Flexible Budget Restless Multi-Armed Bandits: Improving Resource Allocation in Public Health Settings

Anonymous Author(s)
Submission Id: ???

ABSTRACT

Restless multi-armed bandits (RMABs) are widely used to optimize the allocation of limited resources in sequential decision-making settings, particularly in public health systems. However, the assumption of a fixed budget for each step in the planning horizon in classic RMABs may not be appropriate for realistic real-world planning when resources are not necessarily limited at each step or when certain critical steps would require a larger use of resources than other steps. To address this issue, this paper proposes the use of Restless multi-armed bandits with *flexible budgets* (F-RMABs) that allows surplus resources in one round to be distributed to an earlier or later round, leading to more effective and efficient resource allocation. Additionally, the paper highlights the importance of considering critical times, such as peaks in disease contagion, in public health settings where interventions may become more critical and necessary, and delaying or early use of resources can be beneficial. Overall, this paper emphasizes the significance of incorporating criticality in the deployment of RMABs via F-RMABs for optimal resource allocation in public health systems.

KEYWORDS

Multi-armed bandits, resource allocation, public health

ACM Reference Format:
Anonymous Author(s). 2023. Flexible Budget Restless Multi-Armed Bandits: Improving Resource Allocation in Public Health Settings. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 4 pages.

1 INTRODUCTION

Restless multi-armed bandits (RMABs) have gained increasing attention as a model for efficient resource allocation in public health settings, with examples including treatment adherence for tuberculosis [4], maternal health and child care [1], and general adherence dynamics prevalent in many public health intervention problems [3]. However, the classic RMAB approach assumes fixed resource constraints at each step of the planning horizon, limiting its effectiveness in real-world settings where resources are not strictly constrained at each round but over multiple time steps, and certain interventions may need prioritization over others.

To address this issue, we propose the use of flexible budget restless multi-armed bandits (F-RMABs) [6] in public health settings. F-RMABs allow for the total resources to be used to be budgeted over a flexible time window, enabling public health practitioners to adjust their policies based on changing resource availability and prioritize critical interventions. This approach can lead to more efficient resource allocation and improved health outcomes for the population being served.

We further highlight two specific situations where the use of flexible resources can lead to better resource allocation policies in the public health setting: (i) responding to highly transmissible diseases with a wave behavior and (ii) designing health mandates that, while responding to disease outbreaks, reduce social tensions among individuals. By taking a sequential decision-making approach and adapting to changing circumstances, F-RMABs can help optimize resource allocation strategies and reduce the burden of health mandates.

In this paper, we present F-RMABs as a better alternative for sequential planning in public health settings. Unlike classic RMABs, which have a fixed per-round budget constraint, F-RMABs allow for a flexible budget over a time window of length F within the horizon H, where the total cost of all actions over that flexible window must be less than or equal to FB to maintain the per-round budget constraint on average. With the flexibility of F-RMABs, public health practitioners can make better resource allocation decisions and ultimately improve health outcomes for their patients.

In this paper, we propose F-RMABs as a superior approach for sequential resource allocation in public health settings. We begin by introducing the F-RMAB model and then present an algorithm to compute effective F-RMAB policies. To demonstrate the advantages of F-RMABs, we conduct experiments on synthetic domains that are motivated by real-world public health scenarios. Our results indicate that policies with budget flexibility achieve a performance improvement of up to 24% and 11% compared to fixed budget policies on our synthetic domains. Overall, our findings suggest that the use of F-RMABs can significantly enhance resource allocation strategies in public health settings, leading to better health outcomes for the populations being served.

2 FLEXIBLE BUDGET RMABS

Here, we define *restless multi-armed bandits with flexible budget* (F-RMABs) and provide algorithms to solve for reward maximizing policies in this setting. In section 2.1 we give a background on classic RMABs and in section 2.2 we define F-RMABs as a general class of RMABs with flexible per-round budget.

2.1 Background: Restless Multi-Armed Bandits

An RMAB instance consists of N independent Markov decision processes (MDPs), each corresponding to an arm of the instance [5]. Each MDP is defined by the tuple $\{\mathcal{S}, \mathcal{A}, R, \mathcal{P}\}$. $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the set of possible actions, R is the reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, and $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ represents the transition function. We use $P^a_{s,s'}$ to denote the probability of transitioning from state s to state s' under the action a.

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.

We let $s^t = [s_1^t, s_2^t \dots s_N^t]$ denote the vector of states of the N MDPs at time step t. A policy is a mapping $\pi^t : \mathcal{S}^N \to \mathcal{A}^N$ that informs the action to take at a given state, at time step t. We consider the more general multi-action case in which $|\mathcal{A}| \geq 2$ and define an action-cost matrix $\boldsymbol{c}$ of size $N \times |\mathcal{A}|$, i.e., c_{nj} is the cost of taking action $j \in \mathcal{A}$ on arm n. Let $\mathbb{1}_{\pi^t(s^t)}$ be the one-hot encoder of size $N \times |\mathcal{A}|$, where each row n indicates which action to perform on arm n at time step t. The planner's goal is to find reward maximizing policies $\{\pi^t\}_{t=1}^H$ under the budget constraint $\mathbb{1}_{\pi^t(s^t)} \cdot \boldsymbol{c} \leq B$ for each $t \in [H]$. Here H is the horizon length and $\cdot$ is the Frobenius inner product.

The total reward accrued can be measured using discounted, average, or total reward criteria in the finite- or infinite-horizon settings; we consider the total reward criterion in the finite-horizon setting, which enables the clearest analysis of our method. The expected *total reward* from initial state s^0 is defined as $V_\pi^1(s^0) = \mathbb{E}\left[\sum_{t=1}^H \sum_{n=1}^N R(s_n^{t-1}, [\pi^t(s^{t-1})]_n, s_n^t)\right]$ where the next state is drawn according to $s_n^t \sim P^{[\pi^t(s^{t-1})]_n}_{s_n^{t-1}, s_n^t}$. The planner's goal is to find policies $\pi = \{\pi^t\}_{t=1}^H$ that maximize the total reward.

2.2 Definition

In F-RMABs, we define the MDP followed by each arm using the tuple $\{\mathcal{S}, \mathcal{A}, R, \mathcal{P}\}$ just as in the classic RMAB setting. We now consider a flexible-budget time window of length F where $F \leq H$. Our goal is to find optimal policies $\{\pi^t\}_{t=1}^H$ such that $\sum_{t=1}^F (\mathbb{1}_{\pi^t(s^t)} \cdot \boldsymbol{c}) \leq FB$ and $\mathbb{1}_{\pi^t(s^t)} \cdot \boldsymbol{c} \leq B$ for $t = F+1, \dots, H$. That is, we consider an exhaustible budget FB that is available to spend over the flexible window $1, \dots, F$ and think of B as the one-step budget at every time step t after the flexible window $t = F+1, \dots, H$.

3 SOLVING F-RMAB POLICIES

3.1 F-RMAB as an optimization problem

Existing RMAB solution approaches require a fixed budget per round, leading to suboptimal performance. To make use of flexibility, we extend Lagrangian relation to the flexible setting and solve the resulting min-max problem with gradient algorithms. We present this formulation as it is presented and derived in (Rodriguez Diaz et al., 2023).

Recall, the budget constraint for F-RMABs over the flexible window is given by:

$$\sum_{t=1}^F \mathbb{1}_{\pi^t(s^t)} \cdot \boldsymbol{c} \leq FB, \tag{1}$$

where $\mathbb{1}_{\pi^t(s^t)}$ is the one-hot encoded matrix of size $N \times |\mathcal{A}|$, where each row n indicates the action recommended by the policy π^t on arm n, at time step t. Since this budget constraint is over multiple timesteps, formulating the optimal Bellman equation requires expanding the state space of the F-RMAB to capture the budget remaining after a given action is taken. However, this expansion adds an additional layer of combinatorial complexity over that involved in formulating the optimal Bellman equation for classic RMABs. Moreover, it is unclear how to relax this single budget constraint, which covers multiple timesteps, in a convenient or informative manner.

We make the insight that Eq. 1 can be reformulated to the following equivalent constraint structure, which introduces per-round budget variables:

$$\mathbb{1}_{\pi^t(s^t)} \cdot \boldsymbol{c} \leq b_t \qquad \forall t \in \{1, \dots, F\} \tag{2}$$

$$\sum_{t=1}^F b_t \leq FB\,. \tag{3}$$

We will show this reformulated set of constraints is much more convenient to solve. For this constraint structure, each b_t for $t \in \{1, \dots, F\}$ is a variable that we must solve for in the original maximization problem. The key idea is that having a constraint in each round of the problem will allow us to follow a per-round Lagrangian relaxation, enabling us to convert the problem into a more tractable form.

Thus for the finite-horizon problem with total time horizon of length H and flexible time window of length F, the F-RMAB problem can be formulated as the following optimization problem:

$$\max_{\substack{\pi^1,\dots,\pi^H \\ b_1,\dots,b_F}} \mathbb{E}\left[\sum_{t=1}^H \sum_{n=1}^N R(s_n^{t-1}, [\pi^t(s^{t-1})]_n, s_n^t)\right] \tag{4}$$

$$\text{s.t.} \quad \mathbb{1}_{\pi^t(s^t)} \cdot \boldsymbol{c} \leq b_t, \quad \forall t \in \{1, \dots, F\} \tag{5}$$

$$\mathbb{1}_{\pi^t(s^t)} \cdot \boldsymbol{c} \leq B, \quad \forall t \in \{F+1, \dots, H\} \tag{6}$$

$$\sum_{t=1}^F b_t \leq FB \tag{7}$$

Since the optimal policies for all arms are still coupled by budget constraints, this problem is still at least as hard as standard RMABs. However, we carry out a Lagrangian relaxation that gives a new problem that upper bounds Eq. 4, but is in a far more tractable form, as we show in Theorem 3.1. The key benefit of Theorem 3.1 is that, if G is convex, there are efficient algorithms for solving optimization problems with this structure. See [6] for detailed proof.

Theorem 3.1. *The Lagrangian relaxation of Eq. 4 gives a new first-order primal-dual optimization problem which upper bounds Eq. 4 and has structure:*

$$\min_{x \in X} \max_{y \in Y} \langle Kx, y \rangle + G(x) - H^*(y), \tag{8}$$

where X and Y are finite-dimensional vector spaces equipped with inner product $\langle \cdot, \cdot \rangle$. $K : X \to Y$ is a linear operator and $G : X \to \mathbb{R} \cup \{\infty\}$ and $H^* : Y \to \mathbb{R} \cup \{\infty\}$ are convex functions.

Now that we have shown the underlying structure of our problem, in the next section we describe an approach for solving the Lagrangian sub-problem optimally, and using that to derive good policies for the F-RMAB.

3.2 Solving with a gradient algorithm

We now present an algorithm for solving F-RMABs. The key idea is that, for a given state in a given round, the solution will contain information about how budget within a flexible window would be best allocated, and what actions are best to take. We use that information to actually take actions each round in the environment.

The optimization problem from Theorem 3.1 is solved by building from the proximal optimization method of Chambolle and Pock [2], which is desirable for its convergence properties on concave-convex

min-max optimization problems. The key challenge in implementing their approach is in efficiently computing the proximal steps.

Note first that the proximal operator (or proximal mapping) of a convex function F is

$$\mathbf{prox}_{\sigma F}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \left(F(\boldsymbol{u}) + \frac{1}{2\sigma} ||\boldsymbol{u} - \boldsymbol{x}||_2^2 \right).$$

Following the notation in Chambolle and Pock [2], $\mathbf{prox}_{\sigma F}(\boldsymbol{x}) = (\boldsymbol{I} + \sigma \partial F)^{-1}$. Then, the proximal operator of $H^\star(\boldsymbol{b})$ is $\mathbf{prox}^{\sigma}_{H^\star}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \left(\frac{1}{2\sigma} ||\boldsymbol{u} - \boldsymbol{x}||_2^2 \right) = \boldsymbol{x}$. Hence, the proximal operator of the zero function $H^\star$ is the identity. The proximal operator of G does not have any analytical form. However, it is a piecewise-liear function. Since the proximal operator of linear functions is simply $\boldsymbol{x} - \sigma \nabla F(\boldsymbol{x})$, a good approximation of $\mathbf{prox}_{\sigma G}(\boldsymbol{x})$ is $x - \sigma \nabla G(x)$. Though an approximation, as we show next, computing ∇G is convenient, and performs well in practice.

Proposition 3.2. *The gradient of G at $(\boldsymbol{\lambda}, \mu)$ is given by $\nabla G((\boldsymbol{\lambda}, \mu)) = [D^1, D^2, \ldots, D^F, D^{F+1} + B, \ldots, D^H + B, FB]$ where $D^t = \mathbb{E}[\sum_{n \in [N]} -c_n^t]$ is the expected sum of costs over all arms in step t under the optimal policy for $\boldsymbol{\lambda}$.*

The main challenge then is in computing D^t which has no convenient closed form. However, as long as we can compute the optimal policy $\pi^*(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda}$, we can get unbiased samples of each D^t via Monte Carlo simulation of $\pi^*(\boldsymbol{\lambda})$.

Combining each of these steps, we have a complete algorithm for solving FRAMB policies to our desired level of convergence [2]. This approach is called *primal-dual stochastic gradient* (PDSG) and the pseudocode for its implementation is presented in Algorithm 1.

Algorithm 1 PDSG

Input: Flexible window F, horizon H, initial values $b^0 \in \mathbb{R}^F, \lambda^0 \in \mathbb{R}^H, \mu^0 \in \mathbb{R}$, gradient steps $\tau, \sigma > 0$, transition probability P, per-round budget B, and number of gradient samples N_s for each state s

$\bar{\nu}^0 = [\boldsymbol{\lambda}^0, \mu^0]$
while *not converged* **do**
  $b^{n+1} = b^n + \sigma K \bar{\nu}^n$
  $\hat{\nu}^{n+1} = \nu^n - \tau K' b^{n+1}$ $\{\hat{\nu}^{n+1} = [\hat{\boldsymbol{\lambda}}^{n+1}, \hat{\mu}^{n+1}]\}$
  π^{n+1} = FiniteHBellmanLP$(P, \hat{\boldsymbol{\lambda}}^{n+1}, H)$ {LP to compute value func given λ. In appendix.}
  $\nabla G(\hat{\nu}^{n+1})$ = SampleGrads$(N_s, \pi^{n+1}, P, H, F)$
  $\nu^{n+1} = \hat{\nu}^{n+1} + \tau \nabla G(\hat{\nu}^{n+1})$
  $\bar{\nu}^{n+1} = 2\nu^{n+1} - \nu^n$
end while

4 EXPERIMENTAL EVALUATION

We evaluate the algorithm presented in section 3.2 on two synthetic domains motivated by public health situations. Our results show these domains all benefit from per-round budget flexibility.

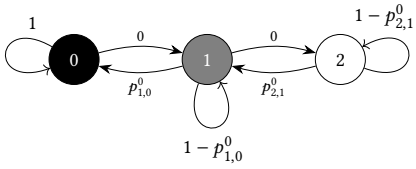
4.1 Synthetic domains

Dropout state. This first domain characterizes settings with potential urgent interventions, such as clinical health settings in which patients are likely to never return after dropping out of a program

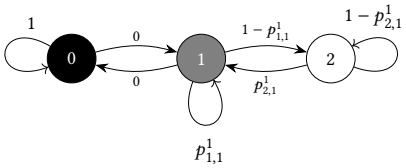
Algorithm 2 SampleGrads

Input: Number of gradient samples N_s, policy $\pi : \mathcal{S}^N \to \mathcal{A}^N$, transition function P, planning horizon H, flexible time window F

for $i \in \{1, ..., N_s\}$ **do**
  $\boldsymbol{x}_i = (x_{1,i}, \ldots, x_{H,i})$ = `MonteCarlo`(π, P) {Simulate π in environment P, return cost at $t \in [H]$}
end for
$\boldsymbol{D} = \frac{1}{N_s} \sum_{i=1}^{N_s} \boldsymbol{x}_i$
return $[\boldsymbol{D}, BF]$ $\{\nabla G\}$



(a) Passive transition probabilities



(b) Active transition probabilities

Figure 1: Drop out state domain with three states: drop out ($s = 0$), risk (1), and safe (2). Passive transition probabilities are presented in Figure (a) and active transition probabilities are shown in Figure (b). We take $p^0_{0,0} \in [0.85, 0.95]$, $p^1_{0,0} = 0$, $p^0_{1,1} \in [0.35, 0.5]$ and $p^1_{1,1} = 1$ in our experiments.

[1]. We consider three states: dropout ($s = 0$), at-risk ($s = 1$), and safe ($s = 2$). We consider a binary action set $\mathcal{A} = \{0, 1\}$ corresponding to a passive action ($a = 0$) and active action ($a = 1$). The reward function $R : \mathcal{S} \to \mathbb{R}$ is defined as $R(0) = 0$ and $R(1) = R(2) = 1$. Once an arm reaches the dropout state, it can not transition to any other state, i.e. $P^a_{0,0} = 1$ for all $a \in \mathcal{A}$. Fig. 1 illustrates the remaining active and passive transition probabilities in this domain.

In this dropout domain, interventions may be more urgent in certain rounds depending on the combined state of all arms. For example, if at time t, k arms are at risk of transitioning to a dropout state, i.e. k arms are in state 1 as shown in Figure 1, acting on these k arms at t is more urgent than acting on them at $t + 1$ or other near future steps. Thus this domain illustrates a key instance when the F-RMAB class introduces essential flexibility.

Two-state process. The two-state process models approaches in health intervention planning such as maternal health care [**?**]. This domain models an entity with two states, a *good* and a *bad* state, with reward $R(1) = 1$ for each arm in the good state and $R(0) = 0$ for the bad state. See Figure 2 in Appendix for a diagram of the model,

with four transition probability parameters. Transition probabilities for each arm $n \in [N]$ are $p^1_{0,1} = p^1_{1,1} = 1$, $p^0_{0,0} \in [0.85, 0.95]$ and $p^0_{1,1} \in [0.5, 0.85]$ are uniformly sampled, and $p^0_{0,1} = 1 - p^0_{0,0}$ and $p^0_{1,0} = 1 - p^0_{1,1}$ as shown in Fig. 2.

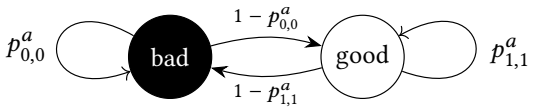


Figure 2: Two-state process domain with $p^a_{s,s'}$ the transition probability from state s to s' after taking action a.

5 RESULTS

We test PDSG (Algorithm 1) and Compress heuristics, a heuristic that translates F-RMABs into classic RMABs by reasoning about consecutive steps, to solve for F-RMABs and compare them against a classic RMAB solution algorithm with fixed per round budget on the three domains described above. For each domain we consider a planning horizon of length H, an initial per round budget of $B = 1$, and vary the length of the flexible time window F.

In Fig. 3 we see that optimal policies that allow for budget flexibility attain higher reward than optimal policies restricted to a fixed budget at every round. The *Hawkins* approach demonstrates the optimal reward achieved with a fixed budget. Our *PDSG* algorithm to solve for optimal policies in F-RMABs attains higher cumulative reward than Hawkins across all settings. Notably, PDSG progressively obtains higher rewards with longer flexible time windows, demonstrating the additional planning power that can be gained with wider windows of flexibility.

As shown in Fig. 3(a), the flexible algorithms *Compress (closing)* and *PDSG-200* obtain a maximum increase in reward of 21.50% and 23.56% respectively for $F = 5$ compared to the reward obtained by *Hawkins*. This increase in reward is obtained by designing a more efficient allocation of resources over a planning horizon of length $H = 30$. This allocation is done by mostly assigning 0 to 2 resources at each step of the planning horizon, even for $F = 3$ and $F = 5$, in

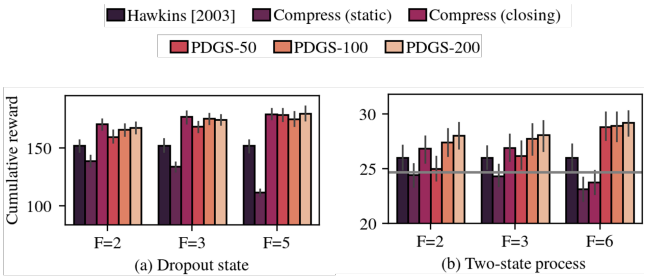


Figure 3: Cumulative reward for (a) dropout state with $H = 30$, $N = 10$, $B = 1$, and (b) two-state process with $H = 6$, $N = 10$, $B = 1$.The cumulative reward axis range in (a) starts at the average value for a policy taking B random actions. The horizontal gray line in (b) denotes this same value for the two-step process domain.

contrast to *Hawkins* which is restricted to use 1 resource at each step.

As shown in Fig. 3(b) for the two-state process domain, our method to solve for policies with flexible budget (*PDSG-200*) attains an increase in reward of 6.71%, 5.66%, and 11.32% for flexible time windows of length $F = 2, 3, 6$ respectively in contrast to the per round budget policy derived by *Hawkins*. We observe that RMABs can also benefit from flexibility in settings with as few as two states, which are relevant settings for health intervention planning, in contrast to the other two domains considering more than two states and having intermediate states that directly characterize waiting steps until reaching a bad state.

6 CONCLUSION

This paper proposes the use of flexible budget restless multi-armed bandits (F-RMABs) as a better alternative for sequential planning in public health settings. F-RMABs allow for the total resources to be budgeted over a flexible time window, enabling public health practitioners to adjust their policies based on changing resource availability and prioritize critical interventions. Our experiments on synthetic domains that are motivated by real-world public health scenarios demonstrated that F-RMAB policies with budget flexibility achieved a significant improvement in performance compared to fixed budget policies. We present these results as a proof of concept of the potential usefulness of flexible budgets in sequential resource allocation in public health settings.

The adoption of F-RMABs can make a substantial contribution to the field of public health, providing a powerful tool for optimizing resource allocation policies and reducing the burden of health mandates. However, there is still much to explore in the application of F-RMABs to real-world data, particularly in the settings of the synthetic domains presented in this paper. Further research is needed to assess the effectiveness of F-RMABs in a broader range of public health settings and to evaluate their impact on health outcomes. Overall, our findings suggest that F-RMABs can significantly enhance resource allocation strategies in public health settings, leading to better health outcomes for the populations being served.

REFERENCES

[1] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. 2021. Learn to Intervene: An Adaptive Learning Policy for Restless Bandits in Application to Preventive Healthcare. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)* (2021). http://arxiv.org/abs/2105.07965 Number: arXiv:2105.07965 arXiv:2105.07965 [cs].

[2] Antonin Chambolle and Thomas Pock. 2011. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision* 40, 1 (May 2011), 120–145. https://doi.org/10.1007/s10851-010-0251-1

[3] Jackson A. Killian, Arshika Lalan, Aditya Mate, Manish Jain, Aparna Taneja, and Milind Tambe. 2023. Adherence Bandits.

[4] Aditya Mate, Jackson A. Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. 2020. Collapsing Bandits and Their Application to Public Health Interventions. *Advances in Neural Information Processing Systems* (July 2020). http://arxiv.org/abs/2007.04432 arXiv: 2007.04432.

[5] Martin Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[6] Paula Rodriguez Diaz, Jackson A. Killian, Lily Xu, Arun Sai Suggala, Aparna Taneja, and Milind Tambe. 2023. Flexible Budgets in Restless Bandits:A Primal-Dual Algorithm for Efficient Budget Allocation.