2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

An Optimal-Control Approach to Infinite-Horizon Restless Bandits: Achieving Asymptotic Optimality with Minimal Assumptions

Chen Yan

***Abstract*— We adopt an optimal-control framework for addressing the undiscounted infinite-horizon discrete-time restless N-armed bandit problem. Unlike most studies that rely on constructing policies based on the relaxed single-armed Markov Decision Process (MDP), we propose relaxing the entire bandit MDP as an optimal-control problem through the certainty equivalence control principle. Our main contribution is demonstrating that the reachability of an optimal stationary state within the optimal-control problem is a sufficient condition for the existence of an asymptotically optimal policy. Such a policy can be devised using an "align and steer" strategy. This reachability assumption is less stringent than any prior assumptions imposed on the arm-level MDP, notably the unichain condition is no longer needed.**

I. Introduction

The Restless Bandit (RB) problem addresses the challenge of optimally allocating limited resources across a set of dynamically evolving alternatives [21]. Each alternative, or "arm", changes state over time according to a Markov Decision Process (MDP), irrespective of whether it is currently being exploited or not, hence the term "restless". This problem encapsulates a broad range of real-world scenarios, from queue management and sensor scheduling to wireless communication and adaptive clinical trials . Despite its theoretical and practical significance, finding optimal solutions remains notoriously challenging [17], driving ongoing research into efficient heuristics and asymptotically optimal policy design [19], [4], [11], [13], [8]. This paper contributes to this vibrant field by proposing an optimal-control framework that offers fresh insights into the asymptotic optimality of policies for the RB problem.

Contributions:

- We propose a novel approach by relaxing the stochastic bandit problem into a deterministic optimal-control problem, diverging from the conventional strategy of relaxation into a single-armed problem (see Figure 1).
- We link asymptotic optimality in the bandit problem to the reachability of an optimal stationary point via feasible control, bypassing the unichain assumption for a broader applicability that includes multichain models.
- We propose the "align and steer" strategy for constructing asymptotically optimal policies, assuming reachability. Our numerical studies highlight the superiority of integrating model predictive control within this strategy.

Notations: To differentiate between the single-armed MDP and the N-armed bandit MDP, we use the letter s to denote the state of the former, which assumes a finite set of S values, and $\mathbf{x}, \mathbf{X}$ for the state of the latter, represented as a population vector within the unit simplex Δ of dimension S upon dividing by N. For the bandit-level problem, capital letters indicate stochastic systems, lowercase for deterministic, and boldface for vectors, treated as row vectors. The subset $\Delta^{(N)}$ of Δ consists of points whose coordinates are multiples of $1/N$. Vector inequality $\mathbf{x} \geq \mathbf{y}$ are defined componentwise. We use *control rule* π for deterministic optimal-control problems and *policy* π^N for stochastic N-armed bandit MDPs. Control mappings are denoted as $\pi(\mathbf{x}) = \mathbf{u}$, with $\mathbf{x}^\pi(t)$ (resp. $\mathbf{u}^\pi(t)$) representing the state (resp. control) after applying π over t steps on an initial state $\mathbf{x}(0)$.

Chen Yan was with Statify, Inria Grenoble and BioSP, INRAE Avignon. He is now in the EECS Department with University of Michigan, Ann Arbor. chenyaa@umich.edu

II. Problem Setup and Literature Review

A. Model Description

Consider the undiscounted infinite-horizon discrete-time Restless Bandit (RB) problem with N homogenous arms. Each arm itself is a Markov Decision Process (MDP) with state space $\mathcal{S} := \{1, 2, \ldots, S\}$ and action space $\mathcal{A} := \{0, 1\}$. There is a budget constraint requiring that at each time step, exactly αN arms can take action 1, with $0 < \alpha < 1$. For simplicity we assume that αN is always an integer. The state space of the N-armed bandit is therefore $\mathcal{S}^N$ and the action space is a subset of $\mathcal{A}^N$. The arms are weakly-coupled, in the sense that they are only linked through the budget constraint, i.e. for a given feasible action $\mathbf{a} \in \mathcal{A}^N$, the bandit transitions from a state $\mathbf{s} \in \mathcal{S}^N$ to state $\mathbf{s}' \in \mathcal{S}^N$ with probability $\mathbb{P}(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) = \prod_{n=1}^N \mathbb{P}(s'_n \mid s_n, a_n) = \prod_{n=1}^N P^{a_n}_{s_n, s'_n}$, where for each action $a_n = a \in \mathcal{A}$, the matrix $\mathbf{P}^a$ is a probability transition matrix of dimension $S \times S$. Upon choosing an action $\mathbf{a}$ in state $\mathbf{s}$, we receive an instant-reward $\sum_{n=1}^N r^{a_n}_{s_n}$, where $r^a_s \in \mathbb{R}$ depends on the state s and action a.

A *Markovian* policy π^N for the N-armed problem chooses at each time t a feasible action $\mathbf{a}(t)$ based solely on the current state $\mathbf{s}(t)$. It is *stationary* if in addition it does not depend on t. Our goal is to maximize the long-term average expected reward from all N arms across all stationary policies, facing an exponentially large state and action space as N increases. [1] Formally, this bandit MDP with a given initial

[1] In contrast to stochastic and adversarial bandits, where the model is not fully known and the emphasis is on minimizing regret compared to a hindsight optimal [16], the current Markovian bandit setting assumes all problem parameters and the system states are known, focusing on the design of efficient and effective algorithms.

state $\mathbf{s}(0)$ is formulated as:

$$\max_{\pi^N} \quad V_{\pi^N}(\mathbf{s}(0)) := \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{N}\mathbb{E}_{\pi^N}\left[\sum_{n=1}^{N} r_{s_n(t)}^{a_n(t)}\right] \quad (1)$$

$$\text{s.t.} \quad \mathbb{P}\left(\mathbf{s}(t+1) \mid \mathbf{s}(t), \mathbf{a}(t)\right) = \prod_{n=1}^{N} P_{s_n(t),s_n(t+1)}^{a_n(t)}, \quad (2)$$

$$\mathbf{a}(t)\cdot\mathbf{1}^\top = \alpha N, \quad \mathbf{a}(t)\in\{0,1\}^N \quad \forall t\geq 0. \quad (3)$$

The difficulty of the RB problem is that the N arms are coupled by the constraints in Equation (3), and the conventional approach begins with relaxing these constraints in Equation (3), which must be met at every time t with probability one, to a single time-averaged constraint in expectation: $\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\pi^N}\left[\mathbf{a}(t)\cdot\mathbf{1}^\top\right] = \alpha N$. This effectively decompose the N-armed problem into N independent single-armed problem, each having the following form in relation to a single-armed policy $\bar{\pi}$ and an initial arm state $s(0)$:

$$\max_{\bar{\pi}} \quad V_{\bar{\pi}}(s(0)) := \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\bar{\pi}}\left[r_{s(t)}^{a(t)}\right] \quad (4)$$

$$\text{s.t.} \quad \mathbb{P}(s(t+1)\mid s(t),a(t)) = P_{s(t),s(t+1)}^{a(t)}, \quad \forall t\geq 0$$

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\bar{\pi}}\left[a(t)\right] = \alpha.$$

This single-armed problem can be equivalently formulated using state-action frequency variables, see [18, Section 8.9.2] and Problem (8) below. We denote by $\bar{\pi}^*$ as one such optimal single-armed policy with an optimal value $V_{\bar{\pi}^*}(s(0))$. Note that the initial arm state $s(0)\in\mathcal{S}$ can be extended to a probability distribution in $\mathcal{S}$, represented by a point $\mathbf{x}(0)\in\Delta$. Under the *unichain* assumption, the optimal value $V_{\bar{\pi}^*}(\mathbf{x}(0))$ of Problem (4) is independent of the initial arm state distribution $\mathbf{x}(0)$ [18].

B. The Approach via Optimal-Control

In this paper, our approach is to approximate Problem (1) with an optimal-control problem via the Certainty Equivalence Control (CEC) principle ([3, Chapter 6]). Throughout this paper, we will not make the blanket assumption that the single-armed MDP is unichain.

1) Arm States Concatenation and the CEC Problem: Given that the N arms are homogeneous, representing the bandit state through the concatenation of arm states can significantly simplify subsequent analysis. To achieve this, denote by $\mathbf{X}\in\Delta^{(N)}$ where X_s is the fraction of arms in state $s\in\mathcal{S}$, normalized by division by N. A similar notation goes for the control $\mathbf{U}$, so that U_s is the fraction of arms in state s taking action 1 under the control $\mathbf{U}$.

Using this arm states concatenation, the Markovian evolution of the bandit state in Equation (2) can be expressed as $\mathbf{X}(t+1)\overset{d}{=}\phi(\mathbf{X}(t),\mathbf{U}(t))+\mathcal{E}(\mathbf{X}(t),\mathbf{U}(t))$, where $\phi(\cdot,\cdot)$ is the deterministic *linear* function:

$$\phi(\mathbf{X},\mathbf{U}) := (\mathbf{X}-\mathbf{U})\cdot\mathbf{P}^0 + \mathbf{U}\cdot\mathbf{P}^1, \quad (5)$$

and $\mathcal{E}(\cdot,\cdot)$ is a Markovian random vector, whose properties are summarized in the following lemma, with a proof utilizing standard probability techniques available in [11, Lemma 1]:

Lemma 1 ([11]): The random vector $\mathcal{E}(\mathbf{X}(t),\mathbf{U}(t))\overset{d}{=}\mathbf{X}(t+1)-\phi(\mathbf{X}(t),\mathbf{U}(t))$ verifies:

$$\mathbb{E}\left[\mathcal{E}(\mathbf{X},\mathbf{U})\mid\mathbf{X},\mathbf{U}\right]=\mathbf{0};$$
$$\mathbb{E}\left[\|\mathcal{E}(\mathbf{X},\mathbf{U})\|_1\mid\mathbf{X},\mathbf{U}\right]\leq\sqrt{S}/\sqrt{N};$$
$$\mathbb{P}\left(\|\mathcal{E}(\mathbf{X},\mathbf{U})\|_1\geq\xi\mid\mathbf{X},\mathbf{U}\right)\leq 2S\cdot e^{-2N\xi^2/S^2}.$$

Given a state $\mathbf{x}\in\Delta$, define the following two control sets:

$$\mathcal{U}(\mathbf{x}) := \left\{\mathbf{u}\mid\mathbf{u}\cdot\mathbf{1}^\top=\alpha \text{ and } \mathbf{0}\leq\mathbf{u}\leq\mathbf{x}\right\};$$
$$\mathcal{U}^{(N)}(\mathbf{x}) := \left\{\mathbf{u}\mid\mathbf{u}\in\mathcal{U}(\mathbf{x}) \text{ and } N\cdot\mathbf{u} \text{ is an integer vector}\right\};$$

as well as the linear instant-reward function:

$$R(\mathbf{x},\mathbf{u}) := (\mathbf{x}-\mathbf{u})\cdot(\mathbf{r}^0)^\top + \mathbf{u}\cdot(\mathbf{r}^1)^\top.$$

Note that $\mathcal{U}(\mathbf{x})$ is always non-empty, and $\mathcal{U}^{(N)}(\mathbf{x})$ is non-empty if $\mathbf{x}\in\Delta^{(N)}$. A *feasible control* π is a map from $\mathbf{x}\in\Delta$ to $\mathcal{U}(\mathbf{x})$, while a *feasible policy* π^N maps $\mathbf{x}$ into $\mathcal{U}^{(N)}(\mathbf{x})$. An equivalent formulation of Problem (1) using arm states concatenation (where $\mathbf{s}(0)$ yields $\mathbf{x}(0)$) is:

$$\max_{\pi^N} \quad V_{\pi^N}(\mathbf{x}(0)) := \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\pi^N}\left[R(\mathbf{X}(t),\mathbf{U}(t))\right] \quad (6)$$

$$\text{s.t.} \quad \mathbf{X}(t+1)\overset{d}{=}\phi(\mathbf{X}(t),\mathbf{U}(t))+\mathcal{E}(\mathbf{X}(t),\mathbf{U}(t)),$$
$$\mathbf{U}(t)\in\mathcal{U}^{(N)}(\mathbf{X}(t)) \ a.s., \quad \forall t\geq 0.$$

The two requirements below Equation (6) result from arm states concatenation of Equations (2) and (3), respectively.

We now link Problem (6) to its corresponding CEC problem, where the uncertainties $\mathcal{E}(\cdot,\cdot)$ are assumed to be identically zero. Specifically, the CEC problem is defined as a maximization task over all stationary control rules π, described as follows:

$$\max_{\pi} \quad V_\pi(\mathbf{x}(0)) := \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} R(\mathbf{x}(t),\mathbf{u}(t)) \quad (7)$$

$$\text{s.t.} \quad \mathbf{x}(t+1) = \phi\left(\mathbf{x}(t),\mathbf{u}(t)\right),$$
$$\mathbf{u}(t)\in\mathcal{U}(\mathbf{x}(t)), \quad \forall t\geq 0.$$

Bearing its resemblance to Problem (6), the above Problem (7) is now deterministic with *uncountable* state and action space. In contrast to the single-armed MDP Problem (4) that relaxes the constraints (3) into a *single* time-averaged expectation constraint, the CEC problem relaxes these constraints into expectation constraints at *every* time step, represented by $\mathbf{u}(t)\cdot\mathbf{1}^\top=\alpha$ for all $t\geq 0$. It is a *linear* control problem where the set of feasible controls depends on the state. Note that there are two distinct paths that naturally lead to consider Problem (7): the first involves taking the large N limit in Problem (6) and referring to Lemma 1 as we previously discussed; the second entails taking the large T limit from the finite-horizon RB relaxation to a linear program, as considered in various works [4], [11].

As Problem (6) is generally intractable [17], our strategy employs a stationary control rule π that optimally solves the more tractable Problem (7). From this, we construct an induced policy π^N that matches as much as possible to π. This is made precise in the following definition:

***Definition** 1:* (Induced Policy π^N from Control Rule π) For a feasible stationary control rule π of Problem (7), a corresponding induced policy π^N for Problem (6) is defined as any stationary policy such that for an input state $\mathbf{X} \in \Delta^{(N)}$ with $\overline{\mathbf{U}} = \pi(\mathbf{X})$, it outputs a control $\pi^N(\mathbf{X}) = \mathbf{U} \in \mathcal{U}^{(N)}(\mathbf{X})$ satisfying $\left\|\mathbf{U} - \overline{\mathbf{U}}\right\|_1 \leq S/N$.

The general observation from Lemma 1 is that if π is optimal, then the induced policy π^N as in Definition 1 should also be close to optimal for large values of N. This will be precisely formulated in Theorem 1 below.

2) Stationary Problems: By definition, a *stationary point* $(\mathbf{x}_e, \mathbf{u}_e)$ of Problem (7) is one such that $\mathbf{u}_e \in \mathcal{U}(\mathbf{x}_e)$ and $\mathbf{x}_e = \phi(\mathbf{x}_e, \mathbf{u}_e)$. A stationary point $(\mathbf{x}^*, \mathbf{u}^*)$ is optimal if it solves the corresponding static problem with (7). This is what we refer to as the *conventional* static problem. In this paper, however, in order to also take into account multichain models, we shall consider the refined static problem with optimal value denoted as $V_e^*(\mathbf{x}(0))$, following [2], [15]:

$$\max_{\mathbf{x},\mathbf{u},\mathbf{h}^0,\mathbf{h}^1} \quad V_e(\mathbf{x}(0)) := R(\mathbf{x},\mathbf{u}) \tag{8}$$

$$\begin{aligned} \text{s.t.} \quad & \mathbf{x} = \phi(\mathbf{x},\mathbf{u}), \\ & \mathbf{u} \in \mathcal{U}(\mathbf{x}), \\ & \mathbf{x} + \mathbf{h}^0 + \mathbf{h}^1 - \mathbf{h}^0 \cdot \mathbf{P}^0 - \mathbf{h}^1 \cdot \mathbf{P}^1 = \mathbf{x}(0), \\ & \mathbf{x} \geq \mathbf{0},\ \mathbf{h}^0 \geq \mathbf{0},\ \mathbf{h}^1 \geq \mathbf{0}. \end{aligned} \tag{9}$$

We recover the conventional static problem if in the above optimization problem Equation (9) is replaced by $\mathbf{x} \cdot \mathbf{1}^\top = 1$ and there are no $\mathbf{h}^0, \mathbf{h}^1$ variables. Problem (8) is a refinement since multiply Equation (9) by $\mathbf{1}^\top$ on the right gives the relation $\mathbf{x} \cdot \mathbf{1}^\top = 1$. The additional variables $\mathbf{h}^0, \mathbf{h}^1$ appearing in Problem (8) can be interpreted as a deviation measure [2]. In fact, by [15, Theorem 10], if the single-armed MDP is unichain, then for any initial condition $\mathbf{x}(0)$ Problem (8) is equivalent to the conventional static problem, so the two formulations make no difference. However, we will illustrate in Section III-E the necessity of considering the refined static problem for the more general multichain models.

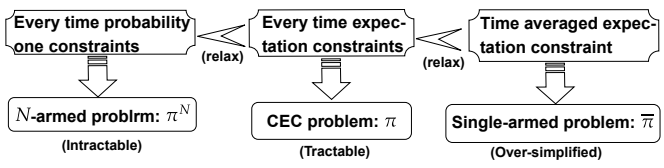

Fig. 1: Relationship of the three optimization problems

3) Value Comparison and Asymptotic Optimality: The links between the three major problems considered in this work are summarized in Figure 1. The relationships among the values of the various optimization problems defined up to this point are as follows:

$$\begin{aligned} & V_{\pi^N}(\mathbf{s}(0)) = V_{\pi^N}(\mathbf{x}(0)) \\ & \leq V_\pi(\mathbf{x}(0)) \leq V_{\bar{\pi}^*}(\mathbf{x}(0)) = V_e^*(\mathbf{x}(0)), \end{aligned} \tag{10}$$

where the first relationship arises from the concatenation of arm states; the second is established in the technical report [22, Section VI], whose proof is based on induction on the horizon; the third results from Problem (4) being a more relaxed formulation than Problem (7); and the final relationship can be deduced, for instance, from [15], by treating $\mathbf{x} - \mathbf{u}$ and $\mathbf{u}$ as state-action frequency variables. Equation (10) leads us to propose the following definition:

***Definition** 2:* (Optimal Stationary Point and Asymptotic Optimality) For a given initial state $\mathbf{x}_{\text{init}}$, we call $(\mathbf{x}^*, \mathbf{u}^*)$ that solves the refined static Problem (8) with $\mathbf{x}(0) = \mathbf{x}_{\text{init}}$ an *optimal stationary point*, and $\mathbf{x}^*$ an optimal stationary state. We call a control rule π of Problem (7) *averaged-reward optimal* if $V_\pi(\mathbf{x}(0)) = V_e^*(\mathbf{x}(0))$; the corresponding stationary policy π^N of Problem (6) in Definition 1 is said to be *asymptotically optimal* for the N-armed RB problem, if it verifies $\lim_{N\to\infty} V_{\pi^N}(\mathbf{x}(0)) = V_e^*(\mathbf{x}(0))$.

C. Comparison with Related Works

Existing literature on infinite-horizon RB problems typically constructs policies π^N based on the single-armed problem in Figure 1. We have demonstrated that the latter essentially conveys the same information as an optimal stationary point within this control problem framework, thus revealing that achieving asymptotic optimality necessitates additional model assumptions.

Firstly, traditional methods assume the unichain condition for the single-armed MDP. Additionally, specific policies like the Whittle index policy [20] and the fluid priority policy [19] necessitate their induced dynamical systems conforming to the Global Attractor Property. Furthermore, to affirm the exponential convergence rate, [11] introduces a more stringent Uniform Global Attractor Property (UGAP). These assumptions on global dynamical system behavior are theoretically challenging to verify. In response, [13] proposes a more easily verifiable Synchronization Assumption for the optimal single-armed policy $\bar{\pi}^*$, achieving $\mathcal{O}(\sqrt{N})$ asymptotic optimality with the Follow-the-Virtual-Advice (FTVA) policy. Extending this approach, [14] further simplifies the criteria, showing that the unichain and aperiodic assumptions on $\bar{\pi}^*$ are sufficient for their ID / Focus Set policies. Lemma 2 will demonstrate that this sufficient condition implies our reachability condition in Definition 4. On the other hand, without the unichain assumption, a stationary optimal single-arm policy may not even exist ([15]; see also Section III-E). Therefore, new ideas are required to construct asymptotically optimal policies. This paper aims to address this challenge.

Roadmap: We overcome this issue by focusing on the dynamical yet still tractable CEC problem. Theorem 1 demonstrates that for the induced policy π^N to achieve asymptotic optimality, the corresponding control rule π must be average-reward optimal, and additionally, a bias-related term (Equation (11)) needs to be a continuous function.

Theorem 2 demonstrates that this can be achieved through the "align and steer" strategy (Algorithm 1) with a linear control for steering, provided that a mild reachability condition, as detailed in Definition 4, is met by the model.

A comprehensive comparison of various existing policies on the undiscounted infinite-horizon discrete-time RB problem is summarized in the table below.

Policies	Assumptions	Rate
Whittle / LP-Priority [11]	UGAP &Regular& Unichain	$e^{-\Omega(N)}$
FTVA [13]	Synchronization & Unichain	$\mathcal{O}(\sqrt{N})$
ID / Focus-Set [14]	Unichain	$\mathcal{O}(\sqrt{N})$
Align and Steer [this work]	Communicating	$\mathcal{O}(1)$

III. Reachability and Asymptotic Optimality

Throughout this section, we fix an initial condition $\mathbf{x}_{\text{init}} \in \Delta$ and a corresponding optimal stationary point $(\mathbf{x}^*, \mathbf{u}^*)$ of Problem (8). Denote by $\mathcal{S}^+ := \{s \in \mathcal{S} \mid x_s^* > 0\}$. It is important to remember that, in the case of multichain models, all quantities we deduce are dependent on the initial state. We will make this dependence explicit whenever possible.

A. The Effective Control Rules

In order to formulate the effectiveness of a control beyond average-reward optimality, we define, for a control rule π, the following possibly unbounded functions for all $\mathbf{x} = \mathbf{x}(0)$:

$$G^\pi(\mathbf{x}) := \sum_{t=0}^{\infty} (\mathbf{x}^\pi(t), \mathbf{u}^\pi(t)) - (\mathbf{x}^*, \mathbf{u}^*), \tag{11}$$

where we recall that $(\mathbf{x}^*, \mathbf{u}^*)$ is an optimal stationary point by solving Problem (8) with initial state $\mathbf{x}$. We define a stationary control rule π for Problem (7) as *effective* if $G^\pi(\mathbf{x})$ is a continuous function over $\mathbf{x} \in \Delta$ under the $\mathcal{L}^1$-norm. This implies that $G^\pi(\cdot)$ is bounded given the compactness of Δ. The concept and justification for an effective control rule are encapsulated in the subsequent theorem:

Theorem 1: (Effective Control Rule Leads to Asymptotically Optimal Policy) Fix an initial state $\mathbf{x}_{\text{init}}$. For a stationary control rule π of Problem (7) with an optimal stationary point $(\mathbf{x}^*, \mathbf{u}^*)$ and optimal value $V_e^*(\mathbf{x}(0))$ defined in Problem (8) with $\mathbf{x}(0) = \mathbf{x}_{\text{init}}$, consider the function $G^\pi(\mathbf{x})$ defined in Equation (11). If $G^\pi(\mathbf{x})$ is a continuous function over $\mathbf{x} \in \Delta$ (under the $\mathcal{L}^1$-norm). Then the induced stationary policy π^N for Problem (6) in Definition 1 is asymptotically optimal: $\lim_{N\to\infty} V_{\pi^N}(\mathbf{x}(0)) = V_e^*(\mathbf{x}(0))$.

The proof of Theorem 1 employs the standard Stein's method ([12]) and is detailed in the technical report [22, Section VII]. Informally, for a control rule π to be averaged-reward optimal, the state-control pairs $(\mathbf{x}^\pi(t), \mathbf{u}^\pi(t))$ need to converge to $(\mathbf{x}^*, \mathbf{u}^*)$ independently of the initial state. To establish and study refined notions of optimality beyond the average-reward criterion, particularly for comparison with the stochastic N-armed problem, we must consider a function of the type $G^\pi(\mathbf{x})$. We refer to [6] for an in-depth discussion of various optimality criteria in this context.

It is important to note that with further regularity of the function $G^\pi(\mathbf{x})$ in the vicinity of $\mathbf{x}^*$, i.e. Lipschitz-continuity or $\mathcal{C}^1$-smoothness, the convergence rate of $V_{\pi^N}(\mathbf{x}(0))$ towards $V_e^*(\mathbf{x}(0))$ can be determined. Similar ideas have been explored in prior research, including [12], [10]. The primary challenge is determining whether such an effective control rule can be established and the methodology for its construction, which we aim to explore subsequently.

B. The Align and Steer Policy

Our idea of constructing an effective control rule can be summarized as "align and steer", which is based on the following observation from the linearity nature of Problem (7): If we decompose $\mathbf{x} \in \Delta$ into a sum of two parts $\mathbf{x} = \mathbf{v}_{\text{align}} + \mathbf{v}_{\text{steer}}$ with $\mathbf{v}_{\text{align}}, \mathbf{v}_{\text{steer}} \geq \mathbf{0}$, then the normalized vectors $\mathbf{x}_{\text{align}} := \mathbf{v}_{\text{align}} / \|\mathbf{v}_{\text{align}}\|_1$ and $\mathbf{x}_{\text{steer}} := \mathbf{v}_{\text{steer}} / \|\mathbf{v}_{\text{steer}}\|_1$ again belong to the simplex Δ. Now take $\mathbf{u}_{\text{align}} \in \mathcal{U}(\mathbf{x}_{\text{align}})$ and $\mathbf{u}_{\text{steer}} \in \mathcal{U}(\mathbf{x}_{\text{steer}})$ as feasible controls for $\mathbf{x}_{\text{align}}$ and $\mathbf{x}_{\text{steer}}$ respectively. The linear combination

$$\|\mathbf{v}_{\text{align}}\|_1 \cdot \mathbf{u}_{\text{align}} + \|\mathbf{v}_{\text{steer}}\|_1 \cdot \mathbf{u}_{\text{steer}}$$

turns out to be a feasible control for $\mathbf{x}$. The key is to split $\mathbf{x}$ so that $\mathbf{v}_{\text{align}}$ is *collinear* with $\mathbf{x}^*$ and possesses the *maximum* possible $\mathcal{L}^1$-norm. This enables the choice of $\mathbf{u}_{\text{align}}$ as $\mathbf{u}^*$ with a maximum alignment (refer to Equation (13) below).

Definition 3: (Maximum Alignment Coefficient with $\mathbf{x}^*$) For $\mathbf{x} \in \Delta$, we call the real constant

$$\delta(\mathbf{x}) := \max\{\delta \geq 0 \mid \mathbf{x} \geq \delta \cdot \mathbf{x}^*\} \tag{12}$$

the *maximum alignment coefficient* of $\mathbf{x}$ with the target $\mathbf{x}^*$.

From this definition, it follows that $0 \leq \delta(\mathbf{x}) \leq 1$, with $\delta(\mathbf{x}) = 0$ if and only if there exists an arm state $s \in \mathcal{S}^+ = \{s \in \mathcal{S} \mid x_s^* > 0\}$ with $x_s = 0$; and $\delta(\mathbf{x}) = 1$ occurs if and only if $\mathbf{x} = \mathbf{x}^*$. For any $\mathbf{x} \in \Delta$, it can be expressed as $\mathbf{x} = \delta(\mathbf{x}) \cdot \mathbf{x}^* + \mathbf{x} - \delta(\mathbf{x}) \cdot \mathbf{x}^*$. Here, the component $\mathbf{v}_{\text{align}} = \delta(\mathbf{x}) \cdot \mathbf{x}^*$ represents the mass in $\mathbf{x}$ already in alignment with stationarity, whereas $\mathbf{v}_{\text{steer}} = \mathbf{x} - \delta(\mathbf{x}) \cdot \mathbf{x}^*$ requires the application of a specific control, π_{steer}, designed to steer the remaining mass into $\mathcal{S}^+$ for subsequent alignment. This concept is further explored in the following subsection.

Now assume that a certain feasible π_{steer} has been specified. The corresponding *align and steer* control rule $\pi_{\text{align\&steer}} : \mathbf{x} \mapsto \mathcal{U}(\mathbf{x})$ is defined as:

$$\pi_{\text{align\&steer}}(\mathbf{x}) := \\ \delta(\mathbf{x}) \cdot \mathbf{u}^* + (1 - \delta(\mathbf{x})) \cdot \pi_{\text{steer}}\left(\frac{\mathbf{x} - \delta(\mathbf{x}) \cdot \mathbf{x}^*}{1 - \delta(\mathbf{x})}\right). \tag{13}$$

We emphasize that $\delta(\mathbf{x})$ plays a crucial role in ensuring that $(\mathbf{x} - \delta(\mathbf{x}) \cdot \mathbf{x}^*)/(1 - \delta(\mathbf{x}))$ is a well-defined state vector in the simplex Δ. Algorithm 1 describes the induced align and steer policy $\pi^N_{\text{align\&steer}} : \mathbf{x} \mapsto \mathcal{U}^{(N)}(\mathbf{x})$ in detail.

An advantage of the align and steer approach is the considerable flexibility in selecting the appropriate steering control, π_{steer}, in Algorithm 1. Throughout the rest of this section, we focus on the linear steering control $\pi_\ell(\mathbf{x}) = \alpha \cdot \mathbf{x}$. Our objective is to introduce a mild reachability assumption, under which $\pi^N_{\text{align\&}\ell}$ is theoretically established to be asymptotically optimal. Subsequently, in the next Section IV, we adopt Model Predictive Control (MPC) as the steering control and conduct numerical studies on $\pi^N_{\text{align\&MPC}}$.

Algorithm 1: The align and steer policy $\pi^N_{\text{align\&steer}}$.

Input: A feasible steering control π_{steer} of Problem (7); An initial state $\mathbf{X}_{\text{init}}$ for the N-armed bandit Problem (6).

Solve the static Problem (8) with initial state $\mathbf{x}(0) = \mathbf{X}_{\text{init}}$ to obtain an optimal stationary state $\mathbf{x}^*$;

Set $\mathbf{X}_{\text{current}} := \mathbf{X}_{\text{init}}$;

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

  Set $\overline{\mathbf{U}}_{\text{current}} := \pi_{\text{align\&steer}}(\mathbf{X}_{\text{current}})$ from Equation (13) ;

  Compute a control $\mathbf{U}_{\text{current}}$ with inputs $\mathbf{X}_{\text{current}}$ and $\overline{\mathbf{U}}_{\text{current}}$ as outlined in Definition 1 ;

  Apply $\mathbf{U}_{\text{current}}$ on $\mathbf{X}_{\text{current}}$ and advance to the next state $\mathbf{X}_{\text{next}}$, then set $\mathbf{X}_{\text{current}} := \mathbf{X}_{\text{next}}$;

end

C. Reachability and a Linear Control Rule

To introduce the key concept of reachability, we derive two observations from the construction in Equation (13). First, for any given $\mathbf{x} \in \Delta$ with $\delta_0 := \delta(\mathbf{x})$ and any $t \geq 0$, the maximum alignment coefficient of $\mathbf{x}^{\pi_{\text{align\&steer}}}(t)$, after applying $\pi_{\text{align\&steer}}$ on $\mathbf{x}$ for t steps, is at least as large as δ_0. Second, defining $\hat{\mathbf{x}} := (\mathbf{x} - \delta_0 \cdot \mathbf{x}^*)/(1 - \delta_0)$, then for a certain time $T_0 \geq 1$, the value of $\delta\left(\mathbf{x}^{\pi_{\text{align\&steer}}}(T_0)\right)$ remains equal to δ_0 if and only if π_{steer} fails to steer mass from $\hat{\mathbf{x}}$ into $\mathcal{S}^+$ for alignment in the preceding $T_0 - 1$ steps. Equivalently, the maximum alignment coefficient of $\hat{\mathbf{x}}^{\pi_{\text{steer}}}(t)$ is 0 for *all* $1 \leq t \leq T_0 - 1$ (note that $\delta(\hat{\mathbf{x}})$ is 0 by construction).

***Definition** 4:* (Reachability of Optimal Stationary State $\mathbf{x}^*$) Fix an initial state $\mathbf{x}_{\text{init}}$. An optimal stationary state $\mathbf{x}^*$ of the refined static Problem (8) with $\mathbf{x}(0) = \mathbf{x}_{\text{init}}$ is called *reachable*, if there exists a feasible and stationary control rule π_{steer}, a positive constant $\theta > 0$ and a finite time $T_0 \geq 1$ such that the maximum alignment coefficient in Definition 3 satisfies $\delta\left(\mathbf{x}^{\pi_{\text{steer}}}(T_0)\right) \geq \theta$, with $\mathbf{x}^{\pi_{\text{steer}}}(0) = \mathbf{x}$ for *all* $\mathbf{x} \in \Delta$. Otherwise we call $\mathbf{x}^*$ *unreachable*.

From this definition, if $\mathbf{x}^*$ is unreachable, then for any feasible control π_{steer} and for any $T_0 \geq 1$, there always exists a counterexample $\mathbf{x} \in \Delta$ along with an arm state $s \in \mathcal{S}^+$, such that the s-th coordinate of $\mathbf{x}^{\pi_{\text{steer}}}(T_0)$ is 0. This situation can arise due to the non-communicating nature and periodicity issues within the single-armed MDP. By definition, a MDP with state space $\mathcal{S}$ is called *weakly communicating* if $\mathcal{S}$ can be partitioned into a closed set $\mathcal{S}^c$ of states in which each state is accessible under some deterministic stationary policy from any other state in the set, plus a possibly empty set of states that are transient under every policy. An arm state $s \in \mathcal{S}$ is *aperiodic* under a single-armed policy $\bar{\pi}$ if $\gcd\{n \in \mathbb{N} \mid (\mathbf{P}^{\bar{\pi}})^n_{ss} > 0\} = 1$, with $\mathbf{P}^{\bar{\pi}}$ the transition matrix of the single-armed Markov chain induced by policy $\bar{\pi}$.

We now consider the *linear* control defined by $\pi_\ell(\mathbf{x}) := \alpha \cdot \mathbf{x}$. This is a feasible control according to the definition of $\mathcal{U}(\mathbf{x})$. For $t \geq 0$, by plugging into Equation (5) this control rule we obtain that $\mathbf{x}^{\pi_\ell}(t) = \mathbf{x} \cdot (\mathbf{P}^\alpha)^t$, where $\mathbf{P}^\alpha := \alpha \cdot \mathbf{P}^1 + (1 - \alpha) \cdot \mathbf{P}^0$. Note that $\mathbf{P}^\alpha$ is also the transition matrix of the single-armed Markov chain induced by policy $\bar{\pi}_\ell$ "always take action 1 with probability α". We argue that a certain communicating and aperiodic condition is sufficient for reachability:

***Lemma** 2:* (Weakly-Communicating and Aperiodic Single-Armed MDP Implies Reachability) Fix an initial state $\mathbf{x}_{\text{init}}$. Let $\mathbf{x}^*$ be an optimal stationary state of the refined static Problem (8) with $\mathbf{x}(0) = \mathbf{x}_{\text{init}}$, and denote by $\mathcal{S}^+ = \{s \in \mathcal{S} \mid x^*_s > 0\}$. If the single-armed MDP in Problem (4) is weakly communicating and the set of arm states $\mathcal{S}^+$ are aperiodic under the single-armed policy $\bar{\pi}_\ell$ "always take action 1 with probability α", then $\mathbf{x}^*$ is reachable.

Proof: Take $\pi_{\text{steer}} = \pi_\ell$ in Definition 4. Since $\mathcal{S}^+ \subset \mathcal{S}^c$, combine with the aperiodicity assumption, there exists $T_0 \geq 1$ such that $\min_{s \in \mathcal{S}, s' \in \mathcal{S}^+} (\mathbf{P}^\alpha)^{T_0}_{ss'} := p_0 > 0$. Consequently for all $\mathbf{x} \in \Delta$, it holds true that

$$\delta\left(\mathbf{x}^{\pi_\ell}(T_0)\right) = \delta\left(\mathbf{x} \cdot (\mathbf{P}^\alpha)^{T_0}\right) \geq \frac{p_0}{\max_{s \in S} x^*_s} := \theta > 0. \tag{14}$$

■

As a clarification, we point out that the condition of "being aperiodic under $\bar{\pi}_\ell$" is less stringent than "being aperiodic under an optimal single-armed policy $\bar{\pi}^*$" assumed in [14]. This is because periodicity is a pure graph-theoretic question, and $\bar{\pi}_\ell$ leads to the maximum number of directed edges in the connectivity graph among all single-armed policies. In addition, we highlight that by refining the notion of reachability in Definition 4, we can also accommodate non-communicating cases within our framework. The reason is that any MDP can be partitioned uniquely into communicating classes plus a (possible empty) class of states which are transient under any policy. Hence the approach in this paper can be directly generalized to the multichain case when initially non arm is in a transient state. This adaptation requires only those $\mathbf{x}$ that possess *the same* positive mass in communicating classes common with the initial state $\mathbf{x}_{\text{init}}$ to be in accordance with the reachability assumption within each class.

D. Reachability Implies Asymptotic Optimality

We are now ready to state the main result of this paper, which demonstrates that $\pi^N_{\text{align\&}\ell}$ is asymptotically optimal under the reachability assumption. The key idea is to compare the control rule $\pi_{\text{align\&}\ell}$ that maximize the alignment whenever possible with another rule $\pi_{\text{delay\&}\ell}$ that *delay* the alignment. We emphasize that, while the function $G^{\pi_{\text{delay\&}\ell}}(\mathbf{x})$ for the latter is easier to handle, the delayed alignment control rule is not stationary and depends on the entire history of the deterministic state trajectory. As such it cannot be used to construct a stationary policy for the N-armed bandit.

***Theorem** 2:* (Reachability of $\mathbf{x}^*$ Implies Asymptotic Optimality) Fix an initial state $\mathbf{x}_{\text{init}}$. Suppose Problem (8) with $\mathbf{x}(0) = \mathbf{x}_{\text{init}}$ possesses an optimal stationary state $\mathbf{x}^*$ that is reachable as defined in Definition 4. Let $\pi_{\text{align\&}\ell}$ represent the align and steer control rule from Equation (13), with the linear steering control rule $\pi_\ell(\mathbf{x}) = \alpha \cdot \mathbf{x}$. In this case $\pi_{\text{align\&}\ell}$

is effective. Thus, in conjunction with Theorem 1, the policy $\pi^N_{\text{align\&}\ell}$ from Algorithm 1 is asymptotically optimal.

Proof: For each finite horizon $T \geq 1$, define

$$G^{\pi}(\mathbf{x}, T) := \sum_{t=0}^{T-1} (\mathbf{x}^{\pi}(t), \mathbf{u}^{\pi}(t)) - (\mathbf{x}^*, \mathbf{u}^*).$$

The function $\mathbf{x} \mapsto G^{\pi}(\mathbf{x}, T)$ is a continuous function, provided that the control rule π is a continuous map, which holds for $\pi_{\text{align\&}\ell}$. Our strategy for proving the continuity of $G^{\pi_{\text{align\&}\ell}}(\mathbf{x})$ is to show that the sequence of continuous functions $G^{\pi_{\text{align\&}\ell}}(\mathbf{x}, T)$ indexed by T converges *uniformly* to $G^{\pi_{\text{align\&}\ell}}(\mathbf{x})$ over all $\mathbf{x} \in \Delta$.

a) Step One: To ease the notation, let us fix $\mathbf{x} \in \Delta$ and write $\delta_0 := \delta(\mathbf{x})$. Consider another feasible control rule $\pi_{\text{delay\&}\ell}$ that delay the alignment in the following sense: We align a fixed amount δ_0 of its mass with $\mathbf{x}^*$ for the first T_0 time steps, where T_0 is defined as in Lemma 2, and constantly apply π_ℓ on the steering part. Formally, the control rule is

$$\pi_{\text{delay\&}\ell}(\mathbf{x}') := \delta_0 \cdot \mathbf{u}^* + (1 - \delta_0) \cdot \pi_\ell \left(\frac{\mathbf{x}' - \delta_0 \cdot \mathbf{x}^*}{1 - \delta_0} \right), \quad (15)$$

for $\mathbf{x}' = \mathbf{x}^{\pi_{\text{delay\&}\ell}}(0) = \mathbf{x}$, and subsequently for $\mathbf{x}' = \mathbf{x}^{\pi_{\text{delay\&}\ell}}(1), \ldots, \mathbf{x}^{\pi_{\text{delay\&}\ell}}(T_0 - 1)$. Denote by $\mathbf{x}^{(1)} := (\mathbf{x} - \delta_0 \cdot \mathbf{x}^*)/(1 - \delta_0)$. Then the system trajectory under the control $\pi_{\text{delay\&}\ell}$ in Equation (15) is

$$\mathbf{x}^{\pi_{\text{delay\&}\ell}}(t) = \delta_0 \cdot \mathbf{x}^* + (1 - \delta_0) \cdot \mathbf{x}^{(1)} \cdot (\mathbf{P}^{\alpha})^t, \ 0 \leq t \leq T_0.$$

As a consequence of Lemma 2, there exists $\mathbf{x}^{(2)} \in \Delta$ such that $\mathbf{x}^{(1)} \cdot (\mathbf{P}^{\alpha})^{T_0} = \theta \cdot \mathbf{x}^* + (1 - \theta) \cdot \mathbf{x}^{(2)}$, with $\theta > 0$ defined in Equation (14). Hence

$$\begin{aligned} &\mathbf{x}^{\pi_{\text{delay\&}\ell}}(T_0) \\ &= (1 - (1 - \delta_0) \cdot (1 - \theta)) \cdot \mathbf{x}^* + (1 - \delta_0) \cdot (1 - \theta) \cdot \mathbf{x}^{(2)}. \end{aligned}$$

Now for time steps $T_0 + 1, T_0 + 2, \ldots, 2T_0$ we repeat the same procedure, except that now we start with an alignment of mass $1 - (1 - \delta_0) \cdot (1 - \theta)$ with $\mathbf{x}^*$. Note that

$$\delta_0 < 1 - (1 - \delta_0) \cdot (1 - \theta) \leq \delta\big(\mathbf{x}^{\pi_{\text{delay\&}\ell}}(T_0)\big).$$

With the same reasoning we deduce that there exists $\mathbf{x}^{(3)} \in \Delta$ such that

$$\begin{aligned} &\mathbf{x}^{\pi_{\text{delay\&}\ell}}(2T_0) \\ &= \big(1 - (1 - \delta_0) \cdot (1 - \theta)^2\big) \cdot \mathbf{x}^* + (1 - \delta_0) \cdot (1 - \theta)^2 \cdot \mathbf{x}^{(3)}. \end{aligned}$$

By a straightforward induction, we infer that for all integer $k \geq 1$, there exists $\mathbf{x}^{(k+1)} \in \Delta$ such that

$$\begin{aligned} &\mathbf{x}^{\pi_{\text{delay\&}\ell}}(k \cdot T_0) \\ &= \big(1 - (1 - \delta_0) \cdot (1 - \theta)^k\big) \cdot \mathbf{x}^* + (1 - \delta_0) \cdot (1 - \theta)^k \cdot \mathbf{x}^{(k+1)}, \end{aligned}$$

and consequently

$$\delta\big(\mathbf{x}^{\pi_{\text{delay\&}\ell}}(t)\big) \geq 1 - (1 - \delta_0) \cdot (1 - \theta)^k \quad (16)$$

for $k \cdot T_0 \leq t < (k + 1) \cdot T_0$.

b) Step Two: Our next observation is that for $t \geq 0$ it holds true that

$$\delta\left(\mathbf{x}^{\pi_{\text{delay\&}\ell}}(t)\right) \leq \delta\left(\mathbf{x}^{\pi_{\text{align\&}\ell}}(t)\right). \quad (17)$$

Indeed, delaying the alignment of a certain mass with $\mathbf{x}^*$ invariably leads to a reduced maximum alignment coefficient in the future, compared to aligning this mass with $\mathbf{x}^*$ at an earlier time. To illustrate, consider $0 < \delta_1 < \delta_0$ and express $\mathbf{x}$ as follows:

$$\mathbf{x} = \delta_1 \cdot \mathbf{x}^* + (\delta_0 - \delta_1) \cdot \mathbf{x}^* + (1 - \delta_0) \cdot \mathbf{x}^{(1)}.$$

Should the mass amounting to $\delta_0 - \delta_1$ remain unaligned, it will be subjected to the linear control rule π_ℓ, along with $\mathbf{x}^{(1)}$. A portion of this mass may eventually become aligned at a later time. However, irrespective of the specific decisions made concerning $(\delta_0 - \delta_1) \cdot \mathbf{x}^*$, what happens on $(1 - \delta_0) \cdot \mathbf{x}^{(1)}$ remains unchanged. Consequently, at any future point, the maximum alignment coefficient achieved by aligning δ_0 at an initial step is always larger than that of aligning δ_1 at the same juncture.

c) Step Three: To conclude the proof, let $\varepsilon > 0$ be fixed. We define a finite horizon $T(\varepsilon) = k(\varepsilon) \cdot T_0$, where the value of $k(\varepsilon) \in \mathbb{N}$ will be chosen later. Then

$$\begin{aligned} &\left\| G^{\pi_{\text{align\&}\ell}}(\mathbf{x}) - G^{\pi_{\text{align\&}\ell}}(\mathbf{x}, T(\varepsilon)) \right\|_1 \\ &\leq \sum_{t=T(\varepsilon)}^{\infty} (1 - \delta(\mathbf{x}^{\pi_{\text{align\&}\ell}}(t))) \\ &\qquad \cdot \left(\left\| \widetilde{\mathbf{x}}(t) - \mathbf{x}^* \right\|_1 + \left\| \pi_\ell\left(\widetilde{\mathbf{x}}(t)\right) - \mathbf{u}^* \right\|_1 \right) \\ &\qquad \left(\text{We abbreviate } \frac{\mathbf{x}^{\pi_{\text{align\&}\ell}}(t) - \delta(\mathbf{x}^{\pi_{\text{align\&}\ell}}(t)) \cdot \mathbf{x}^*}{1 - \delta(\mathbf{x}^{\pi_{\text{align\&}\ell}}(t))} \text{ as } \widetilde{\mathbf{x}}(t). \right) \\ &\leq 2(1 + \alpha) \cdot \sum_{t=T(\varepsilon)}^{\infty} (1 - \delta(\mathbf{x}^{\pi_{\text{delay\&}\ell}}(t))) \ \text{(By Equation (17))} \\ &\leq 2(1 + \alpha)(1 - \delta_0) T_0 \cdot \sum_{k=k(\varepsilon)}^{\infty} (1 - \theta)^k \ \text{(By Equation (16))} \\ &= 2(1 + \alpha)(1 - \delta_0)(1 - \theta)^{k(\varepsilon)} \cdot \frac{T_0}{\theta}. \end{aligned}$$

It suffices to choose $k(\varepsilon) := \left\lceil \log_{1-\theta} \frac{\varepsilon \cdot \theta}{2(1+\alpha)(1-\delta_0) \cdot T_0} \right\rceil$ to deduce that $\left\| G^{\pi_{\text{align\&}\ell}}(\mathbf{x}) - G^{\pi_{\text{align\&}\ell}}(\mathbf{x}, T(\varepsilon)) \right\|_1 \leq \varepsilon$ for all $\mathbf{x} \in \Delta$. Since the choice of $\varepsilon > 0$ is arbitrary, we conclude by the uniform convergence theorem that $G^{\pi_{\text{align\&}\ell}}(\mathbf{x})$ is a continuous function defined over all $\mathbf{x} \in \Delta$. ■

E. Discussion on the Reachability Assumption

We first remark that it is crucial to utilize the refined static Problem (8) for computing the optimal stationary point in multichain models. For instance, consider $\mathbf{P}^0 = \mathbf{P}^1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\mathbf{r}^0 = (1, 0)$, $\mathbf{r}^1 = (0, 1)$, and with any $0 < \alpha < 1$. This model is non-communicating. If we solve the conventional static problem described after Problem (8), then the optimal stationary state is $\mathbf{x}^* = (\alpha, 1 - \alpha)$ and is unreachable unless $\mathbf{x}_{\text{init}} = (\alpha, 1 - \alpha)$; while if we take the initial state $\mathbf{x}_{\text{init}}$ into account and solve the refined static

Problem (8) with $\mathbf{x}(0) = \mathbf{x}_{\text{init}}$, the optimal stationary state is $\mathbf{x}^* = \mathbf{x}_{\text{init}}$ itself and becomes reachable.

We now compare previous methods based on the single-armed MDP with our approach in a multichain model. Set $\mathbf{P}^0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.9 & 0 & 0.1 & 0 \\ 0 & 0 & 0.9 & 0.1 \\ 0 & 0 & 0.1 & 0.9 \end{pmatrix}$, $\mathbf{P}^1 = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.1 & 0.9 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.1 & 0 & 0.9 & 0 \end{pmatrix}$, $\mathbf{r}^0 = (0, 0, 1, 1)$, $\mathbf{r}^1 = (1, 1, 0, 0)$ and $\alpha = 0.5$. The initial state is $\mathbf{x}_{\text{init}} = (0.4, 0, 0.6, 0)$, with the corresponding optimal stationary point $\mathbf{x}^* = (0.25, 0.25, 0.25, 0.25), \mathbf{u}^* = (0.25, 0.25, 0, 0)$. It turns out that a *stationary* optimal single-armed policy does not exist for this problem, due to the state-action frequency constraint. As highlighted in [15], exploring the broader class of Markovian policies is necessary to find an optimal solution. Consequently, approaches based on stationary optimal single-armed policies, as seen in [13], [14], are inadequate. Similarly, for the fluid priority policy [19], any priority "permutation of (1,2) > permutation of (3,4)" falls within this category. However, prioritizing "4 > 3" for this particular initial state $\mathbf{x}_{\text{init}} = (0.4, 0, 0.6, 0)$ is essential for asymptotic optimality; without it, the 0.6 portion of arms in states $\{3, 4\}$ cannot transition to states $\{1, 2\}$. Policies based on the single-armed MDP typically lack the capability to make such critical distinctions. In contrast, since the model is communicating and aperiodic, by Lemma 2, we deduce that $\mathbf{x}^*$ is reachable and consequently by Theorem 2, the policy $\pi^N_{\text{align\&}\ell}$ is asymptotically optimal.

IV. Numerical Experiments

It can be seen that the simple linear control π_ℓ that has played a key role in Theorem 2 may not be the best candidate for the task of steering. Ideally the steering control π_{steer} should be designed to take any vector $\mathbf{x}$ towards $\mathbf{x}^*$ in the most reward-efficient manner. Motivated by the Model Predictive Control (MPC), a such control can be constructed by solving a finite look-ahead window T_w version of Problem (7), which is a linear program ([11]), followed by adopting the first control from this solution, see e.g. [7, Section 3]. We refer to this policy using MPC steering strategy as $\pi^N_{\text{align\&MPC}}$. In this section, we set a look-ahead window of $T_w = 100$, noting that the MPC appears to stabilize at 50 steps ahead on the examples encountered [7]. Simulations are conducted over a horizon of $T = 10000$. As highlighted in Section III-E, previous methods generally fail with multichain models. Hence our focus in this section lies on unichain models and the performance differences for finite N.

We first consider an example with three states that has been studied in [10], [13]. A noticeable feature of this example is that the Whittle index policy, which is actually the best-performed priority policy among all possible priorities, is not asymptotically optimal, as can be inferred from Figure 2. We also plot $\delta(\mathbf{X}(t))$ over a sample run for the three policies with $N = 1000$ once each 5 time steps for the first 200 time steps. The oscillation of π_{priority} presented here is caused by the fact that its dynamics is attracted to a period-2 cycle. The ID policy is introduced in [14]. Both the ID policy and the align and steer policy proposed in the current paper are asymptotically optimal, but the later performs better. We believe that this is because the closed-loop MPC is constantly driving the steering part to align with $\mathbf{x}^*$ in the most reward-efficient way. We note that it is also possible to plot the alignment of the control variable with $\mathbf{u}^*$, which exhibits similar behavior compared to the alignment of the state variable.

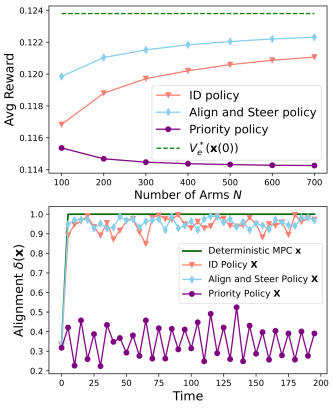

Fig. 2: An example where no priority policy is asymptotically optimal.

We next consider an example with eight states that has been proposed in [13]. A noteworthy feature of this example is that there are a total number of 36 priority policies that are asymptotically optimal: "any permutation of (2,3,4) > 1 > 5 > any permutation of (6,7,8)" are all observed to be asymptotically optimal policies. The performance of these priority policies are actually slightly superior than $\pi^N_{\text{align\&MPC}}$, as can be seen from Figure 3. We visualize that the steering of both policies align with the deterministic MPC, while the steering of the ID policy is different.

From the numerical analysis presented in this section, it is evident that $\pi^N_{\text{align\&MPC}}$ consistently delivers outstanding performance. However, given its computational efficiency and simplicity, a priority policy such as the LP-index policy from [11] should be considered at first hand. Only in instances where the global attractor property does not seem to be fulfilled by these priority policies should we consider resorting to the ID policy or $\pi^N_{\text{align\&MPC}}$. The former necessitates sampling of arm actions and their rectification at every decision epoch, which requires at least $\mathcal{O}(N)$ time; the latter entails solving a linear program being the transient problem of (7) with a horizon of T_w at each time step.

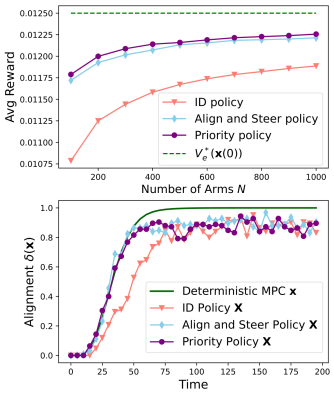

Fig. 3: An example where certain priority policies perform slightly better than $\pi^N_{\text{align\&MPC}}$.

V. Extension and Conclusion

A. On the Generalization to Weakly-Coupled MDPs

The weakly-coupled MDP is a substantial and natural extension of the restless multi-armed bandit, characterized by each (homogeneous) arm having multiple actions (i.e., $|\mathcal{A}| > 2$) and the imposition of multiple budget constraints

on the bandit. To ensure problem feasibility, it is assumed that there exists an action 0 that does not consume any resources, as for the RB. This topic has been explored in a series of studies, including [1], [9], [5], within both finite-horizon and discounted infinite-horizon frameworks, but not within the undiscounted setting addressed in the current paper. A notable aspect of the optimal-control approach adopted in this paper is its straightforward applicability to weakly-coupled MDPs with minimal additional effort required, thus filling an important gap in existing research.

Indeed, the CEC problem for weakly-coupled MDPs can be generally expressed as

$$\begin{aligned} \max_{\pi} \quad & V_\pi(\mathbf{x}(0)) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} R(\mathbf{x}(t), \mathbf{u}(t)) \qquad (18) \\ \text{s.t.} \quad & \mathbf{x}(t+1) = \phi(\mathbf{x}(t), \mathbf{u}(t)), \\ & f(\mathbf{x}(t), \mathbf{u}(t)) = \mathbf{0}, \quad g(\mathbf{x}(t), \mathbf{u}(t)) \le \mathbf{0}, \quad \forall t \ge 0. \end{aligned}$$

Here, the control variable $\mathbf{u}$ is a vector of size $|\mathcal{S}| \times |\mathcal{A}|$; $R(\mathbf{x}, \mathbf{u})$ represents a general linear function denoting the instant-reward; $\phi(\mathbf{x}, \mathbf{u})$ is a linear function describing the expected Markovian transition as in Equation (5); and $f(\mathbf{x}, \mathbf{u})$, $g(\mathbf{x}, \mathbf{u})$ are linear functions related to budget and problem structure constraints. Leveraging results from [2], [15], which focus on undiscounted average-reward multichain MDPs with linear state-action frequency constraints, we can deduce as in the RB case the relationships in Equation (10) for the various optimization problems. Consequently, the approach outlined in this paper can be applied to this more general context as well. The crucial aspect that facilitates this extension is the *linearity* of the CEC optimal-control Problem (18).

B. Conclusion

In this work, we have introduced an optimal-control framework for the undiscounted infinite-horizon N-armed RB problem, focusing on relaxing hard constraints to expected trajectory constraints at each time step, unlike traditional methods that average these constraints over time. This approach, balancing complexity between overly-simplified single-armed MDPs and intractable N-armed RB problems, allows us to derive asymptotically optimal policies by steering the system towards an optimal stationary state within a deterministic framework. Future research directions include:

1) The Lipschitz-Continuity of $G^\pi(\mathbf{x})$: Under the generality considered in this work, the possibility of ensuring Lipschitz-continuity for $G^\pi(\mathbf{x})$, which implies $\mathcal{O}(\sqrt{N})$ convergence rates of the induced policy π^N [10], remains open. The applicability of Lyapunov-function-based proof techniques from single-armed MDPs to multichain scenarios would significantly advance our understanding.

2) The Exponential Turnpike Property and Choice of Lookahead Window: Investigating the exponential turnpike property's role ([7]) in determining the finite lookahead window T_w for MPC controls could greatly impact the efficiency of applying our framework in practice. This exploration could also yield crucial insights into the dynamics of RB optimal-control problems.

ACKNOWLEDGEMENT

The author would like to thank the three anonymous reviewers and the editors for providing thorough feedback and improvements to the paper. This research is supported in part by the NSF under grant number 2207548.

REFERENCES

[1] Daniel Adelman and Adam J. Mersereau. Relaxations of weakly coupled stochastic dynamic programs. *Oper. Res.*, 56(3):712–727, may 2008.

[2] Eitan Altman and Flos Spieksma. The linear program approach in multi-chain markov decision processes revisited. *Zeitschrift für Operations Research*, 42:169–188, 1995.

[3] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific, 2012.

[4] David B. Brown and James E. Smith. Index policies and performance bounds for dynamic selection problems. *Manag. Sci.*, 66:3029–3050, 2020.

[5] David B Brown and Jingwei Zhang. Fluid policies, reoptimization, and performance guarantees in dynamic resource allocation. *Operations Research*, 2023.

[6] Dean A Carlson, Alain B Haurie, and Arie Leizarowitz. *Infinite horizon optimal control: deterministic and stochastic systems*. Springer Science & Business Media, 2012.

[7] Tobias Damm, Lars Grune, Marleen Stieler, and Karl Worthmann. An exponential turnpike theorem for dissipative discrete time optimal control problems. *SIAM Journal on Control and Optimization*, 52(3):1935–1957, 2014.

[8] Jing Fu, Bill Moran, and Peter G Taylor. A restless bandit model for resource allocation, competition, and reservation. *Operations Research*, 70(1):416–431, 2022.

[9] Nicolas Gast, Bruno Gaujal, and Chen Yan. The lp-update policy for weakly coupled markov decision processes. *arXiv preprint arXiv:2211.01961*, 2022.

[10] Nicolas Gast, Bruno Gaujal, and Chen Yan. Exponential asymptotic optimality of whittle index policy. *Queueing Systems*, pages 1–44, 2023.

[11] Nicolas Gast, Bruno Gaujal, and Chen Yan. Linear program-based policies for restless bandits: Necessary and sufficient conditions for (exponentially fast) asymptotic optimality. *Mathematics of Operations Research*, 2023.

[12] Nicolas Gast and Benny Van Houdt. A Refined Mean Field Approximation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* , 1(28), 2017.

[13] Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Restless bandits with average reward: Breaking the uniform global attractor assumption. In *Advances in Neural Information Processing Systems*, 2023.

[14] 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.

[15] Arie Hordijk and Lodewijk CM Kallenberg. Constrained undiscounted stochastic dynamic programming. *Mathematics of Operations Research*, 9(2):276–289, 1984.

[16] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[17] Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queuing network control. *Math. Oper. Res*, pages 293–305, 1999.

[18] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[19] Maaike Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *Annals of Applied Probability*, 26(4):1947–1995, 2016.

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

[21] P. Whittle. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, 25A:287–298, 1988.

[22] Chen YAN. An optimal-control approach to infinite-horizon restless bandits: Achieving asymptotic optimality with minimal assumptions. *arXiv preprint arXiv:2403.11913*, 2024.