6-2023

Avoiding starvation of arms in restless multi-armed bandit

Dexun LI
Singapore Management University, dexunli.2019@phdcs.smu.edu.sg

Pradeep VARAKANTHAM
Singapore Management University, pradeepv@smu.edu.sg

Avoiding Starvation of Arms in Restless Multi-Armed Bandits

Dexun Li
Singapore Management University
Singapore
dexunli.2019@smu.edu.sg

Pradeep Varakantham
Singapore Management University
Singapore
pradeepv@smu.edu.sg

ABSTRACT

Restless multi-armed bandits (RMAB) is a popular framework for optimizing performance with limited resources under uncertainty. It is an extremely useful model for monitoring beneficiaries (arms) and executing timely interventions using health workers (limited resources) to ensure optimal benefit in public health settings. For instance, RMAB has been used to track patients' health and monitor their adherence in tuberculosis settings, ensure pregnant mothers listen to automated calls about good pregnancy practices, etc. Due to the limited resources, typically certain individuals, communities, or regions are starved of interventions, which can potentially have a significant negative impact on the individual/community in the long term. To that end, we first define a soft fairness objective which entails an algorithm never probabilistically favors one arm over another if the long-term cumulative reward of choosing the latter arm is higher. Then we provide a scalable approach to ensure long-term optimality while satisfying the proposed fairness constraints in RMAB. Our method, referred to as *SoftFair*, can balance the trade-off between the goal of having resources uniformly distributed and maximizing cumulative rewards. *SoftFair* also provides theoretical performance guarantees and is asymptotically optimal. Finally, we demonstrate the utility of our approaches on simulated benchmarks and show that the soft fairness objective can be handled without a significant sacrifice on the optimal value.

KEYWORDS

Restless Multi-Armed Bandits; Fairness; Softmax; Whittle indexs

ACM Reference Format:
Dexun Li and Pradeep Varakantham. 2023. Avoiding Starvation of Arms in Restless Multi-Armed Bandits. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 9 pages.

1 INTRODUCTION

Restless Multi-Armed Bandits (RMAB) Process is a generalization of the classical Multi-Armed Bandits (MAB) process, which has been studied since the 1930s [15]. RMAB is a powerful framework for budget-constrained resource allocation tasks in which a decision-maker must select a subset of arms for interventions in each round. Each arm evolves according to an underlying Markov Decision Process (MDP). The overall objective in a RMAB model is to sequentially select arms so as to maximize the expected value of the cumulative rewards collected over all the arms. RMAB is of relevance in public health monitoring scenarios, recommendation systems and many others. Tracking a patient's health or adherence and intervening at the right time is an ideal problem setting for an RMAB [2, 25], where the patient health/adherence state is represented using an arm. Resource limitation constraint in RMAB comes about due to the severely limited availability of healthcare personnel. By developing practically relevant approaches for solving RMAB within severe resource limitations, RMAB can assist patients in alleviating health issues such as diabetes [28]. hypertension [4], tuberculosis [5, 29], depression [23, 27], etc.

While Whittle index based approaches [19, 25] address the RMAB problem with an infinite time horizon by providing an asymptotically optimal solution, they are susceptible to starving arms, which can have severe repercussions in public health scenarios. Owing to the deterministic selection strategy of picking arms that provide the maximum benefit, in many problems, only a small set of arms typically get picked. As shown in our experimental analysis, Figure 1 provides one example, where almost 50% of the arms do not get any interventions using the Whittle index approach. While it is an optimal decision, it should be noted that interventions help educate patients or beneficiaries on potential benefits and starvation of such interventions for many patients can result in a lack of proper understanding of the program and reduce its effectiveness in the long run. Thus, there is a need to not starve arms without significantly sacrificing optimality. Providing such decision support with a fairness mindset can promote acceptability among community [16, 34].

Existing works have proposed different notions of fairness in the context of MAB to prevent starvation by enabling the selection of non-optimal arms. Li et al. [21] study a new Combinatorial Sleeping MAB model with Fairness constraints, called CSMAB-F. Their fairness definition requires algorithm to ensure a minimum selection fraction for each arm. Patil et al. [30] introduce similar fairness constraints in the stochastic MAB problem, where they use a pre-specified vector to denote the guaranteed number of selections. Joseph et al. [13] define fairness as saying that a worse arm should not be picked compared to a better arm, despite the uncertainty on payoffs. Chen et al. [6] form the allocation decision-making problem as the MAB with fairness constraints, where fairness is defined as a minimum rate at which a task or resource is assigned to a user. Since knowing the guaranteed number (or proportion) of selection is difficult to ascertain *a priori*, we generalize these fairness notions for MAB. We build on the notion of fairness introduced by Jabbari et al. [11] for reinforcement learning setting and introduce a *soft fairness* notion for our RMAB setting. Our soft fairness definition requires that an RMAB algorithm never favor an arm probabilistically over another arm, if the long-term cumulative reward of choosing the latter arm is higher.

In summary, our goal is to compute stochastic policy for selecting arms in finite horizon RMAB, which satisfies the soft fairness notion. To that end, we make the following contributions:

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

- A practically relevant algorithm called *SoftFair*, that enforces the soft fairness constraint and thereby avoids starvation of interventions for arms. Unlike the well-known Whittle index algorithm that can only solve the infinite horizon setting, *SoftFair* can also easily handle finite horizon RMAB.
- *SoftFair* provides a trade-off between optimal performance and avoiding intervention starvation for arms. This trade-off is highlighted by the performance bounds and theoretical properties of the *SoftFair* algorithm.
- Detailed experimental results demonstrate that *SoftFair* is competitive with other policies in terms of expected reward, while significantly reducing the starvation of interventions for arms (by increasing the entropy of the stochastic policy).

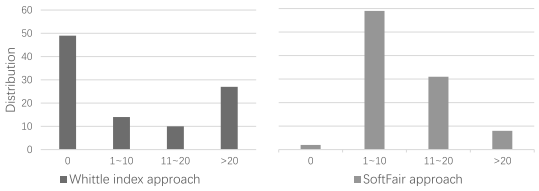

Figure 1: The x-axis is the number of times being selected, and the y-axis is the frequency distribution. We consider the RMAB given in Section 3, with $k = 10$, $n = 100$, $T = 100$. Left: the Whittle index algorithm. Right: *SoftFair* ($c = 2$). As can be noted, without fairness constraints in place, the arm activation frequency is lopsided, and almost 50% of the arms never get activated.

2 RELATED WORK

We focus on two threads of relevant research, the first category is related to approaches for solving RMAB, and the second category is related to fairness definitions and related approaches in decision making.

As one of the most well-studied generalisations of the Multi-Armed Bandit (MAB), RMAB is increasingly used for decision making problems ranging from wireless broadcast [33, 36], job allocation [12], cancer detection [19], wildlife protection [32], recommender systems [26], and health intervention [3, 17]. Whittle [41] considered the Lagrangian relaxation of the RMAB in which arm selection constraint (number of arms selected = k) is enforced on average over the horizon. This policy, referred as the Whittle index policy is asymptotically optimal [40]. Liu and Zhao [22] investigate the application of RMAB in dynamic multichannel access, establish indexability and obtain Whittle index in closed form for both discounted and average reward criteria. In [32], the authors formulate the wildlife protection problem as a RMAB model and present an algorithm that is based on binary search to find Whittle index policy. Mate et al. [25] build a fast algorithm for computing the Whittle index, which provides an order-of-magnitude speedup compared to Qian et al. [32]. Biswas et al. [3] develop a model-free learning method based on Q-learning mechanism and show that it converge to the optimal solution asymptotically. Online RMAB has also raised some attentions in recent years, Wang et al. [39] present a learning policy to construct offline instances in guiding action selection. Xiong et al. [42] propose a generative model based reinforcement learning augmented algorithm toward an index policy.

Another line of work that is closely related to ours is the growing body of literature on ensuring fairness in decision making [7, 11], in particular in the domain of resource allocation [6, 21]. For example, ensuring resources are fairly distributed among the arms is an important design concern in wireless communication systems [8]. In the case of beneficiaries, an arm/patient might consider action/participation fair when participation of a certain patient (i.e., due to receiving an active action) resulted in a greater increase in expected time spent in a adherent state compared to non-participation (i.e., the passive action on the arm/patient) [16]. One widely used fairness notion in MAB literature is to ensure that there is a minimum rate of arm activation for each user (arm) over time [21, 31]. Joseph et al. [13] introduce the study of fairness in MAB problems, where their fairness notion is defined as not giving preference to a worse arm over a better one. The quality of an action is the expected immediate reward for selecting action from current state. However, this notion of fairness can lead to policies favoring short-term rewards and ignoring long-term rewards. Jabbari et al. [11] therefore adapt the fairness notion by defining the quality of an action as its potential long-term reward, and generalize it to a reinforcement learning setting. Li and Varakantham [20] define fairness as a minimum number of times an arm should be activated in a decision period, but their definition requires pre-specified values about the number of activation times and the length of the decision period.

We generalize the fairness notion in [11] to the RMAB setting, and make several advancements in this paper.

3 MODEL: RMAB

In this section, we formally introduce the finite horizon RMAB model with a new objective of computing policies that balance the trade-off between maximizing cumulative rewards while giving a reasonable chance for each arm (proportional to their value) to get selected for intervention. As indicated earlier, this is a property that is of critical importance in public health settings. RMAB is defined using the tuple: $\langle N, \{\mathcal{M}_i\}_{i \in N}, T, k \rangle$. There are N independent arms[1] and each arm i evolves according to an associated Markov Decision Process (MDP), $\mathcal{M}_i$ is characterized by the tuple $\{\mathcal{S}_i, \mathcal{A}_i, \mathcal{P}_i, R_i, \gamma\}$:

- $\mathcal{S}_i$ represents the state space. Typically, in public health settings, $\mathcal{S}_i = \{0, 1\}$. 1 represents patient in the "good" state (patient adheres to the health program), and 0 represents patient in the "bad" state (patient not adhering).
- $\mathcal{A}_i$ represents the action space. $\mathcal{A}_i = \{0, 1\}$ with 1 corresponding to activating or intervening on the arm and 0 action corresponding to not activating the arm or staying passive.
- $\mathcal{P}_i$ represents the action dependent transition dynamics of arm i. Specifically, $P^{a_i}_{s_i, s'_i}$ refers to the probability of transitioning from state s_i to s'_i when the arm i is taking action $a_i \in \{0, 1\}$.
- R_i provides the independent rewards obtained by arm i. We assume a range for these rewards, given by $[R_{min}, R_{max}]$. We use a simple reward function: $R_i(s_i, a_i) = s'_i \in \{0, 1\}$ determined by the

[1] in public health settings, the patients or beneficiaries will be the arms

next state s_i' obtained by taking action a_i when the observed state is s_i for any arm $i \in [N]$. [2] Note that the expected immediate reward will be $\mathbb{E}[R_i(s_i, a_i)] = P^{a_i}_{s_i,1}$.

- γ is the discount factor.

T is the time horizon. k is the resource capacity constraint that limits the number of arms that can be selected at each time step $t \in [T]$, i.e.,:

$$\sum_{i=1}^{N} a_i^t = k \tag{1}$$

Policy, π for the overall RMAB is a mapping from joint states, $\mathbf{s} = [s_1, \ldots, s_N]$ of all arms to joint actions, $\mathbf{a} = [a_1, \ldots, a_N]$. $\pi(\mathbf{s}, \mathbf{a}) \in [0, 1]$ denotes the probability of selecting the joint action $\mathbf{a}$ when the joint state of RMAB is $\mathbf{s}$. Particularly, $\pi_i(s_i, a_i) \in [0, 1]$ denotes the probability of selecting action a_i, with $\sum_{a_i} \pi_i(s_i, a_i) = 1$. We denote the state-action value function for a policy π by

$$Q^{\pi}(\mathbf{s}, \mathbf{a}) = \mathbb{E}^{\pi}[\sum_{t=1}^{T} \gamma^{t-1} R^t(\mathbf{s}^t, \mathbf{a}^t)] = \mathbb{E}^{\pi}[\sum_{t=1}^{T} \gamma^{t-1} \sum_{i=1}^{N} R_i^t(s_i^t, a_i^t)]$$

$Q^{\pi}(\mathbf{s}, \mathbf{a})$ is the expected cumulative discounted long-term reward over all arms when taking action $\mathbf{a}$ in the joint state $\mathbf{s}$. The objective is to find an optimal policy π^* that can satisfy

$$Q^{\pi^*}(\mathbf{s}, \mathbf{a}) = \max_{\pi} Q^{\pi}(\mathbf{s}, \mathbf{a}) = Q^*(\mathbf{s}, \mathbf{a}),$$

where $Q^*(\mathbf{s}, \mathbf{a})$ is the optimal state-action value function:

$$Q^*(\mathbf{s}, \mathbf{a}) = R(\mathbf{s}, \mathbf{a}) + \gamma \sum_{\mathbf{s}'} \Pr(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \max_{\mathbf{a}'} Q^*(\mathbf{s}', \mathbf{a}') \tag{2}$$

Similar to Jabbari et al. [11], we define the fairness using the state-action value function $Q^*(\mathbf{s}, \mathbf{a})$ as follows:

Definition 1. *(Fairness) A stochastic policy, π is fair if for any time step $t \in [T]$, any joint state $\mathbf{s}$ and actions $\mathbf{a}, \mathbf{a}'$, where $\mathbf{a} \neq \mathbf{a}'$:*

$$\pi^t(\mathbf{s}, \mathbf{a}) \geq \pi^t(\mathbf{s}, \mathbf{a}') \text{ only if } Q^*(\mathbf{s}, \mathbf{a}) \geq Q^*(\mathbf{s}, \mathbf{a}') \tag{3}$$

In summary, the objective is to efficiently approximate the maximum cumulative long-term reward while satisfying resource constraints and fairness constraints. Towards this end, the reward maximization problem can be formulated as

$$\underset{\pi}{\text{maximize}}\ \mathbb{E}_{\pi}[\sum_{t=1}^{T} \gamma^{t-1} R^t(\mathbf{s}^t, \mathbf{a}^t)] \tag{4}$$
$$\text{such that Equation. 1, and Equation. 3 are satisfied}$$

We show in Proposition 1 that this fairness notion at the level of joint actions is equivalent to selecting arms with higher probability if their relative importance is higher.

4 *SOFTFAIR* APPROACH

In this section, we design a probabilistically fair (as defined in Definition 1) arm selection algorithm, referred to as *SoftFair*. *SoftFair* builds on softmax value iteration [38, 43] in conjunction with Whittle index. Softmax value iteration is one of the simplest dynamic programming algorithms, which recursively computes the value function through a point-wise update rule [35]. The notations that are frequently used in this paper are summarized in Table 1.

In order to implement the softmax value iteration method in the RMAB setting, we need to compute the relative value of activating an arm (in comparison to not activating the arm) and compute the probability distribution of selecting an arm using a *softmax* function over the relative value. More specifically, during the ep-th iteration, *SoftFair* identifies the estimated value function of the state of each arm $i \in [N]$ at the time step $t \in [T]$, and calculate the difference of state-action value function between the active and passive action.

$$\begin{aligned} Q_i^{t;ep}(s_i^t, a_i^t) &= R_i^t(s_i^t, a_i^t) + \gamma \sum_{s_i^{t+1}} \Pr(s_i^{t+1}|s_i^t, a_i^t) V_i^{t+1;ep}(s_i^{t+1}) \\ \zeta_i^{t;ep}(s_i^t, a_i^t) &= e^{Q_i^{t;ep}(s_i^t, a_i^t) - V_i^{t;ep}(s_i^t)} \\ \lambda_i^{t;ep} &= \log \zeta_i^{t;ep}(s_i^t, a_i^t = 1) - \log \zeta_i^{t;ep}(s_i^t, a_i^t = 0) \end{aligned} \tag{5}$$

Here $V_i^{t;ep}(\cdot)$ is the value function of arm i from time step t till the end of horizon after being updated ep times. Similarly, $Q_i^{t;ep}(s_i^t, a_i^t)$ is the state-action value function of arm i from time step t till the end of horizon during ep-th iteration. Then *SoftFair* maps each arm i's state to a state-specific probability distribution over actions using the following *softmax* expression in the $k = 1$ case.

$$\pi^{t;ep}(\mathbf{s}^t, \mathbf{a}^t = \mathbb{I}_{\{i\}}) = \frac{\exp(c \cdot \lambda_i^{t;ep})}{\sum_{q=1}^{N} \exp(c \cdot \lambda_q^{t;ep})} \tag{6}$$

where $\mathbf{a}^t = \mathbb{I}_{\{i\}}$ denotes the joint action [3] to select the arm i while keeping other arms passive, and $\pi^{t;ep}(\mathbf{s}^t, \mathbf{a}^t = \mathbb{I}_{\{i\}})$ denotes the probability that arm i will be selected under the joint state $\mathbf{s}^t$ during the ep-th iteration. $c \in (0, \infty)$ is the multiplier parameter [4] that can adjust the gap between the probabilities of choosing an arm. If $c = \infty$, *SoftFair* becomes the standard optimal Bellman operations [1] (Refer to Equation 13). When $c = 0$, each arm has the same probability of being selected, and *SoftFair* can make the resources uniformly distributed. Equation 6 shows the probability of selecting an joint action at each time step when $k = 1$.

Unfortunately, this expression does not hold when selecting a subset of arms, i.e., $k > 1$. This is because when the resource constraint $k > 1$, the probability of an arm being selected will also rely on the probability of other arms being selected, henceforth affecting the recursive update of the value function. Let $\mathbf{a}^t = \mathbb{I}_{\{\phi\}}$ denote the action to select arms in the set ϕ. Then, ϕ^- is the set that includes all of the arms except those in set ϕ. After getting the action probability of selecting a single arm, which is the multinomial distribution, formulated as $[\pi^{t;ep}(\mathbf{s}^t, \mathbf{a}^t = \mathbb{I}_{\{1\}}), \pi^{t;ep}(\mathbf{s}^t, \mathbf{a}^t = \mathbb{I}_{\{2\}}), \ldots, \pi^{t;ep}(\mathbf{s}^t, \mathbf{a}^t = \mathbb{I}_{\{N\}})]$. We can then sample from this multinomial distribution without replacement to obtain k arms to activate, which ensures that we meet the resource constraint as well as the fairness constraint. More specifically, we can derive the probability that the arm i is among the k selected arms (active set ϕ),

[2] The reward function over RMAB can be written as $R(\mathbf{s}, \mathbf{a}) = \sum_{i=1}^{N} R_i(s_i, a_i)$

[3] $\mathbb{I}_{\{i\}}$ is the indicator with value 1 at the ith item and value 0 at other places. Equivalently, this means activating arm i while keeping the other arms passive

[4] The updation process of our *Softfair* algorithm will converge to the Bellman Equation 13 with an exponential rate in terms of c [37], and c controls the asymptotic performance [18].

Table 1: Notations

Notation	Description
k, N, T	N:number of all arms in RMAB; k:number of arms can be selected each round; T: time horizon.
c, ep, γ	c: multiplier parameter, ep: iteration times, γ: discount factor.
$s_i, a_i, \mathbf{s}, \mathbf{a}$	s_i, a_i: state and action of arm i, $\mathbf{s}, \mathbf{a}$: joint state vector and joint action vector of RMAB.
$[n], [T]$	We use [n] to represent the set of integers $\{1, \ldots, n\}$ for $n \in \mathbb{N}$ and [T] also has the same interpretation.
$P^a_{s,s'}$	$P^a_{s,s'}$ refers to the probability of transition from state s to s' when an arm is taking action a.
$Pr(a_i\|\mathbf{s})$	$Pr(a_i\|\mathbf{s})$ is the probability that arm i is among the selected arms when the joint state of RMAB is $\mathbf{s}$.
$Q^t_m(s,a)$, $V^t_m(s)$	$Q^t_m(s,a)$: A state-action value function for the subsidy m and state s when taking action a start at time step t followed by optimal policy using Whittle index based approach in the future time steps; $V^t_m(s)$: Value function for the subsidy m and state s start at time step t using Whittle index based approach
$Q^t(s,a)$, $V^t(s)$	$Q^t(s,a)$: The state-action value function when taking action a at time step t with state s $V^t(s)$: The value function at the time step t with state s.

denoted as $\Pr(a^t_i = 1|\mathbf{s}^t)$ [5]. Consider the multinomial distribution, the results of k draws made at random without replacement is a random permutation of all the elements, and this can be computed through the permutation tool. Consequently, we have:

$$\pi^{t;ep}(\mathbf{s}^t, \mathbf{a}^t = \mathbb{I}_{\{\phi\}}) = \prod_{i\in\phi} \Pr(a^t_i = 1|\mathbf{s}^t) \prod_{j\in\phi^-} \left(1 - \Pr(a^t_j = 1|\mathbf{s}^t)\right) \tag{7}$$

For arm i, the value function $V^{t;ep}_i(\cdot)$ at time step t during the ep-th iteration can be written as

$$V^{t;ep}_i(s^t_i) = \sum_{a^t_i \in \{0,1\}} \Pr(a^t_i|\mathbf{s}^t) Q^{t;ep}_i(s^t_i, a^t_i)$$

Please note that $Pr(a^t_i|\mathbf{s}^t)$ is computed based on the sample estimate. The update of value function $V^{t;ep}_i(\cdot)$ during the ep-th iteration for any $t \in [T]$ is:

- For the current state, s^t_i of arm i, the value function is updated in the following way:

$$V^{t;ep+1}_i(s^t_i) = \sum_{a^t_i \in \{0,1\}} \Pr(a^t_i|\mathbf{s}^t) \sum_{s^{t+1}_i \in \{0,1\}} \Pr(s^{t+1}_i|s^t_i, a^t_i) \cdot \left(R(s^t_i, a^t_i) + \gamma V^{t+1;ep}_i(s^{t+1}_i)\right) \tag{8}$$

- For all other states, s^t of arm i we have

$$V^{t;ep+1}_i(s^t) = V^{t;ep}_i(s^t) \tag{9}$$

Similarly, we can also write the equation to update the state-action value function, and we provide it in the Appendix. The overall process of *SoftFair* is summarized in Algorithm 1 and is guaranteed to ensure that an arm is selected in proportion with its λ value, thereby guaranteeing fairness while approximately maximizing the overall value. This guarantee is possible because we can decouple the fairness constraint defined on the joint action to each individual arm. We have the following proposition, which is equivalent to the definition 1.

Proposition 1. *Fairness of a stochastic policy defined in Equation 3 can also be stated in terms of arm selection as follows:*

$$Pr(a^t_i = 1|\mathbf{s}^t) \geq Pr(a^t_j = 1|\mathbf{s}^t) \text{ only if } \lambda^t_i \geq \lambda^t_j \tag{10}$$

[5]Note that $\Pr(a_i = 1|\mathbf{s}) = \pi^{ep}(\mathbf{s}^t, \mathbf{a}^t = \mathbb{I}_{\{i\}}) = softmax_c(c \cdot \lambda_i)$ if $k = 1$

Algorithm 1 SoftFair Value Iteration (*SoftFair*)

Input: Transition matrices $\{\mathcal{P}_i\}_{i\in N}$, time horizon T, set of observed states $\mathbf{s}$, resource constraint k, multiplier parameter c, iteration length I

Output: The value function $V_i(s)$ for arm $i \in [N]$

```
V^t_i(s) ← 0, ∀s, i, t
for iteration ep = 1, ..., I do
    Initialize s^0 = {s^0_1, ..., s^0_N}
    for step t = 0, ..., T do
        for arm i = 1, ..., n do
            Compute Q^{t;ep}_i(s^t_i, a^t_i) and λ^{t;ep}_i(s^t_i, a^t_i) using Equation. 5
            Compute π^{t;ep}(s^t, a^t = I_{i}) using Equation. 6
        end for
        Sample k arms and add them into action set
        for arm i = 1, ..., n do
            Compute Pr(a^t_i = 1|s^t)
            Update V^{t;ep}_i(s) using Equation. 8 and Equation. 9
        end for
        Play the arm in the action set, and observe next state s^{t+1}
    end for
end for
```

Intuitively, this implies an arm, i will not be selected with lower probability than that of arm j if λ value of arm i is higher than that of arm j. The proof showing that the proposition is equivalent to the definition 1 is provided in the appendix.

5 ANALYSIS OF *SOFTFAIR*

In this section, we formally analyze the properties of the *SoftFair* algorithm. We begin by comparing *SoftFair* with the well-known Whittle index algorithm and show why the Whittle index approach is not suitable for our case (Fairness constraint and Finite time horizon), and then provide the performance bound of *SoftFair*.

5.1 *SoftFair vs.* Whittle index based methods

Whittle index policy is known to be the asymptotically optimal solution to RMAB for the *infinite* time horizon. It independently assigns an index value for each arm to measure how attractive it is to activate an arm at a particular state. The index is computed using the concept of a "subsidy" m, which can be viewed as the opportunity cost of remaining passive, and is rewarded to the arm that is passive, in addition to the usual reward. The Whittle index for an arm i is defined as the infimum value of the subsidy, m that must be offered to the algorithm to make the algorithm indifferent between selecting and not selecting the arm. Consider a single arm $i \in [n]$ where the state is s_i^t at time step $t \in [T]$, let $Q_{m;i}^t(s_i^t, a_i^t = 0)$ and $Q_{m;i}^t(s_i^t, a_i^t = 1)$ denote its active and passive state-action value functions under a subsidy m, respectively. For ease of explanation, we drop the subscript i when there is no ambiguity. The value function of an arm in the state s is

$$V_m^t(s^t) = \max\{Q_m^t(s^t, a^t = 0), Q_m^t(s^t, a^t = 1)\}.$$

The Whittle index $W(s^t)$ for the state s^t can be formally written as:

$$W(s^t) = \inf_m \left\{ m^t : Q_m^t(s^t, a^t = 0) = Q_m^t(s^t, a^t = 1) \right\}. \quad (11)$$

After computing the Whittle index for each arm, a policy π will activate those k arms whose current states have the highest indices. In order to use the Whittle index approach, it needs to satisfy a technical condition called *indexability* introduced by Weber and Weiss [40]. The indexability can be expressed in a simple way: Consider an arm with subsidy m, the optimal action is passive, then $\forall m' > m$, the optimal action should remain passive. The RMAB is indexable if every arm is indexable.

However, traditional Whittle index based approaches rely on the assumption of an infinite time horizon, and the performance deteriorates severely when time horizons are finite. Figure 2 shows an illustrative example where Whittle index values are low when an arm's residual time horizon is short, and there is a bias in approximating the Whittle index value under the finite time horizon setting using methods proposed in [24, 32]. Often, real-world phenomena are formalized in a finite time horizon setting, which precludes the direct use of Whittle index based methods. We now demonstrate that a phenomenon called Whittle index decay [20, 24] exists in our problem. All detailed proofs can be found in the Appendix.

Theorem 1. *At any time step $t \in [T]$, the Whittle index m^t for arm i under the observed state s_i^t is the value that satisfies the equation $Q_m^t(s_i^t, a_i^t = 0) = Q_m^t(s_i^t, a_i^t = 1)$. The Whittle index will decay as the value of current time step t increases: $\forall t < T : m^t > m^{t+1} \geq m^T = P_{s,1}^1 - P_{s,1}^0$.*

Proof Sketch. Consider the discount reward criterion with the discount factor γ, we can simply compute m^T and m^{T-1} by solving equations $Q_m^T(s_i^T, a_i^T = 0) = Q_m^T(s_i^T, a_i^T = 1)$[6]. We can find that $m^{T-1} \geq m^T = P_{s,1}^1 - P_{s,1}^0$. Then in order to show $m^t > m^{t+1}$, we first prove a lemma to show value function $V_m^t(s_i^t) > V_m^{t+1}(s_i^t) \geq 0$, and then we can combine this with the definition of m^t to complete the proof. The detailed proof is in the appendix. □

The Whittle index based approach needs to solve the costly finite horizon problem because the index value varies according to the residual time horizon even in the same state, and computing the index value under the finite horizon setting is $(O(|S|^k T)$ time and space complexity [10]. However, as an alternative method, our *SoftFair* can naturally approximate the optimal value function at arbitrary time steps while requiring less memory space than model-free learning methods such as Q-learning. In addition, the optimal condition for approximating the Whittle index value is difficult to satisfy. For example, Mate et al. [24] demonstrate that their proposed approach is optimal under the condition:

$$P_{1,1}^1 - P_{0,1}^1 \leq \left(P_{1,1}^0 - P_{0,1}^0\right)\left(1 + \gamma(P_{1,1}^1 - P_{0,1}^1)\right)(1 - \gamma)$$

Intuitively, consider the case where $P_{0,1}^0 = P_{0,1}^1$ and $P_{1,1}^0 = P_{1,1}^1$ (also considered by Liu and Zhao [22]), this makes such a condition always not satisfied. Furthermore, we will show that the Whittle index based approach fails to address the problem of fair distribution of interventions (the distribution of resources is lopsided). In contrast, our proposed method, *SoftFair* becomes the optimal algorithm when $c \to \infty$ and can control the trade-off between optimal performance and uniform distribution of resources.

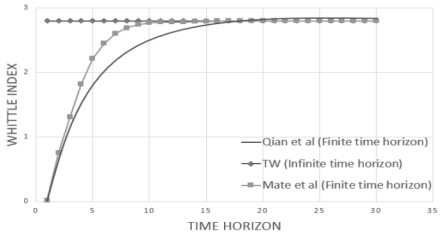

Figure 2: Whittle index value as a function of the residual time horizon. Figure taken from Mate et al. [24]. The grey line is the whittle index value in an infinite time horizon setting, and the others are approximated Whittle index values under a finite time horizon to capture the index decay phenomenon.

Due to the finite time horizon setting in many practical applications, the Whittle index based method can not effectively approximate the whittle index value, and it only concentrates on beneficiaries who can mostly improve the objective in the case of initiatives related to public health. This can result in some beneficiaries never having the opportunity to receive intervention from public health professionals, which may lead to a poor adherence behavior and henceforth a bad state from which improvements may only be marginal even with intervention, preventing them from ever being chosen by the index policy. Refer to Figure 1 to get a better picture of the difference between the Whittle index approach and *SoftFair*. We can see that when using the Threshold Whittle index based method proposed by Mate et al. [24], the activation frequency of the arm is extremely unbalanced, with nearly half of the arms never being selected. Such starvation of interventions may escalate to communities. To avoid such cycle between bad outcomes, the RMAB needs to consider fairness in addition to maximizing cumulative long-term reward when picking arms. We now demonstrate why *SoftFair* can satisfy our proposed fairness constraint while effectively approximating our cumulative reward maximization objective. We begin by providing a theorem showing that *SoftFair* is guaranteed to be optimal when the multiplier parameter $c \to \infty$.

[6] get m^{T-1} by solving $Q_m^{T-1}(s_i^{T-1}, a_i^{T-1} = 0) = Q_m^{T-1}(s_i^{T-1}, a_i^{T-1} = 1)$

Theorem 2. *Choose the top k arms according to the λ value in Equation 6 ($c \to \infty$) is equivalent to maximizing the cumulative long-term reward.*

Proof Sketch. Because when c approaches infinity, *SoftFair* becomes deterministically choosing the arm with the highest λ value. Let ϕ^* to be the set of actions containing the k arms with the highest-ranking of λ value, we need to show $Q(\mathbf{s}, \mathbf{a} = \mathbb{I}_{\{\phi\}}) \geq Q(\mathbf{s}, \mathbf{a}' = \mathbb{I}_{\{\phi'\}})$ for $\forall \phi'$, where ϕ' is the set of any k selected arms, and $\phi' \neq \phi^*$. We first get the expression of $\sum_{i \in \phi} \lambda_i$. Combining the definition of λ in Equation 5 with the fact that $\sum_{i \in \phi^*} \lambda_i \geq \sum_{j \in \phi'} \lambda_j$, we add $\sum_{z \notin \phi^* \wedge z \notin \phi'} Q(s_z, a_z = 0)$ on both sides of the inequality function to show $Q(\mathbf{s}, \mathbf{a} = \mathbb{I}_{\{\phi\}}) \geq Q(\mathbf{s}, \mathbf{a}' = \mathbb{I}_{\{\phi'\}})$. □

When c approaches infinity, *SoftFair* becomes the optimal policy, but it will suffer from the starvation phenomena. As c gets closer to 0, *SoftFair* can ensure that every arm/beneficiary has roughly the same probability of receiving the intervention, which leads to a uniform distribution of resources. Given these facts, c can control the trade-off between ensuring the fair distribution of resources and the objective of maximizing cumulative rewards. In the subsequent theorem, we demonstrate that *SoftFair* satisfies our proposed fairness constraint.

Theorem 3. *SoftFair is fair under our proposed fairness constraint, and c controls the trade-off between fairness and optimal performance.*

Proof Sketch. Similar to the proof of the Theorem 2, we can see that the value of λ is proportional to the state-action value function. According to the Equation 6, the probability of selecting an arm is the softmax function on λ, and it can be guaranteed that the higher the value of λ, the higher the probability of selecting that arm. Therefore *SoftFair* remains fair under our proposed fairness constraints. The trade-off between ensuring a fair distribution of resources and the objective of maximizing cumulative rewards is governed by c, where a larger c means *SoftFair* prefers arms with a higher value of λ, while a small c means that *SoftFair* tends to ensure that resources are uniformly distributed among the arms. □

In the next section we will show how the value of c controls the performance bounds of the *SoftFair* algorithm.

5.2 Performance bound of *SoftFair*

For ease of explanation, we investigate the case of $k = 1$ at each time step, and the multi-selection ($k > 1$) can be viewed as the iteration of the case $k = 1$. Let Ψ_{soft} denote our *Soft* operator at time step $t \in [T]$, we ignore the subscript t here, which is

$$\begin{aligned} Q^{ep+1}(\mathbf{s}, \mathbf{a}) &= \Psi_{soft} Q^{ep}(\mathbf{s}, \mathbf{a}) \\ &= \sum_{\mathbf{s}'} \Pr(\mathbf{s}'|\mathbf{s}, \mathbf{a})(R(\mathbf{s}, \mathbf{a}) + \gamma \sum_{\mathbf{a}'} \Pr(\mathbf{a}'|\mathbf{s}') Q^{ep}(\mathbf{s}', \mathbf{a}')) \\ &= R(\mathbf{s}, \mathbf{a}) + \gamma \sum_{\mathbf{s}'} \Pr(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \sum_{\mathbf{a}'} \Pr(\mathbf{a}'|\mathbf{s}') Q^{ep}(\mathbf{s}', \mathbf{a}'). \end{aligned} \tag{12}$$

Before we derive the performance bound for *SoftFair*, We first bound the state-action value function in the following lemma.

Lemma 1. *The $Q(\mathbf{s}, \mathbf{a})$ is bounded within $[0, n/(1-\gamma)]$.*

Proof Sketch. The upper bound can be obtained by showing that $\forall(\mathbf{s}, \mathbf{a})$, state-action value during the $ep-$th iteration are bounded through induction. □

Corollary 1. *As we have $R_{max} = n$ and $R_{min} = 0$ of RMAB, we can easily derive that $|Q(\mathbf{s}, \mathbf{a}) - Q(\mathbf{s}, \mathbf{a}')| \leq \frac{n}{1-\gamma}$, for $\forall Q$ and $\forall(\mathbf{s}, \mathbf{a})$.*

Following Song et al. [37], we let $\delta(\mathbf{s}) = \sup_Q \max_{\mathbf{a},\mathbf{a}'} |Q(\mathbf{s}, \mathbf{a}) - Q(\mathbf{s}, \mathbf{a}')|$ denote the largest distance between state-action value functions. Then we have the following lemma showing the bound on the difference between two state-action value functions.

Lemma 2. *$\forall Q$ and $\forall \mathbf{s}$, Let $\Pr(\cdot|\mathbf{s}) = [\Pr(\mathbf{a} = \mathbb{I}_{\{1\}}|\mathbf{s}), \ldots, \Pr(\mathbf{a} = \mathbb{I}_{\{n\}}|\mathbf{s})]^\top$ and $Q(\mathbf{s}, \cdot) = [Q(\mathbf{s}, \mathbf{a} = \mathbb{I}_{\{1\}}), \ldots, Q(\mathbf{s}, \mathbf{a} = \mathbb{I}_{\{n\}})]^\top$, here the superscript $\top$ denotes the vector transpose. We have $\frac{\delta(\mathbf{s})}{n \exp[c \cdot \delta(\mathbf{s})]} \leq \max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a}) - (\Pr(\cdot|\mathbf{s}))^\top Q(\mathbf{s}, \cdot) \leq \frac{n-1}{2+c}$.*

Proof Sketch. We first sort $Q(\mathbf{s}, \mathbf{a} = \mathbb{I}_{\{i\}})$ in the ascending order according to the λ value and replace $\Pr(\cdot|\mathbf{s})$ with $Q(\mathbf{s}, \cdot)$. We take advantage of the fact that for any two non-negative sequences $\{x_i\}$ and $\{y_i\}$, $\frac{\sum_i x_i}{1+\sum_i y_i} \leq \sum_i \frac{x_i}{1+y_i}$, combine this fact with the difference between state-action value functions for different actions. Through using Taylor series, we can derive the upper and lower bounds. □

Different from *Soft* Operator Ψ_{soft} in Eq. 12 ,let Ψ denote the Bellman optimality operator, which we have

$$\begin{aligned} Q^{ep+1}(\mathbf{s}, \mathbf{a}) &= \Psi Q^{ep}(\mathbf{s}, \mathbf{a}) \\ &= R(\mathbf{s}, \mathbf{a}) + \gamma \sum_{\mathbf{s}'} \Pr(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \max_{\mathbf{a}'} Q^{ep}(\mathbf{s}', \mathbf{a}') \end{aligned} \tag{13}$$

For the optimal state-action value function, we have $\Psi Q^*(\mathbf{s}, \mathbf{a}) = Q^*(\mathbf{s}, \mathbf{a})$. We have the following theorem showing the performance bound of *SoftFair* compared to the optimal value.

Theorem 4. *Our SoftFair method can achieve the performance bound as $\limsup_{ep \to \infty} V^{ep}(\mathbf{s}) \leq V^*(\mathbf{s})$, where $V^*(\mathbf{s})$ is the optimal value function. More specifically, we have*

$$\limsup_{ep \to \infty} Q^{ep}(\mathbf{s}, \mathbf{a}) \leq Q^*(\mathbf{s}, \mathbf{a}) \quad \textit{and}$$

$$\liminf_{ep \to \infty} Q^{ep}(\mathbf{s}, \mathbf{a}) \geq Q^*(\mathbf{s}, \mathbf{a}) - \frac{n-1}{(2+c)(1-\gamma)}$$

Proof. We derive the performance bound through induction based on Lemma 1 and 2. □

Conjecture 1. *For the cause when multiple arms can be pulled at each time step, i.e., $k > 1$, Our SoftFair method can achieve the bound as $\limsup_{ep \to \infty} \Psi^{ep} V^0(\mathbf{s}) \leq V^*(\mathbf{s})$. More specifically, we have*

$$\limsup_{ep \to \infty} Q^{ep}(\mathbf{s}, \mathbf{a}) = \limsup_{ep \to \infty} \Psi^{ep} Q^0(\mathbf{s}, \mathbf{a}) \leq Q^*(\mathbf{s}, \mathbf{a}) \quad \textit{and}$$

$$\liminf_{ep \to \infty} Q^{ep}(\mathbf{s}, \mathbf{a}) \geq Q^*(\mathbf{s}, \mathbf{a}) - \frac{n-k}{(2+c)(1-\gamma)}$$

6 EXPERIMENTS

In this section, we empirically compare our proposed method *SoftFair* to the baselines on both (a) a realistic patient adherence behavior dataset [14] and (b) a synthetic dataset to represent more general settings enforced by structural constraints on transition

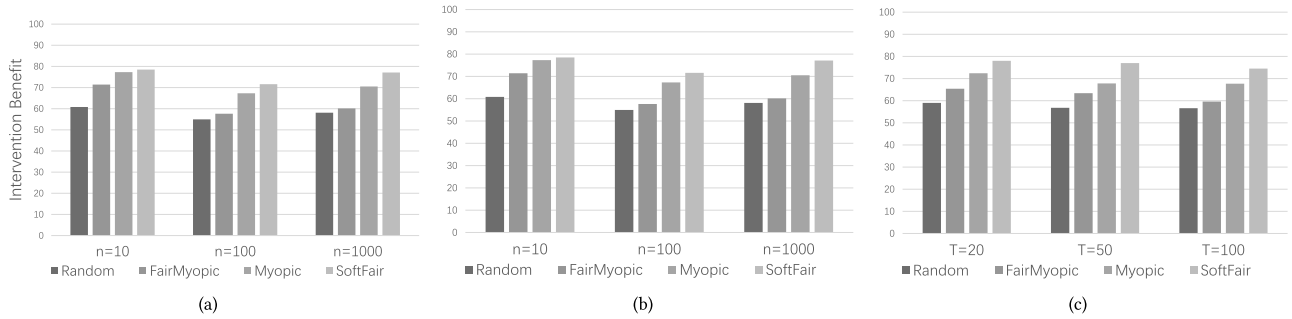

Figure 3: Intervention benefit of *SoftFair* is consistently greater than other baselines. (a) We fix $T = 50$, and $k = 10\% * n$, and let $n = \{10, 100, 1000\}$. (b) We fix $T = 50$, and $n = 100$, and let $k = \{5, 10, 20\}$. (c) We fix $n = 100$, and $k = 10$, and let $T = \{20, 50, 100\}$.

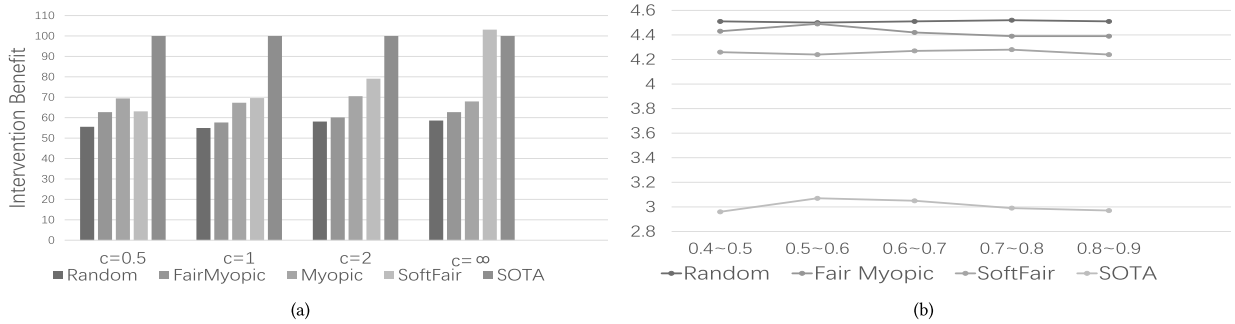

Figure 4: (a) The intervention benefit of different multiplier c. Here $c = \infty$ refers to deterministically selecting the top k arm with the highest cumulative rewards. (b) The action entropy of a single process. We investigate the action entropy for different value of $P^1_{0,1}$ range from 0.4 to 0.9 (at 0.45, 0.55, 0.65, 0.75, 0.85, respectively), and $c = 1$.

matrix (more details in Appendix). We consider the finite time horizon where reward is the undiscounted sum of arms/beneficiaries in the good state over all time steps and set the following scenario for the simulation: $n = \{10, 100, 1000\}$, $k = \{5\%n, 10\%n, 20\%n\}$, $T = \{20, 50, 100\}$. All results are averaged over 50 simulations. In particular, We compare our method against the following baselines:

- **Random**: At each time step, algorithm randomly select k arms to play. This will ensure that each arm has the same probability of being selected.
- **Myopic**: A myopic policy ignores the impact of present actions on future rewards and instead focuses entirely on the predicted immediate returns. It select k arms that maximize the expected reward at the immediate next time step. Formally, this could be described as choosing the k arms with the largest gap $\Delta^t = P^1_{s,1} - P^0_{s,1}$ at time step t under the observed state s.
- **FairMyopic**: After computing Δ^t for each arm, instead of deterministically selecting the arm with the highest immediate reward, we use the *softmax* function over Δ^t to get the probability of each arm being selected. Then we sample the k arms according to the probability.
- **FaWT**: Algorithm proposed by Li and Varakantham [20]. They ensure that each arm will be selected at least η times during any intervention interval of length L. Since this algorithm requires two predefined and extra parameters, the intervention interval length L and the minimum selection times during each interval η, it is not feasible to create a fair comparison against other approaches across all settings. However, for one of the settings we are able to provide a direct comparison with *SoftFair* by doing a brute force search for fair parameter values for FaWT.
- **SOTA**: Algorithm proposed by Mate et al. [24] under the assumption that the states of all arms are fully observable and the transition probabilities are known. We use a sigmoid function to approximate the Whittle index value and select arms deterministically for the finite time horizon setting.

We examine policy performance from two perspectives: (a) Intervention benefit (essentially the solution quality): The intervention benefit is defined as $\frac{\bar{R}_{\text{method}} - \bar{R}_{\text{No intervention}}}{\bar{R}_{\text{SOTA}} - \bar{R}_{\text{No intervention}}} \times 100\%$. It calculates the difference between one algorithm's expected cumulative reward and the cumulative reward when no intervention is involved, then normalized by the difference between the asymptotically optimal but fairness-agnostic SOTA algorithm in baselines (100% intervention benefit) and the reward obtained without intervention (0% intervention benefit) and. (b) Action distribution entropy (representative of the fairness): We calculate the selection frequency distribution across all time steps, and then compute its entropy after normalization through: $Entropy = -\sum_{i \in [n]} P(i) \log P(i)$, where $P(i)$ refers

Table 2: Resutls for CAPA Adherence dataset with $n = 100, k = 10, T = 80$.

Policy	Intervention benefit	Action entropy
Random	79 ± 13	4.56 ± 0.0056
Myopic	98 ± 3.3	2.67 ± 0.0
FairMyopic	83 ± 11	4.5 ± 0.0089
SoftFair	93 ± 7.6	4.27 ± 0.019

to the normalization of the number of times arm i is selected (i.e., the number of times that arm i has been selected divided by $k \cdot T$), and $P(i) \log P(i) = 0$ if an arm is never selected.

Realistic dataset. : Obstructive sleep apnea is one of the most prevalent sleep disorder among adults, and continuous positive airway pressure therapy (CPAP) is a highly effective treatment when it is used consistently for the duration of each sleep bout. But non-adherence to CPAP in patients hinders effective treatment for this type of sleep disorder. Similar to [9], we adapt the Markov model of CPAP adherence behavior in [14] to a two-state system with the clinical adherence criteria. We add a small noise to each transition matrix so that the dynamics of each individual arm is different (See more details about the dataset in Appendix).

In table 2, we report average results for each algorithm. Myopic method has the best performance, which is caused by the specific structure of the underlying transition matrices, since there is not too much difference between n Markovian models, and in this case the Myopic approach is indeed close to optimal. However, the myopic approach has significantly lower action entropy, which is indicative of overall fairness. Meanwhile, our *SoftFair* provides the right trade-off between intervention benefit and having a varied selection of arms (high action entropy) at each time step.

Synthetic dataset. (a) We first test the performance when the number of patients (arms) varies. Figure 3a compares the intervention benefit for $n = \{10, 100, 1000\}$ patients and $k = 10\%$ of n. As shown in Figure 3a, in addition to satisfying the fairness constraints, our *SoftFair* consistently outperforms the Random, Myopic and FairMyopic baselines. (b) We next compare the intervention benefit when the number of arms n is fixed and the resource constraint k is varied. Specifically, we fix $n = 100$ patients, and let $k = \{5, 10, 20\}$. Figure 3b shows that there has been a gradual increase in the intervention benefit as the k increases. One possible reason is that a larger resource budget k can make the arms with higher cumulative rewards more likely to be selected, thereby reducing the performance gap with the SOTA method. (c) The performance of our method is slightly influenced by the time horizon T. As shown in Figure 3c, the common trend is that a smaller T leads to better performance. This means that our method can efficiently solve the RMAB in a finite time horizon, while a larger horizon T will make the convergence slower. Overall, all results demonstrate the our method provides a good trade-off between providing high intervention benefit and preventing starvation for arms (through high action entropy).

Intervention benefit when c changes. We investigate the effect of the multiplier parameter c on performance. Formally, a larger c will

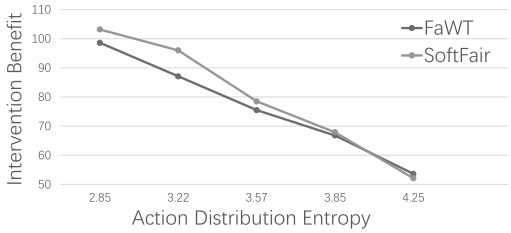

Figure 5: Comparison of performance of FaWT and *SoftFair* when their action distribution entropy values are close.

widen the gap between the probabilities of choosing an arm, leading to better performance as it prefers selecting an arm with a higher cumulative reward. Figure 4 (a) reveals that *SoftFair* performs well empirically as c increases, and if we deterministically choose the top k arms based on the value of λ, it achieves the optimal result.

Action entropy comparison. We also compare the entropy of the action of a process in the synthetic dataset when $P^1_{0,1}$ ranges from 0.4 to 0.9. As shown in Figure 4, the Random policy has the highest value as it requires uniform selection of all arms. Our proposed method, *SoftFair* consistently has a higher action entropy than the SOTA method because we enforce fairness constraints. FairMyopic has a high action entropy value, but it is indeed unfair under our proposed fairness constraints, as it relies on immediate rewards.

SoftFair vs. FaWT. We perform a search in the value space of parameters η and L of FaWT and the value space of multiplier c of *SoftFair*, and we use the value of these parameters which makes the values of the action distribution entropy of these two methods close to each other and compare their performance. We present the result in Figure 5. As shown in the figure, *SoftFair* can better balance the trade-off between the goal of uniform resource distribution and maximizing cumulative rewards. This may be due to the difficulty in satisfying the conditions for optimal performance of FaWT.

Discussion. In some real-world applications, state transitions may not be fully available. In this case, we can learn the transition probabilities online using a learning method based on Thompson Sampling. We provide detailed experiments in the appendix and show that it performs well in real-world situations.

7 CONCLUSION

In this paper, we study fairness constraints in the context of Restless Multi-Arm Bandits model, which is of critical importance for adherence problems in public health (e.g., monitoring the adherence of preventive medicine for Tuberculosis, monitoring engagement of mothers during calls on good practices during pregnancy). To tackle the challenges introduced by the objective, we design a computationally efficient algorithm by integrating the *softmax* value iteration technique in the RMAB setting. Our algorithm can effectively approximate the optimal value function within the proven performance bounds while having fairness guarantees.