Asymptotically Optimal Multi-Armed Bandit Policies under a Cost Constraint

Apostolos N. Burnetas ABURNETAS@MATH.UOA.GR
Department of Mathematics
National and Kapodistrian University
Panepistemiopolis, Athens 15784, Greece

Odysseas Kanavetas OKANAVETAS@SABANCIUNIV.EDU
Department of Industrial Engineering
Sabanci University
Orhanli Tuzla, Istanbul 34956, Turkey

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

Abstract

We develop asymptotically optimal policies for the multi armed bandit (MAB), problem, under a cost constraint. This model is applicable in situations where each sample (or activation) from a population (bandit) incurs a known bandit dependent cost. Successive samples from each population are iid random variables with unknown distribution. The objective is to have a feasible policy for deciding from which population to sample from, so as to maximize the expected sum of outcomes of n total samples or equivalently to minimize the regret due to lack on information of sample distributions, For this problem we consider the class of feasible uniformly fast (f-UF) convergent policies, that satisfy sample path wise the cost constraint. We first establish a necessary asymptotic lower bound for the rate of increase of the regret function of f-UF policies. Then we construct a class of f-UF policies and provide conditions under which they are asymptotically optimal within the class of f-UF policies, achieving this asymptotic lower bound. At the end we provide the explicit form of such policies for the case in which the unknown distributions are Normal with unknown means and known variances.

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

Introduction

Consider the problem of sequential sampling from a finite number of independent statistical populations, where successive samples from a population are iid random variables with unknown distribution.

Consider the problem of sequential sampling from k independent statistical populations, Π^i, $i = 1, \ldots, k$. Successive samples from population i constitute a sequence of i.i.d. random variables $X_1^i, X_2^i, \ldots$ following a univariate distribution with density $f_i(\,|\underline{\theta}_i)$ with respect to a nondegenerate measure v. The density $f_i(\,|\,)$ is known and $\underline{\theta}_i$ is a parameter belonging to some set Θ_i. Let $\underline{\underline{\theta}} = (\underline{\theta}_1, \ldots, \underline{\theta}_k)$ denote the set of parameters, $\underline{\underline{\theta}} \in \Theta$, where $\Theta \equiv \Theta_1 \times \ldots \times \Theta_k$. Given $\underline{\underline{\theta}}$ let $\underline{\mu}(\underline{\underline{\theta}}) = (\mu_1(\underline{\theta}_1), \ldots, \mu_k(\underline{\theta}_k))$ be the vector of expected values, i.e. $\mu_i(\underline{\theta}_i) = E_{\underline{\underline{\theta}}}(X^i)$. The true value $\underline{\underline{\theta}}_0$ of $\underline{\underline{\theta}}$ is unknown. We make the assumption that outcomes from different populations are independent.

Sampling from population Π_i incurs cost c_i per sample, and without loss of generality we assume $c^1 \leq c^2 \leq \ldots \leq c^N$, and not all c^i are equal. Without loss of generality we assume $c^1 \leq c^0 < c^k$. In case $c^0 < c^1$ the problem is infeasible and in the other case where $c^0 \geq c^k$ the cost constraint is redundant. Let $d = \max\{j : c^j \leq c^0\}$. Then $1 \leq d < k$ and $c^d \leq c^0 < c^{d+1}$. We consider adaptive policies which depend only in the past observations of selections and outcomes. Specifically, let A_t, X_t , $t = 1, 2, ...$ denote the population selected and the observed outcome at period t. Let $h_t = (\alpha_1, x_1, ..., \alpha_{t-1}, x_{t-1})$ denote a history of actions and observations available at period t. An adaptive policy is a sequence $\pi = (\pi_1, \pi_2, ...)$ of history dependent probability distributions on $\{1, ..., k\}$, such that $\pi_t(j, h_t) = P(A_t = j|h_t)$. Given h_n, let $T_\pi^\alpha(n)$ denote the number of times population α has been sampled during the first n periods $T_\pi^\alpha(n) = \sum_{t=1}^n 1\{A_t = \alpha\}$. Let $\mathcal{V}_\pi(n)$ and $\mathcal{C}_\pi(n)$ be respectively the total *reward earned* and total *cost incurred* up to period n, i.e.,

$$\mathcal{V}_\pi(n) = \sum_{i=1}^{k} \sum_{t=1}^{T_\pi^i(n)} X_t^i, \tag{1}$$

$$\mathcal{C}_\pi(n) = \sum_{i=1}^{k} \sum_{t=1}^{T_\pi^i(n)} c^i 1\{A_t = i\}. \tag{2}$$

We call an adaptive policy *feasible* if

$$\mathcal{C}_\pi(n)/n \leq c^0, \quad \forall\ n = 1, 2, \ldots \tag{3}$$

The objective is to obtain a feasible policy π that maximizes in some sense $E_{\underline{\underline{\theta}}} \mathcal{V}_\pi(n)$, $\forall \underline{\underline{\theta}} \in \underline{\underline{\Theta}}$. In the next section we will show that this is equivalent to minimizing a *regret function* $R_\pi(\underline{\underline{\theta}}, n)$ that represents the expected loss due to lack on information of sample distributions. For this, we consider the class of feasible policies that are uniformly fast (UF) convergent, in the sense of Burnetas and Katehakis (1996b); we call these polices (f-UF) policies. We first establish in Theorem 1, a necessary asymptotic lower bound for the rate of increase of the regret function of f-UF policies. Then we construct a class of "block f-UF" policies and provide conditions under which they are asymptotically optimal within the class of f-UF policies, achieving this asymptotic lower bound, cf. Theorem 2. At the end we provide the explicit form of an asymptotical optimal f-UF policy, for the case in which the unknown distributions are Normal with unknown means and known variances. These policies form the basis for deriving logarithmic regret polices for more general models, cf. Auer et al. (2002), Auer and Ortner (2010), Cowan et al. (to appear), Cowan and Katehakis (2015a).

The extensive literature on the multi-armed bandit (MAB) problem, includes the following: Lai and Robbins (1985), Katehakis and Robbins (1995), Kleinberg (2004), Mahajan and Teneketzis (2008), Audibert et al. (2009), Auer and Ortner (2010), Honda and Takemura (2011), Bubeck and Slivkins (2012), Cowan and Katehakis (2015b) and references therein. As far as we know, the first formulation of the MAB problem with a side constraint considered herein was given in Burnetas and Katehakis (1998). Tran-Thanh et al. (2010), considered the problem when the cost of activation of each arm is fixed and becomes known after the arm is used once. Burnetas and Kanavetas (2012) considered a version of this problem and constructed a consistent policy (i.e., with regret $R_\pi(n) = o(n)$). In this paper we employ a stricter version of the average cost constraint that requires the average sampling cost not to exceed c^0 at any time period and not only in the limit. Badanidiyuru et al. (2013), considered the problem where there can be more than one side constraints ("knapsack") and showed how to construct polices with sub-linear regret. They also discuss interesting applications of the model, such as to: problems of dynamic pricing Wang et al. (2014), Johnson et al. (2015), dynamic procurement Singla and Krause (2013), and auctions Tran-Thanh et al. (2014). Ding et al. (2013) constructed UF policies (i.e., with regret $R_\pi(n) = o(logn)$) for cases in which activation costs are bandit dependent iid random variables. For other recent related work we refer to: Guha and

Munagala (2007), Tran-Thanh et al. (2012), Thomaidou et al. (2012), Lattimore et al. (2014), Sen et al. (2015).

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

Model description - Preliminaries

The complete information problem, where $\underline{\underline{\theta}}$ is known, and the expected average reward is to be maximized, can be solved via the following linear program (LP-1).

$$\begin{aligned} z^*(\underline{\underline{\theta}}) &= \max \sum_{j=1}^{k} \mu_j(\underline{\theta}_j) x_j \\ & \sum_{j=1}^{k} c^j x_j + y = c^0 \\ & \sum_{j=1}^{k} x_j = 1 \\ & x_j \geq 0, \forall j \; y \geq 0. \end{aligned} \tag{4}$$

The solution is a randomized sampling policy which at each period selects population j with probability x_j, for $j = 1, \ldots, k$, where the randomization probabilities x_j are an optimal solution to the above linear program (LP), cf. Burnetas and Kanavetas (2012); Burnetas and Katehakis (1998). However, such policy may not be feasible in our framework that requires $\mathcal{C}_\pi(n)/n \leq c^0, \ \forall \ n = 1, 2, \ldots,$ because simple randomization may lead to sampling in such a way that $\mathcal{C}_\pi(n)/n$ exceeds c^0, for some periods. However, in the complete information setting, under the assumption that the coefficients c^j are all rational, any optimal solution of LP-1 which is an extreme point is also rational, thus an optimal randomized policy can be implemented as a *periodic sampling policy* within *blocks* of time periods within which the order of sampling can be set so that the sampling cost constraint is never violated, and the sampling frequencies remain equal to x_j. We use generalizations of this idea in the incomplete information framework in the sequel.

We next introduce necessary notation regarding the LP-1. First, its dual problem (DLP-1) is

$$\begin{aligned} z_D^*(\underline{\underline{\theta}}) &= \min \; g + c^0 \lambda \\ & g + c^1 \lambda \geq \mu_1(\underline{\theta}_1) \\ & \qquad \vdots \\ & g + c^k \lambda \geq \mu_k(\underline{\theta}_k) \\ & g \in \mathbf{R}, \lambda \geq 0. \end{aligned}$$

A basic matrix B is of the form $\begin{pmatrix} c^i & c^j \\ 1 & 1 \end{pmatrix}$, for some $i \leq d < j$ or $\begin{pmatrix} c^i & 1 \\ 1 & 0 \end{pmatrix}$ for some $i \leq d$. They correspond to sampling from the pair (i, j) or population i, respectively. We denote the Basic

Feasible Solution (BFS) corresponding to matrix B as $b = \{i, j\}$ or $b = \{i\}$, respectively. Note that in the case of degenerate BFS b, more than one matrices B correspond to the same b.

We use K to denote the set of BFS:

$$K = \{b \ : \ b = \{i, j\}, \ i \leq d \leq j \text{ or } b = \{i\}, \ i \leq d\}.$$

Since the feasible region of Eq. (16) is bounded, K is finite.

For a basic matrix B, let $v^B = (\lambda^B, g^B)$ denote the dual vector corresponding to B, i.e., $v^B = \mu_B(\underline{\underline{\theta}})B^{-1}$, where $\mu_B(\underline{\underline{\theta}}) = (\mu_i(\underline{\theta}_i), \mu_j(\underline{\theta}_j))$, or $\mu_B(\underline{\underline{\theta}}) = (\mu_i(\underline{\theta}_i), 0)$, depending on the form of B.

Regarding optimality, a BFS is optimal if and only if for at least one corresponding basic matrix B the reduced costs (dual slacks) are all nonnegative:

$$\phi^B_\alpha(\underline{\underline{\theta}}) \equiv c^\alpha \lambda^B + g^B - \mu_\alpha(\underline{\theta}_\alpha) \geq 0, \ \alpha = 1, \ldots, k.$$

A basic matrix B satisfying this condition is optimal. It is easy to show that the reduced cost can be expressed as a linear combination of the unknown population means, i.e., $\phi^B_\alpha(\underline{\underline{\theta}}) = \underline{w}^B_\alpha \underline{\mu}(\underline{\underline{\theta}})$, where $\underline{w}^B_\alpha$ is an appropriately defined vector that does not depend on $\underline{\mu}(\underline{\underline{\theta}})$. In the sequel we use the notation $s(\underline{\underline{\theta}})$ to denote the set with optimal solutions of LP-1 for a vector $\underline{\mu}(\underline{\underline{\theta}})$, i.e., $s(\underline{\underline{\theta}}) = \{b \in K : b \text{ corresponds to an optimal BFS}\}$.

We define the loss or regret function of policy π as the finite horizon loss in expected reward with respect to the optimal policy under complete information:

$$\begin{aligned} R_\pi(\underline{\underline{\theta}}, n) &= nz^*(\underline{\underline{\theta}}) - E_{\underline{\underline{\theta}}} \mathcal{V}_\pi(n) \\ &= nz^*(\underline{\underline{\theta}}) - \sum_{j=1}^{k} \mu_j(\underline{\theta}_j) E_{\underline{\underline{\theta}}} T^j_\pi(n) \end{aligned} \tag{5}$$

We next derive an equivalent expression that relates the regret to the solution of the complete information LP. Recall that for any basic matrix B which corresponds to an optimal solution of LP-1, from the DLP-1 program we have that $\forall j$: $z^*(\underline{\underline{\theta}}) = c^0 \lambda^B + g^B$ and $\mu_j(\underline{\theta}_j) = c^j \lambda^B + g^B - \phi^B_j(\underline{\underline{\theta}})$. These relations and Eq. (5) imply:

$$R_\pi(\underline{\underline{\theta}}, n) = \sum_{j=1}^{k} \phi^B_j(\underline{\underline{\theta}}) E_{\underline{\underline{\theta}}} T^j_\pi(n) + \lambda^B \ \sum_{j=1}^{k} (c^0 - c^j) E_{\underline{\underline{\theta}}} T^j_\pi(n), \tag{6}$$

for any $\underline{\underline{\theta}} \in \Theta$ and $B \in s(\underline{\underline{\theta}})$.

We now state:

Definition 1. a) A feasible policy π is called *consistent* if

$$R_\pi(\underline{\underline{\theta}}, n) = o(n), \ n \to \infty, \ \forall \ \underline{\underline{\theta}} \in \Theta.$$

b) A feasible policy π is called *f-uniformly fast* (f-UF) if

$$R_\pi(\underline{\underline{\theta}}, n) = o(n^a), \ n \to \infty, \ \forall \ a > 0, \ \forall \ \underline{\underline{\theta}} \in \Theta.$$

In the sequel we will show that there exist f-UF policies, following the approach of (Burnetas and Katehakis 1996), by construction of a function $M(\underline{\underline{\theta}})$ and a f-UF policy π^0 such that

$$\liminf R_{\pi^0}(\underline{\underline{\theta}}, n)/\log n \leq M(\underline{\underline{\theta}}) \ \ \forall \underline{\underline{\theta}} \in \Theta.$$

The assymptotic optimality of π^0 then follows from Theorem 1. Detailed proofs are provided in the appendix.

Lower Bound for the Regret

Recall that for $b \in K$, b is an optimal solution of linear program LP-1 for some $\underline{\underline{\theta}} \in \Theta$ if and only if for at least one corresponding basic matrix B, $\phi^B_\alpha(\underline{\underline{\theta}}) \geq 0,\ \alpha = 1, \ldots, k$.

For any $b \in s(\underline{\underline{\theta}})$, where $b = \{i, j\}$ or $\{i\}$ and $\alpha \neq i, j$, we define the sets $\Delta\Theta_\alpha(\underline{\underline{\theta}})$ and $D(\underline{\underline{\theta}})$, as follows. The first set contains all values of Θ_α under which the complete information problem under the perturbed $\underline{\underline{\theta}}^{'}$ has a unique optimal BFS which includes population α. The second set $D(\underline{\underline{\theta}})$, contains all populations which are not contained in any optimal solution under parameter set $\underline{\underline{\theta}}$ but, by varying only parameter $\underline{\theta}_\alpha$, a uniquely optimal BFS that contains them can be found. Formally,

$$\Delta\Theta_\alpha(\underline{\underline{\theta}}) = \{\theta^{'}_\alpha \in \Theta_\alpha : s(\underline{\underline{\theta}}^{'}) = \{\{i, \alpha\} \text{ or } \{\alpha, j\} \text{ or } \{\alpha\}\}\},$$

where $\underline{\underline{\theta}}^{'} = (\underline{\theta}_1, \ldots, \underline{\theta}^{'}_\alpha, \ldots, \underline{\theta}_k)$, is a new vector such that only parameter $\underline{\theta}^{'}_\alpha$ is changed from $\underline{\theta}_\alpha$.

$$D(\underline{\underline{\theta}}) = \{\alpha :\ \alpha \notin b \text{ for any } b \in s(\underline{\underline{\theta}}) \text{ and } \Delta\Theta_\alpha(\underline{\underline{\theta}}) \neq \emptyset\},$$

Let $I(\underline{\theta}_\alpha, \underline{\theta}^{'}_\alpha)$ denote the Kullback-Leibler information number, defined as

$$I(\underline{\theta}_\alpha, \underline{\theta}^{'}_\alpha) = \int_{-\infty}^{+\infty} \log \frac{f(x; \underline{\theta}_\alpha)}{f(x; \underline{\theta}^{'}_\alpha)} f(x; \underline{\theta}_\alpha) dv(x).$$

Now we can define the minimum deviation, in the sense of the Kullback-Leibler information number, of parameter $\underline{\theta}^{'}_\alpha$ from $\underline{\theta}_\alpha$ in order to achieve that the population α becomes optimal under $\underline{\theta}^{'}_\alpha$.

$$K_\alpha(\underline{\underline{\theta}}) \quad = \quad \inf\{I(\underline{\theta}_\alpha, \underline{\theta}^{'}_\alpha) :\ \underline{\theta}^{'}_\alpha \in \Delta\Theta_\alpha(\underline{\underline{\theta}})\}.$$

We have:
Lemma 1 For any $\underline{\underline{\theta}}$, and any optimal matrix B under $\underline{\underline{\theta}}$, $\exists\ \rho = \rho(\underline{\underline{\theta}}, \alpha, B)$ such that for any $\underline{\theta}^{'}_\alpha \in \Delta\Theta_\alpha(\underline{\underline{\theta}})$:
(i) $\phi^B_j(\underline{\underline{\theta}}^{'}) = \phi^B_j(\underline{\underline{\theta}}) \geq 0,\ \forall\ j \neq \alpha$ and $\phi^B_\alpha(\underline{\underline{\theta}}^{'}) = \phi^B_\alpha(\underline{\underline{\theta}}) + \mu_\alpha(\underline{\theta}_\alpha) - \mu_\alpha(\underline{\theta}^{'}_\alpha) < 0$,
(ii) $\mu^*_\alpha(\underline{\underline{\theta}}) < \mu_\alpha(\underline{\theta}^{'}_\alpha) < \mu^*_\alpha(\underline{\underline{\theta}}) + \rho$, where $\rho > 0$ and $\mu^*_\alpha(\underline{\underline{\theta}}) = \phi^B_\alpha(\underline{\underline{\theta}}) + \mu_\alpha(\underline{\theta}_\alpha)$.

The above Lemma implies the following form for $K_\alpha(\underline{\underline{\theta}})$ which is necessary for the proof of Lemmas and Theorems of the paper, $K_\alpha(\underline{\underline{\theta}})$ is equal to:

$$\inf\{I(\underline{\theta}_\alpha, \underline{\theta}^{'}_\alpha) :\ \underline{\theta}^{'}_\alpha \in \Theta_\alpha,\ \mu^*_\alpha(\underline{\underline{\theta}}) < \mu_\alpha(\underline{\theta}^{'}_\alpha) < \mu^*_\alpha(\underline{\underline{\theta}}) + \rho\},$$

where $\rho = \rho(\underline{\underline{\theta}}, \alpha, B) > 0$.

Lemma 2 and Proposition 1 below are used to establish the following Lemma 3 from which Theorem 1 for the regret function follows.

First note that in Eq. (6) both terms are nonnegative, the first because of optimality and the second because of feasibility. Therefore it follows that a necessary and sufficient condition for a policy π to be f-UF is that for $\underline{\underline{\theta}} \in \Theta$ and any optimal BFS b under $\underline{\underline{\theta}}$ and for all B corresponds to b.

$$\phi^B_j(\underline{\underline{\theta}}) \lim_{n \to \infty} \frac{E_{\underline{\underline{\theta}}} T^j_\pi(n)}{n^a} = 0, \text{ for all } a > 0,\ j \notin b, \tag{7}$$

and also,

$$\lambda^B \lim_{n \to \infty} \frac{\sum_{j \in b}(c^0 - c^j) E_{\underline{\underline{\theta}}} T^j_\pi(n)}{n^a} = 0. \tag{8}$$

We can now state:

Lemma 2 If there is a uniquely optimal BFS and $B \in s(\underline{\underline{\theta}})$. Then
(i) if $B = \begin{pmatrix} c^i & c^j \\ 1 & 1 \end{pmatrix}$, for some $i \leq d < j \Rightarrow \lambda^B > 0$,
(ii) if $B = \begin{pmatrix} c^i & 1 \\ 1 & 0 \end{pmatrix}$, for some $i \leq d \Rightarrow \lambda^B = 0$.

Proposition 1 For any f-UF policy π and for all $\underline{\underline{\theta}} \in \Theta$ we have that for $\alpha \in D(\underline{\underline{\theta}})$, any $\underline{\underline{\theta}}^{'} \in \Delta(\underline{\underline{\theta}})$ and for all positive sequences: $\beta_n = o(n)$ it is true that

$$P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) < \beta_n] = o(n^{a-1}), \text{ for all } a > 0.$$

Lemma 3 If $P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) < \beta_n] = o(n^{a-1})$, for all $a > 0$ and a positive sequence $\beta_n = o(n)$ then

$$\lim_{n\to\infty} P_{\underline{\underline{\theta}}}[T_\pi^\alpha(n) < \frac{\log n}{K_\alpha(\underline{\underline{\theta}})}] = 0,$$

for all $\underline{\underline{\theta}} \in \Theta$ and $\alpha \in \Delta(\underline{\underline{\theta}})$.

We next define the function $M(\underline{\underline{\theta}})$ and prove the main theorem of this section. Let

$$M(\underline{\underline{\theta}}) = \sum_{j \in D(\underline{\underline{\theta}})} \frac{\phi_j^B(\underline{\underline{\theta}})}{K_j(\underline{\underline{\theta}})}.$$

Theorem 1 If π is an f-UF policy then

$$\liminf_{n\to\infty} \frac{R_\pi(\underline{\underline{\theta}}, n)}{\log n} \geq M(\underline{\underline{\theta}}), \ \forall \underline{\underline{\theta}} \in \Theta.$$

Proof Recall,

$$R_\pi(\underline{\underline{\theta}}, n) = \sum_{j=1}^{k} \phi_j^B(\underline{\underline{\theta}}) E_{\underline{\underline{\theta}}} T_\pi^j(n) + \lambda^B[nc^0 - E_{\underline{\underline{\theta}}} C_\pi(n)],$$

and by Lemma 3, using the Markov inequality, we obtain that if π is f-UF, then

$$\liminf_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T_\pi^j(n)}{\log n} \geq \frac{1}{K_j(\underline{\underline{\theta}})}, \ \forall j \in D(\underline{\underline{\theta}}), \ \forall \underline{\underline{\theta}} \in \Theta.$$

Also, we have from Lemma 2 that $\lambda^B \geq 0$ and from Eq. (3), we have that $nc^0 - E_{\underline{\underline{\theta}}} \mathcal{C}_\pi(n) \geq 0$, for all n. Finally, we have that the optimal populations under $\underline{\underline{\theta}}$ have $\phi_j^B(\underline{\underline{\theta}}) = 0$, thus

$$\liminf_{n\to\infty} \frac{R_\pi(\underline{\underline{\theta}}, n)}{\log n} \geq \sum_{j \in D(\underline{\underline{\theta}})} \frac{\phi_j^B(\underline{\underline{\theta}})}{K_j(\underline{\underline{\theta}})}, \text{ for all } \underline{\underline{\theta}} \in \Theta.$$

Blocks and Block Based Policies

We consider a class of policies such that the sampling is performed in groups of subsequent periods called sampling blocks, of finite length, where the total cost of actions in each block satisfies the cost constraint of Eq. (3) as follows. Define the differences

$$\delta^i \equiv c^i - c^0.$$

δ^i expresses the net effect of a single observation from a population i on the sampling budget. This effect is a cost if $\delta^i > 0$ or a benefit (net savings) if $\delta^i < 0$.

The original problem is equivalent to the transformed problem where $c^i = \delta^i$, $i = 1, ..., k$, $c^0 = 0$ and the sampling constraint is

$$\frac{1}{n}\sum_{t=1}^{n}\delta^{A_t} \leq 0, \ \forall \ n.$$

Since δ^i is assumed to be rational, for each $i = 1, \ldots, k$ and there is a finite number of them we may assume, without loss of generality, that they are all integers.

Let $J \subseteq \{1, ..., k\}$ be the subset of populations sampled within a sampling block. The "cheap" populations in J must be sampled often enough to finance sampling of the "expensive" ones. Mathematically it suffices to find $\{m_j, j \in J\}$ such that each population $j \in J$ is sampled m_j times, and $\sum_{j \in J} m_j \delta^j \leq 0$, $m_j \in \mathbf{N}$, $\forall \ j \in J$. Any block with m_j satisfying the previous properties is called admissible. One possibility is to consider the smallest block, which will be appropriate in the incomplete information case. Thus the minimum length of the sampling block, $\ell(J)$, is the solution of the following linear program

$$\ell(J) = \min\{\sum_{j \in J} m_j \ : \ \sum_{j \in J} m_j \delta^j \leq 0 \ \& \ m_j \in \mathbf{N}, \ \forall \ j \in J\}.$$

An optimal solution of LP-1 specifies randomization probabilities that guarantee maximization of the average reward subject to the cost constraint. The populations into this optimal solution define the set J, and J, δ^i and ℓ are observable constants.

We use the Initial Sampling Block (ISB) and Linear Programming Block (LPB) blocks below to define a class of policies $\tilde{\pi}$ that are feasible, as follows.

a) A policy $\tilde{\pi}$ starts with an ISB block during which all populations $\{1, ..., k\}$ are sampled at least a predetermined number of times n_0, with a sufficient number of samples taken from cheap (small c^i) populations, so that the constraint of Eq. (3) is satisfied sample path-wise. This block is necessary in order to obtain initial estimates of $\mu_j(\underline{\theta}_j)$ for all populations. This block that the ISB block has the minimum length of $\ell(J)$, defined above, with $J = \{1, ..., k\}$.

b) After a completion of an ISB block a $\tilde{\pi}$ policy chooses any BFS (or equivalently a single population $\{i\}$ or a pair of $\{i, j\}$) and continues sampling for a block of time periods LPB=LPB(b) as follows.

i) When $b = \{i\}$, (which means that $c^i \leq c^0$) $\tilde{\pi}$ samples from population i only once. In this case we define the LPB block to have length equal to: $m_i^b = 1$, and its sampling frequency x_i to be equal to 1, $x_i = 1$.

ii) When $b = \{i, j\}$, $\tilde{\pi}$ samples a number of times each population in $\{i, j\}$ in b so as the cost feasibility of $\tilde{\pi}$ is maintained during the block. The latter is accomplished by taking the length of the LPB block to be equal to: $m_i^b + m_j^b = |\delta^j| + |\delta^i|$, where $m_i^b = |\delta^j|$ and $m_j^b = |\delta^i|$, and sampling the least cost population first in such a way that the frequencies are equal to the randomization probabilities:

$$x_i = \frac{|\delta^j|}{|\delta^i| + |\delta^j|}, \ x_j = \frac{|\delta^i|}{|\delta^i| + |\delta^j|},$$

Remark 1 Note that in the second case of an LPB, the randomization probabilities for $\{i, j\}$, and the block length $m_i^b + m_j^b$, are computed without solving LP-1, using the known, cf. Eq. (9), δ's.

Note that a *block based policy* is a well defined adaptive policy. In the sequel we restrict our attention to *block based policies*; for notational simplicity we will simply write π in place of $\tilde{\pi}$, when there is no risk for confusion.

Assume that we have l successive blocks we take $\widetilde{T}_\pi^b(l)$ to be the number of LPB(b) type blocks in first $l \geq 2$ blocks (since for $l = 1$ we start with an ISB block). Thus $\sum_{b \in K} \widetilde{T}_\pi^b(l) = l - 1$. Let $S_\pi(l)$ be the total length of first l blocks and let $L_n = L_{\tilde{\pi}}(n)$ denote the number of blocks in n periods.

We can easily show that

$$T_\pi^\alpha(S_\pi(l)) = \sum_{b:\alpha\in b} m_\alpha^b \, \widetilde{T}_\pi^b(l) + m_\alpha,$$

where m_α^b is the number of samples from population α between a LPB(b) and m_α is the number of samples from population α in the ISB block. Now we can define the regret of blocks

$$\begin{aligned}\widetilde{R}_\pi(\underline{\underline{\theta}}, l) = E_{\underline{\underline{\theta}}} S_\pi(l) \, z^*(\underline{\underline{\theta}}) - E_{\underline{\underline{\theta}}} \sum_{j=1}^{k} \sum_{b\in K} \mu_j(\underline{\theta}_j) \, m_j^b \, \widetilde{T}_\pi^b(l) \\ - \sum_{j=1}^{k} \mu_j(\underline{\theta}_j) m_j.\end{aligned}$$

We note that

$$T_\pi^\alpha(S_\pi(L_n)) \leq T_\pi^\alpha(n) \leq T_\pi^\alpha(S_\pi(L_n)) + M_\alpha, \tag{9}$$

where M_α is the maximum number of times where population α appears in every block. Thus we obtain the following relation for the two types of regret,

$$\begin{aligned}\widetilde{R}_\pi(\underline{\underline{\theta}}, L_n) + (n - E_{\underline{\underline{\theta}}} S_\pi(L_n)) \, z^*(\underline{\underline{\theta}}) - \sum_{j=1}^{k} M_j \, \mu_j(\underline{\theta}_j) \\ \leq R_\pi(\underline{\underline{\theta}}, n) \leq \widetilde{R}_\pi(\underline{\underline{\theta}}, L_n) + (n - E_{\underline{\underline{\theta}}} S_\pi(L_n)) \, z^*(\underline{\underline{\theta}}).\end{aligned} \tag{10}$$

The above and Eq. (10) imply the following relation between the two regret functions,

$$\limsup_{n\to\infty} \frac{R_\pi(\underline{\underline{\theta}}, n)}{\log n} = \limsup_{n\to\infty} \frac{\widetilde{R}_\pi(\underline{\underline{\theta}}, L_n)}{\log L_n}. \tag{11}$$

From Eq. (11), it follows that if we want to find a policy that achieves the lower bound for $R_\pi(\underline{\underline{\theta}}, n)$ it suffices to find a policy that achieves the lower bound for $\widetilde{R}_\pi(\underline{\underline{\theta}}, L_n)$.

Asymptotically Optimal Policies In this section we provide a general method to construct asymptotically optimal policies π^0 that achieve the lower bound for the regret. To state the policy we need some definitions. We define at any block l and for every population α as $\widetilde{\mu}_\alpha$

$$\widetilde{\mu}_\alpha = \sup_{\underline{\theta}_\alpha'} \{\mu_\alpha(\underline{\theta}_\alpha') : I(\hat{\underline{\theta}}_\alpha^l, \underline{\theta}_\alpha') \leq \frac{\log S_\pi(l-1)}{T_\pi^\alpha(S_\pi(l-1))}\},$$

and as $\Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)}$

$$\Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)} = \{\alpha \ : \mu_\alpha^*(\hat{\underline{\underline{\theta}}}^l) < \widetilde{\mu}_\alpha < \mu_\alpha^*(\hat{\underline{\underline{\theta}}}^l) + \rho(\hat{\underline{\underline{\theta}}}^l, \alpha, \hat{B})\}.$$

We recall that if we have an optimal BFS b, where $b = \{i, j\}$ or $\{i\}$ then the optimal solution is $z^b = \mu_i x_i + \mu_j x_j$ or $z^b = \mu_i$.

INFLATED Z-POLICY π^0:

Start with one ISB block in order to have at least one estimate from each population. Then,
Step 1 Assume that at the beginning of block l, $l > 1$, we have the estimates $\hat{\underline{\underline{\theta}}}^l$, from the previous $l-1$ blocks with $\mu_1(\hat{\underline{\theta}}_1^l), ..., \mu_k(\hat{\underline{\theta}}_k^l)$. We take the solution of LP-1:

$$z^{b(\hat{\underline{\underline{\theta}}}^l)} = \max_{\widetilde{b}_i(\hat{\underline{\underline{\theta}}}^l)} \{z^{\widetilde{b}_i(\hat{\underline{\underline{\theta}}}^l)} : \ \widetilde{T}_\pi^{\widetilde{b}_i(\hat{\underline{\underline{\theta}}}^l)}(l) \geq \tau(l-1)\}$$

where $\widetilde{b}_i$ are all the BFS in K and τ is any fixed constant in: $(0, 1/|K|)$.

Step 2 Then for every $\alpha = \{1, \ldots, k\}$, we compute the $\widetilde{\mu}_\alpha$'s and $\Phi_l^{(\hat{B},\hat{\underline{\underline{\theta}}}^l)}$'s.

Then, if $\Phi_l^{(\hat{B},\hat{\underline{\underline{\theta}}}^l)} = \emptyset$, we take $\pi^0(\hat{\underline{\underline{\theta}}}^l) = b(\hat{\underline{\underline{\theta}}}^l))$, otherwise for every $\alpha \in \Phi_l^{(\hat{B},\hat{\underline{\underline{\theta}}}^l)}$ we define the index:

$$u_\alpha(\hat{\underline{\underline{\theta}}}^l, \underline{\theta}'_\alpha) = \max_{\underline{\theta}'_\alpha}\{z^{b_\alpha(\hat{\underline{\underline{\theta}}}^l, \underline{\theta}'_\alpha)} : I(\hat{\underline{\theta}}^l_\alpha, \underline{\theta}'_\alpha) \leq \frac{\log S_\pi(l-1)}{T_\pi^\alpha(S_\pi(l-1))}\},$$

and we take

$$\pi^0(\hat{\underline{\underline{\theta}}}^l) = \ \arg\max \ \{u_\alpha(\hat{\underline{\underline{\theta}}}^l, \underline{\theta}'_\alpha), \quad \alpha \in \Phi_l^{(\hat{B},\hat{\underline{\underline{\theta}}}^l)}\}$$

Remark 2 a) In Step 1 of our policy we have to compute the values of the objective function for finite number of basic feasible solutions. These computations are not complicated because the LP solution only needs the mean values of the populations at this block and the randomization frequencies which are as we know constants and depend only on which populations we have in the BFS. We recall that if we have a BFS b, where $b = \{i, j\}$ or $\{i\}$ then the optimal solution is $z^b = \mu_i x_i + \mu_j x_j$ or $z^b = \mu_i$. Thus, in order to compute the value of the objective function it is not required to solve the LPs but only to compute and compare the corresponding z^b, using these explicit formulas.

The main result of this paper is that under the following conditions policy π^0 is asymptotically optimal.

To state condition C1 we need the definition of the index $J_\alpha(\underline{\underline{\theta}}, \epsilon)$, of population α for any $\underline{\underline{\theta}} \in \Theta$, $\epsilon > 0$, an optimal matrix B under $\underline{\underline{\theta}}$, and a $\rho(\underline{\underline{\theta}}, \alpha, B)$, as in Lemma 1, we define: $\Theta'_\alpha(\epsilon) = \{\underline{\theta}'_\alpha : \mu^*_\alpha(\underline{\underline{\theta}}) - \epsilon < \mu_\alpha(\underline{\theta}'_\alpha) < \mu^*_\alpha(\underline{\underline{\theta}}) + \rho(\underline{\underline{\theta}}, \alpha, B) - \epsilon\}$ and

$$J_\alpha(\underline{\underline{\theta}}, \epsilon) = \inf_{\underline{\theta}'_\alpha \in \Theta'_\alpha(\epsilon)} \{I(\underline{\theta}_\alpha, \underline{\theta}'_\alpha) : z(\underline{\theta}'_\alpha) > z^*(\underline{\underline{\theta}}) - \epsilon\}.$$

(C1) $\forall\, \underline{\underline{\theta}} \in \Theta,\ i \notin s(\underline{\underline{\theta}})$ such that $\Delta\Theta_i(\underline{\underline{\theta}}) = \emptyset$, if $\mu^*_i(\underline{\underline{\theta}}) - \epsilon < \mu_i(\underline{\theta}'_i) < \mu^*_i(\underline{\underline{\theta}}) + \rho(\underline{\underline{\theta}}, i, B) - \epsilon,\ \forall\, \epsilon > 0$, for some $\underline{\theta}'_i \in \Theta_i$, the following relation holds:

$$\lim_{\epsilon \to 0} J_i(\underline{\underline{\theta}}, \epsilon) = \infty.$$

(C2) $\forall i$, $\forall\, \underline{\theta}_i \in \Theta_i$, $\forall\, \epsilon > 0$,

$$P_{\underline{\theta}_i}(|\hat{\underline{\theta}}^t_i - \underline{\theta}_i| > \epsilon) = o(1/t), \text{ as } t \to \infty.$$

(C3) $\forall\, b_\alpha \in s(\underline{\underline{\theta}})$, $\forall i$, $\forall\, \underline{\theta}_i \in \Theta_i$, $\forall\, \epsilon > 0$, as $t \to \infty$

$$P_{\underline{\underline{\theta}}}(z^{b_\alpha(\hat{\underline{\underline{\theta}}}^j, \underline{\theta}'_\alpha)} \leq z^*(\underline{\underline{\theta}}) - \epsilon, , \text{ for some } j \leq t) = o(1/t).$$

Next, we state and prove the main theorem of the paper.

Theorem 2. Under conditions (C1),(C2), and (C3), and policy π^0, defined above, the following holds.

$$\limsup_{n \to \infty} \frac{R_{\pi^0}(\underline{\underline{\theta}}, n)}{\log n} \leq M(\underline{\underline{\theta}}), \text{ for all } \underline{\underline{\theta}} \in \Theta.$$

Proof

To establish the above inequality it is sufficient to show that for policy π^0 the inequalities below hold.

$$\limsup_{n \to \infty} \frac{E_{\underline{\underline{\theta}}} T^j_{\pi^0}(n)}{\log n} \leq \frac{1}{K_j(\underline{\underline{\theta}})}, \ \forall j \in D(\underline{\underline{\theta}}), \tag{12}$$

$$\limsup_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T^j_{\pi^0}(n)}{\log n} = 0, \ \forall j \notin D(\underline{\underline{\theta}}), \tag{13}$$

$$nc^0 - E_{\underline{\underline{\theta}}} C_{\pi^0}(n) = o(\log n). \tag{14}$$

The proof of these inequalities is given in the appendix.

From the definition of index $J_\alpha(\hat{\underline{\underline{\theta}}}^l, \epsilon)$, where $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)}$,

$$J_\alpha(\hat{\underline{\underline{\theta}}}^l, \epsilon) = \inf_{\underline{\theta}'_\alpha} \{ I(\hat{\underline{\theta}}^l_\alpha, \underline{\theta}'_\alpha) : z^{b_\alpha(\hat{\underline{\underline{\theta}}}^l, \underline{\theta}'_\alpha)} > z^*(\underline{\underline{\theta}}) - \epsilon \},$$

we have that $u_\alpha(\hat{\underline{\underline{\theta}}}^l, \underline{\theta}'_\alpha) > z^*(\underline{\underline{\theta}}) - \epsilon$ if and only if $J_\alpha(\hat{\underline{\underline{\theta}}}^l, \epsilon) < \log S_\pi(l-1)/T^\alpha_\pi(S_\pi(l-1))$.

Remark 3 According to Remark $4b$ in (Burnetas and Katehakis 1996) condition (C2) is equivalent to C2' below which is easier to verify.
(C2') $\forall\ \delta > 0$, as $t \to \infty$

$$\sum_{j=1}^{t-1} P_{\underline{\theta}_i}(b(\hat{\underline{\underline{\theta}}}^j) \in s(\underline{\underline{\theta}}), J_i(\hat{\underline{\underline{\theta}}}^j, \epsilon) \leq J_i(\underline{\underline{\theta}}, \epsilon) - \delta) = o(\log t).$$

Normal Distributions with known variances

Assume the observations X^j_α from population α are normally distributed with unknown means $EX^j_\alpha = \theta_\alpha$ and known variances σ^2_α, i.e., $\underline{\theta}_\alpha = \theta_\alpha$, $\mu_\alpha(\underline{\theta}_\alpha) = \theta_\alpha$, and $\Theta_\alpha = (-\infty, +\infty)$. Given history h_l, define

$$\mu_\alpha(\hat{\theta}^l_\alpha) = \frac{\sum_{j=1}^{T^\alpha_{\pi^0}(S_{\pi^0}(l-1))} X^j_\alpha}{T^\alpha_{\pi^0}(S_{\pi^0}(l-1))}.$$

Now from the definition of Θ_α, it follows that $\Delta\Theta_\alpha(\underline{\theta}) = (\theta_\alpha + \phi^B_\alpha(\underline{\theta}), \theta_\alpha + \phi^B_\alpha(\underline{\theta}) + \rho(\underline{\theta}, \alpha, B))$ for any optimal matrix B under $\underline{\theta}$, therefore $D(\underline{\theta}) = \{1, ..., k\}$, $\forall\ \underline{\theta} \in \Theta$. Thus, we can see from the structure of the sets Θ_α and $\Delta\Theta_\alpha(\underline{\theta})$ that the condition (C1) is satisfied.

Also, we have:

$$I(\theta_\alpha, \theta'_\alpha) = \frac{(\theta'_\alpha - \theta_\alpha)^2}{2\sigma^2_\alpha}$$

$$K_\alpha(\underline{\theta}) = \frac{(\phi^B_\alpha(\underline{\theta}))^2}{2\sigma^2_\alpha}.$$

Therefore our indices are

$$u_\alpha(\hat{\underline{\theta}}^l, \theta^{K_\alpha}_\alpha) = z^{b_\alpha(\hat{\underline{\theta}}^l, \theta^{K_\alpha}_\alpha)},$$

where

$$\theta^{K_\alpha}_\alpha = \hat{\theta}^l_\alpha + \sigma_\alpha \left(\frac{2 \log S_{\pi^0}(l-1)}{T^\alpha_{\pi^0}(S_{\pi^0}(l-1))} \right)^{1/2},$$

For example, if $b_\alpha(\hat{\underline{\theta}}^l, \theta^{K_\alpha}_\alpha) = \{\alpha, j\}$ then $z^{b_\alpha(\hat{\underline{\theta}}^l, \theta^{K_\alpha}_\alpha)} = \theta^{K_\alpha}_\alpha x_\alpha + \hat{\theta}^l_j x_j$ and $z^*(\underline{\theta}) = \theta_\alpha x_\alpha + \theta_j x_j$. Therefore for $b_\alpha(\hat{\underline{\theta}}^l, \theta^{K_\alpha}_\alpha) \in s(\underline{\theta})$ and from the structure of $z^{b_\alpha(\hat{\underline{\theta}}^l, \theta^{K_\alpha}_\alpha)}$ the index is a sum of normal distributions which is also normal or a normal distribution and from the tail of normal distribution condition (C3) is satisfied.

According to Remark 3 the next sum of probabilities is equivalent to the condition (C2)

$$\sum_{t=2}^{L_n} P_{\theta_i}(b(\hat{\underline{\theta}}^t) \in s(\underline{\theta}), J_i(\hat{\underline{\theta}}^t, \epsilon) \leq J_i(\underline{\theta}, \epsilon) - \delta)$$

$$= \sum_{t=2}^{L_n} P_{\theta_i}(b(\hat{\underline{\theta}}^t) \in s(\underline{\theta}), |\hat{\theta}^t_i - \theta_i| > \xi), \xi > 0,$$

where the equality follows after some algebra because of the normal distribution and that we know exactly the I's and consequently the properties of J's

$$J_i(\underline{\hat{\theta}}^t, \epsilon) = \inf_{\theta_i^{'}} \{ I(\hat{\theta}_i^t, \theta_i^{'}) : z^{b_i(\underline{\hat{\theta}}^t, \theta_i^{'})} > z^*(\underline{\theta}) - \epsilon \} \leq$$

$$J_i(\underline{\theta}, \epsilon) = \inf_{\theta_i^{'}} \{ I(\theta_i, \theta_i^{'}) : z^{b_i(\underline{\theta}, \theta_i^{'})} > z^*(\underline{\theta}) - \epsilon \} - \delta.$$

Also, we have that $\hat{\theta}_i^t$ is the average of iid random normal variables with mean θ_i thus

$$\begin{aligned} P_{\theta_i}^{\pi^0}(|\hat{\theta}_i^t - \theta_i| > \xi) &\leq P_{\theta_i}^{\pi^0}(|\hat{\theta}_i^l - \theta_i| > \xi, \text{ for some } l \leq t) \\ &\leq \sum_{l=1}^{t} P_{\theta_i}^{\pi^0}(|\hat{\theta}_i^l - \theta_i| > \xi) = o(1/t), \end{aligned}$$

where the last equality follows from is a consequence of the tail inequality $1 - \Phi(x) < \Phi(x)/x$ for the standard normal distribution. Thus, we can see that the condition (C2) holds.

Summary of Policy At the beginning we take an ISB block. Then at the beginning of block l we the take

$$z^{b(\underline{\underline{\hat{\theta}}}^l)} = \max_{\widetilde{b}_i(\underline{\underline{\hat{\theta}}}^l)} \{ z^{\widetilde{b}_i(\underline{\underline{\hat{\theta}}}^l)} : \; \widetilde{T}_\pi^{\widetilde{b}_i(\underline{\underline{\hat{\theta}}}^l)}(l) \geq \tau(l-1) \}$$

and find our indices

$$u_\alpha(\underline{\hat{\theta}}^l, \theta_\alpha^{K_\alpha}) = z^{b_\alpha(\underline{\hat{\theta}}^l, \theta_\alpha^{K_\alpha})},$$

where

$$\theta_\alpha^{K_\alpha} = \hat{\theta}_\alpha^l + \sigma_\alpha \left(\frac{2 \log S_{\pi^0}(l-1)}{T_{\pi^0}^\alpha(S_{\pi^0}(l-1))} \right)^{1/2}. \tag{15}$$

Finally, we choose to employ as block l the $\arg\max_\alpha \{ u_\alpha(\underline{\hat{\theta}}^l, \underline{\theta}_\alpha^{K_\alpha}) \}$.

Remark 4 In the case in which σ_α are unknown, we expect that a (log - rate regret) f-UF policy can be obtained by replacing σ_α in Eq. 15) by a constant times $\hat{\sigma}_\alpha$, as in Auer et al. (2002). This work is not included due to space limitations.

References

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 Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 207–216. IEEE, 2013.

Peter L 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 N Burnetas and Odysseas A Kanavetas. Adaptive policies for sequential sampling under incomplete information and a cost constraint. *N.J. Daras(ed.), Applications of Mathematics and Informatics in Military Science, Springer*, pages 97–112, 2012.

Apostolos N 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 N 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. Sequential allocation problems with side constraints. In *INFORMS Seattle 1998, Annual Meeting, Seattle WA*, 1998.

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. An asymptotically optimal UCB policy for uniform bandits of unknown support. *arXiv preprint arXiv:1505.01918*, 2015a.

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

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, 2015c.

Wesley Cowan, Junya Honda, and Michael N Katehakis. Asymptotic optimality, finite horizon regret bounds, and a solution to an open problem. *Journal of Machine Learning Research, preprint arXiv:1504.05823*, to appear.

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.

Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *AAAI-13*, pages 232–238, 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, Kevin Glazebrook, and Richard R. Weber. *Multi-armed Bandit Allocation Indices.* John Wiley & Sons, West Sussex, U.K., 2011.

Sudipto Guha and Kamesh Munagala. Approximation algorithms for budgeted learning problems. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 104–113. ACM, 2007.

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.

Kris Johnson, David Simchi-Levi, and He Wang. Online network revenue management using Thompson sampling. *Available at SSRN*, 2015.

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: Adoptive Statistical Procedures and Related Topics*, pages 29–39. Institute of Math. Stats., 1986.

Michael N Katehakis and Herbert Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8584, 1995.

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.

Robert D Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704, 2004.

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.

Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Optimal resource allocation with semi-bandit feedback. *arXiv preprint arXiv:1406.3840*, 2014.

Lihong Li, Remi Munos, and Csaba Szepesvári. 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.

Aditya Mahajan and Demosthenis Teneketzis. Multi-armed bandit problems. In *Foundations and Applications of Sensor Management*, pages 121–151. Springer, 2008.

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

Sandip Sen, Anton Ridgway, and Michael Ripley. Adaptive budgeted bandit algorithms for trust development in a supply-chain. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 137–144. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

Adish Singla and Andreas Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1167–1178. International World Wide Web Conferences Steering Committee, 2013.

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.

Stamatina Thomaidou, Michalis Vazirgiannis, and Kyriakos Liakopoulos. Toward an integrated framework for automated development and optimization of online advertising campaigns. *arXiv preprint arXiv:1208.1187*, 2012.

Long Tran-Thanh, Archie Chapman, Munoz De Cote Flores Luna, Jose Enrique, Alex Rogers, and Nicholas R Jennings. Epsilon–first policies for budget–limited multi-armed bandits. In *AAAI-10*, pages 1211–1216, 2010.

Long Tran-Thanh, Archie Chapman, Munoz De Cote Flores Luna, Jose Enrique, Alex Rogers, and Nicholas R Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *AAAI-12*, pages 1134–1140, 2012.

Long Tran-Thanh, Lampros C Stavrogiannis, Victor Naroditskiy, Valentin Robu, Nicholas R Jennings, and Peter Key. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions, 2014.

Zizhuo Wang, Shiming Deng, and Yinyu Ye. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331, 2014.

Appendix: Proofs

Lemma 1 For any $\underline{\underline{\theta}}$, and any optimal matrix B under $\underline{\underline{\theta}}$, $\exists\ \rho = \rho(\underline{\underline{\theta}}, \alpha, B)$ such that for any $\underline{\theta}^{'}_{\alpha} \in \Delta\Theta_{\alpha}(\underline{\underline{\theta}})$:
(i) $\phi^{B}_{j}(\underline{\underline{\theta}}^{'}) = \phi^{B}_{j}(\underline{\underline{\theta}}) \geq 0,\ \forall\ j \neq \alpha$ and $\phi^{B}_{\alpha}(\underline{\underline{\theta}}^{'}) = \phi^{B}_{\alpha}(\underline{\underline{\theta}}) + \mu_{\alpha}(\underline{\theta}_{\alpha}) - \mu_{\alpha}(\underline{\theta}^{'}_{\alpha}) < 0$,
(ii) $\mu^{*}_{\alpha}(\underline{\underline{\theta}}) < \mu_{\alpha}(\underline{\theta}^{'}_{\alpha}) < \mu^{*}_{\alpha}(\underline{\underline{\theta}}) + \rho$, where $\rho > 0$ and $\mu^{*}_{\alpha}(\underline{\underline{\theta}}) = \phi^{B}_{\alpha}(\underline{\underline{\theta}}) + \mu_{\alpha}(\underline{\theta}_{\alpha})$.

Proof (i) It is obvious that $\phi^{B}_{j}(\underline{\underline{\theta}}^{'}) = \phi^{B}_{j}(\underline{\underline{\theta}}) \geq 0,\ \forall\ j \neq \alpha$ because we only change the parameter of population α and $\phi^{B}_{j}(\underline{\underline{\theta}}^{'}) = \phi^{B}_{j}(\underline{\underline{\theta}}) \equiv c^{j}\lambda^{B} + g^{B} - \mu_{j}(\underline{\theta}_{j})$.

For a population $\alpha \in B(\underline{\underline{\theta}})$ we have that $\alpha \notin b$, for any $b \in s(\underline{\underline{\theta}})$. Therefore $\phi^{B}_{\alpha}(\underline{\underline{\theta}}) \equiv c^{\alpha}\lambda^{B} + g^{B} - \mu_{\alpha}(\underline{\theta}_{\alpha}) > 0$, for any B corresponding to b.

Now, any optimal $b \in s(\underline{\underline{\theta}})$ is not optimal under $\underline{\underline{\theta}}^{'} = (\underline{\theta}_{1}, \ldots, \underline{\theta}^{'}_{\alpha}, \ldots, \underline{\theta}_{k})$, for any $\underline{\theta}^{'}_{\alpha} \in \Delta\Theta_{\alpha}(\underline{\underline{\theta}})$, thus $s(\underline{\underline{\theta}}^{'}) = \{b^{'}\}$ where $b^{'} \notin s(\underline{\underline{\theta}})$.

Therefore, for any optimal matrix B under $\underline{\underline{\theta}}$ we have that $\phi^{B}_{\alpha}(\underline{\underline{\theta}}^{'}) \equiv c^{\alpha}\lambda^{B} + g^{B} - \mu_{\alpha}(\underline{\theta}^{'}_{\alpha}) < 0$ because B is not optimal under $\underline{\underline{\theta}}^{'}$.

Now from $\phi^{B}_{\alpha}(\underline{\underline{\theta}}) = c^{\alpha}\lambda^{B} + g^{B} - \mu_{\alpha}(\underline{\theta}_{\alpha})$ we have that $\phi^{B}_{\alpha}(\underline{\underline{\theta}}^{'}) = \phi^{B}_{\alpha}(\underline{\underline{\theta}}) + \mu_{\alpha}(\underline{\theta}_{\alpha}) - \mu_{\alpha}(\underline{\theta}^{'}_{\alpha}) < 0$.

(ii) Consider first the case that $b = \{i, j\}$ is an optimal solution under $\underline{\underline{\theta}}$ with corresponding optimal matrix $B = B(\underline{\underline{\theta}})$. and $b' = \{i, \alpha\}$ is an optimal solution under $\underline{\underline{\theta}}^{'}$ with corresponding optimal matrix $B' = B(\underline{\underline{\theta}}')$. From $i)$ we have that $z^{*}(\underline{\underline{\theta}}^{'}) > z^{*}(\underline{\underline{\theta}})$ iff $\mu_{\alpha}(\underline{\theta}^{'}_{\alpha}) > \mu^{*}_{\alpha}(\underline{\underline{\theta}})$.

Since b' is uniquely optimal under $\underline{\underline{\theta}}^{'}$ we have that $\phi^{B'}_{s}(\underline{\underline{\theta}}^{'}) > 0$, for any $s \neq i, \alpha$. Now in order for that condition to hold we use that $\phi^{B}_{s}(\underline{\underline{\theta}}) > 0$ for any $s \neq i, j$ and we have that for $s > i$ it suffices that $\mu^{*}_{\alpha}(\underline{\underline{\theta}}) < \mu_{\alpha}(\underline{\theta}^{'}_{\alpha})$, but for $s < i$ we must have $\mu^{*}_{\alpha}(\underline{\underline{\theta}}) < \mu_{\alpha}(\underline{\theta}^{'}_{\alpha}) < \mu^{*}_{\alpha}(\underline{\underline{\theta}}) + \rho$, where ρ is a positive constant. Thus, if $\mu^{*}_{\alpha}(\underline{\underline{\theta}}) < \mu_{\alpha}(\underline{\theta}^{'}_{\alpha}) < \mu^{*}_{\alpha}(\underline{\underline{\theta}}) + \rho$ then $\phi^{B'}_{s}(\underline{\underline{\theta}}^{'}) > 0$ for any s.

The other cases where the population α is a population with cost lower than C_0 and the optimal solution under $\underline{\underline{\theta}}^{'}$ has this form $b' = \{\alpha, j\}$ or $b' = \{\alpha\}$ follow the same arguments as in the previous paragraph.
□

Lemma 2 If b is uniquely optimal BFS and B any optimal matrix under $\underline{\underline{\theta}}$. Then
(i) if $B = \begin{pmatrix} c^{i} & c^{j} \\ 1 & 1 \end{pmatrix}$, for some $i \leq d < j \Rightarrow \lambda^{B} > 0$,
(ii) if $B = \begin{pmatrix} c^{i} & 1 \\ 1 & 0 \end{pmatrix}$, for some $i \leq d \Rightarrow \lambda^{B} = 0$.

Proof *(i)* Let $\underline{\underline{\theta}} : s(\underline{\underline{\theta}}) = \{b\}$, $b = (i, j)$ for $i \leq d < j$, then $\lambda^{B} > 0$ because if $\lambda^{B} = 0$ we must have more than one solutions in the primal, which cannot occur because b is uniquely optimal.
(ii) Let $\underline{\underline{\theta}} : s(\underline{\underline{\theta}}) = \{b\}$, $b = (i)$ for $i \leq d$, then $\lambda^{B} = 0$ from the dual solution and $\phi^{B}_{j}(\underline{\underline{\theta}}) > 0$ for all $j \neq i$.
□

We recall for the next Proposition

$$\begin{aligned} z^{*}(\underline{\underline{\theta}}) \quad = \quad & \max \sum_{j=1}^{k} \mu_{j}(\underline{\theta}_{j}) x_{j} \\ & \sum_{j=1}^{k} c^{j} x_{j} + y = c^{0} \\ & \sum_{j=1}^{k} x_{j} = 1 \\ & x_{j} \geq 0, \forall j,\ y \geq 0, \end{aligned} \tag{16}$$

and that a necessary and sufficient condition for a uniformly good policy π is that for $\underline{\underline{\theta}} \in \Theta$ and any optimal BFS b under $\underline{\underline{\theta}}$,

$$\phi_j^B(\underline{\underline{\theta}}) \lim_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T_\pi^j(n)}{n^a} = 0, \text{ for all } a > 0, \; j \notin b, \tag{17}$$

and also,

$$\lambda^B \lim_{n\to\infty} \frac{\sum_{j\in b}(c^0 - c^j) E_{\underline{\underline{\theta}}} T_\pi^j(n)}{n^a} = 0, \text{ for all } B \text{ corresponds to } b. \tag{18}$$

Proposition 1 For any uniformly good policy π and for all $\underline{\underline{\theta}} \in \Theta$ we have that for $\alpha \in D(\underline{\underline{\theta}})$, any $\underline{\underline{\theta}}^{'} \in \Delta(\underline{\underline{\theta}})$ and for all positive $\beta_n = o(n)$ it is true that

$$P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) < \beta_n] = o(n^{a-1}), \text{ for all } a > 0.$$

Proof Let $\alpha \in D(\underline{\underline{\theta}})$, $\theta_\alpha^{'} \in \Delta\Theta_\alpha(\underline{\underline{\theta}})$, because of $\Delta\Theta_\alpha(\underline{\underline{\theta}})$'s definition we must have a $b^{'}$ which is uniquely optimal under $\underline{\underline{\theta}}^{'}$ $(s(\underline{\underline{\theta}}^{'}) = \{b^{'}\})$ and $\alpha \in b^{'}$. Then we have two cases for the uniquely optimal solution $b^{'}$.

For the first case where $b^{'} = \{\alpha\}$ if $b^{'}$ is nondegenerate then the basic matrix $B^{'} = \begin{pmatrix} c^\alpha & 1 \\ 1 & 0 \end{pmatrix}$ and from Lemma 2 for a uniformly good policy $\lambda^B = 0$ thus,

$$E_{\underline{\underline{\theta}}'} T_\pi^j(n) = o(n^a), \text{ for all } a > 0, \text{ for all } j \notin b^{'}.$$

If $b^{'}$ is degenerate then it must be true that $c^\alpha = c^0$ if we consider any matrix $B^{'} = \begin{pmatrix} c^\alpha & c^j \\ 1 & 1 \end{pmatrix}$ then $\lambda_{B'} > 0$ thus $(c^0 - c^j)E_{\underline{\underline{\theta}}'}T_\pi^j(n) + (c^0 - c^\alpha)E_{\underline{\underline{\theta}}'}T_\pi^\alpha(n) = o(n^a)$ and since $c^0 = c^\alpha$ we have that $E_{\underline{\underline{\theta}}'}T_\pi^j(n) = o(n^a)$ also from Eq. (17) $E_{\underline{\underline{\theta}}'}T_\pi^i(n) = o(n^a)$, for all $i \neq j, \alpha$ thus $E_{\underline{\underline{\theta}}'}T_\pi^j(n) = o(n^a)$, for all $j \neq \alpha$.

Therefore,

$$n - E_{\underline{\underline{\theta}}'}T_\pi^\alpha(n) = o(n^a), \text{ for all } a > 0. \tag{19}$$

It is also true that

$$\begin{aligned} E_{\underline{\underline{\theta}}'}T_\pi^\alpha(n) &= \sum_{k=1}^{n} k\, P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) = k] \\ &= \sum_{k=1}^{\lfloor\beta_n\rfloor} k\, P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) = k] + \sum_{k=\lfloor\beta_n\rfloor+1}^{n} k\, P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) = k] \\ &\leq \beta_n P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) \leq \beta_n] + nP_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) > \beta_n] \\ &= n - (n - \beta_n)P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) \leq \beta_n]. \end{aligned}$$

Therefore

$$n - E_{\underline{\underline{\theta}}'}T_\pi^\alpha(n) \geq (n - \beta_n)P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) \leq \beta_n]. \tag{20}$$

From Eq. (19) and Eq. (20) we obtain

$$(n - \beta_n)P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) \leq \beta_n] = o(n^a), \text{ for all } a > 0,$$

thus

$$P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) \leq \beta_n] = o(n^{a-1}), \text{ for all } a > 0.$$

In the case that $c^{\alpha} > c^0$ and $b^{'} = \{j_0, \alpha\}$ (we do not study the case where $c^{\alpha} < c^0$ because we prove a general result which includes this case via population j_0 which has $c^{j_0} < c^0$) we have from Lemma 2 that for a uniformly good policy $\lambda^B > 0$, thus

$$E_{\underline{\underline{\theta}}'} T_{\pi}^{j}(n) = o(n^a), \ \forall \ a > 0, \ \forall \ j \notin b^{'} = \{j_0, \alpha\} \tag{21}$$

and

$$(c^0 - c^{j_0}) E_{\underline{\underline{\theta}}'} T_{\pi}^{j_0}(n) + (c^0 - c^{\alpha}) E_{\underline{\underline{\theta}}'} T_{\pi}^{\alpha}(n) = o(n^a), \ \forall \ a > 0. \tag{22}$$

If we sum Eq. (21) for all $j \neq \alpha, j_0$ it follows that

$$n - E_{\underline{\underline{\theta}}'} T_{\pi}^{j_0}(n) - E_{\underline{\underline{\theta}}'} T_{\pi}^{\alpha}(n) = \varepsilon_n, \text{ where } \varepsilon_n = o(n^a), \ \forall \ a > 0. \tag{23}$$

Dividing Eq. (22) with $c^{\alpha} - c^{j_0}$ and using Eq. (23), we obtain after some algebra the following two equalities

$$\begin{aligned} n x_{j_0}^{'} - E_{\underline{\underline{\theta}}'} T_{\pi}^{j_0}(n) &= o(n^a), \\ n x_{\alpha}^{'} - E_{\underline{\underline{\theta}}'} T_{\pi}^{\alpha}(n) &= o(n^a), \ \forall \ a > 0. \end{aligned} \tag{24}$$

where $x_{j_0}^{'} = \frac{c^{\alpha} - c^0}{c^{\alpha} - c^{j_0}}$ and $x_{\alpha}^{'} = \frac{c^0 - c^{j_0}}{c^{\alpha} - c^{j_0}}$ are the probabilities which correspond to optimal solution $b^{'}$ of linear program Eq. (16) under $\underline{\underline{\theta}}^{'}$.

For any n let

$$\Gamma_n^{\pi} = \sum_{j \neq \alpha, j_0} T_{\pi}^{j}(n), \text{ and } F_n^{\pi} = \sum_{j \neq \alpha, j_0} (c^0 - c^j) T_{\pi}^{j}(n).$$

Thus, it is obvious that

$$F_n^{\pi} \leq \Gamma_n^{\pi} (c^0 - c^1).$$

Furthermore, from Eq. (23)

$$E_{\underline{\underline{\theta}}'} \Gamma_n^{\pi} = o(n^a), \ \forall \ a > 0. \tag{25}$$

Now, we know that

$$n c^0 - C_{\pi}(n) = F_n^{\pi} + (c^0 - c^{\alpha}) T_{\pi}^{\alpha}(n) + (c^0 - c^{j_0}) T_{\pi}^{j_0}(n),$$

and from $n c^0 - C_{\pi}(n) \geq 0, \ \forall \ n$, we have that

$$(c^{\alpha} - c^0) T_{\pi}^{\alpha}(n) \leq F_n^{\pi} + (c^0 - c^{j_0}) T_{\pi}^{j_0}(n),$$

therefore

$$\begin{aligned}
\frac{c^\alpha - c^0}{c^\alpha - c^{j_0}} T_\pi^\alpha(n) &\le \frac{F_n^\pi}{c^\alpha - c^{j_0}} + \frac{c^0 - c^{j_0}}{c^\alpha - c^{j_0}} T_\pi^{j_0}(n) \\
x_{j_0}^{'} T_\pi^\alpha(n) &\le \frac{F_n^\pi}{c^\alpha - c^{j_0}} + x_\alpha^{'} T_\pi^{j_0}(n) \\
(1 - x_\alpha^{'}) T_\pi^\alpha(n) &\le \frac{F_n^\pi}{c^\alpha - c^{j_0}} + x_\alpha^{'} T_\pi^{j_0}(n) \\
T_\pi^\alpha(n) &\le \frac{F_n^\pi}{c^\alpha - c^{j_0}} + x_\alpha^{'} (T_\pi^\alpha(n) + T_\pi^{j_0}(n)) \\
T_\pi^\alpha(n) &\le \frac{F_n^\pi}{c^\alpha - c^{j_0}} + x_\alpha^{'} (n - \Gamma_n^\pi) \\
T_\pi^\alpha(n) &\le n x_\alpha^{'} + \frac{F_n^\pi}{c^\alpha - c^{j_0}} - x_\alpha^{'} \Gamma_n^\pi,
\end{aligned}$$

and we recall $F_n^\pi \le \Gamma_n^\pi (c^0 - c^1)$, thus

$$\begin{aligned}
T_\pi^\alpha(n) &\le n x_\alpha^{'} + \frac{\Gamma_n^\pi (c^0 - c^1)}{c^\alpha - c^{j_0}} - x_\alpha^{'} \Gamma_n \\
T_\pi^\alpha(n) &\le n x_\alpha^{'} + \Gamma_n^\pi \rho(j_0, \alpha)
\end{aligned}$$

where $\rho(j_0, \alpha) = \frac{c^{j_0} - c^1}{c^\alpha - c^{j_0}} \ge 0$.

Finally,

$$n x_\alpha^{'} - T_\pi^\alpha(n) + \Gamma_n^\pi \rho(j_0, \alpha) \ge 0. \tag{26}$$

Thus, from Markov inequality, for any positive $\beta_n = o(n)$

$$\begin{aligned}
&P_{\underline{\underline{\theta}}'}(n x_\alpha^{'} - T_\pi^\alpha(n) + \Gamma_n^\pi \rho(j_0, \alpha) \ge n x_\alpha^{'} - \beta_n) \\
&\le \tfrac{E_{\underline{\underline{\theta}}'}(n x_\alpha^{'} - T_\pi^\alpha(n) + \Gamma_n^\pi \rho(j_0, \alpha))}{n x_\alpha^{'} - \beta_n} \\
&= \tfrac{o(n^a)}{n x_\alpha^{'} - \beta_n} = o(n^{a-1}),\ \forall\ a > 0.
\end{aligned}$$

Therefore

$$P_{\underline{\underline{\theta}}'}(T_\pi^\alpha(n) \le \beta_n) \le P_{\underline{\underline{\theta}}'}(T_\pi^\alpha(n) \le \beta_n + \Gamma_n^\pi \rho(j_0, \alpha)) = o(n^{a-1}), \forall\ a > 0.$$

Substituting $T_\pi^\alpha(n) = n - \Gamma_n^\pi - T_\pi^{j_0}(n)$ into Eq. (26) we have

$$T_\pi^{j_0}(n) - n x_{j_0}^{'} + (1 + \rho(j_0, \alpha)) \Gamma_n^\pi \ge 0,$$

then

$$P_{\underline{\underline{\theta}}'}(T_\pi^{j_0}(n) \le \beta_n) = P_{\underline{\underline{\theta}}'}(Z_n^\pi \le \beta_n - n x_{j_0}^{'} + (1 + \rho(j_0, \alpha)) \Gamma_n^\pi),$$

where

$$Z_n^\pi = T_\pi^{j_0}(n) - n x_{j_0}^{'} + (1 + \rho(j_0, \alpha)) \Gamma_n^\pi \ge 0,$$

and

$$E_{\underline{\underline{\theta}}'} Z_n^\pi = o(n^a),\ \forall\ a > 0 \text{ from } Eq.\,(24) \text{ and } Eq.\,(25).$$

Let,

$$V_n^\pi = \{Z_n^\pi \leq \beta_n - nx_{j_0}^{'} + (1+\rho(j_0,\alpha))\Gamma_n^\pi\},\ \text{then}$$

$$\begin{aligned} P_{\underline{\underline{\theta}}'}(V_n^\pi) &= P_{\underline{\underline{\theta}}'}(V_n^\pi \cap \{\Gamma_n^\pi \leq n\delta\}) + P_{\underline{\underline{\theta}}'}(V_n^\pi \cap \{\Gamma_n^\pi > n\delta\}) \\ &\leq P_{\underline{\underline{\theta}}'}(V_n^\pi \cap \{\Gamma_n^\pi \leq n\delta\}) + P_{\underline{\underline{\theta}}'}(\Gamma_n^\pi > n\delta) \end{aligned} \tag{27}$$

where $0 < \delta < \frac{x_{j_0}^{'}}{1+\rho(j_0,\alpha)}$ and using Eq. (25) we have that

$$\begin{aligned} P_{\underline{\underline{\theta}}'}(\Gamma_n^\pi > n\delta) &\leq \frac{E_{\underline{\underline{\theta}}'}\Gamma_n^\pi}{n\delta} \\ &= \frac{o(n^a)}{n\delta} = o(n^{a-1}),\ \forall\ a > 0. \end{aligned} \tag{28}$$

Let,

$$\begin{aligned} G_n^\pi &= \{V_n^\pi \cap \{\Gamma_n^\pi \leq n\delta\}\} \\ &= \{Z_n^\pi \leq \beta_n - nx_{j_0}^{'} + (1+\rho(j_0,\alpha))\Gamma_n^\pi \text{ and } \Gamma_n^\pi \leq n\delta\} \\ &\subseteq \{Z_n^\pi \leq \beta_n + [(1+\rho(j_0,\alpha))\delta - x_{j_0}^{'}]n\}, \\ &= \{Z_n^\pi \leq \beta_n - \varphi n\}, \end{aligned}$$

where

$$\varphi = x_{j_0}^{'} - (1+\rho(j_0,\alpha))\delta > x_{j_0}^{'} - (1+\rho(j_0,\alpha))\frac{x_{j_0}^{'}}{1+\rho(j_0,\alpha)} = 0.$$

Now for any positive $\beta_n = o(n)$,

$$\exists\ n_0 :\ \beta_n - n\varphi < 0,\ \forall\ n > n_0$$

and we have that

$$P_{\underline{\underline{\theta}}'}(G_n^\pi) = 0, \forall\ n > n_0(\varphi),$$

thus from Eq. (27),Eq. (28)

$$P_{\underline{\underline{\theta}}'}(V_n^\pi) \leq o(n^{a-1}),\ \forall\ a > 0.$$

Finally,

$$P_{\underline{\underline{\theta}}'}(T_\pi^{j_0}(n) \leq \beta_n) = o(n^{a-1}),\ \forall\ a > 0,\ \text{for any positive } \beta_n = o(n).$$

So far we have shown that a necessary condition for a uniformly good policy is that $\forall\ \underline{\underline{\theta}} \in \underline{\Theta}$, and $\forall\ \alpha \in D(\underline{\underline{\theta}})$ it must be true that the number of samples from populations j_0 and α are at least β_n correspondingly, because $P_{\underline{\underline{\theta}}'}(T_\pi^{j_0}(n) \leq \beta_n) = o(n^{a-1})$, $P_{\underline{\underline{\theta}}'}(T_\pi^{\alpha}(n) \leq \beta_n) = o(n^{a-1})$ for any positive sequence of constants $\beta_n = o(n)$.
□

Lemma 3 If $P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) < \beta_n] = o(n^{a-1})$, for all $a > 0$ and positive $\beta_n = o(n)$ then

$$\lim_{n\to\infty} P_{\underline{\underline{\theta}}}[T_\pi^\alpha(n) < \frac{\log n}{K_\alpha(\underline{\underline{\theta}})}] = 0,$$

for all $\underline{\underline{\theta}} \in \Theta$ and $\alpha \in \Delta(\underline{\underline{\theta}})$.

Proof If we take $\beta_n = \frac{\log n}{K_\alpha(\underline{\underline{\theta}})}$then $P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) < \frac{\log n}{K_\alpha(\underline{\underline{\theta}})}] = o(n^{a-1})$ and using a change of measure from $\underline{\underline{\theta}}^{'}$ to $\underline{\underline{\theta}}$ and following the arguments in Burnetas and Katehakis (1996b); Lai and Robbins (1985) we have that

$$\lim_{n\to\infty} P_{\underline{\underline{\theta}}}[T_\pi^\alpha(n) < \frac{\log n}{K_\alpha(\underline{\underline{\theta}})}] = 0.$$

□

We recall for Theorem 2 that

$$1. \quad \limsup_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T_\pi^j(n)}{\log n} \le \frac{1}{K_j(\underline{\underline{\theta}})}, \text{ for all } j \in D(\underline{\underline{\theta}}), \tag{29}$$

$$2. \quad \limsup_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T_\pi^j(n)}{\log n} = 0, \text{ for all } j \notin D(\underline{\underline{\theta}}), \tag{30}$$

$$3. \quad nc^0 - E_{\underline{\underline{\theta}}} C_\pi(n) = o(\log n). \tag{31}$$

From the definition of $T_\pi^\alpha(n)$ we can see that

$$T_\pi^\alpha(S_\pi(L_n)) \le T_\pi^\alpha(n) \le T_\pi^\alpha(S_\pi(L_n)) + M_\alpha, \tag{32}$$

where M_α is the maximum number of times where population α appears in every block.

We derived $\widetilde{T}_\pi^b(L_n)$ as below

$$\begin{aligned} \widetilde{T}_\pi^b(L_n) &= \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^t) \notin s(\underline{\underline{\theta}})\} + \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^t) \in s(\underline{\underline{\theta}})\} \\ &\le \sum_{t=2}^{L_n} 1\{b(\hat{\underline{\underline{\theta}}}^t) \notin s(\underline{\underline{\theta}})\} + \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^t) \in s(\underline{\underline{\theta}})\} \end{aligned} \tag{33}$$

Finally, a policy π is called feasible if

$$\frac{C_\pi(n)}{n} \le c^0, \ \forall \ n = 1, 2, \ldots. \tag{34}$$

Theorem 2 Let policy π^0, under conditions (C1),(C2), and (C3) then

$$\limsup_{n\to\infty} \frac{R_{\pi^0}(\underline{\underline{\theta}}, n)}{\log n} \le M(\underline{\underline{\theta}}), \text{ for all } \underline{\underline{\theta}} \in \Theta.$$

Proof We need to prove Eq. (29), Eq. (30) and Eq. (31). From the Eq. (32), Eq. (33) and Lemmas 4 and 5 we have proved the relations Eq. (29) and Eq. (30). Equation Eq. (31) follows from Eq. (34) the feasibility of π^0 and block policies.

□

Lemma 4 Let policy π^0, under conditions (C1),(C2)

$$\limsup_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} \widetilde{T}_{\pi^0,2}^b(L_n)}{\log L_n} \le \frac{1}{K_i(\underline{\underline{\theta}})}, \text{ for all } i \in D(\underline{\underline{\theta}}), i \in b, b \notin s(\underline{\underline{\theta}}) \text{ and}$$

$$\limsup_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} \widetilde{T}_{\pi^0,2}^b(L_n)}{\log L_n} = 0, \text{ for all } i \notin D(\underline{\underline{\theta}}), i \in b, b \in s(\underline{\underline{\theta}}).$$

Proof We can divide the sum $\widetilde{T}^{b}_{\pi^0,2}(L_n)$ as follows

$$\begin{aligned}
\widetilde{T}^{b}_{\pi^0,2}(L_n) &= \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^{t}) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^{t})\} \\
&= \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^{t}) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^{t}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) > z^*(\underline{\underline{\theta}}) - \epsilon\} \\
&\qquad + \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^{t}) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^{t}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) \leq z^*(\underline{\underline{\theta}}) - \epsilon\}.
\end{aligned}$$

From the relation between the two indices u_i and J_i we have that

$$\begin{aligned}
&\sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^{t}) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^{t}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) > z^*(\underline{\underline{\theta}}) - \epsilon\} \\
&\leq \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^{t}) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^{t}), J_i(\hat{\underline{\underline{\theta}}}^{t}, \epsilon) < \frac{\log S_{\pi^0}(t-1)}{T^i_{\pi^0}(S_{\pi^0}(t-1))}\} \\
&= \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^{t}) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^{t}), \\
&\qquad J_i(\hat{\underline{\underline{\theta}}}^{t}, \epsilon) < \frac{\log S_{\pi^0}(t-1)}{T^i_{\pi^0}(S_{\pi^0}(t-1))}, J_i(\hat{\underline{\underline{\theta}}}^{t}, \epsilon) > J_i(\underline{\underline{\theta}}, \epsilon) - \delta\} \\
&\quad + \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^{t}) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^{t}), \\
&\qquad J_i(\hat{\underline{\underline{\theta}}}^{t}, \epsilon) < \frac{\log S_{\pi^0}(t-1)}{T^i_{\pi^0}(S_{\pi^0}(t-1))}, J_i(\hat{\underline{\underline{\theta}}}^{t}, \epsilon) \leq J_i(\underline{\underline{\theta}}, \epsilon) - \delta\} \\
&\leq \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^{t}) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^{t}), T^i_{\pi^0}(S_{\pi^0}(t-1)) < \frac{\log L_n}{J_i(\underline{\underline{\theta}}, \epsilon) - \delta}\} \\
&\quad + \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^{t}) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^{t}, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^{t}), J_i(\hat{\underline{\underline{\theta}}}^{t}, \epsilon) \leq J_i(\underline{\underline{\theta}}, \epsilon) - \delta\}.
\end{aligned}$$

Now, the first sum of the last inequality for $c = \frac{\log L_n}{J_i(\underline{\underline{\theta}},\epsilon)-\delta}$ and s integer is equal to

$$\begin{aligned}
& \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^t) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t), T^i_{\pi^0}(S_{\pi^0}(t-1)) < c\} \\
\leq \; & \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, T^i_{\pi^0}(S_{\pi^0}(t-1)) < c\} \\
= \; & \sum_{t=2}^{L_n} \sum_{s=0}^{\lfloor c/m_i^b \rfloor} 1\{\pi_t^0 = b, T^i_{\pi^0}(S_{\pi^0}(t-1)) = s\, m_i^b + m_i\} \\
= \; & \sum_{s=0}^{\lfloor c/m_i^b \rfloor} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, T^i_{\pi^0}(S_{\pi^0}(t-1)) = s\, m_i^b + m_i\} \\
\leq \; & \lfloor c/m_i^b \rfloor + 1 \\
\leq \; & \frac{c}{m_i^b} + 1 = \frac{\log L_n}{m_i^b(J_i(\underline{\underline{\theta}},\epsilon) - \delta)} + 1.
\end{aligned}$$

Thus,

$$\begin{aligned}
& E_{\underline{\underline{\theta}}} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^t) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t), T^i_{\pi^0}(S_{\pi^0}(t-1)) < \frac{\log L_n}{J_i(\underline{\underline{\theta}},\epsilon) - \delta}\} \\
& \leq \frac{\log L_n}{m_i^b(J_i(\underline{\underline{\theta}},\epsilon) - \delta)} + 1. \qquad (35)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
& \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^t) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t), J_i(\hat{\underline{\underline{\theta}}}^t, \epsilon) \leq J_i(\underline{\underline{\theta}},\epsilon) - \delta\} \\
& \leq \sum_{t=2}^{L_n} 1\{b(\hat{\underline{\underline{\theta}}}^t) \in s(\underline{\underline{\theta}}), J_i(\hat{\underline{\underline{\theta}}}^t, \epsilon) \leq J_i(\underline{\underline{\theta}},\epsilon) - \delta\}
\end{aligned}$$

Then from (C2) and Remark 3 we have that

$$\begin{aligned}
& E_{\underline{\underline{\theta}}} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^t) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t), J_i(\hat{\underline{\underline{\theta}}}^t, \epsilon) \leq J_i(\underline{\underline{\theta}},\epsilon) - \delta\} \\
& \leq o(\log L_n). \qquad (36)
\end{aligned}$$

Now we have that $u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t) > u_s(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_s^{'})$ for any population s which is contained in an optimal BFS of $\underline{\underline{\theta}}$. Now let $b(\hat{\underline{\underline{\theta}}}^t) = (r, s)$ and obviously $b = (i, s)$,thus we can show the following inequalities

$$\begin{aligned}
&\sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^t) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i^{'}) \le z^*(\underline{\underline{\theta}}) - \epsilon\} \\
&\le \sum_{t=2}^{L_n} 1\{u_s(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_s^{'}) \le z^*(\underline{\underline{\theta}}) - \epsilon\} \\
&\le \sum_{t=2}^{L_n} 1\{u_s(\hat{\underline{\underline{\theta}}}^j, \underline{\theta}_s^{'}) \le z^*(\underline{\underline{\theta}}) - \epsilon, \text{ for some } j \le S_{\pi^0}(t-1)\} \\
&= \sum_{t=2}^{L_n} 1\{|\hat{\underline{\theta}}_s^j - \underline{\theta}_s| > \xi, \text{ for some } j \le S_{\pi^0}(t-1)\}.
\end{aligned}$$

Thus

$$\begin{aligned}
&E_{\underline{\underline{\theta}}} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b(\hat{\underline{\underline{\theta}}}^t) \in s(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i^{'}) \le z^*(\underline{\underline{\theta}}) - \epsilon\} \\
&\le o(\log L_n), \qquad (37)
\end{aligned}$$

because

$$\begin{aligned}
&P_{\underline{\theta}_s}^{\pi^0}(|\hat{\underline{\theta}}_s^j - \underline{\theta}_s| > \xi, \text{ for some } j \le t) \\
&\le \sum_{j=1}^{t} P_{\underline{\theta}_s}^{\pi^0}(|\hat{\underline{\theta}}_s^j - \underline{\theta}_s| > \xi) = o(1/t),
\end{aligned}$$

since policy π^0 at any block t chooses $b(\hat{\underline{\underline{\theta}}}^t) = (r,s)$ when $\widetilde{T}_{\pi^0}^{b(\hat{\underline{\underline{\theta}}}^t)}(t) \ge \tau(t-1)$.

Finally, it follows from Eq. (35), Eq. (36) and Eq. (37) that

$$E_{\underline{\underline{\theta}}} \widetilde{T}_{\pi^0}^b(L_n) \le \frac{\log L_n}{m_i^b(J_i(\underline{\underline{\theta}}, \epsilon) - \delta)} + 1 + o(\log L_n) + o(\log L_n).$$

Now from the definition of $J_i(\underline{\underline{\theta}}, \epsilon)$ and (C1) we have that

$$\lim_{\epsilon \to 0} J_i(\underline{\underline{\theta}}, \epsilon) = K_i(\underline{\underline{\theta}}), \text{ for } i \in D(\underline{\underline{\theta}}) \text{ and } \lim_{\epsilon \to 0} J_i(\underline{\underline{\theta}}, \epsilon) = \infty, \text{ for } i \notin D(\underline{\underline{\theta}}).$$

Thus

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} \widetilde{T}_{\pi^0,2}^b(L_n)}{\log L_n} \le \frac{1}{K_i(\underline{\underline{\theta}})}, \text{ for all } i \in D(\underline{\underline{\theta}}), i \in b, b \notin s(\underline{\underline{\theta}}) \text{ and} \\
&\limsup_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} \widetilde{T}_{\pi^0,2}^b(L_n)}{\log L_n} = 0, \text{ for all } i \notin D(\underline{\underline{\theta}}), i \in b, b \in s(\underline{\underline{\theta}}).
\end{aligned}$$

□

For the next Lemma, let $0 < \varepsilon < \{z^*(\underline{\underline{\theta}}) - \max_{b \notin s(\underline{\underline{\theta}})} z^{b(\underline{\underline{\theta}})}\}/2$ and c a positive integer, then we define for $r = 0, 1, 2, \ldots$

$$A_r = \bigcap_{1 \le j \le |K|} \{\max_{\tau c^{r-1} \le l-1 \le c^{r+1}} |z^{\widetilde{b}_j(\hat{\underline{\underline{\theta}}}^l)} - z^{\widetilde{b}_j(\underline{\underline{\theta}})}| \le \varepsilon\} \text{ and}$$

$$B_r = \bigcap_{b_\alpha \in s(\underline{\underline{\theta}})} \{z^{b_\alpha(\hat{\underline{\underline{\theta}}}^i, \underline{\theta}'_\alpha)} \geq z^*(\underline{\underline{\theta}}) - \varepsilon,\ \text{for all } 1 \leq i \leq \tau(l-1) \text{ and } c^{r-1} \leq l-1 \leq c^{r+1}\},$$

where $0 < \tau < 1/|K|$ is the same as in the π^0.

Lemma 5 Under conditions (C2),(C3)
(i) $P_{\underline{\underline{\theta}}}^{\pi^0}(\overline{A}_r) = o(c^{-r})$, $P_{\underline{\underline{\theta}}}^{\pi^0}(\overline{B}_r) = o(c^{-r})$.
Moreover, if $c > 1/(1-|K|\tau)$ and $r \geq r_0$ then
(ii) on $A_r \cap B_r$, $b(\hat{\underline{\underline{\theta}}}^l) \in s(\underline{\underline{\theta}})$ for all $c^{r-1} \leq l-1 \leq c^{r+1}$.
(iii) $E_{\underline{\underline{\theta}}} \widetilde{T}^b_{\pi^0,1}(L_n) = \sum_{t=2}^{L_n} P_{\underline{\underline{\theta}}}^{\pi^0}(b(\hat{\underline{\underline{\theta}}}^t) \notin s(\underline{\underline{\theta}})) = o(\log L_n)$.

Proof (i) We have that from (C2)

$$P_{\underline{\underline{\theta}}}^{\pi^0}(\max_{\tau c^{r-1} \leq l-1 \leq c^{r+1}} |z^{\widetilde{b}_j(\hat{\underline{\underline{\theta}}}^l)} - z^{\widetilde{b}_j(\underline{\underline{\theta}})}| > \varepsilon) = o(c^{-r}),\ 1 \leq j \leq |K|$$

holds for the sample mean of the estimates $\hat{\underline{\underline{\theta}}}^l = \frac{\hat{\underline{\underline{\theta}}}^1 + \ldots + \hat{\underline{\underline{\theta}}}^l}{l-1}$ thus it follows that $P_{\underline{\underline{\theta}}}^{\pi^0}(\overline{A}_r) = o(c^{-r})$.

Now let q be the smallest positive integer such that $\lfloor c^{r-1}/\tau^q \rfloor \geq c^{r+1}$. For $t = 0, ..., q$ and $l_t = \lfloor c^{r-1}/\tau^t \rfloor$ we define the sets

$$Q_t = \bigcap_{b_\alpha \in s(\underline{\underline{\theta}})} \left\{ z^{b_\alpha(\hat{\underline{\underline{\theta}}}^i, \underline{\theta}'_\alpha)} \geq z^*(\underline{\underline{\theta}}) - \varepsilon,\ \text{for all } 1 \leq i \leq l_t \right\}.$$

Then by (C3),

$$P_{\underline{\underline{\theta}}}^{\pi^0}(\overline{Q}_t) = o(1/l_t) = o(c^{-r}) \text{ for } t = 0, ..., q. \tag{38}$$

Now given that $c^{r-1} \leq l-1 \leq c^{r+1}$ and $1 \leq i \leq \tau(l-1)$, there exists $t \in \{0, ..., q\}$ such that $l_{t+1} > l-1 \geq l_t \geq i$ and therefore for every fix b_α we have that

$$z^{b_\alpha(\hat{\underline{\underline{\theta}}}^l, \underline{\theta}'_\alpha)} \geq z^{b_\alpha(\hat{\underline{\underline{\theta}}}^{l_t}, \underline{\theta}'_\alpha)} \geq z^*(\underline{\underline{\theta}}) - \varepsilon.$$

for every $b_\alpha \in s(\underline{\underline{\theta}})$ on the event $\bigcap_{0 \leq t \leq q} Q_t$. Thus, because of $B_r \supset \bigcap_{0 \leq t \leq q} Q_t$ and Eq. (38) we have that $P_{\underline{\underline{\theta}}}^{\pi^0}(\overline{B}_r) = o(c^{-r})$.

(ii) Let $V^{\pi^0}_{s(\underline{\underline{\theta}})}(l) = \sum_{b \in s(\underline{\underline{\theta}})} \widetilde{T}^b_{\pi^0}(l)$ be the number of times that π^0 samples from $s(\underline{\underline{\theta}})$ up to l sampling block.

We note that

$$\max_{b \in s(\underline{\underline{\theta}})} \widetilde{T}^b_{\pi^0}(l) \geq \frac{V^{\pi^0}_{s(\underline{\underline{\theta}})}(l)}{\# s(\underline{\underline{\theta}})}. \tag{39}$$

Consider that at any block l and $c^{r-1} \leq l-1 \leq c^{r+1}$, we have that $u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^l) \in s(\underline{\underline{\theta}})$, and $u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^l)$ corresponds to an optimal BFS $b_{\alpha^*}(\hat{\underline{\underline{\theta}}}^l)$. Then if $b(\hat{\underline{\underline{\theta}}}^l) \in s(\underline{\underline{\theta}})$ we have the requested. Now, let assume that $b(\hat{\underline{\underline{\theta}}}^l) \notin s(\underline{\underline{\theta}})$, and we have that $b_{\alpha^*}(\hat{\underline{\underline{\theta}}}^l) \in s(\underline{\underline{\theta}})$ which means that on $A_r \cap B_r$ the policy π^0 chooses from $s(\underline{\underline{\theta}})$.

Then since $\widetilde{T}^{b(\hat{\underline{\underline{\theta}}}^l)}_{\pi^0}(l) \geq \tau(l-1)$,

$$z^{b(\hat{\underline{\underline{\theta}}}^l)} \leq \max_{b \notin s(\underline{\underline{\theta}})} z^{b(\underline{\underline{\theta}})} + \varepsilon < z^*(\underline{\underline{\theta}}) - \varepsilon \text{ on } A_r.$$

In the case where $\widetilde{T}_{\pi^0}^{b_{\alpha^*}(\hat{\underline{\underline{\theta}}}^l)}(l) \geq \tau(l-1)$, we have on the event A_r

$$z^*(\underline{\underline{\theta}}) - \varepsilon \leq z^{b_{\alpha^*}(\hat{\underline{\underline{\theta}}}^l)}.$$

In the other case where $\widetilde{T}_{\pi^0}^{b_{\alpha^*}(\hat{\underline{\underline{\theta}}}^l)}(l) < \tau(l-1)$, we have on the event B_r

$$z^*(\underline{\underline{\theta}}) - \varepsilon \leq z^{b_{\alpha^*}(\hat{\underline{\underline{\theta}}}^l)}.$$

On the event $A_r \cap B_r$, since π^0 employs from $s(\underline{\underline{\theta}})$ at block l and $c^{r-1} \leq l-1 \leq c^{r+1}$, and since $c > 1/(1-|K|\tau)$ it follows that

$$V_{s(\underline{\underline{\theta}})}^{\pi^0}(l) \geq \frac{\#s(\underline{\underline{\theta}})}{|K|}(l-1-c^{r-1}-2|K|) > (\#s(\underline{\underline{\theta}}))\tau(l-1) \tag{40}$$

for all $c^{r-1} \leq l-1 \leq c^{r+1}$ and $r \geq r_0$.

From Eq. (39) and Eq. (40), we obtain on $A_r \cap B_r$

$$\max_{b \in s(\underline{\underline{\theta}})} \widetilde{T}_{\pi^0}^b(l) > \tau(l-1) \tag{41}$$

for all $c^{r-1} \leq l-1 \leq c^{r+1}$ if $r \geq r_0$.

We note that for $r \geq r_0$ and $c^{r-1} \leq l-1 \leq c^{r+1}$, on the event $A_r \cap B_r$,

$$\begin{aligned}
&\max\{z^b : \widetilde{T}_{\pi^0}^b(l) \geq \tau(l-1) \text{ and } b \notin s(\underline{\underline{\theta}})\} \\
&\leq \max_{b \notin s(\underline{\underline{\theta}})} z^b + \varepsilon < z^*(\underline{\underline{\theta}}) - \varepsilon \\
&\leq \min\{z^b : \widetilde{T}_{\pi^0}^b(l) \geq \tau(l-1) \text{ and } b \in s(\underline{\underline{\theta}})\}
\end{aligned}$$

the last set is nonempty because of Eq. (41). Hence $b(\hat{\underline{\underline{\theta}}}^l) \in s(\underline{\underline{\theta}})$ for all $c^{r-1} \leq l-1 \leq c^{r+1}$ on the event $A_r \cap B_r$ if $r \geq r_0$.

(iii) Let $c > 1/(1-|K|\tau)$. Then it follows from (i) and (ii) that for $r \geq r_0$ and $c^{r-1} \leq t-1 \leq c^{r+1}$,

$$P_{\underline{\underline{\theta}}}^{\pi^0}(b(\hat{\underline{\underline{\theta}}}^t) \notin s(\underline{\underline{\theta}})) \leq P_{\underline{\underline{\theta}}}^{\pi^0}(\overline{A}_r) + P_{\underline{\underline{\theta}}}^{\pi^0}(\overline{B}_r) = o(c^{-r})$$

and therefore

$$\sum_{c^{r-1} \leq t-1 \leq c^{r+1}} P_{\underline{\underline{\theta}}}^{\pi^0}(b(\hat{\underline{\underline{\theta}}}^t) \notin s(\underline{\underline{\theta}})) = o(1).$$

Hence,

$$\sum_{t=2}^{L_n} P_{\underline{\underline{\theta}}}^{\pi^0}(b(\hat{\underline{\underline{\theta}}}^t) \notin s(\underline{\underline{\theta}})) = o(\log L_n).$$

□