Probability in the Engineering and Informational Sciences, **31**, 2017, 284–310.
doi:10.1017/S026996481600036X

ASYMPTOTICALLY OPTIMAL MULTI-ARMED BANDIT POLICIES UNDER A COST CONSTRAINT

APOSTOLOS BURNETAS
Department of Mathematics,
University of Athens,
Athens, Greece
E-mail: aburnetas@math.uoa.gr

ODYSSEAS KANAVETAS
Faculty of Engineering and Natural Sciences,
Sabanci University,
Istanbul, Turkey
E-mail: okanavetas@sabanciuniv.edu

MICHAEL N. KATEHAKIS
Department of Management Science and Information Systems,
Rutgers University,
NJ, USA
E-mail: mnk@rutgers.edu

We consider the multi-armed bandit problem under a cost constraint. Successive samples from each population are i.i.d. with unknown distribution and each sample incurs a known population-dependent cost. The objective is to design an adaptive sampling policy to maximize the expected sum of n samples such that the average cost does not exceed a given bound sample-path wise. We establish an asymptotic lower bound for the regret of feasible uniformly fast convergent policies, and construct a class of policies, which achieve the bound. We also provide their explicit form under Normal distributions with unknown means and known variances.

Keywords: applied probability, stochastic modelling

1. INTRODUCTION: MODEL DESCRIPTION

Consider the problem of sequential sampling from a finite number of independent statistical populations, where successive samples from a population are i.i.d. random variables with unknown distribution. Specifically, we 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 non-degenerate measure v. The density $f_i(\,|\,)$ is known and $\underline{\theta}_i$ is a parameter belonging to some set Θ_i. Let $\underline{\theta} = (\underline{\theta}_1, \ldots, \underline{\theta}_k)$ denote the set of parameters, $\underline{\theta} \in \Theta$, where $\Theta \equiv \Theta_1 \times \cdots \times \Theta_k$. Given $\underline{\theta}$ let $\underline{\mu}(\underline{\theta}) = (\mu_1(\underline{\theta}_1), \ldots, \mu_k(\underline{\theta}_k))$ be

© Cambridge University Press 2016 0269-9648/16 $25.00

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

the vector of expected values, that is, $\mu_i(\underline{\theta}_i) = E_{\underline{\theta}}(X^i)$. The true value $\underline{\theta}_0$ of $\underline{\theta}$ is unknown. We make the assumption that outcomes from different populations are independent.

Sampling from population Π_i incurs a positive cost c^i per sample, and without loss of generality we assume $c^1 \le c^2 \le \cdots \le c^k$, and not all c^i are equal. The objective is to maximize the expected average reward per period subject to the constraint that the average sampling cost per period does not exceed a given upper bound c^0 for each period. Without loss of generality we assume $c^1 \le c^0 < c^k$. In case where $c^0 < c^1$, the problem is infeasible, while in the other case where $c^0 \ge c^k$ the cost constraint is redundant. Let $d = \max\{j : c^j \le c^0\}$. Then $1 \le d < k$ and $c^d \le c^0 < c^{d+1}$. We consider adaptive policies which depend only in the past observations of selections and outcomes. Specifically, let A_t, X_t, $t = 1, 2, \ldots$, denote the population selected and the observed outcome at period t. Let $H_t = (A_1, X_1, \ldots, A_{t-1}, X_{t-1})$ denote the history of actions and observations available at period t. An adaptive policy is a sequence $\pi = (\pi_1, \pi_2, \ldots)$ of history-dependent probability distributions on $\{1, \ldots, k\}$, such that $\pi_n(j, h_n) = P(A_n = j | h_n)$ for a given realization h_n of H_n. 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, that is,

$$\mathcal{V}_\pi(n) = \sum_{i=1}^{k} \sum_{t=1}^{T_\pi^i(n)} X_t^i, \tag{1.1}$$

$$\mathcal{C}_\pi(n) = \sum_{i=1}^{k} \sum_{t=1}^{T_\pi^i(n)} c^i. \tag{1.2}$$

We call an adaptive policy *feasible* if

$$\mathcal{C}_\pi(n)/n \le c^0, \quad \forall n = 1, 2, \ldots. \tag{1.3}$$

The objective is to obtain a feasible policy π that maximizes in some sense $E_{\underline{\theta}}\mathcal{V}_\pi(n)$, $\forall \underline{\theta} \in \underline{\Theta}$. In the next section, we will show that this is equivalent to minimizing a *regret function* $R_\pi(\underline{\theta}, n)$ that represents the expected loss due to lack of information on the sample distributions. For this, we consider the class of feasible policies that are uniformly fast (UF) convergent, in the sense of Burnetas and Katehakis [10]; we call these polices (f-uniformly fast (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 asymptotically 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, Cesa-Bianchi, and Fischer [2], Auer and Ortner [3], Cowan, Honda, and Katehakis [16], Cowan and Katehakis [18].

The extensive literature on the multi-armed bandit (MAB) problem, includes the following: Lai and Robbins [36], Katehakis and Robbins [31], Burnetas and Katehakis [10], Kleinberg [34], Mahajan and Teneketzis [40], Audibert, Munos and Szepesvári [1], Auer and Ortner [3], Honda and Takemura [27], Bubeck and Slivkins [6], Cowan and Katehakis [17] 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 [12]. Tran-Thanh et al. [47], 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 [7] considered a version of this problem and constructed a consistent policy (i.e., with regret $R_\pi(n) = o(n)$, as $n \to \infty$). In

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

the present 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, Kleinberg, and Slivkins [4], considered the problem where there can be more than one side ("knapsack") constraints 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, Deng, and Ye [50], Johnson, Simchi-Levi, and Wang [28], dynamic procurement Singla and Krause [43], and auctions Tran-Thanh et al. [49]. Ding et al. [21] constructed UF policies (i.e., with regret $R_\pi(n) = O(\log n)$, as $n \to \infty$) for cases in which activation costs are bandit-dependent i.i.d. random variables. Johnson et al. [28], provide a network revenue management application of this specific problem where each Bernoulli bandit is specified by an unknown probability of success (where success means that a customer makes a purchase) that depends on the choice of a different price from a finite set of price (vectors). The constraint(s) correspond to limited resources, for example, number of units of a product available. Another application discussed in the same paper involves a "display advertising example". Such problems arise in online ad platforms (e.g., Google) which employ pay per click systems that work as follows. When a user clicks a displayed ad, the advertiser sponsoring the ad pays some amount of money that is split between Google and the publisher of the ad. In this context, each Bernoulli bandit corresponds to a different ad (or multi-ad configuration) with unknown click probability. The side constraint represents advertisers restrictions, for example, budget. For other recent related work we refer to: Guha and Munagala [26], Tran-Thanh et al. [48], Thomaidou, Vazirgiannis, and Liakopoulos [46], Lattimore, Crammer, and Szepesvári [37], Sen, Ridgway, and Ripley [42].

For other work in this area we refer to Katehakis and Derman [30], Katehakis and Veinott Jr [32], Burnetas and Katehakis [8], Burnetas and Katehakis [9], Lagoudakis and Parr [35], Bartlett and Tewari [5], Tekin and Liu [44], Jouini et al. [29], Dayanik, Powell, and Yamazaki [20], Filippi, Cappé, and Garivier [24], Osband and Van Roy [41]. As well as Burnetas and Katehakis [13], Audibert et al. [1], Auer and Ortner [3], Gittins, Glazebrook, and Weber [25], Bubeck and Slivkins [6], Cappé et al. [15], Kaufmann [33], Li, Munos, and Szepesvári [38], Cowan and Katehakis [17], Cowan and Katehakis [19], and references therein. For dynamic programming extensions we refer to Burnetas and Katehakis [11], Butenko, Pardalos, and Murphey [14], Tewari and Bartlett [45], Audibert et al. [1], Littman [39], Feinberg, Kasyanov, and Zgurovsky [22] and references therein.

2. PRELIMINARIES

The complete information problem where $\underline{\theta}$ is known, the expected average reward is to be maximized, and the expected average cost does not exceed c^0, can be solved via the following linear program $\mathrm{LP}(\underline{\theta})$ which is instrumental in the development of the lower bounds and the asymptotically optimal policy.

$$\begin{aligned} \mathrm{LP}(\underline{\theta}): \quad z^*(\underline{\theta}) = \max & \sum_{j=1}^{k} \mu_j(\theta_j) x_j, \\ & \sum_{j=1}^{k} c^j x_j + y = c^0, \\ & \sum_{j=1}^{k} x_j = 1, \\ & x_j \geq 0, \quad \forall j \; y \geq 0. \end{aligned} \tag{2.1}$$

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

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($\underline{\theta}$), cf. Burnetas and Kanavetas [7], Burnetas and Katehakis [12]. 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($\underline{\theta}$) 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($\underline{\theta}$). First, its dual problem DLP($\underline{\theta}$) is

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

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 F to denote the set of BFS:

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

Since the feasible region of Eq. (2.1) is bounded, F is a finite set.

For a basic matrix B, let $v^B = (\lambda^B, g^B)$ denote the dual vector corresponding to B, that is, $v^B = \mu_B(\underline{\theta})B^{-1}$, where $\mu_B(\underline{\theta}) = (\mu_i(\underline{\theta}_i), \mu_j(\underline{\theta}_j))$, or $\mu_B(\underline{\theta}) = (\mu_i(\underline{\theta}_i), 0)$, 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 non-negative:

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

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, that is, $\phi_\alpha^B(\underline{\theta}) = \underline{w}_\alpha^B \underline{\mu}(\underline{\theta})$, where $\underline{w}_\alpha^B$ is an appropriately defined vector that does not depend on $\underline{\mu}(\underline{\theta})$. In the sequel, we use the notation $s(\underline{\theta})$ to denote the set with optimal solutions of LP($\underline{\theta}$) for a vector $\underline{\mu}(\underline{\theta})$, that is, $s(\underline{\theta}) = \{b \in F : b \text{corresponds to an optimal BFS}\}$.

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

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:

$$R_\pi(\underline{\underline{\theta}}, n) = nz^*(\underline{\underline{\theta}}) - E_{\underline{\underline{\theta}}}\mathcal{V}_\pi(n)$$
$$= nz^*(\underline{\underline{\theta}}) - \sum_{j=1}^{k} \mu_j(\underline{\theta}_j) E_{\underline{\underline{\theta}}} T_\pi^j(n). \tag{2.2}$$

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($\underline{\underline{\theta}}$), from the DLP($\underline{\underline{\theta}}$) program we have that $z^*(\underline{\underline{\theta}}) = c^0\lambda^B + g^B$ and $\mu_j(\underline{\theta}_j) = c^j\lambda^B + g^B - \phi_j^B(\underline{\underline{\theta}})$, $\forall j$. These relations and Eq. (2.2) imply:

$$R_\pi(\underline{\underline{\theta}}, n) = \sum_{j=1}^{k} \phi_j^B(\underline{\underline{\theta}}) E_{\underline{\underline{\theta}}} T_\pi^j(n) + \lambda^B \sum_{j=1}^{k} (c^0 - c^j) E_{\underline{\underline{\theta}}} T_\pi^j(n), \tag{2.3}$$

for any $\underline{\underline{\theta}} \in \Theta$.

We now state:

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

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

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

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

In the sequel, we will show that there exist f-UF policies, following the approach of Burnetas and Katehakis [10], by construction of a function $M(\underline{\underline{\theta}})$ and a f-UF policy π^0 such that

$$\lim_{n\to\infty} R_{\pi^0}(\underline{\underline{\theta}}, n)/\log n \le M(\underline{\underline{\theta}}), \quad \forall \underline{\underline{\theta}} \in \Theta.$$

It will be shown in Theorem 1, that policy π^0 has the much stronger property of asymptotic optimality. Indeed, $M(\underline{\underline{\theta}})$ is also a uniform lower bound on the limit of $R_\pi(\underline{\underline{\theta}}, n)/\log n$, of any f-UF policy.

3. A LOWER BOUND FOR THE REGRET

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

For any $b \in s(\underline{\underline{\theta}})$, where $b = \{i, j\}$ or $\{i\}$ and $\alpha \ne i, j$, we define the sets $\Delta\Theta_\alpha(\underline{\underline{\theta}})$ and $D(\underline{\underline{\theta}})$, as follows. The first set includes all perturbed values $\underline{\theta}_\alpha^{'}$ of $\underline{\theta}_\alpha$ of population α, such that the complete information problem under $\underline{\underline{\theta}}^{'}$ where only $\underline{\theta}_\alpha$ is perturbed to $\underline{\theta}_\alpha^{'}$ 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,

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

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

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 $\underline{\theta}_\alpha$. Then, $D(\underline{\underline{\theta}})$ is the set of populations, which are not optimal under $\underline{\underline{\theta}}$, but become part of a uniquely optimal BFS after a parameter change of $\underline{\theta}_\alpha$ only,

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

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 population α becomes optimal under $\underline{\theta}'_\alpha$.

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

We have:

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

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:

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

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. (2.3) both terms are non-negative, the first because of optimality and the second because of feasibility. Therefore, 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 corresponding to b:

$$\phi_j^B(\underline{\underline{\theta}}) \lim_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T_\pi^j(n)}{n^a} = 0, \quad \text{for all } a > 0, \ j \notin b, \tag{3.1}$$

and

$$\lambda^B \lim_{n\to\infty} \frac{\sum_{j\in b}(c^0 - c^j) E_{\underline{\underline{\theta}}} T_\pi^j(n)}{n^a} = 0. \tag{3.2}$$

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

We can now state:

LEMMA 2: *Assume 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* $i \leq d < j$ *we have* $\lambda^B > 0$,
(ii) if $B = \begin{pmatrix} c^i & 1 \\ 1 & 0 \end{pmatrix}$, *for some* $i \leq d$ *we have* $\lambda^B = 0$.

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*

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

So far we have shown that a necessary condition for a UF policy is that $\forall \underline{\underline{\theta}} \in \underline{\underline{\Theta}}$, 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 a positive sequence* $\beta_n = o(n)$ *then*

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

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

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*

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

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

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

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

Also, we have from Lemma 2 that $\lambda^B \geq 0$ and from Eq. (1.3), $nc^0 - E_{\underline{\underline{\theta}}} C_\pi(n) \geq 0$, for all n. Finally, we have that the optimal populations under $\underline{\underline{\theta}}$ have $\phi_j^B(\underline{\underline{\theta}}) = 0$, thus

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

This completes the proof. ■

4. BLOCKS AND BLOCK-BASED POLICIES

We consider a class of policies such that 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. (1.3) as follows. Define the differences

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

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

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

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

Since δ^i are 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 $J \subseteq \{1, \ldots, k\}$ 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 integer linear program:

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

An optimal solution of LP($\underline{\underline{\theta}}$) 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) 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 $\{1, \ldots, k\}$ 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. (1.3) is satisfied sample-path wise. This block is necessary in order to obtain initial estimates of $\mu_j(\underline{\underline{\theta}}_j)$ for all populations. This means that the ISB block has the minimum length

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

of $\ell(J)$, defined above with one more constraint that $m_j \geq n_0$ for all j, with $J = \{1, \ldots, k\}$.

(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:

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

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($\underline{\theta}$), using the known 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, $\sum_{b \in F} \widetilde{T}_\pi^b(l) = l - 1$. 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 in a LPB(b) block and m_α is the number of samples from population α in the ISB block. Now we can define the regret of the first l blocks

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

$$- \sum_{j=1}^{k} \mu_j(\underline{\theta}_j) m_j.$$

We note that

$$T_\pi^\alpha(S_\pi(L_n)) \leq T_\pi^\alpha(n) \leq T_\pi^\alpha(S_\pi(L_n)) + M_\alpha, \tag{4.1}$$

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

$$\widetilde{R}_\pi(\underline{\theta}, L_n) + (n - E_{\underline{\theta}} S_\pi(L_n)) \, z^*(\underline{\theta}) - \sum_{j=1}^{k} M_j \, \mu_j(\underline{\theta}_j)$$

$$\leq R_\pi(\underline{\theta}, n) \leq \widetilde{R}_\pi(\underline{\theta}, L_n) + (n - E_{\underline{\theta}} S_\pi(L_n)) \, z^*(\underline{\theta}). \tag{4.2}$$

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

The above imply the following relation between the two regret functions,

$$\limsup_{n\to\infty} \frac{R_\pi(\underline{\underline{\theta}}, n)}{\log n} = \limsup_{n\to\infty} \frac{\widetilde{R}_\pi(\underline{\underline{\theta}}, L_n)}{\log L_n}. \tag{4.3}$$

From Eq. (4.3), it follows that in order 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 $\widetilde{R}_\pi(\underline{\underline{\theta}}, L_n)$.

5. 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 Z-UCB policy below we need some definitions. At the beginning of any block l we have the estimates $\hat{\underline{\underline{\theta}}}^l$, which give an optimal solution $b(\hat{\underline{\underline{\theta}}}^l)$ and corresponding optimal basis matrix $\hat{B}$ of of LP$(\hat{\underline{\underline{\theta}}}^l)$. Using the optimal solution of the linear program $b(\hat{\underline{\underline{\theta}}}^l)$ we can compute a corresponding $\rho(\hat{\underline{\underline{\theta}}}^l, \alpha, \hat{B})$, which satisfies the conditions of Lemma 1. We define at any block l and for every population α as $v_\alpha = v_\alpha(\hat{\underline{\underline{\theta}}})$

$$v_\alpha = \sup_{\underline{\theta}'_\alpha} \left\{ \mu_\alpha(\underline{\theta}'_\alpha) : I(\hat{\underline{\theta}}^l_\alpha, \underline{\theta}'_\alpha) \le \frac{\log S_\pi(l-1)}{T^\alpha_\pi(S_\pi(l-1))} \right\}, \tag{5.1}$$

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

$$\Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)} = \{\alpha \ : \mu^*_\alpha(\hat{\underline{\underline{\theta}}}^l) < v_\alpha < \mu^*_\alpha(\hat{\underline{\underline{\theta}}}^l) + \rho(\hat{\underline{\underline{\theta}}}^l, \alpha, \hat{B})\}. \tag{5.2}$$

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 $z^b = \mu_i$. Let $F = \{b\}$ denote the set of all BFSs of LP$(\underline{\underline{\theta}})$. Note that this set depends on the sampling costs of bandits and does not change since our sampling costs are constants and let $|F|$ denote its cardinality ($|F| = k(k-1)$). Finally, we take a fixed constant $\tau \in (0, 1/|F|)$.

Z-UCB Policy π^0

Take one ISB block in order to have one estimate $\hat{\underline{\theta}}_a$ from each population a. We call the vector of all estimates $\hat{\underline{\underline{\theta}}}^2$ and $\widetilde{T}^b_\pi(l)$ be the number of LPB(b) type blocks in first $l \ge 2$ blocks for any $b \in F$.

For blocks $l = 2, \ldots,$ we employ only LPB(b) blocks each of which corresponds to a feasible (not necessarily optimal) solution $b^l = b^l(\hat{\underline{\underline{\theta}}}^l, \{\widetilde{T}^b_\pi(l)\}_{b\in F})$ which is equal to $\pi^0(\hat{\underline{\underline{\theta}}}^l)$ defined below.

Step 1. At the beginning of block $l = 2, \ldots,$ we have the estimates $\hat{\underline{\underline{\theta}}}^l$, and the total the number of LPB(b) type blocks in first $l \ge 2$ blocks $\widetilde{T}^b_\pi(l)$, from the previous $l-1$ blocks. Then we compute:

$$\xi^l = \max_{b\in F}\{z^b(\hat{\underline{\underline{\theta}}}^l) \ : \ \widetilde{T}^b_\pi(l) \ge \tau(l-1)\}.$$

Let b^l_0 be any BFS in F be such that $z^{b^l_0}(\hat{\underline{\underline{\theta}}}^l) = \xi^l$.

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

Step 2. Then for every population $\alpha = \{1, \ldots, k\}$, we compute the v_α's and $\Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}$'s, by Eqs. (5.1) and (5.2).

(*i*) If $\Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)} \neq \emptyset$, for every $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}$ we define the index: $u_\alpha(\hat{\underline{\theta}}^l) = u_\alpha(\hat{\underline{\theta}}^l, s, t)$

$$u_\alpha(\hat{\underline{\theta}}^l, s, t) = \max_{\underline{\theta}'_\alpha} \left\{ z^{b_\alpha(\hat{\underline{\theta}}^l, \underline{\theta}'_\alpha)} : I(\hat{\underline{\theta}}^l_\alpha, \underline{\theta}'_\alpha) \leq \frac{\log s}{t} \right\}, \tag{5.3}$$

where $b_\alpha(\hat{\underline{\theta}}^l, \underline{\theta}'_\alpha)$ is the uniquely optimal solution of LP$(\hat{\underline{\theta}}^l, \underline{\theta}'_\alpha)$, which is obtained from the LP$(\hat{\underline{\theta}}^l)$ when we replace only the parameter of population α by $\underline{\theta}'_\alpha$. Given the history until l block and the statistics $S_{\pi^0}(l-1)$ and $T^\alpha_{\pi^0}(S_{\pi^0}(l-1))$, we compute, for every $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}$, the indices

$$u_\alpha(\hat{\underline{\theta}}^l) = u_\alpha(\hat{\underline{\theta}}^l, s, t) = u_\alpha(\hat{\underline{\theta}}^l, S_{\pi^0}(l-1), T^\alpha_{\pi^0}(S_{\pi^0}(l-1))),$$

and the corresponding uniquely optimal BFS: $b^0_\alpha(\hat{\underline{\theta}}^l) = b_\alpha(\hat{\underline{\theta}}^l, \underline{\theta}^0_\alpha(s,t))$ in Eq. (5.3) cf. discussion in page 5.

The Z-UCB policy π^0 employs the LPB block which corresponds to the index:

$$\pi^0(\hat{\underline{\theta}}^l) = \underset{b^0_\alpha(\hat{\underline{\theta}}^l),\ \alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}}{\arg\max} \{u_\alpha(\hat{\underline{\theta}}^l)\}. \tag{5.4}$$

(*ii*) If $\Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)} = \emptyset$, we take $\pi^0(\hat{\underline{\theta}}^l) = b^l_0$, which means that there are not populations which under an increase can give a better solution, the Z-UCB policy π^0 employs the LPB(b^l_0) block.

Remark 2:

(a) In the computation of ξ^l in Step 1 above, it is not required to solve any LPs but only to compute and compare the corresponding z^b, for which $\widetilde{T}^b_\pi(l) \geq \tau(l-1)$, using the explicit formulas for z^b; recall that for any BFS b, where $b = \{i, j\}$ or $\{i\}$ the optimal solution is $z^b = \mu_i x_i + \mu_j x_j$ or $z^b = \mu_i$.

(b) With apologies for the notation we have used b^l_0 as the initial optimal solution and $b^0_\alpha(\hat{\underline{\theta}}^l)$ as the inflated solutions of the initial solution.

In words, the Z-UCB policy firstly employs an ISB block in order to have estimates for each population. Then it samples LPB blocks, $l = 1, 2, \ldots$, following two steps. In the first step, the Z-UCB policy finds the initial optimal solution b^l_0 (with basis matrix $\hat{B}$) which is the best solution among the BFS ($\{b\}$) that have been sampled sufficiently often: $\widetilde{T}^b_\pi(l) \geq \tau(l-1)$, where τ is any fixed constant in $(0, 1/|F|)$. In the second step, it computes the indices $u_\alpha(\hat{\underline{\theta}}^l, s, t)$ (of Eq. (5.3)) for all populations $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}$, by varying only $\underline{\theta}'_\alpha$. Thus, for any $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}$, we have the corresponding new uniquely optimal BFS: $b^0_\alpha(\hat{\underline{\theta}}^l)$. Then, the Z-UCB policy chooses to employ the LPB block which corresponds to the highest z value among $b^0_\alpha(\hat{\underline{\theta}}^l)$, $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}$.

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

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

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

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

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

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

we have that $u_\alpha(\hat{\underline{\theta}}^l)>z^*(\underline{\theta})-\epsilon$ if and only if $J_\alpha(\hat{\underline{\theta}}^l,\epsilon)<\log S_\pi(l-1)/T_\pi^\alpha(S_\pi(l-1))$.

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

$$\lim_{\epsilon\to0}J_i(\underline{\theta},\epsilon)=\infty.$$

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

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

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

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

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.*

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

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

$$\limsup_{n\to\infty}\frac{E_{\underline{\theta}}T_{\pi^0}^j(n)}{\log n}\le\frac{1}{K_j(\underline{\theta})},\quad\forall j\in D(\underline{\theta}),\tag{5.5}$$

$$\limsup_{n\to\infty}\frac{E_{\underline{\theta}}T_{\pi^0}^j(n)}{\log n}=0,\quad\forall j\notin D(\underline{\theta}),\quad j\notin b,\quad\forall b\in s(\underline{\theta}),\tag{5.6}$$

$$nc^0-E_{\underline{\theta}}C_{\pi^0}(n)=o(\log n).\tag{5.7}$$

The proof of these inequalities is given in the appendix. ■

Remark 3: According to Remark 4*b* in Burnetas and Katehakis [10] condition (C2) is equivalent to C2′ below which is easier to verify.

(C2′) $\forall\delta>0$, as $t\to\infty$

$$\sum_{j=1}^{t-1}P_{\underline{\theta}_i}(b_0^j\in s(\underline{\theta}),J_i(\hat{\underline{\theta}}^j,\epsilon)\le J_i(\underline{\theta},\epsilon)-\delta)=o(\log t).$$

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

6. NORMAL DISTRIBUTIONS WITH KNOWN VARIANCES

Assume the observations X_α^j from population α are normally distributed with unknown means $EX_\alpha^j = \theta_\alpha$ and known variances σ_α^2, that is, $\underline{\theta}_\alpha = \theta_\alpha$, $\mu_\alpha(\underline{\theta}_\alpha) = \theta_\alpha$, and $\Theta_\alpha = (-\infty, +\infty)$. Given history h_l, define

$$\mu_\alpha(\hat{\theta}_\alpha^l) = \frac{\sum_{j=1}^{T_{\pi^0}^\alpha(S_{\pi^0}(l-1))} X_\alpha^j}{T_{\pi^0}^\alpha(S_{\pi^0}(l-1))}.$$

Now from the definition of Θ_α, it follows that $\Delta\Theta_\alpha(\underline{\theta}) = (\theta_\alpha + \phi_\alpha^B(\underline{\theta}), \theta_\alpha + \phi_\alpha^B(\underline{\theta}) + \rho(\underline{\theta}, \alpha, B))$ for any optimal matrix B under $\underline{\theta}$, therefore $D(\underline{\theta}) = \{\alpha : \alpha \notin b \text{ for any } b \in s(\underline{\theta})\}$, $\forall\, \underline{\theta} \in \Theta$. Thus, we can see from the structure of the sets Θ_α and $\Delta\Theta_\alpha(\underline{\theta})$ that condition (C1) is satisfied and that $\Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)} \neq \emptyset$ (we do not have the case (ii) in Step 2 of our policy).

Also, we have:

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

$$K_\alpha(\underline{\theta}) = \frac{(\phi_\alpha^B(\underline{\theta}))^2}{2\sigma_\alpha^2}.$$

It is easy to see that the Z-UCB indices of Eq. (5.3) simplify to:

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

where

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

is the θ_α which satisfies the maximum in the index $u_\alpha(\hat{\underline{\theta}}^l)$, of Eq. (5.3).

Note that $b_\alpha(\hat{\underline{\theta}}^l, \theta_\alpha^{K_\alpha}) = \{\alpha, j\}$ or $\{\alpha\}$; thus, in the first case we have $z^{b_\alpha(\hat{\underline{\theta}}^l, \theta_\alpha^{K_\alpha})} = \theta_\alpha^{K_\alpha} x_\alpha + \hat{\theta}_j^l x_j$ and $z^*(\underline{\theta}) = \theta_\alpha x_\alpha + \theta_j x_j$ and in the second $z^{b_\alpha(\hat{\underline{\theta}}^l, \theta_\alpha^{K_\alpha})} = \theta_\alpha^{K_\alpha}$ and $z^*(\underline{\theta}) = \theta_\alpha$. Therefore, for $b_\alpha(\hat{\underline{\theta}}^l, \theta_\alpha^{K_\alpha}) \in s(\underline{\theta})$ and from the structure of $z^{b_\alpha(\hat{\underline{\theta}}^l, \theta_\alpha^{K_\alpha})}$ the index is a sum of normal distributions, which is also a normal distribution, and from a well-known tail inequality of normal distribution condition (C3) is satisfied.

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

$$\sum_{t=2}^{L_n} P_{\theta_i}(b_0^t \in s(\underline{\theta}), J_i(\hat{\underline{\theta}}^t, \epsilon) \leq J_i(\underline{\theta}, \epsilon) - \delta)$$

$$= \sum_{t=2}^{L_n} P_{\theta_i}(b_0^t \in s(\underline{\theta}), |\hat{\theta}_i^t - \theta_i| > \xi), \xi > 0,$$

where the equality follows after some algebra because of the normal distribution and the explicit form of $I(\hat{\theta}_i^t, \theta_i^{'})$ in this case:

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

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

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

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

$$P_{\theta_i}^{\pi^0}(|\hat{\theta}_i^t - \theta_i| > \xi) \leq P_{\theta_i}^{\pi^0}(|\hat{\theta}_i^l - \theta_i| > \xi, \text{ for some } l \leq t)$$
$$\leq \sum_{l=2}^{t} P_{\theta_i}^{\pi^0}(|\hat{\theta}_i^l - \theta_i| > \xi) = o(1/t),$$

where the last equality follows from a consequence of the tail inequality $1 - \Phi(x) < \Phi(x)/x$ for the standard normal distribution, Feller [23]. Thus, we can see that condition (C2) holds.

6.1. Summary of Z-UCB Policy

Take one ISB block in order to have one estimate $\hat{\theta}_a$ from each population a. We call the vector of all estimates $\hat{\underline{\theta}}^2$ and $\widetilde{T}_\pi^b(l)$ be the number of LPB(b) type blocks in first $l \geq 2$ blocks for any $b \in F$.

For blocks $l = 2, \ldots,$ we employ only LPB(b) blocks each of which corresponds to a feasible solution $b^l = b^l(\hat{\underline{\theta}}^l, \{\widetilde{T}_\pi^b(l)\}_{b\in F})$ which is equal to $\pi^0(\hat{\underline{\theta}}^l)$ defined below.

Step 1. At the beginning of block $l = 2, \ldots,$ we have the estimates $\hat{\underline{\theta}}^l$, and the total number of LPB(b) type blocks $\widetilde{T}_\pi^b(l)$, from the previous $l-1$ blocks. Then we compute:

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

Let b_0^l be any BFS in F such that $z^{b_0^l}(\hat{\underline{\theta}}^l) = \xi^l$.

Step 2. Then for every population $\alpha = \{1, \ldots, k\}$ we compute the indices

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

where $b_\alpha(\hat{\underline{\theta}}^l, \theta_\alpha^{K_\alpha})$ is the solution, which we obtain if we replace only the parameter of population α by $\theta_\alpha^{K_\alpha}$, where

$$\theta_\alpha^{K_\alpha} = \hat{\theta}_\alpha^l + \sigma_\alpha \left(\frac{2\log S_{\pi^0}(l-1)}{T_{\pi^0}^\alpha(S_{\pi^0}(l-1))}\right)^{1/2}. \tag{6.1}$$

Finally, the Z-UCB policy employs as block l the $\arg\max_{b_\alpha}\{u_\alpha(\hat{\underline{\theta}}^l)\}$.

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. (6.1) by a constant times $\hat{\sigma}_\alpha$, as in Auer et al. [2]. This work is currently in progress.

Acknowledgement

We are grateful to an anonymous referee for providing several suggestions, which significantly improved the paper. We also acknowledge support for this project from the National Science Foundation (NSF grant CMMI-1450743), the European Union (European Social Fund – ESF), and Greek national funds through the Operational Program Education and Lifelong Learning of the National Strategic Reference Framework (NSRF) – Research Funding Program: Thales-Athens University of Economics and Business-New Methods in the Analysis of Market Competition: Oligopoly, Networks and Regulation.

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

References

1. Audibert, J.-Y., Munos, R., & Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science* 410(19): 1876–1902.
2. Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47(2–3): 235–256.
3. Auer, P. and Ortner, R. (2010). Ucb revisited: improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica* 61(1–2): 55–65.
4. Badanidiyuru, A., Kleinberg, R., & Slivkins, A. (2013). Bandits with knapsacks. Foundations of Computer Science (FOCS). In *2013 IEEE 54th Annual Symposium on IEEE*, pp. 207–216.
5. Bartlett, P.L. & Tewari, A. (2009). Regal: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, pp. 35–42.
6. Bubeck, S. & Slivkins, A. (2012). The best of both worlds: stochastic and adversarial bandits. arXiv:1202.4473.
7. Burnetas, A.N. & Kanavetas, O.A. (2012). Adaptive policies for sequential sampling under incomplete information and a cost constraint. In N.J. Daras (ed.), *Applications of mathematics and informatics in military science*, Springer, pp. 97–112.
8. Burnetas, A.N. & Katehakis, M.N. (1993). On sequencing two types of tasks on a single processor under incomplete information. *Probability in the Engineering and Informational Sciences* 7(1): 85–119.
9. Burnetas, A.N. & Katehakis, M.N. (1996). On large deviations properties of sequential allocation problems. *Stochastic Analysis and Applications* 14(1): 23–31.
10. Burnetas, A.N. & Katehakis, M.N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* 17(2): 122–142.
11. Burnetas, A.N. & Katehakis, M.N. (1997). Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research* 22(1): 222–255.
12. Burnetas, A.N. & Katehakis, M.N. (1998). Sequential allocation problems with side constraints. INFORMS Seattle 1998, Annual Meeting, Seattle, WA.
13. Burnetas, A.N. & Katehakis, M.N. (2003). Asymptotic Bayes analysis for the finite-horizon one-armed-bandit problem. *Probability in the Engineering and Informational Sciences* 17(01): 53–82.
14. Butenko, S., Murphey, R., & Pardalos, P.M. (eds.). (2013). *Cooperative control: models, applications and algorithms* (Vol. 1). Springer Science & Business Media, New York.
15. Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., & Stoltz, G. (2013). Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics* 41(3): 1516–1541.
16. Cowan, W., Honda, J., & Katehakis, M.N. (2015). Asymptotic optimality, finite horizon regret bounds, and a solution to an open problem. arXiv:1504.05823. *Journal of Machine Learning Research*, to appear.
17. Cowan, W. & Katehakis, M.N. (2015). Asymptotic behavior of minimal-exploration allocation policies: almost sure, arbitrarily slow growing regret. arXiv:1505.02865.
18. Cowan, W. & Katehakis, M.N. (2015). An asymptotically optimal UCB policy for uniform bandits of unknown support. arXiv:1505.01918.
19. Cowan, W. & Katehakis, M.N. (2015*c*). Multi-armed bandits under general depreciation and commitment. *Probability in the Engineering and Informational Sciences* 29(01): 51–76.
20. Dayanik, S., Powell, W.B., & Yamazaki, K. (2013). Asymptotically optimal Bayesian sequential change detection and identification rules. *Annals of Operations Research* 208(1): 337–370.
21. Ding, W., Qin, T., Zhang, X.-D., & Liu, T.-Y. (2013). Multi-armed bandit with budget constraint and variable costs. In *AAAI-13 Conference*, pp. 232–238.
22. Feinberg, E.A., Kasyanov, P.O., & Zgurovsky, M.Z. (2014). Convergence of value iterations for total-cost MDPs and POMDPs with general state and action sets. Adaptive Dynamic Programming and Reinforcement Learning (ADPRL). In *2014 IEEE Symposium on IEEE*, pp. 1–8.
23. Feller, W. (1967). *An introduction to probability theory and its applications*, Vol. 1; 3rd ed, Wiley, New York.
24. Filippi, S., Cappé, O., & Garivier, A. (2010). Optimism in reinforcement learning based on Kullback–Leibler divergence. In *48th Annual Allerton Conference on Communication, Control, and Computing*, pp. 115–122.
25. Gittins, J.C., Glazebrook, K., & Weber, R.R. (2011). *Multi-armed Bandit allocation indices*, West Sussex, UK: John Wiley & Sons.
26. Guha, S. & Munagala, K. (2007). Approximation algorithms for budgeted learning problems. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, ACM, pp. 104–113.

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

27. Honda, J. & Takemura, A. (2011). An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning* 85(3): 361–391.
28. Johnson, K., Simchi-Levi, D., & Wang, H. (2015). Online network revenue management using Thompson sampling. Available at SSRN.
29. Jouini, W., Ernst, D., Moy, C., & Palicot, J. (2009). Multi-armed bandit based policies for cognitive radio's decision making issues. In *Third International Conference on Signals, Circuits and Systems (SCS)*, pp. 1–6.
30. Katehakis, M.N. & Derman, C. (1986). *Computing optimal sequential allocation rules. Clinical Trials*. Vol. 8; Lecture Note Series: Adoptive Statistical Procedures and Related Topics, Institute of Mathematical Statistics, pp. 29–39.
31. Katehakis, M.N. & Robbins, H. (1995). Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America* 92(19): 8584.
32. Katehakis, M.N. & Veinott A.F. Jr. (1987). The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research* 12: 262–268.
33. Kaufmann, E. (2015). Analyse de stratégies Bayésiennes et fréquentistes pour l'allocation séquentielle de ressources. Doctorat, ParisTech.
34. Kleinberg, R.D. (2004). Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems Conference*. pp. 697–704.
35. Lagoudakis, M.G. & Parr, R. (2003). Least-squares policy iteration. *The Journal of Machine Learning Research* 4: 1107–1149.
36. Lai, T.L. & Robbins, H. (1985). 'Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6(1): 4–2.
37. Lattimore, T., Crammer, K., & Szepesvári, C. (2014). Optimal resource allocation with semi-Bandit feedback. arXiv:1406.3840 .
38. Li, L., Munos, R., & Szepesvári, C. (2014). On minimax optimal offline policy evaluation. arXiv:1409.3653.
39. Littman, M.L. (2012). Inducing partially observable Markov decision processes. In *ICGI Conference*, pp. 145–148.
40. Mahajan, A. & Teneketzis, D. (2008). Multi-armed bandit problems. In A.O. Hero, D. Castanon, D. Cocharn, & K. Kastella (eds.), *Foundations and applications of sensor management*, Springer, pp. 121–151.
41. Osband, I. & Van Roy, B. (2014). Near-optimal reinforcement learning in factored MDPs. In *Advances in Neural Information Processing Systems Conference*. pp. 604–612.
42. Sen, S., Ridgway, A., & Ripley, M. (2015). Adaptive budgeted bandit algorithms for trust development in a supply-chain. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, International Foundation for Autonomous Agents and Multiagent Systems*. pp. 137–144.
43. Singla, A. & Krause, A. (2013). Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 1167–1178.
44. Tekin, C. & Liu, M. (2012). Approximately optimal adaptive learning in opportunistic spectrum access. *INFOCOM, 2012 Proceedings IEEE*. IEEE, pp. 1548–1556.
45. Tewari, A. & Bartlett, P.L. (2008). Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Advances in Neural Information Processing Systems Conference*. pp. 1505–1512.
46. Thomaidou, S., Vazirgiannis, M., & Liakopoulos, K. (2012). Toward an integrated framework for automated development and optimization of online advertising campaigns. arXiv:1208.1187 .
47. Tran-Thanh, L., Chapman, A., Luna, M.D.C.F., Enrique, J., Rogers, A., & Jennings, N.R. (2010). Epsilon – first policies for budget – limited multi-armed bandits. In *AAAI-2010 Conference*, pp. 1211–1216.
48. Tran-Thanh, L., Chapman, A., Luna, M.D.C.F., Enrique, J., Rogers, A., & Jennings, N.R. (2012). Knapsack based optimal policies for budget-limited multi-armed bandits. In *AAAI-2012 Conference*, pp. 1134–1140.
49. Tran-Thanh, L., Stavrogiannis, L.C., Naroditskiy, V., Robu, V., Jennings, N.R., & Key, P. (2014). Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions'. University of Southampton, UK, Technical Report.
50. Wang, Z., Deng, S., & Ye, Y. (2014). Close the gaps: a learning-while-doing algorithm for single-product revenue management problems. *Operations Research* 62(2): 318–331.

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

APPENDIX PROOFS

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

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

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

Now, any optimal $b \in s(\underline{\theta})$ is not optimal under $\underline{\theta}' = (\underline{\theta}_1, \ldots, \underline{\theta}'_\alpha, \ldots, \underline{\theta}_k)$, for any $\underline{\theta}'_\alpha \in \Delta\Theta_\alpha(\underline{\theta})$, thus $s(\underline{\theta}') = \{b'\}$ where $b' \notin s(\underline{\theta})$.

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

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

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

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

The other cases where the population α is a population with cost lower than c^0 and the optimal solution under $\underline{\theta}'$ has this form $b' = \{\alpha, j\}$ or $b' = \{\alpha\}$ follows with the same arguments as in the previous paragraph. ■

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

PROOF: *(i)* Let $\underline{\theta}: s(\underline{\theta}) = \{b\}$, $b = (i, j)$ 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 $\underline{\theta}: s(\underline{\theta}) = \{b\}$, $b = (i)$ for $i \leq d$, then $\lambda^B = 0$ from the dual solution and $\phi_j^B(\underline{\theta}) > 0$ for all $j \neq i$. ■

We recall for the next Proposition

$$z^*(\underline{\theta}) = \max \sum_{j=1}^{k} \mu_j(\underline{\theta}_j)x_j,$$

$$\sum_{j=1}^{k} c^j x_j + y = c^0,$$

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

$$\sum_{j=1}^{k} x_j = 1,$$

$$x_j \geq 0, \forall j, y \geq 0,$$

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

$$\phi_j^B(\underline{\underline{\theta}}) \lim_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T_\pi^j(n)}{n^a} = 0, \quad \text{for all } a > 0,\ j \notin b, \tag{A.1}$$

and also,

$$\lambda^B \lim_{n\to\infty} \frac{\sum_{j\in b}(c^0 - c^j) E_{\underline{\underline{\theta}}} T_\pi^j(n)}{n^a} = 0, \quad \text{for all } B \text{ corresponding to } b. \tag{A.2}$$

PROPOSITION 1: *For any 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*

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

PROOF: Let $\alpha \in D(\underline{\underline{\theta}})$, $\theta_\alpha' \in \Delta\Theta_\alpha(\underline{\underline{\theta}})$. Because of the definition of $\Delta\Theta_\alpha(\underline{\underline{\theta}})$ 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 non-degenerate then the basic matrix

$$B' = \begin{pmatrix} c^\alpha & 1 \\ 1 & 0 \end{pmatrix}$$

and from Lemma 2 for a UF policy $\lambda^B = 0$ thus,

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

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 $E_{\underline{\underline{\theta}}'} T_\pi^j(n) = o(n^a)$. Moreover from Eq. (A.1) $E_{\underline{\underline{\theta}}'} T_\pi^i(n) = o(n^a)$, for all $i \neq j, \alpha$, thus $E_{\underline{\underline{\theta}}'} T_\pi^j(n) = o(n^a)$, for all $j \neq \alpha$.

Therefore,

$$n - E_{\underline{\underline{\theta}}'} T_\pi^\alpha(n) = o(n^a), \quad \text{for all } a > 0. \tag{A.3}$$

It is also true that

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

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

Therefore

$$n - E_{\underline{\underline{\theta}}'} T_\pi^\alpha(n) \geq (n - \beta_n) P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) \leq \beta_n]. \tag{A.4}$$

From Eqs. (A.3) and (A.4) we obtain

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

thus

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

We next consider the case $b^{'} = \{j_0, \alpha\}$ with $c^{j_0} < c^0 < c^\alpha$. (The case $c^\alpha < c^0 < c^{j_0}$ is completely analogous.) We have from Lemma 2 that for a UF policy $\lambda^B > 0$, thus

$$E_{\underline{\underline{\theta}}'} T_\pi^j(n) = o(n^a), \quad \forall a > 0, \ \forall j \notin b^{'} = \{j_0, \alpha\} \tag{A.5}$$

and

$$(c^0 - c^{j_0}) E_{\underline{\underline{\theta}}'} T_\pi^{j_0}(n) + (c^0 - c^\alpha) E_{\underline{\underline{\theta}}'} T_\pi^\alpha(n) = o(n^a), \quad \forall a > 0. \tag{A.6}$$

If we sum Eq. (A.5) for all $j \neq \alpha, j_0$ it follows that

$$n - E_{\underline{\underline{\theta}}'} T_\pi^{j_0}(n) - E_{\underline{\underline{\theta}}'} T_\pi^\alpha(n) = \varepsilon_n, \text{ where } \varepsilon_n = o(n^a), \quad \forall a > 0. \tag{A.7}$$

Dividing Eq. (A.6) with $c^\alpha - c^{j_0}$ and using Eq. (A.7), we obtain after some algebra the following two equalities:

$$\begin{aligned} n x_{j_0}^{'} - E_{\underline{\underline{\theta}}'} T_\pi^{j_0}(n) &= o(n^a), \\ n x_\alpha^{'} - E_{\underline{\underline{\theta}}'} T_\pi^\alpha(n) &= o(n^a), \quad \forall a > 0. \end{aligned} \tag{A.8}$$

where

$$x_{j_0}^{'} = \frac{c^\alpha - c^0}{c^\alpha - c^{j_0}} \quad \text{and} \quad x_\alpha^{'} = \frac{c^0 - c^{j_0}}{c^\alpha - c^{j_0}}$$

are the probabilities, which correspond to optimal solution $b^{'}$ of linear program Eq. (2.1) 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. (A.7)

$$E_{\underline{\underline{\theta}}'} \Gamma_n^\pi = o(n^a), \quad \forall a > 0. \tag{A.9}$$

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

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

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

therefore

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

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,

$$n x_\alpha^{'} - T_\pi^\alpha(n) + \Gamma_n^\pi \rho(j_0, \alpha) \ge 0. \tag{A.10}$$

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

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

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. (A.10) 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

$$E_{\underline{\underline{\theta}}'} Z_n^\pi = o(n^a), \ \forall a > 0 \text{ from Eqs. (A.8) and (A.9).}$$

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

Let

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

$$\begin{aligned} P_{\underline{\underline{\theta}}'}(V_n^\pi) &= P_{\underline{\underline{\theta}}'}(V_n^\pi \cap \{\Gamma_n^\pi \le n\delta\}) + P_{\underline{\underline{\theta}}'}(V_n^\pi \cap \{\Gamma_n^\pi > n\delta\}) \\ &\le P_{\underline{\underline{\theta}}'}(V_n^\pi \cap \{\Gamma_n^\pi \le n\delta\}) + P_{\underline{\underline{\theta}}'}(\Gamma_n^\pi > n\delta), \end{aligned} \tag{A.11}$$

where $0 < \delta < \{x_{j_0}^{'}/[1+\rho(j_0,\alpha)]\}$ and using Eq. (A.9) we have that

$$\begin{aligned} P_{\underline{\underline{\theta}}'}(\Gamma_n^\pi > n\delta) &\le \frac{E_{\underline{\underline{\theta}}'}\Gamma_n^\pi}{n\delta} \\ &= \frac{o(n^a)}{n\delta} = o(n^{a-1}),\ \forall a > 0. \end{aligned} \tag{A.12}$$

Let

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

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 $\beta_n = o(n)$,

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

and we have that

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

thus from Eqs. (A.11), (A.12)

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

Finally,

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

This completes the proof. ■

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*

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

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

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

PROOF: If we take $\beta_n = \log n / K_\alpha(\underline{\underline{\theta}})$, then $P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) < \{\log n / K_\alpha(\underline{\underline{\theta}})\}] = o(n^{a-1})$ and using a change of measure from $\underline{\underline{\theta}}'$ to $\underline{\underline{\theta}}$ and following the arguments in Burnetas and Katehakis [10], Lai and Robbins [36] we have that

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

This completes the proof. ■

We recall for Theorem 2 that

$$1.\quad \limsup_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T_\pi^j(n)}{\log n} \le \frac{1}{K_j(\underline{\underline{\theta}})}, \quad \text{for all } j \in D(\underline{\underline{\theta}}), \tag{A.13}$$

$$2.\quad \limsup_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T_\pi^j(n)}{\log n} = 0, \quad \text{for all } j \notin D(\underline{\underline{\theta}}),\ j \notin b,\ \forall b \in s(\underline{\underline{\theta}}), \tag{A.14}$$

$$3.\quad nc^0 - E_{\underline{\underline{\theta}}} C_\pi(n) = o(\log n). \tag{A.15}$$

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

$$T_\pi^\alpha(S_\pi(L_n)) \le T_\pi^\alpha(n) \le T_\pi^\alpha(S_\pi(L_n)) + M_\alpha, \tag{A.16}$$

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

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

$$\begin{aligned}\widetilde{T}_\pi^b(L_n) &= \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b_0^t \notin s(\underline{\underline{\theta}})\} + \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b_0^t \in s(\underline{\underline{\theta}})\} \\ &\le \sum_{t=2}^{L_n} 1\{b_0^t \notin s(\underline{\underline{\theta}})\} + \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b_0^t \in s(\underline{\underline{\theta}})\}.\end{aligned} \tag{A.17}$$

Finally, a policy π is called feasible if

$$\frac{C_\pi(n)}{n} \le c^0, \quad \forall n = 1, 2, \ldots. \tag{A.18}$$

THEOREM 3: *Under conditions (C1),(C2), and (C3), policy π^0 satisfies:*

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

PROOF: We need to prove Eqs. (A.13)–(A.15). From Eqs. (A.16) and (A.17) and Lemmas A.8 and A.9 we have proved the relations Eq. (A.13) and (A.14). Equation (A.15) follows from Eq. (A.18) the feasibility of π^0 and block policies. ■

For the analysis of proofs of Lemmas 4 and 5 we use the index $u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\underline{\theta}}_i')$ at block t, where $\underline{\underline{\theta}}_i'$ is the corresponding inflated $\underline{\underline{\theta}}_i$, to denote the index $u_i(\hat{\underline{\underline{\theta}}}^t)$ of population i. In fact, $\underline{\underline{\theta}}_i' \equiv \underline{\underline{\theta}}_i^0(S_{\pi^0}(t-1), T_{\pi^0}^i(S_{\pi^0}(t-1)))$ according to the definition of our policy.

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

LEMMA 4: *Under conditions (C1),(C2), policy* π^0 *satisfies:*

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

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

PROOF: We decompose $\widetilde{T}^b_{\pi^0,2}(L_n)$ as follows:

$$\begin{aligned}\widetilde{T}^b_{\pi^0,2}(L_n)&=\sum_{t=2}^{L_n}1\{\pi_t^0=b,b_0^t\in s(\underline{\underline{\theta}}),u_i(\hat{\underline{\underline{\theta}}}^t,\underline{\theta}_i^{'})=u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t)\}\\&=\sum_{t=2}^{L_n}1\{\pi_t^0=b,b_0^t\in s(\underline{\underline{\theta}}),u_i(\hat{\underline{\underline{\theta}}}^t,\underline{\theta}_i^{'})=u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t),u_i(\hat{\underline{\underline{\theta}}}^t,\underline{\theta}_i^{'})>z^*(\underline{\underline{\theta}})-\epsilon\}\\&\quad+\sum_{t=2}^{L_n}1\{\pi_t^0=b,b_0^t\in s(\underline{\underline{\theta}}),u_i(\hat{\underline{\underline{\theta}}}^t,\underline{\theta}_i^{'})=u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t),u_i(\hat{\underline{\underline{\theta}}}^t,\underline{\theta}_i^{'})\leq z^*(\underline{\underline{\theta}})-\epsilon\}.\end{aligned}$$

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_0^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\left\{\pi_t^0=b,b_0^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))}\right\}\\&=\sum_{t=2}^{L_n}1\Big\{\pi_t^0=b,b_0^t\in s(\underline{\underline{\theta}}),u_i(\hat{\underline{\underline{\theta}}}^t,\underline{\theta}_i^{'})=u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t),\\&\qquad\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\Big\}\\&\quad+\sum_{t=2}^{L_n}1\Big\{\pi_t^0=b,b_0^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\Big\}\\&\leq\sum_{t=2}^{L_n}1\left\{\pi_t^0=b,b_0^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}\right\}\\&\quad+\sum_{t=2}^{L_n}1\{\pi_t^0=b,b_0^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

$$c=\frac{\log L_n}{J_i(\underline{\underline{\theta}},\epsilon)-\delta}$$

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

and s integer is equal to

$$\begin{aligned}
&\sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b_0^t \in s(\underline{\theta}), u_i(\hat{\underline{\theta}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\theta}}^t), T_{\pi^0}^i(S_{\pi^0}(t-1)) < c\} \\
&\leq \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, T_{\pi^0}^i(S_{\pi^0}(t-1)) < c\} \\
&= \sum_{t=2}^{L_n} \sum_{s=0}^{\lfloor c/m_i^b \rfloor} 1\{\pi_t^0 = b, T_{\pi^0}^i(S_{\pi^0}(t-1)) = s\, m_i^b + m_i\} \\
&= \sum_{s=0}^{\lfloor c/m_i^b \rfloor} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, T_{\pi^0}^i(S_{\pi^0}(t-1)) = s\, m_i^b + m_i\} \\
&\leq \lfloor c/m_i^b \rfloor + 1 \\
&\leq \frac{c}{m_i^b} + 1 = \frac{\log L_n}{m_i^b(J_i(\underline{\theta}, \epsilon) - \delta)} + 1.
\end{aligned}$$

Thus,

$$\begin{aligned}
&E_{\underline{\theta}} \sum_{t=2}^{L_n} 1\left\{\pi_t^0 = b, b_0^t \in s(\underline{\theta}), u_i(\hat{\underline{\theta}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\theta}}^t), T_{\pi^0}^i(S_{\pi^0}(t-1)) < \frac{\log L_n}{J_i(\underline{\theta}, \epsilon) - \delta}\right\} \\
&\leq \frac{\log L_n}{m_i^b(J_i(\underline{\theta}, \epsilon) - \delta)} + 1. \qquad \textbf{(A.19)}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b_0^t \in s(\underline{\theta}), u_i(\hat{\underline{\theta}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\theta}}^t), J_i(\hat{\underline{\theta}}^t, \epsilon) \leq J_i(\underline{\theta}, \epsilon) - \delta\} \\
&\quad \leq \sum_{t=2}^{L_n} 1\{b_0^t \in s(\underline{\theta}), J_i(\hat{\underline{\theta}}^t, \epsilon) \leq J_i(\underline{\theta}, \epsilon) - \delta\}.
\end{aligned}$$

Then from (C2) and Remark 3 we have that

$$\begin{aligned}
&E_{\underline{\theta}} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b_0^t \in s(\underline{\theta}), u_i(\hat{\underline{\theta}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\theta}}^t), J_i(\hat{\underline{\theta}}^t, \epsilon) \leq J_i(\underline{\theta}, \epsilon) - \delta\} \\
&\leq o(\log L_n). \qquad \textbf{(A.20)}
\end{aligned}$$

Now we have that $u_i(\hat{\underline{\theta}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\theta}}^t) > u_s(\hat{\underline{\theta}}^t, \underline{\theta}_s^{'})$ for any population s which is contained in an optimal BFS of $\underline{\theta}$. Now let $b_0^t = (r, s)$ and obviously $b = (i, s)$, thus we can show the following

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

inequalities:

$$\begin{aligned}
&\sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b_0^t \in s(\underline{\theta}), u_i(\hat{\underline{\theta}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\theta}}^t), u_i(\hat{\underline{\theta}}^t, \underline{\theta}_i^{'}) \leq z^*(\underline{\theta}) - \epsilon\} \\
&\quad \leq \sum_{t=2}^{L_n} 1\{u_s(\hat{\underline{\theta}}^t, \underline{\theta}_s^{'}) \leq z^*(\underline{\theta}) - \epsilon\} \\
&\quad \leq \sum_{t=2}^{L_n} 1\{u_s(\hat{\underline{\theta}}^j, \underline{\theta}_s^{'}) \leq z^*(\underline{\theta}) - \epsilon, \text{ for some } j \leq S_{\pi^0}(t-1)\} \\
&\quad = \sum_{t=2}^{L_n} 1\{|\hat{\underline{\theta}}_s^j - \underline{\theta}_s| > \xi, \text{for some } j \leq S_{\pi^0}(t-1)\}.
\end{aligned}$$

Thus

$$\begin{aligned}
&E_{\underline{\theta}} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b_0^t \in s(\underline{\theta}), u_i(\hat{\underline{\theta}}^t, \underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\theta}}^t), u_i(\hat{\underline{\theta}}^t, \underline{\theta}_i^{'}) \leq z^*(\underline{\theta}) - \epsilon\} \\
&\quad \leq o(\log L_n), \qquad \textbf{(A.21)}
\end{aligned}$$

because

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

since policy π^0 at any block t chooses $b_0^t = (r, s)$ when $\widetilde{T}_{\pi^0}^{b_0^t}(t) \geq \tau(t-1)$.

Finally, it follows from Eqs. (A.19)–(A.21) that

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

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

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

Thus,

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

This completes the proof. ■

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

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 $r = 0, 1, 2, \ldots$

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

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

where $\widetilde{b}_j$ any BFS in F and $0 < \tau < 1/|F|$ 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 $c > 1/(1 - |F|\tau)$ *and* $r \ge r_0$ *then*
(ii) on $A_r \cap B_r$, $b_0^l \in s(\underline{\underline{\theta}})$ *for all* $c^{r-1} \le l - 1 \le c^{r+1}$.
(iii) $E_{\underline{\underline{\theta}}} \widetilde{T}^b_{\pi^0,1}(L_n) = \sum_{t=2}^{L_n} P_{\underline{\underline{\theta}}}^{\pi^0}(b_0^t \notin s(\underline{\underline{\theta}})) = o(\log L_n)$.

PROOF: (i) We have that from (C2)

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

holds for the sample mean of the estimates

$$\hat{\underline{\underline{\theta}}}^l = \frac{\hat{\underline{\underline{\theta}}}^1 + \cdots + \hat{\underline{\underline{\theta}}}^l}{l-1},$$

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 $\lfloor c^{r-1}/\tau^q \rfloor \ge c^{r+1}$. For $t = 0, \ldots, q$ and $l_t = \lfloor c^{r-1}/\tau^t \rfloor$ we define the sets

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

Then by (C3),

$$P_{\underline{\underline{\theta}}}^{\pi^0}(\overline{Q}_t) = o(1/l_t) = o(c^{-r}) \text{ for } t = 0, \ldots, q. \tag{A.22}$$

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

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

for every $b_\alpha \in s(\underline{\underline{\theta}})$ on the event $\bigcap_{0 \le t \le q} Q_t$. Thus, because of $B_r \supset \bigcap_{0 \le t \le q} Q_t$ and Eq. (A.22) we have that $P_{\underline{\underline{\theta}}}^{\pi^0}(\overline{B}_r) = o(c^{-r})$.

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

We note that

$$\max_{b \in s(\underline{\underline{\theta}})} \widetilde{T}^b_{\pi^0}(l) \ge \frac{V_{s(\underline{\underline{\theta}})}^{\pi^0}(l)}{|s(\underline{\underline{\theta}})|}, \tag{A.23}$$

where $|s|$ denotes the number of elements of s.

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press

Consider that at any block l and $c^{r-1} \leq l-1 \leq c^{r+1}$, we have that $u_{\alpha^*}(\hat{\underline{\theta}}^l) \in s(\underline{\theta})$, and $u_{\alpha^*}(\hat{\underline{\theta}}^l)$ corresponds to an optimal BFS $b_{\alpha^*}(\hat{\underline{\theta}}^l)$. Then if $b_0^l \in s(\underline{\theta})$ we have the requested. Now, let assume that $b_0^l \notin s(\underline{\theta})$, and we have that $b_{\alpha^*}(\hat{\underline{\theta}}^l) \in s(\underline{\theta})$ which means that on $A_r \cap B_r$ the policy π^0 chooses from $s(\underline{\theta})$.

Then since $\widetilde{T}_{\pi^0}^{b_0^l}(l) \geq \tau(l-1)$,

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

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

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

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

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

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

$$V_{s(\underline{\theta})}^{\pi^0}(l) \geq \frac{|s(\underline{\theta})|}{|F|}(l-1-c^{r-1}-2|F|) > (|s(\underline{\theta})|)\tau(l-1) \tag{A.24}$$

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

From Eqs. (A.23) and (A.24), we obtain on $A_r \cap B_r$

$$\max_{b \in s(\underline{\theta})} \widetilde{T}_{\pi^0}^b(l) > \tau(l-1) \tag{A.25}$$

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 $A_r \cap B_r$,

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

the last set is nonempty because of Eq. (A.25). Hence, $b_0^l \in s(\underline{\theta})$ 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 $c > 1/(1-|F|\tau)$. Then it follows from (i) and (ii) that for $r \geq r_0$ and $c^{r-1} \leq t-1 \leq c^{r+1}$,

$$P_{\underline{\theta}}^{\pi^0}(b_0^t \notin s(\underline{\theta})) \leq P_{\underline{\theta}}^{\pi^0}(\overline{A}_r) + P_{\underline{\theta}}^{\pi^0}(\overline{B}_r) = o(c^{-r})$$

and therefore

$$\sum_{c^{r-1} \leq t-1 \leq c^{r+1}} P_{\underline{\theta}}^{\pi^0}(b_0^t \notin s(\underline{\theta})) = o(1).$$

Hence,

$$\sum_{t=2}^{L_n} P_{\underline{\theta}}^{\pi^0}(b_0^t \notin s(\underline{\theta})) = o(\log L_n).$$

This completes the proof. ■

https://doi.org/10.1017/S026996481600036X Published online by Cambridge University Press