ORIE 4154: Pricing and Market Design **Fall 2026**

Lecture 3 — From Optimal Pricing to Learning

Lecturer: Sid Banerjee *Date: September 1, 2026*

3.1 Overview

In Lecture 2 we represented heterogeneous buyer values by a distribution F and obtained the demand curve

$$D(p) = N(1 - F(p)).$$

We also saw several equivalent ways to characterize a zero-cost monopoly price: in terms of elasticity, hazard rate, and virtual value.

We first add costs and obtain the standard monopoly markup condition. But this only takes us so far. The real question is what happens when the demand curve itself is unknown.

The development has two parts:

known demand: optimize $\longrightarrow$ unknown demand: learn while earning.

The second part introduces the key difficulty of online learning: the seller learns about a price only by actually charging it. Thus information has an opportunity cost.

The main path is

markup $\longrightarrow$ unknown demand $\longrightarrow$ greedy failure $\longrightarrow$ regret.

Here and throughout, arm i should be read as the pricing action "charge p_i." Sections marked * are optional reading.

3.2 Pricing with marginal cost

Suppose each sale incurs marginal cost $c \geq 0$. Expected profit is

$$\pi(p) = (p - c)D(p). \tag{3.1}$$

For an interior optimum p^*, the first-order condition is

$$D(p^*) + (p^* - c)D'(p^*) = 0. \tag{3.2}$$

Define the absolute price elasticity of demand

$$\epsilon(p) := -\frac{pD'(p)}{D(p)}.$$

Substituting into (3.2) gives the *Lerner markup rule*

$$\boxed{\frac{p^* - c}{p^*} = \frac{1}{\epsilon(p^*)}.} \tag{3.3}$$

Thus the optimal percentage markup is not determined by cost alone: it depends on how sensitive demand is to price.

3.2.1 The same condition in virtual-value form

Recall from Lecture 2 that if

$$D(p) = N(1 - F(p)),$$

then the virtual value associated with F is

$$\phi(v) = v - \frac{1 - F(v)}{f(v)}.$$

Since

$$D'(p) = -Nf(p),$$

condition (3.2) becomes

$$1 - F(p^*) - (p^* - c)f(p^*) = 0,$$

or equivalently

$$\boxed{\phi(p^*) = c.} \tag{3.4}$$

So three views of the same optimum are

$$\boxed{\frac{p^* - c}{p^*} = \frac{1}{\epsilon(p^*)} \quad \Longleftrightarrow \quad (p^* - c)h(p^*) = 1 \quad \Longleftrightarrow \quad \phi(p^*) = c.}$$

This is worth remembering: the monopoly pricing rule is a marginal-revenue-equals-marginal-cost condition written in three different coordinate systems.

3.3 Why cost-plus pricing is not a general principle

A common heuristic is to choose a markup factor $\alpha > 0$ and set

$$p = (1 + \alpha)c.$$

But (3.3) says the correct markup depends on the demand curve. Two products with the same cost can have very different optimal prices if buyers have different willingness-to-pay distributions.

Example 3.1 (Exponential values). *Suppose*

$$V \sim \text{Exp}(\lambda), \qquad D(p) = Ne^{-\lambda p}.$$

Then

$$\phi(p) = p - \frac{1}{\lambda}.$$

Hence (3.4) *gives*

$$p^* = c + \frac{1}{\lambda}. \tag{3.5}$$

Here the optimal dollar *markup is constant. This simple additive-markup rule works only because the exponential distribution has a constant hazard rate; it is not a general cost-plus principle.*

Example 3.2 (Uniform values). *Suppose*

$$V \sim \text{Unif}[0, b], \qquad 0 \le c \le b.$$

Then for $0 \le p \le b$,

$$D(p) = N\left(1 - \frac{p}{b}\right),$$

and

$$\phi(p) = 2p - b.$$

Thus

$$p^* = \frac{b + c}{2}. \tag{3.6}$$

A one-dollar increase in marginal cost increases the optimal price by only fifty cents.

The larger lesson is:

Pricing requires knowing something about demand, not just knowing cost.

That raises the question that drives the rest of the notes.

3.4 But how do we know demand?

Suppose we have a finite menu of candidate prices

$$\mathcal{P} = \{p_1, \dots, p_K\}.$$

If we knew the purchase probability at every price,

$$q_i := \mathbb{P}(V \geq p_i),$$

then expected revenue from price p_i would be

$$\mu_i = p_i q_i \tag{3.7}$$

(ignoring marginal cost for notational simplicity).
We would simply choose

$$i^* \in \arg\max_i \mu_i.$$

But suppose q_i is unknown. When customer t arrives and we charge price p_i, we observe

$$Y_t = \mathbf{1}\{\text{customer purchases}\} \in \{0, 1\}$$

and receive revenue

$$X_t = p_i Y_t.$$

Thus, when price p_i is chosen,

$$\mathbb{E}[Y_t] = q_i, \qquad \mathbb{E}[X_t] = \mu_i = p_i q_i.$$

The seller now faces a new problem:

To estimate the revenue of a price, we must actually charge that price.

Unlike an offline statistical exercise, data collection affects current profit.

3.5 A focused probability review: estimating a purchase probability

Suppose we charge a fixed price p_i on n independent customers and observe

$$Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \text{Bernoulli}(q_i).$$

The empirical purchase probability is

$$\widehat{q}_i = \frac{1}{n}\sum_{s=1}^{n} Y_s.$$

Then

$$\mathbb{E}[\widehat{q}_i] = q_i, \qquad \mathrm{Var}(\widehat{q}_i) = \frac{q_i(1-q_i)}{n} \le \frac{1}{4n}.$$

The empirical mean revenue is

$$\widehat{\mu}_i = p_i \widehat{q}_i.$$

The law of large numbers says $\widehat{q}_i \to q_i$, but for sequential decisions we need a *finite-sample* statement.

3.5.1 Hoeffding's inequality

For iid Bernoulli observations,

$$\mathbb{P}\left(|\widehat{q}_i - q_i| \ge r\right) \le 2e^{-2nr^2}. \tag{3.8}$$

Equivalently, for confidence level $1-\delta$,

$$|\widehat{q}_i - q_i| \le \sqrt{\frac{\log(2/\delta)}{2n}} \tag{3.9}$$

with probability at least $1-\delta$.

The width shrinks like $1/\sqrt{n}$. This simple fact will generate the exploration bonus in Lecture 4.

Remark 3.3 (Scaling to revenue). *If price p_i produces rewards in $[0, p_i]$, we can estimate q_i and multiply its confidence radius by p_i. Alternatively, divide every revenue observation by one common upper bound $p_{\max} = \max_i p_i$; this puts all rewards in $[0,1]$ without changing which price maximizes expected revenue. We use this common normalization in the basic bandit development.*

3.6 Why greedy learning can fail

The most obvious adaptive pricing rule is:

> At each round, charge the price with the largest empirical mean revenue so far.

This is the *greedy* rule. It can get trapped by early noise.

Example 3.4 (Two arms and one unlucky observation). *Consider two actions with Bernoulli rewards:*

$$\mu_1 = 0.5, \qquad \mu_2 = 0.6.$$

Suppose we try each action once, and happen to observe

$$X_1 = 1, \qquad X_2 = 0.$$

Then the empirical means are

$$\widehat{\mu}_1 = 1, \qquad \widehat{\mu}_2 = 0.$$

A purely greedy rule now chooses arm 1. If it never returns to arm 2, then it never gets evidence that arm 2 is actually better.

This is not merely a bad estimate. It is a *feedback loop*:

$$\boxed{\text{decision} \longrightarrow \text{data observed} \longrightarrow \text{future decision.}}$$

The arm that looks poor receives little data, so our belief that it is poor may never be corrected.

3.6.1 Linear regret is possible

Suppose the bad initial event above causes the algorithm to play arm 1 for almost all of the remaining T rounds. The per-round loss relative to the best arm is

$$\Delta = \mu_2 - \mu_1 = 0.1.$$

Thus the total expected loss on that event is of order

$$\Delta T = \Theta(T).$$

Since the misleading initial event has nonzero probability, pure greedy can have expected regret that grows linearly with the horizon.

The issue is fundamental: an action can look bad because it is truly bad, or because we have not sampled it enough.

3.7 Exploration versus exploitation

This suggests two competing objectives:

- **Exploit:** choose the action that currently looks best.

- **Explore:** choose an uncertain action to learn whether it might be better.

Exploration sacrifices current expected reward for information that may improve future decisions.

This is the basic multi-armed bandit problem.

3.8 The stochastic multi-armed bandit model

There are K arms and T rounds. Arm i has an unknown reward distribution supported on $[0, 1]$ with mean

$$\mu_i.$$

At round t:

1. the algorithm selects an arm I_t;
2. an independent reward X_t is drawn from that arm's distribution;
3. the algorithm observes X_t but not the rewards it would have obtained from the other arms.

Let

$$\mu^* = \max_i \mu_i, \qquad \Delta_i = \mu^* - \mu_i.$$

The benchmark is a clairvoyant decision maker who knows the best arm in advance and plays it in every round.

3.9 Regret

Define cumulative regret by

$$\mathrm{Reg}(T) = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right]. \tag{3.10}$$

If $N_i(T)$ denotes the number of times arm i is selected in the first T rounds, then

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \sum_i \mu_i \mathbb{E}[N_i(T)].$$

Since $\sum_i N_i(T) = T$,

$$\mathrm{Reg}(T) = \sum_i (\mu^* - \mu_i)\mathbb{E}[N_i(T)] \tag{3.11}$$

$$= \boxed{\sum_i \Delta_i \mathbb{E}[N_i(T)].} \tag{3.12}$$

This decomposition turns the problem into a very concrete question:

How many times must we play each suboptimal arm before we are confident enough to stop?

If $\mathrm{Reg}(T) = o(T)$, then average regret $\mathrm{Reg}(T)/T$ vanishes: asymptotically, almost all decisions are as good as if we knew the best arm from the start.

Here the horizon T is another form of scale: more periods can supply more information. But scale helps only if the policy actually gathers informative samples. A greedy policy may generate an arbitrarily long history while learning nothing new about an arm it abandoned too early.

3.10 A first fix: forced exploration *

One simple solution is to reserve some rounds for exploration.

Explore-then-commit:
sample each arm n times, estimate all means, then play the empirically best arm for the rest of the horizon.

If n is too small, we may commit to the wrong arm. If n is too large, we waste too many rounds exploring obviously inferior arms. Balancing these effects gives a sublinear regret guarantee, but not the logarithmic, instance-dependent behavior we would like.

The problem is that this exploration is *non-adaptive*: we decide in advance how much to learn about every arm, regardless of what the data say.

3.11 Why adaptive exploration comes next

The ingredients are now in place:

1. empirical means estimate unknown rewards;
2. confidence radii shrink like $1/\sqrt{n}$;

3. greedy can fail because uncertain actions may be abandoned too early;
4. regret measures the opportunity cost of learning;
5. non-adaptive exploration wastes samples on arms that are already clearly poor.

The natural question is:

Can we make uncertainty itself a reason to explore?

Lecture 4 answers this using *optimism under uncertainty* and the UCB algorithm.

Suggested reading

For the bandit model, regret, uniform exploration, and confidence-bound viewpoint, see Aleksandrs Slivkins, *Introduction to Multi-Armed Bandits*, Chapter 1, especially Sections 1.1–1.3. The book also gives a useful lower-bound perspective in Chapter 2, explaining why distinguishing nearby reward distributions necessarily requires samples.