ECE 901: Large-scale Machine Learning and Optimization	**Fall 2016**
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 connection 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 $S = \{Z_i = (\boldsymbol{x}_i, y_i) \mid i = 1, 2, \dots, n\}$ are drawn independently from a distribution $Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{D}$. The model is characterized by parameters $\boldsymbol{w}$ and the loss function is $\ell(\boldsymbol{w}; Z)$.

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

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

and the empirical risk as

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

The generalization error is the difference

$$\epsilon = R(\boldsymbol{w}) - \hat{R}(\boldsymbol{w}) \tag{13.3}$$

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

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

Next we define the stability of an algorithm. Stability is a metric to show how the result of an algorithm varies with one sample 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 dataset $S^i = (S \backslash \{Z_i\}) \bigcup \{Z_i'\}$.

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

$$\mathop{\mathbb{E}}_{S,Z_i',Z}[| \ \ell(A(S_i); Z) - \ell(A(S^i); Z) \ |] \le \varepsilon \tag{13.4}$$

The following theorem shows algorithmic stability implies good generalization error.

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

Proof:

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

□

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

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

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

Proof:

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

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

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

According to the property of 2λ-strongly convex functions

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

Thus

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

By Eq. 13.5 and L-Lipschitz of ℓ_i.

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

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