Lyapunov-Based Sample Complexity Analysis for Weakly-Coupled MDPs

Tianhao Wu[1] **Matthew Zurek**[1] **Yudong Chen**[1] **Weina Wang**[2] **Qiaomin Xie**[1]
[1]University of Wisconsin–Madison [2]Carnegie Mellon University
{tianhao.wu, matthew.zurek, qiaomin.xie}@wisc.edu
yudongchen@cs.wisc.edu weinaw@cs.cmu.edu

Abstract

We study learning in average-reward weakly coupled Markov decision processes (WCMDPs) with heterogeneous arms. Naive approaches suffer exponential computation and sample complexity in the number of subsystems. We study a plug-in approach built on an efficient planning algorithm, which attains the first finite-sample (PAC) optimality-gap guarantees with polynomial sample complexity. This result is established under a new framework built on a Lyapunov analysis of a reference policy combined with a Lyapunov drift transfer technique, which can be viewed as a generalization of the classical simulation lemma.

1 Introduction

Weakly-coupled Markov decision processes (WCMDPs) [14] are a natural abstraction for large-scale decision-making systems—from job scheduling [40] and machine maintenance [13] to healthcare [5], surveillance [32], and online advertising [6, 44]. A WCMDP comprises N arms/subproblems, each of which is itself an MDP. In the heterogeneous case, these MDPs may differ across arms. At each timestep, the controller chooses an action for each arm, upon which the arms evolve independently. The arms are coupled by a set of per-period global budget constraints on the actions: for each resource type, the aggregate cost summed over all arms' actions must not exceed a prescribed value.

A baseline sample complexity bound for learning WCMDPs could be derived by ignoring the weakly-coupled structure and treating the system as a single tabular MDP. Recent work has essentially resolved the sample complexity of tabular average-reward MDPs [18, 33, 35, 47, 46] (see Appendix A for more details), in particular showing a $\Theta(|\mathcal{S} \times \mathcal{A}|)$ sample size dependence, where $\mathcal{S} \times \mathcal{A}$ is the state-action space of the tabular MDP. Hence this naive baseline would lead to a dependence on the product of the sizes of the state-action spaces of each subproblem MDP, which is exponential in the number of subsystems N. Moreover, it would anyways be computationally intractable to solve WCMDPs via this exact tabular reduction. This motivates our central question: *how can one learn a near-optimal stationary policy in average-reward WCMDPs without incurring exponential dependence on N?* Our work answers this by exploiting weak coupling structurally, leading to finite-sample guarantees with only *polynomial* dependence on N.

For average-reward WCMDP, most existing work focuses on the planning setting where the MDP model is known. Recent work studies the *learning* setting with finite data, particularly for Restless Bandits (RBs), a special case of WCMDP. The work [3] proves asymptotic convergence of Q-learning guided by Whittle-index, and [2] proposes a Lagrange Index Policy and establishes asymptotic optimality in largely homogeneous RBs. Most related to us is [39], which proposes an index-aware algorithm for multi-action RBs and proves sublinear regret. The *fully heterogeneous* WCMDPs we study allow for multiple per-period budget constraints and are more general then RBs. Moreover, we consider a offline dataset/generative model rather than online trajectories, and our results establish

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

finite-sample PAC optimality-gap guarantees, in contrast to the regret bounds or asymptotic guarantees in existing literature. See Appendix A for additional discussion on related work.

Our Contribution We study average-reward WCMDPs with N fully heterogeneous arms and K per-period budget constraints in the generative model setting. We develop a *plug-in* approach built on an efficient planning algorithm—specifically the ID Policy with Reassignment (Algorithm 1)—that handles heterogeneity. Using n samples drawn from the generative model, we estimate the MDP model for each arm and plug the estimates into the planning algorithm. Under a unichain and mixing assumption on each arm (Assumption 1), we show that our algorithm achieves an optimality gap of $O\left(N\sqrt{\frac{S+\log(SAN/\eta)}{n}}\right) + O\left(\frac{1}{\sqrt{N}}\right)$ with probability $1-\eta$, where SA is the size of the state-action space per arm. This is the first *finite-sample (PAC)* optimality-gap guarantee for WCMDPs whose sample and computational complexities scale *polynomially* in N. To establish this result, we develop a new framework that builds on a Lyapunov analysis of a reference policy combined with a Lyapunov drift transfer technique, generalizing the classical simulation lemma. This framework applies whenever the reference policy admits such a Lyapunov analysis, and hence we believe it is broadly applicable to the analysis of other stochastic systems and reinforcement learning algorithms.

2 Problem setup

A weakly coupled Markov decision process consists of N arms. Each arm $i \in [N]$ is an MDP $\mathcal{M}_i = (\mathcal{S}, \mathcal{A}, p_i, r_i, (c_{k,i})_{k\in[K]})$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces with cardinalities $|\mathcal{S}| = S, |\mathcal{A}| = A$, and $p_i(s_i' \mid s_i, a_i)$ is the transition probability from state s_i to s_i' under action a_i. Denote the joint state and action by $\boldsymbol{s} = (s_1, \ldots, s_N) \in \mathcal{S}^N$ and $\boldsymbol{a} = (a_1, \ldots, a_N) \in \mathcal{A}^N$. The N arms evolve independently given the joint action. If arm i is in state s_i and takes action a_i, it yields reward $r_i(s_i, a_i) \in [0, r_{\max}]$ and incurs costs $c_{k,i}(s_i, a_i) \in [0, c_{\max}]$ for $k \in [K]$. The joint actions $\boldsymbol{a}$ must satisfy budget constraints $\sum_{i\in[N]} c_{k,i}(s_i, a_i) \le \alpha_k N, \forall k \in [K]$. We assume a dummy action $0 \in \mathcal{A}$ exists such that $c_{k,i}(s, 0) = 0$ for all k, i, s, hence the all-0 joint action is always feasible.

2.1 Policy, state and performance criterion

For a stationary policy $\pi : \mathcal{S}^N \to \mathcal{A}^N$, we write $\mathbb{P}^\pi(\cdot)$ and $\mathbb{E}^\pi[\cdot]$ for probability and expectation under the law induced by this policy. Let $S_{i,t}$ denote the state of arm i and $\boldsymbol{S}_t = (S_{i,t})_{i\in[N]}$ the system state; the joint action is $\boldsymbol{A}_t = (A_{i,t})_{i\in[N]}$. Define the one-hot representation $X_{i,t} = (X_{i,t}(s))_{s\in\mathbb{S}} \in \mathbb{R}^S$, where $X_{i,t}(s) = 1$ if $S_{i,t} = s$ and 0 otherwise. Also define the system matrix $X_t \in \mathbb{R}^{N\times S}$ whose i-th row is $X_{i,t}$. We use $\boldsymbol{S}_\infty, \boldsymbol{A}_\infty, X_\infty$ to denote the random variables following the stationary distributions of $\boldsymbol{S}_t, \boldsymbol{A}_t, X_t$. We only consider stationary Markov policies. The long-run average reward a.k.a. gain of a policy π from initial state $\boldsymbol{s}_0$ is defined by $\rho^\pi(\boldsymbol{s}_0) \triangleq \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{N}\sum_{i\in[N]} \mathbb{E}^\pi\left[r(S_{i,t}, A_{i,t})\right]$. A WCMDP is the following optimization problem:

$$\max_{\pi} \quad \rho^\pi(\boldsymbol{s}_0) \quad \text{s.t.} \quad \textstyle\sum_{i\in[N]} c_{k,i}(S_{i,t}, A_{i,t}) \le \alpha_k N, \quad \forall k \in [K], \forall t \ge 0. \tag{1}$$

It is a standard result [28, Theorem 9.1.8] that for finite MDPs, there always exists a stationary Markov policy that attains the optimal average reward, denoted by $\rho^\star$, namely the maximum of (1).

2.2 LP relaxation and optimal single-armed policy

Below we present the linear programming (LP) relaxation of the N-armed RB problem given in [41], which plays a central role in the analysis of RB policies:

$$\max_{(y_i(s,a))_{i\in[N],\, s\in\mathcal{S},\, a\in\mathcal{A}}} \quad \tfrac{1}{N}\textstyle\sum_{i\in[N],\, s\in\mathcal{S},\, a\in\mathcal{A}} y_i(s,a) r_i(s,a) \tag{2a}$$

$$\text{subject to} \quad \tfrac{1}{N}\textstyle\sum_{i\in[N]}\sum_{s\in\mathcal{S},\, a\in\mathcal{A}} y_i(s,a) c_{k,i}(s,a) \le \alpha_k, \quad \forall k \in [K], \tag{2b}$$

$$\textstyle\sum_{s'\in\mathcal{S},\, a'\in\mathcal{A}} p_i(s \mid s', a') y_i(s', a') = \sum_{a\in\mathcal{A}} y_i(s,a), \quad \forall s \in \mathcal{S}, i \in [N], \tag{2c}$$

$$\textstyle\sum_{s'\in\mathcal{S},\, a'\in\mathcal{A}} y_i(s', a') = 1, \quad y_i(s,a) \ge 0, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, i \in [N]. \tag{2d}$$

Letting ρ^{rel} be the optimal value of (2), $\{y^\star(s,a)\}_{s\in\mathcal{S},a\in\mathcal{A}}$ is the corresponding optimal solution. [41] have already shown that $\rho^{\mathrm{rel}} \geq \rho^\star(\boldsymbol{s}_0), \forall \boldsymbol{s}_0 \in \mathcal{S}^N$. This relation allows us to bound the optimality gap of any policy π using the inequality $\rho^\star(\boldsymbol{s}_0) - \rho^\pi(\boldsymbol{s}_0) \leq \rho^{\mathrm{rel}} - \rho^\pi(\boldsymbol{s}_0), \forall \boldsymbol{s}_0 \in \mathcal{S}^N$.

2.3 Learning under a generative model

We assume access to a generative model (or simulator) that enables independent sampling from the transition distribution $p_i(\cdot \mid s,a)$ for any arm i and any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$. While the transition kernel p_i is not explicitly known, we can collect n independent samples $S^1_{s,a},\ldots,S^n_{s,a}$ from $p_i(\cdot \mid s,a)$ for each $i \in \{1,2,\ldots,N\}, (s,a) \in \mathcal{S}\times\mathcal{A}$. Based on these samples, we construct an empirical estimate of the single-arm transition kernel: $\widehat{p}_i(s' \mid s,a) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{S^i_{s,a} = s'\}, \forall s' \in \mathcal{S}$. Accordingly, we can construct a product-form empirical model $(\widehat{P}, r)$ for the N-armed system:

$$\widehat{P}(\boldsymbol{s}' \mid \boldsymbol{s},\boldsymbol{a}) = \widehat{P}(s'_1,\ldots,s'_N \mid s_1,\ldots,s_N,a_1,\ldots,a_N) := \textstyle\prod_{i=1}^N \widehat{p}_i(s'_i \mid s_i,a_i).$$

We define the LP relaxation of the empirical WCMDP problem similarly to the LP (2) but with true models p_i replaced by $\widehat{p}_i$; see (8) in Appendix B. Let $\widehat{\rho}^{\mathrm{rel}}$ be the optimal value of the empirical LP (8), $\{\widehat{y}^\star(s,a)\}_{s\in\mathcal{S},a\in\mathcal{A}}$ be the corresponding optimal solution, $\widehat{\pi}^\star_i$ be the optimal single-armed policy. Let $\widehat{\rho}^\pi \in \mathbb{R}^{S^N}$ be the gain of a policy π under the empirical MDP $(\widehat{P}, r)$.

3 Main result

Before stating the main result, we fix notation and the standing assumption for the single-armed policies, and then we introduce the planning algorithm (Algorithm 1, adapted from [41]). Fixing an arm i, for each stationary Markov policy $\pi : \mathcal{S} \to \mathcal{A}$, the state of arm i evolves as a Markov chain on $\mathcal{S}$ with transition matrix $p^i_\pi = (p^i_\pi(s,s'))_{s,s'\in\mathcal{S}}$. We assume the chain is unichain (one recurrent class, possibly with transient states) and denote its unique stationary distribution $\mu^i_\pi = (\mu^i_\pi(s))_{s\in\mathcal{S}}$. Define the mixing time τ^i_π as $\tau^i_\pi := \max_{s\in\mathcal{S}} \min\left\{t \in \mathbb{N} : \left\|(p^i_\pi)^t(s,\cdot) - \mu^i_\pi(\cdot)\right\|_1 \leq \frac{1}{4}\right\}$, which is finite if the chain is aperiodic. We impose the following assumption throughout the paper:

Assumption 1 (Unichain and Uniform Mixing)**.** *We assume that for each arm i, under any policy π in the single-armed MDP, the induced Markov chain is a unichain. Furthermore, we assume that all such Markov chains mix uniformly, in the sense that their mixing times are uniformly bounded above by a constant τ_0, i.e $\forall i \in [N], \forall \pi, \tau^i_\pi \leq \tau_0$. Define $\tau = (3 + \log_2 S)\tau_0$.*

We consider the optimal single-armed policy $\bar{\pi}^\star_i$ defined in [41]. With Assumption 1, the Markov chain induced by $\bar{\pi}^\star_i$ converges to a unique stationary distribution, denoted by $\mu^\star_i = (\mu^\star_i(s))_{s\in\mathcal{S}}$.

Now we describe the planning algorithm we use. We begin with a one-time preprocessing step: solve the LP in (2) to obtain the optimal state–action frequencies $(y^\star_i(s,a))$, compute the optimal single-armed policies $\bar{\pi}^\star_i$, and reassign arm indices using Algorithm 2. In the online phase at time t, for each arm i we sample an ideal action $A^{\mathrm{ideal}}_{i,t} \sim \bar{\pi}^\star_i(\cdot \mid S_{i,t})$. We then attempt to implement these ideal actions in increasing order of arm IDs, setting the real action of arm 1 to $A^{\mathrm{ideal}}_{1,t}$, then arm 2, and so on, as long as all per-period global budget constraints are satisfied. As soon as any resource budget is exhausted or would be violated, we stop and assign the no-cost action 0 to all remaining arms.

Theorem 1. *Consider any N-armed WCMDP with initial state $\boldsymbol{s}_0$ satisfying Assumption 1. Feed $\widehat{\mathcal{M}}_i = (\mathcal{S},\mathcal{A},\widehat{p}_i,r_i,(c_{k,i})_{k\in[K]})$ into Algorithm 1 and obtain $\widehat{\pi}_{\mathrm{ID}}$. For any $0 < \eta < 1$ and sample size n, there exist constants C', C'' independent of N such that, with probability at least $1-\eta$,*

$$\rho^\star(\boldsymbol{s}_0) - \rho^{\widehat{\pi}_{\mathrm{ID}}}(\boldsymbol{s}_0) \leq C' N \sqrt{\frac{S + \log(SAN/\eta)}{n}} + \frac{C''}{\sqrt{N}}$$

To the best of our knowledge, this is the first work to provide *finite-sample (PAC)* guarantees for *average-reward* weakly coupled MDPs (WCMDPs) whose sample complexity depends *polynomially*—rather than exponentially—on the system dimension (number of subsystems N).

4 Technical Overview

Our framework builds on a Lyapunov analysis of a reference policy combined with a Lyapunov drift transfer technique, which can be viewed as a generalization of the classical simulation lemma. Here

we provide an overview of this framework. We remark that this framework applies whenever the reference policy admits such a Lyapunov analysis, and is therefore more broadly applicable.

Our goal is to upper bound $\rho^\star - \rho^{\widehat{\pi}_{\mathrm{ID}}}$. For both the true MDP and the empirical MDP, one can derive upper bounds on the gain of any policy via a linear program relaxation [41], denoted as ρ^{rel} and $\widehat{\rho}^{\mathrm{rel}}$, respectively. Then

$$\rho^\star - \rho^{\widehat{\pi}_{\mathrm{ID}}} \leq \rho^{\mathrm{rel}} - \rho^{\widehat{\pi}_{\mathrm{ID}}} = \rho^{\mathrm{rel}} - \widehat{\rho}^{\pi_{\mathrm{ID}}} + \widehat{\rho}^{\pi_{\mathrm{ID}}} - \rho^{\widehat{\pi}_{\mathrm{ID}}} \leq \rho^{\mathrm{rel}} - \widehat{\rho}^{\pi_{\mathrm{ID}}} + \widehat{\rho}^{\mathrm{rel}} - \rho^{\widehat{\pi}_{\mathrm{ID}}}. \tag{3}$$

We focus on bounding the term $\rho^{\mathrm{rel}} - \widehat{\rho}^{\pi_{\mathrm{ID}}}$; the term $\widehat{\rho}^{\mathrm{rel}} - \rho^{\widehat{\pi}_{\mathrm{ID}}}$ can be bounded in a similar manner. The standard approach for analyzing the sample complexity of plug-in planning algorithms (e.g. [4, 1, 22, 45]) is to utilize the *simulation lemma*, which is the identity

$$\rho^{\pi_{\mathrm{ID}}} - \widehat{\rho}^{\pi_{\mathrm{ID}}} = \widehat{\mathbb{E}}^{\pi_{\mathrm{ID}}}\left[(P_{\pi_{\mathrm{ID}}} - \widehat{P}_{\pi_{\mathrm{ID}}})h^{\pi_{\mathrm{ID}}}(X_\infty)\right], \tag{4}$$

where $h^{\pi_{\mathrm{ID}}}$ is the relative value function of π_{ID} (see for instance [24, 10, 45]). Because $\rho^{\mathrm{rel}} \geq \rho^{\pi_{\mathrm{ID}}} \geq \rho^{\mathrm{rel}} - O(1/\sqrt{N})$ [41], (4) could be used to control $\rho^{\mathrm{rel}} - \widehat{\rho}^{\pi_{\mathrm{ID}}}$ (and furthermore they are equivalent up to $O(1/\sqrt{N})$). The key problem with this approach is that it is unclear how to control $h^{\pi_{\mathrm{ID}}}$. In the framework below, the role of $h^{\pi_{\mathrm{ID}}}$ is replaced by the Lyapunov function V, which is constructed explicitly by [41] and already enjoys known bounds on its size (i.e. $\|V\|_\infty$).

Lyapunov analysis. We take the ID policy proposed in [41], denoted by π_{ID}, as the reference policy. The optimality gap of π_{ID} is analyzed via a Lyapunov function V, which satisfies two key properties (for some constants β_V, K_V, C_1 and C_2):

- (**C1**) **Drift:** For all states x, $\mathbb{E}^{\pi_{\mathrm{ID}}}[V(X_{t+1}) - V(X_t) \mid X_t = x] \leq -\beta_V V(x) + K_V\sqrt{N}$;
- (**C2**) **Optimality gap dominance:** $\rho^{\mathrm{rel}} - \rho^{\pi_{\mathrm{ID}}} \leq C_1 \mathbb{E}^{\pi_{\mathrm{ID}}} V(X_\infty)/N + C_2/\sqrt{N}$.

Moreover, the dominance property extends to the empirical MDP:

- (**C2′**) **Optimality gap dominance:** $\rho^{\mathrm{rel}} - \widehat{\rho}^{\pi_{\mathrm{ID}}} \leq C_1 \widehat{\mathbb{E}}^{\pi_{\mathrm{ID}}} V(X_\infty)/N + C_2/\sqrt{N}$.

Drift transfer. Let $P_{\pi_{\mathrm{ID}}}$ and $\widehat{P}_{\pi_{\mathrm{ID}}}$ be the transition probability matrices (generators) of the Markov chains under π_{ID} in the true and empirical models, respectively. Treating V as a column vector, we can write $\mathbb{E}^{\pi_{\mathrm{ID}}}\left[V(X_{t+1}) \mid X_t = x\right] = P_{\pi_{\mathrm{ID}}} V(x)$, which is the vector $P_{\pi_{\mathrm{ID}}} V$ evaluated at state x. Then the drift condition (**C1**) can be rewritten as

$$(P_{\pi_{\mathrm{ID}}} - I)V(x) \leq -\beta_V V(x) + K_V\sqrt{N}.$$

By adding and subtracting $\widehat{P}_{\pi_{\mathrm{ID}}}$, this implies a drift condition in the *empirical* system

$$(\widehat{P}_{\pi_{\mathrm{ID}}} - I)V(x) = (P_{\pi_{\mathrm{ID}}} - I)V(x) + (\widehat{P}_{\pi_{\mathrm{ID}}} - P_{\pi_{\mathrm{ID}}})V(x) \leq -\beta_V V(x) + \Delta \tag{5}$$

where $\Delta := K_V\sqrt{N} + (\widehat{P}_{\pi_{\mathrm{ID}}} - P_{\pi_{\mathrm{ID}}})V(x)$ contains an additional term capturing model inaccuracy.

Empirical performance bound Given the empirical drift condition (5), bounding $\rho^{\mathrm{rel}} - \widehat{\rho}^{\pi_{\mathrm{ID}}}$ proceeds analogously to how a bound on $\rho^{\mathrm{rel}} - \rho^{\pi_{\mathrm{ID}}}$ is derived from (**C1**) and (**C2**) [41]: Applying (5) to the stationary random variable X_∞ under the empirical distribution $\widehat{\mathbb{E}}^{\pi_{\mathrm{ID}}}$ and rearranging, we have

$$\beta_V \widehat{\mathbb{E}}^{\pi_{\mathrm{ID}}}\left[V(X_\infty)\right] \leq \widehat{\mathbb{E}}^{\pi_{\mathrm{ID}}}\left[\Delta(X_\infty)\right] = K_V\sqrt{N} + \widehat{\mathbb{E}}^{\pi_{\mathrm{ID}}}\left[(\widehat{P}_{\pi_{\mathrm{ID}}} - P_{\pi_{\mathrm{ID}}})V(X_\infty)\right]. \tag{6}$$

We can then combine (6) with (**C2′**) to obtain

$$\begin{aligned}\rho^{\mathrm{rel}} - \widehat{\rho}^{\pi_{\mathrm{ID}}} &\leq C_1 \widehat{\mathbb{E}}^{\pi_{\mathrm{ID}}} V(X_\infty)/N + C_2/\sqrt{N} \\ &\leq \frac{C_1}{\beta_V N}\widehat{\mathbb{E}}^{\pi_{\mathrm{ID}}}\left[(\widehat{P}_{\pi_{\mathrm{ID}}} - P_{\pi_{\mathrm{ID}}})V(X_\infty)\right] + \left(\frac{C_1 K_V}{\beta_V} + C_2\right)\frac{1}{\sqrt{N}}\end{aligned} \tag{7}$$

as desired. The intrinsic suboptimality of the ID policy, which is still present in the full-information (planning) setting, is captured by the second term of (7), while the first term reflects the statistical error which arises from transferring the drift condition for π_{ID} to the empirical system. This first term bears a close resemblance to the simulation lemma (4), but replaces $h^{\pi_{\mathrm{ID}}}$ with the Lyapunov function V. Overall, we believe this approach has broad potential for analyzing the sample complexity of planning algorithms deployed in a plug-in fashion, so long as the policy output by the planning algorithm admits a Lyapunov-drift-based performance guarantee.

Acknowledgment

T. Wu and Q. Xie are supported in part by the U.S. National Science Foundation (NSF) Grants CNS-1955997, ECCS-2339794, and ECCS-2432546. W. Wang is supported in part by the NSF Grants ECCS-2145713, CCF-2403194, CCF-2428569, and ECCS-2432545. M. Zurek and Y. Chen are supported in part by the NSF Grants CCF-2233152 and DMS-2023239, a Vilas Associate Award, and a Cisco Fellowship.

References

[1] Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal, April 2020. URL `http://arxiv.org/abs/1906.03804`. arXiv:1906.03804 [cs, math, stat] version: 3.

[2] Konstantin Avrachenkov, Vivek S Borkar, and Pratik Shah. Lagrangian index policy for restless bandits with average reward. *arXiv preprint arXiv:2412.12641*, 2024.

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

[4] Mohammad Gheshlaghi Azar, Remi Munos, and Bert Kappen. On the Sample Complexity of Reinforcement Learning with a Generative Model, June 2012. URL `http://arxiv.org/abs/1206.6461`. arXiv:1206.6461 [cs, stat].

[5] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. Learning index policies for restless bandits with application to maternal healthcare. 2021.

[6] Craig Boutilier and Tyler Lu. Budget allocation using weakly coupled, constrained markov decision processes. In *UAI*, 2016.

[7] David B Brown and James E Smith. Index policies and performance bounds for dynamic selection problems. *Management Science*, 66(7):3029–3050, 2020.

[8] David B Brown and Jingwei Zhang. Dynamic programs with shared resources and signals: Dynamic fluid policies and asymptotic optimality. *Operations Research*, 70(5):3015–3033, 2022.

[9] David B Brown and Jingwei Zhang. Fluid policies, reoptimization, and performance guarantees in dynamic resource allocation. *Operations Research*, 73(2):1029–1045, 2025.

[10] X. R. Cao. Single Sample Path-Based Optimization of Markov Chains. *Journal of Optimization Theory and Applications*, 100(3):527–548, March 1999. ISSN 1573-2878. doi: 10.1023/A:1022634422482. URL `https://doi.org/10.1023/A:1022634422482`.

[11] Ibrahim El Shar and Daniel Jiang. Weakly coupled deep q-networks. *Advances in Neural Information Processing Systems*, 36:43931–43950, 2023.

[12] Nicolas Gast, Bruno Gaujal, and Chen Yan. Reoptimization nearly solves weakly coupled markov decision processes. *arXiv preprint arXiv:2211.01961*, 2022.

[13] Kevin D Glazebrook, HM Mitchell, and PS Ansell. Index policies for the maintenance of a collection of machines by a set of repairmen. *European Journal of Operational Research*, 165 (1):267–284, 2005.

[14] Jeffrey Thomas Hawkins. *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. PhD thesis, Massachusetts Institute of Technology, 2003.

[15] Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Achieving exponential asymptotic optimality in average-reward restless bandits without global attractor assumption. *arXiv preprint arXiv:2405.17882*, 2024.

[16] Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Unichain and aperiodicity are sufficient for asymptotic optimality of average-reward restless bandits. *arXiv preprint arXiv:2402.05689*, 2024.

[17] Weici Hu and Peter Frazier. An asymptotically optimal index policy for finite-horizon restless bandits. *arXiv preprint arXiv:1707.00205*, 2017.

[18] Yujia Jin and Aaron Sidford. Towards Tight Bounds on the Sample Complexity of Average-reward MDPs, June 2021. URL `http://arxiv.org/abs/2106.07046`. arXiv:2106.07046 [cs, math].

[19] Jackson A Killian, Arpita Biswas, Sanket Shah, and Milind Tambe. Q-learning lagrange policies for multi-action restless bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 871–881, 2021.

[20] Jackson A Killian, Lily Xu, Arpita Biswas, and Milind Tambe. Restless and uncertain: Robust policies for restless bandits via deep multi-agent reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 990–1000. PMLR, 2022.

[21] Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-optimal sample complexity for mdps via anchoring. *arXiv preprint arXiv:2502.04477*, 2025.

[22] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model. In *Advances in Neural Information Processing Systems*, volume 33, pages 12861–12872. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/96ea64f3a1aa2fd00c72faacf0cb8ac9-Abstract.html`.

[23] Nicolas Meuleau, Milos Hauskrecht, Kee-Eung Kim, Leonid Peshkin, Leslie Pack Kaelbling, Thomas L Dean, and Craig Boutilier. Solving very large weakly coupled markov decision processes. *AAAI/IAAI*, 8:2, 1998.

[24] Carl D. Meyer, Jr. The Condition of a Finite Markov Chain and Perturbation Bounds for the Limiting Probabilities. *SIAM Journal on Algebraic Discrete Methods*, 1(3):273–283, September 1980. ISSN 0196-5212. doi: 10.1137/0601031. URL `https://epubs.siam.org/doi/abs/10.1137/0601031`. Publisher: Society for Industrial and Applied Mathematics.

[25] Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I Hou, Srinivas Shakkottai, et al. Neurwin: Neural whittle index network for restless bandits via deep rl. *Advances in Neural Information Processing Systems*, 34:828–839, 2021.

[26] Khaled Nakhleh, I Hou, et al. Deeptop: Deep threshold-optimal policy for mdps and rmabs. *Advances in Neural Information Processing Systems*, 35:28734–28746, 2022.

[27] Gergely Neu and Nneka Okolo. Dealing with unbounded gradients in stochastic saddle-point optimization, June 2024. URL `http://arxiv.org/abs/2402.13903`. arXiv:2402.13903 [cs, math, stat] version: 2.

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

[29] Francisco Robledo, Vivek Borkar, Urtzi Ayesta, and Konstantin Avrachenkov. Qwi: Q-learning with whittle index. *ACM SIGMETRICS Performance Evaluation Review*, 49(2):47–50, 2022.

[30] Francisco Robledo, Urtzi Ayesta, and Konstantin Avrachenkov. Deep reinforcement learning for weakly coupled mdp's with continuous actions. In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pages 67–80. Springer, 2024.

[31] Adrienne Tuynman, Rémy Degenne, and Emilie Kaufmann. Finding good policies in average-reward Markov Decision Processes without prior knowledge, May 2024. URL `http://arxiv.org/abs/2405.17108`. arXiv:2405.17108 [cs].

[32] Sofía S Villar. Indexability and optimal index policies for a class of reinitialising restless bandits. *Probability in the engineering and informational sciences*, 30(1):1–23, 2016.

[33] Jinghan Wang, Mengdi Wang, and Lin F. Yang. Near Sample-Optimal Reduction-based Policy Learning for Average Reward MDP, December 2022. URL http://arxiv.org/abs/2212.00603. arXiv:2212.00603 [cs].

[34] Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward markov decision processes, 2024. URL https://arxiv.org/abs/2310.08833.

[35] Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal Sample Complexity for Average Reward Markov Decision Processes, February 2024. URL http://arxiv.org/abs/2310.08833. arXiv:2310.08833.

[36] Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of applied probability*, 27(3):637–648, 1990.

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

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

[39] Guojun Xiong, Shufan Wang, and Jian Li. Learning infinite-horizon average-reward restless multi-action bandits via index awareness. *Advances in Neural Information Processing Systems*, 35:17911–17925, 2022.

[40] Zhe Yu, Yunjian Xu, and Lang Tong. Deadline scheduling as restless bandits. *IEEE Transactions on Automatic Control*, 63(8):2343–2358, 2018.

[41] Xiangcheng Zhang, Yige Hong, and Weina Wang. Projection-based lyapunov method for fully heterogeneous weakly-coupled mdps. *arXiv preprint arXiv:2502.06072*, 2025.

[42] Xiangyu Zhang and Peter I. Frazier. Restless Bandits with Many Arms: Beating the Central Limit Theorem, July 2021. URL http://arxiv.org/abs/2107.11911. arXiv:2107.11911 [cs, math].

[43] Xiangyu Zhang and Peter I. Frazier. Near-optimality for infinite-horizon restless bandits with many arms, March 2022. URL http://arxiv.org/abs/2203.15853. arXiv:2203.15853 [cs, math].

[44] Jiahong Zhou, Shunhui Mao, Guoliang Yang, Bo Tang, Qianlong Xie, Lebin Lin, Xingxing Wang, and Dong 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.

[45] Matthew Zurek and Yudong Chen. The Plug-in Approach for Average-Reward and Discounted MDPs: Optimal Sample Complexity Analysis, October 2024. URL http://arxiv.org/abs/2410.07616. arXiv:2410.07616 [cs].

[46] Matthew Zurek and Yudong Chen. Span-agnostic optimal sample complexity and oracle inequalities for average-reward rl. *arXiv preprint arXiv:2502.11238*, 2025.

[47] Matthew Zurek and Yudong Chen. Span-Based Optimal Sample Complexity for Weakly Communicating and General Average Reward MDPs. *Advances in Neural Information Processing Systems*, 37:33455–33504, January 2025. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3acbe9dc3a1e8d48a57b16e9aef91879-Abstract-Conference.html.

Notation

Let $\mathbb{R}$, $\mathbb{N}$, and $\mathbb{N}_+$ denote the sets of real numbers, nonnegative integers, and positive integers, respectively. For $n \leq n'$ with $n, n' \in \mathbb{N}_+$, define the sets $[n] \triangleq \{1, 2, \ldots, n\}$, $[n : n'] \triangleq \{n, n+1, \ldots, n'\}$ and $[0,1]_n = \left\{\frac{i}{n} \mid i \in \mathbb{N}, 0 \leq \frac{i}{n} \leq 1\right\}$. For vectors $u, v \in \mathbb{R}^S$, we use the inner product $\langle u, v\rangle = \sum_{s\in\mathbb{S}} u(s)v(s)$. For each cost type $k \in [K]$, let $c^\star_{k,i}(s) = \sum_{a\in\mathcal{A}} \bar{\pi}^\star_i(a \mid s)c_{k,i}(s,a)$, and let $c^\star_k = (c^\star_{k,i})_{i\in[N]}$ denote the vector of the functions $c^\star_{k,i}$'s. In addition, let $r^\star_i(s) = \sum_{a\in\mathcal{A}} \bar{\pi}^\star_i(a \mid s)r_i(s,a)$, and let $r^\star = (r^\star_i)_{i\in[N]}$ denote the vector of the functions $r^\star_i$'s. We combine these vectors into a set $\mathcal{G} = \{c^\star_1, c^\star_2, \ldots, c^\star_K, r^\star\}$. For a matrix $A \in \mathbb{R}^{m\times n}$, the induced operator norm from ℓ_∞ to ℓ_∞ is defined as $\|A\|_{\infty\to\infty} = \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{1\leq i\leq m} \sum_{j=1}^n |A_{ij}|$.

Symbol	Meaning
$\mathcal{S}, \mathcal{A}$	State and Action space
S, A	Cardinality of the state and action space, $S = \lvert\mathcal{S}\rvert, A = \lvert\mathcal{A}\rvert$
s, a	one-dimension state and action
$\boldsymbol{s}, \boldsymbol{a}$	N-dimension state and action
$p_i(\cdot \mid s, a)$	True transition probability of arm i given (s, a)
$\widehat{p}_i(\cdot \mid s, a)$	Empirical transition probability of arm i given (s, a)
$P(\cdot \mid \boldsymbol{s}, \boldsymbol{a})$	True transition probability of the N-arm MDP given (s, a)
$\widehat{P}(\cdot \mid \boldsymbol{s}, \boldsymbol{a})$	Empirical transition probability of the N-arm MDP given (s, a)
$\mu^\star$	Stationary distribution obtained by running $\bar{\pi}^\star$ under the true MDP
$\widehat{\mu}^\star$	Stationary distribution obtained by running $\widehat{\pi}^\star$ under the empirical MDP
$\widehat{\mu}^{\pi_{\mathrm{ID}}}$	Stationary distribution obtained by running π_{ID} under the empirical MDP
$\mu^{\widehat{\pi}_{\mathrm{ID}}}$	Stationary distribution obtained by running $\widehat{\pi}_{\mathrm{ID}}$ under the true MDP
P_π	One-step transition matrix induced by a policy π
P_π^∞	Limiting transition matrix of the Markov chain under policy π
$\lVert\cdot\rVert_1$	ℓ_1 norm of a vector
$\lVert\cdot\rVert_{\infty\to\infty}$	The induced operator norm from ℓ_∞ to ℓ_∞ of a matrix
$\mathbf{1}$	All-ones vector
$\mathbb{1}$	Indicator function
λ_W	Maximal eigenvalue of a matrix W
$\mathbb{E}^\pi$	Expectation under policy π and the true transition kernel P
$\widehat{\mathbb{E}}^\pi$	Expectation under policy π and the empirical transition kernel $\widehat{P}$
ρ^π	average reward of policy π under true model
$\widehat{\rho}^\pi$	average reward of policy π under empirical model

A Related Work

We review prior work by first considering *planning* and then *learning*, and within each part we contrast the *average–reward* criterion with *finite–horizon or discounted* objectives, moving from Restless Bandits (RBs) to general WCMDPs as assumptions weaken, and we conclude by summarizing recent progress on the sample complexity of *tabular* average–reward MDPs under the generative–model setting.

Under planning with average–reward, the RB literature builds on the Whittle relaxation and the Whittle index policy [37, 36]. Subsequent papers establish asymptotic or $o(1)$ optimality gaps under increasingly relaxed structural conditions, including unichain/aperiodicity and attractor assumptions [17, 16, 15], and explore typed or partially heterogeneous arms [42, 43]. Moving from RBs to general WCMDPs, scalable planning is achieved via linear relaxations, Lagrangian decomposition, and periodic re–optimization [23, 12, 30]. Very recent analyses obtain asymptotic gaps such as $O(1/\sqrt{N})$ for fully heterogeneous systems by Lyapunov–style arguments for ID–type policies [41], though these are still performance guarantees rather than learning bounds.

Turning to planning with finite–horizon or discounted objectives, RB work develops fluid and LP relaxations with performance guarantees and asymptotic optimality in various regimes [7–9, 17]. Analogous WCMDP schemes based on re–optimization are effective computationally [12, 23], yet their optimality gaps often grow super–linearly with the (effective) horizon, which limits direct transfer to average–reward analyses.

Building on these planning foundations, the literature on *learning* with average–reward increasingly leverages planning priors to structure exploration under unknown dynamics. In Restless Bandits (RBs)—a special case of WCMDPs—*index–aware* algorithms for infinite–horizon, average–reward, multi–action models achieve sublinear regret with polynomial dependence on problem size [39]; Whittle–index–guided Q–learning admits convergence in the average–reward setting [3]. Finite–time (non–asymptotic) rates have also been established for Q–Whittle methods, quantifying how estimation error translates into value loss [38]. Beyond indices, works exploit *threshold–optimal* structure to design actor–critic / policy–gradient schemes that can serve as practical surrogates for explicit index computation while retaining theoretical guarantees under structural assumptions [26]. Moving from RBs to *general* weakly coupled systems, weak coupling has been embedded directly into the learning architecture: the tabular WCQL counterpart of WCDQN enjoys *almost–sure convergence*, while the deep variant reports strong empirical gains [11].

Turning next to *finite–horizon or discounted* objectives with global budgets, *Lagrangian relaxation* combined with (deep) Q–learning yields *regret* or *asymptotic* guarantees in RB–type models [19, 20]; index–aware and Q–Whittle–type algorithms also remain effective in this regime, often with an empirical emphasis [25, 29].

Finally we discuss related work on the sample complexity of tabular average-reward MDPs. For the uniformly mixing and (more general) weakly communicating settings, [34] and [47] developed minimax-optimal algorithms, both utilizing a discounted reduction plug-in approach. These results match lower bounds due to [18] and [33]. [47] also developed algorithms and matching lower bounds for the most general setting of multichain (aka general) MDPs. These results require prior knowledge of environmental complexity parameters, a shortcoming which was addressed by [31, 27, 45, 21, 46]. In particular [45] studies a direct plug-in approach for solving average-reward MDPs, without discounted reduction. We also refer to these works and the references therein for further background on the history of this problem.

B LP relaxation of the empirical RB problem

Below we give the LP relaxation of the empirical N-armed RB system constructed from the data.

$$\max_{(y_i(s,a))_{i\in[N],\, s\in\mathcal{S},\, a\in\mathcal{A}}} \frac{1}{N}\sum_{i\in[N]}\sum_{s\in\mathcal{S},\, a\in\mathcal{A}} y_i(s,a) r_i(s,a) \tag{8a}$$

$$\text{subject to} \quad \frac{1}{N}\sum_{i\in[N]}\sum_{s\in\mathcal{S},\, a\in\mathcal{A}} y_i(s,a) c_{k,i}(s,a) \le \alpha_k, \quad \forall k\in[K], \tag{8b}$$

$$\sum_{s'\in\mathcal{S},\, a'\in\mathcal{A}} \widehat{p}_i(s \mid s',a') y_i(s',a') = \sum_{a\in\mathcal{A}} y_i(s,a), \quad \forall s\in\mathcal{S}, \forall i\in[N], \tag{8c}$$

$$\sum_{s'\in\mathcal{S},\, a'\in\mathcal{A}} y_i(s',a') = 1, \quad y_i(s,a)\ge 0, \quad \forall s\in\mathcal{S},\, \forall a\in\mathcal{A},\, \forall i\in[N]. \tag{8d}$$

C Algorithms

Algorithm 1: ID policy with reassignment

Input : N-armed WCMDP instance $(\mathcal{M}_i = (\mathcal{S}, \mathcal{A}, p_i, r_i, (c_{k,i})_{k\in[K]})_{i\in[N]}$

Preprocessing:

Solve the LP in (2) to obtain the optimal state–action frequencies $(y_i^\star(s,a))_{i\in[N],\, s\in\mathcal{S},\, a\in\mathcal{A}}$.

Compute the optimal single-armed policies $(\bar{\pi}_i^\star)_{i\in[N]}$.

Perform ID reassignment using Algorithm 2.

Real-time:

for $t = 0, 1, 2, \ldots$ **do**

  Sample ideal actions $A_{i,t}^{\text{ideal}} \sim \bar{\pi}_i^\star(\,\cdot \mid S_{i,t}\,)$ for all $i \in [N]$.

  $I \leftarrow 1$.

  while $\sum_{i\in[I]} c_{k,i}(S_{i,t},\, A_{i,t}^{\text{ideal}}) \le \alpha_k N,\ \forall\, k \in [K]$ **do**

    Arm I: take action $A_{I,t} \leftarrow A_{I,t}^{\text{ideal}}$; $I \leftarrow I + 1$.

  foreach $i \in \{I, I+1, \ldots, N\}$ **do**

    take action $A_{i,t} \leftarrow 0$.

Algorithm 2: ID reassignment

Input : optimal state–action frequencies $(y_i^\star(s,a))_{i\in[N],\, s\in\mathcal{S},\, a\in\mathcal{A}}$, budgets $(\alpha_k)_{k\in[K]}$

Output : new arm IDs recorded in $\texttt{newID}(i)$ for each $i \in [N]$

Compute $(C_{k,i}^\star)_{i\in[N],\, k\in[K]}$ and the set of active constraints $\mathcal{A}$ using (8).

if $\mathcal{A} = \varnothing$ **then**

  $\texttt{newID}(i) \leftarrow i$ for all $i \in [N]$; `// No need for ID reassignment`

else

  $\mathcal{F} \leftarrow \varnothing$; `// Arms that have been assigned new IDs`

  $\alpha_{\min} \leftarrow \min_{k\in[K]} \alpha_k$; $\delta \leftarrow \alpha_{\min}/4$; $d \leftarrow \left\lceil \dfrac{(c_{\max} - \delta)K}{\alpha_{\min}/2 - \delta} \right\rceil$.

  For each $k \in \mathcal{A}$, set $\mathcal{D}_k \leftarrow \{\, i \in [N] : C_{k,i}^\star \ge \delta \,\}$.

  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 $\sum_{i\in\mathcal{F}} C_{k,i}^\star\, \mathbf{1}\{\texttt{newID}(i) \in \mathcal{I}(\ell)\} < \delta$ **then**

        pick any $i \in \mathcal{D}_k$ and set $\texttt{newID}(i) \leftarrow j$;

        remove i from every $\mathcal{D}_{k'}$ ($k' \in \mathcal{A}$); $\mathcal{F} \leftarrow \mathcal{F} \cup \{i\}$; $j \leftarrow j + 1$.

  For all $i \in [N] \setminus \mathcal{F}$, assign $\texttt{newID}(i)$ randomly from $[N] \setminus \{\texttt{newID}(i') : i' \in \mathcal{F}\}$.