Efficient Resource Allocation with Fairness Constraints in Restless Multi-Armed Bandits

Abstract

Restless Multi-Armed Bandits (RMAB) is an apt model to represent decision-making problems in public health interventions (e.g., tuberculosis, maternal, and child care), anti-poaching planning, sensor monitoring, personalized recommendations and many more. Existing research in RMAB has contributed mechanisms and theoretical results to a wide variety of settings, where the focus is on maximizing expected value. In this paper, we are interested in ensuring that RMAB decision making is also fair to different arms while maximizing expected value. In the context of public health settings, this would ensure that different people and/or communities are fairly represented while making public health intervention decisions. To achieve this goal, we formally define the fairness constraints in RMAB and provide planning and learning methods to solve RMAB in a fair manner. We demonstrate key theoretical properties of fair RMAB and experimentally demonstrate that our proposed methods handle fairness constraints without sacrificing significantly on solution quality.

1 INTRODUCTION

Picking the right time and manner of limited interventions is a problem of great practical importance in tuberculosis [Mate et al., 2020], maternal and child care [Biswas et al., 2021, Mate et al., 2021b], anti-poaching operations [Qian et al., 2016], cancer detection [Lee et al., 2019], and many others. All these problems are characterized by multiple arms (i.e., patients, pregnant mothers, regions of a forest) whose state evolves in an uncertain manner (e.g., medication usage in the case of tuberculosis, engagement patterns of mothers on calls related to good practices in pregnancy) and threads moving to "bad" states have to be steered to "good" outcomes through interventions. The key challenge is that the number of interventions is limited due to a limited set of resources (e.g., public health workers, patrol officers in anti-poaching operations). Restless Multi-Armed Bandits (RMAB), a generalization of Multi-Armed Bandits (MAB) that allows non-active bandits to also undergo the Markovian state transition, has become an ideal model to represent the aforementioned problems of interest as it models uncertainty in arm transitions (to capture uncertain state evolution), actions (to represent interventions) and budget constraint (to represent limited resources).

Existing work [Mate et al., 2020, Biswas et al., 2021, Mate et al., 2021a] has focused on developing theoretical insights and practically efficient methods to solve RMAB. At each decision epoch, RMAB methods identify arms that provide the biggest improvement with an intervention. Such an approach though technically optimal can result in certain arms (or type of arms) getting starved for interventions.

In the case of interventions with regards to public health, RMAB algorithms focus interventions on the top beneficiaries who will improve the objective (public health outcomes) the most. This can result in certain beneficiaries never talking to public health workers and thereby moving to bad states (and potentially also impacting other beneficiaries in the same community) from where improvements can be minor even with intervention and hence never getting picked by RMAB algorithms. As shown in Fig. 1, when using the Threshold Whittle index approach proposed by Mate et al. [2020], the arm activation probability is lopsided, with 30% of arms getting activated more than 50 times and 50% of the arms are never activated. Such starvation of interventions can result in arms moving to a bad state from where interventions cannot provide big improvements and therefore there is further starvation of interventions for those arms. Such starvation can happen to entire regions or communities, resulting in lack of fair support for beneficiaries in those regions/communities. To avoid such cycles between bad outcomes, there is a need for RMAB algorithms to consider fairness in addition to maximizing expected reward when picking arms. Risk sensitive RMAB [Mate et al., 2021b]

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

considers an objective that targets to reduce such starvation, however, they *do not guarantee* that arms (or types of arms) are picked a minimum number of times.

Recent work in Multi-Armed Bandits (MAB) has presented different notions of fairness. For example, Li et al. [2019] study a Combinatorial Sleeping MAB model with Fairness constraints, called CSMAB-F. The fairness constraints ensure a minimum selection fraction for each arm. Patil et al. [2020] introduce similar fairness constraints in the stochastic multi-armed bandit problem, where they use a pre-specified vector to denote the guaranteed number of pulls. Joseph et al. [2016] 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. [2020] define the fairness constraint as a minimum rate that is required when allocating a task or resource to a user. The above fairness definitions are relevant and we generalize from these to propose a fairness notion for RMAB. Unfortunately, approaches developed for fair MAB cannot be utilized for RMAB, due to uncertain state transitions with passive actions as well.

Contributions: To the best of our knowledge, this is the first paper to consider fairness constraints in RMAB. Here are the key contributions:

- We propose a fairness constraint wherein for any arm (or more generally, for a type of arm), we require that the number of decision epochs since the arm (or the type of arm) was activated last time is upper bounded. This will ensure that every arm (or type of arm) gets activated a minimum number of times, thus generalizing on the fairness notions in MAB described earlier.
- We provide a modification to the Whittle index algorithm that is scalable and optimal while being able to handle both finite and infinite horizon cases. We also provide a model-free learning method to solve the problem when the transition probabilities are not known beforehand.
- Experiment results on the generated dataset show that our proposed approaches can achieve good performance while still satisfying the fairness constraint.

2 PROBLEM DESCRIPTION

In this section, we formally introduce the RMAB problem. There are N independent arms, each of which evolves according to an associated Markov Decision Process (MDP). An MDP is characterized by a tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{P}, r\}$, where $\mathcal{S}$ represents the state space, $\mathcal{A}$ represents the action space, $\mathcal{P}$ represents the transition function, and r is the state-dependent reward function. Specifically, each arm has a binary-state space: 1 ("good") and 0 ("bad"), with action-dependent transition matrix $\mathcal{P}$ that is potentially different for each arm. Let $a_t^i \in \{0, 1\}$ denote the action taken at time step t for arm i, and $a_t^i = 1 (a_t^i = 0)$ indicates an active (passive) action for arm i. Due to limited resources, at each

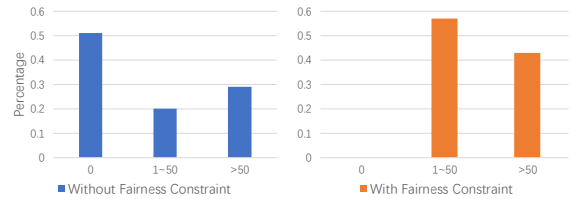

Figure 1: The x-axis is the number of times activated, and the y-axis is the percentage of each frequency range. We consider the RMAB given in Section 2, with $k = 10$, $N = 100$, $T = 1000$ and $L = 50$, $\eta = 2$. Left: the result of using the Whittle index algorithm without considering fairness constraints. Right: the result of when considering fairness constraints. As can be noted, without fairness constraints in place, almost 50% of the arms never get activated.

decision epoch, the decision-maker can activate (or intervene on) at most k out of N arms and receive reward accrued from all arms determined by their states. $\sum_{i=1}^{N} a_t^i = k$ describes this limited resource constraint. Figure 2 provides an example of an arm in RMAB.

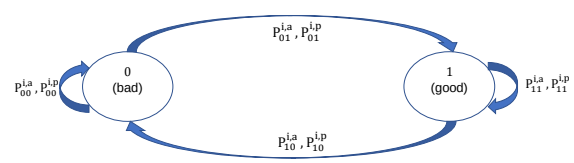

Figure 2: a and p denote the active and passive actions on arm i respectively. $P_{s,s'}^{i,a}$ and $P_{s,s'}^{i,p}$ are the transition probabilities from state s to state s' under action a and p respectively for arm i.

The state of arm i evolves according to the transition matrix $P_{s,s'}^{a,i}$ for the active action and $P_{s,s'}^{p,i}$ for the passive action. We follow the setting in Mate et al. [2020], when the arm i is activated, the latent state of arm i will be fully observed by the decision-maker. The states of passive arms are unobserved by the decision-maker. When considering such partially observable problem, it is sufficient to let the MDP state be the belief state: the probability that the arm is in the "good" state. We need to keep track of the belief state on the current state of the unobserved arm. This can be derived from the decision-maker's partial information which is encompassed by the last observed state and the number of decision time steps since the last activation of the arm. Let $\omega_s^i(u)$ denote the belief state, i.e., the probability that the state of arm i is 1 when it was activated u time steps ago with the observed state s. The belief state in next time step can be obtained by solving the following recursive equations:

$$\omega_s^i(u+1) = \begin{cases} \omega_s^i(u) P_{1,1}^{p,i} + (1 - \omega_s^i(u)) P_{0,1}^{p,i} & \text{passive} \\ P_{s',1}^{i,a} & \text{active} \end{cases} \tag{1}$$

Where s' is the new state observed for arm i when the

active action was taken. The belief state can be calculated in closed form with the given transition probabilities. We let $\omega = \omega_s^i(u+1)$ for ease of explanation when there is no ambiguity.

A policy π maps the belief state vector $\Omega_t = \{\omega_t^1, \cdots, \omega_t^N\}$ at each time step t for all arms to the action vector, $a_t = \{0,1\}^N$. Here ω_t^i is the belief state for arm i at time step t. We want to design an optimal policy to maximize the cumulative long-term reward over all the arms. One widely used performance measure is the expected discounted reward over the horizon T :

$$\mathbb{E}_\pi[\sum_{t=1}^{T} \beta^{t-1} R_t(\Omega_t, \pi(\Omega_t))|\Omega_0]$$

Here $R_t(\Omega_t, \pi(\Omega_t))$ is the reward obtained in slot t under action $a_t = \pi(\Omega_t)$ determined by policy π, β is the discount factor. As we discussed in the introduction, in addition to maximizing the cumulative reward, ensuring fairness among the arms is also a key design concern for many real-world applications. In order to model the fairness requirement, we introduce constraints that ensure that any arm (or kind of arms) is activated at least η times during any decision interval of length L. The overall optimization problem corresponding to the problem at hand is thus given by:

$$\begin{aligned} &\underset{\pi}{\text{maximize}}\ \mathbb{E}_\pi[\sum_{t=1}^{T} \beta^{t-1} R_t(\Omega_t, \pi(\Omega_t))|\Omega_0] \\ &\text{subject to} \sum_{i}^{N} a_t^i = k, \forall t \in \{1, \ldots, T\} \\ &\qquad \sum_{t=u}^{u+L} a_t^i \geq \eta \quad \forall u \in \{1, \ldots, T-L\}, \forall i \in \{1, \ldots, N\}. \end{aligned} \tag{2}$$

η is the minimum number of times an arm should be activated in a decision period of length L. The strength of fairness constraints is thus governed by the combination of L and η. Obviously, this requires $k \times L > N \times (\eta - 1)$ as the fairness constraint should meet the resource constraint. This fairness problem can be formulated at the level of regions/communities by also summing over all the arms, i in a region in the second constraint, i.e.,

$$\sum_{i \in r} \sum_{t=u}^{u+L} a_t^i \geq \eta$$

Our approaches with a simple modification are also applicable to this fairness constraint at the level of regions/communities.

3 BACKGROUND: WHITTLE INDEX

In this section, we describe the Whittle Index algorithm [Whittle, 1988] to solve RMAB. This algorithm at every time step, computes index values (Whittle Index values) for every arm and then activates the arms that have the top "k" index values. Whittle index quantifies how appealing it is to activate a certain arm. This algorithm provides optimal solutions if the underlying RMAB satisfies the indexability property, defined in Definition 1.

Formally[1], the Whittle index of an arm in a belief state ω (i.e., the probability of good state 1) is the minimum subsidy λ such that it is optimal to make the arm passive in that belief state. Let $V_{\lambda,T}(\omega)$ denote the value function for the belief state ω over a horizon T. Then it could be written as

$$V_{\lambda,T}(\omega) = \max\{V_{\lambda,T}(\omega; a=0), V_{\lambda,T}(\omega; a=1)\}, \tag{3}$$

where $V_{\lambda,T}(\omega; a=0)$ and $V_{\lambda,T}(\omega; a=1)$ denote the value function when taking passive and active actions respectively at the first decision epoch followed by optimal policy in the future time steps. Because the expected immediate reward is ω and subsidy for a passive action is λ, we have the value function for passive action as:

$$V_{\lambda,T}(\omega, a=0) = \lambda + \omega + \beta V_{\lambda,T-1}(\tau^1(\omega)), \tag{4}$$

where $\tau^1(\omega)$ is the 1-step belief state update of ω when the passive arm is unobserved for another 1 consecutive slot (see the update rule in Eq. 1). Note that ω is also the expected reward associated with that belief state. For an active action, the immediate reward is ω and there is no subsidy. However, the actual state will be known and then evolve according to the transition matrix for the next step:

$$\begin{aligned} V_{\lambda,T}(\omega, a=1) = \omega + \beta(&\omega V_{\lambda,T-1}(P_{1,1}^a) + \\ &(1-\omega) V_{\lambda,T-1}(P_{0,1}^a)) \end{aligned} \tag{5}$$

Definition 1 *An arm is indexable if the passive set under the subsidy λ given as $\mathcal{P}_\lambda = \{\omega : V_{\lambda,T}(\omega, a=0) \geq V_{\lambda,T}(\omega, a=1)\}$ monotonically increases from $\emptyset$ to the entire state space as λ increases from $-\infty$ to ∞. The RMAB is indexable if every arm is indexable.*

Intuitively, this means that if an arm takes passive action with subsidy λ, it will also take passive action if $\lambda' > \lambda$. Given the *indexability*, $W_T(\omega)$ is the least subsidy, λ that makes it equally desirable to take active and passive actions.

$$W_T(\omega) = \inf_\lambda \{\lambda : V_{\lambda,T}(\omega; a=1) \leq V_{\lambda,T}(\omega; a=0)\} \tag{6}$$

Definition 2 *A policy is a threshold policy if there exists a threshold λ_{th} such that the action is passive $a=0$ if $\lambda > \lambda_{th}$ and $a=1$ otherwise.*

Existing efficient methods for solving RMABs derive these

[1] Since we will only be talking about one arm at a time step, we will abuse the notation by not indexing belief, action and value function with arm id or time index.

threshold policies.

4 FAIRNESS IN RMAB

The key advantage of a Whittle index based approach is scalability without sacrificing solution quality. In this section, we provide Whittle index based approaches to handle fairness constraints under known and unknown transition models, with both infinite and finite horizon settings. We specifically consider partially observable settings[2].

4.1 INFINITE HORIZON

When we need to consider the partial observability of the state of the RMAB problem, it is sufficient to let the MDP state be the belief state: the probability that the arm is in the "good" state [Kaelbling et al., 1998]. As a result, the partially observable RMAB has a large number of belief states [Mate et al., 2020].

Recall that the definition of the Whittle index $W_T(\omega)$ of belief state ω is the smallest λ s.t. it is optimal to make the arm passive in the current state. We can compute the Whittle index value for each arm, and then rank the index value of all N arms and select top k arms at each time step to activate. With fairness constraints, the change to the approach is minimal and intuitive. ***The optimal policy is to choose the arms with the top "k" index values until a fairness constraint is violated for an arm. In that time step, we replace the last arm in top-k with the arm for which fairness constraint is violated.*** We show that this simple change works across the board for the infinite and finite horizon, fully and partially observable settings. We provide the detailed algorithm in Algorithm 1 and also provide sufficient conditions under which the Algorithm 1 is optimal.

We now provide the expression for λ. $V_{\lambda,\infty}(\omega)$ denotes the value that can be accrued from a single-armed bandit process with subsidy λ over infinite time horizon ($T \to \infty$) if the belief state is ω. Therefore, we have:

$$V_{\lambda,\infty}(\omega) = \max \begin{cases} \lambda + \omega + \beta V_{\lambda,\infty}(\tau^1(\omega)) & \text{passive} \\ \omega + \beta \left(\omega V_{\lambda,\infty}(P^a_{1,1}) + (1-\omega) V_{\lambda,\infty}(P^a_{0,1})\right) & \text{active} \end{cases} \tag{7}$$

For any belief state ω, the u-steps belief update $\tau^u(\omega)$ will converge to ω^* as $u \to \infty$, where $\omega^* = \frac{P^p_{0,1}}{1+P^p_{0,1}-P^p_{1,1}}$. It should be noted that this convergence can happen in two ways depending on the state transition patterns:

- Case 1: Positively correlated channel ($P^p_{1,1} \geq P^p_{0,1}$). The belief update process is shown in Figure 3. We can see that for the positively correlated case, they have a monotonous belief update process.

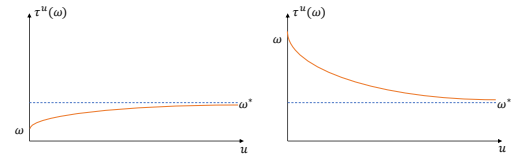

Figure 3: The u-step belief update of an unobserved arm ($P^p_{1,1} \geq P^p_{0,1}$)

We first consider the *non-increasing belief process* as indicated in the right graph. Formally, for $\forall u \in \mathbb{N}^+$, we have $\omega(u) \geq \omega(u+1)$ if the initial belief state ω is above the convergence value. Similarly, for the *increasing belief process* shown in the left graph, we have the initial belief state $\omega < \omega^*$.

- Case 2: Negatively correlated channel ($P^p_{1,1} < P^p_{0,1}$).

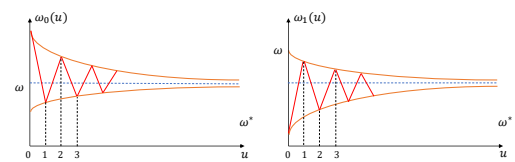

Figure 4: The u-step belief update of an unobserved arm ($P^p_{1,1} < P^p_{0,1}$)

The belief state converges to ω^* from the opposite direction as shown in Figure 4. This case has similar properties and is less common in the real world because it is more likely to remain in a good state than to move from a bad state to a good state. Therefore, we omit the lengthy discussion.

The belief state transition patterns are of particular importance because in proving optimality of Algorithm 1, the belief evolution pattern for the arm (whose fairness constraint will be violated) plays a crucial role.

Theorem 1 *For infinite time horizon ($T \to \infty$) RMAB with Fairness Constraints governed by parameters η and L, Algorithm 1 (i.e., activating arm i at the end of the time period when its fairness constraint is violated) is optimal:*

1. *For $\omega^i \leq \omega^*$ (*increasing belief process*), if*

$$(P^{i,p}_{1,1} - P^{i,p}_{0,1})\left(1 + \frac{\Delta_3}{1-\beta}\right)\left(1 - \beta(P^{i,a}_{1,1} - P^{i,a}_{0,1})\right) \leq (P^{i,a}_{1,1} - P^{i,a}_{0,1}) \tag{8}$$

$\Delta_3 = \min\{(P^{i,p}_{1,1} - P^{i,p}_{0,1}), (P^{i,a}_{1,1} - P^{i,a}_{0,1})\}$.

2. *For $\omega^i \geq \omega^*$ (*non-increasing belief process*), if:*

$$(P^{i,p}_{1,1} - P^{i,p}_{0,1})(1-\beta)\Delta_1 \geq (P^{i,a}_{1,1} - P^{i,a}_{0,1})\left(1 - \beta(P^{i,a}_{1,1} - P^{i,a}_{0,1})\right) \tag{9}$$

[2]We also provide a discussion about fully observable setting in the appendix

Algorithm 1: Fair Whittle Thresholding (FaWT)

Input: Transition matrix $\mathcal{P}$, fairness constraint, η and L, set of belief states $\{\omega^1, \ldots, \omega^N\}$, k

for *each arm i in* 1 *to* N **do**
  Compute the corresponding Whittle index $TW(\omega^i)$ under the infinite horizon using the *Forward and Reverse Threshold* policy;
  if *the activation frequency* η *for arm* i *will not be satisfied at the end of the period of length* L **then**
    Add arm i to the action set ϕ;
    $k = k - 1$;
  if *finite horizon* **then**
    Compute the the index value $W_1(\omega^i)$;
    Compute the Whittle index $W_T(\omega^i)$ using Equation 10;
Add arms with top k highest $TW(\cdot)$ (for infinite horizon case) or $W_T(\cdot)$ (for finite horizon case) values to the action set ϕ Decrease the residual time horizon by $T = T - 1$;

Output: Action set ϕ

$$\Delta_1 = \min\{1, 1 + \beta(P^{i,p}_{1,1} - P^{i,p}_{0,1}) - \beta(P^{i,a}_{1,1} - P^{i,a}_{0,1})\}$$

Proof Sketch. Consider an arm i that has not been activated for $L-1$ time slots. In such a case, Algorithm 1 will select arm i to activate in the next time step $t = L$. Define the intervention effect of activating arm i as

$$V_{\lambda,\infty}(\omega, a = 1) - V_{\lambda,\infty}(\omega, a = 0)$$

Following standard practice and for notational convenience, we do not index the intervention effect and value functions with i. Due to independent evolution of arms, moving active action of arm i does not result in a greater value function for other arms according to the Whittle index algorithm, thus it suffices to only consider arm i. Here is the proof flow:
(1) Algorithm 1 optimality requires that the intervention effect at time step $t = L - 1$ is smaller than intervention effect at $t = L$. Optimality can be established by requiring the partial derivative of the intervention effect w.r.t. time step t is greater than 0.
(2) However, computing this partial derivative $\frac{\partial(V_{\lambda,\infty}(\omega,a=1) - V_{\lambda,\infty}(\omega,a=0))}{\partial t}$ is difficult because value function expression is complex. We use chain rule to get:

$$\frac{\partial(V_{\lambda,\infty}(\omega, a = 1) - V_{\lambda,\infty}(\omega, a = 0))}{\partial \omega} \cdot \frac{\partial(\omega)}{\partial(t)}$$

(3) The sign of second term, $\frac{\partial \omega}{\partial t}$ is based on the belief state transition pattern described before this theorem. We then need to consider the sign of the first term, $\frac{\partial(V_{\lambda,\infty}(\omega,a=1) - V_{\lambda,\infty}(\omega,a=0))}{\partial \omega}$.
(4) We can compute this by deriving the bound on $V_{\lambda,\infty}(\omega_1) - V_{\lambda,\infty}(\omega_2), \forall \omega_1, \omega_2$ as well as bounds on $\frac{\partial V_{\lambda,\infty}(\omega)}{\partial \omega}$. Detailed proof in appendix. □

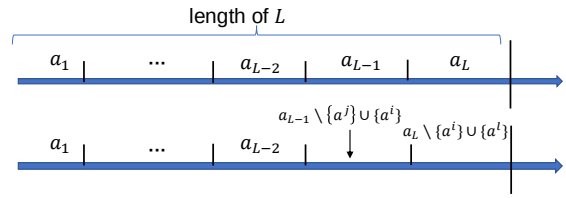

Figure 5: The action vector for RMAB is a_t at time step t. Then we move the action a^i that satisfies fairness constraint to earlier slot and replace k-th ranked action a^j. Action a^l is then added according to the index value at the end.

4.2 FINITE HORIZON

In this part, we demonstrate that the mechanism developed for handling fairness in the infinite horizon setting can also be applied to the finite horizon setting. In showing this, we address two key challenges:

1. Computing the Whittle index under the finite horizon setting in a scalable manner.
2. Showing that Whittle index value reduces as residual horizon decreases. This will assist in showing that it is optimal to activate the fairness violating arm at the absolute last step where a violation will happen and not earlier;

It is costly to compute the index under the finite horizon setting – $O(|\mathcal{S}|^k T)$ time and space complexity [Hu and Frazier, 2017]. Therefore, we take advantage of the fact that the index value has an upper and lower bound, and it will converge to the upper bound as the time horizon $T \to \infty$. Specifically, we use an appropriate functional form to approximate the index value. To do this, we first show gradual Index decay ($\lambda_T > \lambda_{T-1} > \lambda_0$) by improving on the Index decay ($\lambda_T > \lambda_0$) introduced in Mate et al. [2021a].

Theorem 2 *For a finite horizon T, the Whittle index λ_T is the value that satisfies the equation $V_{\lambda_T,T}(\omega, a = 0) = V_{\lambda_T,T}(\omega, a = 1)$ for the belief state ω. Assuming indexability holds, the Whittle index will decay as the value of horizon T decreases: $\forall T > 1 : \lambda_{T+1} > \lambda_T > \lambda_0 = 0$.*

Proof Sketch. We can calculate λ_0 and λ_1 by solving equation $V_{\lambda_0,0}(\omega, a = 0) = V_{\lambda_0,0}(\omega, a = 1)$ and $V_{\lambda_1,1}(\omega, a = 0) = V_{\lambda_1,1}(\omega, a = 1)$ according to Eq. 4 and Eq. 5. We can then derive $\lambda_t > \lambda_{t-1}$ by obtaining $\frac{\partial \lambda_t}{\partial t} > 0$ for $\forall t > 1$ through induction method. The detailed proof can be found in the appendix. □

We can easily compute λ_0, λ_1, and we have $\forall T > 1 : \lambda_{T+1} > \lambda_T > \lambda_0 = 0$ according to Theorem 2, and $\lim_{T\to\infty} \lambda_T \to TW(\omega)$, where $TW(\omega)$ is the Whittle index

value for state ω under infinite horizon. Hence, we can use a sigmoid curve to approximate the index value. One common example of a sigmoid function is the logistic function. This form is also used by Mate et al. [2021a]. Specifically, we let

$$W_T(\omega) = \frac{A}{1 + e^{-kT}} + C, \tag{10}$$

where A and $\frac{A}{2} + C$ are the curve's bounds; k is the logistic growth rate or steepness of the curve. Recall that the definition of the Whittle index $W_T(\omega)$ of belief state ω is the smallest λ s.t. it is optimal to make the arm passive in the current state. We have $W_0(\omega) = 0$ and $W_1(\omega) = \beta(\omega(P^a_{1,1} - P^p_{1,1}) + (1-\omega)(P^a_{0,1} - P^p_{0,1}))$, and $W_\infty(\omega) = TW(\omega)$. By solving these three constraints, we can get the three unknown parameters,

$$C = -TW(\omega), A = 2TW(\omega),$$

$$k = -\log(\frac{2TW(\omega)}{\beta(\omega(P^a_{1,1}-P^p_{1,1})+(1-\omega)(P^a_{0,1}-P^p_{0,1}))+TW(\omega)} - 1)$$

Algorithm 1 shows how to use $W_T(\omega)$ in considering fairness constraint under the finite horizon setting. Next, we show that like in the infinite horizon case, value function and Whittle index decay over time in the case of the finite horizon.

Theorem 3 *Consider the finite horizon RMAB problem with fairness constraint. Algorithm 1 (activating arm i at the end of the time period when its fairness constraint is violated) is optimal:*

1. *When $\omega^i \leq \omega^*$ (increasing belief process), if*

$$(P^{i,p}_{1,1} - P^{i,p}_{0,1}) \left(\Delta_4 \beta \sum_{t=0}^{T-2} [\beta^t] + 1 \right) \leq (P^{i,a}_{1,1} - P^{i,a}_{0,1}) \sum_{t=0}^{T-2} [\beta^t (P^{i,a}_{1,1} - P^{i,a}_{0,1})^t] \tag{11}$$

$\Delta_4 = \min\{(P^{i,p}_{1,1} - P^{i,p}_{0,1}), (P^{i,a}_{1,1} - P^{i,a}_{0,1})\}$*, and T is the residual horizon length.*

2. *When $\omega^i \geq \omega^*$ (non-increasing belief process), if*

$$(P^{i,p}_{1,1} - P^{i,p}_{0,1}) \left(\Delta_2 \beta \sum_{t=0}^{T-2} [\beta^t (P^{i,a}_{1,1} - P^{i,a}_{0,1})^t] + 1 \right) \geq (P^{i,a}_{1,1} - P^{i,a}_{0,1}) \sum_{t=0}^{T-2} \beta^t \tag{12}$$

$\Delta_2 = \min\{(P^{i,p}_{1,1} - P^{i,p}_{0,1}), (P^{i,a}_{1,1} - P^{i,a}_{0,1})\}$.

Proof Sketch. The proof is similar to the infinite horizon case (detailed in Appendix). □

4.3 UNCERTAINTY IN TRANSITION MATRIX

In most real-world applications [Biswas et al., 2021], there may not be adequate information about all the state transitions. In such cases, we don't know how likely a transition is and thus, we won't be able to use the Whittle index approach directly. We provide a mechanism to apply the Thompson sampling based learning mechanism for solving RMAB problems without prior knowledge and where it is feasible to get learning experiences. Thompson sampling [Thompson, 1933] is an algorithm for online decision problems, and can be applied in MDP [Gopalan and Mannor, 2015] as well as Partially Observable MDP [Meshram et al., 2016]. In Thompson sampling, we initially assume that arm has a prior Beta distribution in the transition probability according to the prior knowledge (if there is no prior knowledge available, we assume a prior $Beta(1, 1)$ as this is the uniform distribution on $(0, 1)$). We choose Beta distribution because it is a convenient and useful prior option for Bernoulli rewards [Agrawal and Goyal, 2012].

In our algorithm, referred to as FaWT-U and provided in 2, at each time step, we sample the posterior distribution over the parameters, and then use the Whittle index algorithm to select the arm with the highest index value to play if the fairness constraint is not violated. We can utilize our observations to update our posterior distribution, because playing the selected arms will reveal their current state. Then, the algorithm takes samples from the posterior distribution and repeats the procedure again.

Algorithm 2: Fair Whittle Thresholding with Uncertainty in transition matrix(FaWT-U)

Input: Posterior $Beta$ distribution over the transition matrix $\mathcal{P}$, fairness constraint, η and L, set of belief states $\{\omega^1, \ldots, \omega^N\}$, budget k

1 **for** *each arm i in* 1 *to* N **do**
2 Sample the transition probability parameters independently from posterior;
3 Compute Whittle indices based on the transition matrix and belief state;
4 **if** *the activation frequency η for arm i is not satisfied at the end of the period of length L* **then**
5 Add arm i to the action set ϕ;
6 $k = k - 1$;
7 Add the arms with top k index value into ϕ;
8 Play the selected arms and receive the observations;
9 Update the posterior distribution;

Output: Action set ϕ and updated posterior distribution over parameters

We employ the sampled transition probabilities and belief states $\{\omega^1, \ldots, \omega^N\}$, as well as the residual time horizon T as the input to the Whittle index computation (Line 3 in Algorithm 2).

4.4 UNKNOWN TRANSITION MATRIX

We now tackle the second challenge mentioned, in which the transition matrix is completely unknown. In this case, we can take advantage of the model-free learning method to avoid directly using the whittling index policy.

Q-Learning is most commonly used to solve the sequential decision-making problem, which was first introduced by Watkins [1989] as an early breakthrough in reinforcement learning. It has also been extensively used in RMAB problems Fu et al. [2019], Avrachenkov and Borkar [2020], Biswas et al. [2021] to estimate the expected Q-value, $Q^*(s, a, l)$, of taking action $a \in \{0, 1\}$ after $l \in \{1, \ldots, L\}$ time slots since last observation $s \in \{0, 1\}$. The off-policy TD control algorithm is defined as

$$Q^{t+1}(s_t, a_t, l_t) \leftarrow Q^t(s_t, a_t, l_t) + \alpha_t(s_t, a_t, l_t)\left[R_{t+1} + \gamma \max_a \left(Q^t(s_{t+1}, a, l_{t+1}) - Q^t(s_t, a_t, l_t)\right)\right] \tag{13}$$

Where γ is the discount rate, $\alpha_t(s_t, a_t, l_t) \in [0, 1]$ is the learning rate parameter, i.e., a small $\alpha_t(s_t, a_t, l_t)$ will result in a slow learning process and no update when $\alpha_t(s_t, a_t, l_t) = 0$. While a large $\alpha_t(s_t, a_t, l_t)$ may cause the estimated Q-value to rely heavily on the most recent return, when $\alpha_t(s_t, a_t, l_t) = 1$, the Q-value will always be the most recent return.

We now describe how to use the Whittle index-based Q-Learning mechanism to solve the RMAB problem with fairness constraints. We build on the work by Biswas et al. [2021] for fully observable settings. In addition to considering fairness constraints, our model can be viewed as an extension to the partially observable setting. Due to fairness constraints, l can be a maximum of L time steps. Therefore, belief space is also limited. We are able to use the Q-Learning based approach to effectively compute the Whittle index value and this approach is summarized in Algorithm 3,

One typical form of $\alpha_t(s_t, a_t, l_t)$ could be $1/z(s_t, a_t, l_t)$, where $z(s_t, a_t, l_t) = \left(\sum_{u=0}^{t} \mathbb{I}\{s_u = s, a_u = a, l_u = l\}\right) + 1$ for each initial observed state $s \in \{0, 1\}$, action $a \in \{0, 1\}$ and time length since last activation $l \in \{1, \ldots, L\}$ at the time slot u from the beginning. With such mild form of $\alpha_t(s_t, a_t, l_t)$, we now are able to build the theoretical support for the Q-Learning based Whittle index approach.

Theorem 4 *Selecting the highest-ranking arms according to the $Q_i^*(s, a = 1, l) - Q_i^*(s, a = 0, l)$ till the budget constraint is met is equivalent to maximizing $\left\{\sum_{i=1}^{N} Q_i^*(s, a, l)\right\}$ over all possible action set $\{0, 1\}^N$ such that $\sum_{i=1}^{N} a_i = k$.*

Proof Sketch. A proof based on work by Biswas et al. [2021] is given in Appendix D.3. □

Algorithm 3: Fair Whittle Thresholding based Q-Learning(FaWT-Q)

Input: parameter ϵ and k, and $\alpha_t(s_t, a_t, l_t)$, initial observed state set $\{s\}^N$,

for *each arm i in* 1 *to* N **do**
  Initialize the $Q_i(s, a, l) \leftarrow 0$ for each state $s \in \{0, 1\}$, and each action $a \in \{0, 1\}$ and time length $l \in \{1, \ldots, L\}$;
  For each $s \in \{0, 1\}$ and $l \in \{1, \ldots, L\}$ initialize the Whittle index value set $\lambda_i(s, l) \leftarrow 0$;
for *t from* 1 *to* T **do**
  for *arm i in 1 to N* **do**
    if *the fairness constraint is violated* **then**
      Add arm i to the action set ϕ;
      $k = k - 1$;
  With prob ϵ add random k arms to ϕ and with prob $1 - \epsilon$ add arms with top k $\lambda_i(s, l)$ value ;
  Activate the selected arms and receive rewards and observations;
  for *each arm i in* 1 *to* N **do**
    Update the $Q_i^{t+1}(s, a, l)$ according to Eq. 13;
    if $i \in \phi$ **then**
      Set $l = 1$ and update s_i according to the received observation;
    else
      Set $l = l + 1$;
    Update the new Q-Learning based Whittle index by $\lambda_i^{t+1}(s, l) = Q_i(s, a = 1, l) - Q_i(s, a = 0, l)$

Output: Action set ϕ

Theorem 5 *Stability and convergence: The proposed approach converges to the optimal with probability* 1 *under the following conditions:*
1. The state space and action space are finite;
2. $\sum_{t=1}^{\infty} \alpha_t(s_t, a_t, l_t) = \infty \quad \sum_{t=1}^{\infty} \alpha_t^2(\omega_i(t)) < \infty$

Proof Sketch. The key to the convergence is contingent on a particular sequence of episodes observed in the real process Watkins and Dayan [1992]. Detailed proof is given in Appendix D.4. □

5 EXPERIMENT

To the best of our knowledge, we are the first to explore fairness constraints in RMAB, hence the goal of the experiment section is to evaluate the performance of our approach in comparison to existing baselines:
Random: At each round, decision-maker randomly select k arms to activate.
Myopic: Select k arms that maximize the expected reward at the immediate next round. A myopic policy ignores the impact of present actions on future rewards and instead focuses entirely on predicted immediate returns. Formally, this

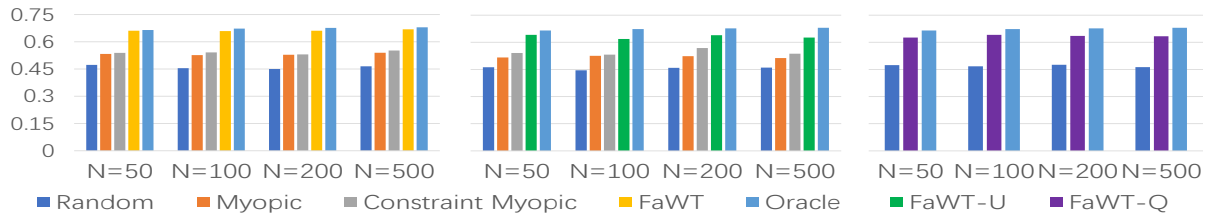

Figure 6: Comparison of performance of our approach and baseline approaches

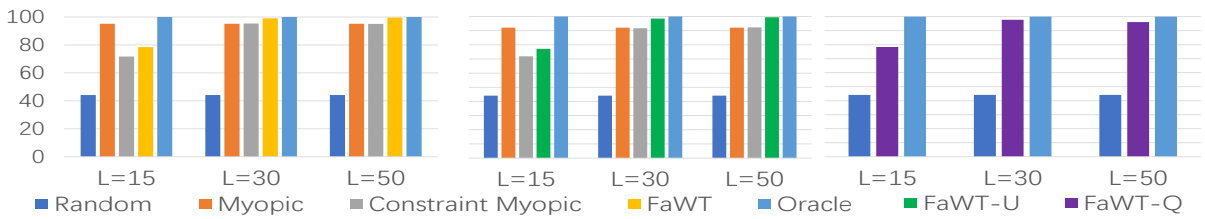

Figure 7: Intervention benefit ratio of our approach and baseline approaches without penalty for the violation of the fairness constraint. We set $N = 100$, $k = 10$, $T = 1000$, $\eta = 2$ and $L = \{15, 30, 50\}$.

could be described as choosing the k arms with the largest gap $\Delta\omega_t = (\omega_{t+1}|a_t = 1) - (\omega_{t+1}|a_t = 0)$ at time t.

Constraint Myopic: It is the same as the Myopic when there is no conflict with fairness constraints, but if the fairness constraint is violated, it will choose the arm that satisfies the fairness constraint to play.

Oracle: Algorithm by Qian et al. [2016] under the assumption that the states of all arms are fully observable and the transition probabilities are known without considering fairness constraints.

To demonstrate the performance of our proposed methods, we test our algorithms on synthetic domains [Mate et al., 2020] and provide numerical results averaged over 50 runs.

Average reward value with penalty: In Figure 6, we show the average reward $\bar{R}$ at each time step received by an arm over the time interval $T = 1000$ for $N = 50, 100, 200, 500$ and $k = 10\% \times N$ with the fairness constraint $L = 20$, and $\eta = 2$. We will receive a reward of 1 if the state of an arm is $s = 1$, and no reward otherwise. We impose a small penalty of -0.01 if the fairness constraint of an arm is not satisfied. The graph on the left shows the performance of FaWT method when assuming the transition matrix is known. The middle graph is the average reward obtained using the FaWT-U approach when the transition model is not fully available. The right graph illustrates the result of FaWT-Q method when the transition model is unknown. As shown in the figure, our approaches consistently outperform the Random and Myopic baselines, and in addition to satisfying the fairness constraints, they have a near-optimal performance with a small difference gap when compared to the Oracle baseline. Note that the Myopic approach may fail in some cases(shown in Mate et al. [2020]), it performs worse than the Random approach.

No penalty for the violation of the fairness constraint: We also investigate the intervention benefit ratio defined as $\frac{\bar{R}_{\text{method}} - \bar{R}_{\text{No intervention}}}{\bar{R}_{\text{Oracle}} - \bar{R}_{\text{No intervention}}} \times 100\%$, where $\bar{R}_{\text{No intervention}}$ denotes the average reward without any intervention involved. Here, we do not employ penalties when the fairness constraint is not satisfied, as we want to evaluate the benefit provided by interventions with our fair policy and policies of other approaches. We provide the intervention benefit ratio for different values of L for all approaches in Figure 7. Again, the left graph shows the result of FaWT approach, the middle graph is the result of FaWT-U approach, and the right graph shows the result of FaWT-Q method. Our proposed approaches can achieve a better intervention benefit ratio compared with the baseline when L is 30 and above. However, for L = 15, where there is a strict fairness constraint (i.e., $\frac{k \times L}{(\eta - 1) \times N}$ is close to 1), it has a significant impact on solution quality. The performances of all our approaches improve when the fairness constraint's strength decreases (L increases). Overall, our proposed methods can handle various levels of fairness constraint strength without sacrificing significantly on solution quality.

We also provide the additional experiment result that studies the influence of intervention level and fairness constraint's strength in the Appendix.

6 CONCLUSION

In this paper, we initiate the study of fairness constraints in Restless Multi-Arm Bandit problems. We define a fairness metric that encapsulates and generalizes existing fairness definitions employed for Multi-Arm Bandit problems. Contrary to expectations, we are able to provide minor modifications to the existing algorithm for RMAB problems in order to handle fairness. We provide theoretical results on how our methods provide the best way to handle fairness without sacrificing solution quality. This is demonstrated empirically as well on benchmark problems from the literature.

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Konstantin Avrachenkov and Vivek S Borkar. Whittle index based q-learning for restless bandits with average reward. *arXiv preprint arXiv:2004.14427*, 2020.

Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. *arXiv preprint arXiv:2105.07965*, 2021.

Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pages 181–190. PMLR, 2020.

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

Kevin D Glazebrook, Diego Ruiz-Hernandez, and Christopher Kirkbride. Some indexable families of restless bandit problems. *Advances in Applied Probability*, 38(3): 643–672, 2006.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Conference on Learning Theory*, pages 861–898. PMLR, 2015.

Morris W Hirsch. Convergent activation dynamics in continuous time networks. *Neural networks*, 2(5):331–349, 1989.

Weici Hu and Peter Frazier. An asymptotically optimal index policy for finite-horizon restless bandits. *arXiv preprint arXiv:1707.00205*, 2017.

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201, 1994.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *arXiv preprint arXiv:1605.07139*, 2016.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Elliot Lee, Mariel S Lavieri, and Michael Volk. Optimal screening for hepatocellular carcinoma: A restless bandit model. *Manufacturing & Service Operations Management*, 21(1):198–212, 2019.

Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2019.

Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, 2010.

Aditya Mate, Jackson A Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health interventions. *arXiv preprint arXiv:2007.04432*, 2020.

Aditya Mate, Arpita Biswas, Christoph Siebenbrunner, and Milind Tambe. Efficient algorithms for finite horizon and streaming restless multi-armed bandit problems. *arXiv preprint arXiv:2103.04730*, 2021a.

Aditya Mate, Andrew Perrault, and Milind Tambe. Risk-aware interventions in public health: Planning with restless multi-armed bandits. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 880–888, 2021b.

Rahul Meshram, Aditya Gopalan, and D Manjunath. Optimal recommendation to users that react: Online learning for a class of pomdps. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 7210–7215. IEEE, 2016.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5379–5386, Apr. 2020. doi: 10.1609/aaai.v34i04.5986.

Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 123–131, 2016.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Sofía S Villar. Indexability and optimal index policies for a class of reinitialising restless bandits. *Probability in the engineering and informational sciences*, 30(1):1–23, 2016.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, pages 287–298, 1988.

A WHITTLE INDEX POLICY

We take advantage of the Fast Whittle Index Computation algorithm introduce in Mate et al. [2020]. They derived a closed form for computing the Whittle index for both average reward and discounted reward criterion, where the objective could also be written as $\bar{R}_\lambda^\pi = \mathbb{E}\sum_\omega f^\pi(\omega) R_{a^\pi}(\omega)$, where $f^\pi(\omega)$ is defined as the fraction of time spent in each belief state ω induced by policy π and $f^\pi(\omega) \in [0,1]$. Their proposed Whittle index computation algorithm can achieve a 3-order-of-magnitude speedup compared to Qian et al. [2016]. In the two-states setting ($s \in \{0,1\}$), they use a tuple $(B_0^{\omega_{th}}, B_1^{\omega_{th}})$ to denote the belief threshold, where $\omega_{th} \in [0,1]$, and $B_0^{\omega_{th}}, B_1^{\omega_{th}} \in 1,\ldots,L$ are the index of the first belief state in each chain where it is optimal to act (i.e., the belief is less than or equal to ω_{th}). The length is at most L long due to our fairness constraints. This is defined as the forward threshold policy, and Mate et al. [2020] used the Markov chain structure to derive the occupancy frequencies for each belief state $\omega_s(t)$, which is as follows,

$$f^{(B_0^{\omega_{th}}, B_1^{\omega_{th}})}(\omega_s(t)) = \begin{cases} a & \text{if } s=0, t \le B_0 \\ b & \text{if } s=1, t \le B_1 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

$$a = \left(\frac{B_1\omega_0(B_0)}{1-\omega_1(B_1)} + B_0\right)^{-1}, b = \left(\frac{B_1\omega_0(B_0)}{1-\omega_1(B_1)} + B_0\right)^{-1} \frac{\omega_0(B_0)}{1-\omega_1(B_1)} \tag{15}$$

These occupancy frequencies do not depend on the subsidy λ. For the forward threshold policy $(B_0^{\omega_{th}}, B_1^{\omega_{th}})$, they use the $R_\lambda^{B_0^{\omega_{th}}, B_1^{\omega_{th}}}$ to denote the average reward, then can decompose the average reward into the contribution of the state reward and the subsidy λ

$$\begin{aligned} R_\lambda^{(B_0^{\omega_{th}}, B_1^{\omega_{th}})} &= \sum_{\omega \in \mathscr{B}} \omega f^{(B_0^{\omega_{th}}, B_1^{\omega_{th}})}(\omega) \\ &+ \omega\left(1 - f^{(B_0^{\omega_{th}}, B_1^{\omega_{th}})}(\omega_1(B_1)) - f^{(B_0^{\omega_{th}}, B_1^{\omega_{th}})}(\omega_0(B_0))\right) \end{aligned} \tag{16}$$

Given the definition of the Whittle index λ, this could be interpreted to two corresponding threshold policies being equally optimal. More specifically, for a belief state $\omega_0(B_0)$, the two adjacent threshold polices $\{(B_0^{\omega_{th}}, B_1^{\omega_{th}}), (B_0^{\omega_{th}}+1, B_1^{\omega_{th}})\}$ would be optimal to be active and passive respectively. recall that the Whittle index is the smallest λ for which the active and the passive actions are both optimal. Thus the subsidy which makes the average reward of those two adjacent polices equal in value must be the Whittle index for the belief state $\omega_0(B_0)$. Formally, this could be calculated through $R_{(\lambda}^{B_0^{\omega_{th}}, B_1^{\omega_{th}})} = R_{(\lambda}^{B_0^{\omega_{th}}, B_1^{\omega_{th}}+1)}$. Similarly, we can obtain the Whittle index for the belief state $\omega_1(B_1)$ through $R_\lambda^{(B_0^{\omega_{th}}, B_1^{\omega_{th}})} = R_\lambda^{(B_0^{\omega_{th}}, B_1^{\omega_{th}}+1)}$. These computations are repeated for every belief states to find the minimum subsidy value while $B_s^{\omega_{th}} \le L$. The main idea of their approach is shown in Fig 8.

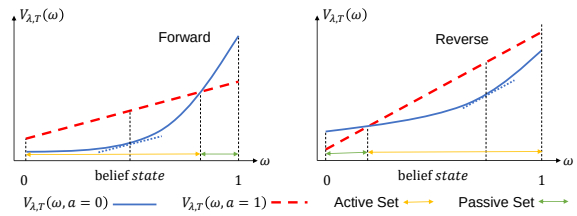

Figure 8: The forward and reverse policy

B FULLY OBSERVABLE SETTING

In order to explain the key idea, we initially consider a fully observable environment and then extend to a partially observable setting. Algorithm 1 provides the algorithm at each iteration for selecting the arms to activate (action set) in fully observable case (for both finite and infinite hoirzon cases).

B.1 INFINITE HORIZON, FULLY OBSERVABLE

We first provide the expression for λ without the fairness constraints. We assume that $V_{\lambda,\infty}(s)$ denotes the value function which can be accrued from a single-armed bandit process with subsidy λ over infinite horizon if the observed state is s. Therefore, for the state 0, we have:

$$\begin{aligned} V_{\lambda,\infty}(0) &= \max \begin{cases} \lambda + \beta(P_{0,1}^p V_{\lambda,\infty}(1) + P_{0,0}^p V_{\lambda,\infty}(0)) & \text{passive} \\ \beta(P_{0,1}^a V_{\lambda,\infty}(1) + P_{0,0}^a V_{\lambda,\infty}(0)) & \text{active} \end{cases} \\ &= \max \begin{cases} \lambda + \beta(P_{0,1}^p (V_{\lambda,\infty}(1) - V_{\lambda,\infty}(0)) + V_{\lambda,\infty}(0)) \\ \beta(P_{0,1}^a (V_{\lambda,\infty}(1) - V_{\lambda,\infty}(0)) + V_{\lambda,\infty}(0)) \end{cases} \end{aligned} \tag{17}$$

Similarly for state 1, we have

$$V_{\lambda,\infty}(1) = \max \begin{cases} \lambda + 1 + \beta(P_{1,1}^p (V_{\lambda,\infty}(1) - V_{\lambda,\infty}(0)) + V_{\lambda,\infty}(0)) & \text{passive} \\ 1 + \beta(P_{1,1}^a (V_{\lambda,\infty}(1) - V_{\lambda,\infty}(0)) + V_{\lambda,\infty}(0)) & \text{active} \end{cases} \tag{18}$$

Recall that the definition of the Whittle index $W_T(s)$ of state s is the smallest λ s.t. it is optimal to make the arm passive in the current state. Therefore, we have $\lambda = \beta\left((P_{0,1}^a - P_{0,1}^p)(V_{\lambda,\infty}(1) - V_{\lambda,\infty}(0))\right)$ for $s=0$ and, $\lambda = \beta\left((P_{1,1}^a - P_{1,1}^p)(V_{\lambda,\infty}(1) - V_{\lambda,\infty}(0))\right)$ for $s=1$. As a result, the Whittle index based approach would rank the index value of all N arms and select top k arms at each time step to activate. With fairness constraints, the change to the approach is minimal and intuitive. The optimal policy is still to choose the arms with the top "k" index values until a fairness constraint is violated for an arm. In that time step, we replace the last arm in top-k with the arm for which fairness constraint is violated. We show that this simple change works across the board for infinite and finite horizon, fully and partially observable settings.

Theorem 6 *Algorithm 1 is optimal for RMAB with Fairness Constraints governed by parameters η and L under certain conditions.*

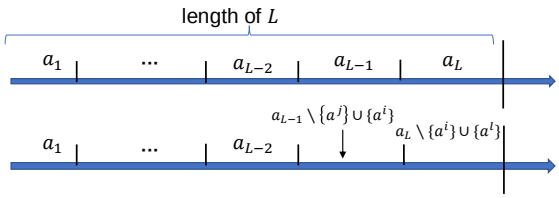

Figure 9: The action is a_t at time step t. Then we move the action a^i that satisfies the fairness constraint to one slot earlier to replace the $k-th$ action a^j, and add the action a^l according to the index value at the end of the time interval.

Proof Sketch. This can be viewed as a special case of the partially observable setting. Please refer to the detailed proof for the partially observable case. □

B.2 FINITE HORIZON, FULLY OBSERVABLE

Existing literature Glazebrook et al. [2006], Villar [2016] on RMAB deals with situations where the time horizon is infinite. In this part, we will demonstrate how the methodology discussed for infinite horizon setting to deal with fairness can also be applied to the finite horizon setting. There are two key challenges: (i) computing the Whittle index under the finite horizon setting; (ii) to show that Whittle index value reduces as residual life time decreases;

It is costly to compute the index under the finite horizon setting ($O(|S|^k T)$ time and space complexity Hu and Frazier [2017]; instead, we can take advantage of the fact that the index value will converge as $T \to \infty$ and $\lambda_0 = 0$. We can use a sigmoid curve to approximate the index value. One common example of a sigmoid function is the logistic function. This form is also used by Mate et al. [2021a]. Specifically, we let

$$W_T(s) = \frac{A}{1+e^{-kT}} + C, \tag{19}$$

where A and $\frac{A}{2} + C$ are the curve's bounds; k is the logistic growth rate or steepness of the curve. We have $W_0(s) = 0$ and $W_1(s) = \beta(s(P^a_{1,1} - P^p_{1,1}) + (1-s)(P^a_{0,1} - P^p_{0,1}))$, and let $W_\infty(s) = TW(s)$, where $TW(s)$ is the Whittle index value for state s under infinite horizon. By solving these three constraints, we can get the three unknown parameters,

$$C = -TW(s), A = 2TW(s),$$

$$k = -\log\left(\frac{2TW(s)}{\beta(s(P^a_{1,1}-P^p_{1,1})+(1-s)(P^a_{0,1}-P^p_{0,1}))+TW(s)} - 1\right)$$

Algorithm 1 shows how to use $W_T(s)$ in considering fairness constraint under the finite horizon setting. As for optimality of Algorithm 1 in case of finite horizon, we have to show that value function and Whittle index decays over time in case of finite horizon, i.e.:

$$V_{\lambda_T,T}(s) > V_{\lambda_{T-1},T-1}(s), \forall s \in \{0,1\}$$

$$\forall T > 1 : \lambda_T > \lambda_{T-1} > \lambda_0 = 0$$

This is to make the same argument as in case of infinite horizon, where we show that activating the fairness ensuring arm one step earlier results in lower value.

We show a lemma first, which can lead to the proof of the Whittle index decay.

Lemma 1 *Intuitively, we can have* $V_{\lambda_T,T}(s) > V_{\lambda_{T-1},T-1}(s)$, *for* $\forall s \in \{0,1\}$.

Proof. For state $s \in \{0,1\}$, we can always find a policy that ensures $V_{\lambda_T,T}(s) > V_{\lambda_{T-1},T-1}(s)$. For example, we assume the optimal policy for the state s with the residual time horizon $T-1$ is π, we can always find a policy π': keep the same policy for the first $T-1$ time slot for the same state s with the residual time horizon T, and then pick the action for the last slot T according to the observed state s'. Because the reward is either 0 or 1, we could have

$$V_{\lambda_T,T;\pi'}(s) = V_{\lambda_{T-1},T-1;\pi}(s) + \beta^T V_{\lambda_1,1}(s') > V_{\lambda_{T-1},T-1;\pi}(s). \tag{20}$$

□

Note this lemma is also suitable for the partially observable case. Then we present the Theorem and corresponding proof.

Theorem 7 *For a finite horizon T, the Whittle index λ_T is the value that satisfies the equation $V_{\lambda_T,T}(s, a=0) = V_{\lambda_T,T}(s, a=1)$ for the observed state s. Assuming indexability holds, the Whittle index will decay as the value of horizon T decreases: $\forall T > 1 : \lambda_T > \lambda_{T-1} > \lambda_0 = 0$.*

Proof Sketch. Because the state s is fully observable, We can easily calculate λ_0 and λ_1 by solving equation $V^p_{\lambda,T}(s) = V^a_{\lambda,T}(s)$. We then could derive $\lambda_T > \lambda_{T-1}$ by obtaining $\frac{\partial \lambda_T}{\partial T} > 0$ for $s \in \{0,1\}$. Detailed proof in appendix. □

Proof. Consider the discount reward criterion with discount factor of $\beta \in [0,1]$ (where $\beta = 1$ corresponds to the average criterion).

$$\begin{aligned} &s + \lambda + \beta(sP^p_{1,1} + (1-s)P^p_{0,1}) = s + \beta(sP^a_{1,1} + (1-s)P^a_{0,1}) \\ &\to \lambda = \beta(s(P^a_{1,1} - P^p_{1,1}) + (1-s)(P^a_{0,1} - P^p_{0,1})) \end{aligned} \tag{21}$$

Because $P^a_{1,1} - P^p_{1,1} > 0$ and $P^a_{0,1} - P^p_{0,1} > 0$, we have $\lambda > 0$. Now we show $\lambda_T > \lambda_{T-1}$. Equivalently, this can be expressed as $\frac{\partial \lambda_T}{\partial T} > 0$. Because the state is fully observable, we first get the close form of λ_T.

- `Case 1`: The state $s = 0$,

$$\begin{aligned}&0 + \lambda_T + \beta(P^p_{0,0}V_{\lambda_{T-1},T-1}(0) + P^p_{0,1}V_{\lambda_{T-1},T-1}(1))\\ =&0 + \beta(P^a_{0,0}V_{\lambda_{T-1},T-1}(0) + P^a_{0,1}V_{\lambda_{T-1},T-1}(1))\\ \rightarrow&\lambda_T = \beta(V_{\lambda_{T-1},T-1}(0)(P^a_{0,0} - P^p_{0,0}) + V_{\lambda_{T-1},T-1}(1)(P^a_{0,1} - P^p_{0,1})).\end{aligned} \tag{22}$$

Intuitively, we can have $V_{\lambda_T,T}(0) > V_{\lambda_{T-1},T-1}(0)$ (see Lemma 1), and $V_{\lambda_T,T}(1) > V_{\lambda_{T-1},T-1}(1)$, we obtain $\frac{\partial V_{\lambda_T,T}(0)}{\partial T} > 0$ and $\frac{\partial V_{\lambda_T,T}(1)}{\partial T} > 0$. Hence, we can get

$$\frac{\partial \lambda_T}{\partial T} = \beta((P^a_{0,0} - P^p_{0,0})\frac{\partial V_{\lambda_{T-1},T-1}(0)}{\partial T} + (P^a_{0,1} - P^p_{0,1})\frac{\partial V_{\lambda_{T-1},T-1}(1)}{\partial T}) > 0 \tag{23}$$

- `Case 2`: The state $s = 1$,

$$\begin{aligned}&1 + \lambda_T + \beta(P^p_{1,0}V_{\lambda_{T-1},T-1}(0) + P^p_{1,1}V_{\lambda_{T-1},T-1}(1))\\ =&1 + \beta(P^a_{1,0}V_{\lambda_{T-1},T-1}(0) + P^a_{1,1}V_{\lambda_{T-1},T-1}(1))\\ \rightarrow&\lambda_T = \beta(V_{\lambda_{T-1},T-1}(0)(P^a_{1,0} - P^p_{1,0}) + V_{\lambda_{T-1},T-1}(1)(P^a_{1,1} - P^p_{1,1})).\end{aligned} \tag{24}$$

Similarly, we have

$$\frac{\partial \lambda_T}{\partial T} = \beta((P^a_{1,0} - P^p_{1,0})\frac{\partial V_{\lambda_{T-1},T-1}(0)}{\partial T} + (P^a_{1,1} - P^p_{1,1})\frac{\partial V_{\lambda_{T-1},T-1}(1)}{\partial T}) > 0 \tag{25}$$

Thus $\forall T > 1 : \lambda_T > \lambda_{T-1} > \lambda_0 = 0$. □

The proof of optimality for Algorithm 1 in case of finite horizon is similar to the infinite horizon setting.

Most of the time, the states $\{s_1, \ldots, s_N\} \in \{0, 1\}^N$ are not observable until the action is taken. The decision-maker, on the other hand, can infer the current states based on the observation history by keeping track of the belief states ω.

C PROOFS FOR THE THEOREM 1 AND THEOREM 3

C.1 BOUNDARY LEMMA

Lemma 2 *For the finite time horizon T, $V_{\lambda_T,T}(\omega_1) - V_{\lambda_T,T}(\omega_2)$ is bounded, where we have*

$$(\omega_1 - \omega_2)\sum_{t=0}^{T-1}\beta^t(P^a_{1,1} - P^a_{0,1})^t \le V_{\lambda_T,T}(\omega_1) - V_{\lambda_T,T}(\omega_2) \le (\omega_1 - \omega_2)\sum_{t=0}^{T-1}\beta^t \tag{26}$$

Proof. We prove the lower bound by induction, and the upper bound can be proven similarly.

When $T = 1$, we start from the definition of the value function $V_{\lambda_T,T}(\omega)$ to have

- passive actions:

$$\begin{aligned}V_{\lambda_1,1}(\omega_1, a = 0) - V_{\lambda_1,1}(\omega_2, a = 0) &= \lambda_1 + \omega_1 - \lambda_1 - \omega_2\\ &= \omega_1 - \omega_2\end{aligned} \tag{27}$$

- active actions:

$$V_{\lambda_1,1}(\omega_1, a = 1) - V_{\lambda_1,1}(\omega_2, a = 1) = \omega_1 - \omega_2 \tag{28}$$

We get $V_{\lambda_1,1}(\omega_1) - V_{\lambda_1,1}(\omega_2) = \omega_1 - \omega_2$. Now we assume $V_{\lambda_T,T}(\omega_1) - V_{\lambda_T,T}(\omega_2) \ge (\omega_1 - \omega_2)\sum_{t=0}^{T-1}\beta^t(P^a_{1,1} - P^a_{0,1})^t$ hold for $\forall T > 1$, then for time horizon $T + 1$, we have

- passive actions:

$$\begin{aligned}&V_{\lambda_{T+1},T+1}(\omega_1, a = 0) - V_{\lambda_{T+1},T+1}(\omega_2, a = 0)\\ =&(\lambda_1 + \omega_1 + \beta V_{\lambda_T,T}(\omega_1(1))) - (\lambda_1 + \omega_2 + \beta V_{\lambda_T,T}(\omega_2(1)))\\ =&\omega_1 - \omega_2 + \beta\left(V_{\lambda_T,T}(\omega_1(1)) - V_{\lambda_T,T}(\omega_2(1))\right)\\ \ge&\omega_1 - \omega_2 + \beta(\omega_1 - \omega_2)\sum_{t=0}^{T-1}\beta^t(P^a_{1,1} - P^a_{0,1})^t\\ \ge&\omega_1 - \omega_2 + (\omega_1 - \omega_2)\sum_{t=1}^{T}\beta^t(P^a_{1,1} - P^a_{0,1})^t \text{ Line } *\\ =&(\omega_1 - \omega_2)\sum_{t=0}^{T}\beta^t(P^a_{1,1} - P^a_{0,1})^t\end{aligned} \tag{29}$$

Line $*$ is because $0 \le P^a_{1,1} - P^a_{0,1} < 1$.

- active actions:

$$\begin{aligned}&V_{\lambda_{T+1},T+1}(\omega_1, a = 1) - V_{\lambda_{T+1},T+1}(\omega_2, a = 1)\\ =&\left(\omega_1 + \beta(\omega_1 V_{\lambda_T,T}(P^a_{1,1}) + (1 - \omega_1)V_{\lambda_T,T}(P^a_{0,1}))\right)\\ &- \left(\omega_2 + \beta(\omega_2 V_{\lambda_T,T}(P^a_{1,1}) + (1 - \omega_2)V_{\lambda_T,T}(P^a_{0,1}))\right)\\ =&\omega_1 - \omega_2 + \beta\left((\omega_1 - \omega_2)(V_{\lambda_T,T}(P^a_{1,1}) - V_{\lambda_T,T}(P^a_{0,1}))\right)\\ =&\omega_1 - \omega_2 + \beta(V_{\lambda_T,T}(\omega_1(1)) - V_{\lambda_T,T}(\omega_2(1)))\\ \ge&\omega_1 - \omega_2 + \beta(\omega_1 - \omega_2)\sum_{t=0}^{T-1}\beta^t(P^a_{1,1} - P^a_{0,1})^t\\ \ge&\omega_1 - \omega_2 + (\omega_1 - \omega_2)\sum_{t=1}^{T}\beta^t(P^a_{1,1} - P^a_{0,1})^t \text{ Line } *\\ =&(\omega_1 - \omega_2)\sum_{t=0}^{T}\beta^t(P^a_{1,1} - P^a_{0,1})^t\end{aligned} \tag{30}$$

Line $*$ is because $0 \le P^a_{1,1} - P^a_{0,1} < 1$. Thus we have $V_{\lambda_T,T}(\omega_1) - V_{\lambda_T,T}(\omega_2) \ge (\omega_1 - \omega_2)\sum_{t=0}^{T-1}\beta^t(P^a_{1,1} - P^a_{0,1})^t$. Similarly, we can prove the lower bound. □

Lemma 3 *For the infinite residual time horizon $T \to \infty$, $V_{\lambda_T,T}(\omega_1) - V_{\lambda_T,T}(\omega_2)$is bounded. Specifically, we have*

$$\frac{\omega_1 - \omega_2}{1 - \beta(P^a_{1,1} - P^a_{0,1})} \le V_{\lambda_T,T}(\omega_1) - V_{\lambda_T,T}(\omega_2) \le \frac{\omega_1 - \omega_2}{1 - \beta} \tag{31}$$

Proof. This can be viewed as a special case of the finite residual time horizon setting where $T \to \infty$. Thus we can

easily derive the lower and upper bound according to the formula for the geometric series:

$$\lim_{T\to\infty}(\omega_1-\omega_2)\sum_{t=0}^{T-1}\beta^t(P^a_{1,1}-P^a_{0,1})^t=\frac{\omega_1-\omega_2}{1-\beta(P^a_{1,1}-P^a_{0,1})}$$

and

$$\lim_{T\to\infty}(\omega_1-\omega_2)\sum_{t=0}^{T-1}\beta^t=\frac{\omega_1-\omega_2}{1-\beta}$$

□

Consider the single-armed bandit process with subsidy λ under the infinite time horizon $T\to\infty$, we have:

$$V_{\lambda,\infty}(\omega)=\max\begin{cases}\lambda+\omega+\beta V_{\lambda,\infty}(\tau^1(\omega)) & \text{passive}\\ \omega+\beta\left(\omega V_{\lambda,\infty}(P^a_{1,1})+(1-\omega)V_{\lambda,\infty}(P^a_{0,1})\right) & \text{active}\end{cases}\tag{32}$$

and we can get

$$\frac{\partial V_{\lambda,\infty}(\omega)}{\partial\omega}=\begin{cases}1+\beta\frac{\partial V_{\lambda,\infty}(\tau^1(\omega))}{\partial\tau^1(\omega)}\frac{\partial\tau^1(\omega)}{\partial\omega} & \text{passive}\\ 1+\beta\left(V_{\lambda,\infty}(P^a_{1,1})-V_{\lambda,\infty}(P^a_{0,1})\right) & \text{active}\end{cases}\tag{33}$$

Similarly, for the finite residual time horizon T we have:

$$\frac{\partial V_{\lambda,T}(\omega)}{\partial\omega}=\begin{cases}1+\beta\frac{\partial V_{\lambda,T-1}(\tau^1(\omega))}{\partial\tau^1(\omega)}\frac{\partial\tau^1(\omega)}{\partial\omega} & \text{passive}\\ 1+\beta\left(V_{\lambda,T-1}(P^a_{1,1})-V_{\lambda,T-1}(P^a_{0,1})\right) & \text{active}\end{cases}\tag{34}$$

Note that for any belief state ω, $\tau^1(\omega)$ is the 1-step belief state update of ω when the passive arm is unobserved for another 1 consecutive slot. According to the Eq. 1, we have $\tau^1(\omega)=\omega P^p_{1,1}+(1-\omega)P^p_{0,1}$, thus

$$0<\frac{\partial\tau^1(\omega)}{\partial\omega}=(P^p_{1,1}-P^p_{0,1})<1\tag{35}$$

Lemma 4 *For the finite residual time horizon T, we have* $\frac{\partial V_{\lambda_T,T}(\omega)}{\partial\omega}\geq\min\{1+\beta(P^p_{1,1}-P^p_{0,1})\sum_{t=0}^{T-2}[\beta^t(P^a_{1,1}-P^a_{0,1})^t],1+\beta(P^a_{1,1}-P^a_{0,1})\sum_{t=0}^{T-2}[\beta^t(P^a_{1,1}-P^a_{0,1})^t]\}$

Proof. According to Eq. 34, for the passive action, we have:

$$\begin{aligned}&1+\beta\frac{\partial V_{\lambda,T-1}(\tau^1(\omega))}{\partial\tau^1(\omega)}\frac{\partial\tau^1(\omega)}{\partial\omega}\\ =&1+\beta\lim_{\delta\to0}\frac{V_{\lambda,T-1}(\tau^1(\omega)+\delta)-V_{\lambda,T-1}(\tau^1(\omega))}{\delta}(P^p_{1,1}-P^p_{0,1})\end{aligned}\tag{36}$$

According to Lemma 2, let $\omega_1=\tau^1(\omega)+\delta$ and $\omega_2=\tau^1(\omega)$, then we have $V_{\lambda,T-1}(\tau^1(\omega)+\delta)-V_{\lambda,T-1}(\tau^1(\omega))\geq(\omega_1-\omega_2)\sum_{t=0}^{T-1}\beta^t(P^a_{1,1}-P^a_{0,1})^t=\delta\sum_{t=0}^{T-1}\beta^t(P^a_{1,1}-P^a_{0,1})^t$. Thus Eq. 36 becomes:

$$\begin{aligned}&1+\beta\frac{\partial V_{\lambda,T-1}(\tau^1(\omega))}{\partial\tau^1(\omega)}\frac{\partial\tau^1(\omega)}{\partial\omega}\\ \geq&1+\beta(P^p_{1,1}-P^p_{0,1})\sum_{t=0}^{T-2}[\beta^t(P^a_{1,1}-P^a_{0,1})^t]\end{aligned}\tag{37}$$

Similarly, for the active action, according to lower bound in Lemma 2, we have:

$$\begin{aligned}&1+\beta\left(V_{\lambda,T-1}(P^a_{1,1})-V_{\lambda,T-1}(P^a_{0,1})\right)\\ \geq&1+\beta(P^a_{1,1}-P^a_{0,1})\sum_{t=0}^{T-2}[\beta^t(P^a_{1,1}-P^a_{0,1})^t]\end{aligned}\tag{38}$$

Therefore, we have $\frac{\partial V_{\lambda_T,T}(\omega)}{\partial\omega}\geq\min\{(P^p_{1,1}-P^p_{0,1}),(P^a_{1,1}-P^a_{0,1})\}\cdot\beta\cdot\sum_{t=0}^{T-2}[\beta^t(P^a_{1,1}-P^a_{0,1})^t]+1$

□

Lemma 5 *For the infinite residual time horizon $T\to\infty$, we have* $\frac{\partial V_{\lambda_T,T}(\omega)}{\partial\omega}\geq\min\{1+\frac{\beta(P^p_{1,1}-P^p_{0,1})}{1-(\beta(P^a_{1,1}-P_{0,1})},\frac{1}{1-\beta(P^a_{1,1}-P^a_{0,1})}\}$

Proof. The proof is similar to the proof for Lemma 4 of the finite setting. We can get the result with assuming $T\to\infty$. □

Lemma 6 *For the finite residual time horizon T, we have* $\frac{\partial V_{\lambda_T,T}(\omega)}{\partial\omega}\leq\min\{1+(P^p_{1,1}-P^p_{0,1})\sum_{t=1}^{T-1}\beta^t,1+(P^a_{1,1}-P^a_{0,1})\sum_{t=1}^{T-1}\beta^t\}$

Proof. The proof is similar to the proof of Lemma 4. According to Eq. 34, we have:

- passive actions:

$$\begin{aligned}&1+\beta\frac{\partial V_{\lambda,T-1}(\tau^1(\omega))}{\partial\tau^1(\omega)}\frac{\partial\tau^1(\omega)}{\partial\omega}\\ =&1+\beta\lim_{\delta\to0}\frac{V_{\lambda,T-1}(\tau^1(\omega)+\delta)-V_{\lambda,T-1}(\tau^1(\omega))}{\delta}(P^p_{1,1}-P^p_{0,1})\end{aligned}\tag{39}$$

According to Lemma 2, let $\omega_1=\tau^1(\omega)+\delta$ and $\omega_2=\tau^1(\omega)$, then we have $V_{\lambda,T-1}(\tau^1(\omega)+\delta)-V_{\lambda,T-1}(\tau^1(\omega))\leq(\omega_1-\omega_2)\sum_{t=0}^{T-1}\beta^t=\delta\sum_{t=0}^{T-1}\beta^t$. Thus Eq. 39 becomes:

$$\begin{aligned}&1+\beta\frac{\partial V_{\lambda,T-1}(\tau^1(\omega))}{\partial\tau^1(\omega)}\frac{\partial\tau^1(\omega)}{\partial\omega}\\ \leq&1+\beta(P^p_{1,1}-P^p_{0,1})\sum_{t=0}^{T-2}[\beta^t]\end{aligned}\tag{40}$$

- active actions, similarly, according to upper bound in

Lemma 2, we have:

$$\begin{aligned} &1+\beta\left(V_{\lambda,T-1}(P^a_{1,1})-V_{\lambda,T-1}(P^a_{0,1})\right) \\ \leq &1+\beta(P^a_{1,1}-P^a_{0,1})\sum_{t=0}^{T-2}[\beta^t] \end{aligned} \tag{41}$$

Therefore, we have $\frac{\partial V_{\lambda_T,T}(\omega)}{\partial\omega} \leq \min\{1+(P^p_{1,1}-P^p_{0,1})\sum_{t=1}^{T-1}\beta^t, 1+(P^a_{1,1}-P^a_{0,1})\sum_{t=1}^{T-1}\beta^t\}$

□

Lemma 7 *For the infinite residual time horizon $T\to\infty$, we have* $\frac{\partial V_{\lambda_T,T}(\omega)}{\partial\omega} \leq \min\{1+\frac{(P^p_{1,1}-P^p_{0,1})}{1-\beta}, 1+\frac{(P^a_{1,1}-P^a_{0,1})}{1-\beta}\}$

Proof. The proof is similar to the proof for Lemma 6 of the finite setting. We can get the result with assuming $T\to\infty$. □

C.2 CONDITION FOR THE OPTIMALITY OF ALGORITHM 1 UNDER FINITE/INFINITE HORIZON

We now give the proof for the Theorem 1 and Theorem 3.

According to the Eq. 1, we can compute the belief gap between ω and 1-time step belief update $\tau^1(\omega)$:

$$\Delta\omega = \tau^1(\omega)-\omega = (P^p_{1,1}-P^p_{0,1}-1)\omega+P^p_{0,1} \tag{42}$$

Remark that we could get a strict condition that depends only on arm A. This is because change from policy π^* to π will only lead to a decrease in the value function for other arms as the optimal actions determined by the Whittle index algorithm will be influenced, henceforth the value will be decreased. Consider the single-arm A, as we discussed earlier, the belief state update process is either monotonically increasing or monotonically decreasing.

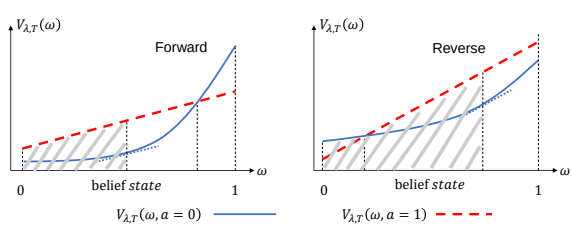

Figure 10: The forward and reverse policy

Case 1 : The belief state monotonically increases as the time passed (shown in left Fig. 3). Formally, this can be expressed as $\frac{\partial\omega_t}{\partial t}>0$, or $\Delta\omega>0$. We now derive the condition for the optimality of our algorithm for the case 1 under finite time horizon T. Consider first (any) period of length L, an arm i has not been activated for the past $(L-1)$ time slots. Thus it needs to be pulled at time step $t=L$ according to our algorithm. Assume that the residual time horizon is h at time step $t=L$, where we have $h+L=T$. We move the active action at time step $t=L$ to one slot earlier, then at time step $t=L-1$, the residual time horizon is $h+1$, and assume that belief state is ω at time step $t=L-1$. We here discuss the finite horizon case because the infinite horizon could be viewed as a special case of the finite horizon setting as $T\to\infty(h\to\infty)$.

Because the belief state will increase as the time passed, thus we define the value gap $\Delta V_{\lambda,h}(\omega) = V_{\lambda,h}(\omega, a=1)-V_{\lambda,h}(\omega, a=0)$ will move from left to the right as the residual time horizon decrease. For the single-arm process, if we can show that the gap difference $\Delta V_{\lambda,h}(\omega)$ increases from left to the right (i.e.,$\Delta V_{\lambda,h}(\omega)$ increases as belief state increases), then this implies that moving the active action that ensuring the fairness at time step $t=L$ to one step earlier (i.e., from right to left) will result in a smaller gap $\Delta V_{\lambda,h}(\omega)$. Thus it is optimal to keep the active action at the end of the period to ensure the fairness constraint. This requires that

$$\frac{\partial V_{\lambda,h}(\omega,a=0)}{\partial\omega} \leq \frac{\partial V_{\lambda,h}(\omega,a=1)}{\partial\omega} \tag{43}$$

According to Eq. 7, we have

$$\frac{\partial V_{\lambda,h}(\omega)}{\partial\omega} = \begin{cases} 1+\beta\frac{\partial V_{\lambda,h-1}(\tau^1(\omega))}{\partial\tau^1(\omega)}\frac{\partial\tau^1(\omega)}{\partial\omega} & \text{passive} \\ 1+\beta\left(V_{\lambda,h-1}(P^a_{1,1})-V_{\lambda,h-1}(P^a_{0,1})\right) & \text{active} \end{cases} \tag{44}$$

As shown in the gray area of the left Fig. 10, at time step $t=L-1$, we derive the technical condition for the optimality of our algorithm in the gray area under the infinite residual time horizon:

$$\begin{aligned} &(P^p_{1,1}-P^p_{0,1})\left(1+\frac{\Delta_3}{1-\beta}\right)\left(1-\beta(P^a_{1,1}-P^a_{0,1})\right) \leq (P^a_{1,1}-P^a_{0,1}) \\ \to&(P^p_{1,1}-P^p_{0,1})\left(1+\frac{\Delta_3}{1-\beta}\right) \leq \frac{P^a_{1,1}-P^a_{0,1}}{1-\beta(P^a_{1,1}-P^a_{0,1})} \text{ Line 1} \\ \to&(P^p_{1,1}-P^p_{0,1})\left(1+\frac{\Delta_3}{1-\beta}\right) \leq V_{\lambda,h-1}(P^a_{1,1})-V_{\lambda,h-1}(P^a_{0,1}) \text{ Line 2} \\ \to&(P^p_{1,1}-P^p_{0,1})\frac{\partial V_{\lambda,h-1}(\tau^1(\omega))}{\partial\tau^1(\omega)} \leq V_{\lambda,h-1}(P^a_{1,1})-V_{\lambda,h-1}(P^a_{0,1}) \text{ Line 3} \\ \to&\frac{\partial\tau^1(\omega)}{\partial\omega}\frac{\partial V_{\lambda,h-1}(\tau^1(\omega))}{\partial\tau^1(\omega)} \leq \left(V_{\lambda,h-1}(P^a_{1,1})-V_{\lambda,h-1}(P^a_{0,1})\right) \text{ Line 4} \\ \to&1+\beta\frac{\partial V_{\lambda,h-1}(\tau^1(\omega))}{\partial\tau^1(\omega)}\frac{\partial\tau^1(\omega)}{\partial\omega} \leq 1+\beta\left(V_{\lambda,h-1}(P^a_{1,1})-V_{\lambda,h-1}(P^a_{0,1})\right) \text{ Line 5} \\ \to&\frac{\partial V_{\lambda,h}(\omega,a=0)}{\partial\omega} \leq \frac{\partial V_{\lambda,h}(\omega,a=1)}{\partial\omega} \text{ Line 6} \end{aligned} \tag{45}$$

Line 1 is obtained via mathematical transformation. Line 2 is obtained from the lower bound in Lemma 2. Line 3 is obtained from the Lemma 7 when assuming $h\to\infty$. Line 4 is obtained from Eq. 35. Line 5 is obtained from the mathematical transformation. Line 6 is obtained from the Eq. 44. And $\Delta_3=\min\{(P^p_{1,1}-P^p_{0,1}),(P^a_{1,1}-P^a_{0,1})\}$

Similarly, we can derive the technical condition for the finite residual time horizon, which is

$$(P^p_{1,1} - P^p_{0,1})\left(\Delta_4 \beta \sum_{t=0}^{h-2}[\beta^t] + 1\right) \leq (P^a_{1,1} - P^a_{0,1}) \sum_{t=0}^{h-2}[\beta^t (P^a_{1,1} - P^a_{0,1})^t] \tag{46}$$

where $\Delta_4 = \min\{(P^p_{1,1} - P^p_{0,1}), (P^a_{1,1} - P^a_{0,1})\}$.

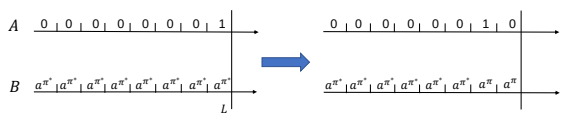

Figure 11: Proof of Theorem 1 and Theorem 3

Case 2 : The belief state monotonically decreases as the time passed (shown in right Fig. 3). Formally, this can be expressed as $\frac{\partial \omega_t}{\partial t} < 0$, or $\Delta\omega < 0$. Similarly, we can derive the condition for the optimality of our algorithm for the case 2 under finite time horizon T. Because the belief state will decrease as the time passed, thus we define the value gap $\Delta V_{\lambda,h}(\omega) = V_{\lambda,h}(\omega, a = 1) - V_{\lambda,h}(\omega, a = 0)$ will move from right to the left as the residual time horizon decrease. For the single-arm process, if we can show that the gap difference $\Delta V_{\lambda,h}(\omega)$ decreases from right to the left (i.e., as belief state decrease, $\Delta V_{\lambda,h}(\omega)$ decreases), then this implies that moving the active action that ensuring the fairness at time step $t = L$ to one step earlier will result in a larger gap $\Delta V_{\lambda,h}(\omega)$. Thus it is optimal to keep the active action at the end of the period, i.e., at time step $t = L$. this requires that

$$\frac{\partial V_{\lambda,h}(\omega, a = 0)}{\partial \omega} \geq \frac{\partial V_{\lambda,h}(\omega, a = 1)}{\partial \omega} \tag{47}$$

As shown in the gray area of the left Fig. 10, at time step $t = L - 1$, we derive the technical condition for the optimality of our algorithm in the gray area under the infinite residual time horizon:

$$\begin{aligned}
&(P^p_{1,1} - P^p_{0,1})(1-\beta)\Delta_1 \geq (P^a_{1,1} - P^a_{0,1})\left(1 - \beta(P^a_{1,1} - P^a_{0,1})\right) \\
\rightarrow &\frac{(P^p_{1,1} - P^p_{0,1})\Delta_1}{1 - \beta(P^a_{1,1} - P^a_{0,1})} \geq \frac{P^a_{1,1} - P^a_{0,1}}{1-\beta} \text{ Line 1} \\
\rightarrow &(P^p_{1,1} - P^p_{0,1})\Delta_1 \lim_{h\to\infty} \sum_{t=0}^{h-2} \beta^t (P^a_{1,1} - P^a_{0,1})^t \geq \frac{P^a_{1,1} - P^a_{0,1}}{1-\beta} \text{ Line 2} \\
\rightarrow &(P^p_{1,1} - P^p_{0,1})\Delta_1 \lim_{h\to\infty} \sum_{t=0}^{h-2} \beta^t (P^a_{1,1} - P^a_{0,1})^t \geq V_{\lambda,h-1}(P^a_{1,1}) - V_{\lambda,h-1}(P^a_{0,1}) \text{ Line 3} \\
\rightarrow &(P^p_{1,1} - P^p_{0,1})\frac{\partial V_{\lambda,h-1}(\tau^1(\omega))}{\partial \tau^1(\omega)} \geq V_{\lambda,h-1}(P^a_{1,1}) - V_{\lambda,h-1}(P^a_{0,1}) \text{ Line 4} \\
\rightarrow &\frac{\partial \tau^1(\omega)}{\partial \omega}\frac{\partial V_{\lambda,h-1}(\tau^1(\omega))}{\partial \tau^1(\omega)} \geq \left(V_{\lambda,h-1}(P^a_{1,1}) - V_{\lambda,h-1}(P^a_{0,1})\right) \text{ Line 5} \\
\rightarrow &1 + \beta\frac{\partial V_{\lambda,h-1}(\tau^1(\omega))}{\partial \tau^1(\omega)}\frac{\partial \tau^1(\omega)}{\partial \omega} \geq 1 + \beta\left(V_{\lambda,h-1}(P^a_{1,1}) - V_{\lambda,h-1}(P^a_{0,1})\right) \text{ Line 6} \\
\rightarrow &\frac{\partial V_{\lambda,h}(\omega, a=0)}{\partial \omega} \geq \frac{\partial V_{\lambda,h}(\omega, a=1)}{\partial \omega} \text{ Line 7}
\end{aligned} \tag{48}$$

Line 1 is obtained via mathematical transformation. Line 2 is obtained from the formula for the geometric series as $\beta(P^a_{1,1} - P^a_{0,1}) < 1$. Line 3 is obtained from the upper bound in Lemma 2. Line 4 is obtained from the Lemma 4 when assuming $h \to \infty$. Line 5 is obtained from Eq. 35. Line 6 is obtained from the mathematical transformation. Line 7 is obtained from the Eq. 44.

Similarly, we can derive the technical condition for the finite residual time horizon, which is

$$(P^p_{1,1} - P^p_{0,1})\left(\Delta_2 \beta \sum_{t=0}^{h-2}[\beta^t (P^a_{1,1} - P^a_{0,1})^t] + 1\right) \geq (P^a_{1,1} - P^a_{0,1}) \sum_{t=0}^{h-2} \beta^t \tag{49}$$

where $\Delta_2 = \min\{(P^p_{1,1} - P^p_{0,1}), (P^a_{1,1} - P^a_{0,1})\}$.

When the belief state is in the white area of the passive set. Then we need to consider arm A and other arms in the active set. We give detailed discussion in Appendix E.2.

D OTHER PROOFS

D.1 DERIVATION OF EQUATION

The belief state can be calculated in closed form with the given transition probabilities. Let $\tau^u_i(\omega_{i,s}(t)) = \omega_{i,s}(t+u)$ denote the u-step belief state update of $\omega_{i,s}(t)$ when the unobserved arm i is updated for u consecutive slots without being selected. Formally,

$$\omega_{i,s}(u) = \tau^{u-1}(P^{a,i}_{s,1}) \text{ where}$$
$$\tau^u_i(\omega) = \frac{P^{p,i}_{0,1} - (P^{p,i}_{1,1} - P^{p,i}_{0,1})^u (P^{p,i}_{0,1} - \omega(1 + P^{p,i}_{0,1} - P^{p,i}_{1,1}))}{(1 + P^{p,i}_{0,1} - P^{p,i}_{1,1})} \tag{50}$$

This is because $\tau^u(\omega) = \omega\tau^u(1) + (1-\omega)\tau^u(0)$, where $\tau^u(1)$ is the u-step transition probability from 1 to 1 when arm is unobserved, and $\tau^u(0)$ is the u-step transition probability from 0 to 1 if the transition matrix $P^a_{s,s'} = P^p_{s,s'} = P_{s,s'}$. From the eigen-decomposition of the transition matrix $\mathcal{P}$, we can have

$$\tau^u(1) = \frac{P_{0,1} + (1 - P_{1,1})(P_{1,1} - P_{0,1})^u}{1 + P_{0,1} - P_{1,1}}$$

and

$$\tau^u(0) = \frac{P_{0,1}(1 - (P_{1,1} - P_{0,1})^u)}{1 + P_{0,1} - P_{1,1}}$$

and solve it to get

$$\tau^u(\omega) = \frac{P_{0,1} - (P_{1,1} - P_{0,1})^u (P_{0,1} - \omega(1 + P_{0,1} - P_{1,1}))}{(1 + P_{0,1} - P_{1,1})}$$

However, as the $P^a_{s,s'} \neq P^p_{s,s'}$, We have $\omega^i_s(1) = P^{a,i}_{o_i,1}$, and $\tau^u(\omega)$ as shown in Eq. 50, which leads to $\omega^i_s(u)$.

D.2 PROOF OF WHITTLE INDEX DECAY UNDER PARTIALLY OBSERVABLE SETTING (THEOREM 2)

Proof. We can use the induction to prove the index decay. Again, λ_0 satisfies: $V_{\lambda_0,0}(\omega, a=0) = V_{\lambda_0,0}(\omega, a=1)$ for any belief state $\omega \in [0,1]$. We can have $\lambda_0 + \omega = \omega$, thus $\lambda_0 = 0$. Similarly, λ_1 can be solved by assuming equation $V_{\lambda_1,1}(\omega, a=0) = V_{\lambda_1,1}(\omega, a=1)$:

$$\begin{aligned}&\omega + \lambda + \beta(\omega P^p_{1,1} + (1-\omega)P^p_{0,1}) = \omega + \beta(\omega P^a_{1,1} + (1-\omega)P^a_{0,1})\\ \rightarrow &\lambda = \beta(\omega(P^a_{1,1} - P^p_{1,1}) + (1-\omega)(P^a_{0,1} - P^p_{0,1}))\end{aligned} \tag{51}$$

As it is true in the real-world that $P^a_{s,1} > P^p_{s,1}$, thus we have $\lambda_1 > 0 = \lambda_0$. Now we assume the hypothesis that $\lambda_u > \lambda_{u-1}$ for $u \in \{1, \ldots, T\}$ holds, we must show: $\lambda_{T+1} > \lambda_T$. There is a similar work done by Mate et al. [2020], however, they only show that $\lambda_T > \lambda_1$ for $\forall T > 1$. Their conclusion is built on the fact that for two non-decreasing, linear functions $f_1(\lambda)$ and $f_2(\lambda)$ of $lambda$ and two points x_1, x_2. Whenever $f_1(x_1) \leq f_2(x_1)$ and $f_1(x_2) \geq f_2(x_2)$, and if $\frac{\partial f_1}{\partial x} \geq \frac{\partial f_2}{\partial x}$, then $x_1 \leq x_2$ is true. We here take advantage of this fact and show a different way to demonstrate the index decay in the partially observable setting such that $\lambda_{T+1} > \lambda_T$.

We first use $V_{\lambda,T+1}(\omega, a=0)$ to denote function f^p_1, and set $f^a_2 = V_{\lambda,T+1}(\omega, a=1)$. Then it is obvious that the value of $f^p_1(\omega, \lambda, t)$ will increase as the exogenous reward λ increase, whereas the value of $f^a_2(\omega, \lambda, t)$ will decrease. Thus we have the following:

$$\frac{\partial f^p_1(\omega, \lambda, t)}{\partial \lambda} > \frac{\partial f^a_2(\omega, \lambda, t)}{\partial \lambda} \tag{52}$$

We can prove $\lambda_{T+1} > \lambda_T$ through contradiction. We first assume that $\lambda_T \geq \lambda_{T-1}$ holds. Then we could have :

$$f^p_1(\omega, \lambda_{T-1}, T-1) = f^a_2(\omega, \lambda_{T-1}, T-1); \tag{53}$$

and

$$f^p_1(\omega, \lambda_T, T-1) > f^p_1(\omega, \lambda_{T-1}, T-1) = f^a_2(\omega, \lambda_{T-1}, T-1) > f^a_2(\omega, \lambda_T, T-1) \tag{54}$$

Similarly, we can also have

$$f^p_1(\omega, \lambda_T, T) = f^a_2(\omega, \lambda_T, T). \tag{55}$$

. We can get

$$\frac{\partial f^p_1(\omega, \lambda_T, t)}{\partial t} < \frac{\partial f^a_2(\omega, \lambda_T, t)}{\partial t} \tag{56}$$

However, we assume that $\lambda_{T+1} \leq \lambda_T$. According to the definition of λ_{T+1}, we have

$$f^p_1(\omega, \lambda_{T+1}, T+1) = f^a_2(\omega, \lambda_{T+1}, T+1); \tag{57}$$

It is obvious that $\frac{\partial f^p_1(\omega,\lambda,t)}{\partial t} > 0$, and $\frac{\partial f^a_2(\omega,\lambda,t)}{\partial t} > 0$ (from Lemma 1). According to Eq. 56 and Eq. 55, we can have

$$f^p_1(\omega, \lambda_T, T+1) < f^a_2(\omega, \lambda_T, T+1) \tag{58}$$

As $\lambda_{T+1} \leq \lambda_T$, according to Eq. 57, we can have

$$f^p_1(\omega, \lambda_T, T+1) \geq f^a_2(\omega, \lambda_T, T+1) \tag{59}$$

We have Eq. 59 and Eq. 58 conflict with each other. Henceforth $\lambda_{T+1} > \lambda_T$. □

D.3 PROOF OF THEOREM 4

Proof. We provide our proof which is based on the work by Biswas et al. [2021]. Let set ϕ^* to be the set of actions containing the k arms with the highest-ranking values of $Q_i(s, a=1, l) - Q_i(s, a=0, l)$, and any k arms that aren't among the top k are included in the set ϕ'. Let $\phi^{-,*}$ and $\phi^{-,\prime}$ denote the set that includes all of the arms except those in set ϕ^* and ϕ', respectively. We add the subscript i here in order to avoid ambiguity in the Q-values of distinct arms i at a given state. We could have:

$$\begin{aligned}&\sum_{i^* \in \phi^*} [Q^*_{i^*}(s_{i^*}, a_{i^*}=1, l_{i^*}) - Q^*_{i^*}(s_{i^*}, a_{i^*}=0, l_{i^*})] \geq\\ &\sum_{j \in \phi'} \left[Q^*_j(s_j, a_j=1, l_j) - Q^*_j(s_j, a_j=0, l_j)\right]\end{aligned} \tag{60}$$

$$\begin{aligned}&\sum_{i^* \in \phi^*} Q^*_{i^*}(s_{i^*}, a_{i^*}=1, l_{i^*}) + \sum_{j \in \phi'} Q^*_j(s_j, a_j=0, l_j) \geq\\ &\sum_{j \in \phi'} Q^*_j(s_j, a_j=1, l_j) + \sum_{i^* \in \phi^*} Q^*_{i^*}(s_{i^*}, a_{i^*}=0, l_{i^*})\end{aligned} \tag{61}$$

Adding $\sum_{i \notin \phi^* \& i \notin \phi'} Q^*_i(s_i, a_i=0, l_i)$ on both sides,

$$\begin{aligned}&\sum_{i^* \in \phi^*} Q^*_{i^*}(s_{i^*}, a_{i^*}=1, l_{i^*}) + \sum_{j \in \phi^{-,*}} Q^*_j(s_j, a_j=0, l_j) \geq\\ &\sum_{i \in \phi'} Q^*_i(s_i, a_i=1, l_i) + \sum_{j \in \phi^{-,\prime}} Q^*_j(s_j, a_j=0, l_j)\end{aligned} \tag{62}$$

Thus from Equation 62, we can see that taking intervention action in the action set As can be seen from Equation 62, adopting intervention action for the arms in the set ϕ^* would maximizes $\left\{\sum_{i=1}^N Q^*_i(s, a, l)\right\}$. □

D.4 PROOF OF THEOREM 5

Proof. The key to the convergence is contingent on a particular sequence of episodes observed in the real process Watkins and Dayan [1992]. The first condition is easy

to be satisfied as to the presence of the fairness constraint. It is a reasonable assumption under the ϵ-greedy action selection mechanism, that any state-action pair can be visited an unlimited number of times as $T \to \infty$. The second condition has been well-studied in Hirsch [1989], Watkins and Dayan [1992], Jaakkola et al. [1994], and it guarantees that when the condition is met, the Q-value converges to the optimal $Q^*(s, a, l)$. As a result, $Q_i(s, a = 1, l) - Q_i(s, a = 0, l)$ converges to $Q_i^*(s, a = 1, l) - Q_i^*(s, a = 0, l)$. Also, $Q_i^*(s, a = 1, l) - Q_i^*(s, a = 0, l)$ is the calculated Q-Learning based Whittle index, and choosing top-ranked arms based on these values would lead to an optimal solution.

□

E GENERAL CONDITION

E.1 A GENERAL CONDITION FOR THEOREM 6

During a time interval, we consider the two arms A and B as shown in Figure 11, B is the k-th ranked arm as shown in algorithm 1. We assume that the optimal policy is π^*, which only considers the fairness constraint when it is violated at the end of interval, implying that arm A has not been activated in the past $(L-1)$ slots as shown in left. We move the action $a_L = 1$ to one slot earlier, where action a_L is to ensure the fairness constraint. We denote this policy as π shown in the right. We first consider the fully observable setting here. Assume that the state of arm A in time step $L-1$ is s_A, and the state of arm B is s_B, and $s_A, s_B \in \{0, 1\}$. We let the corresponding Whittle index for arm A is λ_A, and λ_B for arm B. We can calculate the value function $V_{\lambda_i,\infty}(\omega_i)$ for arm A from the time slot $L-1$ under the policy π^*, we have

$$V_{\lambda_A,\infty}(s_A) = s_A + \lambda_A + \beta\{P_{1,1}^a s_A + P_{0,1}^p(1-s_A) + \beta[(P_{1,1}^a s_A + P_{0,1}^p(1-s_A))V_{\lambda_A,\infty}(P_{1,1}^a) + (1 - P_{1,1}^a s_A - P_{0,1}^p(1-s_A))V_{\lambda_A,\infty}(P_{0,1}^a)]\} \tag{63}$$

Note that $V_{\lambda,\infty}(\omega) = \omega V_{\lambda,\infty}(1) + (1-\omega)V_{\lambda,\infty}(0)$. Similarly, we can get the value function for the arm A under the policy π:

$$V_{\lambda_A,\infty}(s_A) = s_A + \beta(s_A V_{\lambda_A,\infty}(P_{1,1}^a) + (1-s_A)V_{\lambda_A,\infty}(P_{0,1}^a)) \tag{64}$$

Similarly, the value function for the arm B under policy π^*:

$$V_{\lambda_B,\infty}(s_B) = s_B + \beta(s_B V_{\lambda_B,\infty}(P_{1,1}^a) + (1-s_B)V_{\lambda_B,\infty}(P_{0,1}^p)), \tag{65}$$

and the value function for the arm B under policy π:

$$V_{\lambda_B,\infty}(s_B) = \lambda_B + s_B + \beta(V_{\lambda_B,\infty}(s_B P_{1,1}^p + (1-s_B)P_{0,1}^p)), \tag{66}$$

Let the Eq. 63 minus Eq. 64 and Eq. 65 minus Eq. 66, and then sum these two values to let it greater than 0, we then get the general condition for theorem 6. The explicit calculation for the value function can be found in [Liu and Zhao, 2010].

E.2 GENERAL CONDITION FOR THEOREM 1 AND THEOREM 3

This is similar to the fully observable case E.1. Assume the optimal action set that calculated by Whittle index policy is A, and the last several action set that is used to satisfy the fairness constraints is B, where we have action $b \in B$ for arm i. As shown in Figure 12. If we shift the action $b \in B$

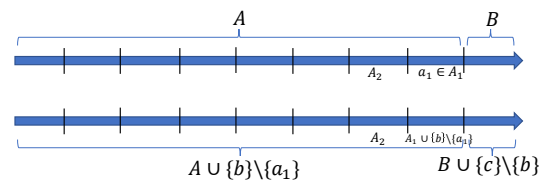

Figure 12: Proof of Theorem 1

at the end of time slot to one slot earlier, and add another action c for arm j into the action set B, then according to the belief state update process, the other following action set except for arm i and j will not be influenced. As a result, we only need to consider arm i and arm j. If we showed that the modified policy will not have greater value function. again, then we can repeatedly move the action b to one slot earlier and show that the value function will never have a higher value than the original policy. Hence, we can conclude that the modified policy is smaller than the optimal policy. The Whittle index approach is still applicable under the fairness constraint unless the last few slots are used to ensure that the fairness criteria are met. The numerical condition is as follows,

We consider the partially observable setting. Assume that the probability that the arm A is in good state at time slot $L-1$ is ω_i, and arm B is ω_j, and we assume the corresponding Whittle index for arm A is λ_i, and λ_j for arm B. We can calculate the value function $V_{\lambda_i,\infty}(\omega_i)$ from the time slot $L-1$ for the policy π^*, we have

$$\begin{aligned} V_{\lambda_i,\infty}(\omega_i) &= \lambda_i + \omega_i + \beta(\omega_i P_{1,1}^p + (1-\omega_i)P_{0,1}^p) \\ &+\beta^2\{(A+B)V_{\lambda_i,\infty}(1) + (C+D)V_{\lambda_i,\infty}(0)\} \end{aligned} \tag{67}$$

where $A = P_{1,1}^p(\omega_i P_{,1}^p + (1-\omega_i)P_{0,1}^p)$, $B = P_{0,1}^a(1 - \omega_i P_{1,1}^p - (1-\omega_i)P_{0,1}^p)$, $C = (1 - P_{1,1}^a)(\omega_i P_{1,1}^p + (1-\omega_i)P_{0,1}^p)$ and $D = (1 - P_{0,1}^a)(1 - \omega_i P_{1,1}^p - (1-\omega_i)P_{0,1}^p)$.

Similarly, we can get the value function for the arm A under the policy π:

$$\begin{aligned} V_{\lambda_i,\infty}(\omega_i) &= \omega_i + \beta(\omega_i P_{1,1}^a + (1-\omega_i)P_{0,1}^a + \lambda_i) \\ &+\beta^2\{(A+B')V_{\lambda_i,\infty}(1) + (C'+D')V_{\infty}(\lambda_i, 0)\}, \end{aligned} \tag{68}$$

where $B' = B - P_{0,1}^a + P_{0,1}^p$, $C' = C + \omega_i(P_{1,1}^a - P_{1,1}^p) + (1-\omega_i)(P_{0,1}^a - P_{0,1}^p)$ and $D' = D + \omega_i(P_{1,1}^p - P_{1,1}^a) + (1-\omega_i)(P_{0,1}^p - P_{0,1}^a) + (P_{0,1}^a - P_{0,1}^p)$. Let Eq. 67 minus Eq. 68,

we can have

$$\lambda_i(1-\beta)+\beta(\omega_i(P^p_{1,1}-P^a_{1,1})+(1-\omega_i)(P^p_{0,1}-P^a_{0,1}))+ \beta^2((P^a_{0,1}-P^p_{0,1})V_{\lambda_i,\infty}(1)+(P^p_{0,1}-P^a_{0,1})V_{\lambda_i,\infty}(0)) \tag{69}$$

Similarly, for the arm B, we can get the change in the value function from the policy π^* to π:

$$\lambda_j(\beta-1)-\beta(\omega_j(P^p_{1,1}-P^a_{1,1})-(1-\omega_j)(P^p_{0,1}-P^a_{0,1}))+ \beta^2((P^a_{0,1}-P^p_{0,1})V_{\lambda_j,\infty}(1)+(P^p_{0,1}-P^a_{0,1})V_{\lambda_j,\infty}(0)) \tag{70}$$

We sum Eq. 69 and Eq. 70 and let it greater than 0, then this is the condition that the optimal policy of theorem 1 for RMAB with the fairness constraint is to select the arm to play when the fairness constraint is violated at the time interval under the partially observable setting. Explicit computation of the value function can be found in [Liu and Zhao, 2010].

F VISUALIZATION OF ALGORITHM

We give a visualization of our proposed Whittle index based approach to solve the fairness constraint in Figure 15.

The belief state MDP works as follows: initially, after an action, the state $s \in \{0,1\}$ of the selected arm is observed. Then the belief state then changes to $P^a_{s,1}$ one slot later, which is represented as the blue node at the head of the chain. Subsequent passive actions cause the belief state evolves according to the initial observation in the same chain. Then if the arm is activated again under the proposed algorithm, it will transit to the head of one of the chains with the probability according to its belief state as shown in the black arrow. If the arm's fairness constraint is not met, i.e., it has not been chosen in the last $L-1$ time slots, it will be activated at the time slot L, and go to the head of one of the chains (as shown by the red dashed arrow).

G ADDITIONAL RESULTS

G.1 FAIRNESS CONSTRAINTS STRENGTH

In this section, we provide the average reward results for different fairness constraints, and see how they influence the overall performance. The strength of fairness restrictions is represented by the combination of L and η. For instance, η is a parameter to determine the lowest bound of the number of times an arm should be activated in a decision period of length L. Smaller L, on the other hand, indicates that a strict fairness constraint should be addressed in a shorter time length. For ease of explanation, we fix the value of $\eta = 2$, which means that an arm will be activated twice in any given time steps of length L. We can change the value of L to measure the fairness constraint level. We investigate three different categories of fairness constraint strength as follows,

- **Strong level**: The strong fairness constraints impose a strict restrictions on the action. Here we assume that the strong fairness constraints L satisfy $\frac{k\times L}{N} = 1.3$, this can translate to at most 30% arms can be engaged twice when before all arms have been pulled previously.
- **Medium level**: We define the medium fairness constraints by solving: $\frac{k\times L}{N} = 2$.
- **Low level**: The low strength of fairness constraints can be interpreted as a low fairness restriction on the distribution of the resources, i.e., we have $\frac{k\times L}{N} = 3$, which means all arms will receive the health intervention before each arm has been activated three times on average.

We provide the average reward results in Figure 13. Again, the left graph shows the performance of Whittle index approach with fairness constraint when the transition model is known, the middle graph presents the result of the Thompson sampling-based approach for Whittle index calculation, the the right graph shows the result for the Q-Learning based Whittle index approach. Our proposed approach can handle fairness constraints at different strength level without sacrificing significantly on the solution quality.

G.2 INTERVENTION LEVEL

In this part, we present the performance results for various resource levels where the fairness constraint L is fixed, and we ensure that $k \times L < N$. Here, we let $L = 30$, and $N = 100$, and we're looking at the performance of the intervention ratio where $\frac{k}{N} = 5\%, 10\%, 20\%, 30\%$ respectively. We can see that our proposed approach to solve the fairness constraint can consistently outperform the Random and Myopic baselines regardless of the intervention strength while does not have significant differences when compared to the optimal value without taking fairness constraints into account.

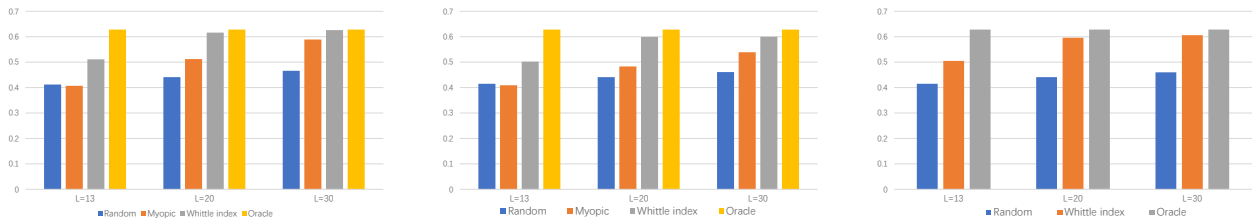

Figure 13: The average reward of each arm over the time length $T = 1000$ with small penalty for the violation of the fairness constraint.

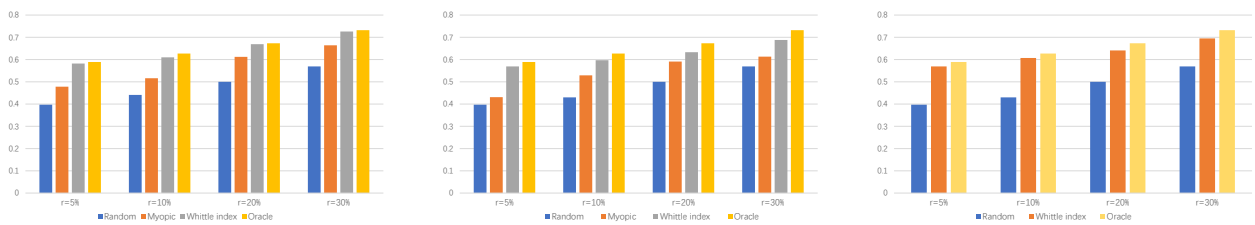

Figure 14: The average reward of each arm over the time length $T = 1000$ with small penalty for the violation of the fairness constraint.

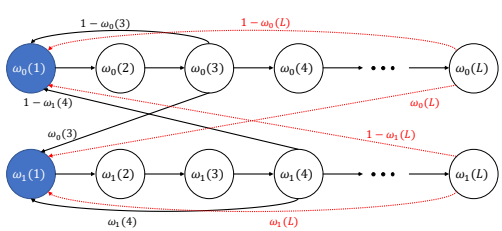

Figure 15: Visualization of Whittle index approach with fairness constraints.