Adherence Bandits

Jackson A. Killian[*],[1] **Arshika Lalan**[*],[2] **Aditya Mate**[*],[1]
Manish Jain,[2] **Aparna Taneja**,[2] **Milind Tambe**[1,2]

[1]Harvard University, [2]Google Research
jkillian@g.harvard.edu, arshikal@google.com, aditya_mate@g.harvard.edu,
{manishjn, aparnataneja, milindtambe}@google.com

Abstract

We define a new subclass of the restless multi-armed bandit framework, that we name Adherence Bandits, designed to capture the dynamics prevalent in many public health intervention problems. We discuss key properties of Adherence Bandits, their real-world motivations, how structures lead to both technical and computational advantages, and natural extensions that have been or can be made to the subclass. We summarize key research works that have contributed to the growing sub-area and finish by highlighting future directions of research.

Introduction

Sequential resource allocation crops up in a multitude of real world settings. Often, the disparity between available resources and the total number of recipients can be high. Some common examples are cancer screening for a high risk population (Lee 2016), fighting wildfires (Chan, Tran-Thanh, and Viswanathan 2021) and communication networks (Blasco and Gündüz 2015). Restless multi-armed bandits (RMABs) (Whittle 1988) are a widely adopted technique to model such scenarios. We focus on a growing conceptual class of RMAB that emphasizes keeping the resource recipient (arms) in a good state, accruing rewards for all arms kept in that state. That is, 'Adhering' to the good state and preventing 'dropout' to the bad state is the hallmark of these problems. Sticking to an education training regime, and habit formation in medication adherence (Killian et al. 2019) are some examples. Problems in the public health domain are especially of this nature and have been studied in detail. In this work, we propose *Adherence RMABs* as a subclass of RMABs designed to capture these common dynamics.

Restless multi-armed bandits (RMABs) consist of N heterogeneous control processes (arms) and an agent who can pull K arms at each time step to accumulate rewards over time. Each arm of an RMAB is composed of a Markov Decision Process (MDP) (Puterman 2014) which can evolve and change states even when an arm is not pulled, as opposed to the classic *stochastic* multi-armed bandit problem (Sutton and Barto 2018). The agent's goal is to plan a policy for the sequential pulling of arms conforming to the K-arm per-round budget constraint in order to maximise the total reward that can be obtained. The reward for each arm depends generally on the states and actions of the corresponding MDP.

Figure 1: A mother enrolled in ARMMAN's maternal health support program receives an automated message. Photo courtesy of ARMMAN.

RMABs are popularly used in application domains such as machine repair and sensor maintenance (Glazebrook, Ruiz-Hernandez, and Kirkbride 2006), planning anti-poaching patrols (Qian et al. 2016), communication systems (Sombabu et al. 2020), web crawling (Avrachenkov and Borkar 2019) and congestion control (Avrachenkov et al. 2013). In recent times, RMABs have been used to model interventions in the public health domain (Mate et al. 2022; Nishtala et al. 2021; Lee 2016; Li and Varakantham 2022). SAHELI is a system to efficiently utilize the limited availability of health workers for improving maternal and child health, and is the first successfully deployed application for RMABs in public health (Verma et al. 2023).

RMABs are an appropriate solution to model many problems relating to adherence, specifically public health intervention resource allocation (Mate et al. 2022; Killian et al. 2019) problems as they can handle several of the key challenges presented by this domain. For instance, resources are limited in such problems, which is captured in the restricted budget K of RMABs. Additionally, the effect of intervention on individuals may vary, which is captured by RMABs allowing heterogeneous arms. Moreover, an intervention will change the future health or adherence state of a person, compared to if the same person did not receive an intervention, which is captured by MDPs of RMABs, but *not* captured by simpler stochastic MABs. Finally, the goal of most public health intervention problems is to maximize health or adherence, naturally aligning with the reward maximization objective of RMABs. Clearly, RMABs are well suited to these challenges and objectives.

[*]These authors contributed equally and are listed in alphabetical order.

Due to their prevalence in modelling adherence-based problems, it is imperative to look at the characteristics of RMAB models in the specific domain. We term this subset of RMABs as 'Adherence Bandits' and define their attributes. Our key contributions are: (i) Establish Adherence Bandits to maximise adherence in public health intervention problems and lay foundation to their definition. (ii) Derive useful technical properties of Adherence Bandits and their extensions (iii) View key relevant related works through the lens of Adherence Bandits (Table 2 in appendix).

Related Works

Monitoring of patient adherence in public healthcare is a significant problem; non-adherence can be a serious threat to a patient's well being (Martin et al. 2005), not only in terms of health, but also economically. (Sullivan 1990)

Previous work highlights RMABs as a good model of sequential resource allocation by allowing the non-intervened arms to also evolve over time to simulate a more realistic system (Whittle 1988). Several works in the healthcare domain have studied patient adherence without RMABs (Tuldrà et al. 1999; Corotto et al. 2013; Killian et al. 2019) however, these models are unable to handle the sequential nature of resource allocation due to their single-shot predictions. Moreover, budget constraints can be hard to model and the pool of beneficiaries labelled as "high-risk" can itself be very large, defeating the purpose of resource allocation when using such single-shot predictors. Other sequential resource allocation methods, such as (Liao et al. 2020), also fail to take into account the limited budget of resources. RMABs consider sequential resource allocation under a specified budget constraint, which makes them an effective tool to model public health interventions.

RMABs have been successfully explored and widely studied in the public health domain (Mate et al. 2022; Li and Varakantham 2022; Lee 2016). They have also been successfully deployed and used by ARMMAN — a maternal health and childcare NGO focused on creating scalable solutions for empowering pregnant women and mothers and enabling healthy children by improving access to preventive information and services (Verma et al. 2023). Their deployed RMAB system, SAHELI, has already reached approximately 100K beneficiaries and is on track to serve 1 million beneficiaries by the end of 2023. Thus, RMABs are an efficient and scalable solution to model problems in the public space.

In the public health setting, RMABs have specific characteristics that must always be followed. Previous work fails to take into account the adherence-specific attributes of RMABs and does not exploit their properties when building models for these problems. (Mintz et al. 2020) attempt to present a useful framework for modelling habituation for healthcare-adherence improving interventions, however they do so using a *rested* bandit model. A rested MAB ignores the evolving state of beneficiaries in the absence of an intervention, which is a crucial characteristic to take into account to prevent dropouts. Our work extends a restless multi-armed bandit model for maximising adherence, and establishes specific properties that are relevant to the domain.

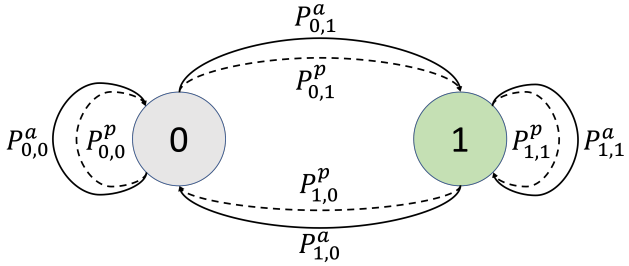

Figure 2: A 2-action MDP with a good (1) and bad (0) state. P^a (solid lines) denotes "active" action (intervention) and P^p (dashed lines) denotes "passive" action (no intervention).

Preliminaries

RMAB

A Restless Multi-armed Bandit (RMAB) model consists of N arms, where each arm evolves as an independent MDP. The i-th arm in an RMAB model is an MDP defined by the tuple $(\mathcal{S}_i, \mathcal{A}_i, R_i, P_i)$. $\mathcal{S}_i$ and $\mathcal{A}_i$ are the state space and action space respectively, and $R_i, P_i : S \times A \times S \to \mathbb{R}$ are the reward and transition functions. $P^{\alpha}_{i,s,s'}$ denotes the transition probability of evolving from state s to state s' given the action α. A policy π is a mapping $\times_{i\in[N]}\mathcal{S}_i \to \times_{i\in[N]}\mathcal{A}_i$ that selects the action to be taken on each arm, given the state of all arms, subject to the resource constraint $||\pi(\boldsymbol{s})||_1 \leq K \;\; \forall \boldsymbol{s} \in \times_{i\in[N]}\mathcal{S}_i$. The optimal policy $\pi^\star$ is the policy that maximises the total reward accrued. Two common criterion to measure the total reward accrued are the discounted reward or average reward criteria, which sum up the individual rewards accrued at each time step t. The discounted reward criteria across T time steps and N arms is $\sum_{t=1}^{T}\gamma^{t-1}\sum_{i\in[N]} r_{t,i}$, where $\gamma \in [0, 1]$ is the discount factor.

Solution Strategy

RMABs suffer from the curse of dimensionality as the state space grows exponentially in the number of arms. Solving for the optimal actions in an RMAB is a P-SPACE hard problem in general (Papadimitriou and Tsitsiklis 1999), even with the transition rules known fully. Existing work has focused on designing index based solution techniques to RMABs. The 'Whittle Index'(Whittle 1988) solution strategy is the most popular solution paradigm for RMABs. The sanctity of this method however hinges on the validity of a technical condition called 'indexability'. Indexability establishes the existence of an index-based solution and also guarantees the asymptotic optimality of its solution. However, even verifying indexability for RMABs has proven to be notoriously difficult, with no known result available on indexability for RMABs in general. In this work, we show indexability guarantees to hold for the basic Adherence Bandits framework.

	2-State, 1-Dimensional	S-State, 1-Dimensional
Full Observability	Indexable/Threshold Optimal	?/?
Collapsing Observability	88% Indexable/Threshold Optimal	?/?
General Observability	?/?	?/?

Table 1: Summary existing technical results for ARs with different structure. Question marks indicated open areas of study. Threshold optimal indicates that a simple class of policies can be used to compute the Whittle index in closed form.

Basics of an Adherence RMAB

Properties of Adherence RMABs

We define Adherence Bandits (ABs) as a subclass of general RMABs with the following basic properties.

1. ABs have two discrete states which capture "adherence" or "engagement" with respect to the health content:
$$s \in \mathcal{S} := \{0, 1\} \tag{1}$$
2. The states of adherence RMABs are fully observable, e.g., planners know whether or not a health message was listened to via call records.
3. Rewards are a simple identity function of the current state. Moreover, rewards are accrued on all arms, regardless of whether the planner takes an action. That is:
$$R(s, a, s') = s \tag{2}$$
Coupled with the reward maximization objective of RMAB, this ensures that policies maximize the total adherence of the cohort over time.
4. Acting is at least as good as not acting. That is, delivering an intervention should not decrease the probability that the arm reaches the adhering state:
$$P^a_{0,1} \geq P^p_{0,1}; P^a_{1,1} \geq P^p_{1,1}; \tag{3}$$
5. An arm that is already adhering is more likely to stay adhering than an arm that was not previously adhering:
$$P^a_{1,1} > P^a_{0,1}; P^p_{1,1} > P^p_{0,1} \tag{4}$$
6. ABs have just two actions, namely $\{$NOT INTERVENE,INTERVENE$\}$, or $\{0,1\}$.
7. The number of interventions is much lesser than the total number of beneficiaries
$$K << N \tag{5}$$

Useful Technical Properties

We show that two-state RMAB instances with fully observable states, are always indexable (proof in appendix).

Theorem 1. *Consider an RMAB instance with each arm representing a fully observable $2-$state MDP with an arbitrary transition matrix P. Each arm of such an RMAB is indexable and consequently, the RMAB is indexable.*

Moreover, the 2-state model has some computational conveniences when considering robust objectives. First, the 2-state model guarantees indexability, reducing the class of policies over which one must search for adversaries (Killian et al. 2023). Additionally, the constraints on the probability space reduce the size of the uncertainty set that must be considered, helping make adversarial search algorithms tractable (Killian et al. 2023).

Useful Applied Considerations

The two-state model is easily interpretable, e.g., a binary yes/no adherence state, which facilitates discussion with domain experts about the dynamics of the problem – one of the most complicated and time-consuming parts of many collaborations, but often overlooked in academic papers.

From the computational side, the two-state model is also convenient for allowing scale-up, since real-world public health problems often have thousands if not millions of arms (Mate et al. 2022). The smaller models also require less data to support, which is important since data is often scarce.

Real Problems as Basic ABs

The Basic AB, and related extensions thereof, capture many real world problems. One important example is the maternal health engagement problem faced by ARMMAN. In this setting mothers must listen to weekly automated messages providing life-saving information about various stages of their pregnancy. ARMMAN has limited health workers who can place service calls to try to improve adherence of mothers with low listenership. We have collaborated with ARMMAN to model this problem as a Basic AB, where the problem meets all the above seven properties. We have conducted field studies demonstrating the positive impact of targeting interventions via Basic ABs (Mate et al. 2022), and have followup work studying the robustness to model uncertainty over the same Basic AB (Killian et al. 2023). Another key example of an AB is that of scheduling chronic care visits (Deo et al. 2013). They also consider an RMAB model that is a special case of the basic 2-state AB with perfect intervention effects. We give an extended comparison to related works under the lens of the AB definition and its extensions in the appendix (due to space constraints).

Extensions of Basic ABs

While the basic AB model captures the key elements crucial to utilizing the RMAB framework for adherence problems in

public health, oftentimes, this there are additional complexities involved that necessitate fundamental advances along several axes as highlighted in the following subsections.

Adherennce Bandits with Collapsing Observability

In many public health settings, especially outside of digital content delivery, full observability of the state may not be possible. As a real-world example, the adherence status of tuberculosis patients may only be observed for patients receiving an intervention (Mate et al. 2020). For arms that are not pulled, there is uncertainty in the true state, which may evolve in the background. However for arms that are pulled, an observation is received and the uncertainty collapses.

Inspired from the tuberculosis medication adherence monitoring use case, Mate et al. (2020) propose the "Collapsing Bandits" model to capture this phenomenon, extending basic ABs by relaxing the full observability assumption. They also adopt the Whittle index based solution approach and utilize the special "collapsing" structure to speed up the index computation. Herlihy et al. (2021) Adopt the collapsing structure to the domain of sleep apnea treatment adherence. Mate et al. (2021) then consider an extension of collapsing observability to a 'streaming' setting, in which new arms arrive and existing arms leave the system. This models health programs in which new enrollees may join on an ongoing basis and existing enrollees finish the health program and leave after a finite stay.

The indexability guarantees of basic ABs however, no longer fully extend to Collapsing Bandits. (Mate, Perrault, and Tambe 2021) derive sufficient conditions in terms of the transition function P, that guarantee indexability. These conditions yield provable indexability guarantees for 88% of all instances of Collapsing Bandits. (Mate, Perrault, and Tambe 2021) also propose an extension to the collapsing observability model to account for 'imperfect' observations. In this setting, the states are still unobserved for arms not pulled, but even for arms that are pulled, the observation may be imprecise and the uncertainty only collapses partially to fixed probability values depending on the observation.

Extending basic ABs to handle general observability, while giving technical guarantees, is an open challenge.

Adherence Bandits with General Reward function

The basic AB model assumes a simple reward function in the state s, equivalent to optimizing for an objective that depends on aggregate statistics, such as the average adherence of the beneficiary cohort. In partial observability settings, this translates to a linear reward function in terms of the belief vector over the true underlying states. However, in many scenarios, a planner may be risk-sensitive (e.g., risk-averse) or may need to account for fairness and need tools that go beyond optimizing for aggregate cohort statistics.

Mate, Perrault, and Tambe (2021) extend basic ABs by proposing a technique that can admit any non-linear, non-decreasing reward functions ρ. This relaxes the reward function assumption to require only a simpler condition: $R(1) > R(0)$. Herlihy et al. (2021) propose an improvement of index policies to guarantee 'probabilistic fairness', defined in terms of a minimum probability of receiving an intervention.

ABs with Multiple actions

In many public health intervention problems, planners have access to more than one type of intervention. In such cases, more tools or structures are needed to develop efficient and well-performing policies, since the commonly-used Whittle Index policy for ABs requires having only two actions.

First, we extend the constraints on P for the multi-action case. We maintain that intervention has a positive effect compared to no intervention. That is: $P_{0,1}^{a_i} \geq P_{0,1}^{a_0}; P_{1,1}^{a_i} \geq P_{1,1}^{a_0} \ \forall i > 0$. Where a_0 is the passive action. However, in general, no relative ordering of the effectiveness of interventions types is assumed, reflecting real-world heterogeneity.

To solve Multi-action ABs, the Lagrangian relaxation can be used, from which the Whittle index policy is derived (Killian, Perrault, and Tambe 2021). However, this involves solving a linear program, which is less efficient than computing indexes. To account for this, previous work developed a method that take advantage of realities in real-world data, namely that many patients have unwavering adherence behavior, regardless of intervention (Killian et al. 2019; Killian, Perrault, and Tambe 2021). In a multi-action AB model, the "value-function" for such unwavering patients has a simple form that can be replaced with a far cheaper representation without sacrificing solution quality. Killian, Perrault, and Tambe (2021) develop a method to automatically detect such patients, work with their bounded forms, and find well-performing policies by focusing computational expenditure only on those patients with more complex dynamics who *need* intervention to maintain good adherence.

Killian et al. (2021) extended these ideas to develop methods for learning policies for multi-action ABs online, and Killian et al. (2022) for operating in robust multi-action AB environments. In such cases, the simple 2-state models, P constraints, and existence of unwavering-style patients all provide sample efficiency benefits that greatly speed up policy learning.

Other Extensions to ABs

We envision other natural extensions to ABs. First, it may be natural to include more than two adherence states, e.g., highly engaged, semi-engaged, or disengaged, to capture more complex dynamics. Second, it may be natural to jointly model adherence and its health effects, i.e., designing a multi-dimensional state space. Third, it may be desirable to plan budget allocations flexibly within certain time periods, e.g., plan weekly allocations with a monthly budget (Diaz et al. 2023). Finally, in some cases, especially in long treatment regimens, adherence transition dynamics may change for each patient over time (non-stationarity of P).

Conclusion

We identify and define 'Adherence Bandits' (AB) as a special subclass of restless multi-armed bandits, that is naturally suited to address prevalent adherence monitoring challenges in public health. We highlight key defining features of a basic AB, discuss existing work extending basic ABs, as well as directions for future work appealed to by existing public health challenges.

References

Avrachenkov, K.; Ayesta, U.; Doncel, J.; and Jacko, P. 2013. Congestion control of TCP flows in Internet routers by means of index policy. *Computer Networks*, 57(17): 3463–3478.

Avrachenkov, K.; and Borkar, V. S. 2019. A learning algorithm for the Whittle index policy for scheduling web crawlers. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1001–1006. IEEE.

Ayer, T.; Zhang, C.; Bonifonte, A.; Spaulding, A. C.; and Chhatwal, J. 2019. Prioritizing hepatitis C treatment in US prisons. *Operations Research*, 67(3): 853–873.

Blasco, P.; and Gündüz, D. 2015. Multi-Access Communications With Energy Harvesting: A Multi-Armed Bandit Model and the Optimality of the Myopic Policy. *IEEE Journal on Selected Areas in Communications*, 33(3): 585–597.

Chan, H.; Tran-Thanh, L.; and Viswanathan, V. 2021. Fighting wildfires under uncertainty: A sequential resource allocation approach. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 4322–4329.

Corotto, P. S.; McCarey, M. M.; Adams, S.; Khazanie, P.; and Whellan, D. J. 2013. Heart failure patient adherence: epidemiology, cause, and treatment. *Heart failure clinics*, 9(1): 49–58.

Deo, S.; Iravani, S.; Jiang, T.; Smilowitz, K.; and Samuelson, S. 2013. Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operations Research*, 61(6): 1277–1294.

Diaz, P. R.; Killian, J. A.; Xu, L.; Suggala, A. S.; Taneja, A.; and Tambe, M. 2023. Flexible Budgets in Restless Bandits:A Primal-Dual Algorithm for Efficient Budget Allocation. In *AAAI Conference on Artificial Intelligence (AAAI)*.

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

Herlihy, C.; Prins, A.; Srinivasan, A.; and Dickerson, J. 2021. Planning to Fairly Allocate: Probabilistic Fairness in the Restless Bandit Setting. *arXiv preprint arXiv:2106.07677*.

Killian, J. A.; Biswas, A.; Shah, S.; and Tambe, M. 2021. Q-learning Lagrange policies for multi-action restless bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 871–881.

Killian, J. A.; Perrault, A.; and Tambe, M. 2021. Beyond" To Act or Not to Act": Fast Lagrangian Approaches to General Multi-Action Restless Bandits. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 710–718.

Killian, J. A.; Wilder, B.; Sharma, A.; Choudhary, V.; Dilkina, B.; and Tambe, M. 2019. Learning to prescribe interventions for tuberculosis patients using digital adherence data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2430–2438.

Killian, J. A.; Xu, L.; Biswas, A.; and Tambe, M. 2022. Restless and Uncertain: Robust Policies for Restless Bandits via Deep Multi-Agent Reinforcement Learning. In *The 38th Conference on Uncertainty in Artificial Intelligence*.

Killian, J. A.; Xu, L.; Biswas, A.; Verma, S.; Nair, V.; Taneja, A.; Hegde, A.; Madhiwalla, N.; Diaz, P. R.; Johnson-Yu, S.; and Tambe, M. 2023. Robust Planning over Restless Groups: Engagement Interventions for a Large-Scale Maternal Telehealth Program. In *AAAI Conference on Artificial Intelligence*.

Lee, E. 2016. *Management of a Chronically Ill Population: An Operations Approach to Liver Cancer Screening.* Ph.D. thesis.

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

Li, D.; and Varakantham, P. 2022. Efficient Resource Allocation with Fairness Constraints in Restless Multi-Armed Bandits. *arXiv preprint arXiv:2206.03883*.

Liao, P.; Greenewald, K.; Klasnja, P.; and Murphy, S. 2020. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1): 1–22.

Martin, L. R.; Williams, S. L.; Haskard, K. B.; and DiMatteo, M. R. 2005. The challenge of patient adherence. *Therapeutics and clinical risk management*, 1(3): 189.

Mate, A.; Biswas, A.; Siebenbrunner, C.; and Tambe, M. 2021. Efficient algorithms for finite horizon and streaming restless multi-armed bandit problems. *arXiv preprint arXiv:2103.04730*.

Mate, A.; Killian, J.; Xu, H.; Perrault, A.; and Tambe, M. 2020. Collapsing Bandits and Their Application to Public Health Intervention. *Advances in Neural Information Processing Systems*, 33: 15639–15650.

Mate, A.; Madaan, L.; Taneja, A.; Madhiwalla, N.; Verma, S.; Singh, G.; Hegde, A.; Varakantham, P.; and Tambe, M. 2022. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12017–12025.

Mate, A.; Perrault, A.; and Tambe, M. 2021. Risk-Aware Interventions in Public Health: Planning with Restless Multi-Armed Bandits. In *AAMAS*, 880–888.

Mintz, Y.; Aswani, A.; Kaminsky, P.; Flowers, E.; and Fukuoka, Y. 2020. Nonstationary bandits with habituation and recovery dynamics. *Operations Research*, 68(5): 1493–1516.

Nishtala, S.; Madaan, L.; Mate, A.; Kamarthi, H.; Grama, A.; Thakkar, D.; Narayanan, D.; Chaudhary, S.; Madhiwalla, N.; Padmanabhan, R.; et al. 2021. Selective Intervention Planning using Restless Multi-Armed Bandits to Improve Maternal and Child Health Outcomes. *arXiv preprint arXiv:2103.09052*.

Papadimitriou, C.; and Tsitsiklis, J. 1999. The complexity of optimal queueing network control." inMathematics of Operations Research.

Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Qian, Y.; Zhang, C.; Krishnamachari, B.; and Tambe, M. 2016. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 123–131.

Sombabu, B.; Mate, A.; Manjunath, D.; and Moharir, S. 2020. Whittle index for AoI-aware scheduling. In *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, 630–633. IEEE.

Sullivan, S. D. 1990. Noncompliance with medication regimens and subsequent hospitalization: a literature analysis and cost of hospitalization estimate. *J Res Pharm Econ*, 2: 19–33.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tuldrà, A.; Ferrer, M. J.; Fumaz, C. R.; Bayés, R.; Paredes, R.; Burger, D. M.; and Clotet, B. 1999. Monitoring adherence to HIV therapy. *Archives of Internal Medicine*, 159(12): 1376–1377.

Verma, S.; Singh, G.; Mate, A.; Verma, P.; Gorantla, S.; Madhiwalla, N.; Hegde, A.; Thakkar, D.; Taneja, A.; Jain, M.; and Milind, T. 2023. Deployed SAHELI: Field Optimization of Intelligent RMAB for Maternal and Child Care.

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

Domain	Description	Status	Basic AB?	Collapsing Extension?	Other AB Extension?
Maternal health (Mate et al. 2022; Verma et al. 2023)	Mothers scheduled for engagement-boosting interventions in telehealth program.	Deployed	Yes	No	Possible, not yet proposed
Community care scheduling (Deo et al. 2013)	Schedule asthma treatments. 2-state health model with all AB properties. Multi-state extensions.	Simulated	Yes	No	Also consider multi-state
Tuberculosis (TB) (Mate et al. 2020)	TB patients scheduled for adherence-boosting interventions. 2-states, observed on action.	Simulated	No	Yes	Possible
Sleep apnea treatment adherence (Herlihy et al. 2021)	Sleep apnea patients scheduled for CPAP adherence-boosting interventions. 2-states, observed on action	Simulated	No	Yes	Fairness constraints
Cancer screening (Lee, Lavieri, and Volk 2019)	Monotone disease progression model. Action reveals hidden state.	Simulated	No	No	Multi-state, specialized collapsing observability
Hepatitis C (Ayer et al. 2019)	Model Hepatitis C progression of prison inmates. Monotone disease state progression model. Action stops progression.	Simulated	No	No	Multi-state
Daily Step Counts (Diabetes) (Mintz et al. 2020)	Arms are actions, with possible increasing or decreasing reward depending on frequency of play.	Simulated	No	No	Continuous state, sub-Gaussian rewards, multi-action

Table 2: Related works under the lens of Adherence Bandits (ABs).

Appendix

Please see Table 2 for a list of published related works as they relate to Adherence Bandits.

Proof of Theorem 1

Theorem 1. *Consider an RMAB instance with each arm representing a fully observable $2-$state MDP with an arbitrary transition matrix P. Each arm of such an RMAB is indexable and consequently, the RMAB is indexable.*

Proof. To show indexability of RMABs with fully observed 2-state MDPs on arms, we prove three useful Lemmas which lead to the Theorem proof.

Lemma 1. *An RMAB arm with transition matrix P, value function $V_m(s)$ for state s, passive subsidy m and a discount factor of β is indexable if it satisfies:*

$$1+\beta\big[(P^a_{s1}-P^p_{s1})(V'_m(0)-V'_m(1))\big]\geq 0 \textit{ for } s\in\{0,1\}$$

Proof. Consider the above condition:

$$\begin{aligned}
&1+\beta\big[(P^a_{s1}-P^p_{s1})(V'_m(0)-V'_m(1))\big]\geq 0\\
\implies &1+\beta\big[V'(1)(P^p_{s1}-P^a_{s1})+(P^a_{s1}-P^p_{s1})V'_m(0)\big]\geq 0\\
\implies &1+\beta\Big[\big(V'(1)P^p_{s1}-V'(0)P^p_{s1}+V'(0)\big)\\
&-\big(V'(1)P^a_{s1}-V'(0)P^a_{s1}+V'(0)\big)\Big]\geq 0\\
\implies &\Big[1+\beta\big(V'(1)P^p_{s1}+V'(0)(1-P^p_{s1})\big)\Big]\\
&-\Big[\beta\big(V'(1)P^a_{s1}+V'(0)(1-P^a_{s1})\big)\Big]\geq 0\\
\implies &\frac{\partial}{\partial m}\Big[m+\beta\Big(V(1)P^p_{s1}+V(0)(1-P^p_{s1})\Big)\Big]\\
&-\frac{\partial}{\partial m}\Big[\beta\Big(V'(1)P^a_{s1}+V'(0)(1-P^a_{s1})\Big)\Big]\geq 0\\
\implies &\frac{\partial}{\partial m}\Big[V^p_m(s)\Big]-\frac{\partial}{\partial m}\Big[V^a_m(s)\Big]\geq 0
\end{aligned}\tag{6}$$

This condition implies that passive value function, $V^p_m(s)$ increases with m at a rate greater than the active value function, $V^a_m(s)$. Thus, if the passive action were optimal for a given $m*$, it implies the passive action will still be optimal $\forall m > m*$. This implies indexability. □

Lemma 2. *Let ΔV denote $V'_m(0)-V'_m(1)$, the difference in the derivatives of value functions for states 0 and 1, with respect to passive subsidy, m. Then,*

$$\Delta V \coloneqq V'_m(0)-V'_m(1)=\frac{\mathbf{1}_{\alpha_0=p}-\mathbf{1}_{\alpha_1=p}}{1-\beta\big(P^{\alpha_1}_{11}-P^{\alpha_0}_{01}\big)}\tag{7}$$

Proof. We solve for ΔV as follows:

$$\begin{aligned}
\Delta V =&V'_m(0)-V'_m(1)\\
=&\frac{\partial}{\partial m}\big(\max\{V^p_m(0),V^a_m(0)\}\big)\\
&-\frac{\partial}{\partial m}\big(\max\{V^p_m(1),V^a_m(1)\}\big)\\
=&\frac{\partial}{\partial m}\big(V^{\alpha_0}_m(0)-V^{\alpha_1}_m(1)\big)\\
&\text{where } \alpha_s \text{ denotes the optimal action at state } s\\
=&\mathbf{1}_{\alpha_0=p}-\mathbf{1}_{\alpha_1=p}+\beta\Big[P^{\alpha_0}_{01}\big(V'_m(1)-V'_m(0)\big)\\
&-P^{\alpha_1}_{11}\big(V'_m(1)-V'_m(0)\big)\Big]\\
&\text{where } \mathbf{1} \text{ denotes the indicator function}\\
=&\mathbf{1}_{\alpha_0=p}-\mathbf{1}_{\alpha_1=p}+\beta\Big[\big(V'_m(0)-V'_m(1)\big)\big(P^{\alpha_1}_{11}-P^{\alpha_0}_{01}\big)\Big]\\
\Delta V =&\mathbf{1}_{\alpha_0=p}-\mathbf{1}_{\alpha_1=p}+\beta\Big[\big(\Delta V\big)\big(P^{\alpha_1}_{11}-P^{\alpha_0}_{01}\big)\Big]
\end{aligned}$$

Re-arranging the terms, we get:

$$\Delta V =\frac{\mathbf{1}_{\alpha_0=p}-\mathbf{1}_{\alpha_1=p}}{1-\beta\big(P^{\alpha_1}_{11}-P^{\alpha_0}_{01}\big)}\tag{8}$$

□

Plugging in the expression for ΔV back in the condition of Lemma 1, the condition for indexability can be written as:

$$1+\frac{\beta\big[\Delta P_s(\mathbf{1}_{\alpha_0=p}-\mathbf{1}_{\alpha_1=p})\big]}{1-\beta\big(P^{\alpha_1}_{11}-P^{\alpha_0}_{01}\big)}\geq 0 \text{ for } s\in\{0,1\}\tag{9}$$

$$\text{i.e. } \frac{1+\beta\big(\Delta P_s\ \Delta\mathbf{1}-\Delta P_a\big)}{1-\beta\ \Delta P_\alpha}\geq 0 \text{ for } s\in\{0,1\}\tag{10}$$

where $\Delta\ \mathbf{1}=\mathbf{1}_{\alpha_0=p}-\mathbf{1}_{\alpha_1=p}$, $\Delta P_\alpha = P^{\alpha_1}_{11}-P^{\alpha_0}_{01}$ and recall that $\Delta P_s = P^a_{s1}-P^p_{s1}$

We finish the theorem proof by showing in Lemma 3 that the condition of Equation 10 holds true.

Lemma 3. *For a 2-state MDP, let $\alpha_s\in\{a,p\}$ denote the optimal action at state s under a passive subsidy of m. Define $\Delta\ \mathbf{1}\coloneqq\mathbf{1}_{\alpha_0=p}-\mathbf{1}_{\alpha_1=p}$, $\Delta P_\alpha\coloneqq P^{\alpha_1}_{11}-P^{\alpha_0}_{01}$. Then,*

$$\frac{1+\beta\big(\Delta P_s\ \Delta\mathbf{1}-\Delta P_\alpha\big)}{1-\beta\ \Delta P_\alpha}\geq 0 \textit{ for } s\in\{0,1\}\tag{11}$$

Proof. We prove this Lemma by showing that both the numerator and denominator of the expression on the left-hand-side above are non-negative. For the denominator:

$$\begin{aligned}
&\|\beta\Delta P_\alpha\|=\|\beta\|.\|\Delta P_\alpha\|<1\\
\implies &0\leq 1-\|\beta\Delta P_\alpha\|\leq 1-\beta\Delta P_\alpha\\
\implies &\text{Denominator is non-negative}
\end{aligned}\tag{12}$$

For the numerator:

Consider the 4 possible cases corresponding to possible combinations of values of α_0 and α_1. We show that in each case the numerator is non-negative.

Case 1. $\alpha_0 = a, \alpha_1 = a$ AND
Case 2. $\alpha_0 = p, \alpha_1 = p$:

$$\text{Numerator} = 1 + \beta(-\Delta P_\alpha) \geq 0$$

Case 3. $\alpha_0 = p, \alpha_1 = a$:

$$\begin{aligned}\text{Numerator} &= 1 + \beta(\Delta P_s - \Delta P_\alpha) \\ &= 1 + \beta(P^a_{s1} - P^p_{s1} - P^a_{11} + P^p_{01}) \\ &= 1 + \beta\big((P^a_{s1} - P^a_{11}) - (P^p_{s1} - P^p_{01})\big)\end{aligned}$$

The former (latter) bracket is 0 if $s = 1$ ($s = 0$).

$$\implies \text{Numerator} \geq 0 \tag{13}$$

Case 4: $\alpha_0 = a, \alpha_1 = p$

$$\begin{aligned}\text{Numerator} &= 1 + \beta(-\Delta P_s - \Delta P_\alpha) \\ &= 1 + \beta(-P^a_{s1} + P^p_{s1} - P^p_{11} + P^a_{01}) \\ &= 1 + \beta\big(- (P^a_{s1} - P^a_{01}) + (P^p_{s1} - P^p_{11})\big)\end{aligned}$$

The former (latter) bracket is 0 if $s = 1$ ($s = 0$).

$$\implies \text{Numerator} \geq 0 \tag{14}$$

This completes the Theorem proof. □