Indexability is Not Enough for Whittle: Improved, Near-Optimal Algorithms for Restless Bandits

Abheek Ghosh
University of Oxford
abheek.ghosh@cs.ox.ac.uk

Dheeraj Nagaraj
Manish Jain
Google Research
{dheerajnagaraj,manishjn}@google.com

Milind Tambe
Google Research
Harvard University
milindtambe@google.com

ABSTRACT

We study the problem of planning restless multi-armed bandits (RMABs) with multiple actions. This is a popular model for multi-agent systems with applications like multi-channel communication, monitoring and machine maintenance tasks, and healthcare. Whittle index policies, which are based on Lagrangian relaxations, are widely used in these settings due to their simplicity and near-optimality under certain conditions. In this work, we first show that Whittle index policies can fail in simple and practically relevant RMAB settings, *even when* the RMABs are indexable. We discuss why the optimality guarantees fail and why asymptotic optimality may not translate well to practically relevant planning horizons.

We then propose an alternate planning algorithm based on the mean-field method, which can provably and efficiently obtain near-optimal policies with a large number of arms, without the stringent structural assumptions required by the Whittle index policies. This borrows ideas from existing research with some improvements: our approach is hyper-parameter free, and we provide an improved non-asymptotic analysis which has: (a) no requirement for exogenous hyper-parameters and tighter polynomial dependence on known problem parameters; (b) high probability bounds which show that the reward of the policy is reliable; and (c) matching sub-optimality lower bounds for this algorithm with respect to the number of arms, thus demonstrating the tightness of our bounds. Our extensive experimental analysis shows that the mean-field approach matches or outperforms other baselines.

KEYWORDS

Restless Bandits; Resource Allocation; Mean-Field; Whittle Index

ACM Reference Format:
Abheek Ghosh, Dheeraj Nagaraj, Manish Jain, and Milind Tambe. 2023. Indexability is Not Enough for Whittle: Improved, Near-Optimal Algorithms for Restless Bandits. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 9 pages.

1 INTRODUCTION

Multi-Armed Bandits have been extensively studied [23, 36, 40] and widely deployed in online decision-making. Restless multi-armed bandit (RMAB) is a specific instance of this, which deals with optimal allocation of limited resources to multiple agents/arms. Each arm corresponds to a known Markov decision process (MDP), making this a planning problem for multi-agent reinforcement learning. While this is computationally hard [31], there are index policy based approximation algorithms that utilize Lagrangian relaxations of the associated integer programming problems. These policies, called the Whittle index policies [43], are simple to implement and are asymptotically optimal under certain structural assumptions on the MDPs. RMABs have been applied across a multitude of domains like multi-channel communication [3, 17, 30, 37, 50], monitoring and machine maintenance [46], security [33], and healthcare [5, 24, 27, 34], including deployments in the field [28].

However, these index policies require assumptions like homogeneity, indexability, irreducibility, infinite-horizon average reward, and global attractor properties of certain McKean-Vlasov type flows [17, 40] associated with the Lagrange relaxations to have asymptotic optimality guarantees. These properties can be hard to even verify and practitioners often apply Whittle index policies for their applications without verification. The general argument for doing so, in the words of [40], is: *... the evidence so far is that counterexamples to the conjecture (asymptotic optimality of Whittle index policy) are rare and that the degree of sub-optimality is very small.* In this work, our first contribution is to show that these assumptions cannot be taken for granted even in natural, practically relevant situations by constructing two-action RMABs where Whittle index based policies perform sub-optimally by a large margin.

The most emphasized property required for using the Whittle index policy is indexability. Even Whittle's seminal paper [43] that introduced RMABs and proposed the Whittle index policy defined the policy only for indexable RMABs. [40], among others, show that without indexability, Whittle index policy can be sub-optimal. However, all our examples *are* of indexable RMABs. Therefore, our work further shows that even with indexability, Whittle index policy can be arbitrarily bad compared to the optimal policy for some natural examples. For these examples, even intuitively, the Whittle index policy seems sub-optimal. The examples we provide are well-motivated and are inspired from practical applications of RMABs, e.g. [21, 28].

Our second contribution is to propose the mean-field planning algorithm (MFP) that works provably, reliably, and efficiently with minimal assumptions in settings with a large number of agents N. MFP is based on mean-field limits, and considers a continuum limit of a system with a large number of similarly behaving agents. This allows us to remove the combinatorial complexity and obtain analytically tractable objects like differential/difference equations (called the McKean-Vlasov type equations [17, 40]), albeit with a small approximation error. Our method first maps the problem for any finite N to a continuum limit, where linear programming based

Full version (with appendix) of the paper: https://arxiv.org/abs/2211.00112.

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

methods can be efficiently deployed to solve the problem. Then, it maps the solution of the continuum limit back to the finite N system to provide us with an approximately optimal solution.

Our third contribution is to provide a tight, non-asymptotic convergence of MFP to the optimal policy as a function of N. In contrast to Whittle index policies, MFP works without any of the restrictive structural assumptions on the MDPs themselves. Our algorithmic approach is similar to the algorithm proposed by [47], and our analysis shares similarities with the analysis in [38]. However, our approach does not require precision hyper-parameter tuning of those two approaches and hence is more generic. Our theoretical analysis improves upon the ones in prior works by polynomial factors of problem dependent parameters.

Our final contribution is to provide an experimental evaluation of MFP in different settings, especially ones of practical importance. Our approach performs as well as or outperforms state-of-the-art index based policies consistently in all the experiments with a large number of agents.

2 RELATED WORK

Mean-Field Limits: Mean field limits are extensively studied in statistical physics and probability theory [7, 19, 29, 41], game theory [10, 15], networks and control theory [8, 9, 11–13, 25, 45], and reinforcement learning [1, 26, 39, 44] in order to understand the behavior of stochastic dynamical systems with multiple interacting particles or agents. Works like [7, 13, 22] consider convergence of multi-agent continuous time control systems to McKean-Vlasov equations under general conditions.

Restless Bandits and Approximation Limits: [32] shows that obtaining the exact solution to the RMAB planning problem is PSPACE-hard. This led to research on approximation algorithms for the optimal policy whenever the number of arms N is large. For the case of two action RMABs, the practically effective 'Whittle index' policy was proposed in [43], and was shown to be asymptotically optimal under certain structural assumptions in [40]. [4] extends this Lagrangian relaxation based approach to a hierarchy of relaxations culminating in the exact solution. Recent works have also extended Langrangian relaxations to consider more complex MDPs with multiple actions [14, 17, 20, 21]. In the two-action case, [38] extends the Whittle index to the non-indexable setting, but with other structural assumptions. [6, 18] consider Lagrangian relaxations for finite horizon two-action MDPs. While these policies do not require indexability, the guarantees in [6, 18] suffer from an exponential dependence on the horizon. In contrast, [48, 49] consider fluid balance policies where the relaxation allows the resource constraint to hold in expectation, but this does not allow for multiple actions.

Our main algorithm MFP is based on the algorithm proposed in [47] which does not require indexability or any other structural assumptions on the MDP. The method in [47] requires additional hyper-parameters to be carefully set in order to ensure that the constraints are satisfied and the relaxation is near-optimal as $N \to \infty$ with an error of $\tilde{O}(\sqrt{N})$. We modify the sub-routine that translates the solution to the mean-field LP to a policy for the RMAB, and obtain the following four improvements: (i) Our method is hyper-parameter free, which is very attractive to the practitioner. (ii) Suppose all arms are not identical, but there are K clusters of arms each associated with a different MDP with $|S|$ states. Our error bounds scale as $\sqrt{K|S|N}$ whereas the results in [47] scale as $\sqrt{K^4|S|^4N}$. Consequently, when $K|S|$ scales as $o(\sqrt{N}) \cap \omega(N^{1/8})$, our bounds still ensure asymptotically optimal policy unlike the prior work.[1] (iii) We obtain high probability confidence bounds for the random discounted reward under the mean-field policy, showing that this policy is at most $O(\sqrt{K|S|N})$ away from the optimal *expected* reward with *high probability*. (iv) We obtain matching $\Omega(\sqrt{N})$ lower bounds for the sub-optimality, which shows that the analysis is tight with respect to the number of arms N.

3 NOTATION AND PRELIMINARIES

We begin by describing a (Multi-Action) Restless Multi-Armed Bandit (RMAB) problem. Let T denote the time horizon for planning, and $[1 : T]$ (shorthanded as $[T]$) denote the set of positive integers $\{1, 2, \ldots, T\}$. In an RMAB, each arm $i \in [N]$ at time $t \in [T]$ corresponds to an MDP $(S, A, R_{t,i}, P_{t,i})$, where S is the set of states, and A is the set of actions. We use $P_{t,i}(s'|s, a)$ to denote arm i's transition probability from state s to s' under action a at time t. Similarly, let $R_{t,i} : S \times A \to \mathbb{R}_{\geq 0}$ and $C_{t,i} : S \times A \to \mathbb{R}_{\geq 0}$ denote the reward and the cost, respectively, of playing action a for arm i in state s at time t. Notice that the rewards and transitions can change with time. We assume that there is an action of cost 0 for each t, i, and s.[2]

We require that the actions must satisfy the budget constraint: $\sum_{i\in[N]} C_{t,i}(s_{t,i}, a_{t,i}) \leq B_t \in \mathbb{R}_{\geq 0}$. A policy $\pi = (\pi_t)_{t\in[T]}$ recommends a joint action $(a_{t,i})_{i\in[N]}$ as a function of the joint state $(s_{t,i})_{i\in[N]}$ i.e., $\pi_t : S^N \to A^N$ for each $t \in [T]$. Note that, the MDPs of all agents collectively form a single, large MDP with state-space S^N. This admits a deterministic optimal policy, which maps the joint state deterministically to a single joint action $(a_{t,i}) \in A^N$ (see [35]). Therefore, we will work with deterministic policies only.

The expected discounted reward starting from state s_1 under the policy π is defined as $V^{\pi}_{\gamma}(s_1) = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} \sum_{i\in[N]} R_{t,i}(s_{t,i}, \pi_t(s_t)_i) \middle| s_1\right]$ where the next state is drawn according to $s_{t+1,i} \sim P_{t,i}(\cdot|s_{t,i}, \pi_t(s_t)_i)$ independently for $i \in [N]$ when conditioned on s_t. $\gamma \in [0, 1]$ is the discount factor, the actions are selected according to the policy π, and $\pi_t(s_t)_i$ is the i-th component of $\pi_t(s_t)$. The planner's goal is to maximize the total expected reward.

3.1 RMABs with Clusters

While known theoretical bounds require homogeneity of the arms (i.e, the associated MDPs are identical), different arms behave differently in practically relevant settings. For example, different patients might have different reactions to the same treatment, as observed in healthcare applications [27, 28]. A clustering approach, where arms are grouped into clusters is often applied in these scenarios [28]. Each arm belongs to exactly one cluster, and the MDPs associated with all arms in a given cluster are identical. This setting allows for certain aspects of the model to scale with the number of clusters and

[1] Assuming that the optimal expected reward is $\Theta(N)$, which is usually the case.

[2] This is without loss of generality for t and i: all the costs and budget can always be reduced by the minimum positive cost to get a zero cost action. While this condition is *with* loss of generality for s, it is not limiting in practice given most RMAB applications have a default zero cost action.

not the arms, and the model itself can be learned in a statistically efficient way from data (e.g. using collaborative filtering) without playing each arm individually multiple times.

We shall use the notation based on clusters throughout the paper, and we describe this notation in Table 1. The state of a cluster can be succinctly described as the vector of counts of arms in that given cluster and a given state, due to the homogeneity of arms (i.e, arms in a given state and a given cluster are following identical MDPs and are indistinguishable from each other in terms of reward maximization). Thus, we represent the state of the cluster as $\widehat{\mu}_t = (\widehat{\mu}_{t,i}(s))_{i\in[K],s\in S}$, and the action at time t as $\widehat{\alpha}_t = (\widehat{\alpha}_{t,i}(s,a))_{i\in[K],s\in S,a\in A}$. A few identities are a direct consequence of our definitions:

$$\sum_{s\in S}\widehat{\mu}_{t,i}(s) = N_i,\ \sum_{a\in A}\widehat{\alpha}_{t,i}(s,a) = \widehat{\mu}_{t,i}(s),\ \forall i\in[K], t\in[T], s\in S.$$

As an illustrative example, suppose that there are four states $S = \{s_1, s_2, s_3, s_4\}$ and three actions $A = \{a_1, a_2, a_3\}$. Say, cluster i has 100 arms, and at time t, the states s_1, s_2, and s_3 have 30 arms each and 10 arms are in the fourth state s_4. Then, the succinctly represented state of cluster i is $\widehat{\mu}_{t,i} = (30, 30, 30, 10)$, where $\widehat{\mu}_{t,i}(s_1) = \widehat{\mu}_{t,i}(s_2) = \widehat{\mu}_{t,i}(s_3) = 30$ and $\widehat{\mu}_{t,i}(s_4) = 10$. Similarly, for the 30 arms in cluster i and state s_1, if we take action a_1 for 15 arms, a_2 for 10 arms, and a_3 for the remaining 5 arms, then $\widehat{\alpha}_{t,i}(s_1,\cdot) = (15, 10, 5)$.

Symbol	Meaning
S	State space, subscripted as s
A	Action space, subscripted as a
T	Time horizon for the RMAB, subscripted as t
K	Number of clusters, subscripted as i
$P_{t,i}(s'\|s,a)$	Transition probability from s to s', given a at time t for an agent in cluster i for $a\in A, s, s'\in S$
$R_{t,i}(s,a)$	Reward in state s for playing action a for an agent in cluster i at time t for $a\in A, s\in S$
$C_{t,i}(s,a)$	Cost in state s for playing action a for an agent in cluster i at time t for $a\in A, s\in S$
N_i	Number of arms in cluster i, where all the N_i add up to N arms
$\widehat{\mu}_{t,i}(s)$	Number of arms in cluster i in state s at time t
$\widehat{\alpha}_{t,i}(s,a)$	Number of times the action a is applied to the set of arms in the state s and cluster i in time t
γ	Discount factor, $\gamma\in[0,1]$

Table 1: Notation based on clustered RMABs.

4 FAILURE EXAMPLES FOR INDEX POLICIES

We now provide natural examples motivated by applications in mobile health care [2, 16, 42] where the Whittle index policy performs much worse than the optimal policy (and the policy found by the mean-field method). In these domains, healthcare workers reach out to patients through mobile apps or phone calls to get them to adhere to treatment regimens. Here, the patients are modeled as arms, with their adherence with the mobile health program being represented as an MDP [21, 28]. The action space of the healthcare workers are the different modes of communication (including no outreach). The number of healthcare workers (resource) is limited and not all beneficiaries can be regularly contacted by them. Patients with similar MDPs are grouped together in clusters, similar to clustered RMABs we described in Section 3.1. We now describe three specific RMAB examples inspired from this domain, and compute the Whittle indices for them.

Example 1. In this example (see Figure 1), the healthcare workers make one of two actions per patient—call or no call. A patient either decides to continue to engage with the program, or not participate further (e.g. delete the application from the mobile device, or block the incoming call number). Here, we model two types of patients: (a) *greedy*: a patient that engages with the program in the first time step but does not continue in the program even after repeated attempts of healthcare workers; (b) *reliable*: a patient who is careful to engage in the beginning but will remain in the program for long periods if contacted by healthcare workers consistently. The RMAB model for this scenario has two types of arms corresponding to the *reliable* and *greedy* types of patients, three states per arm (*start*, *engaged*, and *dropout*), and two actions (*active*, or call and *passive*, or no call).

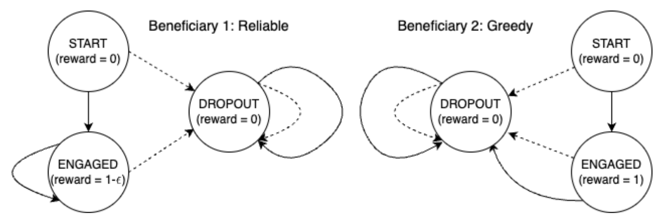


Figure 1: Example 1 - Whittle index policy is sub-optimal. (Solid arrows: active actions; dotted arrows: passive actions.)

The behavior of the arms for each type of patient is graphically shown in Figure 1. The solid lines are transitions for the *active* action ($a = 1$), and the dotted lines are for the *passive* action ($a = 0$). The nodes represent the 3 states of *start* s_s, *engaged* s_e and *dropout* s_d for the patients. For the patient of the *reliable* type, the engaged state gives a reward of $1-\epsilon$ and all other states give a 0 reward. For the patient of the *greedy* type, the engaged state gives a reward of 1 and all other states give a 0 reward. The transition probabilities for all transitions for both patient types are given in Table 2. The parameters for reliable arms are denoted by superscript of r and greedy by g. We initialize the RMAB to have n arms of each type (i.e. total arms $N = 2n$), as well as the budget $B = n = N/2$ for every time step. All agents start at their respective s_s state.

	Reliable Arm	Greedy Arm
Active Action	$P^r(s_e\|s_s, a=1) = 1$ $P^r(s_e\|s_e, a=1) = 1$ $P^r(s_d\|s_d, a=1) = 1$	$P^g(s_e\|s_s, a=1) = 1$ $P^g(s_d\|s_e, a=1) = 1$ $P^g(s_d\|s_d, a=1) = 1$
Passive Action	$P(s_d\|s, a=0) = 1\ \forall s\in\{s_s, s_e, s_d\}$	
All other transition probabilities are 0.		

Table 2: Transition Probabilities for Example 1

The key behavior is that a *reliable* patient moves to the state s_d only after a *passive* action, whereas a *greedy* patient moves to s_d

latest by the second time step regardless of the actions. Also, once in s_d, an agent stays there forever. So, the greedy arms give rewards for only one time-step, while the reliable arms give rewards for forever if we keep on allocating resources to them. If the discount factor γ is not too small, we can intuit that a good policy should allocate resources to the reliable arms. The theorem below shows that the Whittle index does not follow this observation and hence is sub-optimal. We refer to the full version of the paper for its proof.

Theorem 1. *The instance described in Example 1 is indexable. The ratio of optimal reward and the Whittle index policy reward is at least* $\frac{1-\epsilon}{1-\gamma}$*, which goes to* ∞ *as* $\gamma \to 1$ *for any* $0 < \epsilon < 1$ *and* $0 < \gamma < 1$.

Note that the result in Theorem 1 holds for any $n \geq 1$ (and hence $n \to \infty$), thus making the Whittle index policy asymptotically sub-optimal. Example 1 is indexable (Theorem 1) but not homogeneous (i.e, different arms behave differently with respect to the same action). While homogeneity is one of the conditions required by [40] for the optimality of the Whittle index policy, *lack of homogeneity is not the reason why the Whittle index policy is sub-optimal* here. We show this by transforming this example into a homogenous RMAB by combining the reliable and greedy MDPs into a single MDP.

Example 2. As shown in Figure 2, consider only one type of arm with five states: *reliable-start*, *reliable-engaged*, *greedy-start*, *greedy-engaged*, and *dropout*, which are equivalent to the states in Example 1. That is, *reliable-start* state behaves like the *start* state of the *reliable* arm, and so on. Similarly, as in Example 1, we define there to be $2n$ arms (n in each of *reliable-start* and *greedy-start*) and a budget of n.[3]

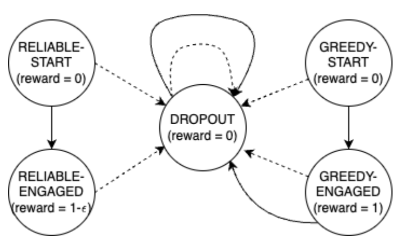


Figure 2: Example 2 - Whittle index policy is sub-optimal. (Solid arrows: active actions; dotted arrows: passive actions.)

As in Example 1, the Whittle index policy prefers the arms in the *greedy-start* state compared to the arms in the *reliable-start* state. The analysis and the results are exactly the same as for Theorem 1: Example 2 is indexable and the Whittle index policy can be arbitrarily worse compared to the optimal policy as $\gamma \to 1$.

Both these examples show that the Whittle index is sub-optimal for indexable and homogeneous RMABs *even as* $N \to \infty$. Thus, the natural question is: *what about the asymptotic optimality proof of [40]*? In addition to these conditions, [40] requires that an RMAB satisfy certain 'irreducibility' and 'global attractor' properties, and proves the optimality of Whittle index for the infinite horizon average reward. However, these conditions often go unverified by designers who implement such policies.

We now further modify Example 2 to make it irreducible[4] for every possible single-MDP policy (in addition to homogeneity and indexability). These properties, together with the global attractor property, guarantee that a Whittle index based policy is optimal for the *infinite horizon average reward* problem. However, most environments of practical interest are either finite time horizon or use discounted infinite horizon reward. We show in Example 3 that Whittle index based policies can be sub-optimal in such practical applications, and should be used after careful analysis and experimentation.

Example 3. In this example, we consider the same *state* and *rewards* as in Example 2, but modify the transition probabilities to ensure irreducibility with respect to any policy. This was not the case previously since, for example, an arm could never transition from s_e^g to s_e^r. Consider the following transition probabilities:

- $P(s_{gs}, 0, s_{ge}) = \eta_s = 1 - P(s_{gs}, 0, s_d)$,
- $P(s_{rs}, 0, s_{re}) = \eta_s = 1 - P(s_{rs}, 0, s_d)$,
- $P(s_{re}, 1, s_d) = \eta_r = 1 - P(s_{re}, 0, s_{re})$,
- $P(s_d, \cdot, s_{gs}) = P(s_d, \cdot, s_{rs}) = \eta_d = (1 - P(s_d, \cdot, s_d))/2$,

where s_d is the dropout state, s_s^g is greedy-start state and so on. η_s, η_r, and η_d are positive constants, and are called ergodicity parameters. We set $\eta_s = 0.05$, $\eta_r = 0.1$, $\eta_d = 0.1$, and $\epsilon = 0.01$.

We prove the indexability of Example 3 graphically (see full version of the paper) by plotting the difference in q-values of active and passive actions as we increase the subsidy for the passive action. Note that the MDP satisfies indexabilitiy since this plot has a negative slope and crosses the x-axis for each state at a unique point, and that the index (x-intercept) of the greedy-start state (s_s^g) is more than the reliable-start (s_s^r) state. While we do this for infinite horizon average reward, the same applies for discounted reward and finite horizon reward as well.

In Table 3, we show that the Whittle index is sub-optimal for either discounted reward or finite horizon for Example 3. The lower bound for the optimal reward is given by the index policy that always prefers s_s^r to s_s^g. The difference is also more significant if the ergodicity parameters are made smaller, e.g., if we set $\eta_s = \eta_r = \eta_d = 0.01$, then even with $\gamma = 0.95$, Whittle policy gets a discounted reward of 3.04 compared to at least 9.32 for the optimal policy.

Setting	Whittle	Optimal
discounted, $\gamma = 0.95$	7.33	≥ 8.65
discounted, $\gamma = 0.8$	1.17	≥ 1.86
finite horizon, $T = 20$	7.41	≥ 9.11

Table 3: Estimated rewards for Example 3.

For Example 3, the Whittle index policy is sub-optimal for finite time or discounted reward because the *mixing time* (time it takes to approximately reach the stationary distribution) is large in comparison to the (effective) planning horizon T or $\frac{1}{1-\gamma}$. Whittle needs the system to be close to the stationary distribution to start

[3] We can also modify this MDP so that every arm starts at the same initial state by adding a dummy *start* state while maintaining indexability and homogeneity. This dummy state transitions to either *reliable-start* or *greedy-start* state with probability 0.5 irrespective of the action played. After one time step, this ensures a similar initial condition as described in Example 2 for large N.

[4] A Markov chain is irreducible if there is a non-zero probability to go from any state to any other state after sufficiently many steps.

acting near-optimally. As the MDP in Example 3 is irreducible for every policy, the mixing time is finite, which ensures optimality of Whittle for non-discounted infinite time reward (given other optimality conditions of [40] are also satisfied). In the full version of the paper, we show that Whittle's reward goes towards optimality if we either increase the effective horizon or decrease the mixing time for Example 3.

We have not explicitly tested for the global attractor property in our examples, which states that the McKean-Vlasov differential equation which gives the time evolution of the empirical distribution of the states as $N \to \infty$ under the Whittle index policy converges to a unique fixed point, for every initial condition [40]. However, verifying this property is hard in general. In our numerical experiments, the system always converged to the fixed point, which makes us believe that our example satisfies the global attractor property. This further motivates the need for an algorithm whose guarantees does not require these multiple pre-conditions.

5 THE MEAN-FIELD METHOD

We now propose the mean-field planning (MFP) method as an approximation technique for finding the optimal policy for an RMAB. The basic idea behind MFP is that when N is large, the space of empirical distributions of the states can be approximated closely by the simplex $\Delta(S)$ (up to a quantization error of $O(1/N)$). When N is large, a 'law of large numbers' type effect shows that the evolution of the empirical measure due to a fixed policy is almost deterministic. Indeed, in MFP, we use this property to approximate the stochastic process of each cluster's state $\widehat{\mu}_t = (\widehat{\mu}_{t,i}(s))_{i\in[K],s\in S}$ by a deterministic process denoted by $\overline{\mu}_t = (\overline{\mu}_{t,i}(s))_{i\in[K],s\in S}$.

We define the mean-field MDP corresponding to the RMAB instance under consideration below and construct a linear program which solves the mean-field MDP exactly. Our MFP algorithm, described in Section 5.3, solves this linear program at each time instant and quantizes the optimal solution in order to compute a feasible action for the RMAB.

5.1 The Mean-Field MDP

Given the state of the RMAB and the action taken at $t = 1$, $\widehat{\mu}_1$ and $\widehat{\alpha}_1$ respectively, the state of the RMAB at $t = 2$ is a random variable $\widehat{\mu}_2 = (\widehat{\mu}_{2,i}(s'))_{i,s'}$ where $\widehat{\mu}_{2,i}(s') = \sum_{s,a}\sum_{\ell\in[\widehat{\alpha}_{1,i}(s,a)]}\mathbb{I}_{\{x_\ell(s,a)=s'\}}$, $x_\ell(s,a)$ is a random state in S picked based on the distribution $P_{1,i}(\cdot|s,a)$, and $\mathbb{I}_{\{\dots\}}$ denotes the indicator function. MFP approximates this *random variable* $\widehat{\mu}_2$ by the *deterministic* $\overline{\mu}_2 = (\overline{\mu}_{2,i}(s))_{i,s}$, where $\overline{\mu}_{2,i}(s') = \mathbb{E}[\widehat{\mu}_{2,i}(s')] = \sum_{s\in S,a\in A} P_{1,i}(s'|s,a)\,\widehat{\alpha}_{1,i}(s,a)$. In a similar fashion, we can approximate the random variable $\widehat{\mu}_t$ by the deterministic $\overline{\mu}_t$ for each $t \in [T]$. As we shall show later, whenever the number of agents is large, the typical value of $\widehat{\mu}_2 - \overline{\mu}_2$ is $O(\sqrt{N})$. Similarly, a mean-field action $\overline{\alpha}_t = (\overline{\alpha}_{t,i}(s,a))_{i,s,a}$ roughly corresponds to the expected number of arms in cluster i and state s for which we play the action a at time t and it can be fractional.

As an illustrative example of the mean-field approximation: Suppose there are two states $S = \{s_1, s_2\}$ and two actions $A = \{a_1, a_2\}$. Let cluster i have 5 arms, and at $t = 1$, all the arms in cluster i are in state s_1, i.e., $\widehat{\mu}_{1,i} = (5,0)$. For all 5 arms in cluster i and state s_1, say we play a_1, i.e., $\alpha_{1,i}(s_1,\cdot) = (5,0)$. Let's assume $P_{1,i}(\cdot|s_1,a_1) = (0.3,0.7)$, i.e., an agent in cluster i and state s_1 subjected to action a_1 moves to state s_1 w.p. (with probability) 0.3 and state s_2 w.p. 0.7. The real-life state at $t = 2$ for cluster i is a random variable $\widehat{\mu}_{2,i}$, where $\widehat{\mu}_{2,i} = (\ell, 5-\ell)$ w.p. $\binom{5}{\ell}(0.3)^\ell(0.7)^{5-\ell}$ for $\ell \in [0:5]$. On the other hand, the mean-field method makes a deterministic approximation of $\widehat{\mu}_{2,i}$ in the form of the mean-field state $\overline{\mu}_{2,i} = \mathbb{E}[\widehat{\mu}_{2,i}] = (1.5, 3.5)$. Notice that $\overline{\mu}_{2,i}$ can take fractional values although $\widehat{\mu}_{2,i}$ can never be fractional. Similarly, $\overline{\alpha}_{2,i}$ can also be fractional, e.g., $\overline{\alpha}_{2,i}(s_1,\cdot) = (1, 0.5)$ is a valid mean-field action at time 2 for this example.

Recall the original MDP defined in Section 3. The mean-field MDP is a deterministic MDP with state space $\mathsf{S}^{\mathsf{MF}} = N\Delta([K]\times S) = \{Nx : x \in \Delta([K]\times S)\}$ and action space $\mathsf{A}^{\mathsf{MF}} \subseteq N\Delta([K]\times S\times A)$, with horizon T. Given an action $a \in A$, let $P_{t,i}(\cdot|\cdot,a)$ denote the $|S| \times |S|$ transition matrix for cluster i under action a at time t. Using $R_{t,i}(\cdot,a)$ and $C_{t,i}(\cdot,a)$ to denote the $|S|$-dimensional reward and cost vectors, the reward and the cost for playing $\overline{\alpha}_{t,i}$ for cluster i at time t can be written as $\sum_{a\in A}\overline{\alpha}_{t,i}(\cdot,a)^\intercal R_{t,i}(\cdot,a)$ and $\sum_{a\in A}\overline{\alpha}_{t,i}(\cdot,a)^\intercal C_{t,i}(\cdot,a)$, respectively, where $x^\intercal$ denotes the transpose of a vector x. An action $\overline{\alpha}_t$ is feasible at time t if $\sum_{i\in[K]}\sum_{a\in A}\overline{\alpha}_{t,i}(\cdot,a)^\intercal C_{t,i}(\cdot,a) \le B_t$ and $\sum_{a\in A}\overline{\alpha}_{t,i}(\cdot,a) = \overline{\mu}_{t,i}(\cdot)$ for every $i \in [K]$. For all $t \in [T-1]$, we define the mean-field state evolution of state $\overline{\mu}_t$ under any feasible action $\overline{\alpha}_t$ as:

$$\overline{\mu}_{t+1,i}(\cdot)^\intercal = \sum_{a\in A}\overline{\alpha}_{t,i}(\cdot,a)^\intercal P_{t,i}(\cdot|\cdot,a).$$

5.2 The Mean-Field LP

For the mean-field MDP defined above, we can find the optimal action by solving the following LP. This LP gives a mean-field action $\overline{\alpha}$ computed using the real-life state at time $t = 1$, $\widehat{\mu}_1$, which is known.

$$\max_{\overline{\alpha}_t,\overline{\mu}_t} \sum_{t,i,a}\gamma^{t-1}\,\overline{\alpha}_{t,i}(\cdot,a)^\intercal R_{t,i}(\cdot,a) \tag{1}$$

$$\text{s.t.} \qquad \overline{\mu}_{1,\mathrm{i}} = \widehat{\mu}_{1,\mathrm{i}}, \qquad \forall i \tag{2}$$

$$\overline{\mu}_{t+1,i}(\cdot)^\intercal = \sum_{a\in A}\overline{\alpha}_{t,i}(\cdot,a)^\intercal P_{t,i}(\cdot|\cdot,a), \qquad \forall t \in [T-1], \forall i \tag{3}$$

$$\sum_{i\in[K]}\sum_{a\in A}\overline{\alpha}_{t,i}(\cdot,a)^\intercal C_{t,i}(\cdot,a) \le B_t, \qquad \forall t \tag{4}$$

$$\sum_{a\in A}\overline{\alpha}_{t,i}(\cdot,a) = \overline{\mu}_{t,i}(\cdot), \qquad \forall t,i \tag{5}$$

$$\overline{\alpha}_{t,i}(\cdot,a) \ge 0, \qquad \forall t,i,a \tag{6}$$

Equation (4) is the budget constraint whereas Equations (5) and (6) guarantee consistency. This LP is similar to the the LP given in [47] with the hyperparameters ϵ set to 0. Next, we describe the mean-field policy computed using this LP - i.e, a policy computed for the real-world RMAB using the mean-field solution $\overline{\alpha}$. Note that this is different from the randomized policy computation in [47].

5.3 RMAB Policy from the Mean-Field LP

We are given the initial real-life state $\widehat{\mu}_1$. For $t = 1, 2, \dots, T$ execute the steps described below:

(1) Solve the above LP with the initial state as $\widehat{\mu}_t$ and the time horizon as $(T-t+1)$, corresponding to the MDP considered from times t to T. Let's denote the mean-field states and

actions computed by solving the LP as $\overline{\mu}^{(t)} = (\overline{\mu}_\tau^{(t)})_{\tau \in [t:T]}$ and $(\overline{\alpha}_\tau^{(t)})_{\tau \in [t:T]}$, respectively. Notice that given an initial state at time t, $\overline{\mu}_t^{(t)} = \widehat{\mu}_t$.

(2) Play the action $\overline{\alpha}_t^{(t)}$. Notice that, by construction, $\sum_a \overline{\alpha}_{t,i}^{(t)}(s,a) = \overline{\mu}_{t,i}^{(t)}(s) = \widehat{\mu}_{t,i}(s)$. But, some of the $\overline{\alpha}_{t,i}^{(t)}(s,a)$ values may not be integers. So, we take a `floor` and play the action a for $\lfloor \overline{\alpha}_{t,i}^{(t)}(s,a) \rfloor$ arms in cluster i and state s at time t. For the remaining arms (in cluster i and state s at time t), which is equal to $\widehat{\mu}_{t,i}(s) - \sum_a \lfloor \overline{\alpha}_{t,i}^{(t)}(s,a) \rfloor$, play the 0 cost action. This ensures that the 'rounded' action stays within the budget B_t.

(3) After playing the action at time t, we get the realization of the real-life state at time $t+1$, $\widehat{\mu}_{t+1}$.

Step 2 introduces a potential *rounding error* in the algorithm. However, in our experiments, we observed that the mean-field actions were generally integral. This is likely because the costs, rewards and budget in our experiments were all integers, and our optimal solution of the resulting LP may be an integral extreme point. Since this might not hold for general RMAB instances, we incorporate and bound the rounding error in our analysis.

In the algorithm above, we solve the LP for each time step, which is an accepted deployment paradigm for practitioners because it allows the algorithm to adjust its policy based on the evolution of the real-life state. If a designer prefers to solve the LP only once, then they may use an alternate algorithm we discuss in the full version of the paper. While this variant may have lower performance in practice, both the algorithms have similar worst-case error bounds. We again refer to the full version of the paper for such a worst-case example.

6 ANALYSIS: NEAR-OPTIMALITY OF MFP

In this section, we prove that the mean-field planning (MFP) method is close to optimal if the number of arms is large compared to $K|S|$. Specifically, we shall show that the reward we get from the real-life process when we play actions computed using the mean-field method is close to the expected reward we can get by playing actions from an optimal real-life policy, with high probability. All omitted proofs are provided in the full version of the paper.

6.1 Intuition Behind the Proof

Lemma 6 presents a simple relaxation based argument to show that the optimal reward for the mean-field MDP is always larger than of the optimal expected reward for the RMAB. Therefore, we reduce the proof to showing that the state-evolution given by the application of the mean-field policy to the RMABs (given by $\widehat{\mu}_t$) 'tracks' the mean-field flow given by $\overline{\mu}_t$, up-to a small error. That is, we want to bound $\| \widehat{\mu}_t - \overline{\mu}_t^{(1)} \|_1$.

To this end, we decompose the error produced in each time step into two types: (1) Rounding Error, and (2) Variance Error. Rounding error occurs since the actions proposed by the mean-field solution, given by $\overline{\alpha}_t^{(t)}$, need not be integers. As proposed in the algorithm, we quantize these actions so that the budget is not exceeded. We can show by a simple counting argument that this error is at-most $O(K|S||A|)$ and independent of N.

The mean-field MDP, computed at time t, assumes that $\overline{\mu}_t^{(t)}$ transitions to $\overline{\mu}_{t+1}^{(t)}$ exactly. However, this is only true in expectation. As we show, the actual empirical state $\widehat{\mu}_{t+1}$ is such that $\| \overline{\mu}_{t+1}^{(t)} - \widehat{\mu}_{t+1} \|_1 = \tilde{O}(\sqrt{K|S|N})$ with high probability due to the inherent randomness present in the system. We call this the variance error. By bounding the accumulation of these two kinds of errors, we show that the mean-field policy applied to the RMAB performs near optimally.

6.2 Preliminaries

In our analysis, we shall use a version the Azuma-Hoeffding inequality for bounds on multinomial distributions.

Lemma 1. *Let $X_1, \ldots, X_n$ be independent random vectors in $\{e_1, \ldots, e_k\}$ (where e_i denotes the i-th standard basis vector in $\mathbb{R}^k$)*

(1) $\mathbb{E}[|| \sum_{i \in [n]} (X_i - \mathbb{E}[X_i]) ||_1] \le \sqrt{kn}$, *and*

(2) $\mathbb{P}[|| \sum_{i \in [n]} (X_i - \mathbb{E}[X_i]) ||_1 \ge \epsilon] \le \delta$ *for all* $0 < \delta < 1$, *where* $\epsilon = \sqrt{2\log(2)kn + 2n\log(\frac{1}{\delta})}$.

To make the analysis concise, let us introduce some additional notation. Let $R_t(\mu_t, \alpha_t) \equiv R_t(\alpha_t) = \sum_{i,s,a} R_{t,i}(s,a)\alpha_{t,i}(s,a)$ denote the total reward at time t for playing an action α_t. Let $\overline{\mathrm{V}}_{t:T}(\mu_t)$ be the optimal objective value of the mean-field LP for time t to T with μ_t as the start state at time t (Section 5.3). Similarly, let $\widehat{\mathrm{V}}^{\pi}_{t:T}(\mu_t)$ denote the real-life reward for using any RMAB policy π for time $[t:T]$ starting with state μ_t. Note that $\widehat{\mathrm{V}}^{\pi}_{t:T}(\mu_t)$ is random. Let $\widehat{\mathrm{V}}_{t:T}(\mu_t) = \max_\pi \mathbb{E}[\widehat{\mathrm{V}}^{\pi}_{t:T}(\mu_t)]$ be the maximum expected reward achieved by any RMAB policy. We shall also use the notation introduced in Section 5.3: $\overline{\pi}$ to denote the mean-field policy; $\overline{\mu}_\tau^{(t)}$ and $\overline{\alpha}_\tau^{(t)}$, for $t \in [T]$, $\tau \in [t:T]$, to denote the states and actions as computed by the LPs in Section 5.3. Let $R_{max} = \max_{t,i,s,a} R_{t,i}(s,a)$.

6.3 Analysis

As mentioned in Section 3, an RMAB is a large but finite MDP and has an optimal deterministic Markov policy. Therefore, we restrict ourselves to comparing the reward of the mean-field policy to that of deterministic Markov policies. It can also be observed from the description of the mean-field policy in Section 5.3 that it is deterministic and Markov. We shall first focus on the finite horizon non-discounted reward, where $\gamma = 1$ and $T < \infty$ and later extend the analysis to infinite horizon discounted rewards.

Theorem 2. *The mean-field policy $\overline{\pi}$ has a non-discounted reward for finite time horizon T that is at most $O(R_{max}T^2\sqrt{K|S|N})$ less than the expected reward of an optimal policy. In particular,*

(1) *lower-bound for expected reward:* $\mathbb{E}[\widehat{\mathrm{V}}^{\overline{\pi}}_{1:T}(\widehat{\mu}_1)] \ge \widehat{\mathrm{V}}_{1:T}(\widehat{\mu}_1) - \frac{T^2 R_{max}}{4}(\sqrt{K|S|N} + 5K|S||A|)$,

(2) *high-probability lower-bound for reward:* $\widehat{\mathrm{V}}^{\overline{\pi}}_{1:T}(\widehat{\mu}_1) \ge \widehat{\mathrm{V}}_{1:T}(\widehat{\mu}_1) - \frac{T^2 R_{max}}{4}(\sqrt{2\log(2)K|S|N + 2N\log(\frac{T}{\delta})} + 5K|S||A|)$ *with probability at least* $(1-\delta)$ *for every* $0 < \delta < 1$.

Proof. We shall prove our result using a sequence of lemmas (Lemmas 2 to 6).

First, we prove a Lipschitz condition on the objective value of the optimal solution of the mean-field LP. In particular, say μ and

μ' are two start states such that the ℓ_1-distance between μ and μ' is small, we shall show the difference between the optimal objective value computed by the LP given these start states is also small.

Lemma 2. *Let μ_t and μ'_t be two arbitrary states at time t with $||\mu_t - \mu'_t||_1 = \delta$. Then the difference in the optimal objective value of the mean-field LP for time steps t to T starting from μ_t and μ'_t is bounded by*

$$|V_{t:T}(\mu_t) - V_{t:T}(\mu'_t)| \leq (T - t + 1)R_{max}\delta/2.$$

In the next lemma, we bound the rounding error introduced in step (2) of the mean-field policy described in Section 5.3.

Lemma 3. $|\widehat{V}^{\overline{\pi}}_{1:T}(\widehat{\mu}_1) - \sum_t R_t(\overline{\alpha}^{(t)}_t)| \leq TK|S||A|R_{max}$ *almost surely.*

Next lemma connects the real-life reward for using the mean-field policy to the objective value of the mean-field LP at time $t = 1$.

Lemma 4. *We have the following bound:* $|\sum_t R_t(\overline{\alpha}^{(t)}_t) - \overline{V}_{1:T}(\overline{\mu}_1)| \leq \sum_{t\in[2:T]} \frac{(T-t+1)R_{max}||\overline{\mu}^{(t)}_t - \overline{\mu}^{(t-1)}_t||_1}{2}$.

Notice that $\overline{\mu}^{(t)}_t$ is the state at time t when playing the mean-field policy and $\overline{\mu}^{(t-1)}_t$ is the mean-field estimate of the state at time t as estimated by the mean-field LP at time $t - 1$. Next two lemmas bound $||\overline{\mu}^{(t)}_t - \overline{\mu}^{(t-1)}_t||_1$.

Lemma 5. *We have the following bounds on $||\overline{\mu}^{(t)}_t - \overline{\mu}^{(t-1)}_t||_1$:*

(1) $\mathbb{E}[||\overline{\mu}^{(t)}_t - \overline{\mu}^{(t-1)}_t||_1] \leq \sqrt{K|S|N} + K|S||A|$,

(2) $\mathbb{P}\big[||\overline{\mu}^{(t)}_t - \overline{\mu}^{(t-1)}_t||_1 \geq \sqrt{2\log(2)K|S|N + 2N\log(1/\delta)} + K|S||A|\big] \leq \delta$, *for every* $0 < \delta < 1$.

Next we show that the objective value of the mean-field LP at time $t = 1$, $\overline{V}_{1:T}(\overline{\mu}_1)$, upper bounds the expected reward of any real-life policy.

Lemma 6. *Let $\widehat{\pi}$ denote any arbitrary Markov policy. The expected reward of this policy is bounded above by the objective value of the mean-field LP at time $t = 1$, i.e., $\mathbb{E}[\widehat{V}^{\widehat{\pi}}_{1:T}(\widehat{\mu}_1)] \leq \overline{V}_{1:T}(\overline{\mu}_1)$, where $\widehat{\mu}_1 = \overline{\mu}_1$ is an arbitrary start state.*

We are now ready to prove our theorem using the lemmas given above. From Lemma 6, we know that $\mathbb{E}[\widehat{V}^{\widehat{\pi}}_{1:T}(\widehat{\mu}_1)] \leq \overline{V}_{1:T}(\overline{\mu}_1)$ for every Markov policy $\widehat{\pi}$. So, the optimal expected real-life reward $\widehat{V}_{1:T}(\widehat{\mu}_1) = \max_{\widehat{\pi}} \mathbb{E}[\widehat{V}^{\widehat{\pi}}_{1:T}(\widehat{\mu}_1)] \leq \overline{V}_{1:T}(\overline{\mu}_1)$. Further, using Lemma 3 and Lemma 4, we have $\overline{V}_{1:T}(\overline{\mu}_1) \leq \widehat{V}^{\overline{\pi}}_{1:T}(\widehat{\mu}_1) + \sum_{t\in[2:T]} \frac{(T-t+1)R_{max}||\overline{\mu}^{(t)}_t - \overline{\mu}^{(t-1)}_t||_1}{2} + TK|S||A|R_{max}$. Putting together, the optimal real-life reward is bounded above by the reward received by the mean-field policy as follows: $\widehat{V}_{1:T}(\widehat{\mu}_1) \leq \widehat{V}^{\overline{\pi}}_{1:T}(\widehat{\mu}_1) + TK|S||A|R_{max} + \sum_{t\in[2:T]} \frac{(T-t+1)R_{max}||\overline{\mu}^{(t)}_t - \overline{\mu}^{(t-1)}_t||_1}{2}$.

Using Lemma 5, we can lower-bound $\mathbb{E}[\widehat{V}^{\overline{\pi}}_{1:T}(\widehat{\mu}_1)]$ as $\mathbb{E}[\widehat{V}^{\overline{\pi}}_{1:T}(\widehat{\mu}_1)] \geq \widehat{V}_{1:T}(\widehat{\mu}_1) - \frac{T^2R_{max}}{4}(5K|S||A| + \sqrt{K|S|N})$, and with high probability bound $\widehat{V}^{\overline{\pi}}_{1:T}(\widehat{\mu}_1)$ as $\mathbb{P}\big[\widehat{V}^{\overline{\pi}}_{1:T}(\widehat{\mu}_1) \geq \widehat{V}_{1:T}(\widehat{\mu}_1) - \frac{T^2R_{max}}{4}(5K|S||A| + \sqrt{2\log(2)K|S|N + 2N\log(T/\delta)})\big] \leq 1 - \delta$, for every $0 < \delta < 1$. □

The next theorem proves the result for infinite horizon discounted reward. When the time horizon is infinite, we run the mean-field algorithm for a suitably selected truncated horizon. This proof is a modification of the proof for Theorem 2.

Theorem 3. *For the infinite horizon RMAB planning problem with discount factor $\gamma < 1$, the mean-field policy $\overline{\pi}$ computed using a suitably truncated horizon has a reward that is at most $O(R_{max}\sqrt{K|S|N}/(1-\gamma)^2)$ less than the expected reward of an optimal policy. In particular,*

(1) *lower-bound for expected reward:* $\widehat{V}_{1:\infty}(\widehat{\mu}_1) \leq \mathbb{E}[\widehat{V}^{\overline{\pi}}_{1:\infty}(\widehat{\mu}_1)] + \frac{R_{max}((2-\gamma)K|S||A|+\gamma\sqrt{K|S|N}))}{2(1-\gamma)^2}$ *solved using a truncated horizon of* $T \geq \frac{2\sqrt{N}}{\sqrt{K|S|}} + 1$,

(2) *high-probability lower-bound for reward:* $\widehat{V}_{1:\infty}(\widehat{\mu}_1) \leq \widehat{V}^{\overline{\pi}}_{1:\infty}(\widehat{\mu}_1) + \frac{R_{max}((2-\gamma)K|S||A|+\gamma\sqrt{2\log(2)K|S|N+2N\log(N/\delta)}}{2(1-\gamma)^2}$ *with probability at least $(1 - \delta)$ for every $0 < \delta < 1$ solved using a truncated horizon of* $T \geq \frac{\sqrt{2N}}{\sqrt{\log(2)K|S|+\log(1/\delta)}} + 1$.

We complement the error upper bounds of the mean-field algorithm proved in Theorems 2 and 3 with a tight error lower bound with respect to N.

7 EXPERIMENTS

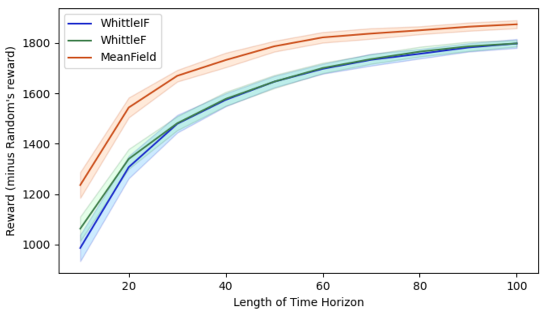
We present results of experiments in two simulation environments based on real-world use cases of RMABs: (i) maternity healthcare [28], and (ii) tuberculosis healthcare [21].[5] Both these experiments show the effectiveness of the MFP algorithm for real-world settings of practical importance for RMABs.

7.1 Mobile Healthcare for Maternal Health

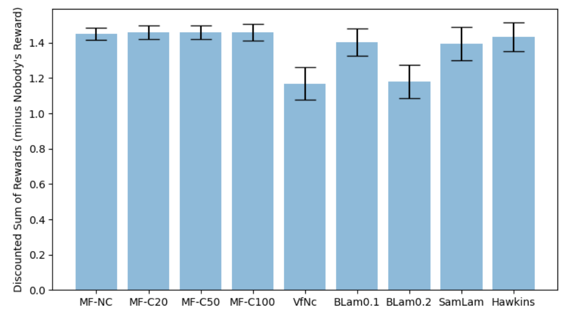
This experiment is based on the real-world data of the mMitra program of the healthcare NGO ARMMAN [2]. Under this program, healthcare workers at ARMMAN make phone calls to enrollees (beneficiaries) of the program to increase engagement and deliver targeted health information. The number of healthcare workers to make phone calls is substantially lower than the number of enrolled beneficiaries, and the healthcare workers have to continually prioritize which beneficiaries to call. [28] used RMABs to solve this resource allocation problem, modeling each beneficiary as an MDP shown in Figure 4. [28] divides the beneficiaries into 40 clusters, estimates the RMAB parameters (transition probabilities) for each cluster offline, and then uses these parameters to find a resource allocation scheme using the Whittle index policy.

The MFP algorithm is particularly suitable for the clustered data of ARMMAN. We communicated with the authors to run our MFP algorithm on this ARMMAN dataset and we present those results in Figure 3a. We compare the performance of the mean-field policy with the Whittle index policy. We use two versions of the Whittle index policy: (i) the usual infinite-horizon policy; (ii) finite-horizon modification of the Whittle index policy, where the infinite horizon q-values are replaced with finite horizon q-values computed by back-tracking. The transition probabilities of 96158 arms are estimated from the real historical data collected by ARMMAN using the

[5] Code: https://github.com/google-research/socialgood/tree/mfp/meanfield.



(a) ARMMAN Experiments Reward Plot



(b) Tuberculosis Experiment Reward Plot (1000 arms)

Figure 3: Experimental Results

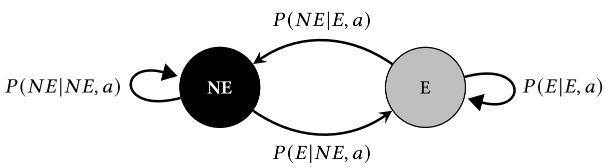


Figure 4: MDP model for ARMMAN mMitra program from [28]. NE implies that a beneficiary is not engaged with the program, whereas E implies engagement; a is the action.

clustering approach of [28]. We use a budget is 1000, i.e. 1000 calls per week from healthcare workers to beneficiaries. The discount factor γ is set to 0.95. The y-axis in Figure 3a is the improvement in the average reward of a policy as compared to the random policy, and the x-axis shows the length of the time horizon. These results are for 40 clusters and averaged over 40 runs.

Figure 3a shows that mean-field performs better than Whittle and has an additional average reward of about 200. MFP gives about 20% higher reward (i.e., number of engaged beneficiaries) per *active* action (i.e. phone call) than Whittle. We also observe that mean-field performs better than Whittle for other cluster sizes. Although the worst case time complexity of the mean-field algorithm depends upon the complexity of solving LPs (polynomial but not linear), we see that for our experiments the mean-field algorithm scaled linearly. This shows that MFP may scale well for real-life instances. See the full version of the paper for further details about the experimental environment and additional plots.

7.2 Tuberculosis Healthcare

We test the mean-field algorithm on a rigorous simulation environment (based on a real-world dataset) developed by [21], motivated by tuberculosis care in India. In this domain, a single healthcare worker manages up to 200 patients encouraging them to adhere to their 6-month TB treatment regimen. Each patient is modeled as an individual arm, with state being a tuple of (adherence level, treatment phase, day of treatment). The actions available to the healthcare worker are (i) do nothing (no cost or impact), (ii) call (moderate cost and impact), and (iii) visit (high cost and impact). We refer to [21] for a detailed description of the RMAB.

We simulate this setting for many parameter combinations. The original simulation framework by [21] did not have any clustering. We ran the mean-field algorithm as: **MF-NC**, without any clustering ($K = N$); **MF-C**K, with K clusters $K = 20, 50, 100$. We compare our algorithms with the ones in [21]: **Hawkins**, which is the multi-action extension of Whittle; **VfNc**, **SamLam** (called SampleLam in [21]), **BLam0.1**, and **BLam0.2**, which are approximations of **Hawkins** as described in [21]. Figure 3b plots the reward of these algorithms for 1000 arms and budget being 10% = 100, averaged over 25 runs (minus the reward of a policy that takes all 0 cost actions called **Nobody**). We see that MFP performs as well as others. In the full version of the paper, we provide results with smaller number of arms; again, MFP's performance is as good as the other algorithms.

Additional Experiments: We also simulated another environment proposed in [21], which is similar to the examples we show in Section 4. This simulation environment also has *greedy* and *reliable* type of agents, along with a third agent type named *easy*. In these experiments (see full version of the paper), MFP significantly outperforms the Whittle index policy.

8 CONCLUSIONS

This work challenges the techniques used in practically deployed solutions to important resource allocation problems and proposes and evaluates an alternative, provably effective method. Specifically, we demonstrate that it is important to read the fine print before applying Whittle index policies for RMABs in sensitive applications like healthcare by constructing simple, practically relevant and indexable instances of RMABs where Whittle index policies perform poorly. We then propose and analyze the mean field planning (MFP) algorithm, which is simple, efficient, and hyper-parameter free. This method provably obtains near-optimal policies with a large number of arms N without any structural assumptions of the Whittle index policies. We then demonstrate via extensive experiments that MFP obtains policies that perform better than the state-of-the-art Lagrangian relaxation based policies—policies that are currently used in practical applications in the healthcare domain.

ACKNOWLEDGMENTS

This work was done while AG was doing an internship at Google Research India. We thank Aditya Mate, Shresth Verma, and Jackson Killian for their help with the data and the code for the experiments.

REFERENCES

[1] Beatrice Acciaio, Julio Backhoff-Veraguas, and René Carmona. 2019. Extended mean field control problems: stochastic maximum principle and transport perspective. *SIAM journal on Control and Optimization* 57, 6 (2019), 3666–3693.

[2] ARMMAN. 2019. Assessing the Impact of Mobile-based Intervention on Health Literacy among Pregnant Women in Urban India. https://armman.org/wp-content/uploads/2019/09/Sion-Study-Abstract.pdf. Accessed: 2022-08-12.

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

[4] Dimitris Bertsimas and José Niño-Mora. 2000. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research* 48, 1 (2000), 80–90.

[5] Biswarup Bhattacharya. 2018. Restless bandits visiting villages: A preliminary study on distributing public health services. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. ACM, New York, NY, USA, 1–8.

[6] David B Brown and James E Smith. 2020. Index policies and performance bounds for dynamic selection problems. *Management Science* 66, 7 (2020), 3029–3050.

[7] Amarjit Budhiraja, Paul Dupuis, and Markus Fischer. 2012. Large deviation properties of weakly interacting processes via weak convergence methods. *The Annals of Probability* 40, 1 (2012), 74–102.

[8] Neil Cammardella, Ana Bušić, Yuting Ji, and Sean Meyn. 2019. Kullback-Leibler-Quadratic optimal control of flexible power demand. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, IEEE, France, 4195–4201.

[9] Neil Cammardella, Ana Bušić, and Sean Meyn. 2020. Kullback-Leibler-Quadratic Optimal Control.

[10] René Carmona, François Delarue, et al. 2018. *Probabilistic theory of mean field games with applications I-II*. Springer, New York, NY.

[11] Yue Chen, Md Umar Hashmi, Joel Mathias, Ana Bušić, and Sean Meyn. 2018. Distributed control design for balancing the grid using flexible loads. In *Energy Markets and Responsive Grids*. Springer, New York, NY, 383–411.

[12] Michael Chertkov, Vladimir Y Chernyak, and Deepjyoti Deka. 2018. Ensemble control of cycling energy loads: Markov decision approach. In *Energy Markets and Responsive Grids*. Springer, New York, NY, 363–382.

[13] Massimo Fornasier and Francesco Solombrino. 2014. Mean-field optimal control. *ESAIM: Control, Optimisation and Calculus of Variations* 20, 4 (2014), 1123–1152.

[14] Kevin D Glazebrook, David J Hodge, and Chris Kirkbride. 2011. General notions of indexability for queueing control and asset management. *The Annals of Applied Probability* 21, 3 (2011), 876–907.

[15] Olivier Guéant, Jean-Michel Lasry, and Pierre-Louis Lions. 2011. Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010*. Springer, New York, NY, 205–266.

[16] Vida Health. 2022. Vida. https://www.vida.com/.

[17] David J Hodge and Kevin D Glazebrook. 2015. On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Advances in Applied Probability* 47, 3 (2015), 652–667.

[18] Weici Hu and Peter Frazier. 2017. An Asymptotically Optimal Index Policy for Finite-Horizon Restless Bandits.

[19] Leo P Kadanoff. 2009. More is the same; phase transitions and mean field theories. *Journal of Statistical Physics* 137, 5 (2009), 777–797.

[20] Jackson A. Killian, Arpita Biswas, Sanket Shah, and Milind Tambe. 2021. Q-Learning Lagrange Policies for Multi-Action Restless Bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 871–881.

[21] Jackson A. Killian, Andrew Perrault, and Milind Tambe. 2021. Beyond "To Act or Not to Act": Fast Lagrangian Approaches to General Multi-Action Restless Bandits. In *AAMAS*. AAMAS, UK, 710–718.

[22] Daniel Lacker. 2017. Limit theory for controlled McKean–Vlasov dynamics. *SIAM Journal on Control and Optimization* 55, 3 (2017), 1641–1672.

[23] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press, Cambridge, Boston.

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

[25] Jr-Shin Li. 2010. Ensemble control of finite-dimensional time-varying linear systems. *IEEE Trans. Automat. Control* 56, 2 (2010), 345–357.

[26] Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. 2018. Efficient Large-Scale Fleet Management via Multi-Agent Deep Reinforcement Learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 1774–1783.

[27] 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* 33 (2020), 15639–15650.

[28] Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. 2022. Field Study in Deploying Restless Multi-Armed Bandits: Assisting Non-profits in Improving Maternal and Child Health. *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (06 2022), 12017–12025.

[29] Sylvie Méléard. 1996. *Asymptotic behaviour of some interacting particle systems; McKean-Vlasov and Boltzmann models*. Springer Berlin Heidelberg, Berlin, Heidelberg, 42–95.

[30] Navikkumar Modi, Philippe Mary, and Christophe Moy. 2019. Transfer restless multi-armed bandit policy for energy-efficient heterogeneous cellular network. *EURASIP Journal on Advances in Signal Processing* 2019, 1 (2019), 1–19.

[31] Christos H Papadimitriou and John N Tsitsiklis. 1999. The Complexity of Optimal Queuing Network Control. *Mathematics of Operations Research* 24, 2 (1999), 293–305.

[32] Christos H. Papadimitriou and John N. Tsitsiklis. 1999. The Complexity of Optimal Queuing Network Control. *Mathematics of Operations Research* 24, 2 (1999), 293–305.

[33] Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. 2016. Restless Poachers: Handling Exploration-Exploitation Tradeoffs in Security Domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 123–131.

[34] Diego Ruiz-Hernández, Jesús M Pinar-Pérez, and David Delgado-Gómez. 2020. Multi-machine preventive maintenance scheduling with imperfect interventions: A restless bandit approach. *Computers & Operations Research* 119 (2020), 104927.

[35] Csaba Szepesvári. 2010. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning* 4, 1 (2010), 1–103.

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

[37] Vishrant Tripathi and Eytan Modiano. 2019. A Whittle Index Approach to Minimizing Functions of Age of Information. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Allerton, Allerton, 1160–1167.

[38] Ina Maria Verloop. 2016. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *The Annals of Applied Probability* 26, 4 (2016), 1947–1995.

[39] Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. 2020. Breaking the Curse of Many Agents: Provable Mean Embedding q-Iteration for Mean-Field Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Baltimore, Article 936, 12 pages.

[40] Richard R Weber and Gideon Weiss. 1990. On an index policy for restless bandits. *Journal of applied probability* 27, 3 (1990), 637–648.

[41] Pierre Weiss. 1907. L'hypothèse du champ moléculaire et la propriété ferromagnétique. *J. Phys. Theor. Appl.* 6, 1 (1907), 661–690.

[42] Inc Welldoc. 2022. Welldoc. https://www.welldoc.com/.

[43] Peter Whittle. 1988. Restless bandits: Activity allocation in a changing world. *Journal of applied probability* 25, A (1988), 287–298.

[44] Jiachen Yang, Xiaojing Ye, Rakshit Trivedi, Huan Xu, and Hongyuan Zha. 2017. Learning Deep Mean Field Games for Modeling Large Population Behavior.

[45] Sarath Yasodharan and Rajesh Sundaresan. 2022. Large deviations of mean-field interacting particle systems in a fast varying environment. *The Annals of Applied Probability* 32, 3 (2022), 1666–1704.

[46] Zhe Yu, Yunjian Xu, and Lang Tong. 2018. Deadline scheduling as restless bandits. *IEEE Trans. Automat. Control* 63, 8 (2018), 2343–2358.

[47] Gabriel Zayas-Caban, Stefanus Jasin, and Guihua Wang. 2019. An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits. *Advances in Applied Probability* 51, 3 (2019), 745–772.

[48] Xiangyu Zhang and Peter I. Frazier. 2021. Restless Bandits with Many Arms: Beating the Central Limit Theorem.

[49] Xiangyu Zhang and Peter I. Frazier. 2022. Near-optimality for infinite-horizon restless bandits with many arms.

[50] Qing Zhao, Bhaskar Krishnamachari, and Keqin Liu. 2008. On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance. *IEEE Transactions on Wireless Communications* 7, 12 (2008), 5431–5440.