6.853 Algorithmic Game Theory and Data Science April 18, 2019

Lecture 18

Lecturer: Costis Daskalakis *Scribe: Govind Ramnarayan*

1 Recap and Motivation

So far, we have discussed how to learn good auctions given sample access to value distributions. This has some issues. One, it is often impractical. Two, typically in a non-truthful auction we can only observe the *bids*, and not the *values*. Hence, to find the underlying value distribution, we would need to "invert" the bid distribution by assuming that the bidders are bidding according to an equilibrium and using equilibrium conditions (which we will see later in the course).

However, it is clear that to do this we need algorithms that learn bid distributions from samples. In this lecture, we will learn about how to learn CDFs from samples. In the following two lectures, we will see how we can also learn PDFs.

2 Learning CDFs

Throughout this lecture, we will use F to denote a CDF, and f to denote its corresponding PDF. Concretely, today's goal is the following.

Goal 1. *Given samples from a single-dimensional CDF F, find $\hat{F}$ such that $|F(z) - \hat{F}(z)| < \varepsilon$ for all z.*

Let us first consider the following "natural" estimator. Given samples $X_1, \ldots, X_n \sim F$ (where we recall that F is a CDF), we consider the "empirical CDF" $\hat{F}$, where

$$\hat{F}(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq z\}} \tag{1}$$

Lecture today will be focused on analyzing the performance of this estimator. We will look at two different approaches for the analysis. The first approach will be based on Rademacher complexity, and the second approach will improve the Rademacher complexity bound using a technique called "chaining."

2.1 First Approach: Rademacher Complexity

Let us consider the hypothesis class $\mathcal{H}$ consisting of all *indicators of half-lines*. More formally, let $\mathcal{H}$ be defined as

$$\mathcal{H} = \{\mathbf{1}_{\{x \leq z\}} \forall z \in \mathbb{R}\} \cup \{\mathbf{1}_{\{x \geq z\}} \forall z \in \mathbb{R}\} \tag{2}$$

For notational convenience going forward, we will define $h_{\leq z} \stackrel{def}{=} \mathbf{1}_{\{x \leq z\}}$ and similarly $h_{\geq z} \stackrel{def}{=} \mathbf{1}_{\{x \geq z\}}$. Similarly, we will denote the vector of samples $X_1, \ldots, X_n$ as X_1^n.

We state the following straightforward claim.

Claim 1. *Let X_i be the random samples used in the construction of the empirical CDF $\hat{F}$. Then*

$$sup_x |F(x) - \hat{F}(x)| = sup_{h \in \mathcal{H}} \left(\mathbb{E}_{X \sim F} [h(X)] - \frac{1}{n} \sum_{i=1}^{n} h(X_i) \right) \tag{3}$$

Proof. Fix a value z. Using the definition of the halfline indicator $h_{\leq z}$, we observe that

$$\begin{aligned} F(z) - \hat{F}(z) &= \Pr_X[f(X) \leq z] - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq z} \\ &= \mathbb{E}_{X\sim F}[h_{\leq z}(X)] - \frac{1}{n}\sum_{i=1}^{n} h_{\leq z}(X_i) \end{aligned}$$

Additionally, we note that flipping $h_{\leq z}$ to $h_{\geq z}$ in the above expression negates it. That is:

$$\mathbb{E}_{X\sim F}[h_{\geq z}(X)] - \frac{1}{n}\sum_{i=1}^{n} h_{\geq z}(X_i) = -1 \cdot \left(\mathbb{E}_{X\sim F}[h_{\leq z}(X)] - \frac{1}{n}\sum_{i=1}^{n} h_{\leq z}(X_i)\right)$$

and so putting these facts together gives us that

$$|F(z) - \hat{F}(z)| = max_{h\in\{h_{\leq z}, h_{\geq z}\}} \left(\mathbb{E}_{X\sim F}\left[h(X)\right] - \frac{1}{n}\sum_{i=1}^{n} h(X_i)\right)$$

which immediately yields the claim. □

So, to analyze the quality of the empirical CDF $\hat{F}$, it suffices to upper bound the right hand side of the equation in Claim 1 on expectation over the samples $X_1, \ldots, X_n \sim F$ using Rademacher complexity. Recall that we have previously used Rademacher complexity to establish that the true loss is close to the empirical loss whenever our hypothesis h comes from a sufficiently simple hypothesis class[1]. In this case, we have no real notion of "loss"; however, we want to bound the expectation of a function h under the true distribution F with the empirical average of h, over all h in some hypothesis class. Hence, despite the absence of a "loss" in this problem, we can still use the machinery of Rademacher complexity (e.g. by considering $\ell(h, z) := h(z)$ in the established framework).

Hence, by applying the Rademacher complexity result from Lecture 15 (Lemma 26.2, [1]), we get that

$$\begin{aligned} \mathbb{E}_{X_1^n\sim F}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{X\sim F}[h(X)] - \frac{1}{n}\sum_i h(X_i)\right)\right] &\leq \mathbb{E}_{X_1^n\sim F}\left[R(\mathcal{H}, X_1^n)\right] \\ &= \mathbb{E}_{X_1^n\sim F}\left[\frac{2}{n}\cdot\mathbb{E}_\sigma\left(sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i h(X_i)\right)\right] \qquad (4) \end{aligned}$$

Now we wish to upper bound (4). To do this, note that for any fixed samples $X_1, \ldots, X_n$, there are $\leq 4n+2$ relevant hypotheses h in $\mathcal{H}$, one of which achieves the supremum for any σ. To see this, note that since $\mathcal{H}$ consists of indicators of half-lines, the only features of $h_{\leq z}$ that are relevant are where z falls in relation to the X_i's. Since the X_i's are all real numbers, we have just $2n+1$ cases:

1. $z = X_i$ for some i (n cases), or
2. z is between two adjacent samples X_i and X_j, in which case wlog we suppose that $z = (X_i + X_j)/2$ ($n-1$ cases), or
3. $z = \pm\infty$ (2 cases)

Combining the possible functions $h_{\leq z}$ and $h_{\geq z}$ yields the claim that we have at most $4n+2$ possibilities overall. Hence, we have a good bound on the size of the hypothesis class, and can apply Massart's Lemma to get

$$(4) \leq \sqrt{\frac{2\log(4n+2)}{n}}$$

[1] See Lecture 15.

and combining with Claim 1, we get that

$$\mathbb{E}_{X_1^n \sim F}[sup_x|F(x) - \hat{F}(x)|] \leq \sqrt{\frac{2\log(4n+2)}{n}}$$

We furthermore note that by applying concentration inequalities to the Rademacher complexity, we can strengthen this expectation bound to a high probability bound, i.e.

$$\Pr_{X_1^n \sim F}\left[sup_x|F(x) - \hat{F}(x)| > \sqrt{\frac{2\log(4n+2)}{n}} + \varepsilon\right] \leq e^{-2n\varepsilon^2}$$

2.2 Second Approach: Dudley's Chaining

We now outline an improvement that can be achieved via a more sophisticated technique, chaining. The end result will be the famous Dvoretzky-Kiefer-Wolfowitz (or DKW) Inequality:

$$\mathbb{E}_{X_1,\ldots,X_n}[sup_x|F(x) - \hat{F}(x)|] \leq \sqrt{\frac{c}{n}} \tag{5}$$

First, we need to set up some definitions. Suppose $\mathcal{A}$ is a set of subsets of $\mathbb{R}^d$. Take $\mathcal{H}$ to be the set of indicator functions for these sets; that is, $\mathcal{H} = \{\mathbf{1}_{x\in A}, \forall A \in \mathcal{A}\}$.

Given a set of subsets $\mathcal{A}$ and points $x_1, \ldots, x_n \in \mathbb{R}^d$, we can define a set of bit vectors as follows:

$$\mathcal{A}(x_{1:n}) = \{b = (b_1, \ldots, b_n) \in \{0,1\}^n : \exists A \in \mathcal{A} : b_i = \mathbf{1}_{x_i \in A}, \forall i \in [n]\}$$

Intuitively, the set $\mathcal{A}(x_{1:n})$ corresponds to the set of all possible intersections of a set $A \in \mathcal{A}$ with the points $(x_1, \ldots, x_n)$.

String distance: Next, we define a notion of distance between bit strings. Given two bit strings $b, c \in \{0,1\}^n$, define their distance by

$$\rho(b,c) = \sqrt{\frac{1}{n}\sum_i \mathbf{1}_{b_i \neq c_i}}$$

You may notice that ρ is just the square root of the normalized Hamming distance on strings.

Cover: Given a subset $B \subseteq \{0,1\}^n$ and radius r, B_r is an r-cover of B if and only if for all $b \in B$, there exists $b' \in B_r$ such that $\rho(b,b') \leq r$.

Covering Number: Let $N(r,B)$ denote the size of the smallest r-cover of B.

Also, let $F(A) = \mathbb{E}_{X\sim F} h_A(X)$ and let $\hat{F}(A) = (1/n)\sum_i h_A(X_i)$ (note that in the case where the sets A are intervals $(-\infty, z)$, these actually do correspond to $F(z)$ and $\hat{F}(z)$ respectively).

Finally we are ready to give Dudley's Chaining Theorem.

Theorem 1 (Dudley's Chaining)**.**

$$\mathbb{E}_{X_1,\ldots,X_n}[sup_{A\in\mathcal{A}}|\hat{F}(A) - F(A)|] \leq \frac{24}{\sqrt{n}} \cdot \max_{x_1,\ldots,x_n \in \mathbb{R}^d} \int_0^1 \sqrt{\log(2N(r, \mathcal{A}(x_{1:n})))}dr$$

Proof. To be given as an exercise. □

Now we outline the proof that Chaining implies the DKW inequality. We restate the DKW theorem.

Theorem 2.

$$\mathbb{E}_{X_1^n \sim F}[sup_x|F(x) - \hat{F}(x)|] \leq \sqrt{\frac{c}{n}}$$

We also note that this can be strengthened to a high probability bound, like we did in Lectures 14 and 15 with PAC learning problems, to get that

$$\Pr[sup_x|F(x) - \hat{F}(x)| > \sqrt{\frac{c}{n}} + \varepsilon] \leq e^{-2n\varepsilon^2}$$

Proof. The proof uses chaining. Let $\mathcal{A} = \{(-\infty, z) : z \in \mathbb{R}\}$, and fix samples $x_1, \ldots, x_n \in \mathbb{R}$, and assume wlog that $x_1 \leq x_2 \leq \ldots \leq x_n$.

In order to apply Dudley's Chaining Theorem to solve the problem, we need a good bound on the covering number $N(r, \mathcal{A}(x_{1:n}))$. To get this, we will describe how to cover $\mathcal{A}(x_{1:n})$ for a desired radius r.

Observe that $\mathcal{A}(x_{1:n})$ consists of strings of the form $(1, 1, \ldots, 1, 0, 0, \ldots, 0)$, and so $|\mathcal{A}(x_{1:n})| \leq n + 1$. Fix $r \in (0, 1)$, and let $k := \lfloor \frac{n}{r^2} \rfloor$. We will cover $\mathcal{A}(x_{1:n})$ by only using strings of the form $\underbrace{11\cdots11}_{\text{multiple of } k}\ 00\cdots00$.

So the cover B_r that we construct contains $n/k = n/(\lfloor nr^2 \rfloor)$ strings. Hence, we get that

$$N(r, \mathcal{A}(x_{1:n})) \leq \frac{n}{\lfloor nr^2 \rfloor} \leq \frac{1}{r^2} + 1$$

And so, we get that

$$\begin{aligned}\int_0^1 \sqrt{\log(2N(r, \mathcal{A}(x_{1:n})))}dr &\leq \int_0^1 \sqrt{\log(\frac{2}{r^2} + 2)}dr \\ &\leq \int_0^1 \sqrt{\log(\frac{4}{r^2})}dr \\ &\leq \sqrt{2\pi}\end{aligned}$$

and we conclude the result by Theorem 1. □

Note that this improves on the vanilla Rademacher bound (by shaving off the root-log factor)!

2.3 A Quick and Dirty Application of DKW to Auctions

Suppose we have one item, and one bidder whose value is $v \sim F$, where F has an unknown support $[0, H]$. Our goal is to take samples from F, then find an auction with revenue $OPT - O(\varepsilon \cdot H)$. To do this efficiently, we will use the DKW inequality.

By the DKW inequality, we can use $O(1/\varepsilon^2)$ samples to find an empirical CDF $\hat{F}$ such that

$$sup_x |F(x) - \hat{F}(x)| \leq \varepsilon \tag{6}$$

We can now use the empirical CDF $\hat{F}$ to compute the price; that is,

$$\hat{p} = argmax_{x \in [0,H]}(x(1 - \hat{F}(x))) \tag{7}$$

By Myerson's Theorem, we know that the optimal revenue OPT is achieved by posting the price

$$p^* = argmax_{x \in [0,H]}(x(1 - F(x)))$$

The question is, how high is the revenue from posting our computed price $\hat{p}$ compared to the optimal revenue from posting p^*?

$$\begin{aligned}\text{Revenue from } \hat{p} &= \hat{p}(1 - F(\hat{p})) \\ &\overset{(6)}{\geq} \hat{p}(1 - \hat{F}(\hat{p}) - \varepsilon) \\ &= \hat{p}(1 - \hat{F}(\hat{p})) - \varepsilon\hat{p} \\ &\overset{(7)}{\geq} p^*(1 - \hat{F}(p^*)) - \varepsilon\hat{p} \\ &\overset{(6)}{\geq} p^*(1 - F(p^*) - \varepsilon) - \varepsilon\hat{p} \\ &= p^*(1 - F(p^*)) - \varepsilon(p^* + \hat{p}) \\ &= OPT - 2\varepsilon H\end{aligned}$$

so $O(1/\varepsilon^2)$ samples suffices to get $OPT - O(\varepsilon \cdot H)$ revenue, improving upon the bound from last time!

References

[1] Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms. *Cambridge University Press, 2014*.