Journal of Machine Learning Research 16 (2015) 3721-3756 Submitted 3/14; Revised 11/14; Published 12/15

Non-Asymptotic Analysis of a New Bandit Algorithm for Semi-Bounded Rewards

Junya Honda HONDA@STAT.T.U-TOKYO.AC.JP
Department of Complexity Science and Engineering
The University of Tokyo
Kashiwa-shi, Chiba, 277-8561, Japan

Akimichi Takemura TAKEMURA@STAT.T.U-TOKYO.AC.JP
Department of Mathematical Informatics
The University of Tokyo
Bunkyo-ku, Tokyo, 113-8561, Japan

Editor: Olivier Teytaud

Abstract

In this paper we consider a stochastic multiarmed bandit problem. It is known in this problem that Deterministic Minimum Empirical Divergence (DMED) policy achieves the asymptotic theoretical bound for the model where each reward distribution is supported in a known bounded interval, say $[0, 1]$. However, the regret bound of DMED is described in an asymptotic form and the performance in finite time has been unknown. We modify this policy and derive a finite-time regret bound for the new policy, Indexed Minimum Empirical Divergence (IMED), by refining large deviation probabilities to a simple non-asymptotic form. Further, the refined analysis reveals that the finite-time regret bound is valid even in the case that the reward is not bounded from below. Therefore, our finite-time result applies to the case that the minimum reward (that is, the maximum loss) is unknown or unbounded. We also present some simulation results which shows that IMED much improves DMED and performs competitively to other state-of-the-art policies.

Keywords: stochastic bandit, finite-time regret, large deviation principle

1. Introduction

In the multiarmed bandit problem a gambler pulls arms of a slot machine sequentially so that the total reward is maximized. There is a tradeoff between exploration and exploitation since he cannot know the most profitable arm unless pulling all arms infinitely many times.

There are two main formulations for this problem: stochastic and nonstochastic bandits. In the stochastic setting rewards of each arm follow an unknown distribution (Agrawal, 1995; Gittins, 1989; Vermorel and Mohri, 2005) whereas the rewards are determined by an adversary in the nonstochastic setting (Auer et al., 2002b). In this paper we consider the K-armed stochastic bandit, where rewards of arm $i \in \{1, 2, \cdots, K\}$ are i.i.d. sequence from unknown distribution $F_i \in \mathcal{F}$ with expectation μ_i for a model $\mathcal{F}$ known to the gambler. For the maximum expectation $\mu^* \equiv \max_i \mu_i$, we call an arm i optimal if $\mu_i = \mu^*$ and suboptimal otherwise. If the gambler knows each μ_i beforehand, it is best to choose optimal arms at every round. A *policy* is a strategy of the gambler for choosing arms based on the past results

©2015 Junya Honda and Akimichi Takemura.

of plays. The performance of a policy is usually measured by *pseudo-regret*, or simply *regret* in short. This is the gap of cumulative expectations between the optimal choice and the actual choice, which is expressed as

$$\mathcal{R}(n) \equiv \sum_{i:\mu_i<\mu^*} (\mu^* - \mu_i) T_i(n)\,,$$

where $T_i(n)$ is the number of plays of arm i through the first n rounds.

1.1 Theoretical Bound and its Achievability

Robbins (1952) first considered this setting and Lai and Robbins (1985) gave a framework for determining an optimal policy by establishing an asymptotic theoretical bound for the regret. Later this theoretical bound was extended to multiparameter or nonparametric models $\mathcal{F}$ by Burnetas and Katehakis (1996). It is proved in their paper that under a mild regularity condition any policy satisfies

$$\mathrm{E}[T_i(n)] \ge \frac{\log n}{D_{\mathrm{inf}}(F_i, \mu^*; \mathcal{F})} - \mathrm{o}(\log n) \tag{1}$$

for any suboptimal arm i, where $D_{\mathrm{inf}}(F, \mu; \mathcal{F})$ is defined in terms of Kullback-Leibler divergence $D(\cdot\|\cdot)$ by

$$D_{\mathrm{inf}}(F, \mu; \mathcal{F}) = \inf_{G \in \mathcal{F}: \mathrm{E}_G[X] > \mu} D(F\|G)\,.$$

The most popular model in the nonparametric setting is the family of distributions with supports contained in a known bounded interval, say $[0, 1]$. For this model, which we denote by $\mathcal{A}_0$, it is known that fine performance can be obtained by policies called Upper Confidence Bound (UCB) (Auer et al., 2002a; Audibert et al., 2009; Cappé et al., 2013). However, although some bounds for regrets of UCB policies have been obtained in a non-asymptotic form, they do not necessarily achieve the asymptotic theoretical bound.

Recently Honda and Takemura (2010) proposed Deterministic Minimum Empirical Divergence (DMED) policy, which chooses arms based on the value of $D_{\mathrm{inf}}(\hat{F}_i, \mu; \mathcal{A}_0)$, or simply written as $D_{\mathrm{inf}}(\hat{F}_i, \mu)$, for empirical distribution $\hat{F}_i$ of arm i. Whereas DMED achieves the asymptotic theoretical bound, the evaluation heavily depends on an asymptotic analysis and any finite-time regret bound has been unknown.

In this paper, we consider the family $\mathcal{A}$ of distributions on $(-\infty, 1]$ instead of the bounded support model $\mathcal{A}_0$. We first show that $D_{\mathrm{inf}}(F, \mu; \mathcal{A}_0) = D_{\mathrm{inf}}(F, \mu; \mathcal{A})$ for all $F \in \mathcal{A}_0$. Thus, any asymptotically optimal policy for the model $\mathcal{A}$ is also asymptotically optimal for $\mathcal{A}_0$, even though the gambler has more candidates for the true distribution of each arm in the model $\mathcal{A}$ than in $\mathcal{A}_0$.

We next propose a policy, the *IMED (Indexed Minimum Empirical Divergence)* algorithm. This is an indexed version of DMED in the sense that IMED simply chooses an arm which minimizes an index at each round whereas DMED requires to keep a list of arms to be pulled. We derive a finite-time regret bound of IMED for any distribution in $\mathcal{A}$ such that moment generating function $\mathrm{E}[\mathrm{e}^{\lambda X}]$ exists in some neighborhood of $\lambda = 0$. The derived bound coincides with the asymptotic theoretical bound and therefore IMED is asymptotically optimal for both $\mathcal{A}$ and $\mathcal{A}_0$. Since nonstochastic bandits inevitably require the boundedness of the support, we see that an advantage of assuming stochastic bandits is that the semi-bounded rewards can be dealt with in this nonparametric setting. Furthermore, we show that the reminder term of the logarithmic regret of IMED is $\mathrm{O}(1)$, whereas they are $\mathrm{O}((\log n)^a)$, $0 < a < 1$, in previously known asymptotically optimal regret bounds.

Note that DMED policy can be implemented without knowledge of the lower bound of the reward and achieves the asymptotic bound if the reward is only bounded from below by some unknown value. In this sense it is intuitively not surprising that DMED or its variant achieves the asymptotic the semi-bounded reward. However, the theoretical analysis for DMED in Honda and Takemura (2010) heavily depends on the boundedness of the support and its extension is not theoretically obvious.

There has also been some research for the nonparametric stochastic bandit with *unbounded* support distributions (Bubeck et al., 2012; Liu and Zhao, 2011). In particular, it is shown in Bubeck et al. (2012) that a logarithmic regret can be achieved if, for some $\epsilon > 0$, $\mathrm{E}_{F_i}[|X|^{1+\epsilon}]$ is bounded by a value *known to the gambler* beforehand. Although our assumption of the existence of the moment generating function $\mathrm{E}_{F_i}[\mathrm{e}^{\lambda X}]$ is more restrictive than the existence of the moment $\mathrm{E}_{F_i}[|X|^{1+\epsilon}]$, IMED does not require any knowledge on the value of $\mathrm{E}_{F_i}[\mathrm{e}^{\lambda X}]$ (or $\mathrm{E}_{F_i}[|X|^{1+\epsilon}]$). Therefore our assumption is not comparable to that in Bubeck et al. (2012).

1.2 Motivation for Semi-bounded Support Model

An example such that the lower bound of the reward is unknown or unbounded is the minimization of the sum of the time-delays in some task such as network routing (Vermorel and Mohri, 2005; Krishnamurthy et al., 2001), where the agent has many sources to obtain the same data. In this case, it may take a long time to complete the task and it is natural to consider that the reward (that is, negative of the time-delay) is not bounded from below. One may wonder that if some time-limit is fixed then the problem becomes a bounded bandit and a good finite-time regret has been already achieved by, for example, kl-UCB in Cappé et al. (2013) (although the regret bound of kl-UCB is not asymptotically optimal for distributions other than Bernoulli distributions). However, the time-limit (or the maximum time-delay) is usually set "conservatively", that is, set to a value much larger than time-delays in usual tries. In such a case, policies based only on empirical means tend to work poorly (see also Audibert et al., 2009). For example, kl-UCB achieves a regret near

$$\sum_{i:\mu_i<\mu^*} \frac{\mu^* - \mu_i}{D(\mathrm{B}(\mu_i)\|\mathrm{B}(\mu^*))} \log n$$

for reward distributions on $[0, 1]$, where $\mathrm{B}(\mu)$ denotes the Bernoulli distribution with mean μ. On the other hand, if the gambler conservatively estimates the lower bound of the reward by $a < 0$ instead of 0, he applies the policy after the rescaling from $[a, 1]$ to $[0, 1]$ and the regret becomes

$$\sum_{i:\mu_i<\mu^*} \frac{\mu^* - \mu_i}{D(\mathrm{B}((\mu_i - a)/(1-a))\|\mathrm{B}((\mu^* - a)/(1-a)))} \log n \,,$$

which goes to infinity as $a \to -\infty$. Audibert et al. (2009) overcame this problem by UCB-V policy, which uses empirical variances as well as empirical means. However, in turn, UCB-V does not necessarily perform well for usual Bernoulli distributions as reported in Cappé et al. (2013). Therefore the IMED policy has an advantage since it always achieves the optimal regret bound, which does not depend on whether the gambler knows the lower bound of the reward or not.

1.3 Outline

This paper is organized as follows. In Sect. 2 we give definitions used throughout this paper and propose the IMED policy as an indexed version of DMED. In Sect. 3, we give the main results of this paper on the finite-time regret bound of IMED for distributions on $(-\infty, 1]$. We discuss relation between IMED and other policies in Sect. 4 and give some simulation results of these policies in Sect. 5. The remaining sections and appendices are devoted to the proof of the main theorems. In Sect. 6, we analyze properties of the function D_{inf} for our model. In Sect. 7, we derive a large deviation probability of an empirical distribution $\hat{F}_t$ measured with D_{inf} in a non-asymptotic form. By using this probability, we derive the finite-time regret bound of IMED in Sect. 8. We conclude this paper with some discussion on the regularity condition assumed throughout the paper in Sect. 9. We evaluate constants used in the finite-time regret bound in Appendix A. We give a proof of a lemma analogous to the bounded-support model in Appendix B. Finally we prove the asymptotic but refined regret bound of IMED in Appendix C.

2. Preliminaries

In this section we introduce notation used throughout this paper and propose the IMED policy.

2.1 Notation

Let $\mathcal{A}_a$, $a \in (-\infty, 1)$, be the family of probability distributions on $[a, 1]$. We denote the family of distributions on $(-\infty, 1]$ by $\mathcal{A}_{-\infty}$ or simply $\mathcal{A}$. For $F \in \mathcal{A}$, the cumulative distribution at a point $x \in \mathbb{R}$ is denoted by $\bar{F}(x) \equiv F((-\infty, x])$, where $F(A)$, $A \subset \mathbb{R}$, denotes the measure of a set A. $\mathrm{E}_F[\cdot]$ denotes the expectation under $F \in \mathcal{A}$. When we write, for example, $\mathrm{E}_F[u(X)]$ for a function $u : \mathbb{R} \to \mathbb{R}$, X denotes a random variable with distribution F. The expectation of F is denoted by $\mathrm{E}(F) \equiv \mathrm{E}_F[X]$.

Let $J(n) \in \{1, 2, \cdots, K\}$ be the arm pulled at the n-th round. We define $T_i(n)$ as the number of times that arm i has been pulled through the first n rounds. Then, we have $T_i(n) = \sum_{l=1}^{n} \mathbb{1}[J(l) = i]$ where $\mathbb{1}[\cdot]$ denotes the indicator function. $\hat{F}_{i,t}$ and $\hat{\mu}_{i,t}$ denote the empirical distribution and the mean of arm i when arm i is pulled t times. $\hat{F}_i(n) \equiv \hat{F}_{i,T_i(n)}$ and $\hat{\mu}_i(n) \equiv \hat{\mu}_{i,T_i(n)}$ denote the empirical distribution and the mean of arm i at the n-th round. The largest empirical mean after the first n rounds is denoted by $\hat{\mu}^*(n) \equiv \max_i \hat{\mu}_i(n)$.

The function D_{inf} defined as

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

```
Algorithm 1 IMED Policy
Initialization: Pull each arm once.
Loop: Choose an arm $i$ minimizing
    $$I_i(n) \equiv T_i(n)D_{\mathrm{inf}}(\hat{F}_i(n), \hat{\mu}^*(n); \mathcal{A}) + \log T_i(n)\,,$$
  where the tie-breaking rule is arbitrary.
```

plays a central role in the DMED policy in Honda and Takemura (2010) and the IMED policy defined below. Let

$$\begin{aligned} L(\nu; F, \mu) &\equiv \mathrm{E}_F[\log(1-(X-\mu)\nu)]\,, \\ L_{\max}(F,\mu) &\equiv \max_{0\le\nu\le\frac{1}{1-\mu}} L(\nu;F,\mu)\,. \end{aligned} \tag{2}$$

Functions L and $L_{\max}$ correspond to the Lagrangian function and the dual problem of $D_{\mathrm{inf}}(F,\mu;\mathcal{A})$, respectively. The following proposition shows that D_{inf} is equal to $L_{\max}$ in the case of the bounded support model $\mathcal{A}_0$. In Sect. 3 we prove that the same result holds for the semi-bounded support model $\mathcal{A}$.

Proposition 1 (Honda and Takemura, 2010, Theorem 5) *For all $F \in \mathcal{A}_0$ and $\mu < 1$ it holds that $D_{\mathrm{inf}}(F,\mu;\mathcal{A}_0) = L_{\max}(F,\mu)$.*

2.2 IMED Policy

In the model $\mathcal{A}_0$, Honda and Takemura (2010) proposed an asymptotically optimal policy, DMED, which maintains the list of arms satisfying

$$T_i(n)D_{\mathrm{inf}}(\hat{F}_i(n), \hat{\mu}^*(n); \mathcal{A}_0) + \log T_i(n) \le \log n \tag{3}$$

where The DMED policy pulls an arm from the list in some order.

In this paper, we use the left-hand side of (3) as the index $I_i(n)$ for choosing an arm. Our proposed policy, Indexed Minimum Empirical Divergence (IMED) policy, is described as Algorithm 1. In the index $I_i(n)$, the first term $T_i(n)D_{\mathrm{inf}}(\hat{F}_i(n), \hat{\mu}^*(n)) \ge 0$ corresponds to the penalty for empirical distributions unlikely to occur from a distribution with expectation larger than $\hat{\mu}^*(n)$ and IMED usually chooses a currently optimal arm i since it satisfies $D_{\mathrm{inf}}(\hat{F}_i(n), \hat{\mu}^*(n)) = 0$. The second term $\log T_i(n)$ is the penalty for arms pulled too many times and corresponds to the exploration function.

Note that here we say that IMED is an index policy in a weaker sense than other index policies. Although both IMED and well known index policies such as Gittins index (Gittins, 1989) and UCB choose an arm which maximizes or minimizes its index at each round, the values of Gittins index and UCB score of each arm can be determined only from samples of the corresponding arm. On the other hand, the index of IMED also requires the maximum empirical mean over all arms, which depends on statistics of other arms. It may seem somewhat unnatural to use such an index for choosing an arm but IMED has an advantage in the computational complexity for this property of the index as discussed in Sect. 4.1.

3. Main Results

We now state the main results of this paper in Theorems 2, 3 and 5. In Theorem 2, we show that the theoretical bound does not depend on knowledge of the lower bound of the support. In Theorem 3, we give a non-asymptotic regret bound of IMED, which shows that the theoretical bound can be achieved by IMED. We give an asymptotic but refined regret bound of IMED in Theorem 5.

Theorem 2 *Let $a \in [-\infty, 1)$ and $F \in \mathcal{A}_a$ be arbitrary.* (i) $D_{\mathrm{inf}}(F,\mu;\mathcal{A}_a) = D_{\mathrm{inf}}(F,\mu;\mathcal{A})$. (ii) *If $\mu < 1$ then*

$$D_{\mathrm{inf}}(F,\mu;\mathcal{A}) = L_{\max}(F,\mu)\,.$$

We prove this theorem in Sect. 6. The part (i) of this theorem means that the theoretical bound does not depend on whether the gambler knows lower bound of the support of distributions or he has to consider the case that the support is not bounded from below. Furthermore, from (ii), we can compute $D_{\mathrm{inf}}(F,\mu;\mathcal{A})$ by using the expression $L_{\max}(F,\mu)$ as in the case of $\mathcal{A}_0$. In view of this theorem we sometimes write $D_{\mathrm{inf}}(F,\mu)$ instead of more precise $D_{\mathrm{inf}}(F,\mu;\mathcal{A}_a)$ or $D_{\mathrm{inf}}(F,\mu;\mathcal{A})$.

Define

$$\nu_i^* \equiv \mathop{\mathrm{argmax}}_{0\le\nu\le\frac{1}{1-\mu^*}} \mathrm{E}_{F_i}[\log(1-(X-\mu^*)\nu)]\,,$$
$$\lambda_{i,\mu} \equiv \sup\left\{\lambda \in \mathbb{R}\cup\{\infty\} : \mathrm{E}_{F_i}\left[\left(\frac{1-X}{1-\mu}\right)^{\lambda}\right] \le 1\right\}\,, \tag{4}$$

where we show that ν_i^* exists uniquely when $\mathrm{E}(F_i) < \mu^*$ in Sect. 6 and show $\lambda_{i,\mu} > 1$ for $\mu < \mu_i$ in Sect. 7. We further define Fenchel-Legendre transforms of cumulant generating functions of random variables X and $\log(1-(X-\mu^*)\nu_i^*)$ as

$$\Lambda_i^*(x) \equiv \sup_{\lambda}\{\lambda x - \log \mathrm{E}_{F_i}[\mathrm{e}^{\lambda X}]\}\,, \tag{5}$$
$$\tilde{\Lambda}_i^*(x) \equiv \sup_{\lambda}\left\{\lambda x - \log \mathrm{E}_{F_i}[(1-(X-\mu^*)\nu_i^*)^{\lambda}]\right\}\,. \tag{6}$$

Then, for[1] $\Delta_i \equiv \mu^* - \mu_i$ and $\mathcal{I}_{\mathrm{opt}} \equiv \{j : \mu_j = \mu^*\} \subset \{1,\cdots,K\}$, the regret of IMED is bounded as follows.

Theorem 3 *Assume that $\mu^* < 1$ and $\mathrm{E}_{F_j}[\mathrm{e}^{\lambda X}] < \infty$ in some neighborhood of $\lambda = 0$ for some $j \in \mathcal{I}_{\mathrm{opt}}$. Then, for any fixed $0 < \delta < \min_{i:\mu_i<\mu^*}\Delta_i/2$, the expected number of pulls of a suboptimal arm $i \notin \mathcal{I}_{\mathrm{opt}}$ is bounded as*

$$\mathrm{E}[T_i(n)] \le \frac{\log n}{D_{\mathrm{inf}}(F_i,\mu^*) - \frac{2\delta}{1-\mu^*}} + \frac{1}{1-\mathrm{e}^{-\tilde{\Lambda}_i^*(D_{\mathrm{inf}}(F_i,\mu^*)-\frac{\delta}{1-\mu^*}))}}$$
$$+ \min_{j\in\mathcal{I}_{\mathrm{opt}}}\left\{\frac{6\mathrm{e}}{(1-1/\lambda_{j,\mu^*-\delta})(1-\mathrm{e}^{-(1-1/\lambda_{j,\mu^*-\delta})\Lambda_j^*(\mu^*-\delta)})^3}\right\}.$$

[1] We often use the subscript i for a suboptimal arm and use j for an optimal arm.

Consequently, the expected regret is bounded as

$$\mathrm{E}[\mathcal{R}(n)] \le \sum_{i:\Delta_i>0} \Delta_i \left(\frac{\log n}{D_{\mathrm{inf}}(F_i,\mu^*) - \frac{2\delta}{1-\mu^*}} + \frac{1}{1-\mathrm{e}^{-\tilde{\Lambda}_i^*(D_{\mathrm{inf}}(F_i,\mu^*)-\frac{\delta}{1-\mu^*}))}} \right) \\ + \left(\sum_{i=1}^{K} \Delta_i \right) \min_{j\in\mathcal{I}_{\mathrm{opt}}} \left\{ \frac{6\mathrm{e}}{(1-1/\lambda_{j,\mu^*-\delta})(1-\mathrm{e}^{-(1-1/\lambda_{j,\mu^*-\delta})\Lambda_j^*(\mu^*-\delta)})^3} \right\}.$$

We prove Theorem 3 in Sect. 8 based on non-asymptotic large deviation probabilities for $D_{\mathrm{inf}}(\hat{F}_i(n), \hat{\mu}^*(n))$ given in Sect. 7. In Appendix A, we discuss simple representations of $(\lambda_{j,\mu},\ \Lambda_i^*(x),\ \tilde{\Lambda}_i^*(x))$ and show that $\lambda_{j,\mu^*-\delta} = 1+\mathrm{O}(\delta)$, $\Lambda_i^*(\mu^*-\delta) = \mathrm{O}(\delta^2)$ and $\tilde{\Lambda}_i^*(D_{\mathrm{inf}}(F_i,\mu^*) - \delta/(1-\mu^*)) = \mathrm{O}(\delta^2)$. The following corollary is straightforward from this observation.

Corollary 4 *Under the assumption of Theorem 3,*

$$\mathrm{E}[\mathcal{R}(n)] = \sum_{i:\mu_i<\mu^*} \frac{\Delta_i \log n}{D_{\mathrm{inf}}(F_i,\mu^*)} + \mathrm{O}((\log n)^{10/11}). \tag{7}$$

Proof From $1-\mathrm{e}^{-\epsilon} = \mathrm{O}(\epsilon)$ and the above observation on $(\lambda_{j,\mu},\ \Lambda_i^*(x),\ \tilde{\Lambda}_i^*(x))$,

$$\mathrm{E}[\mathcal{R}(n)] \le \sum_{i:\mu_i<\mu^*} \frac{\Delta_i \log n}{D_{\mathrm{inf}}(F_i,\mu^*)} + \mathrm{O}(\delta \log n) + \mathrm{O}(\delta^{-2}) + \mathrm{O}(\delta^{-10}).$$

We obtain (7) by letting $\delta = \mathrm{O}((\log n)^{-1/11})$. ■

From this corollary we see that IMED is asymptotically optimal in view of (1). However, the reminder term $\mathrm{O}((\log n)^{10/11})$ is quite larger than those of known asymptotically optimal policies for other models although our model, the semi-bounded support model, is quite complicated. For example, it is shown in Cappé et al. (2013) that the KL-UCB policy achieves the asymptotic bound with reminder term $\mathrm{O}(\sqrt{\log n})$ for a subclass of one-dimensional exponential families and $\mathrm{O}((\log n)^{4/5} \log\log n)$ for the finite support model. The following theorem shows that the reminder term can be much improved in our model.

Theorem 5 *(i) Assume that $\mu^* < 1$ and $\mathrm{E}_{F_i}[\mathrm{e}^{\lambda X}] < \infty$ in some neighborhood of $\lambda = 0$ for all $i \in \{1,2,\cdots,K\}$. Then*

$$\mathrm{E}[\mathcal{R}(n)] = \sum_{i:\mu_i<\mu^*} \frac{\Delta_i \log n}{D_{\mathrm{inf}}(F_i,\mu^*)} + \mathrm{O}(1). \tag{8}$$

(ii) Furthermore, if the distribution of each arm has a bounded support then the reminder term $\mathrm{O}(1)$ in (8) can be replaced with $-\mathrm{O}(\log\log n)$, that is, there exists $C > 0$ such that for all sufficiently large n

$$\mathrm{E}\left[\mathcal{R}(n)\right] \le \sum_{i:\mu_i<\mu^*} \frac{\Delta_i \log n}{D_{\mathrm{inf}}(F_i,\mu^*)} - C \log\log n. \tag{9}$$

The proof of this theorem is much more complicated than that of Theorem 3 and given in Appendix C.

Note that a policy asymptotically optimal for the semi-bounded support model is also asymptotically optimal for the model of finite-support distributions (Honda and Takemura, 2011, Theorem 3). Therefore the regret bound (9) of IMED is asymptotically better than that of KL-UCB in Cappé et al. (2013) for finite-support distributions, of which the reminder term is $\mathrm{O}((\log n)^{4/5} \log\log n)$.

To the best of the authors' knowledge, this is the first result to show that the asymptotic bound (1) is achievable with a reminder term $\mathrm{O}(1)$ instead of $\mathrm{o}(\log n)$. The key to this refined bound is to apply a technique for a stopping-time of a stochastic process, which we evaluate in Lemma 18. The authors think that the regret bounds of other policies can also be improved by using this novel technique.

4. Relation with Other Policies

In the previous sections we showed that IMED achieves the asymptotic bound for the semi-bounded support model. In this section we compare IMED with other policies which achieve a logarithmic regret for some models.

4.1 KL-UCB Policies

Burnetas and Katehakis (1996) proposed a UCB policy for a general class $\mathcal{F}$ which chooses an arm maximizing the index

$$\sup\left\{\mu : T_i(n) D_{\mathrm{inf}}(\hat{F}_i(n), \mu; \mathcal{F}) \le f(n)\right\} \tag{10}$$

for some exploration function $f(n)$. They gave a sufficient condition for the asymptotic optimality of this policy for general model $\mathcal{F}$ and proved that the condition is satisfied for the finite support model and the normal distribution model with known variances. Furthermore Cappé et al. (2013) proved its asymptotic optimality with a finite-time regret bound for the finite support model and a subclass of exponential families. They also proved that this policy where $D_{\mathrm{inf}}(\mu; \mathcal{F})$ is replaced with the Bernoulli divergence

$$D_{\mathrm{inf}}(\hat{F}_i(n), \mu; \mathcal{F}_{\mathrm{Ber}}) = \hat{\mu}_i(n) \log \frac{\hat{\mu}_i(n)}{\mu} + (1 - \hat{\mu}_i(n)) \log \frac{1 - \hat{\mu}_i(n)}{1 - \mu} \tag{11}$$

achieves a logarithmic regret for general distributions with supports in $[0, 1]$. We refer to this policy for general model $\mathcal{F}$ as KL-UCB and the policy with (11) for bounded-support distributions as kl-UCB after Cappé et al. (2013). We can make the KL-UCB policy computationally feasible by using Prop. 1 and Theorem 2 for the bounded support model and the semi-bounded support model, respectively, but the asymptotic optimality for these models has been currently unknown although the authors believe that it can be proved as in IMED by using Theorem 2 and large deviation probabilities evaluated in the next section.

Other than the theoretical guarantee of the asymptotic optimality, the IMED has an advantage in the computational complexity. In the semi-bounded support model (or the bounded support model), the computation of D_{inf} itself involves an optimization and a simple representation of (10) has not been known whereas D_{inf} can be represented as a *univariate* convex optimization as shown in Theorem 2.

Furthermore, since $D_{\mathrm{inf}}(F,\mu)=0$ for $\mathrm{E}(F)=\mu$, IMED does not require the computation of $D_{\mathrm{inf}}(\hat{F}_i(n),\hat{\mu}^*(n))$ for currently optimal arms and the computation of these values for currently suboptimal arms are sufficient. Since any suboptimal arm is pulled at most $\mathrm{O}(\log n)$ times in average, the size of the support of $\hat{F}_i(n)$ is $\mathrm{O}(\log n)$ and the average complexity of IMED at each round becomes $\mathrm{O}(\log n)$. On the other hand, KL-UCB also require the computation of (10) for currently optimal arms and the complexity becomes $\mathrm{O}(n)$ as discussed in Cappé et al. (2013, Sect. 6.2). This advantage of IMED justifies to some extent the use of a somewhat unnatural index which depends on statistics of other arms.

4.2 Bayesian Policies

There have also been some Bayesian policies which are known to achieve the asymptotic bound for some model.

The Bayes-UCB policy (Kaufmann et al., 2012a) is a variant of UCB family obtained by the replacement of $T_i(n)D_{\mathrm{inf}}(\hat{F}_i(n),\mu)$ in (10) with a quantity associated with a posterior probability on the true expectation of the arm. The asymptotic optimality of this policy is proved for the Bernoulli model.

Another Bayesian policy is Thompson sampling (TS) originally proposed in Thompson (1933), which is a randomized algorithm which chooses an arm according to the posterior probability that the arm is optimal. TS is proved to be asymptotically optimal for general one-dimensional exponential families including the Bernoulli model (Kaufmann et al., 2012b; Agrawal and Goyal, 2013; Korda et al., 2013). It is also reported that TS is easily applicable to many models with a state-of-the-art performance (Chapelle and Li, 2012; Russo and Roy, 2013). On the other hand, TS requires random sampling from the posterior which is difficult for models other than exponential families, particularly in the nonparametric models. Although it may become tractable for the semi-bounded support model in nonparametric Bayesian framework, it is not very simple compared to the computation of D_{inf} and it remains unknown whether TS works practically for our model.

4.3 Achievability of Logarithmic Regret for Semi-bounded Support Model

Another question is whether or not there exists a simpler policy than IMED which achieves a (possibly non-optimal) logarithmic regret for the semi-bounded support model. For the bounded support model a logarithmic regret can be achieved by kl-UCB policy as described above. The key property of KL-UCB is

$$D(\mathrm{B}(\mathrm{E}(F))\|\mathrm{B}(\mu))\le D_{\mathrm{inf}}(F,\mu)$$

for $F\in\mathcal{A}_0$, which means that the Bernoulli divergence can be used as a lower bound of $D_{\mathrm{inf}}(F,\mu)$ when the expectation (that is, the first-order moment) of F is specified. However, in the derivation of this inequality a convex function on the support $[0,1]$ is bounded from *above* and the lower and upper bounds of the support are explicitly required (see Sect. 6.1 of Cappé et al. (2013) for detail), which makes difficult to bound $D_{\mathrm{inf}}(F,\mu)$ for general $F\in\mathcal{A}$.

A natural idea to bound $D_{\mathrm{inf}}(F,\mu)$ is to use higher-order moments of F. DMED-M (Honda and Takemura, 2012) is a policy based on this idea and obtained by replacing $D_{\mathrm{inf}}(F,\mu)=D_{\mathrm{inf}}(F,\mu;\mathcal{A}_a)$ for $F\in\mathcal{A}_a$, $a>-\infty$, with

$$D_{\mathrm{inf}}^{(d)}(\boldsymbol{M}^{(d)},\mu;\mathcal{A}_a)\equiv\inf_{G\in\mathcal{A}_a:\mathrm{E}_G[X^m]=\mathrm{E}_F[X^m],\,m=1,2,\cdots,d}D_{\mathrm{inf}}(G,\mu;\mathcal{A}_a)\,,$$

where $\boldsymbol{M}^{(d)}=(\mathrm{E}_F[X],\mathrm{E}_F[X^2],\cdots,\mathrm{E}_F[X^d])$. We can compute $D_{\mathrm{inf}}^{(d)}$ by solving algebraic equations and it is expressed in an explicit form for $d\le 4$ from the theory of Tchebycheff system (Karlin and Studden, 1966). The important point is that $D_{\mathrm{inf}}^{(d)}$ for even d does not depend on the lower bound a of the support (Honda and Takemura, 2012, Theorem 3). This means that DMED-M for even d achieves a logarithmic regret bound without knowledge on the lower bound a of the support whereas a policy using Bernoulli divergence $D_{\mathrm{inf}}^{(1)}(\mathrm{E}(F),\mu;\mathcal{A}_a)$ becomes meaningless for $a\to-\infty$ as discussed in Introduction. Therefore we can expect that DMED-M, or other policies based on $D_{\mathrm{inf}}^{(d)}$, also achieves a logarithmic regret for the semi-bounded support model since the key technique, Tchebycheff system, is extended to semi-bounded support distributions (Karlin and Studden, 1966, Chap. V).

5. Experiment

In this section we give some simulation results for IMED, DMED, Thompson sampling (TS) and KL-UCB family. For the KL-UCB family, we use $f(n)=\log n$ as an exploration function for (10) since the asymptotic optimality is shown in Burnetas and Katehakis (1996) for some models and it is empirically recommended in Cappé et al. (2013) although the latter paper uses $f(n)=\log n+c\log\log n$ for some $c>0$ in the proof of the optimality. The kl-UCB+ and KL-UCB+ (Garivier and Cappé, 2011) are empirical improvements of kl-UCB and KL-UCB, respectively, where $f(n)=\log n$ is replaced with $f(n)=\log(n/T_i(n))$. The optimality analysis of these policies has not been given but a similar version is discussed in Kaufmann (2014, Proposition 2.4) for some models.

Each plot is an average over 10,000 trials. In the four figures given below, IMED and KL-UCB+ performed almost the best. Whereas the complexity of IMED is smaller than KL-UCB family as discussed in Sect. 4.1, the regret of IMED was slightly worse than that of KL-UCB+.

First, Fig. 1 shows simulation results of IMED, DMED, TS, kl-UCB and kl-UCB+ for ten-armed bandit with Bernoulli rewards with $\mu_1=0.1$, $\mu_2=\mu_3=\mu_4=0.05$, $\mu_5=\mu_6=\mu_7=0.02,\mu_8=\mu_9=\mu_{10}=0.01$, which is the same setting as those in[2] Kaufmann et al. (2012b) and Cappé et al. (2013).

Next, we consider the case that the time-delay X_i' for some task by the i-th agent follows an exponential distribution with density $\mathrm{e}^{-x/\mu_i'}/\mu_i'$, $x\ge 0$, and the player tries to minimize the cumulative delay. Since we modeled the reward as a random variable in $(-\infty,1]$, we set

[2] The simulation result for DMED in this paper is different from those in these references where DMED is reported to perform much worse. This is because a policy where (3) is replaced with the condition

$$T_i(n)D_{\mathrm{inf}}(\hat{F}_i(n),\hat{\mu}^*(n);\mathcal{A}_0)\le\log n$$

is used as "DMED" in these references although the optimality proof of DMED is given for (3). This replacement can be interpreted as that of KL-UCB+ with KL-UCB (see also Garivier and Cappé (2011)).

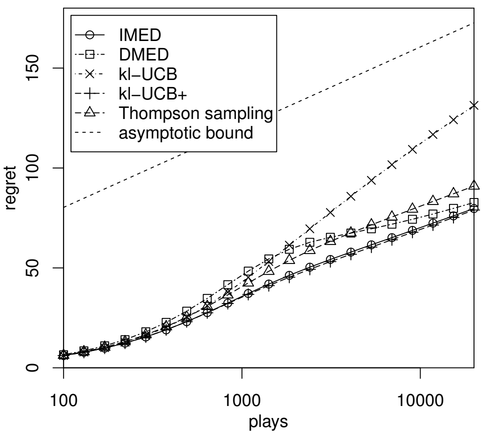

Figure 1: Average regret for 10-armed Bernoulli bandit.

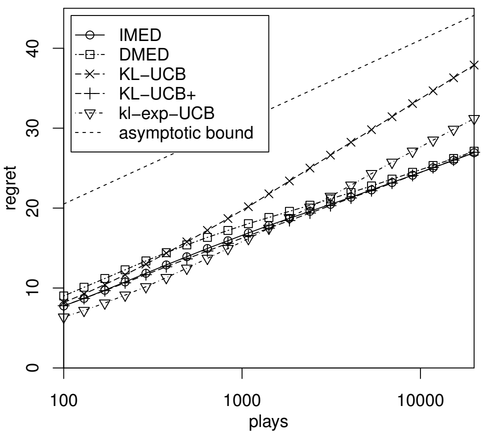

Figure 2: Average regret for 5-armed bandit where the negative reward follows an exponential distribution.

the reward as $X_i = 1 - X'_i$, that is, X_i has density $\mathrm{e}^{-(1-x)/\mu'_i}/\mu'_i = \mathrm{e}^{-(1-x)/(1-\mu_i)}/(1-\mu_i)$, $x \le 1$, with expectation $\mu_i = 1 - \mu'_i$. Fig. 2 shows simulation results for 5-armed bandit with $\mu'_i = 1/5,\ 1/4,\ 1/3,\ 1/2,\ 1$, that is, $\mu_i = 4/5,\ 3/4,\ 2/3,\ 1/2,\ 0$. We used IMED, DMED, KL-UCB, KL-UCB+ for $\mathcal{A}$ and KL-UCB for the (shifted) exponential distributions, which we refer as kl-exp-UCB, where the KL divergence is written as

$$D(\hat{\mu}_i \| \mu) = \frac{1-\hat{\mu}_i}{1-\mu} - 1 - \log\frac{1-\hat{\mu}_i}{1-\mu}\,.$$

The kl-exp-UCB policy explicitly assumes the knowledge that $1 - X_i$ follows an exponential distribution (and under the same assumption TS can also be implemented) whereas the other policies only uses the knowledge on the upper bound of the reward.

Since kl-exp-UCB is asymptotically optimal for exponential distributions, it is theoretically assured that it asymptotically outperforms other policies for this setting. Nevertheless, it seems from the comparison of kl-exp-UCB and KL-UCB that the gap between theoretical bounds for semi-bounded support model and for exponential distributions is not very large, which supports the effectiveness of the nonparametric model.

Finally, Figs. 3 and 4 show results of IMED, DMED, KL-UCB and KL-UCB+ for truncated normal distributions on $[0, 1]$ and $(-\infty, 1]$, respectively, as examples of multiparameter models. The cumulative distribution of each reward is given by

$$\bar{F}_i(x) = \begin{cases} 0, & x < a, \\ \frac{\Phi((x-\mu'_i)/\sigma_i) - \Phi((a-\mu'_i)/\sigma_i)}{\Phi((1-\mu'_i)/\sigma_i) - \Phi((a-\mu'_i)/\sigma_i)}, & a \le x < 1, \\ 1, & 1 \le x, \end{cases}$$

where $a = 0$ or $-\infty$, and Φ is the cumulative distribution function of the standard normal distribution. We also give results of kl-UCB and TS for the Bernoulli bandit for the setting

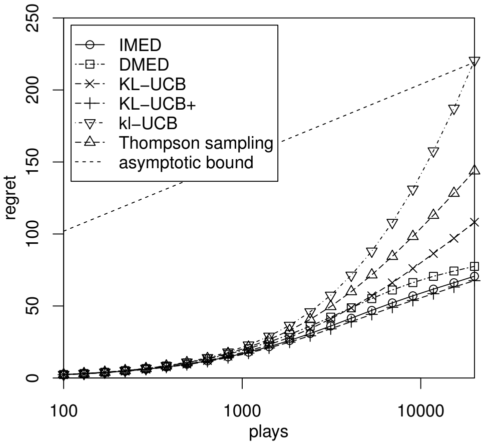

Figure 3: Average regret for 5-armed bandit with truncated normal distributions on $[0, 1]$.

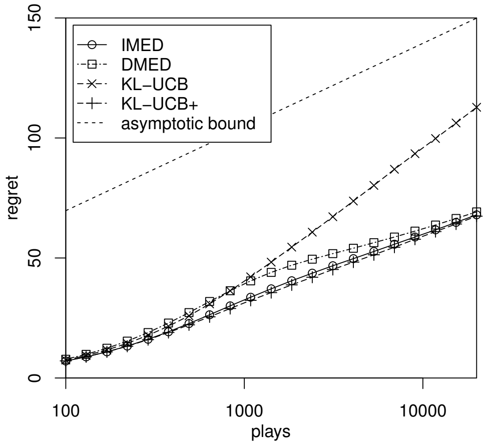

Figure 4: Average regret for 5-armed bandit with truncated normal distributions on $(-\infty, 1]$.

of Fig. 3 where the reward is bounded. For each experiment we set expectations and variances before truncation as $\mu_i' = 0.6,\ 0.5,\ 0.5,\ 0.4,\ 0.4$ and $\sigma_i = 0.4,\ 0.2,\ 0.4,\ 0.2,\ 0.4$. The expectation of each arm after truncation is given by $\mu_i = 0.519,\ 0.5,\ 0.5,\ 0.465, 0.481$ for support $[0, 1]$ and $\mu_i = 0.319,\ 0.390,\ 0.265,\ 0.320,\ 0.206$ for support $(-\infty, 1]$. We see from Fig. 3 that the policies for the nonparametric model work much better than that for the Bernoulli model.

6. Properties of D_{inf} in the Semi-bounded Support Model

In this section we extend some results on $D_{\mathrm{inf}}(F, \mu; \mathcal{A}_0)$ in Honda and Takemura (2010) to model $\mathcal{A} = \mathcal{A}_{-\infty}$ and prove Theorem 2.

The minimization function $D_{\mathrm{inf}}(F, \mu; \mathcal{A})$ is expressed as

$$\begin{aligned} \text{minimize:}\ & \int \left(\log \frac{\mathrm{d}F}{\mathrm{d}G} \right) \mathrm{d}F \\ \text{subject to:}\ & G \in \mathcal{A} \text{ is a positive finite measure on } (-\infty, 1], \\ & \int \mathrm{d}G = 1,\ \int x \mathrm{d}G > \mu\,, \end{aligned}$$

which has an infinite-dimensional variable and finite constraints. An optimization problem of this form is called a *partially-finite convex optimization* and many researches have been conducted (Borwein and Lewis, 1993; Ito et al., 2000). We can prove the relation $D_{\mathrm{inf}}(F, \mu; \mathcal{A}_0) = L_{\max}(F, \mu)$ in Prop. 1 in a generic way for this problem although it is proved in a problem-specific way in Honda and Takemura (2010, Theorem 5). Nevertheless, we were not able to find a result straightforwardly applicable to our target $D_{\mathrm{inf}}(F, \mu; \mathcal{A})$ for the reason below and we analyze this problem in a problem-specific way.

The difficulty in the model $\mathcal{A}$ lies in the fact that $\mathcal{A}$ is not compact and the operator $x : \mathcal{A} \to \mathbb{R} : G \mapsto \int x \mathrm{d}G$ in the constraint is not continuous under the Lévy metric since $f(x) = x$ is not a bounded function on $(-\infty, a]$. For this reason it is necessary to evaluate the effect of tail weights of measures on expectations precisely.

First we consider the function $L(\nu; F, \mu) = \mathrm{E}_F[\log(1 - (X - \mu))\nu]$. The integrand $l(x, \nu) \equiv \log(1 - (x - \mu)\nu)$ is differentiable in $\nu \in (0, (1 - \mu)^{-1})$ for all $x \in (-\infty, 1]$ with

$$\begin{aligned} \frac{\partial l(x,\nu)}{\partial \nu} &= -\frac{x-\mu}{1-(x-\mu)\nu} = \frac{1}{\nu}\left(1 - \frac{1}{1-(x-\mu)\nu}\right), \\ \frac{\partial^2 l(x,\nu)}{\partial \nu^2} &= -\frac{(x-\mu)^2}{(1-(x-\mu)\nu)^2}. \end{aligned}$$

Since they are bounded in $x \in (-\infty, 1]$, the integral $L(\nu; F, \mu)$ is differentiable in ν with

$$L'(\nu; F, \mu) \equiv \frac{\partial L(\nu; F, \mu)}{\partial \nu} = \frac{1}{\nu}\left(1 - \mathrm{E}_F\left[\frac{1}{1-(X-\mu)\nu}\right]\right), \tag{12}$$

$$L''(\nu; F, \mu) \equiv \frac{\partial^2 L(\nu; F, \mu)}{\partial \nu^2} = -\mathrm{E}_F\left[\frac{(X-\mu)^2}{(1-(X-\mu)\nu)^2}\right]. \tag{13}$$

From these derivatives the optimal solution $\nu^* = \nu^*(F, \mu) = \mathrm{argmax}_{0 \le \nu \le (1-\mu)^{-1}} L(\nu; F, \mu)$ of (2) exists uniquely except for the case $X = \mu$ (a.s.) and satisfies the properties in the following lemmas.

Lemma 6 *Assume that* $\mathrm{E}(F) < \mu < 1$ *holds. If* $\mathrm{E}_F[(1-\mu)/(1-X)] < 1$ *then* $\nu^* = (1-\mu)^{-1}$ *and therefore* $\mathrm{E}_F[1/(1-(X-\mu)\nu^*)] < 1$. *Otherwise,* $\nu^* \in (0, (1-\mu)^{-1})$ *and* $\mathrm{E}_F[1/(1-(X-\mu)\nu^*)] = 1$.

Lemma 7 $L_{\max}(F, \mu)$ *is differentiable in* $\mu < \mathrm{E}(F)$ *with*

$$\frac{\mathrm{d}L_{\max}(F,\mu)}{\mathrm{d}\mu} = \nu^*(F,\mu) \le \frac{1}{1-\mu}.$$

Lemma 6 is straightforward from the derivatives (12) and (13). The proof of Lemma 7 is completely analogous to the proof of Honda and Takemura (2011, Theorems 3 (iii)) where the same results is derived for distributions on a finite support. We give the proof for completeness in Appendix B.

Define $F_{(a)} \in \mathcal{A}_a$ as the distribution obtained by transferring the probability of $(-\infty, a)$ under F to $x = a$, that is, the cumulative distribution function of $F_{(a)}$ is defined as

$$\bar{F}_{(a)}(x) \equiv \begin{cases} 0 & x < a, \\ \bar{F}(x) & x \ge a. \end{cases}$$

Recall that $L_{\max}(F, \mu) = \max_{0 \le \nu \le (1-\mu)^{-1}} L(\nu; F, \mu) = \max_{0 \le \nu \le (1-\mu)^{-1}} \mathrm{E}_F[\log(1 - (X - \mu))\nu]$. Now we give the key to extension for the semi-bounded support in the following lemma, which shows that the effect of the tail weight is bounded uniformly if the expectation is bounded from below.

Lemma 8 *Fix arbitrary $\mu, \tilde{\mu} < 1$ and $\epsilon > 0$. Then there exists $a(\epsilon, \mu, \tilde{\mu})$ such that $|L_{\max}(F_{(a)}, \mu) - L_{\max}(F, \mu)| \le \epsilon$ for all $a \le a(\epsilon, \mu, \tilde{\mu})$ and all $F \in \mathcal{A}$ such that $\mathrm{E}(F) \ge \tilde{\mu}$.*

Proof Take sufficiently small $a < \min\{0, \mu\}$ and define $A = (-\infty, a), B = [a, 1]$. Note that $F(A) + F(B) = 1$. First we have

$$\begin{aligned} F(A) &\le \frac{1 - \tilde{\mu}}{1 - a} &(14)\\ \int_A x \mathrm{d}F(x) &\ge \tilde{\mu} - 1 + F(A) &(15) \end{aligned}$$

from

$$\begin{aligned} &\mathrm{E}(F) \le aF(A) + 1 \cdot F(B) = 1 - (1 - a)F(A) \\ &\mathrm{E}(F) \le \int_A x \mathrm{d}F(x) + 1 \cdot F(B) \,, \end{aligned}$$

respectively. Next, $L_{\max}(F, \mu)$ can be written as

$$\begin{aligned} L_{\max}(F, \mu) &= \max_{0 \le \nu \le \frac{1}{1-\mu}} \mathrm{E}_F[\log(1 - (X - \mu)\nu)] \\ &= \max_{0 \le \nu \le \frac{1}{1-\mu}} \left\{ \int_A \log \frac{1 - (x - \mu)\nu}{1 - (a - \mu)\nu} \mathrm{d}F(x) + \int_B \log(1 - (x - \mu)\nu) \mathrm{d}F_{(a)}(x) \right\}. \quad (16) \end{aligned}$$

Since $(1 - (x - \mu)\nu)/(1 - (a - \mu)\nu)$ is increasing in ν for $x \le a$, substituting 0 and $(1 - \mu)^{-1}$ into ν, we can bound the first term as

$$\begin{aligned} 0 &\le \int_A \log \frac{1 - (x - \mu)\nu}{1 - (a - \mu)\nu} \mathrm{d}F(x) \\ &\le \int_A \log \frac{1 - x}{1 - a} \mathrm{d}F(x) \\ &\le F(A) \int_A \log(1 - x) \frac{\mathrm{d}F(x)}{F(A)} & \text{(by } a \le 0\text{)} \\ &\le F(A) \log \left(\int_A (1 - x) \frac{\mathrm{d}F(x)}{F(A)} \right) & \text{(Jensen's inequality)} \\ &\le F(A) \log \frac{1 - \tilde{\mu}}{F(A)} \,. & \text{(by (15))} \end{aligned}$$

From $\lim_{x \to 0} x \log x = 0$ and (14), the first term of (16) converges to 0 as $a \to -\infty$. The second term of (16) equals $L_{\max}(F_{(a)}, \mu)$ and the proof is completed. ■

Now we show Theorem 2 based on the preceding lemmas.

Proof of Theorem 2 (i) Recall that $G_{(a)}$ is the distribution such that the weight of G on $(-\infty, a)$ is transported to the point a. Thus, if $F \in \mathcal{A}_a$ is absolutely continuous with respect to G then $\mathrm{d}F/\mathrm{d}G \ge \mathrm{d}F/\mathrm{d}G_{(a)}$ holds almost everywhere on the support of F and we have $D(F \| G) \ge D(F \| G_{(a)})$. On the other hand if F is not absolutely continuous then

$D(F\|G)=\infty$ and therefore $D(F\|G)\ge D(F\|G_{(a)})$ still holds for this case. Combining them we have

$$\begin{aligned}\inf_{G\in\mathcal{A}:\mathrm{E}(G)>\mu}D(F\|G)&\ge\inf_{G\in\mathcal{A}:\mathrm{E}(G)>\mu}D(F\|G_{(a)})\\&\ge\inf_{G\in\mathcal{A}:\mathrm{E}(G_{(a)})>\mu}D(F\|G_{(a)})\quad\big(\text{by }\mathrm{E}(G)\le\mathrm{E}(G_{(a)})\big)\\&=\inf_{G\in\mathcal{A}_a:\mathrm{E}(G)>\mu}D(F\|G)\,.\end{aligned}$$

On the other hand it holds from $\mathcal{A}_a\subset\mathcal{A}$ that

$$\inf_{G\in\mathcal{A}:\mathrm{E}(G)>\mu}D(F\|G)\le\inf_{G\in\mathcal{A}_a:\mathrm{E}(G)>\mu}D(F\|G)$$

and we obtain $\inf_{G\in\mathcal{A}:\mathrm{E}(G)>\mu}D(F\|G)=\inf_{G\in\mathcal{A}_a:\mathrm{E}(G)>\mu}D(F\|G)$.

(ii) We show $D_{\mathrm{inf}}(F,\mu;\mathcal{A})\le L_{\max}(F,\mu)$ and $D_{\mathrm{inf}}(F,\mu;\mathcal{A})\ge L_{\max}(F,\mu)$ separately. To prove the former inequality, let us consider a measure for any (measurable) set $S\subset\mathbb{R}$

$$G^*(S)\equiv\begin{cases}\int_S\frac{1-\mu}{1-x}\mathrm{d}F+(1-\mathrm{E}_F[\frac{1-\mu}{1-X}])\mathbb{1}[1\in S]\,, & \mathrm{E}_F[\frac{1-\mu}{1-X}]\le 1\,,\\ \int_S\frac{1}{1-(x-\mu)\nu^*}\mathrm{d}F\,, & \mathrm{E}_F[\frac{1-\mu}{1-X}]>1\,.\end{cases}$$

We can see from Lemma 6 that G^* is a probability measure such that $\mathrm{E}(G^*)=\mu$ and $D(F\|G^*)=L(\nu^*;F,\mu)=L_{\max}(F,\mu)$. Therefore the mixture distribution $(1-\epsilon)G^*+\epsilon\delta_1$ satisfies $\mathrm{E}((1-\epsilon)G^*+\epsilon\delta_1)=(1-\epsilon)\mu+\epsilon>\mu$ for any $\epsilon\in(0,1)$ where δ_1 is the point mass measure at 1. As a result,

$$\begin{aligned}D_{\mathrm{inf}}(F,\mu;\mathcal{A})&\le D(F\|(1-\epsilon)G^*+\epsilon\delta_1)\\&\le\int\log\frac{\mathrm{d}F}{\mathrm{d}((1-\epsilon)G^*)}\mathrm{d}F\\&=D(F\|G^*)-\log(1-\epsilon)\\&=L_{\max}(F,\mu)-\log(1-\epsilon)\end{aligned}$$

and we obtain $D_{\mathrm{inf}}(F,\mu;\mathcal{A})\le L_{\max}(F,\mu)$ by letting $\epsilon\downarrow 0$.

Next we show the latter inequality. Let $A=(-\infty,a]$ and $B=(a,1]$, and define F_A and G_A as probability measures such that $F_A(S)=F(S\cap A)/F(A)$ and $G_A(S)=G(S\cap A)/G(A)$. Then, for any probability measure G such that F is absolutely continuous with respect to G, it holds that

$$\begin{aligned}D(F\|G)&=\int_A\log\frac{\mathrm{d}F}{\mathrm{d}G}\mathrm{d}F+\int_B\log\frac{\mathrm{d}F}{\mathrm{d}G}\mathrm{d}F\\&=F(A)\int_A\log\frac{G(A)}{F(A)}\frac{\mathrm{d}F_A}{\mathrm{d}G_A}\mathrm{d}F_A+\int_B\log\frac{\mathrm{d}F}{\mathrm{d}G}\mathrm{d}F\\&=F(A)\int_A\log\frac{G(A)}{F(A)}\mathrm{d}F_A+F(A)\int_A\log\frac{\mathrm{d}F_A}{\mathrm{d}G_A}\mathrm{d}F_A+\int_B\log\frac{\mathrm{d}F}{\mathrm{d}G}\mathrm{d}F\\&=F(A)\log\frac{G(A)}{F(A)}+F(A)D(F_A\|G_A)+\int_B\log\frac{\mathrm{d}F}{\mathrm{d}G}\mathrm{d}F\\&\ge F(A)\log\frac{G(A)}{F(A)}+\int_B\log\frac{\mathrm{d}F}{\mathrm{d}G}\mathrm{d}F\\&=D(F_{(a)}\|G_{(a)})\end{aligned}$$

and therefore,

$$\begin{aligned}\inf_{G\in\mathcal{A}:\mathrm{E}(G)>\mu} D(F\|G) &\geq \inf_{G\in\mathcal{A}:\mathrm{E}(G)>\mu} D(F_{(a)}\|G_{(a)})\\ &\geq \inf_{G\in\mathcal{A}_a:\mathrm{E}(G_{(a)})>\mu} D(F_{(a)}\|G_{(a)}) \quad (\text{by } \mathrm{E}(G)\leq \mathrm{E}(G_{(a)}))\,.\end{aligned}$$

Let $F'_{(a)}$ and $G'_{(a)}$ be the probability distributions of $(X-a)/(1-a)$ when X follows $F_{(a)}$ and $G_{(a)}$, respectively. Then, letting $\epsilon>0$ be arbitrary and $a<\mu$ be sufficiently small, we obtain from invariance of KL divergence under scale transformation that

$$\begin{aligned}\inf_{G\in\mathcal{A}:\mathrm{E}(G)>\mu} D(F\|G) &\geq \inf_{G\in\mathcal{A}:\mathrm{E}(G_{(a)})>\mu} D(F_{(a)}\|G_{(a)})\\ &= \inf_{G\in\mathcal{A}:\mathrm{E}(G'_{(a)})>\frac{\mu-a}{1-a}} D(F'_{(a)}\|G'_{(a)})\\ &= D_{\mathrm{inf}}\left(F'_{(a)},\frac{\mu-a}{1-a};\mathcal{A}_0\right)\\ &= L_{\max}\left(F'_{(a)},\frac{\mu-a}{1-a}\right) \qquad (\text{by Prop. 1})\\ &= L_{\max}\left(F_{(a)},\mu\right) \quad (\text{by expression of } L_{\max} \text{ in (2)})\\ &\geq L_{\max}(F,\mu)-\epsilon \qquad (\text{by Lemma 8})\end{aligned}$$

and we complete the proof by letting $\epsilon\downarrow 0$. ■

7. Large Deviation Probabilities for Empirical Distributions Measured with D_{inf}

It is essential for evaluation of IMED to derive large deviation probabilities on $\hat{F}_{i,t}$ and $\hat{\mu}_{i,t}$. In this section we discuss probabilities on the empirical distribution and the mean from a generic distribution $F\in\mathcal{A}$, for which we write $(\hat{F}_t,\hat{\mu}_t)$ by dropping the subscript i from $(\hat{F}_{i,t},\hat{\mu}_{i,t})$.

The key to the non-asymptotic evaluation lies in the fact that

$$\begin{aligned}D_{\mathrm{inf}}(\hat{F}_t,\mu) &= \max_{0\leq\nu\leq\frac{1}{1-\mu}} \mathrm{E}_{\hat{F}_t}[\log(1-(X-\mu)\nu)]\\ &= \max_{0\leq\nu\leq\frac{1}{1-\mu}} \left\{\frac{1}{t}\sum_{l=1}^{t}\log(1-(X_l-\mu)\nu)\right\},\end{aligned}$$

where each X_l follows distribution F. Although it involves a maximization, it is essentially an empirical mean of one-dimensional random variables $\log(1-(X_l-\mu)\nu)$. By Cramér's theorem below, we can bound the large deviation probability for such an empirical mean in a non-asymptotic form.

Proposition 9 (Dembo and Zeitouni, 1998, Eqs. (2.2.12) and (2.2.13)) *Assume that the moment generating function* $\mathrm{E}_F[\mathrm{e}^{\lambda X}]$ *exists in some neighborhood of* $\lambda=0$*. Then, for* *any $x \in \mathbb{R}$*

$$\frac{1}{t}\log P_F[\hat{\mu}_t \ge x] \le -\sup_{\lambda \ge 0}\left\{\lambda x - \log \mathrm{E}_F[\mathrm{e}^{\lambda X}]\right\} .$$

Also, if $x < \mathrm{E}(F)$ then

$$\frac{1}{t}\log P_F[\hat{\mu}_t \le x] \le -\Lambda^*(x) \tag{17}$$

and if $x > \mathrm{E}(F)$ then

$$\frac{1}{t}\log P_F[\hat{\mu}_t \ge x] \le -\Lambda^*(x) \tag{18}$$

where $\Lambda^*(x) = \sup_\lambda\{\lambda x - \log \mathrm{E}_F[\mathrm{e}^{\lambda X}]\}$.

We prove Props. 10–12 given below by Cramér's theorem.

Proposition 10 *For any $F \in \mathcal{A}, \mu > \mathrm{E}(F)$ and $u < D_{\mathrm{inf}}(F,\mu)$,*

$$P_F[D_{\mathrm{inf}}(\hat{F}_t,\mu) \le u] \le \mathrm{e}^{-t\tilde{\Lambda}^*(u)} ,$$

where $\tilde{\Lambda}^*(x) = \sup_\lambda\{\lambda x - \mathrm{E}_F[(1-(X-\mu)\nu^*)^\lambda]\}$ for $\nu^* = \mathrm{argmax}_{0\le\nu\le(1-\mu)^{-1}} \mathrm{E}_F[\log(1-(X-\mu)\nu)]$.

Proof For $\nu^* = \mathrm{argmax}_{0\le\nu\le(1-\mu)^{-1}} \mathrm{E}_F[\log(1-(X-\mu)\nu)]$ we have

$$\begin{aligned} P_F[D_{\mathrm{inf}}(\hat{F}_t,\mu) \le u] &= P_F\left[\max_{0\le\nu\le(1-\mu)^{-1}} \mathrm{E}_{\hat{F}_t}[\log(1-(X-\mu)\nu)] \le u\right] \\ &\le P_F\left[\mathrm{E}_{\hat{F}_t}[\log(1-(X-\mu)\nu^*)] \le u\right] . \end{aligned}$$

For $X_1, X_2, \cdots$ following distribution F, we can regard $\mathrm{E}_{\hat{F}_t}[\log(1-(X-\mu)\nu^*)]$ as the empirical mean of $Y_i = \log(1-(X_i-\mu)\nu^*)$, $i = 1,\cdots,t$, which has expectation $D_{\mathrm{inf}}(F,\mu)$. Then the theorem follows immediately from (17) of Prop. 9. ■

Proposition 11 *Fix any $F \in \mathcal{A}$ and $\mu < \mathrm{E}(F)$ and assume that the moment generating function $\mathrm{E}_F[\mathrm{e}^{\lambda X}]$ of F exists in some neighborhood of $\lambda = 0$. (i) For $\lambda_\mu = \sup\{\lambda \in \mathbb{R}\cup\{+\infty\} : \mathrm{E}_F[((1-X)/(1-\mu))^\lambda] \le 1\}$, we have $\lambda_\mu > 1$. (ii) For any $u \in \mathbb{R}$,*

$$P_F[D_{\mathrm{inf}}(\hat{F}_t,\mu) \ge u,\ \hat{\mu}_t \le \mu] \quad \le \quad \begin{cases} \mathrm{e}^{-t\Lambda^*(\mu)}, & \text{if } u \le \Lambda^*(\mu)/\lambda_\mu , \\ 2\mathrm{e}(1+\lambda_\mu t)\mathrm{e}^{-t\lambda_\mu u}, & \text{otherwise.} \end{cases}$$

where $\Lambda^*(x) = \sup_\lambda\{\lambda x - \log \mathrm{E}_F[\mathrm{e}^{\lambda X}]\}$ and we define $\lambda \mathrm{e}^{-\lambda} = 0$ for $\lambda = +\infty$.

Remark 1 Since $D_{\inf}(\hat{F}_t, \mu) \ge u$ implies

$$\begin{aligned} D(\hat{F}_t \| F) &\ge D_{\inf}(\hat{F}_t, \mathrm{E}(F)) \\ &\ge D_{\inf}(\hat{F}_t, \mu) \\ &\ge u\,, \end{aligned}$$

it is easy to prove from Sanov's theorem (Dembo and Zeitouni, 1998, Chap. 6.2) that

$$\limsup_{t\to\infty} \frac{1}{t} \log P_F[D_{\inf}(\hat{F}_t, \mu) \ge u,\ \hat{\mu}_t \le \mu] \le -u\,,$$

that is, $P_F[D_{\inf}(\hat{F}_t, \mu) \ge u,\ \hat{\mu}_t \le \mu]$ is roughly bounded by e^{-tu}. Prop. 11 shows that this bound can be refined to $\mathrm{e}^{-t\lambda_\mu u}$ for large u and its coefficient is explicitly bounded by a polynomial $2\mathrm{e}(1 + \lambda_\mu t)$.

Proof of Proposition 11 (i) Since we assume $\mathrm{E}[\mathrm{e}^{\lambda X}] < \infty$ in some neighborhood of $\lambda = 0$,

$$\mathrm{E}_F\left[\left(\frac{1-X}{1-\mu}\right)^{\lambda}\right] = \frac{\mathrm{E}_F[(1-X)^\lambda]}{(1-\mu)^\lambda}$$

is finite and continuous in $\lambda \ge 0$. We obtain $\lambda_\mu > 1$ from

$$\mathrm{E}_F\left[\left(\frac{1-X}{1-\mu}\right)^{1}\right] = \frac{1-\mathrm{E}(F)}{1-\mu} < 1\,.$$

(ii) Fix an arbitrary $\delta > 0$ and let $M_\delta = \lceil 1/(2\delta(1-\mu)) \rceil$. Define $\nu_{(m)}$ for $m = -M_\delta, -M_\delta + 1, \cdots, 0, \cdots, M_\delta$ by

$$\nu_{(m)} = \frac{1 + \frac{m}{M_\delta}}{2(1-\mu)}\,.$$

Then $\{[\nu_{(m)}, \nu_{(m+1)}]\}_{m=-M_\delta,\cdots,M_\delta-1}$ partitions $[0, (1-\mu)^{-1}]$ into intervals with length at most δ. Therefore the event $\{D_{\inf}(\hat{F}_t, \mu) \ge u\}$ can be expressed as

$$\begin{aligned} \{D_{\inf}(\hat{F}_t, \mu) \ge u\} &= \left\{\exists \nu \in \left[0, \tfrac{1}{1-\mu}\right],\ L(\nu; \hat{F}_t, \mu) \ge u\right\} \\ &= \bigcup_{m=-M_\delta}^{-1} \left\{\exists \nu \in \left[\nu_{(m)}, \nu_{(m+1)}\right], L(\nu; \hat{F}_t, \mu) \ge u\right\} \\ &\qquad \cup \bigcup_{m=1}^{M_\delta} \left\{\exists \nu \in \left[\nu_{(m-1)}, \nu_{(m)}\right], L(\nu; \hat{F}_t, \mu) \ge u\right\}. \end{aligned} \tag{19}$$

Since $|\nu_{(m+1)} - \nu_{(m)}| \le \delta$ and $L(\nu; \hat{F}_t, \mu)$ is concave in ν, it holds for $m \le -1$ that

$$\begin{aligned} &\left\{\exists \nu \in \left[\nu_{(m)}, \nu_{(m+1)}\right],\ L(\nu; \hat{F}_t, \mu) \ge u\right\} \\ &\quad \subset \left\{L(\nu_{(m+1)}; \hat{F}_t, \mu) - \delta \min\{0, L'(\nu_{(m+1)}; \hat{F}_t, \mu)\} \ge u\right\} \\ &\quad \subset \left\{L(\nu_{(m+1)}; \hat{F}_t, \mu) - \delta \min\{0, L'(\nu_{(0)}; \hat{F}_t, \mu)\} \ge u\right\}. \end{aligned} \tag{20}$$

Similarly it holds for $m \ge 1$ that

$$\begin{aligned}&\left\{\exists \nu \in \left[\nu_{(m-1)}, \nu_{(m)}\right],\, L(\nu;\hat{F}_t,\mu) \ge u\right\}\\ &\quad\subset \left\{L(\nu_{(m-1)};\hat{F}_t,\mu) + \delta \max\{0, L'(\nu_{(0)};\hat{F}_t,\mu)\} \ge u\right\}. \qquad (21)\end{aligned}$$

Here the derivative L' is expressed from (12) as

$$L'(\nu;\hat{F}_t,\mu) = \frac{1}{\nu} - \frac{1}{\nu}\mathrm{E}_{\hat{F}_t}\left[\frac{1}{1-(X-\mu)\nu}\right].$$

Since $1/(1-(x-\mu)\nu)$ is positive and increasing in $x \le 1$, it is bounded as

$$\frac{1}{\nu} \ge L'(\nu;\hat{F}_t,\mu) \ge \frac{1}{\nu} - \frac{1}{\nu}\frac{1}{1-(1-\mu)\nu} = -\frac{1-\mu}{1-(1-\mu)\nu}.$$

Thus $L'(\nu_{(0)};\hat{F}_t,\mu) = L'(1/(2(1-\mu));\hat{F}_t,\mu)$ is bounded as

$$2(1-\mu) \ge L'(\nu_{(0)};\hat{F}_t,\mu) \ge -2(1-\mu).$$

Combining this with (19), (20) and (21) we obtain

$$P_F[D_{\inf}(\hat{F}_t,\mu) \ge u] \le \sum_{\substack{m\neq 0:\\ -M_\delta \le m \le M_\delta}} P_F\left[L(\nu_{(m)};\hat{F}_t,\mu) \ge u - 2(1-\mu)\delta\right]. \qquad (22)$$

Now recall that

$$\lambda_\mu = \sup\left\{\lambda : \mathrm{E}_F\left[\left(\frac{1-X}{1-\mu}\right)^\lambda\right] \le 1\right\} > 1.$$

Then, by Prop. 9,

$$\begin{aligned}&P_F\left[L(\nu_{(m)};\hat{F}_t,\mu) \ge u - 2(1-\mu)\delta\right]\\ &\le \exp\left(-t\sup_{\lambda\ge 0}\left\{\lambda(u-2(1-\mu)\delta) - \log \mathrm{E}_F[\mathrm{e}^{\lambda\log(1-(X-\mu)\nu_{(m)})}]\right\}\right)\\ &\le \exp\Bigg(-t\sup_{\lambda\ge 1}\Big\{\lambda(u-2(1-\mu)\delta)\\ &\qquad\qquad -\log\left(\mathrm{E}_F[\mathrm{e}^{\lambda\log(1-(X-\mu)\cdot 0)}] \vee \mathrm{E}_F[\mathrm{e}^{\lambda\log(1-(X-\mu)\cdot(1-\mu)^{-1})}]\right)\Big\}\Bigg) \qquad (23)\\ &= \exp\left(-t\sup_{\lambda\ge 1}\left\{\lambda(u-2(1-\mu)\delta) - \log\left(1 \vee \mathrm{E}_F\left[\left(\frac{1-X}{1-\mu}\right)^\lambda\right]\right)\right\}\right)\\ &\le \exp\left(-t\lambda_\mu(u-2(1-\mu)\delta)\right), \qquad (24)\end{aligned}$$

where (23) follows from $0 \le \nu_{(m)} \le (1-\mu)^{-1}$ and the convexity of $\mathrm{E}_F[\mathrm{e}^{\lambda\log(1-(X-\mu)\nu)}]$ in $\nu \in [0,(1-\mu)^{-1}]$ for $\lambda \ge 1$. Therefore we obtain from (22) and (24) that

$$\begin{aligned}P_F[D_{\inf}(\hat{F}_t,\mu) \ge u] &\le 2M_\delta \exp\left(-t\lambda_\mu(u-2(1-\mu)\delta)\right)\\ &\le 2\left(1+\frac{1}{2(1-\mu)\delta}\right)\exp\left(-t\lambda_\mu(u-2(1-\mu)\delta)\right)\end{aligned}$$

and we complete the proof by letting $\delta = 1/(2t\lambda_\mu(1-\mu))$ and combining it with (17). ■

We prove Theorem 3 by the above two propositions. We also use the following proposition on the large deviation probability of $D_{\mathrm{inf}}(\hat{F}_t, \mu)$ under a more general setting for the proof of Theorem 5.

Proposition 12 *Fix any $u, \mu \in \mathbb{R}$ and $F \in \mathcal{A}$ such that $\mathrm{E}(F) < \mu < 1$. Then*

$$P_F[D_{\mathrm{inf}}(\hat{F}_t, \mu) \ge u] \le 2\mathrm{e}(1+t)\exp\left(-t\left(u - \log\frac{1-\mathrm{E}(F)}{1-\mu}\right)\right).$$

Proof Since (22) and (23) also hold for the case of this theorem, we obtain the theorem by letting $\lambda = 1$ and $\delta = 1/(2t(1-\mu))$. ■

8. Regret Analysis for IMED

In this section we prove Theorem 3 by using a technique similar to that for UCB policies. First we prove Lemma 13 below as a fundamental property of the IMED policy on the minimum index $I^*(l) \equiv \min_{i\in\{1,2,\cdots,K\}} I_i(l)$.

Lemma 13 *For any $x > 0$ and arm i,*

$$\sum_{l=1}^{\infty} \mathbb{1}[I^*(l) \le x,\ J(l) = i] \le \mathrm{e}^x .$$

Proof This is straightforward from

$$\begin{aligned}
\sum_{l=1}^{\infty} \mathbb{1}[I^*(l) \le x,\ J(l) = i] &= \sum_{t=1}^{\infty}\sum_{l=1}^{\infty} \mathbb{1}[I^*(l) \le x,\ J(l) = i,\ T_i(l) = t] \\
&\le \sum_{t=1}^{\infty}\sum_{l=1}^{\infty} \mathbb{1}[\log t \le x,\ J(l) = i,\ T_i(l) = t] \\
&\qquad\qquad (J(l) = i \text{ implies } I^*(l) = I_i(l) \ge \log T_i(l)) \\
&= \sum_{t=1}^{\lfloor \mathrm{e}^x \rfloor}\sum_{l=1}^{\infty} \mathbb{1}[J(l) = i,\ T_i(l) = t] \\
&\le \sum_{t=1}^{\lfloor \mathrm{e}^x \rfloor} 1 \qquad (\{J(l) = i,\ T_i(l) = t\} \text{ occurs for at most one } l) \\
&\le \mathrm{e}^x .
\end{aligned}$$

■

We prove Theorem 3 by Lemma 14 below.

Lemma 14 *It holds for any $\mu < \mu^*$ and arm i that*

$$\mathrm{E}\left[\sum_{l=1}^{\infty} \mathbb{1}[\hat{\mu}^*(l) \le \mu,\ J(l) = i]\right] \le \inf_{j\in\mathcal{I}_{\mathrm{opt}}} \left\{ \frac{6\mathrm{e}}{(1 - 1/\lambda_{j,\mu})(1 - \mathrm{e}^{-(1-1/\lambda_{j,\mu})\Lambda_j^*(\mu)})^3} \right\}.$$

Proof Let j be any optimal arm, that is, j such that $\Delta_j = 0$. We will bound the RHS of

$$\begin{aligned}\sum_{l=1}^{\infty} \mathbb{1}[\hat{\mu}^*(l) \le \mu,\, J(l) = i] \;&=\; \sum_{l=1}^{\infty} \mathbb{1}[\hat{\mu}_j(l) \le \hat{\mu}^*(l) \le \mu,\, J(l) = i] \\ &\le\; \sum_{t=1}^{\infty}\sum_{l=1}^{\infty} \mathbb{1}[\hat{\mu}_{j,t} \le \hat{\mu}^*(l) \le \mu,\, T_j(l) = t,\, J(l) = i]\,. \qquad (25)\end{aligned}$$

Since $\{\hat{\mu}_{j,t} \le \hat{\mu}^*(l) \le \mu,\, T_j(l) = t\}$ implies

$$\begin{aligned} I^*(l) &= \min_i I_i(l) \\ &\le I_j(l) \\ &= tD_{\mathrm{inf}}(\hat{F}_{j,t}, \hat{\mu}^*(l)) + \log t \\ &\le tD_{\mathrm{inf}}(\hat{F}_{j,t}, \mu) + \log t\,, \end{aligned}$$

we see from Lemma 13 that $\{\hat{\mu}_{j,t} \le \hat{\mu}^*(l) \le \mu,\, T_j(l) = t,\, J(l) = i\}$ occurs for at most $t\mathrm{e}^{tD_{\mathrm{inf}}(\hat{F}_{j,t},\mu)}$ rounds. Therefore from (25) we obtain

$$\sum_{l=1}^{\infty} \mathbb{1}[\hat{\mu}^*(l) \le \mu,\, J(l) = i] \le \sum_{t=1}^{\infty} \mathbb{1}[\hat{\mu}_{j,t} \le \mu]\, t\mathrm{e}^{tD_{\mathrm{inf}}(\hat{F}_{j,t},\mu)}\,. \qquad (26)$$

Let $P(u) \equiv P_{F_j}[D_{\mathrm{inf}}(\hat{F}_{j,t}, \mu) > u,\, \hat{\mu}_{j,t} \le \mu]$. Simply writing λ_j and Λ_j^* for $\lambda_{j,\mu}$ and $\Lambda_j^*(\mu)$ in (4) and (5), respectively, we have from Prop. 11 that

$$\begin{aligned} &\mathrm{E}\left[\mathbb{1}[\hat{\mu}_{j,t} \le \mu]\, t\mathrm{e}^{tD_{\mathrm{inf}}(\hat{F}_{j,t},\mu)}\right] \\ &= \int_0^\infty t\mathrm{e}^{tu}(-\mathrm{d}P(u)) \\ &= \left[t\mathrm{e}^{tu}(-P(u))\right]_0^\infty + \int_0^\infty t^2\mathrm{e}^{tu}P(u)\mathrm{d}u \qquad \text{(integration by parts)} \\ &\le t\mathrm{e}^{-t\Lambda_j^*} + \int_0^{\Lambda_j^*/\lambda_j} t^2\mathrm{e}^{tu}\cdot\mathrm{e}^{-t\Lambda_j^*}\mathrm{d}u + \int_{\Lambda_j^*/\lambda_j}^\infty t^2\mathrm{e}^{tu}\cdot 2\mathrm{e}(1+\lambda_j t)\mathrm{e}^{-t\lambda_j u}\mathrm{d}u \\ &= t\mathrm{e}^{-t\Lambda_j^*} + t\left[\mathrm{e}^{t(u-\Lambda_j^*)}\right]_0^{\Lambda_j^*/\lambda_j} - 2\mathrm{e}t(1+\lambda_j t)\left[\frac{\mathrm{e}^{-t(\lambda_j-1)u}}{\lambda_j - 1}\right]_{\Lambda_j^*/\lambda_j}^\infty \\ &= t\mathrm{e}^{-t(1-1/\lambda_j)\Lambda_j^*} + 2\mathrm{e}t(1+\lambda_j t)\frac{\mathrm{e}^{-t(1-1/\lambda_j)\Lambda_j^*}}{\lambda_j - 1} \\ &= \left(\frac{1 - 1/\lambda_j + 2\mathrm{e}/\lambda_j}{1 - 1/\lambda_j}\right)\cdot t\mathrm{e}^{-t(1-1/\lambda_j)\Lambda_j^*} + \frac{2\mathrm{e}}{1-1/\lambda_j}\cdot t^2\mathrm{e}^{-t(1-1/\lambda_j)\Lambda_j^*}\,. \qquad (27) \end{aligned}$$

From (26), (27) and formulas

$$\begin{aligned} \sum_{t=1}^{\infty} t\mathrm{e}^{-rt} \;&\le\; \frac{1}{(1-\mathrm{e}^{-r})^2} \le \frac{1}{(1-\mathrm{e}^{-r})^3} \\ \sum_{t=1}^{\infty} t^2\mathrm{e}^{-rt} \;&\le\; \frac{2}{(1-\mathrm{e}^{-r})^3}\,, \end{aligned}$$

it holds that

$$\begin{aligned}\mathrm{E}\left[\sum_{l=1}^{\infty}\mathbb{1}[\hat{\mu}^*(l)\le\mu,\,J(l)=i]\right]&\le\left(\frac{1+(2\mathrm{e}-1)/\lambda_j+4\mathrm{e}}{1-1/\lambda_j}\right)\frac{1}{(1-\mathrm{e}^{-t(1-1/\lambda_j)\Lambda_j^*})^3}\\&\le\left(\frac{1+(2\mathrm{e}-1)+4\mathrm{e}}{1-1/\lambda_j}\right)\frac{1}{(1-\mathrm{e}^{-t(1-1/\lambda_j)\Lambda_j^*})^3}\\&=\frac{6\mathrm{e}}{(1-1/\lambda_j)(1-\mathrm{e}^{-t(1-1/\lambda_j)\Lambda_j^*})^3}\,.\end{aligned}\tag{28}$$

We complete the proof by taking j which minimizes (28) over the optimal arms $j\in\mathcal{I}_{\mathrm{opt}}$. ■

Proof of Theorem 3 First we decompose $T_i(n)$ as

$$\begin{aligned}T_i(n)&=\sum_{l=1}^{n}\mathbb{1}[J(l)=i]\\&=\sum_{l=1}^{n}\mathbb{1}[J(l)=i,\,\hat{\mu}^*(l)\le\mu^*-\delta]+\sum_{l=1}^{n}\mathbb{1}[J(l)=i,\,\hat{\mu}^*(l)\ge\mu^*-\delta]\,.\end{aligned}\tag{29}$$

The summation of the second term of (29) is bounded as

$$\begin{aligned}\sum_{l=1}^{n}\mathbb{1}[J(l)=i,\,\hat{\mu}^*(l)\ge\mu^*-\delta]&=\sum_{t=1}^{n}\mathbb{1}\left[\bigcup_{l=1}^{n}\{J(l)=i,\,T_i(l)=t,\,\hat{\mu}^*(l)\ge\mu^*-\delta\}\right]\\&\le\sum_{t=1}^{n}\mathbb{1}\left[\bigcup_{l=1}^{n}\{I_i(l)=I^*(l),\,T_i(l)=t,\,\hat{\mu}^*(l)\ge\mu^*-\delta\}\right].\end{aligned}$$

Note that $I^*(l)\le\max_{i:\hat{\mu}_i(l)=\hat{\mu}^*(l)}I_i(l)=\max_{i:\hat{\mu}_i(l)=\hat{\mu}^*(l)}\log T_i(l)\le\log n$ for all $l\le n$. Then we have

$$\begin{aligned}&\mathrm{E}\left[\sum_{l=1}^{n}\mathbb{1}[J(l)=i,\,\hat{\mu}^*(l)\ge\mu^*-\delta]\right]\\&\le\mathrm{E}\left[\sum_{t=1}^{n}\mathbb{1}\left[tD_{\mathrm{inf}}(\hat{F}_{i,t},\mu^*-\delta)\le\log n\right]\right]\qquad(\text{by }I^*(l)=I_i(l)\ge tD_{\mathrm{inf}}(\hat{F}_i(l),\hat{\mu}^*(l)))\\&=\sum_{t=1}^{\infty}P_{F_i}\left[tD_{\mathrm{inf}}(\hat{F}_{i,t},\mu^*-\delta)\le\log n\right]\\&=\sum_{t=1}^{\infty}P_{F_i}\left[t\left(D_{\mathrm{inf}}(\hat{F}_{i,t},\mu^*)-\int_{\mu^*-\delta}^{\mu^*}\left.\frac{\mathrm{d}D_{\mathrm{inf}}(\hat{F}_{i,t},\mu)}{\mathrm{d}\mu}\right|_{\mu=u}\mathrm{d}u\right)\le\log n\right]\\&\le\sum_{t=1}^{\infty}P_{F_i}\left[t\left(D_{\mathrm{inf}}(\hat{F}_{i,t},\mu^*)-\int_{\mu^*-\delta}^{\mu^*}\frac{\mathrm{d}u}{1-u}\right)\le\log n\right]\qquad(\text{by Lemma 7})\\&\le\sum_{t=1}^{\infty}P_{F_i}\left[t\left(D_{\mathrm{inf}}(\hat{F}_{i,t},\mu^*)-\frac{\delta}{1-\mu^*}\right)\le\log n\right].\end{aligned}$$

By letting

$$M = \left\lceil \frac{\log n}{D_{\mathrm{inf}}(F_i, \mu^*) - \frac{2\delta}{1-\mu^*}} \right\rceil ,$$

we have

$$\begin{aligned}
&\mathrm{E}\left[\sum_{l=1}^{n} \mathbb{1}[J(l) = i,\ \hat{\mu}^*(l) \geq \mu^* - \delta]\right] \\
&\leq M - 1 + \sum_{t=M}^{\infty} P_{F_i}\left[t\left(D_{\mathrm{inf}}(\hat{F}_{i,t}, \mu^*) - \frac{\delta}{1-\mu^*}\right) \leq \log n\right] \\
&\leq M - 1 + \sum_{t=M}^{\infty} P_{F_i}\left[M\left(D_{\mathrm{inf}}(\hat{F}_{i,t}, \mu^*) - \frac{\delta}{1-\mu^*}\right) \leq \log n\right] \\
&\leq M - 1 + \sum_{t=M}^{\infty} P_{F_i}\left[D_{\mathrm{inf}}(\hat{F}_{i,t}, \mu^*) \leq D_{\mathrm{inf}}(F_i, \mu^*) - \frac{\delta}{1-\mu^*}\right] \\
&\leq M - 1 + \sum_{t=M}^{\infty} \mathrm{e}^{-t\tilde{\Lambda}(D_{\mathrm{inf}}(F_i,\mu^*) - \frac{\delta}{1-\mu^*})} \qquad \text{(by Prop. 10)} \\
&\leq \frac{\log n}{D_{\mathrm{inf}}(F_i, \mu^*) - \frac{2\delta}{1-\mu^*}} + \frac{1}{1 - \mathrm{e}^{-\tilde{\Lambda}_i^*(D_{\mathrm{inf}}(F_i,\mu^*) - \frac{\delta}{1-\mu})}} .
\end{aligned}$$

On the other hand, we can bound the expectation of the first term of (29) by Lemma 14 with $\mu := \mu^* - \delta$, which completes the proof of the theorem. ■

9. Concluding Remarks and Discussion

We considered a nonparametric stochastic bandit where only the upper bound of the reward is known. We proved that the theoretical bound does not depend on the knowledge of the lower bound of the reward. We also showed that the bound can be achieved by the IMED policy, an indexed version of the DMED policy.

A future work is to examine whether the assumption on existence of moment generating functions $\mathrm{E}_{F_i}[\mathrm{e}^{\lambda X}]$ can be weakened to existence of moments $\mathrm{E}_{F_i}[X^m]$. In the analysis of IMED it is important to evaluate tail probabilities of $\hat{\mu}_{i,t}$ and $D_{\mathrm{inf}}(\hat{F}_{i,t}, \mu) = \max_{0 \leq \nu \leq (1-\mu)^{-1}} \mathrm{E}_{\hat{F}_{i,t}}[\log(1 - (X - \mu)\nu)]$. Although the latter one is more essential in the behavior of IMED, this only requires the existence of the moment $\mathrm{E}[\mathrm{e}^{\lambda \log(1-(X-\mu)\nu)}] = \mathrm{E}[(1 - (X - \mu)\nu)^{\lambda}]$ and we assumed the existence of $\mathrm{E}_{F_i}[\mathrm{e}^{\lambda X}]$ only for the evaluation of $\hat{\mu}_{i,t}$. Furthermore, in the most part of evaluations involving $\hat{\mu}_{i,t}$ it suffices to show that

$$\sum_{t=1}^{\infty} t^p \Pr[|\hat{\mu}_{i,t} - \mu_i| > \delta] < \infty \tag{30}$$

for some $p \geq 0$, which we can assure to hold only by assuming $\mathrm{E}_{F_i}[X^{2+p}] < \infty$ (Chow and Lai, 1975). From these reasons we conjecture that the assumption $\mathrm{E}[\mathrm{e}^{\lambda X}] < \infty$ can be weakened by using (30) but it remains as an open problem.

Acknowledgments

This work was supported by JSPS Grant-in-Aid for Scientific Research No. 26106506, 25220001.

Appendix A. Representations of Constants for Large Deviation Probabilities

In Theorem 3, $\lambda_{i,\mu}$, $\Lambda_i^*(x)$ and $\tilde{\Lambda}_i^*(x)$ in (4)–(6) are used in the constant term of the regret. We discuss explicit representations of them in this appendix.

First we evaluate $\Lambda_i^*(x)$ and $\tilde{\Lambda}_i^*(x)$, which are Legendre-Fenchel transforms of cumulant generating functions of random variables X and $Y = \log(1-(X-\mu^*)\nu_i^*)$, respectively, where X follows F_i. If the support of F_i is bounded from below by $a > -\infty$ then by Hoeffding's inequality (Hoeffding, 1963) we have

$$\Lambda_i^*(\mu_i + \delta) \geq \frac{2\delta^2}{(1+a)^2}\,.$$

Similarly, from $Y \in [\log(1-(1-\mu^*)\nu_i^*),\, \log(1-(a-\mu^*)\nu_i^*)]$

$$\tilde{\Lambda}_i^*(D_{\mathrm{inf}}(F_i,\mu^*) - \delta) \geq \frac{2\delta^2}{\left(\log\frac{1-(a-\mu^*)\nu_i^*}{1-(1-\mu^*)\nu_i^*}\right)^2}\,.$$

Furthermore, we can evaluate $\Lambda_i^*(\mu_i+\delta)$ and $\tilde{\Lambda}_i^*(D_{\mathrm{inf}}(F_i,\mu^*)-\delta)$ for general cases including $a = -\infty$ by the following lemma.

Lemma 15 *For sufficiently small $\delta > 0$,*

$$\Lambda_i^*(\mu_i + \delta) \geq \frac{\delta^2}{2\sigma_i^2} + \mathrm{o}(\delta^2)\,, \tag{31}$$

$$\tilde{\Lambda}_i^*(D_{\mathrm{inf}}(F_i,\mu^*) - \delta) \geq \frac{(1-\mu^*)\delta^2}{4(1-\mu_i)} + \mathrm{o}(\delta^2)\,, \tag{32}$$

where $\sigma_i^2 = \mathrm{E}_{F_i}[(X-\mu_i)^2]$ is the variance of F_i.

Proof Since the cumulant generating function of F_i is expressed as

$$\log \mathrm{E}_{F_i}[\mathrm{e}^{\lambda X}] = \mu_i\lambda + \frac{\sigma_i^2\lambda^2}{2} + \mathrm{o}(\lambda^2)\,,$$

we obtain (31) from

$$\begin{aligned}\Lambda_i^*(\mu_i+\delta) &= \sup_\lambda \left\{(\mu_i+\delta)\lambda - \log \mathrm{E}_{F_i}[\mathrm{e}^{\lambda X}]\right\}\\ &= \sup_\lambda \left\{\delta\lambda - \frac{\sigma_i^2\lambda^2}{2} + \mathrm{o}(\lambda^2)\right\}\\ &\geq \frac{\delta^2}{2\sigma_i^2} + \mathrm{o}(\delta^2)\,. \qquad \left(\text{by letting } \lambda := \delta/\sigma_i^2\right)\end{aligned}$$

Similarly, from $\mathrm{E}_{F_i}[Y] = D_{\mathrm{inf}}(F_i, \mu^*)$ we have

$$\begin{aligned}\tilde{\Lambda}_i^*(D_{\mathrm{inf}}(F_i,\mu^*)-\delta) &= \sup_{\lambda}\left\{(D_{\mathrm{inf}}(F_i,\mu^*)-\delta)\lambda - \log \mathrm{E}_{F_i}[\mathrm{e}^{\lambda Y}]\right\}\\ &\ge \frac{\delta^2}{2\tilde{\sigma}_i^2} + \mathrm{o}(\delta^2)\,, \end{aligned} \tag{33}$$

where $\tilde{\sigma}_i^2$ is the variance of $Y = \log(1-(X-\mu^*)\nu_i^*)$. Since Y has expectation $\mathrm{E}_{F_i}[Y] = D_{\mathrm{inf}}(F_i, \mu^*)$, the variance $\tilde{\sigma}_i^2$ is expressed as

$$\begin{aligned}\tilde{\sigma}_i^2 &= \mathrm{E}_{F_i}[(Y - D_{\mathrm{inf}}(F_i,\mu^*))^2]\\ &= \mathrm{E}_{F_i}\left[\left(\log\frac{\mathrm{e}^Y}{\mathrm{e}^{D_{\mathrm{inf}}(F_i,\mu^*)}}\right)^2\right]\,.\end{aligned}$$

Note that $(\log z)^2$ is smaller than z^{-1} for $z \to +0$ and smaller than z for $z \to \infty$. Thus there exists $c_0 > 0$ such that $(\log z)^2 \le c_0(z + z^{-1})$ for all $z > 0$. In fact, this inequality holds for $c_0 \ge 0.533$ (and thus, for $c_0 = 1$). Therefore

$$\begin{aligned}\tilde{\sigma}_i^2 &\le \mathrm{E}_{F_i}\left[\frac{\mathrm{e}^Y}{\mathrm{e}^{D_{\mathrm{inf}}(F_i,\mu^*)}} + \frac{\mathrm{e}^{D_{\mathrm{inf}}(F_i,\mu^*)}}{\mathrm{e}^Y}\right]\\ &\le \mathrm{E}_{F_i}[\mathrm{e}^Y] + \mathrm{e}^{D_{\mathrm{inf}}(F_i,\mu^*)}\mathrm{E}_{F_i}[\mathrm{e}^{-Y}] && (\text{by } D_{\mathrm{inf}}(F_i,\mu^*) \ge 0)\\ &= \mathrm{E}_{F_i}[\mathrm{e}^Y] + \mathrm{e}^{\mathrm{E}_{F_i}[Y]}\mathrm{E}_{F_i}[\mathrm{e}^{-Y}]\\ &\le \mathrm{E}_{F_i}[\mathrm{e}^Y] + \mathrm{E}_{F_i}[\mathrm{e}^Y]\mathrm{E}_{F_i}[\mathrm{e}^{-Y}] && (\text{by Jensen's inequality})\\ &= (1-(\mu_i-\mu^*)\nu_i^*)\cdot\left(1+\mathrm{E}_{F_i}\left[\frac{1}{1-(X-\mu^*)\nu_i^*}\right]\right)\\ &\le \left(1-\frac{\mu_i-\mu^*}{1-\mu^*}\right)\cdot(1+1) && (\text{by Lemma 6})\\ &= \frac{2(1-\mu_i)}{1-\mu^*}\,. \end{aligned} \tag{34}$$

We obtain (32) by combining (34) with (33). ■

Next we bound $\lambda_{i,\mu}$ with an explicit form in the following lemma and we see that $\lambda_{i,\mu_i-\delta} \ge 1 + (1-\mu_i)\delta/\sigma_i^2 + \mathrm{o}(\delta)$.

Lemma 16 *If $\mu < \mu_i < 1$ then*

$$\lambda_{i,\mu} \ge \begin{cases} 1 + \frac{(1-\mu)(\mu_i-\mu)}{\sigma_i^2-(1-\mu_i)(\mu_i-\mu)}, & \text{if } \sigma_i^2 \ge (\mu_i-\mu)(2-\mu_i-\mu),\\ 2, & \text{otherwise.}\end{cases} \tag{35}$$

Proof Since x^λ is convex in λ, we have

$$\begin{aligned}\lambda_{i,\mu} &= \sup\left\{\lambda : \mathrm{E}_{F_i}\left[\left(\frac{1-X}{1-\mu}\right)^\lambda\right] \le 1\right\}\\ &\ge \sup\left\{\lambda \in [1,2] : \mathrm{E}_{F_i}\left[\left(\frac{1-X}{1-\mu}\right)^\lambda\right] \le 1\right\}\\ &\ge \sup\left\{\lambda \in [1,2] : \mathrm{E}_{F_i}\left[(2-\lambda)\left(\frac{1-X}{1-\mu}\right)^1 + (\lambda-1)\left(\frac{1-X}{1-\mu}\right)^2\right] \le 1\right\}\\ &= \sup\left\{\lambda \in [1,2] : (2-\lambda)\frac{1-\mu_i}{1-\mu} + (\lambda-1)\frac{\sigma_i^2+(1-\mu_i)^2}{(1-\mu)^2} \le 1\right\}. \qquad (36)\end{aligned}$$

If $(\sigma_i^2 + (1-\mu_i)^2)(1-\mu)^{-2} \ge 1$, that is, if $\sigma_i^2 \ge (\mu_i-\mu)(2-\mu_i-\mu)$ then λ satisfying

$$(2-\lambda)\frac{1-\mu_i}{1-\mu} + (\lambda-1)\frac{\sigma_i^2+(1-\mu_i)^2}{(1-\mu)^2} = 1$$

is contained in $[1,2]$. Therefore we obtain (35) for this case by solving this equality. In the other case, the condition in (36) is satisfied by $\lambda = 2$ and we have $\lambda_{i,\mu} \ge 2$. ■

Appendix B. Proof of Lemma 7

We prove this lemma by the technique known as sensitivity analysis for optimization problems given below.

Proposition 17 (Fiacco, 1983, Corollary 3.4.3) *For a function $f(x,y) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, let $f^*(y)$ be a local minimum of $f(x,y)$ in some neighborhood of x. Assume that there exists a point x^* such that*

- *$f(x,y)$ is twice continuously differentiable in some neighborhood of $(x^*,0)$,*
- *$\Delta_x f(x,0)|_{x=x^*} = 0$, and*
- *$\Delta_x^2 f(x,0)|_{x=x^*}$ is positive definite.*

Then $\Delta_y f^*(y) = \Delta_y f(x,y)|_{x=x^*}$.

Proof From Lemma 6, for the case $\mathrm{E}_F[(1-\mu)/(1-X)] < 1$ we have $L_{\max}(F,\mu) = \mathrm{E}_F[\log((1-X)/(1-\mu))]$. Therefore,

$$\frac{\partial}{\partial\mu}L_{\max}(F,\nu) = \frac{1}{1-\mu} = \nu^*(F,\mu)$$

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

$$\lim_{\epsilon\downarrow 0}\frac{L_{\max}(F,\mu+\epsilon) - L_{\max}(F,\mu)}{\epsilon} = \frac{1}{1-\mu} = \nu^*(F,\mu)$$

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

Now consider the case $\mathrm{E}_F[(1-\mu)/(1-X)]\ge 1$. In this case, $L_{\max}(F,\mu)=\max_{0\le\nu\le(1-\mu)^{-1}} L(\nu;F,\mu)=\max_\nu L(\nu;F,\mu)$ from $L'(0;F,\mu)=0$, $L'((1-\mu)^{-1};F,\mu)\le 0$ and the convexity of $L(\nu;F,\mu)$. For this unconstrained optimization problem it holds from Prop. 17 that

$$\frac{\mathrm{d}(\max_\nu L(\nu;F,\mu))}{\mathrm{d}\mu}=\left.\frac{\mathrm{d}L(\nu;F,\mu)}{\mathrm{d}\mu}\right|_{\nu=\nu^*}=\nu^*(F,\mu)\,.$$

Therefore, we obtain

$$\frac{\partial}{\partial\mu}L_{\max}(F,\mu)=\nu^*(F,\mu)$$

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

$$\lim_{\epsilon\uparrow 0}\frac{L_{\max}(F,\mu+\epsilon)-L_{\max}(F,\mu)}{\epsilon}=\nu^*(F,\mu)$$

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

Appendix C. Proof of Theorem 5

In this appendix we show Theorem 5 on the refined (asymptotic) regret bound of IMED. We prove the theorem by the following lemma on a stopping time of a stochastic process.

Lemma 18 *Let $\{Y_i\}_{i=1,2,\cdots}$ be i.i.d. random variables such that $\mathrm{E}[Y_1]>0$ and $\mathrm{E}[\mathrm{e}^{Y_1}]<\infty$.* (i) *For $S_t=\sum_{i=1}^t Y_i$ and sufficiently large $M>0$, the stopping time $\tau=\min\{t:S_t>M\}$ satisfies*

$$\mathrm{E}[\tau]\le\frac{M+\log M}{\mathrm{E}[Y_1]}+\mathrm{O}(1)\,.$$

(ii) *Furthermore, if $\operatorname{ess\,sup} Y_i<\infty$, that is, the support of the distribution of Y_i is bounded from above then*

$$\mathrm{E}[\tau]\le\frac{M}{\mathrm{E}[Y_1]}+\mathrm{O}(1)\,.$$

Proof (i) For any $A>0$, define $Y_i'=Y_i\wedge A$ and $S_t'=\sum_{i=1}^t Y_i'$. For simplicity we also define $S_0'=S_0=0$. Since $S_t'\le S_t$ always holds, $\tau'=\min\{t:S_t'>M\}$ satisfies $\tau\le\tau'$.

Since $\tau_n'=n\wedge\tau'$ is a bounded stopping time, it holds from discrete Dynkin's formula (Meyn and Tweedie, 1992, Sect. 4.2) that

$$\begin{aligned}\mathrm{E}[S'_{\tau_n'}]&=\mathrm{E}[S_0']+\mathrm{E}\left[\sum_{i=1}^{\tau_n'}\mathrm{E}[S_i'|S_1',S_2',\cdots,S_{i-1}']-S_{i-1}'\right]\\&=\mathrm{E}\left[\sum_{i=1}^{\tau_n'}\mathrm{E}[Y_i']\right]\\&=\mathrm{E}[Y_i']\mathrm{E}\left[\tau_n'\right]\end{aligned}$$

and therefore

$$\mathrm{E}[\tau_n'] = \frac{\mathrm{E}[S'_{\tau_n'}]}{\mathrm{E}[Y_1']} \le \frac{\mathrm{E}[S'_{\tau_n'-1} + A]}{\mathrm{E}[Y_1']} \le \frac{M+A}{\mathrm{E}[Y_1']} \,. \tag{37}$$

By defining $(x)_+ = 0 \vee x$, we can bound $\mathrm{E}[Y_1']$ by

$$\begin{aligned} \mathrm{E}[Y_1'] &= \mathrm{E}[Y_1 - (Y_1 - A)_+] \\ &\ge \mathrm{E}[Y_1] - \frac{\mathrm{E}[\mathrm{e}^Y]}{\mathrm{e}^{A+1}} \,. \qquad (\text{by } (y-A)_+ \le \mathrm{e}^{y-(A+1)}) \end{aligned} \tag{38}$$

Combining (37) with (38) and letting $A = \log((M+1)\mathrm{E}[\mathrm{e}^{Y_1}]/\mathrm{E}[Y_1]) - 1$, we have

$$\begin{aligned} \mathrm{E}[\tau_n'] &\le \frac{M+1}{M} \frac{M + \log\left(\frac{\mathrm{E}[\mathrm{e}^{Y_1}]}{\mathrm{E}[Y_1]}(M+1)\right) - 1}{\mathrm{E}[Y_1]} \\ &= \frac{M + \log M}{\mathrm{E}[Y_1]} + \mathrm{O}(1) \,. \end{aligned}$$

Finally we complete the proof by

$$\begin{aligned} \mathrm{E}[\tau] &\le \mathrm{E}[\tau'] \\ &= \mathrm{E}\left[\lim_{n\to\infty} \tau_n'\right] \\ &= \lim_{n\to\infty} \mathrm{E}[\tau_n'] \qquad (\text{by monotone convergence theorem}) \\ &= \frac{M + \log M}{\mathrm{E}[Y_1]} + \mathrm{O}(1) \,. \end{aligned}$$

(ii) In the case of $\operatorname{ess\,sup} Y_i < \infty$, we can directly evaluate τ instead of τ' and (37) is replaced with

$$\mathrm{E}[\tau] \le \frac{M + \operatorname{ess\,sup} Y_i}{\mathrm{E}[Y_1]} = \frac{M}{\mathrm{E}[Y_1]} + \mathrm{O}(1) \,. \qquad \blacksquare$$

Proof of Theorem 5 For simplicity we consider the case $K = 2$ and assume $\mu^* = \mu_1 > \mu_2$. We can prove the theorem for the case $K > 2$ in the same way (see Remark 2 below this proof).

First we define three constants independent of n by

$$\xi \equiv \frac{1}{2\log\frac{1-\mu_2}{1-\mu_1}} > 0 \tag{39}$$

$$\rho \equiv \frac{D_{\inf}(F_2, \mu_1)}{3} > 0$$

$$\mu' \equiv \max\left\{\mu_1 - \rho(1-\mu_1),\ \frac{\mu_1+\mu_2}{2}\right\} \in (\mu_2, \mu_1) \,. \tag{40}$$

We also define the following six events for sufficiently small $\delta > 0$

$$\begin{aligned}
A_l &\equiv \{J(l) = 2,\ T_2(l) \ge \xi \log n\}\,, \\
B_l^{(1)} &\equiv \{\hat{\mu}^*(l) \le \mu'\}\,, \\
B_l^{(2)} &\equiv \{\mu' < \hat{\mu}^*(l) \le \mu_1 - \delta\}\,, \\
B_l^{(3)} &\equiv \{\mu_1 - \delta < \hat{\mu}^*(l)\}\,, \\
C_l &\equiv \{\hat{\mu}_2(l) \le \mu'\}\,, \\
D_l &\equiv \left\{D_{\inf}(\hat{F}_2(l), \mu_1) \ge D_{\inf}(F_2, \mu_1) - \rho\right\}\,.
\end{aligned}$$

Since the whole sample space is covered by

$$C_l^c \cup D_l^c \cup B_l^{(1)} \cup (B_l^{(2)} \cap C_l \cap D_l) \cup (B_l^{(3)} \cap C_l)\,,$$

we have

$$\begin{aligned}
T_2(n) &= \sum_{l=1}^n \mathbb{1}[J(l) = 2] \\
&\le \xi \log n + \sum_{l=1}^n \mathbb{1}[A_l] \\
&\le \sum_{l=1}^n \mathbb{1}[A_l \cap C_l^c] + \sum_{l=1}^n \mathbb{1}[A_l \cap D_l^c] + \sum_{l=1}^n \mathbb{1}\left[A_l \cap B_l^{(1)}\right] \\
&\quad + \sum_{l=1}^n \mathbb{1}\left[A_l \cap B_l^{(2)} \cap C_l \cap D_l\right] + \left(\xi \log n + \sum_{l=1}^n \mathbb{1}\left[A_l \cap B_l^{(3)} \cap C_l\right]\right)\,. \qquad (41)
\end{aligned}$$

We bound expectations of these terms in the followings. The essential point is that the only events involving $B_l^{(2)}$ and $B_l^{(3)}$ depend on the small constant δ and the number of rounds of the other events can be bounded independently of δ. We can derive a tight bound for events $B_l^{(2)}$ and $B_l^{(3)}$ with respect to δ by considering these events under C_l and D_l, that is, under the condition that statistics $\hat{\mu}_2(l)$ and $D_{\inf}(\hat{F}_2(l), \mu_1)$ are not very far from the true expectation.

First we have[3]

$$\sum_{l=1}^n \mathbb{1}[A_l \cap C_l^c] \le \sum_{t=\xi \log n}^{\infty} \mathbb{1}\left[\bigcup_{l=1}^n \{J(l) = 2,\ \hat{\mu}_{2,t} > \mu',\ T_2(l) = t\}\right] \qquad (42)$$

[3] The summation $\sum_{t=\xi \log n}^{\infty}$ in (42) should be $\sum_{t=\lceil \xi \log n \rceil}^{\infty}$ to be precise. However we omit the rounding operations $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ in the proof of this theorem for simplicity since these do not affect the asymptotic analysis.

and therefore

$$
\begin{aligned}
\mathrm{E}\left[\sum_{l=1}^{n} \mathbb{1}[A_l \cap C_l^c]\right] &\le \sum_{t=\xi \log n}^{\infty} P_{F_2}[\hat{\mu}_{2,t} > \mu'] \\
&\le \sum_{t=\xi \log n}^{\infty} \mathrm{e}^{-t\Lambda_2^*(\mu')} \qquad \text{(by (18) of Prop. 9)} \\
&= \frac{\mathrm{e}^{-(\xi \log n)\Lambda_2^*(\mu')}}{1-\mathrm{e}^{-\Lambda_2^*(\mu')}} \\
&= \mathrm{O}(\mathrm{e}^{-\mathrm{O}(\log n)}) \\
&= \mathrm{o}(1)\,. \qquad (43)
\end{aligned}
$$

Second, we have

$$
\sum_{l=1}^{n} \mathbb{1}[A_l \cap D_l^c] \le \sum_{t=\xi \log n}^{\infty} \mathbb{1}\left[\bigcup_{l=1}^{n} \Big\{J(l)=2,\ D_{\inf}(\hat{F}_{2,t},\mu_1) < D_{\inf}(F_2,\mu_1)-\rho,\ T_2(l)=t\Big\}\right].
$$

From Prop. 10, its expectation is bounded as

$$
\begin{aligned}
\mathrm{E}\left[\sum_{l=\xi \log n}^{n} \mathbb{1}[A_l \cap D_l^c]\right] &\le \sum_{t=\xi \log n}^{\infty} \mathrm{e}^{-t\tilde{\Lambda}_2^*(D_{\inf}(F_2,\mu_1)-\rho)} \\
&= \frac{\mathrm{e}^{-(\xi \log n)\tilde{\Lambda}_2^*(D_{\inf}(F_2,\mu_1)-\rho)}}{1-\mathrm{e}^{-\tilde{\Lambda}_2^*(D_{\inf}(F_2,\mu_1)-\rho)}} \\
&= \mathrm{o}(1)\,. \qquad (44)
\end{aligned}
$$

Third, we have

$$
\mathrm{E}\left[\sum_{l=\xi \log n}^{n} \mathbb{1}\Big[A_l \cap B_l^{(1)}\Big]\right] = \mathrm{O}(1) \qquad (45)
$$

from Lemma 14 with $\mu := \mu'$ since μ' is a constant independent of δ and n.

Fourth, we have

$$
\begin{aligned}
&\sum_{l=1}^{n} \mathbb{1}\Big[A_l \cap B_l^{(2)} \cap C_l \cap D_l\Big] \\
&\le \sum_{t_2=\xi \log n}^{\infty} \sum_{t_1=1}^{\infty} \mathbb{1}\left[\bigcup_{l=1}^{n} \{J(l)=2,\ T_1(l)=t_1,\ T_2(l)=t_2,\ B_l^{(2)} \cap C_l \cap D_l\}\right].
\end{aligned}
$$

Note that $\{T_2(l)=t_2,\ B_l^{(2)} \cap D_l\}$ implies

$$
\begin{aligned}
I_2(l) &\ge t_2 D_{\inf}(\hat{F}_2(l),\mu') \\
&\ge t_2\Big(D_{\inf}(\hat{F}_2(l),\mu_1)-\rho\Big) \qquad \text{(by (40) and Lemma 7)} \\
&\ge t_2\left(D_{\inf}(F_2,\mu_1)-2\rho\right) \qquad \text{(by definition of } D_l\text{)} \\
&= t_2\rho\,.
\end{aligned}
$$

Furthermore, $J(l)=2$ implies $I_2(l)\le I_1(l)$ and $\{T_1(l)=t_1,\ B_l^{(2)}\cap C_l\}$ implies $I_1(l)=\log t_1$. Combining them, we have

$$\begin{aligned}\sum_{l=1}^{n}\mathbb{1}\left[A_l\cap B_l^{(2)}\cap C_l\cap D_l\right] &\le \sum_{t_2=\xi\log n}^{\infty}\sum_{t_1=1}^{\infty}\mathbb{1}[\rho t_2\le\log t_1,\ \hat{\mu}_{1,t_1}\le\mu_1-\delta]\\ &=\sum_{t_2=\xi\log n}^{\infty}\sum_{t_1=\mathrm{e}^{\rho t_2}}^{\infty}\mathbb{1}[\hat{\mu}_{1,t_1}\le\mu_1-\delta] \qquad (46)\end{aligned}$$

and therefore

$$\begin{aligned}\mathrm{E}\left[\sum_{l=1}^{n}\mathbb{1}\left[A_l\cap B_l^{(2)}\cap C_l\cap D_l\right]\right] &\le \sum_{t_2=\xi\log n}^{\infty}\sum_{t_1=\mathrm{e}^{\rho t_2}}^{\infty}P_{F_1}\left[\hat{\mu}_{1,t_1}\le\mu_1-\delta\right]\\ &\le \sum_{t_2=\xi\log n}^{\infty}\frac{\mathrm{e}^{-\mathrm{e}^{\rho t_2}\Lambda_1^*(\mu_1-\delta)}}{1-\mathrm{e}^{-\Lambda_1^*(\mu_1-\delta)}} \qquad \text{(by (17) of Prop. 9)}\\ &\le \sum_{t_2=\xi\log n}^{\infty}\frac{\mathrm{e}^{-(\frac{(\rho t_2)^3}{3}+\rho t_2)\Lambda_1^*(\mu_1-\delta)}}{1-\mathrm{e}^{-\Lambda_1^*(\mu_1-\delta)}}\\ &\qquad\qquad \left(\text{by } \mathrm{e}^x\ge\frac{x^3}{3}+x \text{ for } x\ge 0\right)\\ &\le \sum_{t_2=\xi\log n}^{\infty}\frac{\mathrm{e}^{-(\frac{(\rho\xi\log n)^3}{3}+\rho t_2)\Lambda_1^*(\mu_1-\delta)}}{1-\mathrm{e}^{-\Lambda_1^*(\mu_1-\delta)}}\\ &=\frac{\mathrm{e}^{-(\frac{(\rho\xi\log n)^3}{3}+\rho\xi\log n)\Lambda_1^*(\mu_1-\delta)}}{(1-\mathrm{e}^{-\Lambda_1^*(\mu_1-\delta)})(1-\mathrm{e}^{-\rho\Lambda_1^*(\mu_1-\delta)})}\\ &=\frac{\mathrm{e}^{-\mathrm{O}(\delta^2(\log n)^3)}}{\mathrm{O}(\delta^4)}\,. \qquad (47)\end{aligned}$$

Finally we evaluate two terms

$$\xi\log n+\sum_{l=1}^{n}\mathbb{1}\left[A_l\cap B_l^{(3)}\cap C_l\right]=\xi\log n+\sum_{t=\xi\log n}^{n}\mathbb{1}\left[\bigcup_{l=1}^{n}\{J(l)=2,\ T_2(l)=t,\ B_l^{(3)}\cap C_l\}\right]$$

in (41). Here note that $\{T_2(l)=t\ge\xi\log n,\ B_l^{(3)}\}$ implies

$$\begin{aligned}I_2(l)&\ge tD_{\mathrm{inf}}(\hat{F}_2,\mu_1-\delta)+\log t\\ &\ge t\left(D_{\mathrm{inf}}(\hat{F}_2,\mu_1)-\frac{\delta}{1-\mu_1}\right)+\log(\xi\log n) \qquad \text{(by Lemma 7)}\end{aligned}$$

and $\{J(l)=2,\ B_l^{(3)}\cap C_l\}$ implies $I_2(l)\le I_1(l)=\log T_1(l)\le\log n$. As a result, we have

$$\xi\log n+\sum_{l=1}^{n}\mathbb{1}\left[A_l\cap B_l^{(3)}\cap C_l\right]$$

$$
\begin{aligned}
&\le \xi\log n + \sum_{t=\xi\log n}^{\infty} \mathbb{1}\left[t\left(D_{\mathrm{inf}}(\hat{F}_2,\mu_1)-\frac{\delta}{1-\mu_1}\right)\le \log n-\log(\xi\log n)\right]\\
&= \sum_{t=1}^{\infty} \mathbb{1}\left[t\left(D_{\mathrm{inf}}(\hat{F}_{2,t},\mu_1)-\frac{\delta}{1-\mu_1}\right)\le \log n-\log(\xi\log n)\right]\\
&\quad + \sum_{t=1}^{\xi\log n} \mathbb{1}\left[t\left(D_{\mathrm{inf}}(\hat{F}_{2,t},\mu_1)-\frac{\delta}{1-\mu_1}\right)> \log n-\log(\xi\log n)\right]. \qquad (48)
\end{aligned}
$$

The expectation of the second term of (48) can be evaluated as

$$
\begin{aligned}
&\mathrm{E}\left[\sum_{t=1}^{\xi\log n} \mathbb{1}\left[t\left(D_{\mathrm{inf}}(\hat{F}_{2,t},\mu_1)-\frac{\delta}{1-\mu_1}\right)> \log n-\log(\xi\log n)\right]\right]\\
&\quad\le \sum_{t=1}^{\xi\log n} P_{F_2}\left[D_{\mathrm{inf}}(\hat{F}_{2,t},\mu_1)>\frac{\log n-\log(\xi\log n)}{\xi\log n}\right]\\
&\quad= \sum_{t=1}^{\xi\log n} P_{F_2}\left[D_{\mathrm{inf}}(\hat{F}_{2,t},\mu_1)>\frac{1}{\xi}-\mathrm{o}(1)\right]\\
&\quad= \sum_{t=1}^{\xi\log n} P_{F_2}\left[D_{\mathrm{inf}}(\hat{F}_{2,t},\mu_1)>2\log\frac{1-\mu_2}{1-\mu_1}-\mathrm{o}(1)\right] \qquad \text{(by (39))}\\
&\quad= \mathrm{O}(1). \qquad \text{(by Prop. 12)} \qquad (49)
\end{aligned}
$$

Putting (41) and (43)–(49) together, we have

$$
\begin{aligned}
\mathrm{E}[T_2(n)] \le{}& \mathrm{E}\left[\sum_{t=1}^{\infty} \mathbb{1}\left[t\left(D_{\mathrm{inf}}(\hat{F}_2,\mu_1)-\frac{\delta}{1-\mu_1}\right)\le \log n-\log(\xi\log n)\right]\right]\\
&+\frac{\mathrm{e}^{-\mathrm{O}(\delta^2(\log n)^3)}}{\mathrm{O}(\delta^4)}+\mathrm{O}(1). \qquad (50)
\end{aligned}
$$

Let $Y_t=\log(1-(X_{2,t}-\mu_1)\nu_2^*)-\delta/(1-\mu_1)$ and define a stochastic process $\{S_t\}_{t=1,2,\cdots}$ by $S_t=\sum_{l=1}^{t} Y_l$. For a stopping time $\tau=\min\{t : S_t>\log n-\log(\xi\log n)\}$, the first term of (50) is bounded by

$$
\begin{aligned}
&\mathrm{E}\left[\sum_{t=1}^{\infty} \mathbb{1}\left[t\left(D_{\mathrm{inf}}(\hat{F}_{2,t},\mu_1)-\frac{\delta}{1-\mu_1}\right)\le \log n-\log(\xi\log n)\right]\right]\\
&\quad\le \mathrm{E}\left[\sum_{t=1}^{\infty}\mathbb{1}[S_t\le \log n-\log(\xi\log n)]\right]\\
&\quad= \mathrm{E}\left[(\tau-1)+\sum_{m=\tau+1}^{n}\mathbb{1}\left[S_\tau+\sum_{l=\tau+1}^{m} Y_l\le \log n-\log(\xi\log n)\right]\right]\\
&\quad\le \mathrm{E}[\tau]+\mathrm{E}\left[\sum_{m=\tau+1}^{n}\mathbb{1}\left[\sum_{l=\tau+1}^{m} Y_l\le 0\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathrm{E}[\tau] + \mathrm{E}\left[\mathrm{E}\left[\sum_{m=\tau+1}^{n} \mathbb{1}\left[\sum_{l=\tau+1}^{m} Y_l \le 0\right] \middle| \tau\right]\right] \\
&= \mathrm{E}[\tau] + \mathrm{E}\left[\sum_{m=\tau+1}^{n} P_{F_2}\left[\sum_{l=\tau+1}^{m} Y_l \le 0 \middle| \tau\right]\right]. \qquad (51)
\end{aligned}
$$

Note that $\mathrm{E}[Y_t] = D_{\mathrm{inf}}(F_2,\mu_1) - \delta/(1-\mu_1)$ and $\mathrm{E}[\mathrm{e}^{Y_t}] = \mathrm{e}^{-\delta/(1-\mu_1)}(1-(\mu_2-\mu_1)\nu_i^*) < \infty$. Then we obtain from (i) of Lemma 18 that

$$
\begin{aligned}
\mathrm{E}[\tau] &\le \frac{\log n - \log(\xi \log n) + \log(\log n - \log(\xi \log n))}{D_{\mathrm{inf}}(F_2,\mu_1) - \frac{\delta}{1-\mu_1}} + \mathrm{O}(1) \\
&= \frac{\log n}{D_{\mathrm{inf}}(F_2,\mu_1) - \frac{\delta}{1-\mu_1}} + \mathrm{O}(1) \\
&= \frac{\log n}{D_{\mathrm{inf}}(F_2,\mu_1)} + \mathrm{O}(\delta \log n) + \mathrm{O}(1). \qquad (52)
\end{aligned}
$$

On the other hand, from Cramér's theorem we obtain

$$
\begin{aligned}
&\mathrm{E}\left[\sum_{m=\tau+1}^{n} P_{F_2}\left[\sum_{l=\tau+1}^{m} Y_l \le 0 \middle| \tau\right]\right] \\
&= \mathrm{E}\left[\sum_{m=\tau+1}^{n} P_{F_2}\left[\frac{1}{m-\tau}\sum_{l=\tau+1}^{m} \log(1-(X_{2,l}-\mu_1)\nu_2^*) \le \frac{\delta}{1-\mu_1} \middle| \tau\right]\right] \\
&\le \mathrm{E}\left[\sum_{m=\tau+1}^{n} \mathrm{e}^{-(m-\tau)\tilde{\Lambda}_2^*(\frac{\delta}{1-\mu_1})}\right] \qquad \text{(by Prop. 9 and definition of } \tilde{\Lambda}_2^* \text{ in (6))} \\
&\le \frac{1}{1-\mathrm{e}^{-\tilde{\Lambda}_2^*(\frac{\delta}{1-\mu_1})}} \\
&= \mathrm{O}(1). \qquad \text{(by Lemma 15)} \qquad (53)
\end{aligned}
$$

By combining (51)–(53) with (50) we have

$$
\mathrm{E}[T_2(n)] \le \frac{\log n}{D_{\mathrm{inf}}(F_2,\mu_1)} + \mathrm{O}(\delta \log n) + \frac{\mathrm{e}^{-\mathrm{O}(\delta^2(\log n)^3)}}{\mathrm{O}(\delta^4)} + \mathrm{O}(1).
$$

We obtain (i) of Theorem 5 by letting $\delta = \mathrm{O}((\log n)^{-1})$.

In the case that each arm has a bounded support we can apply (ii) of Lemma 18. As a result, (52) is replaced with

$$
\begin{aligned}
\mathrm{E}[\tau] &\le \frac{\log n - \log(\xi \log n)}{D_{\mathrm{inf}}(F_2,\mu_1) - \frac{\delta}{1-\mu_1}} + \mathrm{O}(1) \\
&= \frac{\log n}{D_{\mathrm{inf}}(F_2,\mu_1)} + \mathrm{O}(\delta \log n) - \mathrm{O}(\log\log n)
\end{aligned}
$$

and we obtain (ii) of Theorem 5 by this replacement. ■

Remark 2 The proof for $K > 2$ is almost the same as the case $K = 2$. The only different point is the evaluation around (46), wherein the pair $(T_1(l), T_2(l))$ is considered. For $K > 3$ we can proceed the evaluation in the same way by taking the summation over contributions of all pairs $(T_j(l), T_i(l))$, $j \in \mathcal{I}_{\mathrm{opt}}, i \neq j$.

References

Rajeev Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33(6):1926–1951, 1995.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *Proceedings of AISTATS 2010*, volume 31, pages 99–107, 2013.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410: 1876–1902, April 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Jonathan M. Borwein and Adrian S. Lewis. Partially-finite programming in L_1 and the existence of maximum entropy estimates. *SIAM Journal on Optimization*, 3(2):248–267, 1993.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *arXiv*, 2012. URL `http://arxiv.org/abs/1209.1727`.

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

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

Oliver Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Proceedings of NIPS 2011*, volume 24, pages 1252–1260, Granada, Spain, 2012.

Yuan S. Chow and Tze L. Lai. Some one-sided theorems on the tail distribution of sample sums with applications to the last time and largest excess of boundary crossings. *Transactions of the American Mathematical Society*, 208:51–72, 1975.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Applications of Mathematics*. Springer-Verlag, New York, second edition, 1998.

Anthony V. Fiacco. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. Academic Press, New York, 1983.

Aurelien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of COLT 2011*, Budapest, Hungary, 2011.

John C. Gittins. *Multi-armed Bandit Allocation Indices*. Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons, Chichester, 1989.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of COLT 2010*, pages 67–79, Haifa, Israel, 2010.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.

Junya Honda and Akimichi Takemura. Stochastic bandit based on empirical moments. In *Proceedings of AISTATS 2012*, pages 529–537, Canary Islands, Spain, 2012.

S Ito, Y Liu, and K. L. Teo. A dual parametrization method for convex semi-infinite programming. *Annals of Operations Research*, 98(1-4):189–213, 2000.

Samuel Karlin and William J. Studden. *Tchebycheff Systems, with Applications in Analysis and Statistics*. Interscience Publishers New York, 1966.

Emilie Kaufmann. *Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources*. PhD thesis, TELECOM ParisTech, 2014.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *Proceedings of AISTATS 2012*, pages 592–600, 2012a.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: an asymptotically optimal finite-time analysis. In *Proceedings of ALT 2012*, pages 199–213, Berlin, Heidelberg, 2012b. Springer-Verlag.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Proceedings of NIPS 2013*, Lake Tahoe, NV, USA, 2013.

Balachander Krishnamurthy, Craig Wills, and Yin Zhang. On the use and performance of content distribution networks. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, pages 169–182, New York, USA, 2001.

Tze L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Keqin Liu and Qing Zhao. Multi-armed bandit problems with heavy-tailed reward distributions. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 485–492. IEEE, 2011.

Sean P. Meyn and R. L. Tweedie. Stability of Markovian processes I: Criteria for discrete-time chains. *Advances in Applied Probability*, 24:542–574, 1992.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–35, 1952.

Daniel Russo and Benjamin V. Roy. Learning to optimize via posterior sampling. *arXiv*, 2013. URL `http://arxiv.org/abs/1301.2609`.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

Joannès Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *Proceedings of ECML 2005*, pages 437–448, Porto, Portugal, 2005. Springer.