MICHIGAN ROSS

Working Paper

An Asymptotically Optimal Heuristic for General Non-Stationary Finite-Horizon Restless Multi-Armed Multi-Action Bandits

Gabriel Zayas-Caban
Center for Healthcare Engineering and Patient Safety
University of Michigan

Stefanus Jasin
Stephen M. Ross School of Business
University of Michigan

Guihua Wang
Stephen M. Ross School of Business
University of Michigan

Ross School of Business Working Paper
Working Paper No. 1374
March 2017

UNIVERSITY OF MICHIGAN

An Asymptotically Optimal Heuristic for General Non-stationary Finite-Horizon Restless Multi-Armed Multi-Action Bandits

Gabriel Zayas-Cabán
gzayasca@umich.edu
Center for Healthcare Engineering and Patient Safety, University of Michigan, Ann Arbor, MI 48109

Stefanus Jasin, Guihua Wang
sjasin, guihuaw@umich.edu
Stephen M. Ross School of Business, University of Michigan, Ann Arbor, MI 48109

We propose an asymptotically optimal heuristic, which we termed the Randomized Assignment Control (RAC) for restless multi-armed bandit problems with discrete-time and finite states. It is based on a linear programming relaxation to the original stochastic control formulation. In contrast to most of the existing literature, we consider a finite horizon with multiple actions and time-dependent (i.e. non-stationary) upper bound on the total number of bandits that can be activated each time period. 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. Numerical simulations confirms that our proposed policy indeed performs well in the asymptotic setting. Perhaps more surprisingly, these simulations show that RAC performs well in the non-asymptotic setting as well. Finally, we show that RAC is asymptotically optimal for a dynamic population, where bandits can randomly arrive and depart the system, and discuss how our framework extends to more general costs and constraints.

Key words: xxx; xxx

1. Introduction

We present a policy that is asymptotically-optimal for a general finite-horizon restless multi-armed bandit problem. A multi-armed bandits problem (MABP) involves activating competing bandits/arms sequentially over time. A fixed number of bandits have to be activated at any given time, and each bandit evolves according to a controlled stochastic process when it is activated. A solution to a MABP specifies which bandits can be activated at each decision epoch to minimize either the expected discounted or long-run average cost associated with how the bandits evolve over time. In the MABP literature, one of the most celebrated results is the Gittins index policy, introduced by Gittins in 1979. 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 both expected discounted and long-run average costs. About a decade after Gittins (1979), Whittle (1988) generalized the MABP to allow non-active bandits to also change states (dubbed by Whittle as a *changing world* setting), giving rise to the restless multi-armed bandits problem (RMABP).

RMABP is a general modeling framework encompassing many applications: sequential selection of clinical trials in medicine, sensor management, manufacturing systems, queueing networks, and appointment scheduling (e.g., Gittins et al., 2011; Deo et al, 2013). Unfortunately, an optimal policy for a RMABP is rarely a Gittins index policy and is frequently difficult to determine in any tractable manner. 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. He could then define an indexability property that ensured the relaxed problem has an optimal policy similar to a Gittins index policy, that is, one that assigns each bandit an index and activates bandits based on their index. This policy, known as Whittle's index policy, approximates the solution for the original RMABP and reduces to Gittins index policy when bandits do not change states if they are not activated. Whittle conjectured his index policy is asymptotically optimal when the number of bandits that can be activated per period and the population of bandits grow proportionally large. In their seminal work, Weber and Weiss (1990) proved Whittle's conjecture for bandits 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 Whittle's index policy can fail to be asymptotically optimal if the global attractor condition is not satisfied.

In the same asymptotic setting as in Whittle (1998) and Weber and Weiss (1990), we introduce an asymptotically-optimal policy, called *Randomized Assignment Control* (RAC), that does not require an indexability property and applies to general finite-horizon RMABPs. Its control parameters are constructed by formulating a Linear Programming (LP) relaxation of the RMABP and then finding an optimal solution of a perturbation to this LP relaxation. Finite-horizon RMABPs are particularly useful when the dynamics of the bandits are either non-stationary or model parameters need to be re-estimated. Due to this uncertainty, a decision maker may be more concerned with performance in a finite time window than over the long-run. Although the asymptotic optimality of RAC is for the finite horizon problem, we also show that it remains asymptotically optimal if the decision horizon is allowed to grow at a certain rate. Moreover, as we discuss in Remark 1 in Section 4, the decision horizon can be large when a discount factor is included in the cost term.

We also show numerically that RAC performs well even when the number of bandits that can be activated per period and the population size of bandits are relatively small (i.e., the non-asymptotic setting). So, RAC should be practical for a wide range of applications. We then allow bandits to arrive and leave stochastically at the beginning of every period (c.f. Verloop, 2015). Again, RAC is shown to be asymptotically-optimal, thereby extending the utility of RAC to applications that can be modeled as discrete-time queueing systems when decisions are made in batches.

We view our work as having the following four contributions. First, to the best of our knowledge, we are the first to propose an asymptotically optimal heuristic for a general RMABP for fixed and dynamic population models. Second, our proposed heuristic 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. Third, we allow for a non-stationary (or period-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 of a specific class and in a specific state. We emphasize that the analysis of RMABP with more than two actions has remained elusive in the literature unless additional structure is assumed such as dominant action. Our analysis can be easily extended to the setting with non-stationary transition matrices and costs, as well as to multi-class bandits (c.f. Verloop, 2015). Finally, the generality of our modeling framework means that our heuristic can be used in diverse applications including worker scheduling (Gittins et. al, 2011), resource allocation to a population in public health (Brandeau, 2005), and appointment scheduling/capacity management in healthcare where patients' health state follow Markovian dynamics and are fully observable (c.f. Deo et al, 2013).

The remaining of the paper is organized as follows. Section 2 summarizes the related literature and highlights our contributions. Section 3 details the basic model with a fixed population of bandits, along with the corresponding LP relaxation of the stochastic RAMBP. Section 4 provides the definition of RAC policy and analyzes its performance. Section 5 examines simple numerical examples to test the non-asymptotic performance of RAC. Section 6 analyzes the performance of RAC in the setting with dynamic population of bandits and Section 7 concludes the paper.

2. Related Literature

As alluded to in the introduction, RMABPs provide a very general modeling framework for sequential decision-making under uncertainty. It is thus not surprising that they have been used to support decision-making in many applications. Some of the existing literature on RMBAPs assumes that model parameters are given a priori and then focus on characterizing or approximating optimal policies. There is also literature that consider both learning (i.e., parameter estimation) and characterizing/approximating optimal policies. Our work belongs to the first category; we assume that the model parameters are already given and focus only on approximating optimal policies. We 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 (2015) for other related references. Here, we will only review works that are most closely related to ours.

Existing work on RMABP often assumes both a stationary activation budget (i.e., the maximum number of active bandits per period is independent of time) and only two possible actions per bandit (Whittle, 1981; Weiss, 1988; Verloop, 2015). To the best of our knowledge, a non-stationary activation budget has only been considered in Cohen et. al (2014). These authors derived sufficient conditions for the optimality of a myopic policy for a particular problem known as the finite-horizon discounted-cost dynamic spectrum access problem. This problem requires that a decision-maker search for idle channels in a spectrum of multiple channels with the busy/idle state of each channel evolving as a two-state Markov chain. In contrast, our work allows for an arbitrary finite number of actions and states and determines an asymptotically-optimal policy rather than sufficient conditions for optimality of a certain policy. RMABPs with multiple actions are often referred to as super-processes (c.f Gittins et. al, 2011) and were first considered 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, then there is an optimal policy that is indexable. A less strict condition is given in Gittins et. al (2011), but still to ensure there is an optimal policy that is indexable. Multiple actions were also considered in Verloop (2015) for multi-class restless bandits with a long-run average cost criteria. Similar to our work, they developed asymptotically-optimal policies for a fixed population of bandits that can arrive and depart from the system. Their model, however, focuses on generalizing the concept of dominant action. Our proposed RAC policy makes no such restriction to a class of dominant action policies.

Verloop (2015) introduced a broad class of priority policies (i.e., policies that prioritize certain bandits) that do not require indexability, but can still be asymptotically-optimal. However, since she considered the long-run average cost criteria, these conditions require that the differential equation describing the fluid model associated with the RMABP has a globally attracting equilibrium point, coinciding with similar conditions needed in Weiber and Weiss (1990). This condition is needed to guarantee that the equilibrium point induces a priority policy and that the process under this priority policy converges to the equilibrium point independent of 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 uniform integrable. For a fixed population of bandits, these are satisfied when the generated Markov process is unichain (i.e., the Markov chain has no two disjoint closed sets) so that the resulting Markov chain has a unique equilibrium distribution. For a dynamic population, they are satisfied provided that 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. We extend the dynamic population framework in Verloop (2015) in four ways. First, we allow time-varying constraints on the number of active bandits per period. 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 conditions.

In terms of methodology, our work is 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. 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 Ny et. al (2008) considered LP relaxations of the Dynamic Programming (DP) formulation of infinite-horizon RMABPs. An alternative LP relaxation can be derived by considering a fluid model of the RMAPB that only takes into account mean drifts of bandit dynamics (Weber and Weiss, 1990; Verloop, 2015). This fluid approach is also called the *Certainty Equivalent* approach in the broader Operations Research literature and is closest to the LP formulation in this paper.

Finite-horizon MABPs have been studied in Robbins (1972) and Bradt et. al (1956), who focused on the well-studied 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 both an efficient and exact algorithms to compute the index. To the best of our knowledge, we are the first to analyze finite-horizon RMABPs with a non-stationary activation budget and optimal policies that are non-indexable.

3. Basic Model

We consider a finite-horizon, discrete-time model where time $t \in \{1, \ldots, T+1\}$ with $T < \infty$. Let $\mathbb{J} = \{j : 1 \leq j \leq J\}$ denote the set of feasible states and $\mathbb{A} = \{a : 0 \leq a \leq A\}$ denote the set of feasible actions. Our 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, that represent states in which a bandit will incur a high penalization cost at the end of the horizon. 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, we will call a bandit either *active* when a proper action is applied to it or *passive* otherwise. We assume that each bandit transitions from state i to state j under action a according to a probability $p^a_{i,j}$ and that the maximum number of active bandits at time t is $b_t > 0$, called the *activation budget*. Two types of costs can be incurred by the system. First, a cost c^a_j 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 $T+1$, where we allow for at most m bandits to be in an undesirable state at time $T+1$ without being penalized.

The decision-making scenario is as follows. At time t, the decision-maker decides the number of bandits in each state to receive a specific treatment. After receiving treatment, a bandit incurs a cost and transitions to a potentially new state at time $t+1$. The objective is to minimize the expected total treatment and penalty costs. Let Π denote the set of all non-anticipating policies and π denote a feasible policy in Π. Because each bandit has identical transition and cost dynamics, we do need to keep track of individual bandits and instead keep track of the number of bandits in state j that receive treatment a at period t under π, denoted by $X^{\pi,a}_j(t)$, as well as the the number of bandits in state i that receive treatment a at time t under π and then transition to state j, denoted by $Y^{\pi,a}_{i,j}(t)$. Additionally, we assume that $Y^{\pi,a}_i(t) = (Y^{\pi,a}_{i,j}(t))$ is a vector of multinomial random variables and that $n_{j,1}$ is the initial number of arms in state j at time $t = 1$.

If we let V^π denote the total expected costs (both treatment and penalty costs) under policy $\pi \in \Pi$ with certain constraints:

$$
\begin{aligned}
V^\pi &= \mathbb{E}^\pi\left[\sum_{t=1}^{T}\sum_{a=0}^{A}\sum_{j=1}^{J} c^a_j \cdot X^{\pi,a}_j(t) + \phi \cdot \left(\sum_{a=0}^{A}\sum_{j\in\mathbb{U}} X^{\pi,a}_j(T+1) - m\right)^+\right] \\
\text{s.t.} \quad & \sum_{a=0}^{A} X^{\pi,a}_j(t) = \sum_{a=0}^{A}\sum_{i=1}^{J} Y^{\pi,a}_{i,j}(t-1) \qquad \forall j \geq 1,\, t \geq 2 \\
& \sum_{a=0}^{A} X^{\pi,a}_j(1) = n_{j,1} \qquad \forall j \geq 1 \\
& \sum_{a=1}^{A}\sum_{j=1}^{J} X^{\pi,a}_j(t) \leq b_t \qquad \forall t \geq 1,
\end{aligned}
$$

then the stochastic control model can be expressed as:

$$V^S = \min_{\pi \in \Pi} V^\pi, \tag{1}$$

where the constraints hold almost surely.

A few remarks are in order. First, the seemingly simple stochastic control model in (1) is in fact surprisingly general. As alluded, it can be used to model applications in scheduling, capacity management (e.g. admission and/or service rate control), resource allocation, and others. It can also be easily modified to include several features that are commonly used in the queueing control literature such as non-stationary random arrival processes, non-stationary random service processes (e.g. non-stationary budget), random service completions, and random abandonments (see Section 6). Second is that the cost structure and budget constraints for the stochastic control model presented in (1) follow those that are commonly found in the RMABP literature. We have done this for clarity of presentation. However, our analysis and results also hold under more general settings. For example, the cost terms c_j^a can depend on time, so that, for instance, we can include discounted costs. This versatility comes from the fact that we consider a finite-horizon total expected cost criteria instead of infinite-horizon discounted expected or long-run average expected costs as our optimization objective. This frees us from having to impose structural conditions on the model required to guarantee convergence to an equilibrium point (c.f Verloop, 2015). In what follows, for ease of notation, whenever it is clear from context which policy is being used, we will suppress the notational dependency on π.

Rather than solving the stochastic control model in (1), we will instead solve a family of associated Linear Programs (LP) indexed by $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_T) \geq \mathbf{0}$:

$$\begin{aligned}
V^D(\epsilon) \;=\; \min_{x,z} \quad & \sum_{t=1}^{T}\sum_{a=0}^{A}\sum_{j=1}^{J} c_j^a \cdot x_j^a(t,\epsilon) + \phi \cdot z(\epsilon) && (2)\\
\text{s.t.} \quad & \sum_{a=0}^{A} x_j^a(t,\epsilon) = \sum_{a=0}^{A}\sum_{i=1}^{J} x_i^a(t-1,\epsilon)\cdot p_{i,j}^a \qquad \forall j \geq 1,\, t \geq 2 \\
& \sum_{a=0}^{A} x_j^a(1,\epsilon) = n_{j,1} \qquad \forall j \geq 1 \\
& \sum_{a=1}^{A}\sum_{j=1}^{J} x_j^a(t,\epsilon) \leq b_t - \epsilon_t \qquad \forall t \geq 1 \\
& z(\epsilon) \geq \sum_{j \in \mathbb{U}}\sum_{a=0}^{A}\sum_{i=1}^{J} x_i^a(T,\epsilon)\cdot p_{i,j}^a - m \\
& z(\epsilon),\, x_j^a(t,\epsilon) \geq 0 \qquad \forall a \geq 0,\, j \geq 1,\, t \geq 1
\end{aligned}$$

One can arrive at LP (2) by replacing all the random variables in (1) with their expected values, replacing the original budget b_t in (1) with $b_t - \epsilon_t$, and introducing a new variable $z(\epsilon)$ to capture the penalty for bandits in undesirable states at the end of the time horizon. In the special case when $\epsilon = \mathbf{0}$, the LP (2) is simply the deterministic relaxation of the original stochastic problem (1). Otherwise when $\epsilon > \mathbf{0}$, we interpret ϵ as a buffer to not exceed the activation budget as $X_j^{a,\pi}(t)$ randomly deviates from $x_j^a(t)$. The magnitude of ϵ_t is chosen such that we avoid exceeding the activation budget with a high probability.

We end this section with the following result. The proof is straightforward and is therefore omitted.

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

The result tells us that $V^D(\mathbf{0})$ is a lower bound of V^S. As a result, we can use $V^D(\mathbf{0})$ as a proxy for V^S and study the performance of any feasible policy π by analyzing the difference between V^π and $V^D(\mathbf{0})$. In this paper, we will refer to $V^\pi - V^D(\mathbf{0})$ simply as the *loss* of policy π.

4. Randomized Activation Control

We define our policy *Randomized Activation Control* (RAC) for the stochastic control model in (1) using an optimal solution $x_j^{*,a}(t,\epsilon)$ and $z^*(\epsilon)$ of LP (2) for a given $\epsilon \geq \mathbf{0}$. We use this optimal solution, which may not be unique, to introduce categorical random variables $Z_{j,l}(t,\epsilon) \in \mathbb{A}$ with the property that $\mathbb{P}(Z_{j,l}(t,\epsilon) = a) = q_j^{*,a}(t,\epsilon)$ for all $a \in \mathbb{A}$ where

$$n_j^*(t,\epsilon) = \sum_{a=0}^{A} x_j^{*,a}(t,\epsilon)$$

$$q_j^{*,a}(t,\epsilon) = \begin{cases} \frac{x_j^{*,a}(t,\epsilon)}{n_j^*(t,\epsilon)} & \text{if } n_j^*(t,\epsilon) > 0 \\ 1_{\{a=0\}} & \text{if } n_j^*(t,\epsilon) = 0 \end{cases}$$

By definition, $n_j^*(1,\epsilon) = n_{j,1}$ and $q_j^{*,a}(t,\epsilon) \in [0,1]$. These random variables specify which actions are taken by our policy at time t, as described below.

```
Randomized Activation Control (RAC)
1: Pick $\epsilon$ and solve LP (2).
2: Compute $\mathbf{q}^*(\epsilon)$.
3: At time $t$, do:
        a. Randomly pick an arm $l$ that has not been picked before;
        b. If bandit $l$ is in state $j$, randomly generate $Z_{j,l}(t,\epsilon)$;
        c. If total activated bandit so far is smaller than $b_t$, apply action $Z_{j,l}(t,\epsilon)$ to arm $l$;
        d. If total activated bandit so far is exactly $b_t$, apply action 0 to arm $l$;
```

Note that RAC randomizes the order in which bandits are assigned an action at time t. An alternative policy could adjust RAC to prioritize certain bandits over others. For example, bandits in a state $j = 1$ could be assigned an action at time t before bandits in state $j = 2$ are assigned an action. While this is an important consideration, we do not address how changes to bandit order in the algorithm could improve the policy.

The performance of RAC for the stochastic control model in (1) depends crucially on the choice of ϵ. To see this, note that we can decompose the loss of RAC as follows:

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

If ϵ is close to 0, then $V^D(\epsilon) - V^D(\mathbf{0})$ is also close to 0. However, $V^{RAC} - V^D(\epsilon)$ could be large, since deviations of X_j^a from x_j^a is likely to lead RAC to exhaust the activation budget each time period before those bandits that should have been activated are even considered for activation. As a result, many bandits may end up in undesirable states at the beginning of period $T+1$. On the other hand, if ϵ is large, then $V^{RAC} - V^D(\epsilon)$ will be close to 0. In this case, there is a negligible probability that we will reach the per period budget each time period, and we will be able to activate all of bandits that need to be activated to prevent too many bandits from reaching the set of undesirable states at the beginning of period $T+1$. However, $V^D(\epsilon) - V^D(\mathbf{0})$ in this case can be potentially large. It follows that care must be taken when choosing ϵ.

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

Lemma 2. *There exists a constant $M > 0$ independent of T and $\epsilon \geq 0$ satisfying $\epsilon_t \leq b_t$ for all t such that $V^D(\epsilon) - V^D(\mathbf{0}) \leq M \cdot \left[\sum_{t=1}^T \epsilon_t\right]$.*

Lemma 2 says that $V^D(\epsilon) - V^D(\mathbf{0})$ is roughly proportional to $\sum_{t=1}^T \epsilon_t$. The proof of this Lemma can be found in the Appendix and is based on a duality argument. The proof depends on the fact that LP (2) can be transformed into a separable LP whose optimal dual solution is also an optimal dual solution for the original LP (2). We remark that we *cannot* directly apply existing results on LP sensitivity analysis, like the one presented in Schrijver (2000, see Section 10.4) since it yields a bound for $V^D(\epsilon) - V^D(\mathbf{0}) = O(T^2 \, ||\epsilon||_\infty)$ that will be too loose to prove asymptotic-optimality.

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

Lemma 3. *For any $\epsilon \geq 0$, we have*

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

with probability at least

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

If ϵ is small, then the bound in (3) is small and holds with a small probability; if, on the other hand, ϵ is large, the bound in (3) is large and holds with a large probability. Ideally, we would like to choose ϵ that yields a small bound in (3) and a large probability in (4). To do this, we consider an asymptotic setting where $n_{j,1}$ (for all j), b_t (for all t), and m are uniformly scaled by a factor of $\theta > 0$. This is the same asymptotic setting considered in Weber and Weiss (1990) and Verloop (2015). Let V_θ^S and V_θ^π denote the corresponding total expected costs under the optimal policy and a feasible policy $\pi \in \Pi$, respectively. Also, let $V_\theta^D(\epsilon)$ denote the optimal value of the corresponding LP (2). It is not difficult to see that the optimal solution of the LP (2) for $\epsilon = \mathbf{0}$ and $\theta > 0$ is given by $x_{\theta,j}^{*,a}(t,\mathbf{0}) = \theta \cdot x_j^{*,a}(t,\mathbf{0})$ and $z_\theta^*(\mathbf{0}) = \theta \cdot z^*(\mathbf{0})$ so that $V_\theta^D(\mathbf{0}) = \theta \cdot V^D(\mathbf{0})$. We can define $n_{\theta,j}^*(t,\epsilon)$ and $q_{\theta,j}^{*,a}(t,\epsilon)$ analogously to how $n_j^*(t,\epsilon)$ and $q_j^{*,a}(t,\epsilon)$ are defined, but with $x_j^{*,a}(t,\mathbf{0})$ replaced with $x_{\theta,j}^{*,a}(t,\mathbf{0})$. Let $n_{tot} := \sum_{i=1}^J n_{i,1}$ and define $S(t) := \{(a,j) : x_j^{*,a}(t,\mathbf{0}) > 0 \text{ and } c_j^a > 0\}$. We state our main result in this section below.

Theorem 1. *For all t, suppose that $S(t) \neq \emptyset$ and $\epsilon_t = 6 \cdot A \cdot J^2 \cdot \sqrt{d \cdot n_{tot} \cdot \theta \cdot \log\theta}$ for some $d > 0$. If $\theta \cdot b_t \geq \epsilon_t$ for all t, we have:*

$$\frac{V_\theta^{RAC} - V_\theta^S}{V_\theta^S} = O\left(\frac{T}{\theta^d} + \sqrt{\frac{d \cdot \log\theta}{\theta}}\right). \tag{5}$$

Proof. By Lemma 3, under the choice of ϵ in Theorem 1, $C_\theta^{RAC} - V_\theta^D(\epsilon)$ is $O(T \cdot \sqrt{d \cdot \theta \cdot \log\theta})$ with probability at least $1 - \Theta\left(\frac{T}{\theta^d}\right)$ (for the second exponential term in (4), for any given $d > 0$, we can further bound $\sqrt{d \cdot n_{tot} \cdot \theta \cdot \log\theta} \geq d \cdot \log\theta$ for all large θ). Since there is a total of $\theta \cdot n_{tot}$ bandits in each period, $C_\theta^{RAC} - V_\theta^D(\epsilon)$ is at most $\Theta(T \cdot \theta \cdot n_{tot})$ with probability at most $\Theta\left(\frac{T}{\theta^d}\right)$. Putting these two results together, we get $V_\theta^{RAC} - V_\theta^D(\epsilon) = O\left(T \cdot \sqrt{d \cdot \theta \cdot \log\theta} + \frac{T^2}{\theta^{d-1}}\right)$. By Lemma 1, under the choice of ϵ in Theorem 1, $V_\theta^D(\epsilon) - V_\theta^D(\mathbf{0})$ is $O(T \cdot \sqrt{d \cdot \theta \cdot \log\theta})$. Putting the bounds for $V_\theta^{RAC} - V_\theta^D(\epsilon)$ and $V_\theta^D(\epsilon) - V_\theta^D(\mathbf{0})$ together yields $V_\theta^{RAC} - V_\theta^D(\mathbf{0}) = O\left(T \cdot \sqrt{d \cdot \theta \cdot \log\theta} + \frac{T^2}{\theta^{d-1}}\right)$. The proof is complete by noting that $\frac{V_\theta^{RAC} - V_\theta^S}{V_\theta^S} \leq \frac{V_\theta^{RAC} - V_\theta^D(\mathbf{0})}{V_\theta^D(\mathbf{0})}$ and $V_\theta^D(\mathbf{0})$ is $\Theta(T \cdot \theta)$ (because $S(t) \neq \emptyset$ for all t, which implies that in each period we always incur a cost that is at least proportional to θ). ■

The condition $S(t) \neq 0$ for all t is relatively mild and simply means that we always activate some bandits in the deterministic system. As for the condition $\theta \cdot b_t \geq \epsilon_t$ for all t, it is immediately satisfied for all sufficiently large θ. 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, note that the bound depends on T only through the term $\frac{T}{\theta^d}$; technically, 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 $T \sim \theta^n$, then we can choose $d > n$ and RAC is still asymptotically optimal. Thus, despite the fact that our model is set as a finite-horizon model, our result suggests that RAC is quite versatile and can be applied to a problem with a very long decision horizon.

Remark 1 (Discounted Cost Criteria). Suppose that we multiply the cost term at time t with δ^{t-1} for some discount factor $\delta \in (0,1)$. Using the same arguments in the proofs of Lemmas 2 and 3, it can be shown that the bound in Lemma 2 becomes $M \cdot \left[\sum_{t=1}^{T} \delta^{t-1} \cdot \epsilon_t\right]$ for some $M > 0$ independent of T and $\epsilon \geq \mathbf{0}$, and the bound in (3) becomes $\left[\phi + \max_{a,j} c_j^a\right] \cdot \left[\sum_{t=1}^{T} \delta^{t-1} \cdot \epsilon_t\right]$. Suppose that $S(t) \neq 0$ for all t. If we now use $\epsilon_t = 6 \cdot A \cdot J^2 \cdot \sqrt{d \cdot n_{tot} \cdot \ln(t+e-1) \cdot \theta \cdot \ln\theta}$, then using the same arguments in the proof of Theorem 1, it is not difficult to see that

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

with probability at least $1 - \Theta\left(\sum_{t=1}^{T} \frac{1}{(t+e-1)^{d \cdot \ln\theta}}\right)$. Additionally, $C_\theta^{RAC} - V_\theta^D(\epsilon)$ is at most

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

with probability at most $\Theta\left(\sum_{t=1}^{T} \frac{1}{(t+e-1)^{d \cdot \ln\theta}}\right)$. Since $V_\theta^D(\mathbf{0})$ is $\Theta(\sum_{t=1}^{T} \delta^{t-1} \cdot \theta) = \Theta(\theta)$ and

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

for all large θ, we have

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

For this bound to hold, we need $\theta \cdot b_t \geq \epsilon_t$ for all t otherwise the LP (2) has a negative activation budget and hence, is infeasible. Given our choice of ϵ_t, this condition is immediately satisfied for all t and all large θ so long as $T = o\left(\exp\left\{\frac{\theta}{d \cdot \log\theta}\right\}\right)$. Thus, not only RAC is asymptotically optimal in the setting with discounted cost, its relative loss is also *independent* of T for very large T. This constraint on T tightens the bound in Theorem 1, which would otherwise grow exponentially with θ when T is close to $\exp\left\{\frac{\theta}{d \cdot \log\theta}\right\}$.

5. Numerical Experiments

In this section, we test the performance of RAC using two experiments. In the first experiment, we consider an instance of RMABP with 2 states and 2 actions; in the second, we consider an instance of RMABP with 5 states and 5 actions. In each experiment, we use $\epsilon_t = \sqrt{\theta \cdot \log\theta}$ for all t and choices of θ. The details of all other parameters can be found in the Appendix. We report the percentage loss of RAC (i.e., $\frac{V_\theta^{RAC} - V_\theta^D(\mathbf{0})}{V_\theta^D(\mathbf{0})}$) for the two experiments in Tables 1 and 2, respectively.

Table 1 **Percentage loss for 2 states/2 actions**

θ	$T=10$	$T=30$	$T=50$	$T=100$
1	5.40	6.45	5.77	5.14
5	4.90	3.44	5.09	5.40
10	3.76	5.25	4.46	3.80
20	3.35	3.53	4.34	3.98
40	3.32	2.17	2.67	2.72
60	2.37	2.79	2.29	1.83
80	2.49	2.29	1.85	1.66
100	2.14	1.38	2.33	1.92
200	1.32	1.05	1.58	1.48

Table 2 **Percentage loss for 5 states/5 actions**

θ	$T=10$	$T=30$	$T=50$	$T=100$
1	3.12	2.67	2.75	3.01
5	2.80	3.60	3.10	3.18
10	2.40	2.10	2.30	2.68
20	1.70	1.80	1.70	1.65
40	1.20	1.10	0.90	1.02
60	1.00	1.00	1.00	0.77
80	0.80	0.80	0.70	0.73
100	0.70	0.70	0.80	0.70
200	0.50	0.60	0.50	0.49

Note that, as predicted by Theorems 1 and 2, RAC performs better as θ increases. However, what is perhaps surprising, RAC also appears to be performing very well when θ is small (its relative loss is only about 6% when $\theta=1$). These results suggest RAC performs well not only in the asymptotic regime, but for a wide range of θ.

6. Model with Arrivals

We extend the basic model in Section 3 to allow for arrivals. We use the same notation as Section 3 with the addition of a stochastic arrival of $\Lambda_{j,t}$ bandits in state j at time t. We assume that $\Lambda_{j,t}$ is a Poisson random variable with mean $\lambda_{j,t}$. Both the stochastic and deterministic formulations of our new model can be written as follows:

$$V^S = \min_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=0}^{A}\sum_{j=1}^{J} c_j^a \cdot X_j^{\pi,a}(t) + \phi\cdot\left(\sum_{j\in\mathbb{U}} N_j^{\pi}(T+1)-m\right)^{+}\right] \tag{6}$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{a=0}^{A} X_j^{\pi,a}(t) = \sum_{a=0}^{A}\sum_{i=1}^{J} Y_{i,j}^{\pi,a}(t-1) + \Lambda_{j,t} \qquad \forall j\geq 1,\, t\geq 2 \\ & \sum_{a=0}^{A} X_j^{\pi,a}(1) = n_{j,1} + \Lambda_{j,1} \qquad \forall j\geq 1 \\ & \sum_{a=1}^{A}\sum_{j=1}^{J} X_j^{\pi,a}(t) \leq b_t \qquad \forall t\geq 1 \end{aligned}$$

$$\begin{aligned}
V^D(\epsilon) \; = \; \min_{x,z} \;\; & \sum_{t=1}^{T}\sum_{a=0}^{A}\sum_{j=1}^{J} c_j^a \cdot x_j^a(t,\epsilon) \, + \, \phi \cdot z(\epsilon) \qquad (7)\\
\text{s.t.} \;\; & \sum_{a=0}^{A} x_j^a(t,\epsilon) = \sum_{a=0}^{A}\sum_{i=1}^{J} x_i^a(t-1,\epsilon)\cdot p_{i,j}^a \, + \, \lambda_{j,t} \qquad \forall j \geq 1, t \geq 2\\
& \sum_{a=0}^{A} x_j^a(1,\epsilon) = n_{j,1} + \lambda_{j,1} \qquad \forall j \geq 1\\
& \sum_{a=1}^{A}\sum_{j=1}^{J} x_j^a(t,\epsilon) \leq b_t - \epsilon_t \qquad \forall t \geq 1\\
& z(\epsilon) \, \geq \, \sum_{j\in\mathbb{U}}\sum_{a=0}^{A}\sum_{i=1}^{J} x_i^a(T,\epsilon)\cdot p_{i,j}^a - m\\
& z(\epsilon),\, x_j^a(t,\epsilon) \geq 0 \qquad \forall a \geq 0,\, j \geq 1,\, t \geq 1
\end{aligned}$$

Lemmas 1 and 2 still hold (we omit the details), and RAC is defined exactly as it was in Section 4. The following result is the analogue of Lemma 3:

Lemma 4. *For any $\epsilon \geq 0$, we have*

$$C^{RAC} - V^D(\epsilon) \; \leq \; \left[\phi + \max_{a,j} c_j^a\right] \cdot \left[\sum_{t=1}^{T} \epsilon_t\right] \qquad (8)$$

with probability at least

$$\begin{aligned}
& 1 - 3\cdot A\cdot J^2 \cdot \sum_{t=1}^{T} \left[\exp\left\{-\frac{\epsilon_t^2}{48\cdot A^2\cdot J^4\cdot\left[\sum_{\ell=1}^{J} n_{\ell,1}\right]}\right\} + \exp\left\{-\frac{\epsilon_t}{12\cdot A\cdot J^2}\right\}\right]\\
& \quad - 3\cdot A\cdot J^2 \cdot \sum_{t=1}^{T} \exp\left\{-\frac{\epsilon_t^2}{64\cdot A^2\cdot J^4\cdot\left[\sum_{s=1}^{t}\sum_{i=1}^{J}\lambda_{i,s}\right]}\right\}. \qquad (9)
\end{aligned}$$

Note that the last summation in (9) is due to arrivals of new bandits at each time point. If $\lambda_{t,j} = 0$ for all t and j, then the last summation equals zero and the bound in (9) is identical to the bound in (4) except for that the numbers 48 and 12 appear in the expression rather than 12 and 6. Let $\lambda_{tot} := \max_t \sum_{j=1}^{J} \lambda_{j,t}$. We consider the same asymptotic setting as in Section 4, where we also scale the arrival rate $\lambda_{j,t}$ with θ. The following theorem is the analogue of Theorem 1:

Theorem 2. *For all t and i, suppose that $\lambda_{t,i} > 0$, $S(t) \neq \emptyset$, and $\epsilon_t = 12 \cdot A \cdot J^2 \cdot \sqrt{t\cdot d\cdot \max\{n_{tot},\lambda_{tot}\}\cdot\theta\cdot\log\theta}$ for some $d > 0$. If $\theta \cdot b_t \geq \epsilon_t$ for all t, we have:*

$$\frac{V_\theta^{RAC} - V_\theta^S}{V_\theta^S} \; = \; O\left(\frac{T^{3/2}}{\theta^{d/2}} + \sqrt{\frac{d\cdot T\cdot\log\theta}{\theta}}\right). \qquad (10)$$

The proof of Theorem 2 is similar to the proof of Theorem 1, and as a result, we defer the complete argument to the Appendix. Similar to when there were no arrivals, the condition $\theta \cdot b_t \geq \epsilon_t$ for all t is immediately satisfied for large θ, and the bound in Theorem 2 tells us that RAC is asymptotically-optimal, since we do not scale T with θ. However, note that the bound is weaker than the bound in Theorem 1 as the parameter T also shows up in the term $\sqrt{\frac{d \cdot T \cdot \log \theta}{\theta}}$. This term is the consequence of having more randomness in the system (from new arrivals), which requires more conservative buffers. From a practical perspective, this means that RAC generally performs best with new stochastic arrivals when the length of the decision horizon is not too long (mathematically, $T = o\left(\frac{\theta}{\log \theta}\right)$). That said, we want to stress that the bound in Theorem 2 is *very loose* as it does not exploit special structures/properties that may exist in the problem. Below, we provide two examples where imposing additional structure on the problem leads to a tighter bound in Theorem 2.

Example 1. Suppose that bandits arrive and stay in the system throughout the entire time horizon. Such a situation could arise modeling a population of patients with a chronic disease (e.g. type-I diabetes) whose health states can change, but never fully recover. For such a model, it is reasonable to assume that b_t should be proportional to t in order to avoid unbounded population sizes (i.e., the number of bandits that need to be properly treated grows faster than the available budget as the length of horizon gets larger). In addition, if a significant number of bandits are supposed to be properly activated at each period (i.e., there exists $\rho > 0$ such that $\sum_{a=1}^{A} \sum_{j=1}^{J} x_j^{*a}(t, \mathbf{0}) \geq \rho \cdot b_t$ for all t), we immediately get $V_\theta^S \geq V_\theta^D(\mathbf{0}) = \Theta(\theta \cdot \sum_{t=1}^{T} t) = \Theta(T^2 \cdot \theta)$. Using the same choice of ϵ_t as in Theorem 2, it is not difficult to see show that

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

Note that the above bound is stronger than the bound in Theorem 2. Since the parameter T appears in the second term only in the denominator, if $T \sim \theta^n$, we can choose $d > n$ and RAC is asymptotically optimal. Thus, similar to our result in Theorem 1, RAC can also be applied to a problem with a very long decision horizon. ■

Example 2. Similar to Example 1, suppose that each bandit may stay in the system throughout the entire horizon. However, assume also that the probability that a bandit stays exactly L periods in the system decays exponentially with L. Such a situation could arise in a model where bandits wait to receive a proper treatment and immediately leave the system once they receive a proper treatment. If a bandit has not received a proper treatment up until the end of the current period, it will stay in the system for the next period with probability $\alpha < 1$. For this setting, we can replace the summation $\sum_{s=1}^{t}\sum_{j=1}^{J}\lambda_{j,t}$ in bound (9) with $\sum_{s=1}^{t}\alpha^{t-s}\cdot\left[\sum_{j=1}^{J}\lambda_{j,t}\right]$. Again, if the hypotheses in Theorem 2 hold except with $\epsilon_t = 12\cdot A\cdot J^2\cdot\sqrt{(\sum_{s=0}^{t-1}\alpha^s)\cdot d\cdot\max\{n_{tot},\,\lambda_{tot}\}\cdot\theta\cdot\log\theta}$ for some $d>0$, then

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

Similar to the bound provided in Example 1, if $T\sim\theta^n$, we can choose $d>n$ and RAC is asymptotically optimal. This means, for example, that RAC performs well for a large decision horizon T even with Poisson arrivals and/or geometrically distributed waiting times. ■

The preceding examples highlight an important point: specific structure of the problem can be exploited to significantly tighten the bound in Theorem 2. It suggests, for instance, that the poor performance of RAC can be due to poor capacity management. The latter can be remedied by either making sure that we have a sufficiently large activation budget in each period or by limiting the admission of new bandits to the system. Thus, our RAC heuristic can be essentially coupled with either budget optimization, admission control, or both.

Remark 2 (Discounted Cost Criteria). As in Remark 1, suppose that we multiply the cost term in period t with δ^{t-1} for some discount factor $\delta\in(0,1)$. If $S(t)\neq\emptyset$ for all t and $\epsilon_t = 12\cdot A\cdot J^2\cdot\sqrt{t\cdot\log(t+e-1)\cdot d\cdot\max\{n_{tot},\,\lambda_{tot}\}\cdot\theta\cdot\log\theta}$, using similar arguments as those presented in Remark 1 and in the proof of Theorem 2, it is not difficult to show that

$$\frac{V_\theta^{RAC}-V_\theta^S}{V_\theta^S} \le \frac{V_\theta^{RAC}-V_\theta^D(\mathbf{0})}{V_\theta^D(\mathbf{0})} = O\left(\frac{1}{\theta^{d/2}}+\sqrt{\frac{d\cdot\log\theta}{\theta}}\right).$$

However, unlike in Remark 1 where the bound holds for $T=o\left(\exp\left\{\frac{\theta}{d\cdot\log\theta}\right\}\right)$, the above bound holds so long as $T\cdot\log T = o\left(\frac{\theta}{\log\theta}\right)$. This is because we need to ensure that $\theta\cdot b_t\ge\epsilon_t$ for all $t\le T$, which will yield a similar order of T for RAC for the undiscounted problem to be asymptotically optimal (c.f Theorem 2). However, we can still recover the exponential order of T (as in Remark 1) by exploiting additional structure of the problem such as that imposed in Examples 1 and 2.

7. Closing Remarks

We considered discrete-time, finite-horizon Restless Multi-armed Bandit Problems for a population of bandits. To our knowledge, we are the first to simultaneously allow for multiple actions and the maximum number of active bandits per period to depend on time (i.e. a non-stationary model). We propose a heuristic we termed the Randomized Assignment Control (RAC), which is based on a linear programming relaxation to the original stochastic control formulation and showed it is asymptotically- optimal. Similar to Verloop (2015), our heuristic does not depend on any indexability properties. However, in contrast to Verloop (2015), it does not require assumptions on the underlying structure of the Markov process generated by each policy (e.g. unichain) or assumptions on the dynamics on the associated deterministic approximation (e.g. globally stable attractor for the fluid approximation). We also extended our model to include random bandit arrivals and departures, more general costs (i.e. non-stationary), and general constraints. This extension ensures our approach could also apply to discrete-time queueing control models. We again show that RAC is asymptotically-optimal in this more general setting.

There are several avenues for further research. Developing and analyzing alternative policies is of clear interest. Comparing alternative policies with RAC in a numerical study can be considered. There are several model extensions worthy of consideration. 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. Needless to say, 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

Proof of Lemma 2. As noted in the paragraph following Lemma 2, we cannot directly use existing results on LP sensitivity analysis (e.g., Schrijver, 2000). However, we are still able to apply these results after transforming the LP. Define:

$$\begin{aligned} \tilde{V}^D(\epsilon) \; = \; \min_{x,z} \quad & \sum_{t=1}^{T}\sum_{a=0}^{A}\sum_{j=1}^{J} c_j^a \cdot x_j^a(t,\epsilon) + \phi \cdot z(\epsilon) \qquad (11) \\ \text{s.t.} \quad & \sum_{a=0}^{A}\sum_{i=1}^{J} x_i^a(t,\epsilon) \cdot p_{i,j}^a = n_j^*(t+1,\epsilon) \qquad \forall j \geq 1,\, t \geq 1 \\ & \sum_{a=0}^{A} x_j^a(1,\epsilon) = n_{j,1} \qquad \forall j \geq 1 \\ & \sum_{a=1}^{A}\sum_{j=1}^{J} x_j^a(t,\epsilon) \leq b_t - \epsilon_t \qquad \forall t \geq 1 \end{aligned}$$

$$
\begin{aligned}
& z(\epsilon) \geq \sum_{j \in \mathbb{U}} \sum_{a=0}^{A} \sum_{i=1}^{J} x_i^a(T, \epsilon) \cdot p_{i,j}^a - m \\
& z(\epsilon),\, x_j^a(t, \epsilon) \geq 0 \qquad \forall\, a \geq 0,\, j \geq 1,\, t \geq 1
\end{aligned}
$$

Note that LP (11) is separable over t and it has the same optimal solution as LP (2). Moreover, since the set of constraints in LP (2) can also be written in the same format as the set of constraints in LP (11) by simply replacing $n_j^*(t+1, \epsilon)$ with $n_j(t+1, \epsilon)$ and including $n_j(t, \epsilon)$ (for all j and t) as part of decision variables, an optimal dual solution for LP (11) is also an optimal dual solution for LP (2). This observation is important because, according to Schrijver (2000, Section 10.4), $V^D(\epsilon) - V^D(\mathbf{0})$ can be bounded by a dot product of an optimal dual solution and a vector whose elements are either 0 or ϵ_t for some t. Applying the bound in equation (24) in Schrijver (2000) to our setting, the absolute magnitude of dual solution can be bounded by a number that is at least of order order T^2 (in Schrijver's notation, it is $n \cdot \Delta \cdot ||c||_1$, and both n and $||c||_1$ are of order T in our setting). This bound, however, is too loose. To deal with this, we start with LP (11) instead of LP (2) and use the optimal dual solution of LP (11). Since LP (11) is separable over t, the absolute magnitude of dual solution corresponding to the sub-LP for period t is independent of T. As a result, we can uniformly bound the absolute magnitude of dual solution corresponding to any constraint by a constant that is independent of T, which yields the bound in Lemma 2 after applying equation (24) in Schrijver (2000). This completes the proof. ■

Proof of Lemma 3. Consider a modified RAC (call it MRAC) that proceeds in the same manner as RAC, with an exception that it ignores the budget constraint in Steps 3c and 3d (i.e., MRAC continues activating arms regardless of the given budget b_t). Let $\tilde{X}_j^a(t, \epsilon)$ denote the number of bandits in state j being applied action a at period t under MRAC, and let $\tilde{N}_j(t, \epsilon)$ denote the number of bandits in state j at the beginning of period t under MRAC. Define $\tilde{\mathcal{A}}(\epsilon) := \tilde{\mathcal{A}}_1(\epsilon) \cap \tilde{\mathcal{A}}_2(\epsilon) \cap \tilde{\mathcal{A}}_3(\epsilon)$, where

$$
\begin{aligned}
\tilde{\mathcal{A}}_1(\epsilon) &= \left\{ \tilde{X}_j^a(t, \epsilon) - x_j^{*a}(t, \epsilon) \leq \frac{\epsilon_t}{2AJ} \;\; \forall\, a \geq 1,\, j,\, t \right\}, \\
\tilde{\mathcal{A}}_2(\epsilon) &= \left\{ \tilde{X}_j^0(t, \epsilon) - x_j^{*0}(t, \epsilon) \leq \frac{\epsilon_t}{2J} \;\; \forall\, j,\, t \right\}, \quad \text{and} \\
\tilde{\mathcal{A}}_3(\epsilon) &= \left\{ \tilde{N}_j(T+1, \epsilon) - n_j^*(T+1, \epsilon) \leq \frac{\epsilon_T}{|\mathbb{U}|} \;\; \forall\, j \in \mathbb{U} \right\}.
\end{aligned}
$$

Note that $\tilde{\mathcal{A}}_1(\epsilon)$ implies $\sum_{a=1}^{A} \sum_{j=1}^{J} \tilde{X}_j^a(t, \epsilon) \leq b_t$ for all t; so, under the same random realizations, MRAC is technically equivalent to RAC on $\tilde{\mathcal{A}}(\epsilon)$. Moreover, on $\tilde{\mathcal{A}}(\epsilon)$, we also have:

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

where the first inequality holds because $(a-x)^+ - (b-x)^+ \le (a-b)^+$ for all a, b, x. We now need to compute a lower bound on $\mathbb{P}(\tilde{\mathcal{A}}(\epsilon))$. We do this by computing an upper bound for each $\mathbb{P}(\tilde{\mathcal{A}}_i(\epsilon)^c)$, $i = 1, 2, 3$. We start with $\mathbb{P}(\tilde{\mathcal{A}}_1(\epsilon)^c)$. By the sub-additive property of probability,

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

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

$$\tilde{X}_j^a(t,\epsilon) \sim \sum_{i=1}^{J} \mathrm{Bin}(n_{i,1}, v_{i,a,j}(t,\epsilon))$$

where $v_{i,a,j}(t,\epsilon)$ is the probability that 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 $v_{i,a,j}(t,\epsilon)$ in terms of $\{p_{i,j}^a\}$ and $\{q_j^{*a}(t,\epsilon)\}$, but this is not necessary for our purpose.) Note that $x_j^{*a}(t,\epsilon) = \sum_{i=1}^{J} n_{i,1} \cdot v_{i,a,j}(t,\epsilon)$. Let $\tilde{S}_{a,j}(t,\epsilon) = \{i : v_{i,a,j}(t,\epsilon) > 0\}$. Then, we can further bound $\mathbb{P}(\tilde{\mathcal{A}}_1(\epsilon)^c)$ as follows:

$$\begin{aligned}
\mathbb{P}(\tilde{\mathcal{A}}_1(\epsilon)^c) &\le \sum_{t=1}^{T} \sum_{\substack{(a,j):\, \tilde{S}_{a,j}(t,\epsilon)\neq\emptyset \\ i\in\tilde{S}_{a,j}(t,\epsilon)}} \mathbb{P}\left(\mathrm{Bin}(n_{i,1}, v_{i,a,j}(t,\epsilon)) - n_i(1)\cdot v_{i,a,j}(t,\epsilon) > \frac{\epsilon_t}{2|\tilde{S}_{a,j}(t,\epsilon)|AJ}\right) \\
&\le \sum_{t=1}^{T} \sum_{\substack{(a,j):\, \tilde{S}_{a,j}(t,\epsilon)\neq\emptyset \\ i\in\tilde{S}_{a,j}(t,\epsilon)}} \exp\left\{-\frac{\epsilon_t^2}{12\cdot|\tilde{S}_{a,j}(t,\epsilon)|^2\cdot A^2\cdot J^2\cdot n_{i,1}\cdot v_{i,a,j}(t,\epsilon)}\right\} \\
&\quad + \sum_{t=1}^{T} \sum_{\substack{(a,j):\, \tilde{S}_{a,j}(t,\epsilon)\neq\emptyset \\ i\in\tilde{S}_{a,j}(t,\epsilon)}} \exp\left\{-\frac{\epsilon_t}{6\cdot|\tilde{S}_{a,j}(t,\epsilon)|\cdot A\cdot J}\right\} \\
&\le A\cdot J^2\cdot \sum_{t=1}^{T}\left[\exp\left\{-\frac{\epsilon_t^2}{12\cdot A^2\cdot J^4\cdot\left[\sum_{\ell=1}^{J} n_{\ell,1}\right]}\right\} + \exp\left\{-\frac{\epsilon_t}{6\cdot A\cdot J^2}\right\}\right].
\end{aligned}$$

The first inequality follows since $\tilde{S}_{a,j}(t,\epsilon) = \emptyset$ implies $\mathbb{P}(\tilde{X}_j^a(t,\epsilon) - x_j^{*a}(t,\epsilon) > \frac{\epsilon_t}{2AJ}) = 0$ (because we must have $\tilde{X}_j^a(t,\epsilon) = 0$ almost surely); the second inequality follows by application of Chernoff bound for Binomial random variable (specifically, if $X \sim \mathrm{Bin}(n,p)$, then $\mathbb{P}(X - np > \delta) \le \exp\left\{-\frac{\delta^2}{3np}\right\}$ for all $\delta \in [0, np)$ and $\mathbb{P}(X - np > \delta) \le \exp\left\{-\frac{\delta}{3}\right\}$ for all $\delta \ge np$; this implies $\mathbb{P}(X - np > \delta) \le \exp\left\{-\frac{\delta^2}{3np}\right\} + \exp\left\{-\frac{\delta}{3}\right\}$ for all $\delta \ge 0$); and the last inequality follows since $|\tilde{S}_{a,j}(t,\epsilon)| \le J$ and $n_{i,1}\cdot v_{i,a,j}(t,\epsilon) \le \sum_{\ell=1}^{J} n_{\ell,1}$. Similarly, we can also bound $\mathbb{P}(\tilde{\mathcal{A}}_2(\epsilon)^c)$ and $\mathbb{P}(\tilde{\mathcal{A}}_3(\epsilon)^c)$ as follows:

$$\mathbb{P}(\tilde{\mathcal{A}}_2(\epsilon)^c) \le J^2\cdot\sum_{t=1}^{T}\left[\exp\left\{-\frac{\epsilon_t^2}{12\cdot J^4\cdot\left[\sum_{\ell=1}^{J} n_{\ell,1}\right]}\right\} + \exp\left\{-\frac{\epsilon_t}{6\cdot J^2}\right\}\right] \quad \text{and}$$

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

The last bound for $\mathbb{P}(\tilde{\mathcal{A}}_3(\epsilon)^c)$ follows because $|\mathbb{U}| \leq J$, and also by noting that $\tilde{N}_j(T+1,\epsilon)$ can be written as a sum of J independent Binomial random variables (specifically, $\tilde{N}_j(T+1,\epsilon) \sim \sum_{i=1}^J \mathrm{Bin}(n_{i,1}, r_{i,j}(t,\epsilon))$, where $r_{i,j}(t,\epsilon)$ is the probability that an arm starting with state i at the beginning of period 1 ends up with state j at the beginning of period $T+1$).

Putting everything together, we conclude that

$$\begin{aligned}
\mathbb{P}(\tilde{\mathcal{A}}(\epsilon)) &\geq 1 - \mathbb{P}(\tilde{\mathcal{A}}_1(\epsilon)^c) - \mathbb{P}(\tilde{\mathcal{A}}_2(\epsilon)^c) - \mathbb{P}(\tilde{\mathcal{A}}_3(\epsilon)^c) \\
&\geq 1 - 3 \cdot A \cdot J^2 \cdot \sum_{t=1}^T \left[\exp\left\{ -\frac{\epsilon_t^2}{12 \cdot A^2 \cdot J^4 \cdot \left[\sum_{\ell=1}^J n_{\ell,1}\right]} \right\} + \exp\left\{ -\frac{\epsilon_t}{6 \cdot A \cdot J^2} \right\} \right].
\end{aligned}$$

This completes the proof. ■

Proof of Lemma 4. The proof is similar to that of Lemma 3 (unless otherwise noted, all the notations have the same meaning as in the proof of Lemma 3). The difference lies in computing a bound for $\mathbb{P}(\tilde{\mathcal{A}}_i(\epsilon)^c)$ for $i = 1,2,3$. Note that $\tilde{X}_j^a(t,\epsilon)$ is now the sum of J independent Binomial random variables *and* a Poisson random variable (to capture new arrivals), i.e.,

$$\tilde{X}_j^a(t,\epsilon) \sim \sum_{i=1}^J \mathrm{Bin}(n_{i,1}, v_{i,a,j}(t,\epsilon)) + \mathrm{Pois}\left(\sum_{s=1}^t \sum_{i=1}^J \lambda_{i,s} \cdot \tilde{v}_{i,a,j}(s,t,\epsilon) \right)$$

where $\tilde{v}_{i,a,j}(s,t,\epsilon)$ is the probability that a new bandit arriving in state i in period s ends up with state j at the beginning of period t and being applied action a. We will use the following inequality for Poisson random variable: If $X \sim \mathrm{Pois}(\lambda)$, then $\mathbb{P}(X - \lambda > \delta) \leq \exp\{\lambda r^2 - \delta r\}$ for all $r \in [0,1]$. In fact, if $0 \leq \delta \leq 2\lambda$, then $\mathbb{P}(X - \lambda > \delta) \leq \exp\left\{-\frac{\delta^2}{4\lambda}\right\}$ (this can be proved by simply substituting $r = \frac{\delta}{2\lambda}$ in the previous bound). Now, as in the proof of Lemma 3, we can bound:

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

where the last inequality follows since $J \geq 1$ and therefore $1 + |\tilde{S}_{a,j}(t,\epsilon)| \leq 1 + J \leq 2J$.

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

$$\begin{aligned}\mathbb{P}(\tilde{\mathcal{A}}_2(\epsilon)^c) \leq J^2 \cdot \sum_{t=1}^{T} & \left[\exp\left\{-\frac{\epsilon_t^2}{48 \cdot J^4 \cdot \left[\sum_{\ell=1}^{J} n_{\ell,1}\right]}\right\} + \exp\left\{-\frac{\epsilon_t}{12 \cdot J^2}\right\}\right] \\ & + \sum_{t=1}^{T} \exp\left\{-\frac{\epsilon_t^2}{64 \cdot J^4 \cdot \left[\sum_{s=1}^{t}\sum_{i=1}^{J} \lambda_{i,s}\right]}\right\} \quad \text{and}\end{aligned}$$

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

Putting everything together, we conclude that

$$\begin{aligned}\mathbb{P}(\tilde{\mathcal{A}}(\epsilon)) &\geq 1 - \mathbb{P}(\tilde{\mathcal{A}}_1(\epsilon)^c) - \mathbb{P}(\tilde{\mathcal{A}}_2(\epsilon)^c) - \mathbb{P}(\tilde{\mathcal{A}}_3(\epsilon)^c) \\ &\geq 1 - 3 \cdot A \cdot J^2 \cdot \sum_{t=1}^{T} \left[\exp\left\{-\frac{\epsilon_t^2}{48 \cdot A^2 \cdot J^4 \cdot \left[\sum_{\ell=1}^{J} n_{\ell,1}\right]}\right\} + \exp\left\{-\frac{\epsilon_t}{12 \cdot A \cdot J^2}\right\}\right] \\ &\quad - 3 \cdot A \cdot J^2 \cdot \sum_{t=1}^{T} \exp\left\{-\frac{\epsilon_t^2}{64 \cdot A^2 \cdot J^4 \cdot \left[\sum_{s=1}^{t}\sum_{i=1}^{J} \lambda_{i,s}\right]}\right\}.\end{aligned}$$

This completes the proof. ■

Proof of Theorem 2. The proof is similar to that of Theorem 1. Let E denote the event where (8) is satisfied. By Lemma 4 and our choice of ϵ in Theorem 2, we already know that $\mathbb{P}(E)$ is at least $1 - \Theta\left(\frac{T}{\theta^d}\right)$. Now, we consider the event E^c. Unlike in the proof of Theorem 1 where we can simply bound $C_\theta^{RAC} - V_\theta^D(\epsilon)$ with a number that is of order $T \cdot \theta \cdot n_{tot}$, we now have new bandits arriving at each period. At period t, we have at most $\theta \cdot n_{tot} + \sum_{s=1}^{t}\sum_{j=1}^{J} \Lambda_{j,s}^\theta$ bandits in the system, where $\Lambda_{j,s}^\theta$ is Poisson with rate $\theta \cdot \lambda_{j,s}$. So, we can bound:

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

Note that $\sum_{t=1}^{T} t \cdot \left(\sum_{j=1}^{J} \Lambda_{j,t}^\theta\right)$ is stochastically dominated by $X \sim \text{Poisson}(T^2 \cdot \theta \cdot \lambda_{tot})$. Applying this fact together with Cauchy-Schwarz inequality yields:

$$\mathbb{E}\left[(C_\theta^{RAC} - V_\theta^D(\epsilon)) \cdot \mathbf{1}\{E^c\}\right] \leq \mathbb{E}\left[(T \cdot \theta \cdot n_{tot} + X)^2\right]^{1/2} \cdot \mathbb{P}(E^c)^{1/2} = O\left(\frac{T^{5/2}}{\theta^{d/2-1}}\right).$$

Putting the above bound together with $\mathbb{E}\left[(C_\theta^{RAC} - V_\theta^D(\epsilon)) \cdot \mathbf{1}\{E\}\right] = O(T^{3/2} \cdot \sqrt{d \cdot \theta \cdot \log \theta})$ (by Lemma 4, our choice of ϵ, and our previous discussions about $\mathbb{P}(E)$) and the fact that $V_\theta^D(\mathbf{0}) = \Omega(T \cdot \theta)$ (because $S(t) \neq \emptyset$ for all t) gives the desired result. ■

Parameters used in numerical experiments.

References

Bertsimas, D., and Nio-Mora, J. 2000. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. Operations Research, 48(1): 80-90.

Brandeau, M. L. 2005. Allocating resources to control infectious diseases. In Operations Research and Health Care, 443-464. Springer US.

Cohen, K., Zhao, Q., and Scaglione, A. 2014, November. Restless multi-armed bandits under time-varying activation constraints for dynamic spectrum access. In Signals, Systems and Computers, 2014 48th Asilomar Conference on: 1575-1578. IEEE.

Deo, S., Iravani, S., Jiang, T., Smilowitz, K., and Samuelson, S. 2013. Improving health outcomes through better capacity allocation in a community-based chronic care model. Operations Research, 61(6): 1277-1294.

Gittins, J. C. 1979. Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society. Series B (Methodological): 148-177.

Gittins, J., Glazebrook, K., and Weber, R. 2011. Multi-armed bandit allocation indices. John Wiley and Sons.

Larranaga, M., Ayesta, U., and Verloop, I. M. 2015. Asymptotically optimal index policies for an abandonment queue with convex holding cost. Queueing Systems, 81(2-3): 99-169.

Ny, J. L., Dahleh, M., and Feron, E. 2008. A Linear Programming Relaxation and a Heuristic for the Restless Bandit Problem with General Switching Costs. arXiv preprint arXiv:0805.1563.

Schrijver, Alexander. 2000. Theory of Linear and Integer Programming. John Wiley & Sons.

Weber, R. R., and Weiss, G. 1990. On an index policy for restless bandits. Journal of Applied Probability, 27(03): 637-648.

Weiss, G. 1988. Branching bandit processes. Probability in the Engineering and Informational Sciences, 2(03): 269-278.

Whittle, P. 1980. Multi-armed bandits and the Gittins index. Journal of the Royal Statistical Society. Series B (Methodological): 143-149.

Whittle, P. (1981). Arm-acquiring bandits. The Annals of Probability, 9(2): 284-292.

Whittle, P. 1988. Restless bandits: Activity allocation in a changing world. Journal of applied probability, 25(A): 287-298.

Table 3 Parameters Used in Numerical Experiments

Parameter (Notation)	Experiment 1	Experiment 2
Number of states ($\lvert\mathbb{J}\rvert$)	2	5
Undesirable state ($\lvert\mathbb{U}\rvert$)	1	2
Number of arms ($n_{j\in J,1}$)	$\begin{bmatrix}2 & 1\end{bmatrix}$	$\begin{bmatrix}2 & 2 & 1 & 1 & 1\end{bmatrix}$
Number of actions ($\lvert\mathbb{A}\rvert$)	2	5
Penalty threshold (m)	1	2
Penalty cost (ϕ)	20	20
Cost matrix ($c^{a\in A}_{j\in J}$)	$\begin{bmatrix}6 & 4\\ 9 & 5\end{bmatrix}$	$\begin{bmatrix}0.4 & 1.5 & 0.3 & 0.2 & 0.1\\ 2.0 & 2.5 & 0.4 & 1.3 & 0.2\\ 5.5 & 5.0 & 2.5 & 0.4 & 0.3\\ 6.0 & 7.0 & 1.7 & 3.5 & 1.4\\ 9.5 & 7.5 & 1.8 & 1.6 & 4.5\end{bmatrix}$
Transition matrix ($P^{a\in A}_{j\in J,k\in J}$)		
$P^0_{j\in J,k\in J} =$	$\begin{bmatrix}0.9 & 0.1\\ 0.3 & 0.7\end{bmatrix}$	$\begin{bmatrix}0.50 & 0.45 & 0.02 & 0.02 & 0.01\\ 0.45 & 0.48 & 0.04 & 0.02 & 0.01\\ 0.43 & 0.45 & 0.09 & 0.02 & 0.01\\ 0.40 & 0.43 & 0.11 & 0.04 & 0.02\\ 0.38 & 0.40 & 0.10 & 0.05 & 0.07\end{bmatrix}$
$P^1_{j\in J,k\in J} =$	$\begin{bmatrix}0.5 & 0.5\\ 0.2 & 0.8\end{bmatrix}$	$\begin{bmatrix}0.45 & 0.40 & 0.07 & 0.05 & 0.03\\ 0.40 & 0.43 & 0.09 & 0.05 & 0.03\\ 0.38 & 0.40 & 0.14 & 0.05 & 0.03\\ 0.35 & 0.38 & 0.16 & 0.07 & 0.04\\ 0.33 & 0.35 & 0.15 & 0.08 & 0.09\end{bmatrix}$
$P^2_{j\in J,k\in J} =$	$\begin{bmatrix}0.4 & 0.6\\ 0.1 & 0.9\end{bmatrix}$	$\begin{bmatrix}0.40 & 0.35 & 0.12 & 0.08 & 0.05\\ 0.35 & 0.38 & 0.14 & 0.08 & 0.05\\ 0.33 & 0.35 & 0.19 & 0.08 & 0.05\\ 0.30 & 0.33 & 0.21 & 0.10 & 0.06\\ 0.28 & 0.30 & 0.20 & 0.11 & 0.11\end{bmatrix}$
$P^3_{j\in J,k\in J} =$		$\begin{bmatrix}0.35 & 0.30 & 0.17 & 0.11 & 0.07\\ 0.30 & 0.33 & 0.19 & 0.11 & 0.07\\ 0.28 & 0.30 & 0.24 & 0.11 & 0.07\\ 0.25 & 0.28 & 0.26 & 0.13 & 0.08\\ 0.23 & 0.25 & 0.25 & 0.14 & 0.13\end{bmatrix}$
$P^4_{j\in J,k\in J} =$		$\begin{bmatrix}0.30 & 0.25 & 0.22 & 0.14 & 0.09\\ 0.25 & 0.28 & 0.24 & 0.14 & 0.09\\ 0.23 & 0.25 & 0.29 & 0.14 & 0.09\\ 0.20 & 0.23 & 0.31 & 0.16 & 0.10\\ 0.18 & 0.20 & 0.30 & 0.17 & 0.15\end{bmatrix}$
$P^5_{j\in J,k\in J} =$		$\begin{bmatrix}0.25 & 0.20 & 0.27 & 0.17 & 0.11\\ 0.20 & 0.23 & 0.29 & 0.17 & 0.11\\ 0.18 & 0.20 & 0.34 & 0.17 & 0.11\\ 0.15 & 0.18 & 0.36 & 0.19 & 0.12\\ 0.13 & 0.15 & 0.35 & 0.20 & 0.17\end{bmatrix}$