# Chapter 3

# Lower Bounds

[author: chapter revised September 2017, seems pretty polished. Would be nice to add another exercise or two.]

This chapter is about *negative* results: about what bandit algorithms *cannot* do. We are interested in lower bounds on regret which apply to all bandit algorithms. In other words, we want to prove that no bandit algorithm can achieve regret better than this lower bound. We prove the $\Omega(\sqrt{KT})$ lower bound, which takes most of this chapter, and state an instance-dependent $\Omega(\log T)$ lower bound without a proof. These lower bounds give us a sense of what are the best possible *upper* bounds that we can hope to prove. The $\Omega(\sqrt{KT})$ lower bound is stated as follows:

**Theorem 3.1.** *Fix time horizon $T$ and the number of arms $K$. For any bandit algorithm, there exists a problem instance such that $\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT})$.*

This lower bound is "worst-case" (over problem instances). In particular, it leaves open the possibility that an algorithm has low regret for most problem instances. There are two standard ways of proving such lower bounds:

(i) define a family $\mathcal{F}$ of problem instances and prove that any algorithm has high regret on some instance in $\mathcal{F}$.

(ii) define a distribution over $\mathcal{F}$ and prove that any algorithm has high regret in expectation over this distribution.

*Remark* 3.2. Note that (ii) implies (i), is because if regret is high in expectation over problem instances, then there exists at least one problem instance with high regret. Also, (i) implies (ii) if $|\mathcal{F}|$ is a constant: indeed, if we have high regret $H$ for some problem instance in $\mathcal{F}$, then in expectation over a uniform distribution over $\mathcal{F}$ regret is least $H/|\mathcal{F}|$. However, this argument breaks if $|\mathcal{F}|$ is large. Yet, a stronger version of (i) which says that regret is high for a *constant fraction* of the instances in $\mathcal{F}$ implies (ii) (with uniform distribution) regardless of whether $|\mathcal{F}|$ is large.

On a very high level, our proof proceeds as follows. We consider 0-1 rewards and the following family of problem instances, with parameter $\epsilon > 0$ to be adjusted in the analysis:

$$\mathcal{I}_j = \begin{cases} \mu_i = 1/2 & \text{for each arm } i \neq j, \\ \mu_i = (1+\epsilon)/2 & \text{for arm } i = j \end{cases} \qquad \text{for each } j = 1, 2\,,\ \ldots\ , K. \tag{3.1}$$

(Recall that $K$ is the number of arms.) Recall from the previous chapter that sampling each arm $\tilde{O}(1/\epsilon^2)$ times suffices for our upper bounds on regret.[1] Now we prove that sampling each arm $\Omega(1/\epsilon^2)$ times is *necessary* to determine whether this arm is good or bad. This leads to regret $\Omega(K/\epsilon)$. We complete the proof by plugging in $\epsilon = \Theta(\sqrt{K/T})$.

The technical details are quite subtle. We present them in several relatively gentle steps.

## 3.1 Background on KL-divergence

Our proof relies on *KL-divergence*, an important tool from Information Theory. This section provides a simplified introduction to KL-divergence, which is sufficient for our purposes.

[1] It immediately follows from Equation (2.7) in Chapter 2.

Consider a finite sample space $\Omega$, and let $p, q$ be two probability distributions defined on $\Omega$. Then, the Kullback-Leibler divergence or *KL-divergence* is defined as:

$$\mathtt{KL}(p,q) = \sum_{x\in\Omega} p(x) \ln \frac{p(x)}{q(x)} = \mathbb{E}_p \left[ \ln \frac{p(x)}{q(x)} \right].$$

This is a notion of distance between two distributions, with the properties that it is non-negative, 0 iff $p = q$, and small if the distributions $p$ and $q$ are close to one another. However, it is not strictly a distance function since it is not symmetric and does not satisfy the triangle inequality.

KL-divergence is a mathematical construct with amazingly useful properties, see Theorem 3.5 below. As far as this chapter is concerned, the precise definition does not matter as long as these properties are satisfied; in other words, any other construct with the same properties would do just as well for us. Yet, there are deep reasons as to why KL-divergence is defined in this specific way, which are are beyond the scope of this book. This material is usually covered in introductory courses on information theory.

*Remark* 3.3. While we are not going to explain why KL-divergence should be defined this way, let us see some intuition why this definition makes sense. Suppose we have data points $x_1, \ldots, x_n \in \Omega$, drawn independently from some fixed, but unknown distribution $p^*$. Further, suppose we know that this distribution is either $p$ or $q$, and we wish to use the data to estimate which one is more likely. One standard way to quantify whether distribution $p$ is more likely than $q$ is the *log-likelihood ratio*,

$$\Lambda_n := \sum_{i=1}^{n} \frac{\log p(x_i)}{\log q(x_i)}.$$

KL-divergence is the expectation of this quantity, provided that the true distribution is $p$, and also the limit as $n \to \infty$:

$$\lim_{n\to\infty} \Lambda_n = \mathbb{E}[\Lambda_n] = \mathtt{KL}(p,q) \quad \text{if } p^* = p.$$

*Remark* 3.4. The definition of KL-divergence, as well as the properties discussed below, extend to infinite sample spaces. However, KL-divergence for finite sample spaces suffices for this class, and is much easier to work with.

We present several basic properties of KL-divergence that will be needed for the rest of this chapter.[2] Throughout, let $\mathtt{RC}_\epsilon$, $\epsilon \geq 0$, denote a biased random coin with bias $\frac{\epsilon}{2}$, *i.e.,* a distribution over $\{0,1\}$ with expectation $(1+\epsilon)/2$.

**Theorem 3.5.** *KL-divergence satisfies the following properties:*

*(a)* **Gibbs' Inequality***:* $\mathtt{KL}(p,q) \geq 0, \forall p, q$. *Further,* $\mathtt{KL}(p,q) = 0$ *iff* $p = q$.

*(b)* **Chain rule***: Let the sample space* $\Omega$ *be composed as* $\Omega = \Omega_1 \times \Omega_1 \times \cdots \times \Omega_n$. *Further, let p and q be two distributions defined on* $\Omega$ *as* $p = p_1 \times p_2 \times \cdots \times p_n$ *and* $q = q_1 \times q_2 \times \cdots \times q_n$, *such that* $\forall j = 1, \ldots, n$, $p_j$ *and* $q_j$ *are distributions defined on* $\Omega_j$. *Then we have the property:* $\mathtt{KL}(p,q) = \sum_{j=1}^{n} \mathtt{KL}(p_j, q_j)$.

*(c)* **Pinsker's inequality***: for any event* $A \subset \Omega$ *we have* $2\left(p(A) - q(A)\right)^2 \leq \mathtt{KL}(p,q)$.

*(d)* **Random coins***:* $\mathtt{KL}(\mathtt{RC}_\epsilon, \mathtt{RC}_0) \leq 2\epsilon^2$, *and* $\mathtt{KL}(\mathtt{RC}_0, \mathtt{RC}_\epsilon) \leq \epsilon^2$ *for all* $\epsilon \in (0, \frac{1}{2})$.

A typical usage of these properties is as follows. Consider the setting from part (b), where $p_j = \mathtt{RC}_\epsilon$ is a biased random coin, and $q_j = \mathtt{RC}_0$ is a fair random coin for each $j$. Suppose we are interested in some event $A \subset \Omega$, and we wish to prove that $p(A)$ is not too far from $q(A)$ when $\epsilon$ is small enough. Then:

$$\begin{aligned}
2(p(A) - q(A))^2 &\leq \mathtt{KL}(p,q) && \textit{(by Pinsker's inequality)} \\
&= \sum_{j=1}^{n} \mathtt{KL}(p_j, q_j) && \textit{(by Chain Rule)} \\
&\leq n \cdot \mathtt{KL}(\mathtt{RC}_\epsilon, \mathtt{RC}_0) && \textit{(by definition of } p_j, q_j\textit{)} \\
&\leq 2n\epsilon^2. && \textit{(by part (d))}
\end{aligned}$$

[2] We present weaker versions of Chain Rule and Pinsker's inequality which suffice for our purposes.

It follows that $|p(A) - q(A)| \leq \epsilon\sqrt{n}$. In particular, $|p(A) - q(A)| < \frac{1}{2}$ whenever $\epsilon < \frac{1}{2\sqrt{n}}$.

We have proved the following:

**Lemma 3.6.** *Consider sample space $\Omega = \{0,1\}^n$ and two distributions on $\Omega$, $p = \mathtt{RC}_\epsilon^n$ and $q = \mathtt{RC}_0^n$, for some $\epsilon > 0$.*[3] *Then $|p(A) - q(A)| \leq \epsilon\sqrt{n}$ for any event $A \subset \Omega$.*

*Remark* 3.7. The asymmetry in the definition of KL-divergence does not matter for the argument above, in the sense that we could have written $\mathtt{KL}(q,p)$ instead of $\mathtt{KL}(p,q)$. Likewise, it does not matter throughout this chapter.

The proofs of the properties in Theorem 3.5 are not essential for understanding the rest of this chapter, and can be skipped. However, they are fairly simple, and we include them below for the sake of completeness.

*Proof of Theorem 3.5(a).* Let us define: $f(y) = y\ln(y)$. $f$ is a convex function under the domain $y > 0$. Now, from the definition of the KL divergence we get:

$$\begin{aligned}
\mathtt{KL}(p,q) &= \sum_{x\in\Omega} q(x)\,\frac{p(x)}{q(x)}\ln\frac{p(x)}{q(x)} = \sum_{x\in\Omega} q(x) f\left(\frac{p(x)}{q(x)}\right) \\
&\geq f\left(\sum_{x\in\Omega} q(x)\frac{p(x)}{q(x)}\right) && \textit{(by Jensen's inequality (Theorem A.1))} \\
&= f\left(\sum_{x\in\Omega} p(x)\right) = f(1) = 0,
\end{aligned}$$

By Jensen's inequality, since $f$ is not a linear function, equality holds (*i.e.,* $\mathtt{KL}(p,q) = 0$) if and only if $p = q$. □

*Proof of Theorem 3.5(b).* Let $x = (x_1, x_2, \ldots, x_n) \in \Omega$ st $x_i \in \Omega_i, \forall i = 1,\ldots,n$. Let $h_i(x_i) = \ln\frac{p_i(x_i)}{q_i(x_i)}$. Then:

$$\begin{aligned}
\mathtt{KL}(p,q) &= \sum_{x\in\Omega} p(x)\ln\frac{p(x)}{q(x)} \\
&= \sum_{i=1}^n \sum_{x\in\Omega} p(x) h_i(x_i) && \left[\text{since } \ln\frac{p(x)}{q(x)} = \sum_{i=1}^n h_i(x_i)\right] \\
&= \sum_{i=1}^n \sum_{x_i^\star\in\Omega_i} h_i(x_i^\star) \sum_{\substack{x\in\Omega,\\ x_i=x_i^\star}} p(x) \\
&= \sum_{i=1}^n \sum_{x_i\in\Omega_i} p_i(x_i) h_i(x_i) && \left[\text{since } \sum_{x\in\Omega,\ x_i=x_i^\star} p(x) = p_i(x_i^\star)\right] \\
&= \sum_{i=1}^n \mathtt{KL}(p_i,q_i).
\end{aligned}$$

□

*Proof of Theorem 3.5(c).* To prove this property, we first claim the following:

**Claim 3.8.** *For each event $A \subset \Omega$,*

$$\sum_{x\in A} p(x)\ln\frac{p(x)}{q(x)} \geq p(A)\ln\frac{p(A)}{q(A)}.$$

*Proof.* Let us define the following:

$$p_A(x) = \frac{p(x)}{p(A)} \quad \text{and} \quad q_A(x) = \frac{q(x)}{q(A)} \quad \forall x \in A.$$

[3] In other words, $p$ is $n$ independent tosses of a biased coin $\mathtt{RC}_\epsilon$, and $q$ is $n$ independent tosses of a fair coin $\mathtt{RC}_0$.

Then the claim can be proved as follows:

$$\begin{aligned}
\sum_{x\in A} p(x)\ln\frac{p(x)}{q(x)} &= p(A)\sum_{x\in A} p_A(x)\ln\frac{p(A)p_A(x)}{q(A)q_A(x)} \\
&= p(A)\left(\sum_{x\in A} p_A(x)\ln\frac{p_A(x)}{q_A(x)}\right) + p(A)\ln\frac{p(A)}{q(A)}\sum_{x\in A} p_A(x) \\
&\geq p(A)\ln\frac{p(A)}{q(A)}. \qquad \left[\text{since } \sum_{x\in A} p_A(x)\ln\frac{p_A(x)}{q_A(x)} = \mathtt{KL}(p_A,q_A)\geq 0\right] \qquad \square
\end{aligned}$$

Fix $A\subset\Omega$. Using Claim 3.8 we have the following:

$$\begin{aligned}
\sum_{x\in A} p(x)\ln\frac{p(x)}{q(x)} &\geq p(A)\ln\frac{p(A)}{q(A)}, \\
\sum_{x\notin A} p(x)\ln\frac{p(x)}{q(x)} &\geq p(\bar{A})\ln\frac{p(\bar{A})}{q(\bar{A})},
\end{aligned}$$

where $\bar{A}$ denotes the complement of $A$. Now, let $a=p(A)$ and $b=q(A)$. Further, assume $a<b$. Then, we have:

$$\begin{aligned}
\mathtt{KL}(p,q) &= a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b} \\
&= \int_a^b \left(-\frac{a}{x}+\frac{1-a}{1-x}\right)dx \\
&= \int_a^b \frac{x-a}{x(1-x)}dx \\
&\geq \int_a^b 4(x-a)dx = 2(b-a)^2. \qquad \textit{(since } x(1-x)\leq\tfrac{1}{4}\textit{)} \qquad \square
\end{aligned}$$

*Proof of Theorem 3.5(d).*

$$\begin{aligned}
\mathtt{KL}(\mathtt{RC}_0,\mathtt{RC}_\epsilon) &= \tfrac{1}{2}\,\ln(\tfrac{1}{1+\epsilon}) + \tfrac{1}{2}\,\ln(\tfrac{1}{1-\epsilon}) \\
&= -\tfrac{1}{2}\,\ln(1-\epsilon^2) \\
&\leq -\tfrac{1}{2}\,(-2\epsilon^2) \qquad \textit{(as } \log(1-\epsilon^2)\geq -2\epsilon^2 \textit{ whenever } \epsilon^2\leq\tfrac{1}{2}\textit{)} \\
&= \epsilon^2. \\
\mathtt{KL}(\mathtt{RC}_\epsilon,\mathtt{RC}_0) &= \tfrac{1+\epsilon}{2}\ln(1+\epsilon) + \tfrac{1-\epsilon}{2}\ln(1-\epsilon) \\
&= \tfrac{1}{2}\left(\ln(1+\epsilon)+\ln(1-\epsilon)\right) + \tfrac{\epsilon}{2}\left(\ln(1+\epsilon)-\ln(1-\epsilon)\right) \\
&= \tfrac{1}{2}\ln(1-\epsilon^2) + \tfrac{\epsilon}{2}\ln\tfrac{1+\epsilon}{1-\epsilon}.
\end{aligned}$$

Now, $\ln(1-\epsilon^2)<0$ and we can write $\ln\frac{1+\epsilon}{1-\epsilon} = \ln\left(1+\frac{2\epsilon}{1-\epsilon}\right) \leq \frac{2\epsilon}{1-\epsilon}$. Thus, we get:

$$\mathtt{KL}(\mathtt{RC}_\epsilon,\mathtt{RC}_0) < \tfrac{\epsilon}{2}\cdot\tfrac{2\epsilon}{1-\epsilon} = \tfrac{\epsilon^2}{1-\epsilon} \leq 2\epsilon^2. \qquad \square$$

## 3.2 A simple example: flipping one coin

We start with a simple application of the KL-divergence technique, which is also interesting as a standalone result. Consider a biased random coin (*i.e.,* a distribution on $\{0,1\}$) with an unknown mean $\mu\in[0,1]$. Assume that $\mu\in\{\mu_1,\mu_2\}$ for two known values $\mu_1>\mu_2$. The coin is flipped $T$ times. The goal is to identify if $\mu=\mu_1$ or $\mu=\mu_2$ with low probability of error.

Let us make our goal a little more precise. Define $\Omega := \{0,1\}^T$ to be the sample space for the outcomes of $T$ coin tosses. Let us say that we need a decision rule $\texttt{Rule} : \Omega \to \{\texttt{High}, \texttt{Low}\}$ with the following two properties:

$$\Pr[\texttt{Rule}(\mathit{observations}) = \texttt{High} \mid \mu = \mu_1] \geq 0.99, \tag{3.2}$$

$$\Pr[\texttt{Rule}(\mathit{observations}) = \texttt{Low} \mid \mu = \mu_2] \geq 0.99. \tag{3.3}$$

We ask, how large should $T$ be for for such a decision rule to exist? We know that $T \sim (\mu_1 - \mu_2)^{-2}$ is sufficient. What we prove is that it is also necessary. We will focus on the special case when both $\mu_1$ and $\mu_2$ are close to $\frac{1}{2}$.

**Lemma 3.9.** *Let $\mu_1 = \frac{1+\epsilon}{2}$ and $\mu_2 = \frac{1}{2}$. For any decision rule to satisfy properties (3.2) and (3.3) we need $T > \frac{1}{4\,\epsilon^2}$.*

*Proof.* Fix a decision rule which satisfies (3.2) and (3.3), and let $A_0 \subset \Omega$ be the event this rule returns $\texttt{High}$. Then

$$\Pr[A_0 \mid \mu = \mu_1] - \Pr[A_0 \mid \mu = \mu_2] \geq 0.98. \tag{3.4}$$

Let $P_i(A) = \Pr[A \mid \mu = \mu_i]$, for each event $A \subset \Omega$ and each $i \in \{1,2\}$. Then $P_i = P_{i,1} \times \ldots \times P_{i,T}$, where $P_{i,t}$ is the distribution of the $t^{th}$ coin toss if $\mu = \mu_i$. Thus, the basic KL-divergence argument summarized in Lemma 3.6 applies to distributions $P_1$ and $P_2$. It follows that $|P_1(A) - P_2(A)| \leq \epsilon\sqrt{T}$. Plugging in $A = A_0$ and $T \leq \frac{1}{4\epsilon^2}$, we obtain $|P_1(A_0) - P_2(A_0)| < \frac{1}{2}$, contradicting (3.4). □

Remarkably, the proof does not really consider what a given decision rule is doing, and applies to all rules at once!

## 3.3 Flipping several coins: "bandits with prediction"

Let us extend the previous example to flipping multiple coins. We consider a bandit problem with $K$ arms, where each arm corresponds to a biased random coin with unknown mean. More formally, the reward of each arm is drawn independently from a fixed but unknown Bernoulli distribution. After $T$ rounds, the algorithm outputs an arm $y_T$, which is the algorithm's prediction for which arm is optimal (has the highest mean reward). We call this version "bandits with predictions". We will only be concerned with the quality of prediction, rather than regret.

As a matter of notation, recall that with 0-1 rewards, a problem instance can be specified as a tuple $\mathcal{I} = (\mu(a) : \forall a \in \mathcal{A})$, where $\mu(a)$ is the mean reward of arm $a$ and $\mathcal{A}$ is the set of all arms. We will number arms from 1 to $K$.

For concreteness, let us say that a good algorithm for "bandits with predictions" should satisfy

$$\Pr[\text{prediction } y_T \text{ is correct } \mid \mathcal{I}] \geq 0.99 \tag{3.5}$$

for each problem instance $\mathcal{I}$. We will use the family (3.1) of problem instances, with parameter $\epsilon > 0$, to argue that one needs $T \geq \Omega\left(\frac{K}{\epsilon^2}\right)$ for any algorithm to "work", *i.e.,* satisfy property (3.5), on all instances in this family. This result is of independent interest, regardless of the regret bound that we've set out to prove.

In fact, we prove a stronger statement which will also be the crux in the proof of the regret bound.

**Lemma 3.10.** *Consider a "bandits with predictions" problem with $T \leq \frac{cK}{\epsilon^2}$, for a small enough absolute constant $c > 0$. Fix any deterministic algorithm for this problem. Then there exists at least $\lceil K/3 \rceil$ arms $a$ such that*

$$\Pr[y_T = a \mid \mathcal{I}_a] < \tfrac{3}{4}. \tag{3.6}$$

The proof for $K = 2$ arms is particularly simple, so we present it first. The general case is somewhat more subtle. We only present a simplified proof for $K \geq 24$, which is deferred to Section 3.4.

*Proof ($K = 2$ arms).* Let us set up the sample space which we will use in the proof. Let $(r_t(a) : \; a \in \mathcal{A}, t \in [T])$ be mutually independent 0-1 random variables such that $r_t(a)$ has expectation $\mu(a)$.[4] We refer to this tuple as the *rewards table*, where we interpret $r_t(a)$ as the reward received by the algorithm for the $t$-th time it chooses arm $a$. The

[4] We use standard shorthand $[T] = \{1, 2, \ldots, T\}$.

sample space is $\Omega = \{0,1\}^{K\times T}$, where each outcome $\omega \in \Omega$ corresponds to a particular realization of the rewards table. Each problem instance $\mathcal{I}_j$ defines distribution $P_j$ on $\Omega$ as follows:

$$P_j(A) = \Pr[A \mid \mathcal{I}_j] \quad \text{for each } A \subset \Omega.$$

Let $P_j^{a,t}$ be the distribution of $r_t(a)$ under instance $\mathcal{I}_j$, so that $P_j = \prod_{a\in\mathcal{A},\; t\in[T]} P_j^{a,t}$.

Let $A = \{\omega \subseteq \Omega : y_T = 1\}$ be the event that the algorithm predicts arm 1. For the sake of contradiction, assume that (3.6) fails for both arms. Then $P_1(A) \geq \frac{3}{4}$ and $P_2(A) < \frac{1}{4}$, so their difference is at least $\frac{1}{2}$.

To arrive at a contradiction, we use a similar KL-divergence argument as before:

$$\begin{aligned} 2(P_1(A) - P_2(A))^2 &\leq \mathtt{KL}(P_1, P_2) && \textit{(by Pinsker's inequality)} \\ &= \sum_{a=1}^{K}\sum_{t=1}^{T} \mathtt{KL}(P_1^{a,t}, P_2^{a,t}) && \textit{(by Chain Rule)} \\ &\leq 2T\cdot 2\epsilon^2 && \textit{(by Theorem 3.5(d)).} \qquad (3.7) \end{aligned}$$

The last inequality holds because for each arm $a$ and each round $t$, one of the distributions $P_1^{a,t}$ and $P_2^{a,t}$ is a fair coin $\mathtt{RC}_0$, and another is a biased coin $\mathtt{RC}_\epsilon$. Therefore,

$$P_1(A) - P_2(A) \leq 2\epsilon\sqrt{T} < \tfrac{1}{2} \quad \text{whenever } T \leq (\tfrac{1}{4\epsilon})^2. \qquad \square$$

**Corollary 3.11.** *Assume $T$ is as in Lemma 3.10. Fix any algorithm for "bandits with predictions". Choose an arm $a$ uniformly at random, and run the algorithm on instance $\mathcal{I}_a$. Then $\Pr[y_T \neq a] \geq \frac{1}{12}$, where the probability is over the choice of arm $a$ and the randomness in rewards and the algorithm.*

*Proof.* Lemma 3.10 easily implies this corollary for deterministic algorithms, which in turn implies it for randomized algorithms (because any randomized algorithm can be expressed as a distribution over deterministic algorithms). □

Finally, we use Corollary 3.11 to finish our proof of the $\sqrt{KT}$ lower bound on regret. We prove the following:

**Theorem 3.12.** *Fix time horizon $T$ and the number of arms $K$. Fix a bandit algorithm. Choose an arm $a$ uniformly at random, and run the algorithm on problem instance $\mathcal{I}_a$. Then*

$$\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT}), \qquad (3.8)$$

*where the expectation is over the choice of arm $a$ and the randomness in rewards and the algorithm.*

*Proof.* Fix $\epsilon$ in (3.1), to be adjusted later, and assume that $T \leq \frac{cK}{\epsilon^2}$, where $c$ is the constant from Lemma 3.10.

Fix round $t$. Let us interpret the algorithm as a "bandits with predictions" algorithm, where the prediction is simply $a_t$, the arm chosen in this round. We can apply Corollary 3.11, treating $t$ as the time horizon, to deduce that $\Pr[a_t \neq a] \geq \frac{1}{12}$. In words, the algorithm chooses a non-optimal arm with probability at least $\frac{1}{12}$. Recall that for each problem instances $\mathcal{I}_a$, the "badness" $\Delta(a_t) := \mu^* - \mu(a_t)$ is $\epsilon/2$ whenever a non-optimal arm is chosen. Therefore,

$$\mathbb{E}[\Delta(a_t)] = \Pr[a_t \neq a]\cdot \tfrac{\epsilon}{2} \geq \epsilon/24.$$

Summing up over all rounds, we obtain $\mathbb{E}[R(T)] = \sum_{t=1}^{T} \mathbb{E}[\Delta(a_t)] \geq \epsilon T/24$. Using $\epsilon = \sqrt{\frac{cK}{T}}$, we obtain (3.8). □

## 3.4 Proof of Lemma 3.10 for $K \geq 24$ arms

This is the crucial technical argument in the proof of our regret bound. Compared to the case of $K = 2$ arms, we need to handle a time horizon that can be larger by a factor of $O(K)$. The crucial improvement is a more delicate version of the KL-divergence argument, which improves the right-hand side of (3.7) to $O(T\epsilon^2/K)$.

For the sake of the analysis, we will consider an additional problem instance

$$\mathcal{I}_0 = \{\mu_i = \tfrac{1}{2} \text{ for all arms } i\ \},$$

which we call the "base instance". Let $\mathbb{E}_0[\cdot]$ be the expectation given this problem instance. Also, let $T_a$ be the total number of times arm $a$ is played.

We consider the algorithm's performance on problem instance $\mathcal{I}_0$, and focus on arms $j$ that are "neglected" by the algorithm, in the sense that the algorithm does not choose arm $j$ very often *and* is not likely to pick $j$ for the guess $y_T$. Formally, we observe the following:

$$\text{there are at least } \tfrac{2K}{3} \text{ arms } j \text{ such that } \mathbb{E}_0(T_j) \leq \tfrac{3T}{K}, \tag{3.9}$$

$$\text{there are at least } \tfrac{2K}{3} \text{ arms } j \text{ such that } P_0(y_T = j) \leq \tfrac{3}{K}. \tag{3.10}$$

(To prove (3.9), assume for contradiction that we have more than $\frac{K}{3}$ arms with $\mathbb{E}_0(T_j) > \frac{3T}{K}$. Then the expected total number of times these arms are played is strictly greater than $T$, which is a contradiction. (3.10) is proved similarly.) By Markov inequality,

$$\mathbb{E}_0(T_j) \leq \tfrac{3T}{K} \text{ implies that } \Pr[T_j \leq \tfrac{24T}{K}] \geq \tfrac{7}{8}.$$

Since the sets of arms in (3.9) and (3.10) must overlap on least $\frac{K}{3}$ arms, we conclude:

$$\text{there are at least } \tfrac{K}{3} \text{ arms } j \text{ such that } \Pr[T_j \leq \tfrac{24T}{K}] \geq \tfrac{7}{8} \text{ and } P_0(y_T = j) \leq \tfrac{3}{K}. \tag{3.11}$$

We will now refine our definition of the sample space. For each arm $a$, define the $t$-round sample space $\Omega_a^t = \{0,1\}^t$, where each outcome corresponds to a particular realization of the tuple $(r_s(a):\ s \in [t])$. (Recall that we interpret $r_t(a)$ as the reward received by the algorithm for the $t$-th time it chooses arm $a$.) Then the "full" sample space we considered before can be expressed as $\Omega = \prod_{a \in \mathcal{A}} \Omega_a^T$.

Fix an arm $j$ satisfying the two properties in (3.11). We will consider a "reduced" sample space in which arm $j$ is played only $m = \frac{24T}{K}$ times:

$$\Omega^* = \Omega_j^m \times \prod_{\text{arms } a \neq j} \Omega_a^T. \tag{3.12}$$

For each problem instance $\mathcal{I}_\ell$, we define distribution $P_\ell^*$ on $\Omega^*$ as follows:

$$P_\ell^*(A) = \Pr[A \mid \mathcal{I}_\ell] \quad \text{for each } A \subset \Omega^*.$$

In other words, distribution $P_\ell^*$ is a restriction of $P_\ell$ to the reduced sample space $\Omega^*$.

We apply the KL-divergence argument to distributions $P_0^*$ and $P_j^*$. For each event $A \subset \Omega^*$:

$$\begin{aligned}
2(P_0^*(A) - P_j^*(A))^2 &\leq \mathtt{KL}(P_0^*, P_j^*) && \textit{(by Pinsker's inequality)} \\
&= \sum_{\text{arms } a} \sum_{t=1}^{T} \mathtt{KL}(P_0^{a,t}, P_j^{a,t}) && \textit{(by Chain Rule)} \\
&= \sum_{\text{arms } a \neq j} \sum_{t=1}^{T} \mathtt{KL}(P_0^{a,t}, P_j^{a,t}) + \sum_{t=1}^{m} \mathtt{KL}(P_0^{j,t}, P_j^{j,t}) \\
&\leq 0 + m \cdot 2\epsilon^2 && \textit{(by Theorem 3.5(d)).}
\end{aligned}$$

The last inequality holds because each arm $a \neq j$ has identical reward distributions under problem instances $\mathcal{I}_0$ and $\mathcal{I}_j$ (namely the fair coin $\mathtt{RC}_0$), and for arm $j$ we only need to sum up over $m$ samples rather than $T$.

Therefore, assuming $T \leq \frac{cK}{\epsilon^2}$ with small enough constant $c$, we can conclude that

$$|P_0^*(A) - P_j^*(A)| \leq \epsilon\sqrt{m} < \tfrac{1}{8} \quad \text{for all events } A \subset \Omega^*. \tag{3.13}$$

To apply (3.13), we need to make sure that the event $A$ is in fact contained in $\Omega^*$, *i.e.,* whether this event holds is completely determined by the first $m$ samples of arm $j$ (and arbitrarily many samples of other arms). In particular, we cannot take $A = \{y_t = j\}$, which would be the most natural extension of the proof technique from the 2-arms case, because this event may depend on more than $m$ samples of arm $j$. Instead, we apply (3.13) twice: to events

$$A = \{y_T = j \text{ and } T_j \leq m\} \text{ and } A' = \{T_j > m\}. \tag{3.14}$$

Note that whether the algorithm samples arm $j$ more than $m$ times is completely determined by the first $m$ coin tosses!

We are ready for the final computation:

$$\begin{aligned}
P_j(A) &\leq \tfrac{1}{8} + P_0(A) && \textit{(by (3.13))} \\
&\leq \tfrac{1}{8} + P_0(y_T = j) \\
&\leq \tfrac{1}{4} && \textit{(by our choice of arm } j\textit{).} \\
P_j(A') &\leq \tfrac{1}{8} + P_0(A') && \textit{(by (3.13))} \\
&\leq \tfrac{1}{4} && \textit{(by our choice of arm } j\textit{).} \\
P_j(Y_T = j) &\leq P^*_j(Y_T = j \textit{ and } T_j \leq m) + P^*_j(T_j > m) \\
&= P_j(A) + P_j(A') \leq \tfrac{1}{4}.
\end{aligned}$$

This holds for any arm $j$ satisfying the properties in (3.11). Since there are at least $K/3$ such arms, the lemma follows.

## 3.5 Instance-dependent lower bounds (without proofs)

There is another fundamental lower bound on regret, which applies to any given problem instance and asserts $\log(T)$ regret with an instance-dependent constant. This lower bound complements the $\log(T)$ *upper* bound that we proved for algorithms UCB1 and Successive Elimination. We present and discuss this lower bound without a proof.

As before, we focus on 0-1 rewards. For a particular problem instance, we view $\mathbb{E}[R(t)]$ a function of $t$, and we are interested in how this function grows with $t$. We start with a simpler and slightly weaker version of the lower bound:

**Theorem 3.13.** *No algorithm can achieve regret $\mathbb{E}[R(t)] = o(c_{\mathcal{I}} \;\; \log t)$ for all problem instances $\mathcal{I}$, where the "constant" $c_{\mathcal{I}}$ can depend on the problem instance but not on the time $t$.*

This version guarantees at least one problem instance on which a given algorithm has "high" regret. We would like to have a stronger lower bound which guarantees "high" regret for each problem instance. However, such lower bound is impossible because of a trivial counterexample: an algorithm which always plays arm 1, as dumb as it is, nevertheless has 0 regret on any problem instance for which arm 1 is optimal. Therefore, the desired lower bound needs to assume that the algorithm is at least somewhat good, so as to rule out such counterexamples.

**Theorem 3.14.** *Fix $K$, the number of arms. Consider an algorithm such that*

$$\mathbb{E}[R(t)] \leq O(C_{\mathcal{I},\alpha}\, t^{\alpha}) \quad \textit{for each problem instance } \mathcal{I} \textit{ and each } \alpha > 0. \tag{3.15}$$

*Here the "constant" $C_{\mathcal{I},\alpha}$ can depend on the problem instance $\mathcal{I}$ and the $\alpha$, but not on time $t$.*

*Fix an arbitrary problem instance $\mathcal{I}$. For this problem instance:*

$$\textit{There exists time } t_0 \textit{ such that for any } t \geq t_0 \quad \mathbb{E}[R(t)] \geq C_{\mathcal{I}} \ln(t), \tag{3.16}$$

*for some constant $C_{\mathcal{I}}$ that depends on the problem instance, but not on time $t$.*

*Remark* 3.15. For example, Assumption (3.15) is satisfied for any algorithm with $\mathbb{E}[R(t)] \leq (\log t)^{1000}$.

Let us refine Theorem 3.14 and specify how the instance-dependent constant $C_{\mathcal{I}}$ in (3.16) can be chosen. In what follows, $\Delta(a) = \mu^* - \mu(a)$ be the "badness" of arm $a$.

**Theorem 3.16.** *For each problem instance $\mathcal{I}$ and any algorithm that satisfies (3.15),*

*(a) the bound (3.16) holds with*

$$C_{\mathcal{I}} = \sum_{a:\; \Delta(a)>0} \frac{\mu^*(1-\mu^*)}{\Delta(a)}.$$

*(b) for each $\epsilon > 0$, the bound (3.16) holds with*

$$C_{\mathcal{I}} = \sum_{a:\; \Delta(a)>0} \frac{\Delta(a)}{\texttt{KL}(\mu(a),\, \mu^*)} - \epsilon.$$

*Remark* 3.17. The lower bound from part (a) is similar to the upper bound achieved by UCB1 and Successive Elimination: $R(T) \leq \sum_{a:\, \Delta(a)>0} \frac{O(\log T)}{\Delta(a)}$. In particular, we see that the upper bound is optimal up to a constant factor when $\mu^*$ is bounded away from 0 and 1, *e.g.,* when $\mu^* \in [\frac{1}{4}, \frac{3}{4}]$.

*Remark* 3.18. Part (b) is a stronger (*i.e.,* larger) lower bound which implies the more familiar form in part (a). Several algorithms in the literature are known to come arbitrarily close to this lower bound. In particular, a version of Thompson Sampling (another standard algorithm discussed in Chapter 4) achieves regret

$$R(t) \leq (1+\delta)\, C_{\mathcal{I}}\ \ln(t) + C'_{\mathcal{I}}/\epsilon^2, \quad \forall \delta > 0,$$

where $C_{\mathcal{I}}$ is from part (b) and $C'_{\mathcal{I}}$ is some other instance-dependent constant.

## 3.6 Bibliographic notes

The $\Omega(\sqrt{KT})$ lower bound on regret is from Auer et al. (2002b). KL-divergence and its properties is “textbook material” from Information Theory, *e.g.,* see Cover and Thomas (1991). The present exposition — the outline and much of the technical details — is based on Robert Kleinberg’s lecture notes from (Kleinberg, 2007, week 9).

We present a substantially simpler proof compared to (Auer et al., 2002b) and (Kleinberg, 2007, week 9) in that we avoid the general ”chain rule” for KL-divergence. Instead, we only use the special case of independent distributions (Theorem 3.5(b) in Section 3.1), which is much easier to state and to apply. The proof of Lemma 3.10 (for general $K$), which in prior work relies on the general ”chain rule”, is modified accordingly. In particular, we define the “reduced” sample space $\Omega^*$ with only a small number of samples from the “bad” arm $j$, and apply the KL-divergence argument to carefully defined events in (3.14), rather than a seemingly more natural event $A = \{y_T = j\}$.

The proof of the logarithmic lower bound from Section 3.5 is also based on a KL-divergence technique. It can be found in the original paper (Lai and Robbins, 1985), as well as in the survey (Bubeck and Cesa-Bianchi, 2012).

## 3.7 Exercises

*Exercise* 3.1 (lower bound for non-adaptive exploration). Consider an algorithm such that:

- in the first $N$ rounds (“exploration phase”) the choice of arms does not depend on the observed rewards, for some $N$ that is fixed before the algorithm starts;
- in all remaining rounds, the algorithm only uses rewards observed during the exploration phase.

Focus on the case of two arms, and prove that such algorithm must have regret $\mathbb{E}[R(T)] \geq \Omega(T^{2/3})$ in the worst case.

*Hint*: Regret is a sum of regret from exploration, and regret from exploitation. For “regret from exploration”, we can use two instances: $(\mu_1, \mu_2) = (1, 0)$ and $(\mu_1, \mu_2) = (0, 1)$, *i.e.,* one arm is very good and another arm is very bad. For “regret from exploitation” we can invoke the impossibility result for “bandits with predictions” (Corollary 3.11).

*Take-away*: Regret bound for Explore-First cannot be substantially improved. Further, allowing Explore-first to pick different arms in exploitation does not help.