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

Lecture 4 — Learning to Price: Optimism and UCB

Lecturer: Sid Banerjee *Date: September 3, 2026*

4.1 Overview

Lecture 3 introduced the stochastic multi-armed bandit problem. The key difficulty was that the data we observe depend on the actions we choose. Purely greedy decision making can lock onto the wrong arm, while uniform exploration wastes samples on arms that are already clearly inferior.

We now develop a more adaptive principle:

When uncertain, act as if the world might be favorable.

This is *optimism under uncertainty*. It leads to the Upper Confidence Bound (UCB) algorithm.

The mathematical goal is to understand why a suboptimal arm with gap Δ_i is selected only about

$$O\left(\frac{\log T}{\Delta_i^2}\right)$$

times, leading to regret

$$O\left(\sum_{i:\Delta_i>0} \frac{\log T}{\Delta_i}\right).$$

We will then return to pricing and discuss what is special about price as a bandit action.

The main path is

$$\boxed{\text{confidence interval} \longrightarrow \text{UCB} \longrightarrow \Delta_i \leq 2r_i \longrightarrow \text{regret.}}$$

Sections marked *are optional reading.

4.2 Bandit setup and notation

There are K arms. Arm i has iid rewards in $[0,1]$ with unknown mean μ_i. Let

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

Let

$$N_i(t) = \sum_{s=1}^{t} \mathbf{1}\{I_s = i\}$$

be the number of times arm i is selected in the first t rounds, and let $\widehat{\mu}_i(t)$ be its empirical mean based on those observations.

As in Lecture 3,

$$\mathrm{Reg}(T) = \sum_i \Delta_i \mathbb{E}[N_i(T)].$$

Thus the central issue is to control how often suboptimal arms are played.

4.3 Confidence intervals from Hoeffding

For a fixed arm i and a fixed sample size n, Hoeffding's inequality gives

$$\mathbb{P}\left(|\widehat{\mu}_{i,n} - \mu_i| \geq r\right) \leq 2e^{-2nr^2}.$$

Choose

$$r(n) = \sqrt{\frac{2\log T}{n}}. \tag{4.1}$$

Then

$$\mathbb{P}\left(|\widehat{\mu}_{i,n} - \mu_i| > r(n)\right) \leq 2T^{-4}.$$

4.3.1 A clean event

We would like all confidence intervals to be valid simultaneously over all arms and all sample sizes. Define

$$\mathcal{E} = \left\{|\widehat{\mu}_{i,n} - \mu_i| \leq r(n) \text{ for every } i \in [K],\ n \leq T\right\}. \tag{4.2}$$

A union bound gives

$$\mathbb{P}(\mathcal{E}^c) \leq 2KT\,T^{-4}.$$

If $K \leq T$, this is at most $2/T^2$.

On the clean event, we can stop doing probability and reason deterministically using the confidence intervals.

Remark 4.1 (Why sample size is subtle). *In an adaptive algorithm, $N_i(t)$ is random and depends on previous rewards. A clean way to justify the concentration bound is to imagine an infinite "reward tape" for each arm: the nth time arm i is played, the algorithm reveals the nth iid draw on that arm's tape. Concentration is applied for each fixed prefix length n, and then union-bounded over n and i. This is the elementary device used in Slivkins' treatment of stochastic bandits.*

4.4 Upper confidence bounds

Define

$$\mathrm{UCB}_i(t) = \widehat{\mu}_i(t-1) + r(N_i(t-1)), \tag{4.3}$$
$$\mathrm{LCB}_i(t) = \widehat{\mu}_i(t-1) - r(N_i(t-1)). \tag{4.4}$$

The UCB algorithm is:

1. Play each arm once.
2. For $t > K$, choose
$$I_t \in \arg\max_i \mathrm{UCB}_i(t). \tag{4.5}$$

Why is this sensible? An arm has a large UCB for one of two reasons:

- its empirical mean is large — a reason to *exploit*;
- its confidence radius is large — a reason to *explore*.

UCB combines these two motives in a single index.

4.5 The key UCB inequality

Let i^* be an optimal arm. Suppose UCB chooses arm i at round t. By definition,

$$\mathrm{UCB}_i(t) \geq \mathrm{UCB}_{i^*}(t).$$

On the clean event,

$$\mathrm{UCB}_{i^*}(t) \geq \mu^*.$$

Also,

$$\widehat{\mu}_i(t-1) \leq \mu_i + r(N_i(t-1)),$$

so

$$\mathrm{UCB}_i(t) = \widehat{\mu}_i(t-1) + r(N_i(t-1)) \leq \mu_i + 2r(N_i(t-1)).$$

Therefore

$$\mu^* \leq \mu_i + 2r(N_i(t-1)), \tag{4.6}$$

or

$$\boxed{\Delta_i \leq 2r(N_i(t-1)).} \tag{4.7}$$

This is the whole proof in one line:

If UCB still chooses a bad arm, then that arm must still be sufficiently uncertain.

Substituting (4.1) into (4.7),

$$\Delta_i \le 2\sqrt{\frac{2\log T}{N_i(t-1)}}.$$

Rearranging,

$$N_i(t-1) \le \frac{8\log T}{\Delta_i^2}. \tag{4.8}$$

Thus, apart from initialization and the rare bad event, a suboptimal arm cannot be selected after it has accumulated more than order $\log T/\Delta_i^2$ samples. Its final count can exceed the displayed threshold by at most the last selected round.

4.6 Gap-dependent regret

Using the regret decomposition,

$$\mathrm{Reg}(T) = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[N_i(T)].$$

Equation (4.8) suggests

$$\mathbb{E}[N_i(T)] = O\left(1 + \frac{\log T}{\Delta_i^2}\right).$$

Hence

$$\boxed{\mathrm{Reg}(T) = O\left(\sum_{i:\Delta_i>0} \frac{\log T}{\Delta_i}\right).} \tag{4.9}$$

The contribution of the bad event is negligible because regret is at most T while $\mathbb{P}(\mathcal{E}^c) = O(T^{-2})$.

Remark 4.2 (Instance-dependent versus worst-case guarantees). *The bound (4.9) is excellent when every suboptimal arm is separated from the optimum by a nontrivial gap. But it becomes large when some Δ_i is very small. A complementary argument splits arms into "small-gap" and "large-gap" groups and yields an instance-independent bound of order*

$$O\left(\sqrt{KT\log T}\right).$$

Sharp constants are secondary here; the important point is the qualitative difference between logarithmic instance-dependent regret and square-root worst-case regret.

4.7 A second adaptive view: successive elimination *

UCB is not the only way to use confidence intervals. Another natural algorithm repeatedly samples all active arms and removes arm i once

$$\mathrm{UCB}_i(t) < \mathrm{LCB}_j(t)$$

for some other active arm j.

The same clean-event proof implies that the best arm is never eliminated, and an arm with gap Δ_i is discarded after roughly

$$O\left(\frac{\log T}{\Delta_i^2}\right)$$

samples.

This alternate view is useful because it isolates the deeper principle:

adaptive exploration means spending samples only until an action is statistically resolved.

4.8 Why exploration is unavoidable *

Could a clever algorithm avoid trying uncertain arms altogether? No. To know whether an arm with mean $1/2$ is actually slightly better or slightly worse than another arm, one needs enough samples to distinguish nearby probability distributions.

For a Bernoulli coin with mean

$$\frac{1}{2} + \delta$$

versus one with mean

$$\frac{1}{2} - \delta,$$

the standard error after n samples is of order $1/\sqrt{n}$. Distinguishing the two therefore requires roughly

$$n = \Omega\left(\frac{1}{\delta^2}\right)$$

samples. More refined lower bounds use KL divergence, but the core message is already visible here:

learning a gap of size Δ costs on the order of $1/\Delta^2$ observations.

This explains why the UCB sample complexity in (4.8) has exactly that dependence, up to the logarithmic confidence factor.

4.9 Back to pricing: prices as arms

Suppose a seller chooses from a finite price set

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

At each customer arrival, price p_i is offered. The customer purchases with probability

$$q_i = 1 - F(p_i),$$

which is unknown to the seller. Revenue is

$$X_i = \begin{cases} p_i, & \text{with probability } q_i, \\ 0, & \text{otherwise.} \end{cases}$$

Thus

$$\mu_i = p_i q_i.$$

After normalizing by the largest possible price, this is exactly a stochastic bandit.

UCB is a canonical model for the exploration–exploitation phenomenon. It should not be read as a claim that a production pricing system literally treats every price as an unrelated arm; the structured feedback discussed below often supports more specialized algorithms.

So a direct pricing implementation of UCB is simply:

$$\boxed{\text{offer the price with the largest optimistic estimate of expected revenue.}} \tag{4.10}$$

This captures the idea that

$$\boxed{\text{a price is both a revenue decision and an experiment.}}$$

4.10 But pricing has extra structure

Treating each candidate price as an unrelated arm throws away useful information.

Suppose a customer has a fixed valuation V. If we offer price p and observe a sale, then we learn

$$V \geq p.$$

This implies that the customer would also have purchased at every lower price. Conversely, if there is no sale at p, then

$$V < p,$$

so the customer would not have purchased at any higher price.

Thus pricing has *partial feedback* across prices:

- sale at p gives information about all lower prices;
- no sale at p gives information about all higher prices.

There is also functional structure. The mean revenue is not an arbitrary vector:

$$\mu(p) = p(1 - F(p)).$$

Nearby prices typically have related purchase probabilities and revenues.

The basic UCB reduction ignores both forms of structure. More specialized dynamic-pricing algorithms exploit them.

4.11 A glimpse of continuum pricing *

Suppose prices can be any value in an interval, say $p \in [0, 1]$. One simple idea is to discretize into K grid points and run a finite-armed bandit algorithm.

This creates two sources of regret:

1. **learning error:** regret from not knowing which grid price is best;
2. **discretization error:** regret because the true optimal price may lie between grid points.

If the revenue curve is sufficiently smooth, a finer grid reduces discretization error but creates more arms to learn. Choosing the grid therefore involves a bias–variance-like tradeoff.

This is the starting point of the dynamic-pricing literature exemplified by Kleinberg–Leighton: exploit the order and smoothness of prices rather than treating prices as unrelated actions. We will not develop the full theory here; the finite-armed model already gives us the core exploration–exploitation ideas that recur later in the course.

4.12 Proof pattern to remember

The UCB proof illustrates a template we will use again:

1. define an ideal benchmark;
2. construct a high-probability "clean event";
3. show that a wrong decision on the clean event can happen only while some uncertainty remains large;
4. bound the number of such decisions;

5. charge the rare complement of the clean event separately.

More rounds shrink uncertainty only when the decision rule continues to generate informative observations. UCB is one way to convert scale into learning: uncertainty itself forces the samples needed to resolve it.

Later, the Bayes Selector analysis will use a different algorithm and benchmark, but a related high-level idea: control regret by controlling the probability or frequency with which the online action disagrees with an ideal offline decision.

4.13 Takeaways

1. Greedy fails because uncertainty and low quality are confounded.
2. Confidence intervals quantify uncertainty.
3. UCB uses optimism to turn uncertainty into exploration.
4. A suboptimal arm with gap Δ needs only about $O(\log T/\Delta^2)$ samples before it is resolved.
5. This yields logarithmic instance-dependent regret.
6. Dynamic pricing is a bandit problem, but with useful order, smoothness, and partial-feedback structure across prices.

Suggested reading

Aleksandrs Slivkins, *Introduction to Multi-Armed Bandits*, Chapter 1, especially Sections 1.3.1–1.3.3 for confidence bounds, successive elimination, and UCB; Chapter 2 for the information-theoretic lower-bound viewpoint. For pricing as a bandit problem, see the ORIE 6154 notes on Bandits in Revenue Management and the Kleinberg–Leighton dynamic-pricing framework.