Journal of Machine Learning Research 5 (2004) 623-648 Submitted 1/04; Published 6/04

The Sample Complexity of Exploration in the Multi-Armed Bandit Problem

Shie Mannor SHIE@MIT.EDU
John N. Tsitsiklis JNT@MIT.EDU
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editors: Kristin Bennett and Nicolò Cesa-Bianchi

Abstract

We consider the multi-armed bandit problem under the PAC ("probably approximately correct") model. It was shown by Even-Dar et al. (2002) that given n arms, a total of $O\big((n/\varepsilon^2)\log(1/\delta)\big)$ trials suffices in order to find an ε-optimal arm with probability at least $1-\delta$. We establish a matching lower bound on the expected number of trials under any sampling policy. We furthermore generalize the lower bound, and show an explicit dependence on the (unknown) statistics of the arms. We also provide a similar bound within a Bayesian setting. The case where the statistics of the arms are known but the identities of the arms are not, is also discussed. For this case, we provide a lower bound of $\Theta\big((1/\varepsilon^2)(n+\log(1/\delta))\big)$ on the expected number of trials, as well as a sampling policy with a matching upper bound. If instead of the expected number of trials, we consider the maximum (over all sample paths) number of trials, we establish a matching upper and lower bound of the form $\Theta\big((n/\varepsilon^2)\log(1/\delta)\big)$. Finally, we derive lower bounds on the expected regret, in the spirit of Lai and Robbins.

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 whose probability distribution is initially unknown. We try these arms in some order, which may depend on the sequence of rewards that have been observed so far. A common objective in this context is to find a policy for choosing the next arm to be tried, 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 try 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 (1952), which derives policies 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 Berry and Fristedt (1985) 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 under an online policy, the seminal work of Lai and Robbins (1985) provides asymptotically 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 Auer et al. (1995, 2002b), 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 (2000) which shows that for any specific time t, one can choose the reward 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 Even-Dar et al. (2002), a policy, called the median elimination algorithm, was provided which requires $O((n/\varepsilon^2)\log(1/\delta))$ trials, and which finds an ε-optimal arm with probability at least $1-\delta$. A matching lower bound was also derived in Even-Dar et al. (2002), 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 also applies when $\delta > 0$ is arbitrarily small.

Our main result can be viewed as a generalization of a $O((1/\varepsilon^2)\log(1/\delta))$ lower bound provided in Anthony and Bartlett (1999), and Chernoff (1972), for the case of two bandits. The proof in Anthony and Bartlett (1999) is based on a hypothesis interchange argument, and relies critically on the fact there are only two underlying hypotheses. Furthermore, it is limited to “nonadaptive” policies, for which the number of trials is fixed a priori. The technique we use is based on a likelihood ratio argument and a tight martingale bound, and applies to general policies.

A different type of lower bound was derived in Auer et al. (2002b) for the expected regret in an adversarial setup. The bounds derived there 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., Chernoff, 1972; Jennison et al., 1982; Siegmund, 1985). In the language of Siegmund (1985), we provide a lower bound on the expected length of a sequential sampling policy under any adaptive allocation scheme. For the case of two arms, it was shown in Siegmund (1985) (p. 148) that if one restricts to sampling policies 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, Siegmund (1985) 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., Jennison et al., 1982), 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. We further comment that our results extend those of Jennison et al. (1982), in that we consider the case where the reward is not Gaussian.

Paper Outline

The paper is organized as follows. In Section 2, we set up our framework, and since we are mainly 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 $O((n/\varepsilon^2)\log(1/\delta))$ lower bound on the expected number of trials under any policy that finds an ε-optimal coin with probability at least $1-\delta$. In Section 4, we provide a refined lower bound that depends explicitly on the specific (though unknown) biases of the coins. This lower bound has the same $\log(1/\delta)$ dependence on δ; furthermore, every coin roughly contributes a factor inversely proportional to the square difference between its bias and the bias of a best coin, but no more that $1/\varepsilon^2$. In Section 5, we derive a lower bound similar to the one in Section 3, but within a Bayesian setting, under a prior distribution on the set of biases of the different coins.

In Section 6 we provide a bound on the expected regret which is similar in spirit to the bound in Lai and Robbins (1985). The constants in our bounds are slightly worse than the ones in Lai and Robbins (1985), but the different derivation, which links the PAC model to regret bounds, may be of independent interest. Our bound holds for any finite time, as opposed to the asymptotic result provided in Lai and Robbins (1985).

The case where the coin biases are known in advance, but the identities of the coins are not, is discussed in Section 7. We provide a policy that finds an ε-optimal coin with probability at least $1-\delta$, under which the expected number of trials is $O\left((1/\varepsilon^2)(n+\log(1/\delta))\right)$. We show that this bound is tight up to a multiplicative constant. If instead of the expected number of trials, we consider the maximum (over all sample paths) number of trials, we establish a matching upper and lower bounds of the form $\Theta((n/\varepsilon^2)\log(1/\delta))$. Finally, Section 8 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 ℓ is associated with a sequence of identically distributed Bernoulli (i.e., taking values in $\{0,1\}$) random variables X_k^ℓ, $k=1,2,\ldots$, with unknown mean p_ℓ. Here, X_k^ℓ corresponds to the reward obtained the kth time that arm ℓ is tried. We assume that the random variables X_k^ℓ, for $\ell=1,\ldots,n$, $k=1,2,\ldots$, are independent, and we define $p=(p_1,\ldots,p_n)$. 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 the next coin to be tried 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_ℓ be the total number of times that coin ℓ is tried, and let $T = T_1 + \cdots + T_n$ 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

$$\mathbf{P}_p\Big(p_I > \max_\ell p_\ell - \varepsilon\Big) \geq 1 - \delta,$$

for *every* $p \in [0,1]^n$. It was shown in Even-Dar et al. (2002) 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

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

A matching lower bound was also established in Even-Dar et al. (2002), 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

We start with our central result, which can be viewed as an extension of Lemma 5.1 from Anthony and Bartlett (1999), as well as a special case of Theorem 5. We present it here because it admits a simpler proof, but also because parts of the proof will be used later. Throughout the rest of the paper, log will stand 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 1/8 and $e^{-4}/4$, respectively.

Proof Let us consider a multi-armed bandit problem with $n+1$ coins, which we number from 0 to n. We consider a finite set of $n+1$ possible parameter vectors p, which we will refer to as "hypotheses." Under any one of the hypotheses, coin 0 has a known bias $p_0 = (1+\varepsilon)/2$. Under one hypothesis, denoted by H_0, all the coins other than zero have a bias of 1/2,

$$H_0:\ p_0 = \frac{1}{2} + \frac{\varepsilon}{2}, \qquad p_i = \frac{1}{2},\ \text{for } i \neq 0\,,$$

which makes coin 0 the best coin. Furthermore, for $\ell = 1, \ldots, n$, there is a hypothesis

$$H_\ell:\ p_0 = \frac{1}{2} + \frac{\varepsilon}{2}, \qquad p_\ell = \frac{1}{2} + \varepsilon, \qquad p_i = \frac{1}{2},\ \text{for } i \neq 0, \ell\,,$$

which makes coin ℓ the best coin.

We define $\varepsilon_0 = 1/8$ and $\delta_0 = e^{-4}/4$. From now on, we fix some $\varepsilon \in (0, \varepsilon_0)$ and $\delta \in (0, \delta_0)$, and a policy, which we assume to be ($\varepsilon/2$,δ)-correct . If H_0 is true, the policy must have probability at least $1-\delta$ of eventually stopping and selecting coin 0. If H_ℓ is true, for some $\ell \neq 0$, the policy must have probability at least $1-\delta$ of eventually stopping and selecting coin ℓ. We denote by $\mathbf{E}_\ell$ and $\mathbf{P}_\ell$ the expectation and probability, respectively, under hypothesis H_ℓ.

We define t^* by

$$t^* = \frac{1}{c\varepsilon^2}\log\frac{1}{4\delta} = \frac{1}{c\varepsilon^2}\log\frac{1}{\theta}, \tag{1}$$

where $\theta = 4\delta$, and where c is an absolute constant whose value will be specified later.[1] Note that $\theta < e^{-4}$ and $\varepsilon < 1/4$.

Recall that T_ℓ stands for the number of times that coin ℓ is tried. We assume that for some coin $\ell \neq 0$, we have $\mathbf{E}_0[T_\ell] \le t^*$. We will eventually show that under this assumption, the probability of selecting H_0 under H_ℓ exceeds δ, and violates $(\varepsilon/2,\delta)$-correctness. It will then follow that we must have $\mathbf{E}_0[T_\ell] > t^*$ for all $\ell \neq 0$. Without loss of generality, we can and will assume that the above condition holds for $\ell = 1$, so that $\mathbf{E}_0[T_1] \le t^*$.

We will now introduce some special events A and C under which various random variables of interest do not deviate significantly from their expected values. We define

$$A = \{T_1 \le 4t^*\},$$

and obtain

$$t^* \ge \mathbf{E}_0[T_1] \ge 4t^*\mathbf{P}_0(T_1 > 4t^*) = 4t^*(1 - \mathbf{P}_0(T_1 \le 4t^*)),$$

from which it follows that

$$\mathbf{P}_0(A) \ge 3/4.$$

We define $K_t = X_1^1 + \cdots + X_t^1$, which is the number of unit rewards ("heads") if the first coin is tried a total of t (not necessarily consecutive) times. We let C be the event defined by

$$C = \left\{\max_{1\le t\le 4t^*}\left|K_t - \frac{1}{2}t\right| < \sqrt{t^*\log(1/\theta)}\right\}.$$

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

Lemma 2 *We have* $\mathbf{P}_0(C) > 3/4$.

Proof We will prove a more general result:[2] we assume that coin i has bias p_i under hypothesis H_ℓ, define K_t^i as the number of unit rewards ("heads") if coin i is tested for t (not necessarily consecutive) times, and let

$$C_i = \left\{\max_{1\le t\le 4t^*}\left|K_t^i - p_i t\right| < \sqrt{t^*\log(1/\theta)}\right\}.$$

First, note that $K_t^i - p_i t$ is a $\mathbf{P}_\ell$-martingale (in the context of Theorem 1, $p_i = 1/2$ is the bias of coin $i = 1$ under hypothesis H_0). Using Kolmogorov's inequality (Corollary 7.66, in p. 244 of Ross, 1983), the probability of the complement of C_i can be bounded as follows:

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

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

$$\mathbf{P}_\ell(C_i) \ge 1 - \frac{4p_i(1-p_i)}{\log(1/\theta)} > \frac{3}{4}, \tag{2}$$

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

1. In this and subsequent proofs, and in order to avoid repeated use of truncation symbols, we treat t^* as if it were integer.
2. The proof for a general p_i will be useful later.

Lemma 3 *If $0 \leq x \leq 1/\sqrt{2}$ and $y \geq 0$, then*

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

where $d = 1.78$.

Proof A straightforward calculation shows that $\log(1-x) + dx \geq 0$ for $0 \leq x \leq 1/\sqrt{2}$. Therefore, $y(\log(1-x)+dx) \geq 0$ for every $y \geq 0$. Rearranging and exponentiating, leads to $(1-x)^y \geq e^{-dxy}$. □

We now let B be the event that $I = 0$, i.e., that the policy eventually selects coin 0. Since the policy is $(\varepsilon/2,\delta)$-correct for $\delta < e^{-4}/4 < 1/4$, we have $\mathbf{P}_0(B) > 3/4$. We have already shown that $\mathbf{P}_0(A) \geq 3/4$ and $\mathbf{P}_0(C) > 3/4$. Let S be the event that A, B, and C occur, that is $S = A \cap B \cap C$. We then have $\mathbf{P}_0(S) > 1/4$.

Lemma 4 *If $\mathbf{E}_0[T_1] \leq t^*$ and $c \geq 100$, then $\mathbf{P}_1(B) > \delta$.*

Proof We let W be the history of the process (the sequence of coins chosen at each time, and the sequence of observed coin rewards) until the policy terminates. We define the likelihood function L_ℓ by letting

$$L_\ell(w) = \mathbf{P}_\ell(W = w),$$

for every possible history w. Note that this function can be used to define a random variable $L_\ell(W)$. We also let K be a shorthand notation for K_{T_1}, the total number of unit rewards ("heads") obtained from coin 1. Given the history up to time $t-1$, the coin choice at time t has the same probability distribution under either hypothesis H_0 and H_1; similarly, the coin reward at time t has the same probability distribution, under either hypothesis, unless the chosen coin was coin 1. For this reason, the likelihood ratio $L_1(W)/L_0(W)$ is given by

$$\begin{aligned}
\frac{L_1(W)}{L_0(W)} &= \frac{(\frac{1}{2}+\varepsilon)^K(\frac{1}{2}-\varepsilon)^{T_1-K}}{(\frac{1}{2})^{T_1}} \\
&= (1+2\varepsilon)^K(1-2\varepsilon)^K(1-2\varepsilon)^{T_1-2K} \\
&= (1-4\varepsilon^2)^K(1-2\varepsilon)^{T_1-2K}. \qquad (3)
\end{aligned}$$

We will now proceed to lower bound the terms in the right-hand side of Eq. (3) when event S occurs.

If event S has occurred, then A has occurred, and we have $K \leq T_1 \leq 4t^*$, so that

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

We have used here Lemma 3, which applies because $4\varepsilon^2 < 4/4^2 < 1/\sqrt{2}$.

Similarly, if event S has occurred, then $A \cap C$ has occurred, which implies,

$$T_1 - 2K \leq 2\sqrt{t^*\log(1/\theta)} = (2/\varepsilon\sqrt{c})\log(1/\theta),$$

where the equality above made use of the definition of t^*. Therefore,

$$\begin{aligned}(1-2\varepsilon)^{T_1-2K} &\geq (1-2\varepsilon)^{\left(2/\varepsilon\sqrt{c}\right)\log(1/\theta)}\\ &\geq e^{-(4d/\sqrt{c})\log(1/\theta)}\\ &= \theta^{4d/\sqrt{c}}.\end{aligned}$$

Substituting the above in Eq. (3), we obtain

$$\frac{L_1(W)}{L_0(W)} \geq \theta^{(16d/c)+(4d/\sqrt{c})}.$$

By picking c large enough ($c=100$ suffices), we obtain that $L_1(W)/L_0(W)$ is larger than $\theta=4\delta$ whenever the event S occurs. More precisely, we have

$$\frac{L_1(W)}{L_0(W)}1_S \geq 4\delta 1_S,$$

where 1_S is the indicator function of the event S. Then,

$$\mathbf{P}_1(B) \geq \mathbf{P}_1(S) = \mathbf{E}_1[1_S] = \mathbf{E}_0\left[\frac{L_1(W)}{L_0(W)}1_S\right] \geq \mathbf{E}_0[4\delta 1_S] = 4\delta\mathbf{P}_0(S) > \delta,$$

where we used the fact that $\mathbf{P}_0(S) > 1/4$. □

To summarize, we have shown that when $c \geq 100$, if $\mathbf{E}_0[T_1] \leq (1/c\varepsilon^2)\log(1/(4\delta))$, then $\mathbf{P}_1(B) > \delta$. Therefore, if we have an $(\varepsilon/2,\delta)$-correct policy, we must have $\mathbf{E}_0[T_\ell] > (1/c\varepsilon^2)\log(1/(4\delta))$, for every $\ell > 0$. Equivalently, if we have an (ε,δ)-correct policy, we must have $\mathbf{E}_0[T] > (n/(4c\varepsilon^2))\log(1/(4\delta))$, which is of the desired form. □

4. A Lower Bound on the Sample Complexity - General Probabilities

In Theorem 1, we worked with a particular unfavorable vector p (the one corresponding to hypothesis H_0), under which a lot of exploration is necessary. 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 $p_1 \geq p_\ell$ for all ℓ, 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 5 *Fix some $\underline{p} \in (0,1/2)$. There exists a positive constant δ_0, and a positive constant c_1 that depends only on $\underline{p}$, such that for every $\varepsilon \in (0,1/2)$, every $\delta \in (0,\delta_0)$, every $p \in [0,1/2]^n$, and every (ε,δ)-correct policy, we have*

$$\mathbf{E}_p[T] \geq c_1\left\{\frac{(|M(p,\varepsilon)|-1)^+}{\varepsilon^2} + \sum_{\ell\in N(p,\varepsilon)}\frac{1}{(p_*-p_\ell)^2}\right\}\log\frac{1}{8\delta},$$

where $p_* = \max_i p_i$,

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

and

$$N(p,\varepsilon) = \left\{\ell: p_\ell \leq p_* - \varepsilon, \text{ and } p_\ell > \underline{p}, \text{ and } p_\ell \geq \frac{\varepsilon + p_*}{1+\sqrt{1/2}}\right\}. \tag{5}$$

In particular, δ_0 *can be taken equal to* $e^{-8}/8$.

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 again some exploration is needed in order to obtain the required confidence that none of these coins is significantly better than the best coin in $M(p,\varepsilon)$.

(b) The expression $(\varepsilon + p_*)/(1+\sqrt{1/2})$ in Eqs. (4) and (5) 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 $\underline{p}$ is described in Remark 7), but does not provide any new insights.

Proof Let us fix $\delta_0 = e^{-8}/8$, some $\underline{p} \in (0,1/2)$, $\varepsilon \in (0,1/2)$, $\delta \in (0,\delta_0)$, 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:

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

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

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

If hypothesis H_ℓ is true, the policy 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 $\mathbf{E}_\ell$ and $\mathbf{P}_\ell$ to denote the expectation and probability, respectively, under the policy being considered and under hypothesis H_ℓ.

We define $\theta = 8\delta$, and note that $\theta < e^{-8}$. Let

$$t^*_\ell = \begin{cases} \dfrac{1}{c\varepsilon^2}\log\dfrac{1}{\theta}, & \text{if } \ell \in M(p,\varepsilon), \\ \dfrac{1}{c(p_1-p_\ell)^2}\log\dfrac{1}{\theta}, & \text{if } \ell \in N(p,\varepsilon), \end{cases}$$

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

$$A_\ell = \{T_\ell \le 4t^*_\ell\}.$$

As in the proof of Theorem 1, if $\mathbf{E}_0[T_\ell] \le t^*_\ell$, then $\mathbf{P}_0(A_\ell) \ge 3/4$.

We define $K^\ell_t = X^\ell_1 + \cdots + X^\ell_t$, which is the number of unit rewards ("heads") if the ℓ-th coin is tried a total of t (not necessarily consecutive) times. We let C_ℓ be the event defined by

$$C_\ell = \left\{ \max_{1 \le t \le 4t^*_\ell} |K^\ell_t - p_\ell t| < \sqrt{t^*_\ell \log(1/\theta)} \right\}.$$

Similar to Lemma 2, and since $\theta = 8\delta < e^{-8}$, we have[3]

$$\mathbf{P}_0(C_\ell) > 7/8.$$

Let B_ℓ be the event $\{I = \ell\}$, i.e., that the policy eventually selects coin ℓ, and let B^c_ℓ be its complement. Since the policy is (ε,δ)-correct with $\delta < \delta_0 < 1/2$, we must have

$$\mathbf{P}_0(B^c_\ell) > 1/2, \qquad \forall\, \ell \in N(p,\varepsilon).$$

We also have $\sum_{\ell \in M(p,\varepsilon)} \mathbf{P}_0(B_\ell) \le 1$, so that the inequality $\mathbf{P}_0(B_\ell) > 1/2$ can hold for at most one element of $M(p,\varepsilon)$. Equivalently, the inequality $\mathbf{P}_0(B^c_\ell) \le 1/2$ can hold for at most one element of $M(p,\varepsilon)$. Let

$$M_0(p,\varepsilon) = \left\{ \ell \in M(p,\varepsilon) \text{ and } \mathbf{P}_0(B^c_\ell) > \frac{1}{2} \right\}.$$

It follows that $|M_0(p,\varepsilon)| \ge (|M(p,\varepsilon)| - 1)^+$.

The following lemma is an analog of Lemma 4.

Lemma 6 *Suppose that $\ell \in M_0(p,\varepsilon) \cup N(p,\varepsilon)$ and that $\mathbf{E}_0[T_\ell] \le t^*_\ell$. If the constant c in the definition of t^* is chosen large enough (possibly depending on $\underline{p}$), then $\mathbf{P}_\ell(B^c_\ell) > \delta$.*

Proof Fix some $\ell \in M_0(p,\varepsilon) \cup N(p,\varepsilon)$. For future reference, we note that the definitions of $M(p,\varepsilon)$ and $N(p,\varepsilon)$ include the condition $p_\ell \ge (\varepsilon + p_*)/(1+\sqrt{1/2})$. Recalling that $p_* = p_1$, $p_\ell \le 1/2$, and using the definition $\Delta_\ell = p_1 - p_\ell \ge 0$, some easy algebra leads to the conditions

$$\frac{\varepsilon + \Delta_\ell}{1 - p_\ell} \le \frac{\varepsilon + \Delta_\ell}{p_\ell} \le \frac{1}{\sqrt{2}}. \tag{6}$$

We define the event S_ℓ by

$$S_\ell = A_\ell \cap B^c_\ell \cap C_\ell.$$

Since $\mathbf{P}_0(A_\ell) \ge 3/4$, $\mathbf{P}_0(B^c_\ell) > 1/2$, and $\mathbf{P}_0(C_\ell) > 7/8$, we have

$$\mathbf{P}_0(S_\ell) > \frac{1}{8}, \qquad \forall\, \ell \in M_0(p,\varepsilon) \cup N(p,\varepsilon).$$

3. The derivation is identical to Lemma 2 except for Eq. (2), where one should replace the assumption that $\theta < e^{-4}$ with the stricter assumption that $\theta < e^{-8}$ used here.

As in the proof of Lemma 4, 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 $K^\ell_{T_\ell}$, 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 we have used the definition $\Delta_\ell = p_1 - p_\ell$. It follows that

$$\begin{aligned}\frac{L_\ell(W)}{L_0(W)} &= \left(1+\frac{\varepsilon+\Delta_\ell}{p_\ell}\right)^K\left(1-\frac{\varepsilon+\Delta_\ell}{p_\ell}\right)^K\left(1-\frac{\varepsilon+\Delta_\ell}{p_\ell}\right)^{-K}\left(1-\frac{\varepsilon+\Delta_\ell}{1-p_\ell}\right)^{T_\ell-K}\\ &= \left(1-\left(\frac{\varepsilon+\Delta_\ell}{p_\ell}\right)^2\right)^K\left(1-\frac{\varepsilon+\Delta_\ell}{p_\ell}\right)^{-K}\left(1-\frac{\varepsilon+\Delta_\ell}{1-p_\ell}\right)^{T_\ell-K}\\ &= \left(1-\left(\frac{\varepsilon+\Delta_\ell}{p_\ell}\right)^2\right)^K\left(1-\frac{\varepsilon+\Delta_\ell}{p_\ell}\right)^{-K}\left(1-\frac{\varepsilon+\Delta_\ell}{1-p_\ell}\right)^{K(1-p_\ell)/p_\ell}\\ &\qquad\cdot\left(1-\frac{\varepsilon+\Delta_\ell}{1-p_\ell}\right)^{(p_\ell T_\ell-K)/p_\ell}.\end{aligned}\tag{7}$$

We will now proceed to lower bound the right-hand side of Eq. (7) for histories under which event S_ℓ occurs. If event S_ℓ has occurred, then A_ℓ has occurred, and we have $K \le T_\ell \le 4t^*$, so that for every $\ell \in N(\varepsilon, p)$, we have

$$\begin{aligned}\left(1-\left(\frac{\varepsilon+\Delta_\ell}{p_\ell}\right)^2\right)^K &\ge \left(1-\left(\frac{\varepsilon+\Delta_\ell}{p_\ell}\right)^2\right)^{4t^*_\ell}\\ &= \left(1-\left(\frac{\varepsilon+\Delta_\ell}{p_\ell}\right)^2\right)^{(4/c\Delta_\ell^2)\log(1/\theta)}\\ &\overset{a}{\ge} \exp\left\{-d\frac{4}{c}\left(\frac{(\varepsilon/\Delta_\ell)+1}{p_\ell}\right)^2\log(1/\theta)\right\}\\ &\overset{b}{\ge} \exp\left\{-d\frac{16}{c{p_\ell}^2}\log(1/\theta)\right\}\\ &= \theta^{16d/p_\ell^2 c}.\end{aligned}$$

In step (a), we have used Lemma 3 which applies because of Eq. (6); in step (b), we used the fact $\varepsilon/\Delta_\ell \le 1$, which holds because $\ell \in N(\varepsilon, p)$.

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

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

In step (a), we have again used Lemma 3; in step (b), we used the fact $\Delta_\ell/\varepsilon \leq 1$, which holds because $\ell \in M(\varepsilon, p)$.

We now bound the product of the second and third terms in Eq. (7).

If $b \geq 1$, then the mapping $y \mapsto (1-y)^b$ is convex for $y \in [0,1]$. Thus, $(1-y)^b \geq 1-by$, which implies that

$$\left(1-\frac{\varepsilon+\Delta_\ell}{1-p_\ell}\right)^{(1-p_\ell)/p_\ell} \geq \left(1-\frac{\varepsilon+\Delta_\ell}{p_\ell}\right),$$

so that the product of the second and third terms can be lower bounded by

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

We still need to bound the fourth term of Eq. (7). We start with the case where $\ell \in N(p, \varepsilon)$. We have

$$\begin{aligned}
\left(1-\frac{\varepsilon+\Delta_\ell}{1-p_\ell}\right)^{(p_\ell T_\ell - K)/p_\ell} &\stackrel{a}{\geq} \left(1-\frac{\varepsilon+\Delta_\ell}{1-p_\ell}\right)^{(1/p_\ell)\sqrt{t_\ell^*\log(1/\theta)}} && (8)\\
&\stackrel{b}{=} \left(1-\frac{\varepsilon+\Delta_\ell}{1-p_\ell}\right)^{(1/p_\ell\sqrt{c}\Delta_\ell)\log(1/\theta)} \\
&\stackrel{c}{\geq} \exp\left\{-\frac{d}{\sqrt{c}}\cdot\frac{\varepsilon+\Delta_\ell}{\Delta_\ell(1-p_\ell)p_\ell}\log(1/\theta)\right\} && (9)\\
&\stackrel{d}{\geq} \exp\left\{-\frac{2d}{\sqrt{c}(1-p_\ell)p_\ell}\log(1/\theta)\right\} && (10)\\
&\stackrel{e}{\geq} \exp\left\{-\frac{4d}{\sqrt{c}p_\ell}\log(1/\theta)\right\} \\
&= \theta^{4d/(p_\ell\sqrt{c})}.
\end{aligned}$$

Here, (a) holds because we are assuming that the events A_ℓ and C_ℓ occurred; (b) uses the definition of t_ℓ^* for $\ell \in N(p, \varepsilon)$; (c) follows from Eq. (6) and Lemma 3; (d) follows because $\Delta_\ell > \varepsilon$; and (e) holds because $0 \leq p_\ell \leq 1/2$, which implies that $1/(1-p_\ell) \leq 2$.

Consider now the case where $\ell \in M_0(p,\varepsilon)$. Equation (8) holds for the same reasons as when $\ell \in N(p,\varepsilon)$. The only difference from the above calculation is in step (b), where t^*_ℓ should be replaced with $(1/c\varepsilon^2)\log(1/\theta)$. Then, the right-hand side in Eq. (9) becomes

$$\exp\left\{-\frac{d}{\sqrt{c}}\cdot\frac{\varepsilon+\Delta_\ell}{\varepsilon(1-p_\ell)p_\ell}\log(1/\theta)\right\}.$$

For $\ell \in M_0(p,\varepsilon)$, we have $\Delta_\ell \le \varepsilon$, which implies that $(\varepsilon+\Delta_\ell)/\varepsilon \le 2$, which then leads to the same expression as in Eq. (10). The rest of the derivation is identical. Summarizing the above, we have shown that if $\ell \in M_0(p,\varepsilon)\cup N(p,\varepsilon)$, and event S_ℓ has occurred, then

$$\frac{L_\ell(W)}{L_0(W)} \ge \theta^{(4d/p_\ell\sqrt{c})+(16d/p_\ell^2 c)}.$$

For $\ell \in M_0(p,\varepsilon)\cup N(p,\varepsilon)$, we have $\underline{p} < p_\ell$. We can choose c large enough so that $L_\ell(W)/L_0(W) \ge \theta = 8\delta$; the value of c depends only on the constant $\underline{p}$. Similar to the proof of Theorem 1, we have

$$\frac{L_\ell(W)}{L_0(W)}1_{S_\ell} \ge 8\delta 1_{S_\ell},$$

where 1_{S_ℓ} is the indicator function of the event S_ℓ. It follows that

$$\mathbf{P}_\ell(B^c_\ell) \ge \mathbf{P}_\ell(S_\ell) = \mathbf{E}_\ell[1_{S_\ell}] = \mathbf{E}_0\left[\frac{L_\ell(W)}{L_0(W)}1_{S_\ell}\right] \ge \mathbf{E}_0[8\delta 1_{S_\ell}] = 8\delta\mathbf{P}_0(S_\ell) > \delta,$$

where the last inequality relies on the already established fact $\mathbf{P}_0(S_\ell) > 1/8$. □

Since the policy is (ε,δ)-correct, we must have $\mathbf{P}_\ell(B^c_\ell) \le \delta$, for every ℓ. Lemma 6 then implies that $\mathbf{E}_0[T_\ell] > t^*_\ell$ 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 $|M_0(p,\varepsilon)| \ge (|M(p,\varepsilon)|-1)^+$, to conclude the proof of the theorem. □

Remark 7 A close examination of the proof reveals that the dependence of c_1 on $\underline{p}$ is captured by a requirement of the form $c_1 \le c_2\underline{p}^2$, for some absolute constant c_2. 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 larger, 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. In fact, the dependence of a lower bound on ε cannot be $\Theta(1/\varepsilon^2)$ when $\max_i p_i = 1$. To see this, consider the following policy π. Try each coin $O((1/\varepsilon)\log(n/\delta))$ times. If one of the coins always resulted in heads, select it. Otherwise, use some (ε,δ)-correct policy $\tilde{\pi}$. It can be shown that the policy π is (ε,δ)-correct (for every $p \in [0,1]^n$), and that if $\max_i p_i = 1$, then $\mathbf{E}_p[T] = O((n/\varepsilon)\log(n/\delta))$.

Theorem 8 *Fix an integer $s \geq 2$, and some $\underline{p} \in (0,1/2)$. There exists a positive constant c_1 that depends only on $\underline{p}$ such that for every $\varepsilon \in (0, 2^{-(s+2)})$, every $\delta \in (0, e^{-8}/8)$, every $p \in [0, 1-2^{-s}]^n$, and every (ε,δ)-correct policy, we have*

$$\mathbf{E}_p[T] \geq \frac{c_1}{s\eta^2}\left\{\frac{(|M(\tilde{p},\varepsilon\eta)|-1)^+}{\varepsilon^2} + \sum_{\ell\in N(\tilde{p},\eta\varepsilon)} \frac{1}{(p_* - p_\ell)^2}\right\}\log\frac{1}{8\delta},$$

where $p_* = \max_i p_i$, $\eta = 2^{s+1}/s$, $\tilde{p}$ is the vector with components $\tilde{p}_i = 1-(1-p_i)^{1/s}$ (for $i = 1,2,\ldots,n$), and M and N are as defined in Theorem 5.

Proof Let us fix $s \geq 2$, $\underline{p} \in (0,1/2)$, $\varepsilon \in (0, 2^{-(s+2)})$, and $\delta \in (0, e^{-8}/8)$. 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

$$\mathbf{P}_{\tilde{p}}\Big(\tilde{p}_I > \max_i \tilde{p}_i - \eta\varepsilon\Big) \geq 1-\delta, \qquad \forall\, \tilde{p} \in [0,(1/2)+\eta\varepsilon]^n;$$

(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 5, applied to $\tilde{\pi}$, to obtain a lower bound on the sample complexity of π.

The new policy $\tilde{\pi}$ is specified 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

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

and with a slight abuse of notation, let $f(p) = (f(p_1),\ldots,f(p_n))$, and similarly for $g(\tilde{p})$. With our construction, when policy $\tilde{\pi}$ is faced with a bias vector $\tilde{p}$, 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

$$\mathbf{E}^{\tilde{\pi}}_{\tilde{p}}[\tilde{T}] = s\mathbf{E}^{\pi}_{g(\tilde{p})}[T], \qquad \mathbf{E}^{\tilde{\pi}}_{f(p)}[\tilde{T}] = s\mathbf{E}^{\pi}_{p}[T], \tag{11}$$

where the superscript in the expectation operator indicates the policy being used.

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 $\eta = 2^{s+1}/s$ and the assumption $\varepsilon < 2^{-(s+2)}$, we have $\tilde{p}_* \leq (1/2)+(1/2s)$, from which it follows that

$$p_* \leq 1-\left(\frac{1}{2}-\frac{1}{2s}\right)^s = 1-\frac{1}{2^s}\left(1-\frac{1}{s}\right)^s \leq 1-\frac{1}{2^s}\cdot\frac{1}{4} = 1-2^{-(s+2)}.$$

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

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

Thus, the derivative of the inverse mapping g satisfies

$$g'(\tilde{p}_*) \geq \frac{1}{\eta},$$

which implies, using the concavity of g, that

$$g(\tilde{p}_* - \eta\varepsilon) \leq g(\tilde{p}_*) - g'(\tilde{p}_*)\varepsilon\eta \leq g(\tilde{p}_*) - \varepsilon.$$

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 (the superscript in the probability indicates the policy being used)

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

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 5, 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 only requires the policy to be (ε,δ)-correct for $p \in [0,(1/2)+\varepsilon]^n$. This gives a lower bound on $\mathbf{E}^{\tilde{\pi}}_{\tilde{p}}[\tilde{T}]$ which, using Eq. (11), translates to the claimed lower bound on $\mathbf{E}^{\pi}_{p}[T]$. This lower bound applies whenever $p = g(\tilde{p})$, for some $\tilde{p} \in [0,1/2]^n$, and therefore whenever $p \in [0,1-2^{-s}]^n$.
□

5. The Bayesian Setting

There is another variant of the problem which is of interest. 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 the average of the measures $\mathbf{P}_p$ over the prior distribution of p. We also use $\mathbf{E}$ to denote the expectation with respect to $\mathbf{P}$. We then define a policy to be (ε,δ)-correct, for a particular prior and associated measure $\mathbf{P}$, if

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

We then have the following result.

Theorem 9 *There exist positive constants c_1, c_2, ε_0, and δ_0, such that for every $n \geq 2$ and $\varepsilon \in (0,\varepsilon_0)$, there exists a prior for the n-bandit problem such that for every $\delta \in (0,\delta_0)$, and (ε,δ)-correct policy for this prior, we have*

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

In particular, ε_0 and δ_0 can be taken equal to $1/8$ and $e^{-4}/12$, respectively.

Proof Let $\varepsilon_0 = 1/8$ and $\delta_0 = e^{-4}/12$, and let us fix $\varepsilon \in (0, \varepsilon_0)$ and $\delta \in (0, \delta_0)$. Consider the hypotheses $H_0, \ldots, H_n$, introduced in the proof of Theorem 1. Let the prior probability of H_0 be $1/2$, and the prior probability of H_ℓ be $1/2n$, for $\ell = 1, \ldots, n$. Fix an $(\varepsilon/2, \delta)$-correct policy with respect to this prior, and note that it satisfies

$$\mathbf{E}[T] \geq \frac{1}{2}\mathbf{E}_0[T] \geq \frac{1}{2}\sum_{\ell=1}^{n}\mathbf{E}_0[T_\ell]. \tag{12}$$

Since the policy is $(\varepsilon/2, \delta)$-correct, we have $\mathbf{P}(p_I > \max_\ell p_\ell - (\varepsilon/2)) \geq 1 - \delta$.

As in the proof of Theorem 5, let B_ℓ be the event that the policy eventually selects coin ℓ. We have

$$\frac{1}{2}\mathbf{P}_0(B_0) + \frac{1}{2n}\sum_{\ell=1}^{n}\mathbf{P}_\ell(B_\ell) \geq 1 - \delta,$$

which implies that

$$\frac{1}{2n}\sum_{\ell=1}^{n}\mathbf{P}_\ell(B_0) \leq \delta. \tag{13}$$

Let G be the set of hypotheses $\ell \neq 0$ under which the probability of selecting coin 0 is at most 3δ, i.e.,

$$G = \{\ell : 1 \leq \ell \leq n,\ \mathbf{P}_\ell(B_0) \leq 3\delta\}.$$

From Eq. (13), we obtain

$$\frac{1}{2n}(n - |G|)3\delta < \delta,$$

which implies that $|G| > n/3$. Following the same argument as in the proof of Lemma 4, we obtain that there exists a constant c such that if $\delta' \in (0, e^{-4}/4)$ and $\mathbf{E}_0[T_\ell] \leq (1/c\varepsilon^2)\log(1/4\delta')$, then $\mathbf{P}_\ell(B_0) > \delta'$. By taking $\delta' = 3\delta$ and requiring that $\delta \in (0, e^{-4}/12)$, we see that the inequality $\mathbf{E}_0[T_\ell] \leq (1/c\varepsilon^2)\log(1/12\delta)$ implies that $\mathbf{P}_\ell(B_0) > 3\delta$ (here, c is the same constant as in Lemma 4). But for every $\ell \in G$ we have $\mathbf{P}_\ell(B_0) \leq 3\delta$, and therefore $\mathbf{E}_0[T_\ell] \geq (1/c\varepsilon^2)\log(1/12\delta)$. Then, Eq. (12) implies that

$$\mathbf{E}[T] \geq \frac{1}{2}\sum_{\ell \in G}\mathbf{E}_0[T_\ell] \geq |G|\frac{1}{c\varepsilon^2}\log\frac{1}{12\delta} \geq c_1'\frac{n}{\varepsilon^2}\log\frac{c_2}{\delta},$$

where we have used the fact $|G| > n/3$ in the last inequality.

To conclude, we have shown that there exists constants c_1' and c_2 and a prior for a problem with $n+1$ coins, such that any $(\varepsilon/2, \delta)$-correct policy satisfies $\mathbf{E}[T] \geq (c_1' n/\varepsilon^2)\log(c_2/\delta)$. The result follows by taking a larger constant c_1' (to account for having $n+1$ and not n coins, and ε instead of $\varepsilon/2$). $\square$

6. Regret Bounds

In this section we consider lower bounds on the regret of *any* policy, and show that one can derive the $\Theta(\log t)$ regret bound of Lai and Robbins (1985) using the techniques in this paper. The results of Lai and Robbins (1985) are asymptotic as $t \to \infty$, whereas ours deal with finite times t. Our lower bound has similar dependence in t as the upper bounds given by Auer et al. (2002a) for some natural sampling algorithms. As in Lai and Robbins (1985) and Auer et al. (2002a), we also show that when t is large, the regret depends linearly on the number of coins.

Given a policy, let S_t be the total number of unit rewards ("heads") obtained in the first t time steps. The regret by time t is denoted by R_t, and is defined by

$$R_t = t \max_i p_i - S_t \,.$$

Note that the regret is a random variable that depends on the results of the coin tosses as well as of the randomization carried out by the policy.

Theorem 10 *There exist positive constants c_1, c_2, c_3, c_4, and a constant c_5, such that for every $n \geq 2$, and for every policy, there exists some $p \in [0,1]^n$ such that for all $t \geq 1$,*

$$\mathbf{E}_p[R_t] \geq \min\{c_1 t,\ c_2 n + c_3 t,\ c_4 n(\log t - \log n + c_5)\}. \tag{14}$$

The inequality (14) suggests that there are essentially two regimes for the expected regret. When n is large compared to t, the expected regret is linear in t. When t is large compared to n, the regret behaves like $\log t$, but depends linearly on n.

Proof We will prove a stronger result, by considering the regret in a Bayesian setting. By proving that the expectation with respect to the prior is lower bounded by the right-hand side in Eq. (14), it will follow that the bound also holds for at least one of the hypotheses. Consider the same scenario as in Theorem 1, where we have $n+1$ coins and $n+1$ hypotheses $H_0, H_1, \ldots, H_n$. The prior assigns a probability of $1/2$ to H_0, and a probability of $1/2n$ to each of the hypotheses $H_1, H_2, \ldots, H_n$. Similar to Theorem 1, we will use the notation $\mathbf{E}_\ell$ and $\mathbf{P}_\ell$ to denote expectation and probability when the ℓth hypothesis is true, and $\mathbf{E}$ to denote expectation with respect to the prior.

Let us fix t for the rest of the proof. We define T_ℓ as the number of times coin ℓ is tried in the first t time steps. The expected regret when H_0 is true is

$$\mathbf{E}_0[R_t] = \frac{\varepsilon}{2} \sum_{\ell=1}^{n} \mathbf{E}_0[T_\ell],$$

and the expected regret when H_ℓ ($\ell = 1, \ldots, n$) is true is

$$\mathbf{E}_\ell[R_t] = \frac{\varepsilon}{2} \mathbf{E}_\ell[T_0] + \varepsilon \sum_{i \neq 0,\ell} \mathbf{E}_\ell[T_i],$$

so that the expected (Bayesian) regret is

$$\mathbf{E}[R_t] = \frac{1}{2} \cdot \frac{\varepsilon}{2} \sum_{\ell=1}^{n} \mathbf{E}_0[T_\ell] + \frac{\varepsilon}{2} \cdot \frac{1}{2n} \sum_{\ell=1}^{n} \mathbf{E}_\ell[T_0] + \frac{\varepsilon}{2n} \sum_{\ell=1}^{n} \sum_{i \neq 0,\ell} \mathbf{E}_\ell[T_i]. \tag{15}$$

Let D be the event that coin 0 is tried at least $t/2$ times, i.e.,

$$D = \{T_0 \geq t/2\} \,.$$

We consider separately the two cases $\mathbf{P}_0(D) < 3/4$ and $\mathbf{P}_0(D) \geq 3/4$. Suppose first that $\mathbf{P}_0(D) < 3/4$. In that case, $\mathbf{E}_0[T_0] < 7t/8$, so that $\sum_{\ell=1}^{n} \mathbf{E}_0[T_\ell] \geq t/8$. Substituting in Eq. (15), we obtain $\mathbf{E}[R_t] \geq \varepsilon t/32$. This gives the first term in the right-hand side of Eq. (14), with $c_1 = \varepsilon/32$.

We assume from now on that $\mathbf{P}_0(D) \geq 3/4$. Rearranging Eq. (15), and omitting the third term, we have

$$\mathbf{E}[R_t] \geq \frac{\varepsilon}{4}\sum_{\ell=1}^{n}\left(\mathbf{E}_0[T_\ell] + \frac{1}{n}\mathbf{E}_\ell[T_0]\right).$$

Since $\mathbf{E}_\ell[T_0] \geq (t/2)\mathbf{P}_\ell(D)$, we have

$$\mathbf{E}[R_t] \geq \frac{\varepsilon}{4}\sum_{\ell=1}^{n}\left(\mathbf{E}_0[T_\ell] + \frac{t}{2n}\mathbf{P}_\ell(D)\right). \tag{16}$$

For every $\ell \neq 0$, let us define δ_ℓ by

$$\mathbf{E}_0[T_\ell] = \frac{1}{c\varepsilon^2}\log\frac{1}{4\delta_\ell}.$$

(Such a δ_ℓ exists because of the monotonicity of the mapping $x \mapsto \log(1/x)$.) Let $\delta_0 = e^{-4}/4$. If $\delta_\ell < \delta_0$, we argue exactly as in Lemma 4, except that the event B in that lemma is replaced by event D. Since $\mathbf{P}_0(D) \geq 3/4$, the same proof applies and shows that $\mathbf{P}_\ell(D) \geq \delta_\ell$, so that

$$\mathbf{E}_0[T_\ell] + \frac{t}{2n}\mathbf{P}_\ell(D) \geq \frac{1}{c\varepsilon^2}\log\frac{1}{4\delta_\ell} + \frac{t}{2n}\delta_\ell.$$

If on the other hand, $\delta_\ell \geq \delta_0$, then $\mathbf{E}_0[T_\ell] \leq (1/c\varepsilon^2)\log(1/4\delta_0)$, which implies (by the earlier analogy with Lemma 4) that $\mathbf{P}_\ell(D) \geq \delta_0$, and

$$\mathbf{E}_0[T_\ell] + \frac{t}{2n}\mathbf{P}_\ell(D) \geq \frac{1}{c\varepsilon^2}\log\frac{1}{4\delta_\ell} + \frac{t}{2n}\delta_0.$$

Using the above bounds in Eq. (16), we obtain

$$\mathbf{E}[R_t] \geq \frac{\varepsilon}{4}\sum_{\ell=1}^{n}\left(\frac{1}{c\varepsilon^2}\log\frac{1}{4\delta_\ell} + h(\delta_\ell)\frac{t}{2n}\right), \tag{17}$$

where $h(\delta) = \delta$ if $\delta < \delta_0$, and $h(\delta) = \delta_0$ otherwise. We can now view the δ_ℓ as free parameters, and conclude that $\mathbf{E}[R_t]$ is lower bounded by the minimum of the right-hand side of Eq. (17), over all δ_ℓ. When optimizing, all the δ_ℓ will be set to the same value. The minimizing value can be δ_0, in which case we have

$$\mathbf{E}[R_t] \geq \frac{n}{4c\varepsilon}\log\frac{1}{4\delta_0} + \delta_0\frac{\varepsilon}{8}t.$$

Otherwise, the minimizing value is $\delta_\ell = n/2ct\varepsilon^2$, in which case we have

$$\mathbf{E}[R_t] \geq \left(\frac{1}{16c\varepsilon} + \frac{1}{4c\varepsilon}\log(c\varepsilon^2/2)\right)n + \frac{1}{4c\varepsilon}n\log(1/n) + \frac{n}{4c\varepsilon}\log t.$$

Thus, the theorem holds with $c_2 = (1/4c\varepsilon)\log(1/4\delta_0)$, $c_3 = \delta_0\varepsilon/8$, $c_4 = 1/4c\varepsilon$, and $c_5 = (1/4) + \log(c\varepsilon^2/2)$. $\square$

7. Permutations

We now consider the case where the coin 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 $\{1,\ldots,n\}$, and where $q \circ \sigma$ stands for permuting the components of the vector q according to σ, i.e., $(q \circ \sigma)_\ell = q_{\sigma(\ell)}$. We say that a policy is (q,ε,δ)-correct if the coin I eventually selected satisfies

$$\mathbf{P}_{q\circ\sigma}\Big(p_I > \max_\ell q_\ell - \varepsilon\Big) \geq 1-\delta,$$

for every permutation σ of the set $\{1,\ldots,n\}$. We start with a $O\big((n+\log(1/\delta))/\varepsilon^2\big)$ upper bound on the expected number of trials, which is significantly smaller than the bound obtained when the coin biases are completely unknown (cf. Sections 3 and 4). We also provide a lower bound which is within a constant factor of our upper bound.

We then consider a different measure of sample complexity: instead of the expected number of trials, we consider the maximum (over all sample paths) number of trials. We show that for every (q,ε,δ)-correct policy, there is a $\Theta((n/\varepsilon^2)\log(1/\delta))$ lower bound on the maximum number of trials. We note that in the median elimination algorithm of Even-Dar et al. (2002), the length of all sample paths is the same and within a constant factor from our lower bound. Hence our bound is again tight.

We therefore see that for the permutation case, the sample complexity depends critically on whether our criterion involves the expected or maximum number of trials. This is in contrast to the general case considered in Section 3: the lower bound in that section applies under both criteria, as does the matching upper bound from Even-Dar et al. (2002).

7.1 An Upper Bound on the Expected Number of Trials

Suppose we are given a vector $q \in [0,1]^n$, and we are told that the true vector p of coin biases is a permutation of q. The policy in Table 1 takes as input the accuracy ε, the confidence parameter δ, and the vector q. In fact the policy only needs to know the bias of the best coin, which we denote by $q^* = \max_\ell q_\ell$. The policy also uses an additional parameter $\delta' \in (0,1/2]$.

The following theorem establishes the correctness of the policy, and provides an upper bound on the expected number of trials.

Theorem 11 *For every $\delta' \in (0,1/2]$, $\varepsilon \in (0,1)$, and $\delta \in (0,1)$, the policy in Table 1 is guaranteed to terminate after a finite number of steps, with probability 1, and is (q,ε,δ)-correct. For every permutation σ, the expected number of trials satisfies*

$$\mathbf{E}_{q\circ\sigma}[T] \leq \frac{1}{\varepsilon^2}\left(c_1 n + c_2 \log\frac{1}{\delta}\right),$$

for some positive constants c_1 and c_2 that depend only on δ'.

Proof We start with a useful calculation. Suppose that at iteration k, the median elimination algorithm selects a coin I_k whose true bias is p_{I_k}. Then, using the Hoeffding inequality, we have

$$\mathbf{P}(|\hat{p}_k - p_{I_k}| \geq \varepsilon/3) \leq \exp\{-2(\varepsilon/3)^2 m_k\} \leq \frac{\delta}{2^k}. \tag{18}$$

Input: Accuracy and confidence parameters $\varepsilon \in (0,1)$ and $\delta \in (0,1)$; the bias of the best coin q^*.
Parameter: $\delta' \leq 1/2$.
0. $k = 1$; 1. Run the median elimination algorithm to find a coin I_k whose bias is within $\varepsilon/3$ of q^*, with probability at least $1-\delta'$. 2. Try coin I_k for $m_k = \lceil (9/2\varepsilon^2)\log(2^k/\delta)\rceil$ times. Let $\hat{p}_k$ be the fraction of these trials that result in "heads." 3. If $\hat{p}_k \geq q^* - 2\varepsilon/3$ declare that coin I_k is an ε-optimal coin and terminate. 4. Set $k := k+1$ and go back to Step 1.

Table 1: A policy for finding an ε-optimal coin when the bias of the best coin is known.

Let K be the number of iterations until the policy terminates. Given that $K > k-1$ (i.e., the policy did not terminate in the first $k-1$ iterations), there is probability at least $1-\delta' \geq 1/2$ that $p_{I_k} \geq q^* - (\varepsilon/3)$, in which case, from Eq. (18), there is probability at least $1-(\delta/2^k) \geq 1/2$ that $\hat{p}_k \geq q^* - (2\varepsilon/3)$. Thus, $\mathbf{P}(K > k \mid K > k-1) \leq 1-\eta$, with $\eta = 1/4$. Consequently, the probability that the policy does not terminate by the kth iteration, $\mathbf{P}(K > k)$, is bounded by $(1-\eta)^k$. Thus, the probability that the policy never terminates is bounded above by $(3/4)^k$ for all k, and is therefore 0.

We now bound the expected number of trials. Let c be such that the number of trials in one execution of the median elimination algorithm is bounded by $(cn/\varepsilon^2)\log(1/\delta')$. Then, the number of trials, $t(k)$, during the kth iteration is bounded by $(cn/\varepsilon^2)\log(1/\delta') + m_k$. It follows that the expected total number of trials under our policy is bounded by

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbf{P}(K \geq k)t(k) &\leq \frac{1}{\varepsilon^2}\sum_{k=1}^{\infty}(1-\eta)^{k-1}\Big(cn\log(1/\delta') + (9/2)\log(2^k/\delta) + 1\Big) \\
&= \frac{1}{\varepsilon^2}\sum_{k=1}^{\infty}(1-\eta)^{k-1}\big(cn\log(1/\delta') + (9/2)\log(1/\delta) + (9k/2)\log 2 + 1\big) \\
&\leq \frac{1}{\varepsilon^2}(c_1 n + c_2\log(1/\delta)),
\end{aligned}$$

for some positive constants c_1 and c_2.

We finally argue that the policy is (q,ε,δ)-correct. For the policy to select a coin I with bias $p_I \leq q^* - \varepsilon$, it must be that at some iteration k, a coin I_k with $p_{I_k} \leq q^* - \varepsilon$ was obtained, but $\hat{p}_k$ came out larger than $q^* - 2\varepsilon/3$. From Eq. (18), for any fixed k, the probability of this occurring is bounded by $\delta/2^k$. By the union bound, the probability that $p_I \leq q^* - \varepsilon$ is bounded by $\sum_{k=1}^{\infty}\delta/2^k = \delta$. □

Remark 12 The knowledge of q^* turns out to be significant: it enables the policy to terminate as soon as there is high confidence that a coin has been found whose bias is larger than $q^* - \varepsilon$, without having to check the other coins. A policy of this type would not work for the hypotheses considered in the proofs of Theorems 1 and 5: under those hypotheses, the value of q^* is not a priori known. We note that Theorem 11 disagrees with a lower bound in a preliminary version (Mannor and Tsitsiklis, 2003) of this paper. It turns out that the latter lower bound is only valid under an additional restriction on the set of policies, which will be the subject of Section 7.3.

7.2 A Lower Bound

We now prove that the upper bound in Theorem 11 is tight, within a constant.

Theorem 13 *There exist positive constants c_1, c_2, ε_0, and δ_1, such that for every $n \geq 2$ and $\varepsilon \in (0, \varepsilon_0)$, there exists some $q \in [0,1]^n$, such that for every $\delta \in (0, \delta_1)$ and every (q, ε, δ)-correct policy, there exists some permutation σ such that*

$$\mathbf{E}_{q \circ \sigma}[T] \geq \frac{1}{\varepsilon^2}\left(c_1 n + c_2 \log \frac{1}{\delta}\right).$$

Proof Let $\varepsilon_0 = 1/4$ and let $\delta_1 = \delta_0/5$, where δ_0 is the same constant as in Theorem 5. Let us fix some $n \geq 2$ and $\varepsilon \in (0, \varepsilon_0)$. We will establish the claimed lower bound for

$$q = (0.5 + \varepsilon,\ 0.5 - \varepsilon, \ldots, 0.5 - \varepsilon), \tag{19}$$

and for every $\delta \in (0, \delta_1)$. In fact, it is sufficient to establish a lower bound of the form $(c_2/\varepsilon^2)\log(1/\delta)$ and a lower bound of the form $c_1 n/\varepsilon^2$. We start with the former.

Part I. Let us consider the following three hypothesis testing problems. For each problem, we are interested in a δ-correct policy, i.e., a policy whose probability of error is less than δ under any hypothesis. We will show that a δ-correct policy for the first problem can be used to construct a δ-correct policy for the third problem, with the same sample complexity, and then apply Theorem 5 to obtain a lower bound.

- Π_1: We have two coins and the bias vector is either $(0.5 - \varepsilon,\ 0.5 + \varepsilon)$ or $(0.5 + \varepsilon,\ 0.5 - \varepsilon)$. We wish to determine the best coin. This is a special case of our permutation problem, with $n = 2$.

- Π_2: We have a single coin whose bias is either $0.5 - \varepsilon$ or $0.5 + \varepsilon$, and we wish to determine the bias of the coin.[4]

- Π_3: We have two coins and the bias vector can be $(0.5,\ 0.5 - \varepsilon)$, $(0.5 + \varepsilon,\ 0.5 - \varepsilon)$, or $(0.5, 0.5 + \varepsilon)$. We wish to determine the best coin.

Consider a δ-correct policy for problem Π_1 except that the coin outcomes are encoded as follows. Whenever coin 1 is tried, record the outcome unchanged; whenever coin 2 is tried, record the opposite of the outcome (i.e., record a 0 outcome as a 1, and vice versa). Under the first hypothesis in problem Π_1, every trial (no matter which coin was tried) has probability $0.5 - \varepsilon$ of being equal to 1, and under the second hypothesis has probability $0.5 + \varepsilon$ of being equal to 1. With this encoding, it is apparent that the information provided by a trial of either coin in problem Π_1 is the same as the

4. A lower bound for this problem was provided in Lemma 5.1 from Anthony and Bartlett (1999). However, that bound is only established for policies with an a priori fixed number of trials, whereas our policies allow the number of trials to be determined adaptively, based on observed outcomes.

information provided by a trial of the single coin in problem Π_2. Thus, a δ-correct policy for Π_1 translates to a δ-correct policy for Π_2, with the same sample complexity.

In problem Π_3, note that coin 2 is the best coin if and only if its bias is equal to $0.5+\varepsilon$ (as opposed to $0.5-\varepsilon$). Thus, any δ-correct policy for Π_2 can be applied to the second coin in Π_3, to yield a δ-correct policy for Π_3 with the same sample complexity.

We now observe that problem Π_3 involves a set of three hypotheses, of the form considered in the proof of Theorem 5, for the case of two coins. More specifically, in terms of the notation used in that proof, we have $p=(0.5, 0.5-\varepsilon)$, and $N(p,\varepsilon)=\{2\}$. It follows that the sample complexity of any δ-correct policy for Π_3 is lower bounded by $(c_1/\varepsilon^2)\log(1/8\delta)$, where c_1 is the constant in Theorem 5.[5] Because of the relation between the three problems established earlier, the same lower bound applies to any δ-correct policy for problem Π_1, which is the permutation problem of interest.

We have so far established a lower bound proportional to $(1/\varepsilon^2)/\log(1/\delta)$ for problem Π_1, which is the permutation problem we are interested in, with a q vector of the form (19), for the case $n=2$. Consider now the permutation problem for the q vector in (19), but for general n. If we are given the information that the best coin can only be one of the first two coins, we obtain problem Π_1. In the absence of this side-information, the permutation problem cannot be any easier. This shows that the same lower bound holds for every $n\geq 2$.

Part II: We now continue with the second part of the proof. We will establish a lower bound of the form $c_1 n/\varepsilon^2$ for the permutation problem associated with the bias vector q introduced in Eq. (19), to be referred to as problem Π. The proof involves a reduction of Π to a problem $\tilde{\Pi}$ of the form considered in the proof of Theorem 5.

The problem $\tilde{\Pi}$ involves $n+1$ coins (coins $0,1,\ldots,n$) and the following $n+2$ hypotheses:

$$H_0':\ p_0=0.5, \qquad p_i=0.5-\varepsilon, \text{ for } i\neq 0\,,$$

$$H_0'':\ p_0=0.5+\varepsilon, \qquad p_i=0.5-\varepsilon, \text{ for } i\neq 0\,,$$

and

$$H_\ell:\ p_0=0.5, \qquad p_\ell=0.5+\varepsilon, \qquad p_i=0.5-\varepsilon, \text{ for } i\neq 0,\ell.$$

Note that the best coin is coin 0 under either hypothesis H_0' or H_0'', and the best coin is coin ℓ under H_ℓ, for $\ell\geq 1$. This leads us to define H_0 as the hypothesis that either H_0' or H_0'' is true.

We say that a policy for $\tilde{\Pi}$ is δ-correct if it selects the best coin with probability at least $1-\delta$. We will show that if we have a (q,ε,δ)-correct policy π for Π, with a certain sample complexity, we can construct an $(\varepsilon,5\delta)$-correct policy $\tilde{\pi}$ with a related sample complexity. We will then apply Theorem 5 to lower bound the sample complexity of $\tilde{\pi}$, and finally translate to a lower bound for π.

The idea of the reduction is as follows. In problem $\tilde{\Pi}$, if we knew that H_0 is not true, we would be left with a permutation problem with n coins, to which π could be applied. However, if H_0 is true, the behavior of π is unpredictable. (In particular, π might not terminate, or it might terminate with an arbitrary decision: this is because we are only assuming that π behaves properly when faced with the permutation problem Π.) If H_0 is true, we can replace coin 1 with a coin whose bias is $0.5+\varepsilon$, resulting in the bias vector q, in which case π is guaranteed to work properly. But what if we replace coin 1 as above, but some H_ℓ, $\ell\neq 0,1$, happens to be true? In that case, there will be two coins with bias $0.5+\varepsilon$ and π may misbehave. The solution is to run two processes in parallel, one

5. Although Theorem 5 is stated for (ε,δ)-correct policies, it is clear from the proof that the lower bound applies to any policy that has the desired behavior under all of the hypotheses considered in the proof.

with and one without this modification, in which case one of the two will have certain performance guarantees that we can exploit.

Consider the (q,ε,δ)-correct policy π for problem Π. Let t_π be the maximum (over all permutations σ) expected time until π terminates when the true coin bias vector is $q\circ\sigma$.

We define two more bias vectors that will be used below:

$$q_- = (0.5-\varepsilon,\ldots,0.5-\varepsilon)\,,$$

and

$$q_+ = (0.5+\varepsilon,\ 0.5+\varepsilon,\ 0.5-\varepsilon,\ldots,0.5-\varepsilon)\,.$$

Note that if H_0 is true in problem $\tilde{\Pi}$, and π is applied to coins $1,\ldots,n$, then π will be faced with the bias vector q_-. Also, if H_ℓ, is true in problem $\tilde{\Pi}$, for some $\ell\neq 0,1$, and we modify the bias of coin 1 to $0.5+\varepsilon$, then policy π will be faced with the bias vector q_+.

Let us note for future reference that, as in Eq. (18), if we sample a coin with bias $0.5+\varepsilon$ for $m \stackrel{\triangle}{=} \lceil (1/\varepsilon^2)\log(1/\delta)\rceil$ times, the empirical mean reward is larger than 0.5 with probability at least $1-\delta$. Similarly, if we sample a coin with bias $0.5-\varepsilon$ for m times, the empirical mean reward is smaller than 0.5 with probability at least $1-\delta$. Sampling a specific coin that many times, and comparing the empirical mean to 0.5, will be referred to as "validating" the coin.

We now describe policy $\tilde{\pi}$ for problem $\tilde{\Pi}$. The policy involves two parallel processes $\mathcal{A}$ and $\mathcal{B}$: it alternates between the two processes, and each process samples one coin in each one of its turns. The processes continue to sample the coins alternately until one of them terminates and declares one of the coins as the best coin (or equivalently selects a hypothesis). The parameter k below is set to $k=\lceil 18\log(1/\delta)\rceil$.

$\mathcal{A}$: Apply policy π to coins $1,2,\ldots,n$. If π terminates and selects coin ℓ, validate coin ℓ by sampling it m times. If the empirical mean reward is more than 0.5, then $\mathcal{A}$ terminates and declares coin ℓ as the best coin. If the empirical mean reward is less than or equal to 0.5, then $\mathcal{A}$ terminates and declares coin 0 as the best coin. If π has carried out $\lceil t_\pi/\delta\rceil$ trials, then $\mathcal{A}$ terminates and declares coin 0 as the best coin.

$\mathcal{B}$: Sample coin 1 for m times. If the empirical mean reward is more than 0.5, then $\mathcal{B}$ terminates and declares coin 1 as the best coin. Otherwise, replace coin 1 with another coin whose bias is $0.5+\varepsilon$. Initialize a counter N with $N=0$.

Repeat k times the following:

(a) Pick a random permutation τ (uniformly over the set of permutations).

(b) Run a τ-permuted version of π, to be referred to as $\tau\circ\pi$; that is, whenever π is supposed to sample coin i, $\tau\circ\pi$ samples coin $\tau(i)$ instead.

(c) If $\tau\circ\pi$ terminates and selects coin 1 as the best coin, set $N:=N+1$.

If $N>2k/3$, then $\mathcal{B}$ terminates and declares coin 0 as the best coin. Otherwise, wait until process $\mathcal{A}$ terminates.

We first address the issue of correctness of policy $\tilde{\pi}$. Note that $\tilde{\pi}$ is guaranteed to terminate in finite time, because process $\mathcal{A}$ can only run for a bounded number of steps. We consider separately the following cases:

1. Process $\mathcal{A}$ terminates first, H_0 is true. In this case the true bias vector faced by π is q_- rather than q. An error can happen only if π terminates and the coin erroneously passes the validation test. The probability of this occurring is at most δ. In all other cases (validation fails or the running time exceeds the time limit $\lceil t_\pi/\delta \rceil$), process $\mathcal{A}$ correctly declares H_0 to be the true hypothesis.

2. Process $\mathcal{A}$ terminates first, H_ℓ is true for some $\ell \neq 0$. Process $\mathcal{A}$ does not declare the correct H_ℓ if one of the following events occurs: $\mathcal{A}$ fails to select the best coin (probability at most δ, since π is (q,ε,δ)-correct); or the validation makes an error (probability at most δ); or the time limit is exceeded. By Markov's inequality, the probability that the running time T_π of policy π exceeds the time limit satisfies

$$\mathbf{P}_q(T_\pi \geq t_\pi/\delta) \leq \mathbf{E}_q[T_\pi]\delta/t_\pi \leq t_\pi\delta/t_\pi = \delta.$$

   So, the total the probability of errors that fall within this case is bounded by 3δ.

3. Process $\mathcal{B}$ terminates first, H_0 is true. Note that under H_0, process $\mathcal{B}$ is faced with n coins whose bias vector is q. Process $\mathcal{B}$ does not declare H_0 if one of the following events occurs: the initial validation of coin 1 makes an error (probability at most δ); or in k runs, the permuted versions of policy π make at least $k/3$ errors. Each such run has probability at most δ of making an error (since π is (q,ε,δ)-correct), independently for each run (because we use a random permutation before each run). Using Hoeffding's inequality, the probability of at least $k/3$ errors is bounded by $\exp\{-2k(1/3-\delta)^2\}$. Since $\delta < 1/12$, this probability is at most $e^{-k/8}$. So, the total the probability of errors that fall within this case is bounded by $\delta + e^{-k/8}$.

4. Process $\mathcal{B}$ terminates first, H_ℓ is true for some $\ell > 1$. Process $\mathcal{B}$ does not declare H_ℓ if one of the following events occurs: the initial validation of coin 1 makes an error (probability at most δ); or in k runs, the permuted versions of policy π select coin 1 for $N > 2k/3$ times. Since in each of the k runs the policy is faced with the bias vector q_+, there are no guarantees on its behavior. However, since we use a random permutation before each run, and since there are two coins with bias $0.5+\varepsilon$, namely coins 1 and ℓ, the probability that the permuted version of π selects coin 1 at any given run is bounded by 1/2. Using Hoeffding's inequality, the probability of selecting coin 1 more than $2k/3$ times is bounded by $\exp\{-2k((2/3)-(1/2))^2\} = e^{-k/18}$. So, the total the probability of errors that fall within this case is bounded by $\delta + e^{-k/18}$.

5. Process $\mathcal{B}$ terminates first, H_1 is true. Process $\mathcal{B}$ fails to declare coin 1 only if an error is made in the initial validation step, which happens with probability bounded by δ.

To conclude, using the union bound, when H_0 is true, the probability of error is bounded by $2\delta + e^{-k/8}$; when H_1 is true, it is bounded by 4δ; and when H_ℓ, $\ell > 1$ is true, it is bounded by $4\delta + e^{-k/18}$. Since $k = 18\log(1/\delta)$, we see that the probability of error, under any hypothesis, is bounded by 5δ.

We now examine the sample complexity of policy $\tilde{\pi}$. We consider two cases, depending on which hypothesis is true.

1. H_0 is true. In process $\mathcal{A}$, policy π is faced with the bias vector q_- and has no termination guarantees, unless it reaches the time limit $\lceil t_\pi/\delta \rceil$. As established earlier (case 3), Process $\mathcal{B}$

will terminate after the initial validation of coin 1 (m trials), plus possibly k runs of policy π (expected number of trials kt_π), with probability at least $1-e^{-k/8}$. Otherwise, $\mathcal{B}$ waits until $\mathcal{A}$ terminates (at most $1+t_\pi/\delta$ time). Multiplying everything by a factor of 2 (because the two processes alternate), the expected time until $\tilde{\pi}$ terminates is bounded by

$$2(m+kt_\pi)+2e^{-k/8}(m+t_\pi/\delta+1).$$

2. H_ℓ is true for some $\ell \neq 0$. In this case, process $\mathcal{A}$ terminates after the validation time m and the time it takes for π to run. Thus, the expected time until termination is bounded by $2(m+1+t_\pi)$.

We have constructed an $(\varepsilon, 5\delta)$-correct policy $\tilde{\pi}$ for problem $\tilde{\Pi}$. Using the above derived time bounds, and the definitions of k and m, the expected number of trials, under any hypothesis H_ℓ, is bounded from above by

$$4m+36\left(\log\left(\frac{1}{\delta}\right)+3\right)t_\pi.$$

On the other hand, problem $\tilde{\Pi}$ involves hypotheses of the form considered in the proof of Theorem 5, with $p=(0.5,\ 0.5-\varepsilon,\ldots,0.5-\varepsilon)$, and with $N(p,\varepsilon)=\{1,2,\ldots,n\}$. Thus, the expected number of trials under some hypothesis is bounded below by $(c_1 n/\varepsilon^2)\log(c_2/\delta)$, for some positive constants c_1 and c_2, leading to

$$c_1\frac{n}{\varepsilon^2}\log\frac{c_2}{\delta} \leq 4m+36\left(\log\left(\frac{1}{\delta}\right)+3\right)t_\pi.$$

This translates to a lower bound of the form $t_\pi \geq c_2 n/\varepsilon^2$, for some new constant c_2, and for all n larger than some n_0. But for $n \leq n_0$, we can use the lower bound $(c_2/\varepsilon^2)\log(1/\delta)$ that was derived in the first part of the proof. $\square$

7.3 Pathwise Sample Complexity

The sample complexity of the policy presented in Section 7.1 was measured in term of the *expected* number of trials. Suppose, however, that we are interested in a policy for which the number of trials is always low. Let us say that a policy has a *pathwise sample complexity* of t, if the policy terminates after at most t trials, with probability 1. We note that the median elimination algorithm of Even-Dar et al. (2002) is a (q,ε,δ)-correct policy whose pathwise sample complexity is of the form $(c_1 n/\varepsilon^2)\log(c_2/\delta)$. In this section, we show that at least for a certain q, there is a matching lower bound on the pathwise sample complexity of any (q,ε,δ)-correct policy.

Theorem 14 *There exist positive constants c_1, c_2, ε_0, and δ_1 such that for every $n \geq 2$ and $\varepsilon \in (0,\varepsilon_0)$, there exists some $q \in [0,1]^n$, such that for every $\delta \in (0,\delta_1)$ and every (q,ε,δ)-correct policy π, there exists some permutation σ under which the pathwise sample complexity of π is at least*

$$c_1\frac{n}{\varepsilon^2}\log\left(\frac{c_2}{\delta}\right).$$

Proof The proof uses a reduction similar to the one in the proof of Theorem 13. Let $\varepsilon_0 = 1/8$ and $\delta_1 = \delta_0/2$, where $\delta_0 = e^{-8}/8$ is the constant in Theorem 5. Let q be the same as in the proof of Theorem 13 (cf. Eq. (19)), and consider the associated permutation problem, referred to as problem

Π. Fix some $\delta \in (0, \delta_1)$ and suppose that we have a (q,ε,δ)-correct policy π for problem Π whose pathwise sample complexity is bounded by t_π for every permutation σ. Consider also the problem $\tilde{\Pi}$ introduced in the proof of Theorem 13, involving the hypotheses $H_0, H_1, \ldots, H_n$. We will now use the policy π to construct a policy $\tilde{\pi}$ for problem $\tilde{\Pi}$.

We run π on the coins $1, 2, \ldots, n$. If π terminates at or before time t_π and selects some coin ℓ, we sample coin ℓ for $\lceil (1/\varepsilon^2)\log(1/\delta)\rceil$ times. If the empirical mean reward is larger than 0.5 we declare H_ℓ as the true hypothesis. If the empirical mean reward of coin ℓ is less than or equal to 0.5, or if π does not terminate by time t_π, we declare H_0 as the true hypothesis.

We start by showing correctness. Suppose first that H_0 is true. For the policy $\tilde{\pi}$ to make an incorrect decision, it must be the case that policy π selected some coin and the empirical mean reward of this coin was larger than 1/2; using Hoeffding's inequality, the probability of this event is bounded by δ. Suppose instead that H_ℓ is true for some $\ell \geq 1$. In this case, policy π is guaranteed to terminate within t_π steps. Policy $\tilde{\pi}$ will make an incorrect decision if either policy π makes an incorrect decision (probability at most δ), or if policy π makes a correct decision but the selected coin fails to validate (probability at most δ). It follows that policy $\tilde{\pi}$ is $(q, \varepsilon, 2\delta)$-correct.

The number of trials under policy $\tilde{\pi}$ is bounded by $t' = t_\pi + \lceil (1/\varepsilon^2)\log(1/\delta)\rceil$, under any hypothesis. On the other hand, using Theorem 5, the expected number of trials under some hypothesis is bounded below by $(c_1 n/\varepsilon^2)\log(c_2/2\delta)$, leading to

$$c_1 \frac{n}{\varepsilon^2} \log \frac{c_2}{2\delta} \leq t_\pi + \frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right).$$

this translates to a lower bound of the form $t \geq (c_1 n/\varepsilon^2)\log(c_2/\delta)$, for some new constants c_1 and c_2. $\square$

8. Concluding Remarks

We have provided bounds on the number of trials required to identify a near-optimal arm in a multi-armed bandit problem, with high probability. For the problem formulations studied in Sections 3 and 5, the lower bounds match the existing $O((n/\varepsilon^2)\log(1/\delta))$ upper bounds. For the case where the values of the biases are known but the identities of the coins are not, we provided two different tight bounds, depending on the particular criterion being used (expected versus maximum number of trials). 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 Bernoulli rewards as a special case. 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 Kullback-Leibler divergence between the reward distributions associated with different arms is finite (as in Lai and Robbins, 1985).

Acknowledgments

We would like to thank Susan Murphy and David Siegmund for pointing out some relevant references from the sequential analysis literature. We thank two anonymous reviewers for their comments. This research was supported by the MIT-Merrill Lynch partnership, the ARO under grant DAAD10-00-1-0466, and the National Science Foundation under grant ECS-0312921.

References

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002a.

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.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32:48–77, 2002b.

D.A. Berry and B. Fristedt. *Bandit Problems*. Chapman and Hall, 1985.

H. Chernoff. *Sequential Analysis and Optimal Design*. Society for Industrial & Applied Mathematics, 1972.

E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In J. Kivinen and R. H. Sloan, editors, *Fifteenth Annual Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

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*, volume 3, pages 55–86. Academic Press, 1982.

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.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

S. Mannor and J.N. Tsitsiklis. Lower bounds on the sample complexity of exploration in the multi-armed bandit problem. In B. Schölkopf and M. K. Warmuth, editors, *Sixteenth Annual Conference on Computational Learning Theory*, pages 418–432. Springer, 2003.

H. Robbins. Some aspects of sequential design of experiments. *Bulletin of the American Mathematical Society*, 55:527–535, 1952.

S. M. Ross. *Stochastic Processes*. Wiley, 1983.

D. Siegmund. *Sequential analysis: Tests and Confidence Intervals*. Springer Verlag, 1985.