Beyond "To Act or Not to Act": Fast Lagrangian Approaches to General Multi-Action Restless Bandits

Jackson A. Killian
Harvard University
Cambridge, MA, USA
jkillian@g.harvard.edu

Andrew Perrault
Harvard University
Cambridge, MA, USA
aperrault@g.harvard.edu

Milind Tambe
Harvard University
Cambridge, MA, USA
tambe@g.harvard.edu

ABSTRACT
This paper presents new algorithms and theoretical results for solutions to Multi-action Multi-armed Restless Bandits, an important but insufficiently studied generalization of traditional Multi-armed Restless Bandits (MARBs). Though MARBs are popular for modeling many problems, they are restricted to binary actions, i.e., “to act or not to act”. This renders them unable to capture critical complexities faced by planners in real domains, such as a system manager balancing maintenance, repair, and job scheduling, or a health worker deciding among treatments for a given patient. Limited previous work on Multi-action MARBs has only been specialized to subproblems. Here we derive multiple algorithms for use on general Multi-action MARBs using Lagrangian relaxation techniques, leading to the following contributions: (i) We develop BLam, a bound optimization algorithm which leverages problem convexity to quickly and provably converge to the well-performing Lagrange policy; (ii) We develop SampleLam, a fast sampling technique for estimating the Lagrange policy, and derive a concentration bound to investigate its convergence properties; (iii) We derive best and worst case computational complexities for our algorithms as well as our main competitor; (iv) We provide experimental results comparing our algorithms to baselines on simulated distributions, including one motivated by a real-world community health intervention task. Our approach achieves significant, up to ten-fold speedups over more general methods without sacrificing performance and is widely applicable across general Multi-action MARBs. Code is available at https://github.com/killian-34/MAMARB-Lagrange-Policies.

KEYWORDS
Multi-action Restless Bandits; Lagrangian Relaxation; Bound Optimization; Markov Decision Process

ACM Reference Format:

1 INTRODUCTION
MARBs, the state-based generalization of classic Multi-Armed Bandits [30], have been studied extensively for solving a diverse set of problems including machine replacement [10, 29], sensing and wireless network scheduling [2, 3, 7, 22, 37], job scheduling [13, 34, 35], anti-poaching patrol scheduling [28], and healthcare [5, 17, 19]. In classic Multi-Armed Bandits, a planner must select k out of N arms on which to act for each of L rounds in a way that maximizes reward produced by the arms. In MARBs, additional complexity is introduced in that the reward on each arm depends on the action as well as an internal state that evolves according to an independent two-action Markov Decision Process (MDP). It has been shown that this problem is, in general, PSPACE-hard to solve exactly [26], but highly effective heuristics are known to exist [4, 33].

However, a critical limitation of MARB frameworks is they only allow for 2 actions: act or not act. This is restrictive for many real-world cases where planners have various actions at their disposal with varying degrees of cost and effect. For example, a system manager may need to balance preventative maintenance, full repair, and job scheduling each with different costs and effects on throughput [6]. In anti-poaching, the planner could allocate different levels of patrol effort to different targets, where more effort has higher cost and higher deterrent effect on poachers [23]. In public health, a community health worker could have several options for intervening with a patient, such as calling, visiting in person, or escalating patients to a more intense treatment [25]. Traditional MARBs simply cannot model these complexities, restricting planners to a world where their only choices are to, e.g., call or not call. Rather, the planner needs to simultaneously optimize the use of all of the tools in their toolbelt each day, subject to a per-day time or cost budget B. This process is visualized in Fig. 1.

To model such problems, we consider an under-examined generalization of MARBs that allow for multiple action types per arm, which we call Multi-Action MARBs (MA2RBs). Previous work has considered extending the classical MARB notion of indexability and

Figure 1: Multi-action MARB: one planner, many stateful agents, daily budget, many actions with varied costs/effects.
corresponding index policies to (MA)$^2$RBs [9]. In both traditional and Multi-action MARBs, index policies are desirable because: (1) they decompose the problem in a manner that scales well and (2) when indexability holds, they are asymptotically optimal [14, 32]. However, both deriving index policies and verifying indexability is notoriously difficult, and largely requires special problem structure [9, 10]. Our goal is thus to develop fast, well-performing policies for a broader class of (MA)$^2$RBs where no structure is assumed and indexability cannot be readily verified. We bypass the task of deriving index policies by taking a more general Lagrangian relaxation approach that leads to an auxiliary problem of computing a policy that minimizes the Lagrange bound. Computing this “Lagrange policy” is desirable because it recovers the index policies when they exist, but is readily computable regardless of problem structure.\(^1\)

The Weakly Coupled MDP (WCMDP) literature offers a method to compute the Lagrange policy for WCMDPs. Here, we recognize WCMDPs as a generalization of (MA)$^2$RBs and identify that this approach can be used to compute the Lagrange policy for (MA)$^2$RBs. However, the approach relies on solving a large linear program (LP) that scales quadratically in the number of arms and states. Setting up and solving this LP quickly becomes infeasible for large problem sizes as we show later in experiments. To address this issue, we investigate and utilize basic structural properties of general (MA)$^2$RBs to create scalable algorithms for computing the Lagrange policy on any (MA)$^2$RB problem, leading to the following contributions:

(i) **Bound optimization algorithm:** We develop BLam, an iterative bound optimization method for computing the Lagrange policy. BLam leverages problem convexity to derive progressively tighter upper and lower bounds on the Lagrange policy via a series of small LPs. We provide key technical results that prove this method converges to the policy that minimizes the Lagrange bound and provide experimental evaluation of its runtime on various distributions.

(ii) **Sampling algorithm:** We develop a sampling-based algorithm, SampleLam, which trades off the guarantees of BLam for speed. SampleLam chooses a random subset of arms, rapidly computes a statistic about the desirability of allocating budget to each arm, then combines the statistics to construct an estimated Lagrange policy for the full problem. We derive a concentration bound to prove the method converges, then use insights from the bound to inform how the algorithm carries out sampling.

(iii) **Complexity Results:** We derive best and worst case computational complexities for our methods as well as our main competitor. Our exact algorithm, BLam, achieves $\approx \sqrt{N}$ improvement and SampleLam achieves a factor of $N$ improvement in the best case.

(iv) **Experimental evaluation:** We compare our algorithms to baselines on synthetic distributions with different underlying structure, including one motivated by a real-world public health challenge. Our algorithms scale up to ten times better than a more general baseline without sacrificing performance, and readily adapt to each problem with minimal tuning. Thus our work newly makes available multiple avenues for computing well-performing policies on new (MA)$^2$RBs at scale, without the need for the user to first arduously derive a problem-specific index policy.

---

\(^1\)Please see the online appendix for additional discussion https://teamcore.seas.harvard.edu/files/teamcore/files/multi_action_bandits_appendix.pdf

2 RELATED WORK

Previous work extends the traditional MARB notion of indexability to (MA)$^2$RBs [9, 14]. However, their analysis is restricted to a subclass of (MA)$^2$RBs with special monotonic structure, whereas we build algorithms for general (MA)$^2$RBs. “Superprocesses” are an alternative multi-action extension where a primary planner distributes a limited set of sub-planners who act on arms without constraint [18, 31, 36]. This structure does not generally apply to (MA)$^2$RBs since they do not constrain the number of agents that can be acted on each round. Very recent work [20] designs a Monte-Carlo rollout approach for estimating traditional and multi-action MARB policies when a restricted set of “threshold” policies are optimal, but our algorithms do not assume this structure.

Also related are WCMDPs in which a planner operates $N$ independent MDPs subject to a set of arbitrary constraints over actions. [21] derive methods for handling “global” resource constraints over all rounds, whereas we address round-by-round constraints. [12], the main baseline we compare against, derivé a Lagrangian relaxation on the general form of a WCMDP and give an LP for minimizing the Lagrange bound. In contrast, we leverage the single constraint nature of (MA)$^2$RBs to greatly speed up the computation of the Lagrange bound compared to [12]. [1] give an approximate dynamic programming method that achieves a tighter bound and better performing policies than the Lagrange approach to WCMDPs. However, it scales exponentially, restricting it to small problem sizes. [11] develop a Lagrange approach for solving WCMDPs with MDPs that grow exponentially in problem parameters, restricting them to approximation techniques. In contrast, we develop a method that exactly computes the Lagrange policy.

Finally, our work is related to a large body of work developing Lagrangian methods for solving traditional MARBs [4, 10, 24, 33]. We generalize these settings to allow for multiple actions. Moreover, the methods we develop here reduce in the binary action case to the widely-used, well-performing, Whittle index policy [10, 33].

3 PRELIMINARIES

A (MA)$^2$RB consists of a set of $N$ arms, each associated with a Markov Decision Process (MDP) [27]. An MDP $(S, A, r, T, \beta)$ consists of a set of states $S$, a set of actions $A$, a state-dependent bounded reward function $r: S \rightarrow \mathbb{R}$, a transition function $T$, where $T(s, a, s')$ gives the probability of transitioning to state $s'$ when action $a$ is taken from state $s$, and a discount factor $\beta \in [0, 1)$. An MDP policy $\pi: S \rightarrow A$ maps states to actions. The long-term discounted reward starting from state $s_0 = s$ is defined as

$$R^\pi_\beta(s) = E \left[ \sum_{t=0}^{\infty} \beta^t r(s_{t+1} \sim T(s_t, \pi(s_t), s'_t)) | \pi, s_0 = s \right]$$

(1)

Each arm $i$ in a (MA)$^2$RB is an MDP with an action set $A_i$ of size $M_i$ and corresponding action cost vector $C_i \in \mathbb{R}^{M_i}$. We assume all action sets and costs are the same for all arms (and henceforth drop the subscript $i$), but all techniques in this paper extend in a straightforward manner to general action sets and costs. Without loss of generality, we assume that the elements $c_j$ of $C$ are ordered ascending. Also, to align with the standard bandit assumption that an arm can be “not played” at no cost, we set $c_0 = 0$. Each round, the planner must select one action for each of the $N$ arms such
that the sum cost of all actions do not exceed a budget $B$. Formally, the planner must choose a decision matrix $X \in \{0,1\}^{N \times M}$ with elements denoted $x_{i,j}$ such that

$$
\sum_{j=0}^{M-1} x_{i,j} = 1 \ \forall i \in 0...N - 1 \ \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} x_{i,j}c_j \leq B \quad (2)
$$

where the first constraint enforces one action per arm and the second enforces the budget. The planner’s goal is to maximize their discounted reward across the arms over time, subject to these constraints. Let $s = [s^0, s^1, ..., s^{N-1}]$ represent the vector of all arm states. The planner’s goal can be represented by the following constrained Bellman equation

$$
J(s) = \max_X \{ \sum_{i=0}^{N-1} r_i(s^i) + \beta E[J(s')|s,X] | \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} x_{i,j}c_j \leq B \} \quad (3)
$$

While Eq. 3 could be solved directly via value iteration, $J(s) \in \mathcal{S}^N$ and the number of feasible actions over which to take the max for each $J(s)$ is also exponential in $N$, making this approach intractable for non-trivial problem sizes. The key insight, though, is that the value functions and actions are only coupled due to the shared budget constraint over all arms. Therefore to simplify the problem, we relax the budget constraint and add it as a penalty to the objective with a Lagrange multiplier $\lambda$ as follows:

$$
J(s, \lambda) = \max_X \{ \sum_{i=0}^{N-1} r_i(s^i) + \lambda(B - \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} x_{i,j}c_j) + \beta E[J(s')|s,X] \} \quad (4)
$$

The value functions then decouple as desired, giving:

$$
J(s, \lambda) = \frac{\lambda B}{1-\beta} \max \{ \sum_{i=0}^{N-1} V_i(s^i, \lambda), \ \text{where} \ \forall i, \ V_i(s^i, \lambda) = \max_{a_j \in \mathcal{A}} \{ r_i(s^i) - \lambda c_j + \beta \sum_{s'^i} T(s^i, a^i_j, s'^i)V_i(s'^i, \lambda) \} \quad (5)
$$

The above derivation was first given by [12] for WCMDPs, and as suggested therein, clearly one can directly solve Eq. 7 using any LP solver. However, Eq. 7 has $N|M\times|A|$ and $N|M\times|A|$ variables and $N|M\times|A|$ constraints. Further, the current lowest known computational complexity for solving an LP is $O(n^2\times \pi)$, where $n$ is the number of variables [15], implying that directly solving Eq. 7 has computational complexity $\approx O(N^2|S|^3)$ (derived in section 4). The key to our approach will be separating the computation of the $\lambda$ that minimizes Eq. 7, henceforth $\lambda_{\min}$, and the corresponding $V^i$’s that solve Eq. 7 in a way that provides fast speedups. Herein we derive exact and heuristic methods for computing $\lambda_{\min}$, each of which has an improved best case complexity in $N$ by a factor of $\sqrt{N}$ or better.

### 4 BOUND OPTIMIZATION WITH BLAM

BLam is our exact approach to computing the Lagrange policy. We first give an overview, noting theorems where relevant that are derived in the next section. The main idea is rooted in the form of the functions $V^i(s^i, \lambda)$ in Eq. 7, visualized in blue in Fig. 2. To exactly compute Eq. 7 requires adding $|S||A|$ variables and $|S|$ variables to the LP for each of the $N$ arms’ value functions $V^i(s^i, \lambda)$. Instead, we will build special approximations to each $V^i(s^i, \lambda)$ that are represented in the LP each with just one variable and a constant number of constraints, achieving vast speedups. The approximations are constructed by rapidly testing for the slope of $V^i(s^i, \lambda)$ at various test points $\lambda_{\text{test}}$ using value iteration, then creating a piecewise linear combination of the slopes. The key is we construct two special types of approximations: one that upper bounds the slope of $V^i(s^i, \lambda)$ and one that lower bounds it, shown in Fig. 2 in green and red, respectively.

We then use the insight that the $V^i(s^i, \lambda)$ in Eq. 7, are indeed convex decreasing functions of $\lambda$ (Prop. 4.1), implying that Eq. 7 is minimized when the combined per-unit decrease to the objective brought by the convex $V^i(s^i, \lambda)$ functions is equal to or less than the constant per-unit increase to the objective brought by $\frac{\partial B}{1-\beta}$. In other words, $\lambda_{\min}$ is the point where the negative sum of slopes of $V^i(s^i, \lambda)$ is equal to $\frac{B}{1-\beta}$ (Prop. 4.2). Crucially, if we replace any $V^i(s^i, \lambda)$ with a convex function with strictly more negative slope (i.e., a lower bound), the value of $\lambda$ at which the negative sum of slopes equals $\frac{B}{1-\beta}$ could only increase, giving an upper bound on $\lambda_{\min}$. The converse also holds, i.e., replacing with upper bound convex functions gives a lower bound on $\lambda_{\min}$ (Thm. 4.3). This constitutes the core tradeoff in our approach: the more $V^i(s^i, \lambda)$ are replaced with approximations in the LP, the faster it will execute, but the looser the bounds will be. We handle this by first “bounding out”, i.e., replacing $V^i(s^i, \lambda)$ with its approximation, all but a small number $K$ processes to get loose bounds on $\lambda_{\min}$ rapidly. We then iteratively add back $V^i(s^i, \lambda)$ to the LP until the bounds on $\lambda_{\min}$ are with a pre-specified $\epsilon$. With minimal tuning, the test points can
be set to create tight enough bounds that BLam will converge after only a small number of iterations, leading to great speed increases.

The algorithm proceeds in two parts. In BLAMPRECOMPUTE, given in Alg. 1, we compute the upper and lower bound approximations of the arms, passing in the MDP parameters of the arms, along with a list of points $\lambda_{test}$ at which to approximate the slopes. VI in Alg. 1 denotes value iteration and $U_L$ will contain the pieces of the piecewise upper and lower bounds for $V^i(s', \lambda)$ for all arms and states. BLAMPRECOMPUTE runs once at the beginning of simulation.

BLAM, given in Alg. 2, runs on each round of the (MA)$^2$RB to compute $\lambda_{min}$ for the current set of arm states $s$ (line 2 of Alg. 2 selects the bounding functions for the current state of each arm). Using the piecewise bounded versions of $V^i(s', \lambda)$, it constructs a special LP, BLAMLP, given in Eq. 8 below, that produces upper and lower bounds on $\lambda_{min}$ by replacing $V^i(s', \lambda)$ with their bounded counterparts. It loops, replacing successively more $V^i(s', \lambda)$ in lieu of their bounded forms, until the resulting bounds on $\lambda_{min}$ are within $\epsilon$. BLAM terminates by running one final value iteration with the appropriate $\lambda_{min}$, the result of which solves Eq. 7 without constructing or solving the full LP, leading to vast speedups. The resulting value functions will be used to construct a final policy in section 6.

4.1 BLam: Derivation

To bound the slope of $V^i(s', \lambda)$, we rely on it having a convex form.

**Proposition 4.1.** $V^i(s', \lambda)$ is convex decreasing in $\lambda$, and as $\lambda \to \infty$, $dV^i(s', \lambda)/d\lambda \to 0$.

Proof. This follows directly from Eq. 6, but can be shown via induction that since $V^i(s', \lambda)$ is a max over piecewise linear convex functions of $\lambda$, it is also piecewise linear convex, and since $\lambda_{ij} \geq 0$, it must be weakly decreasing in $\lambda$. Furthermore, since $\epsilon_0 = 0$ in (MA)$^2$RBs, as $\lambda \to \infty$, the one time charge $\lambda_{ij}$ of any action $a_j$ s.t. $J > 0$ becomes greater than any long-term achievable reward, therefore the optimal policy must always choose to act. At that point, $V^i(s', \lambda) = E[\sum_{r=0}^{\infty} \beta^r r(s)|\pi(a) = 0, \forall a]$ which does not depend on $\lambda$.

Let $\lambda_0(\lambda)$ correspond to the $\lambda$ which solves Eq. 7 when $V^i(s', \lambda)$ are replaced in the objective by $L(U)$. Note that the lower bound functions $L$ will be used to derive upper bounds on the value of $\lambda_{min}$ and vice versa.

Next, we give a helpful intermediate result.

**Proposition 4.2.** The optimal solution to Eq. 7 will be found at the value of $\lambda$ in which the negative sums of the slopes of $V^i(s', \lambda)$ w.r.t. $\lambda$ become less than or equal to $\frac{B}{\tau - \beta}$.

Proof. Assume $\lambda^*$ corresponds to an optimal solution to Eq. 7 and the negative sums of the slopes of convex decreasing $V^i(s', \lambda)$ are greater than $\frac{B}{\tau - \beta}$. Then $\lambda^*$ can be increased by $\epsilon$ and the objective value would decrease, i.e., $J(s, \lambda^* + \epsilon) < J(s, \lambda^*)$ giving a contradiction. $\Box$

We now can prove our main result:

**Theorem 4.3.** $\lambda_t \leq \lambda_{min} \leq \lambda_u$

Proof. The proof is best seen by considering $\lambda_{min}$ which solves $J(s, \lambda)$, i.e., Eq. 7. We start with $\lambda_{min} \leq \lambda_u$: Let $\mathcal{V}$ denote the set of $V^i(s', \lambda)$ in the objective of Eq. 7. Further, let $\mathcal{V}^b$ denote the set of $V^i(s', \lambda)$ which will be replaced by $L \subset L$. Now replace all $\mathcal{V}^b$ with their corresponding $L^b$, name this new LP $J(\lambda_u(s, \lambda)$ and name its optimal solution $\lambda_u$. By definition, at all values of $\lambda$, the slope of $V^i(s', \lambda)$ is convex decreasing in $\lambda$. Hence, $\lambda_u(s, \lambda)$ is the optimal solution to the new LP $J(\lambda_u(s, \lambda))$. Since $\lambda_{min}$ is the optimal solution to $J(s, \lambda)$, we have $\lambda_{min} \leq \lambda_u(s, \lambda)$. Let any $s$ that maximize $\lambda_{min}$. Then $J(s, \lambda_{min}) = \lambda_{min}(s, \lambda_{min})$. Therefore $\lambda_{min}(s, \lambda_{min})$ is a lower bound on $\lambda_u(s, \lambda)$, and $\lambda_{min} \leq \lambda_u(s, \lambda)$.

Figure 2: Constructing bounds on the slope of $V^i(s', \lambda)$ for two different arms with three test points. Note: bounds are with respect to the slope, not the value of the function.
of all the trailing segments of \( L^i \) is less than \( B/(1 - \beta) \). We then set \( K = \max(K, \sqrt{N}) \) (line 4 Alg. 2).

Once \( \lambda_M \) and \( \lambda_t \) are computed once, we iterate to include \( K_{\text{step}} \) more arms in the LP such that \( K = K_{\text{step}} \) then repeat until the algorithm converges to within a difference \( \epsilon \). A straightforward induction argument shows that as \( K \) grows (and the set of bounded arms shrinks), the bounds become progressively tighter and are guaranteed to be exact when \( K = N \). Once \( \lambda_{\text{min}} \) is determined, we use value iteration to rapidly solve Eq. 7, the result of which we will use to derive feasible policies in Section 6.

### 4.2 BLam: Computational Complexity

In BLamPrecompute, BLam computes \( U^i(G^i_k, *) \) and \( L^i(G^i_k, *) \) for all \( s^i, \lambda \), which requires two runs of value iteration for each arm for each test point \( G^i_k \). Assuming all arms use the same number of test points, states and actions, this scales as \( O(NG^iV(|\mathcal{S}|, |\mathcal{A}|)) \) where \( V(\cdot) \) is the computational complexity of value iteration. While an exact computation of value iteration is elusive, it is known to be much faster than the LP formulation [27]. Thus, its complexity will be dominated by the LP solves that occur in BLam— the same applies for the value iteration that runs at the end of BLam each round.

To compute a policy for each round, BLam constructs Eq. 8 as an LP which has \( K|\mathcal{S}| + (N - K) \) variables, \( K|\mathcal{S}| + |\mathcal{A}| \) constraints with \( |\mathcal{S}| \) terms, and \( (N - K)G^i \) constraints with two terms. Although the constraints associated with \( (N - K) \) auxiliary variables only have two non-zero coefficients, we conservatively assume that the matrix for this LP is dense in order to adopt the best known LP complexity result [15]. In the best case, BLam would provide tight bounds on \( \lambda_{\text{min}} \) after just one iteration. So setting \( K = \sqrt{N} \) and assuming \( G^i \ll N \), the per-round complexity is

\[
\Omega(\sqrt{N}|\mathcal{S}|^2|\mathcal{A}| + N|\mathcal{S}|^2 + N^{3/2}|\mathcal{S}| + N^2)
\]

(9)

Where the first term is the LP setup time to add constraints (which dominates the time to add variables) and the last three terms are the LP solve complexity, which is approximately square in the number of variables. Applying the same reasoning to the direct LP solve approach, which has \( N|\mathcal{S}| \) variables and \( N|\mathcal{S}| |\mathcal{A}| \) constraints gives the following best (and worst) case complexity

\[
O(N|\mathcal{S}|^2|\mathcal{A}| + N^2|\mathcal{S}|^2)
\]

(10)

Thus, BLam has a strictly better best-case complexity in the problem size. However, in the worst case, setting \( K_{\text{step}} = \sqrt{N} \), BLam would require the full \( \sqrt{N} \) iterations to get a tight bound on \( \lambda_{\text{min}} \). In this case, the LP setup time would match the naïve LP approach, but successive solves would become more expensive. Using basic summation, this gives a worst-case complexity of

\[
O(N|\mathcal{S}|^2|\mathcal{A}| + N^2|\mathcal{S}|^2)
\]

(11)

Which, handily, is only \( \sqrt{N} \) worse than the naïve approach. However, we will show in experiments that the typical run time and scaling of BLam is much faster than the naïve approach in practice.

### 5 SAMPLELAM

In some cases, especially very large problem sizes, speed can be more critical than performance. Thus, next, we give an algorithm for
Algorithm 3: SampleLam

Data: T, R, C, N, B, β, \( r_{\text{max}}, \epsilon_{\text{min}} \)

1. \( N_{\text{samples}} = \frac{\log(N) r_{\text{max}}}{\epsilon_{\text{min}}} \)
2. \( \text{inds} = \text{RandomChoice}([1, ..., N], N_{\text{samples}}) \)
3. \( T, R = T[\text{inds}], R[\text{inds}] \)
4. \( \lambda_i^{(1)} = [1] \)
5. for \( i = 1, ..., N_{\text{samples}} \) do
6. \( \lambda_i = \text{QuickLP}(T[i], R[i], \frac{B}{N}, C, \beta) \)
7. \( \lambda_i^{(1)} \text{ list append}(\lambda_i) \)
8. \( V(i, s) = [1] // N \times |S| \) array to hold value functions
9. \( \lambda_{\text{min}} = \text{Mean}(\lambda_i^{(1)}) \)
10. for \( i = 1, ..., N \) do
11. \( R_\lambda = R[i] \)
12. for \( x \in 1, ..., |C| \) do subtract action costs
13. \( R_A[x] = -\lambda_{\text{min}} * C[x] \)
14. \( V[i] = V(i, T[i], R_\lambda, \beta) \)
15. return \( V \)

Figure 3: (a) \( \lambda_i \) is normally distributed about a mean equal to \( \lambda_{\text{min}} \). (b) \( \lambda_{\text{min}} \) is determined by a select few "important" arms, not equal to the sample mean. BLam is better suited for this case.

5.1 SampleLam: Concentration Bound and Complexity

To understand SampleLam’s convergence properties, i.e., convergence to the sample mean of \( \lambda_i \), we derive a concentration bound below. The derivation first relies on showing that the distribution of \( \lambda_i \)’s is sub-Gaussian.

Theorem 5.1. \( \lambda_i \) are \( \frac{r_{\text{max}}}{n} \)-sub-Gaussian with \( \sigma^2 = \frac{1}{4} \left( \frac{r_{\text{max}}}{\epsilon_{\text{min}}(1-\beta)} \right)^2 \)

The proof involves showing \( 0 \leq \lambda_i \leq \frac{r_{\text{max}}}{\epsilon_{\text{min}}(1-\beta)} \) and is given in Appendix 2. We can now use sub-Gaussianity to derive a concentration bound relating the number of samples to a confidence parameter \( 1 - \delta \) on the estimate of the sample mean of \( \lambda_i \).

Theorem 5.2. The number of samples \( n \) needed to estimate the sample mean of \( \lambda_i \) within an error \( \epsilon \) and with confidence \( 1 - \delta \) is lower bounded as:

\[
    n \geq \frac{1}{2 \epsilon^2} \left( \frac{r_{\text{max}}}{\epsilon_{\text{min}}(1-\beta)} \right)^2 \log \left( \frac{1}{\delta} \right)
\]

The proof, given in Appendix 2, uses the Hoeffding bound and Thm. 5.1. This bound gives the insight that the greater the reward to cost ratio, the more samples we need to well-estimate the mean. We account for this by including a factor of \( \frac{r_{\text{max}}}{\epsilon_{\text{min}}(1-\beta)} \) in the setting for \( K \) in the SampleLam algorithm. The drawback of this approach is that \( \lambda_{\text{min}} \) is not guaranteed to be close to the sample mean of \( \lambda_i \) in general. An adversarial case is shown in Fig. 3b in which SampleLam would compute an arbitrarily bad estimate for \( \lambda_{\text{min}} \). Setting \( K = \log(N) \frac{r_{\text{max}}}{\epsilon_{\text{min}}} \), the best and worst case complexity for SampleLam is:

\[
    O \left( \log(N) \frac{r_{\text{max}} |S| |A|}{\epsilon_{\text{min}}} + N \times V I(|S|, |A|) \right)
\]

Where the first term is the cost of setting up \( \log(N) \frac{r_{\text{max}}}{\epsilon_{\text{min}}} \) LPs, which dominates the solve time, and the second term is the cost of the final value iteration.

6 COMPUTING A POLICY

Finally, once \( \lambda_{\text{min}} \) is finalized, and the resulting value functions from Eq. 7 have been computed, we use the value functions to compute the one-step greedy policy implied by the bound. To do this, we expand the value functions to compute the one-step greedy policy, which solves Eq. 7 with a single LP, then creates an estimate of \( \lambda_{\text{min}} \) by taking the sample mean of \( \lambda_i \). Finally, it uses this estimate plus value iteration to solve for Eq. 7, which again we will use to derive feasible policies in section 6.

Although this method is not guaranteed to converge to the value of \( \lambda_{\text{min}} \), it is very fast, and works well in practice on distributions which have an approximately normal distribution of budget across arms under the true \( \lambda_{\text{min}} \) policy. An example of such a distribution and the SampleLam estimate of \( \lambda_{\text{min}} \) is given in Fig. 3a.
using an LP solver, then follows Section 6. We also include several versions of BLam using various stopping criterion \( \epsilon \), noted in each plot as BLam[\( \epsilon \)]. Larger \( \epsilon \) will lead to faster running times but looser bounds on \( \lambda_{\min} \), and thus worse performance in general. All algorithms were implemented in Python 3.6 and use Gurobi version 9.0.3 to solve LPs via the gurobiopy interface. All value iterations were computed using a lightly modified version of pymdptoolbox version 4.0b3 [8].

First, we explore an example distribution where VfNc and Hawkins fail arbitrarily in terms of performance and runtime respectively. In this distribution, there are three types of agents: (1) **Greedy:** Must take increasingly expensive actions to collect increasingly high reward. Once the required action is not taken, the agent never produces reward again. This is modeled with a single chain of states, each with unit-increasing reward, reachable only by an action with unit-increasing cost. Failure to take the next action leads to a dead state. (2) **Reliable:** Must take the cheapest non-zero action every round to achieve reward 1. If the arm is not played for any round, it never produces reward again. This is modeled with a simple 2-state chain, in which the final state recurs with the proper action, otherwise it goes to a dead state. (3) **Easy:** Always gives reward of 1 regardless of action. We make the proportion of (1) and (2) equal and set the budget so that all of (1) or (2) could be played (or some mix), but not more. Clearly, the optimal policy is to always play the Reliable agents since committing to the Greedy agents will eventually leave the planner only collecting reward from the Easy agents. However, the Greedy agents will look most attractive to VfNc since, without accounting for cost constraints, it will wrongly assume it can always pay the future cost to obtain increasingly larger reward. Hawkins and BLam, using their constraint-based reasoning, will instead commit to the Reliable agents. However, BLam will automatically detect the significant structure within the problem, such as the existence of Easy agents as well as a simple form for the \( V^i \) of Reliable agents, to build tight bounds on the Lagrange policy using only a small subset of agents in the coupled LP, leading to a significant speedup over Hawkins. The performance and runtime of the algorithms tested on a population with 0.25, 0.25, 0.5 mix across Greedy, Reliable, and Easy agents with a budget of 0.25\( N \) and 30 actions (subsequently, 31 states) are shown in Fig. 4 and 5 confirming these insights. For BLam, all arms used test points \( G^i = \{0\} \). Here, SampleLam gives good but variable performance in exchange for running approximately twice as fast as BLam.

Finally, we test our algorithms on a more rigorous simulation motivated by a real-world public health care challenge, namely, tuberculosis care in India. In this real-world setting, a single community health worker manages up to 200 patients throughout the course of their 6-month antibiotic regimen, monitoring and encouraging patients to take their daily medications. The health worker has a range of actions they can take on each patient aimed at improving their adherence, each with varying cost and effectiveness: call the patient (cheap), visit the patient in their home (semi-expensive), escalate the patient (very expensive). Because the worker’s time and resources are limited, the number and types of actions they can take each day across all patients are also limited.

We model this problem as follows. In the simulation, each patient state is a tuple of (adherence level, treatment phase, day of treatment). The first entry captures the patient’s previous \( d \) days of adherence. The second entry is binary and captures the “phase” of treatment: the intensive phase which lasts for the first \( IPL \) rounds, and the continuation phase which lasts from round \( IPL \) to the end. During the intensive phase, patients tend to have better adherence and are more responsive to intervention. During the continuation phase, both effects tend to degrade and patients may drop out (i.e., adherence of 0). The final entry captures time and can take any of \( IPL + 2 \) values. The first \( IPL \) values count days in the intensive phase and the next two are recurrent states that represent the continuation phase and the dropout state.

In one relevant dataset that captured daily treatment adherence of TB patients in India over the course of a year [16], patients followed four distinct modes: (1) **High adherence:** Adhere daily regardless of health worker action. This makes up the majority of the data; (2) **Low adherence:** Very low adherence regardless of health worker action. (3) **Receptive patients:** Irregular adherence but can benefit from intervention. On average, their adherence drops during the continuation phase, suggesting that interventions become less effective. (4) **Dropout patients:** Like receptive patients but have probability of dropping out during the continuation phase.

We implement each of these patient types in our simulation and include them in the following mix respectively: 0.64, 0.01, 0.175, 0.175. This mix matches the number of High and Low adherence patients observed in the data, and splits the remaining portions evenly. At the start of simulation, each patient is in the maximum adherence state since in the real world, patients begin treatment in person. We run experiments with \( d = 3, 4, 5 \) adherence levels and

---

**Figure 4:** Ignoring future constraints leads to bad policies.

**Figure 5:** BLam and SampleLam scale better than Hawkins.
set \( I_P^L = 2d \). The health worker’s action types are as follows: (1) No action: Take no action (c=0); (2) Call: Moderately increased probability that patient will increase adherence state by 1 (c=1). (3) Visit: Significant increased probability that patient will return to the maximum adherence state. If a patient is in the dropout state, there is a small probability they return to the continuation phase \( c = B \). Finally, rewards are defined as \((\text{adherence level})/d\), so more rewards are received for patients at higher adherence levels.

We simulate this setting for many parameter combinations. For BLam we report results using test points \( G^l = \{0,0.1,0.2,0.5\} \), though we found that, in general, most sets of 3 or 4 evenly spaced points worked well. Fig. 6 shows the performance and runtime for the dataset with budget of \( 0.1N \) for \( d = 3, 4, 5 \) adherence levels. With such a small budget, the tradeoff between individual actions is important. In Fig. 6 we see that all versions of BLam significantly outperform VfNc. Crucially, all versions of BLam also scale much better than Hawkins. In fact, as the number of states in the underlying problem grows the speed ups become even more dramatic ranging from a 2 times speedup with \( d = 3 \) to a 5 times speedup with \( d = 5 \). This is because the Hawkins LP scales quadratically in the number of states of each arm, while the BLam algorithms are able to identify problem structure that keep the underlying LPs small with its bounding techniques, making speedups more dramatic as the problem size increases.

In Appendix 4, we run experiments varying the budget between \( 0.1N, 0.2N, 0.5N \), with \( d = 4 \) adherence levels. As the budget increases, resources are less constrained, so all methods tend to collapse to the same reward. However, again, BLam’s adaptivity allows it to recognize when the problem is less constrained to automatically converge even more quickly to the optimal solution.

These results demonstrate the exemplary ability for our approach to scale well without sacrificing performance on a dataset whose technical structural conditions have no been established a priori. That is, our algorithm can perform exceptionally with minimal tuning, while avoiding undertaking the considerable effort of deriving an index policy and the existence thereof.

8 CONCLUSION

Our work makes available multiple avenues for computing well-performing policies on new \((MA)^2\)RBs at scale. We demonstrate that our algorithms offer vast speedups and can be readily adapted to new problems without the need for the user to first arduously derive a problem-specific index policy, as was previously the case. These advances make multi-action MARBs newly accessible, laying the groundwork for wider study of this important framework.

ACKNOWLEDGMENTS

This work was supported in part by the Army Research Office by Multidisciplinary University Research Initiative (MURI) grant number W911NF1810208. J.A.K. was supported by an NSF Graduate Research Fellowship under Grant DGE1745303. A.P. was supported by the Harvard Center for Research on Computation and Society.

REFERENCES


