From Restless to Contextual: A Thresholding Bandit Reformulation for Finite-horizon Improvement

Jiamin Xu
Cornell Tech

Ivan Nazarov
Causal Foundry

Aditya Rastogi
Causal Foundry

África Periáñez
Causal Foundry

Kyra Gan
Cornell Tech

Abstract

This paper addresses the poor finite-horizon performance of existing online *restless bandit* (RB) algorithms, which stems from the prohibitive sample complexity of learning a full *Markov decision process* (MDP) for each agent. We argue that superior finite-horizon performance requires *rapid convergence* to a *high-quality* policy. Thus motivated, we introduce a reformulation of online RBs as a *budgeted thresholding contextual bandit*, which simplifies the learning problem by encoding long-term state transitions into a scalar reward. We prove the first non-asymptotic optimality of an oracle policy for a simplified finite-horizon setting. We propose a practical learning policy under a heterogeneous-agent, multi-state setting, and show that it achieves a sublinear regret, achieving *faster convergence* than existing methods. This directly translates to higher cumulative reward, as empirically validated by significant gains over state-of-the-art algorithms in large-scale heterogeneous environments. The code is provided in github. Our work provides a new pathway for achieving practical, sample-efficient learning in finite-horizon RBs.

1 INTRODUCTION

Restless bandits extend classic contextual bandits by modeling state transitions and incorporating per-step budget constraints. In a RB, a set of independent arms (agents) evolves over discrete time steps as MDP, with the goal of dynamically allocating a fixed number of active plays to maximize cumulative rewards. However, this modeling power comes with significant learning challenges. Even in the *offline* setting with known model parameters, RB is PSPACE-hard (Papadimitriou & Tsitsiklis, 1999). This computational intractability has directed much of the research toward policies that provide *asymptotic guarantees*, such as Whittle's Index Policy (WIP) (Whittle, 1988; Weber & Weiss, 1990; Hsu, 2018; Kadota et al., 2016) and linear programming-based methods (Verloop, 2016; Zhang & Frazier, 2021; Brown & Zhang, 2023; Hong et al., 2024; Gast et al., 2024; Xiong et al., 2022a), which are proven to approach optimal as the number of agents approaches infinity.

This complexity carries over into the *online* setting, where transition dynamics are unknown. Prior online RB algorithms largely seek to learn the optimal policy by first estimating the full underlying MDP for each agent. Consequently, online RB methods are largely inspired by, and mirror the standard frameworks of, *reinforcement learning* (RL).

They predominantly fall into two canonical RL categories. *Model-based* approaches explicitly learn transition probabilities, typically within either *episodic finite-horizon* (Jung & Tewari, 2019; Wang et al., 2023b; Liang et al., 2024) or infinite-horizon (Ortner et al., 2012; Wang et al., 2020; Xiong et al., 2022b; Akbarzadeh & Mahajan, 2023; Xiong et al., 2022a) frameworks. *Model-free* methods focus on learning the Q-function specifically in the infinite-horizon setting (Killian et al., 2021; Xiong & Li, 2023; Xiong et al., 2024; Kakarapalli et al., 2024). A common strategy in both categories is to then integrate these learned models with WIP for decision-making.[1] These online methods are designed to converge to their asymptotically optimal offline counterparts when agents are *homogeneous*, where the *convergence rate* is measured by the notion of *regret* (formalized in Sec. 2.2).

This reliance on standard RL frameworks means that

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

[1] We note that while LP-based methods are powerful in the offline setting, they are generally more difficult to adapt for online learning.

the performance guarantees for these online algorithms are inherently *asymptotic* in the number of episodes or time steps. However, this focus constitutes a fundamental limitation in practical applications,[2] where horizons are finite due to budget cycles, program durations, or policy deadlines (del Río et al., 2024a,b). In domains like targeted patient interventions (Villar et al., 2015; Biswas et al., 2021; Mate et al., 2021) or resource allocation (Borkar et al., 2017; Brown & Smith, 2020), systems evolve continuously and cannot be reset, making the standard episodic RL framework impractical.

Consequently, an algorithm's performance is judged by its cumulative reward within a *single, limited-time episode*. An asymptotically optimal policy that requires a long time to learn offers no practical value *if it fails to achieve a high reward within the operational horizon*. This critical limitation of prior work motivates a fundamental shift away from the objective of learning complex MDPs, and towards a reformulation designed for sample-efficient learning in the non-episodic, finite-horizon setting. Our contributions are fourfold:

- **A Novel Reformulation**: we interpret regret as convergence rate (Sec. 2.2) and show that online algorithms may achieve higher cumulative reward in a finite horizon by targeting alternative design objectives, which leads to accelerated convergence (Sec. 2.3). This motivates us to introduce a *budgeted thresholding contextual multi-armed bandit* (BT-CMAB), which simplifies the learning objective of RBs to a contextual bandit problem through carefully designed rewards (Sec. 2.4), enabling sample-efficient learning in finite horizons.
- **First Non-asymptotic Finite-Horizon Optimality Guarantee**: we establish the first finite-horizon non-asymptotic (valid for an arbitrary number of agents) optimality guarantee for a tractable RB setting, proving our oracle greedy policy is optimal for a homogeneous 2-state model (Theorem 3.1). This provides a theoretical foundation absent in prior work.
- **A Fast-Converging Learning Algorithm**: we propose ϵ-GT (ϵ-*Greedy Thresholding*, Algorithm 1, Sec. 4), a novel learning algorithm for RB with *heterogeneous agents*. We establish that it achieves sublinear convergence rate (Theorem 4.4), enabled by the careful choice of decaying ϵ.
- **Empirical Validation**: we numerically demonstrate (in Section 5) that our algorithm consistently outperform existing algorithms in various finite-horizon online RB settings across a wide range of instances *with heterogeneous agents*.

[2]Table A.1 provides key application areas for RBs.

1.1 Additional Related Work

Thresholding Bandit/MDP Our paper relates to regret-minimizing thresholding bandit problems, where regret is defined by the gap between the arm's mean and a predefined threshold when the mean falls below the threshold. Under this setting, minimax optimal (max is taken over all possible bandit instances and min taken over all possible learning algorithms) constant regret is achievable (Tamatsukuri & Takahashi, 2019; Michel et al., 2022; Feng et al., 2024). Our problem setup differs from the traditional thresholding bandit framework by incorporating context, where the arrival of contexts follows an MDP. In contrast, Hajiabolhassan & Ortner (2023) study threshold MDPs under an infinite horizon setting, seeking policies with average rewards exceeding a predefined threshold. We, however, focus on the finite-horizon setting.

Episodic MDP RB can be viewed as a single, massive MDP with a state space that is the combinatorial product of all agent states. Thus, our work also relates to episodic MDP, which focuses on maximizing cumulative rewards across a finite horizon. Let K be the number of episodes, each containing T horizon. Azar et al. (2017) establishes a minimax regret of $\mathcal{O}(\sqrt{T^2K})$ for stationary transitions, and Jin et al. (2018) derives a regret of $\mathcal{O}(\sqrt{T^4K})$ for nonstationary transitions.

MAB Experimental Design Our incremental reward explicitly incorporates the trade-off between reward maximization and treatment effect estimation (i.e., estimating the difference in rewards under $A = 0$ and $A = 1$), previously studied in MAB with Pareto-optimal algorithms provided (Simchi-Levi & Wang, 2025; Zhong et al., 2021; Zhang & Wang, 2024). Unlike prior work that treats these aspects separately, ours inherently combines both aspects through *reward* that depends on intervention (i.e., $A = 1$) and non-intervention (i.e., $A = 0$) *outcomes*.

2 FINITE-HORIZON RESTLESS BANDITS AND OUR REFORMULATION

We introduce restless bandits in Sec. 2.1 and reinterpret regret as convergence in Sec. 2.2. Sec. 2.3 argues that better finite-sample performance requires new design objectives, while maintaining reward-based evaluation. We present our reformulation in Sec. 2.4.

2.1 Restless Bandits

In a finite-horizon RB problem, the learner is provided with M *independent* arms, each representing one independent agent. Each agent m is modeled as an

MDP represented by the tuple $(\mathcal{S}, \mathcal{A}, P_a^m, R^m)$, where $P_a^m(s, s') : \mathcal{S} \times \mathcal{S} \to [0, 1]$ represents the transition probability from state s to s' under action a for agent m, and $R^m(s) : \mathcal{S} \to \mathbb{R}$ represents the *expected reward* that agent m receives at state s. We focus on binary action, i.e., $\mathcal{A} = \{0, 1\}$. We highlight that in classical RBs, the *expected* reward, $R^m(s)$, is independent of the action, and the learner observes the reward *immediately after arriving at* the state.

At each time t, given the state $\boldsymbol{s}_t$, the learner's decision $\boldsymbol{a}_{t+1}$ follows a history-dependent random policy, π_{t+1}. Let $\mathcal{H}_t := \{(\boldsymbol{a}_i, \boldsymbol{s}_i, \boldsymbol{r}_i)\}_{i=1}^t \cup \{\boldsymbol{s}_0, \boldsymbol{r}_0\}$ be the history observed up to time t, and let $\mathcal{H}_t^+ := \mathcal{H}_t \cup \{\boldsymbol{a}_{t+1}\}$ be the augmented history up to taking action $\boldsymbol{a}_{t+1}$. Similarly, denote $\mathcal{F}_t$ as the σ-algebra generated by $\mathcal{H}_t$ and $\mathcal{F}_t^+$ as the σ-algebra generated by $\mathcal{H}_t^+$. Then, $\pi_{t+1} : \mathcal{H}_t \to \Delta\big(\{0, 1\}^M\big)$, where $\Delta\big(\{0, 1\}^M\big)$ denotes the probability simplex over all possible actions. The learner is provided with a budget constraint, allowing B arms to be pulled at each time step, i.e., $\sum_{m \in [M]} a_{t+1}^m = B$. After pulling B arms, the learner observes the next state $s_{t+1}^m \sim P_{a_{t+1}^m}^m \left(s_t^m, s_{t+1}^m\right)$, and receives a feedback $\boldsymbol{r}_{t+1}$. Suppose that the reward $\boldsymbol{r}_{t+1}$ is generated according to a distribution $\mathcal{P}_t$, we will assume stationarity of the *expected* reward received at time $t+1$, $\boldsymbol{r}_{t+1}$, when conditioned on history $\mathcal{H}_t^+$ and the current state.

Assumption 2.1. $\mathbb{E}\left[r_{t+1}^m \mid s_{t+1}^m, \mathcal{H}_t^+\right] = R^m\left(s_{t+1}^m\right).$

We permit non-stationarity in the rewards, provided that Assumption 2.1 is satisfied, offering a relaxation of the standard stationarity assumption typically imposed in literature.

The goal of the learner is to find the policy $\boldsymbol{\pi} := \{\pi_t\}_{t \in [T]}$ that maximizes the expected cumulative reward, where the expectation is further taken over the randomized policy ($\boldsymbol{a}_t \sim \pi_t$), in addition to the states ($s_{t+1}^m \sim P_{a_{t+1}^m}^m(s_t^m, s_{t+1}^m)$), subject to the full budget utilization constraint:[3]

$$\max_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{m=1}^M \sum_{t=0}^T R^m\left(s_t^m\right) \mid \boldsymbol{s}_0 \right]$$
$$\text{s.t.} \sum_{m=1}^M a_t^m = B, \ \forall t \in [T]. \tag{1}$$

We will use $\boldsymbol{\pi}^*$ to denote the optimal policy corresponding to Problem (1). Since $\boldsymbol{\pi}^*$ is intractable and unknown (Papadimitriou & Tsitsiklis, 1999), WIP (Whittle, 1988) and LP-based policies (Hong et al., 2024; Xiong et al., 2022a) are proposed as approximations. These policies are proven to be asymptotically optimal as $M \to \infty$ for infinite-horizon rewards with homogeneous agents, a restrictive setting that often does not occur in practice.

[3]Our model incorporates the standard full budget utilization constraint, consistent with the existing literature (Whittle, 1988; Zhang & Frazier, 2021; Hong et al., 2024).

2.2 Regret As Convergence Rate

A critical shift occurs in online learning: *the homogeneous agent assumption is often dropped to model more realistic, heterogeneous environments.* Under heterogeneity, these different approximation policies are no longer equivalent and will achieve different asymptotic performance levels. Consequently, comparing the regret magnitudes of different online algorithms is not meaningful when they target different reference policies (e.g., one converging to WIP versus another converging to an LP-based policy), as they are measuring convergence to fundamentally different benchmarks. This reality reframes the meaning of regret in RBs. It is not an absolute measure of sub-optimality, but a relative measure of the *convergence rate to a chosen reference policy* $\boldsymbol{\pi}^e$. We formalize this idea below:

Definition 2.2 (Regret as Convergence Rate)**.** Let $\boldsymbol{\pi}^e$ be the reference policy that an online algorithm ω aims to converge to. We define the regret of ω relative to this reference policy over horizon T as $\mathcal{R}_l^{\boldsymbol{\pi}^e}(\boldsymbol{\omega})$:[4]

$$\mathcal{R}_l^{\boldsymbol{\pi}^e}(\boldsymbol{\omega}) := \mathbb{E}_{\boldsymbol{\pi}^e} \left[\sum_{m=1}^M \sum_{t=0}^T l^m(s_t^m, a_{t+1}^m) \mid \boldsymbol{s}_0 \right] - \mathbb{E}_{\boldsymbol{\omega}} \left[\sum_{m=1}^M \sum_{t=0}^T l^m(s_t^m, a_{t+1}^m) \mid \boldsymbol{s}_0 \right],$$

where $l^m(s_t^m, a_{t+1}^m)$ serves as an instantaneous performance metric where the choice of l^m depends on the objective of $\boldsymbol{\pi}^e$, which we elaborate below. For heterogeneous agents with single-episode finite horizons, existing RB oracle policies are not optimal and hence not comparable. Thus, the regret $\mathcal{R}_l^{\boldsymbol{\pi}^e}(\boldsymbol{\omega})$ measures the *convergence rate of the learning algorithm ω to its corresponding oracle (or reference) policy* $\boldsymbol{\pi}^e$, where sublinear regret implies that ω converges to $\boldsymbol{\pi}^e$ within sublinear time.

The choice of the loss function l^m is directly linked to the objective of $\boldsymbol{\pi}^e$. For instance, in MAB, if $\boldsymbol{\pi}^e$ targets reward maximization, l is immediate reward (Auer et al., 2002); if $\boldsymbol{\pi}^e$ targets thresholding, l penalizes action on arms below threshold (Feng et al., 2024). Thus, l always measures how well online algorithms are achieving the corresponding objectives of $\boldsymbol{\pi}^e$.

Remark 2.3. The interpretation of regret as a convergence rate appears implicitly in prior work. For

[4]Appendix B.1 extends this definition to episodic settings, which is common in online RBs.

example, Wang et al. (2023b) use Def. 2.2 with $\boldsymbol{\pi}^e$ as WIP, while Xiong et al. (2022a) use it with $\boldsymbol{\pi}^e$ as the infinite-horizon average-reward optimal policy.

2.3 The Need for Indirect Reward Maximization in Online RB

Building on our reinterpretation of regret as convergence rate, we reconsider Problem (1)'s reward maximization objective. In finite-horizon settings, algorithms targeting alternative objectives with faster convergence to high-quality reference policies may achieve superior empirical performance. This approach addresses the fundamental limitation of direct methods for online RB—including those learning $\boldsymbol{\pi}^*$, WIP (Jung & Tewari, 2019; Wang et al., 2023b), or LP-based policies (Xiong et al., 2022a)—which typically suffer from slow convergence in practice.[5]

This slow convergence is well-characterized theoretically. For algorithms directly targeting reward maximization, Auer et al. (2008) establishes a $\Omega(T)$ lower bound when $\boldsymbol{\pi}^e = \boldsymbol{\pi}^*$ and $l^m = R^m$, while the tightest upper bound for learning WIP or LP-based policies remains $\mathcal{O}(T)$ (Wang et al., 2023b). This linear rate stems from the need to construct confidence intervals for *transition probabilities*, which requires extensive exploration of state-action pairs (s_t, a_t, s_{t+1}) to achieve sufficient visitation counts.

Given these fundamental constraints on estimating transition probabilities, we pursue an alternative approach: proposing an oracle (or reference) policy that enables faster convergence by circumventing the need for direct transition probability estimation. We develop this in Sec. 2.4, focusing on reference policies that avoid this bottleneck while maintaining strong *empirical performance*.

2.4 BT-CMAB Reformulation

Thus motivated, we reformulate restless bandit learning via two key modifications:

1. *Contextual bandit framework*: Transition probabilities need not be explicitly estimated. It embeds nonstationarity into rewards, yielding policies that depend only on the current context. This reduces sample complexity when compared with MDPs (Simchi-Levi & Xu, 2022).
2. *Thresholding objective*: At each step, we identify agents whose intervention benefit exceeds a threshold (vs. optimizing cumulative rewards), further lowering sample complexity (Feng et al., 2024) to achieve fast finite-horizon convergence.

Under suitable assumptions, our approach with known transitions/rewards recovers the optimal policy of *offline* RB for any M (Sec. 3), unlike WIP's asymptotic optimality requiring $M, T \to \infty$.

Regret-Minimizing BT-CMAB Reduction In the reduced contextual bandit problem, the context space is the state space. At time t, the reward ϕ^m that agent m receives at state s_t^m under action a_{t+1}^m is defined as the *expected reward* that the agent will receive in the next time period:

$$\begin{aligned}\phi^m(s_t^m, a_{t+1}^m) &:= \sum_{s_{t+1}^m \in \mathcal{S}} P_{a_{t+1}^m}^m(s_t^m, s_{t+1}^m) R^m\left(s_{t+1}^m\right) \\ &= \sum_{s_{t+1}^m \in \mathcal{S}} \mathbb{P}\left(s_{t+1}^m \mid \mathcal{H}_t^+\right) \mathbb{E}\left[r_{t+1}^m \mid s_{t+1}^m, \mathcal{H}_t^+\right] \\ &= \mathbb{E}\left[r_{t+1}^m \mid \mathcal{H}_t^+\right]. \end{aligned} \tag{2}$$

The second equality follows from conditional stationarity (Assump. 2.1) and the Markovian property. Eq. (2) shows that the immediate reward, $\phi^m(s_t^m, a_{t+1}^m)$, can be estimated without modeling the transitions, consistent with the standard bandit reward formulation.

As in online RBs, at each time t, the learner observes $\boldsymbol{s}_t$ and the corresponding reward $\boldsymbol{r}_t$. The learner then updates the estimated uncertainty associated with the estimated $\phi^m(s_{t-1}^m, a_t^m)$ $\forall m \in [M]$ according to Eq. (2), and uses it to guide the action $\boldsymbol{a}_{t+1}$, with $\sum_{m\in[M]} a_{t+1}^m = B$. The goal of the learner is to learn a policy that identifies and selects B agents whose expected gain from taking the action (i.e., $a_{t+1}^m = 1$) exceeds a predefined threshold, γ, in their current state s_t^m. To quantify this expected gain in state $s \in \mathcal{S}$, we define

$$I^m(s) := \phi^m(s, 1) - \phi^m(s, 0),$$

representing the *incremental reward* an agent receives by taking the action versus not at state s. At time t, we say that an agent m is in a good state if $I^m(s_t^m) \geq \gamma$.

At each time t, let $G_t := \{m \in [M] : I^m(s_t^m) \geq \gamma\}$ denote the set of agents currently in a good state. In this BT-CMAB formulation, the optimal policy, $\boldsymbol{\pi}^b$, involves randomly selecting B agents from G_t for intervention when $|G_t| \geq B$ and selecting B agents according to the ordering of incremental reward otherwise.

In our reformulation, we define regret (Def. 2.2) for any policy $\boldsymbol{\pi}$ by setting the reference policy $\boldsymbol{\pi}^e = \boldsymbol{\pi}^b$ and using the loss function $l^m(s_t^m, a_{t+1}^m) := -c_t^m(a_{t+1}^m)$, where the cost function is:

$$c_t^m\left(a_{t+1}^m\right) := (\gamma - I^m(s_t^m))\, \mathbb{I}\left\{I^m(s_t^m) < \gamma, a_{t+1}^m = 1\right\}. \tag{3}$$

[5] Although Xiong et al. (2022a) claim sublinear convergence for finite-horizon RB, their analysis relies on a generative model that permits restarting the MDP from arbitrary states—an assumption rarely satisfied in practice. Without this, their algorithm would incur linear convergence.

This cost function has an intuitive interpretation: it penalizes allocations to agents in "bad" states (where $I^m(s_t^m) < \gamma$), with the penalty proportional to how far the state index falls below the threshold γ. The baseline policy $\boldsymbol{\pi}^b$, by construction, never incurs these costs since it only allocates to agents in good states.

We simplify notation by writing $\mathcal{R}_c^{\boldsymbol{\pi}^b}(\boldsymbol{\pi})$ as $\mathcal{R}(\boldsymbol{\pi})$. This regret directly measures how frequently $\boldsymbol{\pi}$ deviates from $\boldsymbol{\pi}^b$ by allocating to suboptimal states, thus quantifying the convergence rate to this reference policy. The learner's objective is to minimize cumulative regret $\mathcal{R}(\boldsymbol{\pi})$ over horizon T, which equivalently maximizes the convergence rate to $\boldsymbol{\pi}^b$.

3 OPTIMALITY OF BT-CMAB REDUCTION

Before establishing the sublinear convergence rate of our proposed algorithm for BT-CMAB with *heterogeneous* agent, we first establish the optimality of our BT-CMAB formulation for 2-state homogeneous agents—the first non-asymptotic optimality result for finite-horizon RB. This demonstrates that the contextual bandit provides an exact reformulation of the offline RB problem under these conditions. *We note that Sections 4 and 5 lift these assumptions.*

We assume *binary states* and we further assume that all agents share a good state with higher R^m. Without loss of generality (WLOG), we will use 1 to denote the good state and 0 for the bad state. We first describe the greedy policy. WLOG, suppose that at each time step t, the incremental rewards are ordered decreasingly as $I^1\left(s_t^1\right) \geq I^2\left(s_t^2\right) \geq \cdots \geq I^M\left(s_t^M\right)$. The corresponding greedy policy $\pi_t^{\text{greedy}} : \mathcal{S} \to \{0,1\}^M$ is to give actions to the first B agents according to the order. Formally, with a slight abuse of notation, let $\pi_t^{\text{greedy}}(s_t^i)$ be the ith entry of π_t^{greedy}, defined as:

$$\pi_t^{\text{greedy}}(s_t^i) = \begin{cases} 1 & i \leq B \\ 0 & i > B \end{cases}.$$

Let $\boldsymbol{\pi}^{\text{greedy}} := \{\pi_t^{\text{greedy}}\}_{t\in[T]}$. We note that $\boldsymbol{\pi}^{\text{greedy}} = \boldsymbol{\pi}^b$ with suitable choices of γ. A trivial choice of γ would be $\max_{m,s} I^m(s)$. Now we are ready to show that the greedy policy is optimal for any M.

Theorem 3.1. *Assuming all agents are homogeneous, i.e.,* $P^m = P^{m'}, R^m = R^{m'}, \forall m, m' \in [M]$ *and* $\mathcal{S} = \{0,1\}$, $\boldsymbol{\pi}^{greedy} = \boldsymbol{\pi}^*$ *for any* $0 \leq B \leq M < \infty$.

Theorem 3.1 establishes that our reference policy $\boldsymbol{\pi}^b$—which corresponds to the greedy policy $\boldsymbol{\pi}^{\text{greedy}}$ with appropriate threshold selection—achieves non-asymptotic optimality for homogeneous, 2-state finite-horizon restless bandits. This provides stronger guarantees than prior reference policies like WIP or LP-based methods (Hong et al., 2024; Whittle, 1988; Zhang & Frazier, 2021), which only offer asymptotic optimality for infinite-horizon settings under homogeneous agents.

While our analysis builds on the practical two-state assumption commonly used in recent work (Liang et al., 2024; Killian et al., 2023; Wang et al., 2023a), proving Theorem 3.1 required developing novel analytical techniques. Existing tools are designed for analyzing value function-based policies (Eq. (E.11)) in infinite-horizon settings (Whittle, 1988; Hong et al., 2024), making them unsuitable for our problem setup.

Remark 3.2 (Difficulty of Relaxation)**.** Theorem 3.1 does not extend to heterogeneous agents (Appendix C) or multi-state systems (Appendix D). Our proof technique—which reduces Bellman optimality to conditions on value functions—relies critically on the two-state property that transition dynamics can be fully characterized by the probability of transitioning to state 1. This characterization breaks down in multi-state systems, where transitions require tracking probabilities to multiple states. While we conjecture that appropriately designed greedy policies may achieve constant-factor approximations even in these more general settings, such extensions are beyond the scope of this paper.

Although $\boldsymbol{\pi}^b$ is only guaranteed to be optimal for homogeneous agents and a high threshold, our BT-CMAB reduction is valid for heterogeneous agents and an arbitrary threshold γ. As mentioned in Section 2.4, the threshold is associated with fast convergence. We will show in Section 4.1 that a higher threshold will lead to slower convergence for heterogeneous agents. The detailed discussion on how to choose the threshold is deferred to Remark 4.5.

Theorem 3.1 Proof Sketch The proof consists of two steps: 1) We reformulate the problem as an MDP where the state space is the Cartesian product of the individual state space, i.e., $\mathcal{S}^M$; the action space is the subset of $\{0,1\}^M$ that satisfies the budget constraint, and the transition matrix is the product of P^m as all agents are independent. 2) We show that the proposed greedy policy π_t^{greedy} satisfies the Bellman equation (Eq. (E.21)) for all $t \leq T$ by using backward induction. We prove the induction hypothesis by carefully controlling the difference of the value function (Eq. (E.11)).

4 LEARNING IN BT-CMAB

We now present our learning algorithm (Algorithm 1) for BT-CMAB (as reformulated in Sec. 2.4) in a general setting, and establish that it achieves sublinear regret for this new problem. The sublinear regret bound explicitly characterizes the algorithm's convergence rate, demonstrating its fast convergence in finite sample

when compared with those of existing restless bandit algorithms. We note that *the results hold without agent homogeneity or a binary state space.*

Learning in BT-CMAB framework is challenging because it requires estimating the incremental reward—the outcome difference between acting and not acting—which necessitates observing both actions for each agent-state pair (paired exploration). However, to minimize regret, the policy must also exploit this knowledge by taking the optimal action (which could be either 0 or 1, depending on the state and agent). The tension between these two requirements—thorough exploration for accurate estimation and greedy exploitation for regret minimization—defines the core difficulty of the problem.

ϵ-Greedy Thresholding Algorithm To resolve this, we propose an *ϵ-Greedy Thresholding algorithm*, Algorithm 1, that explicitly balances this trade-off. By using a decaying exploration rate ϵ_t, it ensures we sample both actions for each state-agent pair enough for learning while progressively prioritizing the optimal action to minimize regret, as formalized in our proof.

At each time step t, Algorithm 1 proceeds as follows: With probability ϵ_t, the algorithm explores by uniformly selecting B agents to act upon, regardless of their state or estimated reward. The exploration rate ϵ_t decays with the inverse of the minimum time an agent has spent in its current state, encouraging more exploration in under-sampled contexts. With probability $1-\epsilon_t$, the algorithm exploits its knowledge: it computes the empirical mean incremental reward for each agent in its current state, forms a candidate set $\widetilde{G}_{t+1}$ of agents exceeding the threshold γ. The algorithm then proceeds as follows: if $|\widetilde{G}_{t+1}| \geq B$, it gives actions to the top-B agents within $\widetilde{G}_{t+1}$ (e.g., those with the highest empirical means); if $|\widetilde{G}_{t+1}| < B$, it gives actions to all agents in $\widetilde{G}_{t+1}$ and then allocates the remaining $B-|\widetilde{G}_{t+1}|$ actions by selecting the agents with the highest *upper confidence bound* (UCB) on the incremental reward from the rest agents. After the actions are taken, the algorithm observes the new state vector $\boldsymbol{s}_{t+1}$ and the corresponding reward vector $\boldsymbol{r}_{t+1}$.

Next, the algorithm updates its internal estimates using the newly observed rewards and states. For each agent, we update the cumulative rewards and visitation counts for the state-action pairs that were just observed. These are used to compute a new empirical mean for the incremental reward, $\widehat{I}^m_{t+1}(s)$, for each agent-state pair. Finally, the upper confidence bound is recalculated. The new $\text{UCB}^m_{t+1}(s)$ for the next round adds an optimism bonus to the empirical incremental reward. This bonus is derived from an anytime-valid confidence bound (Howard et al., 2021) and is inversely proportional to the square root of the state-action visit counts, ensuring that estimates for less-explored states have higher uncertainty. The specific function used is $f(t,T) = \log\log(2t + 0.72\log(10.4T))$.

As we address a thresholding contextual bandit with MDP-governed contexts instead of MAB, Alg. 1 differs from Auer et al. (2002). We tailor ϵ_t to decay with the minimum state occupancy time rather than the current timestep, ensuring sufficient exploration across states.

4.1 Theoretical Guarantee

We establish that Algorithm 1 achieves a $\mathcal{O}(\sqrt{T})$ regret upper bound under two assumptions. Assump. 4.1 ensures strictly positive transition probabilities (even if arbitrarily small), while Assump 4.2 guarantees the existence of a policy with sufficient good agent-state pairs at each timestep.

Assumption 4.1 (Positive Transition)**.** $P^m_a(s,s') > 0$ for all $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, and $m \in [M]$.

Assump. 4.1 guarantees that any state $\boldsymbol{s}' \in \mathcal{S}^M$ is reachable from any given state $\boldsymbol{s}$ in a single step. Importantly, we only require these one-step transition probabilities to be positive; we do not assume they are bounded away from zero, and our result *does not* depend on the minimum transition probability.

Assumption 4.2 (Exists A Good Policy)**.** There exists a policy achieving zero cumulative cost.

Assump. 4.2 serves primarily to simplify our theoretical analysis by ensuring problem instances are well-behaved. This condition guarantees that an oracle policy exists which can spend its full budget of B actions per round without being forced to assign actions with negative expected returns. In practice, this assumption can be satisfied for any problem instance by carefully choosing the threshold parameter γ; for example, setting $\gamma = -\infty$ guarantees the condition is met trivially. Thus, γ serves as a *tunable parameter* to be optimized for overall performance, a point we discuss further in Remark 4.5.

When combined with Assump. 4.1, Assump. 4.2 yields a stronger guarantee (Lemma 4.3) that is inherently assumed by any algorithm required to spend the full budget B at each round. This is common in prior RB algorithms (Wang et al., 2020, 2023b).

Let $\mathcal{G} := \{(m,s) : m \in [M], s \in \mathcal{S}, I^m(s) \geq \gamma\}$ be the set of good agent-state pairs. Lemma 4.3 ensures that for any state realization $\boldsymbol{s} \in \mathcal{S}^M$, there are always at least B good agent-state pairs.

Lemma 4.3 (Existence of Sufficient Good Agent-State Pairs)**.** *Under Assumptions 4.2 and 4.1, for any joint state $\boldsymbol{s} \in \mathcal{S}^M$, the number of good agent-*

Algorithm 1 ϵ-Greedy Thresholding (ϵ-GT)

Input: initial state vector $\boldsymbol{s}$, round budget B, exploration function $f(t)$, threshold γ, instance parameter η, scaling constant D, exploration decay rate p
Initialize: initial states $\boldsymbol{s}_0 \leftarrow \boldsymbol{s}$; $\forall (s,a,m) \in (\mathcal{S} \times \mathcal{A} \times [M])$: agent state visit count $N^{(0)}_{s,m} \leftarrow 0$, agent state-action visit count $N^{(0)}_{s,m,a} \leftarrow 0$, incremental reward $\widehat{I}^m_0(s) \leftarrow 0$, cumulative immediate rewards $\hat{\phi}^m(s,a) \leftarrow 0$, UCB of incremental reward $\mathrm{UCB}^m_0(s) \leftarrow \infty$
for $t = 0, 1, \cdots, T$ **do**
  Retrieve $\boldsymbol{s}_t$; identify good agents above the threshold: $\widetilde{G}_{t+1} \leftarrow \left\{ m \in [M] \colon \widehat{I}^m_t(s^m_t) \geq \gamma \right\}$.
  Initialize the candidate set $C \leftarrow \widetilde{G}_{t+1}$;
  Sort the remaining agents: $[M] \setminus \widetilde{G}_{t+1}$, in descending order of $\mathrm{UCB}^m_t(s^m_t)$; append top $\max(0, B - |\widetilde{G}_{t+1}|)$ agents from sorted list to C.
  $\widetilde{C} \leftarrow B$ agents in C with highest $\widehat{I}^m_t(s^m_t)$.
  With probability $\epsilon_t = \min \left\{ 1, \min_m \frac{D}{\left(N^{(t)}_{s^m_t,m}+1\right)^{p} \min\{\eta^2, 1/2\}} \right\}$:
    Random exploration: uniformly choose B agents from $[M]$ and set their $a^m_{t+1} \leftarrow 1$.
  Otherwise greedy: set $a^m_{t+1} = 1$ for $m \in \widetilde{C}$, 0 otherwise
  Observe $(\boldsymbol{s}_{t+1}, \boldsymbol{r}_{t+1})$
  For each agent-state pair $(m,s) \in [M] \times \mathcal{S}$, update the UCB of the incremental rewards as:
    $\forall a \in \mathcal{A} : \hat{\phi}^m(s^m_t, a) = \hat{\phi}^m(s^m_t, a) + r^m_{t+1} \mathbb{I}\left\{a^m_{t+1} = a\right\}$.
    $\forall a \in \mathcal{A} : N^{(t+1)}_{s,m,a} = N^{(t)}_{s,m,a} + \mathbb{I}\left\{a^m_{t+1} = a\right\} \mathbb{I}\left\{s = s^m_t\right\}$
    $N^{(t+1)}_{s,m} = N^{(t)}_{s,m} + \mathbb{I}\{s = s^m_t\}$
    $\widehat{I}^m_{t+1}(s) = \frac{\hat{\phi}^m(s,1)}{N^{(t+1)}_{s,m,1}} - \frac{\hat{\phi}^m(s,0)}{N^{(t+1)}_{s,m,0}}$
    $\mathrm{UCB}^m_{t+1}(s) = \widehat{I}^m_{t+1}(s) + \sqrt{\frac{f\left(N^{(t+1)}_{s,m,1}, T\right)}{N^{(t+1)}_{s,m,1}+2}} + \sqrt{\frac{f\left(N^{(t+1)}_{s,m,0}, T\right)}{N^{(t+1)}_{s,m,0}+2}}$
end for

state pairs observed in $\boldsymbol{s}$ is at least B. That is, $|\{m \in [M] : (m, s^m) \in \mathcal{G}\}| \geq B$.

Despite its strong guarantee, Lemma 4.3 (proof in Appendix F.1) is derived from two relatively mild assumptions. This result is critical for two reasons. First, it ensures that any algorithm required to spend a full budget of B actions is never forced to assign actions whose incremental reward is below the threshold. Second, it guarantees that while choosing a suboptimal agent may affect future state distributions, this effect does not accumulate: sufficient good agents are always available. Thus, a policy achieving zero per-step regret is attainable, which forms the foundation for our algorithm's performance.

Let $\Delta^m_s := |\gamma - I^m(s)|$. Define $(m_{\min}, s_{\min})$ as the minimizer of $I^m(s) - \gamma$ over $(m,s) \in \mathcal{G}$, and set $\Delta := \Delta^{m_{\min}}_{s_{\min}}$. Next, we present our main theorem on the regret upper bound, which characterizes the finite-horizon convergence rate of Algorithm 1:

Theorem 4.4 (Regret Upperbound)**.** *Assume r^m_t is 1-subgaussian random variable for all $t \in [T]$. Let $D := \max\left\{\frac{8M}{B}, \frac{8M}{M-B}, 2\right\}, p = 1, \eta = \Delta$. Under Assump.s 4.2 and 4.1, the regret, $\mathcal{R}(\pi)$, defined by Def. 2.2 with the loss function l^m given by Eq. (3), satisfies:*

$$\mathcal{R}(\pi) = \min\left\{ \mathcal{O}\left(\frac{|\mathcal{S}|\,(M|\mathcal{S}| - |\mathcal{G}|)\,MD}{\Delta} \log T \right), \mathcal{O}\left(\sqrt{MDB\big(|\mathcal{S}|\,(M|\mathcal{S}| - |\mathcal{G}|)\big)\,T \log T} \right) \right\}. \tag{4}$$

Eq. (4) achieves a sublinear convergence rate to $\boldsymbol{\pi}^b$, contrasting the linear rate to WIP in Wang et al. (2023b). This bound is independent of γ and the smallest transition probability, with the constant formalized in Appendix F. Although Theorem 4.4 sets $\eta = \Delta$, empirical results show algorithm stability across η values in terms of average cumulative reward (Appendix G.3).

Theorem 4.4 Proof Sketch Our analysis deviates from standard CMAB analysis because estimating the incremental reward requires pulling suboptimal arms; consequently, we must *lower* bound the number of times suboptimal arms are pulled, in addition to the usual upper bound. The argument proceeds in three steps. 1) Regret equivalence: we show that regret is equivalent to the expected number of interventions given to bad agents. 2) Bounding the interventions: we prove that this expected number can be bounded either by the

overestimation of the incremental reward of bad agents or by the underestimation of that of good agents. We further establish that at each time step, both the number of interventions and non-interventions assigned to agents are lower bounded. 3) Concentration: we apply concentration inequality (Lemma F.2) to obtain the desired result. Importantly, stationarity of the reward is not required here, since the concentration inequality we employ holds without such an assumption.

Remark 4.5 (How to Select the Threshold γ?)**.** In our problem, there is an inherent trade-off between satisfying Assump. 4.2 and maximizing cumulative reward. While our reformulation focuses on the threshold identification problem, maximizing cumulative reward remains our ultimate goal. Specifically, a smaller threshold γ makes it easier to satisfy Assump. 4.2, as more state-agent pairs qualify as good. However, a low γ also makes any arm above the threshold optimal, potentially leading to lower rewards. Empirically, we observe that there exists a universal choice of γ which satisfies the Assump. 4.2 that achieves the highest reward as shown in Appendix G.3.

5 EXPERIMENTS

We numerically evaluate the performance of our algorithm, showing: 1) sublinear regret under the BT-CMAB reformulation (results in Appendix G.1), and 2) the achieved average cumulative reward, as defined in the finite-horizon RB setting, outperforms that of existing online RB algorithms. These results underscore the practicality of the proposed reformulation and algorithms in finite-horizon restless bandit problems.

Benchmarks We compare the average cumulative reward obtained by our algorithm with the following benchmarks in online RB settings: WIP based: 1) WIQL (Biswas et al., 2021), 2) UCWhittlePv (Wang et al., 2023b); LP-based: 3) R(MA)^{2}B (Xiong et al., 2022a); and 4) *Random Policy* (uniformly pick B agents), which follows the experimental setting of Wang et al. (2023b); Xiong et al. (2022a).[6]

With heterogeneous agents, no algorithm has been established as optimal. We incorporate three benchmark algorithms with full MDP knowledge: 5) *Oracle Whittle* (Whittle, 1988), 6) *Oracle Greedy* and an LP-based policy 7) *OMR* (Xiong et al., 2022a). Baseline details and implementation are in Appendix G.2.

Setup We set $M = 50, \mathcal{S} = \{0, 1\}$, and compare the results under three different budget settings: $B = 5, 10, 20$. We use: 1) **Synthetic Instances**: The expected reward is set to be $R^m(s) = s$ and randomly generate 50 instances with different transition matrices, 2) **Wireless Communication Dataset (Prieto-Cerdeira et al., 2010)**: We use the same wireless communication dataset as Wang et al. (2020). Details will be deferred to Section 5.2. We also consider the setting where $\mathcal{S} = [10]$ and randomly generate 50 instances. The results are deferred to Appendix G.2.

In Figure 1, each line is averaged over 50 instances, each with 13 repetitions. We consider two reward settings: 1) a noiseless case, and 2) a noisy setting where the rewards at state $s \in \mathcal{S}$ follow a normal distribution with mean $R^m(s)$ and variance 1 for the synthetic dataset and 0.1 for the real-world dataset. We note that we did not fine-tune our hyperparameters. Specifically, we universally choose a threshold that satisfies Assump. 4.2. We have also shown that our algorithm is robust across different η and we have done ablations on the choice of threshold (Appendix G.3).

5.1 Synthetic Instances

Results Under Noiseless Rewards In Figure 1 (top row), we observe that Alg. 1 consistently outperforms online RB algorithms within a horizon of 1000 in all budget settings. As shown in Figure 1, Alg. 1 quickly converges to a stable yet suboptimal policy, whereas WIQL converges more slowly. However, given a sufficiently long horizon, WIQL will eventually converge to the *Oracle Whittle* policy. Nevertheless, Alg. 1 outperforms WIQL when the horizon is short, even when the budget is low, highlighting the practicality of our algorithm. UCWhittlePv requires extensive exploration to converge, and R(MA)^{2}B performs similarly to random policy.

Results Under Noisy Rewards In Figure 1 (bottom row), we excluded UCWhittlePv and R(MA)^{2}B since they underperformed WIQL in the noiseless setting. In all three budget settings, Alg. 1 consistently outperforms WIQL.

Lastly, we compare *Oracle Whittle*, *Oracle Greedy*, and *OMR*. *Oracle Whittle* consistently outperforms *Oracle Greedy* across all budgets, with the gap narrowing as budget increases. Despite this, Alg. 1 surpasses all online benchmarks, demonstrating the benefits of rapid convergence. *OMR* performed worst, likely because its LP-based construction leverages homogeneity and falters in heterogeneous settings.

5.2 Wireless Communication Dataset

This problem (Prieto-Cerdeira et al., 2010) involves digital video broadcasting satellite services to handheld devices via land mobile satellite. In this problem, the

[6] Our results differ from those of Wang et al. (2023b) due to detected bugs in their code; a detailed discussion is provided in Appendix G.2.

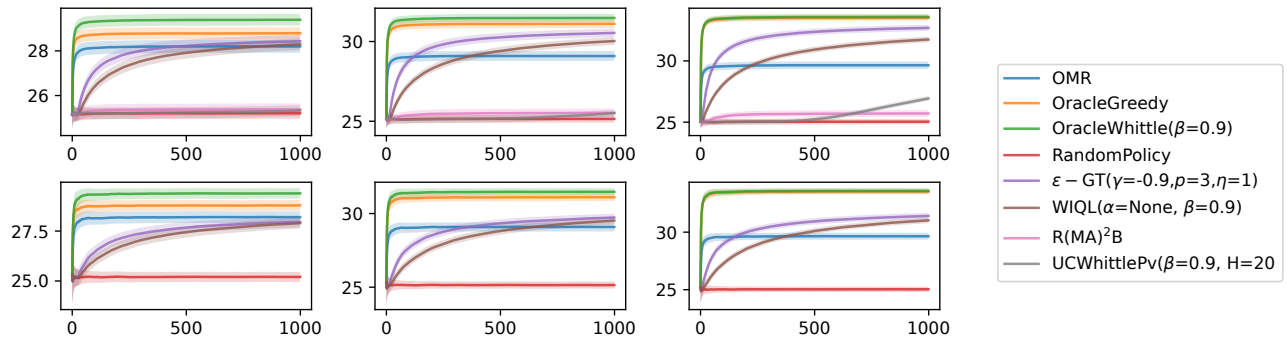

Figure 1: Average (over time) cumulative reward for budgets $B = 5$ (left), 10 (middle), 20 (right), averaged over 50 instances with 13 repetitions each. Top: noiseless reward; bottom: noisy reward. Exclude R(MA)^2)B and UCWhittlePV in the noisy setting since they require known reward.

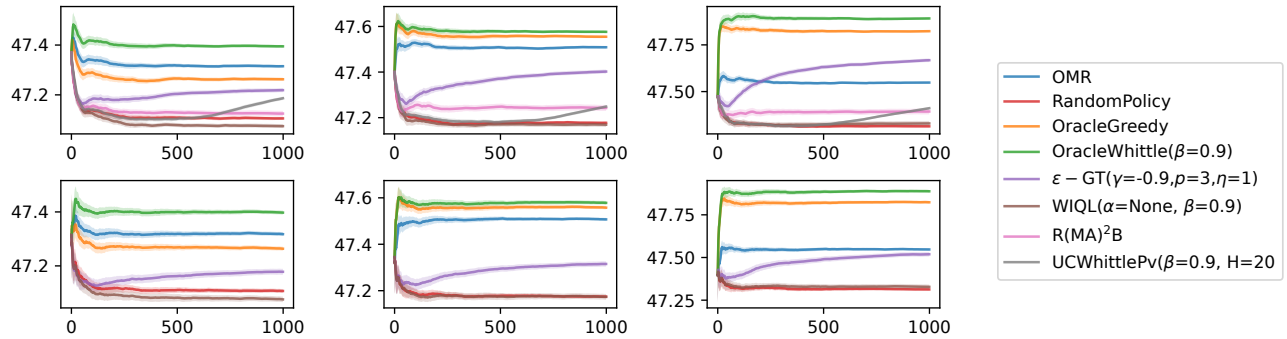

Figure 2: Average (over time) cumulative reward for budgets $B = 5$ (left), 10 (middle), 20 (right), averaged over 100 repetitions of the real-world dataset. Top: noiseless reward; bottom: noisy reward. Exclude R(MA)^2)B and UCWhittlePV in a noisy setting since they require known reward.

environment is modeled as a two-state system ("good" and "bad"), with five different transition models corresponding to different satellite locations. In their original problem, each agent is provided with multiple actions which is modeled as the angle of the satellite. Similar to Wang et al. (2020), we choose two angles as the action: 40 degrees as action 0 and 60 degrees as action 1. The transition probability is chosen according to Table IV in Prieto-Cerdeira et al. (2010). The original reward is dependent on both action and state, but the reward difference between different actions are low (Table III in Prieto-Cerdeira et al. (2010)). Further, the reward difference between the good state and bad state is also small. Therefore, we reconstruct the reward and set the reward for good state as 1 and bad state as 0.9 to mimic the real-world situations where signal is weak. We randomly assign 50 agents with one of the five transition probabilities.

In Figure 2, we observe that in this problem, Our method consistently outperforms all benchmarks within a horizon of 1000 in all budget settings under noiseless and noisy rewards. As shown in Figure 2, Alg. 1 converges to its oracle policy (*OracleGreedy*) quickly while other online RB algorithms converge to their oracle policies more slowly. This translates to the superior performance of Alg. 1, demonstrating our observation that quick convergence translates to better finite sample performance.

Discussions We establish greedy policy optimality in finite-horizon RBs and reformulate online RB as a BT-CMAB with a learning algorithm. Our approach leverages the insight that fast convergence to a sub-optimal policy can outperform slow convergence to a better one in finite horizons. As the BT-CMAB formulation is novel, our work appropriately focuses on establishing its theoretical and empirical value. Deriving a problem-specific lower bound is a natural and important direction for future work. Further, we plan to analyze the optimality of the oracle policy in multi-state setting.

References

Akbarzadeh, N. and Mahajan, A. On learning whittle index policy for restless bandits with scalable regret. *IEEE Transactions on Control of Network Systems*, 2023.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pp. 263–272. PMLR, 2017.

Biswas, A., Aggarwal, G., Varakantham, P., and Tambe, M. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. *arXiv preprint arXiv:2105.07965*, 2021.

Borkar, V. S., Kasbekar, G. S., Pattathil, S., and Shetty, P. Y. Opportunistic scheduling as restless bandits. *IEEE Transactions on Control of Network Systems*, 5:1952–1961, 2017.

Brown, D. B. and Smith, J. E. Index policies and performance bounds for dynamic selection problems. *Management Science*, 66:3029–3050, 2020.

Brown, D. B. and Zhang, J. Fluid policies, reoptimization, and performance guarantees in dynamic resource allocation. *Operations Research*, 2023.

del Río, A. F., Leong, M. B., Saraiva, P., Nazarov, I., Rastogi, A., Hassan, M., Tang, D., and África Periáñez. Adaptive user journeys in pharma e-commerce with reinforcement learning: Insights from swiperx, 2024a.

del Río, A. F., Leong, M. B., Saraiva, P., Nazarov, I., Rastogi, A., Hassan, M., Tang, D., and África Periáñez. Adaptive behavioral ai: Reinforcement learning to enhance pharmacy services, 2024b.

Feng, Q., Ma, T., and Zhu, R. Satisficing exploration in bandit optimization. *arXiv preprint arXiv:2406.06802*, 2024.

Gast, N., Gaujal, B., and Yan, C. Linear program-based policies for restless bandits: Necessary and sufficient conditions for (exponentially fast) asymptotic optimality. *Mathematics of Operations Research*, 49(4):2468–2491, 2024.

Hajiabolhassan, H. and Ortner, R. Online regret bounds for satisficing in mdps. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

Hong, Y., Xie, Q., Chen, Y., and Wang, W. Achieving exponential asymptotic optimality in average-reward restless bandits without global attractor assumption, 2024.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2): 1055 – 1080, 2021.

Hsu, Y.-P. Age of information: Whittle index for scheduling stochastic arrivals. pp. 2634–2638. IEEE Press, 2018.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Jung, Y. H. and Tewari, A. Regret bounds for thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems*, 32, 2019.

Kadota, I., Uysal-Biyikoglu, E., Singh, R., and Modiano, E. Minimizing the age of information in broadcast wireless networks. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 844–851, 2016.

Kakarapalli, P., Kayande, D., and Meshram, R. Faster q-learning algorithms for restless bandits. *arXiv preprint arXiv:2409.05908*, 2024.

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

Killian, J. A., Biswas, A., Xu, L., Verma, S., Nair, V., Taneja, A., Hegde, A., Madhiwalla, N., Diaz, P. R., Johnson-Yu, S., et al. Robust planning over restless groups: engagement interventions for a large-scale maternal telehealth program. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14295–14303, 2023.

Liang, B., Xu, L., Taneja, A., Tambe, M., and Janson, L. A bayesian approach to online learning for contextual restless bandits with applications to public health. *arXiv preprint arXiv:2402.04933*, 2024.

Mate, A., Madaan, L., Taneja, A., Madhiwalla, N., Verma, S., Singh, G., Hegde, A., Varakantham, P., and Tambe, M. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health, 2021.

Michel, T., Hajiabolhassan, H., and Ortner, R. Regret bounds for satisficing in multi-armed bandit problems. *Transactions on Machine Learning Research*, 2022.

Ortner, R., Ryabko, D., Auer, P., and Munos, R. Regret bounds for restless markov bandits. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, ALT'12, pp. 214–228, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 9783642341052.

Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999. ISSN 0364765X, 15265471.

Prieto-Cerdeira, R., Perez-Fontan, F., Burzigotti, P., Bolea-Alamañac, A., and Sanchez-Lago, I. Versatile two-state land mobile satellite channel model with first application to dvb-sh analysis. *International Journal of Satellite Communications and Networking*, 28(5-6):291–315, 2010.

Simchi-Levi, D. and Wang, C. Multi-armed bandit experimental design: Online decision-making and adaptive inference. *Management Science*, 71(6):4828–4846, 2025. doi: 10.1287/mnsc.2023.00492.

Simchi-Levi, D. and Xu, Y. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47:1904–1931, 2022.

Tamatsukuri, A. and Takahashi, T. Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function. *Biosystems*, 180:46–53, 2019.

Verloop, I. M. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. 2016.

Villar, S. S., Bowden, J., and Wason, J. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30: 199, 2015.

Wang, K., Verma, S., Mate, A., Shah, S., Taneja, A., Madhiwalla, N., Hegde, A., and Tambe, M. Scalable decision-focused learning in restless multi-armed bandits with application to maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12138–12146, 2023a.

Wang, K., Xu, L., Taneja, A., and Tambe, M. Optimistic whittle index policy: Online learning for restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10131–10139, 2023b.

Wang, S., Huang, L., and Lui, J. Restless-ucb, an efficient and low-complexity algorithm for online restless bandits. *Advances in Neural Information Processing Systems*, 33:11878–11889, 2020.

Weber, R. R. and Weiss, G. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3): 637–648, 1990.

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

Xiong, G. and Li, J. Finite-time analysis of whittle index based q-learning for restless multi-armed bandits with neural network function approximation. *Advances in Neural Information Processing Systems*, 36:29048–29073, 2023.

Xiong, G., Li, J., and Singh, R. Reinforcement learning augmented asymptotically optimal index policy for finite-horizon restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8726–8734, 2022a.

Xiong, G., Wang, S., and Li, J. Learning infinite-horizon average-reward restless multi-action bandits via index awareness. *Advances in Neural Information Processing Systems*, 35:17911–17925, 2022b.

Xiong, G., Wang, S., Li, J., and Singh, R. Whittle index-based q-learning for wireless edge caching with linear function approximation. *IEEE/ACM Transactions on Networking*, 2024.

Zhang, X. and Frazier, P. I. Restless bandits with many arms: Beating the central limit theorem. *arXiv preprint arXiv:2107.11911*, 2021.

Zhang, Z. and Wang, Z. Online experimental design with estimation-regret trade-off under network interference. *arXiv preprint arXiv:2412.03727*, 2024.

Zhong, Z., Cheung, W. C., and Tan, V. Y. Achieving the pareto frontier of regret minimization and best arm identification in multi-armed bandits. *arXiv preprint arXiv:2110.08627*, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A REAL-WORLD RESTLESS BANDIT EXAMPLES

Table A.1: Examples of good and bad states and corresponding budgeted interventions in various domains

Domain	Good state	Bad state	Budgeted Intervention Example
General Usage	Frequent logins, feature exploration, meaningful use	Rare/brief logins, session abandonment	Reach out, offer help, gather feedback
E-commerce	High purchase frequency, browsing multiple categories	Low/no purchases, high cart abandonment	Provide a personalized discount or free shipping
E-learning	Completes modules, passes quizzes, revisits content	Seldom opens modules, fails quizzes	Progress-based reward (finish by deadline for free certificate)
Loyalty	Completes surveys, redeems rewards, advances loyalty	Not collecting points, ignoring surveys, no rewards	Special bonus for feedback (earn points by reviewing new features)
Patient management	Updates records, completes follow-ups, tracks adherence	Rarely logs updates, misses assessments or follow-ups	Points or bonus for timely follow-ups

We will further use an example to show case how the state, action, reward and budget is defined in practice. SwipeRX is an online platform in Indonesia where pharmacists typically place orders on Mondays. The platform aims to increase engagement and revenue by providing targeted coupons to pharmacists. Since pharmacists operate in different local markets with varying demand patterns and competition from other platforms, we model them as heterogeneous agents, each represented by an MDP.

- **State**: The pharmacist's engagement level with the platform
- **Action**: Binary decision of whether to provide a coupon to a given pharmacist
- **Action-dependent State Transitions**: Based on whether an action is taken or not, the agent's engagement evolves accordingly based on environmental factors and platform interactions
- **Reward**: Engagement metrics (e.g., time spent on platform, order frequency)
- **Budget**: The platform has a budget on the number of coupons available at each day.

B WHITTLE'S INDEX

Whittle's index is computed for each agent. At each time t, given $\boldsymbol{s}_t$, the top B agents with the highest indices are selected. Whittle's index policy has two variants, each aimed at maximizing a different objective in an *infinite* horizon setting: one for maximizing average reward and the other for maximizing discounted reward.

Average Reward: In this case, Whittle's index policy is designed to approximately find the policy $\boldsymbol{\pi}_d$ that maximizes the average reward,

$$\max_{\boldsymbol{\pi}_d} \liminf_{T\to\infty} \frac{1}{TM}\mathbb{E}\left[\sum_{m=1}^{M}\sum_{t=0}^{T} R^m\left(s_t^m\right) \mid \boldsymbol{s}_0\right]$$

$$\text{s.t.} \sum_{m=1}^{M} a_t^m = B, \forall t \in [T]. \tag{B.5}$$

We will denote J^* to be the optimal value of Problem B.5. Instead of directly solving Problem B.5, Whittle deals with the Lagrangian relaxation of Problem B.5, which is stated in Problem B.6

$$\max_{\boldsymbol{\pi}_d} \liminf_{T\to\infty} \frac{1}{TM}\mathbb{E}\left[\sum_{m=1}^{M}\sum_{t=0}^{T} R^m\left(s_t^m\right) + \lambda\left(1 - a_{t+1}^m\right) \mid \boldsymbol{s}_0\right]. \tag{B.6}$$

The key observation to derive Whittle's index policy is that Problem B.6 can be solved by addressing the following dynamic programming separately for each agent

$$\begin{aligned} V_\lambda^m(s) &= \max_{a\in\{0,1\}} Q^m(s,a) \\ Q_\lambda^m(s,a) + \beta_m^* &= R^m(s) - \lambda\mathbb{I}\{a=1\} \\ &\quad + \sum_{s'\in\mathcal{S}} P_a^m(s,s')V^m(s'), \end{aligned}$$

where β_m^* is the maximized average reward for arm m without any constraint, i.e.

$$\beta_m^* := \max \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T} R^m\left(s_t^m\right) - \lambda\mathbb{I}\{a=1\} \mid s_0^m\right].$$

The Whittle's index λ_s^m is defined as the multiplier λ_s^m when it is equally favorable between choosing action 1 and action 0, i.e. $Q_{\lambda_s^m}^m(s,1) = Q_{\lambda_s^m}^m(s,0)$. The Whittle's index policy would then give action to B agents with the highest Whittle's index at each time step t. Denote this as $\boldsymbol{\pi}^{\text{Whittle}}$ and the average reward of Whittle's index policy as J^{Whittle}. The performance of Whittle's index policy is characterized in Theorem B.1.

Theorem B.1 (Theorem 2, Weber & Weiss 1990)**.** *Suppose all arms are homogeneous and assumptions therein, i.e., $P_a^m = P_a^{m'}, R^m = R^{m'}, \forall m \neq m'$, then the Whittle's index policy is asymptotically optimal,*

$$\lim_{M\to\infty} J^{Whittle} = \lim_{M\to\infty} J^*.$$

Discounted Reward: In this case, Whittle's index policy is designed to approximately find the policy $\boldsymbol{\pi}_d$ that maximizes the discounted reward for a fixed discount factor β,

$$\begin{aligned} &\max_{\boldsymbol{\pi}_d} \sum_{t=0}^{\infty} \frac{1}{M}\mathbb{E}\left[\sum_{m=1}^{M} \beta^t R^m\left(s_t^m\right) \mid \boldsymbol{s}_0\right] \\ &\text{s.t.} \sum_{m=1}^{M} a_t^m = B, \forall t \in [T]. \end{aligned} \tag{B.7}$$

Similarly, instead of directly solving Problem B.7, Whittle deals with the Lagrangian relaxation of Problem B.7, which is stated in Problem B.8

$$\max_{\boldsymbol{\pi}_d} \sum_{t=0}^{\infty} \frac{1}{M}\mathbb{E}\left[\sum_{m=1}^{M} \beta^t\left(R^m\left(s_t^m\right) + \lambda\left(1 - a_{t+1}^m\right)\right) \mid \boldsymbol{s}_0\right]. \tag{B.8}$$

Again, the same key observation is that Problem B.8 can be solved by addressing the following dynamic programming separately for each agent

$$\begin{aligned} V_\lambda^m(s) &= \max_{a\in\{0,1\}} Q^m(s,a) \\ Q_\lambda^m(s,a) &= R^m(s) - \lambda\mathbb{I}\{a=1\} \\ &\quad + \beta\sum_{s'\in\mathcal{S}} P_a^m(s,s')V^m(s'). \end{aligned} \tag{B.9}$$

The definition of Whittle's index will be the same in the average reward case.

B.1 Generalized Regret in Online Restless Bandits

Given K episodes each with a horizon length of T and the set of online Whittle's index policy at each episode $k \in [K]$, $\boldsymbol{\omega} := \{\boldsymbol{\omega}_k\}_{k\in[K]}$, the regret $\mathcal{R}^{\boldsymbol{\pi}^{\text{Whittle}}}_{\text{Restless}}(\boldsymbol{\omega})$:

$$\mathcal{R}_f(\boldsymbol{\omega}) := \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_{\boldsymbol{\pi}^{\text{Whittle}}}\left[\sum_{m=1}^{M}\sum_{t=0}^{T} R^m\left(s_t^m\right) \mid \boldsymbol{s}_0\right] - \mathbb{E}_{\boldsymbol{\omega}_k}\left[\sum_{m=1}^{M}\sum_{t=0}^{T} R^m\left(s_t^m\right) \mid \boldsymbol{s}_0\right]\right). \tag{B.10}$$

We note that our problem setting corresponds to the scenario where $K = 1$.

Existing regret bounds on online Whittle's index policy are only meaningful when the number of episodes is sufficiently large. To illustrate this, Wang et al. (2023b) shows that their algorithm has a regret bound of the order $\mathcal{O}\left(T\sqrt{K}\right)$. If we only have 1 episode, this will be equivalent to saying that the regret bound is $\mathcal{O}(T)$. The hardness lies in estimating the MDP: Auer et al. 2008 show that the lower bound of the regret is $\Omega(T\sqrt{K})$. This lower bound indicates that when directly dealing with an MDP, it is impossible to get a regret bound that is sublinear in T.

C COUNTER EXAMPLE FOR HETEROGENEOUS AGENTS

In this section, we will construct an instance involving heterogeneous agents where the greedy policy is not optimal. Say that each agent is associated with two state: State 0 and State 1, and the agent can take either Action 1 (intervention) or Action 0 (no intervention). Assume for now that there are **4** agents with different transition probabilities. The first two agents fall under Type 1: their transition probability under Action 1 is shown in Figure C.3 and their transition probability under Action 0 is shown in Figure C.4. The other two agents fall under Type 2: their transition probability under Action 1 is shown in Figure C.5 and their transition probability under Action 0 is shown in Figure C.6.

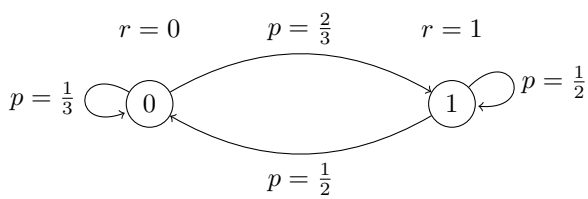

Figure C.3: Dynamic of Agent Type 1 with Intervention

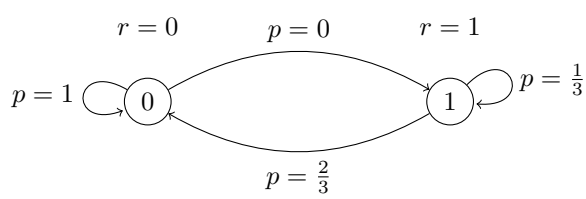

Figure C.4: Dynamic of Agent Type 1 without Intervention

Suppose the budget equals one, meaning that we can only give one agent intervention at each time. We observe that under our problem setup Type 2 agents always have a higher incremental reward $\phi^2(1,1) := \frac{3}{4}\cdot 1, \phi^2(0,1) := \frac{3}{4}\cdot 1, \phi^2(s,0) = 0, I^2(1) = I^2(0) = \frac{3}{4}$, opposed to $\phi^1(0,1) = \frac{2}{3}\cdot 1, \phi^1(1,1) = \frac{1}{2}\cdot 1, \phi^1(0,0) = 0, \phi^1(1,0) = \frac{1}{3}, I(1) = \frac{2}{3}, I(0) = \frac{1}{6}$. However, this does not imply that the optimal policy is to always take Action 1 for Type 2 agents. If all resources are allocated to Type 2 agents, the expected reward is 3/4 per time step summed across all agents. Alternatively, if resources are devoted to Type 1 agents, alternating actions between the two Type 1 agents, the expected reward is $2/3 + 2/3 \times 1/3 = 8/9$ per Type 1 agent over two time periods (where an action is taken in one period but not in the other). To illustrate this point, we ran the following experiment: we set the horizon to

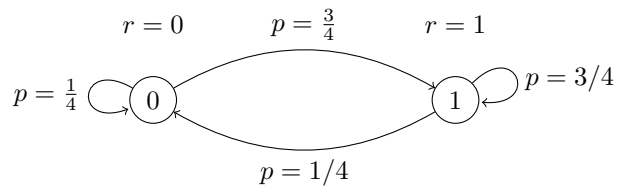

Figure C.5: Dynamic of Agent Type 2 with Intervention

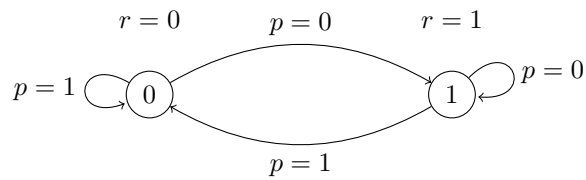

Figure C.6: Dynamic of Agent Type 2 without Intervention

be 20 and repeated the simulation 300 times. We compare the performance of the following two policies in terms of cumulative rewards: Policy 1) choosing one of the Type 2 agents for intervention at each time, and Policy 2) alternating intervention between the two Type 1 agents. The observed cumulative for Policies 1 and 2 are 15.02 and 16.78, respectively. This shows that the greedy policy is not optimal under all heterogeneous agents cases.

D COUNTER EXAMPLE FOR THREE-STATE, HOMOGENEOUS AGENTS

In this section, we will construct an instance involving homogeneous agents with three states where the greedy policy is not optimal. Say that each agent is associated with three states: State 0, State 1 and State 2, and the agent can take either Action 1 (intervention) or Action 0 (no intervention). Assume for now that there are two agents with the same transition probabilities. Their transition probability under Action 1 is shown in Figure D.7 and their transition probability under Action 0 is shown in Figure D.8.

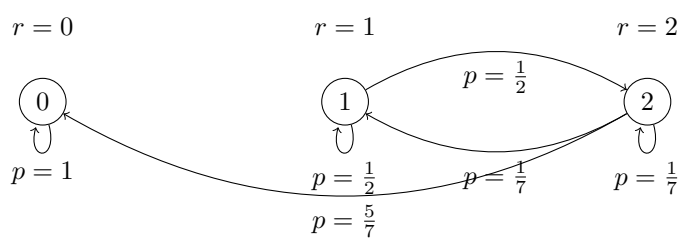

Figure D.7: Dynamic with Intervention

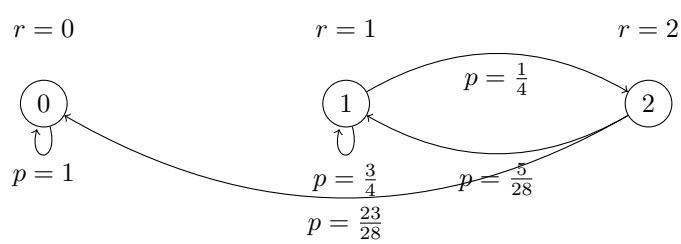

Figure D.8: Dynamic without Intervention

Suppose that the budget equals one, we observe that State 2 and State 1 have the same incremental reward $I(2) = \frac{3}{7} - \frac{5}{28} = \frac{1}{4} = \frac{3}{2} - \frac{5}{4} = I(1)$. However, we will show that when $T = 3$, greedy policy is not optimal.

Specifically, we will show that when $t = 0$ and $\boldsymbol{s}_0 = (1, 2)$, the optimal policy would be giving State 2 Action 1 instead of State 1 Action 1. This will show that greedy policy is not optimal as greedy policy could give State 1 Action 1 by definition. Suppose we are giving State 1 Action 1, then we have

$$\begin{aligned} V((1,2)) &= 3 + \frac{5}{56}V_1^*((1,2)) + \frac{5}{56}V_1^*((1,1)) + \frac{23}{28}V_1^*((1,0)) \\ &= 3 + \frac{5}{56}\frac{47 + 3 * 28}{28} + \frac{5}{56}\frac{19}{4} + \frac{23}{28}\frac{5}{2}. \end{aligned}$$

Similarly, if we are giving State 2 Action 1, then we have

$$\begin{aligned} V((1,2)) &= 3 + \frac{1}{28}V_1^*((2,2)) + \frac{1}{7}V_1^*((1,2)) + \frac{15}{28}V_1^*((1,0)) + \frac{5}{28}V_1^*((2,0)) + \frac{3}{28}V_1^*((1,1)) \\ &= 3 + \frac{1}{28}\left(\frac{17}{28} + 4\right) + \frac{1}{7}\left(3 + \frac{47}{28}\right) + \frac{15}{28}\left(\frac{5}{2}\right) + \frac{5}{28}\left(\frac{3}{7} + 2\right) + \frac{3}{28}\left(\frac{11}{4} + 2\right). \end{aligned}$$

By calculation, we will have the results. This shows that the greedy policy is not optimal under three states setting.

E TECHNICAL DETAILS IN SECTION 3

First, recall that at each time t of an offline restless bandit problem, the learner observes the set of states of M agents, denoted by $\boldsymbol{s}_t := \left(s_t^1, s_t^2, \cdots, s_t^M\right)$, and receives the corresponding reward $\boldsymbol{r}_t := (r_t^1, \cdots, r_t^M)$. Based on the observed states, the learner determines the action $\boldsymbol{a}_{t+1} := \left(a_{t+1}^1, a_{t+1}^2, \cdots, a_{t+1}^M\right)$, according to a deterministic policy $\pi_{t+1}^d(\boldsymbol{s}) : \mathcal{S}^M \to \{0,1\}^M$. The learner is provided with a budget constraint, allowing at most B arms to be pulled at each time step, i.e., $\sum_{m\in[M]} a_{t+1}^m = B$. After pulling up to B arms, the learner observes the next state $\boldsymbol{s}_{t+1}$, generated according to $s_{t+1}^m \sim P_{a_{t+1}^m}^m\left(s_t^m, s_{t+1}^m\right)$, and receives a feedback $\boldsymbol{r}_{t+1}$.

To ease notation, in Section E we will use $q_{ss'}^a(m)$ to denote $P_a^m(s, s')$. Before presenting the proofs, we reformulate the online restless bandit problem as a single large MDP, where the state and action are defined as tuples comprising the composite states and actions of all agents. This reformulation is valid because the agents/arms in the restless bandit problem are independent. This MDP can be described as $\left(\mathcal{S}^M, \mathcal{F}, q, \boldsymbol{R}\right)$, where $\mathcal{F}$ is defined as

$$\mathcal{F} := \left\{ \left(a^1, a^2, \cdots, a^M\right) : \left(a^1, a^2, \cdots, a^M\right) \in \mathcal{A}^M, \sum_{m=1}^M a^m \le B \right\}.$$

Let $q_{\boldsymbol{ss}'}^{\boldsymbol{a}} := \mathbb{P}(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) = P_{\boldsymbol{a}}(\boldsymbol{s}, \boldsymbol{s}') = \prod_{m=1}^M q_{ss'}^a(m)$, and let $\boldsymbol{R}(\boldsymbol{s}) = \sum_{m=1}^M r^m\left(s^m\right)$. Then for each policy $\boldsymbol{\pi}$ and each time-step h, we can define the value function $V_h^{\boldsymbol{\pi}}(\boldsymbol{s})$ as

$$\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=h}^T \boldsymbol{R}(\boldsymbol{s}_t) \mid \boldsymbol{s}_h = \boldsymbol{s}\right]. \tag{E.11}$$

With this definition, Problem (1) is equivalent to

$$\max_{\boldsymbol{\pi}} V_0^{\boldsymbol{\pi}}\left(\boldsymbol{s}_0\right).$$

Again, we use $\boldsymbol{\pi}^* := \{\pi_t^*\}_{t\in[T]}$ to denote the maximizer, and let V_h^* be the value function under policy $\boldsymbol{\pi}^*$. We will use this notation in Appendix E.1, and F.

E.1 Proof of Theorem 3.1

Theorem 3.1. *Assuming all agents are homogeneous, i.e., $P^m = P^{m'}, R^m = R^{m'}, \forall m, m' \in [M]$ and $\mathcal{S} = \{0, 1\}$, $\boldsymbol{\pi}^{greedy} = \boldsymbol{\pi}^*$ for any $0 \le B \le M < \infty$.*

Proof. WLOG, assume that $I(1) \le I(0)$. For simplicity, we first look at the simple case where $M = 2, B = 1$. The generalization to multi-agent case is straightforward and will be deferred to the last part of the proof. First

of all, we will derive some useful equations that are useful in the proof.

$$
\begin{aligned}
&\sum_{s^1_{h+1},s^2_{h+1}} q^0_{1s^1_{h+1}} q^1_{0s^2_{h+1}} V^*_{h+1}\left(\left(s^1_{h+1},s^2_{h+1}\right)\right) - \sum_{s^1_{h+1},s^2_{h+1}} q^1_{1s^1_{h+1}} q^0_{0s^2_{h+1}} V^*_{h+1}\left(\left(s^1_{h+1},s^2_{h+1}\right)\right)\\
&= q^0_{11}q^1_{01}V^*_{h+1}\left((1,1)\right) + q^0_{11}q^1_{00}V^*_{h+1}\left((1,0)\right) + q^0_{10}q^1_{01}V^*_{h+1}\left((1,0)\right) + q^0_{10}q^1_{00}V^*_{h+1}\left((0,0)\right)\\
&- q^1_{11}q^0_{01}V^*_{h+1}\left((1,1)\right) - q^1_{11}q^0_{00}V^*_{h+1}\left((1,0)\right) - q^1_{10}q^0_{01}V^*_{h+1}\left((1,0)\right) - q^1_{10}q^0_{00}V^*_{h+1}\left((0,0)\right)\\
&= q^0_{11}q^1_{01}V^*_{h+1}\left((1,1)\right) - q^0_{11}q^1_{01}V^*_{h+1}\left((1,0)\right) + q^0_{11}V^*_{h+1}\left((1,0)\right)\\
&+ q^1_{01}V^*_{h+1}\left((1,0)\right) - q^0_{11}q^1_{01}V^*_{h+1}\left((1,0)\right)\\
&+ V^*_{h+1}\left((0,0)\right) + q^0_{11}q^1_{01}V^*_{h+1}\left((0,0)\right) - \left(q^0_{11}+q^1_{01}\right)V^*_{h+1}\left((0,0)\right)\\
&- q^1_{11}q^0_{01}V^*_{h+1}\left((1,1)\right) + q^1_{11}q^0_{01}V^*_{h+1}\left((1,0)\right) - q^1_{11}V^*_{h+1}\left((1,0)\right)\\
&- q^0_{01}V^*_{h+1}\left((1,0)\right) + q^1_{11}q^0_{01}V^*_{h+1}\left((1,0)\right)\\
&- V^*_{h+1}\left((0,0)\right) - q^1_{11}q^0_{01}V^*_{h+1}\left((0,0)\right) + \left(q^1_{11}+q^0_{01}\right)V^*_{h+1}\left((0,0)\right)\\
&= \left(q^0_{11}+q^1_{01}-q^1_{11}-q^0_{01}\right)\left(V^*_{h+1}\left((1,0)\right) - V^*_{h+1}\left((0,0)\right)\right)\\
&+ \left(q^0_{11}q^1_{01}-q^0_{01}q^1_{11}\right)\left(V^*_{h+1}\left((1,1)\right) - 2V^*_{h+1}\left((1,0)\right) + V^*_{h+1}\left((0,0)\right)\right).
\end{aligned}
\tag{E.12}
$$

Next, we will simplify the difference between the value function of different states, assuming that the optimal policy is the greedy policy. For simplicity, we will only give the detailed derivation of the equation for $V^*_h\left((1,1)\right) - V^*_h\left((1,0)\right)$. The equation for $V^*_h\left((1,0)\right) - V^*_h\left((0,0)\right)$ holds for similar reasons.

$$
\begin{aligned}
V^*_h\left((1,1)\right) - V^*_h\left((1,0)\right) &= 2R(1) + \sum_{s^1_{h+1},s^2_{h+1}} q^1_{1s^1_{h+1}} q\left(s^2_{h+1} \mid 1,0\right) V^*_{h+1}\left(\left(s^1_{h+1},s^2_{h+1}\right)\right)\\
&- R(1) - R(0) - \sum_{s^1_{h+1},s^2_{h+1}} q^0_{1s^1_{h+1}} q^1_{0s^2_{h+1}} V^*_{h+1}\left(\left(s^1_{h+1},s^2_{h+1}\right)\right)\\
&= R(1) - R(0) + q^1_{11}q^0_{11}V^*_{h+1}\left((1,1)\right) + q^1_{11}q^0_{10}V^*_{h+1}\left((1,0)\right)\\
&+ q^1_{10}q^0_{11}V^*_{h+1}\left((1,0)\right) + q^0_{10}q^1_{10}V^*_{h+1}\left((0,0)\right)\\
&- q^0_{11}q^1_{01}V^*_{h+1}\left((1,1)\right) - q^0_{11}q^1_{00}V^*_{h+1}\left((1,0)\right)\\
&- q^0_{10}q^1_{01}V^*_{h+1}\left((1,0)\right) - q^0_{10}q^1_{00}V^*_{h+1}\left((0,0)\right)\\
&= R(1) - R(0) + q^1_{11}q^0_{11}V^*_{h+1}\left((1,1)\right) + q^1_{11}V^*_{h+1}\left((1,0)\right) - q^1_{11}q^0_{11}V^*_{h+1}\left((1,0)\right)\\
&+ q^0_{11}V^*_{h+1}\left((1,0)\right) - q^1_{11}q^0_{11}V^*_{h+1}\left((1,0)\right)\\
&+ V^*_{h+1}\left((0,0)\right) + q^0_{11}q^1_{11}V^*_{h+1}\left((0,0)\right) - \left(q^1_{11}+q^0_{11}\right)V^*_{h+1}\left((0,0)\right)\\
&- q^0_{11}q^1_{01}V^*_{h+1}\left((1,1)\right) - q^0_{11}V^*_{h+1}\left((1,0)\right) + q^1_{01}q^0_{11}V^*_{h+1}\left((1,0)\right)\\
&- q^1_{01}V^*_{h+1}\left((1,0)\right) + q^0_{11}q^1_{01}V^*_{h+1}\left((1,0)\right)\\
&- V^*_{h+1}\left((0,0)\right) - q^0_{11}q^1_{01}V^*_{h+1}\left((0,0)\right) + \left(q^0_{11}+q^1_{01}\right)V^*_{h+1}\left((0,0)\right)\\
&= R(1) - R(0) + \left(q^1_{11}-q^1_{01}\right)\left(V^*_{h+1}\left((1,0)\right) - V^*_{h+1}\left((0,0)\right)\right)\\
&+ q^0_{11}\left(q^1_{11}-q^1_{01}\right)\left(V^*_{h+1}\left((1,1)\right) - 2V^*_{h+1}\left((1,0)\right) + V^*_{h+1}\left((0,0)\right)\right).
\end{aligned}
\tag{E.13}
$$

Similarly, we can have

$$
\begin{aligned}
V^*_h\left((1,0)\right) - V^*_h\left((0,0)\right) &= R(1) - R(0) + \left(q^0_{11}-q^0_{01}\right)\left(V^*_{h+1}\left((1,0)\right) - V^*_{h+1}\left((0,0)\right)\right)\\
&+ q^1_{01}\left(q^0_{11}-q^0_{01}\right)\left(V^*_{h+1}\left((1,1)\right) - 2V^*_{h+1}\left((1,0)\right) + V^*_{h+1}\left((0,0)\right)\right).
\end{aligned}
\tag{E.14}
$$

We will derive the final useful equation for the incremental reward.

$$
\begin{aligned}
I(1) &= \phi(1,1) - \phi(1,0)\\
&= q^1_{11}R(1) + q^1_{10}R(0) - q^0_{11}R(1) - q^0_{10}R(0) \qquad \text{(E.15)}\\
&= \left(q^1_{11}-q^0_{11}\right)\left(R(1)-R(0)\right). \qquad \text{(E.16)}
\end{aligned}
$$

Similarly, we also have

$$I(0) = \left(q_{01}^1 - q_{01}^0\right)\left(R(1) - R(0)\right). \tag{E.17}$$

In the following, we will consider four cases.

Case 1: $q_{11}^0 \geq q_{01}^0, q_{11}^1 \leq q_{01}^1, q_{11}^1 \geq q_{11}^0$. We will use backward induction to prove the following statements:

$$\forall t \leq T, 0 \leq V_t^*\left((1,0)\right) - V_t^*\left((0,0)\right) \leq \frac{R(1) - R(0)}{q_{10}^0} \tag{E.18}$$

$$0 \leq V_t^*\left((1,1)\right) - V_t^*\left((1,0)\right) \leq R(1) - R(0) \tag{E.19}$$

$$\pi_{t+1}^{\text{greedy}} = \pi_{t+1}^*. \tag{E.20}$$

The statements hold trivially when $t = T$. Suppose the statements hold when $t \geq h+1$, when $t = h$, we will first prove Eq. (E.20) for $t = h$. Recall that the policy satisfying the following Bellman optimality equation is the optimal policy:

$$V_h^*\left((s_h^1, s_h^2)\right) = R(s_h^1) + R(s_h^2) + \max_{a_{h+1}^1, a_{h+1}^2} \sum_{s_{h+1}^1, s_{h+1}^2} q_{s_h^1 s_{h+1}^1}^{a_{h+1}^1} q_{s_h^2 s_{h+1}^2}^{a_{h+1}^2} V_{h+1}^*\left(s_{h+1}^1, s_{h+1}^2\right). \tag{E.21}$$

By the problem formulation, one of the $a_{h+1}^i = 1$. Therefore, it's easy to see that when $s_h^1 = s_h^2$, randomly giving action will satisfy this equation. We will focus on the case where $s_h^1 \neq s_h^2$. In that case, to show Eq. (E.20), we only need to show that

$$\sum_{s_{h+1}^1, s_{h+1}^2} q_{1 s_{h+1}^1}^0 q_{0 s_{h+1}^2}^1 V_{h+1}^*\left(s_{h+1}^1, s_{h+1}^2\right) \geq \sum_{s_{h+1}^1, s_{h+1}^2} q_{1 s_{h+1}^1}^1 q_{0 s_{h+1}^2}^0 V_{h+1}^*\left(s_{h+1}^1, s_{h+1}^2\right). \tag{E.22}$$

By Eq. (E.12), this is equivalent to showing

$$\begin{aligned}&\left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right)\\ &+ \left(q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1\right)\left(V_{h+1}^*\left((1,1)\right) - 2V_{h+1}^*\left((1,0)\right) + V_{h+1}^*\left((0,0)\right)\right) \geq 0.\end{aligned}$$

Then we have

$$\begin{aligned}&\left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right)\\ &+ \left(q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1\right)\left(V_{h+1}^*\left((1,1)\right) - 2V_{h+1}^*\left((1,0)\right) + V_{h+1}^*\left((0,0)\right)\right)\\ &\geq \left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right)\\ &- \left(q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1\right)\left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right)\\ &= \frac{I(0) - I(1)}{R(1) - R(0)}\left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right) - \frac{\left(q_{11}^1 I(0) - q_{01}^1 I(1)\right)}{R(1) - R(0)}\left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right)\\ &= \frac{\left(q_{10}^1 I(0) - q_{00}^1 I(1)\right)}{R(1) - R(0)}\left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right)\\ &\geq 0,\end{aligned}$$

where the first inequality holds because of the induction hypothesis and $q_{11}^0 \geq q_{01}^0, q_{01}^1 \geq q_{11}^1$, the first equality holds because of Eq. (E.17) and Eq. (E.16) and the last inequality holds because $q_{10}^1 \geq q_{00}^1$, $I(0) \geq I(1)$ and the assumption that Eq. (E.18) holds for $t = h+1$.

Next, we first show the LHS of Eq. (E.18) for $t = h$ under the assumption that Eq. (E.20) holds for $t = h$ and Eq. (E.18) holds for $t = h+1$.

$$\begin{aligned}V_h^*\left((1,0)\right) - V_h^*\left((0,0)\right) &= R(1) - R(0) + \left(q_{11}^0 - q_{01}^0\right)\left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right)\\ &\quad + q_{01}^1\left(q_{11}^0 - q_{01}^0\right)\left(V_{h+1}^*\left((1,1)\right) - 2V_{h+1}^*\left((1,0)\right) + V_{h+1}^*\left((0,0)\right)\right)\\ &\geq R(1) - R(0) + \left(q_{11}^0 - q_{01}^0\right)\left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right)\end{aligned}$$

$$
\begin{aligned}
&- q_{01}^1 \left(q_{11}^0 - q_{01}^0\right) \left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right) \\
&\geq R(1) - R(0) + q_{00}^1 \left(q_{11}^0 - q_{01}^0\right) \left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right) \\
&\geq 0,
\end{aligned}
$$

where the first equality holds because of Eq. (E.14) and the last inequality holds since $q_{11}^0 \geq q_{01}^0$. Under the same condition, for the RHS of Eq. (E.18), we have

$$
\begin{aligned}
q_{10}^0 \left(V_h^*\left((1,0)\right) - V_h^*\left((0,0)\right)\right) &= q_{10}^0(R(1) - R(0)) \\
&+ q_{10}^0 q_{00}^1 \left(q_{11}^0 - q_{01}^0\right) \left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right) \\
&+ q_{10}^0 q_{01}^1 \left(q_{11}^0 - q_{01}^0\right) \left(V_{h+1}^*\left((1,1)\right) - V_{h+1}^*\left((1,0)\right)\right) \\
&\leq q_{10}^0(R(1) - R(0)) \\
&+ q_{00}^1 \left(q_{11}^0 - q_{01}^0\right) (R(1) - R(0)) \\
&+ q_{10}^0 q_{01}^1 \left(q_{11}^0 - q_{01}^0\right) (R(1) - R(0)) \\
&\leq q_{10}^0(R(1) - R(0)) \\
&+ q_{00}^1 q_{11}^0 (R(1) - R(0)) + q_{01}^1 q_{11}^0 (R(1) - R(0)) \\
&= R(1) - R(0).
\end{aligned}
$$

Then, we only need to show Eq. (E.19) for $t = h$. Assuming that Eq. (E.20) holds for $t = h$ and Eq. (E.19) holds for $t = h + 1$, for the RHS, we have

$$
\begin{aligned}
V_h^*\left((1,1)\right) - V_h^*\left((1,0)\right) &= R(1) - R(0) + \left(q_{11}^1 - q_{01}^1\right) \left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right) \\
&+ q_{11}^0 \left(q_{11}^1 - q_{01}^1\right) \left(V_{h+1}^*\left((1,1)\right) - 2V_{h+1}^*\left((1,0)\right) + V_{h+1}^*\left((0,0)\right)\right) \\
&= R(1) - R(0) + q_{11}^0 \left(q_{11}^1 - q_{01}^1\right) \left(V_{h+1}^*\left((1,1)\right) - V_{h+1}^*\left((1,0)\right)\right) \\
&+ q_{10}^0 \left(q_{11}^1 - q_{01}^1\right) \left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right) \\
&\leq R(1) - R(0),
\end{aligned}
$$

where the last inequality holds since $q_{11}^1 \leq q_{01}^1$. Under the same condition, for the LHS, we have

$$
\begin{aligned}
V_h^*\left((1,1)\right) - V_h^*\left((1,0)\right) &= R(1) - R(0) + \left(q_{11}^1 - q_{01}^1\right) \left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right) \\
&+ q_{11}^0 \left(q_{11}^1 - q_{01}^1\right) \left(V_{h+1}^*\left((1,1)\right) - 2V_{h+1}^*\left((1,0)\right) + V_{h+1}^*\left((0,0)\right)\right) \\
&= R(1) - R(0) + q_{11}^0 \left(q_{11}^1 - q_{01}^1\right) \left(V_{h+1}^*\left((1,1)\right) - V_{h+1}^*\left((1,0)\right)\right) \\
&+ q_{10}^0 \left(q_{11}^1 - q_{01}^1\right) \left(V_{h+1}^*\left((1,0)\right) - V_{h+1}^*\left((0,0)\right)\right) \\
&\geq R(1) - R(0) + q_{11}^0 \left(q_{11}^1 - q_{01}^1\right) (R(1) - R(0)) \\
&+ \left(q_{11}^1 - q_{01}^1\right) (R(1) - R(0)) \\
&= \left(q_{00}^1 + q_{11}^1 - q_{11}^0 q_{01}^1 + q_{11}^0 q_{11}^1\right) (R(1) - R(0)) \\
&\geq 0,
\end{aligned}
$$

where the last inequality holds since $q_{11}^1 \geq q_{11}^0$.

Case 2: $q_{11}^0 \leq q_{01}^0, q_{11}^1 \leq q_{01}^1$.

First we consider the sub-case where $\boxed{\boldsymbol{q_{11}^1 q_{01}^0 \leq q_{01}^1 q_{11}^0}}$. We will use backward induction to prove the following statements:

$$
\forall t \leq T, 0 \leq V_t^*\left((1,0)\right) - V_t^*((0,0)) \leq R(1) - R(0) \tag{E.23}
$$

$$
0 \leq V_t^*\left((1,1)\right) - V_t^*((1,0)) \leq R(1) - R(0) \tag{E.24}
$$

$$
\pi_{t+1}^{\text{greedy}} = \pi_{t+1}^*. \tag{E.25}
$$

When $t = T$, the statements hold trivially. Suppose the statements hold for $t \geq h + 1$, when $t = h$, we will first prove Eq. (E.25) under the assumption that Eq. (E.23) and Eq. (E.24) holds for $t = h + 1$. Similarly we only

need to show Eq. (E.22).

$$\begin{aligned}
&\left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&+ \left(q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \\
&\geq \frac{\left(q_{10}^1 I(0) - q_{00}^1 I(1)\right)}{R(1) - R(0)}\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&\geq 0,
\end{aligned}$$

where the first inequality holds because the assumption we make and the last inequality holds because $q_{10}^1 \geq q_{00}^1$, $I(0) \geq I(1)$ and the assumption that Eq. (E.23) holds.

Next we want to prove the RHS of Eq. (E.24) under the assumption that Eq. (E.25) holds for $t = h$ and that Eq. (E.23) and Eq. (E.24) hold for $t = h + 1$.

$$\begin{aligned}
V_h^*((1,1)) - V_h^*((1,0)) &= R(1) - R(0) + \left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&+ q_{11}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \\
&= R(1) - R(0) + q_{11}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,1)) - V_{h+1}^*((1,0))\right) \\
&+ q_{10}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&\leq R(1) - R(0).
\end{aligned}$$

Under the same condition, for the LHS of Eq. (E.24), we have

$$\begin{aligned}
V_h^*((1,1)) - V_h^*((1,0)) &= R(1) - R(0) + q_{11}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,1)) - V_{h+1}^*((1,0))\right) \\
&+ q_{10}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&\geq R(1) - R(0) + q_{10}^0\left(q_{11}^1 - q_{01}^1\right)(R(1) - R(0)) \\
&+ q_{11}^0\left(q_{11}^1 - q_{01}^1\right)(R(1) - R(0)) \\
&= (q_{11}^1 + q_{00}^1)(R(1) - R(0)) \\
&\geq 0.
\end{aligned}$$

Similarly, we can show Eq. (E.23) for $t = h$ under the same condition.

Secondly, we consider the sub-case where $\boxed{\boldsymbol{q_{11}^1 q_{01}^0 \geq q_{01}^1 q_{11}^0}}$. We want to use backward induction to prove the following statements:

$$\forall t \leq T, 0 \leq V_t^*((1,0)) - V_t^*((0,0)) \leq R(1) - R(0) \tag{E.26}$$

$$0 \leq V_t^*((1,1)) - V_t^*((1,0)) \leq R(1) - R(0) \tag{E.27}$$

$$V_t^*((1,1)) - V_t^*((1,0)) \leq V_t^*((1,0)) - V_t^*((0,0)) \tag{E.28}$$

$$\pi_{t+1}^{\text{greedy}} = \pi_{t+1}^*. \tag{E.29}$$

When $t = T$, the statements hold trivially. Suppose this holds for $t \geq h + 1$, when $t = h$, we will first show Eq. (E.29) for $t = h$ under the assumption that Eq. (E.26), Eq. (E.27) and Eq. (E.28) hold for $t = h+1$. Similarly, this is equivalent to show Eq. (E.22).

$$\begin{aligned}
&\left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&+ \left(q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \\
&= \frac{I(0) - I(1)}{R(1) - R(0)}\left(V_{h+1}^*(1,0) - V_{h+1}^*(0,0)\right) \\
&+ \left(q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \\
&\geq 0,
\end{aligned}$$

where the first equality holds because of Eq. (E.17) and Eq. (E.16) and the first inequality holds since $q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1 \leq 0$ and the assumption that Eq. (E.28) holds for $t = h + 1$.

Using the same logic as the previous case, we can prove that

$$0 \le V_h^*((1,0)) - V_h^*((0,0)) \le R(1) - R(0)$$
$$0 \le V_h^*((1,1)) - V_h^*((1,0)) \le R(1) - R(0).$$

Therefore, the only thing we need to show is the Eq. (E.28) for $t = h$ under the assumption that Eq. (E.29) holds for $t = h$ and that Eq. (E.28) holds for $t = h+1$. We have

$$\begin{aligned}
&V_h^*((1,1)) - 2V_h^*((1,0)) + V_h^*((0,0)) \\
&= q_{11}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \\
&+ \left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&- q_{01}^1\left(q_{11}^0 - q_{01}^0\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \\
&- \left(q_{11}^0 - q_{01}^0\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&= q_{11}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \\
&- q_{01}^1\left(q_{11}^0 - q_{01}^0\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \\
&+ \frac{I(1) - I(0)}{R(1) - R(0)}\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right).
\end{aligned}$$

Since $I(1) \le I(0), V_{h+1}^*((1,0)) - V_{h+1}^*((0,0)) \ge 0$, we only need to prove that

$$q_{11}^0\left(q_{11}^1 - q_{01}^1\right) \ge q_{01}^1\left(q_{11}^0 - q_{01}^0\right).$$

Since we have

$$q_{11}^1 q_{01}^0 \ge q_{01}^1 q_{11}^0,$$

we have

$$\begin{aligned}
q_{11}^1 q_{01}^0 - q_{11}^1 q_{11}^0 &\ge q_{01}^1 q_{11}^0 - q_{11}^1 q_{11}^0 \\
q_{11}^1\left(q_{01}^0 - q_{11}^0\right) &\ge q_{11}^0(q_{01}^1 - q_{11}^1) \\
q_{01}^1\left(q_{01}^0 - q_{11}^0\right) \ge q_{11}^1\left(q_{01}^0 - q_{11}^0\right) &\ge q_{11}^0(q_{01}^1 - q_{11}^1) \\
q_{01}^1\left(q_{11}^0 - q_{01}^0\right) &\le q_{11}^0(q_{11}^1 - q_{01}^1).
\end{aligned}$$

This will give us the desired result.

Case 3: $q_{11}^1 \ge q_{01}^1, q_{11}^0 \ge q_{01}^0$.

We will use backward induction to prove the following statements:

$$\forall t \le T, \frac{R(1) - R(0)}{q_{10}^1 + q_{01}^1}\left(1 - \left(q_{11}^1 - q_{01}^1\right)^{T-t+1}\right) \le V_t^*((1,0)) - V_t^*((0,0)) \tag{E.30}$$

$$\frac{R(1) - R(0)}{q_{10}^0 + q_{01}^0}\left(1 - \left(q_{11}^0 - q_{01}^0\right)^{T-t+1}\right) \ge V_t^*((1,0)) - V_t^*((0,0)) \tag{E.31}$$

$$\frac{R(1) - R(0)}{q_{10}^1 + q_{01}^1}\left(1 - \left(q_{11}^1 - q_{01}^1\right)^{T-t+1}\right) \le V_t^*((1,0)) - V_t^*((0,0)) \tag{E.32}$$

$$\frac{R(1) - R(0)}{q_{10}^0 + q_{01}^0}\left(1 - \left(q_{11}^0 - q_{01}^0\right)^{T-t+1}\right) \ge V_t^*((1,0)) - V_t^*((0,0)) \tag{E.33}$$

$$\pi_t^{\text{greedy}} = \pi_t^*. \tag{E.34}$$

When $t = T$, this holds trivially, suppose this holds for $t \ge h+1$. When $t = h$, we will first show that Eq. (E.34) holds for $t = h$ under the assumption that Eq. (E.30), Eq. (E.31), Eq. (E.32) and Eq. (E.33) hold for $t = h+1$. This again is equivalent to show that

$$\begin{aligned}
&\left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&+ \left(q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \ge 0.
\end{aligned}$$

We have

$$\begin{aligned}
&\left(q_{11}^0+q_{01}^1-q_{11}^1-q_{01}^0\right)\left(V_{h+1}^*\left((1,0)\right)-V_{h+1}^*\left((0,0)\right)\right)\\
&+\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\left(V_{h+1}^*\left((1,1)\right)-2V_{h+1}^*\left((1,0)\right)+V_{h+1}^*\left((0,0)\right)\right)\\
&=\left(q_{11}^0+q_{01}^1-q_{11}^1-q_{01}^0-q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\left(V_{h+1}^*\left((1,0)\right)-V_{h+1}^*\left((0,0)\right)\right)\\
&+\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\left(V_{h+1}^*\left((1,1)\right)-V_{h+1}^*\left((1,0)\right)\right).
\end{aligned}$$

If $q_{11}^0+q_{01}^1-q_{11}^1-q_{01}^0-q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\geq 0$, then using our assumption that, $V_{h+1}^*\left((1,1)\right)-V_{h+1}^*\left((1,0)\right)\geq 0, V_{h+1}^*\left((1,0)\right)-V_{h+1}^*\left((0,0)\right)\geq 0$, will give us the desired result.

If $q_{11}^0+q_{01}^1-q_{11}^1-q_{01}^0-q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\leq 0$, with our hypothesis, we have

$$\begin{aligned}
&\left(q_{11}^0+q_{01}^1-q_{11}^1-q_{01}^0-q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\left(V_{h+1}^*\left((1,0)\right)-V_{h+1}^*\left((0,0)\right)\right)\\
&+\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\left(V_{h+1}^*\left((1,1)\right)-V_{h+1}^*\left((1,0)\right)\right)\\
&\geq\left(q_{11}^0+q_{01}^1-q_{11}^1-q_{01}^0-q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\\
&\times\frac{R(1)-R(0)}{q_{10}^0+q_{01}^0}\left(1-\left(q_{11}^0-q_{01}^0\right)^{T-h}\right)\\
&+\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\frac{R(1)-R(0)}{q_{10}^1+q_{01}^1}\left(1-\left(q_{11}^1-q_{01}^1\right)^{T-h}\right).
\end{aligned}\tag{E.35}$$

Denote $\frac{R(1)-R(0)}{q_{10}^0+q_{01}^0}\left(1-\left(q_{11}^0-q_{01}^0\right)^{T-h}\right)$ as d_{upper} and let k be defined as $\frac{R(1)-R(0)}{q_{10}^1+q_{01}^1}\left(1-\left(q_{11}^1-q_{01}^1\right)^{T-h}\right)/d_{\text{upper}}$. Then to show that Eq. (E.35) is larger than 0, we only need to show that

$$q_{11}^0+q_{01}^1-q_{11}^1-q_{01}^0-(k-1)\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\geq 0,$$

which is equivalent to show $k\geq\frac{q_{11}^0q_{01}^1-q_{01}^0q_{11}^1-q_{11}^0-q_{01}^1+q_{11}^1+q_{01}^0}{q_{11}^0q_{01}^1-q_{01}^0q_{11}^1}$.

First, we will show

$$\frac{q_{10}^0+q_{01}^0}{q_{10}^1+q_{01}^1}\geq\frac{q_{11}^0q_{01}^1-q_{01}^0q_{11}^1-q_{11}^0-q_{01}^1+q_{11}^1+q_{01}^0}{q_{11}^0q_{01}^1-q_{01}^0q_{11}^1}.$$

To do this, we only need to show

$$\begin{aligned}
&\left(q_{10}^0+q_{01}^0\right)\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\\
&-\left(q_{10}^1+q_{01}^1\right)\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\\
&-\left(q_{10}^1+q_{01}^1\right)\left(q_{11}^1+q_{01}^0-q_{11}^0-q_{01}^1\right)\geq 0.
\end{aligned}$$

We have

$$\begin{aligned}
&\left(q_{10}^0+q_{01}^0\right)\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\\
&-\left(q_{10}^1+q_{01}^1\right)\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\\
&-\left(q_{10}^1+q_{01}^1\right)\left(q_{11}^1+q_{01}^0-q_{11}^0-q_{01}^1\right)\\
&=\left(q_{10}^0+q_{01}^0-q_{10}^1-q_{01}^1\right)\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1\right)\\
&-\left(q_{10}^1+q_{01}^1\right)\left(q_{11}^1+q_{01}^0-q_{11}^0-q_{01}^1\right)\\
&=\left(q_{11}^1+q_{01}^0-q_{11}^0-q_{01}^1\right)\left(q_{11}^0q_{01}^1-q_{01}^0q_{11}^1-q_{10}^1-q_{01}^1\right).
\end{aligned}$$

Since $q_{11}^1+q_{01}^0-q_{11}^0-q_{01}^1=\frac{I(1)-I(0)}{R(1)-R(0)}\leq 0$ and $q_{11}^0q_{01}^1-q_{01}^1\leq 0$, we have the desired result.

Second, we want to show that $1-\left(q_{11}^1-q_{01}^1\right)^{T-h}\geq 1-\left(q_{11}^0-q_{01}^0\right)^{T-h}$. To do this, we only need to show that $\left(q_{11}^1-q_{01}^1\right)^{T-h}\leq\left(q_{11}^0-q_{01}^0\right)^{T-h}$, which comes directly from the fact that $0\leq q_{11}^1-q_{01}^1\leq q_{11}^0-q_{01}^0$.

Then we have

$$k=\frac{q_{10}^0+q_{01}^0}{q_{10}^1+q_{01}^1}\frac{1-\left(q_{11}^1-q_{01}^1\right)^{T-h}}{1-\left(q_{11}^0-q_{01}^0\right)^{T-h}}$$

$$
\begin{aligned}
&\geq \frac{q_{10}^0 + q_{01}^0}{q_{10}^1 + q_{01}^1} \\
&\geq \frac{q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1 - q_{11}^0 - q_{01}^1 + q_{11}^1 + q_{01}^0}{q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1}.
\end{aligned}
$$

This will give us the desired result.

Next, we will first prove Eq. (E.30) for $t = h$ under the assumption that Eq. (E.30), Eq. (E.31), Eq. (E.32) and Eq. (E.33) hold for $t = h + 1$ and Eq. (E.29) holds for $t = h$. We have

$$
\begin{aligned}
V_h^*((1,1)) - V_h^*((1,0)) &= R(1) - R(0) + \left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&+ q_{11}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \\
&= R(1) - R(0) + q_{11}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,1)) - V_{h+1}^*((1,0))\right) \\
&+ q_{10}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&\geq R(1) - R(0) \\
&+ \left(q_{11}^1 - q_{01}^1\right)\frac{R(1) - R(0)}{q_{10}^1 + q_{01}^1}\left(1 - \left(q_{11}^1 - q_{01}^1\right)^{T-h}\right) \\
&= \frac{R(1) - R(0)}{q_{10}^1 + q_{01}^1}\left(1 - \left(q_{11}^1 - q_{01}^1\right)^{T-h+1}\right).
\end{aligned}
$$

Under the same condition, for Eq. (E.31) when $t = h$, we have

$$
\begin{aligned}
V_h^*((1,1)) - V_h^*((1,0)) &= R(1) - R(0) + \left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&+ q_{11}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,1)) - 2V_{h+1}^*((1,0)) + V_{h+1}^*((0,0))\right) \\
&= R(1) - R(0) + q_{11}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,1)) - V_{h+1}^*((1,0))\right) \\
&+ q_{10}^0\left(q_{11}^1 - q_{01}^1\right)\left(V_{h+1}^*((1,0)) - V_{h+1}^*((0,0))\right) \\
&\leq R(1) - R(0) \\
&+ \left(q_{11}^0 - q_{01}^0\right)\frac{R(1) - R(0)}{q_{10}^0 + q_{01}^0}\left(1 - \left(q_{11}^0 - q_{01}^0\right)^{T-h}\right) \\
&= \frac{R(1) - R(0)}{q_{10}^0 + q_{01}^0}\left(1 - \left(q_{11}^0 - q_{01}^0\right)^{T-h+1}\right),
\end{aligned}
$$

where the inequality holds because we have $q_{11}^1 - q_{01}^1 \leq q_{11}^0 - q_{01}^0$. Using the same logic will give us the Eq. (E.33) and Eq. (E.32) for $t = h$.

Case 4: Suppose $q_{11}^1 \leq q_{01}^1, q_{11}^0 \geq q_{01}^0, q_{11}^0 \geq q_{11}^1$. We will directly show that $V_t^{\boldsymbol{\pi}^{\text{greedy}}}$ satisfies the optimal Bellman equation. With the same logic, we only need to prove that Eq. (E.22) holds for each $t \leq T$. We will first assume that the following holds and will prove this later:

$$
V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) + V_t^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) - 2V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) \leq 0, \ \forall t \leq T. \tag{E.36}
$$

Then we have

$$
\begin{aligned}
&\left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) - V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((0,0))\right) \\
&+ \left(q_{01}^1 q_{11}^0 - q_{01}^0 q_{11}^1\right)\left(V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) + V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) - 2V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0))\right) \\
&= \left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) - V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((0,0))\right) \\
&+ \left(q_{11}^1(q_{11}^0 - q_{01}^0) - q_{11}^0(q_{11}^1 - q_{01}^1)\right)\left(V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) + V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) - 2V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0))\right) \\
&\geq \left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) - V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((0,0))\right) \\
&+ \left(q_{01}^1(q_{11}^0 - q_{01}^0) - q_{11}^0(q_{11}^1 - q_{01}^1)\right)\left(V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) + V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) - 2V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0))\right)
\end{aligned}
$$

$$
\begin{aligned}
&= -V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) + 2V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) \\
&\geq 0.
\end{aligned}
$$

Next, we will prove Eq. (E.36). To ease notation, we will let $a := 1 - q_{01}^1, b := 1 - q_{11}^0, c = q_{11}^0 - q_{01}^0, d = q_{11}^1 - q_{01}^1$. By Eq. (E.13) and (E.14), we have the following iterative equations

$$
\begin{aligned}
V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) &= R(1) - R(0) + ac\left(V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) - V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((0,0))\right) \\
&\quad + (1-a)c\left(V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) - V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0))\right) \\
V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) &= R(1) - R(0) + bd\left(V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) - V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((0,0))\right) \\
&\quad + (1-b)d\left(V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) - V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0))\right).
\end{aligned}
$$

Let $\boldsymbol{V}_t = \begin{pmatrix} V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) \\ V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) \end{pmatrix}, A = \begin{pmatrix} ac & (1-a)c \\ bd & (1-b)d \end{pmatrix}$. We have

$$
\boldsymbol{V}_t = (R(1) - R(0))\mathbf{1} + A\boldsymbol{V}_{t+1}.
$$

First we consider the subcase where $I - A$ is not invertible, which is A has eigenvalue 1. This is equivalent to $\boxed{q_{11}^0 = 1, q_{01}^0 = q_{01}^1 = q_{11}^1 = 0}$. Under this case, we have $V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) = R(1) - R(0)$. We could also easily use induction to show that

$$
V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) = R(1) - R(0) + V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) - V_{t+1}^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) \geq R(1) - R(0).
$$

This will give us the desired result.

Next, we assume $I - A$ is invertible, then we have

$$
\begin{aligned}
\boldsymbol{V}_t &= (R(1) - R(0))(I - A)^{-1}\mathbf{1} + (R(1) - R(0))A^{T-t}\left(\mathbf{1} - (I - A)^{-1}\mathbf{1}\right) \\
&= (R(1) - R(0))(I - A)^{-1}\mathbf{1} - (R(1) - R(0))A^{T-t+1}(I - A)^{-1}\mathbf{1} \\
&= \sum_{k=0}^{T-t}(R(1) - R(0))A^k\mathbf{1}.
\end{aligned}
$$

It is easy to see that A has two eigenvalues:

$$
\lambda_1 = \frac{\eta + \sqrt{\eta^2 - 4\delta}}{2}, \lambda_2 = \frac{\eta - \sqrt{\eta^2 - 4\delta}}{2},
$$

where $\eta = ac + (1-b)d, \delta = (a-b)cd$.

Then, we first consider the case $\boxed{\lambda_i \in \mathbb{R}}$. We have

$$
\begin{aligned}
-V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) + 2V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) &= (1,-1)\boldsymbol{V}_t\mathbf{1} \\
&= (R(1) - R(0))(c - d)\sum_{k=0}^{T-t}\frac{\lambda_1^k - \lambda_2^k}{\lambda_1 - \lambda_2}.
\end{aligned}
$$

Since under Case 4, $c \geq d$, we only need to prove that $\sum \frac{\lambda_1^k - \lambda_2^k}{\lambda_1 - \lambda_2} \geq 0$.

When $\eta \geq 0$, $|\lambda_1| \geq |\lambda_2|$. Therefore, for all k, $\lambda_1^k \geq \lambda_2^k$, which will give us the desired result.

When $\eta \leq 0, \delta \geq 0$, we have $-1 \leq \eta \leq \lambda_2 \leq \lambda_1 \leq 0$. First, we will prove $\lambda_1 + \lambda_1^2 \geq \lambda_2 + \lambda_2^2$:

$$
\begin{aligned}
\lambda_1 + \lambda_1^2 - \lambda_2 - \lambda_2^2 &= (\lambda_1 - \lambda_2)(\lambda_1 + \lambda_2 + 1) \\
&= \delta(1 + \eta)
\end{aligned}
$$

$$\geq 0.$$

Therefore, we have $\frac{\lambda_1}{\lambda_2} \leq \frac{1+\lambda_2}{1+\lambda_1}$. Then we can have for each odd k

$$\frac{\lambda_1^k}{\lambda_2^k} \leq \frac{\lambda_1}{\lambda_2} \leq \frac{1+\lambda_2}{1+\lambda_1}.$$

This means for each odd k, $\lambda_1^k + \lambda_1^{k+1} \geq \lambda_2^k + \lambda_2^{k+1}$. Then we have for all $T-t$, $\sum_{k=0}^{T-t} \lambda_1^k - \lambda_2^k \geq 0$. This will give us the desired result.

When $\eta \leq 0, \delta \leq 0$, we can numerically show that $\lambda_2 \geq -1$. Specifically, we use gradient descent with a sufficient small step size (10^{-5}) and multiple initialization by choosing the grid points to calculate the solutions of the KKT conditions. The results shows that the minimum is larger than -1. Then we have $\sum_{k=0}^{T-t} \lambda_2^k \leq 1$. We also have that $\sum_{k=0}^{T-t} \lambda_1^k \geq 1$ since $\lambda_1 \geq 0$. This will give us the desired result.

Finally, we consider the subcase where $\boxed{\lambda_i \in \mathbb{C}}$. We note that this will only happen when $\delta \geq 0$ which means $q_{11}^0 \leq q_{01}^1$. We will first rewrite the eigenvalues using its spherical form, i.e. $\lambda_i = \rho \exp(\pm i\theta)$, where

$$\rho = \sqrt{(a-b)cd}\ , \cos(\theta) = \frac{ac + (1-b)d}{2\sqrt{(a-b)cd}}.$$

Then we have

$$\begin{aligned}
&- V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,1)) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((0,0)) + 2V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0)) = (1,-1)\boldsymbol{V}_t\mathbf{1} \\
&= (R(1) - R(0))(c-d) \sum_{k=0}^{T-t} \frac{\lambda_1^k - \lambda_2^k}{\lambda_1 - \lambda_2} \\
&= (R(1) - R(0)) \frac{c-d}{1 - 2\rho\cos(\theta) + \rho^2} \left(1 - \rho^{T-t} \frac{\sin((T-t+1)\theta - \rho\sin((T-t)\theta))}{\sin(\theta)}\right).
\end{aligned}$$

Since $c - d \geq 0, 1 - 2\rho\cos(\theta) + \rho^2 = |1 - \rho\exp(i\theta)|^2 \geq 0$, we only need to show that

$$1 - \rho^{T-t} \frac{\sin((T-t+1)\theta - \rho\sin((T-t)\theta))}{\sin(\theta)} \geq 0.$$

Since $0 \leq \rho \leq 1$, we only need to show the following:

$$\begin{aligned}
\rho^{T-t} \frac{\sin((T-t+1)\theta - \sin((T-t)\theta))}{\sin(\theta)} &\leq 1 \\
\rho^{T-t} \frac{\sin((T-t+1)\theta)}{\sin(\theta)} &\leq 1.
\end{aligned}$$

We will use induction to prove the following:

$$\begin{aligned}
\rho^{k} \frac{\sin((k+1)\theta - \sin((k)\theta))}{\sin(\theta)} &\leq 1 \\
\rho^{k} \frac{\sin((k+1)\theta)}{\sin(\theta)} &\leq 1.
\end{aligned}$$

When $k = 0$, the statements trivially hold. Suppose it holds for $k = t$, when $k = t+1$, we have

$$\begin{aligned}
\rho^{t+1} \frac{\sin((t+2)\theta - \sin((t+1)\theta))}{\sin(\theta)} &= \rho^{t+1} \frac{\cos((t+1)\theta + \theta/2)}{\cos(\theta/2)} \\
&= \rho^{t+1} \frac{\cos(t\theta + \theta/2)\cos(\theta) - \sin(t\theta + \theta/2)\sin(\theta)}{\cos(\theta/2)} \\
&\leq \rho\cos(\theta) + 2\sin(\theta/2)\rho.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\rho^{t+1}\frac{\sin((t+2)\theta)}{\sin(\theta)} &= \rho^{t+1}\frac{\sin((t+1)\theta)\cos(\theta)+\sin(\theta)\cos((t+1)\theta)}{\sin(\theta)} \\
&\le \rho\cos(\theta)+\rho \\
&= \frac{1}{2}\left(ac+(1-b)d+2\sqrt{(q_{01}^1-q_{11}^1)(q_{01}^1-q_{11}^0)(q_{11}^0-q_{01}^0)}\right) \\
&\le \frac{1}{2}\left((1-q_{01}^1)q_{11}^0+q_{11}^0(q_{11}^1-q_{01}^1)+2(q_{01}^1-q_{11}^1)\right) \\
&\le \frac{1}{2}\left((1-q_{01}^1)q_{11}^0+(2-q_{11}^0)q_{01}^1\right) \\
&\le 1.
\end{aligned}$$

Now we are going to prove the theorem for arbitrary number of agents and budget. We will again prove the result under three cases.

Case 1: Suppose $q_{11}^1 \le q_{01}^1, q_{11}^0 \ge q_{01}^0, q_{11}^1 \ge q_{11}^0$.

We will first define three sets: $\mathcal{S}_1^{M-2}, \mathcal{S}_2^{M-2}, \mathcal{S}_3^{M-2}$.

$$\begin{aligned}
\mathcal{S}_1^{M-2} &:= \{\boldsymbol{s} : \boldsymbol{s} \in \mathcal{S}^{M-2}, n_{\boldsymbol{s}} \ge B\} \\
\mathcal{S}_2^{M-2} &:= \{\boldsymbol{s} : \boldsymbol{s} \in \mathcal{S}^{M-2}, n_{\boldsymbol{s}} = B-1\} \\
\mathcal{S}_3^{M-2} &:= \{\boldsymbol{s} : \boldsymbol{s} \in \mathcal{S}^{M-2}, n_{\boldsymbol{s}} \le B-2\},
\end{aligned}$$

where $n_{\boldsymbol{s}}$ denotes the number of state 0 of $\boldsymbol{s}$.

We will use backward induction to prove that

$$\forall t \le T, \forall \boldsymbol{s} \in \mathcal{S}_3^{M-2}, 0 \le V_t^*((1,0,\boldsymbol{s})) - V_t^*((0,0,\boldsymbol{s})) \le R(1)-R(0) \tag{E.37}$$

$$0 \le V_t^*((1,1,\boldsymbol{s})) - V_t^*((1,0,\boldsymbol{s})) \le R(1)-R(0) \tag{E.38}$$

$$\forall \boldsymbol{s} \in \mathcal{S}_2^{M-2}, 0 \le V_t^*((1,0,\boldsymbol{s})) - V_t^*((0,0,\boldsymbol{s})) \le \frac{R(1)-R(0)}{q_{10}^0} \tag{E.39}$$

$$0 \le V_t^*((1,1,\boldsymbol{s})) - V_t^*((1,0,\boldsymbol{s})) \le R(1)-R(0) \tag{E.40}$$

$$\forall \boldsymbol{s} \in \mathcal{S}_1^{M-2}, 0 \le V_t^*((1,0,\boldsymbol{s})) - V_t^*((0,0,\boldsymbol{s})) \le \frac{R(1)-R(0)}{q_{10}^0} \tag{E.41}$$

$$0 \le V_t^*((1,1,\boldsymbol{s})) - V_t^*((1,0,\boldsymbol{s})) \le \frac{R(1)-R(0)}{q_{10}^0} \tag{E.42}$$

$$\pi_{t+1}^{\text{greedy}} = \pi_{t+1}^*. \tag{E.43}$$

This holds trivially when $t = T$. Suppose $t \ge h+1$, this holds, we will first prove Eq. (E.43) for $t = h$ under the assumption that Eq. (E.37), Eq. (E.38), Eq. (E.39), Eq. (E.40), Eq. (E.41) and Eq. (E.42) hold for $t = h+1$. Similarly we have the following Bellman optimality equation.

$$\begin{aligned}
V_h^*\left((1,0,\boldsymbol{s}_h)\right) &= R(1)+R(0)+\sum_{s\in\boldsymbol{s}_h} R(s) \\
&+ \max_{a_{h+1}^1, a_{h+1}^2, \cdots, a_{h+1}^M} \sum_{s_{h+1}^1, s_{h+1}^2} q_{1,s_{h+1}^1}^{a_{h+1}^1} q_{0,s_{h+1}^2}^{a_{h+1}^2} \sum_{\boldsymbol{s}_{h+1}} q_{\boldsymbol{s}_h,\boldsymbol{s}_{h+1}}^{\boldsymbol{a}_{h+1}} V_{h+1}^*\left((s_{h+1}^1, s_{h+1}^2, \boldsymbol{s}_{h+1})\right).
\end{aligned}$$

We want to show that fix any policy of choosing $a_{h+1}^3, \cdots, a_{h+1}^M$, giving action 1 to agent who is in state 0 will give a higher value, which is

$$\forall \boldsymbol{a}_{h+1}, \sum_{s_{h+1}^1, s_{h+1}^2} q_{1s_{h+1}^1}^0 q_{0s_{h+1}^2}^1 \sum_{\boldsymbol{s}_{h+1}} q_{\boldsymbol{s}_h,\boldsymbol{s}_{h+1}}^{\boldsymbol{a}_{h+1}} V_{h+1}^*\left((s_{h+1}^1, s_{h+1}^2, \boldsymbol{s}_{h+1})\right)$$

$$\geq \sum_{s^1_{h+1}, s^2_{h+1}} q^1_{1s^1_{h+1}} q^0_{0s^2_{h+1}} \sum_{\boldsymbol{s}_{h+1}} q^{\boldsymbol{a}_{h+1}}_{\boldsymbol{s}_h, \boldsymbol{s}_{h+1}} V^*_{h+1}\left((s^1_{h+1}, s^2_{h+1}, \boldsymbol{s}_{h+1})\right). \tag{E.44}$$

Suppose Eq. (E.44) holds, starting from any policy of allocating action that is not greedy, we can find an agent who is in state 1 and gets action 1 while another agent who is in state 0 gets action 0. By using the result, switching the action between them will give us a better policy. Suppose there are less than B agents who are in state 0, then we can keep switching until all agents who are in state 0 are given action 1 , this is the greedy policy. Suppose there are more than B agents who are in state 0, then we can keep switching until B agents who are in state 0 get action 1, which is the greedy policy.

Now, we need to prove Eq. (E.44). Eq. (E.44) is equivalent to

$$\begin{aligned}&\left(q^0_{11} + q^1_{01} - q^1_{11} - q^0_{01}\right)\left(F_{\boldsymbol{a}_{h+1},h+1}((1,1)) - F_{\boldsymbol{a}_{h+1},h+1}((1,0))\right)\\ &+ \left(q^0_{11}q^1_{01} - q^0_{01}q^1_{11}\right)\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,1)\right) - 2F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) + F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right) \geq 0,\end{aligned} \tag{E.45}$$

where $F_{\boldsymbol{a}_{h+1},h+1}\left((s^1_{h+1}, s^2_{h+1})\right) = \sum_{\boldsymbol{s}_{h+1}} q^{\boldsymbol{a}_{h+1}}_{\boldsymbol{s}_h, \boldsymbol{s}_{h+1}} V^*_{h+1}\left((s^1_{h+1}, s^2_{h+1}, \boldsymbol{s}_{h+1})\right)$. The reason why this holds is similar to what we proved in binary agent case. We know from the assumption that Eq. (E.37)-(E.42) hold and the definition of F that

$$\begin{aligned}0 \leq F_{\boldsymbol{a}_{h+1},h+1}((1,1)) - F_{\boldsymbol{a}_{h+1},h+1}((1,0)) \leq \frac{R(1) - R(0)}{q^0_{10}}\\ 0 \leq F_{\boldsymbol{a}_{h+1},h+1}((1,0)) - F_{\boldsymbol{a}_{h+1},h+1}((0,0)) \leq \frac{R(1) - R(0)}{q^0_{10}}.\end{aligned}$$

Then, we have

$$\begin{aligned}&\left(q^0_{11} + q^1_{01} - q^1_{11} - q^0_{01}\right)\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right)\\ &+ \left(q^0_{11}q^1_{01} - q^0_{01}q^1_{11}\right)\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,1)\right) - 2F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) + F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right)\\ &\geq \left(q^0_{11} + q^1_{01} - q^1_{11} - q^0_{01}\right)\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right)\\ &- \left(q^0_{11}q^1_{01} - q^0_{01}q^1_{11}\right)\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right)\\ &= \frac{I(0) - I(1)}{R(1) - R(0)}\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right) - \frac{\left(q^1_{11}I(0) - q^1_{01}I(1)\right)}{R(1) - R(0)}\\ &\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right)\\ &= \frac{\left(q^1_{10}I(0) - q^1_{00}I(1)\right)}{R(1) - R(0)}\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right)\\ &\geq 0.\end{aligned}$$

This shows Eq. (E.44), which, by our argument, shows Eq. (E.43) for $t = h$. Next, we will first define the specific greedy policy instead of randomly choosing agents who have higher incremental reward at time h for different $\boldsymbol{s}$. When $\boldsymbol{s} \in \mathcal{S}_1^{M-2}$, we will only give action 1 to agent $3, 4, \cdots, M$; when $\boldsymbol{s} \in \mathcal{S}_2^{M-2}$, we will give one action 1 to agent $1, 2$; when $\boldsymbol{s} \in \mathcal{S}_3^{M-2}$, we will give two action 1 to agent $1, 2$.

For $\boldsymbol{s} \in \mathcal{S}_1^{M-2}$, we will show Eq. (E.41) holds for $t = h$ under the assumption that Eq.(E.37)-(E.42) hold for $t = h + 1$ and Eq. (E.43) holds for $t = h$. We have

$$\begin{aligned}V^*_h\left((1,0,\boldsymbol{s})\right) - V^*_h\left((0,0,\boldsymbol{s})\right) &= R(1) - R(0) + \left(q^0_{11} - q^0_{01}\right)\left(F^*_{h+1}\left((1,0)\right) - F^*_{h+1}\left((0,0)\right)\right)\\ &+ q^0_{01}\left(q^0_{11} - q^0_{01}\right)\left(F^*_{h+1}\left((1,1)\right) - 2F^*_{h+1}\left((1,0)\right) + F^*_{h+1}\left((0,0)\right)\right)\\ &= R(1) - R(0) + q^0_{01}\left(q^0_{11} - q^0_{01}\right)\left(F^*_{h+1}\left((1,1)\right) - F^*_{h+1}\left((1,0)\right)\right)\\ &+ q^0_{00}\left(q^0_{11} - q^0_{01}\right)\left(F^*_{h+1}\left((1,0)\right) - F^*_{h+1}\left((0,0)\right)\right)\\ &\geq 0.\end{aligned}$$

Next, we will show the RHS of Eq. (E.41) for $t = h$ under the same condition.

$$q^0_{10}\left(V^*_h\left((1,0,\boldsymbol{s})\right) - V^*_h\left((0,0,\boldsymbol{s})\right)\right) = q^0_{10}(R(1) - R(0)) + q^0_{10}\left(q^0_{11} - q^0_{01}\right)\left(F^*_{h+1}\left((1,0)\right) - F^*_{h+1}\left((0,0)\right)\right)$$

$$
\begin{aligned}
&+ q_{10}^0 q_{01}^0 \left(q_{11}^0 - q_{01}^0\right) \left(F_{h+1}^*((1,1)) - 2F_{h+1}^*((1,0)) + F_{h+1}^*((0,0))\right) \\
&= q_{10}^0 (R(1) - R(0)) + q_{10}^0 q_{01}^0 \left(q_{11}^0 - q_{01}^0\right) \left(F_{h+1}^*((1,1)) - F_{h+1}^*((1,0))\right) \\
&+ q_{10}^0 q_{00}^0 \left(q_{11}^0 - q_{01}^0\right) \left(F_{h+1}^*((1,0)) - F_{h+1}^*((0,0))\right) \\
&\leq q_{10}^0 (R(1) - R(0)) + q_{01}^0 q_{11}^0 (R(1) - R(0)) + q_{00}^0 q_{11}^0 (R(1) - R(0)) \\
&= R(1) - R(0).
\end{aligned}
$$

Meanwhile, when $t = h$, again under the same condition, the LHS of Eq. (E.42) satisfies

$$
\begin{aligned}
V_h^*((1,1,\boldsymbol{s})) - V_h^*((1,0,\boldsymbol{s})) &= R(1) - R(0) + \left(q_{11}^0 - q_{01}^0\right) \left(F_{h+1}^*((1,0)) - F_{h+1}^*((0,0))\right) \\
&+ q_{11}^0 \left(q_{11}^0 - q_{01}^0\right) \left(F_{h+1}^*((1,1)) - 2F_{h+1}^*((1,0)) + F_{h+1}^*((0,0))\right) \\
&= R(1) - R(0) + q_{11}^0 \left(q_{11}^0 - q_{01}^0\right) \left(F_{h+1}^*((1,1)) - F_{h+1}^*((1,0))\right) \\
&+ q_{10}^0 \left(q_{11}^0 - q_{01}^0\right) \left(F_{h+1}^*((1,0)) - F_{h+1}^*((0,0))\right) \\
&\geq 0.
\end{aligned}
$$

With the same reasoning, we can show the RHS of Eq. (E.42) under the same condition for $t = h$.

For $\boldsymbol{s} \in \mathcal{S}_3^{M-2}$, we have

$$
\begin{aligned}
V_h^*((1,0,\boldsymbol{s})) - V_h^*((0,0,\boldsymbol{s})) &= R(1) - R(0) + \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,0)) - F_{h+1}^*((0,0))\right) \\
&+ q_{01}^1 \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,1)) - 2F_{h+1}^*((1,0)) + F_{h+1}^*((0,0))\right) \\
&= R(1) - R(0) + q_{01}^1 \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,1)) - F_{h+1}^*((1,0))\right) \\
&+ q_{00}^1 \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,0)) - F_{h+1}^*((0,0))\right) \\
&= \frac{q_{10}^0}{q_{10}^0} \left(R(1) - R(0) + q_{01}^1 \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,1)) - F_{h+1}^*((1,0))\right)\right) \\
&+ \frac{q_{10}^0}{q_{10}^0} \left(q_{00}^1 \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,0)) - F_{h+1}^*((0,0))\right)\right) \\
&\geq \frac{q_{10}^0}{q_{10}^0} \left(R(1) - R(0) + \frac{q_{01}^1}{q_{10}^0} \left(q_{11}^1 - q_{01}^1\right) (R(1) - R(0))\right) \\
&+ \frac{q_{00}^1}{q_{10}^0} \left(q_{11}^1 - q_{01}^1\right) (R(1) - R(0)) \\
&= \frac{1}{q_{10}^0} \left(q_{10}^0 + q_{00}^1 q_{11}^1 - q_{00}^1 q_{01}^1\right) (R(1) - R(0)) \\
&+ \frac{1}{q_{10}^0} \left(q_{01}^1 \left(q_{11}^1 - q_{01}^1\right)\right) (R(1) - R(0)) \\
&\geq \frac{1}{q_{10}^0} \left(1 - q_{01}^1 + q_{11}^1 - q_{11}^0\right) (R(1) - R(0)) \\
&\geq 0.
\end{aligned}
$$

Meanwhile, we also have

$$
\begin{aligned}
V_h^*((1,1,\boldsymbol{s})) - V_h^*((1,0,\boldsymbol{s})) &= R(1) - R(0) + \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,0)) - F_{h+1}^*((0,0))\right) \\
&+ q_{11}^1 \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,1)) - 2F_{h+1}^*((1,0)) + F_{h+1}^*((0,0))\right) \\
&= R(1) - R(0) + q_{11}^1 \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,1)) - F_{h+1}^*((1,0))\right) \\
&+ q_{10}^1 \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,0)) - F_{h+1}^*((0,0))\right) \\
&= \frac{q_{10}^0}{q_{10}^0} \left(R(1) - R(0) + q_{11}^1 \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,1)) - F_{h+1}^*((1,0))\right)\right) \\
&+ \frac{q_{10}^0}{q_{10}^0} \left(q_{10}^1 \left(q_{11}^1 - q_{01}^1\right) \left(F_{h+1}^*((1,0)) - F_{h+1}^*((0,0))\right)\right) \\
&\geq \frac{q_{10}^0}{q_{10}^0} \left(R(1) - R(0) + \frac{q_{11}^1}{q_{10}^0} \left(q_{11}^1 - q_{01}^1\right) (R(1) - R(0))\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{q_{10}^1}{q_{10}^0}\left(q_{11}^1 - q_{01}^1\right)(R(1)-R(0)) \\
&= \frac{1}{q_{10}^0}\left(q_{10}^0 + q_{10}^1 q_{11}^1 - q_{10}^1 q_{01}^1\right)(R(1)-R(0)) \\
&+ \frac{1}{q_{10}^0}\left(q_{11}^1\left(q_{11}^1 - q_{01}^1\right)\right)(R(1)-R(0)) \\
&\geq \frac{1}{q_{10}^0}\left(1 - q_{01}^1 + q_{11}^1 - q_{11}^0\right)(R(1)-R(0)) \\
&\geq 0.
\end{aligned}
$$

For $\boldsymbol{s} \in \mathcal{S}_2^{M-2}$, we can use the same reasoning used in the two agent case to get the result.

Case 2: $q_{11}^0 \leq q_{01}^0, q_{11}^1 \leq q_{01}^1$.

First we consider the sub-case where $\boxed{\boldsymbol{q_{11}^1 q_{01}^0 \leq q_{01}^1 q_{11}^0}}$. We will use backward induction to prove the following statements:

$$
\forall t \leq T, 0 \leq V_t^*\left((1,0,\boldsymbol{s})\right) - V_t^*((0,0,\boldsymbol{s})) \leq R(1) - R(0) \tag{E.46}
$$

$$
0 \leq V_t^*\left((1,1,\boldsymbol{s})\right) - V_t^*((1,0,\boldsymbol{s})) \leq R(1) - R(0) \tag{E.47}
$$

$$
\pi_{t+1}^{\text{greedy}} = \pi_{t+1}^*. \tag{E.48}
$$

They hold trivially when $t = T$. Suppose they hold when $t \geq h+1$, when $t = h$, we will first prove Eq. (E.48) for $t = h$ under the assumption that Eq. (E.46) and Eq. (E.24) hold for $t = h+1$. Using the same reasoning, we only need to show Eq. (E.45). Again, from the assumptions and the definition of F, we know that

$$
\begin{aligned}
0 \leq F_{\boldsymbol{a}_{h+1},h+1}((1,1)) - F_{\boldsymbol{a}_{h+1},h+1}((1,0)) \leq R(1) - R(0) \\
0 \leq F_{\boldsymbol{a}_{h+1},h+1}((1,0)) - F_{\boldsymbol{a}_{h+1},h+1}((0,0)) \leq R(1) - R(0).
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right) \\
&+ \left(q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1\right)\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,1)\right) - 2F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) + F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right) \\
&\geq \left(q_{11}^0 + q_{01}^1 - q_{11}^1 - q_{01}^0\right)\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right) \\
&- \left(q_{11}^0 q_{01}^1 - q_{01}^0 q_{11}^1\right)\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right) \\
&= \frac{I(0) - I(1)}{R(1) - R(0)}\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right) - \frac{\left(q_{11}^1 I(0) - q_{01}^1 I(1)\right)}{R(1) - R(0)} \\
&\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right) \\
&= \frac{\left(q_{10}^1 I(0) - q_{00}^1 I(1)\right)}{R(1) - R(0)}\left(F_{\boldsymbol{a}_{h+1},h+1}\left((1,0)\right) - F_{\boldsymbol{a}_{h+1},h+1}\left((0,0)\right)\right) \\
&\geq 0.
\end{aligned}
$$

This will give us the desired result. Assuming Eq. (E.48) holds for $t = h$, we will select the greedy policy for $\mathcal{S}_i^{M-2}$ as we did in case 1. Again, the only case that is different from the 2-agent case is when $\boldsymbol{s} \in \mathcal{S}_1^{M-2}, \mathcal{S}_3^{M-2}$. Therefore, we will only show Eq. (E.46) and Eq. (E.47) for those two cases when $t = h$ under the assumption that Eq. (E.46) and Eq. (E.47) hold for $t = h+1$ and Eq. (E.48) holds for $t = h$. We will first prove the RHS of Eq. (E.46). When $\boldsymbol{s} \in \mathcal{S}_1^{M-2}$, we have

$$
\begin{aligned}
V_h^*\left((1,0,\boldsymbol{s})\right) - V_h^*\left((0,0,\boldsymbol{s})\right) &= R(1) - R(0) + \left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,0)\right) - F_{h+1}^*\left((0,0)\right)\right) \\
&+ q_{01}^0\left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,1)\right) - 2F_{h+1}^*\left((1,0)\right) + F_{h+1}^*\left((0,0)\right)\right) \\
&= R(1) - R(0) + q_{01}^0\left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,1)\right) - F_{h+1}^*\left((1,0)\right)\right) \\
&+ q_{00}^0\left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,0)\right) - F_{h+1}^*\left((0,0)\right)\right)
\end{aligned}
$$

$$\leq R(1) - R(0).$$

Similarly, when $\boldsymbol{s} \in \mathcal{S}_3^{M-2}$, we have

$$\begin{aligned}
V_h^*\left((1,0,\boldsymbol{s})\right) - V_h^*\left((0,0,\boldsymbol{s})\right) &= R(1) - R(0) + \left(q_{11}^1 - q_{01}^1\right)\left(F_{h+1}^*\left((1,0)\right) - F_{h+1}^*\left((0,0)\right)\right) \\
&+ q_{01}^1\left(q_{11}^1 - q_{01}^1\right)\left(F_{h+1}^*\left((1,1)\right) - 2F_{h+1}^*\left((1,0)\right) + F_{h+1}^*\left((0,0)\right)\right) \\
&= R(1) - R(0) + q_{01}^1\left(q_{11}^1 - q_{01}^1\right)\left(F_{h+1}^*\left((1,1)\right) - F_{h+1}^*\left((1,0)\right)\right) \\
&+ q_{00}^1\left(q_{11}^1 - q_{01}^1\right)\left(F_{h+1}^*\left((1,0)\right) - F_{h+1}^*\left((0,0)\right)\right) \\
&\leq R(1) - R(0).
\end{aligned}$$

Now, we are going to prove the LHS of Eq. (E.46). When $\boldsymbol{s} \in \mathcal{S}_1^{M-2}$, we have

$$\begin{aligned}
V_h^*\left((1,0,\boldsymbol{s})\right) - V_h^*\left((0,0,\boldsymbol{s})\right) &= R(1) - R(0) + \left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,0)\right) - F_{h+1}^*\left((0,0)\right)\right) \\
&+ q_{01}^0\left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,1)\right) - 2F_{h+1}^*\left((1,0)\right) + F_{h+1}^*\left((0,0)\right)\right) \\
&= R(1) - R(0) + q_{01}^0\left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,1)\right) - F_{h+1}^*\left((1,0)\right)\right) \\
&+ q_{00}^0\left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,0)\right) - F_{h+1}^*\left((0,0)\right)\right) \\
&\geq R(1) - R(0) + q_{01}^0\left(q_{11}^0 - q_{01}^0\right)\left(R(1) - R(0)\right) + q_{00}^0\left(q_{11}^0 - q_{01}^0\right)\left(R(1) - R(0)\right) \\
&= (1 + q_{11}^0 + q_{00}^0)(R(1) - R(0)) \\
&\geq 0.
\end{aligned}$$

Using the same calculation, we can show the case for $\boldsymbol{s} \in \mathcal{S}_3^{M-2}$ and Eq. (E.47).

Secondly, we consider the sub-case where $\boxed{\boldsymbol{q_{11}^1 q_{01}^0 \geq q_{01}^1 q_{11}^0}}$. We want to use backward induction to prove the following statements:

$$\forall t \leq T, 0 \leq V_t^*\left((1,0,\boldsymbol{s})\right) - V_t^*\left((0,0,\boldsymbol{s})\right) \leq R(1) - R(0) \tag{E.49}$$

$$0 \leq V_t^*\left((1,1,\boldsymbol{s})\right) - V_t^*\left((1,0,\boldsymbol{s})\right) \leq R(1) - R(0) \tag{E.50}$$

$$V_t^*\left((1,1,\boldsymbol{s})\right) - V_t^*\left((1,0,\boldsymbol{s})\right) \leq V_t^*\left((1,0,\boldsymbol{s})\right) - V_t^*\left((0,0,\boldsymbol{s})\right) \tag{E.51}$$

$$\pi_{t+1}^{\text{greedy}} = \pi_{t+1}^*. \tag{E.52}$$

When $t = T$, the statements hold trivially. Suppose this holds for $t \geq h+1$, when $t = h$, we will first show Eq. (E.52) under the assumption that Eq. (E.49)-(E.51) hold for $t = h+1$. Using the same reasoning, we only need to show Eq. (E.45). Again, from the above assumptions and the definition of F, we know that

$$\begin{aligned}
0 &\leq F_{\boldsymbol{a}_{h+1},h+1}((1,1)) - F_{\boldsymbol{a}_{h+1},h+1}((1,0)) \leq R(1) - R(0) \\
0 &\leq F_{\boldsymbol{a}_{h+1},h+1}((1,0)) - F_{\boldsymbol{a}_{h+1},h+1}((0,0)) \leq R(1) - R(0).
\end{aligned}$$

Then combined with the fact that $q_{11}^1 q_{01}^0 \geq q_{01}^1 q_{11}^0$, it is easy to see that Eq. (E.45) holds.

Using the same logic used in the first sub-case, we can prove Eq. (E.49),(E.50) for $t = h$. The only thing we need to show is Eq. (E.51) for $t = h$. Again, when $\boldsymbol{s} \in \mathcal{S}_2^{M-2}$, this is the same as the two agent case. Therefore, we will only prove the other two cases under the assumption that Eq. (E.49)-(E.51) hold for $t = h+1$ and Eq. (E.52) holds for $t = h$. We will prove the result for $\boldsymbol{s} \in \mathcal{S}_1^{M-2}$, the other cases will hold for similar reasons. We have

$$\begin{aligned}
&V_h^*\left((1,1,\boldsymbol{s})\right) - 2V_h^*\left((1,0,\boldsymbol{s})\right) + V_h^*\left((0,0,\boldsymbol{s})\right) \\
&= \left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,0)\right) - F_{h+1}^*\left((0,0)\right)\right) \\
&+ q_{11}^0\left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,1)\right) - 2F_{h+1}^*\left((1,0)\right) + F_{h+1}^*\left((0,0)\right)\right) \\
&- \left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,0)\right) - F_{h+1}^*\left((0,0)\right)\right) \\
&- q_{01}^0\left(q_{11}^0 - q_{01}^0\right)\left(F_{h+1}^*\left((1,1)\right) - 2F_{h+1}^*\left((1,0)\right) + F_{h+1}^*\left((0,0)\right)\right) \\
&= \left(q_{11}^0 - q_{01}^0\right)^2\left(F_{h+1}^*\left((1,1)\right) - 2F_{h+1}^*\left((1,0)\right) + F_{h+1}^*\left((0,0)\right)\right) \\
&\leq 0,
\end{aligned}$$

where the last inequality holds because of the assumption that Eq. (E.51) holds for $t = h+1$.

Case 3: $q_{11}^1 \geq q_{01}^1, q_{11}^0 \geq q_{01}^0$. Similar to the two agent case, we can prove the following statements using backward induction:

$$\forall t \leq T, \frac{R(1)-R(0)}{q_{10}^1+q_{01}^1}\left(1-\left(q_{11}^1-q_{01}^1\right)^{T-t+1}\right) \leq V_t^*\left((1,0,\boldsymbol{s})\right)-V_t^*((0,0,\boldsymbol{s})) \tag{E.53}$$

$$\frac{R(1)-R(0)}{q_{10}^0+q_{01}^0}\left(1-\left(q_{11}^0-q_{01}^0\right)^{T-t+1}\right) \geq V_t^*\left((1,0,\boldsymbol{s})\right)-V_t^*((0,0,\boldsymbol{s})) \tag{E.54}$$

$$\frac{R(1)-R(0)}{q_{10}^1+q_{01}^1}\left(1-\left(q_{11}^1-q_{01}^1\right)^{T-t+1}\right) \leq V_t^*\left((1,0,\boldsymbol{s})\right)-V_t^*((0,0,\boldsymbol{s})) \tag{E.55}$$

$$\frac{R(1)-R(0)}{q_{10}^0+q_{01}^0}\left(1-\left(q_{11}^0-q_{01}^0\right)^{T-t+1}\right) \geq V_t^*\left((1,0,\boldsymbol{s})\right)-V_t^*((0,0,\boldsymbol{s})) \tag{E.56}$$

$$\pi_{t+1}^{\text{greedy}} = \pi_{t+1}^*. \tag{E.57}$$

The proof will be exactly using the idea used in 2 agent case and the calculation for multiple agents case discussed above, thus will be omitted.

Case 4: Suppose $q_{11}^1 \leq q_{01}^1, q_{11}^0 \geq q_{01}^0, q_{11}^1 \leq q_{11}^0$. We can use the same way used in 2 agent case to directly calculate $2V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,0,\boldsymbol{s})) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((0,0,\boldsymbol{s})) - V_t^{\boldsymbol{\pi}^{\text{greedy}}}((1,1,\boldsymbol{s}))$ and have the result. □

F TECHNICAL DETAILS IN SECTION 4

We will continue to use the notation defined in Appendix E.

F.1 Proof of Lemma 4.3

Lemma 4.3 (Existence of Sufficient Good Agent-State Pairs)**.** *Under Assumptions 4.2 and 4.1, for any joint state $\boldsymbol{s} \in \mathcal{S}^M$, the number of good agent-state pairs observed in $\boldsymbol{s}$ is at least B. That is, $|\{m \in [M] : (m, s^m) \in \mathcal{G}\}| \geq B$.*

Proof. By Bellman equation, we have

$$V_0^{\boldsymbol{\pi}^b}(\boldsymbol{s}_0) = \mathbb{E}\left[\sum_{m=1}^M c_0^m\right] + \mathbb{E}_{\pi_1^b}\left[V_1^{\boldsymbol{\pi}^b}(\boldsymbol{s}_1)\right].$$

Since $V_0^{\boldsymbol{\pi}^b}(\boldsymbol{s}_0) = 0, V_1^{\boldsymbol{\pi}^b}(\boldsymbol{s}_1) \geq 0, P_{\boldsymbol{\pi}^b}(\boldsymbol{s}_0, \boldsymbol{s}_1) > 0$, we have $V_1^{\boldsymbol{\pi}^b}(\boldsymbol{s}_1) = 0$ for all $\boldsymbol{s}_1 \in \mathcal{S}^M$. Since again by Bellman equation, we have $V_1^{\boldsymbol{\pi}^b}(\boldsymbol{s}_1) \geq \sum_{m=1}^M c_1^m\left(\pi_2^b(s_1^m)\right)$, we have this sum should be 0. This will indicate that there are enough good arms for state $\boldsymbol{s}_1$. This will give us the desired result. □

Before proving Theorem 4.4, we first introduce Lemma F.1 which decomposes the cumulative regret to per-timestep regret.

Lemma F.1. *For any $t \leq T$, let per-timestep regret $Regret_t(\pi, \mathcal{H}_t)$ for the policy $\boldsymbol{\pi}$ be defined as:*

$$\begin{aligned} Regret_t(\boldsymbol{\pi}, \mathcal{H}_t) = &\mathbb{E}_{\pi_{t+1}}\left[\sum_{m=1}^M c_t^m(a_{t+1}^m) \mid \mathcal{H}_t\right] + \mathbb{E}_{\pi_{t+1}}\left[\mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^M \sum_{k=t+1}^T c_k^m\left(a_{t+1}^m\right) | \boldsymbol{r}_{t+1}, \boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}, \mathcal{H}_t\right]\right] \\ &- \mathbb{E}_{\pi_{t+1}^b}\left[\sum_{m=1}^M c_t^m\left(a_{t+1}^m\right) \mid \mathcal{H}_t\right] - \mathbb{E}_{\pi_{t+1}^b}\left[\mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^M \sum_{k=t+1}^T c_k^m\left(a_{t+1}^m\right) | \boldsymbol{r}_{t+1}, \boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}, \mathcal{H}_t\right]\right]. \end{aligned}$$

Then the cumulative regret $\mathcal{R}(\pi)$ can be decomposed as following:

$$\mathcal{R}(\pi) = \mathbb{E}\left[\sum_{t=0}^T Regret_t\left(\pi, \mathcal{H}_t\right)\right].$$

F.2 Proof of Lemma F.1

Proof. We will prove the result iteratively from $t = 1$ to T. All we need is calculation. We have

$$\begin{aligned}
\mathcal{R}(\boldsymbol{\pi}) &= \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{m=1}^{M}\sum_{t=0}^{T} c_t^m(a_{t+1}^m)|\boldsymbol{s}_0\right] - \mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^{M}\sum_{t=0}^{T} c_t^m(a_{t+1}^m)|\boldsymbol{s}_0\right] \\
&= \mathbb{E}_{\pi_1}\left[\sum_{m=1}^{M} c_0^m(a_1^m) \mid \mathcal{H}_0\right] + \mathbb{E}_{\pi_1}\left[\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{m=1}^{M}\sum_{k=1}^{T} c_k^m\left(a_{k+1}^m\right)|\boldsymbol{r}_1, \boldsymbol{s}_1, \boldsymbol{a}_1, \mathcal{H}_0\right]\right] \\
&- \mathbb{E}_{\pi_1}\left[\sum_{m=1}^{M} c_0^m(a_1^m) \mid \mathcal{H}_0\right] - \mathbb{E}_{\pi_1}\left[\mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^{M}\sum_{k=1}^{T} c_k^m\left(a_{k+1}^m\right)|\boldsymbol{r}_1, \boldsymbol{s}_1, \boldsymbol{a}_1, \mathcal{H}_0\right]\right] \\
&+ \mathbb{E}_{\pi_1}\left[\sum_{m=1}^{M} c_0^m(a_1^m) \mid \mathcal{H}_0\right] + \mathbb{E}_{\pi_1}\left[\mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^{M}\sum_{k=1}^{T} c_k^m\left(a_{k+1}^m\right)|\boldsymbol{r}_1, \boldsymbol{s}_1, \boldsymbol{a}_1, \mathcal{H}_0\right]\right] \\
&- \mathbb{E}_{\pi_1^b}\left[\sum_{m=1}^{M} c_0^m\left(a_1^m\right) \mid \mathcal{H}_0\right] - \mathbb{E}_{\pi_1^b}\left[\mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^{M}\sum_{k=1}^{T} c_k^m\left(a_{k+1}^m\right)|\boldsymbol{r}_1, \boldsymbol{s}_1, \boldsymbol{a}_1, \mathcal{H}_0\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{m=1}^{M}\sum_{k=1}^{T} c_k^m\left(a_{k+1}^m\right)|\mathcal{H}_1\right] - \mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^{M}\sum_{k=1}^{T} c_k^m\left(a_{k+1}^m\right)|\mathcal{H}_1\right]\right] \\
&+ \operatorname{Regret}\left(\pi, \mathcal{H}_0\right).
\end{aligned}$$

We can continue this trick to $\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{m=1}^{M}\sum_{k=1}^{T} c_k^m\left(a_{k+1}^m\right)|\mathcal{H}_1\right] - \mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^{M}\sum_{k=1}^{T} c_k^m\left(a_{k+1}^m\right)|\mathcal{H}_1\right]$, this will give us the desired result. □

F.3 Proof of Theorem 4.4

Theorem 4.4 (Regret Upperbound)**.** *Assume r_t^m is 1-subgaussian random variable for all $t \in [T]$. Let $D := \max\left\{\frac{8M}{B}, \frac{8M}{M-B}, 2\right\}, p = 1, \eta = \Delta$. Under Assump.s 4.2 and 4.1, the regret, $\mathcal{R}(\pi)$, defined by Def. 2.2 with the loss function l^m given by Eq. (3), satisfies:*

$$\begin{aligned}
\mathcal{R}(\pi) = \min\Bigg\{ & \mathcal{O}\left(\frac{|\mathcal{S}|\left(M|\mathcal{S}| - |\mathcal{G}|\right) MD}{\Delta} \log T\right), \\
& \mathcal{O}\left(\sqrt{MDB\big(|\mathcal{S}|\left(M|\mathcal{S}| - |\mathcal{G}|\right)\big) T \log T}\right)\Bigg\}.
\end{aligned} \tag{4}$$

Before we prove the theorem, we first provide a technical lemma showing the concentration property of the empirical mean estimator $\hat{I}_t^m(s)$, which is a direct result from Howard et al. 2021.

Lemma F.2. *For any $\alpha > 0$, $0 \le t \le T$, we have*

$$\mathbb{P}\left(\left|\hat{I}_t^m(s) - I^m(s)\right| \ge 1.7\sqrt{\frac{\log\log(2N_{s,m}(t)) + 0.72\log(10.4/\alpha)}{N_{s,m}(t)}}\right) \le \alpha.$$

Proof. The proof will proceed by first showing that the cumulative regret can be decomposed into per-timestep regret $\operatorname{Regret}_t(\boldsymbol{\pi}, \mathcal{H}_t)$. Then we will use the standard technique to bound the number of times suboptimal agents are given action 1.

By Lemma F.1, we only need to bound $\operatorname{Regret}_t(\pi, \mathcal{H}_t)$. We have

$$\operatorname{Regret}_t(\pi, \mathcal{H}_t) = \mathbb{E}_{\pi_{t+1}}\left[\sum_{m=1}^{M} c_t^m(a_{t+1}^m) \mid \mathcal{H}_t\right] + \mathbb{E}_{\pi_{t+1}}\left[\mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^{M}\sum_{k=t+1}^{T} c_k^m\left(a_{t+1}^m\right)|\boldsymbol{r}_{t+1}, \boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}, \mathcal{H}_t\right]\right]$$

$$
\begin{aligned}
&- \mathbb{E}_{\pi^b_{t+1}}\left[\sum_{m=1}^{M} c_t^m\left(a_{t+1}^m\right) \mid \mathcal{H}_t\right] - \mathbb{E}_{\pi^b_{t+1}}\left[\mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^{M}\sum_{k=t+1}^{T} c_k^m\left(a_{t+1}^m\right) | \boldsymbol{r}_{t+1}, \boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}, \mathcal{H}_t\right]\right] \\
&= \mathbb{E}_{\pi_{t+1}}\left[\sum_{m=1}^{M} c_t^m(a_{t+1}^m) \mid \mathcal{H}_t\right] + \mathbb{E}_{\pi_{t+1}}\left[\mathbb{E}_{\boldsymbol{\pi}^b}\left[\sum_{m=1}^{M}\sum_{k=t+1}^{T} c_k^m\left(a_{t+1}^m\right) | \boldsymbol{r}_{t+1}, \boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}, \mathcal{H}_t\right]\right] \\
&\leq \mathbb{E}\left[\sum_{I^m(s)\leq\gamma} \Delta_s^m \mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = 1\right\}\right] + \mathbb{E}_{\pi_{t+1}}\left[V_{t+1}^{\boldsymbol{\pi}^b}\left(\boldsymbol{s}_{t+1}\right) \mid \boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}, \mathcal{H}_t\right] \\
&= \mathbb{E}\left[\sum_{I^m(s)\leq\gamma} \Delta_s^m \mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = 1\right\}\right],
\end{aligned}
$$

where the second and the third equality holds by Lemma 4.3. Then we have the cumulative regret is upper bounded by

$$
\sum_{I^m(s)\leq\gamma} \Delta_s^m \mathbb{E}\left[\sum_{t=0}^{T} \mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s\right\}\right].
$$

We have

$$
\mathbb{E}\left[\sum_{t=0}^{T} \mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s\right\}\right] = \mathbb{E}\left[\sum_{t=0}^{T} \mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \text{random}\right\}\right] \tag{F.58}
$$

$$
+ \mathbb{E}\left[\sum_{t=0}^{T} \mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \text{not random}\right\}\right], \tag{F.59}
$$

where random stands for doing random exploration. We will bound the two terms separately. For Eq. (F.58), we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T} \mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \text{random}\right\}\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T} \mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \text{random}, \exists(m', s'), s_t^{m'} = s', N_{s',m'}^{(t)} \leq N_{s_t^{m''},m''}^{(t)}, \forall m''\right\}\right] \\
&\leq \sum_{m'}\sum_{s'} \mathbb{E}\left[\sum_{t=0}^{T} \mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \text{random}, s_t^{m'} = s', N_{s',m'}^{(t)} \leq N_{s_t^{m''},m''}^{(t)}, \forall m''\right\}\right] \\
&\leq \sum_{m'}\sum_{s'} \mathbb{E}\left[\sum_{t=0}^{T} \mathbb{I}\left\{\text{random}, s_t^{m'} = s', N_{s',m'}^{(t)} \leq N_{s_t^{m''},m''}^{(t)}, \forall m''\right\}\right] \\
&\leq \sum_{m'}\sum_{s'} \mathbb{E}\left[\sum_{k=0}^{\infty} \mathbb{I}\left\{\text{random}, N_{s',m'}^{(\tau_k^{m'}(s'))} \leq N_{s_t^{m''},m''}^{(\tau_k^{m'}(s'))}, \forall m'', \tau_k^{m'}(s') \leq T\right\}\right] \\
&\leq \sum_{m'}\sum_{s'} \mathbb{E}\left[\sum_{k=0}^{T} \mathbb{I}\left\{\text{random}, N_{s',m'}^{(\tau_k^{m'}(s'))} \leq N_{s_t^{m''},m''}^{(\tau_k^{m'}(s'))}, \forall m''\right\}\right] \\
&= \sum_{m'}\sum_{s'}\sum_{k=0}^{T} \mathbb{P}\left(\text{random} \mid N_{s',m'}^{(\tau_k^{m'}(s'))} \leq N_{s_t^{m''},m''}^{(\tau_k^{m'}(s'))}, \forall m''\right) \mathbb{P}\left(N_{s',m'}^{(\tau_k^{m'}(s'))} \leq N_{s_t^{m''},m''}^{(\tau_k^{m'}(s'))}, \forall m''\right) \\
&\leq \sum_{m'}\sum_{s'}\sum_{k=0}^{T} \mathbb{P}\left(\text{random} \mid N_{s',m'}^{(\tau_k^{m'}(s'))} \leq N_{s_t^{m''},m''}^{(\tau_k^{m'}(s'))}, \forall m''\right) \\
&\leq \sum_{m'}\sum_{s'}\sum_{k=0}^{T} \frac{D}{k\Delta^2}
\end{aligned}
$$

$$\leq SMD\frac{\log(T)}{\Delta^2}, \tag{F.60}$$

where $\tau_k^m(s) := \inf\{t : N_{s,m}^{(t)} = k\}$.

For Eq. (F.59), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T}\mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \text{not random}\right\}\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T}\mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \text{not random}, \exists (m', s') \in \mathcal{G}, \hat{I}_t^m(s) \geq \hat{I}_t^{m'}(s'), s_t^{m'} = s'\right\}\right] \\
&\leq \sum_{(m',s')\in\mathcal{G}} \mathbb{E}\left[\sum_{t=0}^{T}\mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \hat{I}_t^m(s) \geq \hat{I}_t^{m'}(s'), s_t^{m'} = s'\right\}\right]
\end{aligned}$$

For each $(m', s') \in \mathcal{G}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T}\mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \hat{I}_t^m(s) \geq \hat{I}_t^{m'}(s'), s_t^{m'} = s'\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=0}^{T}\mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \hat{I}_t^m(s) \geq \gamma + \frac{\Delta}{2}, s_t^{m'} = s'\right\}\right] \qquad \text{(F.61)} \\
&+ \mathbb{E}\left[\sum_{t=0}^{T}\mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \hat{I}_t^{m'}(s') \leq \gamma + \frac{\Delta}{2}, s_t^{m'} = s'\right\}\right]. \qquad \text{(F.62)}
\end{aligned}$$

Again, for Eq. (F.61), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T}\mathbb{I}\left\{a_{t+1}^m = 1, s_t^m = s, \hat{I}_t^m(s) \geq \gamma + \frac{\Delta}{2}, s_t^{m'} = s'\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=0}^{T}\mathbb{I}\left\{s_t^m = s, \hat{I}_t^m(s) \geq \gamma + \frac{\Delta}{2}\right\}\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty}\mathbb{I}\left\{s_{\tau_k^m(s)}^m = s, \hat{I}_{\tau_k^m(s)}^m(s) \geq \gamma + \frac{\Delta}{2}\right\}, \tau_k^m(s) \leq T\right] \\
&\leq \mathbb{E}\left[\sum_{k=0}^{T}\mathbb{I}\left\{\hat{I}_{\tau_k^m(s)}^m(s) \geq \gamma + \frac{\Delta}{2}\right\}\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{T}\mathbb{I}\left\{N_{s,m}^{(\tau_k^m(s))}(r) \in [k_1 t_{\text{random}}^k, k_2 t_{\text{random}}^k], \hat{I}_{\tau_k^m(s)}^m(s) \geq \gamma + \frac{\Delta}{2}\right\}\right] \\
&+ \mathbb{E}\left[\sum_{k=0}^{T}\mathbb{I}\left\{N_{s,m}^{(\tau_k^m(s))}(r) \notin [k_1 t_{\text{random}}^k, k_2 t_{\text{random}}^k], \hat{I}_{\tau_k^m(s)}^m(s) \geq \gamma + \frac{\Delta}{2}\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{k=0}^{T}\mathbb{I}\left\{N_{s,m}^{(\tau_k^m(s))}(r) \in [k_1 t_{\text{random}}^k, k_2 t_{\text{random}}^k], \hat{I}_{\tau_k^m(s)}^m(s) \geq \gamma + \frac{\Delta}{2}, t_{\text{random}}^k \geq k_3^k\right\}\right] \\
&+ \mathbb{E}\left[\sum_{k=0}^{T}\mathbb{I}\left\{N_{s,m}^{(\tau_k^m(s))}(r) \notin [k_1 t_{\text{random}}^k, k_2 t_{\text{random}}^k]\right\}\right] + \mathbb{E}\left[\sum_{k=0}^{T}\mathbb{I}\left\{t_{\text{random}}^k \leq k_3^k\right\}\right],
\end{aligned}$$

where $N_{s,m}^{(\tau_k^m(s))}(r) := \sum_{l=1}^{k}\mathbb{I}\left\{a_{\tau_l^m(s)}^m = 0, s_{\tau_l^m(s)}^m = s, \text{random}\right\}$ denotes the total number of times that action 0 is given to agent m when m is in state s and the algorithm is doing random exploration, $t_{\text{random}}^k := \sum_{l=1}^{m}\sum_{l=1}^{k}\mathbb{I}\{\text{random}\}$ denotes the total number of times when agent m is in state s and the algorithm is doing random explorations. We let $k_1 = \max\left\{\frac{M-B}{2M}, \frac{2M-3B}{2M}\right\}, k_2 = \min\left\{\frac{3(M-B)}{2M}, \frac{2M-B}{2(M-B)}, k_3^k = \frac{C\log(k)}{2\min\{\Delta^2, 1/2\}}\right\}$

For the first term, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{T}\mathbb{I}\left\{N_{s,m}^{(\tau_k^m(s))}(r)\in[k_1t_{\text{random}}^k,k_2t_{\text{random}}^k],\hat{I}_{\tau_k^m(s)}^m(s)\geq\gamma+\frac{\Delta}{2},t_{\text{random}}^k\geq k_3^k\right\}\right]\\
&\leq\mathbb{E}\left[\sum_{k=0}^{T}\mathbb{I}\left\{N_{s,m}^{(\tau_k^m(s))}(r)\in[k_1t_{\text{random}}^k,k_2t_{\text{random}}^k],N_{s,m,1}^{(\tau_k^m(s)}\geq(1-k_2)t_{\text{random}}^k,\hat{I}_{\tau_k^m(s)}^m(s)\geq\gamma+\frac{\Delta}{2},t_{\text{random}}^k\geq k_3^k\right\}\right]\\
&\leq\mathbb{E}\left[\sum_{k=0}^{T}\mathbb{I}\left\{N_{s,m}^{(\tau_k^m(s))}(r)\geq k_1k_3^k,N_{s,m,1}^{(\tau_k^m(s)}\geq(1-k_2)k_3^k,\hat{I}_{\tau_k^m(s)}^m(s)\geq\gamma+\frac{\Delta}{2}\right\}\right]\\
&\leq\mathbb{E}\left[\sum_{k=0}^{T}\mathbb{I}\left\{N_{s,m}^{(\tau_k^m(s))}(r)\geq k_1k_3^k,N_{s,m,1}^{(\tau_k^m(s)}\geq(1-k_2)k_3^k,\left|\hat{I}_{\tau_k^m(s)}^m(s)-I^m(s)\right|\geq\frac{\Delta}{2}\right\}\right]\\
&\leq\sum_{k=0}^{T}\frac{1}{k^2}\\
&\leq 2,
\end{aligned}
$$

where the second last inequality holds due to Lemma F.2. Similarly, we can prove the last two terms are upper bounded by 2 using Lemma F.2.

Using the same way, we can upper bound Eq. (F.62) by 6. Therefore, Eq. (F.59) can be upper bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T}\mathbb{I}\left\{a_{t+1}^m=1,s_t^m=s,\text{not random}\right\}\right]\\
&=\sum_{(m',s')\in\mathcal{G}}\mathbb{E}\left[\sum_{t=0}^{T}\mathbb{I}\left\{a_{t+1}^m=1,s_t^m=s,\hat{I}_t^m(s)\geq\hat{I}_t^{m'}(s'),s_t^{m'}=s'\right\}\right]\\
&\leq 12|\mathcal{G}|.
\end{aligned}
\tag{F.63}
$$

Further assume that $k\,|\Delta|\geq\Delta_s^m$, for all $I^m(s)\leq\gamma$. Then combining Eq. (F.60) and Eq. (F.63) will give us the desired result.

We also have

$$
\begin{aligned}
\mathcal{R}(\pi)&\leq kB\Delta T+\frac{k|\mathcal{S}|MD(M|\mathcal{S}|-|\mathcal{G}|)}{\Delta}\log T\\
&\leq k\sqrt{BMD|\mathcal{S}|(M|\mathcal{S}|-|\mathcal{G}|)T\log T}.
\end{aligned}
$$

□

G IMPLEMENTATION DETAIL

All the experiments are conducted on MacBook Pro with Apple M3 Pro chip and 18 GB memory.

G.1 Regret Curve

To numerically verify Theorem 4.4, we set $M=2$, $B=1$, and $\mathcal{S}=\{0,1\}$. We set the expected reward of the state as $R^m(s)=s$. We randomly generated 10 instances and chose three γ values satisfying Assumption 4.2, labeling each γ by the corresponding gap Δ between the threshold and the smallest incremental reward. Specifically, we will randomly generate a number between $[0,1]$ independently and assign this number to be $P_a^m(s,1),\forall a,m,s\in\{0,1\}$.

Without loss of generality, assume that $I^1(1)\geq I^1(0)\geq I^2(1)\geq I^2(0)$. We choose γ to be between $I^1(0)$ and $I^2(0)$. Specifically, let $\theta\in(0,1)$ be a tunable parameter. We set $\gamma:=\theta I^2(1)+(1-\theta)I^2(0)$. To make sure that the gap between γ and $I^2(1)$ is not too small, we choose those instances where $I^2(1)-I^2(0)\geq 0.15$. To generate the three different γ values in Figure G.9, we selected three different choices of θ: $0.1,0.5,0.9$, yielding the observed Δ values.

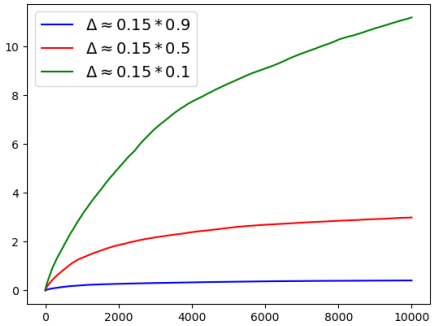

Figure G.9: Regret Curve as horizon T grows.

For each γ, the regret was averaged over 100 repetitions per instance (Figure G.9). Figure G.9 demonstrates sublinear regret for all choices of γ, with regret stabilizing at a constant as the horizon length increases, consistent with Theorem 4.4. Further, as Δ decreases, regret increases as it becomes harder to distinguish whether the incremental reward is above the threshold or not. This matches the results in Theorem 4.4.

G.2 Cumulative Reward

We first describe the benchmarks we chose in detail:

- Oracle Whittle: A variant of Whittle's index policy with access to the ground truth transitions and rewards, optimized for maximizing discounted reward (as described in Appendix B).
- Oracle Greedy: The greedy policy we defined (Algorithm 1) with knowledge of the ground truth incremental reward.
- OMR: A modified LP-based policy for maximizing finite-horizon reward.
- WIQL (Biswas et al., 2021): A Q-learning based approach to learn the value function of each arm at each state by interacting with the RMAB instance. We use the same learning rate (α) as in Biswas et al. (2021).
- UCWhittle-penalty (Wang et al., 2023b): It constructs the confidence interval of the unknown transition probability and then chooses one to calculate Whittle's index policy. Specifically, it constructs a confidence interval and chooses the transition probability that will maximize $V_\lambda^m(s)$ (Eq. (B.9)) for some adaptively updated λ.
- R(MA)^{2}B: It uses the explore-then-commit framework. It first randomly gives action and constructs the confidence interval of the unknown transition. After the exploration period, it uses the estimated transition to calculate the OMR policy.

In all of the settings, the discount factor β was set to 0.9. We note that though $\beta = 1$ was chosen in Biswas et al. (2021), $\beta = 0.9$ has better empirical performance, which aligns with the choice in Wang et al. (2023b).

Then, we will discuss more on the implementation of the algorithms. First of all for the UCWhittle families introduced in Wang et al. (2023b), the algorithms reinitialize the initial state after certain time-steps. We do not include this episodic structure as this contradicts our goal of maximizing the finite-horizon cumulative reward. We keep the rest structures of their algorithms as the same.

We additionally note that the code provided by Wang et al. (2023b) contains errors and could not be used to reproduce the results shown in Wang et al. (2023b). A mistake was found in Line 73 of uc_whittle.py that they provided. Specifically, they use the current state of an agent as the index to get the current action of this agent. As a result, we build our simulation environment from scratch.

Furthermore, we note that the Whittle's index could be negative. Therefore, we make modifications to the binary search routine of getting Whittle's index. Specifically, we let the initial lower bound be -10 to ensure a more accurate calculation of Whittle's index.

We also note that in the original R(MA)^{2}B algorithm, it assumes that they can reinitialize at any state for arbitrary number of times. This makes their exploration criteria, which will only stop when sufficient number of samples are collected for each agent, state, action pair, possible. We think this assumption is not impractical and do not use this structure. Therefore, we also change the algorithm so that the exploration period will end after the same number of time-steps instead of number of samples.

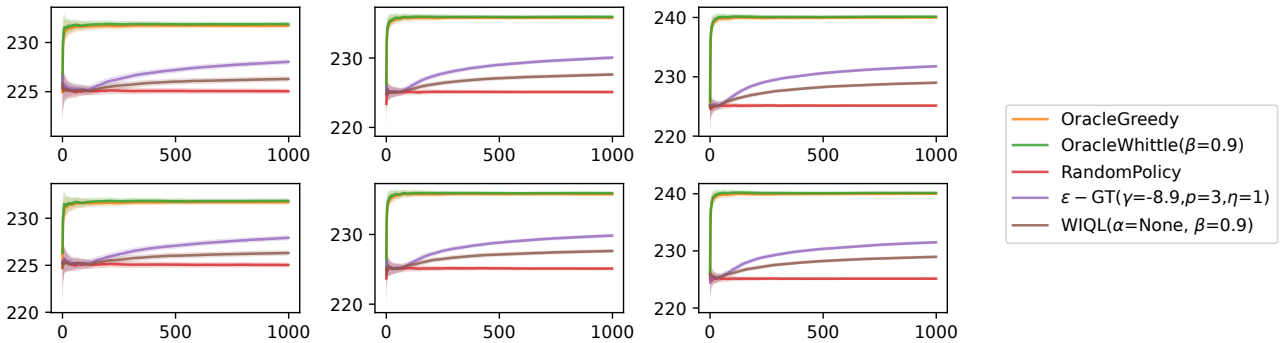

Figure G.10: Average (over time) cumulativere reward for budgets $B = 5$ (left), 10 (middle), 20 (right), averaged over 50 instances with 13 repetitions each. Top: noiseless reward; bottom: noisy reward. Ten states.

Randomly Generated Instances With 10 States In this paragraph, we increase the size of the state space to 10 states to randomly generate 50 instances to test the performance of ϵ-GT comparing to the benchmarks. Similarly, we also test two settings: 1) a noiseless case where the reward is deterministic, and 2) a noisy case where the rewards at state $s \in \mathcal{S}$ follow a normal distribution with mean $R^m(s)$ and variance 0.1. We would like to note that since UCWhittlePv and RRMABB behave like random policy for 2 states, we do not include that in the 10 states setting.

As shown in Figure G.10, ϵ-GT with threshold $\gamma = -8.9$ consistently outperform WIQL in both noiseless and noisy setting. Further, we would like to emphasize that the gap between ϵ-GT and WIQL is larger than the 2 states setting, demonstrating that our algorithm works well in larger states setting and adapting to the real-life setting better.

G.3 Ablation Study

In this section, we will test the performance of Alg. 1 with different choices of η and γ.

Choice of η: We will choose four choices of $\eta \in \{0.1, 0.5, 1, 1.5\}$. We will fix the γ and p to align with the choice of γ used for experiments in Section 5. As shown in Figures G.11-G.13 Alg. 1 is stable over different choices of η for all of the synthetic binary state, ten state instances and real instance we have tested.

Choice of γ In this paragraph, we will compare $\epsilon-$GT with different threshold. We will fix other hyper-parameters to align with choices used in Section 5. As shown in Figure G.14-Figure G.16, for all instances, as γ increases, the convergence rate of Alg. 1 is slower. This translates to a lower average reward. Also, we observe that among all instances, a low threshold that satisfies Assump. 4.2 will achieve the best performance.

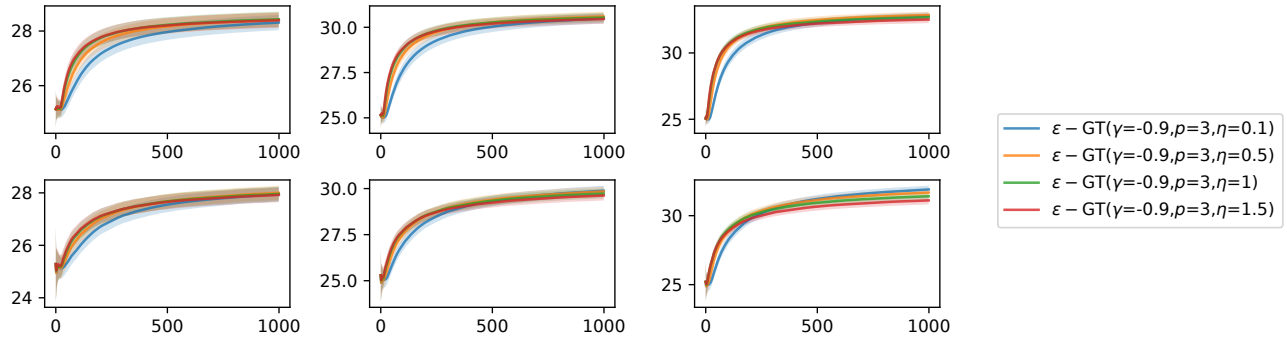

Figure G.11: Average (over time) cumulative reward for budgets $B = 5$ (left), 10 (middle), 20 (right), averaged over 50 binary state instances with 13 repetitions each. Top: noiseless reward; bottom: noisy reward.

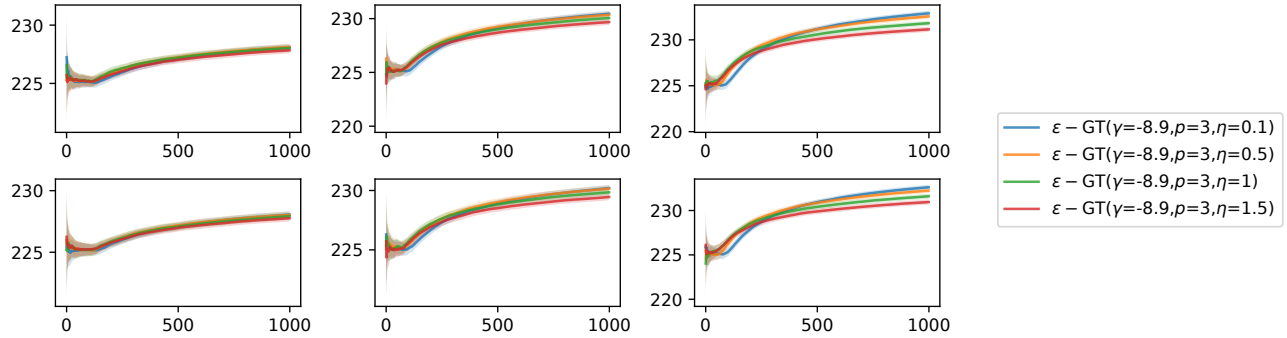

Figure G.12: Average (over time) cumulative reward for budgets $B = 5$ (left), 10 (middle), 20 (right), averaged over 50 10 state instances with 13 repetitions each. Top: noiseless reward; bottom: noisy reward.

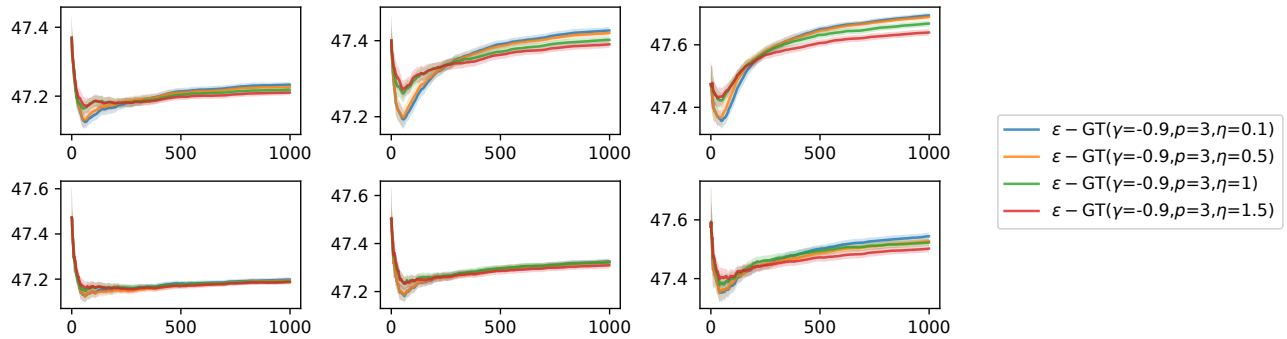

Figure G.13: Average (over time) cumulative reward for budgets $B = 5$ (left), 10 (middle), 20 (right), averaged over 100 repetitions of real dataset. Top: noiseless reward; bottom: noisy reward.

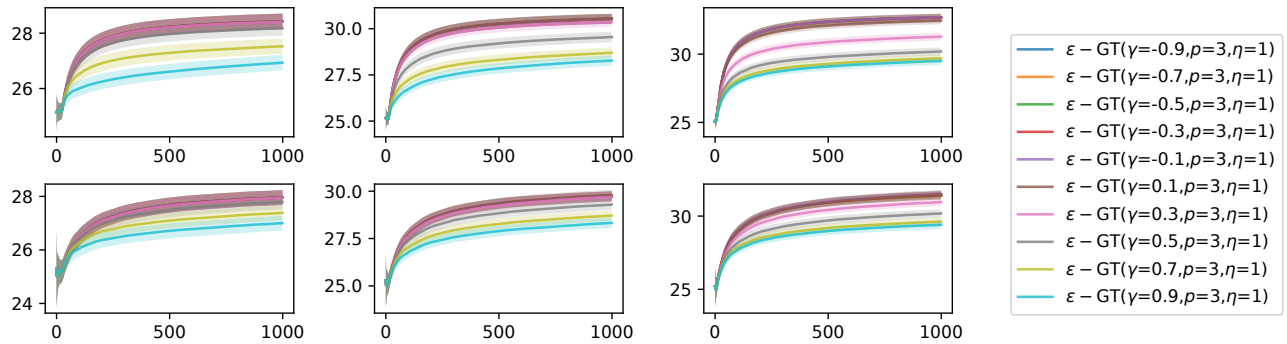

Figure G.14: Average (over time) cumulative reward for budgets $B = 5$ (left), 10 (middle), 20 (right), averaged over 50 binary state instances with 13 repetitions each. Top: noiseless reward; bottom: noisy reward.

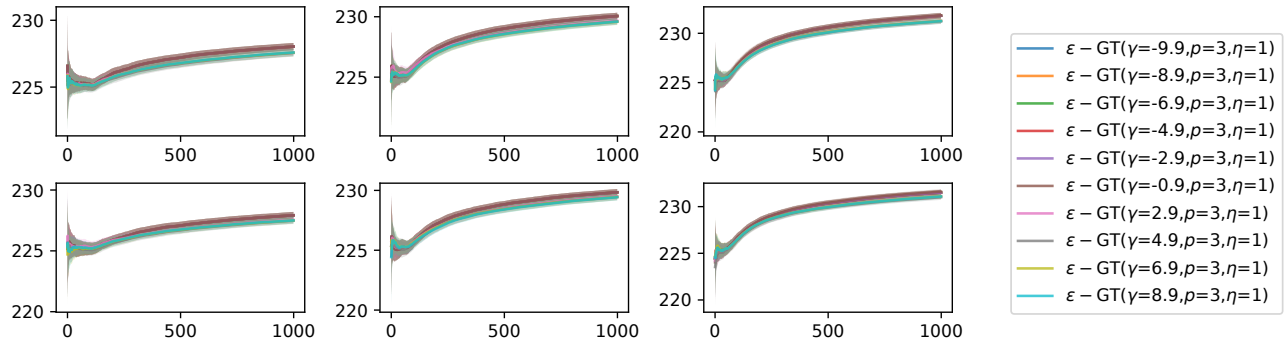

Figure G.15: Average (over time) cumulative reward for budgets $B = 5$ (left), 10 (middle), 20 (right), averaged over 50 ten state instances with 13 repetitions each. Top: noiseless reward; bottom: noisy reward.

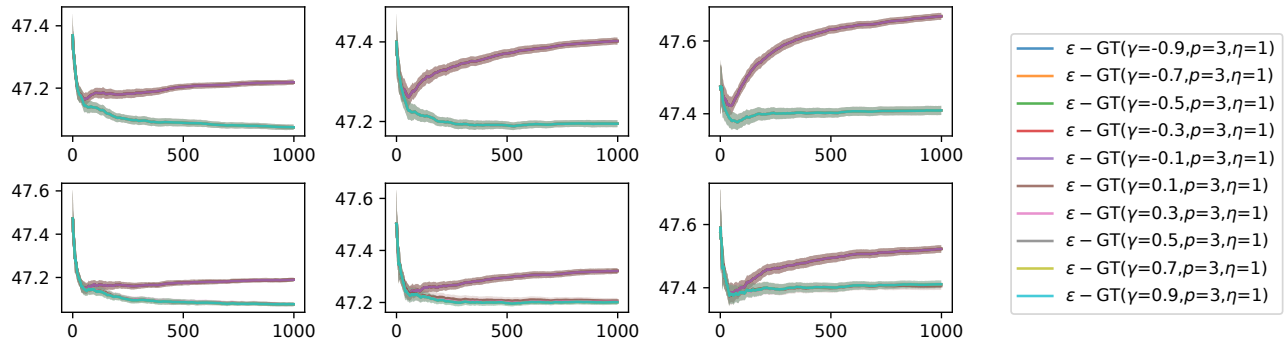

Figure G.16: Average (over time) cumulative reward for budgets $B = 5$ (left), 10 (middle), 20 (right), averaged over 100 repetitions of real dataset. Top: noiseless reward; bottom: noisy reward.