Machine Learning Theory (CS 6783)

Lecture 21: General Statistical Learning and Algorithmic Stabiltity

1 Recap of Statistical Learning

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and consider the learning algorithm $\hat{y} : \bigcup_{t=1}^{\infty} \mathcal{Z}^t \mapsto \mathcal{Y}^{\mathcal{X}}$. We assume that sample $S = \{z_1, \ldots, z_n\}$ drawn iid from fixed distribution $\mathcal{D}$ over $\mathcal{Z}$. We are interested in algorithms that guarantee that:

$$\mathbb{E}_S \left[L(\hat{y}(S)) \right] - \inf_{f \in \mathcal{F}} L(f) \leq \epsilon_{cons}(n)$$

where for any $h \in \mathcal{Y}^{\mathcal{X}}$, $L(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$ and $\epsilon_{cons}(n)$ goes to 0. We will consider a problem learnable if there exists a uniform rate $\epsilon_{cons}(n)$ that goes to 0 with n and a learning algorithm such that attains this upper bound.

2 Stability of a Learning Algorithm

Informally, an algorithm is said to be stable if replacing one sample from the training set does not change outcome by much.

For a sample S, Let $S^{(i)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$ be obtained by replacing the i'th sample by another $z_i' \in \mathcal{Z}$.

Definition 1. *A learning algorithm $\hat{y}$ is said to be Uniform Replace One (URO) stable with rate ϵ_{stable} if for any sample S and any $S^{(t)}$'s, and any $z_1'', \ldots, z_n''$,*

$$\frac{1}{n} \sum_{t=1}^{n} \left| \ell(\hat{y}(S), z_t'') - \ell(\hat{y}(S^{(t)}), z_t'') \right| \leq \epsilon_{\text{stable}}(n)$$

A key observation we make now is that if $\hat{y}$ is URO stable with rate ϵ_{stable} then,

$$\left| \frac{1}{n} \sum_{t=1}^{n} \mathbb{E} \left[\ell(\hat{y}(S), z_t') - \ell(\hat{y}(S^{(t)}), z_t') \right] \right| \leq \epsilon_{\text{stable}}(n) \tag{1}$$

which is got by simply replacing z_t'';s by z_t''s and then using Jensen.

3 Stability + AERM implies Learnability

3.1 Stability Implies Generalization

A simple argument shows us that stable algorithms generalize well. That is:

Lemma 1. *If learning algorithm $\hat{y}$ is URO stable with rate ϵ_{stable} then,*

$$\left| \mathbb{E}_S \left[L(\hat{y}(S)) - \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}(S), z_t) \right] \right| \leq \epsilon_{\text{stable}}(n)$$

That is it generalizes. The vice-versa is also true.

Proof. Note that:

$$\begin{aligned} \mathbb{E}_S \left[\frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}(S), z_t) \right] &= \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_S \left[\ell(\hat{y}(S^{(t)}), z_t) \right] \\ &= \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_S \left[\ell(\hat{y}(S^{(t)}), z'_t) \right] \end{aligned}$$

Also note that

$$\mathbb{E}_S \left[L(\hat{y}(S)) \right] = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{S, z'_1, \ldots, z'_n} \left[\ell(\hat{y}(S), z'_t) \right]$$

Hence we conclude that

$$\left| \mathbb{E}_S \left[L(\hat{y}(S)) - \frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}(S), z_t) \right] \right| = \left| \frac{1}{n} \sum_{t=1}^{n} \mathbb{E} \left[\ell(\hat{y}(S), z'_t) - \ell(\hat{y}(S^{(t)}), z'_t) \right] \right|$$

□

3.2 A Stable Approximate ERM Algorithm has Low Excess Risk

An approximate ERM algorithm is one that returns a hypothesis whose empirical loss is close to that of ERM.

Specifically we will say that a learning rule $\hat{y}$ is an AERM with rate ϵ_{AERM} is for any n,

$$\mathbb{E}_S \left[\frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}(S), z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t) \right] \leq \epsilon_{\text{AERM}}(n)$$

If an algorithm is stable we already showed that it generalizes. If further we have that is is an AERM then notice that

$$\begin{aligned} \mathbb{E}_S \left[L(\hat{y}(S)) \right] - \inf_{f \in \mathcal{F}} L(f) &= \mathbb{E}_S \left[L(\hat{y}(S)) \right] - \inf_{f \in \mathcal{F}} \mathbb{E}_S \left[\frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t) \right] \\ &\leq \mathbb{E}_S \left[L(\hat{y}(S)) \right] - \mathbb{E}_S \left[\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t) \right] \\ &\leq \mathbb{E}_S \left[L(\hat{y}(S)) \right] - \mathbb{E}_S \left[\frac{1}{n} \sum_{t=1}^{n} \ell(\hat{y}(S), z_t) \right] + \epsilon_{\text{AERM}}(n) \\ &\leq \epsilon_{\text{stable}}(n) + \epsilon_{\text{AERM}}(n) \end{aligned}$$

Thus we have shown that any learning algorithm that is stable and minimizes training error approximately, learns. In the following section we will show that the converse is also true.

3.3 Characterization of Learnability Using Stability

We will now show that the converse is also true. That is, if the problem is learnable, or in other words, if there is an algorithm $\hat{y}$ such that for any distribution $\mathcal{D}$, if we have that

$$\mathbb{E}_S\left[L(\hat{y}(S))\right] - \inf_{f\in\mathcal{F}} L(f) \leq \epsilon_{\text{rate}}(n)$$

then there has to be an algorithm that is both an AERM and is uniform RO stable. Specifically we will prove the following theorem.

Theorem 2. *If there exists an algorithm $\hat{y}$ such that for any distribution $\mathcal{D}$, for sample drawn form this distribution:*

$$\mathbb{E}_S\left[L(\hat{y}(S))\right] - \inf_{f\in\mathcal{F}} L(f) \leq \epsilon_{\text{rate}}(n)$$

then, there exists an algorithm $\hat{\hat{y}}$ s.t.

1. *$\hat{\hat{y}}$ is $\epsilon_{\text{stable}}(n) = \frac{2}{\sqrt{n}}$ URO stable*
2. *$\hat{\hat{y}}$ is an AERM with rate $\epsilon_{\text{AERM}}(n) = 2\epsilon_{\text{rate}}(n^{1/4}) + O\left(\frac{1}{\sqrt{n}}\right)$*

To prove this theorem we first introduce the following two lemma's.

Lemma 3. *If there exists an algorithm $\hat{y}$ such that for any distribution $\mathcal{D}$, for sample drawn form this distribution:*

$$\mathbb{E}_S\left[L(\hat{y}(S))\right] - \inf_{f\in\mathcal{F}} L(f) \leq \epsilon_{\text{rate}}(n)$$

then, there exists an algorithm $\hat{\hat{y}}$ s.t.

1. *$\hat{\hat{y}}$ is $2/\sqrt{n}$ URO stable*
2. *$\hat{\hat{y}}$ is such that for any distribution $\mathcal{D}$, for sample drawn form this distribution:*

$$\mathbb{E}_S\left[L(\hat{\hat{y}}(S))\right] - \inf_{f\in\mathcal{F}} L(f) \leq \epsilon_{\text{rate}}(\sqrt{n})$$

3. *Further, we have that*

$$\mathbb{E}_S\left[\left|L(\hat{\hat{y}}(S)) - \frac{1}{n}\sum_{t=1}^{n}\ell(\hat{\hat{y}}(S), z_t)\right|\right] \leq O\left(1/\sqrt{n}\right)$$

Proof. $\hat{\hat{y}}(S)$ simply runs $\hat{y}$ on just the first $\sqrt{n}$ samples. That is, if $S = \{z_1, \ldots, z_n\}$ is the original sample then let $\tilde{S} = \{z_1, \ldots, z_{\sqrt{n}}\}$. Then,

$$\hat{\hat{y}}(S) = \hat{y}(\tilde{S})$$

Now note that by our premise on $\hat{y}$, since $\hat{\hat{y}}$ runs $\hat{y}$ on first $\sqrt{n}$ samples,

$$\mathbb{E}_S\left[L(\hat{\hat{y}}(S))\right] - \inf_{f\in\mathcal{F}} L(f) \leq \epsilon_{\text{rate}}(\sqrt{n})$$

as stated. Further, note that for any $t > \sqrt{n}$,

$$\hat{y}(S) = \hat{y}(S^{(t)}) = \hat{y}(\tilde{S})$$

since $\tilde{S}$ is only the first $\sqrt{n}$ samples. Hence,

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\left|\ell(\hat{y}(S), z_t'') - \ell(\hat{y}(S^{(t)}), z_t'')\right| &= \frac{1}{n}\sum_{t=1}^{\sqrt{n}}\left|\ell(\hat{y}(S), z_t'') - \ell(\hat{y}(S^{(t)}), z_t'')\right| \\ &\leq \frac{2\sqrt{n}}{n} = \frac{2}{\sqrt{n}}\end{aligned}$$

Hence we conclude the second statement. The final statement can follow from either the fact that most of the samples in S are not used by algorithm and hence its error on those samples are unbiased estimate from $\mathcal{D}$ or we can simply use the result we proved earlier that stability which we already have implies generalization and hence the result. Hence we conclude the proof. □

Lemma 4. *If there exists an algorithm $\hat{y}$ such that for any distribution $\mathcal{D}$, for sample drawn form this distribution:*

$$\mathbb{E}_S\left[L(\hat{y}(S))\right] - \inf_{f\in\mathcal{F}} L(f) \leq \epsilon_{\text{rate}}(n),$$

then it is true that for any distribution $\mathcal{D}$,

$$\mathbb{E}_{S\sim\mathcal{D}^n}\left[\left|\inf_{f\in\mathcal{F}} L_{\mathcal{D}}(f) - \frac{1}{n}\sum_{t=1}^{n}\ell(f, z_t)\right|\right] \leq 2\epsilon_{\text{rate}}(n^{1/4}) + O\left(\frac{1}{\sqrt{n}}\right)$$

Proof. Let $\sigma_1, \ldots, \sigma_{n^{1/4}}$ be drawn iid from distribution $\text{Unif}\{[n]\}$. Define sample $S' = \{z_{\sigma_1}, \ldots, z_{\sigma_{n^{1/4}}}\}$ of size $n^{1/4}$. Now note that

$$P\left[\exists i, j \text{ s.t. } \sigma_i = \sigma_j\right] \leq \frac{\sum_{i=1}^{n^{1/4}}(i-1)}{n} \leq \frac{n^{2/4}}{n} = \frac{1}{\sqrt{n}}$$

Hence, we have that:

$$\begin{aligned}\mathbb{E}\left[L_{\mathcal{D}}(\hat{y}(S'))\right] - \inf_{f\in\mathcal{F}} L_{\mathcal{D}}(f) &= \mathbb{E}_{\sigma_1,\ldots,\sigma_{n^{1/4}}}\left[\mathbb{E}_S\left[L_{\mathcal{D}}(\hat{y}(S'))\right]\right] - \inf_{f\in\mathcal{F}} L_{\mathcal{D}}(f) \\ &= \mathbb{E}_{\sigma_1,\ldots,\sigma_{n^{1/4}}}\left[\mathbb{E}_S\left[L(\hat{y}(S'))\right] | \exists i, j \text{ s.t. } \sigma_i = \sigma_j\right] P\left[\exists i, j \text{ s.t. } \sigma_i = \sigma_j\right] \\ &\quad + \mathbb{E}_{\sigma_1,\ldots,\sigma_{n^{1/4}}}\left[\mathbb{E}_S\left[L_{\mathcal{D}}(\hat{y}(S'))\right] | \forall i, j, \sigma_i \neq \sigma_j\right] P\left[\forall i, j, \sigma_i \neq \sigma_j\right] - \inf_{f\in\mathcal{F}} L_{\mathcal{D}}(f) \\ &\leq \frac{1}{\sqrt{n}} + \mathbb{E}_{\sigma_1,\ldots,\sigma_{n^{1/4}}}\left[\mathbb{E}_S\left[L_{\mathcal{D}}(\hat{y}(S'))\right] | \forall i, j, \sigma_i \neq \sigma_j\right] - \inf_{f\in\mathcal{F}} L_{\mathcal{D}}(f)\end{aligned}$$

Now note that if there is no pair i, j such that $\sigma_i = \sigma_j$, then we can conclude that S' is basically a sample set drawn iid from $\mathcal{D}$ of size $n^{1/4}$. Hence, using the premise about $\hat{y}$, we conclude that under this conditioning,

$$\mathbb{E}_{\sigma_1,\ldots,\sigma_{n^{1/4}}}\left[\mathbb{E}_S\left[L_{\mathcal{D}}(\hat{y}(S'))\right] | \forall i, j, \sigma_i \neq \sigma_j\right] - \inf_{f\in\mathcal{F}} L_{\mathcal{D}}(f) \leq \epsilon_{\text{rate}}(n^{1/4})$$

Hence we conclude that:

$$\mathbb{E}\left[L_{\mathcal{D}}(\hat{y}(S'))\right] - \inf_{f\in\mathcal{F}} L_{\mathcal{D}}(f) \leq \epsilon_{\text{rate}}(n^{1/4}) + \frac{1}{\sqrt{n}} \tag{2}$$

On the other hand, given sample S, consider the distribution S' is drawn from. Note that S' is a sample set of size $n^{1/4}$ that is drawn iid from distribution $\text{Unif}\{z_1, \ldots, z_n\}$. Lets call this distribution $\hat{\mathcal{D}}(S)$. Note that, conditioned on sample S, the test loss w.r.t. this distribution is

$$L_{\hat{\mathcal{D}}(S)}(f) = \frac{1}{n}\sum_{t=1}^{n} \ell(f, z_t)$$

That is, conditioned on a sample S, S' is drawn from a distribution whose test loss is exactly the empirical average over S or in other words the training loss w.r.t. S. Now from our premise, conditioned on S, our learning algorithm when applied on S' should yield an excess risk guarantee of the form

$$\mathbb{E}_{S'}\left[L_{\hat{\mathcal{D}}(S)}(\hat{y}(S'))|S\right] - \inf_{f\in\mathcal{F}} L_{\hat{\mathcal{D}}(S)}(f) \leq \epsilon_{\text{rate}}(n^{1/4})$$

But since $L_{\hat{\mathcal{D}}(S)}(f) = \frac{1}{n}\sum_{t=1}^{n} \ell(f, z_t)$ we have,

$$\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{t=1}^{n} \ell(\hat{y}(S'), z_t)|S\right] - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(f, z_t) \leq \epsilon_{\text{rate}}(n^{1/4})$$

Hence taking expectation over S we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \ell(\hat{y}(S'), z_t)\right] - \mathbb{E}_S\left[\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(f, z_t)\right] \leq \epsilon_{\text{rate}}(n^{1/4}) \tag{3}$$

Now we are almost ready to combine Eq. 2 and 3. To do so we note that for any S' of size $n^{1/4}$,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{t=1}^{n}\ell(\hat{y}(S'),z_t)-\frac{1}{|S\setminus S'|}\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)\right| \\
&\qquad=\left|\frac{1}{n}\left(\sum_{t\in S'}\ell(\hat{y}(S'),z_t)+\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)\right)-\frac{1}{|S\setminus S'|}\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)\right| \\
&\qquad\le\frac{1}{n}\left|\sum_{t\in(S\setminus S')^c}\ell(\hat{y}(S'),z_t)\right|+\left|\frac{1}{n}\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)-\frac{1}{|S\setminus S'|}\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)\right| \\
&\qquad\le\frac{1}{n}\left|\sum_{t\in(S\setminus S')^c}\ell(\hat{y}(S'),z_t)\right|+\left|\frac{1}{n}\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)-\frac{1}{|S\setminus S'|}\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)\right| \\
&\qquad\le\frac{n^{1/4}}{n}+\left|\frac{1}{n}\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)-\frac{1}{|S\setminus S'|}\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)\right| \\
&\qquad\le\frac{n^{1/4}}{n}+\left(1-\frac{|S\setminus S'|}{n}\right)\left|\frac{1}{|S\setminus S'|}\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)\right| \\
&\qquad\le\frac{n^{1/4}}{n}+\left(1-\frac{n-n^{1/4}}{n}\right)=\frac{2}{n^{3/4}} \qquad (4)
\end{aligned}$$

In the above, we have used the fact that losses are bounded by 1 and triangle inequality. However, note that $\hat{y}(S')$ only looks at samples in S' and so samples in $S\setminus S'$ are fresh samples drawn from distribution $\mathcal{D}$ not used by $\hat{y}$. Hence,

$$\mathbb{E}_S\left[\left|L_{\mathcal{D}}(\hat{y}(S'))-\frac{1}{|S\setminus S'|}\sum_{t\in S\setminus S'}\ell(\hat{y}(S'),z_t)\right|\right]\le\frac{1}{\sqrt{S\setminus S'}}\le\frac{1}{\sqrt{n-n^{1/4}}} \qquad (5)$$

The above is because if we condition on $\sigma_1,\ldots,\sigma_{n^{1/4}}$ then $S\setminus S'$ can be seen as a sample drawn fresh from $\mathcal{D}$. And for any bounded random variables X_i's drawn iid from a fixed distribution,

$$\mathbb{E}\left[\left|\mathbb{E}[X]-\frac{1}{m}\sum_{i=1}^{m}X_i\right|\right]=\frac{1}{m}\mathbb{E}\left[\left|\sum_{i=1}^{m}(X_i-\mathbb{E}[X])\right|\right]\le\frac{1}{m}\sqrt{\sum_{i=1}^{m}\mathbb{E}\left[(X_i-\mathbb{E}[X])^2\right]}\le\frac{1}{\sqrt{m}}$$

Hence combining Eq. 4 and Eq. 5 we conclude that:

$$\mathbb{E}_S\left[\left|L_{\mathcal{D}}(\hat{y}(S'))-\frac{1}{n}\sum_{t=1}^{n}\ell(\hat{y}(S'),z_t)\right|\right]\le\frac{2}{n^{3/4}}+\frac{1}{\sqrt{n-n^{1/4}}}$$

Using this we can combine Eqns. 2 and 3 and obtain that:

$$\mathbb{E}_{S\sim\mathcal{D}^n}\left[\left|\inf_{f\in\mathcal{F}}L_{\mathcal{D}}(f)-\frac{1}{n}\sum_{t=1}^{n}\ell(f,z_t)\right|\right]\le 2\epsilon_{\text{rate}}(n^{1/4})+\frac{1}{\sqrt{n}}+\frac{2}{n^{3/4}}+\frac{1}{\sqrt{n-n^{1/4}}}$$

□

Finally we are ready to prove Theorem 2.

Proof of Theorem 2. First from Lemma 3 we can conclude that $\hat{\hat{y}}$ described in the Lemma is URO stable with the prescribed rate. Now also note that by the same lemma, the prescribed algorithm generalizes, that is

$$\mathbb{E}_S\left[L(\hat{\hat{y}}(S)) - \frac{1}{n}\sum_{t=1}^{n}\ell(\hat{\hat{y}}(S), z_t)\right] \leq \frac{1}{\sqrt{n}}$$

But then, from previous lemma we have that

$$\inf_{f\in\mathcal{F}} L(f) - \mathbb{E}_S\left[\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f, z_t)\right] \leq 2\epsilon_{\text{rate}}(n^{1/4}) + O\left(\frac{1}{\sqrt{n}}\right)$$

Hence we can conclude that

$$\mathbb{E}_S\left[\frac{1}{n}\sum_{t=1}^{n}\ell(\hat{\hat{y}}(S), z_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f, z_t)\right] \leq 2\epsilon_{\text{rate}}(n^{1/4}) + O\left(\frac{1}{\sqrt{n}}\right)$$

which concludes the proof. □