INTRODUCTORY LECTURE NOTES – OPTIMIZATION UNDER UNCERTAINTY COURSE

Multi-Armed Bandits

8-hour course

Contents

1 The bandit problem **1**

1.1 Actions, rewards, and feedback 2

1.2 The adversarial model and the regret metric 2

1.3 Stochastic model 3

2 The failure of greedy and need for exploration **5**

2.1 Definition of greedy and proof of failure 6

2.2 Full information doesn't rescue greedy 7

3 Concentration of empirical means **7**

3.1 The Chernoff bound: exponential moments control tails 8

3.2 From bounded observations to Hoeffding's inequality 9

3.3 Confidence intervals and sample requirements 10

3.4 Comparing two arms 11

4 Explore then commit **12**

4.1 The algorithm 12

4.2 Exploration cost and commitment cost 12

4.3 Two arms: the role of the exploration budget 14

4.4 A budget that does not require knowing the gap 15

These notes are currently in progress and will be repeatedly updated before the exam.

1 The bandit problem

Many applications require a sequence of decisions whose outcomes are initially uncertain. A news recommender selects an article and observes whether the user clicks [4]; an advertising system chooses an advertisement and observes the user's response [1]. In a clinical trial, a patient receives one treatment, and only the outcome of that treatment is observed for that patient [5]. In agricultural field trials, a farmer tests a crop-management practice on a plot and observes its performance under the growing conditions of that trial [2].

These examples share a central difficulty: a decision determines both the **reward** obtained and the **information** available for future decisions. The learner observes the outcome of

the selected action, but not what would have happened under another choice in the same situation. Choosing an action can therefore serve two purposes: obtaining a good outcome now and learning how to make better choices later.

The multi-armed bandit problem isolates this interaction between learning and decision making. An *arm* is an available action, such as an article, treatment, or crop-management practice. The objective is to collect a large cumulative reward while learning from the outcomes of the selected arms. We first allow rewards to vary arbitrarily over time, then study the simpler stochastic model, where each arm has a fixed reward distribution. The applications above motivate this basic model; personalization, delayed outcomes, and risk considerations can require extensions beyond this introductory course.

These notes are largely inspired by *Bandit Algorithms* by Lattimore and Szepesvári [3]. For further details, see its Chapters 1 and 4 for the bandit framework, Chapter 2 for probability, Chapter 5 for concentration inequalities, and Chapter 6 for explore then commit. More specific reading pointers are provided where these tools are introduced.

1.1 Actions, rewards, and feedback

There are $K \geq 2$ arms, represented by an arm set $\mathcal{A} = \{1, \dots, K\}$, and $T \geq 1$ rounds. The number T is the *horizon*. At each round t, the interaction is as follows:

1. The environment assigns a reward vector $r_t = (r_{t,1}, \dots, r_{t,K}) \in [0,1]^K$, hidden from the learner.
2. The learner selects an arm $A_t \in \mathcal{A}$.
3. The learner receives and observes $Y_t = r_{t,A_t}$.

The other coordinates of r_t remain unobserved: this is the *bandit feedback*. Under *full-information feedback*, the entire vector r_t would instead be revealed after the action is selected. In either case, the current rewards are unknown when making the decision.

The learner bases its decision on the observed history

$$H_{t-1} = (A_1, Y_1, \dots, A_{t-1}, Y_{t-1})$$

and may also use internal randomness, for example to explore an arm at random or to break a tie. A *policy* specifies a distribution $\pi_t(\cdot \mid H_{t-1})$ over the next arm for each possible history. It can use past observations, but cannot use unobserved rewards or future information. A deterministic policy is a special case that assigns probability one to a single arm for every history.

Rewards and losses. We maximize rewards throughout these notes. Equivalently, one can minimize losses $\ell_{t,i} = 1 - r_{t,i}$. The corresponding regret is the same. Fixing one convention avoids reversing inequalities when comparing algorithms.

1.2 The adversarial model and the regret metric

In the *adversarial model*, rewards need not follow a fixed distribution. They can vary arbitrarily from round to round. For our initial definitions, assume that the entire reward table $(r_{t,i})_{1 \leq t \leq T,\, 1 \leq i \leq K}$ is fixed before interaction begins. Such an adversary is called *oblivious*. The learner does not know this table.

More generally, a *non-anticipating adaptive adversary* may use the past interaction to choose r_t, but cannot see the learner's current private random draw before choosing that vector. In these notes we keep only consider the oblivious model for simplicity.

To measure performance, we must choose a comparator. A comparator that picks the best arm separately at every round can exploit knowledge of that round's hidden rewards. For arbitrary reward sequences, a learner cannot generally match such a comparator. We instead compare with the *best fixed arm in hindsight*: the single arm with the largest total reward over all T rounds.

Definition 1 (Adversarial regret)**.** For a reward table and a realized sequence of actions, the regret is

$$\widehat{\mathcal{R}}_T = \max_{i \in \mathcal{A}} \sum_{t=1}^{T} r_{t,i} - \sum_{t=1}^{T} r_{t,A_t}. \tag{1.1}$$

For a fixed reward table, the expected regret is $\mathcal{R}_T^{\mathrm{adv}} = \mathbb{E}[\widehat{\mathcal{R}}_T]$, where the expectation is over the learner's randomization.

The comparator selects one arm and uses it throughout all T rounds. It knows the table when selecting that arm; the learner must make decisions sequentially. The realized regret can be negative: a learner that switches arms may outperform every fixed arm.

We seek guarantees that hold for every admissible reward table. For fixed K, a *no-regret* guarantee bounds the worst-case expected regret by a function $g(T) = o(T)$. Equivalently, for every reward table,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r_{t,A_t}\right] \geq \max_{i \in \mathcal{A}} \frac{1}{T}\sum_{t=1}^{T} r_{t,i} - \frac{g(T)}{T}.$$

Thus the learner's expected reward per round is at least the best fixed arm's average reward, minus an error that tends to zero. The learner may also do better than that arm. A regret bound proportional to T does not give a vanishing error per round.

1.3 Stochastic model

In a *stationary stochastic bandit*, each arm i has an unknown reward distribution ν_i supported on $[0, 1]$. Every pull of that arm produces a new independent sample from ν_i. The distributions remain unchanged over time.

A useful construction assigns an infinite sequence of potential observations to each arm:

$$X_{i,1}, X_{i,2}, \ldots \overset{\text{iid}}{\sim} \nu_i.$$

We take these sequences to be independent across arms as well. The nth pull of arm i reveals $X_{i,n}$, whether that pull occurs early or late. Define

$$N_i(t) = \sum_{s=1}^{t} \mathbf{1}\{A_s = i\}, \qquad N_i(0) = 0.$$

Then $Y_t = X_{A_t, N_{A_t}(t)}$. Because actions depend on earlier rewards, the observed rewards $Y_1, \ldots, Y_T$ generally do not form an iid sample from a single distribution. Define the expected reward of arm i by $\mu_i = \mathbb{E}[X_{i,1}]$. Conditional on the past history and the selected arm, the next reward is an independent draw from that arm's distribution, so

$$\mathbb{E}[Y_t \mid H_{t-1}, A_t] = \mu_{A_t}. \tag{1.2}$$

We further define

$$\mu^\star = \max_{i \in \mathcal{A}} \mu_i, \qquad a^\star \in \arg\max_{i \in \mathcal{A}} \mu_i.$$

An optimal arm maximizes the expected reward, rather than the reward of any particular draw. Several arms may be optimal, and $a^\star$ denotes one of them.

Definition 2 (Gaps and stochastic pseudo-regret)**.** The *suboptimality gap* of arm i is $\Delta_i = \mu^\star - \mu_i$. The expected stochastic pseudo-regret is

$$R_T = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]. \tag{1.3}$$

We will simply call R_T the stochastic regret.

Proposition 1 (Regret decomposition)**.** *For any policy in the stochastic model,*

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{A_t}\right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[N_i(T)]. \tag{1.4}$$

Proof. By Equation (1.2) and the law of total expectation,

$$\mathbb{E}[Y_t] = \mathbb{E}[\mathbb{E}[Y_t \mid H_{t-1}, A_t]] = \mathbb{E}[\mu_{A_t}].$$

Linearity of expectation then gives

$$R_T = \sum_{t=1}^{T} (\mu^\star - \mathbb{E}[Y_t]) = \mathbb{E}\left[\sum_{t=1}^{T} (\mu^\star - \mu_{A_t})\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{A_t}\right].$$

Finally, group the terms according to the selected arm:

$$\sum_{t=1}^{T} \Delta_{A_t} = \sum_{i=1}^{K} \Delta_i \sum_{t=1}^{T} \mathbf{1}\{A_t = i\} = \sum_{i=1}^{K} \Delta_i N_i(T).$$

Taking expectations and using linearity once more proves the second equality in Equation (1.4). □

Probability toolkit

The exam will assume familiarity with the following rules and with elementary conditional probability and independence. For the proof above, the key tools are:

1. **Linearity of expectation.** For integrable random variables $Z_1, \ldots, Z_n$ and constants $c_1, \ldots, c_n$,
$$\mathbb{E}\left[\sum_{j=1}^{n} c_j Z_j\right] = \sum_{j=1}^{n} c_j \mathbb{E}[Z_j].$$
Independence is not required. In particular, a finite sum can be moved inside or outside an expectation even when the actions are adaptive.
2. **Total expectation (tower property).** Averaging a conditional expectation gives the original expectation:
$$\mathbb{E}[\mathbb{E}[Z \mid W]] = \mathbb{E}[Z].$$
Here $W = (H_{t-1}, A_t)$: first average the reward for a fixed history and chosen arm, then average over histories and arm choices.
3. **Indicators and counting.** For an event B, $\mathbb{E}[\mathbf{1}\{B\}] = \mathbb{P}(B)$. Thus
$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}(A_t = i).$$
Counts such as $N_i(T)$ are random; linearity still applies to them.

Probability toolkit (continued)

Also review the union bound and the distinction between linearity and factorization: $\mathbb{E}[ZW] = \mathbb{E}[Z]\mathbb{E}[W]$ holds for independent integrable variables, but not for arbitrary dependent variables.

Reading. Chapter 2 of Lattimore and Szepesvári [3] reviews important probability tools for bandits. Section 4.5 gives the bandit regret decomposition. Focus on applying these identities to finite sums and conditional expectations.

Returning to the bandit problem, Proposition 1 shows that each arm contributes its gap multiplied by its expected number of pulls. Every pull of a suboptimal arm therefore costs its gap *in expectation*. Pulling an optimal arm contributes zero pseudo-regret, even if its observed reward happens to be small.

Example 1 (Two Bernoulli arms)**.** Suppose $\nu_1 = \text{Bernoulli}(3/4)$ and $\nu_2 = \text{Bernoulli}(1/4)$. Arm 1 is optimal, and $\Delta_2 = 1/2$. Therefore $R_T = \frac{1}{2}\mathbb{E}[N_2(T)]$. The goal is to collect just enough evidence to discard arm 2 and mostly play arm 1, it is not necessary to estimate μ_1 and μ_2 with arbitrary accuracy.

What simplifies in the stochastic model? There are fixed means to estimate, at least one fixed optimal arm in expectation, and independent observations from each arm's unchanged distribution. Concentration inequalities can therefore turn samples into reliable comparisons between arms. The decomposition in Equation (1.4) reduces the analysis to counting suboptimal pulls. Under adversarial rewards, an arm's past average need not predict its future rewards, and there is no corresponding fixed mean gap.

Remark 1 (Two different comparators)*.* The stochastic benchmark $T\mu^\star$ is the reward of the best fixed arm *in expectation*. It differs from the expected reward of the arm that happens to be best on a realized random reward table. Indeed,

$$\mathbb{E}\left[\max_i \sum_{t=1}^{T} r_{t,i}\right] \geq \max_i \mathbb{E}\left[\sum_{t=1}^{T} r_{t,i}\right] = T\mu^\star.$$

Thus Equations (1.1) and (1.3) should not be identified by simply declaring the adversarial table random. To be precise, let $B_T \coloneqq \mathbb{E}[\widehat{\mathcal{R}}_T] - R_T$. For independent rewards in $[0,1]$, the exponential-moment bound in Lemma 1 yields

$$0 \leq B_T \leq \sqrt{\frac{T\log K}{2}}, \qquad \frac{\mathbb{E}[\widehat{\mathcal{R}}_T] - R_T}{T} \longrightarrow 0 \quad \text{as } T \to \infty.$$

Therefore, the cumulative quantities need not approach one another. However, B_T does not determine whether a policy is no regret or not, as it is always sublinear. Note that the above bound cannot be improved in general, even in cases where all arms are equal. For instance, it is tight with K independent Bernoulli(1/2) arms, while $R_T = 0$ for every policy. Choosing the best arm after seeing the realized table rewards favorable random fluctuations. From now on, our stochastic results concern R_T.

2 The failure of greedy and need for exploration

If reward averages are unknown, a natural idea is to estimate them and always choose the arm that currently looks best. The difficulty is that an arm which is not selected supplies no new observations. A mistaken estimate can prevent its own correction.

2.1 Definition of greedy and proof of failure

For each arm, let

$$\widehat{\mu}_{i,n} = \frac{1}{n}\sum_{s=1}^{n} X_{i,s}$$

denote its average after n pulls. At a time t when $N_i(t) > 0$, write $\widehat{\mu}_i(t) = \widehat{\mu}_{i,N_i(t)}$. Under arbitrary reward sequences, the same empirical average is defined directly from the rewards observed on the rounds when the arm was selected.

Definition 3 (Greedy policy)**.** Assuming $T \geq K$, first pull each arm once, in the order $1, \ldots, K$. At every subsequent round choose

$$A_t \in \arg\max_{i \in \mathcal{A}} \widehat{\mu}_i(t-1), \tag{2.1}$$

using a rule specified in advance to resolve ties, for example selecting the smallest index or choosing a maximizer uniformly at random.

Initialization makes the averages well defined. It does not ensure that those averages are accurate. After initialization, this policy has no explicit mechanism for revisiting an arm whose estimate is too pessimistic.

This problem can occur even when the reward distributions are stationary. Return to the Bernoulli instance of Example 1, with means $\mu_1 = 3/4$ and $\mu_2 = 1/4$. An unlucky initialization can make the optimal arm look worse than the suboptimal one. The question is whether greedy will ever collect the observations needed to correct that initial mistake.

Proposition 2 (Greedy can have linear stochastic regret)**.** *On the preceding Bernoulli instance, Greedy satisfies*

$$R_T \geq \frac{T-1}{32}, \qquad T \geq 2. \tag{2.2}$$

Proof. Consider the event that the initial pull of the optimal arm returns 0 and the initial pull of the suboptimal arm returns 1:

$$E = \{X_{1,1} = 0,\ X_{2,1} = 1\}, \qquad \mathbb{P}(E) = \left(1 - \frac{3}{4}\right) \cdot \frac{1}{4} = \frac{1}{16}.$$

On E, the empirical means after initialization are $\widehat{\mu}_1(2) = 0$ and $\widehat{\mu}_2(2) = 1$, so greedy selects arm 2 next. As long as arm 1 is not pulled again, its empirical mean stays at 0. However, after any number $n \geq 1$ of pulls of arm 2, it holds that

$$E \Longrightarrow \widehat{\mu}_{2,n} = \frac{1}{n}\sum_{s=1}^{n} X_{2,s} \geq \frac{1}{n} > 0.$$

Therefore, greedy selects arm 2 at every round after initialization, regardless of its later rewards. Consequently, $N_2(T) = T - 1$ on E. On other outcomes the count is nonnegative, so

$$\mathbb{E}[N_2(T)] \geq \mathbb{E}[N_2(T)\mathbf{1}\{E\}] = (T-1)\mathbb{P}(E) = \frac{T-1}{16}.$$

Arm 1 has gap zero and arm 2 has gap $\Delta_2 = 1/2$. Applying Proposition 1 gives

$$R_T = \frac{1}{2}\mathbb{E}[N_2(T)] \geq \frac{T-1}{32}. \qquad \square$$

An unlucky start can therefore cause a permanent mistake. Stationarity makes an arm's samples informative about its mean, but it does not help an algorithm that stops sampling that arm. The lower bound averages over all initial outcomes: a trapping event of fixed positive probability is enough to produce linear expected regret.

Exploration and exploitation. The failure of greedy illustrates the central challenge of bandit problems: the **exploration–exploitation tradeoff**. *Exploitation* means choosing the arm that currently appears best to earn reward now. *Exploration* means gathering information to improve future choices, even if this requires pulling an arm that currently appears worse. In the example above, sampling arm 1 again could correct the initial mistake, but greedy never gives it another chance. Exploration may cost reward in the short term, while insufficient exploration can make the learner repeat a mistake for the rest of the horizon. Continuing to test clearly inferior arms, however, also incurs unnecessary regret. A good bandit algorithm must balance the immediate cost of acquiring information against the future regret that this information can prevent.

2.2 Full information doesn't rescue greedy

The example introduced above exploits the absence of new observations from the arm that greedy abandons. Would greedy work if we could observe the rewards of *all* arms after every round? In the adversarial setting, the answer is negative.

Under full-information feedback, the greedy policy is more commonly known as *Follow-the-Leader* (FTL): select the arm with the largest cumulative reward so far. Choose arm 1 in the first round, when there are no past observations, and for $t \geq 2$ select

$$A_t \in \operatorname*{arg\,max}_{i \in \{1,\dots,K\}} \sum_{s=1}^{t-1} r_{s,i}.$$

Consider the following reward table, fixed before interaction begins:

Round t	1	2	3	4	5	6	$\cdots$
Arm 1 reward	1/2	1	0	1	0	1	$\cdots$
Arm 2 reward	1	0	1	0	1	0	$\cdots$
FTL choice	1	2	1	2	1	2	$\cdots$
FTL reward	1/2	0	0	0	0	0	$\cdots$

This reward design ensures that, at each round, the wrong arm is selected! The cumulative reward of FTL remains stuck at 1/2 forever. Meanwhile, committing to either arm guarantees at least $\frac{T}{2}$ cumulative reward. It follows that

$$\mathcal{R}_T^{\text{adv}} \geq \frac{T-1}{2}.$$

Therefore, even in full information, trading off exploration and exploitation remain essential in the adversarial model. At a high level, some stability in the policy played by the learner is necessary for good performance.

Remark 2 (Two reasons for greedy failure). In the stochastic example, greedy fails to collect the information needed to correct its estimates. In the full-information adversarial example, it observes every past reward, but the past leader is a poor guide to the next round. These are different difficulties: collecting more data helps estimate fixed means, but arbitrary future rewards need not resemble past rewards.

3 Concentration of empirical means

Suppose we collect n independent observations $X_1, \dots, X_n$ from an arm with mean μ, and form $\widehat{\mu}_n = n^{-1} \sum_{s=1}^{n} X_s$. The law of large numbers gives convergence as $n \to \infty$. For a bandit

algorithm, we need a finite-sample statement: how many observations make an error of size ε sufficiently unlikely? The exponential bounds below answer that question. All logarithms in these notes are natural.

3.1 The Chernoff bound: exponential moments control tails

For a random variable Z, its *moment-generating function* is $M_Z(\lambda) = \mathbb{E}[e^{\lambda Z}]$, wherever the expectation is finite. Positive values of λ emphasize large positive values of Z.

Proposition 3 (Markov's inequality)**.** *If $W \geq 0$ and $a > 0$, then $\mathbb{P}(W \geq a) \leq \frac{\mathbb{E}[W]}{a}$.*

Proof. The pointwise inequality $W \geq a\mathbf{1}\{W \geq a\}$ gives

$$\mathbb{E}[W] \geq a\mathbb{E}[\mathbf{1}\{W \geq a\}] = a\mathbb{P}(W \geq a).$$

Divide by $a > 0$. □

Theorem 1 (Chernoff bound)**.** *For any random variable Z and any real x,*

$$\mathbb{P}(Z \geq x) \leq \inf_{\lambda > 0} \left\{ e^{-\lambda x} \mathbb{E}[e^{\lambda Z}] \right\}. \tag{3.1}$$

Values of λ giving an infinite expectation supply no useful bound. Applying the same statement to $-Z$ gives a lower-tail bound.

Proof. Fix $\lambda > 0$. Since the exponential function is increasing,

$$\mathbb{P}(Z \geq x) = \mathbb{P}(e^{\lambda Z} \geq e^{\lambda x}) \leq e^{-\lambda x} \mathbb{E}[e^{\lambda Z}],$$

where the inequality is Proposition 3 applied to the nonnegative variable $e^{\lambda Z}$. Taking the infimum over $\lambda > 0$ gives Equation (3.1). For the lower tail, apply the same argument to $-Z$. □

The Chernoff method consists of bounding the exponential moment and selecting the most favorable value of λ. For independent variables, the exponential moment of a sum factorizes. Specifically, with $S_n = \sum_{s=1}^{n} (X_s - \mathbb{E}[X_s])$,

$$\mathbb{E}[e^{\lambda S_n}] = \mathbb{E}\left[\prod_{s=1}^{n} e^{\lambda (X_s - \mathbb{E}[X_s])} \right] = \prod_{s=1}^{n} \mathbb{E}[e^{\lambda (X_s - \mathbb{E}[X_s])}].$$

Substituting into Theorem 1, for $u > 0$ we obtain

$$\mathbb{P}(S_n \geq u) \leq \inf_{\lambda > 0} \exp\left(-\lambda u + \sum_{s=1}^{n} \log \mathbb{E}[e^{\lambda (X_s - \mathbb{E}[X_s])}] \right). \tag{3.2}$$

Different assumptions on the observations lead to different bounds on these logarithmic exponential moments. For this session, boundedness is sufficient.

3.2 From bounded observations to Hoeffding's inequality

Lemma 1 (Hoeffding's lemma)**.** *If $X \in [a, b]$ almost surely, then for every $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right). \tag{3.3}$$

Proof. If $a = b$, then X is constant and the claim is immediate. Assume $a < b$, and rescale to $U = (X - a)/(b - a) \in [0, 1]$. Write $p = \mathbb{E}[U]$ and $s = \lambda(b - a)$. Convexity of the exponential gives, for every $u \in [0, 1]$ and $s \in \mathbb{R}$,

$$e^{su} \le (1 - u)e^0 + ue^s.$$

Apply this pointwise inequality to the random variable U. Since $p = \mathbb{E}[U]$ and s is a fixed real number, taking expectations gives

$$\begin{aligned}\mathbb{E}[e^{sU}] &\le \mathbb{E}[1 - U + Ue^s] \\ &= 1 - \mathbb{E}[U] + e^s\mathbb{E}[U] \\ &= 1 - p + pe^s.\end{aligned}$$

Here the inequality is preserved by expectation, and the equalities use linearity of expectation. No assumption that U is Bernoulli is needed. Finally, $X - \mathbb{E}[X] = (b-a)(U-p)$, so centering gives

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] = \mathbb{E}[e^{s(U-p)}] = e^{-sp}\mathbb{E}[e^{sU}] \le e^{-sp}(1 - p + pe^s).$$

It remains to bound the logarithm of the right-hand side. Define

$$h(s) = \log(1 - p + pe^s) - ps, \qquad q(s) = \frac{pe^s}{1 - p + pe^s}.$$

Direct differentiation gives $h(0) = h'(0) = 0$ and

$$h'(s) = q(s) - p, \qquad h''(s) = q(s)(1 - q(s)) \le \frac{1}{4},$$

since $q(s) \in [0, 1]$. By Taylor's theorem, for some ξ between 0 and s,

$$h(s) = \frac{s^2}{2}h''(\xi) \le \frac{s^2}{8}.$$

The same argument holds for negative s; at $s = 0$ the bound is immediate. Exponentiating and using $s = \lambda(b - a)$ proves the lemma. □

Hoeffding's lemma supplies the exponential-moment bound needed by the Chernoff method. Combining Theorem 1 and Lemma 1 therefore gives the following corollary. The bound depends only on the ranges of the independent observations; they need not have identical distributions.

Corollary 1 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be independent with $X_s \in [a_s, b_s]$ almost surely and $a_s < b_s$. For $u > 0$,*

$$\mathbb{P}\left(\sum_{s=1}^{n}(X_s - \mathbb{E}[X_s]) \ge u\right) \le \exp\left(-\frac{2u^2}{\sum_{s=1}^{n}(b_s - a_s)^2}\right). \tag{3.4}$$

The same bound holds for the lower tail. In particular, if the observations lie in $[0,1]$ *and all have mean* μ*, then for* $\varepsilon > 0$*,*

$$\mathbb{P}(\widehat{\mu}_n - \mu \geq \varepsilon) \leq e^{-2n\varepsilon^2}, \tag{3.5}$$

$$\mathbb{P}(\widehat{\mu}_n - \mu \leq -\varepsilon) \leq e^{-2n\varepsilon^2}, \tag{3.6}$$

$$\mathbb{P}(|\widehat{\mu}_n - \mu| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}. \tag{3.7}$$

Proof. Let $S_n = \sum_{s=1}^n (X_s - \mathbb{E}[X_s])$ and $V = \sum_{s=1}^n (b_s - a_s)^2 > 0$. Independence and Lemma 1 give

$$\mathbb{E}[e^{\lambda S_n}] = \prod_{s=1}^n \mathbb{E}[e^{\lambda(X_s - \mathbb{E}[X_s])}] \leq \prod_{s=1}^n e^{\lambda^2 (b_s - a_s)^2/8} = e^{\lambda^2 V/8}.$$

By Theorem 1, for every $\lambda > 0$,

$$\mathbb{P}(S_n \geq u) \leq \exp\left(-\lambda u + \frac{\lambda^2 V}{8}\right).$$

The exponent is minimized at $\lambda = 4u/V$. Substitution gives

$$-\lambda u + \frac{\lambda^2 V}{8} = -\frac{4u^2}{V} + \frac{2u^2}{V} = -\frac{2u^2}{V},$$

proving Equation (3.4). Applying the same argument to $-X_s$, whose range has the same length, gives the lower-tail bound.

For $[0,1]$ observations of common mean μ, take $V = n$ and $u = n\varepsilon$ to obtain Equations (3.5) and (3.6). Finally, the event $|\widehat{\mu}_n - \mu| \geq \varepsilon$ is the union of its upper- and lower-tail events. The union bound $\mathbb{P}(B \cup C) \leq \mathbb{P}(B) + \mathbb{P}(C)$ gives Equation (3.7). □

The important feature is exponential decay in $n\varepsilon^2$. To divide the typical error scale by two, we need about four times as many observations. Increasing confidence has a logarithmic sample cost.

3.3 Confidence intervals and sample requirements

For a failure probability $\delta \in (0,1)$, define

$$b_n(\delta) = \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{3.8}$$

With probability at least $1 - \delta$, the interval $[\widehat{\mu}_n - b_n(\delta), \widehat{\mu}_n + b_n(\delta)]$ contains μ. One can intersect this interval with $[0,1]$ without losing coverage. Equivalently, the sufficient sample size for accuracy $\varepsilon > 0$ and failure probability at most δ is

$$n \geq \left\lceil \frac{\log(2/\delta)}{2\varepsilon^2} \right\rceil. \tag{3.9}$$

Indeed, substituting $b_n(\delta)$ into Equation (3.7) gives

$$\mathbb{P}(|\widehat{\mu}_n - \mu| \geq b_n(\delta)) \leq 2e^{-2nb_n(\delta)^2} = \delta.$$

This proves the confidence-interval guarantee. Requiring $b_n(\delta) \leq \varepsilon$ and solving for the integer n gives Equation (3.9).

For example, accuracy 0.1 with confidence 95% is guaranteed by $n = 185$ bounded observations. This is a sufficient distribution-free guarantee, not an exact requirement for every distribution.

If we collect a *fixed* number m of samples from each of K arms, a union bound gives the simultaneous statement

$$\mathbb{P}\left(\forall i \in \mathcal{A} : \ |\widehat{\mu}_{i,m} - \mu_i| \le \sqrt{\frac{\log(2K/\delta)}{2m}}\right) \ge 1 - \delta. \tag{3.10}$$

To verify this, put $b = \sqrt{\log(2K/\delta)/(2m)}$, the radius obtained by using failure probability δ/K for each arm. Then

$$\mathbb{P}(\exists i \in \mathcal{A} : |\widehat{\mu}_{i,m} - \mu_i| > b) \le \sum_{i=1}^{K} \mathbb{P}(|\widehat{\mu}_{i,m} - \mu_i| > b) \le \sum_{i=1}^{K} \frac{\delta}{K} = \delta.$$

The union bound does not require independence between the different arm estimates; independence of the samples within each estimate is sufficient here. Thus enough samples make all arm estimates accurate at once, even though we cannot check their errors directly because the means are unknown.

3.4 Comparing two arms

For selecting an arm, we need to control a wrong ranking rather than estimate every mean to an arbitrary precision.

Corollary 2 (Probability of reversing an arm comparison)**.** *Take m independent samples from each of two arms with means $\mu_\star > \mu_i$ and gap $\Delta_i = \mu_\star - \mu_i$. If the samples lie in $[0, 1]$ and are independent across arms, then*

$$\mathbb{P}(\widehat{\mu}_{i,m} \ge \widehat{\mu}_{\star,m}) \le e^{-m\Delta_i^2}. \tag{3.11}$$

Proof. Define

$$S = \sum_{s=1}^{m} X_{i,s} - \sum_{s=1}^{m} X_{\star,s}, \qquad \mathbb{E}[S] = -m\Delta_i.$$

This is a sum of $2m$ independent variables: the m observations $X_{i,s} \in [0, 1]$ and the m signed observations $-X_{\star,s} \in [-1, 0]$. Each range has length one, so the sum of squared range lengths is $2m$. By Corollary 1,

$$\begin{aligned}
\mathbb{P}(\widehat{\mu}_{i,m} \ge \widehat{\mu}_{\star,m}) &= \mathbb{P}(S \ge 0) \\
&= \mathbb{P}(S - \mathbb{E}[S] \ge m\Delta_i) \\
&\le \exp\left(-\frac{2(m\Delta_i)^2}{2m}\right) = e^{-m\Delta_i^2}.
\end{aligned}$$

□

An alternative is to control the two estimation errors separately. A wrong ranking implies that at least one of

$$\widehat{\mu}_{i,m} - \mu_i \ge \Delta_i/2, \qquad \mu_\star - \widehat{\mu}_{\star,m} \ge \Delta_i/2$$

occurs. A union bound and the two one-sided Hoeffding bounds therefore give $2e^{-m\Delta_i^2/2}$, which is slightly weaker. Either route captures the same essential sample scale: $\Delta_i^{-2}\log(1/\delta)$ observations per arm make a wrong comparison unlikely. Arms with close means are harder to distinguish.

Remark 3 (Fixed sample sizes and adaptive sampling). The preceding bounds concern a fixed sample size. They do not automatically remain valid after replacing n by a random count $N_i(t)$ that depends on the observed rewards. Such uses require an additional argument, for example a union bound over possible counts. Our next algorithm uses a deterministic exploration budget, so this issue does not arise in its arm comparisons.

4 Explore then commit

Greedy can stop collecting information too early. The simplest repair is to reserve a fixed initial period for exploration, and exploit only after every arm has received a prescribed number of samples. This is *explore then commit* (ETC).

Throughout this section we assume independent stochastic rewards in $[0, 1]$, $T \geq K$, and an integer exploration budget $1 \leq m \leq \lfloor T/K \rfloor$. The parameter m counts samples *per arm*; the exploration phase therefore lasts Km rounds.

4.1 The algorithm

Algorithm 4.1 Explore then commit with m samples per arm

Input: Arms $\mathcal{A} = \{1, \dots, K\}$, horizon T, budget $1 \leq m \leq \lfloor T/K \rfloor$

for $s = 1, \dots, m$ **do**
  for $i = 1, \dots, K$ **do**
    Pull arm i and record its reward $X_{i,s}$.
  end for
end for
Compute $\widehat{\mu}_{i,m} = m^{-1} \sum_{s=1}^{m} X_{i,s}$ for every arm.
Select $\widehat{a} \in \arg\max_i \widehat{\mu}_{i,m}$ using a fixed tie-breaking rule.
for $t = Km + 1, \dots, T$ **do**
  Pull arm $\widehat{a}$.
end for

After exploration, the selected arm is fixed. Additional rewards during commitment do not change the decision. If $Km = T$, there is no commitment phase. For $T < K$, one can simply explore for the available rounds; there is not enough time to sample every arm even once.

4.2 Exploration cost and commitment cost

The regret decomposition separates the two phases exactly:

$$R_T = \underbrace{m \sum_{i=1}^{K} \Delta_i}_{\text{exploration cost}} + \underbrace{(T - Km) \sum_{i:\Delta_i > 0} \Delta_i \mathbb{P}(\widehat{a} = i)}_{\text{commitment cost}}. \tag{4.1}$$

Each arm is pulled exactly m times during exploration. During commitment, the entire remaining horizon is spent on one selected arm. An error at that point can therefore be expensive.

If a suboptimal arm is selected, its empirical mean must be at least that of the fixed optimal arm $a^\star$. The comparison guarantee in Corollary 2 controls the probability of this event.

Theorem 2 (Regret bound for ETC with bounded rewards)**.** *For any stochastic bandit with independent rewards in* $[0,1]$ *and any integer* $1 \le m \le \lfloor T/K \rfloor$*, ETC satisfies*

$$R_T \le m \sum_{i=1}^{K} \Delta_i + (T - Km) \sum_{i:\Delta_i > 0} \Delta_i e^{-m\Delta_i^2}. \tag{4.2}$$

In addition, the probability of committing to any suboptimal arm is at most

$$\mathbb{P}(\Delta_{\widehat{a}} > 0) \le \min\left\{1, \sum_{i:\Delta_i>0} e^{-m\Delta_i^2}\right\}. \tag{4.3}$$

Proof. Fix an optimal arm $a^\star$, so $\mu_{a^\star} = \mu^\star$. The idea is to count the cost of exploration exactly, then bound the probability that the arm selected for commitment is suboptimal.

Step 1: count the pulls of each arm. During exploration, every arm is pulled exactly m times. During the remaining $T - Km$ rounds, only $\widehat{a}$ is pulled. Consequently, for every arm i,

$$N_i(T) = m + (T - Km)\mathbf{1}\{\widehat{a} = i\}.$$

The selected arm is random, but the length of each phase is deterministic. Taking expectations therefore yields

$$\mathbb{E}[N_i(T)] = m + (T - Km)\mathbb{P}(\widehat{a} = i).$$

Substitute this expression into the regret decomposition of Proposition 1:

$$\begin{aligned} R_T &= \sum_{i=1}^{K} \Delta_i \mathbb{E}[N_i(T)] \\ &= m \sum_{i=1}^{K} \Delta_i + (T - Km) \sum_{i:\Delta_i>0} \Delta_i \mathbb{P}(\widehat{a} = i). \end{aligned}$$

This proves the exact identity in Equation (4.1). The first term is the exploration cost. In the second term, choosing arm i incurs gap Δ_i for each of the $T - Km$ commitment rounds. Optimal arms have zero gap, so they contribute nothing to that sum.

Step 2: relate a wrong commitment to a wrong comparison. Fix a suboptimal arm i, with $\Delta_i > 0$. By the definition of $\widehat{a}$, selecting i requires its empirical mean to be at least as large as that of every other arm. In particular,

$$\{\widehat{a} = i\} \subseteq \{\widehat{\mu}_{i,m} \ge \widehat{\mu}_{a^\star,m}\}.$$

Only this implication is needed: beating $a^\star$ is necessary for i to be selected, even though it need not be sufficient. The non-strict inequality also covers empirical ties, regardless of the tie-breaking rule. Rewriting the comparison around the true means gives

$$\widehat{\mu}_{i,m} \ge \widehat{\mu}_{a^\star,m} \quad \Longleftrightarrow \quad (\widehat{\mu}_{i,m} - \mu_i) - (\widehat{\mu}_{a^\star,m} - \mu^\star) \ge \Delta_i.$$

Thus the combined estimation errors must overcome the true gap.

Each empirical mean uses exactly m independent observations, and the samples of the two arms are independent. We can therefore apply Corollary 2 to obtain

$$\mathbb{P}(\widehat{a} = i) \le \mathbb{P}(\widehat{\mu}_{i,m} \ge \widehat{\mu}_{a^\star,m}) \le e^{-m\Delta_i^2}. \tag{4.4}$$

This bound concerns the exploration samples before conditioning on which arm is selected. The sample size m is fixed, which is why the concentration result applies directly.

Step 3: insert the error probabilities into the regret. All coefficients in the commitment cost are nonnegative. Replacing each selection probability in Equation (4.1) by its bound in Equation (4.4) gives

$$R_T \le m \sum_{i=1}^{K} \Delta_i + (T - Km) \sum_{i:\Delta_i>0} \Delta_i e^{-m\Delta_i^2},$$

which is Equation (4.2). Notice that we retain the individual gaps: committing to a nearly optimal arm is less costly than committing to a much worse arm.

Step 4: bound the probability of any wrong commitment. Exactly one arm is selected, so the events $\{\widehat{a} = i\}$ are disjoint. Hence

$$\mathbb{P}(\Delta_{\widehat{a}} > 0) = \sum_{i:\Delta_i>0} \mathbb{P}(\widehat{a} = i) \le \sum_{i:\Delta_i>0} e^{-m\Delta_i^2}.$$

A probability is also at most one, giving Equation (4.3). The argument allows several optimal arms: we compare with one fixed optimal arm, and only arms with a positive gap count as errors. If $Km = T$, the commitment cost vanishes and the exploration term is already exact. □

The first term increases linearly with m: exploration spends more time on suboptimal arms. The second contains an error probability that decreases exponentially with m. Too little exploration risks a costly wrong commitment; too much exploration wastes rounds after the ranking is already reliable. An optimal arm has zero gap and contributes to neither sum.

Remark 4 (Comparison with the textbook's constants). Chapter 6 of Lattimore and Szepesvári [3] treats rewards with a more general exponential-moment assumption and normalizes them to be 1-subgaussian. Its corresponding exponential term is $e^{-m\Delta_i^2/4}$. Our stronger boundedness assumption, with range $[0, 1]$, yields the constants in Equation (4.2). The exploration–commitment tradeoff is the same.

4.3 Two arms: the role of the exploration budget

For two arms with a unique optimal arm, write $\Delta > 0$ for the other arm's gap. The bound simplifies to

$$R_T \le m\Delta + (T - 2m)\Delta e^{-m\Delta^2} \le m\Delta + T\Delta e^{-m\Delta^2}. \tag{4.5}$$

Indeed, the optimal arm has gap zero, so the sums in Theorem 2 contain only the other arm's contribution. Exploration pulls that suboptimal arm m times, and commitment lasts $T - 2m$ rounds. The final inequality uses $T - 2m \le T$ and the nonnegativity of the remaining factors.

The horizon matters: an error probability that seems small for one decision can still be costly if the decision is repeated for thousands of rounds.

Example 2 (Comparing exploration budgets)**.** Let $T = 10\,000$ and take arm means 0.6 and 0.4, so $\Delta = 0.2$. The following values use the first bound in Equation (4.5):

m per arm	Exploration cost	Commitment bound	Total bound
50	10.00	267.96	277.96
250	50.00	0.09	50.09
1000	200.00	< 0.01	200.00

A very small budget leaves a large commitment-error term; a very large budget pays too much for exploration. These are upper bounds, not exact regrets or simulation results. The displayed numbers are rounded.

If the gap were known. Suppose, hypothetically, that Δ were available when choosing m. The choice

$$m = \left\lceil \frac{\log T}{\Delta^2} \right\rceil \tag{4.6}$$

makes the comparison-error bound at most $1/T$. Whenever $2m \le T$, it gives

$$R_T \le \frac{\log T}{\Delta} + 2\Delta. \tag{4.7}$$

Derivation of the oracle guarantee. The ceiling in Equation (4.6) gives both

$$m \ge \frac{\log T}{\Delta^2}, \qquad m \le \frac{\log T}{\Delta^2} + 1.$$

The first inequality controls the commitment error:

$$e^{-m\Delta^2} \le e^{-\log T} = \frac{1}{T}, \qquad T\Delta e^{-m\Delta^2} \le \Delta.$$

The second controls the exploration cost: $m\Delta \le (\log T)/\Delta + \Delta$. Combining these two bounds in Equation (4.5) yields

$$R_T \le \frac{\log T}{\Delta} + \Delta + \Delta = \frac{\log T}{\Delta} + 2\Delta. \qquad \square$$

Thus, for a fixed positive gap, a suitable budget gives logarithmic regret. If this budget does not fit in the horizon, the general feasible-budget bound in Theorem 2 still applies, as does the trivial bound $R_T \le T\Delta$.

The caveat is substantial: Δ depends on the unknown means. This is an *oracle tuning rule*, not an implementable rule for an ordinary unknown instance. It describes what the analysis suggests when the difficulty of the comparison is known in advance.

4.4 A budget that does not require knowing the gap

For two bounded arms, $\Delta \le 1$. A gap-independent consequence of Equation (4.5) is

$$R_T \le m + \frac{T}{\sqrt{2em}}, \tag{4.8}$$

To obtain this bound without knowing Δ, first use $m\Delta \le m$. For the second term in Equation (4.5), consider $f_m(x) = xe^{-mx^2}$ for $x \ge 0$. Its derivative is

$$f_m'(x) = e^{-mx^2}(1 - 2mx^2).$$

It is positive before $x = 1/\sqrt{2m}$ and negative afterwards. Therefore

$$\Delta e^{-m\Delta^2} \le \sup_{x \ge 0} f_m(x) = f_m\left(\frac{1}{\sqrt{2m}}\right) = \frac{1}{\sqrt{2em}}.$$

Substituting proves Equation (4.8). The expression remains valid when $\Delta = 0$, in which case the actual regret is zero.

Corollary 3 (A gap-independent ETC guarantee)**.** *With two arms and $T \ge 2$, choose*

$$m = \min\left\{ \left\lfloor \frac{T}{2} \right\rfloor, \left\lceil T^{2/3} \right\rceil \right\}. \tag{4.9}$$

This feasible budget depends only on the horizon and gives $R_T = O(T^{2/3})$, uniformly over stochastic two-arm instances with rewards in $[0, 1]$.

Proof. Set $M = \lceil T^{2/3} \rceil$. First suppose this budget fits in the horizon, so $2M \le T$ and the algorithm uses $m = M$. Since $T^{2/3} \le m \le T^{2/3} + 1$, Equation (4.8) gives

$$R_T \le T^{2/3} + 1 + \frac{T}{\sqrt{2\mathrm{e}}\, T^{1/3}} = \left(1 + \frac{1}{\sqrt{2\mathrm{e}}}\right) T^{2/3} + 1.$$

If $2M > T$, the budget is capped at $m = \lfloor T/2 \rfloor$. Exploration then leaves at most one commitment round: $T - 2m \in \{0, 1\}$. Using $\Delta \le 1$ in Equation (4.5), we obtain

$$R_T \le m\Delta + (T - 2m)\Delta e^{-m\Delta^2} \le m + 1 \le M + 1 \le T^{2/3} + 2.$$

In either case,

$$R_T \le \left(1 + \frac{1}{\sqrt{2\mathrm{e}}}\right) T^{2/3} + 2.$$

The constant does not depend on the gap or the reward distributions, proving the uniform $O(T^{2/3})$ guarantee. □

The $T^{2/3}$ scale balances an exploration cost of order m with a worst-case commitment cost of order $T/\sqrt{m}$. It is sublinear, so this simple policy learns well enough to achieve vanishing average regret in the stochastic setting. The guarantee differs from the logarithmic bound above: it holds without supplying the unknown gap to the algorithm.

What ETC leaves unresolved. A predetermined budget spends the same number of samples on clearly inferior arms and on nearly tied arms. It also makes an irreversible commitment after one exploration phase. If the estimates are already decisive, why keep exploring? If they are still uncertain, why stop? These questions motivate algorithms that allocate exploration using the observations, including confidence-based stopping rules and upper confidence bounds.

Reading. For more background, see Chapters 1 and 4 of Lattimore and Szepesvári [3] for the bandit framework and regret, Chapter 5 for concentration inequalities, and Chapter 6 for explore then commit. The preceding sections include proofs of the regret decomposition, greedy's linear-regret counterexample, the concentration bounds, and the ETC guarantees.

References

[1] Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://papers.nips.cc/paper/2011/hash/e53a0a2978c28872a4505bdb51db06dc-Abstract.html.

[2] Romain Gautron, Dorian Baudry, Myriam Adam, Gatien N. Falconnier, Gerrit Hoogenboom, Brian King, and Marc Corbeels. A new adaptive identification strategy of best crop management with farmers. *Field Crops Research*, 307:109249, 2024. doi: 10.1016/j.fcr.2024.109249.

[3] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[4] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670. ACM, 2010. doi: 10.1145/1772690.1772758. URL https://arxiv.org/abs/1003.0146.

[5] Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30(2):199–215, 2015. doi: 10.1214/14-STS504. URL https://arxiv.org/abs/1507.08025.