10-716: Advanced Machine Learning **Spring 2019**

Lecture 8: February 12

Lecturer: Pradeep Ravikumar *Scribes: Benjamin Eysenbach, Ting Fung Lau, Zinan Lin*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

8.1 Uniform Laws

$$L(\theta, \theta^*) = \mathbb{E}_{x \sim p(\cdot|\theta^*)} \ell(x, \theta) \tag{8.1}$$

$$L_n(\theta, \theta^*) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \theta) \tag{8.2}$$

Empirical Risk Minimization (ERM) is what we actually minimize using samples

$$\hat{\theta} \in \arg \inf_{\theta \in \Theta_0 \subseteq \Theta} L_n(\theta, \theta^*) \tag{8.3}$$

Our "gold standard" is the optimum w.r.t. the true expectation:

$$\theta_0 \in \arg \inf_{\theta \in \Theta_0 \subseteq \Theta} L(\theta, \theta^*) \tag{8.4}$$

To compare these two quantities, we will look at the *excess*:

$$E(\hat{\theta}, \theta_0) = L(\hat{\theta}, \theta^*) - L(\theta_0, \theta^*) \tag{8.5}$$

$$= \underbrace{L(\hat{\theta}, \theta^*) - L_n(\hat{\theta}, \theta^*)}_{T_1} + \underbrace{L_n(\hat{\theta}, \theta^*) - L_n(\theta_0, \theta^*)}_{T_2} + \underbrace{L_n(\theta_0, \theta^*) - L(\theta_0, \theta^*)}_{T_3} \tag{8.6}$$

We know $T_2 \leq 0$ by definition of $\hat{\theta}$ being optimal for L_n.

We can bound T_3 directly using a tail bound:

$$L_n(\theta_0, \theta^*) - L(\theta_0, \theta^*) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \theta_0) - \mathbb{E}[\ell(x_i, \theta_0)] \tag{8.7}$$

For T_3 we assumed the x_i were iid, so each $\ell(x_i, \theta)$ was independent. For T_1, $\hat{\theta}$ depends on x_i, each $\ell(x_i, \hat{\theta})$ is *dependent*. Thus, we cannot directly apply the tail bounds we derived in the last lecture.

$$L_n(\hat{\theta}, \theta^*) - L(\hat{\theta}, \theta^*) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \hat{\theta}) - \mathbb{E}_{x \sim p(\cdot|\theta^*)}[\ell(x, \hat{\theta})] \tag{8.8}$$

$$\leq \sup_{\theta \in \Theta_0} \left| \sum_{i=1}^{n} \ell(x_i, \theta) - \mathbb{E}[\ell(x_i, \theta)] \right| \triangleq \delta_n \tag{8.9}$$

Noting that we can also bound $T_3 \leq \delta_n$, we obtain the following bound on the excess:

$$E(\hat{\theta}, \theta_0) \leq 2\delta_n \tag{8.10}$$

8.1.1 Uniform Laws

We'll begin by defining x^θ as a random variable drawn from $p_\theta(\cdot)$. We are interested in the deviation between the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i^\theta$ and its expectation, $\mathbb{E}[x^\theta]$. In particular, we are interested in the maximum deviation between these quantities, as we vary θ:

$$\sup_\theta |\frac{1}{n}\sum_{i=1}^{n} x_i^\theta - \mathbb{E}[x^\theta]| \tag{8.11}$$

8.1.2 Uniform Laws for CDFs

One early application of uniform laws was to cumulative density functions (CDFs):

$$F(t) \triangleq P(x \leq t) = \mathbb{E}[\mathbb{1}(x \in (-\infty, t))] \tag{8.12}$$

We now define the *empirical* CDF as

$$F_n(t) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \leq t) \tag{8.13}$$

For a fixed t, the Law of Large Numbers tells us that the empirical CDF converges to the true CDF as n goes to infinity:

$$F_n(t) \stackrel{\text{a.s.}}{\to} F(t) \tag{8.14}$$

But we are really interested in the CDF converges simultaneously *for all* t.

Theorem 8.1 (Glivenko-Cantelli) *This theorem tells us that CDFs converge uniformly*

$$\|F_n - F\|_\infty \stackrel{a.s.}{\to} 0 \tag{8.15}$$

where $\|F - G\|_\infty \triangleq \sup_{t\in\mathbb{R}} \|F(t) - G(t)\|$

However, the Glivenko-Cantelli theorem does not tell us about uniform convergence of other quantities. Now, we will prove a generalization of the Glivenko-Cantelli theorem (that will include the result we want to ERM). We consider iid samples $x_i \sim \mathbb{P}$, where each sample belongs to some set: $x_i \in \mathcal{X}$. We consider a set of functions $\mathcal{F}$ defined over the set $\mathcal{X}$. We are interested in the following deviation:

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \underbrace{f(x_i)}_{z^f} - \mathbb{E}[f(x)] \right| \tag{8.16}$$

The Glivenko-Cantelli theorem was a special case, where we considered the following set of functions:

$$\mathcal{F} = \{\mathbb{1}(x \in (-\infty, t])\ ;\ t \in \mathbb{R}\} \tag{8.17}$$

For ERM, we consider another set of functions:

$$\mathcal{F} = \{\ell(\cdot, \theta)\ ;\ \theta \in \Theta\} \tag{8.18}$$

Definition: We define the distance $\|\cdot\|_\mathcal{F}$ as the maximum absolute value over functions in $\mathcal{F}$:

$$\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F} = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}[f(x)] \right| \tag{8.19}$$

Definition: A set of functions $\mathcal{F}$ is a *Glivenko-Cantelli Class* if the following result holds for all distributions $\mathbb{P}$:

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \overset{\text{prob.}}{\to} 0 \tag{8.20}$$

We say that $\mathcal{F}$ is a *strong* Glivenko-Cantelli Class if we have almost-sure convergence:

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \overset{\text{a.s.}}{\to} 0 \tag{8.21}$$

Example: The set $\mathcal{F} = \{\mathbb{1}(x \in S)\ ;\ S \subseteq [0,1]\}$ is not a Glivenko-Cantelli Class. If we draw samples x from a continuous density.

$$\sup_{S \subseteq [0,1]} |\mathbb{E}_n[\mathbb{1}(x \in S)] - \mathbb{E}[\mathbb{1}(x \in S)]| = 1 \neq 0 \tag{8.22}$$

Next, we will look at determining whether a function class is Glivenko-Cantelli. To do this, we will only look at the function evaluations, rather than the functions themselves:

$$\mathcal{F}(x_1^n) \triangleq \{(f(x_1), f(x_2), ..., f(x_n))\ ;\ f \in \mathcal{F}\} \subseteq \mathbb{R}^n \tag{8.23}$$

Intuitively, if the function only takes a few values, that it is more likely that the maximum deviation between the expected value and the empirical average will be small. We recall the definition of the *Rademacher Complexity*:

$$R(S) \triangleq \mathbb{E}_\epsilon \left[\sup_{a \in S} \left| \sum_{i=1}^n \epsilon_i a_i \right| \right] \tag{8.24}$$

If the set S is small, then it is unlikely that we can find a vector $a \in S$ that has high correlation with the noise vector ϵ. As we increase the size of S, we expect that it will be more likely to find a vector in S with high correlation.

We now will look at the Rademacher complexity of the set of function evaluations. The empirical Rademacher complexity is

$$R\left(\frac{\mathcal{F}(x_1^n)}{n}\right) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right| \right] \tag{8.25}$$

We can also look at the population Rademacher complexity by taking an expectation over the samples. This quantity is also called the Rademacher complexity of the function class $\mathcal{F}$:

$$R_n(\mathcal{F}) = \mathbb{E}_{x_1^n} \left[R\left(\frac{\mathcal{F}(x_1^n)}{n}\right) \right] \tag{8.26}$$

Theorem 8.2 *Let a function class $\mathcal{F}$ that is b-uniformly bounded (i.e. $\|f\|_\infty \le b,\ \ \forall f \in \mathcal{F}$) be given. Then, for all $n \ge 1, \delta \ge 0$, we have*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \le 2R_n(\mathcal{F}) + \delta \tag{8.27}$$

with probability at least $1 - \exp(\frac{-n\delta^2}{2b^2})$.

An immediately corollary of this theorem is that if the Rademacher complexity $R_n(\mathcal{F})$ converges to zero, then the function class $\mathcal{F}$ is a GC class.

Theorem 8.3 *Let a b-uniformly bounded function class $\mathcal{F}$ be given. Then, for all $n \ge 1, \delta \ge 0$, we have*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \ge \frac{1}{2} R_n(\mathcal{F}) - \sup_{f \in \mathcal{F}} \frac{|\mathbb{E}[|f|]|}{2\sqrt{n}} - \delta \tag{8.28}$$

with probability at least $1 - \exp(\frac{-n\delta^2}{2b^2})$

Taken together, these results say that the Rademacher complexity gives us both upper and lower bounds for the maximum deviation. Thus, we need to find a way to bound the Rademacher complexity.

8.2 Polynomial Discrimination

Definition: A function class $\mathcal{F}$ has *polynomial discrimination* on the order $v \geq 1$ if, for all $x_1^n \in \mathcal{X}^n$, we have the following bound on the cardinality of function evaluations:

$$\text{card}(\mathcal{F}(x_1^n)) \leq (n+1)^v \tag{8.29}$$

Note that $\mathcal{F}(x_1^n)$ is a set containing length-n vectors. We are counting the number of unique vectors in this set. For example, if each function $f \in \mathcal{F}$ is binary, then there are at most 2^n bit vectors, so

$$\text{card}(\mathcal{F}(x_1^n)) \leq 2^n \tag{8.30}$$

Noting that 2^n is exponential in n, not polynomial, we see that arbitrary binary functions are not polynomial discriminable.

Theorem 8.4 *Let a function class $\mathcal{F}$ that is polynomial discriminable with order v be given. Then we can bound the Rademacher complexity of $\mathcal{F}$ as follows:*

$$R_n(\mathcal{F}) \leq 2\left(\mathbb{E}_{x_1^n}[D(x_1^n)]\right)\sqrt{\frac{v\log(n+1)}{n}} \qquad \textit{where} \qquad D(x_1^n) \triangleq \sup_{f\in\mathcal{F}}\sqrt{\frac{1}{n}\sum_{i=1}^n f^2(x_i)} \tag{8.31}$$

Example: Let's look at the class of CDFs, $\mathcal{F} = \{\mathbb{1}(x \in (-\infty, t])\ ;\ t \in \mathbb{R}\}$. Let's further assume that our samples x_1^n are sorted:

$$x_1 \leq x_2 \leq \cdots \leq x_n \tag{8.32}$$

For a fixed t, we know that $\mathbb{1}(x \in (-\infty, t])$ will be 1 for small i and 0 for large i. Thus, there are $n+1$ possible values for the vector $\mathbb{1}(x_1^n \in (-\infty, t])$.