

Lecture 13 — 10/18

Lecturer: Dimitris Papailiopoulos

Scribe: Huayu Zhang

Note: These lecture notes are still rough, and have only have been mildly proofread.

13.1 Stability of learning algorithms

In this lecture, we try to establish a connect between the stability and generalization of an algorithm. Specifically, algorithmic stability implies a good generalization error. Consider such a learning problem. A set of data $S = \{Z_i = (\mathbf{x}_i, y_i) \mid i = 1, 2, \dots, n\}$ are drawn independently from a distribution $Z_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}$. The model is characterized by parameters \mathbf{w} and the loss function is $\ell(\mathbf{w}; Z)$.

Definition 13.1 (Generalization error). We define the risk as

$$R(\mathbf{w}) = \mathbb{E}_{Z \sim \mathcal{D}} [\ell(\mathbf{w}; Z)] \quad (13.1)$$

and the empirical risk as

$$\hat{R}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{w}; Z_i) \quad (13.2)$$

The generalization error is the difference

$$\epsilon = R(\mathbf{w}) - \hat{R}(\mathbf{w}) \quad (13.3)$$

Let A be the algorithm and $A(S)$ be the output model of the algorithm A on data S . We can write

$$\begin{aligned} R(A(S)) &= \mathbb{E}_{Z \sim \mathcal{D}} [\ell(A(S); Z)] \\ \hat{R}(A(S)) &= \frac{1}{n} \sum_{i=1}^n \ell(A(S); Z_i) \\ \mathbb{E}_S [\epsilon] &= \mathbb{E}_S [R(A(S)) - \hat{R}(A(S))] \end{aligned}$$

Next we define the stability of algorithms. The stability is a metric to show how the result of an algorithm varies with one data changed. On the data set S , replace one data point Z_i with another i.i.d. random variable Z'_i . Then we get a new data set $S^i = (S \setminus \{Z_i\}) \cup \{Z'_i\}$.

Definition 13.2 (ϵ -Stable). Algorithm A is ϵ -Stable if for any $i \in \{1, 2, \dots, n\}$,

$$\mathbb{E}_{S, Z'_i, Z} [\ell(A(S_i); Z) - \ell(A(S^i); Z)] \leq \epsilon \quad (13.4)$$

The following theorem shows algorithmic stability implies good generalization error.

Theorem 13.3. *If A is ε -stable, then $\mathbb{E}_S[|R(A(S)) - \hat{R}(A(S))|] \leq \varepsilon$.*

Proof:

$$\begin{aligned}
 \mathbb{E}_S[\hat{R}(A(S))] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_S[\ell(A(S); Z_i)] \\
 &= \frac{1}{n} \sum_{i=1}^n \{ \mathbb{E}_{S, Z'_i}[\ell(A(S), Z_i)] + \mathbb{E}_{S, Z'_i}[\ell(A(S), Z_i) - \ell(A(S), Z'_i)] \} \\
 &= \mathbb{E}_S[\ell(A(S), Z_i)] = \mathbb{E}_{S^i, Z'_i}[\ell(A(S^i), Z'_i)] \\
 &= \mathbb{E}_S[R(A(S))] + \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{S, Z'_i}[\ell(A(S^i), Z'_i) - \ell(A(S), Z'_i)] \\
 &\stackrel{A \text{ is } \varepsilon\text{-stable}}{\leq} \mathbb{E}_S[R(A(S))] + \varepsilon
 \end{aligned}$$

□

Question 13.4. *What algorithms are ε -stable?*

- $A(S) = \arg \min \sum_{i=1}^n \ell(\mathbf{w}; Z_i)$.
- $A(S)$ = the output of SGD after T iterations.

Case I $A(S) = \arg \min f_S(\mathbf{w})$, $f_S(\mathbf{w}) = L_S(\mathbf{w}) + \lambda \|\mathbf{w}\|^2$, $L_S(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \ell_i(\mathbf{w})$. ℓ is L -Lipschitz. $A(S)$ is $O(\frac{1}{\lambda n})$ -stable.

Proof:

$$\begin{aligned}
 f_S(\mathbf{v}) - f_S(\mathbf{u}) &= L_S(\mathbf{v}) + \lambda \|\mathbf{v}\|^2 - (L_S(\mathbf{u}) + \lambda \|\mathbf{u}\|^2) \\
 &= L_{S^i}(\mathbf{v}) + \lambda \|\mathbf{v}\|^2 - (L_{S^i}(\mathbf{u}) + \lambda \|\mathbf{u}\|^2) + \frac{\ell(\mathbf{v}; z_i) - \ell(\mathbf{v}; z'_i)}{n} - \frac{\ell(\mathbf{u}; z_i) - \ell(\mathbf{u}; z'_i)}{n}
 \end{aligned}$$

Let $\mathbf{v} = A(S^i)$, $\mathbf{u} = A(S)$,

$$\begin{aligned}
 f_S(A(S^i)) - f_S(A(S)) &\leq \frac{\ell(\mathbf{v}; z_i) - \ell(\mathbf{v}; z'_i)}{n} - \frac{\ell(\mathbf{u}; z_i) - \ell(\mathbf{u}; z'_i)}{n} \\
 &\leq \frac{2L}{n} \|A(S) - A(S^i)\|
 \end{aligned}$$

According to the property of 2λ -strongly convex function

$$f_S(\mathbf{w}) - f_S(A(S)) \geq \lambda \|\mathbf{w} - A(S)\|^2$$

Thus

$$\|A(S) - A(S^i)\| \leq \frac{2L}{\lambda n} \quad (13.5)$$

By Eq. 13.5 and L -Lipschitz of ℓ_i .

$$\sup_Z |\ell(A(S), Z) - \ell(A(S^i), Z)| \leq \frac{2L^2}{\lambda n}$$

$A(S)$ is $\frac{2L^2}{\lambda n}$ stable.

□