Mathematical Foundations of Machine Learning(CS 4783/5783)

Lecture 8: Algorithmic Stability and Statistical Learning

1 Algorithmic Stability

Before we talk about stability of a learning algorithm, we need to give a notation for a learning algorithm. Specifically, we define an algorithm $\mathbf{A}$ by a mapping of form $\mathbf{A} : \bigcup_{t=1}^{\infty}(\mathcal{X} \times \mathcal{Y})^t \mapsto \mathcal{Y}^{\mathcal{X}}$. That is a function that takes as input sample (in $\mathcal{X} \times \mathcal{Y}$) of arbitrary length and maps it to a model that maps input instances in $\mathcal{X}$ to outcome $\mathcal{Y}$. In other words, a learning algorithm takes a sample of any length and outputs a model (a model that predicts outcome in $\mathcal{Y}$ given an input in $\mathcal{X}$). Indeed, an algorithm like ERM takes a sample and returns as output a model that minimizes training error on given sample. Now with this definition of an algorithm, we are ready to talk about algorithmic stability.

Informally, an algorithm is said to be stable if deleting a sample from the training set does not change outcome by much. To define such stability, let us first introduce some notation. Given a sample S, let $S^{\backslash i}$ denote the sample got by deleting the i'th sample point from S. That is, $S^{\backslash i} = \{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}$.

Definition 1. *An algorithm $\mathbf{A}$ is said to be stable w.r.t. a distribution $\mathbf{D}$ with rate ϵ_{stable} if:*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S\left[\left|\ell(\mathbf{A}(S)(x_i), y_i) - \ell(\mathbf{A}(S^{\backslash i})(x_i), y_i)\right|\right] \le \epsilon_{\text{stable}}(n)$$

Definition 2. *An algorithm $\mathbf{A}$ is said to be an Approximate ERM (AERM) w.r.t. a class of models $\mathcal{F}$ with rate ϵ_{ERM} if for any sample S of size n,*

$$\widehat{L}_S(\mathbf{A}(S)) \le \min_{f \in \mathcal{F}} \widehat{L}_S(f) + \epsilon_{\text{ERM}}(n)$$

If an algorithm is stable, its test loss and training loss are close (or in other words it generalizes well). If further, the algorithm is an approximate ERM (i.e it approximately minimizes training loss), then such an algorithm has low excess risk in expectation. The following theorem shows that a stable algorithm that is also an AERM, has a low expected excess risk.

Theorem 1. *If a learning algorithm is LOO stable with rate ϵ_{stable} and is also an AERM with rate ϵ_{ERM}, then we have the bound on expected excess risk,*

$$\mathbb{E}_S\left[L_{\mathbf{D}}(\mathbf{A}(S))\right] - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \le \epsilon_{\text{stable}}(n+1) + \epsilon_{\text{ERM}}(n+1)$$

Proof.

$$
\begin{aligned}
\epsilon_{\text{stable}}(n) &\geq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S\left[\left|\ell(\mathbf{A}(S)(x_i), y_i) - \ell(\mathbf{A}(S^{\setminus i})(x_i), y_i)\right|\right] \\
&\geq \mathbb{E}_S\left[\left|\frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{A}(S)(x_i), y_i) - \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{A}(S^{\setminus i})(x_i), y_i)\right|\right] \\
&= \mathbb{E}_S\left[\left|\widehat{L}_S(\mathbf{A}(S)) - \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{A}(S^{\setminus i})(x_i), y_i)\right|\right] \\
&\geq \left|\mathbb{E}_S\left[\widehat{L}_S(\mathbf{A}(S)) - \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{A}(S^{\setminus i})(x_i), y_i)\right]\right| \\
&= \left|\mathbb{E}_S\left[\widehat{L}_S(\mathbf{A}(S)) - \frac{1}{n}\sum_{i=1}^{n}L_{\mathbf{D}}(\mathbf{A}(S^{\setminus i}))\right]\right| \\
&= \left|\mathbb{E}_S\left[\widehat{L}_S(\mathbf{A}(S)) - L_{\mathbf{D}}(\mathbf{A}(S^{\setminus n}))\right]\right| \\
&\geq \mathbb{E}_S\left[L_{\mathbf{D}}(\mathbf{A}(S^{\setminus n})) - \widehat{L}_S(\mathbf{A}(S))\right] \\
&\geq \mathbb{E}_S\left[L_{\mathbf{D}}(\mathbf{A}(S^{\setminus n})) - \min_{f\in\mathcal{F}}\widehat{L}_S(f)\right] - \epsilon_{\text{ERM}}(n) \\
&= \mathbb{E}_S\left[L_{\mathbf{D}}(\mathbf{A}(S^{\setminus n}))\right] - \mathbb{E}_S\left[\min_{f\in\mathcal{F}}\widehat{L}_S(f)\right] - \epsilon_{\text{ERM}}(n) \\
&\geq \mathbb{E}_S\left[L_{\mathbf{D}}(\mathbf{A}(S^{\setminus n}))\right] - \min_{f\in\mathcal{F}}\mathbb{E}_S\left[\widehat{L}_S(f)\right] - \epsilon_{\text{ERM}}(n) \\
&= \mathbb{E}_S\left[L_{\mathbf{D}}(\mathbf{A}(S^{\setminus n}))\right] - \min_{f\in\mathcal{F}}L_{\mathbf{D}}(f) - \epsilon_{\text{ERM}}(n)
\end{aligned}
$$

Hence we have proved that

$$\mathbb{E}_S\left[L_{\mathbf{D}}(\mathbf{A}(S^{\setminus n}))\right] - \min_{f\in\mathcal{F}}L_{\mathbf{D}}(f) \leq \epsilon_{\text{stable}}(n) + \epsilon_{\text{ERM}}(n)$$

However, note that the above says that if we provide the algorithm with sample of size $n-1$ (ie. the last sample deleted), then in expectation we have the excess risk bound of $\epsilon_{\text{stable}}(n) + \epsilon_{\text{ERM}}(n)$. Since n is arbitrary, we can conclude that

$$\mathbb{E}_S\left[L_{\mathbf{D}}(\mathbf{A}(S))\right] - \min_{f\in\mathcal{F}}L_{\mathbf{D}}(f) \leq \epsilon_{\text{stable}}(n+1) + \epsilon_{\text{ERM}}(n+1)$$

□

Remark 1.1. *Note that a simple Markov inequality can convert an expected statement to one that holds with say probability* $1/2$*. From this, using the style of analysis you guys did in Assignment 1, Question 1, you can convert the statement into a high probability one.*

The below theorem shows that the converse is also true. That is, if a problem is learnable with some rate against all distributions, then there always exists a stable AERM.

Theorem 2. *If a learning algorithm* $\mathbf{A}$ *has the following expected excess risk guarantee for all distributions* $\mathbf{D}$

$$\mathbb{E}_S\left[L_{\mathbf{D}}(\mathbf{A}(S))\right] - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \le \epsilon_{\text{rate}}(n),$$

then there always exists a learning algorithm that is both stable and an AERM

We will not prove this statement formally. However, the high level idea for a proof sketch is the following.

1. Given a sample S of size n, instead of running the algorithm $\mathbf{A}$ on S directly, we instead draw n' samples by drawing uniformly with replacement from sample S.

2. Notice that the expected loss from this uniform distribution over S is exactly the training loss on all of S.

3. On the other hand, since we are drawing only n' samples from S, if n' is small compared to n, then with high probability, we never draw the same index/sample twice. Hence such a sample can be seen as sample drawn from distribution $\mathbf{D}$ directly.

4. Now combining the learning guarantee with the above two observations will show that the algorithm is an AERM.

5. Stability of this algorithm follows from the fact that the algorithm only depends on n' samples and so deleting $n - n'$ (ie. most samples) does not affect the outcome at all. Hence stability holds with rate n'/n.

2 Eg. Minimizing Convex loss + norm square regularizer

For this section, we consider the following style of learning algorithms:

$$\hat{\mathbf{w}}_S = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{t=1}^{n} \phi(\mathbf{w}; (x_t, y_t)) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \tag{1}$$

where $\phi : \mathbb{R}^d \times (\mathcal{X} \times \mathcal{Y}) \mapsto \mathbb{R}$ is convex in its first argument and is L-Lipschitz meaning that for any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$,

$$|\phi(\mathbf{w}; (x, y)) - \phi(\mathbf{w}'; (x, y))| \le L \|\mathbf{w} - \mathbf{w}'\|$$

As an example, if $\phi(\mathbf{w}; (x, y)) = \max\{0, 1 - y \cdot \mathbf{w}^\top x\}$, the hinge loss, then ϕ is convex in $\mathbf{w}$ and is L Lipschitz as long as $\|x\| \le L$

Proposition 3. *Define loss* $\ell(\mathbf{w}; (x, y)) = \phi(\mathbf{w}; (x, y)) + \frac{\lambda}{2}\|\mathbf{w}\|^2$. *For this loss, for any sample* S, *it is true that*

$$\hat{L}_S(\mathbf{w}) - \min_{\mathbf{w} \in \mathbb{R}^d} \hat{L}_S(\mathbf{w}) \ge \frac{\lambda}{2} \|\mathbf{w} - \hat{\mathbf{w}}_S\|^2$$

where $\hat{\mathbf{w}}_S = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^d} \hat{L}_S(\mathbf{w})$

Proof. Note that:

$$\hat{L}_S(\mathbf{w}) - \hat{L}_S(\hat{\mathbf{w}}_S) = \frac{1}{n}\sum_{t=1}^{n}\left(\phi(\mathbf{w};(x_t,y_t)) - \phi(\hat{\mathbf{w}}_S;(x_t,y_t))\right) + \frac{\lambda}{2}\left(\|\mathbf{w}\|^2 - \|\hat{\mathbf{w}}_S\|^2\right)$$

by convexity of ϕ we have that $\phi(\mathbf{w};(x_t,y_t)) - \phi(\hat{\mathbf{w}}_S;(x_t,y_t)) \geq \nabla\phi(\hat{\mathbf{w}}_S;(x_t,y_t))^\top(\mathbf{w} - \hat{\mathbf{w}}_S)$

$$\geq \left(\frac{1}{n}\sum_{t=1}^{n}\nabla\phi(\hat{\mathbf{w}}_S;(x_t,y_t))\right)^\top(\mathbf{w} - \hat{\mathbf{w}}_S) + \frac{\lambda}{2}\left(\|\mathbf{w}\|^2 - \|\hat{\mathbf{w}}_S\|^2\right)$$

Notice that since $\hat{\mathbf{w}}_S$ is the minimizer of $\widehat{L}_S$, $\nabla\widehat{L}_S(\hat{\mathbf{w}}_S) = 0$ and so, $-\lambda\hat{\mathbf{w}}_S = \frac{1}{n}\sum_{t=1}^{n}\nabla\phi(\hat{\mathbf{w}}_S;(x_t,y_t))$.

$$\begin{aligned}
&= (-\lambda\hat{\mathbf{w}}_S)^\top(\mathbf{w} - \hat{\mathbf{w}}_S) + \frac{\lambda}{2}\left(\|\mathbf{w}\|^2 - \|\hat{\mathbf{w}}_S\|^2\right)\\
&= -\lambda\hat{\mathbf{w}}_S^\top\mathbf{w} + \lambda\|\hat{\mathbf{w}}_S\|^2 + \frac{\lambda}{2}\left(\|\mathbf{w}\|^2 - \|\hat{\mathbf{w}}_S\|^2\right)\\
&= -\lambda\hat{\mathbf{w}}_S^\top\mathbf{w} + \frac{\lambda}{2}\|\hat{\mathbf{w}}_S\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2\\
&= \frac{\lambda}{2}\|\hat{\mathbf{w}}_S - \mathbf{w}\|^2
\end{aligned}$$

□

Theorem 4. *Define loss $\ell(\mathbf{w};(x,y)) = \phi(\mathbf{w};(x,y)) + \frac{\lambda}{2}\|\mathbf{w}\|^2$. Then, we have the following LOO stability rate for the algorithm described in Equation 1*

$$\mathbb{E}_S\left[|\ell(\hat{\mathbf{w}}_{S^{\setminus t}},(x_t,y_t)) - \ell(\hat{\mathbf{w}}_S,(x_t,y_t))|\right] \leq \frac{4L^2}{\lambda n}$$

That is, the ERM algorithm is stable in expectation.

Proof.

$$\begin{aligned}
\hat{L}_S(\hat{\mathbf{w}}_{S^{\setminus t}}) - \hat{L}_S(\hat{\mathbf{w}}_S) &= \frac{1}{n}\left(\ell(\hat{\mathbf{w}}_{S^{\setminus t}},(x_t,y_t)) - \ell(\hat{\mathbf{w}}_S,(x_t,y_t))\right) + \frac{1}{n}\sum_{s\in[n]\setminus\{t\}}\left(\ell(\hat{\mathbf{w}}_{S^{\setminus t}},z_s) - \ell(\hat{\mathbf{w}}_S,z_s)\right)\\
&= \frac{1}{n}\left(\ell(\hat{\mathbf{w}}_{S^{\setminus t}},(x_t,y_t)) - \ell(\hat{\mathbf{w}}_S,(x_t,y_t))\right) + \frac{n-1}{n}\left(\hat{L}_{S^{\setminus t}}(\hat{\mathbf{w}}_{S^{\setminus t}}) - \hat{L}_{S^{\setminus t}}(\hat{\mathbf{w}}_S)\right)\\
&\leq \frac{1}{n}\left(\ell(\hat{\mathbf{w}}_{S^{\setminus t}},(x_t,y_t)) - \ell(\hat{\mathbf{w}}_S,(x_t,y_t))\right)\\
&\leq \frac{2L}{n}\left\|\hat{\mathbf{w}}_{S^{\setminus t}} - \hat{\mathbf{w}}_S\right\|
\end{aligned}$$

On the other hand, using Proposition 3

$$\hat{L}_S(\mathbf{w}) - \min_{\mathbf{w}\in\mathbb{R}^d}\hat{L}_S(\mathbf{w}) \geq \frac{\lambda}{2}\left\|\mathbf{w} - \hat{\mathbf{w}}_S\right\|^2$$

Hence we conclude that

$$\|\hat{\mathbf{w}}_{S^{\setminus t}} - \hat{\mathbf{w}}_S\| \leq \frac{4L}{\lambda n}$$

Hence we conclude the stability rate that:

$$|\ell(\hat{\mathbf{w}}_{S^{\setminus t}}, (x_t, y_t)) - \ell(\hat{\mathbf{w}}_S, (x_t, y_t))| \leq \frac{4L^2}{\lambda n}$$

□

Now using the above along with theorem 1 we see that:

$$\mathbb{E}_S\left[L_{\mathbf{D}}(\hat{\mathbf{w}}_S)\right] - \min_{\mathbf{w}\in\mathbb{R}^d} L_{\mathbf{D}}(\mathbf{w}) \leq \frac{4L^2}{\lambda n}$$

Using the fact that $\ell(\mathbf{w};(x,y)) = \phi(\mathbf{w};(x,y)) + \frac{\lambda}{2}\|\mathbf{w}\|^2$, we can conclude that:

$$\begin{aligned}\mathbb{E}_S\left[\mathbb{E}_{(x,y)\sim\mathbf{D}}\left[\phi(\hat{\mathbf{w}}_S;(x,y))\right]\right] &\leq \min_{\mathbf{w}\in\mathbb{R}^d}\left\{\mathbb{E}_{(x,y)\sim\mathbf{D}}\left[\phi(\mathbf{w};(x,y))\right] + \frac{\lambda}{2}\|\mathbf{w}\|^2\right\} + \frac{4L^2}{\lambda n} - \frac{\lambda}{2}\|\hat{\mathbf{w}}_S\|^2 \\ &\leq \min_{\mathbf{w}\in\mathbb{R}^d}\left\{\mathbb{E}_{(x,y)\sim\mathbf{D}}\left[\phi(\mathbf{w};(x,y))\right] + \frac{\lambda}{2}\|\mathbf{w}\|^2\right\} + \frac{4L^2}{\lambda n}\end{aligned}$$

Hence we can conclude that for any $\mathbf{w}$,

$$\mathbb{E}_S\left[\mathbb{E}_{(x,y)\sim\mathbf{D}}\left[\phi(\hat{\mathbf{w}}_S;(x,y))\right]\right] - \mathbb{E}_{(x,y)\sim\mathbf{D}}\left[\phi(\mathbf{w};(x,y))\right] \leq \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{4L^2}{\lambda n}$$

That is we have a bound on excess risk in terms of loss ϕ.