

Lower Bounds on the Sample Complexity of Exploration in the Multi-Armed Bandit Problem

Shie Mannor and John N. Tsitsiklis

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology, Cambridge, MA 02139
{shie,jnt}@mit.edu

Abstract. We consider the Multi-armed bandit problem under the PAC (“probably approximately correct”) model. It was shown by Even-Dar et al. [5] that given n arms, it suffices to play the arms a total of $O((n/\varepsilon^2) \log(1/\delta))$ times to find an ε -optimal arm with probability of at least $1 - \delta$. Our contribution is a matching lower bound that holds for any sampling policy. We also generalize the lower bound to a Bayesian setting, and to the case where the statistics of the arms are known but the identities of the arms are not.

1 Introduction

The multi-armed bandit problem is a classical problem in decision theory. There is a number of alternative arms, each with a stochastic reward with initially unknown statistics. We play these arms in some order, which may depend on the sequence of rewards that have been observed so far. The objective is to find a policy for playing the arms, under which the sum of the expected rewards comes as close as possible to the ideal reward, i.e., the expected reward that would be obtained if we were to play the “best” arm at all times. One of the attractive features of the multi-armed bandit problem is that despite its simplicity, it encompasses many important decision theoretic issues, such as the tradeoff between exploration and exploitation.

The multi-armed bandit problem has been widely studied in a variety of setups. The problem was first considered in the 50’s, in the seminal work of Robbins [10], which derives strategies that asymptotically attain an average reward that converges in the limit to the reward of the best arm. The multi-armed bandit problem was later studied in discounted, Bayesian, Markovian, expected reward, and adversarial setups. (See [4] for a review of the classical results on the multi-armed bandit problem.)

Lower bounds for different variants of the multi-armed bandit have been studied by several authors. For the expected regret model, where the regret is defined as the difference between the ideal reward (if the best arm were known) and the reward of an online policy, the seminal work of Lai and Robbins [9] provides tight bounds in terms of the Kullback-Leibler divergence between the distributions of the rewards of the different arms. These bounds grow logarithmically with the

number of steps. The adversarial multi-armed bandit problem (i.e., without any probabilistic assumptions) was considered in [2, 3], where it was shown that the expected regret grows proportionally to the square root of the number of steps. Of related interest is the work of Kulkarni and Lugosi [8]. It was shown there that for any specific time t , one can choose probability distributions so that the expected regret is linear in t .

The focus of this paper is the classical multi-armed bandit problem, but rather than looking at the expected regret, we are concerned with PAC-type bounds on the number of steps needed to identify a near-optimal arm. In particular, we are interested in the expected number of steps that are required in order to identify with high probability (at least $1 - \delta$) an arm whose expected reward is within ε from the expected reward of the best arm. This naturally abstracts the case where one must eventually commit to one specific arm, and quantifies the amount of exploration necessary. This is in contrast to most of the results for the multi-armed bandit problem, where the main aim is to maximize the expected cumulative reward while both exploring and exploiting.

In [5] an algorithm, called the Median Elimination algorithm, was shown to require $O((n/\varepsilon^2) \log(1/\delta))$ arm plays to find, with probability of at least $1 - \delta$, an ε -optimal arm. A matching lower bound was also derived [5], but it only applied to the case where $\delta > 1/n$, and therefore did not capture the case where high confidence (small δ) is desired. In this paper, we derive a matching lower bound which applies for every $\delta > 0$. Let us note here that some results with a similar flavor have been provided in [7]. However, they refer to the case of Gaussian rewards and only consider the case of asymptotically vanishing δ .

Our main result can be viewed as a generalization of a $O(1/\varepsilon^2 \log(1/\delta))$ lower bound provided in [1] for the case of two bandits. The proof in [1] is based on a simple interchange argument, and relies on the fact there are only two underlying hypotheses. The technique we use is based on a likelihood ratio argument and a tight martingale bound. Let us note that the inequalities used in [3] to derive lower bounds for the expected regret in an adversarial setup can also be used to derive a lower bound for our problem, but do not appear to be tight enough to capture the $\log(1/\delta)$ dependence on δ .

Our work also provides fundamental lower bounds in the context of sequential analysis (see, e.g., [12]). In the language of [12], we provide a lower bound on the expected length of a sequential sampling algorithm under any adaptive allocation scheme. For the case of two arms, it was shown in [12] (p. 148) that if one restricts to sampling strategies that only take into account the empirical average rewards from the different arms, then the problems of inference and arm selection can be treated separately. As a consequence, and under this restriction, [12] shows that an optimal allocation cannot be much better than a uniform one. Our results are different in a number of ways. First, we consider multiple hypotheses (multiple arms). Second, we allow the allocation rule to be completely general and to depend on the whole history. Third, unlike most of the sequential analysis literature (see, e.g., [7]), we do not restrict ourselves to the limiting case where

the probability of error converges to zero. Finally, we consider finite time bounds, rather than asymptotic ones.

The paper is organized as follows. In Section 2, we set up our framework, and since we are interested in lower bounds, we restrict to the special case where each arm is a “coin,” i.e., the rewards are Bernoulli random variables, but with unknown parameters (“biases”). In Section 3, we provide a lower bound on the sample complexity. In fact, we show that the lower bound holds true even in the special case where the set of coin biases is known, but the identity of the coins is not. In Section 4, we derive similar lower bounds within a Bayesian setting, where there is a prior distribution on the set of biases of the different coins. In Section 5, we provide a lower bound that depends on the specific (though unknown) biases of the coins. Finally, Section 6 contains some brief concluding remarks.

2 Problem Definition

The exploration problem for multi-armed bandits is defined as follows. We are given n arms. Each arm i is associated with a sequence of identically distributed Bernoulli (i.e., taking values in $\{0, 1\}$) random variables X_k^i , $k = 1, 2, \dots$, with mean p_i . Here, X_k^i corresponds to the reward obtained the k th time that arm i is played. We assume that the random variables X_k^i , for $i = 1, \dots, n$, $k = 1, 2, \dots$, are independent, and we define $p = (p_1, \dots, p_n)$. Given that we restrict to the Bernoulli case, we will use in the sequel the term “coin” instead of “arm.”

A *policy* is a mapping that given a history, chooses a particular coin to be tried next, or selects a particular coin and stops. We allow a policy to use randomization when choosing a coin to sample or when making a final selection. However, we only consider policies that are guaranteed to stop with probability 1, for every possible vector p . (Otherwise, the expected number of steps would be infinite.) Given a particular policy, we let \mathbf{P}_p be the corresponding probability measure (on the natural probability space for this model). This probability space captures both the randomness in the coins (according to the vector p), as well as any additional randomization carried out by the policy. We introduce the following random variables, which are well defined, except possibly on the set of measure zero where the policy does not stop. We let T_i be the total number of times that coin i is tried, and let $T = T_1 + \dots + T_n$ be the total number of trials. We also let I be the coin which is selected when the policy decides to stop.

We say that a policy is (ε, δ) -correct if

$$\mathbf{P}_p\left(p_I > \max_i p_i - \varepsilon\right) \geq 1 - \delta,$$

for every $p \in [0, 1]^n$. It was shown in [5] that there exist constants c_1 and c_2 such that for every n , $\varepsilon > 0$, and $\delta > 0$, there exists an (ε, δ) -correct policy under which

$$\mathbf{E}_p[T] \leq c_1 \frac{n}{\varepsilon^2} \log \frac{c_2}{\delta}, \quad \forall p \in [0, 1]^n.$$

A matching lower bound was also established in [5], but only for “large” values of δ , namely, for $\delta > 1/n$. In contrast, we aim at deriving bounds that capture the dependence of the sample-complexity on δ , as δ becomes small.

3 A Lower bound on the sample complexity

In this section, we present our main results. Theorem 1 below matches the upper bounds in [5]. Throughout, \log stands for the natural logarithm.

Theorem 1. *There exist positive constants c_1 , c_2 , ε_0 , and δ_0 , such that for every $n \geq 2$, $\varepsilon \in (0, \varepsilon_0)$, and $\delta \in (0, \delta_0)$, and for every (ε, δ) -correct policy, there exists some $p \in [0, 1]^n$ such that*

$$\mathbf{E}_p[T] \geq c_1 \frac{n}{\varepsilon^2} \log \frac{c_2}{\delta}.$$

In particular, ε_0 and δ_0 can be taken equal to $1/4$ and $e^{-4}/4$, respectively.

Instead of proving Theorem 1, we will establish a stronger result, which refers to the case where the values of the biases p_i are known up to a permutation. More specifically, we are given a vector $q \in [0, 1]^n$, and we are told that the true vector p of coin biases is of the form $p = q \circ \sigma$, where σ is an unknown permutation of the set $\{1, \dots, n\}$, and where $q \circ \sigma$ stands for permuting the components of the vector q according to σ , i.e., $(q \circ \sigma)_i = q_{\sigma(i)}$. We say that a policy is (q, ε, δ) -correct if

$$\mathbf{P}_{q \circ \sigma} \left(p_I > \max_i q_i - \varepsilon \right) \geq 1 - \delta,$$

for every permutation σ of the set $\{1, \dots, n\}$.

Theorem 2 below establishes a lower bound of the same form as the one in Theorem 1, for the special case discussed here. In essence it provides a lower bound on the number of trials needed in order to identify, with high probability, a coin with bias $(1/2) + \varepsilon$, out of a population consisting of this coin and $n - 1$ fair coins. Theorem 1 is a straightforward corollary of Theorem 2. This is because any (ε, δ) -correct policy must also be (q, ε, δ) -correct for every given q .

Theorem 2. *There exist positive constants c_1 , c_2 , ε_0 , and δ_0 , such that for every $n \geq 2$, $\varepsilon \in (0, \varepsilon_0)$, and $\delta \in (0, \delta_0)$, there exists $q \in [0, 1]^n$ such that every (q, ε, δ) -correct policy satisfies*

$$\mathbf{E}_{\sigma \circ q}[T] \geq c_1 \frac{n}{\varepsilon^2} \log \frac{c_2}{\delta},$$

for every permutation σ . In particular, ε_0 and δ_0 can be taken equal to $1/4$ and $e^{-4}/4$, respectively.

Proof. Let us fix n , $\varepsilon_0 = 1/4$, $\delta_0 = e^{-4}/4$, and some $\varepsilon \in (0, \varepsilon_0)$, and $\delta \in (0, \delta_0)$. Let

$$q = \left(\frac{1}{2} + \varepsilon, \frac{1}{2}, \dots, \frac{1}{2}\right).$$

Suppose that we have a policy which is (q, ε, δ) -correct. We need to show that under such a policy the lower bound in the theorem statement holds.

We introduce a collection of hypotheses H_1, \dots, H_n , defined by

$$H_\ell : \quad p_\ell = \frac{1}{2} + \varepsilon, \quad p_i = \frac{1}{2}, \quad \text{for } i \neq \ell.$$

Since the policy is (q, ε, δ) -correct, the following must be true for every hypothesis H_ℓ : if H_ℓ is true, the policy must have probability at least $1 - \delta$ of eventually stopping and selecting coin ℓ . We denote by \mathbf{E}_ℓ and \mathbf{P}_ℓ the expectation and probability, respectively, under the policy being considered and under hypothesis H_ℓ .

As in Section 2, we denote by T_i the number of times coin i is tried. We also define $K_t^i = X_1^i + \dots + X_t^i$, which is the number of unit rewards (“heads”) if the coin i is tried t times (not necessarily consecutively.) Let $t^* = \frac{1}{c\varepsilon^2} \log \frac{1}{4\delta}$, where c is an absolute constant whose value will be specified later. We also introduce the following events:

- (i) $A_i = \{T_i \leq 4t^*\}$. (Coin i was not tried too many times.)
- (ii) $B_i = \{I = i\}$. (Coin i was selected upon termination.)
- (iii) $C_i = \left\{ \max_{1 \leq t \leq 4t^*} \left| K_t^i - \frac{1}{2}t \right| < \sqrt{t^* \log(1/4\delta)} \right\}$. (Coin i appears fair throughout the process.)

We now establish two lemmas that will be used in the sequel.

Lemma 1. *If $\ell \neq i$, then $\mathbf{P}_\ell(C_i) > 3/4$.*

Proof. First, note that $K_t^i - p_i t$ is a \mathbf{P}_ℓ -martingale, where $p_i = 1/2$ is the bias of coin i under hypothesis H_ℓ . (We are developing the proof for a general p_i , because this will be useful later.) Using Kolmogorov’s inequality (Corollary 7.66, in p. 244 of [11]), the probability of the complement of C_i can be bounded as follows:

$$\mathbf{P}_\ell \left(\max_{1 \leq t \leq 4t^*} \left| K_t^i - p_i t \right| \geq \sqrt{t^* \log(1/4\delta)} \right) \leq \frac{\mathbf{E}_\ell[(K_{4t^*}^i - 4p_i t^*)^2]}{t^* \log(1/4\delta)}.$$

Since $\mathbf{E}_\ell[(K_{4t^*}^i - 4p_i t^*)^2] = 4p_i(1 - p_i)t^*$, we obtain

$$\mathbf{P}_\ell(C_i) \geq 1 - \frac{4p_i(1 - p_i)}{\log(1/4\delta)} > \frac{3}{4},$$

where the last inequality follows because $4\delta < e^{-4}$ and $4p_i(1 - p_i) \leq 1$. □

Lemma 2. *If $0 \leq x \leq 3/4$ and $y \geq 0$, then*

$$(1 - x)^y \geq e^{-dxy},$$

where $d = 2$.

Proof. A straightforward calculation shows that $\log(1-x) + dx \geq 0$ for $0 \leq x \leq 3/4$. Therefore, $y(\log(1-x) + dx) \geq 0$ for every $y \geq 0$. Rearranging and exponentiating leads to $(1-x)^y \geq e^{-dxy}$. \square

The next key lemma is the central part of the proof. It uses a likelihood ratio argument to show that if a coin i is not tried enough times under hypothesis H_ℓ , the probabilities of selecting coin ℓ under hypothesis H_i is substantial.

Lemma 3. *Suppose that $\ell \neq i$ and $\mathbf{E}_\ell[T_i] \leq t^*$. If the constant c in the definition of t^* is larger than 128, then $\mathbf{P}_i(B_\ell) > \delta$.*

Proof. We first observe that

$$t^* \geq \mathbf{E}_\ell[T_i] > 4t^* \mathbf{P}_\ell(T_i > 4t^*) = 4t^*(1 - \mathbf{P}_\ell(T_i \leq 4t^*)),$$

from which it follows that $\mathbf{P}_\ell(A_i) > 3/4$. We now define the event

$$S_i^\ell = A_i \cap B_\ell \cap C_i.$$

Since $\delta < \delta_0 < 1/4$, and the policy is (q, ε, δ) -correct, we must have $\mathbf{P}_\ell(B_\ell) > 3/4$. Furthermore, $\mathbf{P}_\ell(C_i) > 3/4$ (by Lemma 1). It follows that $\mathbf{P}_\ell(S_i^\ell) > \frac{1}{4}$.

We let W be the history of the process until the policy terminates. Thus, W consists of the sequence of observed rewards, and if the policy uses randomization, the sequence of coin choices as well. We define the likelihood function L_i by letting

$$L_i(w) = \mathbf{P}_i(W = w),$$

for every possible history w , and an associated random variable $L_i(W)$. We also let K_i be a shorthand notation for $K_{T_i}^i$, the total number of unit rewards (“heads”) obtained from coin i . Because the underlying hypothesis only affects the coin biases but not the statistics of the policy’s randomizations, we have

$$\begin{aligned} \frac{L_i(W)}{L_\ell(W)} &= \frac{(\frac{1}{2} + \varepsilon)^{K_i} (\frac{1}{2} - \varepsilon)^{T_i - K_i} \frac{1}{2}^{T_\ell}}{\frac{1}{2}^{T_i} (\frac{1}{2} + \varepsilon)^{K_\ell} (\frac{1}{2} - \varepsilon)^{T_\ell - K_\ell}} = \frac{(1 + 2\varepsilon)^{K_i} (1 - 2\varepsilon)^{T_i - K_i}}{(1 + 2\varepsilon)^{K_\ell} (1 - 2\varepsilon)^{T_\ell - K_\ell}} \\ &= \frac{(1 - 4\varepsilon^2)^{K_i} (1 - 2\varepsilon)^{T_i - 2K_i}}{(1 - 4\varepsilon^2)^{K_\ell} (1 - 2\varepsilon)^{T_\ell - 2K_\ell}}. \end{aligned}$$

Since $\varepsilon < \varepsilon_0 = 1/4$, the denominator is smaller than 1, so that

$$\frac{L_i(W)}{L_\ell(W)} \geq (1 - 4\varepsilon^2)^{K_i} (1 - 2\varepsilon)^{T_i - 2K_i}. \quad (1)$$

We will now proceed to lower bound the terms in the right-hand side of Eq. (1) when event S_i^ℓ occurs. If event S_i^ℓ occurs, then A_i occurs, so that $K_i \leq T_i \leq 4t^*$. We therefore have

$$\begin{aligned} (1 - 4\varepsilon^2)^{K_i} &\geq (1 - 4\varepsilon^2)^{4t^*} = (1 - 4\varepsilon^2)^{(4/(c\varepsilon^2)) \log(1/4\delta)} \\ &\geq e^{-(16d/c) \log(1/4\delta)} = (4\delta)^{16d/c}. \end{aligned}$$

(In the inequalities above, we made use of the assumption that $\varepsilon < \varepsilon_0 = 1/4$ and Lemma 2.) Similarly, if the event $A_i \cap C_i$ occurs, then $T_i - 2K_i \leq 2\sqrt{t^* \log(1/4\delta)}$, so that

$$(1 - 2\varepsilon)^{T_i - 2K_i} \geq (1 - 2\varepsilon)^{2\sqrt{t^* \log(1/4\delta)}} \geq e^{-(4d/\sqrt{c}) \log(1/4\delta)} = (4\delta)^{4d/\sqrt{c}}.$$

Using the above inequalities, and by picking c large enough ($c > 128$ will suffice), we obtain that $L_1(W)/L_0(W)$ is at least 4δ , whenever the event S_i^ℓ occurs. More precisely, we have $\frac{L_i(W)}{L_\ell(W)} 1_{S_i^\ell} \geq 4\delta 1_{S_i^\ell}$, where 1_D is the indicator function of an event D . We then have

$$\mathbf{P}_i(B_\ell) \geq \mathbf{P}_i(S_i^\ell) = \mathbf{E}_i[1_{S_i^\ell}] = \mathbf{E}_\ell \left[\frac{L_i(W)}{L_\ell(W)} 1_{S_i^\ell} \right] \geq \mathbf{E}_\ell[4\delta 1_{S_i^\ell}] = 4\delta \mathbf{P}_\ell(S_i^\ell) > \delta,$$

where the last inequality made use of the already established fact $\mathbf{P}_\ell(S_i^\ell) > 1/4$. \square

Since the policy is (q, ε, δ) -correct, we must have $\mathbf{P}_i(B_\ell) \leq \delta$ for every $i \neq \ell$. Lemma 3 then implies that for all $i \neq \ell$ we have that $\mathbf{E}_\ell[T_i] > t^* = \frac{1}{c\varepsilon^2} \log \frac{1}{4\delta}$. By summing over all $i \neq \ell$, we obtain $\mathbf{E}_\ell[T] \geq \frac{n-1}{c\varepsilon^2} \log \frac{1}{4\delta}$, which is of the required form, with $c_1 \geq (n-1)/nc \geq 1/2c$ and $c_2 = 1/4$. \square

4 The Bayesian setup

In this section, we study another variant of the problem, which is based on a Bayesian formulation. In this variant, the parameters p_i associated with each arm are not unknown constants, but random variables described by a given prior. In this case, there is a single underlying probability measure which we denote by \mathbf{P} , and which is a mixture of the measures \mathbf{P}_p . We also use \mathbf{E} to denote expectation with respect to \mathbf{P} . We then define a policy to be (ε, δ) -correct, under a particular prior and associated measure \mathbf{P} , if

$$\mathbf{P} \left(p_I > \max_i p_i - \varepsilon \right) \geq 1 - \delta.$$

We then have the following result.

Theorem 3. *There exist positive constants c_1 , c_2 , ε_0 , and δ_0 , such that for every $n \geq 2$, $\varepsilon \in (0, \varepsilon_0)$, and $\delta \in (0, \delta_0)$, there exists a prior such that every (ε, δ) -correct policy under this prior satisfies*

$$\mathbf{E}[T] \geq c_1 \frac{n}{\varepsilon^2} \log \frac{c_2}{\delta}.$$

In particular, ε_0 and δ_0 can be taken equal to $1/4$ and $e^{-4}/12$, respectively.

Proof. Let us fix some $\varepsilon \in (0, \varepsilon_0)$, $\delta \in (0, \delta_0)$, and some (ε, δ) -correct policy. Consider the hypotheses H_1, \dots, H_n , introduced in the proof of Theorem 2. For $i = 1, \dots, n$, let the prior probability of H_i be equal to $1/n$. It follows that

$$\mathbf{E}[T] = \frac{1}{n} \sum_{\ell=1}^n \sum_{i=1}^n \mathbf{E}_\ell[T_i].$$

Using the same definition for events as in Theorem 2, we have

$$\frac{1}{n} \sum_{\ell=1}^n \mathbf{P}_\ell(B_\ell) = \mathbf{P}\left(p_I > \max_i p_i - \varepsilon\right) \geq 1 - \delta. \quad (2)$$

Let G be the set of coins ℓ for which $\mathbf{P}_\ell(B_\ell) > 1 - 3\delta$. From Eq. (2), we obtain that the cardinality of G satisfies $|G| \geq \lfloor 2n/3 \rfloor$.

We now proceed as in the proof of Theorem 2, except that we replace throughout δ by 3δ . The condition $\delta < e^{-4}/4$ becomes $\delta < e^{-4}/12 = \delta_0$. The analogs of Lemmas 1 and 3 go through. In particular, Lemma 3 implies that if $\ell \neq i$ and $\mathbf{E}_\ell[T_i] \leq (1/c\varepsilon^2) \log(1/12\delta)$, then $\mathbf{P}_i(B_\ell) \geq 3\delta$. However, this can never happen for $\ell \in G$. Hence,

$$\mathbf{E}_\ell[T_i] \geq \frac{1}{c\varepsilon^2} \log \frac{1}{12\delta}, \quad \forall \ell \in G, \forall i \neq \ell.$$

We conclude that

$$\mathbf{E}[T] \geq \frac{1}{n} \sum_{\ell \in G} \sum_{i \neq \ell} \mathbf{E}_\ell[T_i] \geq \frac{|G|(n-1)}{n} \cdot \frac{1}{c\varepsilon^2} \log \frac{1}{12\delta} \geq c_1 \frac{n}{\varepsilon^2} \log \frac{1}{12\delta},$$

where the selection of c_1 is such that the last inequality holds for all $n > 1$. \square

5 A lower bound on the sample complexity - general probabilities

In Theorem 1, we developed a lower bound on the amount of exploration required for any (ε, δ) -correct policy, by exhibiting an unfavorable vector p of coin biases, under which a lot of exploration is necessary. But this leaves open the possibility that for other, more favorable choices of p , less exploration might suffice.

In this section, we refine Theorem 1 by developing a lower bound that explicitly depends on the actual (though unknown) vector p . Of course, for any given vector p , there is an “optimal” policy, which selects the best coin without any exploration: e.g., if $p_* \geq p_i$ for all i , the policy that immediately selects coin 1 is “optimal.” However, such a policy will not be (ε, δ) -correct for *all* possible vectors p .

We start with a lower bound that applies when all coin biases p_i lie in the range $[0, 1/2]$. We will later use a reduction technique to extend the result to a generic range of biases. In the rest of the paper, we use the notational convention $(x)^+ = \max\{0, x\}$.

Theorem 4. *Fix some $p \in (0, 1/2)$. There exists a constant c_1 that depends only on p such that for every $\varepsilon \in (0, 1/2)$, every $\delta \in (0, e^{-4}/4)$, every $p \in [0, 1/2]^n$, and every (ε, δ) -correct policy, we have*

$$\mathbf{E}_p[T] \geq c_1 \left\{ \frac{(|M(p, \varepsilon)| - 3)^+}{\varepsilon^2} + \sum_{\ell \in N(p, \varepsilon)} \frac{1}{(p_* - p_\ell)^2} \right\} \log \frac{1}{4\delta},$$

where $p_* = \max_i p_i$,

$$M(p, \varepsilon) = \left\{ \ell : p_\ell \geq p_* - \varepsilon, \text{ and } p_\ell > \underline{p}, \text{ and } p_\ell \geq \frac{\varepsilon + p_*}{1 + \sqrt{1/2}} \right\}, \quad (3)$$

and

$$N(p, \varepsilon) = \left\{ \ell : p_\ell < p_* - \varepsilon, \text{ and } p_\ell > \underline{p}, \text{ and } p_\ell \geq \frac{\varepsilon + p_*}{1 + \sqrt{1/2}} \right\}. \quad (4)$$

Remarks:

- (a) The lower bound involves two sets of coins whose biases are not too far from the best bias p_* . The first set $M(p, \varepsilon)$ contains coins that are within ε from the best and would therefore be legitimate selections. In the presence of multiple such coins, a certain amount of exploration is needed to obtain the required confidence that none of these coins is significantly better than the others. The second set $N(p, \varepsilon)$ contains coins whose bias is more than ε away from p_* ; they come into the lower bound because some exploration is needed in order to avoid selecting one of these coins.
- (b) The expression $(\varepsilon + p_*)/(1 + \sqrt{1/2})$ in Eqs. (3) and (4) can be replaced by $(\varepsilon + p_*)/(2 - \alpha)$ for any positive constant α , by changing some of the constants in the proof.
- (c) This result actually provides a family of lower bounds, one for every possible choice of \underline{p} . A tighter bound can be obtained by optimizing the choice of \underline{p} , while also taking into account the dependence of the constant c_1 on \underline{p} . This is not hard (the dependence of c_1 on \underline{p} can be extracted from the details of the proof), but is not pursued any further.

Proof. Let us fix some $\underline{p} \in (0, 1/2)$, $\varepsilon \in (0, 1/2)$, $\delta \in (0, e^{-4}/4)$, an (ε, δ) -correct policy, and some $p \in [0, 1/2]^n$. Without loss of generality, we assume that $p_* = p_1$. Let us denote the true (unknown) bias of each coin by q_i . We consider the following hypotheses:

$$H_0 : q_i = p_i, \text{ for } i = 1, \dots, n,$$

and for $\ell = 1, \dots, n$,

$$H_\ell : q_\ell = p_1 + \varepsilon, \quad q_i = p_i, \text{ for } i \neq \ell.$$

If hypothesis H_0 is true, when the policy terminates, it must select some coin i in the set $M(p, \varepsilon)$, in order to have an ε -optimal coin. If hypothesis H_ℓ is true, it must select coin ℓ .

We will bound from below the expected number of times the coins in the sets $N(p, \varepsilon)$ and $M(p, \varepsilon)$ must be tried, when hypothesis H_0 is true. As in Section 3, we use \mathbf{E}_ℓ and \mathbf{P}_ℓ to denote the expectation and probability, respectively, under the policy being considered and under hypothesis H_ℓ .

Let

$$t_\ell^* = \begin{cases} \frac{1}{c\varepsilon^2} \log \frac{1}{4\delta}, & \text{if } \ell \in M(p, \varepsilon), \\ \frac{1}{c(p_1 - p_\ell)^2} \log \frac{1}{4\delta}, & \text{if } \ell \in N(p, \varepsilon), \end{cases}$$

where c is a constant that only depends on \underline{p} , and whose value will be chosen later. Recall that T_i stands for the total number of times that coin i is tried. We define the event

$$A_\ell = \{T_\ell \leq 4t_\ell^*\}.$$

As in the proof of Theorem 2, if $\mathbf{E}_0[T_\ell] < t_\ell^*$, then $\mathbf{P}_0(A_\ell) > 3/4$.

We define $K_t^\ell = X_1^\ell + \dots + X_t^\ell$, which is the number of unit rewards (“heads”) if the ℓ -th coin is tried t times (not necessarily consecutively.) We let C_ℓ be the event defined by

$$C_\ell = \left\{ \max_{1 \leq t \leq 4t_\ell^*} \left| K_t^\ell - p_\ell t \right| < \sqrt{t_\ell^* \log(1/4\delta)} \right\}.$$

We have, similar to Lemma 1, $\mathbf{P}_0(C_\ell) > 3/4$. In that lemma, we had $p_\ell = 1/2$, but the proof is also valid for general p_ℓ .

Let B_ℓ be the event $\{I = \ell\}$, i.e., that the policy eventually selects coin ℓ , and let B_ℓ^c be its complement. Since the policy is (ε, δ) -correct with $\delta < 1/4$, we must have

$$\mathbf{P}_0(B_\ell^c) > 3/4, \quad \forall \ell \in N(p, \varepsilon).$$

We also have $\sum_{\ell \in M(p, \varepsilon)} \mathbf{P}_0(B_\ell) \leq 1$, so that the inequality $\mathbf{P}_0(B_\ell) > 1/4$ can hold for at most three elements of $M(p, \varepsilon)$. Equivalently, the inequality $\mathbf{P}_0(B_\ell^c) < 3/4$ can hold for at most three elements of $M(p, \varepsilon)$. Let

$$M_0(p, \varepsilon) = \left\{ \ell \in M(p, \varepsilon) \text{ and } \mathbf{P}_0(B_\ell^c) \geq \frac{3}{4} \right\}.$$

It follows that $|M_0(p, \varepsilon)| \geq (|M(p, \varepsilon)| - 3)^+$.

The following lemma is an analog of Lemma 3.

Lemma 4. *Suppose that $\mathbf{E}_0[T_\ell] < t_\ell^*$ and that the constant c in the definition of t^* is chosen large enough (possibly depending on \underline{p}). Then $\mathbf{P}_\ell(B_\ell^c) > \delta$, for every $\ell \in M_0(p, \varepsilon) \cup N(p, \varepsilon)$.*

Proof. Fix some $\ell \in M_0(p, \varepsilon) \cup N(p, \varepsilon)$. We define the event S_ℓ by

$$S_\ell = A_\ell \cap B_\ell^c \cap C_\ell.$$

Since $\mathbf{P}_0(A_\ell)$, $\mathbf{P}_0(B_\ell^c)$, $\mathbf{P}_0(C_\ell)$ are all larger than $3/4$, we have $\mathbf{P}_0(S_\ell) > \frac{1}{4}$ for all $\ell \in M_0(p, \varepsilon) \cup N(p, \varepsilon)$.

As in the proof of Lemma 3, we define the likelihood function L_ℓ by letting $L_\ell(w) = \mathbf{P}_\ell(W = w)$, for every possible history w , and use again $L_\ell(W)$ to define the corresponding random variable. Let K be a shorthand notation for $K_{T_\ell}^\ell$, the total number of unit rewards (“heads”) obtained from coin ℓ . We have

$$\begin{aligned} \frac{L_\ell(W)}{L_0(W)} &= \frac{(p_1 + \varepsilon)^K (1 - p_1 - \varepsilon)^{T_\ell - K}}{p_\ell^K (1 - p_\ell)^{T_\ell - K}} = \left(\frac{p_1}{p_\ell} + \frac{\varepsilon}{p_\ell} \right)^K \left(\frac{1 - p_1}{1 - p_\ell} - \frac{\varepsilon}{1 - p_\ell} \right)^{T_\ell - K} \\ &= \left(1 + \frac{\varepsilon + \Delta_\ell}{p_\ell} \right)^K \left(1 - \frac{\varepsilon + \Delta_\ell}{1 - p_\ell} \right)^{T_\ell - K}, \end{aligned}$$

where $\Delta_\ell = p_1 - p_\ell$. It follows that

$$\begin{aligned} & \frac{L_\ell(W)}{L_0(W)} \\ &= \left(1 + \frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^K \left(1 - \frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^K \left(1 - \frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^{-K} \left(1 - \frac{\varepsilon + \Delta_\ell}{1 - p_\ell}\right)^{T_\ell - K} \\ &= \left(1 - \left(\frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^2\right)^K \left(1 - \frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^{-K} \left(1 - \frac{\varepsilon + \Delta_\ell}{1 - p_\ell}\right)^{T_\ell - K}. \end{aligned} \quad (5)$$

Note that the fact that $(p_1 + \varepsilon)/2 < p_\ell$ (cf. Eqs. (3) and (4)) implies that $0 \leq (\varepsilon + \Delta_\ell)/p_\ell \leq 1$.

We will now proceed to lower bound the right-hand side of Eq. (5) for histories under which event S_ℓ occurs.

If event S_ℓ has occurred, then A_ℓ has occurred, and we have $K \leq T_\ell \leq 4t^*$, so that for every $\ell \in N(\varepsilon, p)$, we have

$$\begin{aligned} \left(1 - \left(\frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^2\right)^K &\geq \left(1 - \left(\frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^2\right)^{4t_\ell^*} = \left(1 - \left(\frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^2\right)^{\frac{4}{c\Delta_\ell^2} \log(1/4\delta)} \\ &\stackrel{a}{\geq} \exp\left\{-d\frac{4}{c}\left(\frac{\varepsilon/\Delta_\ell + 1}{p_\ell}\right)^2 \log(1/4\delta)\right\} \\ &\stackrel{b}{\geq} \exp\left\{-d\frac{16}{cp_\ell^2} \log(1/4\delta)\right\} = (4\delta)^{16d/p_\ell^2c}. \end{aligned}$$

In step (a), we have used Lemma 2 and the fact $(\varepsilon + p_1)/(1 + (1/\sqrt{2})) < p_\ell$; in step (b), we used the fact $\varepsilon/\Delta_\ell \leq 1$. (Both facts hold because $\ell \in N(\varepsilon, p)$.)

Similarly, for $\ell \in M(\varepsilon, p)$, we have

$$\begin{aligned} \left(1 - \left(\frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^2\right)^K &\geq \left(1 - \left(\frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^2\right)^{4t_\ell^*} = \left(1 - \left(\frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^2\right)^{\frac{4}{c\varepsilon^2} \log(1/4\delta)} \\ &\stackrel{a}{\geq} \exp\left\{-d\left(\frac{1 + (\Delta_\ell/\varepsilon)}{p_\ell}\right)^2 \frac{4}{c} \log(1/4\delta)\right\} \\ &\stackrel{b}{\geq} \exp\left\{-d\frac{16}{cp_\ell^2} \log(1/4\delta)\right\} = (4\delta)^{16d/p_\ell^2c}. \end{aligned}$$

In step (a), we have used Lemma 2 and the fact $(\varepsilon + p_1)/(1 + (1/\sqrt{2})) < p_\ell$; in step (b), we used the fact $\Delta_\ell/\varepsilon \leq 1$. (Both facts hold because $\ell \in M(\varepsilon, p)$.)

We now bound the product of the second and third terms in Eq. (5). Note that if $p_\ell \leq 1/2$, then $1/p_\ell > 1/(1 - p_\ell)$. It follows that

$$\left(1 - \frac{\varepsilon + \Delta_\ell}{p_\ell}\right)^{-K} \geq \left(1 - \frac{\varepsilon + \Delta_\ell}{1 - p_\ell}\right)^{-K}. \quad (6)$$

We start with the case where $\ell \in N(p, \varepsilon)$. Using Eq. (6) we obtain

$$\begin{aligned}
\left(1 - \frac{\Delta_\ell + \varepsilon}{p_\ell}\right)^{-K} \left(1 - \frac{\varepsilon + \Delta_\ell}{1 - p_\ell}\right)^{T_\ell - K} &\geq \left(1 - \frac{\varepsilon + \Delta_\ell}{1 - p_\ell}\right)^{T_\ell - 2K} \\
&\stackrel{a}{\geq} \left(1 - \frac{\varepsilon + \Delta_\ell}{1 - p_\ell}\right)^{T_\ell - 4p_\ell K} \\
&\stackrel{b}{\geq} \left(1 - \frac{\varepsilon + \Delta_\ell}{1 - p_\ell}\right)^{2\sqrt{t_\ell^* \log(1/4\delta)}} \\
&\stackrel{c}{\geq} \exp\left\{\frac{-2d(\varepsilon + \Delta_\ell)}{\sqrt{c}\Delta_\ell(1 - p_\ell)}\sqrt{\log(1/4\delta)}\right\} \quad (7) \\
&\stackrel{d}{\geq} \exp\left\{\frac{-4d}{(1 - p_\ell)\sqrt{c}}\sqrt{\log(1/4\delta)}\right\} \quad (8) \\
&\stackrel{e}{\geq} \exp\left\{-4d\frac{1}{1 - p_\ell}\frac{1}{\sqrt{c}}\log(1/4\delta)\right\} \\
&\stackrel{f}{\geq} \exp\left\{-8d\frac{1}{\sqrt{c}}\log(1/4\delta)\right\} = (4\delta)^{8d/\sqrt{c}}.
\end{aligned}$$

Here, (a) holds because $p_\ell \leq 1/2$; (b) holds because we are assuming that the event $A_\ell \cap C_\ell$ has occurred; (c) follows by Lemma 2 and by noticing that if $(\varepsilon + p_1)/(1 + \sqrt{1/2}) \leq p_\ell$ and $p_\ell \leq 1/2$, then $(\varepsilon + \Delta_\ell)/(1 - p_\ell) \leq 1/\sqrt{2}$; (d) follows from the fact that $\Delta_\ell > \varepsilon$; (e) follows because the assumption $\delta < e^{-4}/4$ implies that $\sqrt{\log(1/4\delta)} < \log(1/4\delta)$; and (f) holds because $0 \leq p_\ell \leq 1/2$, which implies that $1/(1 - p_\ell) \leq 2$.

Consider now the case where $\ell \in M(p, \varepsilon)$. Developing a bound for the product of the second and third terms in Eq. (5) is similar to the case where $\ell \in N(p, \varepsilon)$. The only difference is that in step (b), t_ℓ^* should be replaced with $1/c\varepsilon^2 \log(1/4\delta)$ (we assume here too that $(\varepsilon + p_1)/(1 + \sqrt{1/2}) \leq p_\ell$). Then, the right-hand side in Eq. (7) becomes $\exp\left\{-d\frac{\varepsilon + \Delta_\ell}{1 - p_\ell} \cdot \frac{2}{\sqrt{c\varepsilon}}\sqrt{\log(1/4\delta)}\right\}$. For $\ell \in M(p, \varepsilon)$, we have $\Delta_\ell \leq \varepsilon$, which implies that $(\varepsilon + \Delta_\ell)/\varepsilon \leq 2$, which then leads to the same expression as in Eq. (8). The rest of the derivation is identical.

Summarizing the above, we have shown that if $\ell \in M(p, \varepsilon) \cup N(p, \varepsilon)$, and event S_ℓ has occurred, then $\frac{L_\ell(W)}{L_0(W)} \geq (4\delta)^{\frac{8d}{\sqrt{c}}}(4\delta)^{\frac{16d}{p_\ell^2 c}}$. For $\ell \in M(p, \varepsilon) \cup N(p, \varepsilon)$, we have $\underline{p} < p_\ell$. We can choose c large enough so that $L_\ell(W)/L_0(W) \geq 4\delta$; the value of c depends only on the constant \underline{p} .

Similar to the proof of Theorem 2, we have $\frac{L_\ell(W)}{L_0(W)}1_{S_\ell} \geq 4\delta 1_{S_\ell}$, where 1_{S_ℓ} is the indicator function of the event S_ℓ . It follows that

$$\mathbf{P}_\ell(B_\ell^c) \geq \mathbf{P}_\ell(S_\ell) = \mathbf{E}_\ell[1_{S_\ell}] = \mathbf{E}_0\left[\frac{L_\ell(W)}{L_0(W)}1_{S_\ell}\right] \geq \mathbf{E}_0[4\delta 1_{S_\ell}] = 4\delta \mathbf{P}_0(S_\ell) > \delta,$$

where the last inequality relies on the already established fact $\mathbf{P}_\ell(S_\ell) > 1/4$. \square

Since the policy is (ε, δ) -correct, we must have $\mathbf{P}_\ell(B_\ell^c) \leq \delta$, for every ℓ . Lemma 4 then implies that $\mathbf{E}_0[T_\ell] \geq t_\ell^*$ for every $\ell \in M_0(p, \varepsilon) \cup N(p, \varepsilon)$. We sum over all $\ell \in M_0(p, \varepsilon) \cup N(p, \varepsilon)$, use the definition of t_ℓ^* , together with the fact $|M_0(p, \varepsilon)| \geq (|M(p, \varepsilon)| - 3)^+$, to conclude the proof of the theorem. \square

Remark: A close examination of the proof reveals that the dependence of c_1 on \underline{p} is captured by a requirement of the form $c_1 \leq \min\{c_2, c_3 \underline{p}^2\}$, for some absolute constants c_2 and c_3 . This suggests that there is a tradeoff in the choice of \underline{p} . By choosing a large \underline{p} , the constant c_1 is made large, but the sets M and N become smaller, and vice versa.

The preceding result may give the impression that the sample complexity is high only when the p_i are bounded by $1/2$. The next result shows that similar lower bounds hold (with a different constant) whenever the p_i can be assumed to be bounded away from 1. However, the lower bound becomes weaker (i.e., the constant c_1 is smaller) when the upper bound on the p_i approaches 1. This is because no $O(1/\varepsilon^2)$ lower bound on $\mathbf{E}_p[T]$ is possible when $\max_i p_i = 1$. In fact, there exists an (ε, δ) -correct algorithm such that $\mathbf{E}_p[T]$ is proportional to $1/\varepsilon$ for every p with $\max_i p_i = 1$.

Theorem 5. *Fix an integer $s \geq 2$, and some $\underline{p} \in (0, 1/2)$. There exists a constant c_1 that depends only on \underline{p} such that for every $\varepsilon \in (0, 2^{-(s+1)})$, every $\delta \in (0, e^{-4}/4)$, every $p \in [0, 1 - 2^{-s}]^n$, and every (ε, δ) -correct policy, we have*

$$\mathbf{E}_p[T] \geq \frac{c_1}{s\eta^2} \left\{ \frac{(|M(\tilde{p}, \varepsilon\eta)| - 3)^+}{\varepsilon^2} + \sum_{\ell \in N(\tilde{p}, \eta\varepsilon)} \frac{1}{(p_* - p_\ell)^2} \right\} \log \frac{1}{4\delta},$$

where $p_* = \max_i p_i$, $\eta = 2^s/s$, \tilde{p} is the vector with components $\tilde{p}_i = 1 - (1 - p_i)^{1/s}$, and M and N are as defined in Theorem 4.

Proof. Let us fix $s \geq 2$, $\underline{p} \in (0, 1/2)$, $\varepsilon \in (0, 2^{-(s+1)})$, and $\delta \in (0, e^{-4}/4)$. Suppose that we have an (ε, δ) -correct policy π whose expected time to termination is $\mathbf{E}_p[T]$, whenever the vector of coin biases happens to be p . We will use the policy π to construct a new policy $\tilde{\pi}$ such that

$$\mathbf{P}\left(p_I > \max_i p_i - \eta\varepsilon\right) \geq 1 - \delta, \quad \forall p \in [0, (1/2) + \eta\varepsilon]^n;$$

(we will then say that $\tilde{\pi}$ is $(\eta\varepsilon, \delta)$ -correct on $[0, (1/2) + \eta\varepsilon]^n$). Finally, we will use the lower bounds from Theorem 4, applied to $\tilde{\pi}$, to obtain a lower bound on the sample complexity of π .

The new policy $\tilde{\pi}$ is described as follows. Run the original policy π . Whenever π chooses to try a certain coin i once, policy $\tilde{\pi}$ tries coin i for s consecutive times. Policy $\tilde{\pi}$ then “feeds” π with 0 if all s trials resulted in 0, and feeds π with 1 otherwise. If \tilde{p} is the true vector of coin biases faced by policy $\tilde{\pi}$, and if policy π chooses to sample coin i , then policy π “sees” an outcome which equals 1 with

probability $p_i = 1 - (1 - \tilde{p}_i)^s$. Let us define two mappings $f, g : [0, 1] \mapsto [0, 1]$, which are inverses of each other by

$$f(p_i) = 1 - (1 - p_i)^{1/s}, \quad g(\tilde{p}_i) = 1 - (1 - \tilde{p}_i)^s,$$

and with a slight abuse of notation, let $f(p) = (f(p_1), \dots, f(p_n))$, and similarly for $g(\tilde{p})$. With our construction, when policy $\tilde{\pi}$ is faced with a bias vector, it evolves in an identical manner as the policy π faced with a bias vector $p = g(\tilde{p})$. But under policy $\tilde{\pi}$, there are s trials associated with every trial under policy π , which implies that $\tilde{T} = sT$ (\tilde{T} is the number of trials under policy $\tilde{\pi}$) and therefore

$$\mathbf{E}_{\tilde{p}}[\tilde{T}] = s\mathbf{E}_{g(\tilde{p})}[T], \quad \mathbf{E}_{f(p)}[\tilde{T}] = s\mathbf{E}_p[T]. \quad (9)$$

We will now determine the ‘‘correctness’’ guarantees of policy $\tilde{\pi}$. We first need some algebraic preliminaries.

Let us fix some $\tilde{p} \in [0, (1/2) + \eta\varepsilon]^n$ and a corresponding vector p , related by $\tilde{p} = f(p)$ and $p = g(\tilde{p})$. Let also $p_* = \max_i p_i$ and $\tilde{p}_* = \max_i \tilde{p}_i$. Using the definition $\eta = 2^s/s$ and the assumption $\varepsilon < 2^{-(s+1)}$, we have $\tilde{p}_* \leq (1/2) + (1/2s)$, from which it follows that

$$p_* \leq 1 - \left(\frac{1}{2} - \frac{1}{2s}\right)^s = 1 - \frac{1}{2^s} \left(1 - \frac{1}{s}\right) \leq 1 - \frac{1}{2^s} \cdot \frac{1}{2} = 1 - 2^{-(s+1)}.$$

The derivative f' of f is monotonically increasing on $[0, 1)$. Therefore,

$$\begin{aligned} f'(p_*) &\leq f'(1 - 2^{-(s+1)}) = \frac{1}{s} \left(2^{-(s+1)}\right)^{(1/s)-1} = \frac{1}{s} 2^{-(s+1)(1-s)/s} \\ &= \frac{1}{s} 2^{s-(1/s)} \leq \frac{1}{s} 2^s = \eta. \end{aligned}$$

Thus, the monotonically decreasing derivative g' of the inverse mapping is at least $1/\eta$ in the set $[0, (1/2) + \eta\varepsilon]$. Hence, $g'(\tilde{p}_*) \geq \frac{1}{\eta}$, which implies that $g(\tilde{p}_* - \eta\varepsilon) \leq g(\tilde{p}_*) - g'(\tilde{p}_*)\eta\varepsilon \leq g(\tilde{p}_*) - \varepsilon$.

Let I be the coin index finally selected by policy $\tilde{\pi}$ when faced with \tilde{p} , which is the same as the index chosen by π when faced with p . We have

$$\begin{aligned} \mathbf{P}(\tilde{p}_I \leq \tilde{p}_* - \eta\varepsilon) &= \mathbf{P}(g(\tilde{p}_I) \leq g(\tilde{p}_* - \eta\varepsilon)) \leq \mathbf{P}(g(\tilde{p}_I) \leq g(\tilde{p}_*) - \varepsilon) \\ &= \mathbf{P}(p_I \leq p_* - \varepsilon) \leq 1 - \delta, \end{aligned}$$

where the last inequality follows because policy π was assumed to be (ε, δ) -correct. We have therefore established that $\tilde{\pi}$ is $(\eta\varepsilon, \delta)$ -correct on $[0, (1/2) + \eta\varepsilon]^n$. We now apply Theorem 4, with $\eta\varepsilon$ instead of ε . Even though that theorem is stated for a policy which is (ε, δ) -correct for all possible p , the proof shows that it also applies to (ε, δ) -correct policies on $[0, (1/2) + \varepsilon]^n$. This gives a lower bound on $\mathbf{E}_{\tilde{p}}[\tilde{T}]$ which, using Eq. (9), translates to the claimed lower bound on $\mathbf{E}_p[T]$. This lower bound applies whenever $p = g(\tilde{p})$, for some $\tilde{p} \in [0, 1/2]^n$, and therefore whenever $p \in [0, 1 - 2^{-s}]^n$. \square

6 Concluding Remarks

We have addressed the problem of deriving lower bounds on the number of steps required to identify a near optimal arm, with high probability, in a multi-armed bandit setting. For the problem formulations studied in Section 3 and 4, the lower bounds match the existing upper bounds of $\Theta((n/\varepsilon^2) \log(1/\delta))$.

Our results have been derived under the assumption of Bernoulli rewards. Clearly, the lower bounds also apply to more general problem formulations, as long as they include the special case of Bernoulli rewards. It would be of some interest to derive similar lower bounds for other special cases of reward distributions. It is reasonable to expect that essentially the same results will carry over, as long as the divergence between the reward distribution associated with different arms is finite (as in [9]).

Acknowledgments. This research was partially supported by a Fulbright postdoctoral fellowship, by the MIT-Merrill Lynch partnership, and by the ARO under grant DAAD10-00-1-0466.

References

1. M. Anthony and P.L. Bartlett. *Neural Network Learning; Theoretical Foundations*. Cambridge University Press, 1999.
2. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proc. 36th Annual Symposium on Foundations of Computer Science*, pages 322–331. IEEE Computer Society Press, 1995.
3. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The non-stochastic multi-armed bandit problem. To appear in *SIAM journal of Computation*, 2002.
4. D.A. Berry and B. Fristedt. *Bandit Problems*. Chapman and Hall, 1985.
5. E. Even-Dar, S. Mannor, and Y. Mansour. PAC Bounds for Multi-Armed Bandit and Markov Decision Processes. In *Fifteenth Annual Conference on Computation Learning Theory*, pages 255–270, 2002.
6. J. Gittins and D. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani, K. Sarkadi, and I. Vincze, editors, *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam, 1974.
7. C. Jennison, I. M. Johnstone and B.W. Turnbull. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. In S. S. Gupta and J. Berger, editors, *Statistical decision theory and related topics III, Vol 2*, pages 55–86. Academic Press, 1982.
8. S. R. Kulkarni and G. Lugosi. Finite-time lower bounds for the two-armed bandit problem *IEEE Trans. Aut. Control*, 45(4):711-714, 2000.
9. T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.
10. H. Robbins. Some aspects of sequential design of experiments. *Bull. Amer. Math. Soc.*, 55:527–535, 1952.
11. S. M. Ross. *Stochastic processes*. Wiley, 1983.
12. D. Siegmund. *Sequential analysis—tests and confidence intervals*. Springer Verlag, 1985.