Stochastic Bandit Based on Empirical Moments

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

Abstract

In the multiarmed bandit problem a gambler chooses an arm of a slot machine to pull considering a tradeoff between exploration and exploitation. We study the stochastic bandit problem where each arm has a reward distribution supported in $[0, 1]$. For this model, there exists a policy which achieves the theoretical bound asymptotically. However the optimal policy requires a computation of a convex optimization which involves the empirical distribution of each arm. In this paper, we propose a policy which exploits the first d empirical moments for arbitrary d fixed in advance. We show that the performance of the policy approaches the theoretical bound as d increases. This policy can be implemented by solving polynomial equations and we derive the explicit solution for d smaller than 5. By choosing appropriate d, the proposed policy realizes a tradeoff between the computational complexity and the expected regret.

1 Introduction

The multiarmed bandit problem is one of the formulations of the tradeoff between exploration and exploitation. This problem is based on an analogy with a gambler playing a slot machine with more than one arm. The gambler pulls arms sequentially so that the total reward is maximized.

We consider a K-armed stochastic bandit problem originally considered in Lai and Robbins (1985). There are K arms and each arm $i = 1, \cdots, K$ has a probability distribution F_i with the expected value μ_i. The gambler chooses an arm to pull based on a policy and receives a reward according to F_i independently in each round. For the largest expected value $\mu^* \equiv \max_i \mu_i$, we call an arm i optimal if $\mu_i = \mu^*$ and suboptimal if $\mu_i < \mu^*$. Then, the goal of the gambler is to maximize the sum of the rewards by pulling optimal arms as often as possible. Much research has been conducted for the stochastic bandit problem (Agrawal, 1995; Even-Dar et al., 2002; Strens, 2000; Vermorel & Mohri, 2005; Gittins, 1989) as well as the non-stochastic bandit problem(Auer et al., 2002b).

In this paper we consider the model $\mathcal{F}$, the family of distributions with supports contained in the bounded interval $[0, 1]$. The gambler knows that each distribution F_i is included in $\mathcal{F}$. For this model Upper Confidence Bound (UCB) policies are popular for their simple form and fine performance (Auer et al., 2002a; Audibert et al., 2009). Recently Honda and Takemura (2010) proposed Deterministic Minimum Empirical Divergence (DMED) policy which satisfies for arbitrary suboptimal arm i that

$$\mathrm{E}[T_i(n)] \leq \frac{1 + \mathrm{o}(1)}{D_{\min}(F_i, \mu^*)} \log n \,, \tag{1}$$

where $T_i(n)$ denotes the number of times that arm i has been pulled over the first n rounds and

$$D_{\min}(F, \mu) \equiv \min_{G \in \mathcal{F}: \mathrm{E}_G[X] \geq \mu} D(F \| G)$$

with Kullback-Leibler divergence $D(\cdot\|\cdot)$. DMED is asymptotically optimal since the coefficient $1/D_{\min}(F_i, \mu^*)$ of $\log n$ on the right-hand side of (1) coincides with the theoretical bound given in Burnetas and Katehakis (1996). However, the complexity of the DMED policy is still larger than e.g. UCB policies, although the computation involved in DMED is formulated as a univariate convex optimization problem. It is mainly because DMED requires the empirical distributions of the arms themselves whereas other popular policies can be computed only by the empirical moments of the arms, such as means and variances.

Now, our question is how we can bring the performance close to the right-hand side of (1) by a pol-

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

icy which only considers the first d empirical moments of the arms at each round. In this paper, we propose *DMED-M* policy which is a variant of DMED and is computable only by the empirical moments of the arms. For arbitrary suboptimal arm i, DMED-M satisfies

$$\mathrm{E}[T_i(n)] \leq \frac{1+\mathrm{o}(1)}{\inf_{F\in\mathcal{F}:\mathrm{E}^{(d)}(F)=\mathrm{E}^{(d)}(F_i)} D_{\min}(F,\mu^*)} \log n\,, \tag{2}$$

where $\mathrm{E}^{(d)}(F) \equiv (\mathrm{E}_F[X], \cdots, \mathrm{E}_F[X^d])$ denotes the first d moments of F and this upper bound approaches (1) as $d \to \infty$.

DMED-M is obtained by an analogy with DMED. Intuitively, DMED exploits the fact that the maximum likelihood that an arm i pulled t times is actually the best is roughly $\exp(-tD_{\min}(\hat{F}_i, \hat{\mu}^*))$, where $\hat{F}_i$ is the empirical distribution of the arm i and $\hat{\mu}^*$ is the currently best sample mean. When ignoring properties of the distribution $\hat{F}_i$ except for its first d moments, we overestimate the maximum likelihood as

$$\exp\Big(-t \inf_{F\in\mathcal{F}:\mathrm{E}^{(d)}(F)=\mathrm{E}^{(d)}(\hat{F}_i)} D_{\min}(F,\hat{\mu}^*)\Big)$$

instead of $\exp(-tD_{\min}(\hat{F}_i, \hat{\mu}^*))$ and the bound (2) appears correspondingly.

In DMED-M, it is necessary to compute the minimum $\inf_{F\in\mathcal{F}:\mathrm{E}^{(d)}(F)=(M_1,\cdots,M_d)} D_{\min}(F,\mu)$ for the argument $(M_1, \cdots, M_d) \in [0,1]^d$ at every round. Classical results on *Tchebycheff systems* and *moment spaces* reveal that the objective function $D_{\min}(\,\cdot\,,\mu)$ is contained in a class in which the optimal solution $\bar{F}$ is determined only by the value of the first d moments $\boldsymbol{M} = (M_1, \cdots, M_d)$. Therefore the infimum is obtained by computing firstly the optimal solution $\bar{F}$ and then the value of the function $D_{\min}(\bar{F},\mu)$. Both are obtained by solving polynomial equations and DMED-M can be computed efficiently for small d.

Note that the above minimization problem is written as a minimization of Kullbuck-Leibler divergence with moment constraints for two distributions, say

$$\inf_{\substack{F,\,G:\\ \mathrm{E}_G[X]\geq\mu\,,\\ \mathrm{E}_F[X^i]=M_i\,,\ i=1,\cdots,d}} D(F\|G)\,. \tag{3}$$

Such a minimization of a divergence function on moment constraints has been researched extensively (Csiszar & Matus, 2009). However, these results do not necessarily give an explicit solution of the minimization although they transform the minimization into a simpler form. Then, our result can be regarded as a special case that the minimization can be solved explicitly by the theory of Tchebycheff systems.

This paper is organized as follows. In Sect. 2, we give definitions used throughout this paper. We propose DMED-M policy in Sect. 3 and consider its practical implementation in Sect. 4. In Sect. 5, we discuss an improvement of DMED-M in terms of the worst case performance. We present some simulation results on DMED-M in Sect. 6. We conclude the paper with some remarks in Sect. 7. We introduce the theory of Tchebycheff systems and moment spaces and give a proof of the main theorem applying this theory in the supplementary material.

2 Preliminaries

Let $\mathcal{F}$ be the family of probability distributions on $[0,1]$ and $F_i \in \mathcal{F}$ be the distribution of the arm $i = 1, \ldots, K$. $\mathrm{E}_F[\cdot]$ denotes the expectation under $F \in \mathcal{F}$. When we write e.g. $\mathrm{E}_F[u(X)]$ for a function $u : \mathbb{R} \to \mathbb{R}$, X denotes a random variable with distribution F. The expected value of arm i is denoted by $\mu_i \equiv \mathrm{E}_{F_i}[X]$ and the optimal expected value is denoted by $\mu^* \equiv \max_i \mu_i$.

Let $T_i(n)$ be the number of times that arm i has been pulled through the first n rounds. $\hat{F}_i(n)$ and $\hat{\mu}_i(n)$ denote the empirical distribution and the mean of arm i after the first n rounds, respectively. The highest empirical mean after the first n rounds is denoted by $\hat{\mu}^*(n) \equiv \max_i \hat{\mu}_i(n)$. We call an arm i a current best if $\hat{\mu}_i(n) = \hat{\mu}^*(n)$.

Now we review results in Honda and Takemura (2010). Define an index for $F \in \mathcal{F}$ and $\mu \in [0,1]$

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

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

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

Under DMED policy proposed in Honda and Takemura (2010), the expectation of $T_i(n)$ for any suboptimal arm i is bounded as

$$\mathrm{E}[T_i(n)] \leq \frac{1+\epsilon}{D_{\min}(F_i,\mu^*)} \log n + \mathrm{O}(1)\,, \tag{4}$$

where $\epsilon > 0$ is arbitrary. The coefficient of the logarithmic term $1/D_{\min}(F_i,\mu^*)$ is the best possible (Burnetas & Katehakis, 1996) and the following property holds for the function $D_{\min}(F,\mu)$.

Proposition 1 (Honda and Takemura, (2010, Theorems 5 and 8))**.** *If* $\mathrm{E}_F[X] \geq \mu$ *then* $D_{\min}(F,\mu) = 0$. *If* $\mathrm{E}_F[X] < \mu = 1$ *then* $D_{\min}(F,\mu) = \infty$. *If* $\mathrm{E}_F[X] <$

$\mu < 1$,

$$D_{\min}(F,\mu) = \max_{0\le\nu\le\frac{1}{1-\mu}} \mathrm{E}_F[\log(1-(X-\mu)\nu)] = \begin{cases} \mathrm{E}_F[\log(1-X)] - \log(1-\mu) & \mathrm{E}_F\left[\frac{1}{1-X}\right] \le \frac{1}{1-\mu}, \\ \max_{0<\nu<\frac{1}{1-\mu}} \mathrm{E}_F[\log(1-(X-\mu)\nu)] & \text{otherwise}, \end{cases} \tag{5}$$

where we define $\log 0 = -\infty$ *and* $1/0 = +\infty$.

Let $\mathrm{E}^{(d)}(F) \equiv (\mathrm{E}_F[X], \cdots, \mathrm{E}_F[X^d])$ denote the first d moments of F. The set of distributions with the first d moments equal to $\boldsymbol{M} = (M_1, \cdots, M_d)$ is defined as $\mathcal{F}(\boldsymbol{M}) \equiv \{F \in \mathcal{F} : \mathrm{E}^{(d)}(F) = \boldsymbol{M}\}$. Now $D^{(d)}_{\min}(\boldsymbol{M},\mu)$ in (3) is written as

$$D^{(d)}_{\min}(\boldsymbol{M},\mu) = \inf_{F\in\mathcal{F}(\boldsymbol{M})} D_{\min}(F,\mu)\,.$$

This function $D^{(d)}_{\min}$ plays a central role throughout this paper.

3 DMED-M Policy

In this section we propose DMED-M policy. This policy determines an arm to pull based on the empirical moments of the arms. DMED-M requires computation of the function $D^{(d)}_{\min}$ and we analyze this function in the next section.

In Algorithm 1, each arm is pulled at most once in one loop. Through the loop, the list of arms pulled in the next loop is determined. 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. The criterion for choosing an arm i is the occurrence of the event $J_i(n)$ given by

$$J_i(n) \equiv \{T_i(n) D^{(d)}_{\min}(\mathrm{E}^{(d)}(\hat{F}_i(n)), \hat{\mu}^*(n)) \le \log n - \log T_i(n)\}\,, \tag{6}$$

where $\mathrm{E}^{(d)}(\hat{F}_i(n))$ represents the first d empirical moments of arm i.

As shown in the algorithm, $|L_C|$ arms are pulled in one loop. At every round, arm i is added to L_N if $J_i(n)$ occurs unless $i \in L_R$, that is, arm i is planned to be pulled in the remaining rounds in the current loop. Note that if arm i is a current best for the n-th round then $J_i(n)$ holds since $D^{(d)}_{\min}(\mathrm{E}^{(d)}(\hat{F}_i(n)), \hat{\mu}^*(n)) = 0$ for this case. Then L_C is never empty. Note

Algorithm 1 DMED-M Policy

Parameter: Integer $d > 0$.
Initialization:
$L_C, L_R := \{1, \cdots, K\}$, $L_N := \emptyset$, $n := K$.
Pull each arm once.
Loop:
1. For $i \in L_C$ in ascending order,
  1.1. $n := n+1$ and pull arm 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 $J_j(n)$ occurs.
2. $L_C, L_R := L_N$ and $L_N := \emptyset$.

that DMED in Honda and Takemura (2010) is obtained by replacing $D^{(d)}_{\min}(\mathrm{E}^{(d)}(\hat{F}_i(n)), \hat{\mu}^*(n))$ in (6) by $D_{\min}(\hat{F}_i(n), \hat{\mu}^*(n))$. In view of Theorem 2 below, DMED can be regarded as DMED-M with $d = \infty$.

Theorem 1. *Under DMED-M policy, for any suboptimal arm i and $\epsilon > 0$ it holds that*

$$\mathrm{E}[T_i(n)] \le \frac{1+\epsilon}{D^{(d)}_{\min}(\mathrm{E}^{(d)}(F_i), \mu^*)} \log n + \mathrm{O}(1) \tag{7}$$

where $\mathrm{O}(1)$ *denotes a constant independent of* n.

We can prove this theorem in a similar way as Theorem 4 of Honda and Takemura (2010) using the fact that $D_{\min}(F,\mu) \ge D^{(d)}_{\min}(\mathrm{E}^{(d)}(F),\mu)$ always holds. However, we omit the proof because it is long and very similar to the proof of Theorem 4 of Honda and Takemura (2010). The bound in Theorem 1 approaches that of DMED given by (4) as $d \to \infty$ from the following theorem, which we show in Appendix A .

Theorem 2. *For arbitrary $F \in \mathcal{F}$ it holds that*

$$\lim_{d\to\infty} D^{(d)}_{\min}(\mathrm{E}^{(d)}(F),\mu) = D_{\min}(F,\mu)\,.$$

Note that this theorem is not necessarily useful practically since the convergent speed is not mentioned. We examine the speed by simulations in Sect 6.

Remark 1. *The significance of this theorem rather lies in the consequence that DMED coincides with DMED-M with $d = \infty$. From this theorem we can infer that DMED-M will be asymptotically optimal among policies exploiting the first d moments of the arms, since it is true at least for $d = \infty$, although we were not able to prove it for finite d.*

4 Practical Implementation

For a computation and a theoretical evaluation of DMED-M, it is essential to analyze the function

$D^{(d)}_{\min}(\boldsymbol{M},\mu)$. In this section we study an explicit representation of this function and the complexity of DMED-M implemented by this representation.

4.1 Explicit Representation of $D^{(d)}_{\min}$

The key to the explicit representation is a theory of Tchebycheff systems (see the supplementary material for detail).

Define the *index* of a positive measure on $[0,1]$ as the size of its support under the special convention that the points $0,1$ are counted as one half. When considering an optimization under the moment constraints, the index plays an important role for the classification of the feasible region $\mathcal{F}(\boldsymbol{M}) = \{F \in \mathcal{F} : \mathrm{E}^{(d)}(F) = \boldsymbol{M}\}$.

First we consider a representation of $D^{(d)}_{\min}$ for a degenerate case.

Lemma 1. *If the index of $F \in \mathcal{F}$ is smaller than $(d+1)/2$ then $D^{(d)}_{\min}(\mathrm{E}^{(d)}(F),\mu) = D_{\min}(F,\mu)$.*

This lemma is straightforward from Prop. 4 in the supplementary material.

Now we consider a general case. Define a *principal representation* of $\boldsymbol{M}$ as a probability measure such that the first d moment is equal to $\boldsymbol{M}$ and its index is $(d+1)/2$. The principal representation is called *upper* if its support contains 1 and *lower* otherwise. There always exist precisely one lower and one upper principal representations except for the case in Lemma 1 (see Prop. 5). These representations are the key to the simple expression of $D^{(d)}_{\min}$.

Theorem 3. *Assume that the index of F is larger than or equal to $(d+1)/2$. Then,* (i) $D_{\min}(F,\mu)$, (ii) $-\mathrm{E}_F[X^{d+1}]$ *and* (iii) $-\mathrm{E}_F[1/(1-X)]$ *are minimized by the upper principal representation $\bar{F}$ over distributions with first d moments equal to $\boldsymbol{M} = \mathrm{E}^{(d)}(F)$. Similarly, they are maximized by the lower principal representation $\underline{F}$ over these distributions.*

We see from (i) (ii) of this theorem that $D_{\min}(F,\mu)$ is minimized by the distribution such that $(d+1)$-st moment is smallest. The meaning of (iii) is described in Sect. 5. An important point of this theorem is that the minimizer $\bar{F}$ of $D_{\min}(F,\mu)$ do not depend on the argument μ. Thus, we can decompose the computation of $D^{(d)}_{\min}(\boldsymbol{M},\mu)$ into two parts: the computation of the upper principal representation $\bar{F}$ and the value of function $D_{\min}(\bar{F},\mu)$.

For the former part, we see from the definition of the upper principal representation that the support and its weight $\{(x_i,\, f_i)\}_{i=1,\cdots,l}$ of $\bar{F}$ are the (unique) solution of

$$\sum_{i=1}^{l} f_i x_i^m = M_m \ (m = 0,\cdots,d),\ x_1 = 0,\ x_l = 1 \quad (8)$$

for odd d and

$$\sum_{i=1}^{l} f_i x_i^m = M_m \ (m = 0,\cdots,d), \qquad x_l = 1 \quad (9)$$

for even d, where $l = \lceil d/2 \rceil + 1$ and the zeroth moment is defined as $M_0 = 1$.

For the latter part, we eventually have to solve the maximization in (5) as in the case of DMED. However, since the argument $\bar{F}$ has a finite support, the optimal solution ν^* attaining the maximum is the solution of l-th degree polynomial equation

$$\frac{\mathrm{d}}{\mathrm{d}\nu}\mathrm{E}_{\bar{F}}[\log(1-(X-\mu)\nu)] = \frac{\sum_{i=1}^{l} f_i(\mu - x_i)\prod_{j\neq i}(1-(x_j-\mu)\nu)}{\prod_{i=1}^{l}(1-(x_i-\mu)\nu)} = 0\,. \quad (10)$$

We give an explicit form of $D^{(d)}_{\min}(\boldsymbol{M},\mu)$ for $d \le 4$ in the following theorem.

Theorem 4. *Assume that $M_1 < \mu < 1$ and the same condition as in Theorem 3 holds. Then $D^{(d)}_{\min}(\boldsymbol{M},\mu)$ is expressed for $d = 1, 2$ as*

$$D^{(1)}_{\min}(\boldsymbol{M},\mu) = (1-M_1)\log\frac{1-M_1}{1-\mu} + M_1\log\frac{M_1}{\mu}\,,$$

$$D^{(2)}_{\min}(\boldsymbol{M},\mu) = \frac{(1-M_1)^2}{1-2M_1+M_2}\log\left(1-\left(\frac{M_1-M_2}{1-M_1}-\mu\right)\nu^{(2)}\right) + \frac{M_2-M_1^2}{1-2M_1+M_2}\log\left(1-(1-\mu)\,\nu^{(2)}\right)$$

where

$$\nu^{(2)} = \frac{(1-M_1)(M_1-\mu)}{(1-M_1)\mu^2-(1-M_2)\mu+M_1-M_2}\,.$$

For $d = 3, 4$, it is expressed as

$$D^{(d)}_{\min}(\boldsymbol{M},\mu) = \sum_{l=1}^{3} f_l^{(d)}\log(1-(x_l^{(d)}-\mu)\nu^{(d)})\,,$$

where

$$\nu^{(d)} = \begin{cases} \frac{-b+\sqrt{b^2+4ac}}{2a}, & a \neq 0, \\ \frac{c}{b}, & a = 0, \end{cases}$$

for

$$a = (x_1^{(d)}-\mu)(x_2^{(d)}-\mu)(x_3^{(d)}-\mu)$$
$$b = (M_2+\mu M_1-2\mu^2) + (x_1^{(d)}+x_2^{(d)}+x_3^{(d)})(\mu-M_1)$$
$$c = \mu - M_1\,,$$

and $\{x_l^{(d)}\}$ *and* $\{f_l^{(d)}\}$ *are given as follows:*

$$\begin{aligned}(x_1^{(3)}, x_2^{(3)}, x_3^{(3)}) &= \left(0, \frac{M_2 - M_3}{M_1 - M_2}, 1\right)\\ f_2^{(3)} &= \frac{(M_1 - M_2)^3}{(M_2 - M_3)(M_1 - 2M_2 + M_3)}\\ f_3^{(3)} &= \frac{M_1 M_3 - M_2^2}{M_1 - 2M_2 + M_3}\\ f_1^{(3)} &= 1 - f_2^{(3)} - f_3^{(3)}, \end{aligned} \tag{11}$$

and

$$\begin{aligned}(x_1^{(4)}, x_2^{(4)}, x_3^{(4)}) &= \left(\frac{\beta - \sqrt{\beta^2 - 4\alpha}}{2}, \frac{\beta + \sqrt{\beta^2 - 4\alpha}}{2}, 1\right)\\ f_1^{(4)} &= \frac{x_2^{(4)}(M_1 - 1) + (M_1 - M_2)}{(x_1^{(4)} - 1)(x_2^{(4)} - x_1^{(4)})}\\ f_2^{(4)} &= \frac{-x_1^{(4)}(M_1 - 1) - (M_1 - M_2)}{(x_2^{(4)} - 1)(x_2^{(4)} - x_1^{(4)})}\\ f_3^{(4)} &= 1 - f_1^{(4)} - f_2^{(4)}, \end{aligned} \tag{12}$$

where

$$\alpha = \frac{-M_4(M_1 - M_2) + M_3(M_1 - M_3) - M_2(M_2 - M_3)}{M_2(M_1 - M_2) - M_1(M_1 - M_3) + (M_2 - M_3)}$$

$$\beta = \frac{M_2(M_2 - M_3) - M_1(M_2 - M_4) + (M_3 - M_4)}{M_2(M_1 - M_2) - M_1(M_1 - M_3) + (M_2 - M_3)}.$$

This theorem is obtained by solving (10) with the solution of (8) and (9) given in Lemma 2 below.

Lemma 2. *Assume that the same condition as in Theorem 4 holds. Then the solution* $(\{x_i^{(d)}\}, \{f_i^{(d)}\})$ *of* (8) *and* (9) *is expressed for* $d = 1, 2$ *as*

$$\begin{aligned}(x_1^{(1)}, x_2^{(1)}) &= (0, 1)\\ (f_1^{(1)}, f_2^{(1)}) &= (1 - M_1, M_1)\\ (x_1^{(2)}, x_2^{(2)}) &= \left(\frac{M_1 - M_2}{1 - M_1}, 1\right)\\ (f_1^{(2)}, f_2^{(2)}) &= \left(\frac{(1 - M_1)^2}{1 - 2M_1 + M_2}, \frac{M_2 - M_1^2}{1 - 2M_1 + M_2}\right), \end{aligned}$$

and is given by (11) *and* (12) *for* $d = 3, 4$.

This lemma can be confirmed by substitution of $(\{x_i^{(d)}\}, \{f_i^{(d)}\})$ into (8) and (9).

Remark 2. *For d more than 4, it is required to solve a polynomial equation of degree more than 2 and the explicit representation of* $D_{\min}^{(d)}$ *is unrealistic or unavailable. However, simulation results in Sect. 6 shows that* $D_{\min}(\mathrm{E}^{(d)}(F), \mu)$ *is very close to* $D_{\min}(F, \mu)$ *for* $d = 3, 4$ *and it seems to be sufficient to consider these degrees to achieve performance near DMED in practice.*

Remark 3. *The expression of* $D_{\min}^{(1)}(\boldsymbol{M}, \mu)$ *coincides with the KL divergence* $D(\mathrm{B}(M_1)\|\mathrm{B}(\mu))$ *between Bernoulli distributions with expectations* M_1 *and* μ*. The divergence between the Bernoulli distributions with these expectations is also considered by Garivier and Cappé (2011). In their paper, it is shown that their KL-UCB policy achieves the same asymptotic bound as* (7) *for* $d = 1$ *with a finite-time regret bound. In the viewpoint of our paper, KL-UCB can be extended formally so that first d moments can be exploited. Since UCB-type policies are generally good at performance for small rounds, these extensions may satisfy both of good performance for small rounds and fine asymptotic behavior.*

4.2 Complexity of DMED-M

In this subsection we first compare the complexity of DMED-M with DMED, i.e. DMED-M with $d = \infty$, and next examine the complexity of DMED-M for varying d.

In DMED, we have to compute

$$\begin{aligned}&D_{\min}(\hat{F}_i(n), \hat{\mu}^*(n))\\ &\quad = \max_{0 \le \nu \le \frac{1}{1-\mu}} \mathrm{E}_{\hat{F}_i(n)}[\log(1 - (X - \hat{\mu}^*(n))\nu)]\end{aligned}$$

for all currently suboptimal arms. Since the objective function is an expectation of the empirical distribution from $T_i(n) = \mathrm{O}(\log n)$ samples, it has complexity at least $\mathrm{O}(K \log n)$ per round.

On the other hand in DMED-M, $D_{\min}^{(d)}(\mathrm{E}^{(d)}(F), \hat{\mu}^*(n))$ is computed in constant time on n. It is because the empirical moments can be computed from the sums $\sum_{t=1}^{T_i(n)} X_{i,t}^m$, $(m = 1, \cdots, d)$ of samples $X_{i,1}, \cdots, X_{i,t}$ from arms i, which can be updated by d additions for each round. Therefore DMED-M has an advantage over DMED since the complexity do not grows with number of rounds.

Next we consider the complexity of DMED-M for varying degree d. The complexity for one computation of the solution of (8) or (9) is denoted by $C_1(d)$ and that of the solution of (10) is denoted by $C_2(d)$. They correspond to the complexity for computing the upper principal representation $\bar{F}$ from $\boldsymbol{M}$ and the value of the function $D_{\min}(\bar{F}, \mu)$, respectively.

It is difficult to formulate $C_1(d)$, the complexity of solving simultaneous polynomial equations for general d. However, since the principal representation only depends on the empirical moments of the arm, we have to compute it at most once per round. Furthermore, the principal representation has to be computed only for currently suboptimal arms. Since suboptimal arms are pulled at most $\mathrm{O}(\log n)$ times, the probability that

a currently suboptimal arm is pulled at n-th round is roughly $\mathrm{O}(1/n)$ from $\mathrm{d}(\log n)/\mathrm{d}n = 1/n$. As a result, the complexity coming from the computation of the principal representation is $\mathrm{O}(C_1(d)/n)$ per round and vanishes as n increases.

Next we consider $C_2(d)$, the complexity for solving (10), which is generally computed by an iterative method, such as Newton's method. Whereas it is necessary to compute for almost all rounds and arms, the arguments $\bar{F}$ and μ do not deviate from those of the last round very much. Then the iteration halts very quickly and the complexity mainly depends on the complexity of computing the objective function, which consists of $l = \lceil d/2 \rceil + 1 = \mathrm{O}(d)$ terms. Therefore $C_2(d) \approx \mathrm{O}(d)$ and the complexity coming from the computation of $D_{\min}(\bar{F}, \mu)$ is roughly $\mathrm{O}(Kd)$ per round. This is the complexity of DMED-M itself after sufficiently large rounds from the argument on $C_1(d)$.

5 Improvement of DMED-M Policy

In DMED-M, $D_{\min}(F, \mu)$ is bounded from below by $D^{(d)}_{\min}(\mathrm{E}^{(d)}(F), \mu)$. When the gap between $D^{(d)}_{\min}$ and $D_{\min}$ is small, DMED-M behaves like the asymptotically optimal policy, DMED. In this section, we propose DMED-MM policy which is obtained by a slight modification to DMED-M. We discuss that DMED-MM works successfully for the case where the gap between $D^{(d)}_{\min}$ and $D_{\min}$ is large.

Define a function $\tilde{D}^{(d)}_{\min}(F, \mu)$ by

$$\tilde{D}^{(d)}_{\min}(F, \mu) \equiv \begin{cases} D_{\min}(F, \mu) & \mathrm{E}_F\left[\frac{1}{1-X}\right] \le \frac{1}{1-\mu}, \\ D^{(d)}_{\min}(\mathrm{E}^{(d)}(F), \mu) & \text{otherwise}, \end{cases}$$

where recall that $D_{\min}(F, \mu) = \mathrm{E}_F[\log(1-X)] - \log(1-\mu)$ for the first case. *DMED-MM (DMED-M Mixed)* is the policy obtained by replacing $D^{(d)}_{\min}(\mathrm{E}^{(d)}(\hat{F}_i(n)), \mu)$ in DMED-M with $\tilde{D}^{(d)}_{\min}(\hat{F}_i(n), \mu)$. Then the criterion for choosing an arm is the same as DMED for the case $\mathrm{E}_{\hat{F}_i(n)}[1/(1-X)] \le 1/(1-\mu)$ and the same as DMED-M otherwise.

Note that the complexity of DMED-MM is almost the same as DMED-M since $\mathrm{E}_{\hat{F}_i(n)}[1/(1-X)]$ and $\mathrm{E}_{\hat{F}_i(n)}[\log(X-\mu)]$ can be computed in constant time from the sums $\sum_t 1/(1-X_{i,t})$ and $\sum_t \log(1-X_{i,t})$, which can be updated in constant time per round.

The relation between DMED, DMED-M and DMED-MM can be illustrated as follows. From Theorem 3 (i), we see that for the lower principal representation $\underline{F}$ DMED-M behaves most differently from DMED among distributions with common moments. On the other hand, from Theorem 3 (iii), the lower principal representation is the most typical measure satisfying $\mathrm{E}_F[1/(1-X)] \le 1$. Once this condition is satisfied, DMED-MM works the same as DMED. In this sense DMED-MM improves the worst case performance of DMED-M.

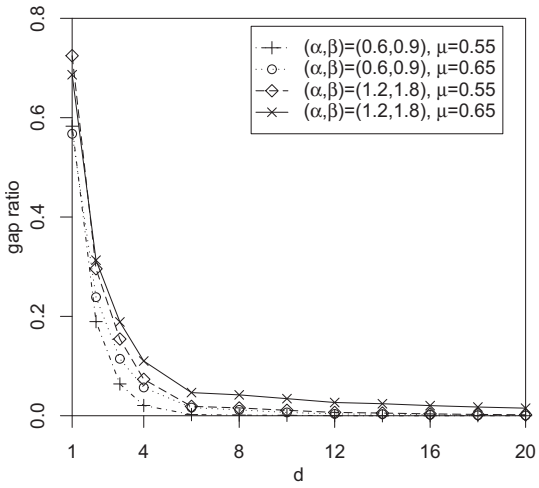

Figure 1: The ratio of gap $D_{\min}(F, \mu) - D^{(d)}_{\min}(\mathrm{E}^{(d)}(F), \mu)$ to $D_{\min}(F, \mu)$ for beta distributions $\mathrm{Be}(\alpha, \beta)$ as d increases.

6 Numerical Experiments

In this section we examine the properties of $D^{(d)}_{\min}$ and the performance of DMED-M and DMED-MM numerically. We use beta distributions $\mathrm{Be}(\alpha, \beta)$ for distributions of the arms since they cover various forms of distributions on $[0, 1]$.

First we examine the speed of the convergence of $D^{(d)}_{\min}(\mathrm{E}^{(d)}(F), \mu)$ to $D_{\min}(F, \mu)$. Fig. 1 shows $(D_{\min}(F, \mu) - D^{(d)}_{\min}(\mathrm{E}^{(d)}(F), \mu))/D_{\min}(F, \mu)$, i.e. the ratio of the gap between $D_{\min}$ and $D^{(d)}_{\min}$, as d increases for various F and μ. We used $\mu = 0.55, 0.65$ and beta distributions $\mathrm{Be}(0.6, 0.9)$, $\mathrm{Be}(1.2, 1.8)$ as F. These beta distributions have the same expectation 0.4, but the density of $\mathrm{Be}(0.6, 0.9)$ has two peaks at 0 and 1 whereas that of $\mathrm{Be}(1.2, 1.8)$ has one peak around its expectation. In these settings, the gap ratio is roughly less than 10% even for $d = 4$, where $D^{(d)}_{\min}$ is expressed explicitly. Note that the speed of the convergence is especially slow for the distribution $\mathrm{Be}(1.2, 1.8)$ and $\mu = 0.65$. This difference seems to come from the fact that only this setting satisfies $\mathrm{E}_F[(1-\mu)/(1-X)] \le 1$ among those in Fig. 1. As discussed in the previous section, we can compute $D_{\min}(F, \mu)$ without optimization for this case.

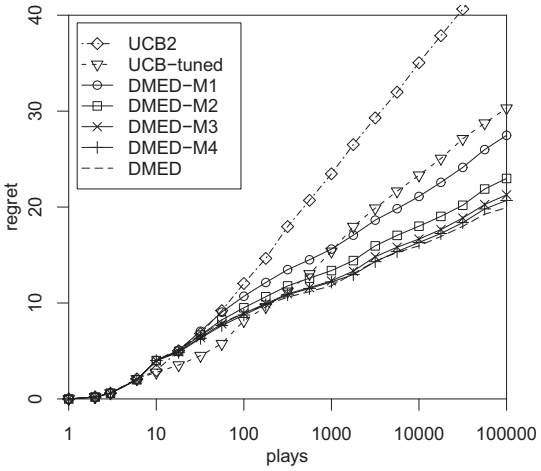

Figure 2: Empirical regrets for beta distributions with heavy weights around $x = 1$.

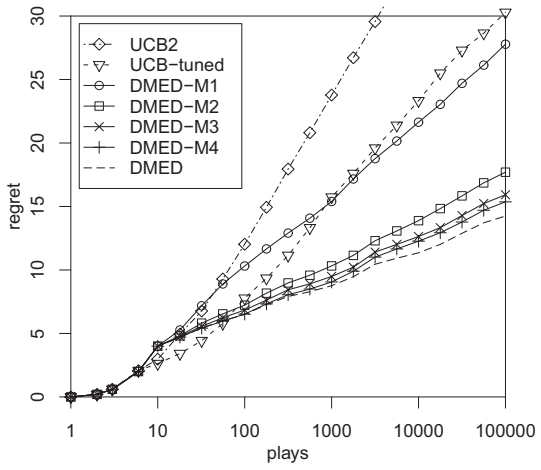

Figure 3: Empirical regrets for beta distributions with small weights around $x = 1$.

Next we compare the average regret of 1–4th degree DMED-M with DMED, UCB2, UCB-tuned (Auer et al., 2002a) for 5-armed bandit problems. Note that UCB2 is an example of a policy which determines an arm to pull based on the empirical means of the arms. Similarly, UCB-tuned is an example of a policy which chooses an arm based on the empirical means and variances.

In Figs. 2 and 3, each plot is an average over 1000 different runs. The vertical axis denotes the regret $\sum_{i:\mu_i<\mu^*}(\mu^* - \mu_i)T_i(n)$, which is the loss due to choosing suboptimal arms. Parameters of beta distributions are $(\alpha, \beta) = (0.45, 0.05), (0.35, 0.15), (0.25, 0.25), (0.15, 0.35), (0.05, 0.45)$ for Fig. 2 and $(\alpha, \beta) = (1.8, 0.2), (1.4, 0.6), (1, 1), (0.6, 1.4), (0.2, 1.8)$ for Fig. 3, where they have the same expectations $\mu = 0.9, 0.7, 0.5, 0.3, 0.1$.

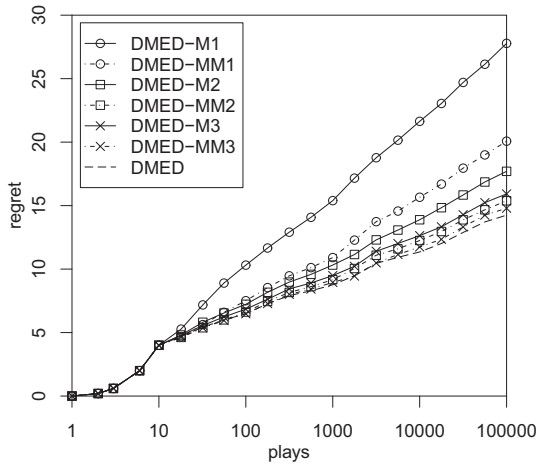

Figure 4: Comparison between DMED-M and DMED-MM.

We see from the figures that DMED-M works better than UCB polices for large rounds and approaches DMED as d increases. Note that the gap between DMED-M and DMED is larger in Fig. 3 than in Fig. 2. It seems to be because the distributions have smaller weights around $x = 1$ and are more different from upper principle representations in Fig. 3.

Finally we confirm that DMED-MM improves the DMED-M for the case that the gap between DMED-M and DMED is large. Fig. 4 shows a simulation result of DMED-M and DMED-MM for the same distributions as in Fig. 3. As shown in the figure, DMED-MM improves DMED-M significantly although the complexities of these policies are very similar.

7 Conclusion

In this paper we proposed DMED-M policy which is computed by the first d empirical moments of the arms. The theoretical bound of DMED-M approaches that of DMED, which is asymptotically optimal, as d increases. The computation involved in DMED-M is represented in an explicit form for $d \leq 4$. We also proposed DMED-MM policy which improves the worst case performance of DMED-M with small increase of the complexity.

An open problem is whether the asymptotic bound of DMED-M is the best for all policies which only consider the empirical moments. We may be able to prove the optimality of DMED-M in this sense under some regularity conditions.

A Proof of Theorem 2

Theorem 2 is proved by a basic result on weak convergence and Lévy distance (see, e.g., Lamperti (1996)). We say that a sequence of probability distributions $\{F_i\}$ converges weakly to F if $\lim_{i\to\infty} \mathrm{E}_{F_i}[u(X)] = \mathrm{E}_F[u(X)]$ for all bounded continuous functions $u(x)$. Define the Lévy distance $L(\cdot,\cdot)$ as

$$L(F,G) = \inf\{h>0 : \forall x, \\ F(x-h)-h \le G(x) \le F(x+h)+h\},$$

where $F(\cdot)$ and $G(\cdot)$ denote cumulative distribution functions. A weak convergence is equivalent to the convergence of the Lévy distance, that is, $\{F_i\}$ converges weakly to F if and only if $\lim_{i\to\infty} L(F_i,F)=0$.

Proposition 2 (Honda and Takemura (2010), Theorem 7)**.** *$D_{\min}(F,\mu)$ is continuous in $F\in\mathcal{F}$ with respect to the Lévy distance.*

Now we prove Theorem 2 by Prop. 2. In the following proof we write $\boldsymbol{M}^{(d)}$ for $(M_1,\cdots,M_d)$ instead of $\boldsymbol{M}$ to clarify the length of the vector.

Proof of Theorem 2. Define

$$L^{(d)}(F) = \sup_{G\in\mathcal{F}(\boldsymbol{M}^{(d)})} L(G,F)$$

for $\boldsymbol{M}^{(d)} = \mathrm{E}^{(d)}(F)$. Since $D^{(d)}_{\min}(\mathrm{E}^{(d)}(F),\mu)$ is bounded as

$$D_{\min}(F,\mu) \ge D^{(d)}_{\min}(\mathrm{E}^{(d)}(F),\mu) \\ \ge \inf_{G:L(G,F)\le L^{(d)}(F)} D_{\min}(G,\mu),$$

it sufficies to show that

$$\limsup_{d\to\infty} L^{(d)}(F) = \limsup_{d\to\infty} \sup_{G\in\mathcal{F}(\boldsymbol{M}^{(d)})} L(G,F) = 0$$

from the continuity of $D_{\min}(F,\mu)$ in F.

Let $\{G_d\in\mathcal{F}(\boldsymbol{M}^{(d)})\}_{d=1,2,\cdots}$ be a sequence such that

$$\limsup_{d\to\infty} \sup_{G\in\mathcal{F}(\boldsymbol{M}^{(d)})} L(G,F) = \limsup_{d\to\infty} L(G_d,F) =: \bar{L}.$$

Since $\mathcal{F}\supset\mathcal{F}(\boldsymbol{M}^{(d)})$ is compact with respect to the Lévy distance, there exist $\bar{G}\in\mathcal{F}$ and a convergent subsequence $\{G_{d_i}\}$ of $\{G_d\}$ such that

$$\lim_{i\to\infty} L(G_{d_i},\bar{G}) = 0, \tag{13}$$

$$\lim_{i\to\infty} L(G_{d_i},F) = \bar{L}, \tag{14}$$

where (13) means that $\{G_{d_i}\}$ converges weakly to $\bar{G}$. From the definition of weak convergence, for all natural numbers $m\in\mathbb{N}$ it holds that $\lim_{i\to\infty}\mathrm{E}_{G_{d_i}}[X^m] = \mathrm{E}_{\bar{G}}[X^m]$. On the other hand, $\mathrm{E}_{G_{d_i}}[X^m] = \mathrm{E}_F[X^m]$ for all $d_i\ge m$ from $G_{d_i}\in\mathcal{F}(\boldsymbol{M}^{(d_i)})$. Therefore we obtain for all $m\in\mathbb{N}$ that

$$\mathrm{E}_F[X^m] = \lim_{i\to\infty}\mathrm{E}_{G_{d_i}}[X^m] = \mathrm{E}_{\bar{G}}[X^m].$$

Note that a sequence of moments $\{\mathrm{E}_F[X^m]\}$ has one-to-one correspondence to a distribution F for the case of the bounded support. Therefore $\bar{G}=F$ and we obtain $\bar{L}=0$ from (13) and (14). □

Acknowledgements

This research is partially supported by the Aihara Project, the FIRST program from JSPS, initiated by CSTP. Junya Honda acknowledges support of JSPS Research Fellowships for Young Scientists.

References

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

Audibert, J.-Y., Munos, R., & Szepesvári, C. (2009). Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, *410*, 1876–1902.

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.

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

Csiszar, I., & Matus, F. (2009). On minimization of multivariate entropy functionals. *Proceedings of IEEE Information Theory Workshop on Networking and Information Theory (ITW 2009)* (pp. 96–100).

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.

Garivier, A., & Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. *Proceedings of COLT 2011*. Budapest, Hungary.

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

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

Karlin, S., & Studden, W. J. (1966). *Tchebycheff systems, with applications in analysis and statistics*. Interscience Publishers New York.

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.

Neumann, J. V. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, *100*, 295–320.

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.