Mathematical Foundations of Machine Learning(CS 4783/5783)

Lecture 3: Uniform Convergence, Symmetrization and Rademacher Complexity

1 Empirical Risk Minimization and Uniform Convergence

Recall from the previous lecture that the ERM algorithm is given by:

$$\hat{f}_{\text{ERM}} \in \arg\min_{f \in \mathcal{F}} \widehat{L}_S(f)$$

That is, find that model in $\mathcal{F}$ that has the smallest training loss. When $\mathcal{F}$ is a very large/complicated set of models, the ERM algorithm can easily fail as it would overfit on the training sample. In the next few lectures, we will try to analyze when this algorithm works well and what "complexity" measure on $\mathcal{F}$ governs how well the ERM performs. We already saw that for finite set of models, we can get a bound that depended only logarithmically on the size of $\mathcal{F}$. How about infinite model sets?

Towards answering this question, we will introduce the tool of uniform convergence. We already saw that, for any $t > 0$,

$$P\left(L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) > 2t\right) \leq P\left(\max_{f \in \mathcal{F}} \left|\widehat{L}_S(f) - L_{\mathbf{D}}(f)\right| > t\right) \tag{1}$$

We will use this observation to provide a concrete upper bound.

Before we proceed we will first write down the statement of the so called McDiarmid's inequality. While its not true that any arbitrary function of n independent random variables concentrates well near its expectation, the McDiarmid's inequality (bounded difference inequality) shows that functions that dont change too much when only one of its n arguments is changed, do concentrate well. Specifically, McDiarmid's inequality theorem is the following.

Theorem 1. *Assume that $\Phi : \mathcal{Z}^n \mapsto \mathbb{R}$ is a function satisfying the condition that: For any $i \in [n]$, and any $z_1, \ldots, z_n \in \mathcal{Z}$ and any $z_i' \in \mathcal{Z}$,*

$$\left|\phi(z_1, \ldots, z_i, \ldots, z_n) - \phi(z_1, \ldots, z_i', \ldots, z_n)\right| \leq \frac{C}{n} \tag{2}$$

Then we have the following concentration result :

$$P\left(|\phi(Z_1, \ldots, Z_n) - \mathbb{E}\left[\phi(Z_1, \ldots, Z_n)\right]| > \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{C^2}\right)$$

where $Z_1, \ldots, Z_n$ are drawn iid from some fixed distribution.

Lemma 2. *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \leq 2\mathbb{E}\left[\max_{f \in \mathcal{F}} \left|\widehat{L}_S(f) - L_{\mathbf{D}}(f)\right|\right] + 2\sqrt{\frac{2\log(2/\delta)}{n}}$$

Proof. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We will apply McDiarmid theorem to the function

$$\phi((x_1, y_1), \ldots, (x_n, y_n)) = \max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right|$$

and as long as losses are bounded by 1, the condition on ϕ given in Eq. 2 is satisfied for $C = 2$. Why?

Well, to see this, let $S^{(i)} = \{(x_1, y_1), \ldots, (x_i', y_i'), \ldots, (x_n, y_n)\}$. That is the sample S where only the i'th sample point is switched. In this case:

$$\begin{aligned}
&\phi((x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)) - \phi((x_1, y_1), \ldots, (x_i', y_i'), \ldots, (x_n, y_n)) \\
&\quad = \max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right| - \max_{f \in \mathcal{F}} \left| \widehat{L}_{S'}(f) - L_{\mathbf{D}}(f) \right|
\end{aligned}$$

max minus max is upper bounded by a single max.

$$\begin{aligned}
&\leq \max_{f \in \mathcal{F}} \left\{ \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right| - \left| \widehat{L}_{S'}(f) - L_{\mathbf{D}}(f) \right| \right\} \\
&\leq \max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - \widehat{L}_{S'}(f) \right| \\
&= \frac{1}{n} \max_{f \in \mathcal{F}} \left| \ell(f(x_i), y_i) - \ell(f(x_i'), y_i') \right| = \frac{2}{n}
\end{aligned}$$

The last line is due to the fact that other than index i, the remaining indices will cancel out. Now, using McDiarmid's inequality for $\max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right|$, we can conclude that, for any $\epsilon > 0$,

$$P\left(\max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right| > \mathbb{E}\left[\max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right| \right] + \epsilon \right) \leq 2 \exp\left(-\frac{n\epsilon^2}{2} \right)$$

Since, for any $t > 0$,

$$P\left(L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) > 2t \right) \leq P\left(\max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right| > t \right),$$

picking $t = \mathbb{E}\left[\max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right| \right] + \epsilon$, we can conclude that:

$$P\left(L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) > 2\mathbb{E}\left[\max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right| \right] + 2\epsilon \right) \leq 2 \exp\left(-\frac{n\epsilon^2}{2} \right)$$

Setting RHS to δ, we can conclude that:

$$P\left(L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) > 2\mathbb{E}\left[\max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right| \right] + 2\sqrt{\frac{2\log(2/\delta)}{n}} \right) \leq \delta$$

In other words, we have that with probability at least $1 - \delta$,

$$L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \leq 2\mathbb{E}\left[\max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right| \right] + 2\sqrt{\frac{2\log(2/\delta)}{n}}$$

This ends the proof. □

Thus one can view the uniform convergence term $\mathbb{E}\left[\max_{f \in \mathcal{F}} \left| \widehat{L}_S(f) - L_{\mathbf{D}}(f) \right| \right]$ as a complexity measure that measures how complex the class of models $\mathcal{F}$ is.

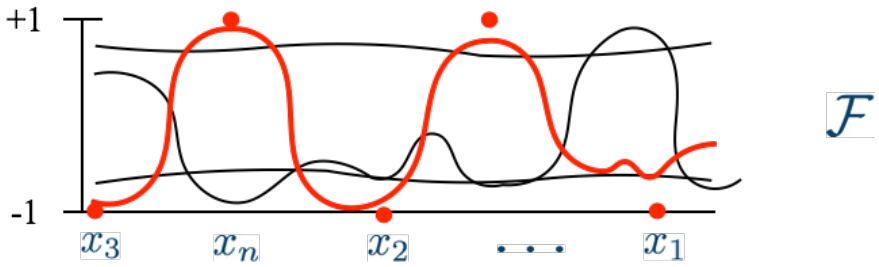
2 Symmetrization and Rademacher Complexity

For any class $\mathcal{F}$,

$$\begin{aligned}
\mathbb{E}_S\left[\max_{f\in\mathcal{F}}\left|L_{\mathbf{D}}(f)-\widehat{L}_S(f)\right|\right] &= \mathbb{E}_S\left[\max_{f\in\mathcal{F}}\left|\mathbb{E}_{S'}\left[\widehat{L}_{S'}(f)\right]-\widehat{L}_S(f)\right|\right] \\
&\le \mathbb{E}_{S,S'}\left[\max_{f\in\mathcal{F}}\left|\widehat{L}_{S'}(f)-\widehat{L}_S(f)\right|\right] \\
&= \mathbb{E}_{S,S'}\left[\max_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}(\ell(f(x'_t),y'_t)-\ell(f(x_t),y_t))\right|\right] \\
&= \mathbb{E}_{S,S'}\mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(\ell(f(x'_t),y'_t)-\ell(f(x_t),y_t))\right|\right] \\
&\le 2\ \mathbb{E}_S\mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right|\right]
\end{aligned}$$

Where in the above each ϵ_t is a Rademacher random variable that is $+1$ with probability $1/2$ and -1 with probability $1/2$. The above is called Rademacher complexity of the loss class $\ell\circ\mathcal{F}$. In general Rademacher complexity of a function class measures how well the function class correlates with random signs. The more it can correlate with random signs the more complex the class is.

$$\text{Example}: \ \mathcal{X}=[0,1],\ \mathcal{Y}=[-1,1]$$



Combining the above Rademacher complexity bound with lemma 2 we get the following corollary.

Corollary 3. *For any class $\mathcal{F}$ and any loss bounded by 1, with probability at least $1-\delta$,*

$$L_{\mathbf{D}}(\hat{f}_{\mathrm{ERM}})-\min_{f\in\mathcal{F}}L_{\mathbf{D}}(f)\le 4\ \mathbb{E}_S\mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right|\right]+2\sqrt{\frac{2\log(2/\delta)}{n}}$$

3 Why Does Symmetrization Help?

The main idea is that once we have introduced the Rademacher variables $\epsilon_1,\ldots,\epsilon_n$, we can look at the Rademacher complexity conditioned on sample S. Specifically, given a sample S, define

$$\mathcal{F}_{|x_1,\ldots,x_n}=\{(f(x_1),\ldots,f(x_n)):f\in\mathcal{F}\}$$

Now note that:

$$\mathbb{E}_S\mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right\}\right]=\mathbb{E}_S\mathbb{E}_\epsilon\left[\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right]$$

Thus we see that we need to bound $\mathbb{E}_\epsilon\left[\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right]$. Since the term $\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)$ has an expectation w.r.t. ϵ's of 0, and because this average converges to 0 with high probability (due to Hoeffding's inequality), using a union bound, we can obtain a bound on the above term that only depends logarithmically on $\left|\mathcal{F}_{|x_1,\ldots,x_n}\right|$. Thus, it is clear that only the cardinality of set $\mathcal{F}_{|x_1,\ldots,x_n}$ matters and not cardinality of all of $\mathcal{F}$. Why does this help?

Think about the threshold example, given n examples, the cardinality restricted to these samples is at most $n+1$. Why?
Well sort any given n points in ascending order, using thresholds, we can get at most n + 1 possible labeling on the n points. Hence for any $x_1,\ldots,x_n$, $|\mathcal{F}_{|x_1,\ldots,x_n}|\leq n+1$

Lemma 4. *For any class $\mathcal{F}$ and any loss bounded by* 1*,*

$$\mathbb{E}_\epsilon\left[\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right|\right]\leq O\left(\sqrt{\frac{\log\left|\mathcal{F}_{|x_1,\ldots,x_n}\right|}{n}}\right)$$

Proof. Since $\mathbb{E}_\epsilon\left[\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right]=0$, using Hoeffding's inequaltity, for any $\epsilon>0$,

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right|>\epsilon\right)\leq 2\exp\left(-\frac{n\epsilon^2}{2}\right)$$

Hence using union bound,

$$P\left(\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right|>\epsilon\right)\leq 2|\mathcal{F}_{|x_1,\ldots,x_n}|\exp\left(-\frac{n\epsilon^2}{2}\right)$$

Now using the fact that for a non-negative RV X, $\mathbb{E}[X]=\int_0^\infty P(X>t)dt$, we get that:

$$\begin{aligned}
&\mathbb{E}_\epsilon\left[\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right|\right]\\
&=\int_0^\infty P\left(\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right|>\epsilon\right)d\epsilon\\
&\leq\sqrt{\frac{2\log(2|\mathcal{F}_{|x_1,\ldots,x_n}|)}{n}}+\int_{\sqrt{\frac{2\log(2|\mathcal{F}_{|x_1,\ldots,x_n}|)}{n}}}^\infty P\left(\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right|>\epsilon\right)d\epsilon\\
&\leq\sqrt{\frac{2\log(2|\mathcal{F}_{|x_1,\ldots,x_n}|)}{n}}+2|\mathcal{F}_{|x_1,\ldots,x_n}|\int_{\sqrt{\frac{2\log(2|\mathcal{F}_{|x_1,\ldots,x_n}|)}{n}}}^\infty\exp\left(-\frac{n\epsilon^2}{2}\right)d\epsilon\\
&=\sqrt{\frac{2\log(2|\mathcal{F}_{|x_1,\ldots,x_n}|)}{n}}+\frac{2\sqrt{2}|\mathcal{F}_{|x_1,\ldots,x_n}|}{\sqrt{n}}\int_{\sqrt{\log(2|\mathcal{F}_{|x_1,\ldots,x_n}|)}}^\infty\exp\left(-x^2\right)dx
\end{aligned}$$

Using upper bound on complementary error function give by $\int_x^\infty e^{-u^2}du \leq e^{-x^2}/2x$ to conclude that:

$$\leq \sqrt{\frac{2\log(2|\mathcal{F}_{|x_1,\ldots,x_n}|)}{n}} + \frac{2\sqrt{2}|\mathcal{F}_{|x_1,\ldots,x_n}|}{\sqrt{n}} \frac{1}{4\sqrt{\log(2|\mathcal{F}_{|x_1,\ldots,x_n}|)}|\mathcal{F}_{|x_1,\ldots,x_n}|}$$

$$= \sqrt{\frac{2\log(2|\mathcal{F}_{|x_1,\ldots,x_n}|)}{n}} + \frac{1}{\sqrt{2n\log(2|\mathcal{F}_{|x_1,\ldots,x_n}|)}}$$

Hence we conclude that:

$$\mathbb{E}_\epsilon \left[\max_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}} \left| \frac{1}{n}\sum_{t=1}^n \epsilon_t \ell(\mathbf{f}[t], y_t) \right| \right] \leq O\left(\sqrt{\frac{\log|\mathcal{F}_{|x_1,\ldots,x_n}|}{n}} \right)$$

□

Using the above lemma with Corollary 3, we conclude that:

Corollary 5. *For any class $\mathcal{F}$ and any loss bounded by 1, for any $\delta > 0$, with probability at least $1-\delta$,*

$$L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \min_{f\in\mathcal{F}} L_{\mathbf{D}}(f) \leq O\left(\sqrt{\frac{\log|\mathcal{F}_{|x_1,\ldots,x_n}|}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right)$$

Example : thresholds
What does $\mathcal{F}_{|x_1,\ldots,x_n}$ for the class of threshold function look like ?
Well sort any given n points in ascending order, using thresholds, we can get at most $n+1$ possible labeling on the n points. Hence $\mathcal{F}_{|x_1,\ldots,x_n} \leq n+1$. From this we conclude that for the learning thresholds problem, for any $\delta > 0$, with probability at least $1-\delta$,

$$L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \min_{f\in\mathcal{F}} L_{\mathbf{D}}(f) \leq O\left(\sqrt{\frac{\log(n/\delta)}{n}} \right)$$

A note for the curious: In the above we used the fact that $\int_x^\infty e^{-u^2}du \leq e^{-x^2}/2x$. If you haven't seen this fact before and are curious how it was proven. Here is the short proof.

First, note that for any $u \geq x$, $u/x > 1$ and so

$$\int_x^\infty e^{-u^2}du \leq \int_x^\infty \frac{u}{x}e^{-u^2}du = \frac{1}{2x}\int_x^\infty 2ue^{-u^2}du = \frac{1}{2x}\int_x^\infty -\frac{d}{du}e^{-u^2}du = \frac{e^{-x^2}}{2x}$$