

Lecture 5: Stochastic Multi-Armed Bandits

Lecturer: *Mengxiao Zhang*

Week 5

1 From Full Information to Bandit Feedback

In Lecture 2, we consider the problem where the learner chose a distribution over experts and then observed the loss of every expert. Even an expert receiving zero weight could be evaluated after the outcome arrived. Return now to the advertising example from Lecture 1. A platform displays one advertisement to a visitor and observes whether the visitor clicks. It does not observe whether the same visitor would have clicked on a different advertisement. Unlike the forecasting example, there is no common label that automatically evaluates every alternative. A seller choosing among several prices faces a similar issue: the observed sale or non-sale gives the revenue at the posted price, but generally does not determine revenue at every other price. Without an additional model relating the actions, these missing outcomes cannot be filled in from the observed one.

This changes the role of a decision. Displaying an advertisement earns a reward, but it also produces a *sample* from that advertisement's unknown response distribution. If an advertisement is never displayed, its performance remains uncertain. Choosing the action that currently looks best can therefore prevent the learner from collecting the evidence needed to discover a better action. The learner must balance *exploitation*, which uses the information already available, and *exploration*, which acquires information that can improve future decisions.

To isolate this difficulty, suppose the environment is stationary: each advertisement has a fixed but unknown click probability, and successive visitors generate independent outcomes from a stable population. We temporarily set aside visitor features, changing preferences, and relationships between advertisements. Even with these simplifications, this learning process remains nontrivial because the learner controls which distributions are sampled. Relative to full information under the same stochastic assumptions, the learner receives less information and must pay for the observations it needs. The stochastic assumption makes the environment more structured than the adversarial models studied earlier; it does not remove the difficulty caused by missing feedback.

How can a learner discover which action is best while losing as little as possible during the discovery process?

2 Stochastic Multi-Armed Bandits (MAB)

2.1 The learning protocol

To handle this, we introduce the following multi-armed bandits (MAB) problem. There are $K \geq 2$ actions, also called *arms*, indexed by $[K] \triangleq \{1, \dots, K\}$. The terminology comes from choosing among slot machines with unknown payoff distributions. We keep the loss convention from the earlier lectures: smaller values are better. A reward $r_{t,i} \in [0, 1]$ can be converted to loss $\ell_{t,i} = 1 - r_{t,i}$

without changing the learning problem. Arm i has an unknown loss distribution ν_i supported on $[0, 1]$, with mean $\mu_i \triangleq \mathbb{E}_{X \sim \nu_i}[X]$.

At each round $t = 1, \dots, T$,

1. the learner selects an arm $I_t \in [K]$ using the previous actions and observations;
2. a fresh loss vector $\ell_t \in [0, 1]^K$ is drawn independently of the past, with $\ell_{t,i} \sim \nu_i$;
3. the learner incurs and observes only ℓ_{t,I_t} .

The loss vectors are independent and identically distributed across rounds, which we call *stochastic MAB*. Their coordinates need not be independent. The learner may randomize, but it cannot use the current loss vector when choosing I_t . We assume the horizon T is known and $T \geq K$. If $T < K$, the trivial regret bound T is available, but there is not enough time even to sample every arm once.

For the advertising example, μ_i is the probability of no click on advertisement i . A stationary bandit model is appropriate when the distribution of visitors remains fixed and a visitor's response is not affected by earlier recommendations.

2.2 The benchmark and the gap decomposition

Define $\mu^* \triangleq \min_{i \in [K]} \mu_i$, $i^* \in \arg \min_{i \in [K]} \mu_i$, $\Delta_i \triangleq \mu_i - \mu^*$. The quantity Δ_i is the *suboptimality gap* of arm i : every pull of this arm costs Δ_i in expected loss relative to an optimal arm. We allow several optimal arms, so a sum over suboptimal arms always means a sum over i with $\Delta_i > 0$. The natural benchmark is the best fixed arm in expectation. Its cumulative expected loss is $T\mu^*$. We therefore study the *pseudo-regret*

$$\overline{\text{Reg}}_T \triangleq \max_{i \in [K]} \mathbb{E} \left[\sum_{t=1}^T (\ell_{t,I_t} - \ell_{t,i}) \right] = \mathbb{E} \left[\sum_{t=1}^T \ell_{t,I_t} \right] - T\mu^*. \quad (1)$$

The expectation includes both the random observations and any randomization used by the learner. This differs from the expected hindsight regret considered earlier:

$$\mathbb{E}[\text{Reg}_T] = \mathbb{E} \left[\sum_{t=1}^T \ell_{t,I_t} - \min_{i \in [K]} \sum_{t=1}^T \ell_{t,i} \right] \geq \overline{\text{Reg}}_T.$$

Here the hindsight comparator can select an arm that was unusually lucky on the realized sample. In contrast, pseudo-regret compares with an arm having the smallest true mean. For example, if all means are equal, every algorithm has zero pseudo-regret, even though its expected hindsight regret can be positive. The two criteria should therefore not be interchanged when comparing bounds.

We can express pseudo-regret more simply by counting how often each arm is selected. Denote by $n_{t,i} \triangleq \sum_{s=1}^t \mathbf{1}\{I_s = i\}$ the number of pulls of arm i through round t , with $n_{0,i} = 0$. This count is random because the selected actions depend on random observations and possibly on the learner's own randomization. Each pull of arm i incurs expected excess loss Δ_i , which gives the following decomposition.

Lemma 2.1. *For every stochastic bandit algorithm, we have*

$$\overline{\text{Reg}}_T = \mathbb{E} \left[\sum_{t=1}^T \Delta_{I_t} \right] = \sum_{i: \Delta_i > 0} \Delta_i \mathbb{E}[n_{T,i}]. \quad (2)$$

Proof. Since the learner chooses I_t before observing the fresh loss vector, the tower property gives $\mathbb{E}[\ell_{t,I_t}] = \mathbb{E}[\mu_{I_t}]$. Subtracting $T\mu^*$ and grouping the terms by the selected arm gives $\overline{\text{Reg}}_T = \mathbb{E}\left[\sum_{t=1}^T (\mu_{I_t} - \mu^*)\right] = \mathbb{E}\left[\sum_{i=1}^K \Delta_i n_{T,i}\right]$. $\square$

The identity turns a regret proof into a counting argument: we need to control how often each suboptimal arm is selected. It also captures the statistical tradeoff. An arm with a large gap is expensive to play but easy to distinguish from an optimal arm; an arm with a small gap is harder to distinguish but less expensive to play.

2.3 Why greedy learning can still fail

In Lecture 2, we saw that Follow-the-Leader can fail in the expert problem because an adversarial loss sequence can repeatedly make the current leader perform poorly. Here, each arm has a fixed loss distribution, so that adversarial construction no longer applies. Might a greedy strategy work in this simpler environment? We could sample every arm once and then choose the arm with the smallest average loss so far. For $n_{t,i} > 0$, the empirical mean of arm i is $\hat{\mu}_{t,i} \triangleq \frac{1}{n_{t,i}} \sum_{s=1}^t \mathbf{1}\{I_s = i\} \ell_{s,i}$. Every term used here is observed. The greedy rule selects $I_t \in \arg \min_i \hat{\mu}_{t-1,i}$. Since the means are fixed, it is tempting to expect these estimates to become accurate and the learner to settle on an optimal arm. But an estimate improves only if its arm is sampled again. Greedy selection can prevent exactly those observations that would correct an early mistake.

Theorem 2.2. *There is a two-arm stochastic bandit instance on which the greedy algorithm, after pulling each arm once, has $\overline{\text{Reg}}_T \geq (T-2)/16$ for every $T \geq 2$. In particular, its pseudo-regret can grow linearly with T .*

Proof. Consider two arms. Arm 1 has Bernoulli loss with mean $1/4$, while arm 2 always has loss $1/2$. Arm 1 is optimal. With probability $1/4$, its first observed loss is 1. After the initialization, the empirical means are then 1 and $1/2$. Greedy selects arm 2 forever: its average stays at $1/2$, and arm 1's average stays at 1. Consequently, for $T \geq 2$,

$$\overline{\text{Reg}}_T \geq \frac{1}{4} \cdot \frac{1}{4}(T-2) = \frac{T-2}{16}.$$

$\square$

The failure now comes from missing feedback rather than an adversarially changing environment. After one unlucky observation, the optimal arm looks worse and is never tried again. Under full information, its observations would continue to arrive and correct the estimate. Under bandit feedback, the learner must deliberately return to that arm to find out whether the initial observation was misleading.

3 Explore-then-Exploit

The greedy example suggests a simple repair: collect enough observations of every arm before deciding which one to use. Choose an integer m with $1 \leq m \leq \lfloor T/K \rfloor$, pull every arm m times, and then use the arm with the smallest empirical mean for the remaining rounds. This is *explore-then-exploit*, also called *explore-then-commit*. It separates the rounds devoted to learning from those devoted to using what has been learned.

Algorithm 1: Explore-then-Exploit

Input: arms $[K]$, horizon T , samples per arm m with $Km \leq T$

for $s = 1, \dots, m$ **do**

 pull each arm once and record its observed loss;

 Compute $\hat{\mu}_i$, the average of the m observations of arm i , for every i ;

 Choose $\hat{i} \in \arg \min_{i \in [K]} \hat{\mu}_i$;

for $t = Km + 1, \dots, T$ **do**

 pull arm $\hat{i}$;

A small m leaves many rounds for exploitation, but an inaccurate comparison may commit the learner to the wrong arm. A large m makes the comparison more reliable, but spends more rounds pulling suboptimal arms during exploration. We need a regret bound that accounts for both costs.

How many exploration rounds are enough? To answer this, we need to relate the number of observations to the accuracy of an empirical mean. Recall Hoeffding's inequality from the first week: for n independent samples $X_1, \dots, X_n \in [0, 1]$ with mean μ ,

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{s=1}^n X_s - \mu \right| > \varepsilon \right) \leq 2e^{-2n\varepsilon^2}. \quad (3)$$

For a desired failure probability $\delta \in (0, 1)$, define $B_\delta \triangleq \ln(2K/\delta)$ and $\rho_\delta(n) \triangleq \sqrt{B_\delta/(2n)}$ for $n \geq 1$. The exploration phase gives exactly m independent observations of each arm, where m is fixed before observing any losses. By Hoeffding's inequality and a union bound over the K arms, with probability at least $1 - \delta$, all empirical means satisfy $|\hat{\mu}_i - \mu_i| \leq \rho_\delta(m)$. Denote this event by

$$\mathcal{H}_\delta \triangleq \{|\hat{\mu}_i - \mu_i| \leq \rho_\delta(m) \text{ for every } i \in [K]\}.$$

Indeed, the failure probability for each arm is at most $2e^{-B_\delta} = \delta/K$, so the probability that any estimate fails is at most δ . We compare the arms only once, at the fixed sample count m , and use the resulting choice throughout exploitation. No union bound over rounds or sample counts is needed.

Theorem 3.1. *For every feasible m and every $\delta \in (0, 1)$, Algorithm 1 satisfies*

$$\overline{\text{Reg}}_T \leq m \sum_{i=1}^K \Delta_i + 2(T - Km) \sqrt{\frac{B_\delta}{2m}} + \delta T. \quad (4)$$

Proof. The exploration phase contributes exactly $m \sum_i \Delta_i$ to pseudo-regret. On $\mathcal{H}_\delta$, the choice of $\hat{i}$ gives

$$\mu_{\hat{i}} \leq \hat{\mu}_{\hat{i}} + \rho_\delta(m) \leq \hat{\mu}_{i^*} + \rho_\delta(m) \leq \mu^* + 2\rho_\delta(m).$$

Thus the gap in each exploitation round is at most $2\rho_\delta(m)$ on $\mathcal{H}_\delta$ and at most 1 otherwise. Since $\mathbb{P}(\mathcal{H}_\delta^c) \leq \delta$, we have $\mathbb{E}[\Delta_{\hat{i}}] \leq 2\rho_\delta(m) + \delta$. The gap decomposition now yields

$$\overline{\text{Reg}}_T = m \sum_{i=1}^K \Delta_i + (T - Km) \mathbb{E}[\Delta_{\hat{i}}] \leq m \sum_{i=1}^K \Delta_i + 2(T - Km) \rho_\delta(m) + \delta T.$$

□

Set $\delta = 1/T$ and denote $B \triangleq \ln(2KT)$. This choice makes the failure contribution δT equal to 1; the dependence on T here comes from choosing δ , not from a union bound over time. Since $\Delta_i \leq 1$, Theorem 3.1 gives

$$\overline{\text{Reg}}_T \leq Km + T\sqrt{\frac{2B}{m}} + 1. \quad (5)$$

More exploration increases the first term and decreases the second. Choose $m = \left\lceil \left(\frac{T^2 B}{K^2}\right)^{1/3} \right\rceil$. Substituting this choice into (5) gives

$$\overline{\text{Reg}}_T = \mathcal{O}\left(K^{1/3}T^{2/3}(\ln(KT))^{1/3}\right) = \mathcal{O}\left(K^{1/3}T^{2/3}(\ln T)^{1/3}\right). \quad (6)$$

The rate reflects the cost of estimating every arm before committing to one of them. Estimation to accuracy ε requires $\mathcal{O}(B/\varepsilon^2)$ observations per arm, giving an exploration cost of $\mathcal{O}(KB/\varepsilon^2)$ and an exploitation cost of $\mathcal{O}(T\varepsilon)$.

4 Arm Elimination

Explore-then-exploit gives every arm the same number of observations. Yet an arm whose mean is far from optimal needs less accurate estimation before it can be ruled out. Can we stop exploring such an arm while continuing to compare the closer competitors?

Example 4.1 (Different arms need different amounts of exploration). Suppose three arms have means 0.20, 0.21, and 0.80. Distinguishing the first two requires much more accurate estimates than ruling out the third. Once the observations show that arm 3 is substantially worse, continuing to sample it wastes rounds that could be used to compare arms 1 and 2.

To decide whether an arm needs more observations, place a confidence interval $[L_i, U_i]$ around each estimated mean. If $L_i > U_j$ for some other arm j , then every mean in arm i 's interval is larger than every mean in arm j 's interval. When the intervals contain the true means, arm i cannot be optimal. We can therefore remove it and spend the remaining rounds sampling arms that could still be best. The stopping decision is now made separately for each arm, based on the observations, rather than by fixing a common exploration budget in advance.

4.1 The algorithm

The elimination rule needs a lower and an upper bound for each active arm. Denote by A_s the active set at the start of sweep s , with $A_1 = [K]$. During each sweep, pull every active arm once. For $i \in A_s$, denote the average of its s observations after the completed sweep by $\tilde{\mu}_{s,i}$, and denote its lower and upper bounds by $L_{s,i}$ and $U_{s,i}$. We will remove arm i if $L_{s,i} > \min_{j \in A_s} U_{s,j}$.

What guarantee do these bounds need? With probability at least $1 - \delta$, we want $L_{s,i} \leq \mu_i \leq U_{s,i}$ for every active arm at every completed sweep. On this event, every elimination is justified by intervals that contain the true means. Would the confidence bound established for explore-then-exploit suffice?

Why adaptive sampling needs a different argument. In explore-then-exploit, each empirical mean uses a fixed number of independent observations and is unbiased. Here, the observations determine which arms survive and receive more samples. The resulting sample count depends on

the sample values, so the empirical mean at that count need not be unbiased. Even though an active arm has exactly s observations after sweep s , conditioning on its survival can change the distribution of those observations.

Example 4.2 (Bias from deciding when to stop). Consider an arm with independent Bernoulli losses of mean $1/2$. Pull it once. If the first loss is 1, stop sampling it; if the first loss is 0, pull it once more and then stop. Use other arms for any remaining rounds. Denote the number of observations of this arm by N and their average by $\bar{X}_N \triangleq N^{-1} \sum_{s=1}^N X_s$. If $N = 1$, the average is necessarily 1. If $N = 2$, the first loss is necessarily 0, so the conditional expected average is $1/4$. Hence

$$\mathbb{E}[\bar{X}_N] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{4} = \frac{5}{8}.$$

The average has upward bias $1/8$. Conditioning on $N = 2$ does not leave two independent Bernoulli samples of mean $1/2$: the condition already reveals that $X_1 = 0$.

The bias need not disappear when more observations are allowed. If the rule instead collects a total of $M \geq 2$ observations after an initial 0, then

$$\mathbb{E}[\bar{X}_N] = \frac{1}{2} + \frac{1}{2} \cdot \frac{M-1}{2M} = \frac{3}{4} - \frac{1}{4M}.$$

The bias approaches $1/4$ as M grows because an initial 1 still causes sampling to stop immediately. Reversing the rule—stopping after an initial 0—produces a downward bias of the same magnitude. These are examples of bias, not universal bounds on its size.

The example shows why we cannot condition on $n_{t,i} = n$ and treat the observations as a fresh sample of size n . Instead, we need a guarantee that holds simultaneously at every possible sample count. Such a guarantee will remain valid when the algorithm chooses a count using the observations. It controls the estimation error with high probability; it does not imply that the empirical mean is unbiased.

Confidence bounds valid throughout the interaction. For each arm i , imagine preparing an infinite sequence $X_{1,i}, X_{2,i}, \dots$ of independent samples from ν_i . Each time the learner pulls arm i , it reveals the next unused sample in this sequence. Thus, whenever $n_{t,i} > 0$, its empirical mean is $\hat{\mu}_{t,i} = n_{t,i}^{-1} \sum_{s=1}^{n_{t,i}} X_{s,i}$. For each fixed n , the first n entries form an independent sample, even though the prefix length chosen by the learner is random.

We can take the sequences independent across arms. Only one arm is observed in each round, so this construction has the same distribution of actions and observed losses as the original protocol, even when coordinates of the original loss vector are correlated.

There are K arms and at most T possible sample counts for each arm. To control all these averages at once, redefine $B_\delta \triangleq \ln(2KT/\delta)$ and $\rho_\delta(n) \triangleq \sqrt{B_\delta/(2n)}$ for elimination and UCB. The extra factor T comes from a union bound over sample counts $n = 1, \dots, T$. This will give the desired guarantee at every round, because each realized count is one of these values.

Lemma 4.3. *For every algorithm, with probability at least $1 - \delta$, the inequality $|\hat{\mu}_{t,i} - \mu_i| \leq \rho_\delta(n_{t,i})$ holds for every arm i and every round $t \leq T$ with $n_{t,i} > 0$. Denote this simultaneous confidence event by*

$$\mathcal{E}_\delta \triangleq \{|\hat{\mu}_{t,i} - \mu_i| \leq \rho_\delta(n_{t,i}) \text{ for all } i \in [K], t \leq T \text{ with } n_{t,i} > 0\}. \quad (7)$$

Proof. For each fixed arm i and sample count $n \leq T$, Hoeffding's inequality applied to its prepared sequence gives

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{s=1}^n X_{s,i} - \mu_i \right| > \rho_\delta(n) \right) \leq 2e^{-B_\delta} = \frac{\delta}{KT}.$$

A union bound over the KT pairs (i, n) shows that all these estimates are accurate simultaneously with probability at least $1 - \delta$. On this event, substitute the realized count $n = n_{t,i}$ to obtain the stated bound for every arm and round. No conditioning on the sample count or on an arm's survival is needed. $\square$

Successive arm elimination. We can now specify the bounds used by the algorithm. After a completed sweep s , define $L_{s,i} \triangleq \tilde{\mu}_{s,i} - \rho_\delta(s)$ and $U_{s,i} \triangleq \tilde{\mu}_{s,i} + \rho_\delta(s)$ for each $i \in A_s$. Lemma 4.3 ensures that $L_{s,i} \leq \mu_i \leq U_{s,i}$ simultaneously for all active arms and completed sweeps on $\mathcal{E}_\delta$. Keep the arms whose lower bounds do not exceed the smallest upper bound:

$$A_{s+1} = \left\{ i \in A_s : L_{s,i} \leq \min_{j \in A_s} U_{s,j} \right\}. \quad (8)$$

All eliminations use the same active set and the same completed sweep. An arm attaining the minimum upper bound always survives because $L_{s,i} \leq U_{s,i}$, so the active set cannot become empty.

Algorithm 2: Successive Arm Elimination

Input: arms $[K]$, horizon T , confidence level $\delta \in (0, 1)$
Initialize $A_1 = [K]$, sweep $s = 1$, and number of completed pulls $t = 0$;
while $t < T$ **do**
 foreach $i \in A_s$ *in a fixed order* **do**
 if $t = T$ **then**
 stop and return;
 pull arm i , record its loss, and set $t \leftarrow t + 1$;
 Compute $\tilde{\mu}_{s,i}$, $L_{s,i}$, and $U_{s,i}$ for all $i \in A_s$;
 Set $A_{s+1} = \{i \in A_s : L_{s,i} \leq \min_{j \in A_s} U_{s,j}\}$;
 Set $s \leftarrow s + 1$;

If only one arm remains, each subsequent sweep simply selects that arm. If the horizon ends in the middle of a sweep, the algorithm stops immediately; it does not spend additional rounds completing the sweep.

Example 4.4 (Eliminating an arm). Empirical means $(0.20, 0.24, 0.60)$ and a common radius 0.05 give confidence intervals $[0.15, 0.25]$, $[0.19, 0.29]$, and $[0.55, 0.65]$. Arm 3 is eliminated because $0.55 > 0.25$. Arms 1 and 2 remain because their intervals do not yet certify which is better. This decision uses only the observations and confidence radii.

Gap-dependent regret. We now analyze the regret. On the simultaneous confidence event, an optimal arm always survives, while a suboptimal arm is removed once its confidence radius is small relative to its gap. This gives the following bound.

Theorem 4.5. *On $\mathcal{E}_\delta$, Algorithm 2 satisfies, for every arm i with $\Delta_i > 0$,*

$$n_{T,i} \leq \min \left\{ T, \left\lfloor \frac{8B_\delta}{\Delta_i^2} \right\rfloor + 1 \right\}. \quad (9)$$

Consequently,

$$\overline{\text{Reg}}_T \leq \sum_{i:\Delta_i>0} \left(\frac{8B_\delta}{\Delta_i} + \Delta_i \right) + \delta T. \quad (10)$$

Proof. First, an optimal arm cannot be eliminated on $\mathcal{E}_\delta$. If $i^* \in A_s$, then $L_{s,i^*} \leq \mu^* \leq \mu_j \leq U_{s,j}$ for every $j \in A_s$. Hence $L_{s,i^*} \leq \min_{j \in A_s} U_{s,j}$, so it survives. Induction from $A_1 = [K]$ shows that every optimal arm remains active.

Next, consider a suboptimal arm $i \in A_s$. The confidence bounds give $L_{s,i} \geq \mu_i - 2\rho_\delta(s)$ and $U_{s,i^*} \leq \mu^* + 2\rho_\delta(s)$. If $4\rho_\delta(s) < \Delta_i$, then $L_{s,i} > U_{s,i^*}$, so arm i is eliminated. This condition is equivalent to $s > 8B_\delta/\Delta_i^2$.

Set $s_i \triangleq \lfloor 8B_\delta/\Delta_i^2 \rfloor + 1$. If arm i remains active through the completed sweep s_i , it is eliminated at the end of that sweep, after exactly s_i pulls. If the horizon ends earlier, including during an incomplete sweep, it has at most s_i pulls. This proves (9).

Multiply the count bound by Δ_i and sum over the suboptimal arms. On $\mathcal{E}_\delta^c$, the sum of the gaps incurred is at most T , and this event has probability at most δ . Taking expectations therefore gives (10). $\square$

Set $\delta = 1/T$ and denote $B \triangleq \ln(2KT^2) = \Theta(\ln T)$ for elimination and UCB. The bound becomes $\overline{\text{Reg}}_T = \mathcal{O}\left(\sum_{i:\Delta_i>0} \frac{\ln T}{\Delta_i}\right)$. The algorithm never uses the gaps. Each gap appears only in the analysis: $\mathcal{O}(B/\Delta_i^2)$ observations, each costing Δ_i , contribute $\mathcal{O}(B/\Delta_i)$ regret.

4.2 Gap-independent bound

The gap-dependent bound contains $1/\Delta_i$. If Δ_i is extremely small, does this mean that the algorithm performs much worse? The answer is no. While a small gap makes an arm harder to distinguish from an optimal arm, it also makes each pull inexpensive. If $\Delta_i = 1/T$, then even pulling that arm in every round contributes only $T\Delta_i = 1$ to pseudo-regret.

This suggests separating the arms at a threshold $\varepsilon > 0$. Arms with gaps at most ε contribute at most $T\varepsilon$ in total, regardless of how often they are selected. For arms with larger gaps, the elimination bound controls their pull counts. Balancing these two contributions gives a bound that does not depend on the individual gaps.

Theorem 4.6. *With $\delta = 1/T$ and $B \triangleq \ln(2KT^2)$, successive elimination satisfies*

$$\overline{\text{Reg}}_T \leq \mathcal{O}(\sqrt{KT \ln T}). \quad (11)$$

Proof. Fix a threshold $\varepsilon > 0$. On $\mathcal{E}_{1/T}$, arms with $\Delta_i \leq \varepsilon$ contribute at most $T\varepsilon$ in total, since there are only T pulls. For arms with $\Delta_i > \varepsilon$, Theorem 4.5 gives

$$\sum_{i:\Delta_i>\varepsilon} \Delta_i n_{T,i} \leq \sum_{i:\Delta_i>\varepsilon} \left(\frac{8B}{\Delta_i} + \Delta_i \right) \leq \frac{8KB}{\varepsilon} + K.$$

Therefore $\overline{\text{Reg}}_T \leq T\varepsilon + 8KB/\varepsilon + K + 1$. Choosing $\varepsilon = \sqrt{8KB/T}$ gives the displayed bound. $\square$

Thus the same procedure gives logarithmic regret on any fixed instance with positive gaps and $\mathcal{O}(\sqrt{KT \ln T})$ regret uniformly over all instances with $K \leq T$. It need not identify every slightly suboptimal arm before the horizon ends. For regret minimization, failing to distinguish nearly equal means can be inexpensive.

5 Upper Confidence Bound (UCB)

Elimination avoids spending the same exploration budget on every arm, but within a sweep it still gives every surviving arm one pull and waits until the sweep is complete before comparing them again. The surviving arms may already have quite different empirical losses. Can we use that information to decide which arm deserves the very next observation? UCB does this by assigning each arm an index that accounts for both its estimated loss and its uncertainty. Every observation can change the next choice, and no arm is permanently removed: it remains available whenever its index becomes competitive. This gives a way to allocate exploration one round at a time while obtaining the same orders of regret as elimination.

5.1 The algorithm

UCB also chooses how often to sample each arm using its previous observations. As the arm-elimination example showed, $\hat{\mu}_{t-1,i}$ need not be an unbiased estimator at this adaptive sample count. We therefore use the same simultaneous confidence guarantee from Lemma 4.3, with $B_\delta \triangleq \ln(2KT/\delta)$ and $\rho_\delta(n) \triangleq \sqrt{B_\delta/(2n)}$. Its union bound over all arms and possible sample counts makes every interval valid throughout the interaction on $\mathcal{E}_\delta$. No new union bound is needed for UCB; the lemma already applies to every sampling algorithm.

After every arm has been sampled at least once, define the round- t lower confidence index $\text{LCB}_{t,i} \triangleq \hat{\mu}_{t-1,i} - \rho_\delta(n_{t-1,i})$. The learner selects $I_t \in \arg \min_{i \in [K]} \text{LCB}_{t,i}$. This is *optimism in the face of uncertainty* (OFU): choose an arm according to a favorable value of its mean that has not yet been ruled out by the observations. The intervals here are allowed to extend beyond $[0, 1]$; this convenient relaxation does not affect their validity or the analysis.

The name *Upper Confidence Bound* comes from the reward convention, where larger values are better. Here we minimize losses and therefore use a lower bound. Indeed, if $r_{t,i} = 1 - \ell_{t,i}$, then $\hat{r}_{t-1,i} = 1 - \hat{\mu}_{t-1,i}$, and

$$\arg \max_i \{\hat{r}_{t-1,i} + \rho_\delta(n_{t-1,i})\} = \arg \min_i \{\hat{\mu}_{t-1,i} - \rho_\delta(n_{t-1,i})\}.$$

We retain the conventional name UCB and use the equivalent loss version throughout.

Algorithm 3: UCB

Input: arms $[K]$, horizon $T \geq K$, confidence level $\delta \in (0, 1)$

Pull each arm once; initialize its count and empirical mean from that observation;

for $t = K + 1, \dots, T$ **do**

choose $I_t \in \arg \min_{i \in [K]} \{\hat{\mu}_{t-1,i} - \rho_\delta(n_{t-1,i})\}$;

pull I_t and observe ℓ_{t,I_t} ;

set $n_{t,I_t} = n_{t-1,I_t} + 1$;

set $\hat{\mu}_{t,I_t} = (n_{t-1,I_t} \hat{\mu}_{t-1,I_t} + \ell_{t,I_t}) / n_{t,I_t}$;

keep $n_{t,i} = n_{t-1,i}$ and $\hat{\mu}_{t,i} = \hat{\mu}_{t-1,i}$ for $i \neq I_t$;

The empirical mean favors arms that have performed well. Subtracting the confidence radius favors arms whose performance remains uncertain.

Example 5.1 (Choosing an uncertain arm). Suppose the empirical losses are $(0.30, 0.38, 0.60)$ and the confidence radii are $(0.04, 0.16, 0.06)$. The lower confidence indices are $(0.26, 0.22, 0.54)$, so UCB selects arm 2 even though arm 1 has the smaller empirical loss. Arm 2's wider interval leaves open

the possibility that it is better. Additional observations will narrow that interval and help resolve the comparison.

With a fixed tie-breaking rule, UCB needs no internal randomization. Its actions are nevertheless random through the observed losses. Unlike greedy selection, it explicitly accounts for the amount of evidence behind each average.

Gap-dependent regret. If UCB selects a suboptimal arm, its lower confidence index must be no larger than the optimal arm's index. On $\mathcal{E}_\delta$, this can happen only while the selected arm's confidence radius is large enough to account for its gap. We use this observation to bound its total number of pulls.

Theorem 5.2. *On $\mathcal{E}_\delta$, every suboptimal arm satisfies*

$$n_{T,i} \leq \min \left\{ T, \left\lfloor \frac{2B_\delta}{\Delta_i^2} \right\rfloor + 1 \right\}. \quad (12)$$

Consequently,

$$\overline{\text{Reg}}_T \leq \sum_{i:\Delta_i>0} \left(\frac{2B_\delta}{\Delta_i} + \Delta_i \right) + \delta T. \quad (13)$$

Proof. On $\mathcal{E}_\delta$, the optimal arm satisfies $\text{LCB}_{t,i^*} \leq \mu^*$. UCB chooses the smallest index, so $\text{LCB}_{t,I_t} \leq \mu^*$ as well. The selected arm's confidence bound gives

$$\mu_{I_t} \leq \hat{\mu}_{t-1,I_t} + \rho_\delta(n_{t-1,I_t}) = \text{LCB}_{t,I_t} + 2\rho_\delta(n_{t-1,I_t}) \leq \mu^* + 2\rho_\delta(n_{t-1,I_t}).$$

Thus, on every round $t > K$,

$$\Delta_{I_t} \leq 2\rho_\delta(n_{t-1,I_t}) = \sqrt{\frac{2B_\delta}{n_{t-1,I_t}}}. \quad (14)$$

If arm i is selected only during initialization, then $n_{T,i} = 1$ and the count bound holds. Otherwise, denote by $t > K$ its final selection time. Just before that selection it has been pulled $n_{T,i} - 1$ times, so (14) gives $n_{T,i} - 1 \leq 2B_\delta/\Delta_i^2$. Since the count is an integer, this proves (12).

On $\mathcal{E}_\delta$, multiply the count bound by Δ_i and sum over the suboptimal arms. On $\mathcal{E}_\delta^c$, the sum of the gaps incurred is at most T . Taking expectations and using $\mathbb{P}(\mathcal{E}_\delta^c) \leq \delta$ gives (13). $\square$

For $\delta = 1/T$, the bound is $\overline{\text{Reg}}_T = \mathcal{O}\left(\sum_{i:\Delta_i>0} \frac{\ln T}{\Delta_i}\right)$, the same logarithmic order as successive elimination. You can also achieve the same $\mathcal{O}(\sqrt{KT \ln T})$ regret by applying a similar argument as in the previous section.