Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Siddarth K
Instructor: Purushottam Kar
Date: February 7, 2018

LECTURE
12

Algorithmic Stability

Introduction

In this lecture, we begin the discussion on algorithmic stability. Till now we have only looked at the properties of the model class, and the results we have shown hold for any model in the model class. Here, we try to analyse a specific model from the model class.

Algorithmic Stability

We first begin with some definitions and notation of the terms.

Definition 12.1. An algorithm $\mathcal{A} : \bigcup_{n=1}^{\infty}(\mathcal{X}, \mathcal{Y})^n \rightarrow \mathcal{F}$ is a mapping from a set of training samples to a model.

Definition 12.2. Given $S = \{(\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^2, \mathbf{y}^2), \cdots, (\mathbf{x}^n, \mathbf{y}^n)\} \in (\mathcal{X}, \mathcal{Y})^n$, $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$, $i \in [n]$, define a perturbed training set $S^i_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})}$ as follows:

$$S^i_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})} = S \setminus \{(\mathbf{x}^i, \mathbf{y}^i)\} \bigcup \{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})\}$$

We abbreviate $S^i_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})}$ by S^i when it is obvious what the point we are swapping out is.

Definition 12.3. A learning algorithm $\mathcal{A}$ is said to be β-URL stable if $\forall S \in (\mathcal{X}, \mathcal{Y})^n, \forall (\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in (\mathcal{X}, \mathcal{Y}), \forall i \in [n]$

$$|l(\hat{f_S}(\mathbf{x}^{'}), \mathbf{y}^{'}) - l(\hat{f_{S^i}}(\mathbf{x}^{'}), \mathbf{y}^{\prime})| \leq \beta(n)$$

, where $\beta(.)$ is some function.

Definition 12.4. Let the generalisation error of the model $\hat{f_S}$ learnt on S be defined as

$$g(S) = |er^l_S \left[\hat{f_S}\right] - er^l_{\mathcal{D}} \left[\hat{f_S}\right]|$$

We would like to show that the generalisation error $g(S)$ does not get too large much of the time. More precisely, we would like to show that the probability that the generalisation error becomes large is very small.

Claim 12.1. $\mathop{\mathbb{P}}_{S} \left[g(S) \geq (n\beta(n) + B)\sqrt{\frac{\log(\frac{1}{\delta})}{n}} \right] \leq \delta$

Proof. We prove a slightly weaker version of the bound, which differs from the claim only by a constant factor.
Furthermore, we only show the proof for the right tail. The proof for the left tail can be derived very similarly.
Let the loss function $l \in [0, B]$ be bounded.
We would like to apply a concentration inequality like McDiarmid's to show this bound. To apply McDiarmid's, we need to show that $g(S)$ is stable.

Lemma 12.2. The function $g(S)$ is $\left(2\beta(n) + \frac{2B}{n}\right)$-stable

Proof. Firstly, note that by triangle inequality

$$g(S) = |er^l_S\left[\hat{f_S}\right] - er^l_{\mathcal{D}}\left[\hat{f_S}\right]| \leq |er^l_S\left[\hat{f_S}\right] - er^l_{S^i}\left[\hat{f_{S^i}}\right]| + |er^l_{\mathcal{D}}\left[\hat{f_S}\right] - er^l_{\mathcal{D}}\left[\hat{f_{s^i}}\right]|$$

$$|er^l_{\mathcal{D}}\left[\hat{f_S}\right] - er^l_{\mathcal{D}}\left[\hat{f_{s^i}}\right]| = \left|\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[l(\hat{f_S}(\mathbf{x}), \mathbf{y}) - l(\hat{f_{S^i}}(\mathbf{x}), \mathbf{y})\right]\right| \leq \beta(n) \tag{1}$$

$$\begin{aligned}
|er^l_S\left[\hat{f_S}\right] - er^l_{S^i}\left[\hat{f_{S^i}}\right]| &= \left|\frac{1}{n}\sum_{j\neq i}(l(\hat{f_S}(\mathbf{x}^j), \mathbf{y}^j) - l(\hat{f_{S^i}}(\mathbf{x}^j), \mathbf{y}^j)) + \frac{1}{n}(l(\hat{f_S}(\mathbf{x}^i, \mathbf{y}^i)) - l(\hat{f}_{S^i}(\mathbf{x}^i), \mathbf{y}^i))\right| \\
&\leq \left|\frac{n-1}{n}\beta(n) + \frac{1}{n}(l(\hat{f_S}(\mathbf{x}^i, \mathbf{y}^i)) - l(\hat{f}_{S^i}(\mathbf{x}^i), \mathbf{y}^i))\right| \\
&\leq \frac{n-1}{n}\beta(n) + \frac{2B}{n} \\
&\Rightarrow |g(S) - g(S^i)| \leq 2\beta(n) + \frac{2B}{n}
\end{aligned}$$

□

Since $g(S)$ is stable, we can apply McDiarmid's to get a concentration bound.

$$\mathop{\mathbb{P}}_{S}\left[g(S) > \mathbb{E}\left[g(S)\right] + \epsilon\right] \leq 2exp\left(-\frac{\epsilon^2}{2nc^2}\right)$$

We take $\delta = 2exp\left(-\frac{\epsilon^2}{2nc^2}\right)$. Solving for ϵ, we get $\epsilon = (n\beta(n) + B)\sqrt{\frac{8\log(\frac{2}{\delta})}{n}}$
Now, the only thing required to prove the result is to bound the value of $\mathbb{E}\left[g(S)\right]$.

$$\mathbb{E}\left[g(S)\right] = \mathbb{E}\left[\left|er^l_S\left[\hat{f_S}\right] - er^l_{\mathcal{D}}\left[\hat{f_S}\right]\right|\right]$$

We will show that both $\mathbb{E}\left[er^l_S\left[\hat{f_S}\right] - er^l_{\mathcal{D}}\left[\hat{f_S}\right]\right]$ and $\mathbb{E}\left[er^l_{\mathcal{D}}\left[\hat{f_S}\right] - er^l_S\left[\hat{f_S}\right]\right]$ are bounded by $\beta(n)$.

Lemma 12.3. $\mathbb{E}\left[g(S)\right] \leq \beta(n)$

Proof.

$$\begin{aligned}
\mathbb{E}\left[er_S^l\left[\hat{f}_S\right] - er_{\mathcal{D}}^l\left[\hat{f}_S\right]\right] &= \mathop{\mathbb{E}}_{S}\left[\frac{1}{n}\sum_i l(\hat{f}_S(\mathbf{x}^i), \mathbf{y}^i) - \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})}\left[l(\hat{f}_S(\mathbf{x}), \mathbf{y})\right]\right] \\
&= \frac{1}{n}\sum_i \mathop{\mathbb{E}}_{S}\left[l(\hat{f}_S(\mathbf{x}^i), \mathbf{y}^i) - \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})}\left[l(\hat{f}_S(\mathbf{x}), \mathbf{y})\right]\right] \\
&= \frac{1}{n}\sum_i \left(\mathop{\mathbb{E}}_{S}\left[l(\hat{f}_S(\mathbf{x}^i), \mathbf{y}^i)\right] - \mathop{\mathbb{E}}_{S,(\mathbf{x},\mathbf{y})}\left[l(\hat{f}_S(\mathbf{x}), \mathbf{y})\right]\right)
\end{aligned}$$

But note that $\mathop{\mathbb{E}}_{S,(\mathbf{x},\mathbf{y})}\left[l(\hat{f}_S(\mathbf{x}), \mathbf{y})\right] = \mathop{\mathbb{E}}_{S^i_{(\mathbf{x},\mathbf{y})},(\mathbf{x}^i,\mathbf{y}^i)}\left[l(\hat{f}_{S^i}(\mathbf{x}^i), \mathbf{y}^i)\right]$, due to symmetrization (since the $n+1$ points $S \cup \{(\mathbf{x}, \mathbf{y})\}$ are all i.i.d).
Substituting it in the equation gives us

$$\begin{aligned}
\frac{1}{n}\sum_i \left(\mathop{\mathbb{E}}_{S}\left[l(\hat{f}_S(\mathbf{x}^i), \mathbf{y}^i)\right] - \mathop{\mathbb{E}}_{S,(\mathbf{x},\mathbf{y})}\left[l(\hat{f}_S(\mathbf{x}), \mathbf{y})\right]\right) &= \frac{1}{n}\sum_i \left(\mathbb{E}\left[l(\hat{f}_S(\mathbf{x}^i), \mathbf{y}^i) - l(\hat{f}_{S^i}(\mathbf{x}^i), \mathbf{y}^i)\right]\right) \\
&\leq \frac{1}{n}\sum_i \beta(n) \\
&= \beta(n)
\end{aligned}$$

Using the same argument, we can show that $\mathbb{E}\left[er_{\mathcal{D}}^l\left[\hat{f}_S\right] - er_S^l\left[\hat{f}_S\right]\right]$ is also bounded by $\beta(n)$. This means that $\mathbb{E}\left[g(S)\right] \leq \beta(n)$ □

Finally, using lemmas 12.2 and 12.3, we can say that

$$\mathop{\mathbb{P}}_{S}\left[g(S) > (n(\beta(n) + B))\sqrt{\frac{8\log(\frac{2}{\delta})}{n}} + \beta(n)\right] \leq \delta$$

□

Note that the aforementioned result holds even if uniform convergence does not hold for the model class $\mathcal{F}$. The result entirely depends on $\beta(n)$, which only depends on the algorithm.

Ensuring algorithmic stability

Till now we have shown that if an algorithm is β-URL stable, then the model generalises what it learns on the training data to the test data. Note that how well this generalisation happens depends on the function $\beta(n)$.
In the concentration bound we have proven, we would like the quantity inside the probability to go to 0 as n tends to ∞. Intuitively, we would like the probability of the generalisation error being large to go down with n. For this to happen, $\beta(n)$ must alteast go down as $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. We will see later on in the course that for many algorithms, $\beta(n)$ falls of as $\mathcal{O}\left(\frac{1}{n}\right)$.

For many commonly used loss functions like the SVM loss function, support vector regression loss, or LASSO, techniques like regularization, perturbation (where one artificially adds noise to the data) or early stopping (helps by preventing the model from latching to a particular point). We will take a closer look at regularization in particular, and compute $\beta(n)$ for a regularized risk minimization setting.

Regularization

The regularized risk minimization setting is defined as follows:

$$\hat{\mathbf{w}}_s = \arg\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} h(\langle \mathbf{w}, \mathbf{x}^i\rangle, y^i)$$

where $h(\langle \mathbf{w}, \mathbf{x}^i\rangle, y^i) = l(\langle \mathbf{w}, \mathbf{x}^i\rangle, y^i) + \lambda \|\mathbf{w}\|_2^2$ is the $L2$-regularized loss function.

Exercise 12.1. Show that $\mathbb{P}\left[\left|er_S^l[\hat{\mathbf{w}}_S] - er_{\mathcal{D}}^l[\hat{\mathbf{w}}_S]\right| > \epsilon\right] \leq \delta$

The loss function l can be any convex and L-Lipschitz loss function, like hinge loss.

Definition 12.5. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be convex if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle$$

Definition 12.6. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be α-strongly convex if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$$

Remark 12.1. If $\mathbf{x}^* \in \arg\min_{\mathbf{x}} g(\mathbf{x})$, then $g(\mathbf{y}) \geq g(\mathbf{x}^*) + \frac{\alpha}{2}\|\mathbf{x}^* - \mathbf{y}\|_2^2$, as $\nabla g(\mathbf{x}^*) = 0$

Exercise 12.2. Show that the function $f(\mathbf{x}) = \frac{\|\mathbf{x}\|_2^2}{2}$ is 1-strongly convex

Exercise 12.3. If the function $f_i(\mathbf{x})$ is α_i-strongly convex $\forall i \in [n]$, then show that the function $f(\mathbf{x}) = \sum_{i\in[n]} f_i(\mathbf{x})$ is $\sum_i \alpha_i$-strongly convex

We would like to show that $\beta_{RRM}(n) \approx \frac{1}{\lambda n}$.

Claim 12.4. $\beta_{RRM}(n) \approx \frac{1}{\lambda n}$

Proof. For simplicity, we assume that $\mathbf{w} \in \mathbb{R}^d$
Firstly define a function $g_S(\mathbf{w}) = \frac{1}{n}\sum_{(\mathbf{x}^i, y^i)\in S} h(\langle \mathbf{w}, \mathbf{x}^i\rangle, y^i)$.
Wlog, assume $\|\mathbf{x}\| \leq 1$, meaning the data has already been normalized.

Exercise 12.4. Show that $g_S(\mathbf{w})$ is λ-strongly convex.

Let $\hat{\mathbf{w}}_S = \arg\min g_S(\mathbf{w})$ and $\hat{\mathbf{w}}_{S^i} = \arg\min g_{S^i}(\mathbf{w})$
Since both functions are λ-strongly convex, we have

$$g_S(\hat{\mathbf{w}}_{S^i}) \geq g_S(\hat{\mathbf{w}}_S) + \frac{\lambda}{2}\|\hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S^i}\|_2^2$$

$$g_S(\hat{\mathbf{w}}_{S^i}) \geq g_S(\hat{\mathbf{w}}_S) + \frac{\lambda}{2}\|\hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S^i}\|_2^2$$

$$\Rightarrow g_S(\hat{\mathbf{w}}_{S^i}) - g_{S^i}(\hat{\mathbf{w}}_{S^i}) \geq g_S(\hat{\mathbf{w}}_S) - g_{S^i}(\hat{\mathbf{w}}_S) + \lambda\|\hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S^i}\|_2^2$$

$$\Rightarrow \frac{1}{n}\left(l(\langle \hat{\mathbf{w}}_{S^i}, \mathbf{x}^i\rangle, y^i) - l(\langle \hat{\mathbf{w}}_{S^i}, \tilde{\mathbf{x}}\rangle, \tilde{y})\right) \geq \frac{1}{n}\left(l(\langle \hat{\mathbf{w}}_S, \mathbf{x}^i\rangle, y^i) - l(\langle \hat{\mathbf{w}}_S, \tilde{\mathbf{x}}\rangle, \tilde{y})\right) + \lambda\|\hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S^i}\|_2^2$$

$$\Rightarrow \frac{1}{n}\left(l(\langle \hat{\mathbf{w}}_{S^i}, \mathbf{x}^i\rangle, y^i) - l(\langle \hat{\mathbf{w}}_S, \mathbf{x}^i\rangle, y^i)\right) \geq \frac{1}{n}\left(l(\langle \hat{\mathbf{w}}_{S^i}, \tilde{\mathbf{x}}\rangle, \tilde{y}) - l(\langle \hat{\mathbf{w}}_S, \tilde{\mathbf{x}}\rangle, \tilde{y})\right) + \lambda\|\hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S^i}\|_2^2$$

Since the loss function is L-Lipschitz, we can say

$$\begin{aligned}\lambda \left\|\hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S^i}\right\|_2^2 &\leq \frac{2L}{n} \left\|\hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S^i}\right\|_2 \\ \Rightarrow \left\|\hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S^i}\right\|_2 &\leq \frac{2L}{\lambda n}\end{aligned}$$

Since the loss function h is Lipschitz, we can say that

$$\begin{aligned}\left|h(\langle \hat{\mathbf{w}}_S, \mathbf{x}\rangle, y) - h(\langle \hat{\mathbf{w}}_{S^i}, \mathbf{x}\rangle, y)\right| &\leq L \left|\langle \hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S^i}, \mathbf{x}\rangle\right| \\ &\leq L \left\|\hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S^i}\right\|_2 \|\mathbf{x}\|_2 \\ &\leq \frac{2L^2}{\lambda n}\end{aligned}$$

Therefore, $\beta_{RRM}(n) = \mathcal{O}\left(\frac{1}{\lambda n}\right)$ □