An Asymptotically Optimal Bandit Algorithm for Bounded Support Models

Junya Honda and Akimichi Takemura
The University of Tokyo, Japan.
{honda,takemura}@stat.t.u-tokyo.ac.jp

Abstract

Multiarmed bandit problem is a typical example of a dilemma between exploration and exploitation in reinforcement learning. This problem is expressed as a model of a gambler playing a slot machine with multiple arms. We study stochastic bandit problem where each arm has a reward distribution supported in a known bounded interval, e.g. $[0, 1]$. In this model, Auer et al. (2002) proposed practical policies called UCB and derived finite-time regret of UCB policies. However, policies achieving the asymptotic bound given by Burnetas and Katehakis (1996) have been unknown for the model. We propose Deterministic Minimum Empirical Divergence (DMED) policy and prove that DMED achieves the asymptotic bound. Furthermore, the index used in DMED for choosing an arm can be computed easily by a convex optimization technique. Although we do not derive a finite-time regret, we confirm by simulations that DMED achieves a regret close to the asymptotic bound in finite time.

1 Introduction

The multiarmed bandit problem is a problem based on an analogy with a gambler playing a slot machine with more than one arm or lever. The objective of the 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.

We consider a K-armed stochastic bandit problem. There are K arms $\Pi_1, \ldots, \Pi_K$ and each Π_j has a probability distribution F_j with the expected value μ_j. The gambler chooses an arm to pull based on a policy and receives a reward according to F_j independently in each round. If the expected values of the arms are known, it is optimal to always pull the arm with the maximum expected value $\mu^* = \max_j \mu_j$. There have been many studies for this problem (Agrawal, 1995; Even-Dar et al., 2002; Meuleau & Bourgine, 1999; Strens, 2000; Vermorel & Mohri, 2005; Yakowitz & Lowe, 1991). There are also many extensions for the problem, such as non-stationary distributions (Gittins, 1989; Ishikida & Varaiya, 1994) and non-stochastic bandit (Auer et al., 2003).

Lai and Robbins (1985) constructed a theoretical framework for determining optimal policies. Burnetas and Katehakis (1996) extended their result to multiparameter or non-parametric models which is relevant to our setting. 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. Π_i is called suboptimal if $\mu_i < \mu^*$. 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 proved the following lower bound for any suboptimal Π_i under consistent policy:

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

with probability tending to one, 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 write

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

A policy is *asymptotically optimal* if the expected value of $T_j(n)$ achieves the right-hand side of (1) as $n \to \infty$. Lai and Robbins (1985) and Burnetas and Katehakis (1996) also proposed policies based on the notion of *upper confidence bound* and proved their optimality for some specific models. Furthermore, Auer et al. (2002) proposed some practical policies called UCB for other models. UCB policies estimate the expectation of each arm in a similar way to upper confidence bound. Note that this framework of the simple K-armed stochastic bandit problem is also important for applications. It is because efficient policies for this simple problem are also bases of some extended framework of bandit problems, such as Kleinberg (2005) for uncountable arms or Kleinberg et al. (2008) for the case that some arms can not be chosen at some rounds.

Now consider our model $\mathcal{A}$, the family of distributions on a known interval, e.g. $[0, 1]$. This model $\mathcal{A}$ represents one of the most basic nonparametric bandit situations. In this model, UCB policies are popular for their simple form and fine performance. However, although the performance of UCB policies is assured theoretically by a non-asymptotic form, their coefficients of the logarithmic term only depend on the expectations and the variances of arms and do not depend on the distributions themselves. Therefore, these theoretical analyses do not necessarily achieve the bound (1).

In this paper we propose Deterministic Minimum Empirical Divergence (DMED) policy. We prove the asymptotic optimality of DMED for our model $\mathcal{A}$. Although we do not give a finite bound of DMED as opposed to UCB policies, we confirm by simulations that DMED achieves performance close to the asymptotic bound in finite time.

Our DMED policy is motivated by a Bayesian viewpoint for the problem (although we do not use a Bayesian framework for theoretical analyses). Consider the case $K = 2$ and assume that Π_1 seems to be the best and $n \approx T_1(n) \gg T_2(n)$ at the n-th round. In this case the maximum likelihood that Π_1 and Π_2 are the best are roughly 1 and $\exp(-T_2(n)D_{\min}(\hat{F}_2, \hat{\mu}^*))$, respectively, where $\hat{F}_2$ is the empirical distribution of rewards from Π_2 and $\hat{\mu}^*$ is the current best sample mean. Then, the posterior expectation of the regret is proportional to $T_2(n) \cdot 1 + n \cdot \exp(-T_2(n)D_{\min}(\hat{F}_2, \hat{\mu}^*))$. DMED tries to minimize this by balancing these two terms. Note that DMED requires a computation of $D_{\min}(\hat{F}_i, \hat{\mu}^*) = \inf_{G \in \mathcal{A}: \mathrm{E}(G) \geq \hat{\mu}^*} D(\hat{F}_i || G)$ at each round. As shown in Theorem 8 below, $D_{\min}$ can be expressed as a univariate convex optimization problem and it can be computed efficiently .

This paper is organized as follows. In Section 2, we give definitions used throughout this paper and recall the asymptotic bound by Burnetas and Katehakis (1996). In Section 3, we propose DMED policy which achieves the asymptotic bound. In Section 4, we analyze $D_{\min}(F, \mu)$ as an optimal value function for a practical implementation and a proof of the optimality of DMED. In Section 5, we prove the asymptotic optimality of DMED by the results of Section 4. Some simulation results are shown in Section 6. We conclude the paper with some remarks in Section 7.

2 Preliminaries

In this section we introduce notation of this paper and present the asymptotic bound for a generic model, which is 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, $j = 1, \ldots, K$. $P_F[\cdot]$ and $\mathrm{E}_F[\cdot]$ denote the probability and the expectation under $F \in \mathcal{F}$, respectively. When we write e.g. $P_F[X \in A]$ $(A \subset \mathbb{R})$ 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 $\mathrm{E}(F) \equiv \mathrm{E}_F[X]$.

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.

The expected value of Π_j is denoted by $\mu_j \equiv \mathrm{E}(F_j)$. We denote the optimal expected value by $\mu^* \equiv \max_j \mu_j$. Let J_n denote 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.

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 rounds, respectively. $\hat{\mu}^*(n) \equiv \max_j \hat{\mu}_j(n)$ denotes the highest empirical mean after the first n rounds. We call Π_j a current best if $\hat{\mu}_j(n) = \hat{\mu}^*(n)$.

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 Kullback-Leibler divergence $D(F||G)$ 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 as follows.

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

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

Consequently

$$\liminf_{N\to\infty} \frac{\mathrm{E}_{\boldsymbol{F}}[T_i(N)]}{\log N} \geq \frac{1}{D_{\inf}(F_i, \mu^*, \mathcal{F})}. \tag{2}$$

3 An Asymptotically Optimal Policy

Let $\mathcal{A} \equiv \{G : \mathrm{supp}(G) \subset [a, b]\}$ be the family of distributions with a bounded support, where $\mathrm{supp}(G)$ is the support of distribution G and a, b are constants known to the player. We assume $a = 0$, $b = 1$ without loss of generality. We consider $\mathcal{A} = \{G : \mathrm{supp}(G) \subset [0, 1]\}$ as a model $\mathcal{F}$ for the rest of this paper.

When we adopt the model $\mathcal{A}$, 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 $D_{\inf}(F, \mu, \mathcal{A}) = \inf_{G\in\mathcal{A}:\mathrm{E}(G)>\mu} D(F||G)$.

Lemma 2 *$D_{\min}(F, \mu, \mathcal{A}) = D_{\inf}(F, \mu, \mathcal{A})$ holds for all $F \in \mathcal{A}$ and $\mu < 0$.*

Proof: $D_{\min}(F, \mu, \mathcal{A}) \leq D_{\inf}(F, \mu, \mathcal{A}) \leq D_{\min}(F, \mu + \epsilon, \mathcal{A})$ holds for arbitrary $\epsilon > 0$ from the definitions of $D_{\min}$ and $D_{\inf}$. $D_{\min}(F, \mu, \mathcal{A}) = D_{\inf}(F, \mu, \mathcal{A})$ follows by letting $\epsilon \downarrow 0$, since we will prove in Theorem 7 that $D_{\min}(F, \mu, \mathcal{A})$ is continuous in $\mu < 0$. ■

We simply write $D_{\min}(F, \mu) \equiv D_{\min}(F, \mu, \mathcal{A})$ when the third argument is obvious from the context. We discuss properties of $D_{\min}$ in Section 4.

Now we introduce Deterministic Minimum Empirical Divergence (DMED) policy and show its asymptotic optimality. We named it "deterministic" because our initial proposal, MED in Honda and Takemura (2010), was a randomized policy.

In the following algorithm, some arms are pulled once in one loop. Through the loop, arms to be pulled in the next loop are chosen and added to a list (a set) of arms. L_C denotes the list of arms to be pulled in the current loop. L_N denotes the list of arms to be pulled in the next loop. $L_R \subset L_C$ denotes the list of remaining arms of L_C which have not yet been pulled in the current loop. Arms are listed in L_N according to the occurrence of the event $J'_n(j)$ given by

$$J'_n(j) \equiv \{T_j(n) D_{\min}(\hat{F}_j(n), \hat{\mu}^*(n)) \leq \log n - \log T_j(n)\}. \tag{3}$$

[Deterministic Minimum Empirical Divergence Policy]

Initialization. $L_C, L_R := \{1, \cdots, K\}$, $L_N := \emptyset$. Pull each arm once. $n := K$.

Loop.

1. For $i \in L_C$ in the ascending order,
   1.1. $n := n + 1$ and pull Π_i. $L_R := L_R \setminus \{i\}$.
   1.2. $L_N := L_N \cup \{j\}$ (without a duplicate) for all $j \notin L_R$ such that $J'_n(j)$ occurs.
2. $L_C, L_R := L_N$ and $L_N := \emptyset$.

As shown above, $|L_C|$ arms are played in one loop. At every round, Π_j is added to L_N if $j \notin L_R$ and $J'_n(j)$ occurs. Note that if Π_j is a current best for the n-th round then $J'_n(j)$ holds since $D_{\min}(\hat{F}_j(n), \hat{\mu}^*(n)) = 0$ for this case. Then L_C is always not empty.

We use only the following fact as a property of DMED policy for our proof of the optimality:

Fact 3 (i) *For any n it holds that $\sum_{m=1}^{n} \mathbb{I}[J_m = j] \leq 2 + \sum_{m=1}^{n} \mathbb{I}[J'_m(j)]$.*
(ii) *If $J'_{n_0}(j)$ occurs for any n_0 then $T_j(n) \geq T_j(n_0) + 1$ for all $n \geq n_0 + K$.*

(i) and (ii) holds from the following reasons: (i) if Π_j is pulled at round $m > 2K$ then there exists a corresponding $n_m < m$ such that $J'_{n_m}(j)$ occurs and j is listed newly in L_N. The constant 2 is the effect of the initialization phase. (ii) There exists only three cases when $J'_{n_0}(j)$ occurs at the n_0-th round: (1) j is listed newly in L_N, (2) j is already listed in L_N, (3) j is listed in L_R. In each case Π_j is pulled at least once through $n_0 + 1, \cdots, n_0 + K$-th rounds and $T_j(n)$ is incremented.

Theorem 4 *Fix $\boldsymbol{F} \in \mathcal{A}^K$ for which there exists j such that $\mu_j = \mu^*$ and $\mu_i < \mu^*$ for all $i \neq j$. Under DMED policy, for any $i \neq j$ and $\epsilon > 0$ it holds that*

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

where $\mathrm{O}(1)$ denotes a constant dependent on ϵ and $\boldsymbol{F}$ but independent of N.

Note that we obtain

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

by dividing both sides by $\log N$, letting $N \to \infty$ and finally letting $\epsilon \downarrow 0$. In view of (2) we see that DMED policy is asymptotically optimal. We prove Theorem 4 in Section 5 by using results on $D_{\min}$ described in Section 4.

Note that the same bound as Theorem 4 can be derived when we substitute $\log T_j(n)$ in the criterion $J'_n(j)$ with an arbitrary constant or 0. However, we adopt the above criterion because simulation results seem better than that of other criteria. Our criterion may be justified by the Bayesian interpretation given in Introduction.

4 Analyses on Minimum Divergence

$D_{\min}(F, \mu)$ is the essential quantity for our DMED policy. In this section we introduce a dual problem $D'_{\min}(F, \mu)$ for $D_{\min}(F, \mu)$, which is computable efficiently. The main goal of this section is to show $D_{\min} = D'_{\min}$ and the continuity of them in F, μ. We discuss differentiability and continuity of $D'_{\min}(F, \mu)$ as a function of F and μ in Section 4.2. We show $D_{\min} = D'_{\min}$ in Section 4.3 by using the results of preceding subsections.

We now endow our model $\mathcal{A}$ with a distance to define the continuities of $D_{\min}(F, \mu)$ and $D'_{\min}(F, \mu)$ in $F \in \mathcal{A}$ and closedness of a subset of $\mathcal{A}$. We adopt Lévy distance

$$L(F, G) \equiv \inf\{h > 0 : F((-\infty, x-h]) - h \leq G((-\infty, x]) \leq F((-\infty, x+h]) + h \text{ for all } x\}$$

for the distance between two distributions. Note that the convergence of the Lévy distance $L(F, F_n) \to 0$ is equivalent to the weak convergence of $\{F_n\}$ to distribution F and we write $F_n \to F$ in this sense (see, e.g., Lamperti (1996) for detail).

4.1 A Dual Problem

For $\mu < 0$, define

$$\begin{aligned} H(\nu, F, \mu) &\equiv \mathrm{E}_F[\log(1 - (X - \mu)\nu)] \\ H'(\nu, F, \mu) &\equiv \frac{\partial H(\nu, F, \mu)}{\partial \nu} = -\mathrm{E}_F\left[\frac{X - \mu}{1 - (X - \mu)\nu}\right] \\ H''(\nu, F, \mu) &\equiv \frac{\partial^2 H(\nu, F, \mu)}{\partial \nu^2} = -\mathrm{E}_F\left[\frac{(X - \mu)^2}{(1 - (X - \mu)\nu)^2}\right] \end{aligned} \tag{4}$$

and

$$D'_{\min}(F, \mu) \equiv \max_{0 \leq \nu \leq \frac{1}{1-\mu}} H(\nu, F, \mu). \tag{5}$$

$D'_{\min}$ corresponds to the Lagrangian dual problem for $D_{\min}$. $D'_{\min}$ is a univariate convex optimization problem and it can be computed efficiently by iterative methods such as Newton's method (see, e.g., Boyd and Vandenberghe (2004) for general methods of convex programming).

We write $H(\nu), H'(\nu), H''(\nu)$ when we regard them as a function of ν and when other arguments are obvious from the context. Note that $H(\nu)$ is concave and strictly concave except for the degenerate case $F(\{\mu\}) = 1$ from (4). Now we define an optimal solution for (5) as

$$\nu^*(F,\mu) \equiv \operatorname*{argmax}_{0\le\nu\le\frac{1}{1-\mu}} H(\nu,F,\mu).$$

Note that $\nu^*(F,\mu)$ is unique except for the case $F(\{\mu\}) = 1$ from the strict concavity of $H(\nu,F,\mu)$ in ν. For the case $F(\{\mu\}) = 1$, $D'_{\min}(F,\mu) = H(\nu,F,\mu)$ holds for all $\nu \in [0,(1-\mu)^{-1}]$ and we define $\nu^*(F,\mu) \equiv (1-\mu)^{-1}$. We write $\nu^*(F)$ or more simply ν^* when other arguments are obvious from the context.

The following theorem is used through proofs in Section 4 and 5.

Theorem 5 *Define* $\mathrm{E}_F[(1-\mu)/(1-X)] = \infty$ *for the case* $F(\{1\}) > 0$. *If* $\mu \le \mathrm{E}(F)$ *then* $D'_{\min}(F,\mu) = 0$. *If* $\mathrm{E}(F) \le \mu$ *and* $\mathrm{E}_F[(1-\mu)/(1-X)] \le 1$ *then* $\nu^* = (1-\mu)^{-1}$ *and* (5) *is simply written as*

$$D'_{\min}(F,\mu) = H\left(\frac{1}{1-\mu}\right) = \mathrm{E}_F\left[\log\frac{1-X}{1-\mu}\right].$$

If $\mathrm{E}(F) \le \mu$ *and* $\mathrm{E}_F[(1-\mu)/(1-X)] \ge 1$ *then* ν^* *satisfies* $H'(\nu^*) = 0$ *and*

$$\mathrm{E}_F\left[\frac{1}{1-(X-\mu)\nu^*}\right] = 1, \qquad \mathrm{E}_F\left[\frac{X}{1-(X-\mu)\nu^*}\right] = \mu. \tag{6}$$

Proof: $D'_{\min}(F,\mu) = 0$ for $\mu \le \mathrm{E}(F)$ follows from $H(0) = 0, H'(0) = \mu - \mathrm{E}(F) \le 0$ and the concavity of $H(\nu)$. $\nu^* = (1-\mu)^{-1}$ for the case $\mathrm{E}_F[(1-\mu)/(1-X)] \le 1$ follows from $H'((1-\mu)^{-1}) = (1-\mu)(1-\mathrm{E}_F[(1-\mu)/(1-X)]) \ge 0$ and the concavity of $H(\nu)$.

Finally we consider the case $\mathrm{E}(F) \le \mu$ and $\mathrm{E}_F[(1-\mu)/(1-X)] \ge 1$. For this case $H'(0) = \mu - \mathrm{E}(F) \ge 0$ and $H'((1-\mu)^{-1}) = (1-\mu)(1-\mathrm{E}_F[(1-\mu)/(1-X)]) \le 0$. Therefore $H'(\nu^*) = 0$ hold from the concavity of $H(\nu)$. (6) follow from

$$\mathrm{E}_F\left[\frac{1}{1-(X-\mu)\nu^*}\right] = \mathrm{E}_F\left[\frac{1-(X-\mu)\nu^*}{1-(X-\mu)\nu^*}\right] + \nu^*\mathrm{E}_F\left[\frac{X-\mu}{1-(X-\mu)\nu^*}\right] = 1 - \nu^* H'(\nu^*) = 1$$

and

$$\mathrm{E}_F\left[\frac{X}{1-(X-\mu)\nu^*}\right] = \mathrm{E}_F\left[\frac{X-\mu}{1-(X-\mu)\nu^*}\right] + \mu\,\mathrm{E}_F\left[\frac{1}{1-(X-\mu)\nu^*}\right] = -H'(\nu^*) + \mu = \mu. \quad \blacksquare$$

4.2 Continuity and Differentiability of the Dual Problem

In this subsection we discuss the differentiability and the continuity of $D'_{\min}(F,\mu)$ in F and μ. We will show $D_{\min} = D'_{\min}$ in the next subsection and the result for $D'_{\min}$ in this subsection also holds for $D_{\min}$.

Theorem 6 $D'_{\min}(F,\mu)$ *is differentiable in* $\mu \in (\mathrm{E}(F),1)$ *for any* $F \in \mathcal{A}$ *with*

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

We omit the proof but it can be proved by Corollary 3.4.3 of Fiacco (1983), which gives the differentiability of an optimal value function with parameters.

Theorem 7 $D'_{\min}(F,\mu)$ *is continuous in* (i) $\mu < 1$ *and* (ii) $F \in \mathcal{A}$.

Proof: (i) The continuity in μ is obvious in the interval $(\mathrm{E}(F),1)$ from the differentiability in Theorem 6. The continuity in $\mu < \mathrm{E}(F)$ is also obvious since $D'_{\min}(F,\mu) = 0$ holds for all $\mu < \mathrm{E}(F)$. Finally we consider the continuity at $\mu = \mathrm{E}(F)$. From (5) and the concavity of $H(\nu)$, it holds that

$$H(0) \le D'_{\min}(F,\mu) \le \max\{H(0),\, H(0) + H'(0)\tfrac{1}{1-\mu}\}$$

or equivalently

$$0 \le D'_{\min}(F,\mu) \le \max\left\{0,\, \tfrac{\mu-\mathrm{E}(F)}{1-\mu}\right\}.$$

Then $\lim_{\mu\to\mathrm{E}(F)} D'_{\min}(F,\mu) = D'_{\min}(F,\mathrm{E}(F)) = 0$ is obtained by letting $\mu \to \mathrm{E}(F)$.

(ii) We consider the lower semicontinuity and the upper semicontinuity separately.

First we show the lower semicontinuity. Fix an arbitrary $\epsilon > 0$. From (5) and the continuity of $H(\nu)$, there exists $\nu_0 \in [0, (1-\mu)^{-1})$ such that $\mathrm{E}_F[\log(1-(X-\mu)\nu_0)] \geq D'_{\min}(F,\mu) - \epsilon$. Then we obtain

$$\begin{aligned} \liminf_{F' \to F} D'_{\min}(F',\mu) &\geq \liminf_{F' \to F} \mathrm{E}_{F'}[\log(1-(X-\mu)\nu_0)] \\ &= \mathrm{E}_F[\log(1-(X-\mu)\nu_0)] \\ &\geq D'_{\min}(F,\mu) - \epsilon. \end{aligned} \tag{7}$$

Note that $\log(1-(x-\mu)\nu_0)$ is continuous and bounded in $x \in [0,1]$ and (7) follows from the definition of weak convergence. The lower semicontinuity holds since ϵ is arbitrary.

Next we prove the upper semicontinuity. First we consider the case $\mathrm{E}(F) > \mu$. In this case, $\mathrm{E}(F') > \mu$ holds for all F' sufficiently close to F. Then $D'_{\min}(F,\mu) = D'_{\min}(F',\mu) = 0$ holds and the upper semicontinuity is obtained.

Next we consider the case $\mathrm{E}_F[(1-\mu)/(1-X)] > 1$ and $\mathrm{E}(F) \leq \mu$. Since $\nu^*(F) < (1-\mu)^{-1}$ in this case, we obtain

$$\begin{aligned} &\limsup_{F' \to F} D'_{\min}(F',\mu) \\ &\leq \limsup_{F' \to F} \left(H(\nu^*(F),F',\mu) + \tfrac{1}{1-\mu}\left|H'(\nu^*(F),F',\mu)\right| \right) \qquad \text{(by the concavity of } H(\nu)) \\ &= H(\nu^*(F),F,\mu) + \tfrac{1}{1-\mu}\left|H'(\nu^*(F),F,\mu)\right| \qquad \text{(by the definition of weak convergence)} \\ &= D'_{\min}(F,\mu) \end{aligned}$$

and the upper semicontinuity is proved for this case.

For the case $\mathrm{E}_F[(1-\mu)/(1-X)] \leq 1$, we omit the proof for lack of space. ■

The proof of the upper semicontinuity is a little complicated for the last case $\mathrm{E}_F[(1-\mu)/(1-X)] \leq 1$. It is because $\nu^* = (1-\mu)^{-1}$ holds for the case and $H(\nu,F,\mu)$ is difficult to analyze at $\nu = (1-\mu)^{-1}$. In fact, in every neighborhood of F, there exists $G \in \mathcal{A}$ such that $H((1-\mu)^{-1},G,\mu) = H'((1-\mu)^{-1},G,\mu) = -\infty$. The upper semicontinuity can be proved by using the definition of the Lévy distance explicitly.

4.3 Equality of Minimum Divergence with the Dual Problem

In this subsection we prove $D_{\min} = D'_{\min}$ in Theorem 8. Therefore we can compute $D_{\min}$ efficiently by solving the univariate convex optimization in $D'_{\min}$. Furthermore, the differentiability and the continuity in Theorem 6 and 7 also hold for $D_{\min}$.

Theorem 8 *$D_{\min}(F,\mu) = D'_{\min}(F,\mu)$ holds for all $F \in \mathcal{A}$ and $\mu < 1$.*

To prove this theorem, we additionally define $\mathcal{A}_f$ and $\mathcal{A}_f(F)$, families of distributions with finite supports by

$$\begin{aligned} \mathcal{A}_f &\equiv \{G \in \mathcal{A} : |\mathrm{supp}(G)| < \infty\}, \\ \mathcal{A}_f(F) &\equiv \{G \in \mathcal{A}_f : \mathrm{supp}(G) \subset \mathrm{supp}'(F)\} \quad (F \in \mathcal{A}_f) \end{aligned}$$

where $\mathrm{supp}'(F) \equiv \{1\} \cup \mathrm{supp}(F)$. Note that $\mathcal{A}_f(F) \subset \mathcal{A}_f \subset \mathcal{A}$ for all $F \in \mathcal{A}_f$.

Lemma 9 *$D_{\min}(F,\mu,\mathcal{A}) = D_{\min}(F,\mu,\mathcal{A}_f(F))$ holds for all $F \in \mathcal{A}_f$.*

We omit the proof but it can be proved by the following fact: if $G(A) \geq G'(A)$ for all $A \subset \mathrm{supp}(F)$ then $D(F||G) \leq D(F||G')$.

Before proving $D_{\min}(F,\mu) = D'_{\min}(F,\mu)$ for general $F \in \mathcal{A}$, we show the equality for $F \in \mathcal{A}_f$ and $\mathrm{E}(F) < \mu < 1$ by the technique of Lagrange multipliers.

Lemma 10 *If $\mathrm{E}(F) < \mu < 1$ and $F \in \mathcal{A}_f$ then $D_{\min}(F,\mu) = D'_{\min}(F,\mu)$ holds.*

Proof (Sketch): 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 $f_i \equiv F(\{x_i\})$. $D_{\min}(F,\mu)$ is expressed as the following parametric convex optimization problem for $G = (g_1, \ldots, g_M)$ from Lemma 9:

$$\text{minimize}: \sum_{i=1}^{M} f_i \log \frac{f_i}{g_i}, \qquad \text{subject to}: g_i \geq 0,\ \forall i,\ \sum_{i=1}^{M} x_i g_i \geq \mu,\ \sum_{i=1}^{M} g_i = 1.$$

It is checked by the technique of Lagrange multipliers (see e.g. Section 28 of Rockafellar (1970)) that the optimal solution is

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

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

$$g_i^* = \begin{cases} 0 & i = 1 \text{ and } f_1 = 0, \\ \frac{f_i}{1-(x_i-\mu)\nu^*} & \text{otherwise} \end{cases}$$

for the case $\mathrm{E}_F[(1-\mu)/(1-X)] \geq 1$ from (6). The lemma is proved immediately from these expressions of $\{g_i^*\}$. ■

Proof of Theorem 8: It is easy to check that $D_{\min}(F,\mu) = D'_{\min}(F,\mu) = 0$ for $\mu \leq \mathrm{E}(F)$. Hence we consider the case $\mathrm{E}(F) < \mu < 1$.

First we prove $D'_{\min}(F,\mu) \geq D_{\min}(F,\mu)$. Define a measure G^* on $[0,1]$ as

$$G^*(A) \equiv \begin{cases} \int_A \frac{1-\mu}{1-x} \mathrm{d}F + (1 - \mathrm{E}_F[(1-\mu)/(1-X)]) \mathbb{I}[0 \in A] & \mathrm{E}_F[(1-\mu)/(1-X)] \leq 1 \\ \int_A \frac{\mathrm{d}F}{1-(x-\mu)\nu^*} & \mathrm{E}_F[(1-\mu)/(1-X)] > 1. \end{cases}$$

To prove $D'_{\min}(F,\mu) \geq D_{\min}(F,\mu)$, it is sufficient to show that G^* is a probability measure with $\mathrm{E}(G^*) \geq \mu$ since $D(F||G^*) = D'_{\min}(F,\mu)$. It is checked easily for the case $\mathrm{E}_F[(1-\mu)/(1-X)] \leq 1$. For the case $\mathrm{E}_F[(1-\mu)/(1-X)] > 1$, it is checked from (6).

Next we prove $D_{\min}(F,\mu) \geq D'_{\min}(F,\mu)$. Take an arbitrary $G \in \mathcal{A}$ satisfying $\mathrm{E}(G) \geq \mu$. Consider a finite partition $\{U_i\}_{i=0,\cdots,n}$ of $[0,1]$:

$$U_i \equiv \begin{cases} \{0\} & i = 0 \\ \left(\frac{i-1}{n}, \frac{i}{n}\right] & i = 1, \cdots, n \end{cases}$$

and define $F^n, G^n \in \mathcal{A}_f$ as

$$F^n\left(\left\{\tfrac{i}{n}\right\}\right) \equiv F(U_i), \qquad G^n\left(\left\{\tfrac{i}{n}\right\}\right) \equiv G(U_i).$$

Then we have

$$\begin{aligned} D(F||G) &\geq D(F^n||G^n) && \text{(by Theorem 2.4.2 of Pinsker (1964))} \\ &\geq D_{\min}(F^n,\mu) && \text{(by } \mathrm{E}(G^n) \geq \mathrm{E}(G) \geq \mu) \\ &= D'_{\min}(F^n,\mu) && \text{(by Lemma 10)} \end{aligned} \tag{8}$$

Note that $L(F^n,F) \leq 1/n$ then $F^n \to F$ as $n \to \infty$. Therefore it holds for any $\epsilon > 0$ that

$$D'_{\min}(F^n,\mu) \geq D'_{\min}(F,\mu) - \epsilon \tag{9}$$

for sufficiently large n from the lower semicontinuity of $D'_{\min}(F,\mu)$ in F.

From (8) and (9) we obtain for all G satisfying $\mathrm{E}(G) \geq \mu$ that

$$D(F||G) \geq D'_{\min}(F,\mu) - \epsilon$$

and

$$D_{\min}(F,\mu) \geq D'_{\min}(F,\mu) - \epsilon.$$

$D_{\min}(F,\mu) \geq D'_{\min}(F,\mu)$ follows since $\epsilon > 0$ is arbitrary. ■

5 A Proof of Theorem 4

Before proving Theorem 4, we show Lemmas 11–14 on properties of $D_{\min}$ and ν^*.

Lemma 11 *$D_{\min}(F,\mu)$ is monotonically increasing in μ.*

This lemma follows immediately from the definition $D_{\min}(F,\mu) = \min_{G \in \mathcal{A}:\mathrm{E}(G) \geq \mu} D(F||G)$. We use this monotonicity in the proof of Theorem 4 implicitly.

Lemma 12 *If $\mathrm{E}(F) < \mu$ then $\nu^* = \nu^*(F,\mu)$ satisfies*

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

Proof: This lemma is easily checked for the case $\mathrm{E}_F[(1-\mu)/(1-X)] \leq 1$ from $\nu^* = (1-\mu)^{-1}$ and we consider the case $\mathrm{E}_F[(1-\mu)/(1-X)] \geq 1$. Define

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

For any fixed $\nu \in [0,(1-\mu)^{-1}]$, $w(x,\nu)$ is convex in $x \in [0,1]$. Therefore

$$\begin{aligned} H'(\nu) &= -\mathrm{E}_F[w(X,\nu)] \\ &\geq -\mathrm{E}_F\left[(1-X)w(0,\nu) + Xw(1,\nu)\right] \\ &= (\mathrm{E}(F)-1)w(0,\nu) - \mathrm{E}(F)w(1,\nu). \end{aligned} \tag{10}$$

The right-hand side of (10) is 0 for $\nu = (\mu - \mathrm{E}(F))/(\mu(1-\mu))$ and therefore we obtain

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

The lemma is proved since $H'(\nu^*) = 0$ holds and H' is monotonically decreasing. ■

Lemma 13 *Fix arbitrary $\mu, \mu' \in (0,1)$ satisfying $\mu' < \mu$. Then there exists $C(\mu,\mu') > 0$ such that*

$$D_{\min}(F,\mu) - D_{\min}(F,\mu') \geq C(\mu,\mu').$$

for all $F \in \mathcal{A}$ satisfying $\mathrm{E}(F) \leq \mu'$.

Proof: Since $D_{\min}(F,\mu)$ is differentiable in $\mu \in (\mathrm{E}(F),1)$ and continuous in $\mu < 1$, we have

$$\begin{aligned} D_{\min}(F,\mu) - D_{\min}(F,\mu') &= \lim_{t \downarrow \mu'} \int_t^\mu \frac{\partial}{\partial u} D_{\min}(F,u)\mathrm{d}u \\ &\geq \lim_{t \downarrow \mu'} \int_t^\mu \frac{u-\mu'}{u(1-u)}\mathrm{d}u \qquad \text{(by Theorem 5 and Lemma 12)} \\ &\geq \frac{(\mu-\mu')^2}{2\mu(1-\mu')} \ \left(=: C(\mu,\mu')\right). \end{aligned}$$

■

Lemma 14 $\sup_{G\in\mathcal{A}} D_{\min}(G,\mu) \leq -\log(1-\mu) < +\infty$ *for all* $0 \leq \mu < 1$.

Proof: By applying Jensen's inequality for

$$D_{\min}(G,\mu) = \max_{0 \leq \nu \leq \frac{1}{1-\mu}} \mathrm{E}_G[\log(1-(X-\mu)\nu)],$$

we obtain

$$\begin{aligned} \sup_{G\in\mathcal{A}} D_{\min}(G,\mu) &\leq \sup_{G\in\mathcal{A}} \max_{0 \leq \nu \leq \frac{1}{1-\mu}} \log(1-(\mathrm{E}(G)-\mu)\nu) \\ &= \sup_{G\in\mathcal{A}} \max\left\{0, \log\frac{1-\mathrm{E}(G)}{1-\mu}\right\} \\ &= -\log(1-\mu). \end{aligned}$$

■

Proof of Theorem 4: We assume $j=1$ and $\mu_2 = \max_{k\neq 1}\mu_k$ without loss of generality. Then $\mu_1 = \mu^*$ and $\mu_k \leq \mu_2$ for $k = 2,\cdots,K$.

Note that $\mu_1 = 1$ is a trivial case and we assume $\mu_1 < 1$ in the following. For the case $\mu_1 = 1$, $F_1(\{1\}) = 1$ and $\mu^*(n)$ is always equal to 0. Therefore $J_i'(n)$ never occurs for sufficiently large n, because $D_{\min}(\hat{F}_i(n),1) = +\infty$ always holds except for the case $\hat{F}_i(n) = F_1$.

We obtain from Fact 3 (i) that

$$T_i(N) \ = \ \sum_{n=1}^N \mathbb{I}[J_n = i] \ \leq \ 2 + \sum_{n=1}^N \mathbb{I}[J_n'(i)].$$

Now we define events A_n and B_n as

$$\begin{aligned} A_n &\equiv \{\hat{\mu}_1(n) \geq \mu_1 - \delta\}, \\ B_n &\equiv \{\hat{\mu}^*(n) \leq \mu_2 + \delta\} = \bigcap_{k=1}^K \{\hat{\mu}_k(n) \leq \mu_2 + \delta\}. \\ C_n &= \bigcup_{k=1}^K \{\hat{\mu}^*(n) = \hat{\mu}_k(n) \cap |\hat{\mu}_k(n) - \mu_k| \geq \delta\} \end{aligned}$$

where $\delta > 0$ is a constant satisfying $\mu_2 < \mu_1 - \delta$ and set sufficiently small in an evaluation on A_n. It is easily checked that $\{A_n^c \cap B_n^c\} \subset C_n$. Therefore $\mathrm{E}_{\boldsymbol{F}}[T_i(N)]$ is bounded as

$$\mathrm{E}_{\boldsymbol{F}}[T_i(N)] \quad \leq \quad 2 + \mathrm{E}_{\boldsymbol{F}}\left[\sum_{n=1}^{N} \mathbb{I}[J_n'(i) \cap A_n]\right] + \mathrm{E}_{\boldsymbol{F}}\left[\sum_{n=1}^{N} \mathbb{I}[B_n]\right] + \mathrm{E}_{\boldsymbol{F}}\left[\sum_{n=1}^{N} \mathbb{I}\left[C_n\right]\right]. \tag{11}$$

In the following Lemmas 15–17 we bound the right-hand side of (11) in this order and they prove the theorem. ■

Lemma 15 *For all $\epsilon > 0$ it holds that*

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

Lemma 16

$$\mathrm{E}_{\boldsymbol{F}}\left[\sum_{n=1}^{N} \mathbb{I}[B_n]\right] = \mathrm{O}(1).$$

Lemma 17

$$\mathrm{E}_{\boldsymbol{F}}\left[\sum_{n=1}^{N} \mathbb{I}[C_n]\right] = \mathrm{O}(1).$$

Before proving these lemmas, we give intuitive interpretations for these terms.

$\sum_{n=1}^{N} \mathbb{I}[J_n'(i) \cap A_n]$ is the main term of $T_i(N)$. Roughly speaking, in DMED policy, Π_i is pulled and $T_i(n)$ is incremented until $T_i(n) D_{\min}(\hat{F}_i(n), \hat{\mu}^*(n))$ and $\log n - \log T_i(n) (\approx \log n)$ in (3) balance. Consider the following two cases on the event A_n:

(1) If A_n happens and $\hat{F}_i$ is sufficiently close to F_i, then $D_{\min}(\hat{F}_i, \hat{\mu}^*) \gtrsim D_{\min}(F_i, \mu^*)$ holds and the above two terms balance when $T_i(n) \lesssim \log n / D_{\min}(F_i, \mu^*)$, which is exactly the asymptotic bound to be achieved.

(2) If A_n and $D_{\min}(\hat{F}_i, \hat{\mu}^*) < D_{\min}(F_i, \mu^*)$ happen, Π_i may be pulled more frequently than case (1). However, as Π_i is pulled, $\hat{F}_i$ approaches F_i and $D_{\min}(\hat{F}_i, \hat{\mu}^*)$ approaches $D_{\min}(F_i, \mu^*)$. Then eventually $D_{\min}(\hat{F}_i, \hat{\mu}^*) < D_{\min}(F_i, \mu^*)$ does not hold and the effect of this event is not large.

The term involving B_n is essential for the consistency of DMED. If B_n occurs then $\hat{\mu}_1(n)$ is not yet close to μ_1. It requires many rounds for Π_1 to be pulled since Π_1 may seem to be suboptimal in this event. Therefore B_n may happen for many n.

On the other hand when C_n occurs, empirical mean $\hat{\mu}_k(n)$ of current best Π_k is not close to the true expectation μ_k. Then Π_k is pulled more frequently and $\hat{\mu}_k(n)$ approaches μ_k. As a result, C_n happens only for a few n.

In the proofs of these three lemmas, we use Theorem 6.2.10 of Dembo and Zeitouni (1998) on the empirical distribution:

Proposition 18 (Sanov's Theorem) *For every closed set Γ of probability distributions (with respect to the Lévy distance),*

$$\limsup_{t \to \infty} \frac{1}{t} \log P_F[\hat{F}_t \in \Gamma] \leq - \inf_{G \in \Gamma} D(G||F).$$

where $\hat{F}_t$ is the empirical distribution of t samples from F.

Proof of Lemma 15 (Sketch): By partitioning the event $J_n'(i)$ according to the value of $T_i(n)$, we obtain

$$\begin{aligned}
&\sum_{n=1}^{N} \mathbb{I}[J_n'(i) \cap A_n] \\
&\quad = \sum_{t=1}^{N} \mathbb{I}\left[\bigcup_{n=1}^{N} \left\{\{t\, D_{\min}(\hat{F}_{i,t}, \hat{\mu}^*(n)) \leq \log n - \log t\} \cap A_n \cap T_i(n) = t\right\}\right]
\end{aligned}$$

$$\leq \frac{(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)} + \sum_{t=\frac{(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)}}^{N} \mathbb{I}\left[\bigcup_{n=1}^{N}\left\{\{tD_{\min}(\hat{F}_{i,t},\hat{\mu}^*(n)) \leq \log n\} \cap A_n \cap T_i(n)=t\right\}\right]$$

$$\leq \frac{(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)} + \sum_{t=\frac{(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)}}^{N} \mathbb{I}\left[\frac{(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)} D_{\min}(\hat{F}_{i,t},\mu_1-\delta) \leq \log N\right]$$

$$\left(\mu_1 - \delta \leq \hat{\mu}^*(n) \text{ on } A_n\right)$$

$$= \frac{(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)} + \sum_{t=\frac{(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)}}^{N} \mathbb{I}\left[D_{\min}(\hat{F}_{i,t},\mu_1-\delta) \leq \frac{D_{\min}(F_i,\mu_1)}{1+\epsilon}\right]. \tag{12}$$

Define $\Gamma^\delta \equiv \{G \in \mathcal{A} : L(F_i,G) \geq \delta\}$. By applying Sanov's Theorem with $F := F_i$ and $\Gamma := \Gamma^\delta$, there exists C_1 such that

$$P_{F_i}[\hat{F}_{i,t} \in \Gamma^\delta] = \mathrm{O}(\exp(-t\,C_1)). \tag{13}$$

Here we use the fact that for sufficiently small $\delta > 0$

$$\left\{D_{\min}(\hat{F}_{i,t},\mu_1-\delta) \leq \frac{D_{\min}(F_i,\mu_1)}{1+\epsilon}\right\} \subset \{\hat{F}_{i,t} \in \Gamma^\delta\} \tag{14}$$

or equivalently $\{L(\hat{F}_{i,t},F_i) < \delta\} \subset \{D_{\min}(\hat{F}_{i,t},\mu_1-\delta) > D_{\min}(F_i,\mu_1)/(1+\epsilon)\}$. This can be proved by the continuity in F and the differentiability in μ of $D_{\min}(F,\mu)$.

From (12), (13) and (14), we obtain

$$\mathrm{E}_{\boldsymbol{F}}\left[\sum_{n=1}^{N}\mathbb{I}[J_n(i) \cap A_n]\right] \leq \frac{(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)} + \sum_{t=\frac{(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)}}^{N} \mathrm{O}(\exp(-t\,C_1))$$

$$= \frac{(1+\epsilon)\log N}{D_{\min}(F_i,\mu_1)} + \mathrm{O}(1). \qquad \blacksquare$$

Proof of Lemma 16: Define $C_2 \equiv C(\mu_1,\mu_2+\delta)/3$ and $Q \equiv \lceil \sup_{G\in\mathcal{A}} D_{\min}(G,\mu_2+\delta)/C_2 \rceil$. $Q < +\infty$ holds from Lemma 14. Take a finite cover $\{S_q\}_{q=1,2,\cdots,Q}$ of $\{G \in \mathcal{A} : \mathrm{E}(G) \leq \mu_2+\delta\}$ as

$$S_q \equiv \{G \in \mathcal{A} : \mathrm{E}(G) \leq \mu_2+\delta,\ (q-1)C_2 \leq D_{\min}(G,\mu_2+\delta) \leq q\,C_2\}.$$

Since $D_{\min}(F,\mu)$ is continuous in F, each S_q is a closed set. By applying Sanov's Theorem with $F := F_1$ and $\Gamma := S_q$, there exists $t_q > 0$ such that for all $t > t_q$

$$\begin{aligned}
P_{F_1}[\hat{F}_{1,t} \in S_q] &\leq \exp\left(-t\Big(\inf_{G\in S_q} D(G||F_1) - C_2\Big)\right) \\
&\leq \exp\left(-t\Big(\inf_{G\in S_q} D_{\min}(G,\mu_1) - C_2\Big)\right) \\
&\leq \exp\left(-t\Big(\inf_{G\in S_q} D_{\min}(G,\mu_2+\delta) + C(\mu_1,\mu_2+\delta) - C_2\Big)\right) \text{ (by Lemma 13)} \\
&\leq \exp\left(-t(q+1)C_2\right). \qquad \left(\text{by } \inf_{G\in S_q} D_{\min}(G,\mu_2+\delta) \geq (q-1)C_2\right)
\end{aligned}$$

Therefore, by defining $t' \equiv \max_{q=1,\cdots,Q} t_q$, it holds for all $t > t'$ that

$$P_{F_1}[\hat{F}_{1,t} \in S_q] \leq \exp\left(-t(q+1)C_2\right). \tag{15}$$

$\sum_{n=1}^{N}\mathbb{I}[B_n]$ is bounded as

$$\sum_{n=1}^{\infty}\mathbb{I}[B_n] \leq \sum_{q=1}^{Q}\sum_{t=1}^{\infty}\sum_{n=1}^{\infty}\mathbb{I}[B_n \cap T_1(n)=t \cap \hat{F}_{1,t} \in S_q] \tag{16}$$

since $\{\hat{F}_1(n) \in \bigcup_{q=1}^{Q} S_q\} = \{\hat{\mu}_1(n) \leq \mu_2+\delta\} \supset B_n$. Now for each t and q we show that

$$\sum_{n=1}^{\infty}\mathbb{I}[B_n \cap T_1(n)=t \cap \hat{F}_{1,t} \in S_q] \leq t\exp(t\,q\,C_2) + K. \tag{17}$$

Assume that $\sum_{n=1}^{\infty} \mathbb{I}[B_n \cap T_1(n) = t \cap \hat{F}_{1,t} \in S_q] \geq t \exp(t\, q\, C_2)$. On this event, we can take an integer $m \geq t \exp(t\, q\, C_2)$ such that the events

$$B_m \cap T_1(m) = t \cap \hat{F}_{1,t} \in S_q \tag{18}$$

and

$$\sum_{n=1}^{m} \mathbb{I}[B_n \cap T_1(n) = t \cap \hat{F}_{1,t} \in S_q] = \lceil t \exp(t\, q\, C_2) \rceil \tag{19}$$

occur. For this m, it holds that

$$\begin{aligned} T_1(m) D_{\min}(\hat{F}_1(m), \hat{\mu}^*(m)) &\leq t \sup_{G \in S_q} D_{\min}(G, \mu_2 + \delta) \qquad \text{(by (18))} \\ &\leq t\, q\, C_2 \\ &\leq \log m - \log t. \qquad \text{(by } m \geq t \exp(t\, q\, C_2)\text{)} \end{aligned}$$

Then $J'_m(1)$ holds and $T_1(n) \geq t+1$ for all $n \geq m + K$ from Fact 3 (ii). Therefore we obtain (17) from

$$\begin{aligned} \sum_{n=1}^{\infty} \mathbb{I}[B_n \cap T_1(n) = t \cap \hat{F}_{1,t} \in S_q] &= \sum_{n=1}^{m+K-1} \mathbb{I}[B_n \cap T_1(n) = t \cap \hat{F}_{1,t} \in S_q] \\ &\leq \lceil t \exp(t\, q\, C_2) \rceil + K - 1 \qquad \text{(by (19))} \\ &\leq t \exp(t\, q\, C_2) + K. \end{aligned}$$

Now we obtain from (15), (16) and (17) that

$$\begin{aligned} \mathrm{E}_{\boldsymbol{F}}\left[\sum_{n=1}^{N} \mathbb{I}[B_n]\right] &\leq \sum_{q=1}^{Q} \sum_{t=1}^{\infty} P_{F_1}[\hat{F}_{1,t} \in S_q]\, (t \exp(t\, q\, C_2) + K) \\ &\leq \sum_{q=1}^{Q} \sum_{t=1}^{t'} (t \exp(t\, q\, C_2) + K) + \sum_{q=1}^{Q} \sum_{t=t'}^{\infty} \exp\left(-t(q+1)C_2\right) (t \exp(t\, q\, C_2) + K) \\ &\leq \mathrm{O}(1) + Q \sum_{t=t'}^{\infty} \Big(t \exp\left(-t\, C_2\right) + K \exp\left(-2\, t\, C_2\right) \Big) = \mathrm{O}(1). \end{aligned}$$

■

Proof of Lemma 17: We obtain from the definition of C_n that

$$\begin{aligned} \sum_{n=1}^{N} \mathbb{I}[C_n] &\leq \sum_{k=1}^{K} \sum_{n=1}^{\infty} \mathbb{I}[\hat{\mu}^*(n) = \hat{\mu}_k(n) \cap |\hat{\mu}_k(n) - \mu_k| \geq \delta] \\ &\leq \sum_{k=1}^{K} \sum_{t=1}^{\infty} \sum_{n=1}^{\infty} \mathbb{I}[\hat{\mu}^*(n) = \hat{\mu}_{k,t} \cap |\hat{\mu}_{k,t} - \mu_k| \geq \delta \cap T_k(n) = t]. \end{aligned}$$

Suppose that $\hat{\mu}^*(n_0) = \hat{\mu}_{k,t} \cap T_k(n_0) = t$ occurs at n_0-th round for the first time. Then Π_k is a current best at the n_0-th round and $J'_{n_0}(k)$ holds. Therefore $T_k(n) \geq t+1$ for all $n \geq n_0 + K$ from Fact 3 (ii). As a result, we obtain

$$\begin{aligned} &\sum_{n=1}^{\infty} \mathbb{I}[\hat{\mu}^*(n) = \hat{\mu}_{k,t} \cap |\hat{\mu}_{k,t} - \mu_k| \geq \delta \cap T_k(n) = t] \\ &\quad = \sum_{n=n_0}^{n_0+K-1} \mathbb{I}[\hat{\mu}^*(n) = \hat{\mu}_{k,t} \cap |\hat{\mu}_{k,t} - \mu_k| \geq \delta \cap T_k(n) = t] \leq K \end{aligned}$$

and

$$\mathrm{E}_{\boldsymbol{F}}\left[\sum_{n=1}^{N} \mathbb{I}[C_n]\right] \leq K \sum_{k=1}^{K} \sum_{t=1}^{\infty} P_{F_k}[\, |\hat{\mu}_{k,t} - \mu_k| \geq \delta].$$

By applying Sanov's Theorem with $F := F_k$ and $\Gamma := \{G \in \mathcal{A} : |\mathrm{E}(G) - \mu_k| \geq \delta\}$, there exists $C_3 > 0$ such that $P_{F_k}[\, |\hat{\mu}_{k,t} - \mu_k| \geq \delta] = \mathrm{O}(\exp(-t\, C_3))$. Now we obtain

$$\mathrm{E}_{\boldsymbol{F}}\left[\sum_{n=1}^{N} \mathbb{I}[C_n]\right] \leq K \sum_{k=1}^{K} \sum_{t=1}^{\infty} \mathrm{O}(\exp(-t\, C_3)) = \mathrm{O}(1).$$

■

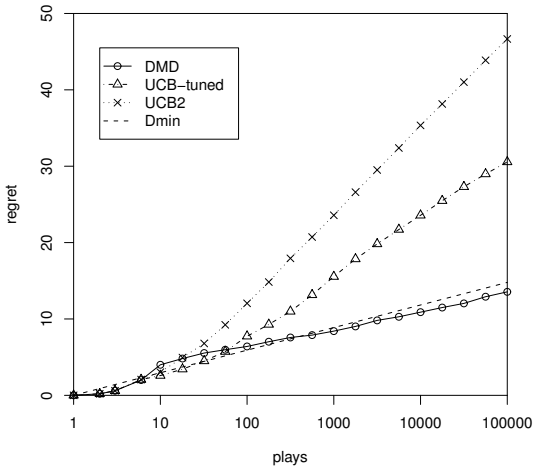

Figure 1: Experiment for beta distributions.

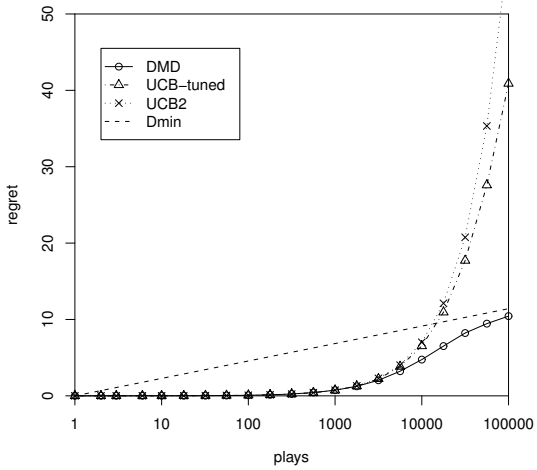

Figure 2: Experiment for distributions, which are hard to distinguish.

6 Experiments

In this section we give experimental results using UCB2, UCB-tuned (Auer et al., 2002) and DMED. In the implementation of DMED, $D_{\min}(\hat{F}_i(n), \hat{\mu}^*(n))$ has to be computed at each round. We can compute it by solving the dual problem discussed in Section 4 with e.g. Newton's method. We omit detailed description of our implementation of DMED but we note that $D_{\min}$ can be computed (or approximated) efficiently as follows:

(1) The optimal solution $\nu^*(\hat{F}_i(n-1), \hat{\mu}_i(n-1))$ of the previous round is a good approximation of the current $\nu^*(\hat{F}_i(n), \hat{\mu}_i(n))$ and the iteration for the optimization halts quickly.

(2) $\hat{\mu}^*(n)$ does not deviate significantly from μ^* for sufficiently large n and $D_{\min}(F, \mu)$ is differentiable in μ from Theorem 6. Therefore, $D_{\min}(\hat{F}_i, \hat{\mu}^*)$ can be approximated accurately by the linear approximation on $\hat{\mu}^*$ as long as $\hat{F}_i$ is not updated. On the other hand, $\hat{F}_i$ is updated only $\mathrm{O}(\log n)$ times through n rounds and its effect on the complexity is small.

Each plot is an average over 1,000 different runs. The labels of each figure are as follows. "regret" denotes $\sum_{i:\mu_i<\mu^*}(\mu^* - \mu_i)T_i(n)$, which is the loss due to choosing suboptimal arms. "Dmin" stands for the asymptotic bound for a consistent policy, $\sum_{i:\mu_i<\mu^*}(\mu^* - \mu_i)\log n / D_{\min}(F_i, \mu^*)$. The asymptotic slope of the regret (in the semi-logarithmic plot) of a consistent policy is more than or equal to that of "Dmin".

Figure 1 is a result for five arms with beta distributions. Beta distribution is an example of a simple continuous distribution on $[0, 1]$. Parameters for beta distributions are $(0.9, 0.1), (7, 3), (0.5, 0.5), (3, 7), (0.1, 0.9)$ and expectations are $\mu_i = 0.9,\ 0.7,\ 0.5,\ 0.3,\ 0.1$. Figure 2 is a result for two arms with discrete distributions

$$F_1(\{0\}) = 0.99, \quad F_1(\{1\}) = 0.01, \quad \mu_1 = 0.01,$$
$$F_2(\{0.008\}) = 0.5, \quad F_2(\{0.009\}) = 0.5, \quad \mu_2 = 0.0085.$$

It is an example of a problem where the optimal arm is hard to distinguish since the suboptimal arm appears to be optimal at first with high probability. We see from these figures that DMED achieves a regret near the asymptotic bound.

7 Conclusion

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

There are many models that $D_{\min}$ can be computed explicitly, such as normal distribution model with unknown mean and variance. We expect that our DMED can be extended to these models.

It is also important to consider the finite horizon case and to derive a finite-time bound of DMED. A finite-time bound may be derived by a non-asymptotic form of Sanov's Theorem in Exercise 6.2.19 of Dembo and Zeitouni (1998). However, its naive application makes the whole discussion extremely longer, e.g. the continuity of $D_{\min}(F, \mu)$ in F has to be of the form "if $L(F, F') \leq \epsilon$ then $|D_{\min}(F, \mu) - D_{\min}(F', \mu)| \leq \delta(\epsilon, F, \mu)$" with explicit $\delta(\cdot, \cdot, \cdot)$. Therefore other approaches may be more realistic.

References

Agrawal, R. (1995). The continuum-armed bandit problem. *SIAM J. Control Optim.*, *33*, 1926–1951.

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

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2003). The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, *32*, 48–77.

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

Burnetas, A. N., & Katehakis, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Adv. Appl. Math.*, *17*, 122–142.

Dembo, A., & Zeitouni, O. (1998). *Large deviations techniques and applications*, vol. 38 of *Applications of Mathematics*. New York: Springer-Verlag. Second edition.

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

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: John Wiley & Sons Ltd. With a foreword by Peter Whittle.

Honda, J., & Takemura, A. (2010). An asymptotically optimal policy for finite support models in the multi-armed bandit problem. Submitted to *Machine Learning*, arXiv:0905.2776v3.

Ishikida, T., & Varaiya, P. (1994). Multi-armed bandit problem revisited. *J. Optim. Theory Appl.*, *83*, 113–154.

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

Kleinberg, R. D., Niculescu-Mizil, A., & Sharma, Y. (2008). Regret bounds for sleeping experts and bandits. *Proceedings of COLT 2008* (pp. 425–436).

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

Lamperti, J. (1996). *Probability; a survey of the mathematical theory*. Wiley Series in Probability Statistics. New York: John Wiley & Sons Ltd. Second edition.

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

Pinsker, M. S. (1964). *Information and information stability of random variables and processes (transl.)*. San Francisco: Holden Day.

Rockafellar, R. T. (1970). *Convex analysis (Princeton Mathematical Series)*. Princeton University Press.

Strens, M. (2000). A bayesian framework for reinforcement learning. *Proceedings of ICML 2000* (pp. 943–950). Morgan Kaufmann, San Francisco, CA.

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

Yakowitz, S., & Lowe, W. (1991). Nonparametric bandit methods. *Ann. Oper. Res.*, *28*, 297–312.