Optimal data driven resource allocation under multi-armed bandit observations

Burnetas, A.N.; Kanavetas, O.; Katehakis, M.N.

Annals of Operations Research
https://doi.org/10.1007/s10479-025-06554-3

ORIGINAL RESEARCH

Optimal data driven resource allocation under multi-armed bandit observations

Apostolos N. Burnetas[1] · **Odysseas Kanavetas**[2] · **Michael N. Katehakis**[3]

Received: 12 September 2024 / Accepted: 25 February 2025

Abstract
This paper introduces the first asymptotically optimal strategy for a multi armed bandit (MAB) model under side constraints. The side constraints model situations in which bandit activations are limited by the availability of certain resources that are replenished at a constant rate. The main result involves the derivation of an asymptotic lower bound for the regret of feasible uniformly fast policies and the construction of policies that achieve this lower bound, under pertinent conditions. Further, we provide the explicit form of such policies for the case in which the unknown distributions are Normal with unknown means and known variances, for the case of Normal distributions with unknown means and unknown variances and for the case of arbitrary discrete distributions with finite support.

Keywords Stochastic bandits · Sequential decision making · Regret minimization · Sequential allocation

1 Introduction

Consider the problem of sequentially activating one of a finite number of independent bandits, where each activation of a bandit incurs a number of bandit dependent resource utilizations, or activation costs. For each resource type the constraint ensures that the total resource utilized (or equivalently cost incurred) at any time does not exceed the current resource availability (budget). It is assumed that following each activation any unused resource amounts can be carried forward for use in future activations. We also make the assumption that successive activations of each bandit yield independent, among different bandits, identically distributed (iid) random rewards with positive means, and distributions that depend on unknown param-

✉ Odysseas Kanavetas
o.kanavetas@math.leidenuniv.nl

Apostolos N. Burnetas
aburnetas@math.uoa.gr

Michael N. Katehakis
mnk@rutgers.edu

[1] Department of Mathematics, National and Kapodistrian University of Athens, Athens, Greece

[2] Leiden University, Mathematical Institute, Einsteinweg 55, 2333 CC Leiden, The Netherlands

[3] Department of Management Science and Information Systems, Rutgers University, Piscataway, NJ 08854, USA

Published online: 27 March 2025

eters. The objective is to obtain a feasible policy that maximizes asymptotically the total expected rewards or equivalently, minimizes asymptotically a regret function. We develop a class of feasible policies that are shown to be asymptotically optimal within a large class of good policies that uniformly fast (UF) convergent, in the sense of Burnetas and Katehakis (1996) and Lai and Robbins (1985). The results in this paper extend the work in Burnetas et al. (2017) which solved the case where there exists only one type of constraint for all bandits. Further, the class of block-UCB (b-UCB) feasible policies developed here, which achieve the asymptotic lower bound in the regret, have a simpler form and are easier to compute than those in Burnetas et al. (2017). We also refer to Burnetas and Kanavetas (2012) where a consistent policy (i.e., with regret $o(n)$) for the case of a single linear constraint was constructed.

There is an extensive literature on the multi-armed bandit (MAB) problem, cf. Mahajan and Teneketzis (2008), Audibert et al. (2009), Auer and Ortner (2010), Honda and Takemura (2011), Bubeck and Slivkins (2012), Lattimore (2018), Cowan and Katehakis (2020), and references therein. MAB models with a finite exploration budget that limits the number of times one can sample (activate) arms during an initial exploration phase, which is used to identify the optimal arm are considered in Bubeck et al. (2009). Tran-Thanh et al. (2010) considers the problem when both the exploration and exploitation phases are limited by a single budget and establish an upper bound for the loss of a budgeted ϵ- first algorithm for this problem. Badanidiyuru et al. (2018) consider the MAB problem with multiple resource constraints and a finite horizon T, assuming that when a resource is exhausted activations stop. They show how to construct policies with regret in the order of $O(\log T)$, where T is the horizon length. Agarwal and Devanur (2014) provided a more general version of Badanidiyuru et al. (2018) which allows arbitrary concave objectives and convex feasibility constraints. Ding et al. (2013) constructed UF policies (i.e., with regret $O(logn)$) for cases in which activation costs are bandit-dependent iid random variables. Applications of MAB models include problems of online revenue management: Ferreira et al. (2018), Wang et al. (2014), Johnson et al. (2015) of dynamic procurement: Singla and Krause (2013), auctions: Tran-Thanh et al. (2014).

One key difference between these finite resource budget models and the model herein is that in the present model is that we do not have a total budget for each resource fixed at the beginning, but rather for each resource its budget is increased at a constant rate at each activation period. Furthermore, the resource constraints must be satisfied at each period in the sense that for each resource its total budget utilization during the first n periods cannot exceed the total n-period budget available for all n. Thus, for each resource there is one constraint at each time period rather than a single constraint for the entire horizon. In this way the bandit activation problem becomes more restricted and it requires a different approach in the activation policies. For example, if a particular bandit consumes a large amount of some resource at each activation, then after one activation the controller may have to wait for several subsequent periods until the budget of this resource is sufficiently replenished so that the bandit may be activated again. Thus the exploration phase is necessarily intertwined with the exploitation phase due to the structure of the resource constraints.

A second key difference is that we construct a new class of feasible UCB policies and we establish their asymptotic optimality. Asymptotic optimality means that our policies achieve the exact asymptotic lower bound in the regret function and not only in terms of order of magnitude $O(\log n)$, as is typical in finite horizon formulations.

On the applications side, the results herein can be used to solve infinite horizon versions of online network revenue management where the retailer must price several unique products, each of which may consume common resources (e.g., inventories of different products)

that have limited availability and are replenished at a constant rate. For versions of such problems with no resource (inventory) replenishment we refer to Ferreira et al. (2018) and references therein. Additional applications include search-based and targeted advertising online learning, cf. Rusmevichientong and Williamson (2006), Agarwal et al. (2014) and references therein.

For other recent related work we refer to: Guha and Munagala (2007), Tran-Thanh et al. (2012), Thomaidou et al. (2012), Lattimore et al. (2014), Sen et al. (2015), Pike-Burke and Grunewalder (2017), Zhou et al. (2018), Spencer and Lopez (2018) and Denardo et al. (2013), Cowan and Katehakis (2015) Pike-Burke et al. (2018), Lattimore and Szepesvári (2018), Pike-Burke and Grunewalder (2017), Cowan et al. (2023). Similar action constrained optimization problems also arise in MDPs cf. Feinberg (1994), Borkar and Jain (2014), queueing Hordijk and Spieksma (1989), many areas c.f. Perakis and Roels (2008), Levi et al. (2015), Babich and Tang (2016), Feng and Shanthikumar (1994), Chen et al. (2021).

In the sequel we first establish in Theorem 5, a necessary asymptotic lower bound for the rate of increase of the regret function of f-UF policies. We then 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 6. For the development of these policies we use the notion of 'blocks of activations', that essentially allow the implementation of necessary randomizations cf. Feinberg (1994), without violating the feasibility constraints. Then, in Sect. 4.1 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, in Sect. 4.2 for the case of Normal distributions with unknown means and unknown variances and in Sect. 4.3 we do the same for case where the unknown distributions are non parametric, discrete with finite support.

2 Model formulation

Consider k independent bandits, where successive activations of a bandit i, constitute a sequence of i.i.d. random variables $X_1^i, X_2^i, \ldots$. For each fixed i, X_t^i, follows a univariate distribution with density $f_i(\,|\underline{\theta}_i)$ with respect to a nondegenerate measure v. The density $f_i(\,|\,)$ is known and $\underline{\theta}_i$ is a vector of parameters belonging to some set Θ_i. Let $\underline{\underline{\theta}} = (\underline{\theta}_1, \ldots, \underline{\theta}_k)$ denote the set of parameters, $\underline{\underline{\theta}} \in \Theta$, where $\Theta \equiv \Theta_1 \times \ldots \times \Theta_k$. Given $\underline{\underline{\theta}}$ let $\underline{\mu}(\underline{\underline{\theta}}) = (\mu_1(\underline{\theta}_1), \ldots, \mu_k(\underline{\theta}_k))$ be the vector of the expected values, i.e. $\mu_i(\underline{\theta}_i) = E_{\underline{\underline{\theta}}}(X^i)$. The true value $\underline{\underline{\theta}}_0$ of $\underline{\underline{\theta}}$ is unknown. We make the assumptions that outcomes from different bandits are independent and all means $\mu_i(\underline{\theta}_i)$ are positive.

Each activation of bandit i incurs L different types of resource utilization (or cost): c_j^i, $j = 1, \ldots, L$. To avoid trivial cases, we will assume that $L < k$. We also assume that there are two classes of bandits, one with $c_j^i < c_j^0$ for each constraint type j and for any bandit i in this class, and the other one with $c_j^i > c_j^0$ for each constraint type j and for any bandit i in this class. Intuitively, this means that there is at least one bandit that can be activated in each period without violating the strict feasibility constraint we define below. Now, by relabeling the bandits we call as bandit $i = 1$ the bandit which has the maximum number of costs c_j^1 that are the minimum among c_j^i in the constraint type j. Similarly, we label as bandit k the bandit which has the maximum number of costs c_j^k that are the maximum among c_j^i in the constraint type j. Note that $c_j^1 < c_j^0$ for each constraint type j, and $c_j^0 < c_j^k$, again for each constraint type j. For simplicity of the mathematical analysis below we assume that there is

no bandit with activating cost c_j^i that is equal to c_j^0. Equivalently, for each constraint type j, there exists d_j, with $1 \le d_j < k$ and $c_j^{d_j} < c_j^0 < c_j^{d_j+1}$ (note that $d_j = \max\{i \;\; c_j^i < c_j^0\}$).

Following standard terminology, adaptive policies depend only on past activations and observed outcomes. Specifically, let $A_t, X_t, t = 1, 2, \ldots$ denote the bandit activated and the observed outcome at period t. Let $h_t = (\alpha_1, x_1, \ldots, \alpha_{t-1}, x_{t-1})$ denote a history of activations 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 $\pi_t(j, h_t) = P(A_t = j \mid h_t)$. Given h_n, let $T_\pi^\alpha(n)$ denote the number of times bandit α has been activated during the first n periods $T_\pi^\alpha(n) = \sum_{t=1}^n 1\{A_t = \alpha\}$. Let $\mathcal{V}_\pi(n)$ and $\mathcal{C}_{j,\pi}(n)$ be respectively the total reward earned and total type j resource utilized (cost incurred) up to period n, i.e.,

$$\mathcal{V}_\pi(n) = \sum_{i=1}^{k} \sum_{t=1}^{T_\pi^i(n)} X_t^i, \tag{2.1}$$

$$\mathcal{C}_{j,\pi}(n) = \sum_{i=1}^{k} \sum_{t=1}^{T_\pi^i(n)} c_j^i. \tag{2.2}$$

We call an adaptive policy feasible if

$$\mathcal{C}_{j,\pi}(n)/n \le c_j^0, \;\; \forall j = 1, \ldots, L, \; \forall n = 1, 2, \ldots \tag{2.3}$$

The objective is to obtain a feasible policy π that maximizes asymptotically $E_{\underline{\underline{\theta}}} \mathcal{V}_\pi(n)$, $\forall \underline{\underline{\theta}} \in \underline{\underline{\Theta}}$, or equivalently, it minimizes asymptotically the regret function $R_\pi(\underline{\underline{\theta}}, n)$ cf. (3.1).

2.1 Optimal solution under known parameters

It follows from standard theory of MDPs cf. Derman (1970), that if all parameters $\underline{\underline{\theta}}$ were known, the optimal activation(s) (the same in all periods) for maximizing the expected average reward are obtained as the solution to the following linear program (LP).

$$\begin{aligned} z^*(\underline{\underline{\theta}}) = \max & \sum_{i=1}^{k} \mu_i(\underline{\theta}_i) x_i \\ \textit{subject to} \quad & \sum_{i=1}^{k} c_1^i x_i + y_1 = c_1^0 \\ & \vdots \\ & \sum_{i=1}^{k} c_L^i x_i + y_L = c_L^0 \\ & \sum_{i=1}^{k} x_i = 1 \\ & x_i \ge 0, \forall i \;\; y_j \ge 0, \forall j, \end{aligned} \tag{2.4}$$

where the variables x_i, for $i = 1, \ldots, k$, represent the activation probabilities for bandit i of an optimal randomized policy.

Thus, a basic matrix B is an $(L+1)\times(L+1)$ matrix that consists of one or at most $L+1$ bandit (and slack) variables x_i (and y_i); recall that $L < k$. Note that any basic feasible solution (BFS) corresponding to such a choice of the matrix B is uniquely determined by the vector corresponding to the choice of basic bandit variables: $b = \{i_1, \ldots, i_j\}$, $j = 1, \ldots, L+1$. For simplicity, the sequel we will not distinguish between B and b, since if one knows one he knows the other. Thus, the vector b uniquely determines a corresponding (possibly randomized) activation policy with randomization probabilities $x_{i_1}, \ldots, x_{i_j}, j = 1, \ldots, L+1$ in b. We use K to denote the set of bandits corresponding to a feasible choice of b, for simplicity written as

$$K = \{b \,:\, b = \{i_1, \ldots, i_j\},\ j = 1, \ldots, L+1\}.$$

Given our assumptions on the $c_j^i s$, it follows that the feasible region of (2.4) is nonempty and bounded, hence K corresponds to a finite number of BFS.

In the sequel it will be more convenient to work with the dual problem DLP stated below.

$$\begin{aligned} z_D^*(\underline{\theta}) = \min\ & c_1^0 g_1 + \ldots + c_L^0 g_L + g_{L+1} \\ subject\ to\quad & c_1^1 g_1 + \ldots + c_L^1 g_L + g_{L+1} \geq \mu_1(\underline{\theta}_1) \\ & \vdots \\ & c_1^k g_1 + \ldots + c_L^k g_L + g_{L+1} \geq \mu_k(\underline{\theta}_k) \\ & g_j \geq 0,\ j = 1, \ldots, L,\ \ g_{L+1} \in \mathbb{R}. \end{aligned}$$

For a basic matrix B of LP, we let $v^B = (g_1^B, \ldots, g_L^B, g_{L+1}^B)$ denote the dual vector corresponding to B, i.e., $v^B = \mu_B(\underline{\theta})B^{-1}$, where $\mu_B(\underline{\theta})$ contains the means of the bandits given by the choice of B.

A BFS is optimal if and only if the reduced costs (dual slacks) for the corresponding basic matrix B are all nonnegative, i.e.,

$$\phi_\alpha^B(\underline{\theta}) \equiv c_1^\alpha g_1^B + \ldots + c_L^\alpha g_L^B + g_{L+1}^B - \mu_\alpha(\underline{\theta}_\alpha) \geq 0,\ \alpha = 1, \ldots, k.$$

Note that it is easy to show that the reduced cost can be expressed as a linear combination of the bandit means, i.e., $\phi_\alpha^B(\underline{\theta}) = \underline{w}_\alpha^B \underline{\mu}(\underline{\theta})$, 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 $O^*(\underline{\theta})$ to denote the set of choices of b corresponding to optimal solutions of the LP for a vector $\underline{\mu}(\underline{\theta})$, i.e., $O^*(\underline{\theta}) = \{b \in K :$ b corresponds to an optimal BFS$\}$.

3 Optimal policies under unknown parameters

3.1 The regret function

In this subsection we consider the case in which $\underline{\theta}$ is unknown and define the regret $R_\pi(\underline{\theta}, n)$ of a policy π as the finite horizon loss in expected reward with respect to the optimal policy π^* corresponding to the case in which $\underline{\theta}$ is known, i.e.,

$$\begin{aligned} R_\pi(\underline{\theta}, n) &= n z^*(\underline{\theta}) - E_{\underline{\theta}} \mathcal{V}_\pi(n) \\ &= n z^*(\underline{\theta}) - \sum_{j=1}^{k} \mu_j(\underline{\theta}_j) E_{\underline{\theta}} T_\pi^j(n). \end{aligned} \tag{3.1}$$

We now state the following. A feasible policy π is called consistent if $R_\pi(\underline{\underline{\theta}}, n) = o(n),\ n \to \infty,\ \forall\, \underline{\underline{\theta}} \in \Theta$, and it is called uniformly fast (f-UF) if $R_\pi(\underline{\underline{\theta}}, n) = o(n^a),\ n \to \infty,\ \forall\, a > 0,\ \forall\, \underline{\underline{\theta}} \in \Theta$.

Following the approach in Burnetas et al. (2017), we will establish in Theorem 5 below a lower bound $M(\underline{\underline{\theta}})$ for the regret of any f-UF policy and construct a block UCB policy π^0 which is f-UF and its regret achieves this lower bound, i.e.,

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

Thus, it will be shown that policy π^0 is asymptotically optimal.

3.2 Lower bound for the regret

For any optimal basis choice $b = \{i_1, \ldots, i_j\} \in O^*(\underline{\underline{\theta}})$, and $\alpha \in b$, we define the sets $\Delta\Theta_\alpha(\underline{\underline{\theta}})$ and $D(\underline{\underline{\theta}})$, as follows

$$\begin{aligned} \Delta\Theta_\alpha(\underline{\underline{\theta}}) &= \{\underline{\theta}'_\alpha \in \Theta_\alpha : O^*(\underline{\underline{\theta}}') = \{b\}\}, \\ D(\underline{\underline{\theta}}) &= \{\alpha :\ \alpha \notin b \text{ for any } b \in O^*(\underline{\underline{\theta}}) \text{ and } \Delta\Theta_\alpha(\underline{\underline{\theta}}) \neq \emptyset\}, \end{aligned} \tag{3.2}$$

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 $\underline{\theta}_\alpha$. Note that the first set consists of all values of Θ_α under which the problem with known parameters under the perturbed $\underline{\underline{\theta}}'$ has a unique optimal solution that includes bandit α. The second set $D(\underline{\underline{\theta}})$, consists of all bandits that do not appear in any optimal solution under parameter set $\underline{\underline{\theta}}$ but, by changing only the parameter vector $\underline{\theta}_\alpha$, there is uniquely optimal solution that contains them.

We next define the minimum distance of a parameter vector $\underline{\theta}_\alpha$ to a new parameter vector $\underline{\theta}'_\alpha$ which makes bandit α to become optimal and hence appear in the unique optimal solution when its parameter becomes $\underline{\theta}'_\alpha$.

$$K_\alpha(\underline{\underline{\theta}}) = \inf\{I(\underline{\theta}_\alpha, \underline{\theta}'_\alpha) :\ \underline{\theta}'_\alpha \in \Delta\Theta_\alpha(\underline{\underline{\theta}})\}, \tag{3.3}$$

where, $I(\underline{\theta}_\alpha, \underline{\theta}'_\alpha)$ denotes the Kullback–Leibler distance for the distributions $f(\cdot, \underline{\theta}_\alpha)$, and $f(\cdot, \underline{\theta}'_\alpha)$ i.e.,

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

The next Lemma establishes lower bounds for the new mean $\mu_\alpha(\underline{\theta}'_\alpha)$ under the changed parameter vector $\underline{\theta}'_\alpha$ in terms of the quantity $\mu^*_\alpha(\underline{\underline{\theta}}) = \phi^B_\alpha(\underline{\underline{\theta}}) + \mu_\alpha(\underline{\theta}_\alpha)$. The proof is specialized and not the focus of this paper, and is relegated to the appendix.

Lemma 1 *For any optimal matrix B under $\underline{\underline{\theta}}$, such that for any $\underline{\theta}'_\alpha \in \Delta\Theta_\alpha(\underline{\underline{\theta}})$ the following is true*

$$\begin{aligned} &\phi^B_j(\underline{\underline{\theta}}') = \phi^B_j(\underline{\underline{\theta}}) \ge 0,\ \forall\ j \neq \alpha, \text{ and} \\ &\phi^B_\alpha(\underline{\underline{\theta}}') = \mu^*_\alpha(\underline{\underline{\theta}}) - \mu_\alpha(\underline{\theta}'_\alpha) < 0. \end{aligned}$$

The above and Eqs. (3.2), (3.3) imply that

$$K_\alpha(\underline{\underline{\theta}}) = \inf\{I(\underline{\theta}_\alpha, \underline{\theta}'_\alpha) :\ \underline{\theta}'_\alpha \in \Theta_\alpha,\ \mu^*_\alpha(\underline{\underline{\theta}}) < \mu_\alpha(\underline{\theta}'_\alpha)\}. \tag{3.4}$$

In order to establish a lower bound on the regret we need to express it as:

$$R_\pi(\underline{\underline{\theta}}, n) = \sum_{j=1}^{k} \phi_j^B(\underline{\underline{\theta}}) E_{\underline{\underline{\theta}}} T_\pi^j(n) + \sum_{i=1}^{L} g_i^B \sum_{j=1}^{k} (c_i^0 - c_i^j) E_{\underline{\underline{\theta}}} T_\pi^j(n), \ \forall \underline{\underline{\theta}} \in \Theta, \tag{3.5}$$

and any optimal basic matrix B, where the above expression follows from the LP and DLP relations since B is an optimal basis: $z^*(\underline{\underline{\theta}}) = \sum_{i=1}^{L} c_i^0 g_i^B + g_{L+1}^B$ and $\phi_i^B(\underline{\underline{\theta}}) = \sum_{j=1}^{L} c_j^i g_j^B + g_{L+1}^B - \mu_i(\underline{\theta}_i)$.

Both terms of the right side of (3.5) are nonnegative, the first due to optimality of B and the second due to the feasibility of B. It follows that a necessary and sufficient condition for a policy π to be f-UF is that for all $\underline{\underline{\theta}} \in \Theta$ and any optimal B under $\underline{\underline{\theta}}$ the following two relations hold.

$$\phi_j^B(\underline{\underline{\theta}}) \overline{\lim}_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T_\pi^j(n)}{n^a} = 0, \ \text{for all } a > 0, \ j \notin b, \tag{3.6}$$

$$\sum_{i=1}^{L} g_i^B \sum_{j\in b} (c_i^0 - c_i^j) \overline{\lim}_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} T_\pi^j(n)}{n^a} = 0 \ \text{for all } a > 0. \tag{3.7}$$

The following lemma and proposition are used to establish in Lemma 4 a lower bound for the activation frequencies of any f-UF policy. They readily imply the lower bound of such polices for the regret in Theorem 5. The proof of the lemma is relegated to the appendix.

Lemma 2 *If there is a uniquely optimal $b \in O^*(\underline{\underline{\theta}})$. Then the following hold.*

(i) If $b = \{i_1, \ldots, i_j\}$, $j = 2, \ldots, L+1$, then $g_i^B > 0$, for $j-1$ out of the L resource constraints, and equal to 0 for the remaining resource constraints.

(ii) When b is a singleton, i.e., $b = \{i_1\}$, then $g_i^B = 0$, for every resource constraint $i = 1, \ldots, L$, i.e., only g_{L+1}^B is positive.

The next proposition establishes that a f-UF policy is such that $\forall \underline{\underline{\theta}} \in \underline{\Theta}$, it must be true that the number of activations from each bandit $\alpha \in D(\underline{\underline{\theta}})$ are at least β_n, for some sequence of positive constants $\beta_n = o(n)$. Its proof is given in the appendix.

Proposition 3 *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*

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

The next Lemma follows using a change of measure from $\underline{\underline{\theta}}'$ to $\underline{\underline{\theta}}$ as done in Burnetas and Katehakis (1996) and in Lai and Robbins (1985).

Lemma 4 *If $P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) < \beta_n] = o(n^{a-1})$, for all $a > 0$ and a positive sequence $\beta_n = o(n)$ then*

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

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

Proof If we take $\beta_n = \log n / K_\alpha(\underline{\underline{\theta}})$ then Proposition 3 implies $P_{\underline{\underline{\theta}}'}[T_\pi^\alpha(n) < \log n / K_\alpha(\underline{\underline{\theta}})] = o(n^{a-1})$. Now, using the change of measure from $\underline{\underline{\theta}}'$ to $\underline{\underline{\theta}}$ and the same arguments as in Burnetas and Katehakis (1996) we have that

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

□

We next define the constant $M(\underline{\underline{\theta}})$ 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 5 *If π is an f-UF policy then*

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

Proof By Lemma 4 and using the Markov inequality, we obtain that if π is f-UF, then

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

Also, we have from Lemma 2 that $g_i^B \geq 0$ and from Eq. (2.3), we have that $nc_i^0 - E_{\underline{\underline{\theta}}} C_{i,\pi}(n) \geq 0$, for all n, i. Finally, we have that the optimal bandits under $\underline{\underline{\theta}}$ have $\phi_j^B(\underline{\underline{\theta}}) = 0$. These observations together with the above two relations suffice to complete the proof if we recall that $R_\pi(\underline{\underline{\theta}}, n) = \sum_{j=1}^k \phi_j^B(\underline{\underline{\theta}}) E_{\underline{\underline{\theta}}} T_\pi^j(n) + \sum_{i=1}^L g_i^B \sum_{j=1}^k (c_i^0 - c_i^j) E_{\underline{\underline{\theta}}} T_\pi^j(n)$. □

3.3 Blocks and block based policies

We consider a class of policies such that activation is performed in groups of periods called *activation blocks* as defined below so as total resource utilization of activations in each block satisfies all the resource constraints of (2.3). For each constraint j we first define the differences

$$\delta_j^i \equiv c_j^i - c_j^0.$$

Note that δ_j^i expresses the net effect of a single activation of bandit i on the corresponding resource c_j^0. This effect is a reduction in the resource c if $\delta_j^i > 0$, and a surplus if $\delta_j^i < 0$, that can be carried over to subsequent periods.

Thus, for any period the feasibility constraint of (2.3) can be written as

$$\frac{1}{n}\sum_{t=1}^n \delta_j^{A_t} \leq 0,\ \forall n,\ j = 1, \ldots, L.$$

Since δ_j^i is assumed to be rational, for each $i = 1, \ldots, k$ and $j = 1, \ldots, L$ and there is a finite number of them we may assume, without loss of generality, that they are all integers, since we have assumed the same for the coefficients and right sides of the constraints.

Let $\mathcal{N} = \{1, \ldots, k\}$ be the set of all bandits. Intuitively, for a specific constraint "low resource utilization" bandits in $\mathcal{N}$ (i.e., bandits with small c_j^i) must be sampled often enough to

accumulate resources (by carrying over 'surplusess') in order to make possible the activation of "high resource utilization" bandits. Mathematically it suffices to find $\{y_j^i, i \in \mathcal{N}\}$ such that bandit $i = 1$ (where the $\delta_j^1 = \min_i \delta_j^i$ for the maximum number of different constraint types $j = 1, \ldots, L$) is sampled y_j^1 times and each bandit $i \in \mathcal{N}\backslash\{1\}$ is sampled once ($y_j^i = 1$, for all $i \in \mathcal{N}\backslash\{1\}$ and $j = 1, \ldots, L$), and $\sum_{i\in\mathcal{N}} y_j^i \delta_j^i \leq 0$, $y_j^i \in \mathbb{N}$, $\forall\, i \in \mathcal{N}$ and $j = 1, \ldots., L$. Any block with $y_j^1 = y_*^1 = \max_j y_j^1$ for all $j = 1, \ldots, L$ satisfying the previous properties is feasible with respect to the constraints of (2.3).

Using the above remarks, we next define the Initial Sampling Block (ISB) and the Linear Programming Blocks (LPBs) to construct a class of block feasible policies $\mathcal{B} = \{\tilde{\pi}\ \tilde{\pi}$ is feasible$\}$ as follows.

ISB block: A policy $\tilde{\pi} \in \mathcal{B}$ starts with an ISB block during which all bandits $\{1, \ldots, k\}$ are sampled at least a predetermined number n_0 of times, while the constraints of (2.3) are satisfied sample path-wise. This block is necessary in order to obtain initial estimates $\mu_j(\hat{\underline{\theta}}_j)$ of $\mu_j(\underline{\theta}_j)$ for all bandits. This ISB block has length of $y_*^1 + (k-1)$, defined above.

LPB blocks: After the completion of an ISB block the $\tilde{\pi}$ policy chooses any b (where $b = \{i_1\}$ or $b = \{i_1, \ldots, i_j\}$) that corresponds to a BFS of the LP and continues activating any of the bandits in b for a block of time periods LPB(b), where LPB(b) is defined as follows.

i) When $b = \{i_1\}$. In this case due to the feasibility of b we must have $c_j^{i_1} < c_j^0$, for all $j = 1, \ldots, L$, and its activating frequency must be equal to 1, i.e., $x_{i_1} = 1$. In this case we define the LPB(b) block to have length equal to 1 and $\tilde{\pi}$ activates bandit i_1 once, i.e., $m_{i_1}^b = 1$.

ii) When $b = \{i_1, \ldots, i_j\}$, we have the positive randomization probabilities $\{x_l\}_{l\in b}$. According to our assumption for rational coefficients in the resource constraints, these randomization probabilities can be written in the form of $x_l = \frac{m_l^b}{X(b)}$, where $X(b) = \sum_{i=i_1}^{i_j} m_{i_j}^b$ is the least common denominator, for integer numbers $m_{i_j}^b$ which we take to be the number of activations of bandit i_j within this LPB(b). In this case we take the length of the LPB(b) block to be equal to: $\sum_{i=i_1}^{i_j} m_{i_j}^b$, and we take $\tilde{\pi}$ to activate m_l^b number of times each bandit $l \in b$, in this way ensuring the constraint feasibility of $\tilde{\pi}$ within the block.

Remark 1 In the above definition of block activations for any b that corresponds to a BFS of the LP we defined the integers m_α^b to be the number of activations from bandit α within a LPB(b). Note that in a computational implementation the solution of an LP may be given in decimals. In this case, one cannot compute an exact least common denominator for the randomization probabilities which is important since the denominator defines the length of the LP block. However, every time one solves a LP one knows the specific constraint equations that correspond to the optimal basic matrix B. Thus one has a subsystem of equations of the form $Bx = c$. Using the determinants expression of the solution one can compute an integer denominator for the randomization probabilities, under the assumption of rational coefficients. Then one can find the least common denominator that is an integer and can be used as the length of the block of activations $\sum_{i=i_1}^{i_j} m_{i_j}^b$ corresponding to B.

The definition of any $\tilde{\pi} \in \mathcal{B}$ policy is completed by continuing activations of bandits in $\mathcal{N}$ by repeating choices of b as above. In the sequel the choices of b will be based on all collected data up to the start of the 'current block' and thus $\tilde{\pi} \in \mathcal{B}$ will be well defined adaptive policies. In what follows we will restrict attention to such policies in $\mathcal{B}$ and for notational simplicity we will simply write π in place of $\tilde{\pi}$, when there is no risk for confusion.

3.4 Regret of block based policies

In this section we define the regret of block based policies and establish its relation with the initial regret of (3.1). 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 in the first $2, \ldots, l$ blocks each corresponds to a single feasible b and we can write $\sum_{b\in K} \widetilde{T}_\pi^b(l) = l - 1$.

Let $S_\pi(l)$ be the total length of first l blocks (including the ISB block) and let $L_n = L_{\tilde{\pi}}(n)$ denote the number of completed blocks in n periods. It can be shown that

$$T_\pi^\alpha(S_\pi(l)) = \sum_{b\in K\,:\alpha\in b} m_\alpha^b\, \widetilde{T}_\pi^b(l) + m_\alpha^0, \tag{3.8}$$

where m_α^b was defined above and m_α^0 is the number of activations of bandit α in the ISB block (i.e., $m_1 = y_*^1$ and $m_\alpha^0 = m_\alpha^0 1$ for all $\alpha \in \mathcal{N}\backslash\{1\}$).

Note that when $\underline{\underline{\theta}}$ is known the quantity $E_{\underline{\underline{\theta}}} S_\pi(l)\, z^*(\underline{\underline{\theta}})$ represents the total expected reward under an optimal policy. When $\underline{\underline{\theta}}$ is unknown the quantity $E_{\underline{\underline{\theta}}} \sum_{j=1}^k \sum_{b\in K} \mu_j(\underline{\theta}_j)\, m_j^b\, \widetilde{T}_\pi^b(l) + \sum_{j=1}^k \mu_j(\underline{\theta}_j) m_j^0$ represents the total expected reward under a block policy π. Thus, we can define the regret of a block policy π as

$$\widetilde{R}_\pi(\underline{\underline{\theta}}, l) = E_{\underline{\underline{\theta}}} S_\pi(l)\, z^*(\underline{\underline{\theta}}) - E_{\underline{\underline{\theta}}} \sum_{j=1}^k \sum_{b\in K:\, j\in b} \mu_j(\underline{\theta}_j)\, m_j^b\, \widetilde{T}_\pi^b(l) \; - \sum_{j=1}^k \mu_j(\underline{\theta}_j) m_j^0. \tag{3.9}$$

Also note that in a period n the length of the completed blocks $S_\pi(L_n))$ is less than or equal to n. When $S_\pi(L_n)) = n$, then the number of activations of bandit α up to period n is equal to the number of activations up to the last completed block, i.e., $T_\pi^\alpha(S_\pi(L_n)) = T_\pi^\alpha(n)$. Otherwise, if $S_\pi(L_n)) < n$ (i.e., period n is within the last block which is uncompleted) then $T_\pi^\alpha(S_\pi(L_n)) < T_\pi^\alpha(n)$. Note that there is a finite constant M_α that is equal to the maximum number of times that bandit α appears in every feasible block (i.e., feasible basis). This number allows one to obtain an upper bound on on $T_\pi^\alpha(n)$ when $S_\pi(L_n)) < n$, i.e., $T_\pi^\alpha(n) \leq T_\pi^\alpha(S_\pi(L_n)) + M_\alpha$. Summarizing the above arguments we have:

$$T_\pi^\alpha(S_\pi(L_n)) \leq T_\pi^\alpha(n) \leq T_\pi^\alpha(S_\pi(L_n)) + M_\alpha, \tag{3.10}$$

The definition of (3.9) and (3.10) yield the following relation for the two types of regret,

$$\begin{aligned} &\widetilde{R}_\pi(\underline{\underline{\theta}}, L_n) + (n - E_{\underline{\underline{\theta}}} S_\pi(L_n))\, z^*(\underline{\underline{\theta}}) - \sum_{j=1}^k M_j\, \mu_j(\underline{\theta}_j) \leq R_\pi(\underline{\underline{\theta}}, n) \\ &\leq \widetilde{R}_\pi(\underline{\underline{\theta}}, L_n) + (n - E_{\underline{\underline{\theta}}} S_\pi(L_n))\, z^*(\underline{\underline{\theta}}). \end{aligned} \tag{3.11}$$

The relations in (3.11) imply the following relation between the two regret functions,

$$\overline{\lim}_{n\to\infty} \frac{R_\pi(\underline{\underline{\theta}}, n)}{\log n} = \overline{\lim}_{n\to\infty} \frac{\widetilde{R}_\pi(\underline{\underline{\theta}}, L_n)}{\log L_n}. \tag{3.12}$$

From (3.12), it follows that to show that a policy achieves the lower bound for $R_\pi(\underline{\underline{\theta}}, n)$ it suffices to show that it achieves the lower bound for $\widetilde{R}_\pi(\underline{\underline{\theta}}, L_n)$.

3.5 Asymptotically optimal block UCB policy

In this section we provide a general method to construct asymptotically optimal policies π^0. We call them Z-UCB policies and such policies 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 $\hat{\underline{\theta}}^l$ which give an optimal solution $b(\hat{\underline{\theta}}^l)$ and corresponding optimal basis matrix $\hat{B}$ of LP$(\hat{\underline{\theta}}^l)$. Using the optimal solution of the linear program $b(\hat{\underline{\theta}}^l)$, we can compute for any block l and for every bandit $\alpha \in \{1, \ldots, k\}$ the inflations $v_\alpha = v_\alpha(\hat{\underline{\theta}}^l)$ of $\mu_\alpha(\hat{\underline{\theta}}^l_\alpha)$ and the set $\Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}$ of all bandits α that according to Lemma 1, after the inflation of the mean only of bandit α to v_α may be in an optimal solution, as follows.

$$v_\alpha(\hat{\underline{\theta}}^l) = \sup_{\underline{\theta}'_\alpha} \{\mu_\alpha(\underline{\theta}'_\alpha) \,:\, I(\hat{\underline{\theta}}^l_\alpha, \underline{\theta}'_\alpha) \leq \frac{\log S_\pi(l-1)}{T_\pi^\alpha(S_\pi(l-1))}\}, \tag{3.13}$$

$$\Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)} = \{\alpha \,:\, \mu^*_\alpha(\hat{\underline{\theta}}^l) = \phi^B_\alpha(\hat{\underline{\theta}}) + \mu_\alpha(\hat{\underline{\theta}}_\alpha) < v_\alpha(\hat{\underline{\theta}}^l)\}. \tag{3.14}$$

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: $u_\alpha(\hat{\underline{\theta}}^l) = u_\alpha(\hat{\underline{\theta}}^l, s, t)$ as follows.

$$u_\alpha(\hat{\underline{\theta}}^l, s, t) = \max_{\underline{\theta}'_\alpha} \{z^{b_\alpha(\hat{\underline{\theta}}^l, \underline{\theta}'_\alpha)} : I(\hat{\underline{\theta}}^l_\alpha, \underline{\theta}'_\alpha) \leq \frac{\log s}{t}\}, \tag{3.15}$$

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 bandit α by $\underline{\theta}'_\alpha$.

We next state the asymptotically optimal Z-UCB policy. It is based on the computation of the quantities $v_\alpha(\hat{\underline{\theta}}^l)$, $\Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}$ and $u_\alpha(\hat{\underline{\theta}}^l, s, t)$ above, where $\hat{\underline{\theta}}^l$ and $\hat{B} = \hat{B}(\hat{\underline{\theta}}^l)$, are updated at the end of each block l. The update for $v_\alpha(\hat{\underline{\theta}}^l)$ requires in general the solution of simple optimization problem with the single convex constraint and objective that depends on the functional relation of $\mu_\alpha(\underline{\theta}_\alpha)$ on $\underline{\theta}_\alpha$. The updates for $\Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}$ and $u_\alpha(\hat{\underline{\theta}}^l, s, t)$ are simple and fast since they do not require solving any complex optimization problems.

Z-UCB POLICY π^0:

Step 1 Employ one ISB block in order to have one estimate $\hat{\underline{\theta}}_a$ from each bandit a. Then, update the vector of estimates $\hat{\underline{\underline{\theta}}}^2$, and the statistics $S_{\pi^0}(1)$ and $T^{\alpha}_{\pi^0}(S_{\pi^0}(1))$.

Step 2 For block l ($l > 1$), employ an LPB($\pi^0(\hat{\underline{\underline{\theta}}}^l)$) block, defined below.
Given the history until the beginning of block l we have the vector of estimates $\hat{\underline{\underline{\theta}}}^l$ and the statistics $S_{\pi^0}(l-1)$ and $T^{\alpha}_{\pi^0}(S_{\pi^0}(l-1))$. Based on these, compute b^l_0, and $z^{b^l_0}(\hat{\underline{\underline{\theta}}}^l)$. Then compute $v_\alpha(\hat{\underline{\underline{\theta}}}^l)$'s by Eq. (3.13) for every bandit $\alpha = \{1, \ldots, k\}$, and $\Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)}$, by Eq. (3.14).
Now there are two cases:
(i) If $\Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)} \neq \emptyset$, for every $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)}$ compute the indices

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

and the corresponding uniquely optimal BFS: $b^0_\alpha(\hat{\underline{\underline{\theta}}}^l) = b_\alpha(\hat{\underline{\underline{\theta}}}^l, \underline{\theta}^0_\alpha(s,t))$ in Eq. (3.15).
The Z-UCB policy π^0 employs the LPB block which corresponds to the index:

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

(ii) If $\Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)} = \emptyset$, take $\pi^0(\hat{\underline{\underline{\theta}}}^l) = b^l_0$, meaning that no bandit offers a better solution under an increase, the Z-UCB policy π^0 employs the LPB(b^l_0) block.

Remark 2 With apologies for the notation we have used b^l_0 as the initial optimal solution of LP($\hat{\underline{\underline{\theta}}}^l$), for the l block, and $b^0_\alpha(\hat{\underline{\underline{\theta}}}^l)$ as the inflated solutions of b^0_l, cf. (3.13) and (3.14).

In words, the Z-UCB policy firstly employs an ISB block in order to have estimates for each bandit. Then, for $l = 1, 2, \ldots$, it employs only LPB blocks according to the following. Firstly, the Z-UCB policy finds the initial optimal solution b^l_0 (with basis matrix $\hat{B}$) which is a solution based on the estimates up to this block. Secondly, it computes the indices $u_\alpha(\hat{\underline{\underline{\theta}}}^l, s, t)$ (of Eq. (3.15)) for all bandits $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)}$, by varying only $\underline{\theta}'_\alpha$. Thus, for any $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)}$, we have the corresponding new uniquely optimal BFS: $b^0_\alpha(\hat{\underline{\underline{\theta}}}^l)$. Then, the Z-UCB policy chooses to employ the LPB block which corresponds to the highest z value among $b^0_\alpha(\hat{\underline{\underline{\theta}}}^l)$, $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\underline{\theta}}}^l)}$.

The main result of this paper is that under the following conditions policy π^0 is asymptotically optimal in the class of f-UF policies.

To state condition C1 below we need the definition of the bandit unobservable quantities: $J_\alpha(\underline{\underline{\theta}}, \epsilon)$, as follows. For any $\underline{\underline{\theta}} \in \Theta$, $\epsilon > 0$, an optimal matrix B under $\underline{\underline{\theta}}$, as in Lemma 1,

we define: $\Theta'_\alpha(\epsilon) = \{\underline{\theta}'_\alpha : \mu^*_\alpha(\underline{\theta}) - \epsilon < \mu_\alpha(\underline{\theta}'_\alpha)\}$ 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 $\alpha \in \Phi_l^{(\hat{B}, \hat{\underline{\theta}}^l)}$,

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

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

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

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

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

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

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

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

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

Theorem 6 *Under conditions (C1),(C2), and (C3), and policy π^0, defined above, the following holds.*

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

Proof From (3.12), to establish the above inequality one can prove the same bound for $\overline{\lim}_{n\to\infty} \widetilde{R}_\pi(\underline{\theta}, L_n)/\log L_n$. Now, from (3.8) and (3.9) we have the following

$$\begin{aligned}\widetilde{R}_\pi(\underline{\theta}, L_n) &= E_{\underline{\theta}}\Big(\sum_{j=1}^k \sum_{b\in K: j\in b} m^b_j\, \widetilde{T}^b_\pi(L_n) + \sum_{j=1}^k m^0_j\Big) z^*(\underline{\theta}) + \sum_{j=1}^k m^0_j - E_{\underline{\theta}} \sum_{j=1}^k \sum_{b\in K: j\in b} \mu_j(\underline{\theta}_j)\, m^b_j\, \widetilde{T}^b_\pi(L_n) \\ &\quad - \sum_{j=1}^k \mu_j(\underline{\theta}_j) m^0_j,\end{aligned}$$

which using the relations: $z^*(\underline{\theta}) = \sum_{i=1}^L c^0_i g^B_i + g^B_{L+1}$ and $\phi^B_i(\underline{\theta}) = \sum_{j=1}^L c^i_j g^B_j + g^B_{L+1} - \mu_i(\underline{\theta}_i)$ and after some algebra can be rewritten as:

$$\widetilde{R}_\pi(\underline{\theta}, L_n) = \sum_{j=1}^k \sum_{b\in K: j\in b} m^b_j\, \phi^B_j(\underline{\theta}) E_{\underline{\theta}} \widetilde{T}^b_\pi(L_n) + \sum_{i=1}^L [E_{\underline{\theta}} S_{\pi^0}(L_n) c^0_i - E_{\underline{\theta}} C_{i,\pi^0}(S_\pi(L_n))] g^B_i.$$

Now, since $\phi^B_j(\underline{\theta}) = 0$ for optimal j, to prove the inequality for the regret it is sufficient to show that for policy π^0 the relations below hold.

$$\overline{\lim}_{n\to\infty} \frac{E_{\underline{\theta}} \widetilde{T}^b_{\pi^0,2}(L_n)}{\log L_n} \le \frac{1}{m^b_i K_i(\underline{\theta})}, \text{ for all } i \in D(\underline{\theta}), i \in b, b \notin O^*(\underline{\theta}), \quad (3.16)$$

$$\overline{\lim}_{n\to\infty}\frac{E_{\underline{\underline{\theta}}}\widetilde{T}^b_{\pi^0,2}(L_n)}{\log L_n}=0,\ \text{for all } i\notin D(\underline{\underline{\theta}}), i\in b, b\notin O^*(\underline{\underline{\theta}}), \tag{3.17}$$

$$\overline{\lim}_{n\to\infty}\frac{E_{\underline{\underline{\theta}}}\widetilde{T}^b_{\pi^0,1}(L_n)}{\log L_n}=0,\ \text{for all } b\notin O^*(\underline{\underline{\theta}}), \tag{3.18}$$

$$\sum_{i=1}^{L}[E_{\underline{\underline{\theta}}}S_{\pi^0}(L_n)c_i^0-E_{\underline{\underline{\theta}}}C_{i,\pi^0}(S_\pi(L_n))]g_i^B=o(\log n), \tag{3.19}$$

where B is an optimal basis under $\underline{\underline{\theta}}$, and $\widetilde{T}^b_{\pi^0,1}(L_n)$ and $\widetilde{T}^b_{\pi^0,2}(L_n)$ can be obtained by the fragmentation of $\widetilde{T}^b_{\pi^0}(L_n)$ as:

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

The proof of the inequalities (3.16), (3.17) and (3.18) is given in Lemma 7 in the appendix. (3.19) follows from Lemma 2 which shows that for the constraints that do not obtain the optimal solution the corresponding g^B are equal to 0 and that a block based policy for the optimal bandits uses the whole amount c^0 of the constraints that give the optimal solution. □

Remark 3 According to Remark 4*b* in Burnetas and Katehakis (1996) condition (C2) is equivalent to C2′ below which is often easier to verify.
(**C2′**) $\forall\ \delta>0$, as $t\to\infty$

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

4 Applications

4.1 Normal distributions with unknown means and known variances

Assume the observations X_α^j from bandit α are normally distributed with unknown means $EX_\alpha^j=\theta_\alpha$ and known variances σ_α^2, i.e., $\underline{\theta}_\alpha=\theta_\alpha, \underline{\underline{\theta}}=\underline{\theta}, \mu_\alpha(\underline{\theta}_\alpha)=\theta_\alpha$, and $\Theta_\alpha=(0,+\infty)$. Given history h_l, define

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

Now from the definition of Θ_α, it follows that $\Delta\Theta_\alpha(\underline{\theta}) = (\theta_\alpha + \phi_\alpha^B(\underline{\theta}), \infty)$ for any optimal matrix B under $\underline{\theta}$, therefore $D(\underline{\theta}) = \{\alpha : \alpha \notin b \text{ for any } b \in O^*(\underline{\theta})\}, \forall\, \underline{\theta} \in \Theta$. 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. (3.15) 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. (3.15).

Note that $b_\alpha(\hat{\underline{\theta}}^l, \theta_\alpha^{K_\alpha}) = \{i_1, \ldots, \alpha, \ldots, i_j\}$; thus, in the first case we have $z^{b_\alpha(\hat{\underline{\theta}}^l, \theta_\alpha^{K_\alpha})} = \hat{\theta}_{i_1}^l x_{i_1} + \ldots + \theta_\alpha^{K_\alpha} x_\alpha + \ldots + \hat{\theta}_{i_j}^l x_{i_j}$ and $z^*(\underline{\theta}) = \theta_{i_1} x_{i_1} + \ldots + \theta_\alpha x_\alpha + \ldots + \theta_{i_j} x_{i_j}$. Therefore, for $b_\alpha(\hat{\underline{\theta}}^l, \theta_\alpha^{K_\alpha}) \in O^*(\underline{\theta})$ 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)

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

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

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:

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

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

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

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

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

where the last equality follows from a consequence of the tail inequality $1 - \Phi(x) < \Phi(x)/x$ for the standard normal distribution, Feller (1967). Thus, we can see that condition (C2) holds.

Summary of Z-UCB Policy π^0:
Step 1 Employ one ISB block in order to have one estimate $\hat{\theta}_a$ from each bandit a. Then, update the vector of estimates $\hat{\underline{\theta}}^2$, and the statistics $S_{\pi^0}(1)$ and $T^{\alpha}_{\pi^0}(S_{\pi^0}(1))$.

Step 2 For block l ($l > 1$) employ an LPB($\pi^0(\hat{\underline{\theta}}^l)$) block defined below. Given the history until the beginning of block l we have the vector of estimates $\hat{\underline{\theta}}^l$ and the statistics $S_{\pi^0}(l-1)$ and $T^{\alpha}_{\pi^0}(S_{\pi^0}(l-1))$. Based on these, compute b^l_0, and $z^{b^l_0}(\hat{\underline{\theta}}^l)$. Then for every bandit $\alpha = \{1, ..., k\}$ we compute the indices

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

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

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

The Z-UCB policy employs as block l the $\pi^0(\hat{\underline{\theta}}^l) = \arg\max_{b_\alpha}\{u_\alpha(\hat{\underline{\theta}}^l)\}$, where ties in the arg max are broken arbitrarily.

4.2 Normal distributions with unknown means and unknown variances

Assume the observations X^j_α from bandit α are normally distributed with unknown means $EX^j_\alpha = \mu_\alpha$ and unknown variances $VarX^j_\alpha = \sigma^2_\alpha$, i.e., $\underline{\theta}_\alpha = (\mu_\alpha, \sigma^2_\alpha)$, and $\Theta_\alpha = \{\underline{\theta}_\alpha : \mu_\alpha > 0,\ \sigma^2_\alpha \geq 0\}$. Given history h_l, define

$$\mu_\alpha(\hat{\underline{\theta}}^l_\alpha) = \frac{\sum_{j=1}^{T^{\alpha}_{\pi^0}(S_{\pi^0}(l-1))} X^j_\alpha}{T^{\alpha}_{\pi^0}(S_{\pi^0}(l-1))} \text{ and } \sigma^2_\alpha(\hat{\underline{\theta}}^l_\alpha) = \frac{\sum_{j=1}^{T^{\alpha}_{\pi^0}(S_{\pi^0}(l-1))} (X^j_\alpha - \mu_\alpha(\hat{\underline{\theta}}^l_\alpha))^2}{T^{\alpha}_{\pi^0}(S_{\pi^0}(l-1))}.$$

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

Also, we have:

$$I(\underline{\theta}_\alpha, \underline{\theta}'_\alpha) = \frac{1}{2} \log \left(1 + \frac{(\mu(\underline{\theta}'_\alpha) - \mu_\alpha)^2}{\sigma^2_\alpha} \right)$$

$$K_\alpha(\underline{\underline{\theta}}) = \frac{1}{2} \log \left(1 + \frac{(\phi^B_\alpha(\underline{\underline{\theta}}))^2}{\sigma^2_\alpha} \right).$$

It is easy to see that the Z-UCB indices of Eq. (3.15) simplify to:

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

where

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

is the mean of the $\underline{\theta}_\alpha$ which satisfies the maximum in the index $u_\alpha(\underline{\hat{\underline{\theta}}}^{l})$, of Eq. (3.15).

Note that $b_\alpha(\underline{\hat{\underline{\theta}}}^{l}, \underline{\theta}_\alpha^{K_\alpha}) = \{i_1, \ldots, \alpha, \ldots, i_j\}$; thus, in the first case we have $z^{b_\alpha(\underline{\hat{\underline{\theta}}}^{l}, \underline{\theta}_\alpha^{K_\alpha})} = \mu_{i_1}(\underline{\hat{\theta}}_{i_1}^{l})x_{i_1} + \ldots + \mu_\alpha(\underline{\theta}_\alpha^{K_\alpha})x_\alpha + \ldots + \mu_{i_j}(\underline{\hat{\theta}}_{i_j}^{l})x_{i_j}$ and $z^*(\underline{\underline{\theta}}) = \mu_{i_1}x_{i_1} + \ldots + \mu_\alpha x_\alpha + \ldots + \mu_{i_j}x_{i_j}$. Therefore, for $b_\alpha(\underline{\hat{\underline{\theta}}}^{l}, \underline{\theta}_\alpha^{K_\alpha}) \in O^*(\underline{\underline{\theta}})$ and from the structure of $z^{b_\alpha(\underline{\hat{\underline{\theta}}}^{l}, \underline{\theta}_\alpha^{K_\alpha})}$ 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. Finally, condition (C2) is satisfied according to the analysis in Cowan et al. (2018).

Summary of Z-UCB Policy π^0:
Step 1 Employ one ISB block in order to have one estimate $\underline{\hat{\theta}}_a$ from each bandit a. Then, update the vector of estimates $\underline{\hat{\underline{\theta}}}^{2}$, and the statistics $S_{\pi^0}(1)$ and $T^\alpha_{\pi^0}(S_{\pi^0}(1))$.

Step 2 For block l $(l > 1)$ employ an LPB$(\pi^0(\underline{\hat{\underline{\theta}}}^{l}))$ block defined below. Given the history until the beginning of block l we have the vector of estimates $\underline{\hat{\underline{\theta}}}^{l}$ and the statistics $S_{\pi^0}(l-1)$ and $T^\alpha_{\pi^0}(S_{\pi^0}(l-1))$. Based on these, compute b_0^l, and $z^{b_0^l}(\underline{\hat{\underline{\theta}}}^{l})$. Then for every bandit $\alpha = \{1, ..., k\}$ we compute the indices

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

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

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

The Z-UCB policy employs as block l the $\pi^0(\underline{\hat{\underline{\theta}}}^{l}) = \arg\max_{b_\alpha}\{u_\alpha(\underline{\hat{\underline{\theta}}}^{l})\}$, where ties in the arg max are broken arbitrarily.

4.3 Discrete distributions with finite support

Assume the observations X_α^j from bandit α are univariate discrete distributions, i.e., $f_\alpha(x, \underline{p}_\alpha) = p_{\alpha x}\mathbf{1}\{X_\alpha = x\}$, $x \in S_\alpha = \{r_{\alpha 1}, \ldots, r_{\alpha d_\alpha}\}$, where the parameters $p_{\alpha x}$ are unknown. The unknown parameters are in $\Theta_\alpha = \{\underline{p}_\alpha \in \mathbb{R}^{d_\alpha} : p_{\alpha x} > 0, \forall x = 1, \ldots, d_\alpha, \sum_x p_{\alpha x} = 1\}$, and $r_{\alpha x}$ are known. Therefore, according to our notation, $\underline{\theta}_\alpha = \underline{p}_\alpha$ and $\underline{\underline{\theta}} = \underline{\underline{p}} = (\underline{p}_1, \ldots, \underline{p}_k)$.

We can compute the mean reward of a bandit as $\mu_\alpha(\underline{\theta}_\alpha) = \mu_\alpha(\underline{p}_\alpha) = \underline{r}_\alpha^{'}\underline{p}_\alpha = \sum_x r_{\alpha x}p_{\alpha x}$, where $\underline{r}_\alpha^{'}$ denotes the transpose of the vector $\underline{r}_\alpha$. Now from the definition

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

Also, we can compute

$$I(\underline{p}_\alpha, \underline{p}_\alpha^{'}) = \sum_{x=1}^{d_\alpha} p_{\alpha x} \log\left(\frac{p_{\alpha x}}{p_{\alpha x}^{'}}\right),$$

$$K_\alpha(\underline{\underline{p}}) = \min_{\underline{p}_\alpha^{'}}\{I(\underline{p}_\alpha, \underline{p}_\alpha^{'}) : \mu_\alpha(\underline{p}_\alpha^{'}) \geq \mu_\alpha(\underline{p}_\alpha) + \phi_\alpha^B(\underline{\underline{p}}), \sum_{x=1}^{d_\alpha} p_{\alpha x}^{'} = 1\}.$$

Now, for any estimators $\hat{\underline{p}}_\alpha^l$ of $\underline{p}_\alpha$, the computation of the index

$$u_\alpha(\hat{\underline{\underline{p}}}^l) = u_\alpha(\hat{\underline{\underline{p}}}^l, s, t) = \max_{\underline{p}_\alpha^{'}}\{z^{b_\alpha(\hat{\underline{\underline{p}}}^l, \underline{p}_\alpha^{'})} : I(\hat{\underline{p}}_\alpha^l, \underline{p}_\alpha^{'}) \leq \frac{\log s}{t}\},$$

involves the solution of $K_\alpha(\hat{\underline{\underline{p}}}^l)$, which is a linear function maximization problem subject to a constraint with convex level sets and a linear constraint, as in Burnetas and Katehakis (1996). Thus, if the minimum in $K_\alpha(\hat{\underline{\underline{p}}}^l)$ is $\underline{p}_\alpha^{K_\alpha}$ then

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

Note that $b_\alpha(\hat{\underline{\underline{p}}}^l, \underline{p}_\alpha^{K_\alpha}) = \{i_1, \ldots, \alpha, \ldots, i_j\}$; thus, in the first case we have $z^{b_\alpha(\hat{\underline{\underline{p}}}^l, \underline{p}_\alpha^{K_\alpha})} = \underline{r}_{i_1}^{'}\hat{\underline{p}}_{i_1}^l x_{i_1} + \ldots + \underline{r}_\alpha^{'}\underline{p}_\alpha^{K_\alpha} x_\alpha + \ldots + \underline{r}_{i_j}^{'}\hat{\underline{p}}_{i_j}^l x_{i_j}$ and $z^*(\underline{\underline{p}}) = \underline{r}_{i_1}^{'}\underline{p}_{i_1} x_{i_1} + \ldots + \underline{r}_\alpha^{'}\underline{p}_\alpha x_\alpha + \ldots + \underline{r}_{i_j}^{'}\underline{p}_{i_j} x_{i_j}$. Thus, the index is just a weighted sum of the estimated means and the inflated one, so in order to prove that the policy satisfies the conditions (C2) and (C3) one can follow exactly the same arguments that are used in Proposition 3 in Burnetas and Katehakis (1996). In Proposition 3 they use arguments based on the properties of the mean rewards of each bandit which hold in our case due to the form of our indices, as we analyzed above.

Summary of Z-UCB Policy π^0:
Step 1 Employ one ISB block in order to have one estimate $\hat{\underline{p}}_a$ from each bandit a. Then, update the vector of estimates $\hat{\underline{\underline{p}}}^2$, and the statistics $S_{\pi^0}(1)$ and $T^{\alpha}_{\pi^0}(S_{\pi^0}(1))$.

Step 2 For block l ($l > 1$) employ an LPB($\pi^0(\hat{\underline{\underline{p}}}^l)$) block defined below.
Given the history until the beginning of block l we have the vector of estimates $\hat{\underline{\underline{p}}}^l$ and the statistics $S_{\pi^0}(l-1)$ and $T^{\alpha}_{\pi^0}(S_{\pi^0}(l-1))$. Based on these, compute b^l_0, and $z^{b^l_0}(\hat{\underline{\underline{p}}}^l)$.
Then for every bandit $\alpha = \{1, ..., k\}$ we compute the indices

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

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

$$\underline{p}^{K_{\alpha}}_{\alpha} = \min_{\underline{p}'_{\alpha}}\{\sum_{x=1}^{d_{\alpha}} p_{\alpha x} \log\left(\frac{p_{\alpha x}}{p'_{\alpha x}}\right) : \underline{r}'_{\alpha}\underline{p}'_{\alpha} = \underline{r}'_{\alpha}\underline{p}_{\alpha} + \phi^B_{\alpha}(\underline{\underline{p}}), \sum_{x=1}^{d_{\alpha}} p'_{\alpha x} = 1\}.$$

The Z-UCB policy employs as block l the $\pi^0(\hat{\underline{\underline{p}}}^l) = \arg\max_{b_{\alpha}}\{u_{\alpha}(\hat{\underline{\underline{p}}}^l)\}$, where ties in the arg max are broken arbitrarily.

5 Conclusions

This paper introduces a novel approach to multi-armed bandit (MAB) problems with side constraints, tackling the challenge of resource-limited activations. By deriving an asymptotic lower bound for regret in feasible policies and designing optimal strategies that achieve this bound, the study advances constrained bandit models. The proposed block-based Upper Confidence Bound (UCB) policies ensure asymptotic optimality while offering computational advantages over existing methods. These findings have broad implications for sequential decision-making in dynamic resource allocation, including applications in online revenue management and targeted advertising. Future research could refine these policies further, particularly in complex environments with evolving constraints and adaptive learning mechanisms.

A Appendix

Proof Lemma 1. It is obvious that $\phi^B_j(\underline{\underline{\theta}}') = \phi^B_j(\underline{\underline{\theta}}) \geq 0,\ \forall\ j \neq \alpha$ because we only change the parameter of bandit α and $\phi^B_j(\underline{\underline{\theta}}') = \phi^B_j(\underline{\underline{\theta}}) \equiv c^j_1 g^B_1 + \ldots + c^j_L g^B_L + g^B_{L+1} - \mu_j(\underline{\theta}_j)$.

For a bandit $\alpha \in B(\underline{\underline{\theta}})$ we have that $\alpha \notin b$, for any $b \in O^*(\underline{\underline{\theta}})$. Therefore $\phi^B_{\alpha}(\underline{\underline{\theta}}) \equiv c^{\alpha}_1 g^B_1 + \ldots + c^{\alpha}_L g^B_L + g^B_{L+1} - \mu_{\alpha}(\underline{\theta}_{\alpha}) > 0$, for any B corresponding to b.

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

Therefore, for any optimal matrix B under $\underline{\underline{\theta}}$ we have that $\phi^B_\alpha(\underline{\underline{\theta}}') \equiv c^\alpha_1 g^B_1 + \ldots + c^\alpha_L g^B_L + g^B_{L+1} - \mu_\alpha(\underline{\theta}'_\alpha) < 0$ because B is not optimal under $\underline{\underline{\theta}}'$.

Now from $\phi^B_\alpha(\underline{\underline{\theta}}) = c^\alpha_1 g^B_1 + \ldots + c^\alpha_L g^B_L + g^B_{L+1} - \mu_\alpha(\underline{\theta}_\alpha)$ we have that $\phi^B_\alpha(\underline{\underline{\theta}}') = \phi^B_\alpha(\underline{\underline{\theta}}) + \mu_\alpha(\underline{\theta}_\alpha) - \mu_\alpha(\underline{\theta}'_\alpha) < 0$. □

Proof Lemma 2. (i) Let $\underline{\underline{\theta}} : O^*(\underline{\underline{\theta}}) = \{b\}$, then $g^B_i > 0$, only for $j-1$ out of L type of constraints because if $g^B_i = 0$ for some of them we must have more than one solutions in the primal, which cannot occur because b is uniquely optimal.

(ii) Let $\underline{\underline{\theta}} : O^*(\underline{\underline{\theta}}) = \{b\}$, then $g^B_i = 0$, for all $i = 1, \ldots, L$ from the dual solution and $\phi^B_j(\underline{\underline{\theta}}) > 0$ for all $j \neq i_1$. Thus, g^B_{L+1} is positive since the means (μ's) are positive. □

Proof Proposition 3. Let $\alpha \in D(\underline{\underline{\theta}})$, $\theta'_\alpha \in \Delta\Theta_\alpha(\underline{\underline{\theta}})$. The definition of $\Delta\Theta_\alpha(\underline{\underline{\theta}})$ and Lemma 1 imply that we must have a b' which is uniquely optimal under $\underline{\underline{\theta}}'$ (i.e., $O^*(\underline{\underline{\theta}}') = \{b'\}$) and $\alpha \in b'$. Then we have two cases for the uniquely optimal solution b' depending on wether b' is a singleton or not.

In the first case $b' = \{\alpha\}$, and Lemma 2 implies that for a f-UF policy $g^{B'}_i = 0, i = 1, \ldots, L$ thus, the definition of a f-UF policy implies that:

$$E_{\underline{\underline{\theta}}'} T^j_\pi(n) = o(n^a), \text{ for all } a > 0, \text{ for all } j \notin b'. \tag{A.1}$$

Therefore,

$$n - E_{\underline{\underline{\theta}}'} T^\alpha_\pi(n) = \sum_{j \notin b'} E_{\underline{\underline{\theta}}'} T^j_\pi(n) = o(n^a), \text{ for all } a > 0. \tag{A.2}$$

Now for any sequence $\beta_n = o(n)$ with $\beta_n < n$ (for all n), we obtain the following.

$$\begin{aligned} E_{\underline{\underline{\theta}}'} T^\alpha_\pi(n) &= \sum_{k=1}^{n} k\, P_{\underline{\underline{\theta}}'}[T^\alpha_\pi(n) = k] \\ &= \sum_{k=1}^{\lfloor \beta_n \rfloor} k\, P_{\underline{\underline{\theta}}'}[T^\alpha_\pi(n) = k] + \sum_{k=\lfloor \beta_n \rfloor + 1}^{n} k\, P_{\underline{\underline{\theta}}'}[T^\alpha_\pi(n) = k] \\ &\leq \beta_n P_{\underline{\underline{\theta}}'}[T^\alpha_\pi(n) \leq \beta_n] + n P_{\underline{\underline{\theta}}'}[T^\alpha_\pi(n) > \beta_n] \\ &= n - (n - \beta_n) P_{\underline{\underline{\theta}}'}[T^\alpha_\pi(n) \leq \beta_n]. \end{aligned}$$

Therefore

$$n - E_{\underline{\underline{\theta}}'} T^\alpha_\pi(n) \geq (n - \beta_n) P_{\underline{\underline{\theta}}'}[T^\alpha_\pi(n) \leq \beta_n]. \tag{A.3}$$

From (A.2) and (A.3) we obtain

$$(n - \beta_n) P_{\underline{\underline{\theta}}'}[T^\alpha_\pi(n) \leq \beta_n] = o(n^a), \text{ for all } a > 0,$$

thus

$$P_{\underline{\underline{\theta}}'}[T^\alpha_\pi(n) \leq \beta_n] = o(n^{a-1}), \text{ for all } a > 0.$$

And the proof is complete for this case.

In the second case $b^{'} = \{i_1, \ldots, i_j\}$, for $j = 2, \ldots, L+1$, where bandit α is one of $i_1, \ldots, i_j$. Then as before (A.1) holds $\forall\, a > 0,\ \forall\, i \notin b^{'}$. It follows from Lemma 2 that for an f-UF policy we must have $g_i^{B^{'}} > 0$, for $j-1$ resource constraints, which we label as $i = s_1, \ldots, s_{j-1}$. Using the last result and and (3.7) we obtain:

$$\sum_{i=s_1}^{s_{j-1}} g_i^{B^{'}} \sum_{l \in b^{'}} (c_i^0 - c_i^l) E_{\underline{\underline{\theta}}^{'}} T_\pi^l(n) = o(n^a),\ \forall\, a > 0. \tag{A.4}$$

If we sum (A.1) for all $i \notin b^{'}$ it follows that

$$n - \sum_{i \in b^{'}} E_{\underline{\underline{\theta}}^{'}} T_\pi^i(n) = \sum_{i \notin b^{'}} E_{\underline{\underline{\theta}}^{'}} T_\pi^i(n) = \varepsilon_n,\ \text{where } \varepsilon_n = o(n^a),\ \forall\, a > 0. \tag{A.5}$$

Now, let $x_{i_1}, \ldots, x_{i_j}$ be the corresponding randomization probabilities, then $\sum_{k \in b^{'}} x_k = 1$ and from (A.5) we have that

$$\sum_{i \in b^{'}} (n x_i - E_{\underline{\underline{\theta}}^{'}} T_\pi^i(n)) = \varepsilon_n,\ \text{where } \varepsilon_n = o(n^a),\ \forall\, a > 0. \tag{A.6}$$

From the definition of z^* we can write $z^*(\underline{\underline{\theta}}^{'}) = \sum_{i_k \neq \alpha} x_{i_k} \mu_{i_k}(\underline{\theta}_{i_k}) + x_\alpha \mu_\alpha(\underline{\theta}_\alpha^{'})$, and from the DLP we have $z^*(\underline{\underline{\theta}}^{'}) = \sum_{k=s_1}^{s_{j-1}} c_k^0 g_k^{B^{'}} + g_{L+1}^{B^{'}}$. Also, from the DLP we obtain that $\phi_i^{B^{'}}(\underline{\underline{\theta}}^{'}) = \sum_{k=s_1}^{s_{j-1}} c_k^i g_k^{B^{'}} + g_{L+1}^{B^{'}} - \mu_i(\underline{\theta}_i)$, for $i \neq \alpha$, and $\phi_\alpha^{B^{'}}(\underline{\underline{\theta}}^{'}) = \sum_{k=s_1}^{s_{j-1}} c_k^i g_k^{B^{'}} + g_{L+1}^{B^{'}} - \mu_i(\underline{\theta}_\alpha^{'})$, where $\phi_i^{B^{'}}(\underline{\underline{\theta}}^{'}) = 0$ for $i \in b^{'}$.

After some algebra we can show that

$$z^*(\underline{\underline{\theta}}^{'}) = (c_{s_1}^0 - c_{s_1}^i) g_{s_1}^{B^{'}} + \ldots + (c_{s_j}^0 - c_{s_{j-1}}^i) g_{s_{j-1}}^{B^{'}} + \mu_i(\underline{\theta}_i),\ \text{for every } i \in b^{'}, \tag{A.7}$$

and

$$\begin{aligned} z^*(\underline{\underline{\theta}}^{'}) &= \sum_{i_k \neq \alpha} x_{i_k} \mu_{i_k}(\underline{\theta}_{i_k}) + x_\alpha \mu_\alpha(\underline{\theta}_\alpha^{'}) \\ &= x_{i_1}(c_{s_1}^{i_1} g_{s_1}^{B^{'}} + \ldots + c_{s_{j-1}}^{i_1} g_{s_{j-1}}^{B^{'}} + g_{L+1}^{B^{'}}) + \ldots + x_{i_j}(c_{s_1}^{i_j} g_{s_1}^{B^{'}} + \ldots + c_{s_{j-1}}^{i_j} g_{s_{j-1}}^{B^{'}} + g_{L+1}^{B^{'}}) \\ &= \sum_{i=s_1}^{s_{j-1}} g_i^{B^{'}} \sum_{l \in b^{'}} c_i^l x_{i_l} + g_{L+1}^{B^{'}}. \end{aligned} \tag{A.8}$$

Using $z^*(\underline{\underline{\theta}}^{'}) = \sum_{k=s_1}^{s_{j-1}} c_k^0 g_k^{B^{'}} + g_{L+1}^{B^{'}}$, and (A.8) we have: $z^*(\underline{\underline{\theta}}^{'}) - g_{L+1}^{B^{'}} = \sum_{k=s_1}^{s_{j-1}} c_k^0 g_k^{B^{'}}$ and $z^*(\underline{\underline{\theta}}^{'}) - g_{L+1}^{B^{'}} = \sum_{i=s_1}^{s_{j-1}} g_i^{B^{'}} \sum_{l \in b^{'}} c_i^l x_{i_l}$ which imply:

$$\sum_{i=s_1}^{s_{j-1}} g_i^{B^{'}} c_i^0 - \sum_{i=s_1}^{s_{j-1}} g_i^{B^{'}} \sum_{l \in b^{'}} c_i^l x_{i_l} = 0. \tag{A.9}$$

In addition, since $\sum_{k \in b^{'}} x_k = 1$ (A.9) can be written as:

$$\sum_{i=s_1}^{s_{j-1}} g_i^{B^{'}} \sum_{l \in b^{'}} c_i^0 x_{i_l} - \sum_{i=s_1}^{s_{j-1}} g_i^{B^{'}} \sum_{l \in b^{'}} c_i^l x_{i_l} = 0,$$

which simplifies into:

$$\sum_{i=s_1}^{s_{j-1}} g_i^{B'} \sum_{l\in b'} (c_i^0 - c_i^l) x_{i_l} = 0.$$

Multiplying both sides of the last equation by n,

$$\sum_{i=s_1}^{s_{j-1}} g_i^{B'} \sum_{l\in b'} (c_i^0 - c_i^l) n x_{i_l} = 0. \tag{A.10}$$

From (A.4) and (A.10) we have that

$$\sum_{i=s_1}^{s_{j-1}} g_i^{B'} \sum_{l\in b'} (c_i^0 - c_i^l)(n x_l - E_{\underline{\underline{\theta}}'} T_\pi^l(n)) = o(n^a),\ \forall\, a > 0. \tag{A.11}$$

Combining, (A.6) and (A.11) we obtain,

$$n x_i - E_{\underline{\underline{\theta}}'} T_\pi^i(n) = o(n^a),\ \forall\, a > 0,\ \forall\, i \in b'. \tag{A.12}$$

For any n let

$$\Gamma_n^\pi = \sum_{l\notin b'} T_\pi^l(n), \text{ and } F_{n,i}^\pi = \sum_{l\notin b'} (c_i^0 - c_i^l) T_\pi^l(n),$$

where i is the $i-$th resource constraint. Now, for each resource constraint i, label as c_i^* the minimum c_i^j for $j = 1, \ldots, k$. With thus defined c_i^* and the above definitions we have:

$$F_{n,i}^\pi \le \Gamma_n^\pi (c_i^0 - c_i^*),\ \text{for all constraints } i = 1, \ldots, L.$$

Furthermore, from (A.5)

$$E_{\underline{\underline{\theta}}'} \Gamma_n^\pi = o(n^a),\ \forall\, a > 0. \tag{A.13}$$

Now, (2.3) and the definition of $F_{n,i}^\pi$ imply the following.

$$\sum_{i=s_1}^{s_{j-1}} g_i^{B'} (n c_i^0 - C_{i,\pi}(n)) = \sum_{i=s_1}^{s_{j-1}} g_i^{B'} F_{n,i}^\pi + \sum_{i=s_1}^{s_{j-1}} g_i^{B'} \sum_{l\in b'} (c_i^0 - c_i^l) T_\pi^l(n). \tag{A.14}$$

For any bandit in b' the following arguments hold. For simplicity we present these arguments only for the specific bandit α for which $\alpha \in D(\underline{\underline{\theta}}), \theta_\alpha' \in \Delta\Theta_\alpha(\underline{\underline{\theta}})$.

Since $g_i^{B'} > 0$ (from the optimality of B') and $n c_i^0 - C_{i,\pi}(n) \ge 0,\ \forall\, n, \forall\, i$ (from the feasibility of π) the right side of (A.14) is nonnegative. Using the non negativity inequality of the right side of (A.14) and moving the term that corresponds to bandit α to the left side in the equation (A.14) (and changing the sign of $(c_\alpha^0 - c_\alpha^l)$) we obtain the inequality below,

$$\sum_{i=s_1}^{s_{j-1}} g_i^{B'} (c_i^\alpha - c_i^0) T_\pi^\alpha(n) \le \sum_{i=s_1}^{s_{j-1}} g_i^{B'} F_{n,i}^\pi + \sum_{i=s_1}^{s_{j-1}} g_i^{B'} \sum_{l\in b', l\neq\alpha} (c_i^0 - c_i^l) T_\pi^l(n).$$

Using (A.7) on both sides of the above inequality (for $i = \alpha$ in the left hand side and for all $i \in b'$, $i \neq \alpha$ in the right side) we obtain:

$$(\mu_\alpha(\underline{\theta}_\alpha') - z^*(\underline{\underline{\theta}}')) T_\pi^\alpha(n) \le \sum_{i=s_1}^{s_{j-1}} g_i^{B'} F_{n,i}^\pi + \sum_{l\in b', l\neq\alpha} (z^*(\underline{\underline{\theta}}') - \mu_l(\underline{\theta}_l)) T_\pi^l(n),$$

which simplifies into:

$$\mu_\alpha(\underline{\theta}_\alpha^{'})T_\pi^\alpha(n) \leq \sum_{i=s_1}^{s_{j-1}} g_i^{B^{'}} F_{n,i}^\pi + z^*(\underline{\underline{\theta}}^{'}) \sum_{l\in b^{'}} T_\pi^l(n) - \sum_{l\in b^{'}, l\neq\alpha} \mu_l(\underline{\theta}_l) T_\pi^l(n).$$

Now from the definition of Γ_n^π we have that $n - \Gamma_n^\pi = \sum_{l\in b^{'}} T_\pi^l(n)$ and recall that $z^*(\underline{\underline{\theta}}^{'}) = x_\alpha \mu_\alpha(\underline{\theta}_\alpha^{'}) + \sum_{l\in b^{'}, l\neq\alpha} x_l \mu_l(\underline{\theta}_l)$ and that by assumption $\mu_l(\underline{\theta}_l) > 0$. Using these and simple algebra the above inequality can be written as,

$$T_\pi^\alpha(n) \leq nx_\alpha + \sum_{i=s_1}^{s_{j-1}} \frac{g_i^{B^{'}}}{\mu_\alpha(\underline{\theta}_\alpha^{'})} F_{n,i}^\pi + \frac{1}{\mu_\alpha(\underline{\theta}_\alpha^{'})} \sum_{l\in b^{'}, l\neq\alpha} (nx_l - T_\pi^l(n))\mu_l(\underline{\theta}_l).$$

The last inequality can be rearranged and written as,

$$nx_\alpha - T_\pi^\alpha(n) + \sum_{i=s_1}^{s_{j-1}} \frac{g_i^{B^{'}}}{\mu_\alpha(\underline{\theta}_\alpha^{'})} F_{n,i}^\pi + \frac{1}{\mu_\alpha(\underline{\theta}_\alpha^{'})} \sum_{l\in b^{'}, l\neq\alpha} (nx_l - T_\pi^l(n))\mu_l(\underline{\theta}_l) \geq 0.$$

For simplicity the above will be written as

$$nx_\alpha - T_\pi^\alpha(n) + A_\pi^1(n) + A_\pi^2(n) \geq 0.$$

where $A_\pi^1(n) = \sum_{i=s_1}^{s_{j-1}} \frac{g_i^{B^{'}}}{\mu_\alpha(\underline{\theta}_\alpha^{'})} F_{n,i}^\pi$ and $A_\pi^2(n) = \frac{1}{\mu_\alpha(\underline{\theta}_\alpha^{'})} \sum_{l\in b^{'}, l\neq\alpha} (nx_l - T_\pi^l(n))\mu_l(\underline{\theta}_l)$.

Note that from (A.12) (and (A.13) respectively) it follows that $A_\pi^i(n) = o(n^a),\ \forall\, a > 0$, and $i = 1, 2$, since all means are positive. From the Markov inequality, for any positive $\beta_n = o(n)$

$$P_{\underline{\underline{\theta}}^{'}}(nx_\alpha - T_\pi^\alpha(n) + A_\pi^1(n) + A_\pi^2(n) \geq nx_\alpha - \beta_n) \leq \frac{E_{\underline{\underline{\theta}}^{'}}(nx_\alpha - T_\pi^\alpha(n) + A_\pi^1(n) + A_\pi^2(n))}{nx_\alpha^{'} - \beta_n}$$
$$= \frac{o(n^a)}{nx_\alpha - \beta_n} = o(n^{a-1}),\ \forall\, a > 0.$$

Using the above we obtain,

$$P_{\underline{\underline{\theta}}^{'}}(T_\pi^\alpha(n) \leq \beta_n) \leq P_{\underline{\underline{\theta}}^{'}}(T_\pi^\alpha(n) \leq \beta_n + A_\pi^1(n) + A_\pi^2(n)) = o(n^{a-1}), \forall\, a > 0.$$

And the proof is complete for this case too. □

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

Lemma 7 *Under conditions (C1),(C2), and (C3) policy π^0 satisfies:*

$$\overline{lim}_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} \widetilde{T}_{\pi^0,2}^b(L_n)}{\log L_n} \leq \frac{1}{m_i^b K_i(\underline{\underline{\theta}})},\ \textit{for all } i \in D(\underline{\underline{\theta}}), i \in b, b \notin O^*(\underline{\underline{\theta}}),$$

$$\overline{lim}_{n\to\infty} \frac{E_{\underline{\underline{\theta}}} \widetilde{T}_{\pi^0,2}^b(L_n)}{\log L_n} = 0,\ \textit{for all } i \notin D(\underline{\underline{\theta}}), i \in b, b \notin O^*(\underline{\underline{\theta}}),$$

$$\overline{lim}_{n\to\infty}\frac{E_{\underline{\underline{\theta}}}\widetilde{T}^{b}_{\pi^0,1}(L_n)}{\log L_n}=0,\ \textit{for all}\ b\notin O^*(\underline{\underline{\theta}}).$$

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

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

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

$$\begin{aligned}
&\sum_{t=2}^{L_n} 1\{\pi^0_t=b, b\notin O^*(\underline{\underline{\theta}}), b\in O^*(\hat{\underline{\underline{\theta}}}^t), u_i(\hat{\underline{\underline{\theta}}}^t,\underline{\theta}_i^{'}) = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t), T^i_{\pi^0}(S_{\pi^0}(t-1)) < c\}\\
&\qquad \le \sum_{t=2}^{L_n} 1\{\pi^0_t=b, T^i_{\pi^0}(S_{\pi^0}(t-1)) < c\}\\
&\qquad = \sum_{t=2}^{L_n}\sum_{s=0}^{\lfloor c/m^b_i\rfloor} 1\{\pi^0_t=b, T^i_{\pi^0}(S_{\pi^0}(t-1)) = s\,m^b_i + m^0_i\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=0}^{\lfloor c/m_i^b \rfloor} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b,\, T_{\pi^0}^i(S_{\pi^0}(t-1)) = s\, m_i^b + m_i^0\} \\
&\le \lfloor c/m_i^b \rfloor + 1 \\
&\le \frac{c}{m_i^b} + 1 = \frac{\log L_n}{m_i^b (J_i(\underline{\underline{\theta}}, \epsilon) - \delta)} + 1.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
E_{\underline{\underline{\theta}}} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b_0^t \in O^*(\underline{\underline{\theta}}), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i') = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t), T_{\pi^0}^i(S_{\pi^0}(t-1)) < \frac{\log L_n}{J_i(\underline{\underline{\theta}}, \epsilon) - \delta}\} \\
\le \frac{\log L_n}{m_i^b (J_i(\underline{\underline{\theta}}, \epsilon) - \delta)} + 1. \qquad \text{(A.15)}
\end{aligned}
$$

Furthermore,

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

Then from (C2) and Remark 3 we have that

$$
\begin{aligned}
E_{\underline{\underline{\theta}}} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b \notin O^*(\underline{\underline{\theta}}), b \in O^*(\hat{\underline{\underline{\theta}}}^t), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i') = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t), J_i(\hat{\underline{\underline{\theta}}}^t, \epsilon) \le J_i(\underline{\underline{\theta}}, \epsilon) - \delta\} \\
\le o(\log L_n). \qquad \text{(A.16)}
\end{aligned}
$$

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

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

Thus from condition (C3)

$$
\begin{aligned}
&E_{\underline{\underline{\theta}}} \sum_{t=2}^{L_n} 1\{\pi_t^0 = b, b \notin O^*(\underline{\underline{\theta}}), b \in O^*(\hat{\underline{\underline{\theta}}}^t), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i') = u_{\alpha^*}(\hat{\underline{\underline{\theta}}}^t), u_i(\hat{\underline{\underline{\theta}}}^t, \underline{\theta}_i') \le z^*(\underline{\underline{\theta}}) - \epsilon\} \\
&\le o(\log L_n). \qquad \text{(A.17)}
\end{aligned}
$$

Finally, it follows from (A.15), (A.16) and (A.17) that

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

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

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

Thus

$$\overline{\lim}_{n\to\infty} \frac{E_{\underline{\underline{\theta}}}\widetilde{T}^b_{\pi^0,2}(L_n)}{\log L_n} \leq \frac{1}{m_i^b\, K_i(\underline{\underline{\theta}})}, \text{ for all } i \in D(\underline{\underline{\theta}}), i \in b, b \notin O^*(\underline{\underline{\theta}}),$$

$$\overline{\lim}_{n\to\infty} \frac{E_{\underline{\underline{\theta}}}\widetilde{T}^b_{\pi^0,2}(L_n)}{\log L_n} = 0, \text{ for all } i \notin D(\underline{\underline{\theta}}), i \in b, b \notin O^*(\underline{\underline{\theta}}) \text{ and}$$

$$\overline{\lim}_{n\to\infty} \frac{E_{\underline{\underline{\theta}}}\widetilde{T}^b_{\pi^0,1}(L_n)}{\log L_n} = 0, \text{ for all } b \notin O^*(\underline{\underline{\theta}}).$$

This completes the proof. □

Acknowledgements We acknowledge support for this work from the National Science Foundation, NSF grant CMMI-16-62629.

Declarations

Conflict of interest The authors declare that there is no Conflict of interest.

References

Agarwal, D., Ghosh, S., Wei, K., & You, S. (2014). Budget pacing for targeted online advertisements at linkedin. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1613-1619). ACM.

Agrawal, S., & Devanur, N. R. (2014). Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation* (pp. 989–1006). ACM.

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.

Auer, P., & Ortner, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica, 61*(1–2), 55–65.

Babich, V., & Tang, C. S. (2016). Franchise contracting: The effects of the entrepreneur's timing option and debt financing. *Production and Operations Management, 25*(4), 662–683.

Badanidiyuru, A., Kleinberg, R., & Slivkins, A. (2018). Bandits with knapsacks. *Journal of the ACM (JACM), 65*(3), 13.

Borkar, V., & Jain, R. (2014). Risk-constrained Markov decision processes. *IEEE Transactions on Automatic Control, 59*(9), 2574–2579.

Bubeck, S. & Slivkins, A. (2012). The best of both worlds: Stochastic and adversarial bandits. arXiv preprint arXiv:1202.4473.

Bubeck, S., Munos, R., & Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory* (pp. 23–37). Springer.

Burnetas, A., & Kanavetas, O. (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* (pp. 97–112). Springer.

Burnetas, A. N., & Katehakis, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics, 17*(2), 122–142.

Burnetas, A. N., Kanavetas, O., & Katehakis, M. N. (2017). Asymptotically optimal multi-armed bandit policies under a cost constraint. *Probability in the Engineering and Informational Sciences, 31*(3), 284–310.

Chen, W., Kanavetas, O., & Katehakis, M. N. (2021). Ameso optimization: A relaxation of discrete midpoint convexity. *Discrete Applied Mathematics, 293*, 177–192.

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

Cowan, W., & Katehakis, M. N. (2020). Exploration-exploitation policies with almost sure, arbitrarily slow growing asymptotic regret. *Probability in the Engineering and Informational Sciences, 34*(3), 406–428.

Cowan, W., Honda, J., & Katehakis, M. N. (2018). Normal bandits of unknown means and variances. *Journal of Machine Learning Research, 18*(154), 1–28.

Cowan, W., Katehakis, M. N., & Ross Sheldon, M. (2023). Halting multi-armed bandits and single payout bandits. *Naval Research Logistics, 70*(7), 639–652.

Denardo, E. V., Feinberg, E. A. & Rothblum, U. G. (2013). The multi-armed bandit, with constraints. In M. N. Katehakis, S. M. Ross, & J. Yang (Ed.), *Cyrus Derman Memorial Volume I: Optimization under Uncertainty: Costs, Risks and Revenues*. Annals of Operations Research, Springer, New York.

Derman, C. (1970). *Finite state Markovian decision processes*. Academic Press.

Ding, W., Qin, T., Zhang, X. D., & Liu, T. Y. (2013). Multi-armed bandit with budget constraint and variable costs. In *AAAI-13*, (pp. 232 – 238).

Feinberg, E. A. (1994). Constrained semi-Markov decision processes with average rewards. *Zeitschrift für Operations Research, 39*(3), 257–288.

Feller, W. (1967). *An introduction to probability theory and its applications* (3rd ed., Vol. 1). Wiley.

Feng, Q., & Shanthikumar, J. G. (1994). How research in production and operations management may evolve in the era of big data. *Production and Operations Management, 27*(8), 1670–1684.

Johnson, K., David Simchi-Levi, F., & Wang, H. (2018). Online network revenue management using thompson sampling. In *Operations research, to appear*.

Guha, S. & Munagala, K. (2007). Approximation algorithms for budgeted learning problems. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, (pp. 104 – 113). ACM.

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

Hordijk, A., & Spieksma, F. M. (1989). Constrained admission control to a queueing system. *Advances in Applied Probability, 21*(2), 409–431.

Johnson, K., Simchi-Levi, D., & Wang, H. (2015). Online network revenue management using Thompson sampling. *Available at SSRN*.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics, 6*(1), 4–2.

Lattimore, T. (2018). Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research, 19*, 1–32.

Lattimore T., & Szepesvári, C. (2018). Bandit algorithms. *preprint*.

Lattimore, T., Crammer, K., & Szepesvári, C. (2014). Optimal resource allocation with semi-bandit feedback. arXiv preprint arXiv:1406.3840.

Levi, R., Perakis, G., & Uichanco, J. (2015). The data-driven newsvendor problem: New bounds and insights. *Operations Research, 63*(6), 1294–1306.

Mahajan, A., & Teneketzis, D. (2008). Multi-armed bandit problems. In *Foundations and Applications of Sensor Management*, (pp. 121–151). Springer.

Perakis, G., & Roels, G. (2008). Regret in the newsvendor model with partial information. *Operations Research, 56*(1), 188–203.

Pike-Burke, C., & Grunewalder, S. (2017). Optimistic planning for the stochastic knapsack problem. In *Artificial Intelligence and Statistics*, (pp. 1114–1122).

Pike-Burke, C., Agrawal, S., Szepesvári, C., & Grunewalder, S. (2018). Bandits with delayed, aggregated anonymous feedback.

Rusmevichientong, P., & Williamson, D. P. (2006). An adaptive algorithm for selecting profitable keywords for search-based advertising services. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, (pp. 260–269). ACM.

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*, (pp. 137 – 144). International Foundation for Autonomous Agents and Multiagent Systems.

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*, (pp. 1167–1178). International World Wide Web Conferences Steering Committee.

Spencer, N. H., & de Lopez, L. K. (2018). Item-by-item sampling for promotional purposes. *Quality Technology & Quantitative Management, 15*(5), 655–662.

Thomaidou, S., Vazirgiannis, M., & Liakopoulos, K. (2012). Toward an integrated framework for automated development and optimization of online advertising campaigns. arXiv preprint arXiv:1208.1187.

Tran-Thanh, L., Chapman, A., De Cote, E. M., Rogers, A., & Jennings, N. R. (2010). Epsilon–first policies for budget–limited multi-armed bandits. In *AAAI-10*, (pp. 1211–1216).

Tran-Thanh, L., Chapman, A., Rogers, A., & Jennings, N. (2012). Knapsack based optimal policies for budget–limited multi–armed bandits. In *AAAI-12*, (pp. 1134 – 1140).

Tran-Thanh, L., Stavrogiannis, L., Naroditskiy, V., Robu, V., Jennings, N. R., & Key, P. (2014). Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions.

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.

Zhou, R., Gan, C., Yan, J., & Shen, C. (2018). Cost-aware cascading bandits. arXiv preprint arXiv:1805.08638.