Rademacher complexity and the uniform convergence theorem

1 Rademacher complexity

We have discussed the Vapnik-Chervonenkis uniform convergence theorem. The theorem bounds the generalization error of an arbitrary hypothesis in a class $\mathcal{H}$ using the empirical error, the VC dimension, and the number of i.i.d. samples. The VC dimension is a measure of a function class' "complexity" or "expressiveness." In this lecture, we define and analyze a related complexity measure for function classes called the *Rademacher complexity*. Then, we bound the generalization error using the empirical error and the Rademacher complexity. We shall see that the Vapnik-Chervonenkis uniform convergence theorem is a consequence of the bound using Rademacher complexity. The double sampling trick is applied again to prove the new bound.

Let $\mathcal{G}$ be a family of functions from some domain $\mathcal{Z}$ to an interval $[a, b]$ on the real line. (In our context, often $\mathcal{Z} = \Omega \times \{0, 1\}$ or $\mathcal{Z} = \Omega \times \{-1, 1\}$.) Let $\mathcal{D}$ be a probability distribution on $\mathcal{Z}$. The distribution $\mathcal{D}$ is implicit in most of the discussions in this lecture note and thus will be omitted whenever there is no confusion.

Definition 1.1 (Empirical Rademacher complexity)**.** Let S be a set of m points from $\mathcal{Z}$. Then, the *empirical Rademacher complexity* of $\mathcal{G}$ (given S) is defined to be

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i) \mid S = \{z_1, \ldots, z_m\}\right],$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$ is a vector of independent Rademacher variables, i.e. $\sigma_i = \pm 1$ with probability $1/2$.

Intuitively, suppose $\mathcal{G}$ is a class of binary classifiers (from Ω to $\{-1, 1\}$), then the expression

$$\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)$$

measures the performance of the best classifier in $\mathcal{G}$ with respect to the labels σ_i on z_i. Thus, overall, $\hat{\mathcal{R}}_S(\mathcal{G})$ is the average performance of the best classifier in $\mathcal{G}$ over all random labellings of the points in S. If $\mathcal{G}$ shatters S, then its empirical Rademacher complexity is 1.

Definition 1.2 (Rademacher complexity)**.** The *Rademacher complexity of* $\mathcal{G}$ is the expected empirical Rademacher complexity of $\mathcal{G}$ over the random choices of S:

$$\mathcal{R}_m(\mathcal{G}) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\hat{\mathcal{R}}_S(\mathcal{G})].$$

We first make a few observations about the Rademacher complexity of a class of binary classifiers and their corresponding 01-loss functions. Let $\mathcal{H}$ be a class of binary classifiers from Ω to $\{-1, 1\}$. (From now on we will use the range $\{-1, 1\}$ instead of $\{0, 1\}$ for technical convenience.) Let $\mathcal{Z} = \Omega \times \{-1, 1\}$, and $\mathcal{D}$ be a (unknown) distribution on $\mathcal{Z}$. Elements of $\mathcal{Z}$ have the form $(\mathbf{x}, y)$ where $\mathbf{x} \in \Omega$ and $y \in \{-1, 1\}$ which is referred to as a label of $\mathbf{x}$.

Now, for every hypothesis $h \in \mathcal{H}$, define the *loss function* $g_h : \mathcal{Z} \to [0, 1]$ for h by $g_h(\mathbf{x}, y) = \mathbf{1}_{h(\mathbf{x}) \neq y}$. Let $\mathcal{G}$ be the class of loss functions for $\mathcal{H}$.

Exercise 1. Prove the following four relationships.

$$\widehat{\text{err}}_S(h) = \frac{1}{m}\sum_{i=1}^{m} g_h(z_i), \text{ for all } h \in \mathcal{H} \tag{1}$$

$$\text{err}(h) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[g_h(z)] \text{ for all } h \in \mathcal{H} \tag{2}$$

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \frac{1}{2}\hat{\mathcal{R}}_S(\mathcal{H}) \tag{3}$$

$$\mathcal{R}(\mathcal{G}) = \frac{1}{2}\mathcal{R}(\mathcal{H}). \tag{4}$$

The objective is to prove the following uniform convergence theorem using Rademacher complexity instead of the VC dimension of $\mathcal{H}$.

Theorem 1.3. *Let $\mathcal{H}$ be a class of functions from Ω to $\{-1, 1\}$. Let $\mathcal{D}$ be an arbitrary distribution on $\mathcal{Z} = \Omega \times \{-1, 1\}$. Then, for any $\delta > 0$,*

$$\mathop{\textit{Prob}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\{\text{err}(h) - \widehat{\text{err}}(h)\} \leq \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}}\right] \geq 1 - \delta,$$

and

$$\mathop{\textit{Prob}}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\{\text{err}(h) - \widehat{\text{err}}(h)\} \leq \widehat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}\right] \geq 1 - \delta.$$

From Exercise 1, it follows that the Lemma stated in the next section implies the above theorem. Basically, we convert a statement about the function class $\mathcal{H}$ into a statement about its loss function class $\mathcal{G}$. We will prove the lemma instead.

2 The main lemma

Lemma 2.1 (Koltchinskii-Panchenko, 2002 [**?**])**.** *Let $\mathcal{G}$ be a class of functions from $\mathcal{Z}$ to $[0, 1]$, and $\mathcal{D}$ be an arbitrary distribution on $\mathcal{Z}$. For notational convenience, we write*

$$\mathbb{E}[g] = \mathop{\mathbb{E}}_{z\in\mathcal{D}}[g(z)]$$

$$\widehat{\mathbb{E}}_S[g] = \frac{1}{m}\sum_{i=1}^{m} g(z_i), \textit{ where } S = \{z_1, \ldots, z_m\}.$$

Then, for any $\delta > 0$,

$$\underset{S\sim\mathcal{D}^m}{\textit{Prob}}\left[\sup_{g\in\mathcal{G}}\{\mathbb{E}[g]-\widehat{\mathbb{E}}_S[g]\}\leq 2\mathcal{R}_m(\mathcal{G})+\sqrt{\frac{\ln(1/\delta)}{2m}}\right]\geq 1-\delta, \tag{5}$$

and

$$\underset{S\sim\mathcal{D}^m}{\textit{Prob}}\left[\sup_{g\in\mathcal{G}}\{\mathbb{E}[g]-\widehat{\mathbb{E}}_S[g]\}\leq 2\widehat{\mathcal{R}}_S(\mathcal{G})+3\sqrt{\frac{\ln(2/\delta)}{2m}}\right]\geq 1-\delta. \tag{6}$$

Proof. The proof of this lemma is similar to the proof of the Vapnik-Chervonenkis uniform convergence theorem. One of the main components of the proof is the double sampling trick. The other component is the swapping each pair of "normal" sample and "ghost" sample. We prove (5) first. Intuitively, the proof contains two main steps.

1. Define $\Phi(S)=\sup_{g\in\mathcal{G}}\{\mathbb{E}[g]-\widehat{\mathbb{E}}_S[g]\}$. Then, $\Phi(S)$ is a random variable (over the random choices of S). We will show that
$$\mathop{\mathbb{E}}_{S}[\Phi(S)]\leq 2\mathcal{R}_m(\mathcal{G}).$$

2. Then, to show (5) we show that $\Phi(S)$ cannot be much more than its expected value with high probability using a concentration inequality called the McDiarmid inequality. The statement and a proof of McDiarmid inequality can be found in Section 4.

Let us start with the first step. We will take m ghost samples $S'=\{z'_1,\ldots,z'_m\}$. Also, let $\sigma=(\sigma_1,\ldots,\sigma_m)$

denote m independent Rademacher variables.

$$\begin{aligned}
\mathop{\mathbb{E}}_{S}[\Phi(S)] &= \mathop{\mathbb{E}}_{S}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}[g]-\widehat{\mathbb{E}}_S[g]\right\}\right] \\
&= \mathop{\mathbb{E}}_{S}\left[\sup_{g\in\mathcal{G}}\left\{\mathop{\mathbb{E}}_{S'}\left[\widehat{\mathbb{E}}_{S'}[g]\right]-\mathop{\mathbb{E}}_{S'}\left[\widehat{\mathbb{E}}_S[g]\right]\right\}\right] \\
&= \mathop{\mathbb{E}}_{S}\left[\sup_{g\in\mathcal{G}}\left\{\mathop{\mathbb{E}}_{S'}\left[\widehat{\mathbb{E}}_{S'}[g]\right]-\widehat{\mathbb{E}}_S[g]\right\}\right] \\
\text{(Jensen ineq., sup is convex)} &\leq \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{S'}\left[\sup_{g\in\mathcal{G}}\left\{\widehat{\mathbb{E}}_{S'}[g]-\widehat{\mathbb{E}}_S[g]\right\}\right]\right] \\
&= \mathop{\mathbb{E}}_{S,S'}\left[\sup_{g\in\mathcal{G}}\left\{\widehat{\mathbb{E}}_{S'}[g]-\widehat{\mathbb{E}}_S[g]\right\}\right] \\
&= \mathop{\mathbb{E}}_{S,S'}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{m}\sum_{i=1}^{m}(g(z'_i)-g(z_i))\right\}\right] \\
\text{(just swapping } z'_i, z_i) &= \mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i(g(z'_i)-g(z_i))\right\}\right] \\
&\leq \mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z'_i)\right\}+\sup_{g\in\mathcal{G}}\left\{\frac{1}{m}\sum_{i=1}^{m}-\sigma_i g(z_i))\right\}\right] \\
&= \mathop{\mathbb{E}}_{S',\sigma}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z'_i)\right\}\right]+\mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i))\right\}\right] \\
&= 2\mathcal{R}_m(\mathcal{G}).
\end{aligned}$$

We next show step 2. If we change $S=(z_1,\ldots,z_m)$ by one point, say we replace z_m by z'_m, then

$$\Phi(S)=\sup_{g\in\mathcal{G}}\left\{\mathbb{E}[g]-\frac{1}{m}\sum_{i=1}^{m}g(z_i)\right\}$$

can only change by at most $1/m$. Hence, by McDiarmid's inequality

$$\text{Prob}[\Phi(S)-2\mathcal{R}_m(\mathcal{G})\geq\epsilon]\leq\text{Prob}[\Phi(S)-\mathbb{E}[\Phi(S)]\geq\epsilon]\leq e^{-2\epsilon^2 m}.$$

For $e^{-2\epsilon^2 m}\leq\delta$, we can set $\epsilon=\sqrt{\frac{\ln(1/\delta)}{2m}}$. This completes the proof of (5).

To prove (6), we apply (5) and McDiarmid's inequality one more time on the function $\widehat{\mathcal{R}}_S(\mathcal{G})$. We leave this as an exercise. □

Exercise 2. Prove (6) by applying (5) and McDiarmid's inequality one more time on the function $\widehat{\mathcal{R}}_S(\mathcal{G})$.

3 Bounding Rademacher complexity by VC-dimension

The following bound shows that Theorem 1.3 implies Vapnik-Chervonenkis uniform convergence theorem.

Theorem 3.1. *Let $\mathcal{H}$ be a class of functions from Ω to $\{-1,1\}$. Let S be m arbitrary points from $\Omega \times \{-1,1\}$. Then,*

$$\widehat{\mathcal{R}}_S(\mathcal{H}) \leq \sqrt{\frac{2\log|\Pi_{\mathcal{H}}(S)|}{m}}.$$

In particular, if $d = \text{VCD}(\mathcal{H})$ then we have

$$\widehat{\mathcal{R}}_S(\mathcal{H}) \leq \sqrt{\frac{2\log|\Pi_{\mathcal{H}}(m)|}{m}} \leq \sqrt{\frac{2d\log(me/d)}{m}},$$

and

$$\mathcal{R}_m(\mathcal{H}) = \mathop{\mathbb{E}}_S[\widehat{\mathcal{R}}_S(\mathcal{H})] \leq \sqrt{\frac{2\log|\Pi_{\mathcal{H}}(m)|}{m}} \leq \sqrt{\frac{2d\log(me/d)}{m}}.$$

The theorem follows almost immediately from a lemma by Massart [**?**]. Using the notion of Gaussian complexity, Bartlett and Mendelson [**?**] proved that $\widehat{\mathcal{R}}_S(\mathcal{H}) = O(\sqrt{d/m})$. Hence, the bound can even be better then what is stated in Theorem 3.1.

Lemma 3.2 (Massart [**?**])**.** *Let $A \subset \mathbb{R}^m$ be a finite set. Let $L = \max_{\mathbf{x}\in A}\|\mathbf{x}\|_2$. Then,*

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{x}\in A}\frac{1}{m}\sum_{i=1}^{m}\sigma_i x_i\right] \leq \frac{L\sqrt{2\log|A|}}{m}.$$

Proof. First of all, there is $1/m$ on both sides which means we can just ignore the factor $1/m$. We use Bernstein's trick. Let $t > 0$ be any real number. Then, using the fact that the exponential function is convex, Jensen's inequality gives us

$$\begin{aligned}
& \exp\left(t\cdot\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) \\
\text{(Jensen ineq.)} \quad &\leq \mathop{\mathbb{E}}_{\sigma}\left[\exp\left(t\cdot\sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right)\right] \\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{x}\in A}\exp\left(t\cdot\sum_{i=1}^{m}\sigma_i x_i\right)\right] \\
\text{(Union bound and linearity of expectation)} \quad &\leq \sum_{\mathbf{x}\in A}\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(t\sum_{i=1}^{m}\sigma_i x_i\right)\right] \\
&= \sum_{\mathbf{x}\in A}\prod_{i=1}^{m}\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(t\sigma_i x_i\right)\right]
\end{aligned}$$

The above expression should look familiar to us. We used to deal with similar expressions in the Tail/Concentration inequalities part of this course. However, the random variables were in $[0,1]$. This time, σ_i ranges from -1 to 1. From Hoeffding's lemma (Lemma 4.1), we get

$$\mathbb{E}[\exp(t\sigma_i x_i)] \leq \exp(t^2 x_i^2/2), \text{ for all } i \in [m].$$

Consequently, for any $t > 0$ we have

$$\exp\left(t \cdot \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) \le \sum_{\mathbf{x}\in A}\prod_{i=1}^{m}\exp(t^2x_i^2/2) = \sum_{\mathbf{x}\in A}\exp(\|\mathbf{x}\|_2^2 t^2/2) \le |A|\exp(L^2t^2/2).$$

Now, taking $\ln$ on both sides we obtain

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \le \frac{\ln|A|}{t} + \frac{L^2 t}{2}.$$

To minimize the upper bound, we choose $t = \frac{\sqrt{2\ln|A|}}{L}$ which completes the proof of the lemma. □

Exercise 3. Prove Theorem 3.1 from Lemma 3.2.

Exercise 4 (Lower Bound)**.** TBD.

4 Hoeffding's lemma and McDiarmid's inequality

Lemma 4.1 (Hoeffding's lemma)**.** *Let $X \in [a, b]$ be a random variable with $\mathbb{E}[X] = 0$. Then, for all $s > 0$*

$$\mathbb{E}[e^{sX}] \le e^{s^2(b-a)^2/8}.$$

Proof. Note that $\mathbb{E}[X] = 0$ implies $a \le 0$ and $b \ge 0$. The function e^{sX} is convex in X. Hence,

$$e^{sX} \le \frac{b-X}{b-a}e^{sa} + \frac{X-a}{b-a}e^{sb}.$$

Taking expectation on both sides, we get

$$\mathbb{E}[e^{sX}] \le \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}.$$

To simplify notations, define $p = \frac{-a}{b-a} \ge 0$ and $t = s(b-a)$. We have

$$\ln\left(\frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}\right) = -pt + \ln(1-p+pe^t) =: f(t).$$

(Now, this form should look really familiar!) Then,

$$\begin{aligned} f'(t) &= = -p + \frac{p}{p+(1-p)e^{-t}} \\ f''(t) &= = \frac{(1-p)pe^t}{((1-p)+pe^t)^2}. \end{aligned}$$

It follows that $f(0) = f'(0) = 0$, and $f''(t) \le 1/4, \forall t$. The second order Taylor's expansion of $f(t)$ above 0 implies, for any $t > 0$, there is some $\zeta \in [0, t]$ such that

$$f(t) = f(0) + tf'(0) + \frac{t^2}{2}f''(\zeta) \le t^2/8.$$

Putting everything together, we obtain the desired inequality:

$$\mathbb{E}[e^{sX}] \leq e^{f(t)} \leq e^{t^2/8} = e^{s^2(b-a)^2/8}.$$

□

McDiarmid [**?**] proved the following inequality. We can prove it by applying Azuma-Hoeffding's inequality to a Doob martingale. Here we are taking a more direct approach.

Theorem 4.2 (McDiarmid's inequality)**.** *Let $X_1, \ldots, X_m$ be m independent random variables on the domain $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$ be a function which maps $X_1, \ldots, X_m$ to a real number. Suppose changing a single coordinate does not change f by much. Specifically, for every $i \in [m]$, $x_1, \ldots, x_m, x_i' \in \mathcal{X}$, we have*

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i \tag{7}$$

where the c_i are some given constants. Then, for any $\epsilon > 0$,

$$\textit{Prob}[f - \mathbb{E}[f] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

Sketch. The proof conceptually is not hard, but it might be confusing if you are not used to conditional expectations. The proof has the following steps.

1. Let $\mathbf{X}^i$ denote the sequence of r.v. $X_1, \ldots, X_i$
2. Define $Z_i = \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}^i]$. (Technically, the sequence Z_i forms a *Doob martingale* but we do not need to know martingale theory here.)
3. Note that $Z_0 = \mathbb{E}[f]$ and $Z_m = f$
4. Consider the random variable $Z_k - Z_{k-1}$
5. Note that $\mathbb{E}[Z_k - Z_{k-1} \mid \mathbf{X}^{k-1}] = 0$
6. Define

$$\begin{aligned} B_k &= \sup_{x_k} \left\{ \mathbb{E}[f \mid \mathbf{X}^{k-1}, X_k = x_k] - \mathbb{E}[f \mid \mathbf{X}^{k-1}] \right\} \\ A_k &= \inf_{x_k} \left\{ \mathbb{E}[f \mid \mathbf{X}^{k-1}, X_k = x_k] - \mathbb{E}[f \mid \mathbf{X}^{k-1}] \right\} \end{aligned}$$

7. Then,

$$\begin{aligned} A_k &\leq Z_k - Z_{k-1} \leq B_k \\ B_k - A_k &\leq c_k \end{aligned}$$

8. Thus, by Hoeffding's lemma for any $t > 0$, and any values assigned to $\mathbf{X}^{k-1}$,

$$\mathbb{E}[e^{t(Z_k - Z_{k-1})} \mid \mathbf{X}^{k-1}] \leq e^{t^2(B_k - A_k)^2/8} \leq e^{t^2 c_k^2/8}.$$

9. Finally, we apply Bernstein's trick and LIE again

$$\begin{aligned}
\text{Prob}[f - \mathbb{E}[f] \geq \epsilon] &= \text{Prob}\left[e^{t(f-\mathbb{E}[f])} \geq e^{t\epsilon}\right] \\
\text{(Markov)} &\leq e^{-t\epsilon}\,\mathbb{E}\left[e^{t(f-\mathbb{E}[f])}\right] \\
\text{(Telescoping)} &= e^{-t\epsilon}\,\mathbb{E}\left[e^{t\sum_{k=1}^{m}(Z_k - Z_{k-1})}\right] \\
\text{(LIE)} &= e^{-t\epsilon}\,\mathbb{E}\left[\mathbb{E}\left[e^{t\sum_{k=1}^{m}(Z_k - Z_{k-1})} \mid \mathbf{X}^{m-1}\right]\right] \\
&= e^{-t\epsilon}\,\mathbb{E}\left[e^{t\sum_{k=1}^{m-1}(Z_k - Z_{k-1})}\,\mathbb{E}\left[e^{t(Z_m - Z_{m-1})} \mid \mathbf{X}^{m-1}\right]\right] \\
&\leq e^{-t\epsilon}e^{t^2c_m^2/8}\,\mathbb{E}\left[e^{t\sum_{k=1}^{m-1}(Z_k - Z_{k-1})}\right] \\
&\leq \cdots \\
&\leq \exp\left(-t\epsilon + (t^2/8)\sum_{k=1}^{m} c_k^2\right)
\end{aligned}$$

10. Finally, we pick t to minimize $-t\epsilon + (t^2/8)\sum_{k=1}^{m} c_k^2$

$$t = \frac{4\epsilon}{\sum_{k=1}^{m} c_k^2}$$

to finish the proof.

□

More details. The following spells out some of the details in the proof. We number the steps as in the sketch.

1. This step is just a definition
2. What do we mean by $Z_i = \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}^i]$? Let's start with Z_0. In this case, $\mathbf{X}^0$ is empty and thus $\mathbf{Z}_0 = \mathbb{E}[f]$. Next,

$$Z_1 = \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}^1] = \mathbb{E}[f(\mathbf{X}) \mid X_1].$$

   This is a random variable which is a function of X_1. More concretely, for any x_1 in the domain of X_1 we have $Z_1(x_1) = \mathbb{E}[f(\mathbf{X}) \mid X_1 = x_1]$. So, the expectation is over the conditional distribution of $X_2, \ldots, X_m$ given X_1. However, since the X_i are all independent, this is simply an expectation of $f(\mathbf{X})$ over $X_2, \ldots, X_m$ with X_1 fixed to be x_1. Similarly,

$$Z_k = \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}^k] = \mathbb{E}[f(\mathbf{X}) \mid X_1, \ldots, X_k]$$

   is a random variable which is a function of $X_1, \ldots, X_k$, and the expectation is over $X_{k+1}, \ldots, X_m$ given $X_1, \ldots, X_k$.
3. This is obvious.
4. The random variable $Z_k - Z_{k-1}$ is a function of $X_1, \ldots, X_k$.

5. Now, suppose we fixed $X_i = x_i$ for $i \in [k-1]$. Then, Z_{k-1} is a number, no longer a random variable; while Z_k is a random variable which is a function of X_k. Hence,

$$\begin{aligned}
\mathbb{E}[Z_k - Z_{k-1} \mid \mathbf{X}^{k-1}] &= \mathbb{E}[Z_k \mid \mathbf{X}^{k-1}] - \mathbb{E}[Z_{k-1} \mid \mathbf{X}^{k-1}] \\
&= \mathbb{E}[\mathbb{E}[f \mid \mathbf{X}^{k-1}, X_k] \mid \mathbf{X}^{k-1}] - \mathbb{E}[f \mid \mathbf{X}^{k-1}] \\
\text{(LIE)} \quad &= \mathbb{E}[f \mid \mathbf{X}^{k-1}] - \mathbb{E}[f \mid \mathbf{X}^{k-1}] \\
&= 0.
\end{aligned}$$

(See section 5 for LIE.)

6. Note that A_k and B_k are actually functions of $X_1, \ldots, X_{k-1}$. So they are themselves random variables.

7. We claim that the inequalities hold for any values of $X_1, \ldots, X_{k-1}$. The fact that $A_k \le Z_k - Z_{k-1} \le B_k$ is obvious. We check the second inequality. Fix arbitrary values of $X_1, \ldots, X_{k-1}$.

$$\begin{aligned}
B_k - A_k &= \sup_x \left\{ \mathbb{E}[f \mid \mathbf{X}^{k-1}, X_k = x] - \mathbb{E}[f \mid \mathbf{X}^{k-1}] \right\} - \inf_y \left\{ \mathbb{E}[f \mid \mathbf{X}^{k-1}, X_k = y] - \mathbb{E}[f \mid \mathbf{X}^{k-1}] \right\} \\
&= \sup_{x,y} \left\{ \mathbb{E}[f \mid \mathbf{X}^{k-1}, X_k = x] - \mathbb{E}[f \mid \mathbf{X}^{k-1}, X_k = y] \right\}
\end{aligned}$$

Now, fix an arbitrary pair x, y (and still fix arbitrary values for $\mathbf{X}^{k-1}$). Let $f_k(\mathbf{X}, x)$ denote $f(\mathbf{X})$ with the first $k-1$ variables fixed to $\mathbf{X}^{k-1}$ and the kth variable fixed to be x. Then, because the variables $X_{k+1}, \ldots, X_m$ are *independent* from the variables $X_1, \ldots, X_k$, we have

$$B_k - A_k = \sup_{x,y} \left\{ \mathop{\mathbb{E}}_{X_{k+1},\ldots,X_m} [\underbrace{f_k(\mathbf{X}, x) - f_k(\mathbf{X}, y)}_{\le c_k}] \right\} \le c_k.$$

8. Self-evident

9. Self-evident

□

Exercise 5. In the above proof, we crucially made use of the fact that the X_i are independent in step 7. If the X_i were not independent we might not have been able to combine the expectations into one because the conditional distribution of the $X_{k+1}, \ldots, X_m$ given $X_k = x$ and $X_k = y$ might be different. Using this idea, find an example of non-independent random variables X_i, a function f satisfying (7), but $B_k - A_k > c_k$ for some k.

Exercise 6 (Hoeffding's inequality from McDiarmid's inequality)**.** Prove the following inequality (Hoeffding's inequality) using McDiarmid's inequality. For $i \in [m]$, let $X_i \in [a_i, b_i]$ be independent random variables. Let $S = \frac{1}{m} \sum_{i=1}^m X_i$. Show that

$$\text{Prob}[S - \mathbb{E}[S] \ge \epsilon] \le \exp\left(\frac{-2\epsilon^2 m^2}{\sum_{i=1}^m (b_i - a_i)^2} \right).$$

5 Law of iterated expectation (LIE)

We have alluded to the following rule (the conditional expectation formula) a few times. Economists like to call this rule a LIE. Let X, Y be any random variables, then

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]].$$

Let us make the expectations more precise:

$$\mathop{\mathbb{E}}_{X}[X] = \mathop{\mathbb{E}}_{Y}\left[\mathop{\mathbb{E}}_{X|Y}[X \mid Y]\right].$$

The outer expectation is over the distribution of Y. The expression $\mathbb{E}_{X|Y}[X \mid Y]$ is a random variable which is a function of Y, and the expectation is over the conditional distribution of X given Y. To technically understand LIE, let us prove it for both continuous distributions and discrete distributions. The two cases are basically identical.

In the discrete case, we can derive LIE as follows.

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]] &= \mathop{\mathbb{E}}_{Y}\left[\mathop{\mathbb{E}}_{X|Y}[X \mid Y]\right] \\
&= \sum_{y} (\mathbb{E}[X \mid Y = y]) \,\text{Prob}[Y = y] \\
&= \sum_{y} \left(\sum_{x} x \cdot \text{Prob}[X = x \mid Y = y] \right) \text{Prob}[Y = y] \\
&= \sum_{x} \sum_{y} x \cdot \text{Prob}[X = x \wedge Y = y] \\
&= \sum_{x} x \sum_{y} \text{Prob}[Y = y \mid X = x] \,\text{Prob}[X = x] \\
&= \sum_{x} x \,\text{Prob}[X = x] \sum_{y} \text{Prob}[Y = y \mid X = x] \\
&= \sum_{x} x \,\text{Prob}[X = x] \\
&= \mathbb{E}[X]
\end{aligned}$$

In the continuous case, we have

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]] &= \mathop{\mathbb{E}}_{Y}\left[\mathop{\mathbb{E}}_{X|Y}[X \mid Y]\right] \\
&= \int_{-\infty}^{\infty} \mathop{\mathbb{E}}_{X|Y}[X \mid Y = y] f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx\right) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy\right) dx \\
&= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y) dy\right) dx \\
&= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) dy\right) dx \\
&= \int_{-\infty}^{\infty} x f_X(x) \underbrace{\left(\int_{-\infty}^{\infty} f_{Y|X}(y|x) dy\right)}_{=1} dx \\
&= \int x f_X(x) dx \\
&= \mathbb{E}[X]
\end{aligned}$$

Exercise 7. Show that for any continuous variables X, Y, Z we have

$$\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Z] = \mathbb{E}[X \mid Z].$$

(This is why the rule is called the law of "iterated" expectation.)

References