Finite-Horizon Single-Pull Restless Bandits: An Efficient Index Policy For Scarce Resource Allocation

Guojun Xiong
Harvard University
Cambridge, United States
gjxiong@g.harvard.edu

Haichuan Wang
Harvard University
Cambridge, United States
haichuan_wang@g.harvard.edu

Yuqi Pan
Harvard University
Cambridge, United States
yuqipan@g.harvard.edu

Saptarshi Mandal
University of Illinois, Urbana-Champaign
Champaign, United States
smandal4@illinois.edu

Sanket Shah
Havard University
Cambridge, United States
sanketshah@g.harvard.edu

Niclas Boehmer
Hasso Plattner Institute
Potsdam, Germany
niclas.boehmer@hpi.de

Milind Tambe
Havard University
Cambridge, United States
tambe@g.harvard.edu

ABSTRACT

Restless multi-armed bandits (RMABs) have been highly successful in optimizing sequential resource allocation across many domains. However, in many practical settings with highly scarce resources, where each agent can only receive at most one resource, such as healthcare intervention programs, the standard RMAB framework falls short. To tackle such scenarios, we introduce Finite-Horizon Single-Pull RMABs (SPRMABs), a novel variant in which each arm can only be pulled once. This single-pull constraint introduces additional complexity, rendering many existing RMAB solutions suboptimal or ineffective. To address this shortcoming, we propose using *dummy states* that expand the system and enforce the one-pull constraint. We then design a lightweight index policy for this expanded system. For the first time, we demonstrate that our index policy achieves a sub-linearly decaying average optimality gap of $\tilde{O}\left(\frac{1}{\rho^{1/2}}\right)$ for a finite number of arms, where ρ is the scaling factor for each arm cluster. Extensive simulations validate the proposed method, showing robust performance across various domains compared to existing benchmarks.

KEYWORDS

Restless multi-armed bandits; scare resource; single-pull constraint; index policy; optimality analysis

ACM Reference Format:
Guojun Xiong, Haichuan Wang, Yuqi Pan, Saptarshi Mandal, Sanket Shah, Niclas Boehmer, and Milind Tambe. 2025. Finite-Horizon Single-Pull Restless Bandits: An Efficient Index Policy For Scarce Resource Allocation. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

1 INTRODUCTION

The restless multi-armed bandits (RMAB) problem [32] is a time slotted game between a decision maker (DM) and the environment. In the standard RMAB model, each “restless” arm is described by a Markov decision process (MDP) [26], and evolves stochastically according to two different transition functions, depending on whether the arm is activated or not. Scalar rewards are generated with each transition. The goal of the DM is to maximize the total expected reward under an instantaneous constraint that at most K out of N arms can be activated at any decision epoch. RMABs have been widely used to model a variety of real-world applications such as problems around congestion control [2], job scheduling [37], wireless communication [5], healthcare [16, 23], queueing systems [19], and cloud computing [36]. One key reason for the popularity of RMABs is their ability to optimize sequential allocation of limited resources to a population of agents in uncertain environments [22].

However, in many real-world scenarios, additional constraints are placed on the allocation. In this paper, we propose and study the new problem of sequentially allocating resources when each agent can only receive a resource in at most one timestep, i.e., we focus on the RMAB problem where no arm can be pulled repeatedly. This constraint is, for instance, prevalent in real-world domains where resources are extremely scarce and there are many more agents than resources, occurring for instance in healthcare, conservation, and machine maintenance. Even in cases where the number of resources and agents are of the same magnitude, organizers might impose single-pull constraints for fairness reasons to ensure an equal treatment of all agents. Lastly, there are also allocation scenarios where an agent only benefits from the first resource assigned to them, for instance, when distributing single-dose vaccines.

Concrete examples where the single-pull constraint is imposed in practice arise in public health, where RMABs are (or can be) used to optimize the allocation of health intervention resources [7, 20, 29]. We present more detailed examples of deployed and emerging applications from healthcare and other domains with the presence of such a single intervention constraint in Section 2. These practical challenges necessitate the development of a new model capable of addressing the unique constraints posed by single-pull scenarios, ensuring efficient allocation of limited resources.

Given the new and urgent requirement of a single pull per arm in many practical domains, we introduce the Finite-Horizon Single-Pull Restless Multi-Armed Bandits (SPRMABs), a novel variant of the RMAB framework where each arm can be pulled at most once. As widely known, the complexity of conventional RMAB lies in the challenge of finding an optimal control strategy to maximize the expected total reward, a problem that is typically intractable [25]. As a result, existing approaches have largely focused on designing efficient heuristic (and at times asymptotically optimal) index-based policies, such as those developed for offline RMABs [5, 18, 28, 32, 38] and reinforcement learning (RL) algorithms for online RMABs [3, 9, 10, 15, 24, 27, 30, 33, 35]. However, the introduction of the single-pull constraint in SPRMABs renders these traditional methods either suboptimal or inapplicable.

Tailoring existing methods, such as the Whittle index policy [32] and fluid linear programming (LP)-based policies [8, 11, 28, 38], to the novel SPRMABs presents significant challenges. Specifically, the Whittle index policy [32] is defined using Lagrange multipliers for activation budget constraints. Introducing the single-pull constraint disrupts this framework, as the Lagrangian formulation becomes ill-defined, causing the method to return highly suboptimal solutions. Similarly, for fluid LP-based methods [8, 11, 28, 38], enforcing the single-pull constraint introduces a new nonlinear constraint, which exponentially increases the complexity as the time horizon and the number of arms grows. The question we tackle in this paper is the following:

> *Is it possible to design a light-weight asymptotically optimal index policy for* SPRMABs*?*

To tackle this challenge, we utilize *dummy states* to duplicate the entire system, ensuring that once an arm is activated, it transitions exclusively between these dummy states. The transitions within the dummy states mirror those of the normal states when no action is taken (i.e., action 0). *Building on this expanded system, we design a lightweight index policy specifically tailored for SPRMABs*, and we demonstrate that our proposed index policy achieves a linear decaying rate in the average optimality gap. Our main contributions can be summarized as follows:

• **Lightweight Index Policy Design:** We leverage the concept of expanding the system through dummy states and develop a lightweight index policy, called single-pull index (SPI) policy, which addresses two challenges that conventional index policies cannot handle. First, in real-world applications, pulling an arm doesn't always guarantee better outcomes, meaning the full budget may not need to be used. Existing index policies often exhaust the budget on the highest indices, leading to suboptimal results. Second, in SPRMABs, each arm can only be pulled once, making activation timing crucial. An arm with a high index now may yield a better reward if pulled later, which current algorithms fail to handle. Dummy states allow deferring decisions without affecting future rewards, conserving resources and tackling both challenges effectively.

• **Optimality Gap:** For the first time, we demonstrate that our proposed index policy achieves a sub-linearly decaying rate of the average optimality gap for a finite number of arms, characterized by the bound $\tilde{O}(\frac{1}{\rho^{1/2}} + \frac{1}{\rho^{3/2}})$, where ρ denotes the scaling factor for each arm cluster.

• **Empirical Simulations:** We conduct extensive simulations to validate the effectiveness of the proposed method, benchmarking it against existing strategies. The results consistently demonstrate robust performance across a variety of domain settings, underscoring the practicality and versatility of our index policy in addressing SPRMABs. This advancement not only enhances the applicability of SPRMABs in equitable resource allocation but also lays a strong foundation for future research in constrained bandit settings.

2 MOTIVATING DOMAINS AND EXAMPLES

The single-pull constraint in RMABs is motivated by multiple real-world domains. We begin by describing examples from public health domains with limited resources [4, 20]. One concrete deployed example RMABs used for a maternal mHealth (mobile health) program in India [22, 29]. This deployment supports an mHealth program of ARMMAN (armman.org), an India-based non-profit that spreads preventative care awareness to pregnant women and new mothers through an automated call service. To reduce dropoffs from the mHealth program, ARMMAN employs health workers to provide live service calls to beneficiaries; however, ARMMAN is faced with a resource allocation challenge because any one time, there are 200K beneficiaries (mothers) enrolled in the program but they have enough staff to only do 1000 live source calls per week. As a result, RMABs are deployed to optimize allocation of their limited live service calls[29], and given the scale of the program, a beneficiary received a maximum of one service call. That is, each RMAB arm represents a mother, and once an arm is pulled, i.e., the mother receives a service call, she does not receive a service call again.

Similarly, in maternal health programs in Uganda [7], RMABs are proposed to be used to allocate scarce wireless vital sign monitors to mothers in maternity wards, where each mother may receive such a monitor only once during her stay. A similar scenario occurs in support programs which can only support a limited number of beneficiaries every week and beneficiaries can only participate in the program once. One such example is malnutrition prevention[17], where a child may be enrolled in a malnutrition program only once. These practical challenges necessitate the development of a new model capable of addressing the unique constraints posed by single-pull scenarios, ensuring an efficient allocation of limited resources.

3 SYSTEM MODEL AND PROBLEM FORMULATION

Consider a finite-horizon RMAB problem with N arms. Each arm n is associated with a specific Unichain Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P_n, r_n, \mathbf{s}_1, T)$, where $\mathcal{S}$ is the finite state space and $\mathcal{A} := \{0, 1\}$ denotes the binary action set. Using the standard terminology from the RMAB literature, we call an arm *passive* when action $a = 0$ is applied to it, and *active* otherwise. $P_n : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ is the transition kernel and $r_n : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function. The total number of activated arms at each time t is constrained by

K, which we call the *activation budget.* The initial state is chosen according to the initial distribution $\mathbf{s}_1$ and $T < \infty$ is the horizon.

At time $t \in [T]$, each arm n is at a specific state $s_n(t) \in \mathcal{S}$ and evolves to $s_n(t+1)$ independently as a controlled Markov process with the controlled transition probabilities $P_n(s_n(t), a_n(t), s_n(t+1))$ when action $a_n(t)$ is taken. The immediate reward earned from activating arm n at time t is denoted by $r_n(t) := r_n(s_n(t), a_n(t))$. Denote the total reward earned at time t by $R(t)$, i.e., $R(t) := \sum_n r_n(t)$. Motivated by the healthcare implementations where each arm (i.e., patient) can only be pulled for once due to resource limitation, now let us consider the scenario where each arm can only be pulled once, and the duration of activation is also one. This is equivalent to the constraint in the following expression

$$\texttt{Single-pull constraint}: \sum_{t=1}^{T} a_n(t) \leq 1, \forall n. \tag{1}$$

Let $\mathcal{F}_t$ denote the operational history until t, i.e., the σ-algebra generated by the random variables $\{s_n(\ell) : n \in [N], \ell \in [t]\}, \{a_n(\ell) : n \in [N], \ell \in [t-1]\}$. Our goal is to derive a policy $\pi : \mathcal{F}_t \mapsto \mathcal{A}^N$ that makes decisions regarding which set of arms are made active at each time $t \in [T]$ so as to maximize the expected value of the cumulative rewards subject to the activation budget and the one-pull constraint in (1), i.e.,

$$\begin{aligned} \texttt{SPRMAB}: \max_{\pi} \mathbb{E}_\pi \left(\sum_{n=1}^{N} \sum_{t=1}^{T} r_n(t) \right) \\ \text{s.t.} \sum_{n=1}^{N} a_n(t) \leq K, \forall t \in [T], \ \sum_{t=1}^{T} a_n(t) \leq 1, \forall n. \end{aligned} \tag{2}$$

where the subscript indicates that the expectation is taken with respect to the measure induced by the policy π. We refer to the problem (2) as the "original problem", which suffers from the "curse of dimensionality", and hence is computationally intractable. We overcome this difficulty by developing a computationally feasible and provably optimal index-based policy.

3.1 Existing Index Policies and Failure Examples

The challenge comes from the "hard" constraints in (2), where the first `budget constraint` must be satisfied at every time step, and the second `single-pull constraint` must be satisfied firmly for all arms. Existing index policy approaches [28, 32, 38] for conventional RMAB problems without the `single-pull` constraint design indices by relaxing the "hard" `activation-budget constraint` $\sum_{n=1}^{N} a_n(t) \leq K, \forall t \in [T]$ to the "relaxed" constraints, i.e., the activation cost at time $t \in [T]$ is limited by K in expectation, which is

$$\texttt{Re-budget constraint}: \mathbb{E}_\pi \left\{ \sum_{n=1}^{N} a_n(t) \right\} \leq K. \tag{3}$$

In the following, we present two typical index polices, one is the Whittle index policy [32], and the other is the LP-based index policy [28, 34, 38].

Whittle Index Policy. Whittle index [32] is designed upon the infinite-horizon average-reward (IHAR) RMAB settings through decomposition. Specifically, Whittle relies on the `Relaxed budget constraint` in (3) and obtains a unconstrained problem for IHAR settings:

$$\texttt{IHAR-RMAB}: \max_{\pi \in \Pi} \liminf_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \sum_{t=1}^{T} \sum_{n=1}^{N} \{ r_n(t) + \lambda (1 - a_n(t)) \},$$

where λ is the Lagrangian multiplier associated with the constraint. The key observation of Whittle is that this problem can be decomposed and its solution is obtained by combining solutions of N independent problems via solving the associated dynamic programming (DP) [33]: $V_n(s) = \max_{a \in \{0,1\}} Q_n(s,a), \forall n \in \mathcal{N}$, where

$$\begin{aligned} Q_n(s,a) + \beta = a \Big(r_n(s,a) + \sum_{s'} p_n(s'|s,1) V_n(s') \Big) \\ + (1-a) \Big(r_n(s,a) + \lambda + \sum_{s'} p_n(s'|s,0) V_n(s') \Big), \end{aligned} \tag{4}$$

where β is unique and equals to the maximal long-term average reward of the unichain MDP, and $V_n(s)$ is unique up to an additive constant, both of which depend on the Lagrangian multiplier λ. The optimal decision a^* in state s then is the one which maximizes the right hand side of the above DP. The Whittle index associated with state s is defined as the value $\lambda_n^*(s) \in \mathbb{R}$ such that actions 0 and 1 are equally favorable in state s for arm n [3, 10], satisfying

$$\begin{aligned} \lambda_n^*(s) := r_n(s,1) + \sum_{s'} p_n(s'|s,1) V_n(s') \\ - r_n(s,0) - \sum_{s'} p_n(s'|s,0) V_n(s'). \end{aligned} \tag{5}$$

Whittle index policy then activates K arms with the largest Whittle indices at each time slot t.

LP-based Index Policy. With the `Relaxed budget constraint` in (3), we can transfer the conventional RMAB problem into an equivalent LP [1] by leveraging the definition of occupancy measure (OM). In particular, the OM μ of a policy π in a finite-horizon MDP is defined as the expected number of visits to a state-action pair (s,a) at each time t, i.e.,

$$\mu = \{ \mu_n(s,a;t) = \mathbb{P}(s_n(t) = s, a_n(t) = a) : \forall n, t | 0 \leq \mu_n(s,a;t) \leq 1 \},$$

which is a probability measure, satisfing $\sum_{s,a} \mu_n(s,a,t) = 1, \forall t \in [T]$. Hence, the associated LP is expressed as

$$\max_{\mu} \sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{(s,a)} \mu_n(s,a;t) r_n(s,a) \tag{6}$$

$$\text{s.t.} \sum_{n=1}^{N} \sum_{s} \mu_n(s,1;t) \leq K, \ \forall t, // \textit{activation constraint} \tag{7}$$

$$\sum_{a} \mu_n(s,a;t) = \sum_{(s',a')} \mu_n(s',a';t-1) P_n(s',a',s), \forall n, s, \quad // \textit{fluid equilibrium} \tag{8}$$

$$\sum_{a} \mu_n(s,a;1) = \mathbf{s}_1(s), \ \forall s, n, // \textit{initial distribution} \tag{9}$$

where (7) is a restatement of the `budget constraint` in (2) for $\forall t \in [T]$, which indicates the activation budget; (8) represents the transition of the occupancy measure from time $t-1$ to time t, $\forall n \in [N]$ and $\forall t \in [T]$; and (9) indicates the initial condition for occupancy measure at time 1, $\forall s \in \mathcal{S}$. Denote the solution to the above LP as $\mu^\star = \{\mu_n^\star(s,a;t) : n \in [N], t \in [T]\}$. A simple index-based policy according to the optimal solution $\mu^\star$ can be designed by dividing the arms at each time slot t into three categories:

(1) High-priority states: $\mu_n^\star(s,0;t) = 0$. (*Pull arms under those states.*)
(2) Medium-priority states: $\mu_n^\star(s,1;t) > 0$, $\mu_n^\star(s,0;t) > 0$.(*Pull arms under those states when remaining budget is available.*)
(3) Low-priority states: $\mu_n^\star(s,1;t) = 0$. (*Do not pull arms under those states.*)

In SPRMAB as shown in (2) where each arm can be activated at most once, the standard Whittle and LP-based index policies may become suboptimal. This is because the these index policies are designed under the assumption that arms can be activated multiple times, and it may not adequately account for the urgency of activating certain arms in a single-pull setting. Below, we present a rigorous example demonstrating how the Whittle and LP-based index policies can fail under these constraints (see Figure 1).

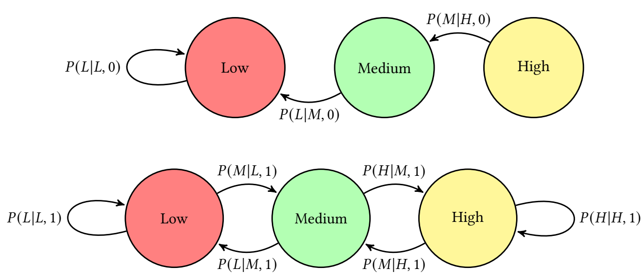

Figure 1: General transition kernels with $a = 0$ in above and $a = 1$ in below for a patient in CPAP example.

Example 1. ***Continuous Positive Airway Pressure Therapy (CPAP).*** *The CPAP [13, 21, 30] is a highly effective treatment when it is used consistently during sleeping for adults with obstructive sleep apnea. Since non-adherence to CPAP in patients hinders the effectiveness, we adapt the Markov model of CPAP adherence behavior to a three-state system with the clinical adherence criteria. To elaborate, three distinct states are defined to characterize adherence levels: Low (L), Medium (M), and High (H) as shown in Figure 1. Generally speaking, when action $a = 0$ is taken, i.e., no intervention, the patient has a probability of* 1 *to move from a higher adherence level to a lower adherence level. While intervention is available, a patient can either transit to a lower adherence lever or a higher adherence level with certain probabilities. In standard CPAP, the reward is set as the* 1 *for state "low adherence",* 2 *for state "medium adherence", and* 3 *for state "high adherence".*

Proposition 1. *The MDP for each patient defined in Example 1 is indexable.*

Proposition 1 indicates that the Whittle index can be employed for the constructed CPAP problem in Example 1. To verify that the Whittle index policy [32] and LP-based policy [34, 38] fail in this example, we construct the following setting. We randomly generate 20 different arms and each arm is duplicated 10 times, whose transition probability matrices are generated randomly. The budget is set to $K = 10$. The objective is to maximize the total adherence level in a finite horizon $T = 10$. More importantly, each arm can only be pulled at most once.

Figure 2 highlights the performance limitations of existing policy strategies—specifically the LP-based method and the Whittle index policy—when `single-pull constraint` presents. It shows the normalized rewards where the optimal policy[1], used as a benchmark, achieves a score of 1.0. The LP-based policy attains 0.76, and the Whittle index policy only 0.46 of the optimal performance. These results underline the ineffectiveness of both the LP-based method and the Whittle index policy in adequately handling the `single-pull constraint`, as neither approach reaches the efficiency of the optimal policy, particularly with the Whittle index-based approach performing less than half as well. This comparison suggests that these methods require modifications or alternative strategies to improve their adaptability and effectiveness under the strict limitations imposed by the `single-pull constraint`.

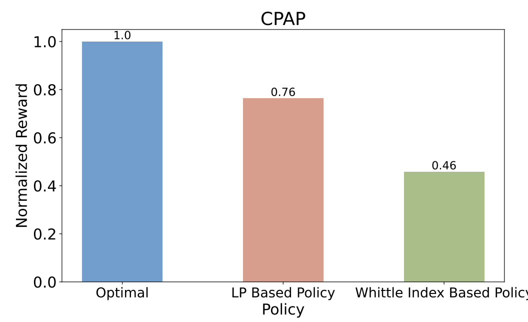

Figure 2: A CPAP setting with 3 **different states,** 20 **different types of arms, each type has** 10 **arms, the budget is set to be** 10 **and the time Horizon is** 10**.**

An intuitive explanation for why existing index policies fail in the single-pull setting is twofold. First, these indices are designed without accounting for the single-pull constraint. Second, traditional index policies pull arms from highest to lowest index until the activation budget is exhausted. However, in the SPRMAB setting, pulling the arm with the highest index at the current time may not lead to better results, as waiting for a future time slot could yield a higher reward. This makes the traditional approach ineffective in such scenarios.

3.2 Challenge for Extending Existing Methods

The limitations of existing index policies in addressing the `single-pull constraint` in (1) become evident in the context of SPRMAB settings. Traditional policies such as the Whittle index fail to accommodate this constraint effectively because the additional dimensional constraint inherent in the single-pull scenario disrupts the foundational principles underpinning the Whittle index's definition, rendering it inapplicable. Consequently, attention shifts to the LP-based index policy. This focus is due to the adaptability of LP approaches, which may allow for the integration of the `single-pull constraint` through modifications to the existing framework. This

[1]Though we usually do not know the performance achieved by optimal policy, we can leverage the optimal value achieved for the LP in (16) to serve as the optimal performance, as it is always an upper bound of the optimal performance. This will be explained in detail in Section 4.

approach requires re-evaluating the LP formulation to ensure it captures the critical aspects of decision-making under the stringent limitations imposed by the single-pull constraint.

Similar to relaxing the "hard" `budget constraint`, we can also relax the `single-pull constraint` in (1) so that the total number of pulls per arm is limited by 1 only in expectation as

$$\texttt{Re-single-pull constraint}:\mathbb{E}_\pi\left\{\sum_{t=1}^{T}a_n(t)\right\}\leq 1,\forall n. \quad (10)$$

Hence, we have the relaxed problem of (2) expressed as

$$\texttt{Re-SPRMAB}:\max_\pi \mathbb{E}_\pi\left(\sum_{n=1}^{N}\sum_{t=1}^{T}r_n(t)\right)$$
$$\text{s.t.}\mathbb{E}_\pi\left\{\sum_{n=1}^{N}a_n(t)\right\}\leq K,\ \mathbb{E}_\pi\left\{\sum_{t=1}^{T}a_n(t)\right\}\leq 1,\forall n. \quad (11)$$

According to the definition of OM μ, the `Re-SPRMAB` in (11) can be reformulated as the following LP [1]:

$$\texttt{SPRMAB-LP}:\ \max_\mu \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{(s,a)}\mu_n(s,a;t)r_n(s,a)$$
$$\text{s.t. } \textit{Constriants } (7)-(9), \sum_{t=1}^{T}\sum_{s}\mu_n(s,1;t)\leq 1,\forall n. \quad (12)$$

It is clear that the `SPRAMB-LP` in (12) achieves an upper bound of the optimal value of `SPRAMB` in (2), which is shown as the following proposition.

Proposition 2. *The optimal value achieved by* `SPRMAB-LP` *in* (12) *is an upper bound of that of* `SPRMAB` *in* (2).

Proof Sketch. Since the `SPRMAB-LP` in (12) is equivalent to the relaxed problem `Re-SPRMAB` in (11) [1], it is sufficient to show that `Re-SPRMAB` in (11) achieves no less average reward than the original problem `SPRMAB` in (2). The proof is straightforward since the constraints in the relaxed problem expand the feasible region of `SPRMAB` in (2). □

One drawback of `SPRMAB-LP` in (12) is that the mapping of `single-pull constraint` from (2) to the one in (12) will make the "hard" constraint relaxed to a significant extent, as the probability of $\mu_n(s,1;t)$ will diffuse to different time steps, which contradicts with the real scenario where arms only be pulled for one particular time slot. This will make the associated index policy designed upon the solution of `SPRMAB-LP` be significantly suboptimal. To make the relaxed problem tighter, we need to add the following constraint: for arbitrary time slot t, if arm n is being activated, the arm should never be activated in other time slots t' with $t'\neq t$, which can be mapped as

$$\sum_s\mu_n(s,1;t)\cdot\sum_s\mu_n(s,1;t')=0. \quad (13)$$

Incorporating the additional constraint in (13) into `SPRMAB-LP` in (12), we have the following optimization problem:

$$\max_\mu \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{(s,a)}\mu_n(s,a;t)r_n(s,a)$$
$$\text{s.t. } \textit{Constraints } (7)-(9),\ (12)$$
$$\sum_s\mu_n(s,1;t)\cdot\sum_s\mu_n(s,1;t')=0. \quad (14)$$

Regarding the novel optimization problem in (14), we have the following proposition.

Proposition 3. *The optimal value achieved by* (14) *lies in the middle of that by* `SPRMAB-LP` *in* (12) *and that of* (2).

Remark 1. *The proof of Proposition 3 follows a similar argument as that in Proposition 2. It indicates that the modified problem in* (14) *achieves a tighter upper bound compared with the* `SPRMAB-LP` *in* (12), *and thus the associated index policy designed upon the solution of* (14) *performs better than that derived from* (12). *However, the above formulation in* (14) *is not an LP any longer due to the last constraint, which leads to outstanding challenge in solving the above revised optimization problem when the state space* $|\mathcal{S}|$ *and time horizon* T *are large.*

4 PROPOSED METHOD

To address the challenge of solving the problem in (14) posed by the nonlinear constraint in (13) (as indicated in Remark 1), we propose a novel method to handle the `single-pull constraint`. This method involves modifying the underlying Markov Decision Processes (MDPs) associated with the arms by introducing the concept of *dummy states*.

In the considered `SPRMABs`, for arbitrary state of each arm $s\in\mathcal{S}$ at current time step t, it transitions to next state $s'\in\mathcal{S}$ at time $t+1$ if a pull is assigned to this arm. After reaching state s' at time $t+1$, the arm will never be pulled again. For every pulled arm, regardless of its state or current time step, the available action set thereafter is restricted to $\{0\}$. Building on the aforementioned observation, we introduce *dummy states* to represent the states reached immediately after an arm is pulled. We enforce that these dummy states have the same transition kernels and reward functions under both actions 0 and 1, identical to those of their corresponding normal states. The formal definition is given as follows.

Definition 1 (Dummy state). *A dummy state* s_d *represents the state* $s\in\mathcal{S}$ *that being transited immediately when an arm is pulled. For a dummy state* s_d *we have the following properties:*

$$P_n(s_d,0,s_d')=P_n(s_d,1,s_d')=P_n(s,0,s'),$$
$$r_n(s_d,0)=r_n(s_d,1)=r_n(s,0),\forall n, \quad (15)$$

i.e., actions 0 *and* 1 *are indifferent in dummy states for all arms.*

Remark 2. *For every state* $s\in\mathcal{S}$*, it has a corresponding dummy state* s_d*. When introducing the dummy states, we duplicate the original state space* $\mathcal{S}$ *and define the dummy state space as* $\mathcal{S}_d$*. As a result, the system now has a new expanded state space* $\mathcal{S}':=\mathcal{S}\bigcup\mathcal{S}_d$*. The intuitive idea behind introducing dummy states and enforcing indifference between actions 0 and 1 for these states is to ensure that the resource budget flows toward arms in non-dummy states, as arms in dummy states yield no gain even if resources are allocated to them. Another key advantage of using dummy states is that they allow us to eliminate the nonlinear constraint in* (13)*. These points will be discussed in more detail in subsequent sections.*

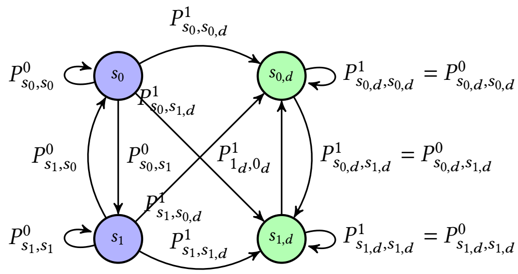

Figure 3: A toy example of SPRAMB with dummy states. The original state space is $\mathcal{S} = \{s_0, s_1\}$, and it leads to a 4-state expanded system as $\mathcal{S}' = \{s_0, s_1, s_{0,d}, s_{1,d}\}$.

To better understand how dummy states work, we present the following toy example.

Example 2. *Consider a setting where the original state space is $\mathcal{S} = \{s_0, s_1\}$. We introduce two corresponding dummy states, $s_{0,d}$ and $s_{1,d}$, which are absorbing states. As a result, the expanded state space becomes $\mathcal{S}' := \{s_0, s_1, s_{0,d}, s_{1,d}\}$. In this setup, once an arm transitions into a dummy state (either $s_{0,d}$ or $s_{1,d}$), it remains in dummy states indefinitely, regardless of the action taken. Hence, both $s_{0,d}$ and $s_{1,d}$ are absorbing states, meaning that no matter which action is chosen (either action 0 or action 1), the transition probabilities from these dummy states remain the same. The transitions between these states are illustrated in Figure 3. Thus, arms in these dummy states provide no additional reward even for positive action assignment, ensuring that resources are directed toward arms in non-dummy states, which can still benefit from positive action assignments.*

Once we provide the new state space $\mathcal{S}'$ containing dummy states and the new transition kernels, we have a new formulation as follows:

$$\begin{aligned}
\max_{\mu} \; & \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}'}\sum_{a\in\mathcal{A}} \mu_n(s,a;t) r_n(s,a) \\
\text{s.t.} \; & \sum_{n=1}^{N}\sum_{s\in\mathcal{S}'} \mu_n(s,1;t) \le K, \ \forall t, \\
& \sum_{a\in\mathcal{A}} \mu_n(s,a;t) = \sum_{s'\in\mathcal{S}'}\sum_{a'\in\mathcal{A}} \mu_n(s',a';t-1) P_n(s',a',s), \forall n, s\in\mathcal{S}', \\
& \sum_{a} \mu_n(s,a;1) = \mathbf{s}_1(s), \ \forall s \in \mathcal{S}, n. \qquad (16)
\end{aligned}$$

Remark 3. *Given the expanded state space $\mathcal{S}'$ with dummy states and the modified transition kernels, we can remove both the* `single-pull constraint` *in* (12) *and the nonlinear constraint in* (13). *The reason we can eliminate the* `single-pull constraint` *is that once an arm is pulled, it transitions to a dummy state, where there is no difference between action 1 and action 0. As a result, when arms in non-dummy states can benefit from a positive action, the active OM naturally flows to those arms. Similarly, the nonlinear constraint to concentrate the OM in* (13) *is no longer required, as it becomes irrelevant without the* `single-pull constraint`.

4.1 The Single-Pull Index Policy

Once we solve (16) and get the optimal $\mu^\star$, we can define the following Markovian policy[2]

$$\chi_n^\star(s,1;t) := \frac{\mu_n^\star(s,1;t)}{\mu_n^\star(s,0;t) + \mu_n^\star(s,1;t)} \in [0,1], \qquad (17)$$

which denotes the probability of selecting action 1 for arm n with state s at time t. Note that the optimal policy (17) is not always feasible for SPRMAB since in the latter at most K units of activation costs can be consumed at a time and the arm can only be pulled once. To this end, we construct our single-pull index (SPI) $\mathcal{I}_n(s_n(t);t)$ associated with arm n at time t as

$$\mathcal{I}_n(s_n(t);t) := \chi_n^\star(s_n(t),1;t) r_n(s_n(t),1), \qquad (18)$$

where $\chi_n^\star(s_n(t),a;t)$ is defined in (17). Notice that our SPI measures the expected reward for activating the arm n in state $s_n(t)$ at time t. Therefore, a higher index indicates a higher expected reward, suggesting that the intuition is to activate arms that contribute more significantly to the accumulated reward. Our index policy then activates arms with SPI indices in a decreasing order. The entire procedure is summarized in Algorithm 1.

Algorithm 1 SPI Index Policy

Input: Initialize $\mathbf{s}_1(s)\ \forall n \in [N]$.

1: Construct the LP according to (16) and solve the occupancy measure μ^*;
2: Compute $\chi_n^\star(s,a,t), \forall s,a,t$ according to (17);
3: Construct the SPI set $\mathcal{I}(t) := \{I_n(s_n(t);t) : n \in [N]\}$ according to (18); and sort $\mathcal{I}(t)$ in a decreasing order;
4: **if** Budget remains **then**
5: Activate arms according to the order in step 3 ;
6: **if** the activated arm is in dummy states **then**
7: Do not pull the arm and let the budget minus one unit;
8: **end if**
9: **end if**

4.2 Asymptotic and Non-Asymptotic Optimality

We now provide results on asymptotic and non-asymptotic optimality for our new index. We begin by showing that our index is asymptotically optimal in the same asymptotic regime as that in Whittle [32] and others [28, 31, 34, 38]. With some abuse of notation, let the number of users be ρN and the resource constraint be ρK in the asymptotic regime with $\rho \to \infty$. In other words, we consider N different classes of users with each class containing ρ users. Let $J^\pi(\rho K, \rho N)$ denote the expected total reward of the original problem (1) under an arbitrary policy π for such a system. Denote the optimal policy for the original problem (2) as $\pi^{Opt} := \{\pi_n^{Opt}, \forall n \in \mathcal{N}\}$.

Theorem 1. *The designed SPI policy (Algorithm 1) is asymptotically optimal, i.e.,*

$$\lim_{\rho\to\infty} \frac{1}{\rho N}\left(J^{\pi^{Opt}}(\rho K, \rho N) - J^{\pi^{SPI}}(\rho K, \rho N)\right) = 0. \qquad (19)$$

[2] If the denominator equals to 0, we direct set $\chi_n^\star(s,1;t) = 0$.

Remark 4. *Theorem 1 indicates that as the number of per-class users goes to infinity, the average gap between the performance achieved by our* SPI *policy π^{SPI} and the optimal policy π^{Opt} tends to be zero. This is a well-established criteria for showing the "optimality" of designed index policies in existing work [14, 28, 32, 34, 38], and the results only hold for $\rho \to \infty$. In the following, we present a more rigorous characterization of the optimality gap under finite scaling factor ρ, which is given by Theorem 2.*

Theorem 2. *For a finite number of $\rho \in \mathbb{R}^+$, with probability at least $1 - \frac{1}{\rho}$ such that the average gap between the performance achieved by our index policy π^{Index} and the optimal policy π^{SPI} is given as*

$$\frac{1}{\rho N}\Big(J^{\pi^{Opt}}(\rho B, \rho N) - J^{\pi^{SPI}}(\rho B, \rho N)\Big) \leq 2r_{max}T\sqrt{\frac{\ln 2\rho}{2N\rho}} + 5r_{max}T\sqrt{\frac{\ln 2\rho}{2\rho^3}}. \tag{20}$$

Corollary 1. *Theorem 2 indicates that the average gap between the performance achieved by our index policy π^{Index} and the optimal policy π^{Opt} is of the order of $O\left(\frac{1}{\rho^{1/2}} + \frac{1}{\rho^{3/2}}\right)$, which is dominated by the first term. We also observe from Theorem 2 that if multiple factor ρ goes to infinity, the performance gap converges to* 0*, which resume the asymptotic optimality in Theorem 1.*

Remark 5. *Our index policy is computationally appealing since it is only based on the "relaxed problem" by solving a LP. Furthermore, if all arms share the same MDP, the LP can be decomposed across arms as in [32], and hence the computational complexity does not scale with the number of arms. More importantly, our index policy is well-defined without the requirement of indexability condition [32]. This is in contrast to most of the existing Whittle index-based policies that are only well defined in the case that the system is indexable, which is hard to verify and may not hold in general. Closest to our work is the parallel work on restless bandits [38], which explores index policies similar to ours, but under the assumption of homogeneous MDPs across arms in the binary action settings, and mainly focus on characterizing the asymptotic optimality gap. There is a branch of work [8, 11, 12, 38] that focuses on analyzing the gap with an explicit relationship involving ρ, but all of them rely on the assumption that rho is large enough for the central limit theorem and mean-field approximation to apply. In contrast, we provide the first characterization that holds for a finite number of ρ. This further differentiates our work with existing literature.*

5 EXPERIMENTS

In this section, we numerically evaluate the proposed SPI policy in three domains: two from real-world applications and one from a synthetic domain, comparing it to state-of-the-art benchmark algorithms. Main results from the two real-world domains are presented here, while the results from the synthetic domain are provided in the supplementary materials.

5.1 Benchmarks

The benchmarks we compare in this paper are listed below:

▷ *Mean-Field LP-based index policy [12]:* The mean-field LP-based index policy is a classic LP-based approach for solving RMAB problems, leveraging mean-field approximation theory when the number of arms is large, as expressed in (6)-(9). However, it does not account for the single-pull constraint when designing the indices.

▷ *Original Whittle index policy [32]:* The Whittle index defined in (5) is the most widely used approach for solving Restless RMABs. It is designed for infinite-horizon problems and does not take the single-pull constraint into account.

▷ *Q-Difference policy [6]:* The Q-difference method designs indices based on the difference between Q-value functions. It is a heuristic approach that can perform well in practice, but it lacks theoretical guarantees, making its performance uncertain in certain scenarios.

▷ *Modified infinite Whittle index policy:* This is a hurestic modification for original Whittle index by considering the dummy states introduced for our proposed SPI policy in Section 4.

▷ *Modified Finite Whittle index policy:* This is a further modification of modified Whittle index by considering finite-horizon time-dependent index.

5.2 Experimental Domains

We briefly introduce the two considered real-world domains below, and relegate the detailed description to Section C.1 in supplementary materials.

5.2.1 Continuous Positive Airway Pressure Therapy (CPAP)[13, 21, 30]. CPAP is a highly effective treatment for adults with obstructive sleep apnea when used consistently during sleep. We model CPAP adherence behavior as a multi-state system, adapting the Markov model with clinical adherence criteria, which reduces to a standard Birth-Death process. In the standard CPAP setting, lower adherence levels yield lower rewards. The objective is to maximize the accumulated reward over time, with the constraint that each patient (arm) can only be pulled (intervened) at most once.

5.2.2 Mobile Healthcare for Maternal Health (MHMH)[12]. In this program, healthcare workers make phone calls to enrollees (beneficiaries) to enhance engagement and provide targeted health information. Since the number of healthcare workers is much smaller than the number of beneficiaries, they must continuously prioritize which beneficiaries to call to maximize the total return. As in [12], we assume two types of beneficiaries, greedy and reliable. We intentionally show two variations of this domain "fixed" and "variable" to illustrate the impact of fixed and varying group sizes on the performance of different index policies.

5.3 Numerical Results

The simulations are conducted for N types of arms, with each type consisting of ρ arms. Each type of arm follows a distinct MDP, and different types of arms have varying MDPs. All results are averaged over 1000 Monte Carlo simulations to ensure robust performance evaluation. Due to the space limitation, we present the main results in this section and more numerical results can be found in Section C in supplementary materials.

Accumulated Reward. We first compare the accumulated reward performance for the proposed SPI index policy with all benchmark polices. Based on the numerical evaluation results presented

Policy	**Birth-Death Process (CPAP)**			**Greedy-Reliable-Fixed (MHMH)**			**Greedy-Reliable-Variable (MHMH)**		
	(20, 5, 10, 10, 10)	(40, 5, 10, 10, 10)	(40, 5, 10, 5, 12)	(10, 3, 25, 50, 10)	(20, 3, 50, 50, 10)	(20, 3, 25, 50, 20)	(20, 3, 1, 2, 20)	(20, 3, 5, 10, 20)	(20, 3, 15, 30, 20)
Upper Bound	200.0	200.0	220.0	174.6	343.6	415.9	16.6	83.2	249.5
SPI	197.5 ± 0.3	200 ± 0	217.4 ± 0.3	172.6 ± 0.3	341.9 ± 0.5	410.6 ± 0.5	14.0 ± 0.2	79.1 ± 0.3	244.5 ± 0.4
Mean Field	187.8 ± 0.7	182.0 ± 0.3	211.4 ± 0.3	156.9 ± 0.4	311.6 ± 0.6	386.0 ± 0.7	13.6 ± 0.2	75.3 ± 0.3	230.3 ± 0.5
Finite Whittle	153.6 ± 0.4	158.0 ± 0.1	153.7 ± 0.4	166.0 ± 0.3	322.1 ± 0.5	402.2 ± 0.5	14.6 ± 0.1	79.2 ± 0.2	240.8 ± 0.4
Infinite Whittle	197.7 ± 0.2	200 ± 0	217.7 ± 0.3	172.1 ± 0.3	325.5 ± 0.5	379.8 ± 0.5	14.0 ± 0.1	75.0 ± 0.3	227.5 ± 0.4
Original Whittle	178.2 ± 0.2	176.7 ± 0.2	199.3 ± 0.3	140.5 ± 0.5	250.3 ± 0.6	339.4 ± 0.6	12.2 ± 0.2	66.8 ± 0.3	202.8 ± 0.5
Q-Difference	111.6 ± 0.4	112.2 ± 0.3	112.0 ± 0.3	165.7 ± 0.4	337.1 ± 0.5	388.5 ± 0.5	14.0 ± 0.1	76.3 ± 0.3	232.3 ± 0.4
Random	140.0 ± 0.5	139.8 ± 0.4	159.8 ± 0.4	26.3 ± 0.3	47.2 ± 0.4	45.2 ± 0.5	1.8 ± 0.1	8.9 ± 0.3	26.9 ± 0.4

Table 1: We present the performance of various policies across different domains and settings. We run each settings for 1000 simulations and present 95% confidence interval. Each setting is denoted by the parameters (number of types N, number of states S, budget K, group size ρ, time horizon T). In each simulation, we consider N types of arms, with each type consisting of ρ arms. Transition probabilities for each type are randomly assigned in every setting. Optimal policies are highlighted in green, and near-optimal policies are highlighted in yellow. Here near-optimal means the gap between it and optimal policy is less than three percent of the upper bound. We use the optimal value achieved for the LP (16) in SPI to serve as the upper bound.

in Table 1, we observe the performance of various policies across different domains and settings, with the proposed SPI policy consistently achieving near-optimal performance. Each setting is described by parameters such as the number of types N, number of states S, budget K, group size ρ, and time horizon T. In all scenarios, SPI policy either matches or comes extremely close to the optimal policy, with a performance gap of less than 3%, demonstrating the robustness of SPI policy across varying settings. Other benchmark policies, such as the Mean Field policy and infinite Whittle index policy, often perform well but exhibit noticeable gaps from the optimal in certain cases. For instance, in the "Greedy-Reliable-Fixed" setting $(20, 3, 50, 50, 20)$, Mean Field index policy achieves 386.0 ± 0.7, significantly lower than the optimal 415.9, indicating sub-optimality. The Random policy consistently underperforms, yielding the lowest rewards across all settings. For example, in the "Greedy-Reliable-Variable" setting $(20, 3, 15, 30, 20)$, Random policy achieves 26.9 ± 0.4, a stark contrast to the optimal 249.5. Both the finite and infinite Whittle index policies show decent performance, but often fall short compared to SPI policy and the optimal policy. In the "Birth-Death Process" setting $(40, 5, 10, 10, 10)$, infinite Whittle index policy achieves the optimal 200 ± 0, while finite Whittle index policy lags behind with 158.0 ± 0.1. Overall, SPI policy consistently demonstrates strong performance, closely matching or achieving optimal rewards across all settings, while other benchmark methods show varying degrees of sub-optimality, with sometimes substantial gaps from the optimal policy.

We also observe that the performance of SPI improves and becomes better as the group size increases. In the "Greedy-Reliable-Variable" setting, when the group size is small, such as $\rho = 2$, policies—including SPI, finite Whittle index, infinite Whittle index, and Q-difference—are either optimal or near-optimal. However, as the group size increases to $\rho = 30$, only SPI policy remains optimal, while finite Whittle index policy becomes sub-optimal and the performance of other policies declines even further.

We consider another domain called *Enhrenfest project* studied in [32]. For domains like the Enhrenfest project, different index policies can achieve very similar near-optimal performance in practice, and the detail is presented in Section C.1.

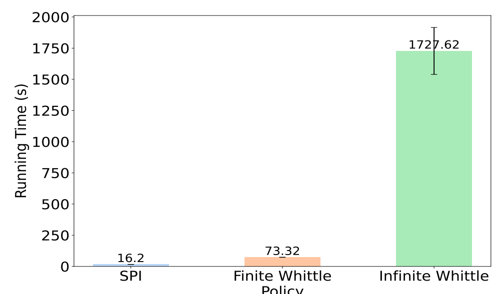

Figure 4: We present the average running time of SPI policy, finite whittle policy, and infinite whittle policy in the CPAP setting $(N, S, K, \rho, T) = (10, 10, 50, 50, 10)$.

Running Time. In Figure 4, we compare the running time of SPI policy with whittle-index-based policies for Birth-Death Process, and we include the running time comparison for randomly generated MDPs as a robustness check in Figure 7 (Section C.2 in supplementary materials). We randomly generate three different Birth-Death Process MDPs, and we take the average running time of each policy under different MDPs. The proposed SPI policy significantly outperforms the Whittle-index-based policy in terms of running time. The Whittle-index-based policy requires first computing the value function and then gradually adjusting the parameter to search the Whittle index, which results in a longer running time due to the curse of dimensionality. In contrast, the SPI policy only needs to solve the LP once when computing the index, which greatly enhances its scalability. This makes SPI particularly well-suited for real-time resource allocation, where non-profit organizations often face tight computation constraints. *Although the infinite Whittle-index policy achieves competitive performance in some domains, its running time is significantly higher than that of the SPI policy, making SPI a more practical and efficient choice in scenarios where rapid decision-making is essential.*

Asymptotic Optimality. We empirically show that SPI is asymptotic optimal, and defer its discussion to Section C.2.3.

ACKNOWLEDGMENTS

This work is supported by the Harvard Data Science Initative.

REFERENCES

[1] Eitan Altman. 2021. *Constrained Markov decision processes*. Routledge.

[2] K Avrachenkov, VS Borkar, and S Pattathil. 2017. Controlling G-AIMD using index policy. In *The 56th IEEE Conference on Decision and Control, Melbourne, December*. 12–15.

[3] Konstantin E Avrachenkov and Vivek S Borkar. 2022. Whittle index based Q-learning for restless bandits with average reward. *Automatica* 139 (2022), 110186.

[4] Turgay Ayer, Can Zhang, Anthony Bonifonte, Anne C Spaulding, and Jagpreet Chhatwal. 2019. Prioritizing hepatitis C treatment in US prisons. *Operations Research* 67, 3 (2019), 853–873.

[5] Saeed Bagheri and Anna Scaglione. 2015. The restless multi-armed bandit formulation of the cognitive compressive sensing problem. *IEEE Transactions on Signal Processing* 63, 5 (2015), 1183–1198.

[6] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. 2021. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. In *Proc. of IJCAI*.

[7] Niclas Boehmer, Yunfan Zhao, Guojun Xiong, Paula Rodriguez-Diaz, Paola Del Cueto Cibrian, Joseph Ngonzi, Adeline Boatin, and Milind Tambe. 2024. Optimizing Vital Sign Monitoring in Resource-Constrained Maternal Care: An RL-Based Restless Bandit Approach. arXiv:2410.08377 [cs.AI] https://arxiv.org/abs/2410.08377

[8] David B Brown and Jingwei Zhang. 2023. Fluid policies, reoptimization, and performance guarantees in dynamic resource allocation. *Operations Research* (2023).

[9] Kobi Cohen, Qing Zhao, and Anna Scaglione. 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, 1575–1578.

[10] Jing Fu, Yoni Nazarathy, Sarat Moka, and Peter G Taylor. 2019. Towards q-learning the whittle index for restless bandits. In *2019 Australian & New Zealand Control Conference (ANZCC)*. IEEE, 249–254.

[11] Nicolas Gast, Bruno Gaujal, and Chen Yan. 2023. Linear program-based policies for restless bandits: Necessary and sufficient conditions for (exponentially fast) asymptotic optimality. *Mathematics of Operations Research* (2023).

[12] Abheek Ghosh, Dheeraj Nagaraj, Manish Jain, and Milind Tambe. 2022. Indexability is not enough for whittle: Improved, near-optimal algorithms for restless bandits. *arXiv preprint arXiv:2211.00112* (2022).

[13] Christine Herlihy, Aviva Prins, Aravind Srinivasan, and John P Dickerson. 2023. Planning to fairly allocate: Probabilistic fairness in the restless bandit setting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 732–740.

[14] Weici Hu and Peter Frazier. 2017. An Asymptotically Optimal Index Policy for Finite-Horizon Restless Bandits. *arXiv preprint arXiv:1707.00205* (2017).

[15] Young Hun Jung and Ambuj Tewari. 2019. Regret bounds for thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[16] Jackson A Killian, Andrew Perrault, and Milind Tambe. 2021. Beyond" To Act or Not to Act": Fast Lagrangian Approaches to General Multi-Action Restless Bandits. In *Proc.of AAMAS*.

[17] Céline Langendorf, Thomas Roederer, Saskia de Pee, Denise Brown, Stéphane Doyon, Abdoul-Aziz Mamaty, Lynda W-M Touré, Mahamane L Manzo, and Rebecca F Grais. 2014. Preventing acute malnutrition among young children in crises: a prospective intervention study in Niger. *PLoS medicine* 11, 9 (2014), e1001714.

[18] Maialen Larrañaga, Urtzi Ayesta, and Ina Maria Verloop. 2014. Index Policies for A Multi-Class Queue with Convex Holding Cost and Abandonments. In *Proc. of ACM Sigmetrics*.

[19] Maialen Larrnaaga, Urtzi Ayesta, and Ina Maria Verloop. 2016. Dynamic Control of Birth-and-Death Restless Bandits: Application to Resource-Allocation Problems. *IEEE/ACM Transactions on Networking* 24, 6 (2016), 3812–3825.

[20] Elliot Lee, Mariel S Lavieri, and Michael Volk. 2019. Optimal Screening for Hepatocellular Carcinoma: A Restless Bandit Model. *Manufacturing & Service Operations Management* 21, 1 (2019), 198–212.

[21] Dexun Li and Pradeep Varakantham. 2022. Towards Soft Fairness in Restless Multi-Armed Bandits. *arXiv preprint arXiv:2207.13343* (2022).

[22] Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. 2022. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12017–12025.

[23] Aditya Mate, Andrew Perrault, and Milind Tambe. 2021. Risk-Aware Interventions in Public Health: Planning with Restless Multi-Armed Bandits. In *Proc.of AAMAS*.

[24] Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I Hou, Srinivas Shakkottai, et al. 2021. NeurWIN: Neural Whittle index network for restless bandits via deep RL. *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021), 828–839.

[25] Christos H Papadimitriou and John N Tsitsiklis. 1999. The complexity of optimal queueing network control. *Mathematics of Operations Research* 24, 2 (1999), 293–305.

[26] Martin L Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

[27] Cem Tekin and Mingyan Liu. 2011. Adaptive Learning of Uncontrolled Restless Bandits with Logarithmic Regret. In *Proc. of Allerton*.

[28] Ina Maria Verloop. 2016. Asymptotically Optimal Priority Policies for Indexable and Nonindexable Restless Bandits. *The Annals of Applied Probability* 26, 4 (2016), 1947–1995.

[29] Shresth Verma, Gargi Singh, Aditya Mate, Paritosh Verma, Sruthi Gorantla, Neha Madhiwalla, Aparna Hegde, Divy Thakkar, Manish Jain, Milind Tambe, et al. 2023. Expanding impact of mobile health programs: SAHELI for maternal and child care. *AI Magazine* 44, 4 (2023), 363–376.

[30] Shufan Wang, Guojun Xiong, and Jian Li. 2024. Online Restless Multi-Armed Bandits with Long-Term Fairness Constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 15616–15624.

[31] Richard R Weber and Gideon Weiss. 1990. On An Index Policy for Restless Bandits. *Journal of applied probability* (1990), 637–648.

[32] Peter Whittle. 1988. Restless Bandits: Activity Allocation in A Changing World. *Journal of applied probability* (1988), 287–298.

[33] Guojun Xiong and Jian Li. 2023. Finite-time analysis of whittle index based Q-learning for restless multi-armed bandits with neural network function approximation. *Advances in Neural Information Processing Systems* 36 (2023), 29048–29073.

[34] Guojun Xiong, Jian Li, and Rahul Singh. 2022. Reinforcement learning augmented asymptotically optimal index policy for finite-horizon restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 8726–8734.

[35] Guojun Xiong, Shufan Wang, and Jian Li. 2022. Learning infinite-horizon average-reward restless multi-action bandits via index awareness. *Advances in Neural Information Processing Systems* 35 (2022), 17911–17925.

[36] Guojun Xiong, Shufan Wang, Gang Yan, and Jian Li. 2022. Reinforcement Learning for Dynamic Dimensioning of Cloud Caches: A Restless Bandit Approach. In *Proc. of IEEE INFOCOM*.

[37] Zhe Yu, Yunjian Xu, and Lang Tong. 2018. Deadline Scheduling as Restless Bandits. *IEEE Trans. Automat. Control* 63, 8 (2018), 2343–2358.

[38] Xiangyu Zhang and Peter I Frazier. 2021. Restless Bandits with Many Arms: Beating the Central Limit Theorem. *arXiv preprint arXiv:2107.11911* (2021).