Achieving Exponential Asymptotic Optimality in Average-Reward Restless Bandits without Global Attractor Assumption

Yige Hong[1] **Qiaomin Xie**[2] **Yudong Chen**[2] **Weina Wang**[1]
[1]Carnegie Mellon University [2]University of Wisconsin-Madison
{yigeh,weinaw}@cs.cmu.edu
qiaomin.xie@wisc.edu yudongchen@cs.wisc.edu

Abstract

We study the infinite-horizon average-reward restless bandit (RB) problem, a representative class of problems within the broader framework of weakly-coupled Markov decision processes (MDPs). Each RB problem consists of N MDPs coupled by a resource constraint. Existing computationally efficient policies either only achieve an $O(1/\sqrt{N})$ optimality gap or require a strong *global attractor assumption* to achieve an exponentially small $O(\exp(-CN))$ optimality gap. In this paper, we propose a novel *two-set policy* that achieves an $O(\exp(-CN))$ optimality gap under the weaker and easily verifiable assumptions of aperiodic unichain, non-degeneracy, and local stability. We further show that dropping *any* of these three assumptions precludes an exponential optimality gap, with local stability playing a particularly fundamental role as demonstrated by our lower bound. Finally, our experimental results confirm that the two-set policy outperforms existing policies when our assumptions are met but not the global attractor assumption, while remaining competitive across general settings.

1 Motivation

The restless bandit (RB) problem [15] is a representative class of problems within the broader framework of weakly-coupled Markov decision processes (MDPs), with a wide range of applications including wireless communication [1], scheduling [2], machine maintenance [6], healthcare [10], among many others [11]. An RB problem consists of multiple small MDPs, referred to as arms, each with two possible actions: activate or keep passive. The overall action is subject to a constraint, thus coupling the dynamics of the arms. At each time step, the decision maker observes the states of the arms and decides which arms to activate, with the goal of maximizing the total rewards from all arms. An RB problem can be viewed as a *planning problem* for a structured large MDP where the model parameters are known.

The state and action spaces of an RB problem grow exponentially with the number of arms, N. In fact, it is known that finding an exact optimal policy for RBs is intractable [12]. However, extensive research has shown that one can design efficiently computable policies that are *asymptotically optimal* as N becomes large. Here we say a policy is asymptotically optimal if its optimality gap is $o(1)$ when $N \to \infty$, where the optimality gap of a policy is defined as the difference between the average reward per arm and that of an optimal policy.

The asymptotic optimality of average-reward RB in the large N regime has been studied for several decades, starting with the seminal papers on the renowned Whittle index policy [15, 14]. Since then, researchers have been weakening the assumptions for asymptotic optimality [13, 7, 8, 16] or

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

improving the order of the optimality gaps [5, 4]. A recent milestone is the *exponential asymptotic optimality* established in [5, 4], where a class of policies called LP-Priority is proved to achieve an optimality gap $O(\exp(-CN))$ for a constant C, under some assumptions. This exponential gap is remarkable because it "beats" the Central Limit Theorem (CLT), an intriguing theoretical property that is not a priori obvious to be possible. In particular, on a high-level, all asymptotic optimality results in the RB literature are based on the concentration of the empirical distribution of the arms' states around a certain optimal distribution. Therefore, an $O(1/\sqrt{N})$ bound owing to CLT was believed to be fundamental. An important observation made in [5, 4] is that if the system has a certain notion of *local linearity*, the optimality gap only depends on the distance between the optimal distribution and the *expected* empirical state distribution, allowing one to go beyond CLT.

Despite the above intuition, it remains unclear what fundamental mechanism leads to exponential asymptotic optimality. Although a simple assumption called non-degeneracy (or non-singularity) suffices to ensure the local linearity property described above, non-degeneracy alone does not guarantee exponential optimality of the LP-Priority policies considered in Gast et al. [5, 4]. In particular, another crucial assumption is needed, namely the Uniform Global Attractor Property (UGAP). UGAP is a stronger version of the global attractor property (GAP) — the latter is assumed in all asymptotic analyses of LP-Priority policies, without which an LP-Priority may even have a constant optimality gap [see 14, 3, 7, for concrete examples]. Unfortunately, UGAP is hard to interpret and verify, as it concerns the global convergence of non-linear difference equations.

In light of the difficulty caused by UGAP, recent work has studied new policies that are asymptotically optimal without assuming UGAP [7, 8, 16]. Different from LP-Priority and Whittle index policies, these new policies actively control the empirical distribution of the states of the arms, driving the distribution to *globally converge* towards a certain optimal distribution, without relying on extraneous assumptions. However, while these results conclude that UGAP is not needed for asymptotic optimality, it remains unclear whether UGAP is fundamental for exponential asymptotic optimality — the best optimality gap proved in this line of work is $O(1/\sqrt{N})$. Given these developments, it is natural to ask if removing UGAP necessarily comes at the cost of degrading optimality gap. Specifically, the goal of this paper is to answer the following question: Is it possible to efficiently find policies that achieve *exponential asymptotic optimality* without assuming UGAP?

2 Problem setup

Consider the discrete-time average-reward restless bandit problem with N homogeneous arms. We label the arms by numbers $[N] \triangleq \{1, 2, \ldots, N\}$, and refer to i as the *ID* of Arm i. Each arm is associated with a Markov decision process (MDP) defined by $(\mathbb{S}, \mathbb{A}, P, r)$, which is called the single-armed MDP. Here $\mathbb{S}$ is a finite state space; $\mathbb{A} = \{0, 1\}$ is the action space, where action 1 is activating/pulling the arm; $P : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to [0, 1]$ is the transition kernel; $r : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ is the reward function. At each time t, a policy π samples the action vector $\boldsymbol{A}_t \triangleq (A_t(i))_{i\in[N]} \in \mathbb{A}^N$ for each arm based on the current state vector $\boldsymbol{S}_t \triangleq (S_t(i))_{i\in[N]} \in \mathbb{S}^N$ and possibly some history states. The objective is to find a policy that maximizes the long-run average reward,

$$R(\pi, \boldsymbol{S}_0) \triangleq \lim_{T\to\infty} \frac{1}{NT} \sum_{t=0}^{T-1} \sum_{i=1}^{N} \mathbb{E}\left[r(S_t(i), A_t(i))\right],$$

subject to a budget constraint $\sum_{i=1}^{N} A_t(i) = \alpha N$ at all times, for some $\alpha \in (0, 1)$. Let $R^*(N, \boldsymbol{S}_0)$ denote the optimal value of this objective. A policy is called *asymptotically optimal* if its optimality gap, $R^*(N, \boldsymbol{S}_0) - R(\pi, \boldsymbol{S}_0)$, vanishes as the system size $N \to \infty$ [13, Definition 4.11]. We focus on the setting where all the model parameters, $\mathbb{S}, \mathbb{A}, P, r, \alpha$, are known.

LP relaxation and assumptions. We follow the Linear Programming (LP) relaxation framework to bound the optimality gap, which is adopted by all prior work to establish asymptotic optimality of RB policies [14, 13, 5, 4, 7]. Specifically, consider the following LP:

$$\begin{aligned} \underset{\{y(s,a)\}_{s\in\mathbb{S},a\in\mathbb{A}}}{\text{maximize}} \quad & \sum_{s\in\mathbb{S},a\in\mathbb{A}} r(s,a)y(s,a) \\ \text{subject to} \quad & \sum_{s\in\mathbb{S}} y(s,1) = \alpha, \end{aligned} \tag{LP}$$

$$\sum_{s'\in\mathbb{S},a\in\mathbb{A}} y(s',a)P(s',a,s) = \sum_{a\in\mathbb{A}} y(s,a),\ \forall s\in\mathbb{S};$$
$$\sum_{s'\in\mathbb{S},a'\in\mathbb{A}} y(s',a') = 1; \quad y(s,a)\geq 0,\ \forall s\in\mathbb{S}, a\in\mathbb{A}.$$

Letting R^{rel} be the optimal value of (LP), it can be show that $R^{\text{rel}} \geq R^*(N, \boldsymbol{S}_0)$ [8].

Given an optimal solution to (LP), $y^* = \{y^*(s,a)\}_{s\in\mathbb{S},a\in\mathbb{A}}$, we can introduce some preliminary definitions to state our assumptions and algorithm. The *optimal single-armed policy*, $\bar{\pi}^*$, is the conditional distribution given by

$$\bar{\pi}^*(a|s) = \begin{cases} \frac{y^*(s,a)}{y^*(s,0)+y^*(s,1)}, & \text{if } y^*(s,0)+y^*(s,1) > 0, \\ 1/2, & \text{if } y^*(s,0)+y^*(s,1) = 0, \end{cases}$$

for $s\in\mathbb{S}$, $a\in\mathbb{A}$. Let $P_{\bar{\pi}^*}$ be the transition matrix induced by the policy $\bar{\pi}^*$ on the single-armed MDP, i.e., $P_{\bar{\pi}^*}(s,s') = \sum_{a\in\mathbb{A}} \bar{\pi}^*(a|s)P(s,a,s')$. The *optimal stationary distribution* $\mu^* \triangleq (\mu^*(s))_{s\in\mathbb{S}}$, where $\mu^*(s) = y^*(s,1) + y^*(s,0)$, is the unique stationary distribution of $P_{\bar{\pi}^*}$ given the aperiodic unichain assumption that we will make next.

We make the following three assumptions. The first two are standard conditions common to prior work: the aperiodic unichain condition or its stronger versions are assumed to achieve asymptotic optimality [14, 13, 5, 4, 7, 8]; non-degeneracy is assumed to achieve exponential asymptotic optimality [5, 4]. Our third assumption, local stability, is strictly weaker than the UGAP assumption in the prior work [5, 4], because it only requires the local dynamics under the LP-Priority policy (captured by the matrix Φ defined below) to be exponentially stable, whereas UGAP requires the global exponential stability of a non-linear dynamic system. All our assumptions are in terms of the *single-armed* MDP primitives (P, r, α) or the LP solution y^*, which makes them easy to verify.

Assumption 2.1 (Aperiodic unichain)**.** The transition probability matrix $P_{\bar{\pi}^*}$ has a simple eigenvalue 1; the other eigenvalues all have a modulus strictly smaller than 1. In other words, $P_{\bar{\pi}^*}$ defines an aperiodic unichain on $\mathbb{S}$ (i.e., with a single-recurrent class, and possibly some transient states).

Assumption 2.2 (Non-degeneracy)**.** For the fixed optimal solution to (LP), y^*, there exists a unique *fluid neutral state*, i.e., a state $\tilde{s}\in\mathbb{S}$ s.t. $y^*(\tilde{s},1) > 0$ and $y^*(\tilde{s},0) > 0$.

Assumption 2.3 (Local stability)**.** Given the non-degenerate condition, define the matrix Φ as:

$$\Phi \triangleq P_{\bar{\pi}^*} - \mathbb{1}^\top \mu^* - (c_{\bar{\pi}^*} - \alpha\mathbb{1})^\top (P_1(\tilde{s}) - P_0(\tilde{s})), \tag{1}$$

where $c_{\bar{\pi}^*} \triangleq (\bar{\pi}^*(1|s))_{s\in\mathbb{S}}$ and $P_a(\tilde{s}) \triangleq (P(\tilde{s},a,s))_{s\in\mathbb{S}}$ are both row vectors; $\mathbb{1}$ is the all-one row vector. We assume that the spectral radius of Φ is strictly less than 1.

3 Algorithm and results

Now we informally state our main policy and results, whose details are given in our full paper [9].

Our main policy, the two-set policy (Algorithm 1), maintains two dynamic subsets of arms and applies two different subroutines to the subsets.

- The first subroutine, Unconstrained Optimal Control, allocates the actions according to the optimal single-armed policy $\bar{\pi}^*$, i.e., let $\bar{\pi}^*(a|s)$ fraction of arms in state s take action s for each $(s,a)\in\mathbb{S}\times\mathbb{A}$. If a subset of arms could persistently follow the Unconstrained Optimal Control, their empirical state distribution would evolve according to the transition matrix $P_{\bar{\pi}^*}$, and finally converge to μ^* by our aperiodic unichain assumption (Assumption 2.1).
- The second subroutine, Optimal Local Control, is a local version of the LP-Priority policy from prior work [13, 5]. This subroutine enforces a local budget constraint, activating an α fraction of arms in its assigned subset (modulo the integer effect). While this subroutine would yield an $O(\exp(-CN))$ optimality gap for some constant $C > 0$ if followed persistently, it has an intrinsic precondition: it is feasible only when the empirical state distribution of its assigned arms is sufficiently close to μ^* in a proper sense.

The two-set policy (Algorithm 1) dynamically updates the subsets controlled by each subroutine. At each time step t, the policy first greedily expands the subset D_t^{OL} where the arms follow

Algorithm 1 Two-Set Policy

Input: N, α, y^*, $\boldsymbol{S}_0$, and $D_{-1}^{\mathrm{OL}} = \emptyset$, $D_{-1}^{\bar{\pi}^*} = \emptyset$

1: **for** $t = 0, 1, \ldots$ **do**
2: **if** $\delta(\boldsymbol{S}_t, [N]) \geq 0$ **then**
3: Let $D_t^{\mathrm{OL}} = [N]$
4: **else if** $\delta(\boldsymbol{S}_t, D_{t-1}^{\mathrm{OL}}) \geq 0$ **then**
5: Let D_t^{OL} be any ϵ_N^{rd}-maximal feasible subset such that $D_t^{\mathrm{OL}} \supseteq D_{t-1}^{\mathrm{OL}}$
6: **else**
7: Let D_t^{OL} be any ϵ_N^{rd}-maximal feasible subset
8: Let $D_t^{\bar{\pi}^*}$ be a subset of $[N] \backslash D_t^{\mathrm{OL}}$ with $|D_t^{\bar{\pi}^*}| = \lfloor \beta(N - |D_t^{\mathrm{OL}}|) \rfloor$, s.t. either $D_t^{\bar{\pi}^*} \supseteq D_{t-1}^{\bar{\pi}^*} \backslash D_t^{\mathrm{OL}}$ or $D_t^{\bar{\pi}^*} \subseteq D_{t-1}^{\bar{\pi}^*} \backslash D_t^{\mathrm{OL}}$
9: Set $A_t(i)$ for $i \in D_t^{\mathrm{OL}}$ using Optimal Local Control
10: Set $A_t(i)$ for $i \in D_t^{\bar{\pi}^*}$ using Unconstrained Optimal Control
11: Set $A_t(i)$ for $i \notin D_t^{\mathrm{OL}} \cup D_t^{\bar{\pi}^*}$ such that $\sum_{i \in [N]} A_t(i) = \alpha N$
12: Apply $A_t(i)$ for $i \in [N]$ and observe $\boldsymbol{S}_{t+1}$

Optimal Local Control, subject to a certain constraint $\delta(\boldsymbol{S}_t, D_t^{\mathrm{OL}}) \geq 0$ that ensures the feasibility of Optimal Local Control. More precisely, D_t^{OL} is chosen to be an ϵ_N^{rd}*-maximal feasible set* —- an approximate notion of being maximal among the collection of subsets $\{D \subseteq [N] \colon \delta(\boldsymbol{S}_t, D) \geq 0\}$, with a certain tolerance level $\epsilon_N^{\mathrm{rd}} = O(1/N)$. We also impose some set-inclusion constraints when expanding D_t^{OL} (Lines 3–7). After choosing D_t^{OL}, the two-set policy chooses an $\min(\alpha, 1-\alpha)$ fraction of the rest of the arm to form $D_t^{\bar{\pi}^*}$ and let them follow the Unconstrained Optimal Policy. Finally, the arms in $(D_t^{\mathrm{OL}} \cup D_t^{\bar{\pi}^*})^c$ take arbitrary actions to meet the budget constraint.

Intuitively, the two-set policy yields the following dynamics: the arms in D_t^{OL} remain close to μ^* in empirical state distribution; the arms in $D_t^{\bar{\pi}^*}$ is driven towards μ^* and gradually merged into D_t^{OL}; the arms in $(D_t^{\mathrm{OL}} \cup D_t^{\bar{\pi}^*})^c$ is gradually merged into $D_t^{\bar{\pi}^*}$. In the long run, D_t^{OL} will contain all arms in the system and remain so for all but $O(\exp(-CN))$ fraction of the time; properties of Optimal Local Control then guarantees an $O(\exp(-CN))$ optimality gap. This intuition is justified in our first main theorem:

Theorem 3.1. *Suppose Assumptions 2.1, 2.2 and 2.3 hold. Let π be the two-set policy in Algorithm 1. Then for some $C > 0$ independent of N:*

$$R^{\mathrm{rel}} - R(\pi, \boldsymbol{S}_0) = O\big(\exp(-CN)\big). \tag{2}$$

Theorem 3.1 establishes the *first* exponential asymptotic optimality without UGAP. In contrast, prior work either requires UGAP to achieve exponential asymptotic optimality or only has an $O(1/\sqrt{N})$ optimality gap. We comment that when the non-degeneracy assumption (Assumption 2.2) and/or the local stability assumption (Assumption 2.3) fail, the two-set policy still achieves an $O(1/\sqrt{N})$ optimality gap. This is because the two-set policy can be viewed as a generalized version of set-expansion [8] and inherits its $O(1/\sqrt{N})$ guarantee under the aperodic unichain condition.

The proof of Theorem 3.1 employs a novel multivariate Lyapunov function, which generalizes the focus-set approach in Hong et al. [8]. This multivariate Lyapunov function enables us to decouple the complex dynamics under the two-set policy, where the states of the arms and the two dynamic subsets are coupled and change simultaneously. We refer the readers to the full paper [9] for the proof.

We argue that Assumptions 2.1, 2.2, and 2.3 in Theorem 3.1 are necessary for proving an exponentially small optimality gap using the LP relaxation framework, i.e., by bounding $R^{\mathrm{rel}} - R(\pi, \boldsymbol{S}_0)$. For the first two standard assumptions, we construct counterexamples in our full paper where violating either Assumptions 2.1 or 2.2 makes an exponentially small $R^{\mathrm{rel}} - R(\pi, \boldsymbol{S}_0)$ unachievable. For the local stability assumption (Assumption 2.3), we establish the following lower bound for a broad class of instances named *regular unstable RBs*. Our assumptions are therefore the weakest possible for proving exponential asymptotic optimality within the predominant LP relaxation framework.

Theorem 3.2. *For every regular unstable RB, every policy π and any initial state vector $\boldsymbol{S}_0$, we have*

$$R^{\mathrm{rel}} - R(\pi, \boldsymbol{S}_0) = \Omega\big(1/\sqrt{N}\big).$$

References

[1] S. Aalto, P. Lassila, and I. Taboada. Whittle index approach to opportunistic scheduling with partial channel information. *Perform. Eval.*, 136:102052, 2019.

[2] T. W. Archibald, D. P. Black, and K. D. Glazebrook. Indexability and index heuristics for a simple class of inventory routing problems. *Oper. Res.*, 57(2):314–326, 2009.

[3] N. Gast, B. Gaujal, and C. Yan. Exponential convergence rate for the asymptotic optimality of Whittle index policy. *arXiv:2012.09064 [cs.PF]*, 2020.

[4] N. Gast, B. Gaujal, and C. Yan. Exponential asymptotic optimality of Whittle index policy. *Queueing Syst.*, 104(1):107–150, 2023.

[5] 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.

[6] 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.

[7] Y. Hong, Q. Xie, Y. Chen, and W. Wang. Restless bandits with average reward: Breaking the uniform global attractor assumption. In *Advances in Neural Information Processing Systems*, volume 36, pages 12810–12844, 2023.

[8] 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.

[9] 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.

[10] A. Mate, L. Madaan, A. Taneja, N. Madhiwalla, S. Verma, G. Singh, A. Hegde, P. Varakantham, and M. Tambe. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. *AAAI Conf. Artificial Intelligence*, 36(11):12017–12025, June 2022.

[11] J. Niño-Mora. Markovian restless bandits and index policies: A review. *Mathematics*, 11(7): 1639, 2023.

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

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

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

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

[16] 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.