Planning to Fairly Allocate: Probabilistic Fairness in the Restless Bandit Setting

Christine Herlihy*
Aviva Prins*
cherlihy@cs.umd.edu
aviva@cs.umd.edu
University of Maryland
College Park, Maryland, USA

Aravind Srinivasan
asriniv1@umd.edu
University of Maryland
College Park, Maryland, USA

John P. Dickerson
johnd@umd.edu
University of Maryland
College Park, Maryland, USA

ABSTRACT

Restless and collapsing bandits are often used to model budget-constrained resource allocation in settings where arms have action-dependent transition probabilities, such as the allocation of health interventions among patients. However, SOTA Whittle-index-based approaches to this planning problem either do not consider fairness among arms, or incentivize fairness without guaranteeing it. We thus introduce ProbFair, a probabilistically fair policy that maximizes total expected reward and satisfies the budget constraint while ensuring a strictly positive lower bound on the probability of being pulled at each timestep. We evaluate our algorithm on a real-world application, where interventions support continuous positive airway pressure (CPAP) therapy adherence among patients, as well as on a broader class of synthetic transition matrices. We find that ProbFair preserves utility while providing fairness guarantees.

CCS CONCEPTS

• **Computing methodologies → Partially-observable Markov decision processes**; *Multi-agent planning*; • **Applied computing** → Health care information systems.

KEYWORDS

probabilistic fairness; resource allocation; restless multi-armed bandits; collapsing multi-armed bandits; Whittle index; intervention planning; sequential decision making; POMDP; healthcare

ACM Reference Format:
Christine Herlihy, Aviva Prins, Aravind Srinivasan, and John P. Dickerson. 2023. Planning to Fairly Allocate: Probabilistic Fairness in the Restless Bandit Setting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3580305.3599467

1 INTRODUCTION

Restless multi-armed bandits (RMABs) are used to model budget-constrained resource allocation tasks in which a decision-maker must select a subset of arms (e.g., projects, patients, assets) to receive a beneficial intervention at each timestep, while the state of each arm evolves over time in an action-dependent, Markovian fashion. Such problems are common in healthcare, where clinicians may be tasked with monitoring large, distributed patient populations and determining which individuals to expend scarce resources on so as to maximize total welfare. RMABs have been proposed to determine which inmates should be prioritized to receive hepatitis C treatment in U.S. prisons [2], and which tuberculosis patients should receive medication adherence support in India [20].

Current state-of-the-art approaches to solving RMABs rely on the indexing work introduced by Whittle [36]. While the Whittle index solves an otherwise PSPACE-complete problem in an asymptotically optimal fashion by decoupling arms [35], it fails to provide any guarantees about how pulls will be distributed *among arms*.

Though the intervention is canonically assumed to be beneficial for *every* arm, the marginal benefit (i.e., relative increase in the probability of a favorable state transition) varies in accordance with each arm's underlying state transition function. Consequently, Whittle index-based maximization of total expected reward *without regard for distributive fairness* empirically allocates all available interventions to a small subset of arms, ignoring the rest [25].

There are many application domains where a bimodal distributive outcome may be perceived as unfair or undesirable by beneficiaries and decision-makers, thus motivating efforts to incentivize or guarantee distributive fairness. In the aforementioned healthcare examples, resource constraints and variation in transition dynamics interact. A practical consequence is that a majority of patients will *never* receive the beneficial intervention(s) in question. This, in turn, means that their clinical outcomes will be strictly worse in expectation than they would be under a policy that guaranteed a non-zero probability of receiving the intervention at each timestep.

To improve distributive fairness, we explore whether it is possible to modify the Whittle index to guarantee each arm at least one pull per user-defined time interval, but find this to be intractable. We then introduce ProbFair, a state-agnostic policy that maps each arm to a fairness-constraint satisfying, stationary probability distribution over actions that takes the arm's transition matrix into account. At each timestep, we then use a dependent rounding algorithm [33] to sample from this probabilistic policy to produce a budget-constraint satisfying discrete action vector.

We evaluate ProbFair on a randomly generated dataset and a realistic dataset derived from obstructive sleep apnea patients tasked with nightly self-administration of continuous positive airway pressure (CPAP) therapy [13, 14].

*Denotes equal contribution to this research.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599467

Our core contributions include:

(i) A novel approach that is both efficiently computable and reward maximizing, subject to the guaranteed satisfaction of budget *and* probabilistic fairness constraints.
(ii) Empirical results demonstrating that ProbFair is competitive vis-à-vis other fairness-inducing policies, and stable over a range of cohort composition scenarios.

2 RESTLESS MULTI-ARMED BANDIT MODEL

Here, we give an overview of the restless multi-armed bandit (RMAB) framework, along with our proposed extension, which takes the form of a fairness-motivated constraint. A restless multi-armed bandit consists of $N \in \mathbb{N}$ independent arms, each of which evolves over a finite time horizon $T \in \mathbb{N}$, according to an associated Markov Decision Process (MDP). Each arm's MDP is characterized by a 4-tuple $(\mathcal{S}, \mathcal{A}, P, r)$ where $\mathcal{S}$ represents the state space, $\mathcal{A}$ represents the action space, P represents an $|\mathcal{S}|\times|\mathcal{A}|\times|\mathcal{S}|$ transition matrix, and $r : \mathcal{S} \longrightarrow \mathbb{R}$ represents a local reward function that maps states to real-valued rewards. Appendix A summarizes notation; note that $[N]$ denotes the set $\{1, 2, \ldots, N\}$.

States, actions, and observability: We specifically consider a discrete two-state system $\mathcal{S} := \{0, 1\}$ where 1 (0) represents being in the "good" ("bad") state, and a set of two possible actions $\mathcal{A} := \{0, 1\}$ where 1 represents the decision to select ("pull") arm $i \in [N]$ at time $t \in [T]$, and 0 represents the choice to be passive (not pull). In the general RMAB setting, each arm's state s_t^i is observable. We consider the partially-observable extension introduced by Mate et al. [20], where arms' states are only observable when they are pulled. Otherwise, an arms' state is replaced with the probabilistic *belief* $b_t^i \in [0, 1]$ that it is in state 1. Such partial observability captures uncertainty regarding patient status and treatment efficacy associated with outpatient or remotely-administered interventions.

Transition matrices: Each arm i is characterized by a set of transition matrices P, where $P^a_{s,s'}$ represents the probability of transitioning from state s to state s' when action a is taken. We assume P to be (a) static and (b) known by the agent at planning time. Assumptions (a) and (b) are likely to be violated in practice; however, they provide a useful modeling foundation, and can be modified to incorporate additional uncertainty, such as the requirement that transition matrices must be learned [12]. Clinical researchers often use longitudinal data to construct risk-adjusted transition matrices that encode cohort-specific transition probabilities. These can guide patient-level decision-making [34].

Consistent with previous literature, we assume strictly positive transition matrix entries, and impose four *structural constraints*: (a) $P^0_{0,1} < P^0_{1,1}$; (b) $P^1_{0,1} < P^1_{1,1}$; (c) $P^0_{0,1} < P^1_{0,1}$; (d) $P^0_{1,1} < P^1_{1,1}$ [20]. These constraints are application-motivated, and imply that arms are more likely to remain in a "good" state than change from a bad state to a good one, and that a pull is helpful when received. In the absence of such constraints, the effect of the intervention may be superfluous or harmful, rather than desirable.

Objective and constraints: In the canonical RMAB setting, the agent's goal is to find a policy π^* that maximizes total expected reward $\arg\max_\pi \mathbb{E}_\pi[R(r(s))]$ while satisfying a *budget constraint*, $k \ll N \in \mathbb{N}$, which allows the agent to select at most k arms at each timestep. We consider a cumulative reward function, $R(\cdot) := \sum_{i\in[N]} \sum_{t\in[T]} \beta^{t-1} r(s_t^i)$, for some discount rate $\beta \in [0, 1]$, and non-decreasing $r(s)$.

We extend this model by introducing a Boolean-valued, distributive fairness-motivated constraint, which may take one of two general forms:

(1) *Time-indexed*: A function $g\left(\cup_{t\in[T]}\{\vec{a}_t\}\right)$ which is satisfied if each arm is pulled at least once within each user-defined time interval $\nu \leq T$ (e.g., at least once every seven days), or a minimum fraction $\psi \in (0, 1)$ of times over the entire time horizon [17].
(2) *Probabilistic*: A function $g'(\vec{p}^i|\vec{a}_t \sim \vec{p}^i\ \forall t)$ which operates on the stationary probability vector $\vec{p}^i$, from which discrete actions are drawn, by requiring the probability that each arm receives a pull at any given t to fall within an interval $[\ell, u]$ where $0 < \ell \leq \frac{k}{N} \leq u \leq 1$.

3 CONTEXT, MOTIVATION & RELATED WORK

In this section, we motivate our ultimate focus on probabilistic fairness by revisiting the distribution of pulls under Whittle-index based policies. We begin by providing background information on the Whittle index, and then proceed to ask: (1) Which arms are ignored, and why does it matter? (2) Is it possible to modify the Whittle index so as to provide a *time-indexed fairness guarantee* for each arm? In response to the latter, we demonstrate that time-indexed fairness guarantees necessitate the coupling of arms, which undermines the indexability of the problem. We then identify prior work at the intersection of algorithmic fairness, constrained resource allocation, and multi-armed bandits, and identify desiderata that characterize our own approach.

3.1 Background: Whittle Index-based Policies

Pre-computing the optimal policy for a given set of restless or collapsing arms is PSPACE-hard in the general case [24]. However, as established by Whittle [36] and formalized by Weber and Weiss [35], if the set of arms associated with a problem are *indexable*, we can decouple the arms and efficiently solve the problem using an asymptotically-optimal heuristic index policy.

Mechanics: At each timestep $t \in [T]$, the value of a pull, in terms of both immediate and expected discounted future reward, is computed for each decoupled arm, $i \in [N]$. This value-computation step relies on the notion of a subsidy, m, which can be thought of as the opportunity cost of passivity. Formally, the Whittle index is the subsidy required to make the agent indifferent between *pulling* and *not pulling* arm i at time t. (Per Section 2, b denotes the probabilistic belief that an arm is in state $s = 1$; for restless arms, $b_t^i = s_t^i \in \{0, 1\}$).

$$W(b_t^i) = \inf_m \left\{m \mid V_m(b_t^i, a_t^i = 0) \geq V_m(b_t^i, a_t^i = 1)\right\} \quad (1)$$

The value function $V_m(b)$ represents the maximum expected discounted reward under passive subsidy m and discount rate β for arm i with belief state $b_t^i \in [0, 1]$ at time t:

$$V_m(b_t^i) = \max \begin{cases} m + r(b_t^i) + \beta V_m\left(b_{t+1}^i\right) & \textit{passive} \\ r(b_t^i) + \beta\left[b_t^i V_m\left(P^1_{1,1}\right) + (1 - b_t^i) V_m\left(P^1_{0,1}\right)\right] & \\ & \textit{active} \end{cases}$$

Once the Whittle index has been computed for each arm, the agent sorts the indices, and the k arms with the greatest index values receive a pull at time t, while the remaining $N - k$ arms are passive. Weber and Weiss [35] give sufficient conditions for *indexability*:

Definition 3.1. An arm is indexable if the set of beliefs for which it is optimal to be passive for a given m, $\mathcal{B}^*(m) = \{b \mid \forall \pi \in \Pi^*_m, \pi(b) = 0\}$, monotonically increases from $\emptyset$ to the entire belief space as m increases from $-\infty$ to $+\infty$. An RMAB is indexable if every arm is indexable.

Indexability is often difficult to establish, and computing the Whittle index can be complex [18]. Prevailing approaches rely on proving the optimality of a *threshold policy* for a subset of transition matrices [23]. A *forward* threshold policy pulls an arm when its state is at or below a given threshold, and makes the arm passive otherwise; the converse is true for a *reverse* threshold policy. Mate et al. [20] give such conditions for this RMAB setting, when $r(b) = b$, and provide an algorithm, Threshold Whittle, that is asymptotically optimal for forward threshold-optimal arms. Mate et al. [21] expand on this work for any non-decreasing $r(b)$ and present the Risk-Aware Whittle algorithm.

3.2 Motivation: Individual Welfare & Whittle

Bimodal allocation: Existing theory does not offer any guarantees about how the sequence of actions will be distributed over arms under Whittle index-based policies, nor about the probability with which a given arm can expect to be pulled at any particular timestep. Prins et al. [25] demonstrate that Whittle-based policies tend to allocate all pulls to a small number of arms, neglecting most of the population. We present similar findings in Appendix B.

This bimodal distribution is a consequence of how the Whittle index prioritizes arms. Whittle favors arms for whom a pull is most beneficial to achieving sustained occupancy in the "good" state, regardless of whether this results in the same subset of arms repeatedly receiving pulls. While the structural constraints in Sec. 2 ensure that a pull is beneficial for every arm, marginal benefit varies. Since reward is a function of each arm's underlying state, arms whose trajectories are characterized by a relative—*but not absolute*—indifference to the intervention are likely to be ignored.

Ethical implications: This zero-valued lower bound on the number of pulls an arm can receive aligns with a *utilitarian* approach to distributive justice, in which the decision-maker seeks to allocate resources so as to maximize total expected utility [3, 19]. This may be incompatible with competing pragmatic and ethical desiderata, including *egalitarian* and *prioritarian* notions of distributive fairness, in which the decision-maker seeks to allocate resources equally among arms (e.g., Round-Robin), or prioritize arms considered to be worst-off under the status quo, for some quantifiable notion of *worst-off* that induces a partial ordering over arms [28, 32]. We consider the *worst off* to be arms who would be deprived of algorithmic attention (e.g., not receive any pulls), or, from a probabilistic perspective, would have a *zero-valued lower bound* on the probability of receiving a pull at any given timestep.

Why algorithmic attention? This choice is motivated by our desire to improve *equality of opportunity* (i.e., access to the beneficial intervention) rather than *equality of outcomes* (i.e., observed adherence). The agent directly controls who receives the intervention, but has only indirect control (via actions) over the sequence of state transitions an arm experiences. Additionally, proclivity towards adherence may vary widely in the absence of restrictive assumptions about cohort homogeneity, and focusing on equality of outcomes could thus entail a significant loss of total welfare.

Distributive fairness and algorithmic acceptability: To realize the benefits associated with an algorithmically-derived resource allocation policy, practitioners tasked with implementation must find the policy to be acceptable (i.e., in keeping with their professional and ethical standards), and potential beneficiaries must find participation to be rational.

With respect to *practitioners*, many clinicians report experiencing mental anguish when resource constraints force them to categorically deny a patient access to a beneficial treatment, and may resort to providing improvised and/or sub-optimal care [6]. Providing fairness-aware decision support can improve acceptability [15, 27] and minimize the loss of reward associated with ethically-motivated deviation to a sub-optimal but equitable approach such as Round-Robin [8, 9]. For *beneficiaries*, we posit that an arm may consider participation rational when it results in an increase in expected time spent in the adherent state relative to non-participation (e.g., due to receiving a strictly positive number of pulls in expectation).

3.3 Time-indexed Fairness and Indexability

We now consider whether it is possible to modify the Whittle index to guarantee time-indexed fairness while preserving our ability to decouple arms. Unfortunately, the answer is no—we provide an overview here and a detailed discussion in Appendix D.1. Recall that structural constraints ensure that when an arm is considered in isolation, the optimal action will *always* be to pull, and that a Whittle-index approach computes the infimum subsidy, m, an arm requires to accept passivity at time t. Whether or not arm i is *actually* pulled at time t depends on how the subsidy of one arm *compares* to the infimum subsidies required by other arms. Thus, any modification intended to *guarantee* time-indexed fairness must be able to alter the ordering *among* arms, such that any arm i which would otherwise have a subsidy with rank $> k$ when sorted in descending order will now be in the top-k arms. Even if we could construct such a modification for a single arm without requiring time-stamped system information, if *every* arm had this same capability, then a new challenge would arise: we would be unable to distinguish among arms, and arbitrary tie-breaking could again jeopardize fairness constraint satisfaction.

3.4 Additional Related Work

While multi-armed bandit problems are canonically framed from the perspective of the decision-maker, interest in individual and group fairness in this setting has grown in recent years [7, 11, 17].

In the *stochastic* multi-armed bandit setting, each arm is characterized by a fixed but unknown average reward rather than by an MDP. The decision-maker thus faces uncertainty about the true utility of each arm and must balance exploration (i.e., pulling arms to gain information about their reward distributions) with exploitation (i.e., pulling the optimal arm(s)) to maximize expected reward. Joseph et al. [11] examine fairness among arms in this setting, and introduce a definition that requires the decision-maker to favor (i.e., select) arms with higher average reward over arms with lower average reward, even in the face of uncertainty. As the authors note, this definition *is* consistent with reward maximization, but imposes a cost in terms of per-round regret when *learning* the optimal policy, due to the fact that arms with overlapping confidence intervals are chained until they can be separated with high confidence.

Prior work in other non-restless bandit settings demonstrates that alternative definitions—i.e., those which center *distributive fairness among arms* as opposed to the principle that arms with similar average rewards should be treated similarly [10], generally entail deviation from optimal behavior. Li et al. [17] study the combinatorial *sleeping* bandit setting, in which arms are stochastic but may be unavailable at any given timestep. They introduce the minimum selection fraction constraint, which we adapt and refer to as time-indexed fairness (see Section 2). Chen et al. [7] consider the *contextual* bandit setting, and propose an algorithm that guarantees each arm a minimum probability of selection at each timestep.

In the *restless* setting that we consider, prior works have tended toward opposite ends of the reward-fairness spectrum by either: (1) redistributing pulls without providing arm-level guarantees [16, 21]; or (2) guaranteeing time-indexed fairness without providing optimality guarantees [25]. Recent work has also considered the adjacent problem of fairness among intervention *providers* (i.e., workers) [5]. In contrast to prior work, we aim to *guarantee* rather than incentivize fairness, without incurring an exponential dependency on the time horizon or sacrificing optimality guarantees. We thus seek an efficient policy that is reward maximizing, subject to the satisfaction of both budget and probabilistic fairness constraints.

4 METHODOLOGICAL APPROACH

Here we introduce ProbFair, an approximately optimal solution to a relaxed version of the allocation task in which we guarantee the satisfaction of *probabilistic* rather than *time-indexed* fairness, along with the budget constraint. This relaxation is necessary for tractability, as it allows us to precompute a stationary, *state-agnostic* probability vector, $\vec{p}^i$, from which constraint-satisfying discrete actions are drawn.

ProbFair maps each arm i to an arm-specific, stationary probability distribution over atomic actions, such that for each timestep t, $P[a_t^i = 1] = p_i$ and $P[a_t^i = 0] = 1 - p_i$, where $p_i \in [\ell, u]$ for all $i \in [N]$ and $\sum_i p_i = k$. Here, ℓ and u are user-defined fairness parameters satisfying $0 < \ell \leq \frac{k}{N} \leq u \leq 1$, per Section 2. Note that ℓT and uT can be interpreted as lower and upper bounds on the expected number of pulls an arm will receive over the time horizon.

In Section 4.1, we describe how to construct the p_i's so as to efficiently approximate our constrained reward-maximization objective within a multiplicative factor of $(1-\epsilon)$, for any given constant $\epsilon > 0$. We use a dependent rounding approach detailed in Section 4.2 to sample from this distribution at each timestep *independently*, to produce a discrete action vector, $\vec{a}_t \in \{0,1\}^N$, which is guaranteed to satisfy the budget constraint, k [33].

To motivate our approach, note that when we take the union of each arm's stationary probability vector, we obtain a system-level policy, $\pi_{PF} : \{i \mid i \in N\} \rightarrow [1 - p_i, p_i]^N$. Regardless of the system's initial state, repeated application of this policy will result in convergence to a steady-state distribution in which (WLOG) arm i is in the adherent state (i.e., state 1) with probability $x_i \in [\ell, u]$, and the non-adherent state (i.e., state 0) with probability $1 - x_i \in [0, 1]$.

By definition, for any arm i, x_i will satisfy the equation:

$$x_i \left[(1-p_i)P^0_{1,1} + p_i P^1_{1,1}\right] + (1-x_i)[(1-p_i)P^0_{0,1} + p_i P^1_{0,1}] = x_i. \quad (2)$$

Thus, $x_i = f_i(p_i)$, where

$$f_i(p_i) = \frac{(1-p_i)P^0_{0,1} + p_i P^1_{0,1}}{1 - (1-p_i)P^0_{1,1} - p_i P^1_{1,1} + (1-p_i)P^0_{0,1} + p_i P^1_{0,1}} \quad (3)$$

We seek the policy which maximizes total expected reward, where reward is non-decreasing in s (i.e., with time spent in the adherent state). Thus, ProbFair is defined as:

$$\pi_{\mathrm{PF}} = arg\,max_{p_i \in [\ell,u]} \sum_i f_i(p_i) \text{ s.t. } \sum_i p_i \leq k \quad (4)$$

Solving this constrained maximization problem is thus consistent with maximizing the expected number of timesteps each arm will spend in the adherent state, subject to satisfying the budget *and* probabilistic fairness constraints. We emphasize that our construction process takes the transition matrices of each arm i into account via f_i (Equation 3).

4.1 Computing the p_i's: Algorithmic Approach

Overview: To construct π_{PF}, we: (1) partition the arms based on the shapes of their respective f_i functions (Eq. 3); (2) perform a grid search over possible ways to allocate the budget, k, between the two subsets of arms; (2a) solve each sub-problem to produce a probabilistic policy for the arms in that subset; (2b) compute the total expected reward of the policy; (3) take the argmax over this set of grid search values to determine the approximately optimal budget allocation; and (4) form π_{PF} by taking the union over the policies produced by evaluating each sub-problem at its approximately optimal share of the budget. Figure 1 visualizes; the remainder of this section provides technical details.

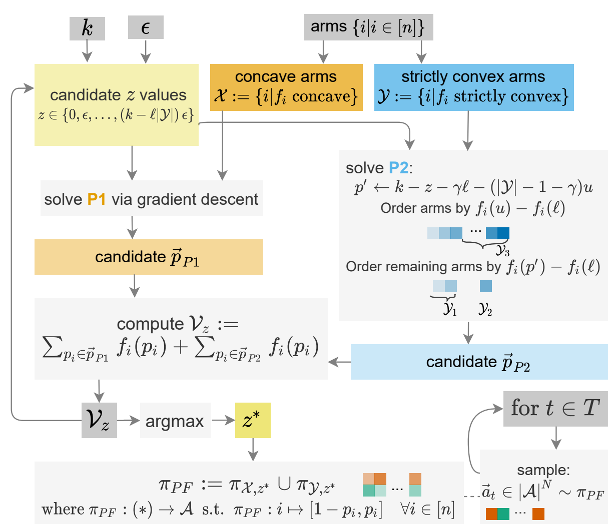


Figure 1: ProbFair: constructing and sampling from π_{PF}

To begin, we introduce two theorems (see App. E for full proofs):

Theorem 4.1. *For every arm $i \in [N]$, $f_i(p_i)$ is either* ***concave*** *or* ***strictly convex*** *in all of $p_i \in [0,1]$.*

Proof Sketch. WLOG, fix an arm $i \in [N]$. For notational convenience, let us define the following constants derived from the arm's transition matrix P^i: $c_1 := P^0_{0,1}$, $c_2 := P^1_{0,1} - P^0_{0,1}$, $c_3 := 1 - P^0_{1,1} + P^0_{0,1}$, and $c_4 := P^0_{1,1} - P^1_{1,1} - P^0_{0,1} + P^1_{0,1}$. Then

$$f_i^{''}(p_i) = \frac{2c_4^2\left(c_1 - \frac{c_2 c_3}{c_4}\right)}{(c_3 + c_4 p_i)^3}. \tag{5}$$

$c_3 + c_4 p_i \in (0,1)$ for all $p_i \in [0,1]$. Thus, the sign of $f_i^{''}(p_i)$ is determined by $c_1 - \frac{c_2 c_3}{c_4}$, which does not depend on p_i. □

Theorem 4.2. *For each arm $i \in [N]$, the structural constraints introduced in Section 2 ensure that $f_i(p_i)$ is monotonically non-decreasing in p_i over the interval $[0,1]$.*

Proof Sketch. WLOG, fix an arm $i \in [N]$. Theorem 4.2 follows directly from the second derivative. c_1, c_2, c_3, and c_4 are constants.

$$\frac{df_i}{dp_i} = \frac{c_2 c_3 - c_1 c_4}{(c_3 + c_4 p_i)^2} \tag{6}$$

By the structural constraints $P^0_{1,1} < P^1_{1,1}$ and $P^0_{0,1} < P^1_{0,1}$, the numerator is positive; the denominator is always positive. □

By Theorem 4.1, the arms can be partitioned into two disjoint sets: $\mathcal{X} = \{i \mid f_i \text{ is concave}\}$ and $\mathcal{Y} = \{i \mid f_i \text{ is strictly convex}\}$.

- **(P1)** maximize $\sum_{i \in \mathcal{X}} f_i(p_i)$ subject to: $p_i \in [\ell, u]$ for all $i \in \mathcal{X}$, and $\sum_{i \in \mathcal{X}} p_i = z$
- **(P2)** maximize $\sum_{i \in \mathcal{Y}} f_i(p_i)$ subject to: $p_i \in [\ell, u]$ for all $i \in \mathcal{Y}$, and $\sum_{i \in \mathcal{Y}} p_i = k - z$

Then, π_{PF} is the union of the solutions to **P1** and **P2** at the optimal grid search value $z^* = arg\,max_z \sum_{i \in \mathcal{X}} f_i(p_i) + \sum_{i \in \mathcal{Y}} f_i(p_i)$. Algorithm 1 provides pseudocode.

Algorithm 1 ProbFair

procedure ProbFair($[N], k, \epsilon, \ell, u$)
  $\mathcal{X} \leftarrow \{i \mid f_i$ is **concave** in all of $p_i \in [0,1]\}$
  $\mathcal{Y} \leftarrow \{i \mid f_i$ is **strictly convex** in all of $p_i \in [0,1]\}$
  `grid_search_vals` $\leftarrow \{\epsilon j \mid j \in [0, k - \ell|\mathcal{Y}|]\}$
  for $z \in$ `grid_search_vals` **do**
    $\pi_{\mathcal{X},z} \leftarrow$ SolveP1($\mathcal{X}, k, z, \ell, u$)
    $\pi_{\mathcal{Y},z} \leftarrow$ SolveP2($\mathcal{Y}, k, z, \ell, u$)
    $\mathcal{V}_z \leftarrow \sum_{p_i \in \pi_{\mathcal{X},z}} f_i(p_i) + \sum_{p_i \in \pi_{\mathcal{Y},z}} f_i(p_i)$
  $z^* \leftarrow arg\,max_z \mathcal{V}_z$
  $\pi_{PF} \leftarrow \pi_{\mathcal{X},z^*} \cup \pi_{\mathcal{Y},z^*}$
  return π_{PF}

P1 is a concave-maximization problem that can be solved efficiently via gradient descent. The computational complexity is $O\left(\frac{|\mathcal{X}|}{\epsilon^2}\right)$ [22]. To solve **P2**, we begin by introducing a lemma that we prove in Appendix E:

Lemma 4.3. ***P2*** *has an optimal solution in which $p_i \in (\ell, u)$ for at most one $i \in \mathcal{Y}$.*

Proof Sketch. Suppose for contradiction there exists some optimal solution $\vec{p}$ with distinct indices $i, j \in \mathcal{Y}$ such that $p_i, p_j \in (\ell, u)$. Let us compare $\vec{p}$ with a perturbed solution $p_i := p_i + \epsilon$ and $p_j := p_j - \epsilon$. Using a Taylor series expansion, the change in objective must be $(\epsilon^2/2) \cdot (f_i''(p_i) + f_j''(p_j)) + O(\epsilon^3)$. Since f_i and f_j are strictly convex, $f_i''(p_i) + f_j''(p_j) > 0$. Thus, the objective increases regardless of the sign of (tiny) ϵ, a contradiction. □

Given this structure, an optimal solution $\{p_i^* \mid i \in \mathcal{Y}\}$ will set some number of arms $\gamma \in \mathbb{Z}^+$ to ℓ, at most one arm to $p' \in (\ell, u]$, and the remaining $|\mathcal{Y}| - \gamma - 1$ arms to u. We represent these subsets by $\mathcal{Y}_1, \mathcal{Y}_2$, and $\mathcal{Y}_3$, respectively. Let $\gamma = \left\lfloor \frac{|\mathcal{Y}|u - (k-z)}{u - \ell} \right\rfloor$, and $p' = k - z - |\mathcal{Y}_1|\ell - |\mathcal{Y}_3|u \in (\ell, u]$. Intuitively, when the remaining budget $k - z$ allows us to set all arms in $\mathcal{Y}$ to u, $\gamma = |\mathcal{Y}_1| = 0$. Conversely, when there is only enough budget left to satisfy the fairness constraint for arms in $\mathcal{Y}$, $\gamma = |\mathcal{Y}_1| = |\mathcal{Y}|$. With the cardinality of each subset thus established, per Theorem 4.4 (see below), we use Algorithm 2 to optimally partition the arms in $\mathcal{Y}$.

Algorithm 2 SolveP2

Note: all sorts are ascending; arrays are zero-indexed.

procedure SolveP2($\mathcal{Y} \subseteq N, k, z, \ell, u$)
  $\gamma \leftarrow \left\lfloor \frac{|\mathcal{Y}|u - (k-z)}{u - \ell} \right\rfloor$
  $p' \leftarrow k - z - \gamma\ell - (|\mathcal{Y}| - 1 - \gamma)u$
  if $|\mathcal{Y}| - \gamma - 1 > 0$ **then**
    $\Delta_{\mathcal{Y}} = \texttt{sort}([f_i(u) - f_i(\ell)\ \forall i \in \mathcal{Y}])$
    $\mathcal{Y}_3 \leftarrow \{(\Delta_{\mathcal{Y}})\,[(|\mathcal{Y}| - \gamma - 1):]\}$
  else $\mathcal{Y}_3 \leftarrow \emptyset$
  $\Delta_{\mathcal{Y} \setminus \mathcal{Y}_3} = \texttt{sort}([f_i(p') - f_i(\ell)\ \forall i \in \mathcal{Y} \setminus \mathcal{Y}_3])$
  $\mathcal{Y}_1 \leftarrow \{(\Delta_{\mathcal{Y} \setminus \mathcal{Y}_3})\,[:\gamma]\}$
  $\mathcal{Y}_2 \leftarrow \{(\Delta_{\mathcal{Y} \setminus \mathcal{Y}_3})\,[\gamma]\}$
  $\pi_{\mathcal{Y}} := i \mapsto (\ell | i \in \mathcal{Y}_1) \vee (p' | i \in \mathcal{Y}_2) \vee (u | i \in \mathcal{Y}_3)$
  return $\pi_{\mathcal{Y}}$

Theorem 4.4. *Alg. 2 yields the mapping from arms in $\mathcal{Y}$ to subsets in $\{\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3\}$ which maximizes $\sum_{i \in \mathcal{Y}} f_i(p_i)$ s.t. $\sum_{i \in \mathcal{Y}} p_i = k - z$. (See Appendix E for the complete proof).*

Proof Sketch. By Lemma 4.3, there exists *at most one arm* with optimal value $p_i^* \in (\ell, u)$. By Lemma E.1, $\gamma := |\mathcal{Y}_1| = \left\lfloor \frac{|\mathcal{Y}|u - (k-z)}{u - \ell} \right\rfloor$ and $p' = k - z - |\mathcal{Y}_1|\ell - |\mathcal{Y}_3|u \in (\ell, u]$. Then, we can rewrite Equation 4 as an optimization problem over set assignment:

$$arg\,max_{\{\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3\}} \sum_{i \in \mathcal{Y}_1} f_i(\ell) + f_j(p') + \sum_{i'' \in \mathcal{Y}_3} f_{i''}(u)$$

$$\text{s.t. } |\mathcal{Y}_1| = \gamma,\ \mathcal{Y}_2 = \{j\},\quad \bigcap_{x=1}^{3} \mathcal{Y}_x = \emptyset, \text{ and } \bigcup_{x=1}^{3} \mathcal{Y}_x = \mathcal{Y}.$$

By algebraic manipulation, assigning the $|\mathcal{Y}| - \gamma - 1$ arms with maximal values of $f_i(u) - f_i(\ell)$ to $\mathcal{Y}_3$ produces a maximal solution. Similarly, we assign $j \in \mathcal{Y}_2$ if $f_j(p') - f_j(\ell)$ is maximal among the remaining arms. By definition, $\mathcal{Y}_1^* = \mathcal{Y} \setminus \left(\mathcal{Y}_2 \bigcup \mathcal{Y}_3^*\right)$, which completes the proof. □

Corollary 4.5. *Alg. 2 has time complexity* $O(|\mathcal{Y}| \log |\mathcal{Y}|)$.

With our solutions to **P1** and **P2** so defined, the cost of finding our probabilistic policy in this way is $O\left(\frac{k-\ell|\mathcal{Y}|}{\epsilon}\left(\frac{|\mathcal{X}|}{\epsilon^2} + |\mathcal{Y}| \log|\mathcal{Y}|\right)\right)$, which is at worst $O\left(\frac{kN}{\epsilon^3}\right)$ when all N arms are in $\mathcal{X}$.

4.2 Sampling Approach

For problem instances with feasible solutions, Algorithm 1 returns π_{PF}, a mapping from the set of arms to a set of stationary probability distributions over actions, such that for each arm i, the probability of receiving a pull at any given timestep is in $[\ell, u]$. By virtue of the fact that $\ell > 0$, this policy guarantees probabilistic fairness constraint satisfaction for all arms. We use a linear-time algorithm introduced by Srinivasan [33] and detailed in Appendix E.2 to sample from π_{PF} at each timestep, such that the following properties hold: (1) with probability one, we satisfy the budget constraint by pulling exactly k arms; and (2) any given arm i is pulled with probability p_i. Formally, each time we draw a vector of binary random variables $(X_1, X_2 \ldots X_N)$ from the distribution π_{PF}, $\Pr[|i : X_i = 1| = k] = 1$ and $\forall i, \Pr[X_i = 1] = p_i$.

5 EXPERIMENTAL EVALUATION

In this section, we empirically demonstrate that ProbFair enforces the probabilistic fairness constraint introduced in Section 2 with minimal loss in total expected reward, relative to fairness-aware alternatives. We begin by identifying our comparison policies, evaluation metrics, and datasets. We then present results from three experiments: (1) ProbFair versus fairness-inducing alternative policies, holding the cohort fixed and considering fairness-aligned sets of hyperparameters; (2) ProbFair evaluated on a breadth of cohorts representing different types of patient populations; and (3) ProbFair when fairness is *not* enforced (i.e., $\ell = 0$), to examine the cost of state agnosticism.[1]

5.1 Experimental Setup

Policies: In our experiments, we compare ProbFair against a subset of the following baseline§ and fairness-{inducing†, guaranteeing‡, and agnostic*} policies:

Random§	Select k arms uniformly at random at each t.
Round-Robin§,‡	Select k arms at each t in fixed, sequential order.
TW-based heuristics‡	Select top-k arms based on Whittle index values. Available arms vary based on time-indexed fairness constraint satisfaction [25].
Risk-Aware TW (RA-TW)†	Select top-k arms based on Whittle index values. Incentivizes fairness via concave reward function [21].
Threshold Whittle (TW)*	Select top-k arms based on Whittle index values [20, 36].

We specifically consider three **Threshold Whittle**-based heuristics: H_{First}, H_{Last}, and H_{Rand}. These heuristics partition the k pulls available at each timestep into (un)constrained subsets, where a pull is *constrained* if it is executed to satisfy a time-indexed fairness constraint. During constrained pulls, only arms that have not yet been pulled the required number of times within a ν-length interval are available; other arms are excluded from consideration, unless *all* arms have already satisfied their constraints. H_{First}, H_{Last}, and H_{Rand} position constrained pulls at the beginning, end, or randomly within each interval of length ν, respectively. Appendix F.1 provides pseudocode.

Objective: In all experiments, we assign equal value to the adherence of a given arm over time. Thus, we set our objective to reward occupancy in the "good" state: a simple local reward $r_t(s_t^i) := s_t^i \in \{0, 1\}$ and undiscounted cumulative reward function, $R(r(s)) := \sum_{i\in[N]} \sum_{t\in[T]} r(s_t^i)$.

Evaluation metrics: We are interested in comparing policies along two dimensions: reward maximization and fairness (i.e., with respect to the distribution of algorithmic attention). To this end, we rely on two performance metrics: (a) intervention benefit and (b) earth mover's distance.

Intervention benefit (IB) is the total expected reward of an algorithm, normalized between the reward obtained with no interventions (0% intervention benefit) and the asymptotically optimal but fairness-agnostic Threshold Whittle algorithm (100%) [20]. Formally,

$$\text{IB}_{\text{NoAct,TW}}(\text{ALG}) := \frac{\mathbb{E}_{\text{ALG}}[R_{\text{ALG}}(\cdot)] - \mathbb{E}_{\text{NoAct}}[R(\cdot)]}{\mathbb{E}_{\text{TW}}[R(\cdot)] - \mathbb{E}_{\text{NoAct}}[R(\cdot)]} \tag{7}$$

Per Lemma F.2 (App. F.2), the *price of fairness (PoF)* metric [4] is inversely proportional to intervention benefit. We thus report IB.

Earth mover's distance (EMD) is a metric that allows us to compute the minimum cost required to transform one probability distribution into another [30]. We use it to compare algorithms with respect to fairness—i.e., how evenly a set of pulls are allocated among arms. (Other metrics that may measure individual distributive fairness are discussed in Appendix F.2.)

For each algorithm, we consider a discrete distribution F of observed pull counts, where each bucket, $j \in \{0 \ldots T\}$, corresponds to a feasible number of total pulls that an arm could receive, and $F[j] \in \{0 \ldots N\}$ corresponds to the number of arms whose observed pull count is equal to j. For example, $F[0]$ corresponds to the quantity of arms never pulled, and $F[T]$ corresponds to the quantity of arms pulled at every timestep. Each algorithm produces kT total pulls, so the distributions have the same total mass.

We use Round-Robin as a fair reference algorithm since it distributes pulls evenly among arms. We then compute the minimum cost required to transform each algorithm's distribution, F_{ALG}, into that of Round-Robin's, F_{RR}.

For our application this is equivalent to:

$$\text{EMD}_{\text{RR}}(\text{ALG}) := \left|\sum_{h=0}^{T}\sum_{j=0}^{h} F_{\text{ALG}}[j] - F_{\text{RR}}[j]\right| \tag{8}$$

Unless otherwise noted, we normalize EMD such that the maximum distance we encounter, that of TW, is one:

$$\frac{\text{EMD}_{\text{RR}}(\text{ALG}) - \text{EMD}_{\text{RR}}(\text{RR})}{\text{EMD}_{\text{RR}}(\text{TW}) - \text{EMD}_{\text{RR}}(\text{RR})} = \frac{\text{EMD}_{\text{RR}}(\text{ALG})}{\text{EMD}_{\text{RR}}(\text{TW})} \tag{9}$$

Datasets: We evaluate performance on two datasets: (a) a realistic patient adherence behavior model and (b) a general set of randomly generated synthetic transition matrices.

[1]Code to reproduce our empirical results is provided at https://github.com/crherlihy/prob_fair_rmab.

CPAP Adherence. Obstructive sleep apnea (OSA) is a common condition that causes interrupted breathing during sleep [26]; when used throughout the entirety of sleep, continuous positive airway pressure therapy (CPAP) eliminates nearly 100% of obstructive apneas for the majority of treated patients [31]. However, poor adherence behavior in using CPAP reduces its beneficial outcomes. CPAP non-adherence affects an estimated 30-40% of patients [29].

We derive the CPAP dataset that we use in our experiments from the work of Kang et al. [13, 14], who model the dynamics and patterns of patient adherence behavior as a basis for designing effective and economical interventions. In particular, we adapt their Markov model of CPAP adherence behavior (a three-state system based on hours of nightly CPAP usage) to a two-state system using the clinical standard for adherence–at least four hours of CPAP machine usage per night [31]. Kang et al. [13] find, via expectation-maximization on CPAP usage patterns, that patients can be divided into two groups based on this clinical standard. Though patients in the first cluster occasionally miss a night, these patients utilize a CPAP machine for more than four hours every night without assistance, while patients in the second cluster do not. We refer to the latter cluster as the *non-adherent* cohort in our analysis.

Kang et al. [14] consider many intervention effects. We specifically consider an intervention effect, $\alpha_{\text{interv}} = 1.1$, that broadly characterizes supportive interventions such as telemonitoring and phone support, which are associated with a moderate 0.70 hours (95% CI ± 0.35) increase in device usage per night [1]. We add random $\sigma = 1$ logistic noise to the transition matrices so that there is some variance in individual arm dynamics. To prevent overlap with the general cohort we consider for contrast, added noise can only *hinder* the probability of adherence in the non-adherent cohort.

Synthetic. In addition, we construct a synthetic dataset of randomly generated arms such that the structural constraints outlined in Section 2 are preserved. We conjecture that forward (reverse) threshold-optimal arms are a subset of concave (strictly convex) arms (see Appendix F.3).

5.2 ProbFair vs. Fairness-aware Alternatives

Here we compare ProbFair to policies which either *induce* or *guarantee* fairness. The former includes Risk-aware Whittle (RA-TW), which incentivizes fairness via concave reward $r(b)$ [21]. We use the authors' suggested reward function $r(b) = -e^{\lambda(1-b)}$, $\lambda = 20$. This imposes a large negative utility on lower belief values, which motivates preemptive intervention. However, RA-TW does not *guarantee* time-indexed or probabilistic fairness for individual arms. The latter includes Round-Robin and the First, Last, and Random heuristics, which guarantee time-indexed fairness but do *not* provide any optimality guarantees.

In Table 1, we report average results for each policy, along with margins of error for 95% confidence intervals, computed over 100 simulation seeds for a synthetic cohort of 100 collapsing arms, with $k = 20$ and $T = 180$. To facilitate meaningful comparisons between ProbFair and the heuristics, we consider combinations of values for ℓ and ν that produce equivalent, integer-valued lower bounds on the number of pulls any arm can expect to receive—i.e., $\min_i \mathbb{E}[\sum_t \mathbb{1}(a_t^i = 1)] = \ell \times T = \frac{T}{\nu}$.

$\min_i \mathbb{E}$[# pulls]	Policy		$\mathbb{E}$[IB] (%)	$\mathbb{E}$[EMD] (%)
10, $\ell = 0.056$, $\nu = 18$	PF	ℓ	88.73 ± 0.26	81.78 ± 0.18
	H_{First}	ν	86.11 ± 0.26	71.53 ± 0.13
	H_{Last}	ν	87.37 ± 0.28	**70.48 ± 0.12**
	H_{Rand}	ν	**90.79 ± 0.22**	74.12 ± 0.15
18, $\ell = 0.1$, $\nu = 10$	PF	ℓ	80.80 ± 0.30	59.96 ± 0.19
	H_{First}	ν	76.62 ± 0.30	49.54 ± 0.09
	H_{Last}	ν	77.95 ± 0.30	**49.26 ± 0.08**
	H_{Rand}	ν	**81.53 ± 0.30**	52.73 ± 0.10
30, $\ell = 0.167$, $\nu = 6$	PF	ℓ	**66.12 ± 0.35**	23.61 ± 0.12
	H_{First}	ν	63.58 ± 0.31	**18.98 ± 0.03**
	H_{Last}	ν	64.63 ± 0.34	19.47 ± 0.04
	H_{Rand}	ν	65.21 ± 0.32	19.64 ± 0.04
comparison	RA-TW		85.12 ± 0.42	**95.80 ± 0.42**
	TW		**100.00 ± 0.00**	100.00 ± 0.00
baseline	Random		50.02 ± 0.35	10.08 ± 0.10
	NoAct		0.00 ± 0.00	73.48 ± 0.13
	RR		**56.96 ± 0.33**	**0.00 ± 0.00**

Table 1: Expected intervention benefit and normalized earth mover's distance by policy and fairness bracket.

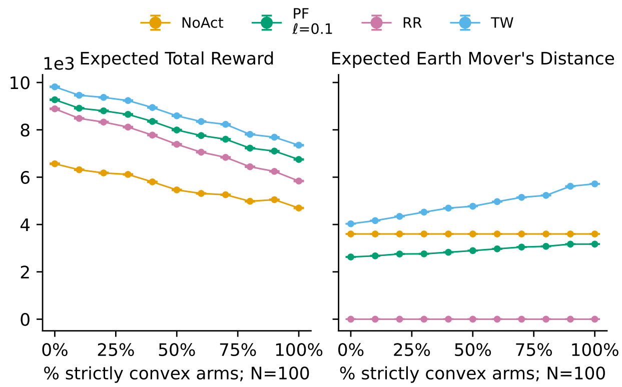
Key findings from this experiment include:

- *Fairer* hyperparameter values (i.e, $\ell \uparrow$, $\nu \downarrow$), correspond to decreases in $\mathbb{E}$[IB] and $\mathbb{E}$[EMD], reflecting improved individual fairness at the expense of total expected reward.
- ProbFair is competitive with respect to RA-TW, outperforming on both metrics when $\ell = 0.056$, and incurring a slight loss in $\mathbb{E}$[IB] but improvement in $\mathbb{E}$[EMD] for $\ell = 0.1$.
- For each (ℓ, ν) combination, ProbFair performs competitively with respect to the best-performing heuristic (which, like TW, are state-aware, see Section 5.4).

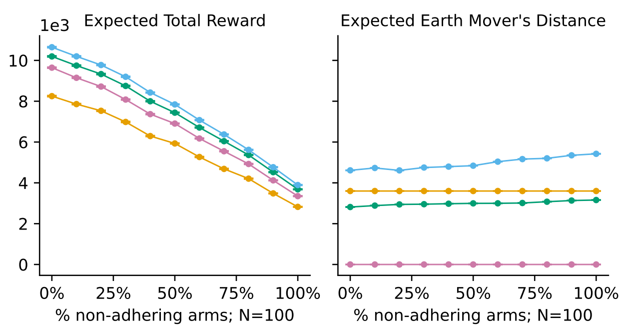
5.3 ProbFair on a Breadth of Cohorts

In this section we conduct sensitivity analysis with respect to cohort composition. For each dataset, we identify a transition matrix characteristic that can be modified during the generation process to produce a subset of arms that will exhibit less favorable transition dynamics than their peers. For the synthetic dataset, this characteristic is *strict convexity*. For the CPAP dataset, it is *non-adherence*, a mnemonic coined by Kang et al. [13] to characterize a cluster of study participants, and contrast this to a model fit on the general patient population.

For each dataset, we generate ten different cohorts, each of which is characterized by the percentage of unfavorable arms that it contains. We use a seed to control the generation process such that each cohort contains 100 collapsing arms in total. A sliding window of the unfavorable arms we can generate with this seed are included as we increase the cardinality of the unfavorable subset. For ease of interpretation, we present unnormalized results over 100 simulation seeds with $k = 20$ and $T = 180$ in Figure 2, and then proceed to summarize normalized performance.



(a) Synthetic results



(b) CPAP results

Figure 2: Expected total reward (left) and unnormalized earth mover's distance (right) on a breadth of cohorts.

Key findings from this experiment include:

- Per Figure 2, for each dataset, expected total reward predictably declines for all policies as the percentage of unfavorable arms increases, while unnormalized EMD increases for TW and ProbFair.
  - *Synthetic*: As the proportion of strictly convex arms increases, ProbFair's allocation of resources tends towards the bimodality of TW.
  - *CPAP*: As the proportion of non-adherent arms increases, the level of intervention required to improve trajectories rises, but the budget constraint is static.
- For each dataset, ProbFair's normalsnipized performance remains stable even as cohort composition is varied:
  - *Synthetic*: With respect to IB (EMD), ProbFair achieves an average (over all cohorts) of averages (over 100 simulations per cohort) of 80.69% ± 1.42% (58.98% ± 1.29%).
  - *CPAP*: The values for IB (EMD) are: 79.84% ± 0.68% (59.68% ± 1.08%).

5.4 ProbFair: Price of State Agnosticism

Here, we investigate the cost associated with ProbFair's state agnosticism, relative to state-aware Threshold Whittle. To ensure a fair comparison, we set $\ell = 0$ and $u = 1$, effectively constructing a version of ProbFair in which probabilistic fairness is not enforced. (Recall that TW is fairness-agnostic; in the previous results, we do not expect ProbFair to obtain the same total reward as TW).

Although ProbFair incorporates each arm's structural information (i.e., transition matrices), it produces a set of *stationary* probability distributions over actions from which all discrete actions are subsequently drawn. TW, in contrast, ingests each arm's current state at each timestep, and is thus able to exploit *realized* sequences of state transitions.

While we thus expect ProbFair to incur some loss in intervention benefit, our results (computed over 100 simulation seeds, with $k = 20$, $N = 100$, and $T = 180$) indicate that this loss is acceptable rather than catastrophic. Relative to TW, $\text{ProbFair}_{\ell=0}$ obtains 97.41% ± 0.26 of $\mathbb{E}[\text{IB}]$ and incurs an increase of only 4.56% ± 0.19 with respect to $\mathbb{E}[\text{EMD}]$.

6 CONCLUSION AND FUTURE WORK

In this paper, we introduce ProbFair, a novel, probabilistically fair algorithm for constrained resource allocation. Our theoretical results prove that this policy is reward-maximizing, subject to the guaranteed satisfaction of both budget and tunable probabilistic fairness constraints. Our empirical results demonstrate that ProbFair preserves utility while providing fairness guarantees. Promising future directions include: (1) extending ProbFair to address larger state and/or action spaces; and (2) relaxing the requirement for stationarity in the construction of π_{PF}.

ACKNOWLEDGMENTS

Christine Herlihy was supported by the National Institute of Standards and Technology's (NIST) Professional Research Experience Program (PREP). John Dickerson and Aviva Prins were supported in part by NSF CAREER Award IIS-1846237, NSF D-ISN Award #2039862, NSF Award CCF-1852352, NIH R01 Award NLM-013039-01, NIST MSE Award #20126334, DARPA GARD #HR00112020007, DoD WHS Award #HQ003420F0035, and a Google Faculty Research award. Aravind Srinivasan was supported in part by NSF awards CCF-1749864 and CCF-1918749, as well as research awards from Adobe, Amazon, and Google. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of U.S. government or funding agencies. We thank Samuel Dooley, Dr. Furong Huang, Naveen Raman, and Daniel Smolyak for helpful input and feedback.

REFERENCES

[1] Kathleen Askland, Lauren Wright, Dariusz R Wozniak, Talia Emmanuel, Jessica Caston, and Ian Smith. 2020. Educational, Supportive and Behavioural Interventions to Improve Usage of Continuous Positive Airway Pressure Machines in Adults with Obstructive Sleep Apnoea. *Cochrane Database of Systematic Reviews* 4 (2020).
[2] Turgay Ayer, Can Zhang, Anthony Bonifonte, Anne C Spaulding, and Jagpreet Chhatwal. 2019. Prioritizing Hepatitis C Treatment in US Prisons. *Operations Research* 67, 3 (2019), 853–873.
[3] Jeremy Bentham. 1781. *An Introduction to the Principles of Morals and Legislation.* Technical Report. McMaster University Archive for the History of Economic Thought.
[4] Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. 2011. The Price of Fairness. *Operations research* 59, 1 (2011), 17–31.
[5] Arpita Biswas, Jackson A Killian, Paula Rodriguez Diaz, Susobhan Ghosh, and Milind Tambe. 2023. Fairness for Workers Who Pull the Arms: An Index Based

Policy for Allocation of Restless Bandit Tasks. In *22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. London, UK.

[6] Catherine R. Butler, Susan P. Y. Wong, Aaron G. Wightman, and Ann M. O'Hare. 2020. US Clinicians' Experiences and Perspectives on Resource Limitation and Patient Care During the COVID-19 Pandemic. *JAMA Network Open* 3, 11 (11 2020), e2027315–e2027315. https://doi.org/10.1001/jamanetworkopen.2020.27315

[7] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. 2020. Fair Contextual Multi-Armed Bandits: Theory and Experiments *(Proceedings of Machine Learning Research, Vol. 124)*, Jonas Peters and David Sontag (Eds.). PMLR, Virtual, 181–190. http://proceedings.mlr.press/v124/chen20a.html

[8] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[9] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2011. Fairness Through Awareness. *CoRR* abs/1104.3913 (2011). arXiv:1104.3913 http://arxiv.org/abs/1104.3913

[11] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 325–333. http://papers.nips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.pdf

[12] Young Hun Jung and Ambuj Tewari. 2019. Regret Bounds for Thompson Sampling in Episodic Restless Bandit Problems. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 9007–9016. http://papers.nips.cc/paper/9102-regret-bounds-for-thompson-sampling-in-episodic-restless-bandit-problems.pdf

[13] Yuncheol Kang, Vittaldas V Prabhu, Amy M Sawyer, and Paul M Griffin. 2013. Markov models for treatment adherence in obstructive sleep apnea. *Age* 49 (2013), 11–6.

[14] Yuncheol Kang, Amy M Sawyer, Paul M Griffin, and Vittaldas V Prabhu. 2016. Modelling Adherence Behaviour for the Treatment of Obstructive Sleep Apnoea. *European journal of operational research* 249, 3 (2016), 1005–1013.

[15] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key Challenges for Delivering Clinical Impact with Artificial Intelligence. *BMC medicine* 17, 1 (2019), 195.

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

[17] Fengjiao Li, Jia Liu, and Bo Ji. 2019. Combinatorial Sleeping Bandits with Fairness Constraints. *CoRR* abs/1901.04891 (2019). arXiv:1901.04891 http://arxiv.org/abs/1901.04891

[18] Keqin Liu and Qing Zhao. 2010. Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access. *IEEE Transactions on Information Theory* 56, 11 (Nov 2010), 5547–5567. https://doi.org/10.1109/tit.2010.2068950

[19] Elliot Marseille and James G. Kahn. 2019. Utilitarianism and the Ethical Foundations of Cost-Effectiveness Analysis in Resource Allocation for Global Health. *Philosophy, Ethics, and Humanities in Medicine* 14, 1 (2019), 1–7. https://doi.org/10.1186/s13010-019-0074-7

[20] Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. 2020. Collapsing Bandits and Their Application to Public Health Intervention. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020).

[21] Aditya Mate, Andrew Perrault, and Milind Tambe. 2021. Risk-Aware Interventions in Public Health: Planning with Restless Multi-Armed Bandits. In *20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. London, UK.

[22] Yurii Nesterov et al. 2018. *Lectures on convex optimization*. Vol. 137. Springer.

[23] José Niño-Mora. 2020. A Verification Theorem for Threshold-Indexability of Real-State Discounted Restless Bandits. *Mathematics of Operations Research* 45, 2 (2020), 465–496.

[24] Christos H Papadimitriou and John N Tsitsiklis. 1994. The Complexity of Optimal Queueing Network Control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*. IEEE, 318–322.

[25] Aviva Prins, Aditya Mate, Jackson A Killian, Rediet Abebe, and Milind Tambe. 2020. Incorporating Healthcare Motivated Constraints in Restless Bandit Based Resource Allocation. *preprint* (2020). https://teamcore.seas.harvard.edu/publications/incorporating-healthcare-motivated-constraints-restless-bandit-based-resource

[26] Naresh M Punjabi. 2008. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society* 5, 2 (2008), 136–143.

[27] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of internal medicine* 169, 12 (2018), 866–872.

[28] John Rawls. 1971. *A Theory of Justice* (1 ed.). Belknap Press of Harvard University Press, Cambridge, Massachussets.

[29] Brian W Rotenberg, Dorian Murariu, and Kenny P Pang. 2016. Trends in CPAP adherence over twenty years of data collection: a flattened curve. *Journal of Otolaryngology-Head & Neck Surgery* 45, 1 (2016), 1–9.

[30] Yossi Rubner, Carlo Tomasi, and Leonidas Guibas. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision* 40 (11 2000), 99–121. https://doi.org/10.1023/A:1026543900054

[31] Amy M Sawyer, Nalaka S Gooneratne, Carole L Marcus, Dafna Ofer, Kathy C Richards, and Terri E Weaver. 2011. A systematic review of CPAP adherence across age groups: clinical and empiric insights for developing CPAP adherence interventions. *Sleep medicine reviews* 15, 6 (2011), 343–356.

[32] Leslie Scheunemann and Douglas White. 2011. The Ethics and Reality of Rationing in Medicine. *Chest* 140 (12 2011), 1625–32. https://doi.org/10.1378/chest.11-0622

[33] A. Srinivasan. 2001. Distributions on Level-Sets with Applications to Approximation Algorithms. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science (FOCS '01)*. IEEE Computer Society, USA, 588.

[34] Lauren N. Steimle and Brian T. Denton. 2017. *Markov Decision Processes for Screening and Treatment of Chronic Diseases*. Springer International Publishing, Cham, 189–222. https://doi.org/10.1007/978-3-319-47766-4_6

[35] Richard R Weber and Gideon Weiss. 1990. On an Index Policy for Restless Bandits. *Journal of Applied Probability* (1990), 637–648.

[36] Peter Whittle. 1988. Restless Bandits: Activity Allocation in a Changing World. *Journal of applied probability* 25, A (1988), 287–298.

A APPENDIX

For the full version of this paper complete with appendixes including additional experimental, modeling, and theoretical results, we invite the reader to the full paper at arXiv:2106.07677. Code to reproduce our empirical results is provided at https://github.com/crherlihy/prob_fair_rmab.