Paradigmatic problem

- **Seller** with limited supply: \( k \) identical items to sell
- In each round \( t = 1 \ldots T \), a new **customer** arrives
  - seller offers 1 item @ price \( p_t \in [0,1] \)
  - customer accepts or rejects
- Until **no more items or no more customers**

**Goal:** adjust price over time, to maximize reward.  
No bonus for leftover items!

- \( S(p) = \text{Pr}[\text{sale @ price } p] \)  
  - **demand curve**
  - fixed but unknown to seller
  - Compete with **best fixed price**
What’s going on: Economics

- interpretation: sale $\iff p_t \leq v_t$, where $v_t$ is customer’s value.
- Where do values come from?
  - worst-case (CS) view: values chosen adversarially
    … often leads to weak positive results.
  - Bayesian (Econ) view: from a known distribution
    … strong assumption, sometimes unrealistic.
- Prior-independent mechanisms are a compromise
  - Private values are sampled IID from unknown distribution
  - Goal: be competitive against the optimal mechanism that knows the distribution (perhaps restricting it to a large, natural class).
What’s going on: Economics

Why Posted Price Mechanisms? (PPM)

- PPMs are widely used in practice.
- Customers do not need to know their exact value, only need to evaluate a single price offer.
- Each customer reveals very little information to the seller (revealed info may hurt her in the future)

Moreover:

- PPMs are truthful and group-strategy proof.
- The optimal online mechanism for known distribution is a PPM
What is going on: Bandits

- First intuition: we want to sell at (unknown) “best price”
  - offered price too low ⇒ likely sale, wasted item
  - offered price too high ⇒ likely no sale, wasted customer
- … but we learn something about the demand distribution
- “explore-exploit tradeoff”, “learn-and-earn”

- With limited supply, our learning ability is handicapped:
  - can’t afford to sell too many items while trying “low” prices
- Without parametric assumptions, no long-range inference
Outline

✓ Intro

☑ Unlimited supply

Limited supply:

☑ some observations
☑ explore-then-exploit
☑ lower bound
☑ a better algorithm (overview)
Unlimited supply \((k = T)\)

- Multi-armed bandit problem
  - *arms* (prices) with fixed but unknown expected rewards
  - *bandit feedback*: only for chosen price
- Special feature: sale @price \(p\) \(\Rightarrow\) sale @any lower price
- Reward function \(\text{Rew}(p) = p \cdot S(p)\) \((\text{expected per-round reward})\)
- Best fixed price \(p^*\): maximizes \(\text{Rew}(p)\)
- Algorithm’s performance measured by \(\text{Regret}(T) = T \cdot \text{Rew}(p^*) - E[\text{algorithm's total reward}]\)
Reduction to bandits

- Uniform discretization $U$, then run a bandit algorithm on $U$

- Regret = Regret$_U + OPT - OPT_U$

- Round down $p^*$ to the nearest price $U$, call it $q^*$
  Selling @$q^*$ loses at most $\epsilon$ per each sale.
  So, discretization error $\leq \epsilon n$

- Pick $\epsilon$ in advance to optimize regret
Reduction to bandits

- simple algorithm: *explore-then-exploit*
  - pick arm u.a.r. from $U$ for $T_0$ rounds, then pick est. “best arm” and stick with it
  - pick $\epsilon, T_0$ in advance to optimize regret
- better approach: *adaptive exploration*
  - *adapt* to observations to zoom in on better arms, e.g., UCB1 or Thompson Sampling
  - pick $\epsilon$ in advance to optimize regret

Regret $\tilde{O}(T^{3/4})$

Optimal regret $\tilde{O}(T^{2/3})$
Lower bound on regret

- **Theorem**: \( \text{Regret} \geq \Omega(T^{2/3}) \)
- **Family of problem instances**:
  - values \( v_t \) are only multiples of \( \epsilon \)
    (note: suffices to consider prices \( p_t \) of the same form)
  - *needle-in-a-haystack*: choose sale probabilities \( S(p) \) so that
    \[
    \text{Rew}(p) \equiv p \cdot S(p) = \begin{cases} 
      1/4 + \epsilon/2, & p = p_0 \\
      1/4, & p \geq 1/4 \\
      0, & p < 1/4
    \end{cases}
    \]
- **Bandit problem with IID rewards**, \( n = \frac{3}{4 \epsilon} \) arms, one “needle”
  \( \implies \) Any algorithm has regret \( \Omega(n/\epsilon) \) for some \( p_0 \).
- **Plug in** \( \epsilon = T^{-1/3} \)
Outline

☑ Intro

☑ Unlimited supply

Limited supply:

☑ some observations
☑ explore-then-exploit
☑ lower bound
☑ a better algorithm (overview)
Limited supply \((k < T)\)

- [even with explore-then-exploit] exploitation is limited by \#items left after exploration
- maximizing expected *per-round* reward is not the right goal. need to think about expected *total* reward
- Best fixed price \(p^* = \arg\max_p \text{Rew}(p)\)
  \(\text{Rew}(p)\) expected total reward for fixed price \(p\)
- Regret\((k, T) = \text{Rew}(p^*) - \text{Rew}(\text{algorithm})\)
  now want regret sublinear in \(k = \#\text{items}\)
Explore-then-exploit fails for $k < \sqrt{T}$

- Uniform discretization, $T_0$ rounds of exploration (u.a.r.)
  then pick “est. best arm” & stick with it.

- If $T_0 \leq \frac{T}{k}$, Two problem instances
  
  \[
  \nu_t = \begin{cases} 
  1 & \text{w. prob } \frac{k}{T} \\
  0 & \text{othw}
  \end{cases}
  \]

  Best exp. reward = $k$

  with const prob., $\{\nu_t = 0 \text{ for all } t \leq T_0\}$

  $\implies$ with const prob., cannot tell apart the two instances $\implies$ regret $\frac{k}{2}$

- If $T_0 \geq \frac{T}{k}$ then for problem instance with value $\nu_t \equiv 1$,
  all items are sold in exploration, at average price $\frac{1}{2}$ $\implies$ regret $\frac{k}{2}$

\[
\nu_t = \begin{cases} 
  \frac{1}{2} & \text{w. prob } \frac{k}{T} \\
  0 & \text{othw}
  \end{cases}
\]

Best exp. reward = $k/2$
**Tool: fractional reward**

- Total expected reward $Rew(p) = p \cdot E[\#\text{sales @}p]$
  
difficult to work with directly, use approximation
- Easy upper bound: $E[\#\text{sales @}p] \leq \min(k, T \cdot S(p))$
- Use $f(p) = p \cdot \min(k, T \cdot S(p))$ “fractional reward”
- **Claim**: $f(p) - O(p\sqrt{k \log k}) \leq Rew(p) \leq f(p)$
- **Proof** ($\leq$): sell at price $p$, let $X_t = 1(\text{sale @round } t)$, $X = \sum_t X_t$.
  - $\mu := E[X] = T \cdot S(p)$
  - by Chernoff Bounds, $|X - \mu| \leq O(\sqrt{\mu \log k})$ WHP
  - $\#\text{sales} = p \cdot \min(k, X) \geq \min(k, \mu - O(\sqrt{\mu \log k}))$
    $\geq \min(k, \mu - O(\sqrt{k \log k}))$
  - $Rew(p) = p \cdot \#\text{sales} \geq p \cdot \min(k, \mu) - O(p)\sqrt{\mu \log k}$
Outline

✓ Intro
✓ Unlimited supply

Limited supply:

✓ some observations
☐ explore-then-exploit
☐ lower bound
☐ a better algorithm (overview)
Explore-then-exploit for $k \sim T$

- Uniform discretization $U$, $T_0 \leq k$ rounds of exploration (u.a.r.) then pick the “exploit arm” $p_0$ & stick with it.
- Here’s how we pick $p_0$ after exploration:
  - Est. sales prob $\hat{S}(p) = \frac{\#\text{sales @}p}{\#\text{rounds @}p}$
  - Est. fractional reward $\hat{f}(p) = p \cdot \min\left(k, T \cdot \hat{S}(p)\right)$.
  - Pick exploit arm: $p_0 = \arg\max_{p \in U} \hat{f}(p)$
- **Theorem**: Regret $\leq O\left(T^{3/4} \log T\right)$
Theorem: Regret $\leq O(T^{3/4}\log T)$

Assume: $|\hat{S}(p) - S(p)| \leq \delta \quad \forall p \in U$, where $\delta = O(\log T)/\sqrt{\epsilon T_0}$

- $|\hat{f}(p) - f(p)| \leq 2pT\delta$
  $\hat{f}(p) = p \cdot \min(k, T \cdot \hat{S}(p))$

- $p^*$: best arm; pick closest arm $q^* \leq p^*$ in discretization

- By defn of exploit arm: $\hat{f}(p_0) \geq \hat{f}(q^*)$

- $f(p_0) \geq f(q^*) - 4T\delta$
  $\geq Rew(q^*) - 4T\delta$
  $\geq Rew(p^*) - 4T\delta - \epsilon T$

- $\{ Rew(p_0) \}$ in exploitation $\geq Rew(p_0) - T_0$
  $\geq f(p_0) - T_0 - \gamma$, \quad $\gamma = O(\sqrt{k \log k})$
  $\geq Rew(p^*) - (T_0 + \gamma + 4T\delta + \epsilon T)$. (plug in prev. eq.)

- Take $\epsilon = T^{-1/4}$ and $T_0 = T^{3/4} \implies$ done
Explore-then-exploit: analysis

**Claim:** \( \{|\hat{S}(p) - S(p)| \leq \delta\} \text{ WHP} \)

\[
\delta = O(\log T)/\sqrt{\epsilon T_0}
\]

- \( T_0 \leq k \implies \text{cannot run out of supply during exploration} \)

- Fix price \( p \in U \). Let \( N = \#\text{rounds} @p \) during exploration.

- \( E[N] = \frac{T_0}{\#\text{prices}} = \epsilon T_0 \). Chernoff Bounds \( \implies N > \epsilon T_0/2 \) WHP.

- \( X_j = 1(\text{sale at the } j\text{-th time price } p \text{ is chosen}) \)
  \( Y_m = X_1 + \cdots + X_m \) sum of IID random variables in \([0,1]\)
  Chernoff Bounds \( \implies \left| \frac{Y_m}{m} - S(p) \right| \leq \delta(m) \) WHP

- Take union bound over all \( m \).

  Then: \( \left| \frac{Y_N}{N} - S(p) \right| \leq \delta \left( \frac{\epsilon T_0}{2} \right) \)

  done because \( \hat{S}(p) = Y_N/N \).
Lower bound on regret

(old) cannot have regret $o(T^{2/3})$ for unlimited supply.
(new) cannot have regret $o(k^{2/3})$ for arbitrarily large $k, T$.

Proof idea: reduce to the LB for unlimited supply.

- Suppose we have an algorithm $\mathcal{A}$ which breaks (new).
  Construct algorithm which breaks (old).

- Suffices to break (old) for arb. large time horizon $T'$. Take any $k, T$ for which $\mathcal{A}$ achieves regret $o(k^{2/3})$.
  We solve any unlimited supply instance $I_0$ with $T' = k/4$.

- Algorithm: use $\mathcal{A}$ on a simulated problem instance $I_{\text{sim}}$:
  - in each round, with prob. $k/2T$, ask next customer from $I_0$; else, just return “no sale”.

- $I_{\text{sim}}$ can’t run out of supply, so its restriction to $I_0$ “works”
Outline

✓ Intro
✓ Unlimited supply

Limited supply:
  ✓ some observations
  ✓ explore-then-exploit
  ✓ lower bound
  ❏ a better algorithm (overview)
UCB algorithm for total rewards

- Uniform discretization $U$

- Want: in each round, pick price
  \[
  \arg\max_{p \in U} \text{UCB}(\text{REW}(p))
  \]

- Approximate
  \[
  \text{REW}(p) \approx p \cdot \min(k, T \cdot S(p)) := f(p)
  \]

- Algorithm: in each round, pick price $p \in U$ with maximal
  \[
  \text{Index}(p) = p \cdot \min(k, n \cdot \text{UCB}(S(p)))
  \]

  \[
  \text{approx. #sales @p} \quad \text{ave. sales rate + conf. term}
  \]
References


- (limited supply: treatment of explore-then-exploit in this lecture) Alex Slivkins, unpublished.