**CS 419M Introduction to Machine Learning** **Spring 2021-22**

Lecture 9: Stability of Learning Algorithms

*Lecturer: Abir De* *Scribe: Group 2*

## 9.1 Characterization of Stability of an Algorithm

Let $A$ be a Learning Algorithm and $S$ be the Data set which is fed into the Learning algorithm.The outcome/output of the learning algorithm is $A(S)$.(We can think of $A(S)$ as a vector to define a norm)

**Definition 9.1.** *(Stability) A learning algorithm A is said to be stable iff*

$$\|A(S) - A(S')\| \leq \mathcal{O}\left(\frac{1}{|S|}\right)$$

*For every $S$ and $S'$ such that $|S \backslash S'| = |S' \backslash S| = 1$.*

The condition on $S$ and $S'$ means that there is only a one element mismatch between the sets.

Consider instead, what happens if we just delete one element $e$ from the set and take the norm of the difference:(Stability towards single element deletions)

$$\|A(S) - A(S \backslash e)\|$$

We want to find the relation of the above with the previously defined notion of stability. This is dealt with in the following theorem.

**Proposition 9.2.** *Let $A$ be a Learning Algorithm, $S = \{(x_i, y_i)\}$ be a data set and $e$ be a single data point, $e = (x_r, y_r)$ for some $r$ such that $e \in S$.The following is a sufficient condition for the Algorithm to be stable:*

$$\|A(S) - A(S \backslash e)\| = \mathcal{O}\left(\frac{1}{|S|}\right) \quad \forall e, S$$

*Proof.* Consider set $S$ and $S'$ such that $|S \backslash S'| = |S' \backslash S| = 1$. This means that there exists $e$ and $e'$ such that $S \backslash e = S' \backslash e'$. We shall also be using Triangle inequality. Let us start with the expression in the definition of stability:

$$\|A(S) - A(S')\| = \|A(S) - A(S \backslash e) + A(S' \backslash e') - A(S')\|$$

(We can do this since $S \backslash e = S' \backslash e'$).Now applying Triangle inequality to the right hand side:

$$\|A(S) - A(S')\| \leq \|A(S) - A(S \backslash e)\| + \|A(S' \backslash e') - A(S')\|$$

But we already have :

$$\|A(S) - A(S\backslash e)\| = \mathcal{O}\left(\frac{1}{|S|}\right)$$

$$\|A(S'\backslash e') - A(S')\| = \mathcal{O}\left(\frac{1}{|S'|}\right)$$

Using this, we have:

$$\|A(S) - A(S')\| \leq \mathcal{O}\left(\frac{1}{|S|}\right) + \mathcal{O}\left(\frac{1}{|S'|}\right)$$

Since $|S| = |S'|$ :

$$\|A(S) - A(S')\| \leq \mathcal{O}\left(\frac{1}{|S|}\right)$$

□

**Note:** If we add noise to $x_i$ then accuracy will decrease, but our model will become more stable.

**Proposition 9.3.** *Let $F_w(s) = \sum_{i \in S}[l(W^T x_i, y_i) + \lambda||W||^2]$ be the loss function of a Learning Algorithm, $S = \{(x_i, y_i)\}$ be a data set where $l$ is convex (i.e., all eigenvalues of $\frac{\partial^2 l(W)}{\partial W^2}$ are positive) and Lipschitz continuous (i.e., $||l(s) - l(s')|| \leq ||s - s'||$). Then the Algorithm will be stable i.e.,*

$$\|W^*(S) - W^*(S')\| = \mathcal{O}\left(\frac{1}{|S|}\right)$$

*where*

$$W^* = \arg\min_W F_W(S)$$

*Proof.* If we are somehow able to prove below inequality

$$K\|W*(S) - W*(S')\|^2 \leq F_{w*(s)}(S) - F_{w*(s')}(S') \leq K'\|W*(S) - W*(S')\|$$

Then

$$K\|W*(S) - W*(S')\|^2 \leq K'\|W*(S) - W*(S')\|$$

$$\|W*(S) - W*(S')\| \leq K'/K$$

and now if we prove

$$K'/K = \mathcal{O}\left(\frac{1}{|S|}\right)$$

then as our requirement we are at the end of proof i.e.

$$\|W^*(S) - W^*(S')\| = \mathcal{O}\left(\frac{1}{|S|}\right)$$

To begin with that take a function g and assume g is convex.
So, from Taylor Series

$$g(W) = g(Wmin) + (dg/dW_{min})^T(W - W_{min}) + (W - W_{min})^T\mathcal{H}(W - W_{min})$$

As

$$(dg/dW_{min}) = 0$$

and

$$\mathcal{H} \leq Eigen_{min}$$

we have

$$g(W) - g(Wmin) \leq Eigen_{min}(W - W_{min})$$

Now take g is Lipschitz Continuous,

$$||g(W) - g(W_{min})|| \leq L||W - W_{min}||$$

Here,

$$F_w(s) = \sum_{i \in S}[l(W^T x_i, y_i) + \lambda||W||^2]$$

So,

$$\mathcal{H} \geq 2\lambda S$$

and K is related to $Eigen_{min} \geq \mathcal{H}$ i.e. $K \approx 2\lambda S$

Now, we again start with

$$\begin{aligned} & F_{w*(s)}(S) - F_{w*(s')}(S') \\ = & F_{w*(s)}(S) - F_{w*(s)}(S') + F_{w*(s)}(S') - F_{w*(s')}(S') \\ \leq & F_{w*(s')}(S) - F_{w*(s')}(S') \\ = & l(W*(S')^T x_i, y_i) - l(W*(S)^T x_i', y_i') \\ \leq & L\|W*(S') - W*(S)\| \end{aligned}$$

by applying the triangular inequality and using the Lipschitz condition □

## 9.2 Group Details and Individual Contribution

- 200110055 Keshav Patel Keval: Definition 9.1 and Proposition 9.2
- 19D070017 Bhavishya: Proposition 9.3
- 200100127 Rahul: Part of the Proof for Proposition 9.3
- 19D070046 Phansalkar Ishan Shrirang: Completed Proof for Proposition 9.3