Today's topic is auction with samples.

1 Introduction to auctions

Definition 1. *In a single item auction, there are k bidders bid for one item, auction design is the problem of designing a truthful mechnism to maximize certain objective.*

For different objectives, there are different results.

- Social Welfare (benefit the society): VCG mechanism will be optimal. In single item case, it will be the second price auction. It is prior-free, i.e., does not require any assumption about the distribution of bidders' value (not the cretiria of this lecture).

- Revenue (benefit the auctioneer): Myerson's algorithm (1981) is optimal for average case. In this case, there is guarantee only for Bayesian setting, i.e., we assume the bidders' values v_i are drawn from certain distribution, $(v_1, v_2, \ldots, v_k) \sim \mathcal{F}$.

In this lecture, we will discuss revenue maximization. Also we assume the bidders' values are independenlty drawn from $\mathcal{F}_i$, i.e. $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \mathcal{F}_k$.

The Myerson's auction can be phrased as VCG mechanism that maximizes virtual values. The virtual value of a bidder with value v is defined to be $\phi(v) = v - \frac{1-F(v)}{f(v)}$ when the distribution with c.d.f. F, p.d.f. f is *regular* ($\phi(v)$ is strictly-increasing).

The main goal will be understanding the empirical version of Myerson's auction, namely, instead of knowing the distribution $\mathcal{F}$, we can only draw samples from it.

This problem is hard in general without assumptions. For example, on a point with a large value on which the distribution assigns a very tiny probability, with certain number of samples we cannot observe the large value and thus cannot provide any guarantees.

Therefore, we need assumptions about the class of distributions.

2   Auction with Samples, First natural approach

There are two criteria for auctions with samples:

- Sample complexity, i.e., the number of samples required.

- Computational complexity, i.e., the running time of the algorithm.

Information theory provides a lower bound of sample complexity $S(n, \epsilon, \mathcal{D})$. Here $\mathcal{D}$ is the class which we assume $\mathcal{F}$ is in.

The algorithm we designed should have sample complexity in $\text{poly}(S)$ and computational complexity in $\text{poly}(S)$.

Open problem: efficient algorithm that has sample complexity $\tilde{O}(S)$.

Here we have our first approach:

1. Take samples from each $\mathcal{F}_i$;

2. Get the empirical distribution $\bar{\mathcal{F}}_i$;

3. Run Myerson's auction with $\bar{\mathcal{F}}_i$.

This approach does not work but a variant of it will.

3   Learning distributions

The problem of learning distributions is defined as:

Definition 2. *Given a family of distributions D, and i.i.d. samples from an distribution $p \in D$. Output a distribution h, s.t. with high probability $d(h, p) \leq \epsilon$. Here $d(h, p)$ measures the distance between two distributions.*

Since the definition is generic, there are three problems we need to address.

- The distance functions we will discuss: total variation distance, Kolmogorov distance.

- Number of samples required $m(\epsilon, D, d)$.

- For the "with high probability", the randomness is from the samples and the algorithm.

3.1 Learning discrete distribution subject to total variation distance

We assume D to be all discrete distributions over domain $[n]$.

For two discrete distributions with p.m.f. p, q, the total variation distance between them is defined as

$$d_{tv} = \max_{A \in [n]} |p(A) - q(A)| = \frac{1}{2} |p - q|_1$$

With m samples $S_1, \ldots, S_m$, the optimal algorithm is just using the empirical distribution, let $N_i = |\{j \in [m] | S_j = i\}|$:

$$\hat{p}_m(i) = \frac{N_i}{m}.$$

When m is large enough, we will have $d_{tv}(p, \hat{p}_m) \leq \epsilon$. To find a lower bound of m, since $d_{tv}(p, \hat{p}_m)$ is a random variable, we can first find the m that makes $\mathbb{E}[d_{tv}(p, \hat{p}_m)] \leq \epsilon/C$, then use Markov inequality to bound the tail probability:

$$\mathbf{Pr}[d_{tv}(p, \hat{p}_m) > \epsilon] \leq \frac{1}{C}.$$

In the following proof, we will ignore the constant C since it is just a scale factor of ϵ.

Theorem 3. *When $m \in \Omega(\frac{n}{\epsilon^2})$, the expected total variation distance between p and $\hat{p}_m$ is upper bounded by ϵ.*

Proof. Since $N_i \sim Binom(m, p(i))$, we have $\mathbb{E}[N_i] = mp(i), \mathrm{Var}[N_i] = mp(i)(1 - p(i))$.

$$
\begin{aligned}
\mathbb{E}[d_{tv}(p, \hat{p}_m)] &\leq \frac{1}{m} \sum_{i=1}^{n} \mathbb{E}[|N_i - mp(i)|] \\
&\leq \frac{1}{m} \sum_{i=1}^{n} \sqrt{\mathbb{E}[(N_i - mp(i))^2]} \\
&= \frac{1}{\sqrt{m}} \sum_{i=1}^{n} \sqrt{p(i)(1 - p(i))} \\
&\leq \frac{1}{\sqrt{m}} \sum_{i=1}^{n} \sqrt{p(i)} \\
&\leq \sqrt{\frac{n}{m}}
\end{aligned}
$$

The second and last inequality is due to Cauchy-Schwarz inequality.

When $m \geq \frac{n}{\epsilon^2}$, we have $\mathbb{E}\left[d_{tv}(p, \hat{p}_m)\right] \leq \epsilon$. $\hfill\square$

The bound of $\frac{n}{\epsilon^2}$ is tight, here we provide a constructive proof.

Proof. We will construct a distribution family D by perturbing the uniform distribution over $[n]$. For each pair $(2i - 1, 2i)$, set their probability to be either $(\frac{1-\epsilon}{n}, \frac{1+\epsilon}{n})$ or $(\frac{1+\epsilon}{n}, \frac{1-\epsilon}{n})$. Thus there will be $2^{n/2}$ different distributions in the family.

To learn a distribution h from samples of $p \in D$, s.t. $d_{tv}(p, h) \leq \epsilon/4$, we need to find the correct perturbing direction for at least $\frac{n}{4}$ pairs. Finding the correct perturbation direction for each pair requires $O(\frac{1}{\epsilon^2})$ samples (by information theory). Thus in total, at least $O(\frac{n}{\epsilon^2})$ samples are required. $\hfill\square$

When n goes to infinity, there is no chance to use finite number of samples to learn a distribution and bound the total variation distance.

However, for certain family of distributions, we can learn distributions subject to the total variation distance less than ϵ, with sample complexity $poly(1/\epsilon, \log n)$

- Distribution class $\mathcal{D}$ is all log-concave distributions over $[n]$. Then we can learn with $\Theta(1/\epsilon^{5/2})$.

- The *hazard rate* of a distribution is $f(x)/(1 - F(X))$ where f is the PDF of the distribution and F is the CDF. If we have monotone none-decreasing hazard rate, we can learn it with $O(log(n)/\epsilon^3)$ samples.

If I make no assumptions about the distribution, to get error probability δ I need $\frac{1}{\delta}\frac{n}{\epsilon^2}$ samples using our previous error analysis and markov's inequality. With more careful analysis, we can show that we only need

$$\Theta\left(\frac{n + \log\frac{1}{\delta}}{\epsilon^2}\right)$$

4   Learning distribution subject to Kolmogorov Distance

Instead of making stronger assumptions on the distribution to achieve sample complexity independent of sample support, we turn towards changing the metric we are using to learn from total variation distance to *Kolmogorov Distance.*

We have a random variable X supported over $\mathbb{R}$. Let F be the CDF of X.

$$F(u) = Pr[X \leq u]$$

Learning using Kolmogorov Distance means

$$d_K(X, Y) = \max_{u \in \mathbb{R}} |F_X(u) - F_Y(u)|$$

Theorem 4. *DKW Inequality For any X over $\mathbb{R}$, we can learn with $O(\frac{1}{\epsilon^2})$ up to $d_K(X, Y) \leq \epsilon$.*

Proof. Consider the empirical. $\hat{F}(u) = \frac{1}{m} \sum_{i=1}^{n} \mathbb{1}\{x_i \leq u\}$

$$\max_{u \in \mathbb{R}} |\hat{F}(u) - F(u)| \leq \frac{1}{\sqrt{m}}$$

with probability at least $\frac{3}{10}$. Using chernoff bounds, $\frac{1}{\epsilon^2} log(\frac{10}{\epsilon})$. $\square$

We can also prove using martingales or chaining.

5 Auction with Samples

Suppose we have a single bidder, single item auction. Also, assume that the distribution of the bidder value is over $[0, 1]$. We will show that with $O(\frac{1}{\epsilon^2})$ samples from F I can achieve revenue at least $OPT - \epsilon$.

By DKW using $O(\frac{1}{\epsilon^2})$ samples form F, I can find $\hat{F}$ such that $\sup_{u \in \mathbb{R}} |F(u) - \hat{F}(u)| \leq \epsilon$. We will use this $\hat{F}$ to compute price $\hat{p}$ that maximizes $\hat{p} \in argmax_{u \in [0,1]} u(1 - \hat{F}(u))$. Recall that if we know F, optimal p^* is exactly $argmax_{u \in [0,1]} u(1 - F(u))$ and this p^* achieves optimal revenue OPT.

$$
\begin{aligned}
OPT &= p^*(1 - F(p^*)) \\
Rev(\hat{p}) &= \hat{p}(1 - F(\hat{p}) \\
&\geq \hat{p}(1 - \hat{F}(\hat{p}) - \epsilon) \\
&= \hat{p}(1 - \hat{F}(\hat{p})) - \epsilon\hat{p} \\
&\geq p^*(1 - \hat{F}(p^*)) - \epsilon\hat{p} \\
&\geq p^*(1 - F(p^*) - \epsilon) - \epsilon\hat{p} \\
&= p^*(1 - F(p^*)) - \epsilon(\hat{p} + p^*) \\
&= OPT - 2\epsilon
\end{aligned}
$$

With the last step using the fact that prices are bounded within the interval $[0, 1]$. But what if the prices are not bounded? Consider the following distribution over prices.

The prices p is set to M^2 with probability $\frac{1}{M}$ and 0 otherwise. In order to avoid sample complexity dependent on M, I need an assumption on the distribution.

We now generalize to the single item, k bidder model. Assume that F is a product distribution so $F = F_1 \times F_2 \cdots \times F_k$ with $v^{(1)}...v^{(k)} \sim F$.

Theorem 5. *In the single item auction with k bidders and independent regular distributions, if $m = \widetilde{\Omega}\left(\frac{k^{10}}{\epsilon^7}\right)$ then there is a m-sample auction computable in polynomial-time in k and $1/\epsilon$ with expected revenue at least $OPT - \epsilon$.*

Theorem 6. *Any such auction requires some polynomial in k/ϵ samples.*

A *regular distribution* is one where the virtual value $x - \frac{1-F(x)}{f(x)}$ is non-decreasing. Let f_i be the pdf of the distribution over prices for the ith bidder and F_i the corresponding CDF. Then the virtual value v_i of bidder i is $\frac{1-F_i(v)}{f_i(v)}$.

Fact: If F_is are regular then Myerson computes the VCG rule on virtual values.

- If all virtual values are less than zero, no one gets it.

- Otherwise the item goes to the bidder with the highest virtual value.