CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 17: Stability Analysis (October 31)

Lecturer: Csaba Szepesvári *Scribes: Kushagra Chandak*

Some algorithms do not explore the whole function space of losses. For such cases, capacity measures of classes like log cardinality, metric entropy, and Rademacher complexity do not give the best generalization bounds. We can exploit other properties of algorithms, like *stability*, to get better generalization bounds.

Previously, we got oracle inequalities of the form

$$Pg_n \le \inf_{g\in\mathcal{G}} Pg + O\left(\frac{\ln|\mathcal{G}|/\delta}{n}\right).$$

These kind of bounds were for ERM. Now consider (stochastic) gradient descent for ERM.

Generalization bounds are used to get guarantees for test error using training samples. But it doesn't say anything about the performance of the algorithm.

Edge cases. If η is 0, then SGD is not doing anything and in the generalization bounds we expect no $\log|\mathcal{G}|$ factor. That is bounds for training loss minus test loss, or

$$Pg_n - P_n g_n\,,$$

should bet better and not have the $\log|\mathcal{G}|$ factor. But do we also expect it for the oracle inequalities? Maybe, if $\mathcal{A}$ is not doing too much and it gets lucky to get close to the best possible loss. So we have two concerns: optimization (oracle inequalities) and generalization (train - test loss). In this section, we will only try to get better generalization error for "stable" algorithms.

Definition 17.1 (Hamming distance)**.** For $z, z' \in \mathcal{Z}^n$, the Hamming distance is defined as

$$H(z,z') = \sum_{i=1}^{n} \mathbb{I}(z_i \neq z'_i)\,.$$

Definition 17.2. An algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathbb{R}^{\mathcal{Z}}$ is ε-uniformly stable if for all $z, z' \in \mathcal{Z}^n$ such that $H(z,z') \le 1$, we have

$$\|\mathcal{A}(z) - \mathcal{A}(z')\|_\infty \le \varepsilon\,.$$

Note that the map $\mathcal{Z} : (\mathcal{Z}^n, H) \to (\mathbb{R}^{\mathcal{Z}}, \|\cdot\|_\infty)$ is ε-Lipschitz.

Remark 17.3. For the binary loss class, an algorithm is ε-uniformly stable only if it outputs the same value (constant) for all inputs z, z'. Therefore, it's important to remember that the notion of ε-uniform stability is not very interesting for the binary loss class.

For randomized algorithms, we define ε-uniform stability in the following way.

Definition 17.4. Let $K : \mathcal{Z}^n \to \mathcal{M}_1(\mathbb{R}^{\mathcal{Z}})$ be a probability kernel. Let $\mu_K : \mathcal{Z}^n \to \mathbb{R}^{\mathcal{Z}}$ be the deterministic "mean" map such that $\mu_K(z)(z') = \int g(z')K(dg|z)$. Then K is ε-uniformly stable if μ_K is ε-uniformly stable.

Let $Z \sim P$ and $G \sim K(\cdot|Z)$. Then the mean map $\mu_K(Z)$ can also be seen as the conditional expectation $\mathbb{E}G|Z$. The mean loss of K is written as $P\mu_K(Z)$ and the empirical loss as $P_n\mu_K(Z)$. (Think of effectively replacing the algorithm K with the algorithm $\mu(K)$).

Proposition 17.5. *Let $P \in \mathcal{M}_1(\mathcal{Z})$ and $K : \mathcal{Z}^n \to \mathcal{M}_1(\mathbb{R}^{\mathcal{Z}})$ be an ε-uniformly stable algorithm. Further let $Z \sim P^{\otimes n}$, $G \sim K(\cdot|Z)$ and $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ be the empirical distribution. Then*

$$\mathbb{E}PG \leq \mathbb{E}P_nG + \varepsilon\,.$$

Proof. We first write $\mathbb{E}[PG]$ in terms of samples so that we can write $\mathbb{E}PG$ and $\mathbb{E}P_nG$ in a similar way and use the stability property. Let $Z' \sim P^{\otimes n}$. Further, let $Z^{(i)} = (Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n)$ and $G^{(i)} \sim K(\cdot|Z^{(i)})$. Now since $P_{(Z_1',Z)} = P_{(Z_i,Z^{(i)})}$ because of independence, which gives

$$P_{(Z_1',G)} = P_{(Z_i,G^{(i)})}\,. \tag{$*$}$$

Therefore for all $i \in [n]$, we have

$$\mathbb{E}[PG] = \mathbb{E}[G(Z_1')] = \mathbb{E}[G^{(i)}(Z_i)]\,,$$

where the second equality uses $(*)$. So taking empirical average on both sides gives

$$\mathbb{E}[PG] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[G^{(i)}(Z_i)]\,.$$

Finally we have

$$\begin{aligned}
\mathbb{E}[PG] - \mathbb{E}[P_nG] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[G^{(i)}(Z_i) - G(Z_i)] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{E}[G^{(i)}(Z_i) - G(Z_i)|Z, Z']] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mu_K(Z^{(i)}) - \mu_K(Z_i)] \\
&\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\varepsilon] \leq \varepsilon\,.
\end{aligned}$$

□

Remark 17.6. Note that $\mathbb{E}[PG] = \mathbb{E}[\mathbb{E}[PG|Z]] = \mathbb{E}[P\mathbb{E}[G|Z]] = \mathbb{E}[P\mu_K(Z)]$. We see that the expected loss of the random choice of the algorithm is the same as the expected loss of the "mean" map. So rather than thinking about the whole distribution over all the loss functions, we can only think about the mean loss function.

The previous result gave us a generalization error bound in expectation. The next result shows that we can also get the result in expectation.

Theorem 17.7. *Let K be ε-uniformly stable, $(Z, Z') \sim P^{\otimes 2n}$ and $\alpha \in (0, 1]$. Assume that for all $\delta \in (0, 1)$, w.p. $1 - \delta$*

$$\alpha\mathbb{E}[P\mu_K(Z)] \leq P_n'\mu_K(Z) + \varepsilon_n(\delta)\,. \tag{$*$}$$

Then w.p. $1 - \delta$,

$$P\mu_K(Z) \leq P_n\mu_K(Z) + (1-\alpha)\mathbb{E}P_n\mu_K(Z) + \varepsilon\left(\frac{\delta}{2}\right) + \varepsilon \cdot (2 + 5\lceil\log_2 n\rceil)\ln\left(\frac{2}{\delta}\right) + (3-\alpha)\varepsilon\,.$$

Discussion. The condition in $(*)$ in Theorem 17.7 is a benign condition that can be easily obtained from a concentration inequality. For example, if $\mu_K \subseteq [0,1]^{\mathcal{Z}}$ then $\varepsilon_n(\delta) = O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{n}}\right)$ with using Hoeffding's inequality with α chosen as 1. The result gives us a high probability bound on the random test error. Further, for large n, we expect ε to behave like $\sim \frac{1}{\sqrt{n}}$ or better.

Definition 17.8 (Centered leave-one-out loss estimate)**.** Let $g : \mathcal{Z}^n \to \mathbb{R}^{\mathcal{Z}}$ be an algorithm and $Z_{1:n+1} \sim P^{\otimes(n+1)}$. Further, let $Z^{(i)} = (Z_1, \ldots, Z_{i-1}, Z_{n+1}, Z_{i+1}, \ldots, Z_n) \in \mathcal{Z}^n$. Then the leave-one-out loss estimate is defined as

$$\bar{g}(Z_{1:n+1}) := \frac{1}{n}\sum_{i=1}^{n} g(Z^{(i)}; Z_i) - \int g(\tilde{z}; Z_i) P^{\otimes n(d\tilde{z})} \,,$$

where $g(Z^{(i)}; Z_i) := g(Z^{(i)})(Z_i)$ is the loss of the algorithm obtained from training on $Z^{(i)}$ and evaluating on Z_i.

Lemma 17.9. *Let $g : \mathcal{Z}^n \to \mathbb{R}^{\mathcal{Z}}$ be an ε-uniformly stable algorithm and for all $z_{1:n}$, $Pg(z_{1:n}) = 0$. Let $\bar{g}$ be the centered leave-one-out loss estimate. Then for all $\delta \in (0,1)$, w.p. $1-\delta$,*

$$\bar{g}(Z_{1:n+1}) \leq \lceil \log_2 n \rceil \, \varepsilon \left(1 + 2.5 \ln \frac{1}{\delta}\right) .$$

Proof of Theorem 17.7. □

We can use uniform stability to derive high probability generalization results like Theorem 17.7. We can also use a more refined notion of stability, called *leave-one-out* stability, to get better bounds in some cases.