Mach Learn (2011) 85:361–391
DOI 10.1007/s10994-011-5257-4

An asymptotically optimal policy for finite support models in the multiarmed bandit problem

Junya Honda · Akimichi Takemura

Received: 4 June 2009 / Accepted: 5 June 2011 / Published online: 2 July 2011

Abstract In the multiarmed bandit problem the dilemma between exploration and exploitation in reinforcement learning is expressed as a model of a gambler playing a slot machine with multiple arms. A policy chooses an arm in each round so as to minimize the number of times that arms with suboptimal expected rewards are pulled. We propose the minimum empirical divergence (MED) policy and derive an upper bound on the finite-time regret which meets the asymptotic bound for the case of finite support models. In a setting similar to ours, Burnetas and Katehakis have already proposed an asymptotically optimal policy. However, we do not assume any knowledge of the support except for its upper and lower bounds. Furthermore, the criterion for choosing an arm, minimum empirical divergence, can be computed easily by a convex optimization technique. We confirm by simulations that the MED policy demonstrates good performance in finite time in comparison to other currently popular policies.

Keywords Bandit problems · Finite-time regret · MED policy · Convex optimization

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 round. There is a dilemma between exploration and exploitation: the gambler cannot

Editor: Nicolo Cesa-Bianchi.

J. Honda (✉)
Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-shi Chiba 277–8561, Japan
e-mail: honda@stat.t.u-tokyo.ac.jp

A. Takemura
Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo-ku Tokyo 113-8656, Japan
e-mail: takemura@stat.t.u-tokyo.ac.jp

tell whether an arm is optimal unless he pulls it many times, but he also sustains a loss if he pulls a suboptimal arm many times.

We consider an infinite-horizon K-armed bandit problem. There are K arms $\Pi_1, \ldots, \Pi_K$ and the arms are pulled an infinite number of times. Arm Π_j provides a reward with a probability distribution F_j with expected value μ_j. The player receives a reward according to F_j independently in each round. If the expected values are known, it is optimal to pull the arm with the maximum expected value $\mu^* = \max_j \mu_j$ every time. Π_i is called suboptimal if $\mu_i < \mu^*$. A policy is an algorithm to choose the next arm to pull based on the results of past rounds.

This problem was first considered by Robbins (1952), and since then, many studies have been undertaken (Agrawal 1995b; Even-Dar et al. 2002; Meuleau and Bourgine 1999; Strens 2000; Vermorel and Mohri 2005; Yakowitz and Lowe 1991). There are also many extensions of the problem. For example, Auer et al. (2002b) removed the assumption that rewards are stochastic. In the stochastic setting, the cases of non-stationary distributions (Gittins 1989; Ishikida and Varaiya 1994; Katehakis and Veinott 1987), or an infinite (possibly uncountable) number of arms (Agrawal 1995a; Kleinberg 2005) have been considered.

In our setting, Lai and Robbins (1985) established a theoretical framework for determining optimal policies, and Burnetas and Katehakis (1996) extended their result to multiparameter or non-parametric models. Consider a model $\mathcal{F}$, that is, a generic family of distributions. The player knows $\mathcal{F}$ and that each F_j is an element of $\mathcal{F}$. Let $T_j(n)$ denote the number of times that Π_j has been pulled over the first n rounds. A policy is *consistent* on model $\mathcal{F}$ if $\mathrm{E}[T_i(n)] = \mathrm{o}(n^a)$ for all suboptimal arms Π_i and all $a > 0$.

Burnetas and Katehakis (1996) proved the following lower bound for any suboptimal arm Π_i under a consistent policy. With probability tending to one,

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

where $\mathrm{E}(G)$ is the expected value of distribution G and $D(\cdot \| \cdot)$ denotes the Kullback-Leibler divergence. Under mild regularity conditions on $\mathcal{F}$,

$$\inf_{G \in \mathcal{F}: \mathrm{E}(G) > \mu} D(F \| G) = \inf_{G \in \mathcal{F}: \mathrm{E}(G) \geq \mu} D(F \| G)$$

and we define

$$D_{\min}(F, \mu) = \inf_{G \in \mathcal{F}: \mathrm{E}(G) \geq \mu} D(F \| G).$$

We sometimes write $D_{\min}(F, \mu, \mathcal{F})$ for $D_{\min}(F, \mu)$ when we want to specify the feasible region explicitly.

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 Lai and Robbins (1985) and Burnetas and Katehakis (1996), policies achieving the above bound are proposed. These policies are based on the notion of an *upper confidence bound*, which can be interpreted as the upper confidence limit for the expectation of each arm with significance level $1/n$.

Although policies based on the upper confidence bound are optimal, they are often hard to compute in practice. Thus, Auer et al. (2002a) proposed some policies, called UCB policies, which estimate the expectation of each arm in a similar way to the upper confidence bound. UCB policies are practical because of their simple form and fine performance. In particular, "UCB-tuned" is widely used because of its excellent simulation results. However, UCB-tuned has not been analyzed theoretically and the consistency of the policy is unknown.

Theoretical analyses of other UCB policies have been given, but the coefficients of their logarithmic terms do not necessarily achieve the bound (1).

In this paper we propose the minimum empirical divergence (MED) policy. We derive a finite-time regret for this policy which achieves the asymptotic bound when $\mathcal{F}$ is the family of all distributions with finite support included in some fixed interval, say [0, 1]. This is larger than the model used in Burnetas and Katehakis (1996) where the supports of distributions come from a common finite set known to the player. The optimality of the MED policy is stronger than that of the policy proposed by Burnetas and Katehakis (1996) because the MED policy achieves the same asymptotic bound in spite of weaker knowledge about the distributions of rewards (see Remark 2 for details). We also show some simulation results for the MED policy. These results are comparable to those of UCB policies.

Our MED policy is motivated by an observation following from (1). When a policy achieving (1) is used, a suboptimal arm Π_i waits roughly $\exp(n_i D_{\min}(F_i, \mu^*))$ rounds to be pulled after the n_i-th play of Π_i. Thus, we expect that a policy pulling Π_i with probability proportional to $\exp(-n_i D_{\min}(F_i, \mu^*))$ will achieve the bound in (1). The MED policy is obtained by plugging $\hat{F}_i, \hat{\mu}^*$ into F_i, μ^* in $D_{\min}$, where $\hat{F}_i$ is the empirical distribution of rewards from Π_i and $\hat{\mu}^*$ is the current best sample mean.

The MED policy has a strong connection with the DMED policy in Honda and Takemura (2010), which is a modification of the MED policy with a deterministic algorithm. The DMED policy is asymptotically optimal for the model of bounded support distributions, which is a more general setting than the finite support model assumed for the MED policy. However, the evaluation of the regret for the DMED policy is completely dependent on asymptotic arguments and the finite-time regret is unknown. We derive a bound for the finite-time regret for the MED policy by using the finiteness of the support.

The MED policy requires the computation of $D_{\min}(\hat{F}_i, \hat{\mu}^*) = \min_{G \in \mathcal{F}: \mathrm{E}(G) \geq \hat{\mu}^*} D(\hat{F}_i \| G)$ at each round whereas the upper confidence bound by Burnetas and Katehakis (1996) requires the computation of

$$\max_{G \in \mathcal{F}: D_{\min}(\hat{F}_i \| G) \leq \frac{\log n}{n_i}} \mathrm{E}(G). \tag{2}$$

$D_{\min}$ and the expression in (2) are quantities that are dual to each other but the former has two advantages in practical implementation. First, $D_{\min}(\hat{F}_i, \hat{\mu}^*)$ is smooth in $\hat{\mu}^*$ which converges to μ^*. Therefore, the value in the previous round can be used as a good approximation of $D_{\min}$ for the current round. On the other hand (2) continues to increase with n and it has to be computed many times. Second, as shown in Theorem 3 below, $D_{\min}$ can be expressed as a *univariate* convex optimization problem for our model. Although (2) is also a convex optimization problem, the nonlinear constraint $D(\hat{F}_i \| G) \leq \frac{\log n}{n_i}$ is harder to handle.

The MED policy is categorized as a probability matching method (see, e.g., Vermorel and Mohri 2005 for classification of policies). In this method each arm is pulled according to a probability reflecting how likely the arm is to be optimal. For example, Wyatt (1997) proposed probability matching policies for Boolean and Gaussian models using a Bayesian approach with prior/posterior distributions. In our approach the probability assigned to each arm is determined by the (normalized) maximum likelihood instead of the posterior probability.

This paper is organized as follows. In Sect. 2 we give the definitions used throughout this paper and introduce the asymptotic bound by Burnetas and Katehakis (1996) that is satisfied by any consistent policy. In Sect. 3 we describe the MED policy and prove its asymptotic optimality for finite support models. We also discuss practical implementation issues for the minimization problem involved in the MED policy. In Sect. 4 some simulation results are shown. We conclude the paper with some remarks in Sect. 5.

2 Preliminaries

In this section, we introduce the notation used in this paper and present the asymptotic bound for a generic model, which was established by Burnetas and Katehakis (1996).

Let $\mathcal{F}$ be a generic family of probability distributions on $\mathbb{R}$ and let $F_j \in \mathcal{F}$ be the distribution of Π_j, for $j = 1, \ldots, K$. $P_F[\cdot]$ and $\mathrm{E}_F[\cdot]$ denote the probability and the expectation under $F \in \mathcal{F}$, respectively. When we write, for example, $P_F[X \in A]$ $(A \subset \mathbb{R})$ or $\mathrm{E}_F[\theta(X)]$ $(\theta(\cdot)$ is a function $\mathbb{R} \to \mathbb{R})$, X denotes a random variable with distribution F. We define $F(A) \equiv P_F[X \in A]$ and $\mathrm{E}(F) \equiv \mathrm{E}_F[X]$. The expected value of Π_j is denoted by $\mu_j \equiv \mathrm{E}(F_j)$. We denote the optimal expected value by $\mu^* \equiv \max_j \mu_j$.

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 J_n be the arm chosen in the n-th round. Then

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

where $\mathbb{I}[\cdot]$ denotes the indicator function. Regret after the n-th round is given by

$$\mathrm{Regret}(n) = \sum_{j=1}^{K} \Delta_j T_j(n) \tag{3}$$

where $\Delta_j \equiv \mu^* - \mu_j$. For notational convenience we write $T'_j(n) \equiv T_j(n-1)$, which is the number of times that Π_j has been pulled prior to the n-th round.

Let $\hat{F}_{j,t}$ and $\hat{\mu}_{j,t} \equiv \mathrm{E}(\hat{F}_{j,t})$ be the empirical distribution and the mean of the first t rewards from Π_j, respectively. Similarly, let $\hat{F}_j(n) \equiv \hat{F}_{j,T'_j(n)}$ and $\hat{\mu}_j(n) \equiv \hat{\mu}_{j,T'_j(n)}$ be the empirical distribution and the mean of Π_j after the first $n-1$ rounds, respectively. $\hat{\mu}^*(n) \equiv \max_j \hat{\mu}_j(n)$ denotes the highest empirical mean after $n-1$ rounds. We call Π_j a current best if $\hat{\mu}_j(n) = \hat{\mu}^*(n)$.

For an event A, 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 the more precise $P_{\boldsymbol{F}}[\{J_n = j\} \cap \{T'_j(n) = t\}]$.

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

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

where the Kullback-Leibler divergence or relative entropy $D(F\|G)$ is given by

$$D(F\|G) \equiv \begin{cases} \mathrm{E}_F\left[\log \frac{\mathrm{d}F}{\mathrm{d}G}\right] & \frac{\mathrm{d}F}{\mathrm{d}G} \text{ exists,} \\ +\infty & \text{otherwise.} \end{cases}$$

$D_{\inf}$ 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 $D_{\inf}(F, \mu, \mathcal{F}) = +\infty$.

Theorem 2 of Lai and Robbins (1985) gave a lower bound for $\mathrm{E}[T_i(n)]$ for any suboptimal Π_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 (1996) extended the bound to general non-parametric models. Their bound is given by the following theorem.

Theorem 1 (Proposition 1 of Burnetas and Katehakis 1996) *Fix a consistent policy and* $\boldsymbol{F} \in \mathcal{F}^K$. *If* $\mathrm{E}(F_i) < \mu^*$ *and* $0 < D_{\inf}(F_i, \mu^*, \mathcal{F}) < \infty$, *then for any* $\epsilon > 0$

$$\lim_{N\to\infty} P_{\boldsymbol{F}}\left[T_i(N) \geq \frac{(1-\epsilon)\log N}{D_{\inf}(F_i, \mu^*, \mathcal{F})}\right] = 1. \tag{4}$$

Consequently,

$$\liminf_{N\to\infty} \frac{\mathrm{E}_{\boldsymbol{F}}[T_i(N)]}{\log N} \geq \frac{1}{D_{\inf}(F_i, \mu^*, \mathcal{F})}. \tag{5}$$

Note that by Markov's inequality

$$\mathrm{E}_{\boldsymbol{F}}[T_i(N)] \geq \frac{(1-\epsilon)\log N}{D_{\inf}(F_i, \mu^*, \mathcal{F})}\, P_{\boldsymbol{F}}\left[T_i(N) \geq \frac{(1-\epsilon)\log N}{D_{\inf}(F_i, \mu^*, \mathcal{F})}\right] \tag{6}$$

and (5) follows straightforwardly from (4) and (6), by dividing both sides of (6) by $\log N$, letting $N \to \infty$ and finally letting $\epsilon \downarrow 0$.

3 Asymptotically optimal policy for finite support models

Let $\mathcal{A} \equiv \{F : |\mathrm{supp}(F)| < \infty, \mathrm{supp}(F) \subset [a, b]\}$ be a family of distributions with a *finite* bounded support, where $\mathrm{supp}(F)$ is the support of distribution F, and a and b are constants known to the player. We assume $a = 0$ and $b = 1$ without loss of generality. We write $\mathrm{supp}'(F) \equiv \{1\} \cup \mathrm{supp}(F)$ and $\mathcal{A}_{\mathcal{X}} \equiv \{F \in \mathcal{A} : \mathrm{supp}(F) \subset \mathcal{X}\}$ where $\mathcal{X}$ is an arbitrary subset of $[0, 1]$.

We consider $\mathcal{A}$ as a model $\mathcal{F}$, and in this section, we propose a policy that we call the minimum empirical divergence (MED) policy. We prove in Theorem 2 that the proposed policy achieves the bound given in Theorem 1. Then, we describe a univariate convex optimization technique to compute the value of $D_{\min}$ used in the policy.

Remark 1 The finiteness of the support can not be determined from finite samples and every policy for $\mathcal{A}$ is also applicable to $\{F : \mathrm{supp}(F) \subset [a, b]\}$. However, our proof of the optimality in this paper is for the above $\mathcal{A}$. The advantage of assuming finiteness is that we can employ the method of types in the large deviation technique (see Appendix A). This enables us to consider all empirical distributions obtained from each arm.

In our model it is convenient to use

$$D_{\min}(F, \mu, \mathcal{A}) \equiv \inf_{G\in\mathcal{A}:\mathrm{E}(G)\geq\mu} D(F\|G)$$

instead of $D_{\inf}(F, \mu, \mathcal{A}) \equiv \inf_{G\in\mathcal{A}:\mathrm{E}(G)>\mu} D(F\|G)$. We write "$D_{\min}$" since we will show through the proof of Theorem 3 that a minimizer G^* exists.

Lemma 1 $D_{\min}(F, \mu, \mathcal{A}) = D_{\inf}(F, \mu, \mathcal{A})$ *holds for all* $F \in \mathcal{A}$ *and* $\mu < 1$.

Proof Let $\epsilon > 0$ be arbitrary. Then

$$\inf_{G\in\mathcal{A}:\mathrm{E}(G)\geq\mu} D(F\|G) \leq \inf_{G\in\mathcal{A}:\mathrm{E}(G)>\mu} D(F\|G) \leq \inf_{G\in\mathcal{A}:\mathrm{E}(G)\geq\mu+\epsilon} D(F\|G)$$

or equivalently

$$D_{\min}(F,\mu,\mathcal{A}) \le D_{\inf}(F,\mu,\mathcal{A}) \le D_{\min}(F,\mu+\epsilon,\mathcal{A}).$$

Since $\epsilon > 0$ is arbitrary, we obtain

$$D_{\inf}(F,\mu,\mathcal{A}) \le \liminf_{\epsilon\downarrow 0} D_{\min}(F,\mu+\epsilon,\mathcal{A}). \tag{7}$$

We will prove in Lemma 6 that $D_{\min}(F,\mu+\epsilon,\mathcal{A}) \le D_{\min}(F,\mu,\mathcal{A}) + \epsilon/(1-\mu-\epsilon)$ for $\epsilon < 1-\mu$. Therefore, the right-hand side of (7) equals $D_{\min}(F,\mu,\mathcal{A})$. □

3.1 Optimality of the minimum empirical divergence policy

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

[Minimum Empirical Divergence Policy]
Initialization. Pull each arm once.
Loop. For the nth round,

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

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

Note that

$$\frac{1}{K} \le p_j(n) \le 1 \tag{8}$$

for any currently best Π_j since $\hat{D}_j(n) = 0$. As a result, it holds for all j that

$$\frac{1}{K}\exp(-T_j'(n)\hat{D}_j(n)) \le p_j(n) \le \exp(-T_j'(n)\hat{D}_j(n)). \tag{9}$$

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

Note that our policy is a randomized policy. Therefore, probability statements below for the MED policy also involve this randomization. However, for simplicity we do not explicitly show this randomization in the notation.

Now we present the main theorem of this paper.

Theorem 2 *Fix $\boldsymbol{F} \in \mathcal{A}^K$ for which there exists a single arm Π_j such that $\mu_j = \mu^*$ and $\mu_i < \mu^*$ for all $i \neq j$. Under the MED policy, the expected regret after the Nth round is bounded as*

$$\mathrm{E}_{\boldsymbol{F}}[\mathrm{Regret}(N)] \le \sum_{i\neq j} \frac{\Delta_i(1+\epsilon)\log N}{D_{\min}(F_i,\mu^*,\mathcal{A})} + c(\epsilon,\boldsymbol{F}) \tag{10}$$

where $\epsilon > 0$ is arbitrary and $c(\epsilon,\boldsymbol{F})$ is a constant independent of N and is specifically given in (36).

We obtain the tightest bound by taking the infimum over ϵ. Also, note that we obtain

$$\limsup_{N\to\infty} \frac{\mathrm{E}_{\boldsymbol{F}}[\mathrm{Regret}(N)]}{\log N} \leq \sum_{i\neq j} \frac{\Delta_i}{D_{\min}(F_i, \mu^*, \mathcal{A})},$$

by dividing both sides by $\log N$, letting $N \to \infty$ and finally letting $\epsilon \downarrow 0$. In view of (3) and (5) we see that the MED policy is asymptotically optimal. We give a proof of Theorem 2 in Sect. 3.3.2.

Remark 2 The optimality of the MED policy in Theorem 2 is stronger than the optimality of the policy proposed by Burnetas and Katehakis (1996). Let $\mathcal{X} \subset [0,\ 1]$ be an arbitrary known finite subset of $[0,\ 1]$ such that $1 \in \mathcal{X}$. Then $\mathcal{A}_{\mathcal{X}} = \{F \in \mathcal{A} : \mathrm{supp}(F) \subset \mathcal{X}\}$ is the model used by Burnetas and Katehakis (1996). Since $\mathcal{A}_{\mathcal{X}} \subset \mathcal{A}$, $D_{\min}(F, \mu, \mathcal{A}) \leq D_{\min}(F, \mu, \mathcal{A}_{\mathcal{X}})$ holds and there is a possibility that $\mathrm{E}_{\boldsymbol{F}}[T_i(n)]$ for $\boldsymbol{F} \in \mathcal{A}_{\mathcal{X}}^K$ achieved by an asymptotically optimal policy for $\mathcal{A}$ may be worse than that achieved by an asymptotically optimal policy for $\mathcal{A}_{\mathcal{X}}$. However, it is easily checked using Lemma 2 below that $D_{\min}(F, \mu, \mathcal{A}) = D_{\min}(F, \mu, \mathcal{A}_{\mathcal{X}})$ for $F \in \mathcal{A}_{\mathcal{X}}$. Therefore, the MED policy achieves the same asymptotic bound as that by Burnetas and Katehakis (1996) for $\boldsymbol{F} \in \mathcal{A}_{\mathcal{X}}^K$ in spite of the weaker knowledge about the distributions.

3.2 Computation of $D_{\min}$ and its properties

For implementing the MED policy it is essential to compute the minimum empirical divergence $D_{\min}(\hat{F}_j(n), \hat{\mu}^*(n), \mathcal{A})$ for each round efficiently. In addition, for proofs of Lemma 1 and Theorem 2, we need to understand the behavior of $D_{\min}(F, \mu, \mathcal{A})$ as a function of μ. In this subsection, we clarify the nature of the convex optimization involved in the computation of $D_{\min}(\hat{F}_j(n), \hat{\mu}^*(n), \mathcal{A})$ and show how the minimum can be computed efficiently.

First we prove that it is sufficient to consider $\mathcal{A}_{\mathrm{supp}'(F)}$ (recall that $\mathrm{supp}'(F) = \mathrm{supp}(F) \cup \{1\}$) for the computation of $D_{\min}(F, \mu, \mathcal{A})$:

Lemma 2 *$D_{\min}(F, \mu, \mathcal{A}) = D_{\min}(F, \mu, \mathcal{A}_{\mathrm{supp}'(F)})$ holds for any $F \in \mathcal{A}$.*

Proof Take an arbitrary $G \in \mathcal{A} \setminus \mathcal{A}_{\mathrm{supp}'(F)}$ such that $\mathrm{E}(G) \geq \mu$, $D(F\|G) < +\infty$ and $G(\mathrm{supp}'(F)) = p < 1$. Define $G' \in \mathcal{A}_{\mathrm{supp}'(F)}$ as

$$G'(\{x\}) \equiv \begin{cases} G(\{1\}) + (1-p) & x = 1 \\ G(\{x\}) & x \neq 1,\ x \in \mathrm{supp}(F) \\ 0 & \text{otherwise.} \end{cases}$$

$\mathrm{E}(G)$ and $D(F\|G)$ are bounded by

$$\begin{aligned} \mathrm{E}(G) &\leq 1 \cdot G(\{1\}) + \sum_{x\in\mathrm{supp}(G)\cap\mathrm{supp}(F)\setminus\{1\}} x G(\{x\}) + \sum_{x\in\mathrm{supp}(G)\setminus(\mathrm{supp}(F)\cup\{1\})} 1 \cdot G(\{x\}) \\ &= \big(G'(\{1\}) - (1-p)\big) + \sum_{x\in\mathrm{supp}(G)\cap\mathrm{supp}(F)\setminus\{1\}} x G'(\{x\}) + (1-p) \\ &= \sum_{x\in\mathrm{supp}'(G)\cap\mathrm{supp}'(F)} x G'(\{x\}) \\ &= \mathrm{E}(G') \quad (\text{as } \mathrm{supp}(G') \subset (\mathrm{supp}'(G) \cap \mathrm{supp}'(F))) \end{aligned}$$

and

$$\begin{aligned} D(F\|G) &= \sum_{x\in\mathrm{supp}(F)} F(\{x\})\log\frac{F(\{x\})}{G(\{x\})} \\ &\geq \sum_{x\in\mathrm{supp}(F)} F(\{x\})\log\frac{F(\{x\})}{G(\{x\})+(1-p)\mathbb{I}[x=1]} \\ &= D(F\|G'), \end{aligned}$$

respectively. Therefore, we obtain

$$\inf_{G\in\mathcal{A}:\mathrm{E}(G)\geq\mu} D(F\|G) \geq \inf_{G'\in\mathcal{A}_{\mathrm{supp}'(F)}:\mathrm{E}(G')\geq\mu} D(F\|G').$$

The converse inequality is obvious from $\mathcal{A}_{\mathrm{supp}'(F)}\subset\mathcal{A}$. □

In view of this lemma, we simply write $D_{\min}(F,\mu)$ instead of $D_{\min}(F,\mu,\mathcal{A}) = D_{\min}(F,\mu,\mathcal{A}_{\mathrm{supp}'(F)})$ for the rest of this paper.

Let $M \equiv |\mathrm{supp}'(F)|$ and denote the finite symbols in $\mathrm{supp}'(F)$ by $x_1,\ldots,x_M$: i.e., $\{1\}\cup\mathrm{supp}(F) = \{x_1,\ldots,x_M\}$. We assume $x_1 = 1$ and $x_i < 1$ for $i > 1$ without loss of generality and write $f_i \equiv F(\{x_i\})$.

Now the computation of $D_{\min}(F,\mu)$ can be formulated as the following convex optimization problem for $G = (g_1,\ldots,g_M)$ using Lemma 2:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{M} f_i\log\frac{f_i}{g_i} \\ &\text{subject to} && g_i\geq 0,\ \forall i, \quad \sum_{i=1}^{M} x_i g_i \geq \mu, \quad \sum_{i=1}^{M} g_i = 1, \end{aligned} \tag{11}$$

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

It is obvious that $G = F$ is the optimal solution with the optimal value 0 when $1\geq\mathrm{E}(F)\geq\mu$. Also $G=\delta_1$, the unit point mass at 1, is the unique feasible solution if $\mu = 1$. For $\mu > 1$ the problem is infeasible. Since these cases are trivial, we consider the case $\mathrm{E}(F) < \mu < 1$ in the following.

Define a function $h(\nu;F,\mu)$ on ν with parameters F,μ by

$$h(\nu;F,\mu) \equiv \mathrm{E}_F[\log(1-(X-\mu)\nu)] = \sum_{i=1}^{M} f_i\log(1-(x_i-\mu)\nu), \tag{12}$$

where we define $\log x\equiv-\infty$ for $x\leq 0$. Then, wherever $h(\nu;F,\mu)$ is finite, the derivatives h', h'' on ν exist and

$$h'(\nu;F,\mu) = \frac{\partial}{\partial\nu}h(\nu;F,\mu) = -\sum_{i=1}^{M}\frac{f_i(x_i-\mu)}{1-(x_i-\mu)\nu}, \tag{13}$$

$$h''(\nu;F,\mu) = \frac{\partial^2}{\partial\nu^2}h(\nu;F,\mu) = -\sum_{i=1}^{M}\frac{f_i(x_i-\mu)^2}{(1-(x_i-\mu)\nu)^2}. \tag{14}$$

We write $h(\nu), h'(\nu), h''(\nu)$ when F, μ are obvious from the context. Now we show in Theorem 3 that the computation of $D_{\min}$ can be expressed as the maximization of $h(\nu)$. Since $h(\nu)$ is concave, this is a univariate convex optimization problem. Therefore, $D_{\min}$ can be computed easily by iterative methods such as Newton's method (see, e.g., Boyd and Vandenberghe (2004) for general methods of convex programming).

Theorem 3 *Define* $\mathrm{E}_F[(1-\mu)/(1-X)] = \infty$ *for the case* $F(\{1\}) = f_1 > 0$. *Then the following three properties hold for* $\mathrm{E}(F) < \mu < 1$:

(i) $D_{\min}(F, \mu)$ *can be written as*

$$D_{\min}(F, \mu) = \max_{0 \le \nu \le \frac{1}{1-\mu}} h(\nu) \tag{15}$$

and the optimal solution $\nu^* \equiv \operatorname{argmax}_{0 \le \nu \le \frac{1}{1-\mu}} h(\nu)$ *is unique*.

In particular for the case $\mathrm{E}[(1-\mu)/(1-X)] \le 1$, *we have* $\nu^* = 1/(1-\mu)$ *and* (15) *can be simply written as*

$$D_{\min}(F, \mu) = h\left(\frac{1}{1-\mu}\right) = \sum_{i=2}^{M} f_i \log\left(\frac{1-x_i}{1-\mu}\right). \tag{16}$$

(ii) ν^* *satisfies*

$$\nu^* \ge \frac{\mu - \mathrm{E}(F)}{\mu(1-\mu)}.$$

(iii) $D_{\min}(F, \mu)$ *is differentiable in* $\mu \in (\mathrm{E}(F), 1)$ *and*

$$\frac{\partial}{\partial \mu} D_{\min}(F, \mu) = \nu^*.$$

Note that

$$\lim_{\nu \uparrow \frac{1}{1-\mu}} h'(\nu) = (1-\mu)\left(1 - \mathrm{E}_F\left[\frac{1-\mu}{1-X}\right]\right) \tag{17}$$

including the case that $\mathrm{E}_F[(1-\mu)/(1-X)] = +\infty$ for $F(\{1\}) > 0$. Intuitively, the special case $\mathrm{E}_F[(1-\mu)/(1-X)] \le 1$ in Theorem 3 (i) arises because $h'(\nu) \ge \lim_{\nu \uparrow 1/(1-\mu)} h'(\nu) \ge 0$, i.e., $h(\nu)$ is monotonically increasing in $\nu \in [0, 1/(1-\mu)]$ for this case.

We sometimes write $\nu^*(F, \mu)$ instead of ν^* when we need to emphasize that ν^* depends on F and μ. We give a proof of Theorem 3 in Sect. 3.3.1.

Remark 3 Parts (ii) and (iii) of the theorem are useful not only for proofs but also for a practical implementation: (ii) can be used for obtaining the lower bound of the optimal solution ν^*, and (iii) can be used for linear approximation if the variation of μ is small.

3.3 Proofs of Theorems 2 and 3

In this section we give proofs of Theorems 2 and 3. First we prove Theorem 3; then we prove Theorem 2 using Theorem 3.

3.3.1 A proof of Theorem 3

(i) Equality $h''(\nu) = h''(\nu; F, \mu) = 0$ holds only for the degenerate case that $f_i = 1$ at $x_i = \mu$, and this case does not satisfy the assumption $\mathrm{E}(F) < \mu$. Therefore, $h''(\nu) < 0$ wherever $h(\nu) > -\infty$ and $h(\nu)$ is strictly concave. The uniqueness of $\nu^*(F, \mu)$ follows from the strict concavity.

Now we derive results (15) and (16) by the technique of Lagrange multipliers (see Theorem 4 in Appendix B). Note that the problem (11) is a convex optimization problem, and therefore, a local minimizing point is a global minimizing point.

The Lagrangian function for (11) is

$$L(\{g_i\}, \{\lambda_i\}, \nu, \xi) = \sum_{i=1}^{M} f_i \log \frac{f_i}{g_i} - \sum_{i=1}^{M} \lambda_i g_i + \nu \left(\mu - \sum_{i=1}^{M} x_i g_i \right) + \xi \left(1 - \sum_{i=1}^{M} g_i \right)$$

and condition (a) in Theorem 4 is

$$\begin{aligned}
&\lambda_i^* g_i^* = 0, \quad \lambda_i^* \geq 0, \quad g_i^* \geq 0, \quad i = 1, \ldots, M, \\
&\nu^* \left(\mu - \sum_{i=1}^{M} x_i g_i^* \right) = 0, \quad \nu^* \geq 0, \quad \sum_{i=1}^{M} x_i g_i^* \geq \mu, \\
&\sum_{i=1}^{M} g_i^* = 1, \\
&-\frac{f_i}{g_i^*} - \lambda_i^* - \nu^* x_i - \xi^* = 0, \quad i = 1, \ldots, M,
\end{aligned}$$

For condition (b) in Theorem 4, extracting (B.1) and (B.2), it suffices to show

$$\sum_{i=1}^{M} z_i^2 \frac{f_i}{(g_i^*)^2} > 0 \quad \text{for all } \boldsymbol{z} = (z_1, \ldots, z_M) \neq 0 \text{ such that } \sum_{i=1}^{M} z_i = 0, \tag{18}$$

where $f_i/(g_i^*)^2 \equiv 0$ for $f_i = 0$.

First we consider the case $\mathrm{E}_F[(1-\mu)/(1-X)] \leq 1$. We show

$$g_i^* = \begin{cases} f_i \frac{1-\mu}{1-x_i} & i \neq 1 \\ 1 - \sum_{j=2}^{M} f_j \frac{1-\mu}{1-x_j} & i = 1, \end{cases} \tag{19}$$

$\lambda_i^* = 0$, $\nu^* = 1/(1-\mu)$ and $\xi^* = -1/(1-\mu)$ satisfy the second-order sufficient conditions for a strict local minimizing point in Theorem 4. Note that $\mathrm{E}_F[(1-\mu)/(1-X)] \leq 1$ implies $f_1 = 0$ and therefore

$$\begin{aligned}
\sum_{i=1}^{M} x_i g_i^* &= \left(1 - \sum_{j=2}^{M} f_j \frac{1-\mu}{1-x_j} \right) + \sum_{i=2}^{M} x_i f_i \frac{1-\mu}{1-x_i} \\
&= 1 - \sum_{j=2}^{M} f_j (1-x_j) \frac{1-\mu}{1-x_j} = \mu
\end{aligned}$$

and

$$g_1^* = 1 - \sum_{j=2}^{M} f_j \frac{1-\mu}{1-x_j} = 1 - \mathrm{E}_F\left[\frac{1-\mu}{1-X}\right] \geq 0.$$

The other conditions in (a) are easily checked. By (18), condition (b) is checked in the following way. The conditions $\boldsymbol{z} \neq 0$ and $\sum_{i=1}^{M} z_i = 0$ imply that $z_i \neq 0$ holds for more than one i. On the other hand, $f_i/(g_i^*)^2 > 0$ holds for all $i \neq 1$. Therefore, $z_i^2 f_i/(g_i^*)^2 > 0$ holds for at least one i and (18) is satisfied. From this we obtain (16) and then (15) follows from the concavity of $h(\nu)$ and $h'(1/(1-\mu)) \geq 0$ by (17).

Now we consider the remaining case, $\mathrm{E}_F[(1-\mu)/(1-X)] > 1$. Since $h'(0) > 0$ and $\lim_{\nu \uparrow 1/(1-\mu)} h'(\nu) < 0$ by (17), $\nu^* = \operatorname{argmax}_{0 \leq \nu \leq 1/(1-\mu)} h(\nu)$ satisfies

$$-h'(\nu^*) = \sum_{i=1}^{M} f_i \frac{x_i - \mu}{1 - (x_i - \mu)\nu^*} = 0.$$

Therefore, we obtain

$$\sum_{i=1}^{M} \frac{f_i}{1-(x_i-\mu)\nu^*} = \sum_{i=1}^{M} f_i \frac{1-(x_i-\mu)\nu^*}{1-(x_i-\mu)\nu^*} + \nu^* \sum_{i=1}^{M} f_i \frac{x_i-\mu}{1-(x_i-\mu)\nu^*} = 1 \tag{20}$$

and

$$\sum_{i=1}^{M} \frac{f_i x_i}{1-(x_i-\mu)\nu^*} = \sum_{i=1}^{M} f_i \frac{x_i-\mu}{1-(x_i-\mu)\nu^*} + \mu \sum_{i=1}^{M} \frac{f_i}{1-(x_i-\mu)\nu^*} = \mu. \tag{21}$$

It can be checked using (20) and (21) that

$$g_i^* = \frac{f_i}{1-(x_i-\mu)\nu^*}, \tag{22}$$

$$\lambda_i^* = \begin{cases} 0 & f_i > 0 \\ 1-(x_i-\mu)\nu^* & f_i = 0, \end{cases}$$

$\xi^* = 1 + \mu\nu^*$ and ν^* satisfy the second-order sufficient conditions (a). Condition (b) is checked in the same way as for the case $\mathrm{E}_F[(1-\mu)/(1-X)] \leq 1$, and (15) is obtained.

Remark 4 From the existence of $\nu^* = \operatorname{argmax}_{0 \leq \nu \leq 1/(1-\mu)} h(\nu)$, $\{g_i^*\}$ in (19) and (22) always exists. Furthermore, it is assured from the second-order conditions that $\{g_i^*\}$ is a minimizer for $D_{\min}(F, \mu)$. As a result, $D_{\min}(F, \mu) \equiv \inf_{G \in \mathcal{A}: \mathrm{E}(G) \geq \mu} D(F\|G)$ can be written as $D_{\min}(F, \mu) = \min_{G \in \mathcal{A}: \mathrm{E}(G) \geq \mu} D(F\|G)$.

(ii) The claim is obviously true for the case $\mathrm{E}_F[(1-\mu)/(1-X)] \leq 1$ and we consider the case $\mathrm{E}_F[(1-\mu)/(1-X)] > 1$. Note that we can assume $\mathrm{E}(F) > 0$ for this case since $\mathrm{E}(F) = 0$ implies $F(\{0\}) = 1$ and $\mathrm{E}_F[(1-\mu)/(1-X)] = 1 - \mu < 1$. Define a function $w(x, \nu)$ on $[0, 1] \times [0, 1/(1-\mu))$ as

$$w(x, \nu) \equiv \frac{x-\mu}{1-(x-\mu)\nu}.$$

Since

$$\frac{\partial^2 w(x,\nu)}{\partial x^2} = \frac{2\nu}{(1-(x-\mu)\nu)^3} \geq 0,$$

$w(x,\nu)$ is convex in $x \in [0,1]$ for any fixed $\nu \in [0, 1/(1-\mu))$. Therefore,

$$\begin{aligned} h'(\nu) &= -\sum_{i=1}^{M} f_i w(x_i,\nu) \\ &\geq -\sum_{i=1}^{M} f_i \big((1-x_i)w(0,\nu) + x_i w(1,\nu)\big) \\ &= (\mathrm{E}(F)-1)w(0,\nu) - \mathrm{E}(F)w(1,\nu) \\ &= \frac{\mu(1-\mathrm{E}(F))}{1+\mu\nu} - \frac{\mathrm{E}(F)(1-\mu)}{1-(1-\mu)\nu}. \end{aligned} \tag{23}$$

The right-hand side of (23) is 0 for $\nu = (\mu - \mathrm{E}(F))/(\mu(1-\mu)) \in [0, 1/(1-\mu))$ and so

$$h'\left(\frac{\mu-\mathrm{E}(F)}{\mu(1-\mu)}\right) \geq 0.$$

Since $h'(\nu)$ is monotonically decreasing from the concavity of $h(\nu)$, the inequality $\nu^* \geq (\mu - \mathrm{E}(F))/(\mu(1-\mu))$ is proved.

(iii) It is obvious that $\frac{\partial}{\partial\mu} D_{\min}(F,\nu) = 1/(1-\mu) = \nu^*$ for $\mathrm{E}_F[(1-\mu)/(1-X)] < 1$ and

$$\lim_{\epsilon \downarrow 0} \frac{D_{\min}(F,\mu+\epsilon) - D_{\min}(F,\mu)}{\epsilon} = \frac{1}{1-\mu}$$

for $\mathrm{E}_F[(1-\mu)/(1-X)] = 1$.

Now consider the case $\mathrm{E}_F[(1-\mu)/(1-X)] \geq 1$. Define an unconstrained optimization problem $D'_{\min}(F,\mu) \equiv \max_\nu h(\nu;F,\mu) = \max_\nu h(\nu)$ with parameter μ. Since $h'(0) \geq 0$ and $\lim_{\nu\uparrow 1/(1-\mu)} h'(\nu) \leq 0$, $D'_{\min}(F,\mu) = \max_{0\leq\nu\leq 1/(1-\mu)} h(\nu) = D_{\min}(F,\mu)$ for this case. Now we apply Theorem 5 in Appendix B to $D'_{\min}(F,\mu) = -\min_\nu(-h(\nu))$. For the unconstrained optimization problem, the second-order sufficient condition that ν^* is a strict local minimization point of $-h(\nu)$ is simply written as $-h'(\nu^*) = 0$ and $-h''(\nu^*) > 0$, which correspond to conditions (a) and (*b*), respectively. We can check them easily and we obtain from Theorem 5 that $D'_{\min}(F,\mu)$ is differentiable in μ with

$$\frac{\partial}{\partial\mu} D'_{\min}(F,\mu) = \frac{\partial}{\partial\mu} h(\nu;F,\mu)\bigg|_{\nu=\nu^*(F,\mu)} = \nu^*(F,\mu).$$

Therefore, we obtain

$$\frac{\partial}{\partial\mu} D_{\min}(F,\mu) = \frac{\partial}{\partial\mu} D'_{\min}(F,\nu) = \nu^*(F,\mu)$$

for $\mathrm{E}_F[(1-\mu)/(1-X)] > 1$ and

$$\lim_{\epsilon \downarrow 0} \frac{D_{\min}(F, \mu - \epsilon) - D_{\min}(F, \mu)}{-\epsilon} = \lim_{\epsilon \downarrow 0} \frac{D'_{\min}(F, \mu - \epsilon) - D'_{\min}(F, \mu)}{-\epsilon}$$
$$= \nu^*(F, \mu) = \frac{1}{1 - \mu}$$

for $\mathrm{E}_F[(1-\mu)/(1-X)] = 1$.

3.3.2 A proof of Theorem 2

For the proof of Theorem 2, it is necessary to measure the distance between a distribution F_i and the corresponding empirical distribution $\hat{F}_i$. We adopt the variational distance

$$\|F - G\| \equiv \frac{1}{2} \sum_{x \in \mathrm{supp}(F) \cup \mathrm{supp}(G)} |F(\{x\}) - G(\{x\})|$$

for the distance between two distributions $F, G \in \mathcal{A}$. Although variational distance is almost meaningless when F_i has a continuous support, it does allow us to use the following two lemmas that are helpful in our finite support model for deriving a bound on the finite-time regret.

Lemma 3 (Lemma 11.6.1 of Cover and Thomas 2006) *For arbitrary $F, G \in \mathcal{A}$*

$$2\|F - G\|^2 \le D(F\|G).$$

This inequality is sometimes called Pinsker's inequality.

Lemma 4 *Let $F, G \in \mathcal{A}$ and $\theta : [0,1] \to \mathbb{R}$ be arbitrary. Then*

$$\left|\mathrm{E}_F[\theta(X)] - \mathrm{E}_G[\theta(X)]\right| \le \left(\max_x \theta(x) - \min_x \theta(x)\right) \|F - G\|.$$

Proof Let $\bar{\theta} = \max_x \theta(x)$, $\underline{\theta} = \min_x \theta(x)$ and $\theta_0 = (\bar{\theta} + \underline{\theta})/2$. Note that $\mathrm{E}_F[\theta(X)] - \mathrm{E}_G[\theta(X)] = \mathrm{E}_F[\theta(X) - \theta_0] - \mathrm{E}_G[\theta(X) - \theta_0]$. Then we obtain

$$\left|\mathrm{E}_F[\theta(X)] - \mathrm{E}_G[\theta(X)]\right| = \left| \sum_{x \in \mathrm{supp}(F) \cup \mathrm{supp}(G)} (\theta(x) - \theta_0)(F(\{x\}) - G(\{x\})) \right|$$
$$\le \frac{1}{2} \sum_{x \in \mathrm{supp}(F) \cup \mathrm{supp}(G)} (\bar{\theta} - \underline{\theta}) \cdot |F(\{x\}) - G(\{x\})|$$
$$= (\bar{\theta} - \underline{\theta}) \|F - G\|,$$

where the inequality follows from the fact that $|\theta(x) - \theta_0| \le (\bar{\theta} - \underline{\theta})/2$ for all x. □

We now give three lemmas on properties of $D_{\min}$.

Lemma 5 *$D_{\min}(F, \mu)$ is monotonically increasing in μ.*

This result follows immediately from the definition $D_{\min}(F, \mu) \equiv \inf_{G \in \mathcal{A}: \mathrm{E}(G) \ge \mu} D(F\|G)$. We use this monotonicity implicitly in the proof of Theorem 2.

Lemma 6 *For* $\mu' \leq \mu < 1$

$$D_{\min}(F,\mu) - D_{\min}(F,\mu') \leq \frac{\mu - \mu'}{1-\mu}. \tag{24}$$

Furthermore, if $\mu' > \mathrm{E}(F)$ *then*

$$D_{\min}(F,\mu) - D_{\min}(F,\mu') \geq \frac{(\mu-\mu')^2}{2}. \tag{25}$$

Proof First we show (24) holds for $\mu' \leq \mathrm{E}(F)$. Since $D_{\min}(F,\mu') = 0$ for $\mu' \leq \mathrm{E}(F)$, we obtain (24) by

$$\begin{aligned} D_{\min}(F,\mu) - D_{\min}(F,\mu') &= D_{\min}(F,\mu) \\ &\leq h(0) + h'(0)\frac{1}{1-\mu} \quad \text{(by (15) and the concavity of } h(\nu)) \\ &= \frac{\mu - \mathrm{E}(F)}{1-\mu} \\ &\leq \frac{\mu-\mu'}{1-\mu}. \end{aligned}$$

Next we show (24) and (25) hold for $\mu' > \mathrm{E}(F)$. $D_{\min}(F,u)$ is differentiable in $u > \mathrm{E}(F)$ from Theorem 3 (iii) and

$$D_{\min}(F,\mu) - D_{\min}(F,\mu') = \int_{\mu'}^{\mu} \frac{\partial}{\partial u} D_{\min}(F,u)\mathrm{d}u = \int_{\mu'}^{\mu} \nu^*(F,u)\mathrm{d}u.$$

Note that

$$\nu^*(F,u) \leq \frac{1}{1-u} \leq \frac{1}{1-\mu} \tag{26}$$

holds from the definition of $\nu^*(F,u)$ and

$$\nu^*(F,u) \geq \frac{u - \mathrm{E}(F)}{u(1-u)} \geq u - \mu' \tag{27}$$

holds from Theorem 3 (ii). We obtain (24) and (25) by integrating the right-hand sides of (26) and (27) over $u \in [\mu', \mu]$. □

Lemma 7 *Define* $d(\tau; F, \mu)$ *for* $\tau > 0$ *by*

$$d(\tau; F, \mu) \equiv \begin{cases} \frac{\tau^2(1-\mu)}{4D_{\min}(F,\mu)} & \nu^*(F,\mu) = \frac{1}{1-\mu} \\ \frac{\tau(1-(1-\mu)\nu^*(F,\mu))}{\nu^*(F,\mu)} & \nu^*(F,\mu) < \frac{1}{1-\mu}. \end{cases} \tag{28}$$

Then

$$D_{\min}(F',\mu) \geq D_{\min}(F,\mu) - \tau. \tag{29}$$

holds for all $F' \in \mathcal{A}$ *satisfying* $\|F' - F\| < d(\tau; F, \mu)$.

Note that $\liminf_{F' \to F} D_{\min}(F', \mu) \geq D_{\min}(F, \mu)$ follows immediately from this lemma, which means that $D_{\min}(F, \mu)$ is lower-semicontinuous in F.

Proof Let $F' \in \mathcal{A}$ satisfy $\|F' - F\| < d(\tau; F, \mu)$. First we show for arbitrary $\nu \in [0, 1/(1-\mu))$ that

$$D_{\min}(F', \mu) \geq \mathrm{E}_F[\log(1 - (X - \mu)\nu)] - \frac{\nu\, d(\tau; F, \mu)}{1 - (1-\mu)\nu}. \tag{30}$$

Since, for $\nu \in [0, 1/(1-\mu))$ and $x \in [0, 1]$, we have

$$\log(1 - (1-\mu)\nu) \leq \log(1 - (x - \mu)\nu) \leq \log(1 - (0 - \mu)\nu),$$

it holds that

$$\begin{aligned}
&\mathrm{E}_F[\log(1 - (X - \mu)\nu)] - \mathrm{E}_{F'}[\log(1 - (X - \mu)\nu)] \\
&\quad \leq d(\tau; F, \mu)\big(\log(1 - (0 - \mu)\nu) - \log(1 - (1 - \mu)\nu)\big) \quad \text{(by Lemma 4)} \\
&\quad = d(\tau; F, \mu) \log\left(1 + \frac{\nu}{1 - (1-\mu)\nu}\right) \\
&\quad \leq d(\tau; F, \mu) \frac{\nu}{1 - (1-\mu)\nu}.
\end{aligned} \tag{31}$$

Now we have (30) since

$$\begin{aligned}
D_{\min}(F', \mu) &= \max_{0 \leq \nu' \leq \frac{1}{1-\mu}} \mathrm{E}_{F'}[\log(1 - (X - \mu)\nu')] \\
&\geq \mathrm{E}_{F'}[\log(1 - (X - \mu)\nu)] \\
&\geq \mathrm{E}_F[\log(1 - (X - \mu)\nu)] - \frac{\nu\, d(\tau; F, \mu)}{1 - (1-\mu)\nu} \quad \text{(by (31))}.
\end{aligned}$$

We obtain (29) for the case $\nu^*(F, \mu) < 1/(1-\mu)$ by letting $\nu := \nu^*(F, \mu)$.

Now we consider the case $\nu^*(F, \mu) = 1/(1-\mu)$. Since $h(\nu) = \mathrm{E}_F[\log(1 - (X - \mu)\nu)]$ is concave in ν, $h(\nu)$ is bounded from below for $\nu \in [0, 1/(1-\mu))$:

$$\begin{aligned}
\mathrm{E}_F[\log(1 - (X - \mu)\nu)] = h(\nu) &\geq \frac{\frac{1}{1-\mu} - \nu}{\frac{1}{1-\mu}} h(0) + \frac{\nu}{\frac{1}{1-\mu}} h\left(\frac{1}{1-\mu}\right) \\
&= \nu(1-\mu) D_{\min}(F, \mu)
\end{aligned} \tag{32}$$

as $h(0) = 0$ and $h(1/(1-\mu)) = D_{\min}(F, \mu)$. Therefore, for arbitrary $\nu \in [0, 1/(1-\mu))$, we obtain

$$D_{\min}(F', \mu) \geq \nu(1-\mu) D_{\min}(F, \mu) - \frac{\nu}{1 - (1-\mu)\nu} \frac{\tau^2 (1-\mu)}{4 D_{\min}(F, \mu)}$$

from (28), (30) and (32). By letting

$$\nu := \begin{cases} \frac{1}{1-\mu}\left(1 - \frac{\tau}{2D_{\min}(F,\mu)}\right) & \tau \leq 2D_{\min}(F, \mu) \\ 0 & \tau > 2D_{\min}(F, \mu), \end{cases}$$

we obtain

$$D_{\min}(F',\mu) \geq \begin{cases} D_{\min}(F,\mu) - \tau + \frac{\tau^2}{4D_{\min}(F,\mu)} & \tau \leq 2D_{\min}(F,\mu) \\ 0 & \tau > 2D_{\min}(F,\mu) \end{cases}$$
$$\geq D_{\min}(F,\mu) - \tau. \qquad \square$$

We now give a proof of Theorem 2.

Proof of Theorem 2 Without loss of generality, we assume that $j=1$ and $\mu_2 = \max_{i=2,\ldots,K} \mu_i$, that is, Π_1 is the optimal and Π_2 is the second optimal arm. Then $\mu_1 = \mu_2 + \Delta_2 > \mu_2 \geq \mu_i$ for $i = 2,\ldots,K$. For notational convenience we denote the event that Π_i is pulled at the n-th round by $J_n(i) \equiv \{J_n = i\}$. Expectations and probabilities under $\boldsymbol{F}$ and the randomization in the policy are simply written as $\mathrm{E}[\cdot]$ and $P[\cdot]$.

We define events $A_n(i)$, B_n, C_n, D_n as follows.

$$A_n(i) \equiv \left\{ \hat{D}_i(n) \geq \frac{D_{\min}(F_i,\mu_1)}{1+\epsilon} \right\}, \quad i = 2,\ldots,K$$
$$B_n \equiv \{\hat{\mu}_1(n) \geq \mu_1 - \delta\}$$
$$C_n \equiv \left\{ \hat{\mu}_1(n) < \mu_1 - \delta \cap \max_{i=2,\ldots,K} \hat{\mu}_i(n) < \mu_1 - \delta \right\}$$
$$D_n \equiv \left\{ \hat{\mu}_1(n) < \mu_1 - \delta \cap \max_{i=2,\ldots,K} \hat{\mu}_i(n) \geq \mu_1 - \delta \right\},$$

where $\delta > 0$ is a constant satisfying

$$\delta < \min\left\{ \Delta_2, \frac{\epsilon(1-\mu_1)}{1+\epsilon} \min_{i\neq 1}\{D_{\min}(F_i,\mu_1)\} \right\}. \tag{33}$$

Now the regret can be written as

$$\mathrm{Regret}(N) = \sum_{n=1}^{N} \sum_{i=2}^{K} \Delta_i \mathbb{I}[J_n(i)]$$
$$= \sum_{n=1}^{N} \left(\sum_{i=2}^{K} \Delta_i \mathbb{I}[J_n(i) \cap B_n] + \sum_{i=2}^{K} \Delta_i \mathbb{I}[J_n(i) \cap B_n^c] \right),$$

where each term is bounded from above by

$$\sum_{i=2}^{K} \Delta_i \mathbb{I}[J_n(i) \cap B_n] \leq \sum_{i=2}^{K} \Delta_i \left(\mathbb{I}[J_n(i) \cap A_n(i)] + \mathbb{I}[J_n(i) \cap A_n^c(i) \cap B_n] \right) \tag{34}$$

and

$$\sum_{i=2}^{K} \Delta_i \mathbb{I}[J_n(i) \cap B_n^c] \leq \left(\max_i \Delta_i \right) \sum_{i=2}^{K} \mathbb{I}[J_n(i) \cap B_n^c]$$
$$\leq \left(\max_i \Delta_i \right) \mathbb{I}[B_n^c]$$

$$\leq \left(\max_i \Delta_i\right)(\mathbb{I}[C_n] + \mathbb{I}[D_n]) \quad (\text{as } B_n^c = C_n \cup D_n). \tag{35}$$

In Lemmas 8–11, which follow this proof, we bound the expected values of sums of the four terms on the right-hand sides of (34) and (35). From these lemmas we obtain

$$\begin{aligned}\mathrm{E}_{\boldsymbol{F}}[\mathrm{Regret}(N)] \leq{}& \sum_{i=2}^{K} \frac{\Delta_i(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)} + \sum_{i=2}^{K} \Delta_i\,(1+\gamma(|\mathrm{supp}(F_i)|, \eta_i(\epsilon,\delta)))\\ &+ \left(\max_i \Delta_i\right)\left(K\gamma\left(|\mathrm{supp}(F_1)|, \frac{\delta^2}{2}\right) + \frac{K^2}{1-\exp(-2(\Delta_2-\delta)^2)}\right)\end{aligned} \tag{36}$$

where $\gamma(x,y)$ $(x,y>0)$ is given by

$$\gamma(x,y) \equiv \sum_{t=1}^{\infty}(t+1)^x \exp(-yt)$$

and $\eta_i(\epsilon,\delta)$ is defined by (37) in Lemma 9 below. We complete the proof by defining $c(\epsilon, \boldsymbol{F})$ in Theorem 2 as the infimum over δ of the sum of the second and third terms of the right-hand side of (36). □

Remark 5 We assumed in Theorem 2 that there exists a single optimal arm. For the case of more than one optimal arms, optimality can be proved in a similar way by substituting $\hat{\mu}_1(n)$ in B_n, C_n with $\max_{j:\mu_j=\mu^*}\hat{\mu}_j(n)$. However, this makes the proofs even longer and so we do not give them in this paper.

Lemma 8 *For $i=2,\ldots,K$ it holds that*

$$\mathrm{E}\left[\sum_{n=1}^{N}\mathbb{I}[J_n(i)\cap A_n(i)]\right] \leq \frac{1+\epsilon}{D_{\min}(F_i,\mu_1)}\log N + 1$$

Lemma 9 *Define*

$$\eta_i(\epsilon,\delta) \equiv 2\left(d\left(\frac{\epsilon}{1+\epsilon}D_{\min}(F_i,\mu_1) - \frac{\delta}{1-\mu_1}; F_i, \mu_1\right)\right)^2, \tag{37}$$

where $d(\cdot;\cdot,\cdot)$ is given in (28). Then it holds for $i=2,\ldots,K$ that

$$\mathrm{E}\left[\sum_{n=1}^{N}\mathbb{I}[J_n(i)\cap A_n^c(i)\cap B_n]\right] \leq \gamma(|\mathrm{supp}(F_i)|, \eta_i(\epsilon,\delta)).$$

Lemma 10

$$\mathrm{E}\left[\sum_{n=1}^{N}\mathbb{I}[C_n]\right] \leq K\gamma\left(|\mathrm{supp}(F_1)|, \frac{\delta^2}{2}\right).$$

Lemma 11

$$\mathrm{E}\left[\sum_{n=1}^{N}\mathbb{I}[D_n]\right] \leq \frac{K^2}{1-\exp(-2(\Delta_2-\delta)^2)}.$$

Before proving these lemmas, we give intuitive interpretations for the four indicator functions in Lemmas 8–11.

$A_n(i)$ represents the event that the estimator $\hat{D}_i(n) = D_{\min}(\hat{F}_i(n), \hat{\mu}^*(n))$ of $D_{\min}(F_i, \mu^*)$ is already close to $D_{\min}(F_i, \mu^*)$ and Π_i is pulled with a small probability. After sufficiently many rounds $A_n(i)$ occurs with probability close to 1 and the term $\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap A_n(i)]$ is the main term for the regret.

The other terms represent events that each estimator is not yet close to the true value. The term involving C_n is essential for the consistency of the MED policy.

$A_n^c(i) \cap B_n$ represents the event that $\hat{D}_i(n)$ has not converged because $\hat{F}_i(n)$ is not close to F_i although $\hat{\mu}^*(n)$ is already close to μ_1. In this event Π_i is pulled and $\hat{F}_i(n)$ is updated more frequently. As a result, $A_n^c(i) \cap B_n$ happens only for a few n.

Similarly, D_n represents the event that $\hat{\mu}_i$ happens to be large for some $i \neq 1$. In this event $\hat{F}_i(n)$ is updated more frequently and D_n also happens only for a few n.

On the other hand, C_n represents the event that $\hat{\mu}_1$ is not yet close to μ_1. It requires many rounds for Π_1 to be pulled since Π_1 seems to be suboptimal in this event. Therefore, C_n may happen for many n.

Proof of Lemma 8 By partitioning $\mathbb{I}[J_n(i) \cap A_n(i)]$ according to the number of occurrences of the event $J_m(i) \cap A_m(i)$ before the n-th round (i.e., $\sum_{m=1}^{n-1} \mathbb{I}[J_m(i) \cap A_m(i)]$), we have

$$\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap A_n(i)] \leq \frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} + \sum_{n=1}^{N} \mathbb{I}\left[J_n(i) \cap A_n(i) \cap \left\{ \sum_{m=1}^{n-1} \mathbb{I}[J_m(i) \cap A_m(i)] > \frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} \right\} \right].$$

Since $\sum_{m=1}^{n-1} \mathbb{I}[J_m(i) \cap A_m(i)] \leq \sum_{m=1}^{n-1} \mathbb{I}[J_m(i)] = T_i'(n)$, we obtain

$$\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap A_n(i)] \leq \frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} + \sum_{n=1}^{N} \mathbb{I}\left[J_n(i) \cap A_n(i) \cap T_i'(n) > \frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} \right].$$

Taking the expected value, we have

$$\begin{aligned}
&\mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[J_n(i) \cap A_n(i)] \right] \\
&\leq \frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} + \sum_{n=1}^{N} P\left[J_n(i) \cap A_n(i) \cap T'(n) > \frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} \right] \\
&\leq \frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} + \sum_{n=1}^{N} P\left[J_n(i) \,\middle|\, A_n(i) \cap T_i'(n) > \frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} \right] \\
&\leq \frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} + N \exp\left(-\frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} \frac{D_{\min}(F_i, \mu_1)}{1+\epsilon} \right) \qquad \text{(by (9))} \\
&= \frac{(1+\epsilon)\log N}{D_{\min}(F_i, \mu_1)} + 1.
\end{aligned}$$

□

Proof of Lemma 9 Assume that $A_n^c(i) \cap B_n$ holds. Then

$$\begin{aligned}\frac{D_{\min}(F_i,\mu_1)}{1+\epsilon} &> D_{\min}(\hat{F}_i(n),\hat{\mu}^*(n))\\ &\geq D_{\min}(\hat{F}_i(n),\mu_1-\delta)\\ &\geq D_{\min}(\hat{F}_i(n),\mu_1)-\frac{\delta}{1-\mu_1} \quad \text{(by (24))}\end{aligned}$$

or equivalently

$$D_{\min}(\hat{F}_i(n),\mu_1)-D_{\min}(F_i,\mu_1) < -\left(\frac{\epsilon}{1+\epsilon}D_{\min}(F_i,\mu_1)-\frac{\delta}{1-\mu_1}\right). \tag{38}$$

From (33), the right-hand side of (38) is negative and

$$\begin{aligned}D(\hat{F}_i(n)\|F_i) &\geq 2\|\hat{F}_i(n)-F_i\|^2 \quad \text{(by Lemma 3)}\\ &\geq 2\left(d\left(\frac{\epsilon}{1+\epsilon}D_{\min}(F_i,\mu_1)-\frac{\delta}{1-\mu_1};F_i,\mu_1\right)\right)^2=\eta_i(\epsilon,\delta),\end{aligned}$$

where the last inequality follows since Lemma 7 implies that if $D_{\min}(F',\mu)-D_{\min}(F,\mu)<-\tau$, then $\|F'-F\|\geq d(\tau;F,\mu)$. Now we evaluate $\mathbb{I}[J_n(i)\cap A_n^c(i)\cap B_n]$ as

$$\begin{aligned}\sum_{n=1}^{N}\mathbb{I}[J_n(i)\cap A_n^c(i)\cap B_n] &\leq \sum_{t=1}^{\infty}\sum_{n=1}^{\infty}\mathbb{I}[J_n(i)\cap T_i'(n)=t\ \cap\ A_n^c(i)\cap B_n]\\ &\leq \sum_{t=1}^{\infty}\sum_{n=1}^{\infty}\mathbb{I}[J_n(i)\cap T_i'(n)=t\ \cap\ D(\hat{F}_{i,t}\|F_i)\geq\eta_i(\epsilon,\delta)]\\ &\leq \sum_{t=1}^{\infty}\mathbb{I}[D(\hat{F}_{i,t}\|F_i)\geq\eta_i(\epsilon,\delta)], \end{aligned}\tag{39}$$

where (39) follows because there is at most one n such that $J_n(i)\cap T_i'(n)=t$. Finally, we obtain

$$\begin{aligned}\mathrm{E}\left[\sum_{n=1}^{N}\mathbb{I}[J_n(i)\cap A_n^c(i)\cap B_n]\right] &\leq \sum_{t=1}^{\infty}P_{F_i}[D(\hat{F}_{i,t}\|F_i)\geq\eta_i(\epsilon,\delta)]\\ &\leq \sum_{t=1}^{\infty}(t+1)^{|\mathrm{supp}(F_i)|}\exp(-t\,\eta_i(\epsilon,\delta))\\ &\qquad \text{(by Lemmas 13 and 14 in Appendix A)}\\ &= \gamma(|\mathrm{supp}(F_i)|,\eta_i(\epsilon,\delta)).\end{aligned}$$

□

Proof of Lemma 10 From Lemma 14 in Appendix A it holds for any type $Q\in\mathcal{L}_t\subset\mathcal{A}$ that

$$P_{F_1}[\hat{F}_{1,t}=Q]\leq\exp(-t\,D(Q\|F_1))\leq\exp(-t\,D_{\min}(Q,\mu_1)). \tag{40}$$

Let $R_1 < \cdots < R_m$ be the smallest m integers in $\{n : T_1'(n) = t \cap C_n\}$. $(R_1, \ldots, R_m)$ is well defined on the event $m \le \sum_{n=1}^{\infty} \mathbb{I}[T_1'(n) = t \cap C_n]$. Note that

$$\left\{ \sum_{n=1}^{\infty} \mathbb{I}[T_1'(n) = t \cap C_n] \ge m \right\} \subset \left\{ T_1'(R_1) = \cdots = T_1'(R_m) \right\} \subset \bigcap_{l=1}^{m-1} \{ J_{R_l} \ne 1 \} \tag{41}$$

since $J_{R_l} = 1$ implies $T_1'(R_l) + 1 = T_1'(R_l + 1) \le T_1'(R_{l+1})$. Let $(r_1, \ldots, r_{m-1}) \in \mathbb{N}^{m-1}$ be a realization of $(R_1, \ldots, R_{m-1})$. (Recall that we write an event as, e.g., "$J_{r_k} \ne 1 \cap R_k = r_k$" instead of "$\{J_{r_k} \ne 1\} \cap \{R_k = r_k\}$".) Then we obtain for any $(r_1, \ldots, r_{m-1})$ that

$$\begin{aligned}
&P\left[\left\{ \sum_{n=1}^{\infty} \mathbb{I}[T_1'(n) = t \cap C_n] \ge m \right\} \cap (R_1, \ldots, R_{m-1}) = (r_1, \ldots, r_{m-1}) \cap \hat{F}_{1,t} = Q \right] \\
&\le P\left[\left\{ \bigcap_{l=1}^{m-1} \{ J_{r_l} \ne 1 \} \right\} \cap (R_1, \ldots, R_{m-1}) = (r_1, \ldots, r_{m-1}) \cap \hat{F}_{1,t} = Q \right] \\
&= P_{F_1}[\hat{F}_{1,t} = Q] \prod_{l=1}^{m-1} \left(P\left[R_l = r_l \,\middle|\, \bigcap_{k=1}^{l-1} \{ J_{r_k} \ne 1 \cap R_k = r_k \} \cap \hat{F}_{1,t} = Q \right] \right. \\
&\quad \left. \times P\left[J_{r_l} \ne 1 \,\middle|\, R_l = r_l \cap \bigcap_{k=1}^{l-1} \{ J_{r_k} \ne 1 \cap R_k = r_k \} \cap \hat{F}_{1,t} = Q \right] \right) \\
&\le P_{F_1}[\hat{F}_{1,t} = Q] \prod_{l=1}^{m-1} \left(P\left[R_l = r_l \,\middle|\, \bigcap_{k=1}^{l-1} \{ J_{r_k} \ne 1 \cap R_k = r_k \} \cap \hat{F}_{1,t} = Q \right] \right. \\
&\quad \left. \times \left(1 - \frac{1}{K} \exp(-t D_{\min}(Q, \mu_1 - \delta)) \right) \right) \qquad \text{(by (9) and } \hat{\mu}^*(R_l) < \mu_1 - \delta) \\
&= P_{F_1}[\hat{F}_{1,t} = Q] \left(1 - \frac{1}{K} \exp(-t D_{\min}(Q, \mu_1 - \delta)) \right)^{m-1} \\
&\quad \times \prod_{l=1}^{m-1} P\left[R_l = r_l \,\middle|\, \bigcap_{k=1}^{l-1} \{ J_{r_k} \ne 1 \cap R_k = r_l \} \cap \hat{F}_{1,t} = Q \right].
\end{aligned}$$

By taking the disjoint union of $(r_1, \ldots, r_{m-1}) \in \mathbb{N}^{m-1}$, we have

$$\begin{aligned}
&P\left[\left\{ \sum_{n=1}^{\infty} \mathbb{I}[T_1'(n) = t \cap C_n] \ge m \right\} \cap \hat{F}_{1,t} = Q \right] \\
&\le P_{F_1}[\hat{F}_{1,t} = Q] \left(1 - \frac{1}{K} \exp(-t D_{\min}(Q, \mu_1 - \delta)) \right)^{m-1}.
\end{aligned} \tag{42}$$

Note that $\mathrm{E}[X] = \sum_{m=1}^{\infty} P[X \geq m]$ for any nonnegative integer random variable X. Then we have

$$
\begin{aligned}
&\mathrm{E}\left[\sum_{n=1}^{\infty} \mathbb{I}[T_1'(n) = t \cap C_n]\right] \\
&= \sum_{Q \in \mathcal{L}_t : \mathrm{E}(Q) < \mu_1 - \delta} \sum_{m=1}^{\infty} P\left[\left\{\sum_{n=1}^{\infty} \mathbb{I}[T_1'(n) = t \cap C_n] \geq m\right\} \cap \hat{F}_{1,t} = Q\right] \\
&\leq \sum_{Q \in \mathcal{L}_t : \mathrm{E}(Q) < \mu_1 - \delta} \sum_{m=1}^{\infty} \exp(-t D_{\min}(Q, \mu_1)) \left(1 - \frac{1}{K} \exp(-t D_{\min}(Q, \mu_1 - \delta))\right)^{m-1} \\
&\quad \text{(by (40) and (42))} \\
&= K \sum_{Q \in \mathcal{L}_t : \mathrm{E}(Q) < \mu_1 - \delta} \exp\Big(-t\big(D_{\min}(Q, \mu_1) - D_{\min}(Q, \mu_1 - \delta)\big)\Big) \\
&\leq K \sum_{Q \in \mathcal{L}_t : \mathrm{E}(Q) < \mu_1 - \delta} \exp\left(-\frac{t\delta^2}{2}\right) \quad \text{(by (25))} \\
&\leq K (t+1)^{|\mathrm{supp}(F_1)|} \exp\left(-\frac{t\delta^2}{2}\right) \quad \text{(by Lemma 13 in Appendix A).}
\end{aligned} \tag{43}
$$

We complete the proof by

$$
\begin{aligned}
\mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[C_n]\right] &\leq \mathrm{E}\left[\sum_{t=1}^{\infty} \sum_{n=1}^{\infty} \mathbb{I}[T_1'(n) = t \cap C_n]\right] \\
&\leq \sum_{t=1}^{\infty} K (t+1)^{|\mathrm{supp}(F_1)|} \exp\left(-\frac{t\delta^2}{2}\right) \quad \text{(by (43))} \\
&= K \gamma\left(|\mathrm{supp}(F_1)|, \frac{\delta^2}{2}\right).
\end{aligned}
$$

□

Finally we prove Lemma 11 using Hoeffding's inequality (see, e.g,. Appendix B of Pollard 1984).

Lemma 12 (Hoeffding's Inequality) *Let $X_1, \ldots, X_n$ be i.i.d. random variables with supports in $[0, 1]$. Then it holds for all $a > 0$ that*

$$
\Pr\left[\frac{1}{n} \sum_{t=1}^{n} X_t - \mathrm{E}[X_1] \geq a\right] \leq \exp(-2na^2).
$$

Proof of Lemma 11 Since $D_n \subset \bigcup_{i=2}^{K} \{\hat{\mu}_i(n) = \hat{\mu}^*(n) > \mu_1 - \delta\}$, it holds that

$$\sum_{n=1}^{\infty} \mathbb{I}[D_n] \le \sum_{i=2}^{K} \sum_{n=1}^{\infty} \mathbb{I}[\hat{\mu}_i(n) = \hat{\mu}^*(n) > \mu_1 - \delta]$$

$$= \sum_{i=2}^{K} \sum_{t=1}^{\infty} \sum_{n=1}^{\infty} \mathbb{I}[T_i'(n) = t \cap \hat{\mu}_{i,t} = \hat{\mu}^*(n) > \mu_1 - \delta]. \tag{44}$$

Now we use a reasoning similar to that used to derive (42). Let $R_1 < \cdots < R_m$ be the smallest m integers in $\{n : T_i'(n) = t \cap \hat{\mu}_{i,t} = \hat{\mu}^*(n) > \mu_1 - \delta\}$. $(R_1, \ldots, R_m)$ is well defined on the event $m \le \sum_{n=1}^{\infty} \mathbb{I}[T_i'(n) = t \cap \hat{\mu}_{i,t} = \hat{\mu}^*(n) > \mu_1 - \delta]$. Note that

$$\left\{ \sum_{n=1}^{\infty} \mathbb{I}[T_i'(n) = t \cap \hat{\mu}_{i,t} = \hat{\mu}^*(n) > \mu_1 - \delta] \ge m \right\} \subset \bigcap_{l=1}^{m-1} \left\{ J_{R_l} \ne i \right\}$$

by the same argument as (41). Then we have

$$\begin{aligned}
&P\left[\sum_{n=1}^{\infty} \mathbb{I}[T_i'(n) = t \cap \hat{\mu}_{i,t} = \hat{\mu}^*(n) > \mu_1 - \delta] \ge m \right] \\
&= P_{F_i}[\hat{\mu}_{i,t} > \mu_1 - \delta] \\
&\quad \times P\left[\sum_{n=1}^{\infty} \mathbb{I}[T_i'(n) = t \cap \hat{\mu}_{i,t} = \hat{\mu}^*(n) > \mu_1 - \delta] \ge m \,\middle|\, \hat{\mu}_{i,t} > \mu_1 - \delta \right] \\
&\le P_{F_i}[\hat{\mu}_{i,t} > \mu_1 - \delta]\, P\left[\bigcap_{l=1}^{m-1} \left\{ J_{R_l} \ne i \right\} \,\middle|\, \hat{\mu}_{i,t} > \mu_1 - \delta \right] \\
&\le P_{F_i}[\hat{\mu}_{i,t} > \mu_1 - \delta] \left(1 - \frac{1}{K} \right)^{m-1} \quad (\text{by } \hat{\mu}_i(R_l) = \hat{\mu}^*(R_l) \text{ and (8)}).
\end{aligned} \tag{45}$$

Therefore,

$$\begin{aligned}
&\mathrm{E}\left[\sum_{n=1}^{\infty} \mathbb{I}[T_i'(n) = t \cap \hat{\mu}_{i,t} = \hat{\mu}^*(n) > \mu_1 - \delta] \right] \\
&= \sum_{m=1}^{\infty} P\left[\sum_{n=1}^{\infty} \mathbb{I}[T_i'(n) = t \cap \hat{\mu}_{i,t} = \hat{\mu}^*(n) > \mu_1 - \delta] \ge m \right] \\
&\le K\, P_{F_i}[\hat{\mu}_{i,t} > \mu_1 - \delta] \quad (\text{by (45)}) \\
&\le K\, P_{F_i}[\hat{\mu}_{i,t} > \mu_i + \Delta_2 - \delta] \quad (\text{by } \mu_1 = \mu_2 + \Delta_2 \ge \mu_i + \Delta_2) \\
&\le K \exp(-2t(\Delta_2 - \delta)^2) \quad (\text{by Hoeffding's inequality and } \Delta_2 - \delta > 0 \text{ from (33)}).
\end{aligned} \tag{46}$$

From (44) and (46) we obtain

$$\mathrm{E}\left[\sum_{n=1}^{N} \mathbb{I}[D_n] \right] \le \sum_{i=2}^{K} \sum_{t=1}^{\infty} K \exp(-2t(\Delta_2 - \delta)^2)$$

$$\le \frac{K^2}{1 - \exp(-2(\Delta_2 - \delta)^2)}. \qquad \square$$

4 Experiments

In this section, we present some simulation results for our MED policy and the UCB policies of Auer et al. (2002a).

First we give an algorithm for computing $D_{\min}(F,\mu)$ and ν^* with parameters ν_0, r, which we denote by $D_{\min}(F,\mu;\nu_0,r)$. Here ν_0 is an initial value of ν for the optimization in Theorem 3 and r is a stopping criterion for iterations. Recall that h, h' and h'' are defined in (12), (13) and (14).

[Computation of $D_{\min}(F,\mu;\nu_0,r)$]

Require: $r>0$, $\nu_0 \geq 0$;
  if $f_1 = 0$ and $\sum_{i\neq 1} f_i \frac{1-\mu}{1-x_i} \leq 1$ **then**
    return $\left(h\left(\frac{1}{1-\mu}\right), \frac{1}{1-\mu}\right)$;
  end if
  $\underline{\nu}, \nu := \frac{\mu - \mathrm{E}(F)}{\mu(1-\mu)}$; $\overline{\nu} := \frac{1}{1-\mu}$; $\nu_{\mathrm{prev}} := \infty$;
  if $\nu_0 \in (\underline{\nu}, \overline{\nu})$ **then**
    $\nu := \nu_0$;
  end if
  while $|\nu - \nu_{\mathrm{prev}}| > r$ **do**
    if $h'(\nu) > 0$ **then**
      $\underline{\nu} := \nu$;
    else
      $\overline{\nu} := \nu$;
    end if
    $\nu_{\mathrm{prev}} := \nu$; $\nu := \nu - h'(\nu)/h''(\nu)$;
    if $\nu \notin (\underline{\nu}, \overline{\nu})$ **then**
      $\nu := \frac{\underline{\nu}+\overline{\nu}}{2}$;
    end if
  end while
  return $\left(\max_{\nu' \in \{\underline{\nu}, \overline{\nu}, \nu\}} h(\nu'), \operatorname{argmax}_{\nu' \in \{\underline{\nu}, \overline{\nu}, \nu\}} h(\nu')\right)$;

In this algorithm, a lower and an upper bound of ν^* are given by $\underline{\nu}$ and $\overline{\nu}$, 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 $\underline{\nu}$, $\overline{\nu}$, the next point is determined by the bisection method, $\nu := (\underline{\nu} + \overline{\nu})/2$. The iteration stops if the current ν is close to the previous value of ν, given by ν_{prev}.

The computations of h, h' and h'' in the while loop are summations over $|\mathrm{supp}(F)|$ terms and are the main contributors to the complexity of this algorithm. In particular, they require $\mathrm{O}(T_i(n))$ ($\approx \mathrm{O}(\log n)$) computations for a continuous support model since $|\mathrm{supp}(\hat{F}_{i,t})| \leq t$. On the other hand, $D_{\min}(F,\mu)$ is differentiable with respect to μ (with slope ν^*) and the argument μ converges to μ^* after sufficiently many rounds. Therefore, it is reasonable to approximate $D_{\min}(F,\mu)$ by the previous value of $D_{\min}(F,\mu;\nu_0,r)$ until the variation of μ is small. From this point of view, we implemented our MED policy for our simulations in the following way:

[Linearly Approximated MED policy]
Parameter: Real $r, s > 0$.
Initialization:

1. Pull each arm once.

2. Set $(\hat{D}_i, \nu_i) := D_{\min}(\hat{F}_{i,1},\ \hat{\mu}^*(K+1); 0, r)$ and $m_i := \hat{\mu}^*(K+1)$ for each $i = 1, \ldots, K$.

Loop: For the n-th round,

1. Update variables for each i:
   - If $J_{n-1} \neq i$ and $|\hat{\mu}^*(n) - m_i| < s$ then $\hat{D}_i := \hat{D}_i + \nu_i(\hat{\mu}^*(n) - m_i)$.
   - Otherwise $(\hat{D}_i, \nu_i) := D_{\min}(\hat{F}_i(n),\ \hat{\mu}^*(n); \nu_i, r)$ and $m_i := \hat{\mu}^*(n)$.
2. Choose arm Π_j according to the probability

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

In this algorithm, $\hat{D}_i$ is an approximation of the current $D_{\min}(\hat{F}_i(n), \hat{\mu}^*(n))$, and m_i, ν_i are the values of $\hat{\mu}^*$, ν^* at the last round in which $\hat{D}_i$ is computed by the algorithm $D_{\min}(\cdot, \cdot; \cdot, \cdot)$. $\hat{D}_i$ is computed (without iteration) by the linear approximation using the derivative in Theorem 3 (iii) when the current $\mu^*(n)$ is close to m_i. Parameter s is the criterion for the approximation.

Now we describe the setting of our experiments. We used the (linearly approximated) MED, UCB-tuned and UCB2 policies. Each plot is an average over 1,000 different runs. The parameter α for UCB2 is set to 0.001; however, the choice of α does not have an important impact on the performance (see Auer et al. 2002a). First we check the effect of the choice of the parameters r and s. Then the MED and UCB policies are compared.

Table 1 gives the list of distributions used in the experiments. They cover various situations that affect the computation of $D_{\min}$ and change how distinguishable the optimal arm is. Distributions 1–4 are examples of 2-armed bandit problems.

In Distribution 1, $\nu^* \geq (\mu - \mathrm{E}(F))/(\mu(1-\mu))$ in Theorem 3 always holds with equality because $\mathrm{supp}(F_i) \subset \{0, 1\}$. Therefore the exact solution can be obtained by $D_{\min}(F, \mu;\ \nu_0, r)$ regardless of ν_0 and r. Also, in Distribution 2 $D_{\min}(F, \mu; \nu, r)$ does not require iteration after sufficiently many rounds since $\mathrm{E}_{F_2}[(1-\mu_1)/(1-X)] < 1$. On the other hand, in Distribution 3 $D_{\min}$ has to be computed numerically in almost all rounds since $\mathrm{E}_{F_2}[(1-\mu_1)/(1-X)] > 1$. As a result, for Distributions 1–3, the behavior of the Linearly Approximated MED policy is closest to the ideal MED policy in Distribution 1 and furthest from it in Distribution 3.

Distribution 4 is an example of a difficult problem where the optimal arm is hard to distinguish since the suboptimal arm appears to be optimal at first with high probability. Distributions 5 and 6 are examples of more general problems where the numbers of arms K and the support sizes are large. $\mathrm{Be}(\alpha, \beta)$ $(\alpha, \beta > 0)$ in Distribution 6 denotes the beta distribution which has the density function

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} \quad \text{for } x \in [0, 1]$$

where $\mathrm{B}(\alpha, \beta)$ is the beta function. Note that beta distributions have continuous support and are not included in $\mathcal{A}$ and, therefore, the performance of the MED policy is not assured theoretically. However, the MED policy is still formally applicable since the supports are bounded.

In Figs. 1–7 the label "regret" denotes $\sum_{i:\mu_i<\mu^*} \Delta_i T_i(n)$, which is the loss due to choosing suboptimal arms, while "regret per round" denotes $(1/n) \cdot \sum_{i:\mu_i<\mu^*} \Delta_i T_i(n)$, which is

Table 1 Distributions for experiments

Distribution 1:	
$F_1(\{0\}) = 0.45$, $F_1(\{1\}) = 0.55$	$\mathrm{E}(F_1) = 0.55$
$F_2(\{0\}) = 0.55$, $F_2(\{1\}) = 0.45$	$\mathrm{E}(F_2) = 0.45$
Distribution 2:	
$F_1(\{0.4\}) = 0.5$, $F_1(\{0.8\}) = 0.5$	$\mathrm{E}(F_1) = 0.6$
$F_2(\{0.2\}) = 0.5$, $F_2(\{0.6\}) = 0.5$	$\mathrm{E}(F_2) = 0.4$
Distribution 3:	
$F_1(\{x\}) = 0.08$ for $x = 0, 0.1, \ldots, 0.9$, $F_1(\{1\}) = 0.2$	$\mathrm{E}(F_2) = 0.56$
$F_2(\{x\}) = \frac{1}{11}$ for $x = 0, 0.1, \ldots, 0.9, 1$	$\mathrm{E}(F_2) = 0.5$
Distribution 4:	
$F_1(\{0\}) = 0.99$, $F_1(\{1\}) = 0.01$	$\mathrm{E}(F_1) = 0.01$
$F_2(\{0.008\}) = 0.5$, $F_2(\{0.009\}) = 0.5$	$\mathrm{E}(F_2) = 0.0085$
Distribution 5:	
$F_1(\{x\}) = 0.08$ for $x = 0, 0.1, \ldots, 0.9$, $F_1(\{1\}) = 0.2$	$\mathrm{E}(F_1) = 0.56$
$F_i(\{x\}) = \frac{1}{11}$ for $x = 0, 0.1, \ldots, 0.9, 1$	$\mathrm{E}(F_i) = 0.5$
	for $i = 2, 3, 4, 5$
Distribution 6:	
$F_1 = \mathrm{Be}(0.9, 0.1)$	$\mathrm{E}(F_1) = 0.9$
$F_2 = \mathrm{Be}(7, 3)$	$\mathrm{E}(F_2) = 0.7$
$F_3 = \mathrm{Be}(0.5, 0.5)$	$\mathrm{E}(F_3) = 0.5$
$F_4 = \mathrm{Be}(3, 7)$	$\mathrm{E}(F_4) = 0.3$
$F_5 = \mathrm{Be}(0.1, 0.9)$	$\mathrm{E}(F_5) = 0.1$

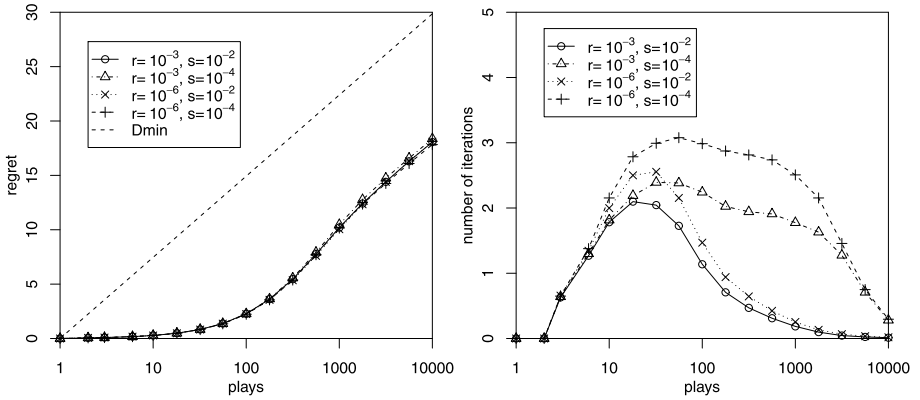

Fig. 1 Comparison of results for different parameters in the MED policy

suitable for observations of the regret at small rounds. The label "number of iterations" denotes the number of iterations of the while loop executed in $D_{\min}(\hat{F}_i(n), \hat{\mu}^*(n); v_i, r)$ at each round. The number is 0 when D_i is computed by the linear approximation. "Dmin" stands for the asymptotic bound for a consistent policy, $\sum_{i:\mu_i<\mu^*} \Delta_i \log n / D_{\min}(F_i, \mu^*)$. The asymptotic slope of the regret (in the semi-logarithmic plot) of a consistent policy is greater than or equal to that of "Dmin".

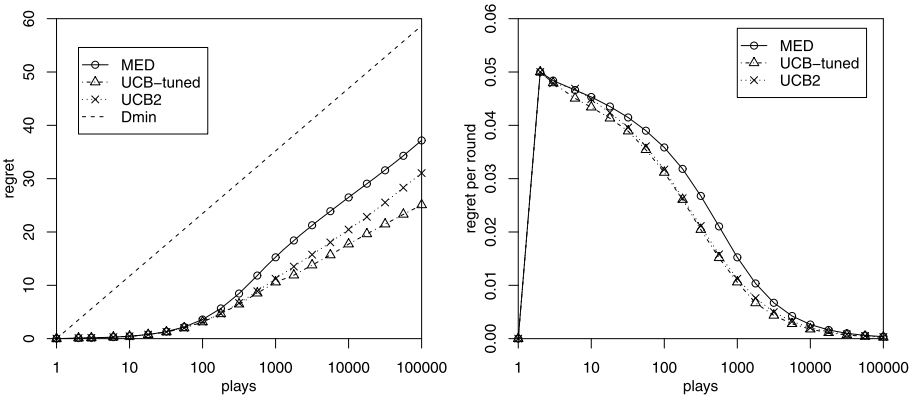

Fig. 2 Simulation results for Distribution 1 (Bernoulli distributions)

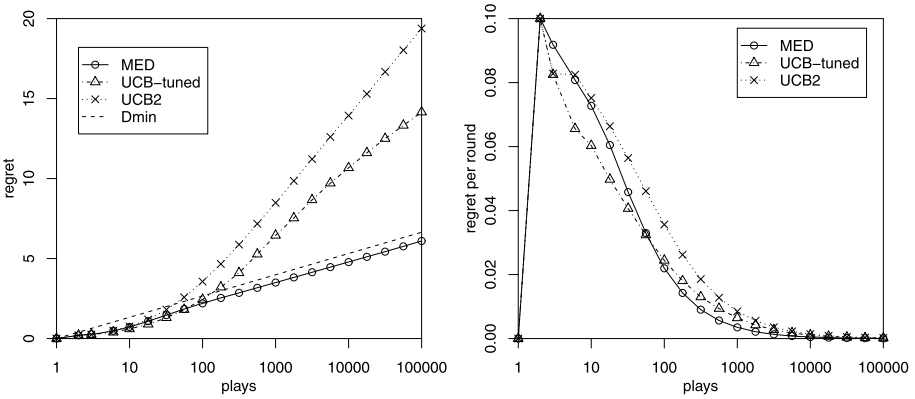

Fig. 3 Simulation results for Distribution 2 (uniform distributions with different supports)

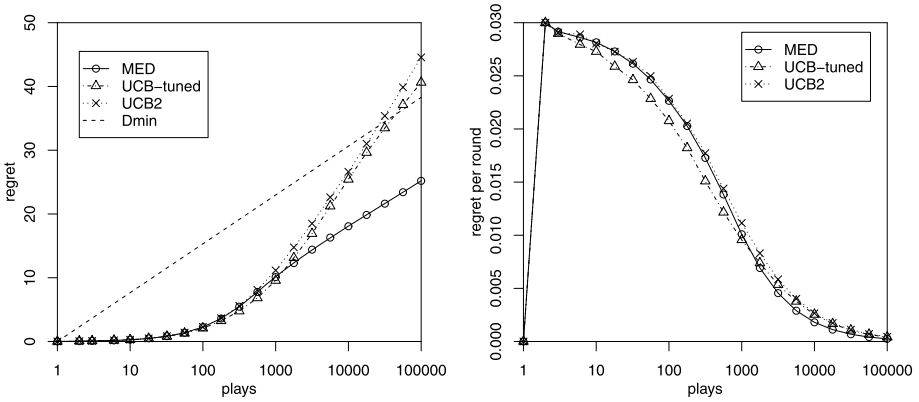

Fig. 4 Simulation results for Distribution 3 (distributions where $D_{\min}$ is computed by the while loop in the algorithm "Computation of $D_{\min}(F, \mu; \nu_0, r)$")

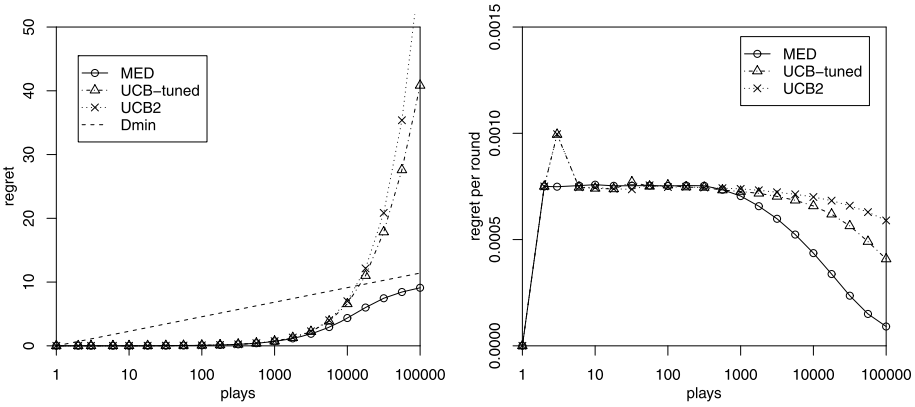

Fig. 5 Simulation results for Distribution 4 (very confusing distributions)

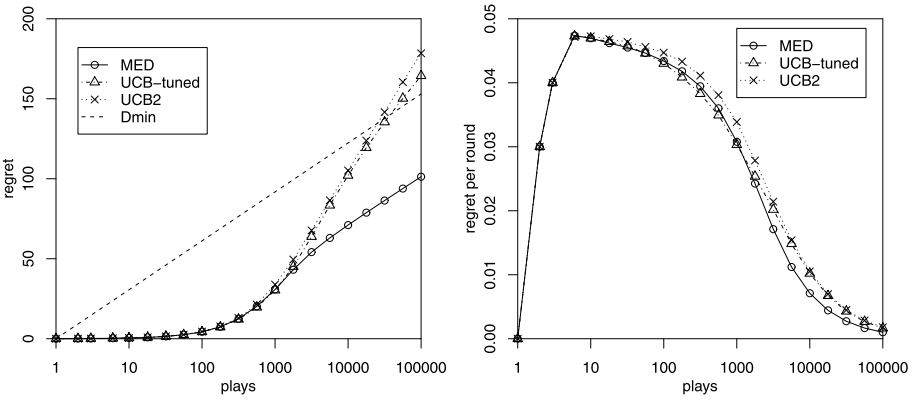

Fig. 6 Simulation results for Distribution 5 (5 arms with a wide support)

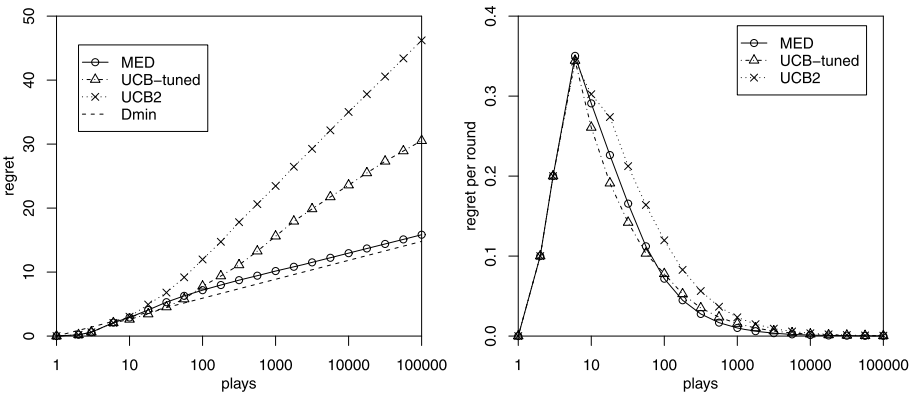

Fig. 7 Simulation results for Distribution 6 (beta distributions)

Table 2 Comparison of theoretical upper bounds for expected regrets

Distribution	MOSS	UCB1	UCB2	MED
1	1800	737	2.2×10^{13}	1.4×10^{13}
2	1219	369	1.1×10^{13}	3.3×10^{9}
3	2216	1228	3.7×10^{13}	2.9×10^{62}
4	6930	49121	1.5×10^{15}	1.5×10^{19}
5	10957	4913	1.5×10^{14}	9.7×10^{62}
6	8571	776	2.3×10^{13}	–

Figure 1 shows an experiment investigating the effects of the choice of the parameters r and s in the MED policy for Distribution 3. The Linearly Approximated MED policy approaches the ideal MED policy as $r, s \to 0$. However, we see from the left figure that the regret is not sensitive to the choice of r or s. In view of the computational complexity, the linear approximation with the criterion s seems to be effective (see the right-hand figure). On the other hand, the effect of varying r seems to be small compared to changing s. This may be because the initial value ν_i in $D_{\min}(\hat{F}_i(n),\ \hat{\mu}^*(n); \nu_i, r)$ is already sufficiently close to the optimal solution and the computation is usually completed in one iteration regardless of r. Based on these results, we use $s = 0.01$ and $r = 0.001$ in the remaining experiments.

Now we summarize the remaining experiments which compare the different policies (Figs. 2–7).

- The MED policy always seems to achieve the asymptotic bound even for continuous support distributions, since the asymptotic slope of the regret is close to that of "Dmin".
- The UCB-tuned works best in most cases when n is small.
- The MED policy eventually performs best, except for Distribution 1 where it performs worst. However, consistency is not proved for the UCB-tuned unlike for the MED and UCB2 policies. It appears that the UCB-tuned policy might not be consistent, because the asymptotic slope of the regret seems to be smaller than that of "Dmin". Note that the theoretical logarithmic terms of the regret are very close for the MED and UCB2 policies with Distribution 1 ($4.983 \log n$ and $5.025 \log n$, respectively). Therefore, this result can be interpreted as follows: the MED policy achieves the asymptotic bound but needs some improvement in the constant term of the regret compared to the UCB2 policy.

Finally, we mention that the upper bound of the expected regret given by (10) and (36) is very inaccurate because of the constant term $c(\epsilon, \boldsymbol{F})$. Table 2 denotes the theoretical upper bound of the expected regrets after the 10000th round of MOSS (Audibert and Bubeck 2009), UCB1 (Auer et al. 2002a), UCB2 and MED. The parameter α in UCB2 is set $\alpha = 0.001$, which is the same as in the simulations. The bound for MED is the infimum of (10) over ϵ. We see from the table that the theoretical upper bound of MED is quite large compared to the other policies, especially for the case that support size $|\mathrm{supp}(F)|$ is large. On the other hand, UCB1 and MOSS assure relatively reasonable regrets, although the coefficients of the logarithmic terms are larger than that of MED and UCB2.

From the simulation results in Figs. 2–6, we can conjecture that the expected regret of MED does not have such a huge value. Therefore, it is still important to derive a realistic finite-time regret for the MED policy.

5 Concluding remarks

We have proposed the minimum empirical divergence (MED) policy, and proved that it 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, we assumed the finiteness of the support although the MED policy also worked well in the simulations for distributions with continuous bounded support. We conjecture that the optimality of the MED policy holds also for the continuous bounded support model as it does for the DMED policy in Honda and Takemura (2010). In addition, there are many models for which $D_{\min}$ can be computed explicitly, such as the normal distribution model with unknown mean and variance. We expect that our MED policy can be extended to these models.

It is important to consider a finite-time regret for the finite horizon case. Although we derived a finite-time regret for the MED policy, it is still very inaccurate. Therefore, it is important to derive a better method for evaluating regret. Furthermore, the MED policy itself should be improved for the special setting of a finite horizon in which the number of rounds is given in advance. In this setting, the value of "exploration" becomes smaller and a current best arm should be pulled more often as the number of remaining rounds becomes smaller.

Acknowledgements We thank the reviewers for helpful comments, which have led to improvements in both our results and presentation. Junya Honda gratefully acknowledges support of JSPS Research Fellowships for Young Scientists. Akimichi Takemura acknowledges support of Aihara Project, the FIRST program from JSPS.

Appendix A: Method of types

Let $\mathcal{X} \subset \mathbb{R}$ be an arbitrary finite set and let F be an arbitrary probability distribution on $\mathcal{X}$, i.e., $\mathrm{supp}(F) \subset \mathcal{X}$. An empirical distribution $\hat{F}_n$ of n independent samples from F is called a *type*. The set of all possible types from n samples is denoted by $\mathcal{L}_n$.

Lemma 13 (Theorem 11.1.1 of Cover and Thomas 2006) $|\mathcal{L}_n| \leq (n+1)^{|\mathcal{X}|}$.

Lemma 14 (Theorem 11.1.4 of Cover and Thomas 2006) *For any type* $Q \in \mathcal{L}_n$,

$$(n+1)^{-|\mathcal{X}|} \exp(-nD(Q\|F)) \leq P_F[\hat{F}_n = Q] \leq \exp(-nD(Q\|F)).$$

Appendix B: Nonlinear optimization and sensitivity analysis

Consider an optimization problem

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) \geq 0, \quad i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0, \quad j = 1, \ldots, p \end{aligned}$$

for variable $\boldsymbol{x} \in \mathbb{R}^n$.

Theorem 4 (Lemma 3.2.1 of Fiacco 1983) *Assume that f, g and h are twice continuously differentiable in a neighborhood of $\boldsymbol{x}^*$. Then $\boldsymbol{x}^*$ is a strict local minimizing point of the problem if there exist Lagrange multiplier vectors $\boldsymbol{u}^* \in \mathbb{R}^m$ and $\boldsymbol{w}^* \in \mathbb{R}^p$ satisfying the second-order sufficient conditions for a strict local minimizing point, given by* (a) *and* (b) *below.*

(a) *First-order KKT conditions*:

$$\begin{aligned} u_i^* g_i(\boldsymbol{x}^*) = 0, \quad u_i^* \geq 0, \quad g_i(\boldsymbol{x}^*) \geq 0, \quad i = 1, \ldots, m, \\ h_j(\boldsymbol{x}^*) = 0, \quad j = 1, \ldots, p, \\ \nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{w}^*) = 0, \end{aligned}$$

where the Lagrangian function is given by

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{w}) \equiv f(\boldsymbol{x}) - \sum_{i=1}^{m} u_i g_i(\boldsymbol{x}) + \sum_{j=1}^{p} w_j h_j(\boldsymbol{x}).$$

(b)

$$\begin{aligned} \boldsymbol{z}^T \nabla_{\boldsymbol{x}}^2 L(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{w}^*) \boldsymbol{z} > 0 \text{ for all } \boldsymbol{z} \neq 0 \text{ such that} \qquad \text{(B.1)} \\ \nabla_{\boldsymbol{x}} g_i(\boldsymbol{x}^*) \boldsymbol{z} \geq 0 \text{ for all } i, \text{ where } g_i(\boldsymbol{x}^*) = 0, \\ \nabla_{\boldsymbol{x}} g_i(\boldsymbol{x}^*) \boldsymbol{z} = 0 \text{ for all } i, \text{ where } u_i^* > 0, \\ \nabla_{\boldsymbol{x}} h_j(\boldsymbol{x}^*) \boldsymbol{z} = 0 \text{ for all } j. \qquad \text{(B.2)} \end{aligned}$$

Now we regard $f(\boldsymbol{x})$ as $f(\boldsymbol{x}, \boldsymbol{\epsilon})$ and consider an unconstrained minimization problem P($\boldsymbol{\epsilon}$) with the optimal value $f^*(\boldsymbol{\epsilon}) = \min_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{\epsilon})$ for parameter $\boldsymbol{\epsilon} \in \mathbb{R}^k$.

Theorem 5 (Sensitivity analysis for unconstrained problem, Corollary 3.4.3 of Fiacco 1983) *Assume that $f(\boldsymbol{x}, \boldsymbol{\epsilon})$ is twice continuously differentiable in $(\boldsymbol{x}, \boldsymbol{\epsilon})$ in a neighborhood of $(\boldsymbol{x}^*, 0)$. If $\boldsymbol{x}^*$ satisfies the above second-order sufficient conditions for problem* P(0) *then, in a neighborhood of $\boldsymbol{\epsilon} = 0$, $f^*(\boldsymbol{\epsilon})$ is differentiable with respect to $\boldsymbol{\epsilon}$ and*

$$\nabla_{\boldsymbol{\epsilon}} f^*(\boldsymbol{\epsilon}) = \nabla_{\boldsymbol{\epsilon}} f(\boldsymbol{x}, \boldsymbol{\epsilon})\big|_{\boldsymbol{x}=\boldsymbol{x}^*}.$$

References

Agrawal, R. (1995a). The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, *33*, 1926–1951.

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

Audibert, J.-Y., & Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *Proceedings of COLT 2009*. Montreal: Omnipress.

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

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, *32*, 48–77.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

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

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd edn.). New York: Wiley-Interscience.

Even-Dar, E., Mannor, S., & Mansour, Y. (2002). Pac bounds for multi-armed bandit and Markov decision processes. In *Proceedings of COLT 2002* (pp. 255–270). London: Springer.

Fiacco, A. V. (1983). *Introduction to sensitivity and stability analysis in nonlinear programming*. New York: Academic Press.

Gittins, J. C. (1989). *Multi-armed bandit allocation indices. Wiley-Interscience Series in Systems and Optimization*. Chichester: Wiley.

Honda, J., & Takemura, A. (2010). An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of COLT 2010*, Haifa, Israel (pp. 67–79).

Ishikida, T., & Varaiya, P. (1994). Multi-armed bandit problem revisited. *Journal of Optimization Theory and Applications*, *83*, 113–154.

Katehakis, M. N., & Veinott, A. F. Jr. (1987). The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, *12*, 262–268.

Kleinberg, R. (2005). Nearly tight bounds for the continuum-armed bandit problem. In *Proceedings of NIPS 2005* (pp. 697–704). New York: MIT Press.

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

Meuleau, N., & Bourgine, P. (1999). Exploration of multi-state environments: Local measures and back-propagation of uncertainty. *Machine Learning*, *35*, 117–154.

Pollard, D. (1984). *Convergence of stochastic processes*. *Springer Series in Statistics*. New York: Springer.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*, 527–535.

Strens, M. (2000). A Bayesian framework for reinforcement learning. In *Proceedings of ICML 2000* (pp. 943–950). San Francisco: Kaufmann.

Vermorel, J., & Mohri, M. (2005). Multi-armed bandit algorithms and empirical evaluation. In *Proceedings of ECML 2005*, Porto, Portugal (pp. 437–448). Berlin: Springer.

Wyatt, J. (1997). *Exploration and inference in learning from reinforcement*. Doctoral dissertation, Department of Artificial Intelligence, University of Edinburgh.

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