IFT 6085 - Lecture 10
Stability and generalization: Part II

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribe(s): Isabela Albuquerque and Nithin Vasisth **Instructor:** Ioannis Mitliagkas

1 Summary

In the previous lecture we defined the concept of uniform stability for an algorithm $\mathcal{A}$. We also derived a bound for the expected value of the defect $D[h_S]$ of a hypothesis h_S learned by a β-uniformly stable algorithm $\mathcal{A}$

$$-\beta \leq \mathbb{E}_S[D[h_S]] \leq \beta.$$

In this lecture, we review the PAC learning setup, some definitions and the result presented in the previous lecture and go further with the analysis of stable learning algorithms by applying McDiarmid's inequality to find a bound with high probability to the actual value of $D[h_S]$.

2 PAC Learning Setup

In the PAC (probably approximately correct) learning setup, we consider the following definitions.

Definition 1 (Training dataset)**.**

$$S = \{z_1, z_2, \ldots, z_n\},$$

where all $z_i = (x_i, y_i)$ are i.i.d. sampled from the unknown data distribution $\mathcal{D}$.

In standard learning problems, x_i represents a feature vector and y_i the label of the i-th sample.

Definition 2 (Hypothesis class)**.** *Represents the set of all possible hypothesis for a fixed architecture.*

$$\mathcal{H} = \{h_1, h_2, \ldots\}.$$

Definition 3 (Loss Function)**.**

$$\ell : \mathcal{Y} \times \mathcal{Y} \longrightarrow [0, M].$$

Note that in some cases we overloaded the notation for ℓ as follows:

$$\ell(h, z) = \ell(h(x), y).$$

Definition 4 (Population risk)**.**

$$R[h] = \mathbb{E}_{z\sim\mathcal{D}}\left[\ell(h(x), y)\right].$$

Definition 5 (Empirical risk)**.**

$$\hat{R}_S[h_S] = \frac{1}{n}\sum_{i=1}^{n} \ell(h_S(x_i), y_i).$$

3 Stability

In order to define the concept of stability, we first introduce the definition of perturbed dataset.

Definition 6 (Perturbed dataset)**.** *Given a dataset S and an index i, the perturbed dataset $S^{i,z}$ is defined as*

$$S^{i,z} = \{z_1, z_2, \ldots, z_{i-1}, z, z_{i+1} \ldots z_n\}.$$

The perturbed dataset $S^{i,z}$ is defined when a sample z replaces the i-th element in the dataset S.

All bounds previously shown in class were *algorithm agnostic*, i.e. they were just dependent on the hypothesis class. To introduce the notion of stability, we define the learning algorithm $\mathcal{A}$ as follows

Definition 7 (Algorithm)**.** *A learning algorithm $\mathcal{A}$ is defined as the following mapping*

$$\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}.$$

Here, h_S represents the hypothesis class that was trained using the algorithm $\mathcal{A}$ on the dataset S, i.e. $h_S = \mathcal{A}(S)$.

Definition 8 (Uniform stability)**.** *An algorithm $\mathcal{A}$ is β-uniformly stable with respect to the loss function ℓ for all $(S, z) \in \mathcal{Z}^{n+1}$ and for all $i \in \{1, 2, \ldots, n\}$ if*

$$\sup_{z' \in \mathcal{Z}} |\ell(h_S, z') - \ell(h_{S^{i,z}}, z')| \leq \beta.$$

Basically, if we say that an algorithm is β-uniformly stable, we mean that any outlier in the dataset does not affect the loss ℓ drastically. It is important to emphasize that this is a uniform bound because it is valid for all datasets S and indexes i.

In the last lecture, the definition of defect was presented and Theorem 10 was proven. We will review those results before showing a bound with high probability for $D[h_S]$.

Definition 9 (Defect)**.**

$$D[h_S] = R[h_S] - \hat{R}_S[h_S].$$

Theorem 10 (Bounding the expectation of the defect)**.** *If $\mathcal{A}$ is a β-uniformly stable algorithm, then*

$$-\beta \leq \mathbb{E}_S[D[h_S]] \leq \beta.$$

Proof was given in the previous class.

Property 11 (Relation between the empirical and the population risk)**.**

$$\mathbb{E}_S[R[h_S]] \leq \mathbb{E}_S[\hat{R}_S[h]] + \beta.$$

Proof. Follows directly from Definition 9 and Theorem 10. □

Note this is a bound in the *expectation*, which can be considered not strong in the sense that if $R[h_S]$ and $\hat{R}_S[h_S]$ match in expectation, it does not imply that one does not exceed the other for all possible h_S. The main goal of this lecture is to extend the latter result to a high probability bound for actual values of $R[h_S]$ and $\hat{R}_S[h_S]$.
In the following theorem we introduce McDiarmid's inequality.

Theorem 12 (McDiarmid inequality)**.** *Let $V_1, V_2, \ldots, V_n \in \mathcal{V}$ be independent random variables and $v_1, v_2, \ldots v_n$ be samples drawn from these random variables. If a function $F : \mathcal{V}^n \to \mathbb{R}$ has the following property for all $i \in \{1, \ldots, n\}$:*

$$\sup_{v_1, v_2, \ldots, v_n, v_i'} |F(v_1, v_2, \ldots, v_n) - F(v_1, v_2, \ldots, v_{i-1}, v_i', v_{i+1}, \ldots, v_n)| \leq c_i,$$

then

$$P(|F(V_1, V_2, \ldots, V_n) - \mathbb{E}[F(V_1, V_2, \ldots, V_n)]| > \epsilon) \leq 2 \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof. See Appendix D.2 of [1]. □

4 Bounding the defect with high probability

Theorem 13 (Bound for the defect)**.** *Consider a β-uniformly stable algorithm $\mathcal{A}$ with respect to a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, M]$. The absolute difference of the defect calculated on a dataset S and on a perturbed version of this dataset $S^{i,z}$ is bounded by*

$$|D[h_S] - D[h_{S^{i,z}}]| \leq 2\beta + \frac{M}{n}.$$

Proof. Let us expand the following value:

$$|D[h_S] - D[h_{S^{i,z}}]| = |R[h_S] - \hat{R}_S[h_S] - R[h_{S^{i,z}}] + \hat{R}_S[h_{S^{i,z}}]|. \tag{1}$$

Using the triangle inequality, we find the following upper bound for the RHS of Eq. 1:

$$|R[h_S] - \hat{R}_S[h_S] - R[h_{S^{i,z}}] + \hat{R}_S[h_{S^{i,z}}| \leq |\hat{R}_S[h_{S^{i,z}}] - \hat{R}_S[h_S]| + |R[h_S] - R[h_{S^{i,z}}]|.$$

Replacing this result in Eq. 1, we obtain:

$$|D[h_S] - D[h_{S^{i,z}}]| \leq |\hat{R}_S[h_{S^{i,z}}] - \hat{R}_S[h_S]| + |R[h_S] - R[h_{S^{i,z}}]|. \tag{2}$$

Now, we use the β-uniform stability of $\mathcal{A}$ with respect to ℓ to find a bound for $|R_S[h_S] - R_S[h_{S^{i,z}}]|$:

$$\begin{aligned} |R[h_S] - R[h_{S^{i,z}}]| &= \mathbb{E}_{z'}[\ell(h_S, z')] - \mathbb{E}_{z'}[\ell(h_{S^{i,z}}, z')], \\ &= \mathbb{E}_{z'}[\ell(h_S, z') - \ell(h_{S^{i,z}}, z'))], \\ &= \int_{z' \in \mathcal{Z}} [\ell(h_S, z') - \ell(h_{S^{i,z}}, z')] \, d\mu(z'), \\ &\leq \beta. \end{aligned} \tag{3}$$

Using the result obtained in Eq. 3 and writing $\hat{R}[h_{S^{i,z}}]$ and $\hat{R}[h_S]$ according to their respective definitions, it is possible to rewrite Eq. 2 as

$$\begin{aligned} |D[h_S] - D[h_{S^{i,z}}]| &\leq \beta + \left| \frac{1}{n} \sum_{j=1}^{n} \ell(h_{S^{i,z}}, z_j) - \ell(h_S, z_j) \right|, \\ &= \beta + \frac{1}{n} \left| \ell(h_{S^{i,z}}, z) - \ell(h_S, z_i) \right| + \frac{1}{n} \sum_{j \neq i} \left| \ell(h_{S^{i,z}}, z_j) - \ell(h_S, z_j) \right|, \end{aligned} \tag{4}$$

Reminding the fact that $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, M]$ and using the β-uniform stability property of $\mathcal{A}$ in Eq. 4 we obtain the desired bound:

$$\begin{aligned} |D[h_S] - D[h_{S^{i,z}}]| &\leq \beta + \frac{M}{n} + \frac{(n-1)\beta}{n}, \\ &\leq 2\beta + \frac{M}{n}. \end{aligned} \tag{5}$$

□

Theorem 14 (Bound for the population risk of a β-uniformly stable algorithm)**.** *Consider a β-uniformly stable algorithm $\mathcal{A}$ with respect to a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, M]$ and a hypothesis h_S with $|S| = n$. The following bound holds with probability $1 - \delta$:*

$$R[h_S] \leq \hat{R}_S[h_S] + \beta + \left(n\beta + \frac{M}{2} \right) \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}.$$

Proof. Using theorem 13, we state McDiarmid's inequality for $D[h_S]$ and then use this result to find a high probability bound for $D[h_S]$:

$$\sup_{S,i,z} |D[h_S] - D[h_{S^{i,z}}]| \leq 2\beta + \frac{M}{n}, \tag{6}$$

then

$$\begin{aligned} P(|D[h_S] - \mathbb{E}[D[h_S]]| > \epsilon) &\leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n}\left(2\beta + \frac{M}{n}\right)^2}\right), \\ &= 2\exp\left(\frac{-2n\epsilon^2}{(2n\beta + M)^2}\right), \\ &= 2\exp\left(\frac{-2n\epsilon^2}{4\left(n\beta + \frac{M}{2}\right)^2}\right), \\ &= 2\exp\left(\frac{-n\epsilon^2}{2\left(n\beta + \frac{M}{2}\right)^2}\right). \end{aligned} \tag{7}$$

Denoting $\delta = 2\exp\left(\frac{-n\epsilon^2}{2\left(n\beta+\frac{M}{2}\right)^2}\right)$ and solving this equation for ϵ, we obtain:

$$\begin{aligned} \delta = 2\exp\left(\frac{-n\epsilon^2}{2\left(n\beta + \frac{M}{2}\right)^2}\right) &\Rightarrow n\epsilon^2 = 2\log\frac{2}{\delta}\left(n\beta + \frac{M}{2}\right)^2, \\ &\Rightarrow \epsilon = \left(n\beta + \frac{M}{2}\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}}. \end{aligned} \tag{8}$$

Thus, with probability $1 - \delta$

$$\begin{aligned} |D[h_S] - \mathbb{E}[D[h_S]]| &\leq \epsilon, \\ D[h_S] &\leq \mathbb{E}[D[h_S]] + \epsilon, \\ D[h_S] &\leq \beta + \epsilon. \end{aligned}$$

Replacing ϵ by the result previously obtained in Eq. 8

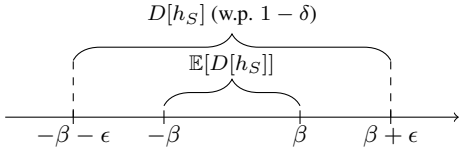
$$D[h_S] \leq \beta + \left(n\beta + \frac{M}{2}\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}}, \tag{9}$$

and, thus, using Definition 9 we finally get the desired result

$$R[h_S] \leq \hat{R}_S[h_S] + \beta + \left(n\beta + \frac{M}{2}\right)\sqrt{\frac{2\log\frac{2}{\delta}}{n}}. \tag{10}$$

□

The following illustration represents a summary of the bounds stated in Theorem 10 and 14 (in terms of the defect):



One can observe that despite the fact the bound for $\mathbb{E}[D[h_S]]$ is tighter, we have no guarantees that the actual of $D[h_S]$ lies in the interval $[-\beta, \beta]$. On the other hand, it possible to assure with probability $1 - \delta$ will be in $[-\beta - \epsilon, \beta + \epsilon]$.

References

[1] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2012.