X}_1^n\right]\right] \\
&= \frac{1}{n}\mathsf{E}\left[\mathsf{E}\left[\sup_{v\in\mathcal{F}(\mathbf{X}_1^n)} \sum_{j=1}^n \sigma_j v_j \,\middle|\, \mathbf{X}_1^n\right]\right] \\
&\le \sqrt{\frac{2\log \Pi_{\mathcal{F}}(n)}{n}} && \text{(Lemma 1)} \\
&\le \sqrt{\frac{2V\log\frac{en}{V}}{n}}, && \text{(Sauer's Lemma)},
\end{aligned}$$

where the first inequality bounded the conditional expectation via Massart's finite class lemma (Lemma 1) with $r = \sqrt{n}$.

By combining the above result with Theorem 3, we have the following theorem.

Theorem 4. *Let $\mathcal{F}$ be a VC class of VC dimension V. Let $\hat{f}$ be any ERM estimator and let f^* be the risk minimizer over $\mathcal{F}$. Then with probability at least $1-\delta$,*

$$R(\hat{f}) \le R(f^*) + 2\sqrt{\frac{2V\log\frac{en}{V}}{n}} + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}.$$

The above result is equivalent to the excess risk bound we previously derived for VC classes. Using a technique called chaining, which sadly will not be covered in this course, one can actually obtain an improved bound on the Rademacher complexity of VC classes which in turn yields an improved excess risk bound:

Theorem 5. *Let $\mathcal{F}$ be a VC class of VC dimension V. Then*

$$\mathcal{R}_n(\mathcal{F}) = O\left(\sqrt{\frac{V}{n}}\right).$$

Note that the logarithmic factor present in our earlier bound has now been eliminated. The above bound thus serves as a real improvement and yields a strictly better excess risk bound than what we had derived a few lectures ago. The bound provided by Theorem 5 is based on a much more intricate analysis; the techniques used are outside of the scope of this course.

3.1 A computable upper bound

Remarkably, it is possible to obtain a fully computable upper bound based only the empirical Rademacher complexity of a class. This result follows by observing that, under our bounded loss assumption, $\widehat{\mathcal{R}}_n(\mathcal{L}_{\mathcal{F}})$ satisfies the bounded diffferences property with $c_j = \frac{b}{n}$ for $j \in [n]$. Therefore, with probability at least $1-\delta$, it holds that

$$\mathcal{R}_n(\mathcal{L}_{\mathcal{F}}) \le \widehat{\mathcal{R}}_n(\mathcal{L}_{\mathcal{F}}) + b\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Also, *precisely the same argument can be applied* if we now draw a sample $\sigma_1, \dots, \sigma_n$ of independent Rademacher random variables and avoid taking the expectation over $\sigma_1, \dots, \sigma_n$; that is, with probability at least $1 - \delta$ (now also over the draw of $\sigma_1, \dots, \sigma_n$), it holds that

$$\mathcal{R}_n(\mathcal{L}_\mathcal{F}) \le \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j \ell_f(Z_j) + b\sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

By applying Theorem 2 and the above inequality, both with their respective δ set to $\delta/2$, we have the following fully observable and in-principle-computable risk bound.

Theorem 6. *Let $\hat{f}$ be any estimator. Assume that $\ell_f(Z)$ takes values in $[0, b]$ for all $f \in \mathcal{F}$. Then with probability at least $1 - \delta$,*

$$\mathsf{E}\left[\ell_f(Z)\right] \le \frac{1}{n} \sum_{j=1}^n \ell_f(Z_j) + 2\widehat{\mathcal{R}}_n(\mathcal{L}_\mathcal{F}) + b\sqrt{\frac{2 \log \frac{2}{\delta}}{n}}.$$

In addition, if $\sigma_1, \dots, \sigma_n$ are independent Rademacher random variables, then with probability at least $1 - \delta$ over both the training sample and $\sigma_1, \dots, \sigma_n$,

$$\mathsf{E}\left[\ell_f(Z)\right] \le \frac{1}{n} \sum_{j=1}^n \ell_f(Z_j) + 2 \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j \ell_f(Z_j) + b\sqrt{\frac{2 \log \frac{2}{\delta}}{n}}.$$

Applying our arguments from above in the case of classification, the first result holds in the case of classification with $2\widehat{\mathcal{R}}_n(\mathcal{L}_\mathcal{F})$ replaced by $\widehat{\mathcal{R}}_n(\mathcal{F})$ and $b = 1$ (and with the analogous modification for the second result).

3.2 Estimating Rademacher complexity from data

Consider the classification variant of the second part of Theorem 6, so that we have with probability at least $1 - \delta$,

$$\mathsf{E}\left[\mathbf{1}\left[\hat{f}(X) \ne Y\right]\right] \le \frac{1}{n} \sum_{j=1}^n \mathbf{1}\left[\hat{f}(X_j) \ne Y_j\right] + \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j f(X_j) + \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}.$$

If we can efficiently compute (or at least approximate) the supremum, then we can precisely quantify (at least with high probability) the gap between the risk of ERM and the risk of f^*.

This is indeed possible, at least if we can efficiently compute ERM itself. To see this, observe that

$$\begin{aligned}
\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j f(X_j) &= 2 \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \left(\frac{1}{2} - \frac{1 - \sigma_j f(X_j)}{2}\right) \\
&= 1 + 2 \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n -\frac{1 - \sigma_j f(X_j)}{2} \\
&= 1 - 2 \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \frac{1 - \sigma_j f(X_j)}{2} \\
&= 1 - 2 \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \mathbf{1}\left[f(X_j) \ne \sigma_j\right].
\end{aligned}$$

Computing the infimum is of course equivalent to computing the empirical risk of ERM with the random signs taking the role of the labels.

4 Properties of Rademacher complexity

We now study a few properties of Rademacher complexity which will be useful when we analyze in the coming lectures, including our analysis of support vector machines.

Property 1 (Affine transformations)**.** For a set A and $c, b \in \mathbb{R}$, define $cA + b$ as the set $\{ca + b : a \in A\}$. Then

$$\widehat{\mathcal{R}}_n(c\mathcal{F} + b) = a\widehat{\mathcal{R}}_n(\mathcal{F}).$$

The proof is simple and is left as an exercise.

Property 2 (Convex hull)**.** Let $\operatorname{conv}(\mathcal{F})$ be the convex hull of $\mathcal{F}$, the set of all convex combinations of functions from $\mathcal{F}$. Then

$$\widehat{\mathcal{R}}_n(\operatorname{conv}(\mathcal{F})) = \widehat{\mathcal{R}}_n(\mathcal{F}).$$

For simplicity, we prove this for the case of finite $\mathcal{F}$ with $|\mathcal{F}| = N$. Denote by Δ the $(N-1)$-simplex over N outcomes, defined as $\{\alpha \in \mathbb{R}^N_+ : \sum_{i=1}^N \alpha_i = 1\}$. Observe that

$$\begin{aligned}
\widehat{\mathcal{R}}_n(\operatorname{conv}(\mathcal{F})) &= \mathsf{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \operatorname{conv}(\mathcal{F})} \frac{1}{n}\sum_{j=1}^n \sigma_j f(X_j)\right] \\
&= \mathsf{E}_{\boldsymbol{\sigma}}\left[\sup_{\alpha \in \Delta} \frac{1}{n}\sum_{j=1}^n \sigma_j \sum_{i=1}^N f_i(X_j)\right].
\end{aligned}$$

Now, for any $\boldsymbol{\sigma}$,

$$\begin{aligned}
\sup_{\alpha \in \Delta} \frac{1}{n}\sum_{j=1}^n \sigma_j \cdot \left(\sum_{i=1}^N \alpha_i f_i(X_j)\right) &= \sup_{\alpha \in \Delta} \sum_{i=1}^N \alpha_i \cdot \left(\frac{1}{n}\sum_{j=1}^n \sigma_j f_i(X_j)\right) \\
&= \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{j=1}^n \sigma_j f(X_j) \\
&= \widehat{\mathcal{R}}_n(\mathcal{F}).
\end{aligned}$$

Property 3 (Sums)**.** For any classes $\mathcal{F}$ and $\mathcal{G}$,

$$\widehat{\mathcal{R}}_n(\mathcal{F} + \mathcal{G}) \le \widehat{\mathcal{R}}_n(\mathcal{F}) + \widehat{\mathcal{R}}_n(\mathcal{G}).$$

The proof is straightforward.

Property 4 (Composition with Lipschitz functions)**.** Let $\phi_1, \ldots, \phi_n$ be a functions from $\mathbb{R}$ to $\mathbb{R}$ with respective Lipschitz constants $L_1, \ldots, L_n$.

$$\mathsf{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{j=1}^n \sigma_j \phi_j(f(X_j))\right] \le \mathsf{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{j=1}^n \sigma_j L_j f(X_j)\right].$$

We typically will be interested in the example $\phi_j(f(x_j)) = \ell(y_j, f(x_j))$, which is L-Lipschitz whenever ℓ is L-Lipschitz in its second argument; note that ϕ_j carries the information about label y_j. This example is so important that we provide it its own property:

Property 5 (Lipschitz losses)**.** If $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is L-Lipschitz in its second argument, then

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

Example 1 (Squared loss)**.** Let ℓ be squared loss, so that $\ell(y, f(x)) = (y - f(x))^2$. Assume that x and y satisfy $|y| \leq C$ and $|f(x)| \leq C$ for all $f \in \mathcal{F}$. Then for any $f, g \in \mathcal{F}$,

$$\begin{aligned} \left|(y - f(x))^2 - (y - g(x))^2\right| &= \left|f^2(x) - g^2(x) - 2y(f(x) - g(x))\right| \\ &= \left|(f(x) - g(x))\,(f(x) + g(x) - 2y)\right| \\ &\leq 4C\,|f(x) - g(x)|\,, \end{aligned}$$

and so ℓ is L-Lipschitz in its second argument with $L = 4C$.

The historical predecessor of the last two properties is a result of Ledoux and Talagrand (1991) known as the Ledoux-Talagrand contraction inequality. The scope of their result is much wider (and, in particular, it applies to a stronger notion of Rademacher complexity that takes the absolute value of the Rademacher-weighted summation), but consequently their result has a multiplicative factor of 2 on the RHS and it requires the stronger assumption that, for each $j \in [n]$, we have $\phi_j(0) = 0$. The version shown here, namely Property 4, is due to Meir and Zhang (2003).[2]

At the end of these lectures notes is a proof of Property 4; the proof is taken from Meir and Zhang (2003) but is less terse than the version appearing in that paper.

5 Bounding Rademacher complexity for linear predictors

In light of Example 1, for linear regression with squared loss and bounded data and bounded predictions, we almost have an explicit risk bound. The only remaining step is to bound the (empirical) Rademacher complexity of the class of predictors.

Lemma 2. *Let $\mathcal{F}$ be a class of linear predictors, defined as $\mathcal{F} = \{f_w : w \in \mathbb{R}^d, \|w\|_2 \leq R\}$, with $f_w(x) = \langle w, x\rangle$. Let $x_1, \ldots, x_n \in \mathbb{R}^d$, and define $R := \max_{j \in [n]} \|x_j\|_2$. Then*

$$\widehat{\mathcal{R}}_n(\mathcal{F}) \leq \frac{BR}{\sqrt{n}}.$$

[2] Their result seems to be less well-known in the machine learning community and has been re-invented by a few authors even 10 years after it was originally proved!

Proof.

$$\begin{aligned}
\widehat{\mathcal{R}}_n(\mathcal{F}) &= \mathsf{E}_{\boldsymbol{\sigma}} \sup_{f_w \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^{n} \sigma_j f_w(x_j) \\
&= \mathsf{E}_{\boldsymbol{\sigma}} \sup_{w:\|w\|_2 \le B} \frac{1}{n} \sum_{j=1}^{n} \sigma_j \langle w, x_j \rangle \\
&= \frac{1}{n}\, \mathsf{E}_{\boldsymbol{\sigma}} \sup_{w:\|w\|_2 \le B} \left\langle w, \sum_{j=1}^{n} \sigma_j x_j \right\rangle \\
&= \frac{B}{n}\, \mathsf{E}_{\boldsymbol{\sigma}} \left\| \sum_{j=1}^{n} \sigma_j x_j \right\|_2 && \text{(take w in the direction $\sum_{j=1}^{n} \sigma_j x_j$)} \\
&\le \frac{B}{n} \left(\mathsf{E}_{\boldsymbol{\sigma}} \left\| \sum_{j=1}^{n} \sigma_j x_j \right\|_2^2 \right)^{1/2} && \text{(Jensen's inequality + concavity of $x \mapsto \sqrt{x}$)} \\
&\le \frac{B}{n} \left(\mathsf{E}_{\boldsymbol{\sigma}} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_i \sigma_j \langle x_i, x_j \rangle \right)^{1/2} \\
&\le \frac{B}{n} \left(\sum_{j=1}^{n} \|x_j\|_2^2 \right)^{1/2} && \text{($\sigma_j^2 = 1$ and independence of σ_j and σ_i for $i \ne j$)} \\
&\le \frac{BR}{\sqrt{n}}.
\end{aligned}$$

□

Proof (of Property 4). For legibility, let $\Phi_j(f)$ denote $\phi_j(f(X_j))$ and let $\Psi_j(f)$ denote $f(X_j)$. We claim that for any function $c\colon \mathcal{F} \to \mathbb{R}$,

$$\mathsf{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}}\left\{c(f)+\sum_{j=1}^{n}\sigma_j\Phi_j(f)\right\}\right] \le \mathsf{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}}\left\{c(f)+\sum_{j=1}^{n}\sigma_j L_j\Psi_j(f)\right\}\right].$$

The proof is by induction. The result clearly holds for $k=0$. Now, assume that the result holds for $k=n-1$. We prove that the result holds for n. Observe that

$$\begin{aligned}
&\mathsf{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}}\left\{c(f)+\sum_{j=1}^{n}\sigma_j\Phi_j(f)\right\}\right]\\
&=\frac{1}{2}\,\mathsf{E}_{\sigma_1,\ldots,\sigma_{n-1}}\left[\sup_{f_1\in\mathcal{F}}\left\{c(f_1)+\sum_{j=1}^{n-1}\sigma_j\Phi_j(f_1)+\Phi_n(f_1)\right\}\right]\\
&\quad+\frac{1}{2}\,\mathsf{E}_{\sigma_1,\ldots,\sigma_{n-1}}\left[\sup_{f_2\in\mathcal{F}}\left\{c(f_2)+\sum_{j=1}^{n-1}\sigma_j\Phi_j(f_2)-\Phi_n(f_2)\right\}\right]\\
&=\mathsf{E}_{\sigma_1,\ldots,\sigma_{n-1}}\left[\sup_{f_1,f_2\in\mathcal{F}}\left\{\frac{c(f_1)+c(f_2)}{2}+\sum_{j=1}^{n-1}\sigma_j\frac{\Phi_j(f_1)+\Phi_j(f_2)}{2}+\frac{\Phi_n(f_1)-\Phi_n(f_2)}{2}\right\}\right].
\end{aligned}$$

Now, for any (f_1, f_2) pair, if $\Phi_n(f_1) < \Phi_n(f_2)$, swapping them to ensures that $\Phi_n(f_1) \ge \Phi_n(f_2)$. Thus, the above is equal to

$$\mathsf{E}_{\sigma_1,\ldots,\sigma_{n-1}}\left[\sup_{f_1,f_2\in\mathcal{F}}\left\{\frac{c(f_1)+c(f_2)}{2}+\sum_{j=1}^{n-1}\sigma_j\frac{\Phi_j(f_1)+\Phi_j(f_2)}{2}+\frac{|\Phi_n(f_1)-\Phi_n(f_2)|}{2}\right\}\right]. \tag{3}$$

Next, unpacking definitions and by the L_j-Lipschitz property of ϕ_j, we have

$$|\Phi_n(f_1)-\Phi_n(f_2)| \le L_j\,|\Psi(f_1)-\Psi(f_2)|\,,$$

and so (3) is at most

$$\mathsf{E}_{\sigma_1,\ldots,\sigma_{n-1}}\left[\sup_{f_1,f_2\in\mathcal{F}}\left\{\frac{c(f_1)+c(f_2)}{2}+\sum_{j=1}^{n-1}\sigma_j\frac{\Phi_j(f_1)+\Phi_j(f_2)}{2}+L_n\frac{|\Psi_n(f_1)-\Psi_n(f_2)|}{2}\right\}\right],$$

which, by the earlier swapping argument for (f_1, f_2), is equal to

$$\mathsf{E}_{\sigma_1,\ldots,\sigma_{n-1}}\left[\sup_{f_1,f_2\in\mathcal{F}}\left\{\frac{c(f_1)+c(f_2)}{2}+\sum_{j=1}^{n-1}\sigma_j\frac{\Phi_j(f_1)+\Phi_j(f_2)}{2}+L_n\frac{\Psi_n(f_1)-\Psi_n(f_2)}{2}\right\}\right],$$

which is equal to

$$\mathsf{E}_{\sigma_n}\left[\mathsf{E}_{\sigma_1,\ldots,\sigma_{n-1}}\left[\sup_{f\in\mathcal{F}}\left\{(c(f)+\sigma_nL_n\Psi_n(f))+\sum_{j=1}^{n-1}\sigma_j\Phi_j(f)\right\}\right]\right].$$

The result holds by the induction hypothesis. □

References

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.