1 Recap

We had started looking at the case where the data points and labels came in pairs, and were drawn from some distribution $D$. Given a sample of i.i.d. random variables $X_1, \ldots, X_m$, with $X_i \in [0, 1]$, we defined $p = \mathbb{E}[X_i]$ and $\hat{p} = \frac{1}{m} \sum_{i=1}^{m} X_i$ and sought to show that $\hat{p}$ converges uniformly to $p$. To do this, we proved Hoeffding’s Inequality:

$$\Pr[\hat{p} \geq p + \varepsilon] = \Pr[\hat{p} \leq p - \varepsilon] \leq e^{-2\varepsilon^2 m}$$

and found stricter bounds for these quantities using Relative Entropy $^1$ $^2$:

$$\Pr[\hat{p} \geq p + \varepsilon] \leq e^{-\text{RE}(p+\varepsilon\|p)m}$$

$$\Pr[\hat{p} \leq p - \varepsilon] \leq e^{-\text{RE}(p-\varepsilon\|p)m}$$

2 McDiarmid’s Inequality

We now look at a generalization of Hoeffding’s Inequality — McDiarmid’s Inequality. While constructing Hoeffding’s Inequality, we had considered $\hat{p} = \frac{1}{m} \sum_{i=1}^{m} X_i$, and had shown that $\frac{1}{m} \sum_{i=1}^{m} X_i = \hat{p} \rightarrow p = \mathbb{E}[X_i] = \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} X_i\right]$. Suppose we now wanted to consider a general case where we replace $\hat{p}$ by some function of the sample, $f(X_1, \ldots, X_m)$. Could we always claim that $f(X_1, \ldots, X_m) \rightarrow \mathbb{E}[f(X_1, \ldots, X_m)]$? For this to hold, we need a special property that changing one input to the function $f$ does not change its value by much. Formally, we assume that $\forall i, \forall x_1, \ldots, x_m$ and $x'_i$ (where $x_1, \ldots, x_m, x'_i$ are possible values for the input variables of the function $f$)

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x'_i, \ldots, x_m)| \leq c_i$$

where $c_i$ is some constant.

**Theorem 1** (McDiarmid’s Inequality). Assume $X_1, \ldots, X_m$ are independent (not necessarily identical) random variables, and $f$ is some function that satisfies the property above. Then,

$$\Pr[f(X_1, \ldots, X_m) \geq \mathbb{E}[f(X_1, \ldots, X_m)] + \varepsilon] \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{m} c_i^2}\right).$$

Hoeffding’s Inequality is a special case of McDiarmid’s Inequality. We require that the random variables $X_1, \ldots, X_m$ are i.i.d, and $X_i \in [0, 1]$. Then, we define $f(X_1, \ldots, X_m) = \frac{1}{m} \sum_{i=1}^{m} X_i$. Note that because the $X_i$’s are constrained to be either 0 or 1, changing one of these values will change the value of $f(X_1, \ldots, X_m)$ by at most $\frac{1}{m}$. So, we set $c_i = \frac{1}{m}$ in McDiarmid’s Inequality to get the required result.

$^1$While the second inequality was not proven in class, its proof resembles that for the first inequality, using the random variables $1 - X_1, \ldots, 1 - X_m$ instead of $X_1, \ldots, X_m$.

$^2$Hoeffding’s inequality is a special case of the first inequality, using the identity that $\text{RE}(p + \varepsilon \| p) \geq 2\varepsilon^2$.  

3 Learning in a Finite Hypothesis Space

Theorem 2. Let $|\mathcal{H}| < \infty$. Given a sample of $m$ points $S = (x_1,\ldots,x_m)$ from some distribution $D$, we have that with probability $\geq 1 - \delta$, $\forall h \in \mathcal{H}$,

$$|\text{err}(h) - \hat{\text{err}}(h)| \leq \varepsilon$$

if $m \geq \frac{\ln 2|\mathcal{H}| + \ln \frac{1}{\delta}}{2\varepsilon^2}$.

Proof. For a fixed hypothesis $h \in \mathcal{H}$, Hoeffding’s inequality gives us that $\Pr[|\hat{p} - p| > \varepsilon] \leq 2e^{-2\varepsilon^2 m}$. As we are dealing with a finite hypothesis space, we can use the Union Bound:

$$\Pr[\exists h \in \mathcal{H} : |\text{err}(h) - \hat{\text{err}}(h)| > \varepsilon] \leq 2|\mathcal{H}|e^{-2\varepsilon^2 m}$$

Setting the RHS to $\delta$ gives us that

$$m = O\left(\frac{\ln 2|\mathcal{H}| + \ln \frac{1}{\delta}}{2\varepsilon^2}\right). \quad (1)$$

Equivalently, we can say that with probability $\geq 1 - \delta$, $\forall h \in \mathcal{H}$,

$$\text{err}(h) \leq \hat{\text{err}}(h) + O\left(\sqrt{\frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{m}}\right). \quad (2)$$

Note that we dropped the two-sided inequality in favour of the one-sided inequality in (2) because for our purposes, it suffices to consider only the direction shown here.

We observe the following from the bounds above:

- The error reduces at a rate of $O\left(\frac{1}{\sqrt{m}}\right)$ in (2), compared to a rate of $O\left(\frac{1}{m}\right)$ when working with consistent hypotheses.

- The amount of data needed in (1) increases from being $O\left(\frac{1}{\varepsilon}\right)$ when working with consistent hypotheses, to $O\left(\frac{1}{\varepsilon^2}\right)$. This is reflected in the Relative Entropy version of the inequality as well — when $p$ is close to $\frac{1}{2}$, $\text{RE}(p + \varepsilon \parallel p)$ is close to $\frac{1}{\varepsilon^2}$; otherwise it is close to $\frac{1}{\varepsilon}$ when $p$ is close to 0 or 1.

This distinction arises because of the difference in the upper bounds we are using — we used $e^{-\varepsilon m}$ in the consistency model, while we use $e^{-2\varepsilon^2 m}$ in this case.

Now suppose that we were encoding the hypothesis space $\mathcal{H}$ by bits. Then, we can replace $\ln |\mathcal{H}|$ in the error bound with $|h|$. In this scenario, the inequality (2) manages to capture the three required properties for learning:

- Simplicity versus Complexity: the lower the value of $|h|$, the lower the generalization error $\text{err}(h)$.

- Large amount of data: the higher the value of $m$, the lower the error.

- Good fit to the dataset: the lower the training error $\hat{\text{err}}(h)$, the lower the generalization error.
Figure 1: Trade-off between complexity and error

We can plot the errors as a function of the complexity of the hypotheses, and get the graph above\(^3\). We can see that increasing the complexity reduces the training error — with more complex hypotheses, we can fit the training data better. It reduces the generalization error as well initially; however, we fall prey to overfitting as the complexity increases, causing an increase in generalization error.

4 Learning in an Infinite Hypothesis Space

In previous lectures, we have used complexity measures such as the growth function and VC-dimension to help us prove learnability in an infinite hypothesis space. However, we will now look at a new measure of complexity that subsumes those that we’ve seen previously — namely, the Rademacher Complexity.

4.1 Rademacher Complexity

We start with a sample of \( m \) points \( S = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle \) where \( x_1, \ldots, x_m \in \mathcal{X} \) and \( y_1, \ldots, y_m \in \{-1, 1\} \), drawn from some distribution \( \mathcal{D} \). Then we can use the training error to measure how well a fixed hypothesis \( h \) fits the training data:

\[
\hat{\mathcal{E}}_{\mathcal{T}}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq y_i\}
\]

\[
= \frac{1}{m} \sum_{i=1}^{m} \frac{1 - y_i h(x_i)}{2}
\]

\[
= \frac{1}{2} - \frac{1}{2m} \sum_{i=1}^{m} y_i h(x_i)
\]

\[
\Longrightarrow \; \frac{1}{m} \sum_{i=1}^{m} y_i h(x_i) = 1 - 2\hat{\mathcal{E}}_{\mathcal{T}}(h)
\]

\(^3\)Reference: http://www.cs.princeton.edu/courses/archive/spring18/cos511/scribe_notes/0305.pdf
This shows that $\frac{1}{m} \sum_{i=1}^{m} y_i h(x_i)$ can be used as a measure of how well $h$ fits the data set, and that this measure is equivalent to training error.

As the best hypothesis in $H$ minimizes $\hat{\text{err}} (h)$, we can measure how well the entire hypothesis space $H$ fits the sample using

$$\max_{h \in H} \frac{1}{m} \sum_{i=1}^{m} y_i h (x_i).$$

Now, consider the following experiment: suppose the labels $y_i$ are given at random. We are interested in finding how well $H$ will fit pure noise. Formally, we replace the labels $y_i$ with independent random variables $\sigma_i$ (also known as Rademacher random variables) such that

$$\sigma_i = \begin{cases} 1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5. \end{cases}$$

Define

$$R = \mathbb{E}_\sigma \left[ \max_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h (x_i) \right].$$

Intuitively, we can see that if a hypothesis class is rich enough, then it is more likely to fit the random labels, and hence, have a higher value of $R$. However, this also exposes us to the dangers of overfitting the given sample.

Let us consider some extreme cases to check for the values of $R$:

- Suppose $H = \{h\}$. Then

  $$R = \mathbb{E}_\sigma \left[ \max_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h (x_i) \right]$$

  $$= \mathbb{E}_\sigma \left[ \frac{1}{m} \sum_{i=1}^{m} \sigma_i h (x_i) \right]$$

  $$= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_\sigma \left[ \sigma_i \right]$$

  $$= 0$$

  This is the minimum possible value of $R$ as $\mathbb{E} \left[ \max_{f} f \right] \geq \max_{f} \mathbb{E} \left[ f \right]$ (by the argument given last lecture) — so $R$ can never be negative.

- Suppose $S$ is shattered by $H$. Then, we know that for any labelling $\sigma$, there exists a hypothesis $h \in H$ such that $h (x_i) = \sigma_i$ for $i = 1, \ldots, m$. In this case, $R = 1$. This is the maximum value $R$ can take.

We will study these topics in a more general and abstract setting. Assume now that we have a family $F$ of real-valued functions where $f : Z \rightarrow \mathbb{R}$ for some set $Z$. Let $S = \langle z_1, \ldots, z_m \rangle$ where $z_1, \ldots, z_m \in Z$ are independently drawn from some distribution $D$. We define the empirical Rademacher Complexity as

$$\hat{R}_S (F) = \mathbb{E}_\sigma \left[ \sup_{f \in F} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f (z_i) \right].$$

4
Note that we are now using the supremum (the least upper bound) instead of the maximum in our definition, and that the empirical Rademacher Complexity is defined with respect to a particular sample $S$. Similarly, we define the expected Rademacher Complexity as

$$R_m(F) = \mathbb{E}_S[\hat{R}_S(F)].$$

We want to prove that $\forall f \in F, \frac{1}{m} \sum_{i=1}^{m} f(z_i) \rightarrow \mathbb{E}_{z \sim D}[f(z)]$ with high probability. We will make use of the shorthand $\hat{E}_S[f] = \frac{1}{m} \sum_{i=1}^{m} f(z_i)$ and $\mathbb{E}[f] = \mathbb{E}_{z \sim D}[f(z)]$.

**Theorem 3.** Let $F$ be a family of functions $f : Z \rightarrow [0, 1]$, and suppose $S = (z_1, \ldots, z_m)$ where $z_i \sim D$. Then, with probability $\geq 1 - \delta$,

$$\forall f \in F : \mathbb{E}[f] \leq \hat{E}_S[f] + 2R_m(F) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right).$$

In terms of the empirical Rademacher Complexity, we have

$$\forall f \in F : \mathbb{E}[f] \leq \hat{E}_S[f] + 2\hat{R}_S(F) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right).$$

**Proof.** We define

$$\Phi(S) = \sup_{f \in F} \left( \mathbb{E}[f] - \hat{E}_S[f] \right)$$

and see that it suffices to consider a bound for $\Phi(S)$, as that would apply for all $f \in F$. The proof consists of three steps. We will show the first step and introduce the second step, but the proof will be completed in the next lecture.

**Step 1**

$\Phi(S)$ is a random variable that is cumbersome to work with. We would prefer to use the constant $\mathbb{E}_S[\Phi(S)]$. In order to do so, we need to first prove that with probability $\geq 1 - \delta$,

$$\Phi(S) \leq \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\ln \frac{1}{\delta}}{m}}.$$

This inequality can be proven using McDiarmid’s Inequality. However, we need to first check whether the conditions for McDiarmid’s Inequality are satisfied:

- The inputs to $\Phi$ must be independent random variables: As $\Phi(S) = \Phi(z_1, \ldots, z_m)$, and $z_1, \ldots, z_m$ are independently distributed from $D$, this condition is satisfied.

- A change in the input should not change the value of $\Phi$ by much: If we change $z_i$ for some $i \in \{1, \ldots, m\}$, $\mathbb{E}[f]$ does not change. $\hat{E}_S[f] = \frac{1}{m} \sum_{i=1}^{m} f(z_i)$, and $z_i \in [0, 1]$, so the value of $\hat{E}_S[f]$ changes by at most $\frac{1}{m}$, and hence, changes the value of $\Phi(S)$ by at most $\frac{1}{m}$. Therefore, setting $c_i = \frac{1}{m}$ in McDiarmid’s Inequality gives us the required bound.
Step 2
The term $\mathbb{E}[f]$ is cumbersome to work with as well; we will make use of the double-sampling trick to help us find a replacement. Suppose $S' = \langle z'_1, \ldots, z'_m \rangle$ where $z'_i$ are independently chosen from $\mathcal{D}$. We want to replace $\mathbb{E}[f]$ with $\hat{\mathbb{E}}_{S'}[f]$. In order to do so, we will prove that

$$
\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S \left[ \sup_{f \in \mathcal{F}} \left( \mathbb{E}[f] - \hat{\mathbb{E}}_S[f] \right) \right] 
\leq \mathbb{E}_{S, S'} \left[ \sup_{f \in \mathcal{F}} \left( \hat{\mathbb{E}}_{S'}[f] - \hat{\mathbb{E}}_S[f] \right) \right]
$$