FALL 2026 BAIS 7900

Lecture 10: Inventory Control and Dynamic Pricing

Lecturer: *Mengxiao Zhang* Week 12

In the previous several lectures, we developed the theory of online convex optimization, stochastic and adversarial multi-armed bandits, and contextual bandits. In this lecture, we apply these tools to two sequential decision problems: inventory control and dynamic pricing. Both problems involve an unknown distribution and observations that depend on the learner's decisions. Their different feedback and cost structures determine which learning algorithms we can use.

1 Inventory Control

A retailer replenishes a product before observing the day's demand. Ordering too much leaves unsold inventory, while ordering too little loses sales. There are two types of products. *Perishable products*, such as fresh bread and prepared meals, cannot be sold after the current period. *Nonperishable products*, such as canned food and household supplies, can be carried into later periods. We first identify the optimal policy when the demand distribution is known, then study how to learn from sales alone.

1.1 Problem Formulation

Fix a horizon $T > 2$, a maximum demand $D > 0$, and costs $h, b > 0$. Denote by $d_t \in [0, D]$ the demand on round t. The demands are drawn independently from an unknown distribution with continuous cumulative distribution function F. The constant h is the cost per unsold unit, interpreted as a disposal cost for perishable products and a holding cost for nonperishable products; b is the penalty per unit of unmet demand. We assume immediate replenishment, no ordering cost, and initial inventory $x_1 = 0$. At each round $t = 1, \dots, T$,

1. the retailer observes the available inventory x_t and chooses a stocking level $y_t \in [x_t, D]$, ordering $y_t - x_t$ units;
2. demand d_t arrives, and the retailer observes only sales $S_t \triangleq \min\{y_t, d_t\}$;
3. the retailer incurs the cost

$$\ell(y_t, d_t) \triangleq h(y_t - d_t)_+ + b(d_t - y_t)_+, \tag{1}$$

   where $(u)_+ \triangleq \max\{u, 0\}$, and the next inventory is

$$x_{t+1} = \begin{cases} 0, & \text{for perishable products,} \\ y_t - S_t = (y_t - d_t)_+, & \text{for nonperishable products.} \end{cases}$$

For nonperishable products, the constraint $y_t \geq x_t$ means that the retailer can add stock but cannot remove unsold units. Unmet demand is lost, so it does not become an order to fill later. If $S_t < y_t$,

we observe the entire demand. If $S_t = y_t$, we know only that demand is at least the stocking level. This is *censored demand feedback*. In particular, the retailer does not observe the shortage quantity or the realized cost $\ell(y_t, d_t)$ when sales are clipped by the available stock.

Denote by $C(y) \triangleq \mathbb{E}_{d\sim F}[\ell(y, d)]$ the expected cost of stocking y units. Integrating the probabilities of leftover stock and unmet demand gives

$$C(y) = h \int_0^y F(u)\, du + b \int_y^D (1 - F(u))\, du.$$

We compare the learner with a policy that knows F and chooses each order before the current demand arrives.

1.2 The Optimal Policy

We now consider what is the optimal inventory policy that minimizes the expected cost $C(y)$. The function $\ell(\cdot, d)$ is convex for every d, so C is convex. To identify its minimizer, since demand is continuous,

$$C'(y) = hF(y) - b(1 - F(y)) = (h + b)F(y) - b.$$

Denote by $\alpha \triangleq b/(h + b)$ the *critical ratio* and choose a stocking level $y^\star \in [0, D]$ satisfying

$$F(y^\star) = \alpha. \tag{2}$$

Such a level exists by continuity of F. It satisfies $C'(y^\star) = 0$, so convexity implies that it minimizes C. Thus $y^\star$ is a Bayes-optimal stocking level: it minimizes the cost averaged over the demand distribution. The corresponding *base-stock policy*, which orders the products to a fixed inventory level at each round, orders up to $y^\star$ at each round. We next verify that this policy minimizes the expected total cost for both perishable and nonperishable products.

Theorem 1.1. *Fix $h, b, D > 0$ and an i.i.d. demand distribution on $[0, D]$ with continuous cumulative distribution function F. Starting from $x_1 = 0$, the base-stock policy with target $y^\star$ is feasible and has expected total cost $TC(y^\star)$. Every policy has expected total cost at least $TC(y^\star)$.*

Proof. Under the base-stock policy, the next inventory is either zero or $(y^\star - d_t)_+ \le y^\star$. Thus replenishing to $y^\star$ is feasible on every round. For any policy, y_t is chosen before the independent demand d_t, so its conditional expected cost is $C(y_t) \ge C(y^\star)$. Summing and taking expectations gives the lower bound $TC(y^\star)$, which the base-stock policy attains. □

When F is unknown, we measure the cost of learning by

$$\mathrm{Reg}_T^{\mathrm{inv}} \triangleq \mathbb{E}\left[\sum_{t=1}^T \ell(y_t, d_t)\right] - TC(y^\star) = \mathbb{E}\left[\sum_{t=1}^T \big(C(y_t) - C(y^\star)\big)\right]. \tag{3}$$

In the following, we discuss how to design an online learning algorithm that achieves sublinear regret in this definition.

1.3 Online Learning with Perishable Inventory and Gradient Feedback

We start with the simpler perishable case. For perishable products, $x_t = 0$ and every $y_t \in [0, D]$ is feasible. Although demand is censored, recall that the gradient information is enough for online convex optimization, and the sales observation indeed supplies a gradient:

$$g_t \triangleq (h+b)\mathbf{1}\{S_t < y_t\} - b = \nabla_y \ell(y_t, d_t). \tag{4}$$

Indeed, unsold stock means that demand is below the stocking level, giving $g_t = h$; selling out gives $g_t = -b$. The equality case $d_t = y_t$ has probability zero because demand is continuous and independent of the past. Therefore, censored demand is enough for gradient feedback even though the realized loss may be unknown.

In Lecture 3, we established the regret guarantee for online gradient descent (OGD) with gradient feedback. Here, the absolute gradient is at most $\max\{h, b\}$. Starting from any $y_1 \in [0, D]$, use

$$y_{t+1} = \Pi_{[0,D]}(y_t - \eta g_t), \tag{5}$$

where $\Pi_{[0,D]}$ denotes projection onto $[0, D]$. The OGD bound from Lecture 3 gives

$$\sum_{t=1}^{T} \big(\ell(y_t, d_t) - \ell(y^\star, d_t)\big) \le \frac{D^2}{2\eta} + \frac{\eta \max\{h, b\}^2 T}{2}. \tag{6}$$

Choosing $\eta = D/(\max\{h, b\}\sqrt{T})$ and taking expectations yields $\mathrm{Reg}_T^{\mathrm{inv}} \le D \max\{h, b\}\sqrt{T}$. Thus the perishable problem is a direct application of online convex optimization.

1.4 Nonperishable Inventory and Gradient Descent

We now extend the gradient method to nonperishable inventory. The target in equation (5) can fall below the stock already on hand. For example, if $y_t = 0.8D$ and $d_t = 0.1D$, the next round begins with $x_{t+1} = 0.7D$, so a target below $0.7D$ cannot be implemented. We therefore need to distinguish the target chosen by gradient descent from the actual stocking level.

Denote by $z_t \in [0, D]$ the target inventory and replenish to

$$y_t = \max\{x_t, z_t\}. \tag{7}$$

If the available inventory exceeds the target, we place no order. Otherwise, we order up to z_t. The gradient update acts on z_t, and we separately control the cost caused by $y_t > z_t$. The good news is that although we can not realize z_t sometimes, the target gradient is still *always* observable from sales. Since $y_t \ge z_t$, the events $S_t < z_t$ and $d_t < z_t$ are equivalent. Thus, we have

$$\widehat{g}_t \triangleq (h+b)\mathbf{1}\{S_t < z_t\} - b = \nabla_y \ell(z_t, d_t). \tag{8}$$

We can therefore apply the same OGD update to the targets, even when the actual stock exceeds them.

Algorithm 1: Gradient Descent for Nonperishable Inventory

Input: horizon $T > 2$, maximum demand $D > 0$, holding cost $h > 0$, penalty cost $b > 0$
Set $\eta = D/(\max\{h, b\}\sqrt{T})$ and initialize $z_1 = x_1 = 0$;
for $t = 1, \ldots, T$ **do**
  replenish to $y_t = \max\{x_t, z_t\}$ and observe sales S_t;
  compute $\widehat{g}_t = (h+b)\mathbf{1}\{S_t < z_t\} - b$;
  set $z_{t+1} = \Pi_{[0,D]}(z_t - \eta \widehat{g}_t)$ and $x_{t+1} = y_t - S_t$;

Denote by $\mu \triangleq \mathbb{E}[d_1] > 0$ the mean demand. The algorithm uses the same step size as in the perishable case and does not need to know μ.

Theorem 1.2. *Fix $T > 2$, $h, b, D > 0$, and an i.i.d. demand distribution on $[0, D]$ with continuous cumulative distribution function and mean $\mu > 0$. Algorithm 1 guarantees that*

$$\mathrm{Reg}_T^{\mathrm{inv}} \le \frac{D^2 \max\{h, b\}}{\mu} \sqrt{T}. \tag{9}$$

Proof. Projection ensures that $z_t \in [0, D]$. Since $x_1 = 0$ and $0 \le x_{t+1} \le y_t$, induction using equation (7) gives $0 \le x_t \le y_t \le D$. Thus the algorithm is feasible, and both the target and actual stock are chosen before the independent demand d_t arrives.

We first bound the cost of the targets. Similar to the perishable case, we have

$$\ell(z_t, d_t) - \ell(y^\star, d_t) \le \widehat{g}_t(z_t - y^\star) \le \frac{(z_t - y^\star)^2 - (z_{t+1} - y^\star)^2}{2\eta} + \frac{\eta \max\{h, b\}^2}{2},$$

where $|\widehat{g}_t| \le \max\{h, b\}$. Summing over t telescopes the squared-distance terms. Taking expectations and using independence of each demand from its target yields

$$\sum_{t=1}^{T} \mathbb{E}[C(z_t) - C(y^\star)] \le \frac{D^2}{2\eta} + \frac{\eta \max\{h, b\}^2 T}{2}. \tag{10}$$

This is the same OGD argument as in the perishable setting.

It remains to account for stock above the target. When $x_t > z_t$, the retailer places no order and the actual stocking level is $y_t = x_t$. Denote by $e_t \triangleq y_t - z_t \ge 0$ the amount by which the actual stock exceeds the target, with $e_1 = 0$. We first relate this difference to cost. For every demand value d,

$$\ell(y_t, d) - \ell(z_t, d) = h\big((y_t - d)_+ - (z_t - d)_+\big) + b\big((d - y_t)_+ - (d - z_t)_+\big) \le h(y_t - z_t) = he_t.$$

Indeed, $y_t \ge z_t$ implies that the shortage term is nonpositive, while the difference in unsold quantities is between zero and $y_t - z_t$. Taking expectation over $d \sim F$ gives $C(y_t) - C(z_t) \le he_t$. It therefore suffices to bound the cumulative quantity $\sum_{t=1}^{T} \mathbb{E}[e_t]$.

To see how much excess stock carries over to the next round, recall that $y_{t+1} = \max\{x_{t+1}, z_{t+1}\}$ and $x_{t+1} = (y_t - d_t)_+$. Since $z_{t+1} \ge 0$, these two identities give, for $t < T$,

$$e_{t+1} = (x_{t+1} - z_{t+1})_+ = (y_t - z_{t+1} - d_t)_+.$$

The target decreases only when $\widehat{g}_t = h$, in which case $z_{t+1} = \max\{z_t - h\eta, 0\}$. Thus $y_t - z_{t+1} = e_t + (z_t - z_{t+1}) \le e_t + h\eta$. We also have $y_t - z_{t+1} \le D$ because $y_t \le D$ and $z_{t+1} \ge 0$. Denote by $a_t \triangleq \min\{D, e_t + h\eta\}$ the resulting upper bound on the stock above the next target. It depends only on the excess e_t known before demand arrives.

Substituting $y_t - z_{t+1} \le a_t$ into the identity for e_{t+1} gives

$$e_{t+1} \le (a_t - d_t)_+ = a_t - \min\{a_t, d_t\} \le a_t \left(1 - \frac{d_t}{D}\right), \tag{11}$$

where the last inequality uses $\min\{a_t, d_t\} \ge a_t d_t / D$ for $a_t, d_t \in [0, D]$. Since a_t is independent of the current demand d_t and $\mathbb{E}[d_t] = \mu$, taking expectations in equation (11) yields

$$\mathbb{E}[e_{t+1}] \le \mathbb{E}\left[a_t \left(1 - \frac{d_t}{D}\right)\right] = \left(1 - \frac{\mu}{D}\right) \mathbb{E}[a_t] \le \left(1 - \frac{\mu}{D}\right) (\mathbb{E}[e_t] + h\eta).$$

Starting from $e_1 = 0$ and repeatedly substituting this inequality, we obtain

$$\mathbb{E}[e_t] \le h\eta \sum_{j=1}^{t-1} \left(1 - \frac{\mu}{D}\right)^j \le h\eta \frac{D-\mu}{\mu}.$$

Summing over $t = 2, \dots, T$ gives

$$\sum_{t=1}^{T} \mathbb{E}[e_t] \le \frac{D-\mu}{\mu} h\eta(T-1) \le \frac{D-\mu}{\mu} h\eta T. \tag{12}$$

Finally, combining $C(y_t) - C(z_t) \le h e_t$ with equation (10) and equation (12) gives

$$\mathrm{Reg}_T^{\mathrm{inv}} = \sum_{t=1}^{T} \mathbb{E}[C(z_t) - C(y^\star)] + \sum_{t=1}^{T} \mathbb{E}[C(y_t) - C(z_t)] \le \frac{D^2}{2\eta} + \frac{\eta \max\{h,b\}^2 T}{2} + \frac{h^2(D-\mu)}{\mu}\eta T.$$

Substituting $\eta = D/(\max\{h,b\}\sqrt{T})$ and using $h^2 \le \max\{h,b\}^2$ gives equation (9). In particular, the regret is $\mathcal{O}(\sqrt{T})$ for fixed costs and a fixed demand distribution. □

1.5 Nonperishable Inventory and Lazy Updates

The gradient method updates its target after every observation, and the preceding analysis controls the inventory above that target. We next give a binary search method that changes its target at most $\lceil \log_2 T \rceil$ times. It uses confidence intervals for the critical quantile and bounds the time needed to reduce inventory after each downward change.

The idea is to keep a target fixed until the observations determine which side of it contains $y^\star$. Suppose that $y^\star$ lies in an interval $[L, U]$, and set $m = (L+U)/2$. Before sampling at this new target, the retailer places no order on rounds with $x_t \ge m$, so $y_t = x_t$. Once $x_t < m$, the retailer replenishes to m. On each subsequent round in the same phase, the stock left from the preceding round is at most m, so replenishing to m remains feasible.

On the nth sampling round of the phase, denote by $Z_n \triangleq \mathbf{1}\{S_t < m\}$ the indicator of unsold stock. Since the retailer stocks m units, this observation has mean $F(m)$. Denote by $\widehat{F}_n \triangleq n^{-1} \sum_{i=1}^{n} Z_i$ the sample mean and by $\rho_n \triangleq \sqrt{2 \log T / n}$ its confidence radius. If $\widehat{F}_n + \rho_n < \alpha$, the test places $F(m)$ below $\alpha = b/(h+b)$, so we move L to m. If $\widehat{F}_n - \rho_n > \alpha$, we move U to m. Otherwise we take another sample at the same target. Each new phase starts its own sample count at $n = 1$. After $M = \lceil \log_2 T \rceil$ halvings, the interval has width at most D/T, and we use its midpoint for the remaining rounds.

To control every test that Algorithm 2 could perform, fix the inventory grid before the first round,

$$\mathcal{M} \triangleq \left\{ \frac{jD}{2^M} : j = 1, \dots, 2^M - 1 \right\}.$$

In phase k, the interval endpoints are consecutive multiples of $D/2^{k-1}$, so its midpoint is a multiple of $D/2^k$ strictly between 0 and D. For $k = 1, \dots, M$, every such midpoint lies in $\mathcal{M}$. As in the stochastic bandit analysis of Lecture 5, we can generate the demands in advance. For each $m \in \mathcal{M}$, prepare a row of T independent demands $d_{m,1}, \dots, d_{m,T}$ from F. Use the next demand in row m each time the algorithm tests that target, and use a separate sequence of demands from F on all other rounds. All these draws are independent. The inventory and target determine which sequence

Algorithm 2: Lazy Confidence Bisection for Nonperishable Inventory

Input: horizon $T > 2$, maximum demand $D > 0$, holding cost $h > 0$, shortage penalty $b > 0$

Set $\alpha = b/(h+b)$, $M = \lceil \log_2 T \rceil$, $L = 0$, $U = D$, $t = 1$, and $x_1 = 0$;
At the end of each round, set $x_{t+1} = y_t - S_t$ and advance t by one. Stop after round T;
for $k = 1, \ldots, M+1$ **do**
  set $m = (L+U)/2$;
  while $x_t \geq m$ **do**
    place no order, set $y_t = x_t$, and observe sales S_t;
  if $k = M+1$ **then**
    replenish to $y_t = m$ and observe S_t in each remaining round;
  else
    for $n = 1, 2, \ldots$ **do**
      replenish to $y_t = m$, observe S_t, and record $Z_n = \mathbf{1}\{S_t < m\}$;
      compute $\widehat{F}_n = n^{-1} \sum_{i=1}^{n} Z_i$ and $\rho_n = \sqrt{2 \log T / n}$;
      if $\widehat{F}_n + \rho_n < \alpha$ **then**
        set $L = m$ and end this phase;
      else if $\widehat{F}_n - \rho_n > \alpha$ **then**
        set $U = m$ and end this phase;

to use before the current demand arrives, so the demand on each round still has distribution F independently of the past. On a test round the retailer stocks m units, so unsold stock means $d_{m,i} < m$. Denote by $\widehat{F}_n(m) \triangleq n^{-1} \sum_{i=1}^{n} \mathbf{1}\{d_{m,i} < m\}$ the frequency of unsold stock in the first n draws of row m. Whenever the algorithm performs its nth test at m, its sample mean $\widehat{F}_n$ equals $\widehat{F}_n(m)$.

Event 1. Denote by $\mathcal{E}$ the event that the estimate at every candidate target is accurate at every sample count,

$$\mathcal{E} \triangleq \bigcap_{m \in \mathcal{M}} \bigcap_{n=1}^{T} \left\{ \left| \widehat{F}_n(m) - F(m) \right| \leq \rho_n \right\}.$$

Then $\Pr[\mathcal{E}] \geq 1 - \frac{4}{T^2}$.

Proof. Fix $m \in \mathcal{M}$ and $n \in \{1, \ldots, T\}$. By continuity of F, each indicator $\mathbf{1}\{d_{m,i} < m\}$ has mean $F(m)$. The first n indicators in row m are independent, so Hoeffding's inequality and $\rho_n = \sqrt{2 \log T / n}$ give

$$\Pr\left[\left| \widehat{F}_n(m) - F(m) \right| > \rho_n \right] \leq 2 \exp(-2n\rho_n^2) = \frac{2}{T^4}.$$

There are $2^M - 1$ candidate targets and T possible sample counts for each. Taking a union bound over these pairs gives

$$\Pr[\mathcal{E}^c] \leq \sum_{m \in \mathcal{M}} \sum_{n=1}^{T} \frac{2}{T^4} = \frac{2(2^M - 1)T}{T^4} \leq \frac{4}{T^2},$$

where $2^M - 1 < 2T$ because $M = \lceil \log_2 T \rceil$. □

The number of target changes is logarithmic, but we also need to bound the time required to remove excess stock. As in Theorem 1.2, the mean demand μ appears in the analysis but is not an input to the algorithm.

Theorem 1.3. *Fix $T > 2$, $h, b, D > 0$, and an i.i.d. demand distribution on $[0, D]$ with continuous cumulative distribution function and mean $\mu > 0$. Algorithm 2 guarantees that*

$$\mathrm{Reg}_T^{\mathrm{inv}} \le \mathcal{O}\left((h+b)\left(D\sqrt{T\log T} + \frac{D^2 \log T}{\mu}\right)\right). \tag{13}$$

Proof. We first bound the sum of cost gaps in equation (3) on Event 1. On $\mathcal{E}$, every confidence interval computed by the algorithm contains $F(m)$. If $\widehat{F}_n(m) + \rho_n < \alpha$, then $F(m) < \alpha = F(y^\star)$, so $m < y^\star$ and we can move L to m. If $\widehat{F}_n(m) - \rho_n > \alpha$, then $F(m) > \alpha$, so $m > y^\star$ and we can move U to m. Hence every update keeps $y^\star$ in $[L, U]$.

Suppose that the algorithm implements m_k in phase k, and denote by n_k the number of sampling rounds at this inventory level. The interval in phase k has width $w_k \triangleq D/2^{k-1}$ and contains $y^\star$, so its midpoint m_k satisfies $|m_k - y^\star| \le w_k/2$. If $n_k \ge 2$, the test after $n_k - 1$ samples did not end the phase. Its confidence interval therefore contains α, and on $\mathcal{E}$ it also contains $F(m_k)$. Hence

$$|F(m_k) - \alpha| \le |F(m_k) - \widehat{F}_{n_k-1}(m_k)| + |\widehat{F}_{n_k-1}(m_k) - \alpha| \le 2\rho_{n_k-1}.$$

All n_k sampling rounds have the same cost gap. By convexity and $C'(m_k) = (h+b)(F(m_k) - \alpha)$,

$$\begin{aligned} n_k\big(C(m_k) - C(y^\star)\big) &\le n_k C'(m_k)(m_k - y^\star) \le (h+b)\frac{w_k n_k}{2}|F(m_k) - \alpha| \\ &\le (h+b) w_k n_k \rho_{n_k-1} \le 2(h+b) w_k \sqrt{n_k \log T}, \end{aligned}$$

where the last inequality uses $n_k - 1 \ge n_k/2$. For $n_k = 1$, the same final bound follows from $|C'(y)| \le \max\{h, b\}$ and $|m_k - y^\star| \le w_k/2$.

Denote by K the number of phases in which the algorithm sampled a midpoint. Since $n_k \le T$ and $\sum_{k=1}^{K} w_k \le 2D$, the total sampling gap satisfies

$$\sum_{k=1}^{K} n_k\big(C(m_k) - C(y^\star)\big) \le 2(h+b)\sqrt{T\log T}\sum_{k=1}^{K} w_k \le 4(h+b)D\sqrt{T\log T}.$$

After M halvings the interval has width at most D/T, so replenishing to its midpoint for the remaining rounds contributes at most $T\max\{h, b\}D/(2T) = \max\{h, b\}D/2$.

The retailer begins with zero stock, so it waits only after a downward change of target, and there are at most M such changes. After such a change, the retailer orders nothing and lets demand consume the stock until it falls below the new target. Denote by $\mathcal{W}_k$ the waiting rounds after the kth change, with $\mathcal{W}_k = \emptyset$ if the target does not fall. The stock at the start of these rounds is at most D. Before the last waiting round the stock has not yet fallen below the target, so the total demand in the earlier waiting rounds is less than D. The demand in the last waiting round is at most D, so $\sum_{t\in\mathcal{W}_k} d_t < 2D$. The demands are i.i.d. with mean μ, and whether round t is a waiting round is decided before d_t arrives, so Wald's identity gives

$$\mu\,\mathbb{E}\big[|\mathcal{W}_k|\big] = \mathbb{E}\Big[\sum_{t\in\mathcal{W}_k} d_t\Big] \le 2D.$$

Summing over the at most M downward changes, the expected total number of waiting rounds is at most $2MD/\mu$. Multiplying by the maximum cost gap $\max\{h, b\}D$ per waiting round gives an expected transition cost gap of at most $2\max\{h, b\}D^2M/\mu$.

On $\mathcal{E}^c$, the total gap from sampling and from replenishing to the final midpoint is at most $\max\{h, b\}DT$. Its contribution to the expectation in equation (3) is at most $\max\{h, b\}DT\Pr[\mathcal{E}^c] \le 4\max\{h, b\}D/T$. Adding this to the bounds on $\mathcal{E}$ for those rounds and the expected transition cost gap proves equation (13). □

2 Dynamic Pricing

Inventory control chooses how much stock to make available at a fixed cost structure. Dynamic pricing instead chooses a price and observes whether a customer purchases. We assume an unlimited supply, so each round's selling decision has no inventory constraint. The central issue is to learn a profitable price from purchase decisions alone.

2.1 Problem Formulation

At each round $t = 1, \dots, T$, a new customer has value $V_t \in [0, 1]$, drawn independently from an unknown distribution. The seller posts $p_t \in [0, 1]$ before observing the value. The customer buys one unit exactly when $V_t \ge p_t$, so the seller observes $Y_t \triangleq \mathbf{1}\{V_t \ge p_t\}$ and receives revenue p_tY_t.

Denote by $q(p) \triangleq \Pr[V \ge p]$ the probability of a sale and by $r(p) \triangleq pq(p)$ the expected revenue. If the distribution is known, choose $p^\star \in \arg\max_{p\in[0,1]} r(p)$ and charge this price on every round. Every history-dependent policy has conditional expected revenue $r(p_t) \le r(p^\star)$, so the optimal expected total revenue is $Tr(p^\star)$. For an unknown distribution, define pricing regret by

$$\mathrm{Reg}_T^{\mathrm{price}} \triangleq Tr(p^\star) - \mathbb{E}\left[\sum_{t=1}^{T} p_tY_t\right] = \mathbb{E}\left[\sum_{t=1}^{T}\big(r(p^\star) - r(p_t)\big)\right]. \tag{14}$$

A purchase at price p_t tells us that the value is at least p_t; a rejection tells us that it is below p_t. We begin with the case in which this comparison is noiseless across customers.

2.2 A Fixed Unknown Value

Suppose that $V_t = v$ on every round for a fixed unknown $v \in [0, 1]$. The optimal price is v, and each observation compares our price with the same number. Binary search therefore gives a simple algorithm: start from $[a, u] = [0, 1]$, post its midpoint, and move the lower endpoint to the midpoint after a purchase or the upper endpoint after a rejection. After $\lceil \log_2 T \rceil$ comparisons, the interval has width at most $1/T$. Each comparison loses at most one unit of revenue, and subsequently charging a loses at most one in total. Thus binary search gives $\mathrm{Reg}_T^{\mathrm{price}} \le \mathcal{O}(\log T)$.

We can improve this bound because the two kinds of error have different costs. A price just below v loses only the price difference, while a price above v loses the entire sale. The *cautious search* method increases the price in small increments, allowing many inexpensive successful comparisons between rejections.

Initialize $[a, u] = [0, 1]$ and $\epsilon = 1/2$. Within a phase, post $a + \epsilon, a + 2\epsilon, \dots$ up to u, stopping at the first rejection. If price p is rejected, the last accepted lower bound is $p - \epsilon$, so replace the interval

by $[p-\epsilon, p]$ and set the next increment to ϵ^2. If the upper endpoint is accepted, then $v = u$ and we charge u afterward. Once the interval width is at most $1/T$, charge its lower endpoint for the remaining rounds.

Theorem 2.1. *Fix $T > 2$ and an unknown value $v \in [0,1]$. Cautious search guarantees that $\mathrm{Reg}_T^{\mathrm{price}} \le \mathcal{O}(\log\log T)$.*

Proof. The first phase has constant regret. Every later phase begins with an interval of width w and uses increment w^2. It has at most $1/w$ accepted offers, each with regret at most w, and at most one rejection, with regret at most one. Thus each phase contributes at most two, even if the horizon ends during that phase. After a rejection, the width becomes w^2, so the successive widths are $1/2, 1/4, 1/16, \ldots$. At most $\mathcal{O}(\log\log T)$ phases occur before the width is at most $1/T$. Charging the final lower endpoint contributes at most one additional regret. □

The improvement comes from reducing costly rejections. The algorithm can make many offers, but the successful offers are sufficiently close to v that their total loss within a phase is constant.

2.3 General Stochastic Values

When customers have independent values, a rejection does not identify an upper bound on the next customer's value. Binary search for a single threshold no longer applies. Instead, we can treat candidate prices as arms and use the bandit algorithms from previous lectures.

Fix an integer $m \ge 1$ and denote by $\mathcal{P}_m \triangleq \{0, 1/m, \ldots, 1\}$ the price grid. The essential approximation property follows from the fact that lowering a price cannot reduce its sale probability. For an optimal price $p^\star$, choose the grid price $\bar{p} = \lfloor mp^\star \rfloor / m$. Its expected revenue satisfies

$$r(\bar{p}) = \bar{p}\Pr[V \ge \bar{p}] \ge \bar{p}\Pr[V \ge p^\star] = r(p^\star) - (p^\star - \bar{p})\Pr[V \ge p^\star] \ge r(p^\star) - \frac{1}{m}. \tag{15}$$

Thus rounding the optimal price downward loses at most one grid spacing in expected revenue. This argument applies to every value distribution, including distributions with atoms. Rounding upward can instead lose all customers whose values equal $p^\star$.

In Lecture 7, we established that Tsallis-INF with K arms guarantees regret at most $\mathcal{O}(\sqrt{KT})$. Apply it to the $K = m+1$ grid prices, supplying the observed loss $1 - p_t Y_t \in [0,1]$. Conditional on the entire value sequence, this is a bandit loss sequence of the type considered in Lecture 7, so its expected guarantee also holds after averaging over the values.

Algorithm 3: Dynamic Pricing on a Uniform Grid

Input: horizon $T > 2$ and grid size parameter $m \ge 1$
Initialize Tsallis-INF from Lecture 7 on $\mathcal{P}_m = \{0, 1/m, \ldots, 1\}$ with $\beta = 1/2$ and $\eta = 1/\sqrt{T}$;
for $t = 1, \ldots, T$ **do**
  use the arm selected by Tsallis-INF as the price p_t;
  post p_t and observe the purchase indicator Y_t;
  supply the loss $1 - p_t Y_t$ to Tsallis-INF for the selected arm;

Theorem 2.2. *Fix $T > 2$ and any i.i.d. value distribution on $[0,1]$. Algorithm 3 guarantees that*

$$\mathrm{Reg}_T^{\mathrm{price}} \le \mathcal{O}\left(\frac{T}{m} + \sqrt{(m+1)T}\right). \tag{16}$$

In particular, choosing $m = \lceil T^{1/3} \rceil$ gives $\mathrm{Reg}_T^{\mathrm{price}} \le \mathcal{O}(T^{2/3})$.

Proof. Compare first with the rounded price $\bar{p}$ in equation (15). The grid approximation and the bandit guarantee give

$$\mathrm{Reg}_T^{\mathrm{price}} = T\big(r(p^\star) - r(\bar{p})\big) + \mathbb{E}\left[\sum_{t=1}^{T}\big(r(\bar{p}) - r(p_t)\big)\right] \le \frac{T}{m} + \mathcal{O}\left(\sqrt{(m+1)T}\right).$$

Substituting the stated value of m proves the final claim. □

2.4 Smooth Revenue

Can we improve the $T^{2/3}$ rate when the value distribution has more structure? We now show how smoothness of the revenue improves the grid approximation.

Assumption 1. The customer values are i.i.d. on $[0,1]$ with a continuously differentiable density f with bounded $f(v)$ and $f'(v)$: $|f(v)|, |f'(v)| \le L$ for all $v \in [0,1]$.

Under Assumption 1, $q(p) = \int_p^1 f(v)\,dv$ and $r(p) = p \cdot q(p)$. Both endpoints give zero revenue, whereas $q(0) = 1$ and continuity give positive revenue at all sufficiently small positive prices. Thus every optimal price lies in $(0,1)$. Fix any such $p^\star$. Differentiating gives $r'(p) = q(p) - pf(p)$, so $r'(p^\star) = 0$.

How much revenue do we lose by posting a price p close to $p^\star$? Denote this loss by $\Delta(p) \triangleq r(p^\star) - r(p)$. Since $r''(p) = -2f(p) - pf'(p)$, Assumption 1 gives $|r''(p)| \le 3L$. Taylor's theorem at $p^\star$ therefore gives

$$0 \le \Delta(p) \le \frac{3L}{2}(p - p^\star)^2. \tag{17}$$

In particular, a grid price within $1/m$ of $p^\star$ loses at most $3L/(2m^2)$ per round.

Algorithm 3 incurs $\mathcal{O}(\sqrt{(m+1)T})$ regret relative to the best of the $m+1$ grid prices. Adding the grid approximation error gives

$$\mathrm{Reg}_T^{\mathrm{price}} \le \frac{3LT}{2m^2} + \mathcal{O}\left(\sqrt{(m+1)T}\right).$$

Choosing $m = \lceil T^{1/5} \rceil$ gives $\mathcal{O}(T^{3/5})$ regret for fixed L.