MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

An Asymptotically Optimal Policy for Finite Support Models in the Multiarmed Bandit Problem

Junya HONDA and Akimichi TAKEMURA

METR 2009–19 May 2009

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

The METR technical reports are published as a means to ensure timely dissemination of scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the authors or by other copyright holders, notwithstanding that they have offered their works here electronically. It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may not be reposted without the explicit permission of the copyright holder.

An Asymptotically Optimal Policy for Finite Support Models in the Multiarmed Bandit Problem

Junya HONDA and Akimichi TAKEMURA

Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo
{honda,takemura}@stat.t.u-tokyo.ac.jp

May, 2009

Abstract

We propose minimum empirical divergence (MED) policy for the multiarmed bandit problem. We prove asymptotic optimality of the proposed policy for the case of finite support models. In our setting, Burnetas and Katehakis [3] has already proposed an asymptotically optimal policy. For choosing an arm our policy uses a criterion which is dual to the quantity used in [3]. Our criterion is easily computed by a convex optimization technique and has an advantage in practical implementation. We confirm by simulations that MED policy demonstrates good performance in finite time in comparison to other currently popular policies.

1 Introduction

The multiarmed bandit problem is a problem based on an analogy with playing a slot machine with more than one arm or lever. Each arm has a reward distribution and the objective of a gambler is to maximize the collected sum of rewards by choosing an arm to pull for each play. There is a dilemma between exploration and exploitation, namely the gambler can not tell whether an arm is optimal unless he pulls it many times, but it is also a loss to pull an inferior (i.e. non-optimal) arm many times.

We consider K-armed bandit problem with K arms $\Pi_1, \ldots, \Pi_K$. Π_j has a probability distribution F_j with the expected value μ_j and the player receives a reward according to F_j independently in each play. If the expected values are known, it is optimal to always pull the arm with the maximum expected value $\mu^* \equiv \max_j \mu_j$. A policy is an algorithm to choose the next arm to pull based on the past results of plays. This problem is first considered by Robbins [9]. Then Lai and Robbins [8] established a theoretical framework for determining optimal policies, and Burnetas and Katehakis [3] extended their result

to multiparameter or non-parametric models. There are many other extensions for this problem, for example, to the case of non-stationary distributions [5], or to the case of infinite (possibly uncountable) arms [1, 7].

Consider a model $\mathcal{F}$ which contains distributions of all arms. Let $T_j(n)$ denote the number of times that Π_j has been pulled over the first n plays. A policy is *consistent* on model $\mathcal{F}$ if $\mathrm{E}[T_i(n)] = \mathrm{o}(n^a)$ for any inferior arm Π_i and arbitrary $a > 0$.

Lai and Robbins [8] proved the following lower bound for any inferior arm i under consistent policy:

$$T_i(n) \geq \left(\frac{1}{D(F_i||F^*)} + \mathrm{o}(1) \right) \log n$$

with probability tending to one, where F^* is the distribution of the optimal arm and $D(\cdot||\cdot)$ denotes the Kullback-Leibler divergence. Later Burnetas and Katehakis [3] extended this bound to non-parametric models by

$$T_i(n) \geq \left(\frac{1}{\inf_{G \in \mathcal{F}: \mathrm{E}(G) > \mu^*} D(F_i||G)} + \mathrm{o}(1) \right) \log n \tag{1}$$

where $\mathrm{E}(G)$ is the expected value of distribution G. A policy is asymptotically optimal if the expected value of $T_j(n)$ achieves the right-hand side of (1) as $n \to \infty$. In [8] and [3], policies achieving the above bound are also proposed. These policies are based on the notion of *upper confidence bound*. It can be interpreted as the upper confidence limit for the expectation of each arm with the significance level $1/n$.

Although policies based on upper confidence bounds are optimal, upper confidence bounds are often hard to compute in practice. Then, Auer et al. [2] proposed some policies called UCB. UCB can be interpreted as a policy based on an approximate upper confidence bound obtained by normal approximation. They are practical policies for their simple form and fine performance. Especially, "UCB-tuned" is widely used because of its excellent simulation results. However, UCB-tuned has not been analyzed theoretically and it is unknown whether the policy has consistency. Theoretical analyses of other UCB policies have been given, but their coefficients of the logarithmic term do not necessarily achieve the bound (1).

In this paper we propose minimum empirical divergence (MED) policy. We prove the asymptotic optimality of MED for non-parametric models with finite supports. We note that MED itself can be always used without the assumption of finite supports. MED policy is based on the notion of minimum empirical divergence, which is a quantity dual to upper confidence bounds. It can be computed easily by a convex optimization technique. We also demonstrate simulation results of MED policy comparable to UCB policies.

This paper is organized as follows. In section 2, we give definitions used throughout this paper and show the asymptotic bound by [3], which is satisfied by any consistent policy. In section 3, we propose MED policy and prove that it is asymptotically optimal for finite support models. We also discuss practical implementation issues of minimization problem involved in MED. In section 4, some simulation results are shown. We conclude the paper with some remarks in section 5.

2 Preliminaries

In this section we introduce notations of this paper and present the asymptotic bound by Burnetas and Katehakis [3].

Let $\mathcal{F}$ be a family of probability distributions and let $F_j \in \mathcal{F}$ be the distribution of Π_j, $j = 1, \ldots, K$. A set of probability distributions for K arms is denoted by $\boldsymbol{F} \equiv (F_1, \ldots, F_K) \in \mathcal{F}^K \equiv \prod_{j=1}^K \mathcal{F}$. The joint probability and the expected value under $\boldsymbol{F}$ are denoted by $P_{\boldsymbol{F}}[\cdot]$, $\mathrm{E}_{\boldsymbol{F}}[\cdot]$, respectively.

Let $\mathrm{E}(F)$ denote the expected value of the distribution F. The expected value of Π_j is denoted by $\mu_j \equiv \mathrm{E}(F_j)$. We denote the optimal expected value by $\mu^* \equiv \max_j \mu_j$. Then we define a set of probability distributions

$$\begin{aligned} \mathcal{F}_j^{K*} &\equiv \{\boldsymbol{F} \in \mathcal{F}^K : \mu_j > \max_{i \neq j} \mu_i\} \qquad (F_j \text{ is the unique best}) \\ \mathcal{F}_j^{K} &\equiv \{\boldsymbol{F} \in \mathcal{F}^K : \mu_j < \mu^*\} \qquad (F_j \text{ is not best}). \end{aligned}$$

Let j_n be the arm chosen in the n-th play. Then

$$T_j(n) = \sum_{i=1}^{n} \mathbb{I}[j_n = j],$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. For notational convenience we write

$$T'_j(n) = T_j(n-1),$$

which is the number of times the arm j has been pulled prior to the n-th play. Let $\hat{F}_j(n)$ be the empirical distribution of $T'_j(n)$ observations from Π_j and let $\hat{\mu}_j(n) \equiv \mathrm{E}[\hat{F}_j(n)]$ be the empirical mean of $T'_j(n)$ observations from Π_j. $\hat{\mu}^*(n) \equiv \max_j \hat{\mu}_j(n)$ denotes the highest empirical mean after $n-1$ plays. We call j the current best if $\hat{\mu}_j(n) = \hat{\mu}^*(n)$. Note that $\hat{F}_j(n)$ and $\hat{\mu}_j(n)$ depend only on the past $t = T'_j(n)$ observations from Π_j. When we want to make this clear we write $\hat{F}_j(n)$ and $\hat{\mu}_j(n)$ as $\hat{F}_{j|T'_j(n)=t}$ and $\hat{\mu}_{j|T'_j(n)=t}$ or more simply as $\hat{F}_{j|T'=t}$ and $\hat{\mu}_{j|T'=t}$.

Let Ω denote the whole sample space. For an event $A \subset \Omega$, the complement of A is denoted by A^C. The joint probability of two events A and B under $\boldsymbol{F}$ is written as $P_{\boldsymbol{F}}(A \cap B)$. For notational simplicity we often write, e.g., $P_{\boldsymbol{F}}[j_n = j \cap T'_j(n) = t]$ instead of more precise $P_{\boldsymbol{F}}[\{j_n = j\} \cap \{T'_j(n) = t\}]$.

Finally we define an index for $F \in \mathcal{F}$ and $\mu \in \mathbb{R}$

$$D_{\inf}(F, \mu) \equiv \inf_{G \in \mathcal{F} : \mathrm{E}(G) > \mu} D(F||G).$$

$D_{\inf}$ represents how distinguishable F is from distributions having expectations larger than μ. If $\{G \in \mathcal{F} : \mathrm{E}(G) > \mu\}$ is empty, we define $D_{\inf}(F, \mu) \equiv +\infty$. The total variation distance between two distributions F, G is denoted by $\|F - G\|$.

Theorem 2 of [8] gave a lower bound for $\mathrm{E}[T_i(n)]$ for any inferior i when a consistent policy is adopted. However their result was hard to apply for multiparameter models and more general non-parametric models. Later Burnetas and Katehakis [3] extended the bound to general non-parametric models. Their bound is given as follows.

Theorem 1 (Proposition 1 of [3])**.** *Fix a consistent policy and* $\boldsymbol{F} \in \mathcal{F}_i^K$*. If* $0 < D_{\inf}(F_i, \mu^*) < \infty$*, then for any* $\epsilon > 0$

$$\lim_{n\to\infty} P_{\boldsymbol{F}}\left[T_i(n) \geq \frac{(1-\epsilon)\log n}{D_{\inf}(F_i, \mu^*)}\right] = 1. \tag{2}$$

Consequently

$$\liminf_{n\to\infty} \frac{\mathrm{E}_{\boldsymbol{F}}[T_i(n)]}{\log n} \geq \frac{1}{D_{\inf}(F_i, \mu^*)}. \tag{3}$$

3 Asymptotically Optimal Policy for Finite Support Models

In this section we propose a policy which we call the minimum empirical divergence (MED) policy. For the model $\mathcal{F}$ consisting of all distributions over a fixed *finite* set $\mathcal{X}$, we prove in Theorem 2 that the proposed policy achieves the bound given in the previous section. Then, we describe a convex optimization technique to compute $D_{\inf}$ used in the policy.

Before presenting our MED policy, we discuss consequences of the assumption of finite $\mathcal{X}$. The advantage of assuming the finiteness is that we can employ the method of types for large deviation techniques. This enables us to consider all distributions over $\mathcal{X}$. Since $\mathcal{X}$ is finite, without loss of generality, we assume that the maximum value in $\mathcal{X}$ is 0:

$$\max_{x\in\mathcal{X}} x = 0.$$

Our model $\mathcal{F}$ can be identified with the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$ and $\mathcal{F}$ is compact. For each $\mu < 0$, $\{G \in \mathcal{F} : \mathrm{E}(G) \geq \mu\}$ and $\{G \in \mathcal{F} : \mathrm{E}(G) > \mu\}$ are not empty. $\{G \in \mathcal{F} : \mathrm{E}(G) \geq \mu\}$ is a compact subset of $\mathcal{F}$, because $\mathrm{E}(G) = \sum_{x\in\mathcal{X}} x P_G(\{x\})$ is continuous in G. By compactness argument it is then easy to show that $D_{\inf}(F, \mu) = D_{\min}(F, \mu)$ for $\mu < 0$, where

$$D_{\min}(F, \mu) \equiv \min_{G\in\mathcal{F}:\mathrm{E}(G)\geq\mu} D(F||G). \tag{4}$$

Properties of the minimizer G^* of the right-hand side will be discussed in section 3.2.

3.1 Optimality of the Minimum Empirical Divergence Policy

We now introduce our MED policy. In MED an arm is chosen randomly in the following way:

[**Minimum Empirical Divergence Policy**]

Initialization. Pull each arm once.

Loop.

1. For each j compute $\hat{D}_j^*(n) \equiv D_{\min}(\hat{F}_j(n), \hat{\mu}^*(n))$.
2. Choose arm j according to the probability

$$p_j(n) \equiv \frac{\exp(-T_j'(n)\hat{D}_j^*(n))}{\sum_{i=1}^K \exp(-T_i'(n)\hat{D}_i^*(n))}. \tag{5}$$

Note that

$$\frac{1}{K}\exp(-T_j'(n)\hat{D}_j^*(n)) \le p_j(n) \le \exp(-T_j'(n)\hat{D}_j^*(n))$$

holds. In particular,

$$\frac{1}{K} \le p_j(n) \le 1$$

for the currently best j, since $D_j^*(n) = 0$.

Intuitively, $p_j(n)$ for currently not the best arm j corresponds to the maximum likelihood that j is actually the best arm. Therefore in MED an arm j is pulled with probability proportional to this likelihood.

Consider the event that the arm j is pulled only finite number of times. Then there exists some t and n_0 such that $T_j'(n) = t$ for all $n \ge n_0$. In this case $\hat{F}_j(n)$ stays the same for all $n \ge n_0$ and $(1/K)\exp(-t\hat{D}_j^*(n))$ stays bounded away from zero, since $\hat{\mu}_j(n) \le \hat{\mu}^*(n) \le 1$, $\forall n \ge n_0$. Then almost surely the arm j will be eventually pulled. This argument shows that under MED policy, every arm is pulled infinitely often almost surely.

Note that our policy is a randomized policy. Therefore probability statements below on MED also involves this randomization. However for notational simplicity we omit denoting this randomization.

Now we present the main theorem of this paper.

Theorem 2. *Fix $\boldsymbol{F} \in \mathcal{F}_j^{K*}$. Then, for any $i \neq j$, and $\epsilon > 0$*

$$\mathrm{E}_{\boldsymbol{F}}[T_i(N)] \le \frac{1+\epsilon}{D_{\min}(F_i, \mu^*)} \log N,$$

for all sufficiently large N.

Note that we obtain

$$\limsup_{N\to\infty} \frac{\mathrm{E}_{\boldsymbol{F}}[T_i(N)]}{\log N} \le \frac{1}{D_{\min}(F_i, \mu^*)},$$

by dividing both sides by $\log N$, letting $N \to \infty$ and finally letting $\epsilon \downarrow 0$. In view of (3) we see that MED policy is asymptotically optimal.

We give a proof of Theorem 2 in section 3.3.

3.2 Computation of $D_{\min}$ and Properties of the Minimizer

For implementing MED policy it is essential to efficiently compute the minimum empirical divergence $D_{\min}(\hat{F}_j(n), \hat{\mu}^*(n))$ for each play. In this subsection, we clarify the nature of the convex optimization involved in $D_{\min}(\hat{F}_j(n), \hat{\mu}^*(n))$ and show how the minimization can be computed efficiently. In addition, for the proof of Theorem 2, we need to clarify the behavior of $D_{\min}(F, \mu)$ as a function of F and μ.

Denote the finite symbols in $\mathcal{X}$ by $x_1, \ldots, x_M$, i.e. $\mathcal{X} = \{x_1, \ldots, x_M\}$. We assume $x_1 = 0$ and $x_i < 0$ for $i > 1$ without loss of generality. Write $F_i \equiv P_F(\{x_i\})$ for $F \in \mathcal{F}$.

Computation of $D_{\min}(F, \mu)$ is formulated as the following optimization problem for $(G_1, \ldots, G_M)$:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{M} F_i \log \frac{F_i}{G_i} \\ \text{subject to} \quad & -G_i \le 0, \ \forall i, \quad \mu - \sum_{i=1}^{M} x_i G_i \le 0, \quad \sum_{i=1}^{M} G_i = 1, \end{aligned}$$

where we define $0 \log 0 \equiv 0$, $0 \log \frac{0}{0} \equiv 0$, and $\frac{1}{0} \equiv +\infty$.

It is obvious that $G = F$ is the optimal solution with the optimal value 0 when $0 \ge \mathrm{E}(F) \ge \mu$. Also $G = \delta_0$, the point mass at 0, is the unique feasible solution if $\mu = 0$. For $\mu > 0$ the problem is infeasible. Since these cases are trivial, we consider the case $\mathrm{E}(F) < \mu < 0$ in the following.

The objective function is convex and the feasible region is a closed convex set. Moreover, for the case $\mu < 0$, there is at least one point that the objective function is finite. Therefore G^* is an optimal solution if and only if G^* satisfies the KKT conditions [10]. The Lagrangian function is written as

$$\sum_{i=1}^{M} F_i \log \frac{F_i}{G_i} - \sum_{i=1}^{M} \lambda_i G_i - \nu \sum_{i=1}^{M} x_i G_i + \xi \sum_{i=1}^{M} G_i,$$

where $\{\lambda_i\}$, ν and ξ are Lagrange multipliers. Then the KKT conditions are written as

$$\begin{aligned} -\frac{F_i}{G_i} - \lambda_i - x_i \nu + \xi &= 0, \ \forall i \qquad (6) \\ G_i &\ge 0, \ \forall i, \quad \sum_{i=1}^{M} x_i G_i \ge \mu, \quad \sum_{i=1}^{M} G_i = 1, \\ \lambda_i G_i &= 0, \ \forall i \qquad (7) \\ \lambda_i &\ge 0, \ \forall i, \quad \nu \left(\mu - \sum_{i=1}^{M} x_i G_i \right) = 0, \quad \nu \ge 0. \end{aligned}$$

Let $(G_i^*, \lambda_i^*, \nu^*, \xi^*)$ be the values satisfying the KKT conditions. By taking the summation of (6) $\times G_i$ for i, we obtain the following condition

$$-1 - \nu^* \mathrm{E}(G^*) + \xi^* = 0. \qquad (8)$$

Lemma 3. *If $F_i > 0$ then $G_i^* > 0$, $\lambda_i^* = 0$ and*

$$-\frac{F_i}{G_i^*} - x_i\nu^* + \xi^* = 0. \tag{9}$$

Proof. The objective function is $+\infty$ in the case that $F_i > 0$ and $G_i^* = 0$. Therefore $G_i^* > 0$. Then, $\lambda_i^* = 0$ holds from the condition (7). (9) follows easily from (6). □

Lemma 4. *$\nu^* > 0$, $\sum_{i=1}^M x_i G_i^* = \mu^*$ and*

$$-1 - \nu^*\mu + \xi^* = 0. \tag{10}$$

Proof. We argue by contradiction. Assume $\nu^* = 0$. Then $\xi^* = 1$ from (8) and $F_i/G_i^* = \xi^* - \lambda_i^* = 1 - \lambda_i^*$ for each i from (6). If $F_i > 0$ then $\lambda_i^* = 0$ from Lemma 3, and $G_i^* = F_i$ for all i with $F_i > 0$. Now

$$1 = \sum_{i:F_i>0} G_i^* + \sum_{i:F_i=0} G_i^* = \sum_{i:F_i>0} F_i + \sum_{i:F_i=0} G_i^* = 1 + \sum_{i:F_i=0} G_i^*,$$

implies $F = G^*$. However for the case $\mathrm{E}(F) < \mu$ this G^* does not satisfy the KKT conditions. Therefore by contradiction we have shown that $\nu^* > 0$ and $\sum_{i=1}^M x_i G_i^* = \mu$. (10) follows easily from (8). □

Lemma 5. *$F_j = 0$ and $G_j^* > 0$ implies $j = 1$. Furthermore if $F_1 = 0$ and $G_1^* > 0$, then*

$$G_i^* = \begin{cases} \frac{\mu F_i}{x_i} & i \neq 1 \\ 1 - \sum_{i=2}^M \frac{\mu F_i}{x_i} & i = 1. \end{cases}$$

Proof. If $G_j^* > 0$ then $\lambda_j^* = 0$ and therefore

$$\xi^* = x_j\nu^* \tag{11}$$

from (6). From (10) and (11), ξ^*, ν^* are solved as

$$\nu^* = \frac{1}{x_j - \mu}, \qquad \xi^* = \frac{x_j}{x_j - \mu}. \tag{12}$$

To satisfy conditions $\nu^*, \xi^* \geq 0$, $j = 1$ is required.

Now we prove $G_i^* = \frac{\mu}{x_i} F_i$ for $i \neq 1$. It is obvious for the case $F_i = 0$. For the case $F_i > 0$, it is derived from (9) and (12). □

Now we summarize our result on the optimal $G^* = (G_1^*, \ldots, G_M^*)$ in (4). Recall that we assumed $\mathcal{X} = \{x_1, \ldots, x_M\}$ with $x_1 = 0$, $x_i < 0, \forall i \geq 2$, $\mathrm{E}[F] < \mu < 0$. Let

$$\mathrm{E}_F[1/X] = \begin{cases} +\infty & F_1 > 0 \\ \sum_{i=2}^M \frac{F_i}{x_i} & F_1 = 0. \end{cases} \tag{13}$$

Theorem 6. *If $\mu \mathrm{E}_F[1/X] \leq 1$, then the optimal solution G^* is given by*

$$G_i^* = \begin{cases} \frac{\mu F_i}{x_i} & i \neq 1 \\ 1 - \mu \mathrm{E}_F[1/X] & i = 1, \end{cases} \tag{14}$$

and the optimal value $D_{\min}(F, \mu) = D(F||G^*)$ is

$$\left(\sum_{i \neq 1} F_i \log(-x_i) \right) - \log(-\mu). \tag{15}$$

If $\mu \mathrm{E}_F[1/X] > 1$, then the optimal value is expressed as

$$\max_{\nu} \sum_{i=1}^{M} F_i \log(1 - (x_i - \mu)\nu). \tag{16}$$

Note that (16) is a convex optimization problem with one parameter, so it can be computed easily by iterative methods such as Newton's method.

The benefit to replace the original minimization problem with the maximization problem (16) is not only the efficiency of the computation but also an assurance of consistency.

As discussed in section 3.1 $p_j(n)$ in MED corresponds to the maximum likelihood that the currently not the best j is actually the best. If the probability of choosing j is smaller than $p_j(n)$, it requires exponential number of plays in the expectation for j to regain the current best position and consistency is lost. On the other hand, if the probability even slightly larger than p_j, then consistency still holds although the coefficient of logarithmic term may become larger. A feasible solution of (16) is always less than or equal to the optimal value. Therefore, MED with $D_{\min}(F, \mu)$ obtained numerically by Theorem 6 has consistency even in the presence of some numerical error.

Proof of Theorem 6. For the first case it is obvious that G^* in (14) satisfies the KKT conditions from Lemma 5. (15) follows easily from (14).

Now we consider the second case. In this case, if $F_i = 0$ then $G_i^* = 0$. From (9) and (10), the optimal value is expressed as

$$\sum_{i=1}^{M} F_i \log(1 - (x_i - \mu)\nu^*),$$

where ν^* satisfies

$$\sum_{i=1}^{M} \frac{F_i}{1 - (x_i - \mu)\nu^*} = 1, \qquad \sum_{i=1}^{M} \frac{F_i x_i}{1 - (x_i - \mu)\nu^*} = \mu.$$

Now we define $\hat{\nu}$ as ν maximizing

$$\sum_{i=1}^{M} F_i \log(1 - (x_i - \mu)\nu). \tag{17}$$

By differentiation of (17) it can be easily checked that $\hat{\nu} = \nu^*$. □

3.3 A Proof of Theorem 2

In this section we give a proof of Theorem 2. We first prove the following two lemmas on the property of $D_{\min}(F, \mu)$.

Lemma 7. *$D_{\min}(F, \mu)$ is monotonically increasing and continuous in μ, and strictly increasing for $\mu > \mathrm{E}(F)$. Moreover, $D_{\min}(F, \mu)$ is continuous at each $F \in \mathcal{F}$.*

Proof. It is obvious that $D_{\min}$ is monotonically increasing on μ.

Next we prove that $D_{\min}$ is strictly increasing for the case $\mathrm{E}(F) < \mu$. It is obvious for the case $0 = \mu > \mathrm{E}(F)$ because $D_{\min}(F, 0) = +\infty$. When $\mathrm{E}(F) < \mu < 0$, the optimal solution $D_{\min}$ satisfies $\mathrm{E}(G) = \mu$ from Lemma 4. Then, $D_{\min}$ is strictly increasing for this case.

Finally we prove the continuity by the theory of stability [6]. The feasible region $\{G : \mathrm{E}(G) \geq \mu\}$ is continuous in μ. The objective function $D(F||G)$ is lower semicontinuous in F everywhere in the region where the objective function is finite. Furthermore, $D(F||G)$ and $\{G : \mathrm{E}(G) \geq \mu\}$ are continuous in μ and F since they do not depend on μ and F, respectively. From above facts $D_{\min}(F, \mu)$ is continuous in μ, and lower semicontinuous in F, that is, $\liminf_{F' \to F} D_{\min}(F', \mu) \geq D_{\min}(F, \mu)$.

Now we prove $D_{\min}(F, \mu)$ is upper semicontinuous in F, i.e.

$$\limsup_{F' \to F} D_{\min}(F', \mu) \leq D_{\min}(F, \mu).$$

Let G^* be an optimal solution satisfying $D(F||G^*) = D_{\min}(F, \mu)$ and let $F' \in \mathcal{F}$ be an arbitrary distribution which satisfies $||F' - F|| \leq \delta$. Without loss of generality assume $\mathcal{X} \subset [-1, 0]$. Then $\mu \geq -1$. Consider a distribution

$$G_i^{*\prime} \equiv \begin{cases} (1 - \epsilon - \delta S)G_i^* + \delta F_i & i \neq 1 \\ (1 - \epsilon - \delta S)G_i^* + \delta F_i + \epsilon & i = 1 \end{cases}$$

where $\delta F_i \equiv \max\{0, F_i' - F_i\}$, $\delta S \equiv \sum_i \delta F_i \, (\leq \delta)$ and $\epsilon \equiv -(1 + 1/\mu)\delta S \geq 0$. Then

$$\mathrm{E}(G^{*\prime}) \geq (1 - \epsilon - \delta S)\mu - \delta S = \mu.$$

It can be easily proved that $D(F'||G^{*\prime}) \to D(F||G^*)$ as $\delta \to 0$. Therefore

$$\limsup_{F' \to F} D_{\min}(F', \mu) \leq \limsup_{F' \to F} D(F'||G^{*\prime}) = D(F||G^*) = D_{\min}(F, \mu).$$

□

Lemma 8. *For any μ_1, μ_2, μ^* satisfying $\mu_1 < \mu^*$ and $\mu_2 < \mu^*$, it holds that*

$$\inf_{G \in \mathcal{F} : \mathrm{E}(G) \leq \mu_1} \{D_{\min}(G, \mu^*) - D_{\min}(G, \mu_2)\} > 0.$$

Proof. Assume $\inf_{G\in\mathcal{F}:\mathrm{E}(G)\le\mu_1}\{D_{\min}(G,\mu^*)-D_{\min}(G,\mu_2)\}=0$ and take a sequence $\{G^1,G^2,\dots\}$ along which the infimum is approached. Since $\mathcal{F}$ is compact, there is a subsequence $\{G^{n_i}\}$ which converge to a distribution H with $\mathrm{E}(H)\le\mu_1$. On the other hand, $D_{\min}(G,\mu)$ is continuous in G, then

$$\begin{aligned} 0 &= \liminf_{n\to\infty}\left(D_{\min}(G^{n_i},\mu^*)-D_{\min}(G^{n_i},\mu_2)\right) \\ &= D_{\min}(H,\mu^*)-D_{\min}(H,\mu_2), \end{aligned}$$

which contradicts the fact that $D_{\min}(H,\mu)$ is strictly increasing for $\mu>\mu_1$. □

We now begin proving Theorem 2. We define some more notations used in the following proof. We fix $j=1$ and let $L\equiv\{2,\dots,K\}$. Then, $\mu^*=\mu_1$ and $\mu_i<\mu_1$ for $i\in L$. Write $\mu_L^*\equiv\max_{j\in L}\mu_j$ and $\hat{\mu}_L^*(n)\equiv\max_{i\in L}\hat{\mu}_i(n)$. For notational convenience we denote

$$J_n(i)\equiv\{j_n=i\},$$

We simply write $\mathrm{E}[\cdot], P[\cdot]$ as an expectation and a probability under $\boldsymbol{F}$ and the randomization in the policy. Now we define events A_n, B_n, C_n, D_n, E_n as follows:

$$\begin{aligned} A_n &\equiv \{\hat{D}_i^*(n)\ge(1-\epsilon_1)D_i^*\} \\ B_n &\equiv \{\hat{\mu}_1(n)\ge\mu_1-\delta\} \\ C_n &\equiv \{\hat{\mu}_1(n)<\hat{\mu}_L^*(n)<\mu_L^*+\delta\} \\ D_n &\equiv \{\hat{\mu}_1(n)<\hat{\mu}_L^*(n)\ \cap\ \mu_L^*+\delta\le\hat{\mu}_L^*(n)\} \\ E_n &\equiv \{\hat{\mu}_1(n)\ge\hat{\mu}_L^*(n)\ \cap\ \hat{\mu}_1(n)<\mu_1-\delta\} \end{aligned}$$

where $D_i^*\equiv D_{\min}(F_i,\mu^*)$. Note that $C_n\cup D_n=\{\hat{\mu}_1(n)<\hat{\mu}_L^*(n)\}$ and

$$\left(\{\hat{\mu}_1(n)\ge\hat{\mu}_L^*(n)\}\cap B_n\right)\ \cup\ E_n=\{\hat{\mu}_1(n)\ge\hat{\mu}_L^*(n)\}=(C_n\cup D_n)^C.$$

Therefore $B_n\cup C_n\cup D_n\cup E_n=\Omega$ and each $\mathbb{I}[J_n(i)]$ in the sum $T_j(N)=\sum_{n=1}^N\mathbb{I}[J_n(i)]$ is bounded from above by

$$\begin{aligned} \mathbb{I}[J_n(i)] \le\ & \mathbb{I}[J_n(i)\cap A_n]+\mathbb{I}[J_n(i)\cap C_n] \\ & +\mathbb{I}[J_n(i)\cap A_n^C\cap B_n]+\mathbb{I}[J_n(i)\ \cap\ D_n]+\mathbb{I}[J_n(i)\ \cap\ E_n]. \end{aligned} \tag{18}$$

We will show that the sum $\sum_{n=1}^N\mathbb{I}[J_n(i)\cap A_n]$ is the main term of $T_j(N)$ and the expectations of all other sums are $\mathrm{O}(1)$. Terms involving C_n is more difficult than others. In the following five lemmas we bound the expected values of sums of five terms on the right-hand side of (18) in this order.

Lemma 9. *Let $\epsilon>0$ be arbitrary. For sufficiently large N*

$$\mathrm{E}\left[\sum_{n=1}^N\mathbb{I}[J_n(i)\cap A_n]\right]\le\frac{1+\epsilon}{D_{\min}(F_i,\mu^*)}\log N.$$

Proof. By partitioning the event $J_n(i) \cap A_n$ by the value of $T'_i(n) = t$ we have

$$\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap A(n)] = \sum_{n=1}^{N} \sum_{t=0}^{n-1} \mathbb{I}[J_n(i) \cap A(n) \cap T'_i(n) = t] \tag{19}$$

Note that

$$j_n = i \ \cap \ T'_i(n) = t \quad \Leftrightarrow \quad T_i(n) = t+1 \ \cap \ T'_i(n) = t.$$

Therefore the right-hand side of (19) is written as

$$\sum_{n=1}^{N} \sum_{t=0}^{n-1} \mathbb{I}[T_i(n) = t+1 \cap A(n) \cap T'_i(n) = t] = \sum_{t=0}^{N-1} \sum_{n=t+1}^{N} \mathbb{I}[T_i(n) = t+1 \cap A(n) \cap T'_i(n) = t].$$

Now for each t, we note that there is at most one n such that $T'_i(n) = t$ and $T_i(n) = t+1$. This n is the time when the arm i is played for the $(t+1)$-st time. Therefore for each t

$$\sum_{n=t+1}^{N} \mathbb{I}[T_i(n) = t+1 \ \cap \ A(n) \ \cap T'_i(n) = t] \leq 1.$$

It follows that

$$\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap A(n)] \leq \frac{1+\epsilon}{D_i^*} \log N + \sum_{t=\frac{1+\epsilon}{D_i^*} \log N}^{N} \sum_{n=t+1}^{N} \mathbb{I}[J_n(i) \cap A(n) \cap T'_i(n) = t].$$

Taking the expected value we have

$$\begin{aligned}
\mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap A(n)]\right] &\leq \frac{1+\epsilon}{D_i^*} \log N + \sum_{t=\frac{1+\epsilon}{D_i^*} \log N}^{N} \sum_{n=t+1}^{N} P[J_n(i) \cap A(n) \cap T'_i(n) = t] \\
&\leq \frac{1+\epsilon}{D_i^*} \log N + \sum_{t=\frac{1+\epsilon}{D_i^*} \log N}^{N} \sum_{n=t+1}^{N} P[J_n(i) \mid A(n) \cap T'_i(n) = t] \\
&\leq \frac{1+\epsilon}{D_i^*} \log N + \sum_{t=\frac{1+\epsilon}{D_i^*} \log N}^{N} N \exp(-t(1-\epsilon_1)) D_1^*) \\
&\leq \frac{1+\epsilon}{D_i^*} \log N + C_1 N^{-(1+\epsilon)(1-\epsilon_1)+1},
\end{aligned}$$

where C_1 is a constant. By setting ϵ_1 sufficiently small, we have $C_1 N^{-(1+\epsilon)(1-\epsilon_1)+1} = \mathrm{o}(1)$. □

Lemma 10.

$$\mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap C_n]\right] = \mathrm{O}(1).$$

Proof. First, we have

$$\begin{aligned}\sum_{n=1}^{N}\mathbb{I}[J_n(i)\cap C_n] &\leq \sum_{n=1}^{N}\mathbb{I}[j_n\in L\cap C_n]\\ &\leq \sum_{t=1}^{N}\sum_{n=1}^{\infty}\mathbb{I}[j_n\in L\cap T_1'(n)=t\cap C_n]. \end{aligned}\tag{20}$$

From the technique of type [4], it holds for any type $Q\in\mathcal{F}$ that

$$P_{F_1}[\hat{F}_{1|T_1'(n)=t}=Q]\leq\exp(-tD(Q||F_1))\leq\exp(-tD_{\min}(Q,\mu_1)).\tag{21}$$

Let $\boldsymbol{R}=(R_1,\ldots,R_m)$ be the first m integers in $\{n:T_1'(n)=t\cap C_n\}$. $\boldsymbol{R}$ is well defined on the event $\sum_{n=1}^{\infty}\mathbb{I}[j_n\in L\cap T_1'(n)=t\cap C_n]\geq m$. Let $\boldsymbol{r}=(r_1,\ldots,r_m)\in\mathbb{N}^m$ be a realization of $\boldsymbol{R}$. Then, it holds that

$$\begin{aligned}&\left\{\sum_{n=1}^{\infty}\mathbb{I}[j_n\in L\cap T_1'(n)=t\cap C_n]\geq m\right\}\cap\boldsymbol{R}=\boldsymbol{r}\cap\hat{F}_1=Q\\ &\quad\subset\bigcap_{l=1}^{m}\{j_{r_l}\in L\}\cap\boldsymbol{R}=\boldsymbol{r}\cap\hat{F}_1=Q\end{aligned}$$

On the other hand, for any $\boldsymbol{r}$,

$$\begin{aligned}&P\left[\bigcap_{l=1}^{m}\{j_{r_l}\in L\}\cap\boldsymbol{R}=\boldsymbol{r}\cap\hat{F}_1=Q\right]\\ &= P_{F_1}[\hat{F}_1=Q]\prod_{l=1}^{m}P\left[R_l=r_l\;\middle|\;\bigcap_{k=1}^{l-1}\{j_{r_k}\in L\cap R_k=r_k\cap\hat{F}_1=Q\}\right]\\ &\qquad\times P\left[j_{r_l}\in L\;\middle|\;R_l=r_l\cap\bigcap_{k=1}^{l-1}\{j_{r_k}\in L\cap R_k=r_k\cap\hat{F}_1=Q\}\right]\\ &\leq P_{F_1}[\hat{F}_1=Q]\prod_{l=1}^{m}P\left[R_l=r_l\;\middle|\;\bigcap_{k=1}^{l-1}\{j_{r_k}\in L\cap R_k=r_k\cap\hat{F}_1=Q\}\right]\\ &\qquad\times\left(1-\frac{1}{K}\exp(-tD_{\min}(Q,\,\mu_L^*+\delta))\right)\\ &= P_{F_1}[\hat{F}_1=Q]\left(1-\frac{1}{K}\exp(-tD_{\min}(Q,\,\mu_L^*+\delta))\right)^m\\ &\qquad\times\prod_{l=1}^{m}P\left[R_l=r_l\;\middle|\;\bigcap_{k=1}^{l-1}\{j_{r_k}\in L\cap R_k\in r_l\cap\hat{F}_1=Q\}\right].\end{aligned}$$

Therefore, by taking the disjoint union of $\boldsymbol{r}$, we have

$$P\left[\left\{\sum_{n=1}^{\infty}\mathbb{I}[j_n\in L\cap T_1'(n)=t\cap C_n]\geq m\right\}\cap\hat{F}_1=Q\right]$$

$$
\begin{aligned}
&\leq P\left[\bigcap_{l=1}^{m}\{j_{R_l} \in L\} \cap \hat{F}_1 = Q\right] \\
&\leq P[\hat{F}_1 = Q]\left(1 - \frac{1}{K}\exp(-tD_{\min}(Q,\, \mu_L^* + \delta))\right)^m . \qquad (22)
\end{aligned}
$$

From (21) and (22), we have

$$
\begin{aligned}
&P\left[\sum_{n=1}^{\infty}\mathbb{I}[j_n \in L \cap T_1'(n) = t \cap C_n] \geq m\right] \\
&\leq \sum_{Q:\mathrm{E}(Q)\geq\hat{\mu}_L^*}\exp(-tD_{\min}(Q,\mu_1))\left(1 - \frac{1}{K}\exp(-tD_{\min}(Q,\, \mu_L^* + \delta))\right)^m
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\mathrm{E}\left[\sum_{n=1}^{\infty}\mathbb{I}[j_n \in L \cap T_1'(n) = t \cap C_n]\right] \\
&\leq \sum_{m=0}^{\infty}\sum_{Q:\mathrm{E}(Q)\geq\hat{\mu}_L^*}\exp(-tD_{\min}(Q,\mu_1))\left(1 - \frac{1}{K}\exp(-tD_{\min}(Q,\, \mu_L^* + \delta))\right)^m \\
&= K\sum_{Q:\mathrm{E}(Q)\geq\hat{\mu}_L^*}\exp(-tD_{\min}(Q,\mu_1))\exp(tD_{\min}(Q,\, \mu_L^* + \delta)) \\
&\leq K\sum_{Q:\mathrm{E}(Q)\geq\hat{\mu}_L^*}\exp(-t\inf_{Q:\mathrm{E}(Q)\geq\hat{\mu}_L^*}\{D_{\min}(Q,\, \mu_1) - D_{\min}(Q,\, \mu_L^* + \delta)\}). \qquad (23)
\end{aligned}
$$

From Lemma 8,

$$
(23) \leq K\sum_{Q:\mathrm{E}(Q)\geq\hat{\mu}_L^*}\exp(-tC_2), \qquad (24)
$$

where $C_2 > 0$ is a constant.

Since there are at most $(t+1)^{|\mathcal{X}|}$ combinations as a type of t observations, we have

$$
(24) \leq K(t+1)^{|\mathcal{X}|}\exp(-tC_2).
$$

Finally we obtain

$$
(20) \leq \sum_{t=1}^{N}K(t+1)^{|\mathcal{X}|}\exp(-tC_2) = \mathrm{O}(1).
$$

□

In the proofs of remaining three lemmas, we use a theorem on the empirical distribution [4].

Theorem 11 (Sanov's Theorem)**.** *For every set* $\Gamma \subset \mathcal{F}$

$$\limsup_{t\to\infty} \frac{1}{t} \log P_F[\hat{F}_{|T'=t} \in \Gamma] \leq - \inf_{G\in\Gamma} D(G||F).$$

Lemma 12.

$$\mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap A_n^C \cap B_n]\right] = \mathrm{O}(1).$$

Proof. Define $\Gamma_1 \subset \mathcal{F}$ as

$$\Gamma_1 \equiv \{G \in \mathcal{F} : ||G - F_i|| \geq \delta_1\}.$$

Since $\inf_{G\in\Gamma_1} D(G||F_i) > 0$, there exists $C_2 > 0$ such that

$$P_{F_i}[\hat{F}_{i|T'=t} \in \Gamma_1] \leq \exp(-C_2 t)$$

for sufficiently large t. Therefore, it holds that

$$1 - P_{F_i}\left[\bigcap_{l=t}^{\infty} \{\hat{F}_{i|T'=t} \notin \Gamma_1\}\right] = \mathrm{O}(\exp(-C_2 t))).$$

Now we show

$$\{A_n^C \cap B_n\} \subset \{\hat{F}_i(n) \in \Gamma_1\}. \tag{25}$$

If $\hat{F}_i(n) \notin \Gamma_1$ and B_n, then

$$\hat{D}_i(n) = D_{\min}(\hat{F}_i(n), \hat{\mu}^*) \geq D_{\min}(\hat{F}_i(n), \hat{\mu}_1) \geq D_{\min}(\hat{F}_i(n), \mu_1 - \delta)$$

from the monotonicity in μ of $D_{\min}$ in Lemma 7. Since $D_{\min}(F_i, \mu^* - \delta) > 0$, we obtain for sufficiently small δ

$$D_{\min}(\hat{F}_i,\ \mu_1 - \delta) \geq (1 - \epsilon_1/2) D_{\min}(F_i, \mu^* - \delta)$$

from the continuity in F of $D_{\min}$ in Lemma 7. Moreover, from the continuity of $D_{\min}$ in μ, it holds for sufficiently small δ that

$$(1 - \epsilon_1/2) D_{\min}(F_i, \mu^* - \delta) \geq (1 - \epsilon_1) D_{\min}(F_i, \mu^*) = (1 - \epsilon_1) D_i^*.$$

Then A_n holds and (25) is proved.

Since

$$\left\{\sum_{i=1}^{N} \mathbb{I}[j_n(i) \cap A_n^C \cap B_n] \geq t\right\} \subset \left\{\sum_{i=1}^{N} \mathbb{I}[j_n(i) \cap \{\hat{F}_i(n) \in \Gamma_1\}] \geq t\right\}$$

$$\subset \left\{ \bigcap_{l=t}^{\infty} \{\hat{F}_{i|T'=t} \notin \Gamma_1\} \right\}^C,$$

we obtain

$$\begin{aligned} \mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap A_n^C \cap B_n]\right] &= \sum_{t=1}^{N} P\left[\sum_{i=1}^{N} \mathbb{I}[j_n(i) \cap A_n^C \cap B_n] \geq t\right] \\ &\leq \sum_{t=1}^{N} \left\{1 - P\left[\bigcap_{l=t}^{\infty} \{\hat{F}_{i|T'=l} \notin \Gamma_1\}\right]\right\} \\ &= \sum_{t=1}^{N} \mathrm{O}(\exp(-C_2 t)) = \mathrm{O}(1). \end{aligned}$$

□

Lemma 13.

$$\mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap D_n]\right] = \mathrm{O}(1). \tag{26}$$

Proof. First we simply bound $\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap D_n]$ by

$$\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap D_n] \leq \sum_{n=1}^{N} \mathbb{I}[D_n].$$

Since $D_n \subset \bigcup_{k \in L} \{\hat{\mu}_k(n) = \hat{\mu}^*(n) > \mu_L^* + \delta\}$, it holds that

$$\begin{aligned} \sum_{n=1}^{N} \mathbb{I}[D_n] &\leq \sum_{k \in L} \sum_{n=1}^{N} \mathbb{I}[\hat{\mu}_k(n) = \hat{\mu}^*(n) > \mu_L^* + \delta] \\ &= \sum_{k \in L} \sum_{t=1}^{N} \sum_{n=1}^{N} \mathbb{I}[\hat{\mu}_{k|T'=t} = \hat{\mu}^*(n) > \mu_L^* + \delta \cap T_k'(n) = t]. \end{aligned}$$

Note that $P[J_n(k)] \geq 1/K$ whenever $\hat{\mu}_k(n) = \hat{\mu}^*(n) > \mu_L^* + \delta$. Therefore, by the same discussion as (22), we have

$$\mathrm{E}\left[\sum_{n=1}^{\infty} \mathbb{I}[\hat{\mu}_{k|T'=t} = \hat{\mu}^*(n) \cap T_k'(n) = t] \;\middle|\; \hat{\mu}_{k|T'=t} > \mu_L^* + \delta\right] \leq K. \tag{27}$$

On the other hand, it holds from Sanov's theorem that for a constant $C_3 > 0$

$$P_{F_k}[\hat{\mu}_{k|T'=t} > \mu_L^* + \delta] = \mathrm{O}(\exp(-C_3 t)) \tag{28}$$

by setting $\Gamma_2 \equiv \{G \in \mathcal{F} : \mathrm{E}(G) > \mu_L^* + \delta\}$. From (27) and (28), we have

$$\begin{aligned}
&\mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[\hat{\mu}_{k|T'=t}(n) = \hat{\mu}^*(n) > \mu_L^* + \delta \cap T_k'(n) = t]\right] \\
&= P_{F_k}[\hat{\mu}_{k|T'=t} > \mu_L^* + \delta]\ \mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[\hat{\mu}_{k|T'=t}(n) = \hat{\mu}^*(n) \cap T_k'(n) = t] \;\middle|\; \hat{\mu}_{k|T'=t} > \mu_L^* + \delta\right] \\
&= \mathrm{O}(\exp(-C_3 t)).
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[D_n]\right] &\leq \mathrm{E}\left[\sum_{k\in L}\sum_{t=1}^{N}\sum_{n=1}^{\infty} \mathbb{I}[\hat{\mu}_{k|T'=t}(n) = \hat{\mu}^*(n) > \mu_L^* + \delta \cap T_k'(n) = t]\right] \\
&= \sum_{k\in L}\sum_{t=1}^{N} \mathrm{O}(\exp(-C_2 t)) = \mathrm{O}(1).
\end{aligned}$$

□

Lemma 14.

$$\mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap E_n]\right] = \mathrm{O}(1).$$

Proof. $\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap E_n]$ is bounded from above by

$$\begin{aligned}
\sum_{n=1}^{N} \mathbb{I}[E_n] &= \sum_{n=1}^{N} \mathbb{I}[\hat{\mu}_1(n) \geq \hat{\mu}_L^*(n) \cap \hat{\mu}_1(n) < \mu_1 - \delta] \\
&\leq \sum_{t=1}^{N}\sum_{n=1}^{N} \mathbb{I}[\hat{\mu}_{1|T'=t} \geq \hat{\mu}_L^*(n) \cap \hat{\mu}_{1|T'=t} < \mu_1 - \delta \cap T_1'(n) = t].
\end{aligned}$$

By the same argument as in Lemma 13, the expected value of the right-hand side is $\mathrm{O}(1)$. □

We have now completed the proof of Theorem 2.

4 Experiments

In this section, we present some simulation results on our MED and UCB policies.

To compute $D_{\min}(F, \mu)$, the following algorithm is adopted which combines Newton's method and bisection method:

[Computation of $D_{\min}(F, \mu)$]

Require: $r > 0$, $\nu_0 \geq 0$;

if $F_1 = 0$ and $\mu \sum_{i=2}^{M} \frac{F_i}{x_i} \leq 1$ **then**
  return $D := \left(\sum_{i \neq 1} F_i \log(-x_i)\right) - \log(-\mu)$, $\nu := -\frac{1}{\mu}$;
end if
if $\nu_0 < \frac{1}{x_+ - \mu}$ and $h(\nu_0) \geq 0$ **then**
  $\nu := \nu_0$;
else
  $\nu := 0$;
end if
$a := 0$, $b := 1/(x_+ - \mu)$;
for $t := 1$ to r **do**
  if $h'(\nu) > 0$ **then**
    $a := \nu$;
  else
    $b := \nu$;
  end if
  $\nu := \nu - h'(\nu)/h''(\nu)$;
  if $\nu \leq a$ or $b \leq \nu$ **then**
    $\nu := \frac{a+b}{2}$;
  end if
end for
return $D := \max_{\nu' \in \{a,b,\nu\}} h(\nu')$, $\nu := \mathrm{argmax}_{\nu' \in \{a,b,\nu\}} h(\nu')$;

$x_+ \equiv \max_{i:F_i>0} x_i$ is the largest symbol on the support of F. r is a repetition number determined in advance. ν_0 is set to 0 in the first play. After that, ν_0 is set to the output ν in the last play. The output D is the value of $D_{\min}(F, \mu)$ used in our policy. h, h', h'' denote

$$
\begin{aligned}
h(\nu) &\equiv \sum_{i=1}^{M} F_i \log(1 - (x_i - \mu)\nu), \\
h'(\nu) &\equiv -\sum_{i=1}^{M} \frac{x_i - \mu}{1 - (x_i - \mu)\nu} F_i, \\
h''(\nu) &\equiv -\sum_{i=1}^{M} \frac{(x_i - \mu)^2}{(1 - (x_i - \mu)\nu)^2} F_i.
\end{aligned}
$$

In this algorithm, a lower and an upper bound of $\hat{\nu}$ is evaluated by a and b, respectively. In each step, the next point is determined based on Newton's method by $\nu := \nu - h'(\nu)/h''(\nu)$. When ν does not improve the bounds a nor b, the next point is determined by bisection method, $\nu := \frac{a+b}{2}$. The complexity of the algorithm is given by $\mathrm{O}(r\,|\mathcal{X}|)$.

Now we describe the setting of our experiments. We used our MED, UCB-tuned and UCB2. Each plot is an average over 100 different runs. The repetition number r is set to 5 in MED.

We note that there is a slight difference between models considered in MED and UCB. UCB-tuned and UCB2 treat models that the minimum and maximum symbol are assumed

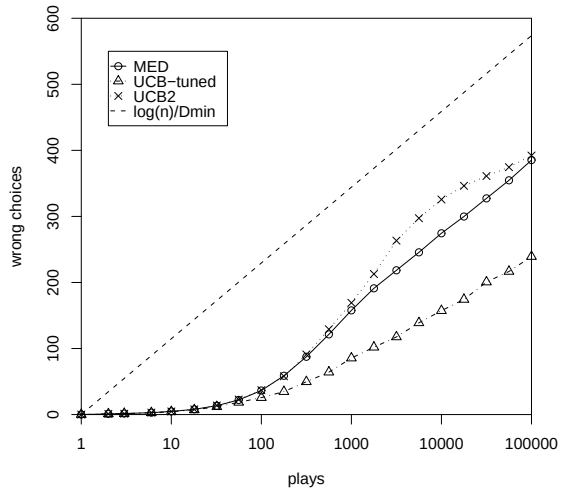

Figure 1: Simulation result for Bernoulli distributions

to be known (in [2], they are set to 0,1). On the other hand, MED treats models that the support is finite and the maximum symbol is assumed to be known. Especially, the maximum symbol is assumed to be 0 in above algorithm for convenience. We use a model in this experiment in which $|\mathcal{X}|$ is finite and $\mathcal{X} \subset [0,1]$ is known. In MED, all symbols are passed to computation after 1 is subtracted from them.

Figure 1 is an experiment for a simple problem that arms have the same support:

$$F_1[X=0] = 0.45, \quad F_1[X=1] = 0.55, \quad \mathrm{E}(F_1) = 0.55,$$
$$F_2[X=0] = 0.55, \quad F_2[X=1] = 0.45, \quad \mathrm{E}(F_2) = 0.45.$$

"wrong choices" denotes $T_j(n)$ s.t. $\mu_j < \mu$, that is, $T_2(n)$ in this model. "log(n)/Dmin" stands for the asymptotic bound for a consistent policy, $\log(n)/D_{\min}(F_2, \mu^*)$. We can see from the figure that UCB-tuned is the best for this model. However, it appears that UCB-tuned may not be consistent, because the asymptotic slope seems to be smaller than $1/D_{\min}(F_2, \mu^*)$. On the other hand, the slope of MED is almost the same as $1/D_{\min}(F_2, \mu^*)$, which coincides with the theoretical evaluation.

Figure 2 is an experiment for another simple problem where uniform distributions have different supports:

$$F_1[X=0.4] = 0.5, \quad F_1[X=0.8] = 0.5, \quad \mathrm{E}(F_1) = 0.6,$$
$$F_2[X=0.2] = 0.5, \quad F_2[X=0.6] = 0.5, \quad \mathrm{E}(F_2) = 0.4.$$

We can see from the figure that MED is the best and seems to be achieving the asymptotic bound.

In Figure 3, we used distributions

$$F_1[X=0.0] = 0.99, \quad F_1[X=1] = 0.01, \quad \mathrm{E}(F_1) = 0.01,$$

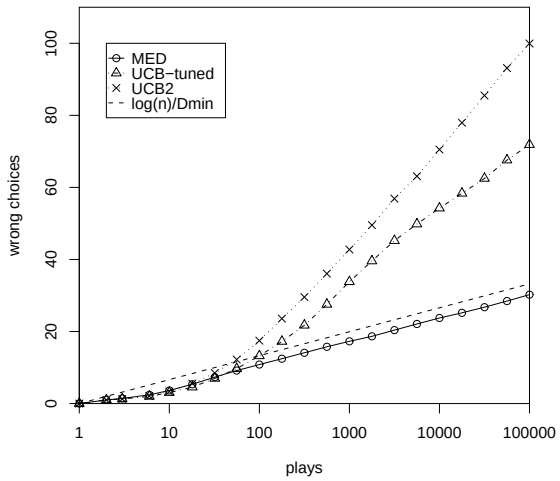

Figure 2: Simulation result for uniform distributions with different supports

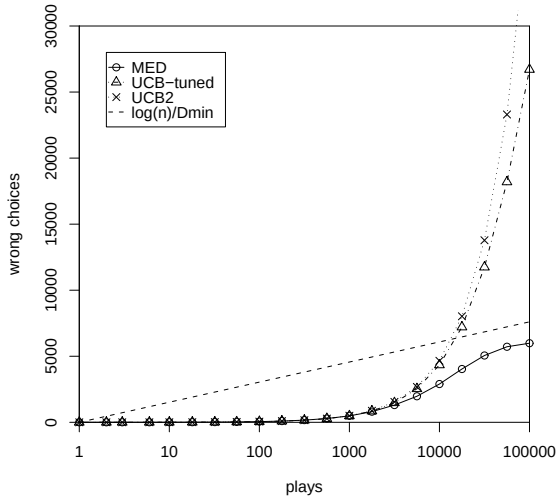

Figure 3: Simulation result for very confusing distributions

$$F_2[X = 0.08] = 0.5, \quad F_2[X = 0.09] = 0.5, \quad \mathrm{E}(F_2) = 0.085.$$

In this setting, it is difficult to distinguish the optimal arm because the inferior arm seems to be optimal at first with high probability. In spite of the difficulty, MED distinguishes the optimal arm quickly and finally converges to the asymptotic bound.

5 Concluding remarks

We proposed a policy, MED, and proved that our policy achieves the asymptotic bound for finite support models. We also showed that our policy can be implemented efficiently by a convex optimization technique.

In the theoretical analysis of this paper, models are assumed to satisfy $|\mathcal{X}| < \infty$ although the assumption is not used in the implementation. Therefore the most important work is to remove the assumption and derive a bound-achieving policy for models with bounded infinite support. In addition, there are many models that $D_{\min}$ can be computed practically, such as normal distribution model with unknown mean and variance. We expect that that our MED can be extended to these models.

References

[1] R. Agrawal, "The continuum-armed bandit problem", *SIAM J. Control and Optimization*, 33:1926–1951, 1995.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multi-armed bandit problem", *Machine Learning*, 47:235–256, 2002.

[3] A. Burnetas and M. Katehakis, "Optimal adaptive policies for sequential allocation problems", *Advances in Applied Mathematics*, 17:122–142, 1996.

[4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., Springer, 1998.

[5] J. C. Gittins, *Multiarmed Bandits Allocation Indices*, Wiley, New York, 1989.

[6] W. W. Hogan, "Point-to-set maps in mathematical programming", *SIAM Review*, 15, pp. 591–603, 1973.

[7] R. Kleinberg, "Nearly tight bounds for the continuum-armed bandit problem", In *NIPS-2004*, 697–704, 2004.

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

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

[10] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1972.