Projection-based Lyapunov method for fully heterogeneous weakly-coupled MDPs

Xiangcheng Zhang[1*†] **Yige Hong**[2*] **Weina Wang**[2]

[1]Weiyang College, Tsinghua University
[2] Computer Science Department, Carnegie Mellon University
xc-zhang21@mails.tsinghua.edu.cn
{yigeh,weinaw}@cs.cmu.edu

Abstract

Heterogeneity poses a fundamental challenge for many real-world large-scale decision-making problems but remains largely understudied. In this paper, we study the *fully heterogeneous* setting of a prominent class of such problems, known as weakly-coupled Markov decision processes (WCMDPs). Each WCMDP consists of N arms (or subproblems), which have distinct model parameters in the fully heterogeneous setting, leading to the curse of dimensionality when N is large. We show that, under mild assumptions, an efficiently computable policy achieves an $O(1/\sqrt{N})$ optimality gap in the long-run average reward per arm for fully heterogeneous WCMDPs as N becomes large. This is the *first asymptotic optimality result* for fully heterogeneous average-reward WCMDPs. Our main technical innovation is the construction of projection-based Lyapunov functions that certify the convergence of rewards and costs to an optimal region, even under full heterogeneity.[1]

1 Introduction

Heterogeneity poses a fundamental challenge for many real-world decision-making problems, where each problem consists of a large number of interacting components. However, despite its practical significance, heterogeneity remains largely understudied in the literature. In this paper, we study *heterogeneous* settings of a prominent class of such problems, known as weakly-coupled Markov decision processes (WCMDPs) [23]. A WCMDP consists of N *arms* (or *subproblems*), where each arm itself is a Markov decision process (MDP). In a heterogeneous setting, the MDPs could be distinct. At each time step, the decision-maker selects an action for each arm, which affects the arm's transition probabilities and reward, and then the arms make state transitions independently. However, these actions are subject to a set of global *budget constraints*, where each constraint limits one type of total cost across all arms at each time step. The objective is to find a policy that maximizes the long-run *average reward* over an infinite time horizon. We focus on the *planning* setting, where all the model parameters (reward function, cost functions, budget, and transition kernel) are known.

WCMDPs have been used to model a wide range of applications, including online advertising [7, 57], job scheduling [48], healthcare [5], surveillance [40], and machine maintenance [21]. A faithful modeling of these applications calls for *heterogeneity*. For instance, in [5], arms are beneficiaries of a health program and they could react to interventions differently; in [40], arms are targets of

*These authors contributed equally to this work.
†Work done during a visit at Carnegie Mellon University.
[1]A full version of this paper [54] can be found at https://www.arxiv.org/abs/2502.06072v5.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

surveillance who have different locations and probabilities to be exposed; in [21], arms are machines that could require distinct repair schedules.

Although heterogeneity is crucial in the modeling of these applications, most existing work on average-reward WCMDPs establishes asymptotic optimality only for the homogeneous setting where all arms share the same set of model parameters [14, 15, 22, 24–27, 39, 42, 47]. Only a few exceptions [25, 39, 46] address heterogeneity, but in highly specialized settings. Among these, a common approach to handle heterogeneity is to consider the *typed heterogeneous* setting, where the N arms are divided into a constant number of types as N scales up, with each type having distinct model parameters. While heterogeneous WCMDPs have been studied under the finite-horizon total-reward and discounted-reward criteria, these results do not extend to the average-reward setting we consider. We review related work in more detail at the end of this section and also in Appendix A.

The key distinction between the homogeneous (or typed heterogeneous) setting and the fully heterogeneous setting is whether the arms can be divided into a *constant* number of homogeneous groups. In the former, the system dynamics depends only on the fraction of arms in each state in each homogeneous group. Thus, the effective dimension of the state space is *polynomial* in N. In contrast, in the fully heterogeneous setting, the state space grows *exponentially* in N, making the problem truly high-dimensional.

Our contribution. In this paper, we study *fully heterogeneous* WCMDPs. We propose a policy we call the *ID policy with reassignment*, which generalizes the ID policy in the literature [26]. In our policy, we first perform an ID reassignment algorithm to reorder the arms, which ensures proper arm prioritization during policy execution. We then run a variant of the ID policy adapted to handle heterogeneity, which consists of two phases. The first phase is a pre-processing phase, where we compute an *optimal single-armed policy* for each arm (denoted as $\bar{\pi}_i^*$ for the i-th arm) that prescribes the *ideal action* the arm would take at each state. The second phase is the real-time phase. At each time step, the policy iterates over the arms according to their reassigned IDs, and it lets as many arms as possible follow their respective ideal actions while satisfying the budget constraints. Unlike the original ID policy, which has only one optimal single-armed policy due to homogeneous arms, our policy computes N optimal single-armed policies, one for each arm. Our proposed policy is efficiently computable, with computational complexity polynomial in N.

We prove that the proposed ID policy with reassignment achieves an $O(1/\sqrt{N})$ optimality gap under mild assumptions as the number of arms N becomes large. Here, the optimality gap refers to the difference between the long-run average reward per arm under our policy and that under the optimal policy. *This is the first result establishing asymptotic optimality for fully heterogeneous average-reward WCMDPs.*

We remark that the original ID policy was designed for a special case of WCMDPs known as restless bandits, and it is for the homogeneous setting. While the generalization in our proposed policy is natural, identifying the appropriate generalization and establishing its optimality gap in the fully heterogeneous setting are technically challenging and require new theoretical approaches.

Technical novelty. The main technical innovation of the paper is the introduction of a novel Lyapunov function for fully heterogeneous WCMDPs. Specifically, to prove the asymptotic optimality of a policy, a key step is to show that the system state is globally attracted to an *optimal region* where most arms can follow the ideal actions generated by their respective optimal single-armed policies $\bar{\pi}_i^*$'s. In the homogeneous setting, we can prove such convergence using state aggregation techniques that rely on the symmetry of arms. However, in the heterogeneous setting, states of different arms cannot be aggregated since arms are no longer symmetric. Our technique is to *project* arm states onto a set of carefully selected feature vectors, and define the Lyapunov function based on these projections. These feature vectors encode the minimal amount of information needed to evaluate the relevant functions of the system state (e.g., instantaneous reward or cost) and predict their future expectations. This projection-based Lyapunov function provides a principled way to measure deviations of the system state from the optimal region in a fully heterogeneous setting. A more detailed discussion of this approach can be found in Section 4.

Beyond WCMDPs, our techniques have the potential to be applied to more general heterogeneous large stochastic systems. Heterogeneity has been a topic of strong interest in these systems, but it is known to be a challenging problem with limited theoretical results. Only recently have there been

notable breakthroughs. [1, 2] extended the popular mean-field analysis to a class of heterogeneous large stochastic systems for the first time, but the results are only for transient distributions. Another line of work [55, 56] studied heterogeneous load-balancing systems. They first analyzed the transient distributions and then used interchange-of-limits arguments to extend the results to steady state. Our method provides a more direct framework for steady-state analysis and has the potential to generalize to a broader range of heterogeneous stochastic systems.

Related work. WCMDPs have been extensively studied with a rich body of literature. Here we briefly overview the most relevant work and refer the reader to Appendix A for a detailed survey.

We first focus on the *average-reward* criterion. As mentioned earlier, most existing work considers the *homogeneous setting*. Early work on WCMDPs primarily focuses on a special case known as the *restless bandit (RB)* problem, where each arm's MDP has a binary action space (active and passive actions) and there is only one budget constraint that limits the total number of active actions across all arms at each time step. The seminal work by Whittle [44] introduced the RB problem and the celebrated Whittle index policy, which was later shown to achieve an $o(1)$ optimality gap as $N \to \infty$ under a set of conditions [42]. Subsequent work on RBs has focused on designing policies that achieve asymptotic optimality under more relaxed conditions [25, 26, 39, 47], as well as improving the optimality gap to $O(1/\sqrt{N})$ [25, 26] or $O(\exp(-cN))$ [14, 15, 27]. Among these papers, [25, 39] address heterogeneous RBs. However, [39] focuses on the *typed heterogeneous* setting, where the N arms are divided into a constant number of types as $N \to \infty$. The paper [25] includes an extension to the fully heterogeneous setting. However, for their result to yield asymptotic optimality, there need to be further assumptions on the orders of the so-called synchronization times in the paper. The policies in both papers cannot be straightforwardly extended to general WCMDPs, which have multiple actions, multiple budget constraints, and state-dependent cost functions.

Beyond RBs, work on general average-reward WCMDPs is scarce. The closest to ours are [24, 39, 46], which studied WCMDPs with a single budget constraint and established $o(1)$ optimality gaps, but again in the homogeneous setting [24] or the *typed heterogeneous* setting [39, 46]. More recently, [22] proved the first $o(1)$ optimality gap result for WCMDPs with general budget constraints, but in the *homogeneous* setting.

Under the *finite-horizon total-reward* or *discounted-reward* criteria, there has been more work on heterogeneous settings, including both the typed heterogeneous setting [12, 17] and, more recently, the fully heterogeneous setting [8–10, 50]. However, the optimality gap in these papers generally grows *super-linearly* with the (effective) time horizon, except under restrictive conditions. Consequently, it is difficult to extend these results to the average-reward setting and still achieve asymptotic optimality.

Although our work focuses on the planning setting where all model parameters are known, there has been growing interest in developing reinforcement learning algorithms for the learning setting with unknown parameters [3, 6, 13, 29, 30, 32, 33, 36, 45, 46]. Many of these approaches rely on well-designed planning policies as a foundation to achieve learning efficiency. In this context, our results can serve as an important building block for developing model-based learning algorithms for fully heterogeneous WCMDPs.

General notation. Let $\mathbb{R}$, $\mathbb{N}$, and $\mathbb{N}_+$ denote the sets of real numbers, nonnegative integers, and positive integers, respectively. Let $[N] \triangleq \{1, 2, \ldots, N\}$ for any $N \in \mathbb{N}_+$ and $[n_1 : n_2] \triangleq \{n_1, n_1 + 1, \ldots, n_2\}$ for $n_1, n_2 \in \mathbb{N}_+$ with $n_1 \leq n_2$. Let $[0, 1]_N = \{i/N : i \in \mathbb{N}, 0 \leq i/N \leq 1\}$, the set of integer multiples of $1/N$ in $[0, 1]$. For a matrix $A \in \mathbb{R}^{d \times d}$, we denote its operator norm as $\|A\|_p = \sup_{x \neq \mathbf{0}} \|Ax\|_p / \|x\|_p$. We use boldface letters to denote matrices, and regular letters to denote vectors and scalars. We write $\mathbb{R}^{\mathbb{S}}$ for the set of real-valued vectors indexed by elements of $\mathbb{S}$, or equivalently, the set of real-valued functions on $\mathbb{S}$; for each $v \in \mathbb{R}^{\mathbb{S}}$, let $v(s)$ to denote its element corresponding to $s \in \mathbb{S}$.

2 Problem setup

We consider a weakly-coupled Markov decision process (WCMDP) that consists of N arms. Each arm has an ID $i \in [N]$ and is associated with a smaller MDP denoted as $\mathcal{M}_i = \big(\mathbb{S}, \mathbb{A}, \mathbb{P}_i, r_i, (c_{k,i})_{k \in [K]}\big)$. Here $\mathbb{S}$ and $\mathbb{A}$ are the state space and the action space, respectively, both assumed to be finite; $\mathbb{P}_i$ describes the transition probabilities with $\mathbb{P}_i(s' \mid s, a)$ being the transition probability from state s

to state s' when action a is taken. The state transitions of different arms are independent given the actions. When arm i is in state s and we take action a, a reward $r_i(s,a)$ is generated, as well as K types of costs $c_{k,i}(s,a), k \in [K]$. We assume that the costs are nonnegative, i.e., $c_{k,i}(s,a) \geq 0$ for all $i \in \mathbb{N}_+, k \in [K], s \in \mathbb{S}$, and $a \in \mathbb{A}$. Note that we allow the arms to be *fully heterogeneous*, i.e., the $\mathcal{M}_i$'s can be *all distinct*.

When taking an action for each arm in this N-armed system, we are subject to budget constraints. Specifically, suppose each arm i is in state s_i. Then the actions, a_i's, should satisfy the constraints:

$$\sum_{i \in [N]} c_{k,i}(s_i, a_i) \leq \alpha_k N, \quad \forall k \in [K], \tag{1}$$

where each $\alpha_k > 0$ is a constant independent of N, and $\alpha_k N$ is referred to as the *budget* for type-k cost. We assume that there exists an action $0 \in \mathbb{A}$ that does not incur any type of cost for any arm at any state, i.e., $c_{k,i}(s,0) = 0$ for all $k \in [K], i \in [N], s \in \mathbb{S}$. This assumption guarantees that there always exist valid actions (e.g., taking action 0 for every arm) regardless of the states of the arms.

Policy and system state. A policy π for the N-armed problem specifies the action for each of the N arms, in a possibly history-dependent way. Under policy π, let $S_{i,t}^\pi$ denote the state of the ith arm at time t, and we refer to $\boldsymbol{S}_t^\pi \triangleq (S_{i,t}^\pi)_{i \in [N]}$ as the *system state*. Similarly, let $A_{i,t}^\pi$ denote the action applied to arm i at time t, and we refer to $\boldsymbol{A}_t^\pi \triangleq (A_{i,t}^\pi)_{i \in [N]}$ as the *system action*. In this paper, we also use an alternative representation of the system state, denoted as $\boldsymbol{X}_t^\pi$ and defined as follows. Let $X_{i,t}^\pi = (X_{i,t}^\pi(s))_{s \in \mathbb{S}} \in \mathbb{R}^{|\mathbb{S}|}$ be a row vector where the entry corresponding to state s is given by $X_{i,t}^\pi(s) = \mathbb{1}\{S_{i,t}^\pi = s\}$; i.e., $X_{i,t}^\pi$ is a one-hot row vector whose s's entry is 1 if $S_{i,t}^\pi = s$ and is 0 otherwise. Then let $\boldsymbol{X}_t^\pi$ be an $N \times |\mathbb{S}|$ matrix whose ith row is $X_{i,t}^\pi$. It is easy to see that $\boldsymbol{X}_t^\pi$ contains the same information as $\boldsymbol{S}_t^\pi$, and we refer to both of them as the system state. In this paper, we often encounter vectors like $X_{i,t}^\pi = (X_{i,t}^\pi(s))_{s \in \mathbb{S}}$, whose entries correspond to different states in $\mathbb{S}$. For such vectors, say u and v, we use the inner product to write a sum for convenience $\langle u, v \rangle \triangleq \sum_{s \in \mathbb{S}} u(s)v(s)$. We sometimes omit the superscript π when it is clear from context.

Maximizing average reward. Our objective is to maximize the long-run time-average reward subject to the budget constraints. To be more precise, we follow the treatment for maximizing average reward in [35]. For any policy π and an initial state $\boldsymbol{S}_0$ of the N-armed system, consider the *limsup* average reward $R^+(\pi, \boldsymbol{S}_0)$ and the *liminf* average $R^-(\pi, \boldsymbol{S}_0)$, defined as $R^+(\pi, \boldsymbol{S}_0) = \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}\left[r_i(S_{i,t}^\pi, A_{i,t}^\pi)\right]$ and $R^-(\pi, \boldsymbol{S}_0) = \liminf_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}\left[r_i(S_{i,t}^\pi, A_{i,t}^\pi)\right]$. If $R^+(\pi, \boldsymbol{S}_0) = R^-(\pi, \boldsymbol{S}_0)$, then the average reward of policy π under the initial condition $\boldsymbol{S}_0$ exists and is defined as

$$R(\pi, \boldsymbol{S}_0) = R^+(\pi, \boldsymbol{S}_0) = R^-(\pi, \boldsymbol{S}_0) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}\left[r_i(S_{i,t}^\pi, A_{i,t}^\pi)\right]. \tag{2}$$

Note that these reward notions divide the total reward from all arms by the number of arms, N, measuring the reward *per arm*. The WCMDP problem is to solve the following optimization problem:

$$\underset{\text{policy } \pi}{\text{maximize}} \quad R^-(\pi, \boldsymbol{S}_0) \tag{3a}$$

$$\text{subject to} \quad \sum_{i \in [N]} c_{k,i}(S_{i,t}^\pi, A_{i,t}^\pi) \leq \alpha_k N, \quad \forall k \in [K], \forall t \geq 0. \tag{3b}$$

Let the optimal value of this problem be denoted as $R^*(N, \boldsymbol{S}_0)$. Note that since the WCMDP is an MDP with finite state and action space, if we replace the $R^-(\pi, \boldsymbol{S}_0)$ in the objective (3a) with $R^+(\pi, \boldsymbol{S}_0)$, the optimal value stays the same [35, Proposition 9.1.6].

Asymptotic optimality. Recall that exactly solving the WCMDP problem is PSPACE-hard [34]. In this paper, our goal is to design a policy π that is *efficiently computable* and *asymptotically optimal* as $N \to \infty$, with the following notion for asymptotic optimality. For any policy π, we define its *optimality gap* as $R^*(N, \boldsymbol{S}_0) - R^-(\pi, \boldsymbol{S}_0)$. We say the policy π is *asymptotically optimal* if

$$R^*(N, \boldsymbol{S}_0) - R^-(\pi, \boldsymbol{S}_0) = o(1) \quad \text{as} \quad N \to \infty. \tag{4}$$

When we take this asymptotic regime as $N \to \infty$, we keep the number of constraints, K, as well as the budget coefficients, $\alpha_1, \alpha_2, \ldots, \alpha_K$, fixed. We assume that the reward functions and cost functions are uniformly bounded, i.e., $\sup_{i\in\mathbb{N}_+} \max_{s\in\mathbb{S},a\in\mathbb{A}} |r_i(s,a)| \triangleq r_{\max} < \infty$ and $\sup_{i\in\mathbb{N}_+} \max_{k\in[K],s\in\mathbb{S},a\in\mathbb{A}} c_{k,i}(s,a) \triangleq c_{\max} < \infty$. This notion for asymptotic optimality is consistent with that in the existing literature (e.g., [39, Definition 4.11]). We are interested in not only achieving asymptotic optimality but also characterizing the *order* of the optimality gap.

In the remainder of this paper, we focus on stationary Markov policies, which are sufficient for achieving the optimal value because the WCMDP problem is an MDP with finite state and action spaces [35, Theorem 9.1.8]. Under any stationary Markovian policy, the long-run reward $R(\pi, \boldsymbol{S}_0) = R^+(\pi, \boldsymbol{S}_0) = R^-(\pi, \boldsymbol{S}_0)$ is always well-defined [35, Theorem 8.1.1].

LP relaxation and an upper bound on optimality gap. We consider the linear program (LP) below, which will play a critical role in performance analysis and policy design:

$$R_N^{\text{rel}} \triangleq \underset{(y_i(s,a))_{i\in[N],s\in\mathbb{S},a\in\mathbb{A}}}{\text{maximize}} \quad \frac{1}{N}\sum_{i\in[N]}\sum_{s\in\mathbb{S},a\in\mathbb{A}} y_i(s,a) r_i(s,a) \tag{5a}$$

$$\text{subject to} \quad \frac{1}{N}\sum_{i\in[N]}\sum_{s\in\mathbb{S},a\in\mathbb{A}} y_i(s,a) c_{k,i}(s,a) \le \alpha_k, \ \forall k \in [K], \tag{5b}$$

$$\sum_{s'\in\mathbb{S},a'\in\mathbb{A}} \mathbb{P}_i(s \mid s',a') y_i(s',a') = \sum_{a\in\mathbb{A}} y_i(s,a), \ \forall s \in \mathbb{S}, \forall i \in [N], \tag{5c}$$

$$\sum_{s'\in\mathbb{S},a'\in\mathbb{A}} y_i(s',a') = 1, \ y_i(s,a) \ge 0, \ \forall s \in \mathbb{S}, \forall a \in \mathbb{A}, \forall i \in [N]. \tag{5d}$$

Lemma 1 below establishes a connection between this LP and the WCMDP.

Lemma 1. *The optimal value of any N-armed WCMDP problem is upper bounded by the optimal value of the corresponding linear program in (5), i.e.,*

$$R^*(N, \boldsymbol{S}_0) \le R_N^{\text{rel}}, \quad \forall N, \forall \boldsymbol{S}_0.$$

An immediate implication of Lemma 1 is that for any policy π, its optimality gap is upper bounded as

$$R^*(N, \boldsymbol{S}_0) - R^-(\pi, \boldsymbol{S}_0) \le R_N^{\text{rel}} - R^-(\pi, \boldsymbol{S}_0). \tag{6}$$

Therefore, to derive an upper bound for the optimality gap, it suffices to control $R_N^{\text{rel}} - R^-(\pi, \boldsymbol{S}_0)$, which we will show is $O(1/\sqrt{N})$ in Theorem 1.

To see the intuition of Lemma 1, we interpret the optimization variable $y_i(s,a)$ as the long-run fraction of time arm i spends in state s and takes action a. We refer to $y_i(s,a)$ as arm i's *state-action frequency* for the state-action pair (s,a). Then the constraints in (5b) of the LP can be viewed as relaxations of the budget constraints in (3b) for the WCMDP. The constraints in (5c)–(5d) guarantee that $y_i(s,a)$'s are proper stationary time fractions. Therefore, the LP is a relaxation of the WCMDP and thus achieves a higher optimal value. The proof of Lemma 1 is provided in Appendix C of [54].

Our LP (5) serves a similar role to the LP used in previous work on restless bandits and WCMDPs (see, e.g., [15, 22, 24, 25, 39, 42, 46]), but with different forms and dependencies on N. Both our LP and the LP in previous work relax the hard budget constraints to time-average constraints. However, in the homogeneous arm setting [15, 22, 24, 42], the LP has only one set of state-action frequencies $y(s,a)$, and the LP is independent of N. As a result, both the optimal value of the LP and the complexity of solving it are independent of N. Prior work for the typed-heterogeneous setting [25, 39, 46] divides the arms into a constant number K of types and defines a set of state-action frequency $y_k(s,a)$ for each type k. The optimal value of the resulting LP and the complexity of solving it depend on the number of types K. Hong et al. [25] includes a generalization to the fully heterogeneous setting, resulting in an LP similar to ours, although restricted to restless bandits.

In our LP, we define a separate set of state-action frequencies $y_i(s,a)$ for each arm $i \in [N]$, making the LP explicitly depend on N. Therefore, the optimal value R_N^{rel} depends on N, and the complexity of solving the LP grows with N. Nevertheless, because the number of variables and constraints scales linearly with N, our LP can still be solved in polynomial time.

Algorithm 1 ID reassignment

Input: optimal state-action frequencies $(y_i^*(s,a))_{i\in[N],s\in\mathbb{S},a\in\mathbb{A}}$, budgets $(\alpha_k)_{k\in[K]}$
Output: new arm ID, recorded in newID(i), for each arm with old ID $i \in [N]$
Compute $(C^*_{k,i})_{i\in[N],k\in[K]}$ and the set of active constraints $\mathcal{A}$ using (8)
if $\mathcal{A} = \emptyset$ **then**
  newID$(i) \leftarrow i$ for all $i \in [N]$ ▷ *No need for ID reassignment*
else
  Initialize $\mathcal{F} \leftarrow \emptyset$ ▷ *Set of arms that have been assigned new IDs*
  Initialize $\mathcal{D}_k \leftarrow \{i \in [N] \colon C^*_{k,i} \geq \delta\}$ for all $k \in \mathcal{A}$
  $\delta \leftarrow \alpha_{\min}/4 \triangleq \min_{k\in[K]} \alpha_k/4$; $d \leftarrow \left\lceil \frac{(c_{\max}-\delta)K}{\alpha_{\min}/2-\delta} \right\rceil$
  for $\ell = 0, 1, \ldots, \lfloor N/d \rfloor - 1$ **do**
    $\mathcal{I}(\ell) \leftarrow [\ell d + 1 : (\ell+1)d]$; $j \leftarrow \ell d + 1$
    for $k \in \mathcal{A}$ **do**
      if $\sum_{i\in\mathcal{F}} C^*_{k,i} \mathbb{1}\{\text{newID}(i) \in \mathcal{I}(\ell)\} < \delta$ **then**
        Pick any i from $\mathcal{D}_k$ and set newID$(i) \leftarrow j$; remove i from $\mathcal{D}_{k'}$ for all k'; add i to $\mathcal{F}$
        $j \leftarrow j + 1$
  For all $i \in [N]\backslash\mathcal{F}$, assign values to their newID(i)'s randomly from $[N]\backslash\{\text{newID}(i') \colon i' \in \mathcal{F}\}$

3 ID policy with reassignment

In this section, we introduce the ID policy with reassignment, generalized from the ID policy designed for homogeneous restless bandits in the literature [26]. Our policy first performs an ID reassignment procedure, and then proceeds to run a variant of the ID policy adapted to handle heterogeneity. We begin by introducing a building block of our policy, referred to as optimal single-armed policies, followed by the ID reassignment algorithm and the execution of the adapted ID policy.

Optimal single-armed policies. Once we obtain a solution to the LP in (5), we can construct a policy for each arm i, which we refer to as an *optimal single-armed policy* for arm i. In particular, let $(y_i^*(s,a))_{i\in[N],s\in\mathbb{S},a\in\mathbb{A}}$ be an arbitrary optimal solution to the LP in (5). Then for arm i, the optimal single-armed policy, $\bar{\pi}_i^*$, is defined as

$$\bar{\pi}_i^*(a \mid s) = \begin{cases} \frac{y_i^*(s,a)}{\sum_{a\in\mathbb{A}} y_i^*(s,a)}, & \text{if } \sum_{a\in\mathbb{A}} y_i^*(s,a) > 0, \\ \frac{1}{|\mathbb{A}|}, & \text{if } \sum_{a\in\mathbb{A}} y_i^*(s,a) = 0, \end{cases} \tag{7}$$

where $\bar{\pi}_i^*(a \mid s)$ is the probability of taking action a given that the arm's current state is s. Note that due to heterogeneity, this optimal single-armed policy $\bar{\pi}_i^*$ can be different for different arms.

The rationale behind these policies is as follows. If each arm i individually follows its optimal single-armed policy $\bar{\pi}_i^*$, then the average reward per arm (total reward divided by N) achieves the upper bound R_N^{rel} given by the LP. However, this strategy only guarantees that the budget constraints are satisfied in a *time-average* sense, rather than conforming to the *hard* constraints in the original N-armed WCMDP. Thus, having each arm follow its optimal single-armed policy is not a valid policy for the original N-armed problem. Nevertheless, these optimal single-armed policies $\bar{\pi}_i^*$'s serve as a guide for how the arms should ideally behave to maximize rewards. The ID policy uses the $\bar{\pi}_i^*$'s as a reference. It is then designed to ensure that even under the hard budget constraints, most arms follow their optimal single-armed policies most of the time, yielding a diminishing gap to R_N^{rel} in reward.

ID reassignment. We first define a few quantities that will be used in the ID reassignment algorithm. For each arm $i \in [N]$ and each cost type $k \in [K]$, the expected cost under the optimal single-armed policy is defined as $C^*_{k,i} = \sum_{s\in\mathbb{S},a\in\mathbb{A}} y_i^*(s,a) c_{k,i}(s,a)$. Based on $C^*_{k,i}$'s, we divide the budget constraints into *active* constraints and *inactive* constraints as follows. For each cost type $k \in [K]$, we say the type-k budget constraint is *active* if

$$\sum_{i\in[N]} C^*_{k,i} \geq \frac{\alpha_k}{2} N, \tag{8}$$

and *inactive* otherwise. Let $\mathcal{A} \subseteq [K]$ denote the set of cost types corresponding to active constraints. Note that replacing $\alpha_k N/2$ with any constant and strict fraction of $\alpha_k N$ will not change the results.

Algorithm 2 ID policy with reassignment

Input: N-armed WCMDP instance $(\mathcal{M}_i)_{i\in[N]}$
Preprocessing:
  Solve the LP in (5) and obtain the optimal state-action frequencies $(y_i^*(s,a))_{i\in[N],s\in\mathbb{S},a\in\mathbb{A}}$
  Calculate the optimal single-armed policies $(\bar{\pi}_i^*)_{i\in[N]}$ using (7)
  Perform ID reassignment using Algorithm 1
Real-time:
for $t = 0, 1, 2, \cdots$ **do**
  Sample ideal actions $\widehat{A}_{i,t} \sim \bar{\pi}_i^*(\cdot \mid S_{i,t})$ for all $i \in [N]$
  $I \leftarrow 1$
  while $\sum_{i\in[I]} c_{k,i}(S_{i,t}, \widehat{A}_{i,t}) \le \alpha_k N, \forall k \in [K]$ **do**
    For arm I, take action $A_{I,t} = \widehat{A}_{I,t}$; $I \leftarrow I + 1$
  For each arm $i \in \{I, I+1, \ldots, N\}$, take action $A_{i,t} = 0$

Based on the costs $C_{k,i}^*$'s and the active constraints, the ID reassignment algorithm rearranges arms so that the cost incurred by each contiguous segment of arms is approximately proportional to the length of the segment. We give a brief explanation of how the reassigned IDs affect the execution of the policy at the end of this section. The algorithm is formally described in Algorithm 1, with more details and properties provided in Appendix D of [54]. In the rest of the paper, we use the *reassigned IDs* to refer to arms, i.e., arm i refers to the arm whose new ID assigned by Algorithm 1 is i.

Constructing ID policy. We are now ready to describe our generalized ID policy, formally described in Algorithm 2. The policy begins with a one-time preprocessing phase: we solve the associated LP, construct the optimal single-armed policies, and reassign arm IDs using the ID reassignment algorithm (Algorithm 1). After the preprocessing, the policy proceeds at each time step t as follows. For each arm i (where i is the reassigned ID), we first sample an action $\widehat{A}_{i,t}$, referred to as an *ideal action*, from the optimal single-armed policy $\bar{\pi}_i^*(\cdot \mid S_{i,t})$. We then attempt to execute these ideal actions, i.e., set the real actions equal to the ideal actions, in ascending order of arm IDs, starting from $i = 1$, then $i = 2$, and so on. We continue the attempt until we have used up at least one type of cost budget, at which point we let the remaining arms take action 0 (the no-cost action).

This ID policy is a natural generalization of the ID policy designed for homogeneous restless bandits [26]. At a high level, using arm IDs to decide the priority order for executing ideal actions guarantees that a subset of arms (those with smaller IDs) can persistently follow their ideal actions. This persistency gives these arms time to converge to their optimal state-action frequencies, which in turn allows their instantaneous costs to converge to steady-state values. This convergence creates slack in the budget constraints, thereby allowing more arms to follow their ideal actions. In contrast, if we do not use IDs but instead randomly select a subset of arms to follow their ideal actions, the convergence may be disrupted. Indeed, [25] provides an example where this randomized strategy fails to achieve asymptotic optimality in restless bandits.

How the reassigned IDs affect the policy execution. In the heterogeneous setting, arms differ in their cost consumptions under their optimal single-armed policies. The ID reassignment algorithm is designed to prevent "plateaus" in cumulative cost as we progress from smaller to larger IDs. If such a plateau exists, the subset of arms allowed to follow their ideal actions (determined by the budget constraints) can become sensitive to the randomness in action sampling, potentially leading to performance instability. The ID reassignment algorithm ensures a regularity property of the cost consumptions, stated in Lemma 2 in Appendix D of [54], which eliminates such plateaus.

4 Main results and technical overview

Before we present the main results, we first state our main assumption. This assumption is for the optimal single-armed policies $\bar{\pi}_i^*$'s. Note that each $\bar{\pi}_i^*$ is a stationary Markov policy. Therefore, under this policy, the state of arm i forms a Markov chain. Let the transition probability matrix of this Markov chain be denoted as $P_i = (P_i(s,s'))_{s\in\mathbb{S},s'\in\mathbb{S}}$, where the row index is the current state s and

the column index is the next state s'. Then $P_i(s, s')$ can be written as

$$P_i(s, s') = \sum_{a \in \mathbb{A}} \mathbb{P}_i(s' \mid s, a) \bar{\pi}_i^*(a \mid s). \tag{9}$$

One can verify that the stationary distribution of this Markov chain is $\mu_i^* = (\mu_i^*(s))_{s \in \mathbb{S}}$ with $\mu_i^*(s) = \sum_{a \in \mathbb{A}} y_i^*(s, a)$, which we refer to as the *optimal state distribution* for arm i. Let τ_i be the *mixing time* of this Markov chain, defined as

$$\tau_i = \max_{s \in \mathbb{S}} \min \left\{ t \in \mathbb{N} \colon \left\| P_i^t(s, \cdot) - \mu_i^*(\cdot) \right\|_1 \leq 1/e \right\}, \tag{10}$$

where P_i^t is the t-step transition probability matrix. The mixing time τ_i is finite if the Markov chain P_i is unichain (one recurrent class, possibly with transient states) and aperiodic.

Assumption 1. *For each arm $i \in \mathbb{N}_+$, the induced Markov chain under the optimal single-armed policy $\bar{\pi}_i^*$ is an aperiodic unichain. Furthermore, the mixing times of these Markov chains have a uniform upper bound; i.e., there exists a positive τ such that for all $i \in \mathbb{N}_+$,*

$$\tau_i \leq \tau. \tag{11}$$

We remark that in the homogeneous or typed heterogeneous settings, once we make the aperiodic unichain assumption in Assumption 1, the uniform upper bound on mixing times automatically exists.

Next, we state our main theorem, Theorem 1, whose proof is provided in Appendix E of [54].

Theorem 1. *Consider an N-armed WCMDP problem satisfying Assumption 1, with initial system state $\boldsymbol{S}_0$. Let policy π be the ID policy with reassignment (Algorithm 2). Then the optimality gap of π is bounded as*

$$R^*(N, \boldsymbol{S}_0) - R(\pi, \boldsymbol{S}_0) \leq \frac{C_{\text{ID}}}{\sqrt{N}},$$

where C_{ID} is a positive constant independent of N.

We re-emphasize that our proposed ID policy with reassignment is the first efficiently computable policy that achieves an $O(1/\sqrt{N})$ optimality gap for fully heterogeneous average-reward WCMDPs. In contrast, the best-known optimality gap for efficiently computable policies for average-reward WCMDPs is $o(1)$, achieved only under restrictive budget constraints and typed-heterogeneity.

We comment that the primary goal of this paper is to characterize the optimality gap in terms of its order in N, which is in line with the main focus of the large body of prior work on restless bandits and WCMDPs. While our analysis also gives an explicit expression for the constant C_{ID}, which shows that $C_{\text{ID}} = O(K^5 \max\{r_{\max}, c_{\max}\}^7 \tau^4 / \alpha_{\min}^6)$, we have not attempted to optimize its dependence on other problem parameters, either through refined analysis or alternative policy design.

Remark 1 (Generalization of result). Our result can be generalized to the setting where a $g(N)$ fraction of arms have unbounded mixing times, and the mixing times of the remaining arms scale with N with an upper bound τ. In this case, we can modify the ID policy by reassigning this $g(N)$ fraction of arms the largest IDs, effectively ignoring these arms. Applying Theorem 1 to the remaining arms then implies $R^*(N, \boldsymbol{S}_0) - R(\pi, \boldsymbol{S}_0) \leq O\big(K^5 \max\{r_{\max}, c_{\max}\}^7 \tau^4 / (\alpha_{\min}^6 \sqrt{N}) + r_{\max} g(N)\big)$. Consequently, this modified policy is asymptotically optimal when $g(N) = o(1)$ and $\tau = o(N^{1/8})$.

Adapting the proof of Theorem 1 also gives the following finite-time bound (see Appendix I of [54]).

Proposition 1 (Finite-time bound)**.** *Under the same conditions as Theorem 1, we have for any $T \geq 1$,*

$$R^*(N, \boldsymbol{S}_0) - \frac{1}{TN} \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{E}\left[r_i(S_{i,t}^\pi, A_{i,t}^\pi) \right] \leq \frac{C_{\text{ID}}}{\sqrt{N}} + \frac{C_{\text{finite}}}{T}, \tag{12}$$

where C_{finite} is another positive constant independent of N.

Technical overview

Our technical approach uses the Lyapunov drift method, which has found widespread applications in queueing systems, Markov decision processes, reinforcement learning, and so on. While the basic

framework of the drift method is standard, the *key challenge* lies in constructing the right Lyapunov function with the desired properties, where the difficulty is exacerbated by the full heterogeneity of the problem under study. Our construction of such a Lyapunov function is highly novel, yet still natural. We reiterate that fully heterogeneous, high-dimensional stochastic systems are poorly understood in the existing literature. Our approach opens up the possibility of analyzing the steady-state behavior of such systems through the Lyapunov drift method.

In the remainder of this section, we consider the ID policy, also referred to as the policy π. Let $\boldsymbol{X}_t$ denote the system state under it, with the superscript π omitted for brevity. To make this overview more intuitive, here let us assume that $\boldsymbol{X}_t$ converges to its steady state $\boldsymbol{X}_\infty$ in a proper sense such that taking expectations in steady state is the same as taking time averages. However, note that our formal results do not need this assumption and directly work with time averages. We call a function V a Lyapunov/potential function if it maps each possible system state to a nonnegative real number.

General framework of the drift method. Here we briefly describe the general framework of the drift method when applied to our problem. The goal is to construct a Lyapunov function V such that

- **(C1)** $R_N^{\text{rel}} - R(\pi, \boldsymbol{S}_0) \le C_1 \mathbb{E}\left[V(\boldsymbol{X}_\infty)\right]/N + O(1/\sqrt{N})$ for some constant C_1;
- **(C2)** (Drift condition) $\mathbb{E}\left[V(\boldsymbol{X}_{t+1}) \mid \boldsymbol{X}_t\right] - V(\boldsymbol{X}_t) \le -C_2 V(\boldsymbol{X}_t) + O(\sqrt{N})$ for a constant C_2.

The drift condition requires that on average, the value of V approximately decreases (ignoring the additive $O(\sqrt{N})$) after a time step. The drift condition implies a bound on $\mathbb{E}\left[V(\boldsymbol{X}_\infty)\right]$. To see this, let $\boldsymbol{X}_t$ follow the steady-state distribution, which means $\boldsymbol{X}_{t+1}$ also follows the steady-state distribution, and take expectations on both sides of the inequality. Then we get $0 = \mathbb{E}\left[V(\boldsymbol{X}_{t+1})\right] - \mathbb{E}\left[V(\boldsymbol{X}_t)\right] \le -C_2 \mathbb{E}\left[V(\boldsymbol{X}_t)\right] + O(\sqrt{N})$, which implies $\mathbb{E}\left[V(\boldsymbol{X}_\infty)\right] = \mathbb{E}\left[V(\boldsymbol{X}_t)\right] = O(\sqrt{N})$. Combining this with (**C1**) proves the desired $O(1/\sqrt{N})$ upper bound on the optimality gap.

Key challenge: constructing Lyapunov function. We highlight this challenge by contrasting the homogeneous setting and the heterogeneous setting. In the *homogeneous* setting, there is only one optimal state distribution, μ^*. The Lyapunov function in [26] is defined based on the distance between the *empirical state distribution* across arms and μ^*. Specifically, it is based on a set of functions $(h(\boldsymbol{X}_t, D))_{D \subseteq [N]}$ defined as:

$$h(\boldsymbol{X}_t, D) = \left\| \boldsymbol{X}_t(D) - m(D)\mu^* \right\|, \tag{13}$$

where $\boldsymbol{X}_t(D) = (X_t(D, s))_{s \in \mathbb{S}}$ denotes within D, the number of arms in each state s, divided by N; $m(D) = |D|/N$; and the norm $\|\cdot\|$ is a properly defined norm. The idea is that if all arms in D follow the optimal single-armed policy, the state distribution of each arm in D gets closer to μ^*, and thus $\boldsymbol{X}_t(D)$ gets closer to $m(D)\mu^*$ over time.

In the *heterogeneous* setting, we also want to construct a Lyapunov function $h(\boldsymbol{X}_t, D)$ to witness the convergence of any set of arms D if they follow the optimal single-armed policies. However, unlike the homogeneous setting, now it no longer makes sense to aggregate arm states into an empirical state distribution, since each arm's dynamics is distinct. Instead, our Lyapunov function considers $X_{i,t} - \mu_i^*$, where recall $X_{i,t}(s)$ is the indicator that arm i's state is s at time t. A naive first attempt is to construct the Lyapunov function from the pointwise distances, $\|X_{i,t} - \mu_i^*\|$ for each arm i, with a properly defined norm $\|\cdot\|$. However, the pointwise distances are very noisy: $\|X_{i,t} - \mu_i^*\|$ could be large even when the state of arm i independently follows the distribution μ_i^* for each i, a situation when we should view the set of arms as already converged.

Intuitively, to make the Lyapunov function properly reflect the convergence of the set of arms (referred to as "the system" in the rest of the section) following the optimal single-armed policies, we would like it to depend less strongly on the state of each individual arm and focus more on the collective properties of the whole system. Our idea is to *project* the system state onto a properly selected set of *feature vectors*, and construct the Lyapunov function based on how far these projections are from the projections of the optimal state distributions $(\mu_i^*)_{i \in [N]}$. Then what features of the system state do we need to determine whether it has converged in a proper sense? The first feature we consider is the instantaneous reward of the system, $\sum_{i \in D} \langle X_{i,t}, r_i^* \rangle$, where $r_i^* \in \mathbb{R}^{\mathbb{S}}$ is the reward function of arm i under $\bar{\pi}_i^*$, and recall that the inner product is defined between two vectors whose entries correspond to states in $\mathbb{S}$. We also want to keep track of the *ℓ-step ahead expected reward*, $\sum_{i \in D} \left\langle X_{i,t} P_i^\ell, r_i^* \right\rangle$,

for each $\ell \in \mathbb{N}_+$. Intuitively, if $\sum_{i\in D} \left\langle (X_{i,t} - \mu_i^*)P_i^\ell, r_i^* \right\rangle$ is small for each $\ell \in \mathbb{N}$, the reward of the system should remain close to that under the optimal state distributions $(\mu_i^*)_{i\in[N]}$ for a long time; conversely, if the state of each arm i independently follows μ_i^*, each of these features should be small as well. We also consider the *ℓ-step ahead expected type-k cost* for each $\ell \in \mathbb{N}$ and $k \in [K]$ as features, defined analogously.

Combining the above ideas, for any set of arms D, we let the Lyapunov function $h(\boldsymbol{X}_t, D)$ be the supremum of the differences between $\boldsymbol{X}_t$ and μ^* in all the features directions defined above, under proper weightings:

$$h(\boldsymbol{X}_t, D) = \max_{g\in\mathcal{G}} \sup_{\ell\in\mathbb{N}} \left| \sum_{i\in D} \left\langle (X_{i,t} - \mu_i^*)P_i^\ell/\gamma^\ell, g_i \right\rangle \right|, \tag{14}$$

where $\gamma = \exp(-1/(2\tau))$ for τ defined in Assumption 1; each element $g \in \mathcal{G}$ is either $g = (r_i^*)_{i\in[N]}$, or corresponds to the type-k cost for some $k \in [K]$ (See Appendix E of [54] for the definition of $\mathcal{G}$). Note that *dividing* each term by powers of γ is another interesting trick, which induces a negative drift in $h(\boldsymbol{X}_t, D)$ under the optimal single-armed policies (See the proof of Lemma 3 in [54]).

Now with the set of functions $(h(\boldsymbol{X}_t, D))_{D\subseteq[N]}$ defined, we generalize the idea of focus sets in [26] to convert $(h(\boldsymbol{X}_t, D))_{D\subseteq[N]}$ into a Lyapunov function $V(\boldsymbol{X}_t)$. We prove that V satisfies **(C1)** and **(C2)** using the structure of $(h(\boldsymbol{X}_t, D))_{D\subseteq[N]}$.

Remark 2. The idea for constructing $h(\boldsymbol{X}_t, D)$ is potentially useful for analyzing other heterogeneous stochastic systems. At a high level, projecting the system state onto a set of feature vectors (and their future expectations) can be roughly viewed as aggregating system states whose relevant performance metrics *remain close for a sufficiently long time*. This idea provides a new way to measure the distance between two system states in a heterogeneous system, and this distance notation enjoys similar properties as that in a homogeneous system, without resorting to symmetry.

5 Experiments

In this section, we perform two sets of experiments to illustrate the numerical performance of the proposed ID policy for fully heterogeneous WCMDPs.

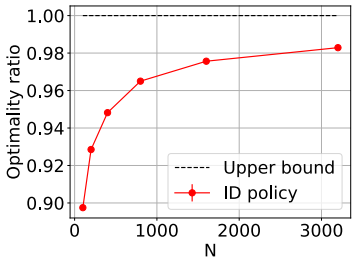

Figure 1: Asymptotic optimality of ID policy.

In the first set of experiments, we demonstrate the asymptotic optimality of the ID policy. We increase the number of arms as $N \in \{100, 200, 400, 800, 1600, 3200\}$. Each arm's MDP has 10 states and 4 actions, with parameters generated uniformly at random in a proper sense. The N-armed problem has 4 budget constraints, with cost functions also generated randomly. More details are provided in Appendix B.1. We simulate the policy for 2×10^4 time steps over 4 replications for each N. To illustrate the performance more clearly, we measure the *optimality ratio*, defined as the ratio between the long-run average reward achieved by a policy and the LP relaxation upper bound R_N^{rel}. Confidence intervals are calculated using the batch means method with a batch size of 4000, but they are typically too small to be visible on figures. Figure 1 shows that the optimality ratio of the ID policy becomes increasingly close to 1 as N increases.

In the second set of experiments, we compare the ID policy with the ERC policy proposed in [46]. As discussed in Section 1, only a few prior papers address heterogeneous WCMDPs. Among them, [46] considers the most general setting, but is still limited to a single-budget constraint, state-independent costs, and typed heterogeneity. To make a fair comparison, we evaluate both policies under this special case, while keeping all other settings the same as in the first set of experiments. ID policy turns out to have a slight improvement over ERC policy in some instances, one of which is shown in Figure 2, and has comparable performance in others. Importantly, unlike ERC, the ID policy applies to far more general classes of WCMDPs.

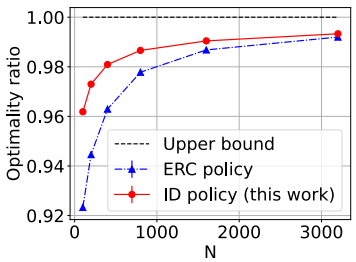

Figure 2: ID policy vs ERC policy.

References

[1] S. Allmeier and N. Gast. Mean field and refined mean field approximations for heterogeneous systems: It works! *Proc. ACM Meas. Anal. Comput. Syst.*, 6(1), Feb. 2022.

[2] S. Allmeier and N. Gast. Accuracy of the graphon mean field approximation for interacting particle systems. *arXiv:2405.08623 [math.PR]*, 2024.

[3] K. E. Avrachenkov and V. S. Borkar. Whittle index based Q-learning for restless bandits with average reward. *Automatica*, 139:110186, 2022.

[4] D. Bertsimas and J. Niño Mora. Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Math. Oper. Res.*, 21(2):257–306, 1996.

[5] A. Biswas, G. Aggarwal, P. Varakantham, and M. Tambe. Learning index policies for restless bandits with application to maternal healthcare. In *Proc. Int. Jt. Conf. Artificial Intelligence (IJCAI)*, pages 1467–1468, 2021.

[6] A. Biswas, G. Aggarwal, P. Varakantham, and M. Tambe. Learning index policies for restless bandits with application to maternal healthcare. In *Proc. Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS)*, page 1467–1468, 2021.

[7] C. Boutilier and T. Lu. Budget allocation using weakly coupled, constrained Markov decision processes. In *Conf. Uncertainty in Artificial Intelligence (UAI)*, pages 52–61, 2016.

[8] D. B. Brown and J. E. Smith. Index policies and performance bounds for dynamic selection problems. *Manage. Sci.*, 66(7):3029–3050, 2020.

[9] D. B. Brown and J. Zhang. Dynamic programs with shared resources and signals: Dynamic fluid policies and asymptotic optimality. *Oper. Res.*, 70(5):3015–3033, 2022.

[10] D. B. Brown and J. Zhang. Fluid policies, reoptimization, and performance guarantees in dynamic resource allocation. *Oper. Res.*, 73(2):1029–1045, 2025.

[11] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, 2019.

[12] J. C. D'Aeth, S. Ghosal, F. Grimm, D. Haw, E. Koca, K. Lau, H. Liu, S. Moret, D. Rizmie, P. C. Smith, G. Forchini, M. Miraldo, and W. Wiesemann. Optimal hospital care scheduling during the SARS-CoV-2 pandemic. *Manage. Sci.*, 69(10):5923–5947, 2023.

[13] I. El Shar and D. Jiang. Weakly coupled deep Q-networks. *Conf. Neural Information Processing Systems (NeurIPS)*, 36, 2023.

[14] N. Gast, B. Gaujal, and C. Yan. Exponential asymptotic optimality of Whittle index policy. *Queueing Syst.*, 104:107–150, 2023.

[15] N. Gast, B. Gaujal, and C. Yan. Linear program-based policies for restless bandits: Necessary and sufficient conditions for (exponentially fast) asymptotic optimality. *Math. Oper. Res.*, 49(4): 2468–2491, 2024.

[16] N. Gast, B. Gaujal, and C. Yan. Reoptimization nearly solves weakly coupled Markov decision processes. *arXiv:2211.01961 [math.OC]*, 2024.

[17] A. Ghosh, D. Nagaraj, M. Jain, and M. Tambe. Indexability is not enough for Whittle: Improved, near-optimal algorithms for restless bandits. In *Proc. Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1294–1302, 2023.

[18] J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[19] J. C. Gittins. Bandit processes and dynamic allocation indices. *J. Roy. Stat. Soc. B Met.*, 41(2): 148–164, 1979.

[20] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani, editor, *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam, 1974.

[21] K. D. Glazebrook, H. M. Mitchell, and P. S. Ansell. Index policies for the maintenance of a collection of machines by a set of repairmen. *Eur. J. Oper. Res.*, 165(1):267–284, 2005.

[22] D. Goldsztajn and K. Avrachenkov. Asymptotically optimal policies for weakly coupled Markov decision processes. *arXiv:2406.04751 [math.OC]*, 2024.

[23] J. T. Hawkins. *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. PhD thesis, Operations Research Center, Massachusetts Institute of Technology, 2003.

[24] D. J. Hodge and K. D. Glazebrook. On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Adv. Appl. Probab.*, 47(3):652–667, 2015.

[25] Y. Hong, Q. Xie, Y. Chen, and W. Wang. Restless bandits with average reward: Breaking the uniform global attractor assumption. In *Conf. Neural Information Processing Systems (NeurIPS)*, 2023.

[26] Y. Hong, Q. Xie, Y. Chen, and W. Wang. Unichain and aperiodicity are sufficient for asymptotic optimality of average-reward restless bandits. *arXiv:2402.05689 [cs.LG]*, 2024.

[27] Y. Hong, Q. Xie, Y. Chen, and W. Wang. Achieving exponential asymptotic optimality in average-reward restless bandits without global attractor assumption. *arXiv:2405.17882 [cs.LG]*, 2024.

[28] W. Hu and P. Frazier. An asymptotically optimal index policy for finite-horizon restless bandits. *arXiv:1707.00205 [math.OC]*, 2017.

[29] J. A. Killian, A. Biswas, S. Shah, and M. Tambe. Q-learning Lagrange policies for multi-action restless bandits. In *Proc. Ann. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, pages 871–881, 2021.

[30] J. A. Killian, L. Xu, A. Biswas, and M. Tambe. Restless and uncertain: Robust policies for restless bandits via deep multi-agent reinforcement learning. In *Conf. Uncertainty in Artificial Intelligence (UAI)*, pages 990–1000, 2022.

[31] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[32] K. Nakhleh and I.-H. Hou. DeepTOP: Deep threshold-optimal policy for MDPs and RMABs. In *Conf. Neural Information Processing Systems (NeurIPS)*, pages 28734–28746, 2022.

[33] K. Nakhleh, S. Ganji, P.-C. Hsieh, I.-H. Hou, and S. Shakkottai. NeurWIN: Neural Whittle index network for restless bandits via deep RL. In *Conf. Neural Information Processing Systems (NeurIPS)*, pages 828–839, 2021.

[34] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Math. Oper. Res.*, 24(2):293–305, 1999.

[35] M. L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2005.

[36] F. Robledo, V. Borkar, U. Ayesta, and K. Avrachenkov. QWI: Q-learning with Whittle index. *ACM SIGMETRICS Perform. Evaluation Rev.*, 49(2):47–50, 2022.

[37] J. N. Tsitsiklis. A short proof of the Gittins index theorem. *Ann. Appl. Probab.*, 4(1):194 – 199, 1994.

[38] P. Varaiya, J. Walrand, and C. Buyukkoc. Extensions of the multiarmed bandit problem: The discounted case. *IEEE Trans. Autom. Control*, 30(5):426–439, 1985.

[39] I. M. Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *Ann. Appl. Probab.*, 26(4):1947–1995, 2016.

[40] S. S. Villar. Indexability and optimal index policies for a class of reinitialising restless bandits. *Probab. Eng. Inf. Sci.*, 30(1):1–23, 2016.

[41] R. Weber. On the Gittins index for multiarmed bandits. *Ann. Appl. Probab.*, 2(4):1024 – 1033, 1992.

[42] R. R. Weber and G. Weiss. On an index policy for restless bandits. *J. Appl. Probab.*, 27(3): 637–648, 1990.

[43] P. Whittle. Multi-armed bandits and the Gittins index. *J. Roy. Stat. Soc. B Met.*, 42(2):143–149, 12 1980.

[44] P. Whittle. Restless bandits: activity allocation in a changing world. *J. Appl. Probab.*, 25:287 – 298, 1988.

[45] G. Xiong and J. Li. Finite-time analysis of Whittle index based Q-learning for restless multi-armed bandits with neural network function approximation. In *Conf. Neural Information Processing Systems (NeurIPS)*, pages 29048–29073, 2023.

[46] G. Xiong, S. Wang, and J. Li. Learning infinite-horizon average-reward restless multi-action bandits via index awareness. In *Conf. Neural Information Processing Systems (NeurIPS)*, pages 17911–17925, 2022.

[47] C. Yan. An optimal-control approach to infinite-horizon restless bandits: Achieving asymptotic optimality with minimal assumptions. In *Proc. IEEE Conf. Decision and Control (CDC)*, pages 6665–6672, 2024.

[48] Z. Yu, Y. Xu, and L. Tong. Deadline scheduling as restless bandits. *IEEE Transactions on Automatic Control*, 63(8):2343–2358, 2018.

[49] G. Zayas-Cabán, S. Jasin, and G. Wang. An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits. *Advances in Applied Probability*, 51(3):745–772, 2019.

[50] J. Zhang. Leveraging nondegeneracy in dynamic resource allocation. *Available at SSRN*, 2024.

[51] X. Zhang and P. I. Frazier. Restless bandits with many arms: Beating the central limit theorem. *arXiv:2107.11911 [math.OC]*, July 2021.

[52] X. Zhang and P. I. Frazier. Near-optimality for infinite-horizon restless bandits with many arms. *arXiv:2203.15853 [cs.LG]*, 2022.

[53] X. Zhang, Y. Hong, and W. Wang. Projection-based Lyapunov method for fully heterogeneous weakly-coupled MDPs. GitHub repository, `https://github.com/YigeHong/wcmdp-fully-hetero/`, 2025.

[54] X. Zhang, Y. Hong, and W. Wang. Projection-based Lyapunov method for fully heterogeneous weakly-coupled MDPs. *arXiv:2502.06072v5 [cs.LG]*, 2025.

[55] Z. Zhao and D. Mukherjee. Optimal rate-matrix pruning for heterogeneous systems. *ACM SIGMETRICS Perform. Evaluation Rev.*, 51(4):26–27, 2024.

[56] Z. Zhao, D. Mukherjee, and R. Wu. Exploiting data locality to improve performance of heterogeneous server clusters. *Stoch. Syst.*, 14(3):229–272, 2024.

[57] J. Zhou, S. Mao, G. Yang, B. Tang, Q. Xie, L. Lin, X. Wang, and D. Wang. RL-MPCA: A reinforcement learning based multi-phase computation allocation approach for recommender systems. In *Proceedings of the ACM Web Conference 2023*, pages 3214–3224, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contribution claimed in the abstract and the introduction is an asymptotic optimality theorem; it corresponds to Theorem 1 in the main results section (Section 4).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 4, after stating Theorem 1, we pointed out that we did not attempt to optimize the optimality gap bound's dependence on any parameters other than N.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have stated our assumption (Assumption 1) in Section 4 right before the main theorem (Theorem 1); we have added a cross-reference to the assumption in the theorem's statement. The proofs are given in the appendix of the arXiv version, and we have referred to the corresponding sections in the arXiv version in the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Section 5 and Appendix B, we have discussed in detail the method of generating random instances, the simulation setting, and the baseline policy.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have cited the link to our GitHub where the simulation code is available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 5 and Appendix B, we have discussed in detail the method of generating random instances, the simulation setting, and the baseline policy. We have also cited the link to the GitHub where the code is available.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have plotted the confidence intervals in the figures, which turn out to be negligibly small.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our experiments are small-scaled and can be completed on a standard PC (6-Core Intel Core i7) within one day.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Since this is a theoretical paper, we believe that there is no potential harms caused by our research process, and our result does not have negative social impact and harmful consequences.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our result is theoretical and we see no direct way that it can be linked to a technology with negative social impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper does not pose such risks, since it is mainly theoretical and only has small-scale simulation experiments.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Our paper is mainly mathematical with only small-scale experiments. It does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have uploaded our simulation code to GitHub.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This is a theory paper and it does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This is a theory paper and it does not involve crowdsourcing or research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development of the paper does not involve LLM.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Detailed review on related work

In this section, we provide a more detailed, though still non-exhaustive, review of the literature. We mainly focus on theoretical work with formal performance guarantees, leaving out the extensive body of work with empirical results. We begin by surveying papers with the same reward criterion as ours, i.e., infinite-horizon average-reward criterion. In this setting, we first review existing work on restless bandits (RBs), which is an extensively studied special case of WCMDPs. We then give a more detailed review of existing results on WCMDPs. Next, we turn to work that considers other reward criteria: the finite-horizon total-reward criterion and the infinite-horizon discounted-reward criterion. Finally, we briefly mention other problems that are related to WCMDPs.

Infinite-horizon average-reward RBs. For the homogeneous arm setting, the first asymptotic optimality result for average-reward homogeneous RBs is established by Weber and Weiss [42]: they show that the *Whittle index policy* [44] achieves an $o(1)$ optimality gap as the number of arms N goes to infinity. There are three key assumptions in [42]: indexability, the global attractor property, and the aperiodic-unichain condition. These assumptions have been gradually relaxed in the subsequent papers. In particular, [39] proposes a class of priority policies based on an LP relaxation. This class of policies, later referred to as the *LP-Priority policies*, generalizes the Whittle index policy. Each LP-Priority policy achieves an $o(1)$ optimality gap without requiring indexability. The work [25] is the first one that breaks the global attractor property assumption. The authors propose a policy named Follow-the-Virtual-Advice (FTVA), which achieves an $O(1/\sqrt{N})$ optimality gap under an assumption named the Synchronization Assumption; there exist problem instances that satisfy the Synchronization Assumption but do not satisfy the global attractor property. Later work [26] further relaxes the conditions and only requires the aperiodic-unichain condition to achieve an $O(1/\sqrt{N})$ optimality gap. More recently, Yan [47] proposes the *align-and-steer policy*, which further weakens the aperiodic-unichain condition and achieves an $o(1)$ optimality gap.

Parallel to relaxing the assumptions for asymptotic optimality, another line of work has focused on improving the optimality gap beyond $O(1/\sqrt{N})$ under slightly stronger assumptions [14, 15, 27]. Specifically, Gast et al. [14] show that the Whittle index policy has an $O(\exp(-cN))$ optimality gap for some constant $c > 0$. In addition to indexability and the aperiodic-unichain condition, [14] also requires a stronger version of the global attractor property named Uniform Global Attractor Property (UGAP), and a condition called non-singularity. Subsequently, Gast et al. [15] show that LP-Priority policies achieve $O(\exp(-cN))$ optimality gaps assuming the aperiodic-unichain condition, UGAP, and a non-degenerate condition that is equivalent to non-singularity. More recently, Hong et al. [27] propose a *two-set policy* that also achieves an $O(\exp(-cN))$ optimality gap while replacing UGAP of [15] with a much weaker condition named local stability.

Among the aforementioned work, [39] and [25] have addressed the heterogeneous arm setting. The setting studied in [39] is the typed heterogeneous setting, where the N arms are divided into a constant number of types as $N \to \infty$. The paper [25] includes an extension to the fully heterogeneous setting. In particular, the proposed FTVA policy generalizes to the fully heterogeneous setting and leads to an optimality gap of $O(\overline{\tau}^{\text{sync}}_{\max}/\sqrt{N})$, where $\overline{\tau}^{\text{sync}}_{\max}$ is the maximum of a quantity called the synchronization time across all arms. Therefore, for this result to yield asymptotic optimality, there needs to be a further assumption that $\overline{\tau}^{\text{sync}}_{\max} = o(\sqrt{N})$. We re-emphasize that the FTVA policy does not generalize to WCMDPs. The main reason is that FTVA heavily relies on the fact that an RB only constrains the number of pulls, while a WCMDP has budget constraints on cost functions each depending on both the state and action of an arm.

Infinite-horizon average-reward WCMDPs. Work on average-reward WCMDPs remains relatively scarce, and to our knowledge, fully heterogeneous WCMDPs have yet to be addressed. Compared to RBs, a WCMDP allows multiple actions for each arm and multiple cost constraints, where each cost function is a function of both the state and the action of an arm. The line of research [22, 24, 39, 46] has generalized the action space and cost model of RBs to WCMDPs, and some of them allow for typed heterogeneity. In particular, Hodge and Glazebrook [24] generalize the Whittle index policy to homogeneous WCMDPs with a single constraint and multiple actions, where each action represents a different activation level and has a different cost. Verloop [39] extends the LP-Priority policies to typed heterogeneous WCMDPs with a single constraint and multiple actions, but requires each action to have the same cost. Then Xiong et al. [46] propose another index policy

for typed heterogeneous WCMDPs with a single constraint, and allow each action to have a different cost. In the three papers above [24, 39, 46], $o(1)$ optimality gaps have been proved under a similar set of assumptions as in most restless bandit papers, i.e., aperiodic-unichain (or irreducibility) condition, global attractor property, and a generalized indexability condition if the policy is Whittle index. Finally, Goldsztajn and Avrachenkov [22] consider homogeneous WCMDPs with multiple actions and multiple cost constraints with general cost functions, and propose a class of policies with $o(1)$ optimality gaps under a weaker-than-standard aperiodic-unichain condition.

Finite-horizon total-reward RBs and WCMDPs. Next, we review the asymptotic optimality results for finite-horizon total-reward RBs and WCMDPs. The finite-horizon setting is better understood than the average-reward setting, partly because the analysis in the finite horizon is not hindered by the technical conditions arising in average-reward MDPs, such as the unichain condition and the global attractor property. On the other hand, computing asymptotically optimal policies for the finite-horizon setting is more complicated, requiring a careful optimization of the transient sample paths.

Hu and Frazier [28] propose the first asymptotically optimal policy for finite-horizon homogeneous RBs, which achieves an $o(1)$ optimality gap without any assumptions.[2] Since then, researchers have established asymptotic optimality in more general settings [8, 10, 12, 17, 49]. Among these papers, the most general setting is addressed by Brown and Zhang [10], where the authors consider fully heterogeneous WCMDPs; they obtain $O(1/\sqrt{N})$ optimality gaps for a naive fluid policy and a reoptimization-based fluid policy, among a few other results to be reviewed in the next paragraphs. Notably, there is also a further generalization of fully heterogeneous WCMDPs, which involves an exogenous state that affects all arms' transitions, rewards, and constraints; Brown and Zhang [9] propose this setting, where they achieve an $O(1/\sqrt{N})$ optimality gap using a dynamic fluid policy.

Another line of work has improved the optimality gap beyond the order $O(1/\sqrt{N})$ by making an additional assumption called non-degeneracy. Specifically, Zhang and Frazier [51] establish an $O(1/N)$ optimality gap in non-degenerate homogeneous RBs. Gast et al. [15] then propose a different policy for the same setting that improves the optimality gap to $O(\exp(-cN))$. Later, Gast et al. [16] and Brown and Zhang [10] establish $O(1/N)$ optimality gaps for homogeneous and typed heterogeneous WCMDPs, respectively, assuming non-degeneracy. More recently, Zhang [50] proposes a policy for fully heterogeneous WCMDPs; the optimality gap bound of the policy interpolates between $O(1/\sqrt{N})$ and $O(1/N)$ as the degree of non-degeneracy varies, unifying the performance bounds in the degenerate and non-degenerate cases.

Despite the generality of the settings and the fast diminishing rate of the optimality gaps as $N \to \infty$, most of the optimality gaps in the finite-horizon setting depend *super-linearly* on the time horizon, so they do not carry over to the infinite-horizon average-reward setting. There are two exceptions, [10, 16], which achieve optimality gaps that depend linearly on the time horizon under some special conditions: [10] requires all entries of the transition kernels to be bounded away from zero; [16] assumes an ergodicity property, which requires two arms in any different states to synchronize in a fixed number of steps under *any sequence of actions* with a positive probability. However, without these conditions, the optimality gaps in [10, 16] depend quadratically on the time horizon. Apart from having distinct optimality gap bounds, all existing algorithms in the finite-horizon setting need to (sometimes repeatedly) solve LPs whose number of variables scales with the time horizon, so they cannot be directly adapted to the infinite-horizon average-reward setting.

Infinite-horizon discounted-reward RBs and WCMDPs. Asymptotic optimality has also been established for RBs and WCMDPs under the infinite-horizon discounted-reward criterion. In particular, Brown and Smith [8] establish an $O(N^{\log_2(\sqrt{\gamma})})$ optimality gap for fully heterogeneous WCMDPs when $\gamma \in (1/2, 1)$. Subsequently, Ghosh et al. [17], Zhang and Frazier [52] obtain $O(1/\sqrt{N})$ optimality gaps for homogeneous and typed heterogeneous RBs, and Brown and Zhang [10] establish the same order of optimality gap for fully heterogeneous WCMDPs. Similar to the finite-horizon setting, most of these optimality gaps depend super-linearly on the effective time horizon $1/(1-\gamma)$

[2] Here, we measure the optimality gap in terms of the reward per arm, to be consistent with our convention. However, in the papers on the finite-horizon total-reward setting, it is also common to measure the optimality gap in terms of the total reward of all arms, which differs from ours by a factor of N. We also adopt the same convention when reviewing the papers on the infinite-horizon discounted-reward setting.

unless special conditions hold [10], so they do not carry over to the infinite-horizon average-reward setting. The policies here also require solving LPs whose complexities scale with the effective time horizon.

Restful bandits, stochastic multi-armed bandits. A special case of RB is the restful bandit (also referred to as nonrestless bandits, rested bandits, or Markovian bandits), where an arm's state does not change if it is not pulled. The restful bandit problem has been widely studied, where the celebrated Gittins index policy is proven to be optimal [4, 19, 20, 37, 38, 41, 43]. We refer the readers to [18] for a comprehensive review of Gittins index and restful bandits. Another related topic is the stochastic multi-armed bandit (MAB) problem, which has been extensively studied; see the book [31] for a comprehensive overview. The key distinction between MABs and RBs is that arms are stateless in MABs, but stateful in RBs. Consequently, MAB becomes trivial with known model parameters, whereas RB is still non-trivial.

B Experimental details

In this appendix, we provide details of the two WCMDP instances considered in Section 5, and the definition of the baseline policy, the ERC policy from [46]. The complete code for these experiments is available on GitHub [53], and all results can be reproduced within 24 hours on a standard PC (e.g., 6-Core Intel Core i7).

B.1 WCMDP instance generation

Details of the WCMDP instance 1 (In Figure 1). $|\mathbb{S}| = 10$, $|\mathbb{A}| = 4$, $K = 4$. For each $i \in [N]$, $s \in \mathbb{S}$, and $a \in \mathbb{A}$, $r_i(s, 0) = 0$, and $r_i(s, a)$ is independently sampled from the uniform distribution over $[0, 1]$ for each $a \neq 0$; $P_i(s, a, \cdot)$ is independently sampled from the uniform distribution over the probability simplex. As for the cost function, for each $i \in [N]$, $s \in \mathbb{S}$, $a \in \mathbb{A}$, and $k \in [K]$, we have $c_{k,i}(s, 0) = 0$, and $c_{k,i}(s, a)$ is independently sampled from the uniform distribution over $[0, 1]$ for each $a \neq 0$. For each $k \in [K]$, α_k is uniformly sampled from $\{0.05, 0.1, 0.15, \ldots, 0.45\}$, i.e., the uniform distribution over integer multiples of 0.05 in the interval $(0, 0.5)$.

Details of the WCMDP instance 2 (In Figure 2). This instance is typed heterogeneous with 10 types, with equal fraction of arms in each type. Each arm has $|\mathbb{S}| = 10$ states, $|\mathbb{A}| = 4$ actions. For each type, the reward function and transition kernel are generated in the same way as each arm in ithe first instance. The budget is also sampled from the same distribution as the first instance. There is a single budget constraint. The cost function depends only on the action, consistent with the setting in [46]; the cost of each of action $a \in \mathbb{A}$ is sampled from the uniform distribution over $[0, 1]$.

B.2 Defin of ERC policy

The ERC policy [46] solves the same LP in (5) to define the single-armed policies $\bar{\pi}_i^*(a|s)$ as in (7). At each time step, the policy computes an index for each arm-state pair $(i, s) \in [N] \times \mathbb{S}$:

$$\mathcal{I}(i, s) = \sum_{a \in \mathbb{A}} \bar{\pi}_i^*(a|s) r_i(s, a).$$

It then iterates through the arms in descending order of these indices. For each arm, it samples an action $a \sim \bar{\pi}_i^*(\cdot \mid s)$ and applies it only if the budget is sufficient. If not, it defaults to action 0.