Dynamic Pricing in Different Valuation Models

Jianyu Xu

Outline

- Problem setup
- Pricing with stochastic valuations
- Pricing with adversarial valuations
- Pricing with a fixed valuation

Problem Setting

For $t = 1,2, \dots, n$:

- Customers create a *valuation* $y_t \in [\mathbf{0}, \mathbf{1}]$ secretly.
- We propose a *price* $p_t \in [\mathbf{0}, \mathbf{1}]$.
- Customers make a *decision* $1_t = 1[p_t \le y_t]$.
- We get a *reward* $r_t(p_t) = p_t \cdot 1[p_t \le y_t]$.

How to define a regret?

Problem Setting -- Regret

Regret is defined as the difference between:

1. Max cumulative reward of a **fixed** price, i.e.

$$\max_p \sum_{t=1}^{n} r_t(p)$$

2. Cumulative reward of our algorithm, i.e.

$$\sum_{t=1}^{n} r_t(p_t)$$

Problem Setting -- Valuations

The series $\{y_t\}_{t=1}^{n}$ can be drawn from 3 models:

- **Identical**: $y_t \equiv p,\ \ t = 1,2,\dots,n.$
- **Stochastic**: $\{y_t\}_{t=1}^{n}$ are i.i.d. samples from a fixed distribution
- **Adversarial** (worst-case).

Problem Setting -- Recap

For $t = 1,2, \ldots, n$:

- Customers create a *valuation* $y_t \in [0,1]$ secretly.
- We propose a *price* p_t.
- Customers make a *decision* $1_t = 1[p_t \leq y_t]$.
- We get a *reward* $r_t(p_t) = p_t \cdot 1[p_t \leq y_t]$.

Regret:

$$\max_p \sum_{t=1}^{n} r_t(p) - \sum_{t=1}^{n} r_t(p_t)$$

Valuation models:

- Identical (fixed)
- Stochastic (i.i.d.)
- Adversarial (worst-case)

Different from bandits

- The feedback is more informative: prices are sequential.
  - If p_t is accepted, then $\forall p \le p_t$ should be accepted.
  - If p_t is rejected, then $\forall p \ge p_t$ should be rejected.
  - While in MAB, the reward of one arm does not indicate the others.
- The actions *can be* continuous.
  - For MAB, there are only K actions.
  - Even though prices are discrete in practice, we usually treat it as continuous in theoretical analysis.

Which is harder?

Pricing v.s. Bandits: which is harder?

I reduced them to each other.

- **Theorem 0.1**: A dynamic pricing problem with K prices can be reduced to a multi-armed bandit problem with K actions.
  - The proof is trivial.

- **Theorem 0.2**: A multi-armed bandit problem with K actions can be reduced to a dynamic pricing problem with $poly(K)$ prices.

Pricing v.s. Bandits: which is harder?

Theorem 0.2: A multi-armed bandit problem with K actions can be reduced to a dynamic pricing problem with $poly(K)$ actions.

Proof sketch: let $K = 2$ as an example:

- In the multi-armed bandit, assume $r_1 = a, r_2 = b, 1/2 < a < b < 1$ without losing of generality.
- We reduce it to 2 dynamic pricing problems:

1. Valuation $\Pr[y = 2b] = \Pr[y = 0] = 1/2$, prices $p_1 = 2a, p_2 = 2b$;
2. Valuation $\Pr[y = 2b] = \frac{a}{2b}, \Pr[y = 2a] = \frac{b^2-a^2}{2ab}, \Pr[y = 0] = \frac{2a-b}{2a}$, prices $p_1 = 2a, p_2 = 2b$

Pricing v.s. Bandits: which is harder?

- Valuation $\Pr[y=2b]=\Pr[y=0]=1/2$, prices $p_1=2a, p_2=2b$;
  - $\mathbb{E}[r(p_1)]=a, \mathbb{E}[r(p_2)]=b.$
- Valuation $\Pr[y=2b]=\frac{a}{2b}, \Pr[y=2a]=\frac{b^2-a^2}{2ab}, \Pr[y=0]=\frac{2a-b}{2a}$, prices $p_1=2a, p_2=2b.$
  - $\mathbb{E}[r(p_1)]=b, \mathbb{E}[r(p_2)]=a.$
- Thus, we have got rid of the sequence of prices.
- Therefore, the bandit problem is reduced to a pricing problem with discrete prices.
- $\Rightarrow$ Pricing with continuous prices $\succcurlyeq$ multi-armed bandits

Shall we treat it as bandits?

- Pros: we are familiar with bandits.
  - Especially for non-parametric models.
- Cons: we will suffer:
  - **Interior** regret caused by discrete prices: a $\sqrt{K}$ factor
  - **Exterior** regret: intervals between discrete prices.

Outline

- Problem setup
- Pricing with stochastic valuations
- Pricing with adversarial valuations
- Pricing with a fixed valuation

Stochastic Valuation: main idea

- Main idea: discretization + stochastic bandits
- How to discretize prices?
  - Uniformly divide into K prices: $\{\frac{1}{K}, \frac{2}{K}, \dots, \frac{i}{K}, \dots, 1 - \frac{1}{K}, 1\}$
- Which bandit algorithm to use?
  - In this paper, they use UCB-1.
- How to bound the regret?
  - Exploit the distance-dependent regret of UCB-1.
  - Carefully select K to balance interior and exterior regret.

Demand Curve

- For any price $x \in [0,1]$, define a "demand function" as:

$$D(x) := \Pr_{\mathrm{y}}[x \le y]$$

- Define an *expected revenue* $f(x) := xD(x)$.
- Denote $\mu_i := f\left(\frac{i}{K}\right)$, and $\mu^* := \max_i \mu_i$, $\Delta_i = \mu^* - \mu_i$.

Assumptions

We make 2 assumptions:

- Assumption 1: the *expected revenue* $f(x) \coloneqq xD(x)$ has a unique global maximum at $x^* \in (0,1)$.

- Assumption 2: $f''(x^*) < 0$. (Local concavity)

Stochastic Valuation: theorem

Based on these two assumptions, we have:

Theorem 3.14. *Assuming that the function $f(x) = xD(x)$ has a unique global maximum $x^* \in (0,1)$, and that $f''(x^*)$ is defined and strictly negative, the strategy* UCB1 *with $K = \lceil (n/\log n)^{1/4} \rceil$ achieves expected regret $O(\sqrt{n \log n})$.*

Here n is T in our notations.

- UCB-1:

> Play machine j that maximizes $\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$, where $\bar{x}_j$ is the average reward obtained from machine j, n_j is the number of times machine j has been played so far, and n is the overall number of plays done so far.

Stochastic Valuation: proof

Theorem 3.14. *Assuming that the function $f(x) = xD(x)$ has a unique global maximum $x^* \in (0,1)$, and that $f''(x^*)$ is defined and strictly negative, the strategy* UCB1 *with $K = \lceil (n/\log n)^{1/4} \rceil$ achieves expected regret $O(\sqrt{n \log n})$.*

We decompose the reward as 4 stages:

1. Reward of UCB-1;
2. Reward of x^*, where $x^* = argmax_x\, f(x)$;
3. Reward of $\frac{j^*}{K}$, where $j^* = \mathrm{argmin_j}|\mathrm{x}^* - \frac{\mathrm{j}^*}{\mathrm{K}}|$;
4. Reward of p^*, where $p^* = argmax_p \sum_t p \cdot 1[p \le y_t]$.

Note: p^* is random.

- $\mathbb{E}[1] \le \mathbb{E}[3] \le \mathbb{E}[2] \le \mathbb{E}[4]$, and 3 regrets in between.

Bandits Regret

Theorem 3.14. *Assuming that the function $f(x) = xD(x)$ has a unique global maximum $x^* \in (0,1)$, and that $f''(x^*)$ is defined and strictly negative, the strategy* UCB1 *with $K = \lceil (n/\log n)^{1/4} \rceil$ achieves expected regret $O(\sqrt{n \log n})$.*

Regret Part 1: UCB-1 v.s. $\frac{j^*}{K}$ closest to x^*

- $\leq$ Regret of UCB-1

Theorem 1. *For all $K > 1$, if policy* UCB1 *is run on K machines having arbitrary reward distributions $P_1, \ldots, P_K$ with support in $[0, 1]$, then its expected regret after any number n of plays is at most*

$$\left[8 \sum_{i:\mu_i<\mu^*} \left(\frac{\ln n}{\Delta_i}\right)\right] + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{j=1}^{K} \Delta_j\right)$$

- $\Delta_i = \mu^* - \mu_i$, where $\mu_i = f\left(\frac{i}{K}\right), \mu^* = \max_i \mu_i$

Note: we may assume $\mu_{j^*} = \mu^*$ without losing of generality.

Bandits Regret

$$\left[8\sum_{i:\mu_i<\mu^*}\left(\frac{\ln n}{\Delta_i}\right)\right]+\left(1+\frac{\pi^2}{3}\right)\left(\sum_{j=1}^{K}\Delta_j\right) \quad \rightarrow \quad O(\sqrt{n\log n})$$

- How to bound Δ_i ? ------ 2 assumptions: unique x^*, negative $f''(x^*)$

Lemma 3.11. *There exist constants C_1, C_2 such that $C_1(x^* - x)^2 < f(x^*) - f(x) < C_2(x^* - x)^2$ for all $x \in [0,1]$.*

Corollary 3.12. $\Delta_i \geq C_1(x^* - i/K)^2$ *for all i. If $\tilde{\Delta}_0 \leq \tilde{\Delta}_1 \leq \ldots \leq \tilde{\Delta}_{K-1}$ are the elements of the set $\{\Delta_1, \ldots, \Delta_k\}$ sorted in ascending order, then $\tilde{\Delta}_j \geq C_1(j/2K)^2$.*

Corollary 3.13. $\mu^* > x^* D(x^*) - C_2/K^2$.

See Notes

Discretization Error

Theorem 3.14. *Assuming that the function $f(x) = xD(x)$ has a unique global maximum $x^* \in (0,1)$, and that $f''(x^*)$ is defined and strictly negative, the strategy* UCB1 *with $K = \lceil (n/\log n)^{1/4} \rceil$ achieves expected regret $O(\sqrt{n \log n})$.*

Regret Part 2: $\frac{j^*}{\mathrm{K}}$ closest to x^* v.s. x^*

Corollary 3.13. $\mu^* > x^* D(x^*) - C_2/K^2$.

- Cumulative error $\le \frac{C_2}{K^2} \cdot n = O(\sqrt{n \log n})$

Ex ante regret and ex *post* regret

Regret Part 3: x^* v.s. $p^* = argmax_p \sum_t p \cdot 1[p \le y_t]$

- i.e., max expected revenue v.s. expected max revenue

- $f(x^*)$ is called ***ex ante*** revenue
  - which is optimal *before* knowing y_t.

- $\frac{1}{n}\sum_t p^* \cdot 1[p^* \le y_t]$ is called ***ex post*** revenue
  - Which is optimal *after* knowing all y_t.

- $\mathbb{E}\left[\max_p \frac{1}{n}\sum_t p \cdot 1[p \le y_t]\right] \ge f(x^*)$

- *Ex ante* regret → training; ex *post* regret → testing

Ex ante regret and ex *post* regret

Theorem 3.14. *Assuming that the function $f(x) = xD(x)$ has a unique global maximum $x^* \in (0,1)$, and that $f''(x^*)$ is defined and strictly negative, the strategy* UCB1 *with $K = \lceil (n/\log n)^{1/4} \rceil$ achieves expected regret $O(\sqrt{n \log n})$.*

Regret Part 3: x^* v.s. $p^* = argmax_p \sum_t p \cdot 1[p \leq y_t]$

- Define: $\rho(x) = \sum_{t=1}^{n} x \cdot 1[x \leq y_t]$
  - $\mathbb{E}[\rho(x^*)] = f(x^*)$
- $\Rightarrow \rho(x) \geq \rho(p^*) - n(p^* - x), \forall x < p^*$.
- $\Rightarrow \int_0^1 \Pr[\rho(x) - \rho(x^*) > \lambda]dx \geq \frac{\lambda}{n} \Pr[\rho(p^*) - \rho(x^*) > 2\lambda]$
- Chernoff Bound: $\Pr[\rho(x) - \rho(x^*) > \lambda] < \exp\{-\lambda^2/2n\}$
  - for martingale
- $\Rightarrow \Pr[\rho(p^*) - \rho(x^*) > 2\lambda] < \min\{1, \frac{n}{\lambda}\exp\{-\lambda^2/2n\}\}$

See Notes

Ex ante regret and ex *post* regret

Theorem 3.14. *Assuming that the function $f(x) = xD(x)$ has a unique global maximum $x^* \in (0,1)$, and that $f''(x^*)$ is defined and strictly negative, the strategy* UCB1 *with $K = \lceil (n/\log n)^{1/4} \rceil$ achieves expected regret $O(\sqrt{n \log n})$.*

$$\Pr[\rho(p^*) - \rho(x^*) > 2\lambda] < \min\{1, \frac{n}{\lambda}\exp\{-\lambda^2/2n\}\}$$

$$\Rightarrow \mathbb{E}[\rho(p^*) - \rho(x^*)] \le \int_0^{+\infty} \Pr[\rho(p^*) - \rho(x^*) > y]\, dy$$

$$< \int_0^{+\infty} \min\left\{1, \frac{2n}{y}\exp\left\{-\frac{y^2}{2n}\right\}\right\} dy$$

$$< \int_0^{\sqrt{4n\log n}} dy + \int_{\sqrt{4n\log n}}^{+\infty} \frac{2n}{\sqrt{4n\log n}} \exp\left\{-\frac{y^2}{2n}\right\} dy$$

$$= O(\sqrt{n\log n})$$

Recap: stochastic valuations

- 2 Methods:
  - Discretization: K uniformly
  - Bandit algorithm: UCB-1
- 3 steps of regret bounds:
  - Regret of UCB-1
  - Error of discretization
  - *Ex post* revenue – *ex ante* revenue
- Skills of proving:
  - Smoothness & Strong concavity → quadratic bounds
  - Distance-dependent regret of UCB-1
  - 2nd definition of expectation

Outline

- Problem setup
- Pricing with stochastic valuations
- Pricing with adversarial valuations
- Pricing with a fixed valuation

Adversarial Valuation: main idea

- Main idea: discretization + **adversarial** bandits
- How to discretize prices?
  - Uniformly divide into K prices: $\{\frac{1}{K}, \frac{2}{K}, \dots, \frac{i}{K}, \dots, 1-\frac{1}{K}, 1\}$
- Which bandit algorithm to use?
  - In this paper, they use **EXP-3**.
- How to bound the regret?
  - Carefully select K to balance interior and exterior regret.

Adversarial Bandits

- The reward $r_i(t)$ of choosing action i at time t is arbitrarily determined in advance, but in secret.

- Regret: compare with the optimal fixed action.
  - Here ex *ante* regret = ex *post* regret.

- Therefore: requires active explorations.
  - Randomness of algorithm.
  - In comparison, UCB-1 has passive explorations.

EXP-3

Algorithm Exp3
Parameters: Real $\gamma \in (0, 1]$.
Initialization: $w_i(1) = 1$ for $i = 1, \ldots, K$.

For each $t = 1, 2, \ldots$

1. Set

$$p_i(t) = (1 - \gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K} \qquad i = 1, \ldots, K.$$

2. Draw i_t randomly accordingly to the probabilities $p_1(t), \ldots, p_K(t)$.
3. Receive reward $x_{i_t}(t) \in [0, 1]$.
4. For $j = 1, \ldots, K$ set

$$\hat{x}_j(t) = \begin{cases} x_j(t)/p_j(t) & \text{if } j = i_t, \\ 0 & \text{otherwise,} \end{cases}$$

$$w_j(t+1) = w_j(t) \exp\left(\gamma \hat{x}_j(t)/K\right) .$$

- First efficient algorithm for adversarial bandits.

Regret Bound

THEOREM 3.1. *For any $K > 0$ and for any $\gamma \in (0, 1]$,*

$$G_{\max} - \mathbf{E}[G_{\mathbf{Exp3}}] \le (e-1)\gamma G_{\max} + \frac{K \ln K}{\gamma}$$

- Let $\gamma = \min\left\{1, \sqrt{\frac{K \ln K}{(e-1)n}}\right\}$, and RHS $\le 2\sqrt{e-1}\sqrt{nK \ln K}$

Also, the discretization error $\le n \cdot \frac{1}{K} = \frac{n}{K}$.

- To balance $\sqrt{nK \ln K}$ and $\frac{n}{K}$, let $K = \left[\frac{n}{\ln n}\right]^{1/3}$, then the regret bound is $O(n^{2/3}(\ln n)^{1/3})$.

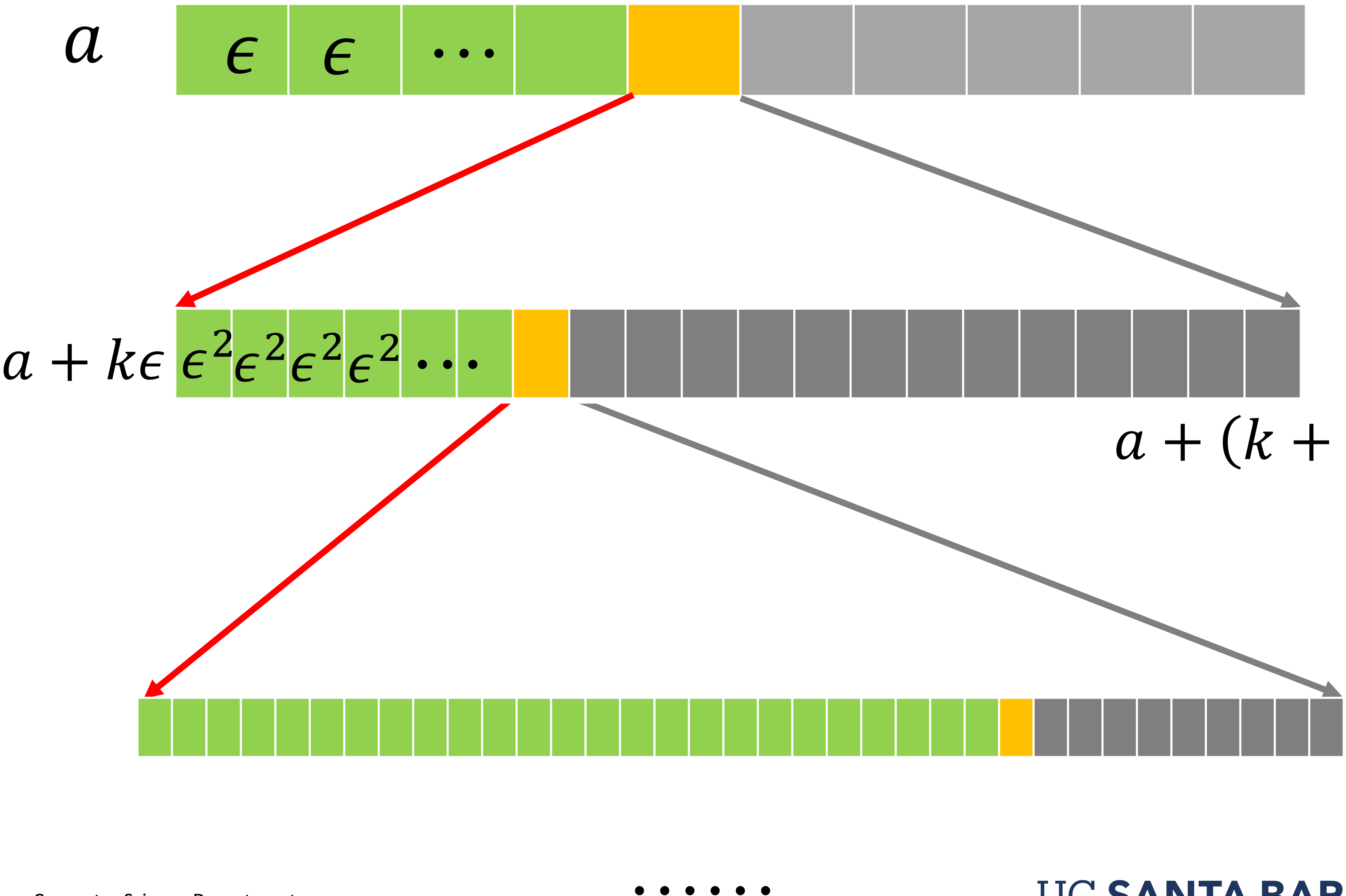
Outline

- Problem setup
- Pricing with stochastic valuations
- Pricing with adversarial valuations
- Pricing with a fixed valuation

Fixed Valuation

- Method: search a feasible interval $[a, b]$ with ϵ-length steps: $a, a+\epsilon, a+2\epsilon, \dots, b-\epsilon, b$.
  - Initialization: a=0, b=1, ϵ=1/2,
- If $a+k\epsilon$ is accepted, but $a+(k+1)\epsilon$ is not, then:
  - $a \leftarrow a+k\epsilon$
  - $b \leftarrow a+(k+1)\epsilon$
  - $\epsilon \leftarrow \epsilon^2$
- Terminal: when $b-a<1/n$, always choose a afterwards.
  - First explore then exploit.

Squaring search



Fixed Valuation: regret bound

Theory: this algorithm achieves regret $O(\log\log n)$.

- Proof sketch: we call each update of $[a, b]$ a *phase*.

1. As ϵ from $\frac{1}{2}$ to $\frac{1}{n}$, there are $O(\log\log n)$ phases.
2. Only one rejection in each phase.

   -- regret of rejection = $O(\log\log n)$.
3. Within each phase, $b - a = \sqrt{\epsilon}$, at most $\frac{\sqrt{\epsilon}}{\epsilon} = \frac{1}{\sqrt{\epsilon}}$ buys.
4. Within each phase, regret is at most $\sqrt{\epsilon} \times \frac{1}{\sqrt{\epsilon}} = 1$.

   -- regret of acceptance = $O(\log\log n)$

Why not binary search?

- Binary search is most informative.
  - But what is "informative"?
  - Do we need "informative"?

Claim: a binary search will suffer from $\Theta(\log n)$ regret.

- For $O(\log n)$, the claim is trivial.
- For $\Omega(\log n)$, consider the case where valuation $= \frac{1}{2}$.
  - Round 1: x=1/2 → accepted.
  - Afterwards: always rejected until stopping explorations.
  - Times of explorations: 1/2 → 1/n, $O(\log n)$
  - Regret of each explorations: 1/2.

Take-home ideas

- Different settings of dynamic pricing problems.
  - Fixed/stochastic/adversarial valuations.
  - Regret: $O(\log\log n), \tilde{O}\left(\sqrt{n\log n}\right), \tilde{O}(n^{2/3})$.
- Approach: discretization + multi-armed bandits.
  - Stochastic bandits: UCB-1, with distance-dependent regret.
  - Adversarial bandits: EXP-3.

UC SANTA BARBARA