An Asymptotically Optimal Policy for Uniform Bandits of Unknown Support

Wesley Cowan  
Department of Mathematics  
Rutgers University  
110 Frelinghuysen Rd., Piscataway, NJ 08854, USA  
CWCOWAN@MATH.RUTGERS.EDU

Michael N. Katehakis  
Department of Management Science and Information Systems  
Rutgers University  
100 Rockafeller Rd., Piscataway, NJ 08854, USA  
MNK@RUTGERS.EDU

Abstract

Consider the problem of a controller sampling sequentially from a finite number of \( N \geq 2 \) populations, specified by random variables \( X^i_k, i = 1, \ldots, N, \) and \( k = 1, 2, \ldots \); where \( X^i_k \) denotes the outcome from population \( i \) the \( k \)th time it is sampled. It is assumed that for each fixed \( i \), \( \{X^i_k\}_{k \geq 1} \) is a sequence of i.i.d. uniform random variables over some interval \( [a_i, b_i] \), with the support (i.e., \( a_i, b_i \)) unknown to the controller. The objective is to have a policy \( \pi \) for deciding, based on available data, from which of the \( N \) populations to sample from at any time \( n = 1, 2, \ldots \) so as to maximize the expected sum of outcomes of \( n \) samples or equivalently to minimize the regret due to lack of information of the parameters \( \{a_i\} \) and \( \{b_i\} \). In this paper, we present a simple inflated sample mean (ISM) type policy that is asymptotically optimal in the sense of: its regret achieving the asymptotic lower bound of Burnetas and Katehakis (1996b). Additionally, finite horizon regret bounds are given.

Keywords: Inflated Sample Means, Upper Confidence Bound, Multi-armed Bandits, Sequential Allocation

1. Introduction

Let \( \mathcal{F} \) be a known family of probability densities on \( \mathbb{R} \), each with finite mean. We define \( \mu(f) \) to be the expected value under density \( f \), and \( \text{Sp}(f) \) to be the support of \( f \). Consider the problem of sequentially sampling from a finite number of \( N \geq 2 \) populations or ‘bandits’, where measurements from population \( i \) are specified by an i.i.d. sequence of random variables \( \{X^i_k\}_{k \geq 1} \) with density \( f_i \in \mathcal{F} \). We take each \( f_i \) as unknown to the controller. It is convenient to define, for each \( i \), \( \mu_i = \mu(f_i) = \int_{\text{Sp}(f_i)} xf(x)dx \), and \( \mu^* = \mu^*(\{f_i\}) = \max_i \mu(f_i) \). Additionally, we take \( \Delta_i = \mu^* - \mu_i \geq 0 \), the discrepancy of bandit \( i \).

We note, but for simplicity will not consider explicitly, that both discrete and continuous distributions can be studied when one takes \( \{X^i_k\}_{k \geq 1} \) to be i.i.d. with density \( f_i \), with respect to some known measure \( \nu_i \).

For any adaptive, non-anticipatory policy \( \pi \), \( \pi(t) = i \) indicates that the controller samples bandit \( i \) at time \( t \). Define \( T^i_n = \sum_{t=1}^{n} 1\{\pi(t) = i\} \), denoting the number of times bandit \( i \) has been sampled during the periods \( t = 1, \ldots, n \) under policy \( \pi \); we take, as a convenience, \( T^i_n(0) = 0 \) for all \( i, \pi \). The value of a policy \( \pi \) is the expected sum of the first \( n \) outcomes under \( \pi \), which we define

©Wesley Cowan, and Michael N. Katehakis.
to be the function $V_\pi(n)$:

$$V_\pi(n) = E \left[ \sum_{i=1}^{N} \sum_{k=1}^{T_i^*} X_k^i \right] = \sum_{i=1}^{N} \mu_i E \left[ T_i^* \right],$$

(1)

where for simplicity the dependence of $V_\pi(n)$ on the unknown densities $\{f_i\}$ is suppressed. The regret of a policy is taken to be the expected loss due to ignorance of the underlying distributions by the controller. Had the controller complete information, she would at every round activate some bandit $i^*$ such that $\mu_{i^*} = \mu^* = \max_i \mu_i$. For a given policy $\pi$, we define the expected regret of that policy at time $n$ as

$$R_\pi(n) = n \mu^* - V_\pi(n) = \sum_{i=1}^{n} \Delta_i E \left[ T_i^* \right].$$

(2)

We are interested in policies for which $V_\pi(n)$ grows as fast as possible with $n$, or equivalently that $R_\pi(n)$ grows as slowly as possible with $n$.

2. Preliminaries - Background

We restrict $\mathcal{F}$ in the following way:

**Assumption 1.** Given any set of bandit densities $\{f_i\}_{i=1}^{N}$, for any sub-optimal bandit $i$, i.e., $\mu(f_i) \neq \mu^*(\{f_i\})$, there exists some $\tilde{f}_i \in \mathcal{F}$ such that $\text{Sp}(\tilde{f}_i) \supset \text{Sp}(f_i)$, and $\mu(\tilde{f}_i) > \mu^*(\{f_i\})$.

Effectively, this ensures that at any finite time, given a set of bandits under consideration, for any bandit there is a density in $\mathcal{F}$ that would both potentially explain the measurements from that bandit, and make it the unique optimal bandit of the set. Hence the optimal bandit almost surely cannot be identified in finite time.

The focus of this paper is on $\mathcal{F}$ as the set of uniform densities over some unknown support. Let $I(f, g)$ denote the Kullback-Liebler divergence of density $f$ from $g$,

$$I(f, g) = \int_{\text{Sp}(f)} \ln \left( \frac{f(x)}{g(x)} \right) f(x) dx = E_f \left[ \ln \left( \frac{f(X)}{g(X)} \right) \right].$$

(3)

It is a simple generalization of a classical result (part 1 of Theorem 1) of Burnetas and Katehakis [1996b] that if a policy $\pi$ is uniformly fast (UF), i.e., $R_\pi(n) = o(n^\alpha)$ for all $\alpha > 0$ and for any choice of $\{f_i\} \subset \mathcal{F}$, then, the following bound holds:

$$\liminf_n \frac{R_\pi(n)}{\ln n} \geq M_{\text{BK}}(\{f_i\}), \text{ for all } \{f_i\} \subset \mathcal{F},$$

(4)

where the bound $M_{\text{BK}}(\{f_i\})$ itself is determined by the specific distributions of the populations:

$$M_{\text{BK}}(\{f_i\}) = \sum_{i: \mu_i \neq \mu^*} \Delta_i \inf_{g \in \mathcal{F}} \left\{ I(f_i, g) : \mu(g) \geq \mu^* \right\}.$$  

(5)

For a given set of densities $\mathcal{F}$, it is of interest to construct policies $\pi$ such that

$$\lim_n \frac{R_\pi(n)}{\ln n} = M_{\text{BK}}(\{f_i\}), \text{ for all } \{f_i\} \subset \mathcal{F}.$$
Such policies achieve the slowest (maximum) regret (value) growth rate possible among UF policies. They have been called UM or asymptotically optimal or efficient, cf. Burnetas and Katehakis (1996b).

For a given \( f \in \mathcal{F} \), let \( \hat{f}_k \in \mathcal{F} \) be an estimator of \( f \) based on the first \( k \) samples from \( f \). Burnetas and Katehakis (1996b) showed that that under sufficient conditions on \( \{ \hat{f}_k \} \), asymptotically optimal (‘UM’) policies could be constructed by initially sampling each bandit some number of \( n_0 \) times, and then for \( n > N * n_0 \), following the index policy:

\[
\pi^0(n+1) = \arg \max_i \{ u^i(n, T^a_\pi(n)) \},
\]

where the indices \( u^i(n, t) \) are ‘inflations of the current estimates for the means’ (ISM), were specified as:

\[
u^i(n, t) = u^i_{\text{BK}}(n, t, \hat{f}_i^t) = \sup_{g \in \mathcal{F}} \left\{ \mu(g) : I(\hat{f}_i^t, g) < \frac{\ln n}{t} \right\}. \tag{7}
\]

The sufficient conditions on the estimators \( \{ \hat{f}_k \} \) are as follows:

Defining

\[
J(f, c) = \inf_{g \in \mathcal{F}} \{ I(f, g) : \mu(g) > c \},
\]

for all choices of \( \{ f_i \} \subset \mathcal{F} \) and all \( c > 0 \), the following hold for each \( i \), as \( k \to \infty \).

1. \( C1: P \left( J(\hat{f}_k, \mu^* - \varepsilon) < J(f_i, \mu^* - \varepsilon) - \delta \right) = o(1/k). \)
2. \( C2: P \left( u^i_{\text{BK}}(k, j, \hat{f}_i^j) \leq \mu_i - \varepsilon \right) \text{ for each } j \in \{ n_0, \ldots, k \} = o(1/k). \)

These conditions correspond to Conditions A1-A3 given in Burnetas and Katehakis (1996b). However under the stated Assumption 1 on \( \mathcal{F} \) given here, Condition A1 therein is automatically satisfied. Conditions A2 (see also Remark 4(b) in Burnetas and Katehakis (1996b)) and A3 are as given in C1 and C2, respectively. Note, Condition (C1) is essentially satisfied as long as \( \hat{f}_i^t \) converges to \( f_i \) (and hence \( J(\hat{f}_k, \mu^* - \varepsilon) \to J(f_i, \mu^* - \varepsilon) \)) sufficiently quickly with \( k \). This can often be verified easily with standard large deviation principles. The difficulty in proving the optimality of policy \( \pi^0 \) is often in verifying that Condition (C2) holds.

**Remark 1** The above discussion is a parameter-free variation of that in Burnetas and Katehakis (1996b), where \( \mathcal{F} \) was taken to be parametrizable, i.e., \( \mathcal{F} = \{ f_0 : \theta \in \Theta \} \), taking \( \theta \) as a vector of parameters in some parameter space \( \Theta \). Further, Burnetas and Katehakis (1996b) considered potentially different parameter spaces (and therefore potentially different parametric forms) for each bandit \( i \). There, Conditions A1-A3 (hence C1, C2 herein) and the corresponding indices were stated in terms of estimates for the bandit parameters, \( \hat{\theta}_i^t(t) \) an estimate of the parameters \( \theta_i^* \) of bandit \( i \), given \( t \) samples. In particular, Eq. (7) appears essentially as

\[
u^i(n, t) = u^i_{\text{BK}}(n, t, \hat{\theta}_i^t(t)) = \sup_{g \in \Theta} \left\{ \mu(g) : I(\hat{f}_i^t(t), g) < \frac{\ln n}{t} \right\} \tag{8}
\]

Early, fundamental work in this area includes Thompson (1933), Robbins (1952), and Gittins (1979), Weiber (1992). The problem of asymptotically minimizing the increase rate of the regret among uniformly fast (UF) policies \( \pi \) (i.e., \( R_\pi(n) = o(n^\alpha) \) for all \( \alpha > 0 \) ) was first considered in Lai and Robbins (1985), who for single parameter models derived the simplified version of the theoretical lower of Eq. (6). \( M_{LR}(\{ \theta_i \}) = \sum_{i: \mu_i(\theta_i) \neq \mu^*} \Delta_i / I(\theta_i, \mu^*) \), where \( \theta^* \) is any of the \( \theta_i \) such that
\( \mu(\theta^*) = \mu^* = \max_i \mu(\theta_i) \). In addition, Lai and Robbins (1983), constructed policies that achieve the asymptotic lower bound for some exponential families including normal distributions with known variances, Bernoulli, Poisson, and the Laplace distribution (which does not belong to the exponential families). Further, for the Bernoulli distribution with unknown success probabilities \( \theta_i \), following Agrawal and Goyal (2011), it was recently established that Thomson sampling archives the lower bound \( M_{LR} \) by Kaufmann et al. (2012), Korda et al. (2013).

For multi-parameter distributions, asymptotically optimal policies have been developed for an arbitrary discrete distributions of known support in Burnetas and Katehakis (1996b). For the same problem Honda and Takemura (2011) and Honda and Takemura (2010) derived optimal policies, cyclic and randomized, that are simpler to implement than those considered in Burnetas and Katehakis (1996b) were constructed. The problem of constructing optimal policies for Normal distributions with unknown means and variances remained open, until recently when Honda and Takemura (2013) established that a form of Thompson sampling with certain priors on \( \mu, \sigma^2 \) achieves the asymptotic lower bound \( M_{BK} \). More recently in Cowan et al. (2015), asymptotically optimal policies of inflated sample mean structure ISM, were given for this problem.

For other work in this area we refer to Katehakis and Derman (1986), Katehakis and Veinott Jr (1987), Burnetas and Katehakis (1993, 1996a, 1996b), Lagoudakis and Parr (2003), Bartlett and Tewari (2009), Tekin and Liu (2012), Jouini et al. (2009), Dayanik et al. (2013), Filippi et al. (2010), and others. As well as Burnetas and Katehakis (2003), Audibert et al. (2009), Auer and Ortner (2010), Gittins et al. (2011), Bubeck and Slivkins (2012), Cappe et al. (2013), Kaufmann (2015), Li et al. (2014), Cowan and Katehakis (2015a, 2015b), and references therein. For dynamic programming extensions we refer to Burnetas and Katehakis (1997), Feinberg et al. (2003), Tewari and Bartlett (2008), Audibert et al. (2009), Littman (2012), and references therein.

### 3. The B-K Lower Bound and Sample Mean Inflation Factors

In this section we take \( \mathcal{F} \) as the set of probability densities on \( \mathbb{R} \) uniform over some finite interval, taking \( f \in \mathcal{F} \) as uniform over \([a_f, b_f] \). Note, as the family of densities is parametrizable, this largely falls under the scope of Burnetas and Katehakis (1996b). However, the results to follow seem to demonstrate a gap in that general treatment of the problem.

Note, some care with respect to support must be taken in applying Burnetas and Katehakis (1996b) to this case, to ensure that the integrals remain well defined. But for this \( \mathcal{F} \), we have that for a given \( f \in \mathcal{F} \), for any \( g \in \mathcal{F} \) such that \( \text{Sp}(f) \subset \text{Sp}(g) \), i.e., \( a_g \leq a_f \) and \( b_f \leq b_g \),

\[
I(f, g) = E_f \left[ \ln \left( \frac{f(X)}{g(X)} \right) \right] = \ln \left( \frac{b_g - a_g}{b_f - a_f} \right). \tag{9}
\]

If \( \text{Sp}(f) \) is not a subset of \( \text{Sp}(g) \), we take \( I(f, g) \) as infinite.

For notational convenience, given \( \{f_i\} \subset \mathcal{F}, \) for each \( i \), we take \( f_i \in \mathcal{F} \) as supported on some interval \([a_i, b_i] \). Note then, \( \mu_i = (a_i + b_i)/2 \).

Given \( t \) samples from bandit \( i \), \( \{X^i_{t'}\}_{t'=1}^t \), we take

\[
\hat{a}^i_t = \min_{t' \leq t} X^i_{t'}, \quad \hat{b}^i_t = \max_{t' \leq t} X^i_{t'}, \tag{10}
\]

the maximum-likelihood estimators of \( a_i \) and \( b_i \) respectively. We may then define \( \hat{f}^i_t \in \mathcal{F} \) as the uniform density over the interval \([\hat{a}^i_t, \hat{b}^i_t] \). Note, \( \hat{f}^i_t \) is the maximum-likelihood estimate of \( f_i \).
We can now state and prove the following.

**Lemma 2** Under Assumption 1 the following are true.

\[
M_{BK}(\{f_i\}) = \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i}{\ln \left(1 + \frac{2\Delta_i}{b_i - a_i}\right)}. \tag{11}
\]

\[
u_{BK}^i(n, t, \hat{f}_i^t) = \hat{a}_i^t + \frac{1}{2} \left(\hat{b}_i^t - \hat{a}_i^t\right) n^{1/t}. \tag{12}
\]

**Proof** Eq. (11) follows from Eq. (5) and the observation that in this case:

\[
\inf_{g \in F} \{I(f_i, g) : \mu(g) \geq \mu^*\} = \ln \left(\frac{2\mu^* - 2a_i}{b_i - a_i}\right) = \ln \left(1 + \frac{2\mu^* - 2\mu_i}{b_i - a_i}\right).
\]

For Eq. (12) we have:

\[
u_{BK}^i(n, t, \hat{f}_i^t) = \sup_{g \in F} \left\{\mu(g) : I(\hat{f}_i^t, g) < \frac{\ln n}{t}\right\}
= \sup_{a \leq \hat{a}_i^t, b \geq \hat{b}_i^t} \left\{\frac{a + b}{2} : \ln \left(\frac{b - a}{\hat{b}_i^t - \hat{a}_i^t}\right) < \frac{\ln n}{t}\right\}
= \frac{\hat{a}_i^t + 1}{2} \left(\hat{b}_i^t - \hat{a}_i^t\right) n^{1/t}. \tag{13}
\]

We are interested in policies \(\pi\) such that \(\lim_n R_\pi(n)/\ln n\) achieves the lower bound indicated above, for every choice of \(\{f_i\} \subset F\). Following the prescription of Burnetas and Katehakis (1996b), i.e. Eq. (12), would lead to the following policy,

**Policy BK-ISM :** \(\pi_{BK}\). At each \(n = 1, 2, \ldots\):

i) For \(n = 1, 2, \ldots, 2N\), sample each bandit twice, and

ii) for \(n \geq 2N\), let \(\pi_{BK}(n+1)\) be equal to:

\[
\arg \max_i \left\{\hat{a}_i^t T_{BK}(n) + \frac{1}{2} \left(\hat{b}_i^t T_{BK}(n) - \hat{a}_i^t\right) n^{\frac{1}{2}}\right\}, \tag{14}
\]

breaking ties arbitrarily.

It is easy to demonstrate that the estimators \(\hat{\theta}_i(t) = (\hat{a}_i^t, \hat{b}_i^t)\) converge sufficiently quickly to \((a_i, b_i)\) in probability that Condition (C1) above is satisfied for \(\hat{f}_i^t\). Proving that Condition (C2) is satisfied, however, is much more difficult, and in fact we conjecture that (C2) does not hold for policy \(\pi_{BK}\). While this does not indicate that that \(\pi_{BK}\) fails to achieve asymptotic optimality, it does imply that the standard techniques are insufficient to verify it. However, asymptotic optimality may provably be achieved by a (seemingly) negligible modification, via the following policy.
4. Asymptotically Optimal ISM Policy

We propose the following policy:

**Policy ISM-Uniform:** \( \pi_{CK} \). At each \( n = 1, 2, \ldots \):

i) For \( n = 1, 2, \ldots, 3N \) sample each bandit three times, and

ii) for \( n \geq 3N \), let \( \pi_{CK}(n+1) \) be equal to:

\[
\arg \max_i \left\{ \hat{a}_i^{T \pi_{CK}(n)} + \frac{1}{2} \left( \hat{b}_i^{T \pi_{CK}(n)} - \hat{a}_i^{T \pi_{CK}(n)} \right) n^{-\frac{i}{2}} \right\},
\]

breaking ties arbitrarily.

In the remainder of this paper, we verify the asymptotic optimality of \( \pi_{CK} \) (Theorem 4), and additionally give finite horizon bounds on the regret under this policy (Theorem 3, 5). Further, while Theorem 5 bounds the order of the remainder term as \( O((\ln n)^{3/4}) \), this is refined somewhat in Theorem 7 to \( o((\ln n)^{2/3+\beta}) \).

5. The Optimality Theorem and Finite Time Bounds

For the work in this section it is convenient to define the bandit spans, \( S_i = b_i - a_i \). We take \( S^* \) to be the minimal span of any optimal bandit, i.e.,

\[
S^* = \min_{i: \mu_i = \mu^*} S_i.
\]

Recall that \( \Delta_i = \mu^* - \mu_i = \max_j \left\{ \frac{a_j + b_j}{2} \right\} - \frac{a_i + b_i}{2} \). The primary result of this paper is the following.

**Theorem 3** For each sub-optimal \( i \) (i.e., \( \mu_i \neq \mu^* \)), let \( (\epsilon_i, \delta_i) \) be such that \( 0 < \epsilon_i < S^*, 0 < \delta_i < S_i \), and \( \epsilon_i + \delta_i < \Delta_i \). For \( \pi_{CK} \) as defined above, for all \( n \geq 3N \):

\[
R_{\pi_{CK}}(n) \leq \sum_{i: \mu_i \neq \mu^*} \Delta_i \ln \left( \frac{1 + \frac{2\Delta_i}{S_i}}{1 - \frac{\epsilon_i + \delta_i}{\Delta_i}} \right) \ln n
+ \sum_{i: \mu_i \neq \mu^*} \left( \frac{S_i}{\delta_i} + 3 \frac{S_i^3}{8 \epsilon_i^3} + 18 \right) \Delta_i.
\]

The proof of Theorem 3 is the central proof of this paper. We delay it briefly, to present two related results that can be derived from the above. The first is that \( \pi_{CK} \) is asymptotically optimal.

**Theorem 4** For \( \pi_{CK} \) as defined above, \( \pi_{CK} \) is asymptotically optimal in the sense that

\[
\lim_{n} \frac{R_{\pi_{CK}}(n)}{\ln n} = M_{BK}(\{f_i\}) = \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i}{\ln \left( 1 + \frac{2\Delta_i}{S_i} \right)}.
\]

**Proof** Fix the \( (\epsilon_i, \delta_i) \) as feasible in the hypotheses of Theorem 3. In that case, we have

\[
\lim_{n} \sup \frac{R_{\pi_{CK}}(n)}{\ln n} \leq \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i}{\ln \left( 1 + \frac{2\Delta_i}{S_i} \right) \left( 1 - \frac{\epsilon_i + \delta_i}{\Delta_i} \right)}.
\]
Taking the infimum as $\epsilon_i + \delta_i \to 0$ yields

$$\limsup_n \frac{R_{\pi_{\text{CK}}}(n)}{\ln n} \leq \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i}{\ln \left(1 + \frac{2\Delta_i}{S_i}\right)}.$$  (19)

This, combined with the previous observation about the $\liminf$ in Eq. (11) completes the result. \(\blacksquare\)

We next give an ‘$\epsilon$-free’ version of the previous bound, which demonstrates the remainder term on the regret under $\pi_{\text{CK}}$ is at worst $O((\ln n)^{3/4})$.

**Theorem 5** For each sub-optimal $i$ (i.e., $\mu_i \neq \mu^*$), let $G_i = \min\left(S_* - S_i, \frac{1}{2} \Delta_i\right)$. For all $n \geq 3N$,

$$R_{\pi_{\text{CK}}}(n) \leq \left(\sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i}{\ln \left(1 + \frac{2\Delta_i}{S_i}\right)}\right) (\ln n)$$

$$+ \sum_{i: \mu_i \neq \mu^*} \left(\frac{8G_i \Delta_i}{S_i + 2\Delta_i} \ln \left(1 + \frac{2\Delta_i}{S_i}\right)^2 + \frac{3S_i^3 \Delta_i}{8G_i^3} (\ln n)^{3/4}\right)$$

$$+ \sum_{i: \mu_i \neq \mu^*} \left(\frac{S_i \Delta_i}{G_i}\right) (\ln n)^{1/4} + 18 \sum_{i: \mu_i \neq \mu^*} \Delta_i.$$  (20)

**Proof** [Proof of Theorem 5] Let $0 < \epsilon < 1$, and for each $i$ let $\epsilon_i = \delta_i = G_i \epsilon$. Hence,

$$\ln \left(1 + \frac{2\Delta_i}{S_i} \left(1 - \frac{(\epsilon_i + \delta_i)}{\Delta_i}\right)\right) = \ln \left(1 + \frac{2\Delta_i}{S_i} \left(1 - \frac{2G_i}{\Delta_i}\right)\right).$$  (21)

Define

$$D_i = \frac{1}{\ln \left(1 + \frac{2\Delta_i}{S_i} \left(1 - \frac{2G_i}{\Delta_i}\right)\right)} - \frac{1}{\ln \left(1 + \frac{2\Delta_i}{S_i}\right)}.$$  (22)

Note the following bound, that

$$D_i \leq \left(\frac{2G_i \epsilon}{\Delta_i - 2G_i \epsilon}\right) \frac{2\Delta_i}{(S_i + 2\Delta_i) \ln \left(1 + \frac{2\Delta_i}{S_i}\right)^2}$$

$$\leq \left(\frac{2G_i \epsilon}{2\Delta_i}\right) \frac{2\Delta_i}{(S_i + 2\Delta_i) \ln \left(1 + \frac{2\Delta_i}{S_i}\right)^2}$$

$$= \frac{8G_i \epsilon}{(S_i + 2\Delta_i) \ln \left(1 + \frac{2\Delta_i}{S_i}\right)^2}.$$  (23)

This first inequality is proven separately as Proposition 9 in the Appendix. The second inequality is simply the observation that $2G_i \epsilon \leq 2G_i \leq \frac{1}{2} \Delta_i$. Applying this bound to Theorem 5 yields the
following bound,

\[
R_{\pi_{\text{CK}}}(n) \leq \left( \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i}{\ln \left( 1 + \frac{2\Delta_i}{S_i} \right)} \right) (\ln n) \\
+ 8 \left( \sum_{i: \mu_i \neq \mu^*} \frac{G_i \Delta_i}{(S_i + 2\Delta_i) \ln \left( 1 + \frac{2\Delta_i}{S_i} \right)^2} \right) \epsilon \ln n \\
+ \left( \sum_{i: \mu_i \neq \mu^*} \frac{S_i \Delta_i}{G_i} \right) \epsilon^{-1} + \frac{3}{8} S^3 \left( \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i}{G_i} \right) \epsilon^{-3} \\
+ 18 \left( \sum_{i: \mu_i \neq \mu^*} \Delta_i \right).
\]

(24)

Taking \(\epsilon = (\ln n)^{-1/4}\) completes the proof. \(\blacksquare\)

**Proof** [Proof of Theorem 1] For any \(i\) such that \(\mu_i \neq \mu^*\), recall that bandit \(i\) is taken to be uniformly distributed on the interval \([a_i, b_i]\). Let \((\epsilon_i, \delta_i)\) be as hypothesized. In this proof, we take \(\pi = \pi_{\text{CK}}\) as defined above. For any event \(A\) we let \(\overline{A}\) denote its complement. Recall that for each \(i\) we let \(\hat{b}_i^j = \max_{t \leq k} X_i^t\) and \(\hat{a}_i^j = \min_{t \leq k} X_i^t\).

We next define the following:

i) The index function \(u_i(k, j) = u_i(k, j, \hat{a}_i^j, \hat{b}_i^j)\):

\[
u_i(k, j) = \hat{a}_i^j + \frac{1}{2} (\hat{b}_i^j - \hat{a}_i^j) k^1/2.
\]

(25)

ii) The following events of interest, \(\mathcal{J}_i^t = \{u_i(t, T^i_\pi(t)) \geq \mu^* - \epsilon_i\}\) and \(\mathcal{K}_i^t = \{\hat{a}_i^t \leq a_i + \delta_i\}\).

iii) For \(n \geq 3N\), the following quantities

\[
n_1^i(n, \epsilon_i, \delta_i) = \sum_{t=3N}^{n} 1\{\pi(t + 1) = i, \mathcal{J}_i^t, \overline{\mathcal{K}_i^t(T^i_\pi(t))}\} \\
n_2^i(n, \epsilon_i, \delta_i) = \sum_{t=3N}^{n} 1\{\pi(t + 1) = i, \mathcal{J}_i^t, \overline{\mathcal{K}_i^t(T^i_\pi(t))}\} \\
n_3^i(n, \epsilon_i, \delta_i) = \sum_{t=3N}^{n} 1\{\pi(t + 1) = i, \mathcal{J}_i^t\}.
\]

(26)

For \(n \geq 3N\), we have the following relationship

\[
T^i_\pi(n + 1) = 3 + \sum_{t=3N}^{n} 1\{\pi(t + 1) = i\} = 3 + n_1^i(n, \epsilon_i, \delta_i) + n_2^i(n, \epsilon_i, \delta_i) + n_3^i(n, \epsilon_i, \delta_i).
\]

(27)

The proof proceeds by bounding, in expectation, each of the three terms.
Observe that, by the structure of the index function \( u_i \),

\[
1 \{ \pi(t + 1) = i, \mathcal{T}_{T_{\pi}^{-1}}^i, K_{T_{\pi}^{-1}(i)} \} \\
\leq 1 \left\{ \pi(t + 1) = i, a_i + \delta_i + \frac{1}{2} (b_i - a_i) T_{\pi}^{-1}(t) \geq \mu^* - \epsilon_i \right\} \\
= 1 \left\{ \pi(t + 1) = i, T_{\pi}^i(t) \leq \frac{\ln t}{2 \mu^* - 2 a_i - 2 \epsilon_i - 2 \delta_i} + 2 \right\} \\
\leq 1 \left\{ \pi(t + 1) = i, T_{\pi}^i(t) \leq \frac{\ln n}{2 \mu^* - 2 a_i - 2 \epsilon_i - 2 \delta_i} + 2 \right\}.
\]

(28)

Hence,

\[
n_i^1(n, \epsilon_i, \delta_i) \leq \\
\sum_{t=3N}^{n} 1 \left\{ \pi(t + 1) = i, T_{\pi}^i(t) \leq \frac{\ln n}{2 \mu^* - 2 a_i - 2 \epsilon_i - 2 \delta_i} + 2 \right\} \\
\leq \sum_{t=1}^{n} 1 \left\{ \pi(t + 1) = i, T_{\pi}^i(t) \leq \frac{\ln n}{2 \mu^* - 2 a_i - 2 \epsilon_i - 2 \delta_i} + 2 \right\} \\
\leq \frac{\ln n}{2 \mu^* - 2 a_i - 2 \epsilon_i - 2 \delta_i} + 2 + 2.
\]

(29)

The last inequality follows, observing that \( T_{\pi}^i(t) \) may be expressed as the sum of \( \pi(t) = i \) indicators, and seeing that the additional condition bounds the number of non-zero terms in the above sum. The additional +2 term simply accounts for the possibilities that \( \pi(1) = i \) and \( \pi(n + 1) = i \).

Note, this bound is sample-path-wise.

For the second term,

\[
n_i^2(n, \epsilon_i, \delta_i) \leq \sum_{t=3N}^{n} 1 \{ \pi(t + 1) = i, \mathcal{K}_{T_{\pi}^{-1}}^i \} \\
= \sum_{t=3N}^{n} \sum_{k=2}^{t} 1 \{ \pi(t + 1) = i, \mathcal{K}_{k}, T_{\pi}^i(t) = k \} \\
= \sum_{t=3N}^{n} \sum_{k=2}^{t} 1 \{ \pi(t + 1) = i, T_{\pi}^i(t) = k \} 1 \{ \mathcal{K}_{k} \} \\
\leq \sum_{k=2}^{n} 1 \{ \mathcal{K}_{k} \} \sum_{t=k}^{n} 1 \{ \pi(t + 1) = i, T_{\pi}^i(t) = k \} \\
\leq \sum_{k=2}^{n} 1 \{ \mathcal{K}_{k} \} \\
= \sum_{k=2}^{n} 1 \{ \hat{a}^i_k > a_i + \delta_i \}.
\]

(30)
The last inequality follows as, for fixed \( k \), \( \{\pi(t + 1) = i, T_{\pi}^i(t) = k\} \) may be true for at most one value of \( t \). It follows then that

\[
\mathbb{E} \left[ n_2^i(n, \epsilon_i, \delta_i) \right] \leq \sum_{k=2}^{n} \frac{\pi(t+1)}{k} \frac{\pi(t)}{k} \left( \frac{a_i + \delta_i}{b_i - a_i} \right)^k
\]

(31)

To bound the \( n_3^i \) term, observe that in the event \( \pi(t + 1) = i \), from the structure of the policy it must be true that \( u_i(t, T_{\pi}^i(t)) = \max_{j} u_j(t, T_{\pi}^j(t)) \). Thus, if \( i^* \) is some bandit such that \( \mu_i = \mu^* \), \( u_{i^*}(t, T_{\pi}^{i^*}(t)) \leq u_i(t, T_{\pi}^i(t)) \). In particular, we take \( i^* \) to be the optimal bandit realizing the minimal span \( b_{i^*} - a_{i^*} \). It follows,

\[
n_3^i(n, \epsilon_i, \delta_i) \leq \sum_{t=3N}^{n} \mathbb{P} \left\{ u_{i^*}(t, s) < \mu^* - \epsilon_i \text{ for some } 3 \leq s \leq t \right\}
\]

(32)

The last step follows as for \( t \) in this range, \( 3 \leq T_{\pi}^{i^*}(t) \leq t \). Hence

\[
\mathbb{E} \left[ n_3^i(n, \epsilon_i, \delta_i) \right] \leq \sum_{t=3N}^{n} \mathbb{P} \left\{ u_{i^*}(t, s) < \mu^* - \epsilon_i \text{ for some } 3 \leq s \leq t \right\}
\]

(33)

Here we may make use of the following result:

**Lemma 6** Let \( X_1, X_2, \ldots \) be i.i.d. \( \text{Unif}[a, b] \) random variables, with \( a < b \), and \( a \) and \( b \) finite. For \( k \geq 2 \), let \( W_k = \max_{t \leq k} X_t \) and \( V_k = \min_{t \leq k} X_t \). In that case, the joint density of \( (W_k, V_k) \) is given by:

\[
f_k(w, v) = \begin{cases} 
  k(k-1)(b-a)^{-k}(w-v)^{k-2} & \text{if } a \leq v \leq w \leq b \\
  0 & \text{else}.
\end{cases}
\]

(34)
We therefore have that

\[
P \left( u^* \left( t, s \right) < \mu^* - \epsilon \right) \\
= P \left( \hat{a}_s^* + \frac{1}{2} \left( \hat{a}_s^* - a_s^* \right) \right) t^{1/(s-2)} < \mu^* - \epsilon \left( t, s \right) \\
= \int_{a^*}^{\mu^*-\epsilon} \int_v^{\min \left( b_s^*, v + 2 \frac{(\mu^* - \epsilon) - a^*_i}{s} \right)} f_s(w, v) \, dw \, dv \\
\leq \int_{a^*}^{\mu^*-\epsilon} \int_v^{v + 2 \frac{(\mu^* - \epsilon) - a^*_i}{s}} f_s(w, v) \, dw \, dv \\
= \frac{1}{2} \left( \frac{\mu^* - \epsilon}{b^*_i - a^*_i} \right)^s \\
= \frac{1}{2} t^{-1} \left( 1 - \frac{2 \epsilon}{b^*_i - a^*_i} \right)^s.
\]

The last step is simply the observation that \( \mu^* = (a^*_i + b^*_i)/2 \). For convenience, let \( \alpha = 2\epsilon_i/(b^*_i - a^*_i) \). We therefore have that

\[
\sum_{s=3}^{t} P \left( u^* \left( t, s \right) < \mu^* - \epsilon \right) \leq \sum_{s=3}^{t} \frac{1}{2} t^{-1} t^{-1/(s-2)} (1 - \alpha)^s \\
\leq \sum_{s=1}^{t-2} \frac{1}{2} t^{-1} t^{-1/s} (1 - \alpha)^{s+2} \\
\leq \frac{1}{2} t^{-1} (1 - \alpha)^2 \sum_{s=1}^{\infty} t^{-1/s} (1 - \alpha)^s.
\]

Hence, from Eq. (33) and the above,

\[
E \left[ n^i_3(n, \epsilon_i, \delta_i) \right] \leq \sum_{t=6}^{n} \frac{1}{2} t^{-1} (1 - \alpha)^2 \sum_{s=1}^{\infty} t^{-1/s} (1 - \alpha)^s \\
\leq \frac{1}{2} (1 - \alpha)^2 \sum_{t=6}^{n} t^{-1} \sum_{s=1}^{\infty} t^{-1/s} (1 - \alpha)^s \\
\leq (1 - \alpha)^2 \left( 15 + \frac{3}{\alpha^3} \right).
\]

The last step is a bound proved separately as Proposition 3 in the Appendix. Observing further that \( 1 - \alpha \leq 1 \), we have finally that

\[
E \left[ n^i_3(n, \epsilon_i, \delta_i) \right] \leq 15 + \frac{3}{\alpha^3} = 15 + \frac{3}{8} \frac{(b_i^* - a_i^*)^3}{\epsilon_i^3}.
\]

Observing that \( T^i_\pi(n) \leq T^i_\pi(n + 1) \), bringing the three terms together we have that

\[
E \left[ T^i_\pi(n) \right] \leq \ln n \left( \frac{2\mu^* - 2a_i - 2\epsilon_i - 2\delta_i}{b_i - a_i} \right) + \frac{b_i - a_i}{\delta_i} + \frac{3}{8} \frac{(b_i^* - a_i^*)^3}{\epsilon_i^3} + 18.
\]

The result then follows from the definition of regret, Eq. (2), and the observation again that \( \mu_i = (b_i + a_i)/2 \).
At various points in the results so far, choices of convenience were made with the purpose of keeping associated constants and coefficients ‘nice’. The techniques and results above may actually be refined slightly to present a somewhat stronger result on the remainder term, at the cost of more complicated coefficients. In particular,

**Theorem 7** For any $\beta > 0$,

$$R_{\pi_{CK}}(n) \leq \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i \ln n}{(1 + \frac{2\Delta_i}{S_i})^2} + o((\ln n)^{2/3 + \beta}).$$  \hspace{1cm} (40)

**Proof** Note that, given the result of Theorem 5, it suffices to take $\beta \leq 1/12$.

Building on the proof of Theorem 3, taking $\alpha = \frac{2\epsilon_i}{(b_{i^*} - a_{i^*})} = \frac{2\epsilon_i}{S_i}$ where $i^*$ is the optimal bandit that realizes the smallest value of $b_{i^*} - a_{i^*}$, we have that

$$\mathbb{E}[T_{\pi}(n)] \leq \frac{\ln n}{\ln \left(\frac{2\mu^* - 2a_{i^*} - 2\Delta_i}{b_{i^*} - a_{i^*}}\right)} + \frac{b_{i^*} - a_{i^*}}{\delta_i}$$

$$+ \frac{1}{2}(1 - \alpha)^2 \sum_{t=6}^{n} t^{-1} \sum_{s=1}^{\infty} t^{-1/s}(1 - \alpha)^s + 3$$

$$\leq \frac{\ln n}{\ln \left(1 + \frac{2\Delta_i}{S_i}\frac{1 - \frac{\delta_i}{\Delta_i}}{(1 + \frac{\delta_i}{\Delta_i})}\right)} + \frac{S_i}{\delta_i}$$

$$+ \frac{1}{2} \sum_{t=6}^{n} t^{-1} \sum_{s=1}^{\infty} t^{-1/s}(1 - \alpha)^s + 3. \hspace{1cm} (41)$$

The proof of Theorem 3 then proceeded to bound the above double sum using Proposition 8. Utilizing the proof of Proposition 8 (but without choosing specific values of $p < 1, q > 1$ to render ‘nice’ coefficients), we have

$$\sum_{t=6}^{n} t^{-1} \sum_{s=1}^{\infty} t^{-1/s}(1 - \alpha)^s$$

$$\leq \left( \frac{1 + q}{\epsilon 1 - p} \right)^{\frac{1 + \frac{\epsilon}{\alpha}}{q - 1}} + \frac{1}{\alpha} \left( \frac{1 + q}{\epsilon \alpha p} \right)^{\frac{\epsilon}{q - 1}} \frac{1}{q - 1} \hspace{1cm} (42)$$

$$= \alpha^{-1 - \frac{\epsilon}{q}} C_1(p, q) + C_2(p, q)$$

$$= \epsilon_i^{-1 - \frac{\epsilon}{q}} \left( \frac{S_i}{2} \right)^{1 + \frac{\epsilon}{q}} C_1(p, q) + C_2(p, q).$$

Where for convenience we are defining $C_1, C_2$ as the associated functions of $p, q$. Note, they are finite for $p < 1, q > 1$. Let $0 < \epsilon < 1$ and define $G_i = \min \left( S_i, S_i, \frac{1}{2} \Delta_i \right)$ as in Theorem 5. Taking $\epsilon_i = \delta_i = \epsilon G_i$, we have the following bound (utilizing Proposition 9 as in the proof of Theorem 5):

$$\mathbb{E}[T_{\pi}(n)] \leq \frac{\ln n}{\ln \left(1 + \frac{2\Delta_i}{S_i}\right)^2} + \frac{8G_i \epsilon \ln n}{(S_i + 2\Delta_i) \ln \left(1 + \frac{2\Delta_i}{S_i}\right)^2}$$

$$+ \frac{S_i}{G_i} \epsilon^{-1} + \epsilon^{-1 - \frac{\epsilon}{q}} \left( \frac{S_i}{2G_i} \right)^{1 + \frac{\epsilon}{q}} C_1(p, q) + C_2(p, q) \hspace{1cm} (43)$$

$$+ C_2(p, q) + 3.$$
At this point, taking $\epsilon = (\ln n)^{-p/(2p+q)}$ yields the following

$$E \left[ T^i(n) \right] \leq \frac{\ln n}{\ln \left( 1 + \frac{2\Delta_i}{S_i} \right)} + \frac{8G_i(\ln n)^{\frac{p+q}{2p+q}}}{(S_i + 2\Delta_i) \ln \left( 1 + \frac{2\Delta_i}{S_i} \right)^2}$$

$$+ \frac{S_i}{G_i}(\ln n)^{\frac{p}{2p+q}} + (\ln n)^{\frac{p+q}{2p+q}} \left( \frac{S_i}{2G_i} \right)^{1+p} C_1(p, q)$$

$$+ C_2(p, q) + 3,$$

(44)

or more conveniently,

$$E \left[ T^i(n) \right] \leq \frac{\ln n}{\ln \left( 1 + \frac{2\Delta_i}{S_i} \right)} + O((\ln n)^{\frac{p+q}{2p+q}}) + O((\ln n)^{\frac{p}{2p+q}}).$$

(45)

Taking $q = \gamma p$, where $(\gamma, p)$ is chosen such that $1/\gamma < p < 1$, the above yields (via the definition of regret, Eq. (2)):

$$R_\pi(n) \leq \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i \ln n}{\ln \left( 1 + \frac{2\Delta_i}{S_i} \right)} + O((\ln n)^{\frac{p+q}{2p+q}}) + O((\ln n)^{\frac{p}{2p+q}}).$$

(46)

At this point, note that taking $\gamma = 2$ recovers the remainder order given in Theorem 5. For a given $1/12 > \beta > 0$, taking $\gamma < (1+6\beta)/(1-3\beta)$ yields $(1+\gamma)/(2+\gamma) < 2/3+\beta$, and completes the proof.

Acknowledgements

We would like to acknowledge support for this project from the National Science Foundation (NSF grant CMMI-14-50743).

Appendix A: Additional Proofs

**Proposition 8** For $0 < \alpha < 1$, for all $n \geq 6$,

$$\sum_{t=6}^{n} \prod_{s=1}^{\infty} t^{-1/s}(1-\alpha)^s \leq 30 + \frac{6}{\alpha^3}.$$  

(47)
Proof [Proof of Proposition 8] Let $1 > p > 0$. We have

$$\sum_{s=1}^{\infty} t^{-1/s}(1 - \alpha)^s$$

$$= \sum_{s=1}^{[\ln(t)^p]} t^{-1/s}(1 - \alpha)^s + \sum_{s=[\ln(t)^p]}^{\infty} t^{-1/s}(1 - \alpha)^s$$

$$\leq \sum_{s=1}^{[\ln(t)^p]} t^{-1/s} + \sum_{s=[\ln(t)^p]}^{\infty} (1 - \alpha)^s$$

(48)

$$\leq \ln(t)^p t^{-1/\ln(t)^p} + \frac{1}{\alpha} (1 - \alpha)^{\ln(t)^p}$$

$$= \ln(t)^p e^{-\ln(t)^p} + \frac{1}{\alpha} (1 - \alpha)^{\ln(t)^p}.$$  

Here we may make use of the following bounds, that for $x \geq 0$, $q > 0$,

$$x^p e^{-x^p} \leq \left( \frac{1}{e} \right)^{\frac{1}{e - 1} - p} x^{-q}$$

$$\alpha \leq \left( \frac{1}{e \ln(1 - \alpha)} \right)^{\frac{x}{p}} x^{-q} \leq \left( \frac{1}{e \alpha p} \right)^{\frac{x}{p}} x^{-q}.$$ (49)

Applying these to the above,

$$\sum_{s=1}^{\infty} t^{-1/s}(1 - \alpha)^s \leq \left( \left( \frac{1}{e} \right)^{\frac{1}{e - 1} - p} + \frac{1}{\alpha} \left( \frac{1}{e \alpha p} \right)^{\frac{x}{p}} \right) \ln(t)^{-q}.$$ (50)

Hence, taking $q > 1$,

$$\sum_{l=6}^{n} \frac{1}{l} \sum_{s=1}^{\infty} t^{-1/s}(1 - \alpha)^s$$

$$\leq \left( \left( \frac{1}{e} \right)^{\frac{1}{e - 1} - p} + \frac{1}{\alpha} \left( \frac{1}{e \alpha p} \right)^{\frac{x}{p}} \right) \sum_{l=6}^{n} \frac{1}{l} \ln(t)^{-q}$$

$$\leq \left( \left( \frac{1}{e} \right)^{\frac{1}{e - 1} - p} + \frac{1}{\alpha} \left( \frac{1}{e \alpha p} \right)^{\frac{x}{p}} \right) \int_{e}^{n} \frac{1}{l} \ln(t)^{-q} dt$$

$$= \left( \left( \frac{1}{e} \right)^{\frac{1}{e - 1} - p} + \frac{1}{\alpha} \left( \frac{1}{e \alpha p} \right)^{\frac{x}{p}} \right) \frac{1 - \ln(n)^{1-q}}{q - 1}$$

$$\leq \left( \left( \frac{1}{e} \right)^{\frac{1}{e - 1} - p} + \frac{1}{\alpha} \left( \frac{1}{e \alpha p} \right)^{\frac{x}{p}} \right) \frac{1}{q - 1}.$$ (51)

At this point, taking $q = 2p$ and $p = 0.55$ yields

$$\sum_{l=6}^{n} \frac{1}{l} \sum_{s=1}^{\infty} t^{-1/s}(1 - \alpha)^s \leq 29.9628 + \frac{5.41341}{\alpha^3},$$ (52)

which, rounding up, completes the result.
Proposition 9  For $Q > 0$, and $0 \leq \epsilon < 1$, the following bound holds:

$$\frac{1}{\ln(1 + Q(1 - \epsilon))} \leq \frac{1}{\ln(1 + Q)} + \frac{\epsilon}{1 - \epsilon (1 + Q) \ln(1 + Q)^2}. \tag{53}$$

Proof [Proof of Proposition 9] Let $A(Q, \epsilon)$ denote the RHS of the above, $B(Q, \epsilon)$ denote the left. We adopt the physicists’ convention of denoting the partial derivative of $F$ with respect to $x$ as $F_x$.

Note, $A(Q, 0) \leq B(Q, 0)$. Hence, it suffices to demonstrate that $A_\epsilon \leq B_\epsilon$ over this range or, since they are both positive,

$$\frac{A_\epsilon}{B_\epsilon} = \frac{(1 + Q)(1 - \epsilon)^2 \ln(1 + Q)^2}{(1 + Q(1 - \epsilon)) \ln(1 + Q(1 - \epsilon))} \leq 1. \tag{54}$$

We take, for convenience, $\delta = 1 - \epsilon$, and want to show that for $0 \leq \delta \leq 1$:

$$\frac{(1 + Q)\delta^2 \ln(1 + Q)^2}{(1 + Q\delta) \ln(1 + Q\delta)^2} \leq 1. \tag{55}$$

The above inequality holds when $\delta = 1$. Taking $C(\delta, Q)$ as the above simplified ratio, it suffices to show that $C_\delta \geq 0$. Simplifying this inequality and canceling the positive factors, it is equivalent to show that $-2Q\delta + (2 + Q\delta) \ln(1 + Q\delta) \geq 0$, or taking $x = Q\delta > 0$,

$$\ln(1 + x) \geq \frac{2x}{2 + x}. \tag{56}$$

This is a fairly standard and easily verified inequality for the function $\ln()$. This completes the proof.

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. arXiv preprint arXiv:1111.1797, 2011.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science, 410(19):1876–1902, 2009.

Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. Periodica Mathematica Hungarica, 61(1-2):55–65, 2010.

Peter Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pages 35–42. AUAI Press, 2009.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. arXiv preprint arXiv:1202.4473, 2012.

Apostolos Burnetas and Michael N Katehakis. On sequencing two types of tasks on a single processor under incomplete information. Probability in the Engineering and Informational Sciences, 7(1):85–119, 1993.

Apostolos Burnetas and Michael N Katehakis. On large deviations properties of sequential allocation problems. Stochastic Analysis and Applications, 14(1):23–31, 1996a.
Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996b.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Apostolos N Burnetas and Michael N Katehakis. Asymptotic bayes analysis for the finite-horizon one-armed-bandit problem. *Probability in the Engineering and Informational Sciences*, 17(01):53–82, 2003.

Sergiy Butenko, Panos M Pardalos, and Robert Murphey. *Cooperative Control: Models, Applications, and Algorithms*. Kluwer Academic Publishers, 2003.

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

Wesley Cowan and Michael N Katehakis. Asymptotic behavior of minimal-exploration allocation policies: Almost sure, arbitrarily slow growing regret. arXiv preprint arXiv:1505.02865, Jul. 31 2015a.

Wesley Cowan and Michael N Katehakis. Multi-armed bandits under general depreciation and commitment. *Probability in the Engineering and Informational Sciences*, 29(01):51–76, 2015b.

Wesley Cowan, Junya Honda, and Michael N Katehakis. Asymptotic optimality, finite horizon regret bounds, and a solution to an open problem. arXiv preprint arXiv:1504.05823, 2015.

Savas Dayanik, Warren B Powell, and Kazutoshi Yamazaki. Asymptotically optimal Bayesian sequential change detection and identification rules. *Annals of Operations Research*, 208(1):337–370, 2013.

Eugene A Feinberg, Pavlo O Kasyanov, and Michael Z Zgurovsky. Convergence of value iterations for total-cost mdps and pomdps with general state and action sets. In *Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2014 IEEE Symposium on*, pages 1–8. IEEE, 2014.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning based on Kullback Leibler divergence. In *48th Annual Allerton Conference on Communication, Control, and Computing*, 2010.

John C. Gittins. Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Stat. Soc. Ser. B*, 41:335–340, 1979.

John C. Gittins, Kevin Glazebrook, and Richard R. Weber. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, West Sussex, U.K., 2011.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79. Citeseer, 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. Optimality of Thompson sampling for Gaussian bandits depends on priors. arXiv preprint arXiv:1311.1894, 2013.

Wassim Jouini, Damien Ernst, Christophe Moy, and Jacques Palicot. Multi-armed bandit based policies for cognitive radio’s decision making issues. In *3rd international conference on Signals, Circuits and Systems (SCS)*, 2009.
Michael N Katehakis and Cyrus Derman. Computing optimal sequential allocation rules. In *Clinical Trials*, volume 8 of *Lecture Note Series: Adaptive Statistical Procedures and Related Topics*, pages 29–39. Institute of Math. Stats., 1986.

Michael N Katehakis and Arthur F Veinott Jr. The multi-armed bandit problem: decomposition and computation. *Math. Oper. Res.*, 12:262–68, 1987.

Emilie Kaufmann. Analyse de stratégies Bayésiennes et fréquentistes pour l’allocation séquentielle de ressources. *Doctorat*, ParisTech., Jul. 31 2015.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1448–1456. Curran Associates, Inc., 2013.

Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107–1149, 2003.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Lihong Li, Remi Munos, and Csaba Szepesvari. On minimax optimal offline policy evaluation. arXiv preprint arXiv:1409.3653, 2014.

Michael L Littman. Inducing partially observable Markov decision processes. In *ICGI*, pages 145–148, 2012.

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored mdps. In *Advances in Neural Information Processing Systems*, pages 604–612, 2014.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Monthly*, 58:527–536, 1952.

Cem Tekin and Mingyan Liu. Approximately optimal adaptive learning in opportunistic spectrum access. In *INFOCOM, 2012 Proceedings IEEE*, pages 1548–1556. IEEE, 2012.

Ambuj Tewari and Peter L Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. In *Advances in Neural Information Processing Systems*, pages 1505–1512, 2008.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Richard R Weber. On the Gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4):1024–1033, 1992.