Adv. Appl. Prob. **51,** 745–772 (2019)
*doi:*10.1017/apr.2019.29

AN ASYMPTOTICALLY OPTIMAL HEURISTIC FOR GENERAL NONSTATIONARY FINITE-HORIZON RESTLESS MULTI-ARMED, MULTI-ACTION BANDITS

GABRIEL ZAYAS-CABÁN,* *University of Wisconsin-Madison*
STEFANUS JASIN,** *** AND
GUIHUA WANG,** **** *University of Michigan*

Abstract

We propose an asymptotically optimal heuristic, which we term *randomized assignment control* (RAC) for a restless multi-armed bandit problem with discrete-time and finite states. It is constructed using a linear programming relaxation of the original stochastic control formulation. In contrast to most of the existing literature, we consider a finite-horizon problem with multiple actions and time-dependent (i.e. nonstationary) upper bound on the number of bandits that can be activated at each time period; indeed, our analysis can also be applied in the setting with nonstationary transition matrix and nonstationary cost function. The asymptotic setting is obtained by letting the number of bandits and other related parameters grow to infinity. Our main contribution is that the asymptotic optimality of RAC in this general setting does not require indexability properties or the usual stability conditions of the underlying Markov chain (e.g. unichain) or fluid approximation (e.g. global stable attractor). Moreover, our multi-action setting is not restricted to the usual dominant action concept. Finally, we show that RAC is also asymptotically optimal for a *dynamic* population, where bandits can randomly arrive and depart the system.

Keywords: Restless bandit; asymptotic optimality; finite horizon; nonstationary; arm acquiring; nonindexable bandit

2010 Mathematics Subject Classification: Primary 90C40
Secondary 68M20; 90B36

1. Introduction

We present a heuristic control/policy that is asymptotically optimal for a finite-horizon restless multi-armed, multi-action bandit problem with a *fixed* population (i.e. the set of bandits remains constant throughout the entire horizon) and *dynamic* population (i.e. bandits can arrive or depart the system). A multi-armed bandit problem (MABP) involves activating competing bandits/arms sequentially over time. In its original form, a fixed number of bandits need to be activated at any given time, and the state of each bandit evolves according to a controlled stochastic process when it is activated. A solution to an MABP specifies which bandits need

Received 21 September 2017; revision received 31 March 2019.
The supplementary material for this article can be found at https://doi.org/10.1017/apr.2019.29.
* Postal address: Mechanical Engineering Building, University of Wisconsin-Madison, 1513 University Avenue, Room 3011 Madison, WI 53706-1691, USA. Email address: zayascaban@wisc.edu
** Postal address: Stephen M. Ross School of Business, University of Michigan, 701 Tappan Street, Ann Arbor, MI 48109, USA.
*** Email address: jasin@umich.edu
**** Email address: guihuaw@umich.edu

to be activated at each decision epoch to minimize either the expected discounted or long-run average cost associated with how the states of the bandits evolve over time. In the MABP literature, a celebrated result is the so-called *Gittins index policy*, introduced in Gittins and Jones (1974) for a fixed population bandit problem. This policy assigns each bandit an index as a function of its current state and then activates the bandit(s) with the largest indices. When only one bandit can be activated at each period, this policy optimizes the infinite-horizon expected discounted costs and infinite-horizon long-run average cost. Whittle (1988) generalized the MABP to allow nonactive bandits to also change their states (dubbed by Whittle as a *changing world* setting), still assuming a fixed population setting, giving rise to the restless multi-armed bandit problem (RMABP).

The RMABP is a general modeling framework encompassing applications in sequential selection of clinical trials in medicine, sensor management, manufacturing systems, queueing networks, appointment scheduling, and capacity management in healthcare. We refer interested readers to the classic text by Gittins *et al.* (2011) and to Bertsimas and Nino-Mora (2000) (and the references therein) for a more detailed discussion of applications of RMABPs in clinical trials, manufacturing systems, and queueing networks, among others; to Washburn (2008), Mahajan and Teneketzis (2008), and Ahmad *et al.* (2009) for applications in sensor management; and to Deo *et al.* (2013), Ayer *et al.* (2016), and Lee *et al.* (2018) for applications in healthcare. The optimal policy for an RMABP is rarely an index policy and it is frequently difficult to determine in any tractable manner (cf. Papadimitriou and Tsitsiklis (1999)). As a result, Whittle (1988) proposed to solve a relaxed version of the RMABP in which the number of activated bandits per period is no longer fixed, but has an upper bound given by the total number of bandits. He then defined an indexability property to ensure the relaxed problem has an optimal policy that is similar to a Gittins index policy, i.e. one that assigns each bandit an index and activates bandits based on their index. This policy, famously known as *Whittle's index policy*, approximates the solution of the original RMABP and reduces to the Gittins index policy when nonactivated bandits do not change their states. Whittle conjectured that this index policy is asymptotically optimal when the number of bandits that can be activated per period and the size of the population of bandits grow proportionally large. This conjecture was proved in the seminal work of Weber and Weiss (1990) for the case of bandits that are governed by the same probability transition matrix, as long as the differential equation corresponding to the fluid approximation of the index policy has a globally stable attractor. Weber and Weiss also showed that Whittle's index policy can fail to be asymptotically optimal if the global attractor condition is not satisfied.

In the same asymptotic setting considered in Whittle (1988) and Weber and Weiss (1990), we introduce an asymptotically optimal heuristic control, called *randomized assignment control (RAC)*, that does *not* require an indexability property and applies to general, finite-horizon fixed population *and* dynamic population RMABPs. RAC works as follows. At each time period, we check the state of each bandit. Depending on the state of the bandit under evaluation, we then randomize the action to be applied to the bandit according to a certain probability distribution, as long as we have not exceeded a prespecified budget for that period. Otherwise, we do not activate the bandit. The control parameters under RAC (i.e. the probability of applying a certain action to a bandit in a given state) are computed using the optimal solution of a linear programming (LP) relaxation of the RMABP. Although the asymptotic optimality of RAC in this paper is proved for the finite-horizon problem, our analysis shows that RAC remains asymptotically optimal when the length of the decision horizon grows at a certain rate. We explain in detail this important point in discussions at several places throughout the paper (cf. Propositions 1 and 2 in Sections 4 and 6, respectively).

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

In addition to providing a theoretical analysis for the asymptotic optimality of RAC, we also show numerically that RAC performs sufficiently well in most cases considered in the numerical study, even for instances when the number of bandits that can be activated at each period and the size of the population of bandits are relatively small (i.e. the nonasymptotic setting). This suggests that RAC can be used in many applications. As mentioned, we consider both fixed and dynamic population RMABPs. In the dynamic population case, we allow bandits to arrive (and leave) stochastically at the beginning of every period (cf. Verloop (2016)). Again, RAC is shown to be asymptotically optimal, thereby extending the utility of RAC to applications that can be modeled as discrete-time controlled Markov processes, including server scheduling problems and flow and server control in discrete-time queueing systems, when decisions are made in batches (cf. Nain and Ross (1986) and Altman (1999, Chapters 5 and 10)).

We view our work as having the following contributions.

1. To the best of our knowledge, we are the first to propose an asymptotically optimal heuristic control for a general finite-horizon RMABP for fixed and dynamic population models. Our LP relaxation approach provides a tractable alternative to solving a dynamic program (DP) exactly, which is usually not tractable.

2. Our proposed heuristic control does *not* rely on any structural assumptions made in the existing literature. These assumptions include indexability properties, as well as assumptions on bandit dynamics, such as the global attractor property referenced above and/or assumptions regarding the recurrence structure of the underlying Markov chain—see Section 2 for a more detailed discussion.

3. We allow for a nonstationary (or time-dependent) bandit activation budget and an arbitrary finite number of actions, without having to restrict to policies that implement a so-called *dominant action*, i.e. the optimal policy always chooses the same action for each activated bandit from a specific class and in a specific state. (Indeed, our analysis can also be applied to the setting with nonstationary transition matrix and nonstationary cost function.) We emphasize that the analysis of the RMABP with more than two actions has remained elusive in the literature unless additional structure is assumed, such as the dominant action. Our analysis can be easily extended to the setting with multi-class bandits (cf. Verloop (2016)).

The remainder of the paper is organized as follows. In Section 2 we summarize related literature and highlight our contributions. Section 3 details the basic model with a fixed population of bandits, along with the corresponding LP relaxation of the stochastic RAMBP, and some preliminary results. In Section 4 we provide the definition of RAC and analyze its performance for the fixed population model under both the total and discounted expect cost criteria. In Section 5 we provide results from simple numerical examples to test the nonasymptotic performance of RAC. In Section 6 we analyze the performance of RAC in the setting with a dynamic population of bandits. Section 7 concludes the paper.

2. Related literature

Here we only review papers that are most closely related to our study and refer interested readers to the classic text by Gittins *et al.* (2011) for a systematic and comprehensive treatment of MABPs and to the recent paper by Verloop (2016) for other related references.

Existing literature on the MABP and RMABP often assumes a stationary activation budget (i.e. the maximum number of bandits that can be activated per period is independent of time)

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

and only two possible actions per bandit (Verloop (2016)). To the best of our knowledge, a nonstationary activation budget has only been considered in Cohen *et al.* (2014). The contrast between Cohen *et al.* (2014) and our work is that we allow for an arbitrary finite number of actions and states, and we focus on developing an asymptotically optimal policy instead of deriving sufficient conditions for optimality of a certain policy form.

RMABPs with multiple actions are often referred to as superprocesses and were first considered for the MABP in the setting where only one bandit can be activated per period (Whittle (1980)). For this setting, it was shown that, under the condition that each state has a dominant action, there is an optimal policy that is indexable. A less strict condition is given in Gittins *et al.* (2011), but still has the same purpose of ensuring that there is an optimal policy that is indexable. Multiple actions were also considered in Verloop (2016) and generalize the concept of dominant action to their setting. Our proposed RAC makes no such restriction to a class of dominant action policies.

Verloop (2016) analyzed multi-class restless bandits with a long-run average cost criteria. A class of priority policies that do not require indexability are introduced, and their asymptotic optimality is analyzed for a fixed and dynamic population of bandits that can arrive and depart from the system. Conditions for asymptotic optimality require that the differential equation that describes the fluid model associated with the RMABP have a globally attracting equilibrium point, similar to the conditions needed in Weber and Weiss (1990). This condition guarantees that the equilibrium point induces a priority policy and that the process under this priority policy converges to the equilibrium point independent of the initial conditions. To guarantee the condition holds, the family of processes that scales to the fluid model must each have a unique invariant probability distribution with finite first moment, and the collection of these unique invariant probability distributions must be tight and uniformly integrable. For a fixed population of bandits, these conditions are satisfied when the generated Markov process is unichain (i.e. the state space for the Markov chain has a single recurrent class C and there is no other absorbing set disjoint from C), so that the resulting Markov chain has a unique equilibrium distribution. For a dynamic population of bandits, they are satisfied when the generated Markov process is irreducible and state 0 (i.e. the 'empty' state) is positive recurrent for any bandit that is never activated, i.e. inactivated bandits eventually leave the system.

Our dynamic population framework differs from that in Verloop (2016) in four ways. First, we allow time-varying constraints on the number of active bandits per period (again, our results can also be extended to the setting with nonstationary transition matrices and nonstationary costs). Second, we consider a finite-horizon setting, so we do not require a global attractor condition. Third, we make no assumptions about the transition probability matrices or the underlying dynamics generated by the policies of consideration, such as being unichain. Fourth, we propose a new type of policy that is neither a priority policy nor an index policy, but is still asymptotically optimal under certain relatively mild conditions.

Kelly (1981) considered an expected long-run average reward criterion for a finite population bandit. Each arm generates a sequence of Bernoulli random variables whose parameters are themselves random variables, and are independent with common distribution satisfying a regularity condition. He showed that, as the discount factor approaches 1, the optimal policy converges to that which pulls the arm(s) that has(ve) incurred the least number of 'failures'. Moreover, as the discount factor approaches 1, the performance of the optimal discounted expected reward policy as judged with respect to the expected average reward criterion approaches the best possible.

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

Our work is also related to the literature that uses LP relaxation to approximate RMABPs. Bertsimas and Nino-Mora (2000) were the first to consider a sequence of LP relaxations to obtain a primal-dual index policy for the infinite-horizon discounted-cost RMABP. Le Ny *et al.* (2008) extended the work of Bertsimas and Nino-Mora (2000) to include switching times/costs between activating bandits. Both Bertsimas and Nino-Mora (2000) and Le Ny *et al.* (2008) considered LP relaxations of the dynamic program formulation of infinite-horizon RMABPs. An alternative LP relaxation can be derived by considering a fluid model of the RMABP that only takes into account mean drifts of bandit dynamics (Weber and Weiss (1990) and Verloop (2016)). This fluid approach is also called the *certainty equivalent* approach in the operations research literature and is closest to the LP formulation in this paper.

Finite-horizon MABPs have been studied in Robbins (1952) and Bradt *et al.* (1956), who focused on the case where engaging a project corresponds to sampling from a Bernoulli population with unknown success probability and the objective is to maximize the expected number of successes over a finite number of plays. We refer interested readers to the monograph by Berry and Fristedt (1985) for additional references on finite-horizon MABPs. More recently, Caro and Gallien (2007), considered the setting motivated by a dynamic assortment problem in the fashion retail industry. Nino-Mora (2011) (see also the references therein) considered a class of finite-horizon discrete-state bandit problems whose optimal policy is known to be of index type (i.e. the counterpart of the Gittins index for a finite-horizon discrete-state bandit), and proposed an efficient and exact algorithm to compute the index.

Finite horizon RMABPs have been less studied. To the best of our knowledge, the work by Hu and Frazier (2017) is the closest to the current study. Hu and Frazier (2017) considered the RMABP with a finite number of independent and identically distributed bandits which can be either activated or not, except over a finite horizon and with a fixed and stationary number M of multiple pulls per period. They proposed an index policy for this model and showed that it is asymptotically optimal. Their asymptotically optimal index policy was obtained in three main steps. First, they constructed a Lagrangian relaxation of the original RMABP, which provides an upper bound to the optimal value of the RMABP, and computed Lagrange multipliers that minimize the value of the RMABP relaxation using subgradient descent. Second, they decomposed the relaxation into smaller Markov decision processes (MDPs), one for each bandit, and each of which can be optimally solved using backward induction. For each smaller MDP, they obtained an 'index' for each time period. They then constructed an index policy for the original RMABP by activating the M largest indices corresponding to the smaller MDPs in each time period. One potential issue is that there can be ties between indices so that it may not be possible to choose exactly M indices. As a result, the third step is a tie-breaking rule that activates the remaining subprocesses between tied states according to a probability distribution induced by an optimal policy that is obtained by solving a linear program with respect to occupation measures and that depends on the Lagrange multipliers.

In addition to our population and multi-action setting, there are three other components of our approach that distinguish it from that of Hu and Frazier (2017) and which make their approach not directly applicable to ours. First is the objective. We introduce a set of undesirable states, which are the states in which a bandit will incur a high penalty cost of at the end of the horizon. In particular, a cost is incurred for each excess bandit that ends up in an undesirable state immediately after the end of the horizon and we allow for at most a fixed number of bandits to be in any of the undesirable states at this time without being penalized. The dependency between the bandits introduced by this terminal cost implies that the Lagrangian relaxation and the subsequent decomposition into sub-MDPs in Hu and Frazier (2017) is not

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

directly applicable. The second difference is that we allow for the number of bandits to be activated in each period to be upper bounded by a time-dependent constant. Because the latter constraint is nonbinding, the Lagrangian relaxation of Hu and Frazier is again not directly applicable. Finally, the index policy in Hu and Frazier (2017) is based on aggregating 'indices' from smaller MDPs based on the Lagrangian relaxation of the original RMABP and on a tie-breaking rule, akin to our formulation of our stochastic control problem, based on an LP formulation for state occupation measures and the Lagrange multipliers. By contrast, ours is based on an LP relaxation where the random variables in the original RMABP are replaced by their means.

Another potential distinction from Hu and Frazier (2017) is that our analysis can be extended to include nonstationary transition probability matrices and costs and state-dependent terminal costs. It can also be generalized to the setting where we have several constraints for several subsets of bandits or one constraint on the total number of activated bandits over the entire finite-time horizon. The proposed approach would also work for the multi-action setting where each action contributes a different weight to the budget. To the best of our knowledge, we are the first to analyze finite-horizon RMABPs with a nonstationary activation budget and asymptotically optimal policies that are nonindexable.

3. Fixed population model

We now describe the setting of a fixed population model where the number of bandits in the system remains constant throughout the horizon (i.e. no new bandits arrive and no existing bandits depart); this setting will be assumed throughout Sections 3 to 5. We consider a finite-horizon, discrete-time model, where time $t \in \{1, \ldots, T+1\}$ with $T < \infty$. Let $\mathbb{J} = \{j\colon 1 \le j \le J\}$ with $J < \infty$ denoting the set of feasible states, and let $\mathbb{A} = \{a\colon 0 \le a \le A\}$ with $A < \infty$ denoting the set of feasible actions. Our results and analysis will still apply when each state has its own set of feasible actions; so, without loss of generality, we will simply assume that all states share the same set of feasible actions $\mathbb{A}$. We also introduce a set of states $\mathbb{U} \subseteq \mathbb{J}$, called the *undesirable* states, which are the states in which a bandit will incur a high penalty cost at the end of the horizon. Undesirable states can represent machine failure in maintenance and replacement models, customers that did not complete service before the end of the horizon or abandoned the system before completing service in queueing scheduling applications, or deteriorated/critical health conditions in healthcare applications. We refer to action $a = 0$ as *no action* (or *no treatment*) and any action $a > 0$ as a *proper action* (or *proper treatment*). Moreover, using the standard terminology from the RMABP literature, we will call a bandit *active* when a proper action is applied to it and *passive* otherwise. We assume that the state of each bandit transitions from state i to state j under action a according to a probability p_{ij}^a and that the maximum number of active bandits at time t is $b_t > 0$, which we call the *activation budget*. Two types of costs can be incurred by the system. First, a cost c_j^a is incurred each time action a is applied to a bandit in state j. Second, a cost ϕ is incurred for each excess bandit that ends up in an undesirable state at time $T+1$; we allow for at most $m \ge 0$ bandits to be in any of the undesirable states at time $T+1$ without being penalized. Finally, we use n_{tot} to denote the total number of bandits in the system. For ease of exposition, we have used a time-independent transition matrix and cost function, but remark that our analysis can be applied to the setting with a time-dependent transition matrix and cost function.

The decision-making scenario is as follows. At the beginning of time t, the decision maker views the state of each bandit in the system and then decides the action/treatment to be applied to each bandit. After receiving the treatment, each bandit incurs a cost and its state transitions

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

to a potentially new state at the beginning of time $t+1$. The objective of the decision maker is to minimize the expected total treatment and penalty costs, or simply, the expected total *costs*.

Let Π denote the set of all nonanticipating controls, and let π denote a feasible control in Π. For a given control π, since all bandits share identical transition dynamics and costs, we do need to keep track of the individual bandits; instead, it is sufficient to simply keep track of the number of bandits in state j that receive treatment a at time t, which we denote by $X_j^{\pi,a}(t)$, as well as the number of bandits in state i that receive treatment a at time t whose states transition to state j at time $t+1$, which we denote by $Y_{ij}^{\pi,a}(t)$. Let n_j denote the number of bandits in state j at time $t=1$, and let V^S denote the expected total costs under an optimal control π^*. Then

$$V^S = \min_{\pi\in\Pi} \mathbb{E}^{\pi}\left[\sum_{t=1}^{T}\sum_{a=0}^{A}\sum_{j=1}^{J} c_j^a \cdot X_j^{\pi,a}(t) + \phi \cdot \left(\sum_{a=0}^{A}\sum_{j\in\mathbb{U}} X_j^{\pi,a}(T+1) - m\right)^+\right] \tag{1a}$$

such that

$$\sum_{a=0}^{A} X_j^{\pi,a}(t) = \sum_{a=0}^{A}\sum_{i=1}^{J} Y_{ij}^{\pi,a}(t-1) \quad \text{for all } j \geq 1,\ t \geq 2, \tag{1b}$$

$$\sum_{a=0}^{A} X_j^{\pi,a}(1) = n_j \quad \text{for all } j \geq 1, \tag{1c}$$

$$\sum_{a=1}^{A}\sum_{j=1}^{J} X_j^{\pi,a}(t) \leq b_t \quad \text{for all } t \geq 1, \tag{1d}$$

where the constraints hold almost surely (i.e. with probability 1).

We remark that the seemingly simple stochastic control model (1) is in fact surprisingly general. As alluded to in Section 1, it can be used for machine maintenance and replacement problems or capacity management and/or resource allocation in healthcare, among others. Additionally, this model can be modified to include dynamic populations (i.e. 'birth' and 'death' RMABP) by including *nonstationary* random arrival processes, *nonstationary* random service capacity (e.g. nonstationary budget), random service completions, and random abandonments. Lastly, we remark that the formulation (1) includes the budget constraints explicitly, which are typically included in the definition of the class Π of admissible controls. In what follows, and for ease of notation, whenever it is clear from the context which policy is being used, we will suppress the notational dependency on π.

In theory, the stochastic control problem (1) can be solved using the standard DP algorithm. Unfortunately, this approach suffers from the well-known curse of dimensionality. Thus, rather than solving (1) exactly, we instead construct and analyze a provably near-optimal heuristic control. To describe our heuristic control in Section 4, we need to introduce the following family of linear programs indexed by $\boldsymbol{\epsilon} := (\epsilon_1, \epsilon_2, \ldots, \epsilon_T) \geq \mathbf{0} = (0, 0, \ldots, 0)$:

$$V^D(\boldsymbol{\epsilon}) = \min_{x,z} \sum_{t=1}^{T}\sum_{a=0}^{A}\sum_{j=1}^{J} c_j^a \cdot x_j^a(t, \boldsymbol{\epsilon}) + \phi \cdot z(\boldsymbol{\epsilon}) \tag{2a}$$

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

such that

$$\sum_{a=0}^{A} x_j^a(t,\boldsymbol{\epsilon}) = \sum_{a=0}^{A}\sum_{i=1}^{J} x_i^a(t-1,\boldsymbol{\epsilon})\cdot p_{ij}^a \quad \text{for all } j\geq 1,\ t\geq 2, \tag{2b}$$

$$\sum_{a=0}^{A} x_j^a(1,\boldsymbol{\epsilon}) = n_j \quad \text{for all } j\geq 1, \tag{2c}$$

$$\sum_{a=1}^{A}\sum_{j=1}^{J} x_j^a(t,\boldsymbol{\epsilon}) \leq b_t - \epsilon_t \quad \text{for all } t\geq 1, \tag{2d}$$

$$z(\boldsymbol{\epsilon}) \geq \sum_{j\in\mathbb{U}}\sum_{a=0}^{A}\sum_{i=1}^{J} x_i^a(T,\boldsymbol{\epsilon})\cdot p_{ij}^a - m, \tag{2e}$$

$$z(\boldsymbol{\epsilon}),\ x_j^a(t,\boldsymbol{\epsilon}) \geq 0 \quad \text{for all } a\geq 0,\ j\geq 1,\ t\geq 1. \tag{2f}$$

One can arrive at linear program (2) by replacing all the random variables in (1) with their expected values, replacing the original activation budget b_t in (1) with $b_t - \epsilon_t$, and introducing a new variable $z(\boldsymbol{\epsilon})$ to capture the number of excess bandits in undesirable states at the end of the time horizon. In the special case when $\boldsymbol{\epsilon} = \mathbf{0}$, linear program (2) is simply the deterministic relaxation of the stochastic control problem (1). Otherwise, when $\boldsymbol{\epsilon} > \mathbf{0}$, we interpret vector $\boldsymbol{\epsilon}$ as a buffer for making sure that the activation budgets are not exceeded when a heuristic control constructed using the solution of (2) is implemented in the original stochastic system.

We end this section with the following result.

Lemma 1. *It holds that* $V^D(\mathbf{0}) \leq V^S$.

Proof. Let π^* denote an optimal policy for our original stochastic control problem (1), and let $x_j^a(t,\mathbf{0}) = \mathbb{E}[X_j^{\pi^* a}(t)]$ for all a, j, t. By Jensen's inequality,

$$\mathbb{E}\left[\left(\sum_{a=0}^{A}\sum_{j\in U} X_j^{\pi^* a}(T+1) - m\right)^+\right] \geq \sum_{a=0}^{A}\sum_{j\in U} x_j^a(T+1,\mathbf{0}) - m.$$

Thus, let

$$z(\mathbf{0}) = \begin{cases} \displaystyle\sum_{a=0}^{A}\sum_{j\in U} x_j^a(T+1,\mathbf{0}) - m & \text{if } \sum_{a=0}^{A}\sum_{j\in U} x_j^a(T+1,\mathbf{0}) \geq m, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that $x_j^a(t,\mathbf{0})$ for all a, j, t and $z(\mathbf{0})$ provide a feasible solution to (2) with $\boldsymbol{\epsilon} := (\epsilon_1, \epsilon_2, \ldots, \epsilon_T) = \mathbf{0} = (0, 0, \ldots, 0)$ having total costs in (2) smaller than the total expected costs in (1) under π^*. This yields $V^D(\mathbf{0}) \leq V^S$. □

Lemma 1 allows us to use $V^D(\mathbf{0})$ as a proxy for V^S and study the performance of any feasible control π by analyzing the difference between V^π and $V^D(\mathbf{0})$, as claimed.

4. Randomized activation control

We now discuss our heuristic control, which we call *randomized activation control* (RAC). To do this, we first need to introduce some notation. For a given $\boldsymbol{\epsilon} \geq \mathbf{0}$, let $x_j^{*,a}(t,\boldsymbol{\epsilon})$ and $z^*(\boldsymbol{\epsilon})$

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

denote an optimal solution of linear program (2). Define $n_j^*(t, \boldsymbol{\epsilon})$ and $q_j^{*,a}(t, \boldsymbol{\epsilon})$ as

$$n_j^*(t, \boldsymbol{\epsilon}) = \sum_{a=0}^{A} x_j^{*,a}(t, \boldsymbol{\epsilon}), \qquad q_j^{*,a}(t, \boldsymbol{\epsilon}) = \begin{cases} \dfrac{x_j^{*,a}(t, \boldsymbol{\epsilon})}{n_j^*(t, \boldsymbol{\epsilon})} & \text{if } n_j^*(t, \boldsymbol{\epsilon}) > 0, \\ 1\{a = 0\} & \text{if } n_j^*(t, \boldsymbol{\epsilon}) = 0. \end{cases}$$

Note that, by definition, we have $n_j^*(1, \boldsymbol{\epsilon}) = n_j$ for all j. Note that $n_j^*(t, \boldsymbol{\epsilon})$ is the total number of bandits in state j at time t (in the deterministic system) and $q_j^{*,a}(t, \boldsymbol{\epsilon})$ is the fraction of bandits in state j at time t that receive action a. Let $Z_{jl}(t, \boldsymbol{\epsilon})$ be a categorical random variable on the set $\mathbb{A}$ with the property that

$$\mathbb{P}\left(Z_{jl}(t, \boldsymbol{\epsilon}) = a\right) = q_j^{*,a}(t, \boldsymbol{\epsilon}) \quad \text{for all } a \in \mathbb{A},$$

where ℓ denotes a bandit. Also, let $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_{n_{\text{tot}}})$ denote a permutation (or ordering) of all of the bandits in the system, where $n_{\text{tot}} = \sum_{j=1}^{J} n_j$, and let $j(t, \ell)$ denote the state of bandit ℓ at time t before any action is applied. The description of RAC is given below.

Algorithm 1. (*RAC.*)

1. Pick $\boldsymbol{\epsilon}$ and solve linear program (2).
2. Compute $q_j^{*,a}(t, \boldsymbol{\epsilon})$ for all a, j, and t.
3. Pick a permutation $\boldsymbol{\varphi}$.
4. For time $t = 1$ to T, do:
   (a) set $k = 1$;
   (b) randomly generate $Z_{j(t,\varphi(k)),k}(t, \boldsymbol{\epsilon})$;
   (c) if $\sum_{m=1}^{k} 1\{Z_{j(t,\varphi(m)),m}(t, \boldsymbol{\epsilon}) \neq 0\} \geq b_t$, apply $a = 0$ to bandit $\varphi(k)$; otherwise, apply action $Z_{j(t,\varphi(k)),k}(t, \boldsymbol{\epsilon})$ to bandit $\varphi(k)$;
   (d) update $k = k + 1$;
   (e) while $k \leq n_{\text{tot}}$, go back to step (b).

The idea behind RAC is as follows. At each time t, we check the bandits in the order prescribed by φ. Suppose that we are currently checking bandit ℓ. If its current state is j then we apply action a to ℓ with probability $q_j^{*,a}(t, \epsilon)$ as long as we have not exceeded the allowed budget in period t; otherwise, we simply do not apply a proper treatment to ℓ. It turns out that the order in which we assess the bandits in RAC (i.e. the permutation $\boldsymbol{\varphi}$) does not affect our analysis and results. Thus, for simplicity, we will simply use $\varphi(k) = k$ for all k. The choice of $\boldsymbol{\epsilon}$, on the other hand, significantly affects the performance of RAC. To see this, note that we can decompose the loss of RAC as

$$0 \leq V^{\text{RAC}} - V^D(\mathbf{0}) = [V^{\text{RAC}} - V^D(\boldsymbol{\epsilon})] + [V^D(\boldsymbol{\epsilon}) - V^D(\mathbf{0})].$$

First note that the first inequality follows from Lemma 1 since $V^{\text{RAC}} \geq V^S$. Next, observe that if all components of $\boldsymbol{\epsilon}$ are close to 0, then $V^D(\boldsymbol{\epsilon}) - V^D(\mathbf{0})$ is also close to 0. However, $V^{\text{RAC}} - V^D(\boldsymbol{\epsilon})$ could be large. This is so because the deviations of $X_j^a(t)$ from $x_j^{*,a}(t, \boldsymbol{\epsilon})$ (due to the stochastic nature of the system) are likely to lead RAC to exhaust the activation budget at time t before those bandits that should have been activated are even considered for activation.

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

As a result, many bandits may end up in undesirable states at the beginning of period $T+1$. On the other hand, if the components of $\boldsymbol{\epsilon}$ are large, $V^{\mathrm{RAC}} - V^D(\boldsymbol{\epsilon})$ will be close to 0. This is because there is a negligible probability that we will exhaust the activation budget in each period and we will be able to activate all bandits that need to be activated. However, the term $V^D(\boldsymbol{\epsilon}) - V^D(\mathbf{0})$ can be large (see Lemma 2 below). It follows that care must be taken when choosing $\boldsymbol{\epsilon}$.

The following lemma provides an upper bound for $V^D(\boldsymbol{\epsilon}) - V^D(\mathbf{0})$.

Lemma 2. *There exists a constant $M > 0$, independent of T, and a vector $\boldsymbol{\epsilon} \geq \mathbf{0}$ satisfying $\epsilon_t \leq b_t$ for all t such that*

$$V^D(\boldsymbol{\epsilon}) - V^D(\mathbf{0}) \leq M \cdot \left[\sum_{t=1}^{T} \epsilon_t \right].$$

Lemma 2 says that $V^D(\boldsymbol{\epsilon}) - V^D(\mathbf{0})$ is roughly proportional to $\sum_{t=1}^{T} \epsilon_t$. The proof of this lemma can be found in Appendix A and utilizes a standard duality argument for LP sensitivity analysis.

Let C^{RAC} denote a realization of total costs under RAC. By definition, we have $\mathbb{E}[C^{\mathrm{RAC}}] = V^{\mathrm{RAC}}$. The following lemma tells us that $C^{\mathrm{RAC}} - V^D(\epsilon)$ too is roughly proportional to $\sum_{t=1}^{T} \epsilon_t$, with a positive probability.

Lemma 3. *For any $\boldsymbol{\epsilon} \geq \mathbf{0}$, we have*

$$C^{\mathrm{RAC}} - V^D(\boldsymbol{\epsilon}) \leq [\phi + \max_{a,j} c_j^a] \cdot \left[\sum_{t=1}^{T} \epsilon_t \right] \tag{3}$$

with probability at least

$$1 - 3 \cdot A \cdot J^2 \cdot \sum_{t=1}^{T} \exp\left\{ - \frac{\epsilon_t^2}{12 \cdot A^2 \cdot J^4 \cdot [\sum_{\ell=1}^{J} n_\ell]} \right\} - 3 \cdot A \cdot J^2 \cdot \sum_{t=1}^{T} \exp\left\{ - \frac{\epsilon_t}{6 \cdot A \cdot J^2} \right\}. \tag{4}$$

If the components of $\boldsymbol{\epsilon}$ are uniformly small then the bound in (3) is small and holds with a small probability. If, on the other hand, the components of $\boldsymbol{\epsilon}$ are uniformly large, the bound in (3) is large and holds with a large probability. Ideally, we would like to choose $\boldsymbol{\epsilon}$ that yields a small bound in (3) that holds with a large probability.

To better characterize the impact of the choice of $\boldsymbol{\epsilon}$ on the performance of RAC, we now turn to an asymptotic setting where n_j (for all j), b_t (for all t), and m are uniformly scaled by a factor of $\theta > 0$. This is the standard asymptotic setting considered in the RMABP literature (e.g. Weber and Weiss (1990) and Verloop (2016)). Let V_θ^S and V_θ^π denote the expected total costs under the optimal control and a feasible control π, respectively, when we use the scaling factor θ. Also, let $V_\theta^D(\boldsymbol{\epsilon})$ denote the optimal value of the corresponding linear program (2). It is not difficult to see that the optimal solution of linear program (2) for $\boldsymbol{\epsilon} = \mathbf{0}$ and $\theta > 0$ is given by

$$x_{\theta,j}^{*,a}(t, \mathbf{0}) = \theta \cdot x_j^{*,a}(t, \mathbf{0}) \quad \text{and} \quad z_\theta^*(\mathbf{0}) = \theta \cdot z^*(\mathbf{0}),$$

which implies that $V_\theta^D(\mathbf{0}) = \theta \cdot V^D(\mathbf{0})$. We also define the terms $n_{\theta,j}^*(t, \boldsymbol{\epsilon})$ and $q_{\theta,j}^{*,a}(t, \boldsymbol{\epsilon})$ analogously to how $n_j^*(t, \boldsymbol{\epsilon})$ and $q_j^{*,a}(t, \boldsymbol{\epsilon})$ are defined, with $x_j^{*,a}(t, \mathbf{0})$ being replaced with $x_{\theta,j}^{*,a}(t, \mathbf{0})$.

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

Let $n_{\text{tot}} := \sum_{i=1}^{J} n_i$, and define $S(t) := \{(a, j) : x_j^{*,a}(t, \mathbf{0}) > 0 \text{ and } c_j^a > 0\}$. We state our first theorem below.

Theorem 1. *Let d be a positive number. Suppose that $S(t) \neq \varnothing$ for all t, and let*

$$\epsilon_t = 6 \cdot A \cdot J^2 \cdot \sqrt{d \cdot n_{\text{tot}} \cdot \theta \cdot \ln \theta}$$

for all time periods t. For all $\theta > \theta^*$ where

$$\frac{\theta^*}{\ln \theta^*} = \frac{36 \cdot A^2 \cdot J^4 \cdot d \cdot n_{\text{tot}}}{\min_t b_t}, \tag{5}$$

there exists a constant $M > 0$ independent of T, d, and θ such that

$$\frac{V_\theta^{\text{RAC}} - V_\theta^S}{V_\theta^S} \leq M \cdot \left[\frac{T}{\theta^d} + \sqrt{\frac{d \cdot \ln \theta}{\theta}} \right].$$

Proof. Fix $d > 0$. It is not difficult to see that condition (5) is equivalent to $\theta \cdot b_t > \epsilon_t$ for all t. Its role is to guarantee that the right-hand side of budget constraints in linear program (2) is positive and, therefore, the linear program has a feasible solution. Now, note that $\sqrt{d \cdot n_{\text{tot}} \cdot \theta \cdot \ln \theta} \geq d \cdot \ln \theta$ for all large θ. Thus, by Lemma 3, under the choice of $\boldsymbol{\epsilon}$ in Theorem 1,

$$C_\theta^{\text{RAC}} - V_\theta^D(\boldsymbol{\epsilon}) = O(T \cdot \sqrt{d \cdot \theta \cdot \ln \theta})$$

with probability at least

$$1 - \Theta\left(\frac{T}{\theta^{3d}}\right) - \Theta\left(\frac{T}{\theta^d}\right) = 1 - \Theta\left(\frac{T}{\theta^d}\right).$$

Since there is a total of $\theta \cdot n_{\text{tot}}$ bandits in each period, $C_\theta^{\text{RAC}} - V_\theta^D(\boldsymbol{\epsilon})$ is at most $\Theta(T \cdot \theta \cdot n_{\text{tot}})$ with probability at most $\Theta(T/\theta^d)$. We conclude that

$$V_\theta^{\text{RAC}} - V_\theta^D(\boldsymbol{\epsilon}) = O\left(T \cdot \sqrt{d \cdot \theta \cdot \ln \theta} + \frac{T^2}{\theta^{d-1}}\right).$$

By Lemma 2, under the choice of $\boldsymbol{\epsilon}$ in Theorem 1, we have

$$V_\theta^D(\boldsymbol{\epsilon}) - V_\theta^D(\mathbf{0}) = O(T \cdot \sqrt{d \cdot \theta \cdot \ln \theta}).$$

Putting the bounds for $V_\theta^{\text{RAC}} - V_\theta^D(\boldsymbol{\epsilon})$ and $V_\theta^D(\boldsymbol{\epsilon}) - V_\theta^D(\mathbf{0})$ together yields

$$V_\theta^{\text{RAC}} - V_\theta^D(\mathbf{0}) = O\left(T \cdot \sqrt{d \cdot \theta \cdot \ln \theta} + \frac{T^2}{\theta^{d-1}}\right).$$

The proof is complete by noting that $(V_\theta^{\text{RAC}} - V_\theta^S)/V_\theta^S \leq (V_\theta^{\text{RAC}} - V_\theta^D(\mathbf{0}))/V_\theta^D(\mathbf{0})$ and $V_\theta^D(\mathbf{0})$ is $\Theta(T \cdot \theta)$ (because $S(t) \neq \varnothing$ for all t, which implies that in each period we always incur a total cost that is at least proportional to θ). □

Remarks 1. (*Theorem 1.*) (a) The condition $S(t) \neq 0$ for all t simply means that, in each period, we always activate some bandits in the deterministic system. This condition is quite mild and is immediately satisfied, for example, when $c_j^a > 0$ for all a, j. It is also satisfied when, for each action a, c_j^a is monotone nondecreasing in j so that lower (higher) states correspond to 'better' ('worse') states. The latter condition is typically satisfied in the context of machine maintenance and replacement problems and capacity management/resource allocation problems in healthcare.

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

(b) The function $f(\theta) = \theta/\ln\theta$ is increasing on $[0, \infty)$ with $f(0) = 0$ and $\lim_{\theta\to\infty} f(\theta) = \infty$. As a result, the parameter θ^* in (5) is well defined.

(c) Since we do *not* scale T in our asymptotic setting, the bound in Theorem 1 tells us that RAC is asymptotically optimal as long as $d > 0$. However, the bound depends on T only through the term T/θ^d. This means that we can also consider an alternative asymptotic setting where T is allowed to grow as a function of θ. For example, if $T \sim \theta^n$, where n is an arbitrary natural number, then we can choose $d > n$ and RAC is still asymptotically optimal. Thus, while our model considers a finite-horizon setting, RAC is versatile in the sense that it can be applied to problems with a very long decision horizon.

The bound in Theorem 1 is for the expected total costs criteria, but our analysis can also be applied to the expected total discounted cost criteria. This is detailed in Proposition 1 below, which is the direct analogue of Theorem 1.

Proposition 1. *Let d be a positive number. Suppose that we multiply the cost at time t with δ^{t-1} for some discount factor $\delta \in (0, 1)$. Suppose that $S(t) \neq 0$ for all t. Let*

$$\epsilon_t = 6 \cdot A \cdot J^2 \cdot \sqrt{d \cdot n_{\text{tot}} \cdot \ln(t + e - 1) \cdot \theta \cdot \ln\theta}$$

for all t (note that ϵ_t now depends on t). For all $\theta > \theta^*$ where

$$\frac{\theta^*}{\ln\theta^*} = \frac{36 \cdot A^2 \cdot J^4 \cdot d \cdot n_{\text{tot}} \cdot \log(T + \mathrm{e} - 1)}{\min_t b_t},$$

there exists a constant $M > 0$ independent of T and θ such that

$$\frac{V_\theta^{\text{RAC}} - V_\theta^S}{V_\theta^S} \leq \frac{V_\theta^{\text{RAC}} - V_\theta^D(\mathbf{0})}{V_\theta^D(\mathbf{0})} \leq M \cdot \left[\frac{1}{\theta^d} + \sqrt{\frac{d \cdot \ln\theta}{\theta}}\right].$$

Proof. The proof is akin to the proof of Theorem 1. Here, we will only provide its outline. Note that, when we multiply the cost at time t with δ^{t-1} for some discount factor $\delta \in (0, 1)$, the bound in Lemma 2 becomes

$$V^D(\boldsymbol{\epsilon}) - V^D(\mathbf{0}) \leq M \cdot \left[\sum_{t=1}^{T} \delta^{t-1} \cdot \epsilon_t\right]$$

for some $M > 0$ independent of T and $\boldsymbol{\epsilon} \geq \mathbf{0}$. The bound can be shown using exactly the same arguments as used in the proof of Lemma 2. See Remark 4 in Appendix A.1. Similarly, the bound in (3) becomes

$$C^{\text{RAC}} - V^D(\boldsymbol{\epsilon}) \leq \left[\phi + \max_{a,j} c_j^a\right] \cdot \left[\sum_{t=1}^{T} \delta^{t-1} \cdot \epsilon_t\right],$$

which can be shown using exactly the same arguments as used in the proof of Lemma 3. See Appendix A.2. Next, observe that, akin to the proof of Theorem 1, our choice of ϵ_t implies that

$$C_\theta^{\text{RAC}} - V_\theta^D(\boldsymbol{\epsilon}) = O\left(\sum_{t=1}^{T} \delta^{t-1} \cdot \sqrt{d \cdot \ln(t + \mathrm{e} - 1) \cdot \theta \cdot \ln\theta}\right) = O(\sqrt{d \cdot \theta \cdot \ln\theta})$$

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

with probability at least

$$1-\Theta\Bigg(\sum_{t=1}^{T}\frac{1}{\theta^{3d\cdot\ln(t+\mathrm{e}-1)}}\Bigg)-\Theta\Bigg(\sum_{t=1}^{T}\frac{1}{\theta^{d\cdot\ln(t+\mathrm{e}-1)}}\Bigg)=1-\Theta\Bigg(\sum_{t=1}^{T}\frac{1}{(t+\mathrm{e}-1)^{d\cdot\ln\theta}}\Bigg).$$

Additionally, $C_\theta^{\mathrm{RAC}}-V_\theta^D(\boldsymbol{\epsilon})$ is at most

$$\Theta\Bigg(\sum_{t=1}^{T}\delta^{t-1}\cdot\theta\cdot n_{\mathrm{tot}}\Bigg)=\Theta(\theta\cdot n_{\mathrm{tot}})$$

with probability at most $\Theta(\sum_{t=1}^{T}1/(t+\mathrm{e}-1)^{d\cdot\ln\theta})$. We conclude that

$$V_\theta^{\mathrm{RAC}}-V_\theta^D(\mathbf{0})=O\Bigg(\sqrt{d\cdot\theta\cdot\ln\theta}+\sum_{t=1}^{T}\frac{\theta}{(t+\mathrm{e}-1)^{d\cdot\ln\theta}}\Bigg).$$

Since $V_\theta^D(\mathbf{0})$ is $\Theta(\sum_{t=1}^{T}\delta^{t-1}\cdot\theta)=\Theta(\theta)$ (by our assumption that $S(t)\neq\varnothing$ for all t) and

$$\sum_{t=1}^{T}\frac{1}{(t+\mathrm{e}-1)^{d\cdot\ln\theta}}\leq\frac{1}{\mathrm{e}^{d\cdot\ln\theta}}+\int_1^\infty\frac{\mathrm{d}x}{(x+\mathrm{e}-1)^{d\cdot\ln\theta}}\leq\frac{2}{\mathrm{e}^{d\cdot\ln\theta}}=\frac{2}{\theta^d}$$

for large θ, we have the desired bound. □

Remark 2. (*Discounted cost.*) Akin to Theorem 1, in order for the bound in Proposition 1 to hold, the condition $\theta>\theta^*$ is equivalent to $\theta\cdot b_t\geq\epsilon_t$ holding for all t. This is important for making sure that the right-hand side of linear program (2) is positive. Given our choice of ϵ_t, this condition is immediately satisfied for all t and all large θ so long as $T=o(\exp\{\theta/d\cdot\ln\theta\})$. Thus, not only is RAC asymptotically optimal for discounted expected cost criteria, its relative loss is also *independent* of T for very large T.

5. Numerical experiments

5.1. RAC performance

In this section, we test the performance of RAC using two experiments. In the first experiment, we consider an instance of the RMABP with two states and two actions; in the second experiment we consider an instance of the RMABP with five states and five actions. In each experiment, we use $\epsilon_t=\sqrt{\theta\cdot\log\theta}$ for all t and a wide range of θ. For both experiments, we randomly generated the cost coefficients according to a uniformly distributed random variable $U(0,10)$. We also generated the transition probabilities by first generating a uniformly distributed random variable $U(0,1)$, and then normalizing the generated rows to obtain transition probability matrix P. We used the same budget $b_t=1$ for all t for both experiments. The remaining details regarding all other parameters can be found in the online supplement Zayas-Cabán *et al.* (2019). In what follows, we report the percentage losses of RAC (i.e. $(V_\theta^{\mathrm{RAC}}-V_\theta^D(\mathbf{0}))/V_\theta^D(\mathbf{0})$) and their corresponding confidence intervals out of 1000 Monte Carlo runs for the two experiments in Tables 1 and 2, respectively. Additional simulation scenarios and their corresponding results can be found in the online supplement Zayas-Cabán *et al.* (2019).

As predicted by Theorems 1 and 2, RAC performs better as θ grows large. This confirms the asymptotic optimality of RAC. The performance of RAC for small θ depends on the problem

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

TABLE 1: Percentage loss for two states and two actions.

θ	$T = 10$	$T = 30$	$T = 50$	$T = 100$
1	5.42 (3.72,7.14)	5.35 (3.68,7.19)	5.51 (3.87,6.96)	5.45 (4.01,7.18)
5	2.50 (1.75,3.41)	2.44 (1.6,3.06)	2.58 (1.82,3.43)	2.37 (1.61,3.23)
10	1.82 (1.28,2.35)	1.75 (1.17,2.25)	1.74 (1.20,2.36)	1.79 (1.12,2.38)
20	1.28 (0.87,1.62)	1.27 (0.83,1.66)	1.28 (0.89,1.74)	1.24 (0.85,1.64)
40	0.88 (0.65,1.18)	0.90 (0.58,1.16)	0.89 (0.62,1.21)	0.87 (0.56,1.12)
60	0.76 (0.48,1.03)	0.76 (0.56,0.98)	0.73 (0.52,0.99)	0.74 (0.51,0.91)
80	0.62 (0.43,0.84)	0.64 (0.38,0.84)	0.63 (0.42,0.86)	0.67 (0.42,0.91)
100	0.57 (0.39,0.79)	0.58 (0.39,0.75)	0.56 (0.36,0.78)	0.57 (0.34,0.76)

TABLE 2: Percentage loss for five states and five actions.

θ	$T = 10$	$T = 30$	$T = 50$	$T = 100$
1	142.46 (125.77,157.68)	142.70 (126.05,159.27)	143.72 (130.39,160.71)	142.34 (125.34,158.97)
5	52.93 (45.03,60.50)	51.86 (45.34,60.25)	52.54 (46.44,58.25)	53.01 (46.73,59.61)
10	31.30 (26.20,35.94)	31.48 (26.20,37.44)	31.09 (26.04,35.90)	31.24 (25.90,35.99)
20	17.84 (13.62,22.29)	17.79 (14.08,22.78)	16.89 (12.51,21.15)	17.18 (13.35,22.25)
40	9.40 (6.63,12.36)	8.96 (6.24,11.72)	9.30 (6.03,13.12)	9.07 (6.20,11.59)
60	6.09 (3.31,8.31)	6.38 (3.61,8.61)	6.19 (4.01,9.55)	6.20 (2.75,9.05)
80	4.82 (2.18,7.02)	4.76 (2.78,6.79)	4.62 (4.01,9.55)	4.83 (2.90,7.65)
100	3.97 (1.89,6.28)	3.85 (1.97,6.14)	4.34 (1.77,6.60)	3.42 (0.94,6.24)

parameters (see the online supplement Zayas-Cabán *et al.* (2019)). However, in most cases, the percentage loss of RAC is only a single digit number starting from $\theta = 10$, which suggests the robustness of RAC even for problems with relatively small θ. Finally, we remark that while the expected percentage loss should be positive, it is possible to have negative percentage loss in simulations. This is due to sampling error in estimating the average costs using Monte Carlo simulations, and most of the negative percentage losses are very close to 0.

5.2. Comparing RAC with an index policy

In this section, we compare the performance of the RAC with a strict priority rule (or index policy). In the two examples below, we consider two states and two actions (i.e. treatment/active or no treatment/inactive) and assume that there is a budget of 1 (i.e. a maximum of one active bandit) in each time period. Our priority rule works as follows: we use our heuristic to calculate the probability of activating an arm in each state and then activate the arm with the highest probability directly. Note that, since we are activating an arm in each period, we will use the entire budget in each period. We then compare the costs for this priority rule with our heuristic. The results are summarized in Tables 3 and 4. Note that RAC outperforms the specific priority rule. Moreover, it does so without always using the full budget in every period. (We report in Table 5 the percentage budget used in RAC. That is, if the budget in each period is b and the number of periods under consideration is T, then the total budget allowed is $b \times T$. If the total budget actually used during the T time periods is B, then the percentage budget is calculated as $B/(b \times T)$.) This is perhaps not too surprising.

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

TABLE 3: Percentage loss for two states and one action example using RAC.

θ	$T = 10$	$T = 30$	$T = 50$	$T = 100$
1	56.66	56.25	58.47	55.82
5	18.11	19.25	18.20	19.25
10	10.26	10.58	9.94	11.06
20	5.96	5.64	5.49	5.17
40	3.68	3.09	3.68	3.16
60	2.55	2.56	2.70	2.93
80	2.02	2.47	2.29	2.29
100	2.12	2.06	1.83	1.96

TABLE 4: Percentage loss for two states and one action using priority rule.

θ	$T = 10$	$T = 30$	$T = 50$	$T = 100$
1	85.76	83.02	87.60	84.47
5	66.33	65.09	65.38	64.68
10	63.84	63.96	64.43	64.20
20	65.00	64.35	64.43	64.31
40	64.12	64.03	64.11	63.85
60	64.19	64.72	64.26	64.79
80	63.95	63.92	64.60	64.16
100	64.46	64.03	64.15	64.30

TABLE 5: Percentage budget used for two states and one action.

θ	$T = 10$	$T = 30$	$T = 50$	$T = 100$
1	9.03	2.98	1.82	0.90
5	10.66	3.55	2.14	1.07
10	11.07	3.69	2.22	1.11
20	11.41	3.79	2.27	1.14
40	11.61	3.86	2.32	1.16
60	11.71	3.89	2.34	1.17
80	11.76	3.92	2.35	1.17
100	11.80	3.92	2.36	1.18

Since a treatment could be costly, it may not be optimal to always use up all the budgets; it is also important to take into account the trade-off between the cost of a treatment and its benefit.

6. Dynamic population model

In this section, we allow new bandits to arrive in each time period, and we also consider random departures (e.g. random service completions or random abandonments). The same

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

notation as before will be used, with an addition that we also use Λ_{jt} to denote the (random) arrival of bandits in state j at time t, where Λ_{jt} is a Poisson random variable with mean $\lambda_{jt} > 0$. The stochastic control formulation of our problem can now be written as

$$V^S = \min_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a=0}^{A} \sum_{j=1}^{J} c_j^a \cdot X_j^{\pi,a}(t) + \phi \cdot \left(\sum_{a=0}^{A} \sum_{j \in \mathbb{U}} X_j^{\pi,a}(T+1) - m \right)^+ \right]$$

such that

$$\sum_{a=0}^{A} X_j^{\pi,a}(t) = \sum_{a=0}^{A} \sum_{i=1}^{J} Y_{ij}^{\pi,a}(t-1) + \Lambda_{jt} \quad \text{for all } j \geq 1,\ t \geq 2,$$

$$\sum_{a=0}^{A} X_j^{\pi,a}(1) = n_j + \Lambda_{j1} \quad \text{for all } j \geq 1,$$

$$\sum_{a=1}^{A} \sum_{j=1}^{J} X_j^{\pi,a}(t) \leq b_t \quad \text{for all } t \geq 1.$$

The corresponding deterministic formulation is given by

$$V^D(\epsilon) = \min_{x,z} \sum_{t=1}^{T} \sum_{a=0}^{A} \sum_{j=1}^{J} c_j^a \cdot x_j^a(t, \boldsymbol{\epsilon}) + \phi \cdot z(\boldsymbol{\epsilon}) \tag{6a}$$

such that

$$\sum_{a=0}^{A} x_j^a(t, \boldsymbol{\epsilon}) = \sum_{a=0}^{A} \sum_{i=1}^{J} x_i^a(t-1, \boldsymbol{\epsilon}) \cdot p_{ij}^a + \lambda_{jt} \quad \text{for all } j \geq 1,\ t \geq 2, \tag{6b}$$

$$\sum_{a=0}^{A} x_j^a(1, \boldsymbol{\epsilon}) = n_j + \lambda_{j1} \quad \text{for all } j \geq 1, \tag{6c}$$

$$\sum_{a=1}^{A} \sum_{j=1}^{J} x_j^a(t, \boldsymbol{\epsilon}) \leq b_t - \epsilon_t \quad \text{for all } t \geq 1, \tag{6d}$$

$$z(\boldsymbol{\epsilon}) \geq \sum_{j \in \mathbb{U}} \sum_{a=0}^{A} \sum_{i=1}^{J} x_i^a(T, \boldsymbol{\epsilon}) \cdot p_{ij}^a - m, \tag{6e}$$

$$z(\boldsymbol{\epsilon}),\ x_j^a(t, \boldsymbol{\epsilon}) \geq 0 \quad \text{for all } a \geq 0,\ j \geq 1,\ t \geq 1. \tag{6f}$$

Both Lemmas 1 and 2 still hold in the new setting (we omit the details), and the definition of RAC is still the same as that presented in Section 4, with one difference: since the number of bandits in the system potentially changes at every time period, the permutation $\boldsymbol{\varphi}$, which provides the order in which we assess the state of existing bandits in step 3 in the original definition of RAC, must now be selected at the beginning of *every* time period. That said, as noted in Section 4, the exact choice of $\boldsymbol{\varphi}$ does not affect our asymptotic analysis and thus we can choose an arbitrary permutation $\boldsymbol{\varphi}$ that randomizes the order of existing bandits at each time period. The following result is the analogue of Lemma 3 and the proof can be found in Appendix B.

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

Lemma 4. *For any* $\boldsymbol{\epsilon} \geq \mathbf{0}$*, we have*

$$C^{\mathrm{RAC}} - V^D(\epsilon) \leq \left[\phi + \max_{a,j} c_j^a\right] \cdot \left[\sum_{t=1}^{T} \epsilon_t\right] \tag{7}$$

with probability at least

$$1 - 3 \cdot A \cdot J^2 \cdot \sum_{t=1}^{T} \exp\left\{-\frac{\epsilon_t^2}{48 \cdot A^2 \cdot J^4 \cdot [\sum_{\ell=1}^{J} n_\ell]}\right\} - 3 \cdot A \cdot J^2 \cdot \sum_{t=1}^{T} \exp\left\{-\frac{\epsilon_t}{12 \cdot A \cdot J^2}\right\}$$
$$- 3 \cdot A \cdot J^2 \cdot \sum_{t=1}^{T} \exp\left\{-\frac{\epsilon_t^2}{64 \cdot A^2 \cdot J^4 \cdot [\sum_{s=1}^{t} \sum_{i=1}^{J} \lambda_{is}]}\right\}. \tag{8}$$

The last summation in (8) is due to the arrivals of new bandits at each time period. If $\lambda_{tj} = 0$ for all t and j then the last summation equals 0 and the bound in (8) is almost identical to the bound in (4), with the exception that the numbers 12 and 6 that appear in the denominator in the original bound have now been replaced with 48 and 12 in the denominator above. Let $\lambda_{\mathrm{tot}} := \max_t \sum_{j=1}^{J} \lambda_{jt}$. We consider the same asymptotic setting as in Section 4, where we also scale the arrival rate λ_{jt} by θ. Let n_{tot} and $S(t)$ be as defined in Section 3. The following theorem is the analogue of Theorem 1.

Theorem 2. *Let* d *be a positive number. Suppose that* $S(t) \neq \varnothing$ *for all* t*, and let*

$$\epsilon_t = 12 \cdot A \cdot J^2 \cdot \sqrt{t \cdot d \cdot \max\{n_{\mathrm{tot}},\ \lambda_{\mathrm{tot}}\} \cdot \theta \cdot \ln \theta}$$

for all t*. For all* $\theta > \theta^*$ *where*

$$\frac{\theta^*}{\ln \theta^*} = 144 \cdot A^2 \cdot J^4 \cdot d \cdot \max\{n_{\mathrm{tot}}, \lambda_{\mathrm{tot}}\} \cdot \left[\max_t \frac{t}{b_t}\right], \tag{9}$$

there exists a constant $M > 0$ *independent of* T *and* M *such that*

$$\frac{V_\theta^{\mathrm{RAC}} - V_\theta^S}{V_\theta^S} \leq M \cdot \left[\frac{T^{3/2}}{\theta^{d/2}} + \sqrt{\frac{d \cdot T \cdot \ln \theta}{\theta}}\right]. \tag{10}$$

Proof. The proof is similar to that of Theorem 1. First, condition (9) is equivalent to $\theta \cdot b_t > \epsilon_t$ for all t. It is important to guarantee that the right-hand side of the budget constraints in linear program (6) is positive and, therefore, the linear program has a feasible solution.

Let E denote the event where (7) is satisfied. By Lemma 4 and our choice of $\boldsymbol{\epsilon}$, $\mathbb{P}(E)$ is at least $1 - \Theta(T/\theta^d)$. Now, we consider the event E^c. Unlike in the proof of Theorem 1 where we can simply bound $C_\theta^{\mathrm{RAC}} - V_\theta^D(\boldsymbol{\epsilon})$ with a number that is of order $\Theta(T \cdot \theta \cdot n_{\mathrm{tot}})$, we now need to take into account new bandits that arrive at each time period. At time t, we have at most $\theta \cdot n_{\mathrm{tot}} + \sum_{s=1}^{t} \sum_{j=1}^{J} \Lambda_{js}^\theta$ bandits present in the system, where Λ_{js}^θ is a Poisson random variable with rate $\theta \cdot \lambda_{js}$. As a result, we can bound:

$$\mathbb{E}[(C_\theta^{\mathrm{RAC}} - V_\theta^D(\boldsymbol{\epsilon})) \cdot \mathbf{1}\{E^c\}]$$
$$\leq \left[\phi + \max_{a,j} c_j^a\right] \cdot \mathbb{E}\left[\left(T \cdot \theta \cdot n_{\mathrm{tot}} + \sum_{t=1}^{T} t \cdot \left(\sum_{j=1}^{J} \Lambda_{j,t}^\theta\right)\right) \cdot \mathbf{1}\{E^c\}\right]. \tag{11}$$

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

Note that $\sum_{t=1}^{T} t \cdot (\sum_{j=1}^{J} \Lambda_{jt}^{\theta})$ is stochastically dominated by the random variable $X \sim \text{Pois}(T^2 \cdot \theta \cdot \lambda_{\text{tot}})$. By the Cauchy–Schwarz inequality, we can further bound (11) as

$$\begin{aligned}\mathbb{E}[(C_{\theta}^{\text{RAC}} - V_{\theta}^{D}(\boldsymbol{\epsilon})) \cdot \mathbf{1}\{E^c\}] &\leq \left[\phi + \max_{a,j} c_j^a\right] \cdot \mathbb{E}[(T \cdot \theta \cdot n_{\text{tot}} + X)^2]^{1/2} \cdot \mathbb{P}(E^c)^{1/2} \\ &= O\left(\frac{T^{5/2}}{\theta^{d/2-1}}\right) \end{aligned} \tag{12}$$

where the last equality follows since

$$\mathbb{P}(E^c)^{1/2} = O\left(\frac{T^{1/2}}{\theta^{d/2}}\right) \quad \text{and} \quad \mathbb{E}[X^2]^{1/2} = T^2 \cdot \theta \cdot \lambda_{\text{tot}},$$

which implies that $\mathbb{E}[(T \cdot \theta \cdot n_{\text{tot}} + X)^2]^{1/2} = O(T^2 \cdot \theta)$.

Putting (12) together with

$$\mathbb{E}[(C_{\theta}^{\text{RAC}} - V_{\theta}^{D}(\boldsymbol{\epsilon})) \cdot \mathbf{1}\{E\}] = O(T^{3/2} \cdot \sqrt{d \cdot \theta \cdot \log \theta})$$

(by Lemma 4 and our choice of $\boldsymbol{\epsilon}$) and the fact that $V_{\theta}^{D}(\mathbf{0}) = \Omega(T \cdot \theta)$ (because $S(t) \neq \varnothing$ for all t) yields the desired results. This completes the proof of Theorem 2. □

Remark 3. (*Theorem 2.*) Unlike the choice of ϵ_t in Theorem 1, our choice of ϵ_t in Theorem 2 scales with $\sqrt{t}$. This is crucial for dealing with additional randomness (due to random arrivals of bandits) in each time period, which requires a more conservative buffer. This scaling results in a weaker bound for RAC in Theorem 2 compared to that in Theorem 1. However, we want to emphasize that the bound in (10) holds for very general settings and does not exploit any special structure that might arise in specific instances of this model (e.g. control of queueing systems). For example, if bandits are leaving the system at a geometric rate then it is possible to derive a much tighter bound for the relative loss of RAC. We discuss this in more detail below.

The following proposition is similar to Proposition 1 in Section 4.

Proposition 2. *Let d be a positive number. Suppose that we multiply the cost at time t with δ^{t-1} for some discount factor $\delta \in (0, 1)$. Suppose that $S(t) \neq 0$ for all t. Let*

$$\epsilon_t = 12 \cdot A \cdot J^2 \cdot \sqrt{t \cdot \log(t + \mathrm{e} - 1) \cdot d \cdot \max\{n_{\text{tot}}, \lambda_{\text{tot}}\} \cdot \theta \cdot \ln \theta}$$

for all t (note that ϵ_t now depends on t). For all $\theta > \theta^*$ where

$$\frac{\theta^*}{\ln \theta^*} = 144 \cdot A^2 \cdot J^4 \cdot d \cdot \max\{n_{\text{tot}}, \lambda_{\text{tot}}\} \cdot \log(T + \mathrm{e} - 1) \cdot \left[\max_t \frac{t}{b_t}\right],$$

there exists a constant $M > 0$ independent of T and θ such that

$$\frac{V_{\theta}^{\text{RAC}} - V_{\theta}^{S}}{V_{\theta}^{S}} \leq \frac{V_{\theta}^{\text{RAC}} - V_{\theta}^{D}(\mathbf{0})}{V_{\theta}^{D}(\mathbf{0})} \leq M \cdot \left[\frac{1}{\theta^{d/2}} + \sqrt{\frac{d \cdot \ln \theta}{\theta}}\right].$$

The proof of Proposition 2 is very similar to the proof of Theorem 2 and, therefore, is omitted. Unlike in Remark 2, where the bound in Proposition 1 holds for $T = o(\exp\{\theta/(d \cdot \ln \theta)\})$, the bound in Proposition 2 holds so long as

$$T \cdot \log T = o\left(\frac{\theta}{\log \theta}\right).$$

In what follows, we discuss a special case of the dynamic population model in which bandits can arrive and leave randomly over time.

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

6.1. Bandit departure and abandonment

Suppose now that each bandit waits to receive a proper treatment ($a \geq 1$) and immediately leaves the system once it receives a proper treatment (i.e. a service completion). Moreover, suppose that, if a bandit has not received a proper treatment up until the end of the current time period, it will stay in the system for the next period with probability $\alpha \in (0, 1)$ and will leave (i.e. abandon) the system with probability $1 - \alpha$. These features can be included in the current modeling framework by introducing an extended state variable that includes the type of treatment a bandit received in the previous time period and a tracking variable that indicates whether the bandit is still in the system at the current time period. For example, we can let $j_t = (j'_t, a_{t-1}, \ell_t)$ be our extended state variable, where j'_t is the current state/condition of the bandit, a_{t-1} is the treatment received in the previous time period, and $\ell_t \in \{0, 1\}$ an indicator of the bandit's presence in the system with 0 (1) denoting it left (remains in) the system at the beginning of period t. The one-step transition probabilities for this model now satisfy the following conditions:

$$p^{a''}_{(j',a',0),(j'',a'',1)} = 0 \quad \text{for all } j', a', j'', a'', \tag{13}$$

$$p^{a''}_{(j',0,1),(j'',a'',1)} = 0 \quad \text{for all } j', j'', a'' \geq 1, \tag{14}$$

$$\sum_{j''} p^{0}_{(j',0,1),\,(j'',0,1)} = \alpha \quad \text{for all } j'. \tag{15}$$

Condition (13) simply says that a bandit that leaves the system never returns; condition (14) says that a bandit immediately leaves the system after receiving a proper treatment; and condition (15) says that the probability a bandit stays in the system for the next time period given that it has not received a proper treatment is α.

For simplicity, suppose that, at the end of the horizon, we only get penalized for the number of bandits who leave the system without being properly treated (i.e. abandonments). This requires keeping track of the number of abandonments at each time period. To do this, we set $c^{a}_{(j',0,1)} = M$ for all j' and $a \geq 1$, where M is a sufficiently large number, and $c^{0}_{(j',0,1)} = 0$ for all j'. Since applying a proper action to a nonexisting bandit is very costly, this setup implies that, once a bandit leaves the system without being properly treated, the last two components of its state remains (0, 1) throughout the horizon. As a result, the total number of abandonments throughout the horizon simply equals $\sum_{a=0}^{A} \sum_{j'} X^{a}_{(j',0,1)}(T+1)$ and its corresponding total penalty costs is

$$\phi \cdot \left(\sum_{a=0}^{A} \sum_{j'} X^{a}_{(j',0,1)}(T+1) - m \right)^{+},$$

where $m \geq 0$ can be interpreted as the maximum number of acceptable abandonments.

The setting with service completions and abandonments discussed above is a special case of the more general setting considered in Theorem 2. As a result, the bound in (10) still holds in this setting. In fact, as previously noted, this bound can be significantly tightened by exploiting the special structure of the problem. That is, we have the following result.

Proposition 3. *Let d be a positive number. Consider the setting with bandit departure and abandonment described above. Suppose that $S(t) \neq \varnothing$ for all t. If we let*

$$\epsilon_t = \frac{12 \cdot A \cdot J^2}{\min_t b_t} \cdot \sqrt{\left(\sum_{s=0}^{t-1} \alpha^s \right) \cdot d \cdot \max\{n_{\text{tot}}, \lambda_{\text{tot}}\} \cdot \theta \cdot \ln \theta}$$

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

for all t, then

$$\frac{V_\theta^{\mathrm{RAC}} - V_\theta^S}{V_\theta^S} = O\left(\frac{1}{1-\alpha} \cdot \frac{T^{1/2}}{\theta^{d/2}} + \sqrt{\frac{d \cdot \ln \theta}{(1-\alpha)\cdot\theta}}\right).$$

Proof. The proof of Proposition 3 is very similar to the proof of Theorem 2. Here, we will only provide its outline. Note that we can strengthen the bound in Lemma 4 by replacing the term $\sum_{s=1}^{t}\sum_{j=1}^{J}\lambda_{jt}$ in the third summation in (8) with $\sum_{s=1}^{t}\alpha^{t-s}\cdot[\sum_{j=1}^{J}\lambda_{jt}]$ (see Remark 5 in Appendix B.1). Let E denote the event where (7) is satisfied. By our choice of $\boldsymbol{\epsilon}$, we have

$$\mathbb{E}[(C_\theta^{\mathrm{RAC}} - V_\theta^D(\boldsymbol{\epsilon}))\cdot \mathbf{1}\{E\}] = O\left(T \cdot \sqrt{\frac{d\cdot\theta\cdot\ln\theta}{1-\alpha}}\right).$$

As for $\mathbb{E}[(C_\theta^{\mathrm{RAC}} - V_\theta^D(\boldsymbol{\epsilon}))\cdot \mathbf{1}\{E^c\}]$, since each bandit stays in the system with probability at most α, we can upper bound (in the stochastic dominance sense) the number of initial bandits that are still present in the system at time t by Binomial$(\theta\cdot n_{\mathrm{tot}}, \alpha^{t-1})$. Similarly, we can also upper bound the number of bandits that arrived in the system at time s in state j and are still present in the system at time t by Poisson $(\theta\cdot\lambda_{js}\cdot\alpha^{t-s})$. Since $\mathbb{P}(E^c) = O(T/\theta^d)$, by similar arguments as in the proof of Theorem 2 (using the Cauchy–Schwarz inequality), we have

$$\mathbb{E}[(C_\theta^{\mathrm{RAC}} - V_\theta^D(\boldsymbol{\epsilon}))\cdot \mathbf{1}\{E^c\}] = O\left(\frac{T\cdot\theta}{1-\alpha}\right)\cdot O\left(\frac{T^{1/2}}{\theta^{d/2}}\right).$$

The proof is complete by noting that $V_\theta^D(\mathbf{0}) = \Omega(T\cdot\theta)$ (because $S(t)\neq\varnothing$ for all t). □

Note that, if $T\sim\theta^n$, where n is a natural number, then we can choose $d > n$ and RAC is asymptotically optimal. This illustrates that RAC performs well for a large decision horizon T with Poisson arrivals and geometrically distributed waiting times.

7. Closing remarks

In this paper, we consider discrete-time, finite-horizon, RMABPs for fixed and dynamic populations of bandits. To our knowledge, we are the first to simultaneously allow for multiple actions and a nonstationary number of active bandits. We proposed a heuristic control, RAC, that is based on an LP relaxation of the original stochastic control formulation and showed that it is asymptotically optimal. Similar to Verloop (2016), our heuristic does not depend on any indexability properties. In contrast to Verloop (2016), it does not require assumptions on the underlying structure of the Markov process generated by each policy or assumptions on the dynamics on the associated deterministic approximation.

As mentioned, we can include nonstationary transition probability matrices and nonstationary costs. We can also generalize to the setting when there are several constraints for several subsets of bandits or one constraint on the total number of activated bandits over the entire finite-time horizon. Since we consider an asymptotic regime where the initial budgets and the number of initial bandits are proportionally scaled, it can be shown that these constraints (one or multiple constraints) will not be violated with a 'high' probability. Also, having state-dependent terminal costs will not change the asymptotic order. Finally, the proposed approach would also work for the multi-action setting where each action contributes a different weight to the budget since once the linear program is solved, we can show that the budget constraints will not be violated with a very high probability. This holds regardless of whether each action contributes the same or different weights to the budget.

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

There are several avenues for further research. The approach in Hu and Frazier (2017) does not directly translate to our setting, but our approach may be applicable to their model since we make no idexability assumptions. One topic for further exploration is to test our heuristic using their data in the specific examples considered in their study. Developing and analyzing alternative policies is of clear interest. Comparing alternative policies with RAC in a numerical study can be considered. One might consider the same model but allow for states to be observed only when a treatment is applied. This would extend the model considered by Deo *et al.* (2013) that is motivated by capacity management in healthcare. This model requires keeping track of additional information (i.e. the time between interventions) and, hence, will require a larger state space. In this case, RAC may no longer be asymptotically optimal and, more broadly, the model would present significant technical challenges for the analysis and computation of good control policies. Another consideration is the possibility of allowing more general (i.e. nonlinear) cost structure and constraints. For instance, to capture the fact that, for many settings (e.g. EMS response, humanitarian logistics), the number of available servers is random, the budget b_t may be assumed to be a random variable for each t. Notwithstanding, RMABPs are a very useful modeling framework that can be applied to a broad range of problems; they provide significant technical challenges for optimal control, which are a bright research direction.

Appendix A. Proofs of the results in Section 4

A.1. Proof of Lemma 2

Note that linear program (2) can be equivalently written as

$$V^D(\boldsymbol{\epsilon}) = \min_{x,z} \sum_{t=1}^{T} \sum_{a=0}^{A} \sum_{j=1}^{J} c_j^a \cdot x_j^a(t, \boldsymbol{\epsilon}) + \phi \cdot z(\boldsymbol{\epsilon})$$

such that

$$\sum_{a=0}^{A} \sum_{i=1}^{J} x_i^a(t, \boldsymbol{\epsilon}) \cdot p_{ij}^a = n_j(t+1, \boldsymbol{\epsilon}) \quad \text{for all } j \geq 1,\ t \geq 1,$$

$$\sum_{a=0}^{A} x_j^a(t, \boldsymbol{\epsilon}) = n_j(t, \boldsymbol{\epsilon}) \quad \text{for all } j \geq 1,\ t \geq 2,$$

$$\sum_{a=0}^{A} x_j^a(1, \boldsymbol{\epsilon}) = n_j \quad \text{for all } j \geq 1,$$

$$\sum_{a=1}^{A} \sum_{j=1}^{J} x_j^a(t, \boldsymbol{\epsilon}) \leq b_t - \epsilon_t \quad \text{for all } t \geq 1,$$

$$z(\boldsymbol{\epsilon}) \geq \sum_{j \in \mathbb{U}} \sum_{a=0}^{A} \sum_{i=1}^{J} x_i^a(T, \boldsymbol{\epsilon}) \cdot p_{ij}^a - m,$$

$$z(\boldsymbol{\epsilon}),\ x_j^a(t, \boldsymbol{\epsilon}),\ n_j(t, \boldsymbol{\epsilon}) \geq 0 \quad \text{for all } a \geq 0,\ j \geq 1,\ t \geq 1.$$

Let $\boldsymbol{\mu}^*(\boldsymbol{\epsilon}) = (\mu_1^*(\boldsymbol{\epsilon}), \mu_2^*(\boldsymbol{\epsilon}), \ldots, \mu_T^*(\boldsymbol{\epsilon}))$ denote the optimal dual solution corresponding to constraints $\sum_{a=1}^{A} \sum_{j=1}^{J} x_j^a(t, \boldsymbol{\epsilon}) \leq b_t - \epsilon_t$ for all $t \geq 1$. By the standard duality argument

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

(see, for example, Schrijver (2000, Section 10.4, Equation (24))), we can bound

$$V^D(\boldsymbol{\epsilon}) - V^D(\mathbf{0}) \leq \left| \sum_{t=1}^{T} \mu_t^*(\mathbf{0}) \cdot \epsilon_t \right|.$$

Since, by definition, $\mu_t^*(\mathbf{0})$ is trivially independent of $\boldsymbol{\epsilon}$, we simply need to show that $\mu_t^*(\boldsymbol{\epsilon})$ is independent of T (i.e. there exists a constant $M > 0$ independent of T such that $\mu_t^*(\mathbf{0}) \leq M$ for all t). We do this by first writing the optimal value of linear program (2) as a sum of the optimal values of T separable linear programs and then argue that the optimal dual solution for linear program (2) is also optimal for these linear programs, and vice versa. Specifically,

$$V^D(\boldsymbol{\epsilon}) = \sum_{t=1}^{T} V_t^D(\boldsymbol{\epsilon}),$$

where $V_t^D(\boldsymbol{\epsilon})$ is defined as

$$V_t^D(\boldsymbol{\epsilon}) = \min_{x,z} \sum_{a=0}^{A} \sum_{j=1}^{J} c_j^a \cdot x_j^a(t, \boldsymbol{\epsilon}) \tag{16a}$$

such that

$$\sum_{a=0}^{A} \sum_{i=1}^{J} x_i^a(t, \boldsymbol{\epsilon}) \cdot p_{ij}^a = n_j^*(t+1, \boldsymbol{\epsilon}) \quad \text{for all } j \geq 1, \tag{16b}$$

$$\sum_{a=0}^{A} x_j^a(t, \boldsymbol{\epsilon}) = n_j^*(t, \boldsymbol{\epsilon}) \quad \text{for all } j \geq 1, \tag{16c}$$

$$\sum_{a=1}^{A} \sum_{j=1}^{J} x_j^a(t, \boldsymbol{\epsilon}) \leq b_t - \epsilon_t, \tag{16d}$$

$$x_j^a(t, \boldsymbol{\epsilon}) \geq 0 \quad \text{for all } a \geq 0,\ j \geq 1, \tag{16e}$$

for $t \leq T - 1$ and

$$V_T^D(\boldsymbol{\epsilon}) = \min_{x,z} \sum_{a=0}^{A} \sum_{j=1}^{J} c_j^a \cdot x_j^a(T, \boldsymbol{\epsilon}) + \phi \cdot z(\boldsymbol{\epsilon}) \tag{17a}$$

such that

$$\sum_{a=0}^{A} \sum_{i=1}^{J} x_i^a(T, \boldsymbol{\epsilon}) \cdot p_{ij}^a = n_j^*(T+1, \boldsymbol{\epsilon}) \quad \text{for all } j \geq 1, \tag{17b}$$

$$\sum_{a=0}^{A} x_j^a(T, \boldsymbol{\epsilon}) = n_j^*(T, \boldsymbol{\epsilon}) \quad \text{for all } j \geq 1, \tag{17c}$$

$$\sum_{a=1}^{A} \sum_{j=1}^{J} x_j^a(T, \boldsymbol{\epsilon}) \leq b_T - \epsilon_T, \tag{17d}$$

$$z(\boldsymbol{\epsilon}) \geq \sum_{j \in \mathbb{U}} \sum_{a=0}^{A} \sum_{i=1}^{J} x_i^a(T, \boldsymbol{\epsilon}) \cdot p_{ij}^a - m, \tag{17e}$$

$$z(\boldsymbol{\epsilon}),\ x_j^a(T, \boldsymbol{\epsilon}) \geq 0 \quad \text{for all } a \geq 0,\ j \geq 1. \tag{17f}$$

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

Observe that the union of all the constraints in linear programs (16) (for all $t \le T-1$) and (17) constitutes all the constraints in linear program (2), with a minor difference that we replace the variable $n_j(t, \boldsymbol{\epsilon})$ with $n_j^*(t, \boldsymbol{\epsilon})$. Since $n_j(t, \boldsymbol{\epsilon}) = n_j^*(t, \boldsymbol{\epsilon})$ is optimal for linear program (2), it is not difficult to see that the optimal solution for linear program (2) is also optimal for linear programs (16) (for all $t \le T-1$) and (17), and vice versa. Moreover, the optimal dual solution corresponding to each constraint in linear programs (16) (for $t \le T-1$) and (17) is also optimal for the corresponding constraint in linear program (2) (this can be easily checked by considering the Karush–Kuhn–Tucker (KKT) conditions at the optimal solution; that is, by checking that complementary slackness holds). This observation has an important consequence. We can indirectly compute $\mu_t^*(\mathbf{0})$ by calculating the optimal dual solution corresponding to constraint $\sum_{a=1}^{A}\sum_{j=1}^{J} x_j^a(t, \mathbf{0}) \le b_t$ in either linear program (16) (for $t \le T-1$) or (17), with $\boldsymbol{\epsilon} = \mathbf{0}$. The proof is complete by noting that the optimal dual solution for either linear program (16) (for $t \le T-1$) or (17) cannot possibly scale up with T (see Remark 4 below); hence, they can be uniformly bounded by a constant that is independent of T.

Remark 4. For any generic linear program $\max\{c'x\colon Ax \le b\}$, the optimal dual solution y^* must satisfy $A'y^* = c'$, so $||y^*||_1 \le n \cdot \Delta \cdot ||c||_1$, where n is the dimension of x and Δ is an upper bound of the absolute values of the components of B^{-1} for all invertible submatrices B of A (see, for example, Schrijver (2000, Section 10.4)). Applied to our setting, the value of the corresponding terms $n \cdot \Delta \cdot ||c||_1$ in linear programs (16) and (17) obviously do not depend on T. So, $\mu_t^*(\mathbf{0})$ is independent of T. Under the discounted cost setting discussed in Remark 3, the term $||c||_1$ will now be multiplied by δ^{t-1}. So, there exists a constant $M > 0$ such that $\mu_t^*(\mathbf{0}) \le M \cdot \delta^{t-1}$ for all t.

A.2. Proof of Lemma 3

To prove Lemma 3, it is more convenient to work with a modified version of RAC defined below instead of the original RAC.

Algorithm 2. (*Modified randomized assignment control (MRAC).*)

1. Pick $\boldsymbol{\epsilon}$ and solve linear program (2).
2. Compute $q_j^{*,a}(t, \boldsymbol{\epsilon})$ for all a, j, and t.
3. Pick a permutation $\boldsymbol{\varphi}$.
4. For time $t = 1$ to T, do:
   (a) set $k = 1$;
   (b) randomly generate $Z_{j(t,\varphi(k)),k}(t, \boldsymbol{\epsilon})$;
   (c) apply action $Z_{j(t,\varphi(k)),k}(t, \boldsymbol{\epsilon})$ to bandit $\varphi(k)$;
   (d) update $k = k + 1$;
   (e) while $k \le n_{\text{tot}}$, go back to step (b).

Note that MRAC proceeds in the same manner as RAC, with an exception that it ignores the budget constraint in the original step 4c (i.e. MRAC continues activating bandits regardless of the given budget b_t has been exhausted). Let $\tilde{X}_j^a(t, \boldsymbol{\epsilon})$ denote the number of bandits in state j being applied action a at period t under MRAC, and let $\tilde{N}_j(t, \boldsymbol{\epsilon})$ denote the number of bandits in state j at the beginning of period t under MRAC. Define four events $\tilde{\mathcal{A}}_1(\boldsymbol{\epsilon})$, $\tilde{\mathcal{A}}_2(\boldsymbol{\epsilon})$, $\tilde{\mathcal{A}}_3(\boldsymbol{\epsilon})$,

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

and $\tilde{\mathcal{A}}(\boldsymbol{\epsilon})$ as follows:

$$\tilde{\mathcal{A}}_1(\boldsymbol{\epsilon}) = \left\{ \tilde{X}_j^a(t, \boldsymbol{\epsilon}) - x_j^{*a}(t, \boldsymbol{\epsilon}) \le \frac{\epsilon_t}{2AJ} \quad \text{for all } a \ge 1,\ j,\ t \right\},$$

$$\tilde{\mathcal{A}}_2(\boldsymbol{\epsilon}) = \left\{ \tilde{X}_j^0(t, \boldsymbol{\epsilon}) - x_j^{*0}(t, \boldsymbol{\epsilon}) \le \frac{\epsilon_t}{2J} \quad \text{for all } j,\ t \right\},$$

$$\tilde{\mathcal{A}}_3(\boldsymbol{\epsilon}) = \left\{ \tilde{N}_j(T+1, \epsilon) - n_j^*(T+1, \boldsymbol{\epsilon}) \le \frac{\epsilon_T}{|\mathbb{U}|} \quad \text{for all } j \in \mathbb{U} \right\},$$

$$\tilde{\mathcal{A}}(\boldsymbol{\epsilon}) = \tilde{\mathcal{A}}_1(\boldsymbol{\epsilon}) \cap \tilde{\mathcal{A}}_2(\boldsymbol{\epsilon}) \cap \tilde{\mathcal{A}}_3(\boldsymbol{\epsilon}).$$

Note that the conditions in $\tilde{\mathcal{A}}_1(\boldsymbol{\epsilon})$ collectively imply that $\sum_{a=1}^{A} \sum_{j=1}^{J} \tilde{X}_j^a(t, \boldsymbol{\epsilon}) \le b_t$ for all t; thus, under the same sample path realizations, MRAC is equivalent to RAC on $\tilde{\mathcal{A}}(\boldsymbol{\epsilon})$. Moreover, on $\tilde{\mathcal{A}}(\boldsymbol{\epsilon})$, we also have

$$C^{\mathrm{MRAC}} - V^D(\boldsymbol{\epsilon}) \le \left[\max_{a,j} c_j^a \right] \cdot \left[\sum_{t=1}^{T} \epsilon_t \right] + \phi \cdot \epsilon_T \le \left[\phi + \max_{a,j} c_j^a \right] \cdot \left[\sum_{t=1}^{T} \epsilon_t \right],$$

where the first inequality holds because $(a-x)^+ - (b-x)^+ \le (a-b)^+$ for all a, b, x. Since MRAC is equivalent to RAC on $\tilde{\mathcal{A}}(\boldsymbol{\epsilon})$, to prove Lemma 3, we simply need to show that the lower bound stated in Lemma 3 holds for $\mathbb{P}(\tilde{\mathcal{A}}(\boldsymbol{\epsilon}))$ under MRAC.

Note that we can bound

$$\mathbb{P}(\tilde{\mathcal{A}}(\boldsymbol{\epsilon})) \ge 1 - \mathbb{P}(\tilde{\mathcal{A}}_1(\boldsymbol{\epsilon})^c) - \mathbb{P}(\tilde{\mathcal{A}}_2(\boldsymbol{\epsilon})^c) - \mathbb{P}(\tilde{\mathcal{A}}_3(\boldsymbol{\epsilon})^c). \tag{18}$$

We will now compute an upper bound for each $\mathbb{P}(\tilde{\mathcal{A}}_i(\boldsymbol{\epsilon})^c)$, $i = 1, 2, 3$. We start with $\mathbb{P}(\tilde{\mathcal{A}}_1(\boldsymbol{\epsilon})^c)$. By the subadditive property of probability,

$$\mathbb{P}(\tilde{\mathcal{A}}_1(\boldsymbol{\epsilon})^c) \le \sum_{t=1}^{T} \sum_{a=1}^{A} \sum_{j=1}^{J} \mathbb{P}\left(\tilde{X}_j^a(t, \boldsymbol{\epsilon}) - x_j^{*a}(t, \epsilon) > \frac{\epsilon_t}{2AJ} \right). \tag{19}$$

We make an important observation: For all a and j, the random variable $\tilde{X}_j^a(t, \boldsymbol{\epsilon})$ can be written as a sum of J independent binomial random variables. Specifically,

$$\tilde{X}_j^a(t, \boldsymbol{\epsilon}) \sim \sum_{i=1}^{J} \mathrm{Bin}(n_i, v_{iaj}(t, \boldsymbol{\epsilon})), \tag{20}$$

where $v_{iaj}(t, \boldsymbol{\epsilon})$ is the probability that, under MRAC, a bandit that starts with state i at the beginning of period 1 ends up with state j at the beginning of period t and then being applied action a in period t. (It is possible to give an explicit expression of $v_{iaj}(t, \boldsymbol{\epsilon})$ in terms of the transition probabilities $\{p_{ij}^a\}$ and the activation probabilities $\{q_j^{*a}(t, \boldsymbol{\epsilon})\}$; but, this is not necessary for our purpose.) Note that $x_j^{*a}(t, \boldsymbol{\epsilon}) = \sum_{i=1}^{J} n_i \cdot v_{iaj}(t, \boldsymbol{\epsilon})$. Let $\tilde{S}_{aj}(t, \boldsymbol{\epsilon}) = \{i\colon v_{iaj}(t, \boldsymbol{\epsilon}) > 0\}$. Using our observation in (20), we can further bound $\mathbb{P}(\tilde{\mathcal{A}}_1(\boldsymbol{\epsilon})^c)$ as

$$\mathbb{P}(\tilde{\mathcal{A}}_1(\boldsymbol{\epsilon})^c) \le \sum_{t=1}^{T} \sum_{\substack{(a,j)\colon \tilde{S}_{aj}(t,\boldsymbol{\epsilon}) \neq \varnothing \\ i \in \tilde{S}_{aj}(t,\boldsymbol{\epsilon})}} \mathbb{P}\left(\mathrm{Bin}(n_i, v_{iaj}(t, \boldsymbol{\epsilon})) - n_i \cdot v_{iaj}(t, \boldsymbol{\epsilon}) > \frac{\epsilon_t}{2 \cdot |\tilde{S}_{aj}(t, \boldsymbol{\epsilon})| \cdot A \cdot J} \right)$$

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

$$\leq \sum_{t=1}^{T} \sum_{\substack{(a,j):\, \tilde{S}_{aj}(t,\boldsymbol{\epsilon})\neq\varnothing \\ i\in\tilde{S}_{aj}(t,\boldsymbol{\epsilon})}} \exp\left\{ -\frac{\epsilon_t^2}{12\cdot|\tilde{S}_{aj}(t,\boldsymbol{\epsilon})|^2\cdot A^2\cdot J^2\cdot n_i\cdot v_{iaj}(t,\boldsymbol{\epsilon})}\right\}$$

$$+\sum_{t=1}^{T} \sum_{\substack{(a,j):\, \tilde{S}_{aj}(t,\boldsymbol{\epsilon})\neq\varnothing \\ i\in\tilde{S}_{aj}(t,\boldsymbol{\epsilon})}} \exp\left\{ -\frac{\epsilon_t}{6\cdot|\tilde{S}_{aj}(t,\boldsymbol{\epsilon})|\cdot A\cdot J}\right\}$$

$$\leq A\cdot J^2\cdot\sum_{t=1}^{T}\exp\left\{-\frac{\epsilon_t^2}{12\cdot A^2\cdot J^4\cdot[\sum_{\ell=1}^{J} n_\ell]}\right\}$$

$$+A\cdot J^2\cdot\sum_{t=1}^{T}\exp\left\{-\frac{\epsilon_t}{6\cdot A\cdot J^2}\right\}. \tag{21}$$

The first inequality in (21) follows since $\tilde{S}_{a,j}(t,\boldsymbol{\epsilon})=\varnothing$ implies that

$$\mathbb{P}\left(\tilde{X}_j^a(t,\boldsymbol{\epsilon})-x_j^{*a}(t,\boldsymbol{\epsilon})>\frac{\epsilon_t}{2AJ}\right)=0.$$

The second inequality in (21) follows by application of the Chernoff bound for the binomial random variable, specifically, if $X\sim\text{Bin}(n,p)$ then

$$\mathbb{P}(X-np>\delta)\leq\exp\left\{-\frac{\delta^2}{3np}\right\}\quad\text{for all }\delta\in[0,np),$$

$$\mathbb{P}(X-np>\delta)\leq\exp\left\{-\frac{\delta}{3}\right\}\quad\text{for all }\delta\geq np,$$

which implies that

$$\mathbb{P}(X-np>\delta)\leq\exp\left\{-\frac{\delta^2}{3np}\right\}+\exp\left\{-\frac{\delta}{3}\right\}\quad\text{for all }\delta\geq 0.$$

The last inequality in (21) follows since $|\tilde{S}_{a,j}(t,\boldsymbol{\epsilon})|\leq J$ and $n_i\cdot v_{iaj}(t,\epsilon)\leq\sum_{\ell=1}^{J} n_\ell$.

Similarly, we can also bound $\mathbb{P}(\tilde{\mathcal{A}}_2(\boldsymbol{\epsilon})^c)$ as

$$\mathbb{P}(\tilde{\mathcal{A}}_2(\boldsymbol{\epsilon})^c)\leq J^2\cdot\sum_{t=1}^{T}\left[\exp\left\{-\frac{\epsilon_t^2}{12\cdot J^4\cdot[\sum_{\ell=1}^{J} n_\ell]}\right\}+\exp\left\{-\frac{\epsilon_t}{6\cdot J^2}\right\}\right]. \tag{22}$$

As for $\mathbb{P}(\tilde{\mathcal{A}}_3(\boldsymbol{\epsilon})^c)$, note that $\tilde{N}_j(T+1,\boldsymbol{\epsilon})$ can also be written as a sum of J independent binomial random variables. Specifically,

$$\tilde{N}_j(T+1,\boldsymbol{\epsilon})\sim\sum_{i=1}^{J}\text{Bin}(n_i,r_{ij}(t,\boldsymbol{\epsilon})),$$

where $r_{ij}(t,\boldsymbol{\epsilon})$ is the probability that, under MRAC, a bandit that starts with state i at the beginning of period 1 ends up with state j at the beginning of period $T+1$. Applying the Chernoff bound to $\mathbb{P}(\tilde{\mathcal{A}}_3(\boldsymbol{\epsilon})^c)$, together with the fact that $|\mathbb{U}|\leq J$, yields

$$\mathbb{P}(\tilde{\mathcal{A}}_3(\boldsymbol{\epsilon})^c)\leq J^2\cdot\left[\exp\left\{-\frac{\epsilon_T^2}{3\cdot J^4\cdot[\sum_{\ell=1}^{J} n_\ell]}\right\}+\exp\left\{-\frac{\epsilon_T}{3\cdot J^2}\right\}\right]. \tag{23}$$

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

Putting all the bounds in (21), (22), and 23) back into (18), and noting that the bound in (21) is larger than the bounds in (22) and (23), completes the proof.

Appendix B. Proof of the results in Section 6

B.1. Proof of Lemma 4

The proof is similar to the proof of Lemma 3 (unless otherwise noted, all notation used here have the same meaning as those used in the proof of Lemma 3). The difference lies in computing a bound for $\mathbb{P}(\tilde{\mathcal{A}}_i(\boldsymbol{\epsilon})^c)$ for $i = 1, 2, 3$. Note that $\tilde{X}_j^a(t, \boldsymbol{\epsilon})$ is now the sum of J independent binomial random variables *and* a Poisson random variable, i.e.

$$\tilde{X}_j^a(t, \boldsymbol{\epsilon}) \sim \sum_{i=1}^{J} \text{Bin}(n_i, v_{iaj}(t, \boldsymbol{\epsilon})) + \text{Pois}\left(\sum_{s=1}^{t} \sum_{i=1}^{J} \lambda_{is} \cdot \tilde{v}_{iaj}(s, t, \boldsymbol{\epsilon}) \right),$$

where $\tilde{v}_{iaj}(s, t, \boldsymbol{\epsilon})$ is the probability that a new bandit that arrives with state i at time s ends up with state j at the beginning of time t and being applied action a at time t. We will use the following inequality for the Poisson random variable: if $X \sim \text{Pois}(\lambda)$ then

$$\begin{aligned} \mathbb{P}(X - \lambda > \delta) &\leq \frac{\mathbb{E}[\exp\{r \cdot (X - \lambda)\}]}{\exp\{r\delta\}} \\ &= \exp\{\lambda \cdot (\mathrm{e}^r - r - 1) - \delta r\} \\ &\leq \exp\{\lambda r^2 - \delta r\} \end{aligned}$$

for all $r \in [0, 1]$; specifically, if $0 \leq \delta \leq 2\lambda$, then $\mathbb{P}(X - \lambda > \delta) \leq \exp\{-\delta^2/4\lambda\}$ (this can be proved by simply substituting $r = \delta/2\lambda$ in the previous bound). Now, as in the proof of Lemma 3, we can bound

$$\begin{aligned} &\mathbb{P}(\tilde{\mathcal{A}}_1(\boldsymbol{\epsilon})^c) \\ &\leq \sum_{t=1}^{T} \sum_{\substack{(a,j):\, \tilde{S}_{a,j}(t,\boldsymbol{\epsilon}) \neq \varnothing \\ i \in \tilde{S}_{a,j}(t,\boldsymbol{\epsilon})}} \mathbb{P}\left(\text{Bin}(n_i, v_{iaj}(t, \boldsymbol{\epsilon})) - n_i \cdot v_{iaj}(t, \boldsymbol{\epsilon}) > \frac{\epsilon_t}{2 \cdot (1 + |\tilde{S}_{a,j}(t, \boldsymbol{\epsilon})|) \cdot A \cdot J} \right) \\ &\quad + \sum_{t=1}^{T} \mathbb{P}\left(\text{Pois}\left(\sum_{s=1}^{t} \sum_{i=1}^{J} \lambda_{si} \cdot \tilde{v}_{iaj}(s, t, \boldsymbol{\epsilon}) \right) \right. \\ &\quad \left. - \sum_{s=1}^{t} \sum_{i=1}^{J} \lambda_{si} \cdot \tilde{v}_{iaj}(s, t, \boldsymbol{\epsilon}) > \frac{\epsilon_t}{2 \cdot (1 + |\tilde{S}_{a,j}(t, \boldsymbol{\epsilon})|) \cdot A \cdot J} \right) \\ &\leq A \cdot J^2 \cdot \sum_{t=1}^{T} \exp\left\{ - \frac{\epsilon_t^2}{48 \cdot A^2 \cdot J^4 \cdot [\sum_{\ell=1}^{J} n_\ell]} \right\} \\ &\quad + A \cdot J^2 \sum_{t=1}^{T} \exp\left\{ - \frac{\epsilon_t}{12 \cdot A \cdot J^2} \right\} \\ &\quad + \sum_{t=1}^{T} \exp\left\{ - \frac{\epsilon_t^2}{64 \cdot A^2 \cdot J^4 \cdot [\sum_{s=1}^{t} \sum_{i=1}^{J} \lambda_{is}]} \right\}, \end{aligned} \tag{24}$$

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

where the last inequality follows since $J \geq 1$ implies that $1 + |\tilde{S}_{a,j}(t, \boldsymbol{\epsilon})| \leq 1 + J \leq 2J$ and the simple fact that $\sum_{s=1}^{t} \sum_{i=1}^{J} \lambda_{si} \cdot \tilde{\nu}_{iaj}(s, t, \boldsymbol{\epsilon}) \leq \sum_{s=1}^{t} \sum_{i=1}^{J} \lambda_{si}$.

Similarly, we can also bound $\mathbb{P}(\tilde{\mathcal{A}}_2(\boldsymbol{\epsilon})^c)$ and $\mathbb{P}(\tilde{\mathcal{A}}_3(\boldsymbol{\epsilon})^c)$ as follows:

$$\mathbb{P}(\tilde{\mathcal{A}}_2(\boldsymbol{\epsilon})^c) \leq J^2 \cdot \sum_{t=1}^{T} \left[\exp\left\{ -\frac{\epsilon_t^2}{48 \cdot J^4 \cdot [\sum_{\ell=1}^{J} n_\ell]} \right\} + \exp\left\{ -\frac{\epsilon_t}{12 \cdot J^2} \right\} \right] + \sum_{t=1}^{T} \exp\left\{ -\frac{\epsilon_t^2}{64 \cdot J^4 \cdot [\sum_{s=1}^{t} \sum_{i=1}^{J} \lambda_{is}]} \right\}, \tag{25}$$

and

$$\mathbb{P}(\tilde{\mathcal{A}}_3(\boldsymbol{\epsilon})^c) \leq J^2 \cdot \left[\exp\left\{ -\frac{\epsilon_T^2}{12 \cdot J^2 \cdot [\sum_{\ell=1}^{J} n_\ell]} \right\} + \exp\left\{ -\frac{\epsilon_T}{6 \cdot J} \right\} \right] + \exp\left\{ -\frac{\epsilon_T^2}{16 \cdot J^4 \cdot [\sum_{s=1}^{T} \sum_{i=1}^{J} \lambda_{is}]} \right\}. \tag{26}$$

Putting the bounds in (24), (25), and (26) together with $\mathbb{P}(\tilde{\mathcal{A}}(\boldsymbol{\epsilon})) \geq 1 - \mathbb{P}(\tilde{\mathcal{A}}_1(\boldsymbol{\epsilon})^c) - \mathbb{P}(\tilde{\mathcal{A}}_2(\boldsymbol{\epsilon})^c) - \mathbb{P}(\tilde{\mathcal{A}}_3(\boldsymbol{\epsilon})^c)$ yields the desired result (simply note that the bound in (24) is larger than the bounds in (25) and (26)). This completes the proof of Lemma 4.

Remark 5. For the setting with bandit departure and abandonment as discussed in Section 6, bounds (24), (25), and (26) can be significantly tightened by replacing the term $\sum_{s=1}^{t} \sum_{i=1}^{J} \lambda_{is}$ with $\sum_{s=1}^{t} \alpha^{t-s} \cdot [\sum_{j=1}^{J} \lambda_{jt}]$. This is so because $\tilde{\nu}_{iaj}(s, t, \boldsymbol{\epsilon}) \leq \delta^{t-s}$ for all $t > s$, which implies that $\sum_{s=1}^{t} \sum_{i=1}^{J} \lambda_{si} \cdot \tilde{\nu}_{iaj}(s, t, \boldsymbol{\epsilon}) \leq \sum_{s=1}^{t} \sum_{i=1}^{J} \delta^{t-s} \cdot \lambda_{si}$.

References

AHMAD, S. H. A. *et al.* (2009). Optimality of myopic sensing in multichannel opportunistic access. *IEEE Trans. Inf. Theory* **55,** 4040–4050.

ALTMAN, E. (1999). *Constrained Markov Decision Processes*. Chapman & Hall/CRC, Boca Raton, FL.

AYER, T. *et al.* (2016). Prioritizing hepatitis C treatment in U.S. prisons. Preprint. Available at SSRN: https://ssrn.com/abstract=2869158.

BERRY, D. A. AND FRISTEDT, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Chapman & Hall, London.

BERTSIMAS, D. AND NIÑO-MORA, J. (2000). Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operat. Res.* **48,** 80–90.

BRADT, R. N., JOHNSON, S. M. AND KARLIN, S. (1956). On sequential designs for maximizing the sum of *n* observations. *Ann. Math. Statist*. **27**, 1060–1074.

CARO, F. AND GALLIEN, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Manag. Sci*. **53**, 276–292.

COHEN, K., ZHAO, Q. AND SCAGLIONE, A. (2014). Restless multi-armed bandits under time-varying activation constraints for dynamic spectrum access. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, IEEE, pp. 1575-1578.

DEO, S. *et al.* (2013). Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operat. Res.* **61,** 1277–1294.

GITTINS, J. C. AND JONES D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (Budapest, 1972; Colloq. Math. Soc. János Bolyai **9**), North-Holland, Amsterdam, pp. 241–266.

GITTINS, J., GLAZEBROOK, K. AND WEBER, R. (2011). *Multi-Armed Bandit Allocation Indices*, 2nd edn. John Wiley, Chichester.

HU, W. AND FRAZIER, P. (2017). An asymptotically optimal index policy for finite-horizon restless bandits. Preprint. Available at https://arxiv.org/abs/1707.00205v1.

https://doi.org/10.1017/apr.2019.29 Published online by Cambridge University Press

KELLY, F. P. (1981). Multi-armed bandits with discount factor near one: the Bernoulli case. *Ann. Statist.* **9,** 987–1001.

LE NY, J., DAHLEH, M. AND FERON, E. (2008). A linear programming relaxation and a heuristic for the restless bandit problem with general switching costs. Preprint. Available at https://arxiv.org/abs/0805.1563v1.

LEE, E., LAVIERI, M. S. AND VOLK, M. (2018). Optimal screening for hepatocellular carcinoma: a restless bandit model. *Manufacturing Service Operat. Manag.* **21**.

MAHAJAN, A. AND TENEKETZIS, D. (2008). Multi-armed bandit problems. In *Foundations Applications of Sensor Management*, Springer, Boston, MA, pp. 121–151.

NAIN, P. AND ROSS, K. W. (1986). Optimal priority assignment with hard constraint. *IEEE Trans. Automatic Control* **31,** 883–888.

NIÑO-MORA, J. (2011). Computing a classic index for finite-horizon bandits. *INFORMS J. Comput.* **23,** 254–267.

PAPADIMITRIOU, C. H. AND TSITSIKLIS, J. N. (1999). The complexity of optimal queuing network control. *Math. Operat. Res.* **24,** 293–305.

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc*. **58**, 527–535.

SCHRIJVER, A. (2000). *Theory of Linear and Integer Programming*. John Wiley, New York.

VERLOOP, I. M. (2016). Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *Ann. Appl. Prob.* **26,** 1947–1995.

WASHBURN, R. B. (2008). Application of multi-armed bandits to sensor management. In *Foundations and Applications of Sensor Management*, Springer, Boston, MA, pp. 153–175.

WEBER, R. R. AND WEISS, G. (1990). On an index policy for restless bandits. *J. Appl. Prob.* **27,** 637–648.

WHITTLE, P. (1980). Multi-armed bandits and the Gittins index. *J. R. Statist. Soc. B* **42,** 143–149.

WHITTLE, P. (1988). Restless bandits: activity allocation in a changing world. In *A Celebration of Applied Probability* (J. Appl. Prob. Spec. Vol. **25(A)**), ed. J. Gani, Applied Probability Trust, Sheffield, pp. 287–298.

ZAYAS-CABÁN, G., JASIN, S. AND WANG, G. (2019). An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits. Supplementary material. Available at https://doi.org/10.1017/apr.2019.29