Copyright © 2002 IFAC
15th Triennial World Congress, Barcelona, Spain

MULTIARMED BANDITS IN THE WORST CASE

Nicolò Cesa-Bianchi

Dept. of Information Technologies
University of Milan, Italy
cesa-bianchi@dti.unimi.it

Abstract: We present a survey of results on a recently formulated variant of the classical (stochastic) multiarmed bandit problem in which no assumption is made on the mechanism generating the rewards. We describe randomized allocation policies for this variant and prove bounds on their regret as a function of the time horizon and the number of arms. These bounds hold for any assignment of rewards to the arms and are tight to within logarithmic factors.

1. INTRODUCTION

The multiarmed bandit (Robbins, 1952; Berry and Fristedt, 1985; Presman *et al.*, 1990) is a stochastic adaptive control problem in which the goal is to maximize the return $X_{i_1,1} + X_{i_2,2} + \ldots$, where $X_{i_t,t} \in \mathbb{R}$ is the reward at time t associated with the control $i_t \in \{1,\ldots,N\}$, and $N > 1$ is a fixed parameter. In the classical formulation of this problem, the controller is gambler, who repeatedly pulls the arms of a N-armed slot machine. For each arm $i \in \{1,\ldots,N\}$, the rewards $X_{i,1}, X_{i,2}, \ldots$ are assumed to be i.i.d. random variables with unknown distribution (rewards are also assumed independent across i). The gambler's goal is to maximize his return by pulling, as often as possible, the arm with the highest reward expectation. The strategy used by the gambler to choose which arm to pull next based on past observed rewards is called an *allocation policy*. We will denote with $I_1, I_2, \ldots$ the sequence of arms pulled by a given allocation policy.

The essence of the bandit problem lies in the need of balancing, as accurately as possible, exploitation with exploration. Exploitation corresponds to pulling the arm with the highest reward estimate. Exploration corresponds to pulling other arms in order to reveal, by refining current reward estimates, arms with a better reward expectation. Any allocation policy for the bandit problem must somehow deal with this tradeoff.

[1] Partial support from ESPRIT Working Group EP 27150 (NeuroCOLT II) is gratefully acknowledged.

The performance of a policy is measured with respect to a given horizon model. In the finite horizon model the goal is to minimize the policy's expected *regret at horizon T*, defined by

$$\left(\max_{1\le i\le N} \mu_i\right) T - \mathbb{E}\left[\sum_{t=1}^{T} X_{I_t,t}\right] .$$

Here $\mu_1,\ldots,\mu_N$ are the expected rewards of the arms. That is, $\mathbb{E}[X_{i,t}] = \mu_i$ for each i and t. Hence, the regret measures how much the allocation strategy lost, on average, for not playing consistently the arm with the highest reward expectation.

Lai and Robbins were the first ones to show that, under mild assumptions on the reward distributions, the expected regret for the optimal policy must eventually grow logarithmically in the size T of the horizon (Lai and Robbins, 1985). In the same paper, they also give examples of allocation policies achieving, for reward distributions in the exponential family this optimal logarithmic rate. These policies typically work by estimating the reward expectation of each arm via upper confidence bound estimators. Such estimators use the reward sample average biased by the length of its one-sided confidence interval — see also (Agrawal, 1995; Burnetas and Katehakis, 1996; Yakowitz and Lowe, 1991). We now illustrate this technique in the simple case of rewards bounded in the $[0,1]$ interval — see (Auer *et al.*, 2002). Let $\overline{X}_{i,t}$ be the sample average for the reward of arm i at time t, and let $S_{i,t}$ be the number of times arm i was pulled in the first t time

steps. Then, at time $t+1$, the policy pulls the arm k maximizing the index

$$C_{k,t} = \overline{X}_{k,t} + \sqrt{\frac{\alpha \ln t}{S_{k,t}}}$$

where $\alpha > 0$ is a parameter. The idea behind this policy is very simple. For $[0,1]$-valued independent random variables $X_{k,1}, X_{k,2}, \ldots$, Chernoff-Hoeffding bounds (Chernoff, 1952) state that

$$\mathbb{P}\left(\overline{X}_{k,t} + \sqrt{\frac{\ln t}{2S_{k,t}}} < \mu_i\right) \leq \frac{1}{t} .$$

This ensures that the index $C_{i^*,t}$ of the best arm i^* (i.e. such that $\mu_{i^*} \geq \max_j \mu_j$) is smaller than the arm's true expected reward μ_{i^*} with probability at most $1/t$. This is in turn used to show that a nonzero regret at time t occurs only with probability $1/t$. When summed over T trials, this bound yields the desired logarithmic regret. In the next sections, we will see another application of estimators based on upper confidence bounds.

In view of introducing our worst-case bandit model, we now look more in detail at lower bounds on the expected regret. Consider the case of a N-armed bandit problem with Bernoulli rewards. Assign to $N-1$ arms a Bernoulli distribution with parameter $1/2$ and to the remaining arm a Bernoulli distribution with parameter $1/2+\varepsilon$. Now, if we choose ε sufficiently small, i.e. $\varepsilon \approx \sqrt{N/T}$, then *any* allocation policy will suffer a regret of order $\sqrt{TN}$. A proof of this fact, shown in (Auer *et al.*, 2002), is included in the appendix.

The result above states that whenever the reward distributions can be chosen as a function of T, then the best achievable regret is of order $\sqrt{T}$. In the next sections, we will describe allocation policies that achieve this square-root regret also in a nonstochastic bandit setting. In particular, we will describe allocation policies with square-root regret even in the case where the rewards are not i.i.d. random variables, but they are chosen *deterministically* in a totally arbitrary (and thus possibly malicious) way.

2. THE WORST-CASE MODEL

In this section we describe allocation policies with regret bounds that hold for any deterministic assignment of rewards, included the the worst possible assignment of rewards for the policy being considered.

Our worst-case bandit model is parametrized by a finite number $N > 1$ of arms and by an unknown *reward assignment* specifying, for each $1 \leq i \leq N$ and for each $t \geq 1$, the deterministic real reward $x_{i,t}$ obtained by pulling arm i at time t. At each time t, the gambler only knows the rewards $x_{I_1,1}, \ldots, x_{I_{t-1},t-1}$ obtained in the past $t-1$ rounds. After the arm I_t is pulled, the gambler observes the reward $x_{I_t,t}$ according to the underlying reward assignment. We will always assume that each reward belongs to a known and bounded interval of the reals, say $[0,1]$ for simplicity. Other than this restriction on the range, the reward assignment is arbitrary.

We will use $G = x_{I_1,1} + \ldots + x_{I_T,T}$ to denote the return at horizon T of a given allocation policy and $G_{\max}$ to denote the return at horizon T of the best arm, i.e.

$$G_{\max} = \max_{1 \leq i \leq N} \sum_{t=1}^{T} x_{i,t} .$$

As our allocation policies are randomized, they induce a probability distribution over the set of all arm sequences $(i_1, i_2, \ldots)$. We will use $\mathbb{E}[G]$ to denote the expected return of such a randomized policy, where the expectation is taken with respect to the induced distribution. Our main measure of performance for a policy is the *expected regret* against the best arm, defined by $G_{\max} - \mathbb{E}[G]$.

3. THE BASIC RESULT

In this section we describe our randomized allocation policy `Exp3` and give bounds on its performance. All results from this section are from (Auer *et al.*, 2002). The randomized policy `Exp3` maintains a weight $w_{i,t}$ for each arm $i = 1, \ldots, N$. Initially, the weights are set to 1, i.e. $w_{i,1} = 1$ for all i. At each time $t = 1, 2, \ldots$ an action i_t is drawn according to the distribution $p_{1,t}, \ldots, p_{N,t}$ assigning to arm i probability

$$p_{i,t} = (1-\gamma)\frac{w_{i,t}}{\sum_{j=1}^{N} w_{j,t}} + \frac{\gamma}{N}$$

where $0 < \gamma \leq 1$ is an input parameter. Let $x_{i_t,t}$ be the reward received. Then, the weights are updated as follows: For $j = 1, \ldots, N$ set

$$\widehat{X}_{j,t} = \begin{cases} x_{j,t}/p_{j,t} & \text{if } j = i_t \\ 0 & \text{otherwise,} \end{cases}$$

$$w_{j,t+1} = w_{j,t} \exp(\gamma \widehat{X}_{j,t}/N) . \qquad (1)$$

Note that $\widehat{X}_{i,s}$ is an unbiased estimate of the actual reward $x_{i,s}$. In fact, as one can easily check,

$$\mathbb{E}\left[\widehat{X}_{i,s} \mid I_1, \ldots, I_{s-1}\right] = x_{i,s} \qquad (2)$$

where the expectation is conditioned on the outcomes of the past $s-1$ randomized pulls. Note further that

$$w_{i,t} = \exp(\gamma(\widehat{X}_{i,1} + \ldots + \widehat{X}_{i,t-1})/N)$$

This shows how the probabilities $p_{i,t}$ address the exploration/exploitation trade-off by first assigning to each arm a probability $w_{i,t}/\left(\sum_{j=1}^{N} w_{j,t}\right)$ exponential

in the estimated current return for the arm (exploitation), and then mixing this probability with the uniform distribution $1/N$ over all arms (exploration). The tuning of the mixing coefficient γ will turn out to be crucial. We start the analysis of Exp3 by stating a lower bound on its total expected return that holds for each choice of the parameter γ.

Theorem 1. For any $N > 0$ and for any $\gamma \in (0,1]$, the expected return of algorithm Exp3 satisfies

$$G_{\max} - \mathbb{E}[G] \leq (e-1)\,\gamma G_{\max} + \frac{N \ln N}{\gamma}$$

for any reward assignment and for any $T > 0$.

A suitable tuning of γ reveals that the regret of Exp3 comes close to the lower bound $\Omega\left(\sqrt{TN}\right)$.

Corollary 2. For any $T > 0$, suppose that Exp3 is run with input parameter

$$\gamma = \min\left\{1, \sqrt{(N \ln N)/((e-1)T)}\right\} .$$

Then, for any reward assignment the expected return of algorithm Exp3 satisfies

$$G_{\max} - \mathbb{E}[G] \leq 2\sqrt{e-1}\sqrt{TN \ln N} .$$

PROOF. If $T \leq (N \ln N)/(e-1)$, then the bound is trivial since the expected regret cannot be more than T. Otherwise, by Theorem 1, the expected regret is at most $(e-1)\,\gamma G_{\max} + (N \ln N)/\gamma$. Plugging our choice of γ completes the proof. □

Note that the bound of Corollary 2 implies that the per-round regret of the policy approaches zero at rate bounded by $2\sqrt{e-1}\sqrt{(N \ln N)/T}$. However, this rate bound was obtained via a tuning of the mixing coefficient γ that depends on the horizon T. This horizon-dependent tuning can be avoided (at the expense of a slightly worse leading constant in the rate bound) using a meta-policy that runs Exp3 with $\gamma = \gamma(T_{\text{guess}})$ where T_{guess} grows geometrically (e.g., T_{guess} is doubled whenever the number T of played rounds is larger than T_{guess}). With this trick we obtain a rate bound of order $\sqrt{(N \ln N)/T}$ that holds uniformly over the time horizon T.

We close this section with the proof of the main theorem.

PROOF (of Theorem 1). Choose $T \geq 1$ and let $i_1, \ldots, i_T$ be an arbitrary sequence of actions chosen by Exp3. Let $W_t = w_{1,t} + \ldots + w_{N,t}$. We now compute upper and lower bounds on $\ln(W_{T+1}/W_1)$ and the take expectations over the policy's random choices. For the upper bound, note that $p_{i,t} \geq \gamma/N$ implies $(\gamma/N)\widehat{X}_{i,t} \leq 1$ for all i,t. Thus we get

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i=1}^{N} \frac{w_{i,t}}{W_t} \exp\left(\frac{\gamma}{N}\widehat{X}_{i,t}\right) \\
&= \sum_{i=1}^{N} \frac{p_{i,t} - \gamma/N}{1-\gamma} \exp\left(\frac{\gamma}{N}\widehat{X}_{i,t}\right) \\
&\leq \sum_{i=1}^{N} \frac{p_{i,t} - \gamma/N}{1-\gamma} \left[1 + \frac{\gamma}{N}\widehat{X}_{i,t} + (e-2)\left(\frac{\gamma}{N}\widehat{X}_{i,t}\right)^2\right] \\
&\qquad \text{as } e^z \leq 1 + z + (e-2)z^2 \text{ for } z \leq 1 \\
&\leq 1 + \frac{\gamma/N}{1-\gamma} x_{i_t,t} + \frac{(e-2)(\gamma/N)^2}{1-\gamma} \sum_{i=1}^{N} \widehat{X}_{i,t}
\end{aligned}$$

where in the last step we used

$$\sum_{i=1}^{N} p_{i,t}\widehat{X}_{i,t} = x_{i_t,t} \quad \text{and} \quad p_{j,t}\widehat{X}_{j,t}^2 \leq \widehat{X}_{j,t}$$

for each $j = 1, \ldots, N$. Taking logarithms and using $\ln(1+x) \leq x$ gives

$$\ln \frac{W_{t+1}}{W_t} \leq \frac{\gamma/N}{1-\gamma} x_{i_t,t} + \frac{(e-2)(\gamma/N)^2}{1-\gamma} \sum_{i=1}^{N} \widehat{X}_{i,t} .$$

Summing over t we then get

$$\ln \frac{W_{T+1}}{W_1} \leq \frac{\gamma/N}{1-\gamma} G + \frac{(e-2)(\gamma/N)^2}{1-\gamma} \sum_{t=1}^{T}\sum_{i=1}^{N} \widehat{X}_{i,t} .$$

On the other hand, for any action j,

$$\ln \frac{W_{T+1}}{W_1} \geq \ln \frac{w_{j,T+1}}{W_1} = \frac{\gamma}{N} \sum_{t=1}^{T} \widehat{X}_{j,t} - \ln N .$$

Combining the upper and lower bounds, we get that the return G on the sequence $i_1, \ldots, i_T$ is at least

$$(1-\gamma) \sum_{t=1}^{T} \widehat{X}_{j,t} - \frac{N \ln N}{\gamma} - (e-2)\frac{\gamma}{N} \sum_{t=1}^{T}\sum_{i=1}^{N} \widehat{X}_{i,t} .$$

We now take expectation with respect to $i_1, \ldots, i_T$. Using (2), the expected return of Exp3 is at least

$$(1-\gamma) \sum_{t=1}^{T} x_{j,t} - \frac{N \ln N}{\gamma} - (e-2)\frac{\gamma}{N} \sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t} .$$

where j is chosen arbitrarily. Using

$$\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t} \leq N G_{\max}$$

we get the statement of the theorem. □

4. CONFIDENCE BOUNDS ON THE RETURN

In Section 1 we have shown that the expected regret of algorithm Exp3 after T plays in the N-armed bandit problem is at most order of $\sqrt{TN \ln N}$. In this section we look more closely at the regret distribution. Following (Auer *et al.*, 2002), we would like to argue that the actual regret $G_{\max} - G$ is close to its expected value $G_{\max} - \mathbb{E}[G]$ with high probability. In this respect, algorithm Exp3 is not good. In fact, the variance of each random variable $\widehat{X}_{i,t}$ is about $1/p_{i,t} = 1/\gamma = T^{1/2}$ (ignoring the dependence on N). Over T plays, the

variance of the return is thus $T^{3/2}$ which implies a potential regret of $T^{3/4}$. To fix this problem we replace the estimator $\widehat{X}_{i,t}$ used in Exp3 with the corrected estimator

$$\widehat{X}_{i,t} + \frac{1}{p_{i,t}} \sqrt{\frac{\ln(TN/\delta)}{TN}} .$$

The term added to $\widehat{X}_{i,t}$ plays the role of an upper confidence bound similar to the upper confidence bounds used by the allocation policies for the stochastic bandit problem. Let $\widehat{G}_i$ be sum, over T plays, of these corrected estimates for the rewards of arm i. Then, along the lines of the proof of Corollary 2, we can prove that for any sequence $i_1, \ldots, i_T$ of plays, the modified Exp3 algorithm achieves a return of at least

$$\max_{1 \leq i \leq N} \widehat{G}_i - c\sqrt{TN \ln(TN/\delta)}$$

for some constant $c > 0$. It can then be proven that the return G_i is at most $\widehat{G}_i$ with probability at least $1-\delta$ simultaneously over all arms i. This shows that the actual (as opposed to expected) regret $G_{\max} - G$ is at most $c\sqrt{TN \ln(TN/\delta)}$ with probability at least $1-\delta$ with respect to the policy's randomization.

5. REGRET AGAINST OMNISCIENT POLICIES

So far we have bounded the policy's regret for not always choosing the single globally best arm, i.e. the arm i maximizing $\sum_t x_{i,t}$. More generally, one could also bound the regret for not choosing a particular *sequence* of arms $j^T = (j_1, j_2, \ldots, j_T)$.

Bounding the regret and simultaneously with respect to all arm sequences, with no restrictions, is clearly hopeless. Yet, we can get a result by allowing the regret to scale with a quantity measuring the "hardness" of the sequence. A good definition for hardness of a sequence j^T is $1+n$, where n is the number of times the arm being played must be changed in order to pull the arms in the order given by the sequence j^T.

We start by analyzing the performance of Exp3 with respect to this new criterion. Let $G_{j^T} = x_{j_1,1} + x_{j_2,2} + \ldots + x_{j_T,T}$ be the return at horizon T of an arbitrary sequence $j^T = (j_1, j_2, \ldots, j_T)$. We want to upper bound $G_{j^T} - G$ irrespective to the underlying reward assignment. If the hardness of j^T is S, then we can think of partitioning j^T in S consecutive segments so that the played action does not change within each segment. For example, if $T = 6$ and $j^T = (3,3,1,1,1,1)$, then the hardness of j^T is 2 and we partition it in segments $(3,3)$ and $(1,1,1,1)$. Now we can measure the regret of Exp3 within each segment just as we did in Theorem 1 and then sum up the regrets over the segments. Recall that the main trick in the proof of Theorem 1 was to upper bound the log-ratio of final to initial weight sums. Here, this translates to upper and lower bound the quantity $\ln(W_{T_{k+1}}/W_{T_k})$, where T_k is the play where the k-th segment begun. Whereas the upper bound is derived very much along the lines of Theorem 1, we have some trouble for the lower bound, as the sum W_{T_k} of the weights at the beginning of the segment is unknown. To fix this problem, we slightly alter Exp3 by replacing the weight update step (1) with

$$w_{j,t+1} = w_{j,t} \exp(\gamma \widehat{X}_{j,t}/N) + \frac{\alpha}{N} \sum_{i=1}^{N} w_{i,t}$$

where α is a new input parameter. This amounts to sharing among all weights a fraction α/N of the total weight. Using this trick, we can prove that for each $1 \leq i \leq N$

$$\ln \frac{W_{T_{k+1}}}{W_{T_k}} \geq \ln \frac{\alpha}{N} + \frac{\gamma}{N} G_i(T_k)$$

where $G_i(T_k)$ is the return of arm i during the k-th segment.

This argument leads us to prove a regret bound of the form

$$G_{j^T} - \mathbb{E}[G] \leq O\left(S\sqrt{TN \ln N}\right)$$

where S is the hardness of j^T. This bound holds uniformly over *all* arm sequences j^T. If one wants a result that holds for a set of arm sequences of hardness at most S', then, by tuning α in terms of S', the above bound improves to

$$G_{j^T} - \mathbb{E}[G] \leq O\left(\sqrt{S'TN \ln N}\right) .$$

6. LINEAR EVALUATION FUNCTIONS

In the worst-case bandit problem, the strongest possible notion of regret is

$$\sum_{t=1}^{T} \max_{1 \leq i \leq N} x_{i,t} - \mathbb{E}[G] . \qquad (3)$$

This is the regret for not having played the *best possible* arm sequence, that is the sequence $(j_1, \ldots, j_T)$ where $j_t = \operatorname{argmax}_{1 \leq i \leq N} x_{i,t}$ for $1 \leq t \leq T$. The results of Section 5 imply that a meaningful bound on this regret is achievable at horizon T whenever the reward assignment is such that the hardness of the best possible sequence is small compared to T. We now consider a slight variant of the worst-case bandit problem, called the *linear evaluation function* problem (Long, 1997), in which a bound on (3) can be proven using a different notion of hardness for an arm sequence. In this variant, at the beginning of each time t the policy observes a real *feature vector* $\mathbf{z}_{i,t} \in \mathbb{R}^d$ associated to each arm i, where d is a fixed parameter and $\left\|\mathbf{z}_{i,t}\right\| \leq 1$ for all i,t. These feature vectors provide

additional information that can be used by the policy to estimate the rewards associated to each arm. In particular, the policy can be thought of betting on a linear relationship between a feature vector $\mathbf{z}_{i,t}$ and its corresponding reward $x_{i,t}$. Accordingly, the bound on the regret (3) will scale with the *approximation error*, defined by

$$A = \min_{\mathbf{u}_1,\ldots,\mathbf{u}_N} \sum_{i=1}^{N} \sum_{t=1}^{T} |x_{i,t} - \mathbf{u}_i \cdot \mathbf{z}_{i,t}| \qquad (4)$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_N \in \mathbb{R}^d$ and the minimization has the additional constraints $\|\mathbf{u}_i\| \leq 1$ for $1 \leq i \leq N$. The approximation error accounts, cumulatively over all arms, for the amount of nonlinearity in the relationship between the rewards $x_{i,t}$ and the feature vectors $\mathbf{z}_{i,t}$.

We now describe a randomized policy for this problem. The policy, which we call `Lep` (Linear Evaluation Player), keeps a weight vector $\mathbf{w}_i \in \mathbb{R}^d$ for each arm i. These vectors are used to compute linear estimates $\widetilde{X}_{i,t} = \mathbf{w}_i \cdot \mathbf{z}_{i,t}$ for the rewards associated to the observed feature vectors. Let $\widetilde{X}_t^*$ the largest of these estimates for time t. Each arm i is then drawn with probability $p_{i,t}$ inversely proportional to the difference $\widetilde{X}_t^* - \widetilde{X}_{i,t}$ (with care for the case where i achieves $\widetilde{X}_t^*$). After observing the reward $x_{k,t}$ corresponding to the actually drawn arm k, the weight $\mathbf{w}_{k,t}$ is updated by adding a term proportional to $\mathbf{z}_{k,t}/p_{k,t}$ if $\widetilde{X}_{k,t} < x_{k,t}$ and by subtracting the same term if $\widetilde{X}_{k,t} \geq x_{k,t}$. The purpose of this update, which is similar to the weight updates in the Perceptron (Rosenblatt, 1958) and Widrow-Hoff (Widrow and Hoff, 1960) learning rules, is to learn on-line the best linear approximation for the reward associated to each arm. However, the weight update is not carried out in those rounds t where $|x_{k,t} - \widetilde{X}_{k,t}| \leq \widetilde{X}_t^* - \widetilde{X}_{k,t}$. This helps to trade-off the random choice of an arm k, whose reward estimate $\widetilde{X}_{k,t}$ is significantly smaller than the best estimate $\widetilde{X}_t^*$, with the information gained by oberving a reward $x_{k,t}$ very different from its estimate $\widetilde{X}_{k,t}$.

For any assignment of rewards, the regret bound for `Lep` is of the form

$$\sum_{t=1}^{T} \max_{1\leq i\leq N} x_{i,t} - \mathbb{E}[G] \leq A + c_1 N\sqrt{T} + c_2 (AN^2T)^{2/3}$$

where A is the approximation error (4) and c_1 and c_2 are positive constants.

7. REFERENCES

Agrawal, R. (1995). Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability* **27**, 1054–1078.

Auer, P. (2000). Using upper confidence bounds for online learning. In: *Proceedings of the 41st Annual Symposium on the Foundations of Computer Science*. IEEE Press.

Auer, P., N. Cesa-Bianchi and P. Fischer (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning* **47**(2/3), 235–256.

Auer, P., N. Cesa-Bianchi, Y. Freund and R.E. Schapire (2002). The non-stochastic multi-armed bandit problem. Submitted for Publication. A preliminary version appeared in: *Proceedings of the 36th Annual Symposium on the Foundations of Computer Science*. IEEE press. pp. 322–331.

Berry, D.A. and B. Fristedt (1985). *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall.

Burnetas, A.N. and M.N. Katehakis (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* **17**(2), 122–142.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics* **23**, 493–507.

Cover, T.M. and J.A. Thomas (1991). *Elements of Information Theory*. John Wiley and Sons.

Gittins, J.C. (1989). *Multi-Armed Bandit Allocation Indices*. Wiley-Interscience series in Systems and Optimization. John Wiley and Sons.

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

Long, P.M. (1997). On-line evaluation and prediction using linear functions. In: *Proceedings of the 10th Annual Conference on Computational Learning Theory*. ACM Press. pp. 21–31.

Presman, E.L., I.N. Sonin, E.A. Medova-Dempster and M.A Dempster (1990). *Sequential Control With Incomplete Information : The Bayesian Approach to Multi-Armed Bandit Problems*. Academic Press.

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

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **65**, 386–408.

Widrow, B. and M.E. Hoff (1960). Adaptive switching circuits. In: *1960 IRE WESCON Conv. Record*. pp. 96–104.

Yakowitz, S. and W. Lowe (1991). Nonparametric bandit methods. *Annals of operations Research* **28**, 297–312.

Appendix A. LOWER BOUND

Theorem 3. For any $N > 2$ and for any $T \geq 1$, consider the N-armed bandit problem such that $N-1$ arms have Bernoulli rewards with parameter $1/2$ and one arm has Bernoulli rewards with parameter $(1/2)\left(1+\min\{\sqrt{TN},1\}/2\right)$. Then, the expected regret of any allocation policy for this bandit problem is at least $(1/20)\min\{\sqrt{TN},T\}$.

PROOF. Write the parameter of the best arm as $1/2+\varepsilon$, where $\varepsilon = \min\{\sqrt{TN},1\}/4$, and assume that the best arm is drawn at random from the N arms. We will use K to denote the random index of this best arm.

Fix an arbitrary allocation policy. We will prove that the expected regret of the policy after T pulls, taking also into account the initial randomized choice of the best arm, is at least the bound stated in the theorem. We use $\mathbb{P}$ and $\mathbb{E}$ to denote probabilities and expectations with respect to the sample space including the choice of the best arm in $\{1,\ldots,N\}$ and the choice of the rewards in $\{0,1\}^T$. We also write $\mathbb{P}_i = \mathbb{P}(\cdot \mid K=i)$ and $\mathbb{E}_i = \mathbb{E}[\cdot \mid K=i]$ to denote conditioning on K.

We will compare the probability of certain events computed according to $\mathbb{P}$ with their probability computed according to $\mathbb{P}'$, which uses the same parameter $1/2$ for all arms. Expectations with respect to $\mathbb{P}'$ will be denoted with $\mathbb{E}'$.

Let T_i be the number of times arm i was pulled by the policy in the T rounds. The core of the proof is the observation that a small increase of ε in the Bernoulli parameter of one arm, when all arms have initial parameter $1/2$, does not increase dramatically the expected number of times this arm is pulled by any allocation strategy. To prove this fact note that $T_i \leq T$ and therefore

$$\begin{aligned}
&\mathbb{E}_i[T_i] - \mathbb{E}'[T_i] \leq T \sum_{x^T\in\{0,1\}^T} \left(\mathbb{P}_i(x^T) - \mathbb{P}'(x^T)\right) \\
&\leq T \sum_{x^T\in\{0,1\}^T} \left(\mathbb{P}_i(x^T) - \mathbb{P}'(x^T)\right) \mathbb{I}_{\mathbb{P}_i(x^T)\geq\mathbb{P}'(x^T)} \\
&= \frac{T}{2} \sum_{x^T\in\{0,1\}^T} \left|\mathbb{P}_i(x^T) - \mathbb{P}'(x^T)\right| \\
&= \frac{T}{2} \left\|\mathbb{P}_i - \mathbb{P}'\right\|_1 \leq \frac{T}{\sqrt{2}} \sqrt{D(\mathbb{P}'\,\|\,\mathbb{P}_i)}
\end{aligned}$$

where in the last step we used a standard inequality between the variational distance $\|\cdot\|_1$ and the Kullback-Leibler distance

$$D\big(\mathbb{P}'\,\|\,\mathbb{P}_i\big) = \sum_{x^T\in\{0,1\}^T} \mathbb{P}'(x^T) \ln \frac{\mathbb{P}'(x^T)}{\mathbb{P}_i(x^T)} \,.$$

We now apply the chain rule for the Kullback-Leibler distance — see, e.g., Theorem 2.5.3 (Cover and Thomas, 1991). As usual, let I_t the index of the arm pulled at time t and $X_{I_t,t}$ the reward obtained. Let also $X^t = (X_{I_1,1},\ldots,X_{I_t,t})$. We have

$$\begin{aligned}
&D\big(\mathbb{P}'\,\|\,\mathbb{P}_i\big) \\
&= \sum_{t=1}^T D\Big(\mathbb{P}'(X_{I_t,t} \mid X^{t-1}) \,\|\, \mathbb{P}_i(X_{I_t,t} \mid X^{t-1})\Big) \\
&= \sum_{t=1}^T \mathbb{P}'(I_t \neq i)\, D(1/2\,\|\,1/2) \\
&\quad + \sum_{t=1}^T \mathbb{P}'(I_t = i)\, D(1/2\,\|\,1/2+\varepsilon) \\
&= \sum_{t=1}^T \mathbb{P}'(I_t = i) \left(\frac{1}{2}\ln\frac{1}{1-4\varepsilon^2}\right) \\
&= \mathbb{E}'[T_i] \left(\frac{1}{2}\ln\frac{1}{1-4\varepsilon^2}\right) \,.
\end{aligned}$$

Hence,

$$\mathbb{E}_i[T_i] \leq \mathbb{E}'[T_i] + \frac{T}{2}\sqrt{\mathbb{E}'[T_i]\ln\frac{1}{1-4\varepsilon^2}} \,.$$

Note also that, by definition of reward distributions, $\mathbb{E}_i[G] = T/2 + \varepsilon\,\mathbb{E}_i[T_i]$. We are now ready to compute an upper bound on the expected return $\mathbb{E}[G]$ of the allocation policy,

$$\begin{aligned}
\mathbb{E}[G] &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}_i[G] \\
&\leq \frac{T}{2} + \frac{\varepsilon}{N}\sum_{i=1}^N \left(\mathbb{E}'[T_i] + \frac{T}{2}\sqrt{\mathbb{E}'[T_i]\ln\frac{1}{1-4\varepsilon^2}}\right) \\
&\leq \frac{T}{2} + \varepsilon\left(\frac{T}{N} + \frac{T}{2}\sqrt{\frac{T}{N}\ln\frac{1}{1-4\varepsilon^2}}\right)
\end{aligned}$$

where we used the facts:

$$\sum_{i=1}^N \mathbb{E}'[T_i] = T \quad \text{and} \quad \frac{1}{N}\sum_{i=1}^N \sqrt{\mathbb{E}'[T_i]} \leq \sqrt{\frac{T}{N}} \,.$$

To get a lower bound on the expected regret it is now enough to observe that

$$\max_{1\leq j\leq N} \mathbb{E}[G_j] = \frac{T}{2} + \varepsilon T$$

and, therefore,

$$\begin{aligned}
&\max_{1\leq j\leq N} \mathbb{E}[G_j] - \mathbb{E}[G] \\
&= \varepsilon\left(T - \frac{T}{N} - \frac{T}{2}\sqrt{\frac{T}{N}\ln\frac{1}{1-4\varepsilon^2}}\right) \,.
\end{aligned}$$

Replacing our choice for ε and using $-\ln(1-x) \leq 4x\ln(4/3)$ for $0 \leq x \leq 1/4$ yields the statement of the theorem.