Learning from Samples

Wednesday, March 22, 2017 8:50 AM

▷ Last time we saw:

- Vickrey with a different reserve for each player is a 2-approx. to Myerson in non i.i.d. regular environments.
- Vickrey with a single reserve $r \sim F$ is a $\left(1-\frac{1}{n+1}\right)$-approx for i.i.d regular where each $v_i \sim F$.

▷ So if we are given a single sample r from the value distribution then we can construct a mechanism M_r s.t.

$$\mathbb{E}_{r, \vec{v} \sim F^n}\left[Rev_{M_r}(\vec{v}) \right] \geq \frac{1}{2} \underbrace{\sup_{M} \mathbb{E}_{\vec{v} \sim F^n}\left[Rev_M(\vec{v}) \right]}_{\substack{\parallel \\ OPT(F) = \mathbb{E}_{\vec{v} \sim F^n}\left[Rev_{Myerson}(\vec{v}) \right]}}$$

▷ What if we had more samples?

$$S = \{r_1, \dots, r_m\} \sim F^m$$

Can we give a mechanism M_S such that:

$$\mathbb{E}_{S, \vec{v}}\left[Rev_{M_S}(\vec{v}) \right] \geq OPT(D) - \epsilon(m)$$

with $\epsilon(m) \to 0$?

▷ What about w.h.p. i.e. w.p. $1-\delta$

over S:

$$E_{\vec{v}}\left[Rev_{M_S}(\vec{v}) \right] \geq OPT(D) - \epsilon(m, \delta)$$

with $\epsilon(m, \delta) \to 0$ as $m \to \infty$ for any fixed δ.

▷ In the non-iid setting: what if we had samples from bid vectors of players in past auctions, i.e.

if $F = F_1 \times \dots \times F_n$ and

$$S = \{ \vec{b}^1, \dots, \vec{b}^m \} \sim F^m$$

$$\vec{b}^1 = (b_1^1, \dots, b_1^n)$$

Then can we construct a mechanism M_S s.t.

$$E_{S \sim F^m, \vec{v} \sim F}\left[Rev_{M_S}(\vec{v}) \right] \geq OPT(F) - \varepsilon(m)$$

or. w.p. $1-\delta$ over S:

$$E_{\vec{v} \sim F}\left[Rev_{M_S}(\vec{v}) \right] \geq OPT(F) - \varepsilon(m, \delta)$$

▷ This is exactly the type of question that PAC learning theory asks

Probably Approximately Correct (Valiant '84)

▷ The general setting of learning was first

analyzed by Vapnik and Chervonenkis '71
(Vapnik '82, 92, 95, 98)

- We have a hypothesis space $\mathcal{H}$ over which we want to learn and optimize
  (the space of dominant strategy truthful mechanisms)
- We have a distribution D over data points $z \in Z$
  (the distribution F over bid/value vectors $\vec{b} \in [0, H]^n$)
- We have an objective function in mind: (loss or reward) we will transition to losses

$$\ell : \mathcal{H} \times Z \to \mathbb{R}$$

  (negative of the revenue of a mechanism M on a value vector $\vec{b}$)

- We want to optimize expected loss

$$\inf_{h \in \mathcal{H}} \mathop{E}_{z \sim D} [\ell(h, z)] = OPT(D)$$

  Known under many names: generalization error, risk, true error

- We have a set of iid samples from D, $S = \{z_1, \ldots, z_n\}$
  (the sample bid vectors).

Goal | Given S find a hypothesis h_S

s.t. w.p. $1-\delta$ over the sample draws:

$$\underbrace{\mathop{E}_{z\sim D}\left[\ell(h_S, z)\right]}_{\text{Denote as } L_D(h_S)} \le OPT(D) + \varepsilon$$

<u>Sample Complexity</u>: For any ε, δ is there an $m(\varepsilon,\delta)$ s.t. $\exists$ alg. that with $m > m(\varepsilon,\delta)$ sample set S, we guarantee that w.p. $1-\delta$: $L_D(h_S) \le OPT(D)+\varepsilon$

$m(\varepsilon,\delta)$ is the sample complexity of the problem.

Ideally, we want it to be $\mathrm{poly}\left(\frac{1}{\varepsilon}, \log(1/\delta)\right)$
Ideally, we want the algorithm to be poly time computable
→ Valiant's definition. '84

▷ Now we will just ask: is $m(\varepsilon,\delta)$ finite for all ε,δ. Then we will say that the problem is <u>learnable</u>.

<u>Candidate algorithm</u>

- Since samples are drawn from the target distribution, the average loss on the samples is a good proxy for the expected loss:

$$\mathcal{L}_S(h) = \frac{1}{m} \sum_{t=1} \ell(h, z_t)$$

- So why not just output the hypothesis h_S which minimizes $\mathcal{L}_S(h)$

$$h_S = \arg\inf_{h \in H} \mathcal{L}_S(h)$$

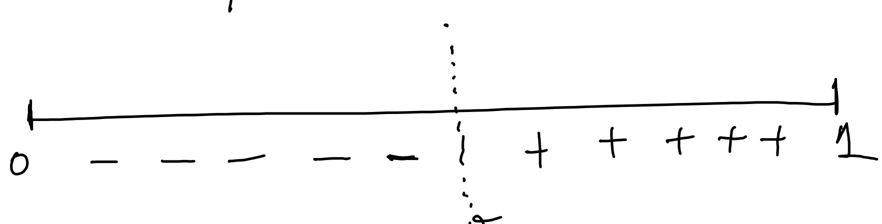
- If $\underbrace{\mathcal{L}_S(\cdot)}$ is close to $\underbrace{\mathcal{L}_D(\cdot)}$
  - loss on samples
  - training error
  - empirical loss

  - loss on true dist.
  - generalization error
  - true error

  then $\mathcal{L}_D(h_s)$ should also be small
- This algorithm is called Empirical Risk Minimization or ERM.
- It is equivalent to the Follow-the-Leader Algorithm!

ERM gone bad

- Let's look at a classification problem in 1-d.
- $z = (x, y) \qquad x \in [0,1], \quad y \in \{0,1\}$



- $x \sim U(0,1) \qquad y \mid x$ is 1 if $x \geq \frac{1}{2}$ and 0 o.w.
- H: space of all mappings from $[0,1] \to \{0,1\}$
- Loss: Prediction error; $\ell(h,(x,y)) = \mathbb{1}\{h(x)=y\}$

- ERM:

$$h_S = \arg\inf_{h \in \mathcal{H}} \frac{1}{m} \sum_{t=1}^{m} \mathbb{1}\{h(x_t) = y_t\}$$

- For instance: (memoization)

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] \text{ s.t. } x_i = x \\ 0 & \text{o.w.} \end{cases}$$

- Obviously: $\mathcal{L}_S(h_S) = 0$ (zero loss on sample)

- But:

$$L_D(h_S) = \mathbb{E}_{(x,y)\sim D}\left[\ell(h_S, (x,y))\right] = \frac{1}{2}$$

since h_S predicts 1 only on a finite (zero-measure) x's

- So we will never achieve $\varepsilon < \frac{1}{2}$ error no matter how large m is!!
- ERM fails! Overfitting to samples!

▶ Is it a problem of ERM??

- Not in this case.
- You can show that if the space $\mathcal{X}$ where the x "lives" is infinite, then the hypothesis space $\mathcal{H}$ of all functions from $\mathcal{X} \to \{0,1\}$ is not <u>learnable</u> by <u>any algorithm</u>
- We need to restrict $\mathcal{H}$!!
- For instance: if $\mathcal{H}$ was the space of all threshold functions,

$$\mathcal{H} = \left\{ f_\theta(x) = \begin{cases} 1 & \text{if } x \geq \theta \\ 0 & \text{o.w.} \end{cases} \text{ for } \theta \in [0,1] \right\}$$

Then this would fix our problem (see

at the end of next class).

Learnability from uniform convergence

▷ More generally the problem was that $|\mathcal{L}_S(h) - \mathcal{L}_D(h)|$ was not uniformly close for all h.

Uniform Convergence

- Suppose that $\forall \varepsilon, \delta$: $\exists\, m_{\mathcal{U}}(\varepsilon, \delta)$ s.t. $\forall\ m > m_{\mathcal{U}}(\varepsilon, \delta)$, w.p. $1-\delta$

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| \leq \varepsilon$$

- i.e. we know that the worst case error of the empirical loss goes to zero uniformly over all hypotheses.

Thm Uniform Convergence $\Rightarrow$ Learnability

Pf

Let: $h_D^* = \arg\inf_{h \in \mathcal{H}} \mathcal{L}_D(h)$

Consider $m \geq m_{\mathcal{U}}(\frac{\varepsilon}{2}, \delta)$, then w.p. $1-\delta$

$$L_D(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(h_D^*) + \frac{\varepsilon}{2}$$

$$L_D(h_S) \le L_S(h_S) + \frac{\varepsilon}{2} \le L_S(h_D^*) + \frac{\varepsilon}{2}$$
$$\le L_D(h_D^*) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2}.$$

this is where our example failed
Empirical loss deceivingly small.

Finite $\mathcal{H}$ and Uniform Convergence

Thm If $\mathcal{H}$ is finite and $\ell(h,\delta) \in [\alpha, b]$

$$m_{\mathcal{U}}(\varepsilon, \delta) \le \left\lceil \frac{\log(2|\mathcal{H}|/\delta)(b-\alpha)^2}{2\varepsilon^2} \right\rceil$$

so $\mathcal{H}$ is learnable and with sample complexity $\mathrm{poly}\left(\frac{1}{\varepsilon}, \log(1/\delta), \log(|\mathcal{H}|)\right)$

Pf For a given m, we want to show:

$$P\left(\sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| \le \varepsilon\right) \ge 1 - \delta$$

Equiv.

$$P\left(\sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| > \varepsilon\right) < \delta$$

Equiv

<u>Equiv</u>

$$P\left(\exists h \in \mathcal{H} : |\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon\right) < \delta$$

<u>By union bound</u>

$$P(\exists h \in \mathcal{H} : |\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon) \leq$$

$$\sum_{h \in \mathcal{H}} \underbrace{P(|\mathcal{L}_S(h) - \mathcal{L}_D(h)| > \varepsilon)}_{\text{Let's bound this!}}$$

<u>For any fixed h:</u>

$$L_D(h) = \mathbb{E}_{z \sim D}\left[\ell(h, z)\right]$$

$$L_S(h) = \frac{1}{m} \sum_{t=1}^{m} \underbrace{\ell(h, z_t)}_{\substack{\text{Random variable with} \\ \mathbb{E}_{z_t \sim D}[\ell(h,z)] = \mathcal{L}_D(h).}}$$

- So $L_S(h)$ is the average of i.i.d. random variables each with expected value $L_D(h)$.
- We need to bound the prob. that this average deviates a lot from its mean!

(similar to Law of Large Numbers but not asymptotic).

Hoeffding's Inequality

Let $\theta_1, \dots, \theta_m$ be i.i.d random variables with $\theta_i \in [a, b]$ and $\mathbb{E}[\theta_i] = \mu$. Then

$$\Pr\left[\left| \frac{1}{m} \sum_{t=1}^{m} \theta_t - \mu \right| > \varepsilon \right] \le 2 \exp\left(- \frac{2m\varepsilon^2}{(b-a)^2} \right)$$

(one of the most useful inequalities in any large deviation analysis)

Back to main thm. we get:

$$P\left(\left| L_S(h) - L_D(h) \right| > \varepsilon \right] \le 2 \exp\left(- \frac{2m\varepsilon^2}{(b-a)^2} \right)$$

Picking $m = \dfrac{\log(2|H|/\delta)\,(b-a)^2}{2\varepsilon^2}$

above is at most: $\dfrac{\delta}{|H|}$

So when we sum over H we get at most δ. $\blacksquare$

Bonus pf of Hoeffding

Bonus pf of Hoeffding

Consider $Z_i = \theta_i - \mu_i \in [a,b]$ Show:

$$P\left(\left|\frac{1}{m}\sum_t Z_t\right| > \varepsilon\right) \le 2\exp\left(-\frac{2m\varepsilon^2}{(b-a)^2}\right)$$

$$P\left(\frac{1}{m}\sum_t Z_t > \varepsilon\right) = P\left(e^{\frac{\lambda}{m}\sum_t Z_t} > e^{\lambda\varepsilon}\right)$$

$$\le \frac{\mathbb{E}\left[e^{\frac{\lambda}{m}\sum_t Z_t}\right]}{e^{\lambda\varepsilon}} \quad \text{(Markov)}$$

$$= \frac{\mathbb{E}\left[\prod e^{\frac{\lambda}{m} Z_t}\right]}{e^{\lambda\varepsilon}}$$

$$= \frac{\prod \mathbb{E}\left[e^{\frac{\lambda}{m} Z_t}\right]}{e^{\lambda\varepsilon}} \quad \text{(independence)}$$

For any $X \in [a,b]$: $\mathbb{E}[e^X] \le e^{\frac{(b-a)^2}{8}}$

Apply to $X = \frac{\lambda}{m} Z_t$:

$$\le \frac{e^{\frac{\lambda^2 (b-a)^2}{8m}}}{e^{\lambda\varepsilon}}$$

$$= e^{\frac{\lambda^2 (b-a)^2}{8m} - \lambda\varepsilon}$$

Pick $\lambda = \frac{4m\varepsilon}{(b-a)^2}$?

Pick $\lambda = \frac{4m\varepsilon}{(b-a)^2}$

$$\leq e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$$

$\blacksquare$

Intuitition! e^X random variable punishes very large deviations of X above zero. So if we can upper bound $\mathbb{E}[e^X]$ we should be able to show great concentration bounds for X.

Independence decouples $\mathbb{E}[e^X] = \left(\mathbb{E}[e^{\frac{1}{m}X_i}]\right)^m$

if $X = \frac{1}{m}\sum_i X_i$

So a bound on $\mathbb{E}[e^{X_i}]$ is amplified by raising it to the m-th power.