LP-based policies for restless bandits: necessary and sufficient conditions for (exponentially fast) asymptotic optimality

Nicolas Gast, Bruno Gaujal, Chen Yan

LP-based policies for restless bandits: necessary and sufficient conditions for (exponentially fast) asymptotic optimality

Nicolas GAST
Univ. Grenoble Alpes, Inria, nicolas.gast@inria.fr

Bruno GAUJAL
Univ. Grenoble Alpes, Inria, bruno.gaujal@inria.fr

Chen YAN
Univ. Grenoble Alpes, Inria, chen.yan@inria.fr

We provide a framework to analyse control policies for the restless Markovian bandit model, under both finite and infinite time horizon. We show that when the population of arms goes to infinity, the value of the optimal control policy converges to the solution of a linear program (LP). We provide necessary and sufficient conditions for a generic control policy to be: i) asymptotically optimal; ii) asymptotically optimal with square root convergence rate; iii) asymptotically optimal with exponential rate. We then construct the LP-index policy that is asymptotically optimal with square root convergence rate on all models, and with exponential rate if the model is non-degenerate in finite horizon, and satisfies a uniform global attractor property in infinite horizon. We next define the LP-update policy, which is essentially a repeated LP-index policy that solves a new linear program at each decision epoch. We provide numerical experiments to compare the efficiency of LP-based policies. We compare the performance of the LP-index policy and the LP-update policy with other heuristics. Our result demonstrates that the LP-update policy outperforms the LP-index policy in general, and can have a significant advantage when the transition matrices are wrongly estimated.

Key words: restless bandits, linear programming, Markov decision process

1. Introduction

In this paper we investigate the famous Markovian restless bandit problem (termed RB for short) over a finite and an infinite horizon. In this problem, a decision maker faces a bandit with N arms, where each arm can be seen as a Markov decision process with two actions: active and passive. At each decision epoch, the decision maker chooses which αN of these N arms to activate, with the goal of maximizing the expected total reward over a finite (or infinite) time-horizon. All transition kernels and state-dependent rewards are assumed to be known. The arms produce rewards and evolve independently, but are coupled through the budget constraint on the number of arms that can be activated at each decision epoch. The word "restless" refers to the transition kernel under the passive action being not necessarily the identity matrix, hence generalizes the classical rested bandit model.

This problem arises in various domains and has numerous applications (see Zhang and Frazier [18] and the references therein for examples). Solving the problem exactly has been shown to be PSPACE-hard in Papadimitriou and Tsitsiklis [11]. Consequently, there has been substantial interest in developing approximate solutions whose performance are provably close to optimal, and at the same time require computations that do not grow exponentially with the number of arms N. We shall focus on the asymptotic regime where the arm population N grows and the activation budget at each epoch, αN, is proportional to N. This regime was first studied in Whittle [15] and has been of longstanding theoretical and practical interest.

Literature review

The pioneering work on this problem appears in Whittle [16], who proposed the famous Whittle index policy on infinite horizon problems, and conjectured that the policy is asymptotically optimal, meaning that the optimality gap (the difference between the performance of the optimal policy and of the Whittle index policy) converges to 0 when N goes to infinity. This conjecture has been proven to be true in Weber and Weiss [14], under an additional uniform global attractor property (termed UGAP for short); but is false in general, as shown by the 4 state counter-example provided in the same paper Weber and Weiss [14]. A later work in Gast et al. [5] actually showed that the optimality gap converges to 0 exponentially fast with N in almost all cases, which provides a theoretical explanation to the empirical good performance of the Whittle index policy.

One potential drawback of the Whittle index policy is that it requires the technical condition of indexability on the RB. Many works have been devoted to computing the indices or testing indexability, e.g. Niño-Mora [9, 10], Gast et al. [4], which makes Whittle index policy easily computable for indexable problems. Yet, we can not apply this policy if the RB is non-indexable. To circumvent this weakness, another approach, based on solving linear programs, is proposed in Verloop [13], where a set of LP-priority policies is defined from the solution of a linear program, and is shown to be all asymptotically optimal (assuming again the UGAP), regardless of indexability. The Whittle index policy is inside this set of LP-priority policies, if the RB is indexable. Later we show in our paper that the asymptotic optimality proven in Verloop [13] occurs at exponential rate under a mild condition.

Studying the problem under infinite horizon is theoretically interesting, but all these asymptotic optimality results mentioned previously rely on the UGAP, which in most cases can only be verified numerically, and may very well not be satisfied on certain problems Gast et al. [5]. This motivates another research direction that considers the corresponding finite horizon model using the linear program approach. To the best of our knowledge, this idea first appears in Hu and Frazier [6], that applies time-dependent Lagrange multipliers to define a LP-based index policy, and shows subsequently that it is asymptotically optimal (i.e. achieving an $o(1)$ optimality gap). Note that for finite-horizon problem, the asymptotically optimal policies are no-longer priority policies. Later in Zayas-Cabán et al. [17] the problem is generalized to multi-actions (instead of the two actions active and passive), and the policy proposed therein achieves an $\mathcal{O}(\log N/\sqrt{N})$ optimality gap. In Brown and Smith [3] the same problem setting as in Hu and Frazier [6] is studied, and their policies are shown to achieve $\mathcal{O}(\frac{1}{\sqrt{N}})$ optimality gap. However, as suggested by the authors of Brown and Smith [3], the convergence appears to be faster than $\mathcal{O}(\frac{1}{\sqrt{N}})$ on certain problems. Indeed, later in Zhang and Frazier [18] the authors proposed a class of fluid-priority policies that incorporate the policies in the two previous works Brown and Smith [3] and Hu and Frazier [6], and show that they achieve $\mathcal{O}(\frac{1}{\sqrt{N}})$ optimality gap in general, and can be improved to $\mathcal{O}(\frac{1}{N})$ if a *non-degeneracy* condition holds on the RB. By refining the policies, we later show in our paper that this $\mathcal{O}(\frac{1}{N})$ rate can actually be further improved to be $e^{-\mathcal{O}(N)}$.

Summary of contributions

In this paper, we provide a generic framework to study the relationship between restless bandit problem and the LP relaxations introduced in Hu and Frazier [6] for the finite horizon and in Verloop [13] for the infinite horizon. In the aforementioned papers, it is shown that the value of the stochastic control problem with N arms converges to the solution of this LP as N goes to infinity. We go further and make the following contributions:

i) The first contribution is to provide a new general framework to study the asymptotic performance of any continuous control policies for finite horizon RB. In this framework, any admissible policy is a *deterministic* map from arms distribution vectors to decision vectors, which is independent to the arm population N. This dependence is only restored later by applying a randomized

rounding technique, discussed in Section 2.3. The advantage of this approach is that it allows us to analyse the asymptotic optimality together with the convergence rate of any policy, by simply investigating properties of these deterministic maps. More precisely, we show that
   a) A *continuous* policy is asymptotically optimal if and only if it is LP-*compatible* (defined in Section 3.2).
   b) If in addition the policy is *Lipschitz continuous*, then the asymptotic optimality occurs at rate $\mathcal{O}(\frac{1}{\sqrt{N}})$.
   c) If in addition the policy is *locally linear* around the LP solution, then the asymptotic optimality occurs at rate $e^{-\mathcal{O}(N)}$.

   These properties show that the asymptotic performance of a control policy is intimately linked with the LP relaxation.

ii) We use the above characterization to provide sufficient conditions for the existence of LP-compatible policies, and to provide an effective construction of such policies. In particular:
   a) For any finite horizon RB, there always exists a LP-compatible Lipschitz-continuous policy.
   b) We show that to ensure the local linearity around the optimal LP solution as in (c), it is necessary and sufficient for the RB to be *non-degenerate*, a condition already introduced in Zhang and Frazier [18] and defined in Section 4.1. Moreover, we show a degenerate example for which no policy converges to the LP solution exponentially fast.

   We also show that the non-degeneracy property is almost equivalent to a property that we call *rankability*, and that implies the existence of an asymptotically optimal priority policy.

iii) The above results show that there exist many policies that are asymptotically optimal. Yet, for a finite number of arms N, not all will perform equality good. To provides the best policy for small N, we study two improvements: (1) the LP-index policy introduced in Hu and Frazier [6], and (2) we introduce a novel LP-based policy that we call LP-update. The later is a completely different approach from all policies considered in the existing literature and consists in frequently updating the control policy by solving a new LP. We show the $\mathcal{O}(\frac{1}{\sqrt{N}})$ rate of asymptotic optimality on this policy. We demonstrate its advantage to previous LP-based policies, both theoretically and practically.

iv) Our last contribution is to analyse the convergence rate of LP-based policies for RB in the *infinite horizon* case. Under the UGAP, we prove that the policy introduced in Verloop [13] has an exponential convergence rate if the RB is non-degenerate. Our proof uses similar techniques as in Gast et al. [5].

Note that the new approach we proposed here first defines a deterministic map using the linear program solution, and the policy is constructed from this deterministic map. The asymptotic optimality, as well as the convergence rate are then transformed into studying properties of this map. The advantage of our approach is that it allows us to capture essential properties of the RB that are necessary and sufficient conditions for any policy being asymptotically optimal with certain rate (square root or exponential).

Outline

The rest of the paper is organized as follows: Section 2 defines the finite horizon RB model as well as the admissible policy. Section 3 introduces a hierarchy of admissible policies, and prove asymptotic optimality (with convergence rate if possible) inside each of the hierarchy. Section 4 provides concrete constructions for the polices discussed in Section 3, and gives necessary and sufficient conditions for exponential convergence rate. Section 5 describes the LP-update policy. Section 6 deals with the infinite horizon case. Section 7 provides numerical studies and finally Section 8 concludes our work.

2. Model description

This paper is mainly focused on discrete time *finite horizon restless bandit* (RB) models. The infinite horizon RB models will be considered in Section 6. We first describe the model in Section 2.1. We

introduce the LP relaxation in Section 2.2. We define the admissible policy and the randomized rounding procedure in Section 2.3, and we list our notational convention in Section 2.4.

2.1. Finite horizon RB

A finite horizon RB model is composed of N statistically identical arms. Each arm can be considered as a Markov decision process (MDP) with a finite state space $\mathcal{S}=\{1\dots d\}$. The state of the nth arm at the *discrete* time $t\geq 0$ is denoted by $S_n(t)\in\{1\dots d\}$. The state of all the arms at time t is denoted by $\mathbf{S}(t)=\big(S_1(t),\dots,S_N(t)\big)$. At each time t, a decision maker observes $\mathbf{S}(t)$ and chooses a fraction $0<\alpha<1$ of the N arms to be activated. In the literature, some researchers study the problem under the non-binding constraint that *at most* a fraction α of arms can be activated at each time (e.g. Brown and Smith [3], Verloop [13]). By adding αN dummy arms that never change states and give zero rewards, we transform the non-binding setting into the binding setting since, for a given set of active arms, activating additional dummy arms does not modify the behavior of the system. Conversely, if we replace the active rewards R_s^1 by R_s^1+R' with a large enough overall positive constant R', we retrieve the non-binding setting from the binding one.

Note that in our model we do not assume αN to be an integer. If it is not, then a coin is tossed at the beginning of each decision epoch and the decision maker has to activate $\lfloor\alpha N\rfloor+1$ arms with probability $\{\alpha N\}=\alpha N-\lfloor\alpha N\rfloor$, and $\lfloor\alpha N\rfloor$ arms with probability $1-\{\alpha N\}$. We denote the action vector at time t by $\mathbf{A}(t)=\big(A_1(t),\dots,A_N(t)\big)$. For each arm that is in state s and whose action is a, the decision maker earns an immediate reward $R_s^a\in\mathbb{R}$.

Given $S_n(t)=s$ and $A_n(t)=a$, the arm n makes a Markovian transition to a state s' with probability $P^a_{s,s'}$. Those transitions are independent among all arms: for given states $\mathbf{s},\mathbf{s}'$ and activation vector $\mathbf{a}$, one has:

$$\mathbb{P}\left(\mathbf{S}(t+1)=\mathbf{s}'\mid\mathbf{S}(t),\mathbf{A}(t),\dots,\mathbf{S}(0),\mathbf{A}(0)\right)=\mathbb{P}\left(\mathbf{S}(t+1)=\mathbf{s}'\mid\mathbf{S}(t)=\mathbf{s},\mathbf{A}(t)=\mathbf{a}\right)=\prod_{n=1}^{N}P^{a_n}_{s_n,s'_n}. \quad (1)$$

By construction, the arms are exchangeable: two arms in the same state and for which the same action is chosen provide the same reward and have the same transition probabilities. This implies that the problem can be expressed by counting the number of arms in each state and the number of arms activated in each state. For a given state s, we denote by $M_s^{(N)}(t)$ the *fraction* of arms in state s at time t, and by $Y_{s,a}^{(N)}(t)$ the *fraction* of arms in state s at time t for which decision $a\in\{0,1\}$ is taken. We denote the corresponding vectors as $\mathbf{M}^{(N)}(t)\in\Delta^d$ and $\mathbf{Y}^{(N)}(t):=\big(Y_{s,1}^{(N)}(t),Y_{s,0}^{(N)}(t)\big)_{s\in\{1\dots d\}}\in\Delta^{2d}$, where Δ^d (and Δ^{2d}) are the d-dimensional (and $2d$-dimensional) simplex of probability vectors.

We denote by $V_{\text{opt}}^{(N)}(\mathbf{m}(0),T)$ the maximal expected gain (per arm) that can be obtained by the decision maker:

$$\begin{aligned}
V_{\text{opt}}^{(N)}(\mathbf{m}(0),T)=\max_{\mathbf{Y}\geq\mathbf{0}}\quad & \mathbb{E}\Big[\sum_{t=0}^{T-1}\sum_{s,a}R_s^aY_{s,a}^{(N)}(t)\Big] && \text{(2a)}\\
\text{s.t.}\quad & \text{Arms follow the Markovian evolution (1)}, && \text{(2b)}\\
& Y_{s,0}^{(N)}(t)+Y_{s,1}^{(N)}(t)=M_s^{(N)}(t)\quad\forall t,s, && \text{(2c)}\\
& \sum_s Y_{s,1}^{(N)}(t)=\begin{cases}(\lfloor\alpha N\rfloor+1)/N, & \text{with probability } \{\alpha N\}\\ \lfloor\alpha N\rfloor/N, & \text{otherwise.}\end{cases}\quad\forall t, && \text{(2d)}\\
& M_s^{(N)}(0)=m_s(0)\quad\forall s, && \text{(2e)}
\end{aligned}$$

where $\mathbf{m}(0)\in\Delta^d$ is the empirical measure of initial state vector: $m_s(0)=\frac{1}{N}\sum_{n=1}^N\mathbf{1}_{\{s_n(0)=s\}}$ for all $s\in\{1\dots d\}$. Note that (2d) represent the constraints that αN of the N arms must be activated at each time, and (2e) correspond to the initial condition.

2.2. LP relaxation

The key difficulty in the above optimization problem (2) is the constraint (2d) that couples the evolution of all arms. The idea is to replace it by the relaxed constraint requiring that the *expected* proportion of activated arms is α for all time steps t:

$$\sum_s \mathbb{E}_\pi\left[Y_{s,1}^{(N)}(t)\right] = \alpha, \quad \forall t. \tag{3}$$

The key property that makes this relaxed problem simpler is that it can then be rewritten entirely by using only the variables $y_{s,a}(t) := \mathbb{E}\left[Y_{s,a}^{(N)}(t)\right]$. To see that, we will show later in Lemma 1 that the Markovian evolution (7) implies that

$$\mathbb{E}\left[M_s^{(N)}(t+1) \mid \mathbf{Y}^{(N)}(t) = \mathbf{y}\right] = \sum_{s',a} y_{s',a} P_{s',s}^a.$$

This implies that (2b) and (2c) can be replaced by (4b) in the optimization problem below. The rest of the costs and constraints then depend only on the expected number of arms in each state. We can therefore write the relaxed optimization problem as a linear problem with value $V_{\mathrm{rel}}(\mathbf{m}(0), T)$:

$$\begin{aligned}
V_{\mathrm{rel}}(\mathbf{m}(0),T) = \max_{\mathbf{y}\geq \mathbf{0}} \quad & \sum_{t=0}^{T-1}\sum_{s,a} R_s^a y_{s,a}(t) && && \text{(4a)}\\
\text{s.t.} \quad & y_{s,0}(t+1) + y_{s,1}(t+1) = \sum_{s',a} y_{s',a}(t) P_{s's}^a && \forall s,t, && \text{(4b)}\\
& \sum_s y_{s,1}(t) = \alpha && \forall t, && \text{(4c)}\\
& y_{s,0}(0) + y_{s,1}(0) = m_s(0) && \forall s. && \text{(4d)}
\end{aligned}$$

In the above optimization problem, the constraints (4c) are the relaxation of the constraints (2d). They impose that the expected fraction of activated arms is α at all time. The constraints (4b) correspond to the expected behavior of the Markovian evolution of the system. Similarly, (4d) correspond to the initial condition (2e).

Note that the optimization problem (4) does not depend on the arm population N. Moreover, as it is a relaxation of (2), it should be clear that $V_{\mathrm{opt}}^{(N)}(\mathbf{m}(0),T) \leq V_{\mathrm{rel}}(\mathbf{m}(0),T)$. Since finding an optimal policy for $V_{\mathrm{opt}}^{(N)}(\mathbf{m}(0),T)$ is impractical, our strategy is to obtain information from optimal solutions to the linear program (4) to construct policies whose values converge quickly to $V_{\mathrm{rel}}(\mathbf{m}(0),T)$ as N goes to infinity. As $V_{\mathrm{opt}}^{(N)}(\mathbf{m}(0),T) \leq V_{\mathrm{rel}}(\mathbf{m}(0),T)$, this will imply that they become asymptotically optimal as N goes to infinity.

2.3. Admissible policies and randomized rounding

A policy determines which arms are made active at each decision epoch. In what follows, we focus on Markovian policies: such a policy is a sequence of decision rules $\pi = (\pi_0 \dots \pi_{T-1})$ such that the decision rule $\pi_t : \Delta^d \to \Delta^{2d}$ specifies the fraction of arms for each action: if $\mathbf{y} = \pi_t(\mathbf{m})$, then when the empirical state vector at time t is $\mathbf{m}$, a fraction $y_{s,a}$ among the m_s arms in state s take action a. We say that a policy is *admissible* if for all times t, all states $\mathbf{m} \in \Delta^d$ and $\mathbf{y} = \pi_t(\mathbf{m})$, we have

$$y_{s,a} \geq 0, \quad \sum_s y_{s,1} = \alpha, \quad \text{and} \quad \sum_a y_{s,a} = m_s \quad \forall s,a. \tag{5}$$

We also say that a policy is continuous (respectively Lipschitz continuous) if for all t, π_t is continuous (respectively Lipschitz continuous).

Note that the definition of an admissible policy is independent of the arm population N and does not assume that if $\mathbf{y} = \pi_t(\mathbf{m})$, then $Ny_{s,a}$ should be an integer. Hence, to make a policy applicable to the original problem with N arms, we use a procedure that we call *randomized rounding* that activates $Ny_{s,1}$ arms in state s *in expectation* and that works as follows:

- In a first pass, one activates $\lfloor Ny_{s,1}\rfloor$ arms in state s, and we let $z_s := Ny_{s,1} - \lfloor Ny_{s,1}\rfloor$;
- In a second pass, one activates an extra $Z_s \in \{0,1\}$ arm in state s, such that for all s, Z_s are random variables that satisfy $\mathbb{E}[Z_s] = z_s \in [0,1)$, and $\sum_s Z_s = \sum_s z_s := h$ (almost surely).

Note that by definition, $h = \lfloor \alpha N\rfloor - \sum_s \lfloor Ny_{s,1}\rfloor$ or $h = \lfloor \alpha N\rfloor + 1 - \sum_s \lfloor Ny_{s,1}\rfloor$ and is therefore an integer. To do the second pass, one cannot simply generate the random variables Z_s independently, because such variables Z_s may not sum to exactly h. An efficient algorithm to solve the above problem can be found in Section 5.2.3 of Ioannidis and Yeh [7]. It has time complexity $\mathcal{O}(hd\cdot\log d)$.

2.4. Notation convention

Throughout our presentation, a bold letter (e.g. $\mathbf{y}$, $\mathbf{m}$) denotes a vector whereas a normal letter (e.g. $y_{s,a}(t)$, $m_s(t)$) denotes a scalar. The bold letter $\mathbf{m}$ always denotes a state vector (that lives in $\Delta^d \subset \mathbb{R}^d$) whereas $\mathbf{y} = (\mathbf{y}_{\cdot,1}, \mathbf{y}_{\cdot,0})$ denotes a state-action pair vector (that lives in $\Delta^{2d} \subset \mathbb{R}^{2d}$). For a vector $\mathbf{m} \in \mathbb{R}^d$, we denote by $\|\mathbf{m}\|_1 = \sum_s |m_s|$ the L_1 norm of $\mathbf{m}$, and $\mathcal{B}(\mathbf{m}^*, \varepsilon) := \{\mathbf{m} \mid \|\mathbf{m} - \mathbf{m}^*\|_1 \le \varepsilon\}$ is the ball centered at $\mathbf{m}^*$ of radius ε. Apart from rare cases, capital letters (e.g. $\mathbf{Y}$, $\mathbf{M}$) denotes random variables whereas small letters denote deterministic values (e.g. $\mathbf{y}$, $\mathbf{m}$). We write $\mathbf{Y}^{(N)}$, $\mathbf{M}^{(N)}$ to emphasize the dependence on arm population N so that each of its coordinate is of the form k/N with $k \in \mathbb{N}$. The function $\mathbf{1}_E$ is a random variable that equals 1 if the event E occurs and 0 otherwise. For a set $\mathcal{S}$, we use $|\mathcal{S}|$ to denote its cardinal.

3. A hierarchy of policies

In this section we introduce a hierarchy of admissible policies having increasingly desirable properties. We first give some preliminary results in Section 3.1. In Section 3.2, we define the notion of LP-compatible policy and show that a continuous admissible policy is asymptotically optimal if and only if it is LP-compatible. If furthermore the policy is Lipschitz continuous, then we obtain a square root convergence rate. In Section 3.3, we show that if the policy is locally linear around one optimal LP solution, then the convergence rate can be improved to be exponential. Proofs of Lemma 1, Theorem 2 and Theorem 3 are given respectively in Section 3.4.1, 3.4.2 and 3.4.3.

3.1. Evolution of $M^{(N)}(\cdot)$ for a given policy

Assume that an admissible policy π is given. To analyse the performance of such a policy, we will analyse how this policy makes the state evolve from $M^{(N)}(t)$ to $M^{(N)}(t+1)$. This evolution is decomposed in three steps: first the policy specifies $\mathbf{Y}(t) = \pi_t(M^{(N)}(t))$, which indicates the proportion of arms that should be activated *on average*, then the randomized rounding procedure produces $Y^{(N)}(t)$, which indicates how many arms should be activated. Lastly, a new state $M^{(N)}(t+1)$ is generated from $Y^{(N)}(t)$. This is summarized in the following diagram:

$$\mathbf{M}^{(N)}(t) \xrightarrow[\text{policy } \pi_t(\cdot)]{\text{admissible}} \mathbf{Y}(t) \xrightarrow[\text{rounding}]{\text{randomized}} \mathbf{Y}^{(N)}(t) \xrightarrow[\text{Markovian transition (1)}]{\text{each arm follows the}} \mathbf{M}^{(N)}(t+1). \tag{6}$$

. In this section, we analyse the Markovian transition that generates $\mathbf{M}^{(N)}(t+1)$ from $\mathbf{Y}^{(N)}(t)$. To do so, we define the function $\phi : \Delta^{2d} \to \Delta^d$ that maps a vector $\mathbf{y} \in \Delta^{2d}$ to a vector $\phi(\mathbf{y}) = \big((\phi(\mathbf{y}))_1, \ \ldots, \ (\phi(\mathbf{y}))_d\big) \in \Delta^d$ whose sth component is

$$(\phi(\mathbf{y}))_s = \sum_{s',a} y_{s,a} P^a_{s',s}. \tag{7}$$

The following lemma shows that $\mathbf{M}^{(N)}(t+1)$ is approximately equal to $\phi(\mathbf{Y}^{(N)}(t))$ when N is large (this is implied by (9)), with an error that decreases as $\mathcal{O}(\frac{1}{\sqrt{N}})$. This observation will be used to show that a continuous admissible policy is optimal if and only if it is LP-compatible. Equation (8) shows that given $\mathbf{Y}^{(N)}(t)$, $\mathbf{M}^{(N)}(t+1)$ is equal to $\phi(\mathbf{Y}^{(N)}(t))$ on average. This fact, combined with the Hoeffding-type inequality (10) and the fact that ϕ is linear, will be critically used in the proof of the exponential rate.

Lemma 1 *Let* $\mathbf{E}^{(N)}(t) = \mathbf{M}^{(N)}(t+1) - \phi(\mathbf{Y}^{(N)}(t))$, *where* $\phi(\cdot)$ *is given in* (7). *We have:*

$$\mathbb{E}\left[\mathbf{E}^{(N)}(t) \mid \mathbf{Y}^{(N)}(t)\right] = \mathbf{0}, \tag{8}$$

$$\mathbb{E}\left[\left\|\mathbf{E}^{(N)}(t)\right\|_1 \mid \mathbf{Y}^{(N)}(t)\right] \leq \frac{\sqrt{d}}{\sqrt{N}}, \tag{9}$$

$$\mathbb{P}\left(\left\|\mathbf{E}^{(N)}(t)\right\|_1 \geq \epsilon \mid \mathbf{Y}^{(N)}(t)\right) \leq 2de^{-2N\epsilon^2/d^2}. \tag{10}$$

A detailed proof of this result is provided in Section 3.4.

3.2. LP-compatibility and asymptotic optimality

For a given admissible policy π, we define $V_\pi^{(N)}(\mathbf{m}(0), T)$ as the expected reward (per arm) when the system has N arms and the policy π is used. For a policy π, we also define $V_\pi(\mathbf{m}(0), T) := \sum_{t=0}^{T-1}\sum_{a,s} R_s^a y_{s,a}^\pi(t)$, where $\mathbf{y}^\pi(t)$ is given by:

$$\begin{aligned}\mathbf{y}^\pi(t) &= \pi_t(\mathbf{m}^\pi(t))\\ \mathbf{m}^\pi(t+1) &= \phi(\mathbf{y}(t)).\end{aligned}$$

We say that a policy π is *LP-compatible* if there exists an optimal solution $\{\mathbf{y}^*(t)\}_{0\leq t\leq T-1}$ of the LP (4), such that $\pi_t(\mathbf{m}^*(t)) = \mathbf{y}^*(t)$ for all $0 \leq t \leq T-1$, where $m_s^*(t) = y_{s,0}^*(t) + y_{s,1}^*(t)$. Following the above definition, an admissible policy is LP-compatible if and only if $V_\pi(\mathbf{m}(0), T) = V_{\text{rel}}(\mathbf{m}(0), T)$.

The following result makes the formal link between LP-compatible policy and asymptotically optimal policies for the N-arms bandit problem. In particular, it shows that a continuous policy π is asymptotically optimal if and only if it is LP-compatible. In addition, the rate of convergence is $\mathcal{O}(\frac{1}{\sqrt{N}})$ when the policy is Lipschitz continuous. Note that this result alone provides necessary and sufficient conditions for asymptotically optimal policy, but does not guarantee the existence of such policies. We will show later in Section 4 that for all finite horizon RB, there always exists a LP-compatible Lipschitz continuous policy that can be easily constructed.

Theorem 2 *Let* $\pi = \{\pi_t\}_{0\leq t\leq T-1}$ *be an admissible and continuous policy. Then:*

$$\lim_{N\to\infty} V_\pi^{(N)}(\mathbf{m}(0), T) = V_\pi(\mathbf{m}(0), T). \tag{11}$$

If in addition π *is Lipschitz continuous, then there exists a constant* $C > 0$ *independent of* N *such that*

$$\left|V_\pi^{(N)}(\mathbf{m}(0), T) - V_\pi(\mathbf{m}(0), T)\right| \leq \frac{C}{\sqrt{N}}. \tag{12}$$

In particular, this implies that:

1. *If* π *is LP-compatible, then* $\lim_{N\to\infty} V_\pi^{(N)}(\mathbf{m}(0), T) = \lim_{N\to\infty} V_{\text{opt}}^{(N)}(\mathbf{m}(0), T) = V_{\text{rel}}(\mathbf{m}(0), T)$.
2. *If* π *is not LP compatible, then* $\limsup_{N\to\infty} V_\pi^{(N)}(\mathbf{m}(0), T) < V_{\text{rel}}(\mathbf{m}(0), T)$.
3. *If* π *is LP-compatible and Lipschitz continuous, then there exists* $C' > 0$ *independent of* N *such that*

$$\left|V_\pi^{(N)}(\mathbf{m}(0), T) - V_{\text{opt}}^{(N)}(\mathbf{m}(0), T)\right| \leq \frac{C'}{\sqrt{N}}.$$

Sketch of proof. A detailed proof is presented in Section 3.4. We give here the main ideas. Recall that $V_\pi^{(N)}(\mathbf{m}(0), T) = \mathbb{E}\left[\sum_{t,a,s} R_s^a Y_{s,a}^{\pi,(N)}(t)\right]$. By using the definition of $V_\pi(\mathbf{m}(0), T)$ and the linearity of expectation, we have:

$$V_\pi^{(N)}(\mathbf{m}(0), T) - V_\pi(\mathbf{m}(0), T) = \sum_{t,a,s} R_s^a \left(\mathbb{E}\left[Y_{s,a}^{\pi,(N)}(t)\right] - y_{s,a}^\pi(t)\right). \tag{13}$$

Consequently, showing that $V_\pi^{(N)}(\mathbf{m}(0), T)$ is close to V_π is equivalent to showing that $\mathbb{E}\left[Y_{s,a}^{\pi,(N)}(t)\right]$ is close to $y_{s,a}^\pi$. In the detailed proof, we show it by recurrence on t using two facts:

- The continuity of π guarantees that if $\mathbf{m}^\pi(t)$ and $\mathbf{M}^{\pi,(N)}(t)$ are close, then so are $\mathbf{y}^\pi(t)$ and $\mathbf{Y}^{\pi,(N)}(t)$.
- Lemma 1 shows that $\mathbf{M}^{\pi,(N)}(t+1) \approx \phi(\mathbf{Y}^{\pi,(N)}(t))$, which implies that if $\mathbf{y}^\pi(t)$ and $\mathbf{Y}^{\pi,(N)}(t)$ are close then so are $\mathbf{m}^\pi(t+1)$ and $\mathbf{M}^{\pi,(N)}(t+1)$. □

3.3. Locally linear policy and exponential convergence rate

As we have shown before, the LP-compatibility is a necessary and sufficient condition for a continuous policy to be asymptotically optimal. In this section, we show that when the policy is locally linear around an optimal solution, then this policy becomes optimal exponentially fast. Note that although LP-compatible policies always exist, this is not always the case for locally linear policies, as we shall see later in Section 4.

We say that an LP-compatible policy $\pi = \{\pi_t\}_{0\le t\le T-1}$ is *locally linear* if there exists a solution $\{\mathbf{y}^*(t)\}_{0\le t\le T-1}$ of (4) such that for all $0 \le t \le T-1$, there exists $\varepsilon_t > 0$ such that $\pi_t(\cdot)$ is *linear* on the ball of radius ε_t centered at $\mathbf{m}^*(t)$, where $m^*_s(t) := y^*_{s,0}(t) + y^*_{s,1}(t)$ for all s.

Theorem 3 *Consider a LP-compatible locally linear policy $\pi = \{\pi_t\}_{0\le t\le T-1}$. There exist two constants $C_1, C_2 > 0$ independent of N such that*

$$\left| V_\pi^{(N)}(\mathbf{m}(0), T) - V_{\text{opt}}^{(N)}(\mathbf{m}(0), T) \right| \le C_1 e^{-C_2 N}$$

We remark that the result of exponential convergence rate in Theorem 3 is much stronger than the general square root rate given in Theorem 2. This is due to the locally linear condition. This local linearity around the optimal trajectory plays a key role in the proof of Theorem 3, as it is used in (18) to justify the interchange of taking expectation with applying a linear function, in order to obtain (19). Our later discussion in Section 4.3.2 actually indicates that the local linearity is essentially necessary to obtain the exponential rate. A second key ingredient in the proof is the concentration inequality (16), which relies on the fact that the N arms are exchangeable. For the more general model where each arm of the bandit has its own state space (this has been considered in Brown and Smith [3] and Hu and Frazier [6]), it is an interesting open question to see if we can formulate an exponential convergence type result in such generic case.

3.4. Proof of results in Section 3

3.4.1. Proof of Lemma 1 For simplicity of notation, let us denote by $\mathbf{y} := \mathbf{Y}^{(N)}(t)$. There are $Ny_{s,a}$ arms in state s and whose action is a and each of these arms makes a transition to state s' with probability $P^a_{s,s'}$. This shows that $M^{(N)}(t+1)$ can be written as a sum of independent random variables as follows:

$$M^{(N)}_{s'}(t+1) = \frac{1}{N} \sum_{s,a} \sum_{i=1}^{Ny_{s,a}} \mathbf{1}_{\{U_{s,a,i} \le P^a_{s,s'}\}},$$

where the variables $U_{s,a,i}$ are i.i.d uniform random variable in $[0,1]$. Taking expectation then gives $\mathbb{E}\left[M^{(N)}_{s'}(t+1) \mid \mathbf{Y}^{(N)}(t)\right] = (\phi(\mathbf{Y}^{(N)}(t)))_{s'}$, which gives (8). It also implies that

$$\begin{aligned}\mathbb{E}\left[|E^{(N)}_{s'}(t+1)|^2 \mid \mathbf{Y}^{(N)}(t) = \mathbf{y}\right] &= \text{var}\left[M^{(N)}_{s'}(t+1) \mid \mathbf{Y}^{(N)}(t) = \mathbf{y}\right] \\ &= \frac{1}{N^2} \sum_{s,a} N y_{s,a} P^a_{s,s'}(1 - P^a_{s,s'}) \le \frac{\sum_{s,a} y_{s,a} P^a_{s,s'}}{N}.\end{aligned}$$

This shows that

$$\mathbb{E}\left[\left\|\mathbf{E}^{(N)}(t+1)\right\|_1 \mid \mathbf{Y}^{(N)}(t) = \mathbf{y}\right] \le \sqrt{d} \frac{\sqrt{\sum_{s'} \sum_{s,a} y_{s,a} P^a_{s,s'}}}{\sqrt{N}} = \frac{\sqrt{d}}{\sqrt{N}},$$

where the first inequality comes from Cauchy-Schwarz, and this gives (9).

Equation (10) is a direct consequence of Hoeffding's inequality. Indeed, one has

$$\mathbb{P}\left(|E_s^{(N)}(t)| \geq \varepsilon/d \,|\, \mathbf{Y}^{(N)}(t)\right) \leq 2e^{-N\varepsilon^2/d^2}.$$

By using the union bound, this implies that

$$\mathbb{P}\left(\left\|\mathbf{E}^{(N)}(t)\right\|_1 | \geq \varepsilon \,|\, \mathbf{Y}^{(N)}(t)\right) \leq d \cdot \mathbb{P}\left(|E_s^{(N)}(t)| \geq \varepsilon/d \,|\, \mathbf{Y}^{(N)}(t)\right) \leq 2de^{-N\varepsilon^2/d^2}.$$

3.4.2. Proof of Theorem 2 Let π be a continuous policy. We will first show by induction on t that $\mathbf{M}^{\pi,(N)}(t)$ converges to $\mathbf{m}^\pi(t)$ in probability as N goes to infinity. This clearly holds for $t=0$ because $\mathbf{m}^\pi(0) = \mathbf{M}^{\pi,(N)}(0) = \mathbf{m}(0)$. Assume that this holds for some $t \geq 0$, and let us show that this implies $\mathbf{Y}^{\pi,(N)}(t)$ also converges to $\mathbf{y}^\pi(t)$ in probability. Indeed, we have

$$\|\mathbf{y}^\pi(t) - \mathbf{Y}^{\pi,(N)}(t)\| \leq \|\pi_t(\mathbf{m}^\pi(t)) - \pi_t(\mathbf{M}^{\pi,(N)}(t))\| + \|\pi_t(\mathbf{M}^{\pi,(N)}(t)) - \mathbf{Y}^{\pi,(N)}(t)\|. \tag{14}$$

By construction of randomized rounding, $\|\pi_t(\mathbf{M}^{\pi,(N)}(t)) - \mathbf{Y}^{\pi,(N)}(t)\| \leq d/N$. This shows that, by continuity of $\pi_t(\cdot)$, if $\mathbf{M}^{\pi,(N)}(t)$ converges in probability to $\mathbf{m}^\pi(t)$, then $\mathbf{Y}^{\pi,(N)}(t)$ also converges to $\mathbf{y}^\pi(t)$ in probability.

For $\mathbf{M}^{\pi,(N)}(t+1)$ and $\mathbf{m}^\pi(t+1)$, we have

$$\|\mathbf{m}^\pi(t+1) - \mathbf{M}^{\pi,(N)}(t+1)\| \leq \|\phi(\mathbf{y}^\pi(t)) - \phi(\mathbf{Y}^{\pi,(N)}(t))\| + \|\mathbf{E}^{(N)}(t)\| \tag{15}$$

As ϕ is continuous and $\mathbf{E}^{(N)}(t)$ converges to $\mathbf{0}$ in probability, this implies that $\mathbf{M}^{\pi,(N)}(t+1)$ converges to $\mathbf{m}^\pi(t+1)$ in probability. This concludes the induction step. Consequently, $\mathbf{Y}^{\pi,(N)}(t)$ converges in probability to $\mathbf{y}^\pi(t)$. As $Y_{s,a}^{\pi,(N)}(t) \in [0,1]$ are bounded, the dominated convergence theorem implies that $\lim_{N\to\infty} \mathbb{E}_\pi\left[Y_{s,a}^{\pi,(N)}(t)\right] = y_{s,a}^\pi(t)$, which by (13) implies (11).

Assume now that for all t, π_t is Lipschitz continuous. As ϕ is linear, ϕ is also Lipschitz continuous. Let L be an upper bound on the Lipschitz constants of π and ϕ. Applying (15), Lemma 1 and (14), we have:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{m}^\pi(t+1) - \mathbf{M}^{\pi,(N)}(t+1)\|\right] &\leq \mathbb{E}\left[\|\phi(\mathbf{y}^\pi(t)) - \phi(\mathbf{Y}^{\pi,(N)}(t))\|\right] + \mathbb{E}\left[\|\mathbf{E}^{(N)}(t)\|\right] \\
&\leq L\mathbb{E}\left[\|\mathbf{y}^\pi(t) - \mathbf{Y}^{\pi,(N)}(t)\|\right] + \sqrt{\frac{d}{N}} \\
&\leq L^2\mathbb{E}\left[\|\mathbf{m}^\pi(t) - \mathbf{M}^{\pi,(N)}(t)\|\right] + \frac{Ld}{N} + \sqrt{\frac{d}{N}}.
\end{aligned}$$

By a direct induction on t (which is essentially the discrete Gronwall's lemma), this implies that $\mathbb{E}\left[\|\mathbf{m}^\pi(t+1) - \mathbf{M}^{\pi,(N)}(t+1)\|\right] = \mathcal{O}(\frac{1}{\sqrt{N}})$. Note however that the hidden constant in the $\mathcal{O}(\cdot)$ grows exponentially with time t. By (13), this implies (12).

To conclude the proof, one should note that a policy π is LP-compatible if and only if $V_\pi(\mathbf{m}(0), T) = V_{\text{rel}}(\mathbf{m}(0), T)$. □

3.4.3. Proof of Theorem 3 Let $\varepsilon := \min_t \varepsilon_t$, and let $F_t : \Delta^d \to \Delta^{2d}$ be the linear function such that $\pi_t(\mathbf{m}) = F_t(\mathbf{m})$ for $\mathbf{m} \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon)$. Denote by $\ell > 0$ the Lipschitz constant of the linear map $\phi(\cdot)$, and by $L_t > 0$ the Lipschitz constant of F_t and write $L := \max_t L_t$.

Let $\delta := \varepsilon/(2\,(1 + \ell L + \cdots + (\ell L)^T))$, and let us denote by $\mathcal{E}(\delta)$ the event:

$$\mathcal{E}(\delta) := \left\{\text{for all } 0 \leq t \leq T-1\text{: } \|\mathbf{E}^{(N)}(t) \leq \delta\|\right\},$$

where $\mathbf{E}^{(N)}(t)$ is defined as in Lemma 1, and let $\overline{\mathcal{E}(\delta)}$ be the complementary of the event $\mathcal{E}(\delta)$.

By (10) of Lemma 1, we have

$$\mathbb{P}\left(\overline{\mathcal{E}(\delta)}\right) \leq 2dT \cdot e^{-2N\delta^2/d^2}. \tag{16}$$

Assume that event $\mathcal{E}(\delta)$ holds. By definition of $\mathbf{E}^{(N)}(t)$ and (6), we have

$$\begin{aligned}
\left\|\mathbf{M}^{(N)}(t+1) - \mathbf{m}^*(t+1)\right\|_1 &= \left\|\phi(\mathbf{Y}^{(N)}(t)) + \mathbf{E}^{(N)}(t) - \phi(\pi_t(\mathbf{m}^*(t)))\right\|_1 \\
&\leq \left\|\phi(\mathbf{Y}^{(N)}(t)) - \phi(\mathbf{Y}(t))\right\|_1 + \left\|\phi(\mathbf{Y}(t)) - \phi(\pi_t(\mathbf{m}^*(t)))\right\|_1 + \left\|\mathbf{E}^{(N)}(t)\right\| \\
&= \left\|\phi(\mathbf{Y}^{(N)}(t)) - \phi(\mathbf{Y}(t))\right\|_1 \\
&\quad + \left\|\phi(\pi_t(\mathbf{M}^{(N)}(t))) - \phi(\pi_t(\mathbf{m}^*(t)))\right\|_1 + \left\|\mathbf{E}^{(N)}(t)\right\| \\
&\leq \frac{2d\ell}{N} + \ell L \cdot \left\|\mathbf{M}^{(N)}(t) - \mathbf{m}^*(t)\right\|_1 + \delta.
\end{aligned} \tag{17}$$

A direct induction until $t = 0$ then implies

$$\left\|\mathbf{M}^{(N)}(t+1) - \mathbf{m}^*(t+1)\right\|_1 \leq (1 + \ell L + \cdots + (\ell L)^t) \cdot (\delta + \frac{2d\ell}{N}).$$

This implies that $\mathbf{M}^{(N)}(t)$ is inside $\mathcal{B}(\mathbf{m}^*(t), \varepsilon)$ for all $0 \leq t \leq T-1$ and $N \geq 2d\ell/\delta$. As a side note, the term $2d\ell/N$ in (17) and the assumption $N \geq 2d\ell/\delta$ will not appear, if the locally linear policy can be constructed as a time-dependent priority policy, as in Proposition 5 for rankable finite horizon RB, since then no randomized rounding is needed anywhere and $\mathbf{Y}^{(N)}(t) = \mathbf{Y}(t)$ always holds.

Consequently, we get:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - \mathbf{y}^*(t) &= \mathbb{E}\left[F_t(\mathbf{M}^{(N)}(t))\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - F_t(\mathbf{m}^*(t)) \\
&= \mathbb{E}\left[F_t\left(\phi(\mathbf{Y}^{(N)}(t-1)\mathbf{1}_{\{\mathcal{E}(\delta)\}})\right)\right] - F_t\left(\phi(\mathbf{y}^*(t-1))\right) \\
&= F_t \circ \phi\left(\mathbb{E}\left[\mathbf{Y}^{(N)}(t-1)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - \mathbf{y}^*(t-1)\right),
\end{aligned} \tag{18}$$

where on the last equality (18) we have interchanged the expectation $\mathbb{E}_\pi[\cdot]$ with $F_t \circ \phi(\cdot)$, which is possible since the later is a linear map. A direct induction on t then implies that

$$\begin{aligned}
\left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - \mathbf{y}^*(t)\right\|_1 &\leq L' \left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t-1)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - \mathbf{y}^*(t-1)\right\|_1 \\
&\leq (L')^T \left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(0)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - \mathbf{y}^*(0)\right\|_1.
\end{aligned} \tag{19}$$

where L' is an upper bound on the Lipschitz constants of maps $F_t \circ \phi(\cdot)$ for $0 \leq t \leq T-1$. Moreover by (16), we have

$$\left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right] - \mathbb{E}\left[\mathbf{Y}^{(N)}(t)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right]\right\|_1 \leq 2d \cdot \mathbb{P}\left(\bar{\mathcal{E}}(\delta)\right) \leq 4d^2 T e^{-C_2 N}, \tag{20}$$

where $C_2 := -2\varepsilon^2/((1 + \cdots + L^{T-1})^2 d^2)$. Combining (19) and (20) gives

$$\left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right] - \mathbf{y}^*(t)\right\|_1 \leq C_1 e^{-C_2 N},$$

where we may choose $C_1 := 4d^2T^2(1 + (L')^T)$. Consequently, by (13), all locally linear LP-compatible policies are asymptotically optimal with exponential rate, and this concludes our proof. □

4. Existence and construction of policies

In this section we provide constructions of Lipschitz continuous policies and locally linear policies, defined in the previous Section 3. In Section 4.1 we define the non-degenerate and rankable RB. In Section 4.2, we introduce the idea of "water-filling", and show that the policies induced by "water-filling" are LP-compatible Lipschitz continuous policies, and are furthermore locally linear policies if the RB is non-degenerate. We compare the non-degenerate condition with the rankable condition in Section 4.3.1. In Section 4.3.2, we construct a degenerate 2-dimensional RB over which no policy converges asymptotically fast to the LP solution. This implies that non-degeneracy is a necessary condition for the exponential convergence rate in general. Proofs of Theorem 5 and Lemma 6 are given respectively in Section 4.3.3 and 4.3.4.

4.1. Non-degenerate and rankable RB

Let $\{\mathbf{y}^*(t)\}_{0\le t\le T-1}$ be an optimal solution of the LP relaxed problem (4). For each time t, we partition the set $\mathcal{S}$ into four sets $\mathcal{S}^+(t)$, $\mathcal{S}^0(t)$, $\mathcal{S}^-(t)$ and $\mathcal{S}^\emptyset(t)$ as follows:

$$\begin{aligned}
\mathcal{S}^+(t) &:= \{s\in\mathcal{S} \mid y^*_{s,1}(t)>0 \text{ and } y^*_{s,0}(t)=0\};\\
\mathcal{S}^0(t) &:= \{s\in\mathcal{S} \mid y^*_{s,1}(t)>0 \text{ and } y^*_{s,0}(t)>0\};\\
\mathcal{S}^-(t) &:= \{s\in\mathcal{S} \mid y^*_{s,1}(t)=0 \text{ and } y^*_{s,0}(t)>0\};\\
\mathcal{S}^\emptyset(t) &:= \{s\in\mathcal{S} \mid y^*_{s,1}(t)=0 \text{ and } y^*_{s,0}(t)=0\}.
\end{aligned}$$

The intuition behind this partition is as follows: For the optimal relaxed solution $\mathbf{y}^*$, at time t, it is optimal to activate all arms whose state is in $\mathcal{S}^+(t)$, a fraction of those whose state is in $\mathcal{S}^0(t)$, and none of those whose state is in $\mathcal{S}^-(t)$. Also note that the optimal solution is such that at time t, there are no arms whose state is in $\mathcal{S}^\emptyset(t)$: for all $s\in\mathcal{S}^\emptyset(t)$, we have $m^*_s(t)=y^*_{s,0}(t)+y^*_{s,1}(t)=0$.

Following this intuitive definition, we construct below a LP-compatible Lipschitz continuous policy that activates in priority the arms in set $\mathcal{S}^+(t)$, then the ones in $\mathcal{S}^0(t)$ and then the ones in $\mathcal{S}^-(t)$. As we shall see below, one has to be careful on how to deal with the arms in $\mathcal{S}^0(t)$.

Before defining the water-filling policy, and for reasons that will become clear in Theorem 5 and Theorem 7, we introduce two definitions:

1. A RB is *rankable* if there exists an optimal solution of (4) for which $|\mathcal{S}^0(t)|\le 1$ for all t. Otherwise we call this RB *non-rankable*.
2. A RB is *non-degenerate* if there exists an optimal solution $\{\mathbf{y}^*(t)\}_{0\le t\le T-1}$ of (4) for which $|\mathcal{S}^0(t)|\ge 1$ for all t. Otherwise we call this RB *degenerate*. This definition coincides with the one in Zhang and Frazier [18].

At first glance it appears that rankable and non-degenerate RB's are complementary to each other. Surprisingly, it turns out that in practice these two conditions are *almost* equivalent, as stated by the next result, that we prove and comment in Section 4.3.1.

Proposition 4 *Consider a RB for which the LP problem (4) has a unique solution. If this RB is not rankable, then it is degenerated.*

We say that a policy is a *(time-dependent) priority policy* if for all time t, there exists a permutation $\sigma=\sigma_1\dots\sigma_d$ of the states (that depends on t) such that the policy activates first the arms in state σ_1, then the ones in state σ_2, etc. up to activating a fraction α of arms. In other words, if the arm configuration vector at time t is $\mathbf{m}\in\Delta^d$, then the policy will activate $y_{s,1}$ arms in state s, where for all $i\in\{1\dots d\}$, $y_{\sigma_i,1}$ is defined as:

$$y_{\sigma_i,1} := \pi^{\text{priority}(\sigma)}_{\sigma_i,1}(\mathbf{m}) = \min(m_{\sigma_i}, \alpha-\sum_{j=1}^{i-1} y_{\sigma_j,1}). \tag{21}$$

The next theorem justifies the notion of rankable RB.

Theorem 5 *A RB is rankable if and only if there exists a time-dependent priority policy that is asymptotically optimal.*

The proof of this result is postponed to Section 4.3.3.

As we shall see later, one can use any order inside $\mathcal{S}^+(t)$ or $\mathcal{S}^-(t)$ and still obtain an asymptotically optimal policy (although some orders are better than others as we elaborate in Section 5.1 and Section 7.1). Theorem 5 shows that one has to be careful on dealing with the states in $\mathcal{S}^0(t)$: if the RB is non-rankable, i.e. if $|\mathcal{S}^0(t)|>1$ for some t, one cannot simply use a fixed priority order between those states at time t to obtain an asymptotically optimal policy. To do so, we shall introduce the idea of "water-filling".

4.2. The water-filling policy

At time t, the water-filling policy observes $\mathbf{M}^{(N)}(t) \in \Delta^d$ and decides $\mathbf{Y}(t) \in \Delta^{2d}$, where $Y_{s,1}(t)$ is the expected fraction of arms that are in state s and should be activated (recall that $\mathbf{Y}^{(N)}(t)$ is then generated from $\mathbf{Y}(t)$ by applying randomized rounding). This policy works as follows. For ease of notation, we drop momentarily the t from the notations and we assume that the states are ordered so that the first $|\mathcal{S}^+|$ states are in $\mathcal{S}^+$, the next $|\mathcal{S}^0|$ states are in $\mathcal{S}^0$, the next $|\mathcal{S}^-|$ states are in $\mathcal{S}^-$, and finally the rest are in $\mathcal{S}^\emptyset$. We view the states as d buckets enumerated from 1 to d, where bucket number $1 \le s \le d$ has capacity $M_s^{(N)}$ and α is the total quantity of water that needs to be poured into these buckets. We fill the buckets one by one in *increasing* order of their numbers, except for the first pass in $\mathcal{S}^0$ as we describe next:

1. We first activate all arms in $\mathcal{S}^+$ by using a strict priority order on the states $1, \cdots, |\mathcal{S}^+|$. The only constraint is to activate no more than what we have, i.e. $Y_{s,1} \le M_s^{(N)}$ for $s \in \mathcal{S}^+$.
2. If there is still some water left, we then activate states in $\mathcal{S}^0$ by using a *reversed* priority order on the states, namely $|\mathcal{S}^+| + |\mathcal{S}^0|, \cdots, |\mathcal{S}^+| + 1$ with the constraint that $Y_{s,1} \le \min(M_s^{(N)}, y_{s,1}^*)$ for $s \in \mathcal{S}^0$.
3. If there is still some water left, we then complete by activating states in $\mathcal{S}^0$ and then in $\mathcal{S}^-$ and then in $\mathcal{S}^\emptyset$ by using the priority order $|\mathcal{S}^+| + 1, \cdots, d$.

Note that if for all t we have $|\mathcal{S}^0(t)| \le 1$, the water-filling policy becomes a time-dependent priority policy.

The next lemma shows that the water-filling policy is LP-compatible Lipschitz continuous, and is furthermore locally linear if the RB is non-degenerate.

Lemma 6 *For any finite horizon RB, the water-filling policy described above is a LP-compatible Lipschitz continuous policy. Moreover, if the RB is non-degenerate, i.e. if for all t, $|\mathcal{S}^0(t)| \ge 1$, then the water-filling policy is a LP-compatible locally linear policy. And if the RB is degenerate, then there is no LP-compatible locally linear policy.*

The proof of Lemma 6 is postponed to Section 4.3.4. A direct consequence of this lemma, combined with Theorem 2 and Theorem 3 is that the water-filling policy is asymptotically optimal at rate at least $\mathcal{O}(\frac{1}{\sqrt{N}})$.

Theorem 7 *For any finite horizon RB, there exists a policy π (constructed by the water-filling procedure) and $C > 0$ such that for any N:*

$$\left| V_\pi^{(N)}(\mathbf{m}(0), T) - V_{\text{opt}}^{(N)}(\mathbf{m}(0), T) \right| \le \frac{C}{\sqrt{N}}. \tag{22}$$

Moreover, if the problem is non-degenerate, then there exists a policy π and $C_1, C_2 > 0$ such that:

$$\left| V_\pi^{(N)}(\mathbf{m}(0), T) - V_{\text{opt}}^{(N)}(\mathbf{m}(0), T) \right| \le C_1 e^{-C_2 N}. \tag{23}$$

Lemma 6 shows that the non-degenerate condition is necessary and sufficient for the existence of a LP-compatible locally linear policy. Theorem 7 is less precise in the sense that we only show that non-degeneracy is sufficient to obtain an exponentially asymptotically optimal policy. In Section 4.3.2, we provide an example of a RB that is degenerate and for which there are no exponentially fast asymptotically optimal policy. Although we do not prove it, we conjecture that this holds in general so that the non-degeneracy is also a necessary condition for (23) to hold.

Remark 8 *Note that the authors of Zhang and Frazier [18] introduce a class of fluid-priority policies (in their Algorithm 1) that is very close to our definition of water-filling policy. In fact, when $|\mathcal{S}^0(t)| \le 1$, both definition coincide and they both correspond to the same priority policy. When $|\mathcal{S}^0(t)| \ge 2$, there are two differences between their algorithm and ours:*

- *When $Ny^*_{s,1}(t)$ is not an integer: the authors choose to round fractional number of arms into integer numbers in the water-filling procedure, e.g. no more than $\lfloor Ny^*_{s,1} \rfloor$ arms can be activated in state $s \in \mathcal{S}^0$, whereas we consider the water-filling procedure as a map from any vector $\mathbf{m} \in \Delta^d$ into the decision vector $\mathbf{y} \in \Delta^{2d}$, and apply the randomized rounding technique afterwards to avoid rounding errors.*
- *When one needs to activate more than $Ny^*_{s,1}(t)$ arms in state $s \in \mathcal{S}^0(t)$, we do a second pass of water-filling algorithm by using a reversed order on $\mathcal{S}^0(t)$ as in the first pass, whereas in Algorithm 1 of Zhang and Frazier [18] the two passes are done in the same order. Using a reversed order allows us to establish the local linearity of π around $\mathbf{m}^*$, which would not be the case if the two passes were done in the same order. This is essential in our proof of the exponential convergence rate in the non-degenerate case.*

Note that in Zhang and Frazier [18] the authors only obtain the $\mathcal{O}(\frac{1}{N})$ convergence rate for their algorithm. We believe that this is mainly due to their rounding procedure.

4.3. Proof of results in Section 4

4.3.1. Proof of Proposition 4 We can actually prove the slightly more general result that claims that for any RB the optimization problem (4) has an optimal solution $\{\mathbf{y}^*(t)\}_{0 \le t \le T-1}$ satisfying

$$\sum_{t=0}^{T-1} \left|\mathcal{S}^0(t)\right| \le T. \tag{24}$$

Indeed, similar to our formulation of the optimization problem as a MDP that we later detail in Equation (33), we can transform the optimization problem (4) into a *constraint* MDP, where the T constraints come from (4c). We then apply Theorem 3.8 of Altman [1], which states that for a feasible infinite horizon discounted MDP with T inequality constraints, there exists an optimal stationary policy such that the total number of randomization that it uses is at most T. Since finite horizon MDP is a sub-class of infinite horizon discounted MDP, and one number of randomization corresponds exactly to one tuple (s,t) such that $s \in \mathcal{S}^0(t)$, our claim in (24) follows.

Proposition 4 is then a direct consequence of Equation (24): if there exists a unique solution, it must satisfy (24), which by the pigeonhole principle implies that either $|\mathcal{S}^0(t)| \le 1$ for all t (the problem is rankable) or there exists t such that $|\mathcal{S}^0(t)| = 0$ (the problem is degenerate). □

The above result implies that under the assumption of a unique solution, a problem that is non-rankable cannot be non-degenerate. This leaves two questions. First, is there a problem that is both rankable and degenerate? The answer is yes and we provide a small example below. Second, what happens when the LP has multiple solutions? The answer to this question is harder and is left for future work. Our view is that, except for very particular problems that have a lot of symmetries, the solution to the LP is mostly unique. If there are multiple solutions, Equation (24) implies that one can always construct a solution such that $|\mathcal{S}^0(t)| \le 1$ for all t (i.e. the problem rankable), or otherwise there always exists t such that $|\mathcal{S}^0(t)| = 0$ for those solutions. Yet, verifying that there are no other solutions is difficult in general.

Example 1 (A rankable and degenerate problem). Let us consider a two states RB with a proportion of activation $\alpha = 0.5$. The initial condition is $\mathbf{m}(0) = [0.5, 0.5]$, the rewards are $\mathbf{R}^0 = [0,0]$ and $\mathbf{R}^1 = [1,0]$, and the matrices are identity matrices: $\mathbf{P}^0 = \mathbf{P}^1 = \mathbf{I}$. The solution to the LP is clearly unique and consists of activating all arms in state 1 and no arms in state 2. Hence, $|\mathcal{S}^0(t)| = 0$ for all t. This example is rankable and is also degenerate.

4.3.2. Necessary condition for exponential convergence rate Consider a two states RB with horizon $T = 2$ and proportion of activation $\alpha = 0.5$. The initial condition is $\mathbf{m}(0) = [0.5, 0.5]$. The rewards are $\mathbf{R}^0 = [0,0]$, $\mathbf{R}^1 = [1,0]$. The transition matrices are

$$\mathbf{P}^1 = \begin{pmatrix} p_1 & 1-p_1 \\ p_2 & 1-p_2 \end{pmatrix}, \mathbf{P}^0 = \begin{pmatrix} q_1 & 1-q_1 \\ q_2 & 1-q_2 \end{pmatrix},$$

with $0 \le p_1, p_2, q_1, q_2 \le 1$. Let us first establish a sufficient condition on the four parameters p_1, p_2, q_1 and q_2 so that the RB is degenerate. For this simple model, solving the linear program (4) amounts to finding the optimal value $0 \le \beta \le 0.5 = \alpha$ as the proportion of activation of arms in state 1 at decision epoch $t = 0$. At decision epoch $t = 1$, there will then be $\beta p_1 + (0.5-\beta)q_1 + (0.5-\beta)p_2 + \beta q_2$ arms in state 1, and the optimal value of (4) is

$$\begin{aligned} & \beta + \min\{0.5,\ \beta p_1 + (0.5-\beta)q_1 + (0.5-\beta)p_2 + \beta q_2\} \\ = \ & \beta + \min\{0.5,\ \beta(p_1+q_2) + (0.5-\beta)(q_1+p_2)\} \end{aligned}$$

By definition, the RB is degenerate if

$$\arg\max_{0\le\beta\le 0.5} \{\beta + \min\{0.5,\ \beta(p_1+q_2) + (0.5-\beta)(q_1+p_2)\}\} \neq 0,\ 0.5, \tag{25}$$

since then $\mathcal{S}^0(0) = \{1, 2\}$. A sufficient condition for (25) to hold is

$$q_1 + p_2 > 1 + p_1 + q_2, \tag{26}$$

under which the arg max of (25) is $\beta^* = 0.5 \times \frac{q_1+p_2-1}{(q_1+p_2)-(p_1+q_2)}$ and $\mathbf{m}^*(1) = [0.5, 0.5]$, so we activate exactly all the proportion $0.5 = \alpha$ of arms in state 1 at decision epoch $t = 1$. Note that we get $|\mathcal{S}^0(0)| = 2$ and $|\mathcal{S}^0(1)| = 0$.

We next consider a stochastic model with a population of N arms, where the 2-dimensional RB satisfies (26) so that it is degenerate. For any LP-compatible policy, our only choice is to activate $\beta^* N$ arms in state 1, $(0.5-\beta^*)N$ arms in state 2 at decision epoch $t = 0$ (apply randomized rounding if necessary); and by the specific choice of values for rewards $\mathbf{R}^0$, $\mathbf{R}^1$, we need to activate as many arms as possible in state 1 at decision epoch $t = 1$. The expected average reward under this policy is then $\beta^* + \mathbb{E}[\min\{0.5, G_N\}]$, where the random variables G_N (indexed by N) inside the bracket are

$$G_N := \frac{bin(\beta^* N, p_1) + bin((0.5-\beta^*)N, q_1) + bin((0.5-\beta^*)N, p_2) + bin(\beta^* N, q_2)}{N}.$$

We have $\mathbb{E}[G_N] = 0.5$ by definition of the value β^*. Moreover, by elementary probability theory, one has

$$\sqrt{N} \cdot \mathbb{E}[0.5 - \min\{0.5, G_N\}] \xrightarrow[N\to\infty]{} C > 0.$$

Since the optimal value of (4) is $\beta^* + 0.5$, this implies that the square root of N convergence with respect to this relaxed upper-bound value can not be improved on this degenerate RB, and it is not due to the problem at decision epoch $t = 0$ with $|\mathcal{S}^0(0)| > 1$, but due to the fact that at $t = 1$ one has $|\mathcal{S}^0(1)| = 0$, and the optimal trajectory is on the boundary of two zones, namely $\left\{\mathbf{m} \in \Delta^d \mid \sum_{s\in\mathcal{S}^+(1)} \le \alpha\right\}$ and $\left\{\mathbf{m} \in \Delta^d \mid \sum_{s\in\mathcal{S}^+(1)} \ge \alpha\right\}$. Note that this example implies in particular that the $\mathcal{O}(\frac{1}{\sqrt{N}})$ convergence rate in Theorem 2 is tight.

Generally speaking, for a degenerate RB, there exists some t for which $|\mathcal{S}^0(t)| = 0$. This implies that $\sum_{s\in\mathcal{S}^+(t)} m^*_s(t) = \alpha$, which means at time t the optimal trajectory is on the boundary of two zones $\left\{\mathbf{m} \in \Delta^d \mid \sum_{s\in\mathcal{S}^+(t)} \le \alpha\right\}$ and $\left\{\mathbf{m} \in \Delta^d \mid \sum_{s\in\mathcal{S}^+(t)} \ge \alpha\right\}$. It is exactly this phenomenon that may prevent an exponentially fast convergence rate.

4.3.3. Proof of Theorem 5 Assume first that the RB is rankable and let $\{\mathbf{y}^*(t)\}_{0\le t\le T-1}$ be an optimal solution of the LP-problem. For each time t, we consider a permutation $\sigma(t)$ that orders the state by starting from the states in $\mathcal{S}^+(t)$, then the only state in $\mathcal{S}^0$, then the states in $\mathcal{S}^-(t)$ and finally the states in $\mathcal{S}^\emptyset$. Let π^{priority} be the time-dependent priority policy that activates at time t the states following the order $\sigma(t)$. By (21), this policy is piecewise affine (with finitely many pieces) and continuous. It is therefore Lipschitz continuous.

We now show that π^{priority} is such that $\pi^{\text{priority}}(\mathbf{m}^*(t)) = \mathbf{y}^*(t)$. By definition of $\mathcal{S}^+(t)$, for all $s \in \mathcal{S}^+(t)$, $y^*_{s,1}(t) = m^*_s(t)$. Let s_0 be the only state in $\mathcal{S}^0(t)$. As $\sum_s y^*_{s,1}(t) = \alpha$, this implies that $\sum_{s\in\mathcal{S}^+(t)} y^*_{s,1}(t) < \alpha$ and therefore that $y^*_{s_0,1} = \alpha - \sum_{s\in\mathcal{S}^+(t)} y^*_{s,1}(t)$. This shows that $y^*_{s,1}(t)$ satisfies the definition of the time-varying policy (21). Note that if $\mathbf{m}$ is such that $0 \le \alpha - \sum_{s\in\mathcal{S}^+(t)} m_s(t) \le m_{s_0}$, then one has:

$$\pi^{\text{priority}}_s(\mathbf{m}) = \begin{cases} m_s & \text{if } s \in \mathcal{S}^+(t) \\ \alpha - \sum_{s\in\mathcal{S}^+(t)} m_s(t) & \text{if } s \in \mathcal{S}^0(t) \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

As a byproduct (which is not used in this proof but will be used later), this also implies that π^{priority} is locally linear if $|\mathcal{S}^0(t)| = 1$ for all t.

Assume now that the RB is non-rankable and let π be a time-dependent priority policy. By construction, at any time t, π activates the states following a permutation $\sigma(t)$. Hence, if there exists at most one state $s = \sigma_i(t)$ such that $\pi_{s,0}(\mathbf{m}^*(t)) > 0$, $\pi_{s,1}(\mathbf{m}^*(t)) > 0$, and for all $j < i$, $\pi_{\sigma_j,0}(\mathbf{m}^*(t)) = 0$, and for all $j > i$, $\pi_{\sigma_j,1}(\mathbf{m}^*(t)) = 0$. This shows that for all t, $|\{s : \pi_{s,0}(\mathbf{m}^*(t)) > 0 \text{ and } \pi_{s,1}(\mathbf{m}^*(t)) > 0\}| \le 1$. Hence, π cannot be LP-compatible because all solutions of (4) are such that there exists a time t such that $|\mathcal{S}^0(t)| \ge 2$, which is implied by the assumption that the RB is non-rankable.

4.3.4. Proof of Lemma 6 Fix $(\mathbf{M}^{(N)}, \mathbf{y}^*)$ as the input for the "water-filling" in dimension d, and let $\mathbf{Y} \in \alpha \cdot \Delta^d$ be the corresponding output. Suppose that the states are sorted so that the first s_+ states are $\mathcal{S}^+ := \{s^+_1, \cdots, s^+_{s_+}\}$, the next s_0 states are $\mathcal{S}^0 := \{s^0_1, \cdots, s^0_{s_0}\}$, the next s_- states are $\mathcal{S}^- := \{s^-_1, \cdots, s^-_{s_-}\}$, and the rest $s_\emptyset$ states are $\mathcal{S}^\emptyset := \{s^\emptyset_1, \cdots, s^\emptyset_{s_\emptyset}\}$. So in total $s_+ + s_0 + s_- + s_\emptyset = d$.

In what follows, we show how the water-filling policy can be viewed as a fixed priority policy over a larger state-space. To see that, we define an auxiliary set of states $\widehat{\mathcal{S}}$ with cardinal $\widehat{d} := s_+ + (2s_0 - 1) + s_- + s_\emptyset$ in which we duplicate all states in $\mathcal{S}^0(t)$ except one:

$$\widehat{\mathcal{S}} := \Big\{ s^+_1, \cdots, s^+_{s_+}, \underbrace{\overline{s}^0_{s_0}, \cdots, \overline{s}^0_2}_{\overline{\mathcal{S}^0}}, s^0_1, \underbrace{\underline{s}^0_2, \cdots, \underline{s}^0_{s_0}}_{\underline{\mathcal{S}^0}}, s^-_1, \cdots, s^-_{s_-}, s^\emptyset_1, \cdots, s^\emptyset_{s_\emptyset} \Big\}, \tag{28}$$

and we define the state $\widehat{\mathbf{M}^{(N)}}$ as:

$$\widehat{M^{(N)}_s} := \begin{cases} M^{(N)}_s, & \text{if } s \in \mathcal{S}^+ \bigcup \mathcal{S}^- \bigcup \mathcal{S}^\emptyset \bigcup \{s^0_1\} \\ \min(M^{(N)}_{s^0_i}, y^*_{s^0_i,1}), & \text{if } s = \overline{s}^0_i \in \overline{\mathcal{S}^0} \\ M^{(N)}_{s^0_i} - \min(M^{(N)}_{s^0_i}, y^*_{s^0_i,1}), & \text{if } s = \underline{s}^0_i \in \underline{\mathcal{S}^0}. \end{cases} \tag{29}$$

Let $\widehat{\mathbf{Y}}$ be the output of a strict priority policy with the input vector $\widehat{\mathbf{M}^{(N)}}$ and where the states activated following the order as in (28). Let $\mathbf{Y}$ be defined as in

$$Y_s := \begin{cases} \widehat{Y}_s, & \text{if } s \in \mathcal{S}^+ \bigcup \mathcal{S}^- \bigcup \mathcal{S}^\emptyset \bigcup \{s^0_{s_0}\} \\ \widehat{Y_{\overline{s}^0_i}} + \widehat{Y_{\underline{s}^0_i}}, & \text{if } s = s^0_i \text{ with } 1 \le i \le s_0 - 1. \end{cases} \tag{30}$$

By construction, the vector $\mathbf{Y}$ corresponds to the vector obtained by the water-filling algorithm constructed in Section 4.2.

Now, consider the map chain

$$(\mathbf{M}^{(N)}, \mathbf{y}^*) \xrightarrow{(29)} (\widehat{\mathbf{M}^{(N)}}) \xrightarrow{\text{strict priority}} \widehat{\mathbf{Y}} \xrightarrow{(30)} \mathbf{Y}. \tag{31}$$

It should be clear that (29) and (30) are Lipschitz continuous functions. As a strict priority policy is Lipschitz continuous, this shows that the water-filling policy is Lipschitz continuous.

Moreover, if $|\mathcal{S}^0| \geq 1$, then (29) is locally linear (and by (27), the strict priority policy used is also locally linear). As (31) is locally linear, this implies that when the RB is non-degenerate, the water-filling policy constructed from this solution is therefore locally linear.

We now show by contradiction that the non-degenerate condition is necessary to obtain a locally linear policy. Assume that the problem is degenerate and consider a solution y^* of the LP problem (4). As the problem is degenerate, there exists t such that $\mathcal{S}^0(t)$ is empty. In the following this t is fixed and omitted from the notation for simplicity.

At time t, we have $\sum_{s\in\mathcal{S}^+} m_s^* = \alpha$. Let us consider an arbitrary function from Δ^d to Δ^{2d} that is locally linear in a small neighborhood of $\mathbf{m}^*$, and we shall show that the policy induced by this function cannot be admissible. Indeed, this linear function is defined by a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ so that $\mathbf{y}_{\cdot,1} = \mathbf{m}\cdot\mathbf{A}$ for any $\mathbf{m}$ in this neighborhood of $\mathbf{m}^*$, and in particular $\mathbf{y}^*_{\cdot,1} = \mathbf{m}^*\cdot\mathbf{A}$. Denote by $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ a small perturbation vector so that $\mathbf{m}^* + \boldsymbol{\varepsilon} \in \Delta^d$ remains in the neighborhood. The assumption of admissibility yields

$$\mathbf{0} \leq (\mathbf{m}^* + \boldsymbol{\varepsilon})\cdot\mathbf{A} = \mathbf{y}^*_{\cdot,1} + \boldsymbol{\varepsilon}\cdot\mathbf{A} \leq \mathbf{m}^* + \boldsymbol{\varepsilon}, \tag{32}$$

where the inequalities are considered componentwise.

Consider now a state $i \in \mathcal{S}^+$, one has $y^*_{i,1} = m^*_i$, hence (32) implies that $(\boldsymbol{\varepsilon}\cdot\mathbf{A})_i \leq \varepsilon_i$. We next replace $\boldsymbol{\varepsilon}$ by $-\boldsymbol{\varepsilon}$, note that this is possible since we are considering a neighbourhood of $\mathbf{m}^*$, and we obtain the inequality in the other direction: $(\boldsymbol{\varepsilon}\cdot\mathbf{A})_i \geq \varepsilon_i$. Consequently, $(\boldsymbol{\varepsilon}\cdot\mathbf{A})_i = \varepsilon_i$ for $i \in \mathcal{S}^+$. Similarly, for a state $i \in \mathcal{S}^-$, using the same idea we obtain $(\boldsymbol{\varepsilon}\cdot\mathbf{A})_i = 0$. This implies that $A_{ij} = \delta_{ij}$ for $i,j \in \mathcal{S}^+$, and $A_{ij} = 0$ for $i,j \in \mathcal{S}^-$. In particular, this matrix $\mathbf{A}$ tells us to activate all arms in $\mathcal{S}^+$ for any $\mathbf{m}$ in a small neighbourhood of $\mathbf{m}^*$. However, since $\sum_{s\in\mathcal{S}^+} m^*_s = \alpha$, in any neighbourhood of $\mathbf{m}^*$, there always exists $\mathbf{m}$ such that $\sum_{s\in\mathcal{S}^+} m_s > \alpha$. This leaves us a contradiction, since we are forced to activate strictly more than α arms for this $\mathbf{m}$. Hence the non-degeneracy is necessary for the existence of a locally linear policy. □

5. Improvements for finite values of N

In the previous section, we constructed a family of policies that are all asymptotically optimal as N converges to infinity. In this section, we discuss two directions that can be used to improve the performance for small values of N. The first one is to use the Lagrangian-optimal index of Brown and Smith [3] – that we call simply the LP indices. The second one is a new policy that we call the LP update policy. We will compare their performance in the numerical section.

5.1. The LP indices

The water-filling policy constructed in the previous section is asymptotically optimal regardless of the order used within the sets $\mathcal{S}^+(t)$ and $\mathcal{S}^-(t)$, and it is possible to use a default priority order. This approach is for instance used Zhang and Frazier [18], as well as in Definition 4.4 of Verloop [13] for the infinite horizon problem. Note that as mentioned in Section 8.1 of Verloop [13], how to set priority ordering within $\mathcal{S}^+$ and $\mathcal{S}^-$ is left open in that paper. In this section, we define the notion of LP indices, that can serve as a tie-breaking rule among $\mathcal{S}^+$ and $\mathcal{S}^-$. Our later numerical experiments suggest that tie solving in $\mathcal{S}^+$ and $\mathcal{S}^-$ has a clear influence on the performance of the policy and that the LP-indices perform very well.

Consider the linear program (4). By strong duality, there exist Lagrange multipliers $\gamma^*_0, \ldots, \gamma^*_{T-1}$ corresponding to the constraints (4c), such that $\{\mathbf{y}^*(t)\}_{0\leq t\leq T-1}$ is also an optimal solution of the following problem:

$$\max_{\mathbf{y}\geq\mathbf{0}} \quad \sum_{t=0}^{T-1}\sum_{s,a}(R^a_s - a\gamma^*_t)y_{s,a}(t) \tag{33a}$$

$$\text{s.t.} \quad y_{s,0}(t+1) + y_{s,1}(t+1) = \sum_{s',a} y_{s',a}(t)P^a_{s's} \qquad \forall s,t, \tag{33b}$$

$$y_{s,0}(0) + y_{s,1}(0) = m_s(0) \qquad \forall s. \tag{33c}$$

The above linear program (33) can be cast into a MDP X with horizon T, state space $\mathcal{S}$ and action space $\{0,1\}$. The reward in state $s \in \mathcal{S}$ under action $a \in \{0,1\}$ is $\widetilde{R_s^a} := R_s^a - a\gamma_t^*$. The transition probabilities are $\mathbb{P}\big(X(t+1)=y \,\big|\, X(t)=x, \text{action}=a\big) = P_{xy}^a$. The initial condition is $X(0) \sim \mathbf{m}(0)$, by interpreting $\mathbf{m}(0)$ as a probability vector. The theory of stochastic dynamic programming Puterman [12] shows that there exists an optimal policy which is Markovian.

Let $Q_{s,a}(t)$ be the Q-values of this policy. We define the LP-indices as

$$I_s(t) := Q_{s,1}(t) - Q_{s,0}(t). \tag{34}$$

The *LP-index policy* is then defined as the water-filling policy, by using the values $I_s(t)$ in (34) as a priority score to rank states within $\mathcal{S}^+(t)$, $\mathcal{S}^-(t)$ and $\mathcal{S}^0(t)$ for the water-filling procedure, at each decision epoch t. Note that these indices coincide with the "optimal Lagrangian index" in Brown and Smith [3]. The LP-indices will also be defined in the infinite horizon case later in Section 6.

The next result justifies the notion of LP-indices. In particular, it implies that when the problem is rankable, the LP-indices can be used to construct directly an asymptotically optimal time-dependent priority policy by ordering the states via decreasing LP indices. Note that when the problem is not rankable, it is really important to use the correct tie-breaking rule among the states such that $I_s(t) = 0$ (for instance by using water-filling). Using another tie-breaking rule is in general sub-optimal, see e.g. Brown and Smith [3].

Lemma 9 *The LP-indices are such that $I_s(t) \geq 0$ for all $s \in \mathcal{S}^+(t)$, $I_s(t) \leq 0$ for all $s \in \mathcal{S}^-(t)$ and $I_s(t) = 0$ for all $s \in \mathcal{S}^0(t)$.*

Proof. Let ψ^* be an optimal Markovian stationary policy of (33) formulated as a Markov decision process X, so that $\psi_{s,a}^*(t)$ is the probability of choosing action a if $X(t) = s$. Our previous discussion shows that

$$y_{s,a}^*(t) = \mathbb{P}^{\psi^*}\big(X(t)=s\big) \cdot \psi_{s,a}^*(t).$$

Hence

- $s \in \mathcal{S}^+(t) \Rightarrow y_{s,0}^*(t) = 0 \Rightarrow \psi_{s,1}^*(t) = 1$ and $\psi_{s,0}^*(t) = 0 \Rightarrow I_s(t) > 0$;
- $s \in \mathcal{S}^-(t) \Rightarrow y_{s,1}^*(t) = 0 \Rightarrow \psi_{s,1}^*(t) = 0$ and $\psi_{s,0}^*(t) = 1 \Rightarrow I_s(t) < 0$;
- $s \in \mathcal{S}^0(t) \Rightarrow 0 < y_{s,0}^*(t) < 1$ and $0 < y_{s,1}^*(t) < 1 \Rightarrow 0 < \psi_{s,1}^*(t) < 1$ and $0 < \psi_{s,0}^*(t) < 1 \Rightarrow I_s(t) = 0$.

□

5.2. The LP-update policy

One potential drawback of the Lipschitz continuous policies with their $\mathcal{O}(\frac{1}{\sqrt{N}})$ convergence rate proven in Theorem 2 is that, the constant $C > 0$ in inequality (22) grows exponentially with the horizon T. Hence, for large T we may need N to be extremely large in order to keep $C/\sqrt{N}$ small. Intuitively, a LP-compatible policy is such that $\pi_t(\cdot)$ satisfies $\pi_t(\mathbf{m}^*(t)) = \mathbf{y}^*(t)$. Hence, if the stochastic vector $\mathbf{M}^{(N)}(t)$ is close to $\mathbf{m}^*(t)$, the decision vector $\mathbf{Y}(t) = \pi_t(\mathbf{M}^{(N)}(t))$ recommended by $\pi_t(\cdot)$ should be close to optimal. Yet, if $\mathbf{M}^{(N)}(t)$ is far from $\mathbf{m}^*(t)$ (this could happen, albeit with a small probability), the decision vector recommended by $\pi_t(\cdot)$ could be far from optimal. To overcome this problem, in this section we introduce a new policy called the *LP-update policy*, that recomputes a new LP-compatible policy periodically. It works as follows:

At decision epoch t, we solve a relaxed LP (4) with parameters $\{\mathbf{M}^{(N)}(t), T-t\}$, where the initial state is $\mathbf{M}^{(N)}(t)$ (as we observe at time t), and the time horizon is $T-t$. We choose the decision vector at time t as given by this LP solution. The *LP-update* policy is to apply this procedure at every decision epoch $0 \leq t \leq T-1$.

Note that solving the LP problem (4) at each time steps can be quite costly. Hence, as a compromise one might do update only from time to time, and apply the water-filling policy obtained from the most recent solution of LP between two updates. For the sake of simplicity, we discuss in the following

the LP-update policy that updates at every decision epoch. The following result demonstrates that the LP-update policy is asymptotically optimal with rate $\mathcal{O}(\frac{1}{\sqrt{N}})$, as any LP-compatible Lipschitz continuous policy does.

Theorem 10 *Let the LP-update policy be defined as above, and denote by* $V^{(N)}_{\mathrm{LP-update}}(\mathbf{m}(0),T)$ *the value of LP-update policy on a RB with parameter set* $\{\mathbf{m}(0),T\}$. *Then there exists a constant* $C'>0$ *independent of* N *such that*

$$\left|V_{\mathrm{rel}}(\mathbf{m}(0),T)-V^{(N)}_{\mathrm{LP-update}}(\mathbf{m}(0),T)\right|\leq\frac{C'}{\sqrt{N}}.$$

Consequently the LP-update policy is asymptotically optimal with rate $\mathcal{O}(\frac{1}{\sqrt{N}})$.

Proof. Denote by $\mathbf{y}^{t*}$ the solution of the LP (4) with parameter set $\left\{\mathbf{M}^{(N)}(t),T-t\right\}$ at decision epoch t. Write similarly $\mathbf{m}^{t*}$ where $m^{t*}_s(t')=y^{t*}_{s,0}(t')+y^{t*}_{s,1}(t')$ for $t\leq t'\leq T-1$ and $s\in\mathcal{S}$. Bellman's principle of optimality gives

$$V_{\mathrm{rel}}(\mathbf{M}^{(N)}(t),T-t)=\sum_{s,a}y^{t*}_{s,a}(t)R^a_s+V_{\mathrm{rel}}(\mathbf{m}^{t*}(t+1),T-(t+1)),\tag{35}$$

and the value of the LP-update policy on parameter set $\left\{\mathbf{M}^{(N)}(t),T-t\right\}$ is

$$V^{(N)}_{\mathrm{LP-update}}(\mathbf{M}^{(N)}(t),T-t)=\sum_{s,a}y^{t*}_{s,a}(t)R^a_s+\mathbb{E}\left[V^{(N)}_{\mathrm{LP-update}}(\mathbf{M}^{(N)}(t+1),T-(t+1))\right].\tag{36}$$

Denote by $Z(t):=V^{(N)}_{\mathrm{LP-update}}(\mathbf{M}^{(N)}(t),T-t)-V_{\mathrm{rel}}(\mathbf{M}^{(N)}(t),T-t)$ the difference between (35) and (36), one has $Z(T)=0$ and for all $t\in\{1\dots T-1\}$:

$$\begin{aligned}\mathbb{E}\left[Z(t)\right]&=\mathbb{E}\left[V^{(N)}_{\mathrm{LP-update}}(\mathbf{M}^{(N)}(t+1),T-(t+1))-V_{\mathrm{rel}}(\mathbf{m}^{t*}(t+1),T-(t+1))\right]\\&=\mathbb{E}\left[Z(t+1)\right]+\mathbb{E}\left[V_{\mathrm{rel}}(\mathbf{M}^{(N)}(t+1),T-t+1)-V_{\mathrm{rel}}(\mathbf{m}^{t*}(t+1),T-(t+1))\right].\end{aligned}$$

From the general theory of linear programming (see for instance Section 5.6.2 of Boyd and Vandenberghe [2]), the function $V_{\mathrm{rel}}(\,\cdot\,,t):\Delta^d\to\mathbb{R}$ is Lipschitz continuous with a constant denoted K_t. Let $K:=\max_t K_t$. We have:

$$\left|V^{(N)}_{\mathrm{LP-update}}(\mathbf{m}(0),T)-V_{\mathrm{rel}}(\mathbf{m}(0),T)\right|=\mathbb{E}\left[Z(0)\right]\leq\sum_{t=0}^{T-1}\mathbb{E}\left[K_t\left\|\mathbf{M}^{(N)}(t+1)-\mathbf{m}^{t*}(t+1)\right\|_1\right].$$

By Lemma 1 we have

$$\begin{aligned}\mathbf{M}^{(N)}(t+1)&=\phi(\mathbf{Y}^{(N)}(t))+\mathbf{E}^{(N)}(t),\\\mathbf{m}^{t*}(t+1)&=\phi(\mathbf{y}^{t*}(t)).\end{aligned}$$

Moreover, by construction $\left\|\mathbf{Y}^{(N)}(t)-\mathbf{y}^{t*}(t)\right\|_1\leq 2d/N$ where the term $2d/N$ is caused by randomized rounding and is of order $\mathcal{O}(\frac{1}{N})$. Recall also that $\phi(\cdot)$ is a Lipschitz function with Lipschitz constant ℓ. The dominating error hence comes from $\mathbb{E}\left[\mathbf{E}^{(N)}(t)\mid\mathbf{Y}^{(N)}(t)\right]\leq c_\phi/\sqrt{N}$, where $c_\phi>0$ is a constant independent of T and N. We therefore can bound:

$$\left|V^{(N)}_{\mathrm{LP-update}}(\mathbf{m}(0),T)-V_{\mathrm{rel}}(\mathbf{m}(0),T)\right|\leq\frac{2KTc_\phi}{\sqrt{N}}.\tag{37}$$

Consequently we may choose $C':=2KTc_\phi$ and our proof is complete. □

Note how by applying the idea of updates we have reduced the growth rate of $(\ell L)^T$ in (22) into a rate of $2KTc_\phi$ in (37), where K is an upper-bound on the Lipschitz constant $\{K_t\}_{t\geq 0}$ of the sequence of functions $\{V_{\text{rel}}(\,\cdot\,,t)\}_{t\geq 0}$. Numerical evidence suggests that the sequence $\{K_t\}_{t\geq 0}$ is in general bounded by a constant independent of T. If this is true, then the constant C' of (37) grows linearly with time, which is much smaller than the exponential growth of the one in (22). This suggests that the LP-update policy should perform better than its non-update counterpart. It is therefore an interesting question to ask whether the LP-update policy becomes optimal exponentially fast on non-degenerate RB models. We leave this as a future research topic. We discuss the comparison between the two approaches in more details in our numerical experiments.

6. Infinite horizon case

In this section we study the discrete time Markovian *infinite horizon RB* model, which is defined with the parameters $\left\{(\mathbf{P}^0,\mathbf{P}^1,\mathbf{R}^0,\mathbf{R}^1);\alpha,N\right\}$. Since the analysis follows the same line as in the finite horizon case, we shall be brief and highlight mainly the differences. In particular, we will discuss the uniform global attractor property in Theorem 12, and compare the LP indices with the classical Whittle indices in Proposition 13.

6.1. Infinite-horizon LP relaxation and non-degenerate RB

The analogue of (2) in the infinite horizon case is

$$V_{\text{opt}}^{(N)}(\infty) = \max_{\pi\in\Pi} \quad \lim_{T\to\infty}\frac{1}{T}\mathbb{E}_\pi\Big[\sum_{t=0}^{T-1} Y_{s,a}^{(N)}(t)R_s^a\Big] \tag{38a}$$

$$\text{s.t.} \quad \sum_s Y_{s,1}^{(N)}(t) = \begin{cases} (\lfloor \alpha N\rfloor+1)/N, & \text{with probability } \{\alpha N\} \\ \lfloor \alpha N\rfloor/N, & \text{otherwise.}\end{cases} \quad \forall t, \tag{38b}$$

$$\text{Arms follow the Markovian evolution (1)} \tag{38c}$$

Here Π is the set of Markovian stationary policies. To ease the discussion, we assume that the infinite horizon RB is such that when one arm considered as a MDP is *unichain*, which means that under any policy in consideration, the corresponding Markov chain contains a single recurrent class.

We next relax the constraints in (38b) into the following single constraint

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_s \mathbb{E}_\pi\left[Y_{s,1}^{(N)}(t)\right] = \alpha, \tag{39}$$

and define variables $y_{s,a}$ for $s\in\mathcal{S}$, $a\in\{0,1\}$ as

$$y_{s,a} := \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_\pi\left[Y_{s,a}^{(N)}(t)\right].$$

We then obtain the following linear program as the analogue of (4):

$$V_{\text{rel}}(\infty) = \max_{\mathbf{y}\geq\mathbf{0}} \quad \sum_{s,a} R_s^a y_{s,a} \tag{40a}$$

$$\text{s.t.} \quad \sum_s y_{s,1} = \alpha, \tag{40b}$$

$$y_{s,0}+y_{s,1} = \sum_{s',a} y_{s',a}P_{s's}^a \qquad \forall s, \tag{40c}$$

$$\sum_{s,a} y_{s,a} = 1. \tag{40d}$$

Denote by $\mathbf{y}^*$ an optimal solution of (40). Similarly to the finite-horizon case, we define the following four sets, which form a partition of $\mathcal{S}$.

$$\begin{aligned}\mathcal{S}^+ &:= \left\{s \in \mathcal{S} \mid y^*_{s,1} > 0 \text{ and } y^*_{s,0} = 0\right\}\\ \mathcal{S}^0 &:= \left\{s \in \mathcal{S} \mid y^*_{s,1} > 0 \text{ and } y^*_{s,0} > 0\right\}\\ \mathcal{S}^- &:= \left\{s \in \mathcal{S} \mid y^*_{s,1} = 0 \text{ and } y^*_{s,0} > 0\right\}\\ \mathcal{S}^\emptyset &:= \left\{s \in \mathcal{S} \mid y^*_{s,1} = 0 \text{ and } y^*_{s,0} = 0\right\}.\end{aligned}$$

Compared to the sets before, these sets do not dependent on t. Note that the unichain assumption implies that $\mathcal{S}^\emptyset$ is empty.

As before, we say that an infinite RB is *non-degenerate* if there exists a solution $\mathbf{y}^*$ of (40) such that $|\mathcal{S}^0| \geq 1$, and is *rankable* if there exists a solution $\mathbf{y}^*$ with $|\mathcal{S}^0| \leq 1$. Similar to Equation (24), we prove that

Proposition 11 *For any infinite horizon RB, the optimization problem (40) has an optimal solution $\mathbf{y}^*$ satisfying $|\mathcal{S}^0| \leq 1$.*

The proof of this claim is similar to its finite horizon counter-part around Equation (24), except that this time we apply Theorem 4.4 of Altman [1], which is the same type of result stated for constrained MDP using the expected average cost criteria. Consequently, any infinite horizon RB is rankable.

6.2. Asymptotic optimality of LP-priority policy with exponential rate

Following Definition 4.4 of Verloop [13], we define the set of LP-priorities as $\Sigma := \bigcup_{\mathbf{y}^*} \Sigma(\mathbf{y}^*)$, where $\Sigma(\mathbf{y}^*)$ is the set of permutations $\sigma = \sigma_1 \dots \sigma_d$ of the d states such that any state in $\mathcal{S}^+$ appears before any state in $\mathcal{S}^0$, and any state in $\mathcal{S}^0$ appears before any state in $\mathcal{S}^-$. We call the corresponding policy a *LP-priority policy*.

By Proposition 11, there exists $\mathbf{y}^*$ such that $|\mathcal{S}^0| \leq 1$. We shall choose this $\mathbf{y}^*$ and fix $\sigma^* \in \Sigma(\mathbf{y}^*)$. Denote by $V^{(N)}_{\mathrm{LP}}(\infty)$ the value of the corresponding LP-priority policy. Clearly we have $V^{(N)}_{\mathrm{LP}}(\infty) \leq V^{(N)}_{\mathrm{opt}}(\infty) \leq V_{\mathrm{rel}}(\infty)$. We wish to show the convergence of $V^{(N)}_{\mathrm{LP}}(\infty)$ to $V_{\mathrm{rel}}(\infty)$ as N goes to infinity, and provide similar rates of convergence. However, in the infinite horizon case, an additional important assumption on the model, which does not appear in the finite horizon case, must be assumed in order for the convergence to hold, for which we discuss next.

As a LP-priority policy is a strict priority policy, one can show that the following map (the analogue of (6))

$$\Psi : \mathbf{M}^{(N)}(t) \xrightarrow[\text{policy}]{\text{LP priority}} \mathbf{Y}(t) = \mathbf{Y}^{(N)}(t) \xrightarrow[\text{Markovian transition (1)}]{\text{each arm follows the}} \phi(\mathbf{Y}^{(N)}(t)) \tag{41}$$

is a piecewise affine and continuous function from Δ^d to Δ^d, with d affine pieces (see Lemma 3.1 of Gast et al. [5]). Define the t-th iteration of maps $\Psi_{t\geq 0}(\cdot)$ as $\Psi_0(\mathbf{m}) = \mathbf{m}$, $\Psi_{t+1}(\mathbf{m}) = \Psi\big(\Psi_t(\mathbf{m})\big)$.

(Uniform Global Attractor Property (UGAP)) The vector $\mathbf{m}^* \in \Delta^d$ given by the optimal solution of (40) is a uniform global attractor of $\Psi_{t\geq 0}(\cdot)$, i.e. for all $\epsilon > 0$, there exists $T(\epsilon) > 0$ such that for all $t \geq T(\epsilon)$ and all $\mathbf{m} \in \Delta^d$, one has $\|\Psi_t(\mathbf{m}) - \mathbf{m}^*\|_1 \leq \epsilon$.

The next theorem is a refinement of the asymptotic optimality result in Verloop [13] (Proposition 4.14), proving the exponential convergence rate under the additional non-degeneracy condition on the infinite horizon RB.

Theorem 12 *Consider an infinite horizon RB which is unichain and satisfies the UGAP. Then the LP-priority policy induced by σ^* is asymptotically optimal. Moreover, if the RB is non-degenerate, then the convergence rate can be shown to be exponential.*

Proof. The proof of this theorem is similar to Theorem 3.2 of Gast et al. [5]. We briefly comment on the necessary conditions for the two theorems. Note that the latter being proved for the Whittle's index policy, as a preliminary, the infinite horizon RB needs to be indexable, whereas we do not need any assumption on indexability for our result here. The non-singularity condition in Gast et al. [5] plays the same role as the non-degenerate condition here, and as the example of Remark 3.1 in Gast et al. [5] shows, in general this condition is necessary to ensure the exponential rate. However, unlike the non-degenerate condition in Theorem 7 for the finite horizon case, this condition in infinite horizon is almost always satisfied. On the other hand, as discussed in length in Section 6 of Verloop [13], the UGAP is a necessary and tricky assumption that poses some technical challenges. For instance, it can often be verified only numerically, and it is an open question as how to design (non-priority) policies that are asymptotically optimal without this property. Note that the proof for the $\mathcal{O}(\frac{1}{\sqrt{N}})$ rate convergence in the degenerate case is also an open question. □

6.3. The infinite-horizon LP indices and the Whittle indices

Similar to the LP indices discussed in Section 5.1 for the finite horizon RB, we can also define those indices in the infinite horizon case as follows: By strong duality, there exists Lagrange multiplier $\gamma^* \in \mathbb{R}$ such that $\mathbf{y}^*$ is also an optimal solution to the following linear program:

$$\max_{\mathbf{y} \geq \mathbf{0}} \quad \sum_{s,a} (R_s^a - a\gamma^*) y_{s,a} \tag{42a}$$

$$\text{s.t.} \quad y_{s,0} + y_{s,1} = \sum_{s',a} y_{s',a} P_{s's}^a \qquad \forall s, \tag{42b}$$

$$\sum_{s,a} y_{s,a} = 1 \tag{42c}$$

We again transform the problem (42) into a MDP, with the modified rewards $\widetilde{R_s^a} := R_s^a - a\gamma^*$. The value function V_s^* for state s satisfies the Bellman equation

$$\begin{aligned} g(\gamma^*) + V_s^* &= \max_a \left\{ \widetilde{R_s^a} + \sum_{s'} V_{s'}^* \cdot P_{ss'}^a \right\} \\ &= \max \left\{ R_s^0 + \sum_{s'} V_{s'}^* \cdot P_{ss'}^0,\ R_s^1 - \gamma^* + \sum_{s'} V_{s'}^* \cdot P_{ss'}^1 \right\} \\ &= \max \left\{ Q_s^0,\ Q_s^1 \right\}, \end{aligned}$$

where $g(\gamma^*)$ is the optimal value of the linear program (42). The LP indices for the infinite horizon RB is then defined as $I_s := Q_s^1 - Q_s^0$ for state s. The *LP-index policy* is the strict priority policy by using the values I_s as a priority order to rank states within $\mathcal{S}^+$, $\mathcal{S}^-$ and $\mathcal{S}^0$ at each decision epoch.

We next recall the classical definition of Whittle indices and the concept of indexability for an infinite horizon RB (see for instance Weber and Weiss [14] and Niño-Mora [9] for a general discussion on this topic). For each value $\gamma \in \mathbb{R}$, the value function $V_s(\gamma)$ for state s satisfies a similar Bellman equation

$$g(\gamma) + V_s(\gamma) = \max_a \left\{ R_s^a - a\gamma + \sum_{s'} V_{s'}(\gamma) \cdot P_{ss'}^a \right\}. \tag{43}$$

Define

$$\mathcal{S}(\gamma) := \left\{ s \in \mathcal{S} \,\middle|\, R_s^1 - \gamma + \sum_{s'} V_{s'}(\gamma) \cdot P_{ss'}^1 > R_s^0 + \sum_{s'} V_{s'}(\gamma) \cdot P_{ss'}^0 \right\}.$$

In other words, $\mathcal{S}(\gamma)$ is the set of states for which the arg max in (43) is $a = 1$. The infinite horizon RB is *indexable* if $\mathcal{S}(\gamma)$ expands monotonically from $\emptyset$ to the full set $\mathcal{S}$ when γ is decreased from $+\infty$ to $-\infty$. The Whittle index γ_s for state s is defined to be the supremum value of γ for which s

belongs to $\mathcal{S}(\gamma)$: $\gamma_s := \sup\{\gamma \in \mathbb{R} \mid s \in \mathcal{S}(\gamma)\}$. The *Whittle index policy* is the strict priority policy by using the values γ_s as a priority score to rank states within $\mathcal{S}^+$, $\mathcal{S}^-$ and $\mathcal{S}^0$ at each decision epoch. The next result shows that both the LP-index policy and the Whittle index policy are LP-priority policies.

Proposition 13 *Assume that the infinite horizon RB is unichain, so that $\mathcal{S}^\emptyset = \emptyset$. Then*

1. *$s \in \mathcal{S}^+ \Rightarrow I_s > 0$; $s \in \mathcal{S}^- \Rightarrow I_s < 0$; $s \in \mathcal{S}^0 \Rightarrow I_s = 0$.*
2. *If we assume furthermore that the infinite horizon RB is indexable in Whittle's sense, then their Whittle indices γ_s satisfy: $s \in \mathcal{S}^+ \Rightarrow \gamma_s > \gamma^*$; $s \in \mathcal{S}^- \Rightarrow \gamma_s < \gamma^*$; $s \in \mathcal{S}^0 \Rightarrow \gamma_s = \gamma^*$.*

Proof.

1. The proof of this claim is analogue to Lemma 9.
2. We first show that for any state $s \in \mathcal{S}^0$ (if there are any), its Whittle index γ_s is exactly γ^*, the Lagrange multiplier in (42). Indeed, by definition of indexability, for any $\gamma > \gamma_s$, one has $s \notin \mathcal{S}(\gamma)$; and for any $\gamma < \gamma_s$, $s \in \mathcal{S}(\gamma)$. So γ_s is the unique value of γ that satisfies the equality

$$R_s^1 - \gamma + \sum_{s'} V_{s'}(\gamma) \cdot P_{ss'}^1 = R_s^0 + \sum_{s'} V_{s'}(\gamma) \cdot P_{ss'}^0.$$

On the other hand, by item 2 of Proposition 13, the states in $\mathcal{S}^0$ are the states with null LP index, so the above equality are satisfied with $\gamma = \gamma^*$. Consequently the Whittle index γ_s for $s \in \mathcal{S}^0$ is γ^*. The other two implications then follow similarly. □

7. Numerical experiments

In this numerical part, we first demonstrate that tie-solving within $\mathcal{S}^+$ and $\mathcal{S}^-$ for the Lipschitz continuous policies using water-filling is important in Section 7.1. We next show the advantage of the LP-update policy to the LP-index policy on the applicant screening problem in Section 7.2, a model proposed in Brown and Smith [3].

7.1. Tie-solving within $\mathcal{S}^+$ and $\mathcal{S}^-$

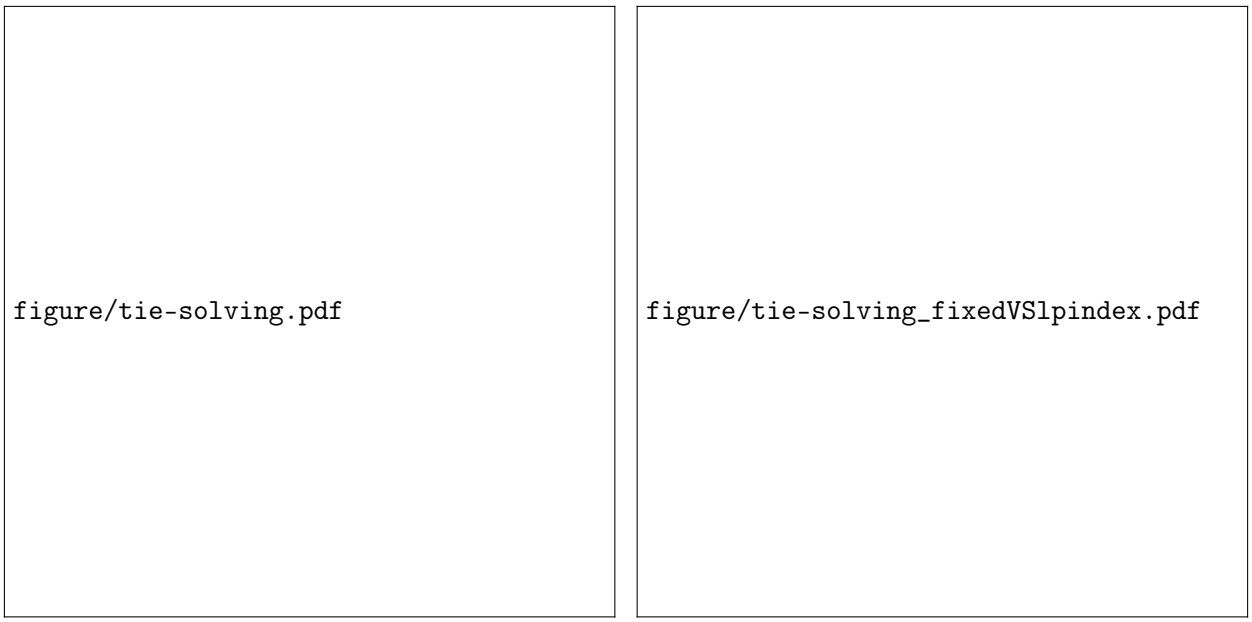
The water-filling policy defined in Section 4.2 is not uniquely defined as it depends on the tie-breaking rule within $\mathcal{S}^+$ and $\mathcal{S}^-$. In Figure 1, we compare the two tie-breaking rules:

- LP-index: Give priority to the highest LP-index first, defined in Section 5.1;
- Random tie-solving: Ties within $\mathcal{S}^+$ and $\mathcal{S}^-$ are solved according to a random priority order that is drawn at the beginning of each simulation. The reported number for this policy is the average among 100 priority orders.

We emphasize that these two policies are LP-compatible policies: to apply them, we first solve the LP to define $\mathcal{S}^+$ and $\mathcal{S}^-$ and apply a water-filling policy. The above tie-breaking rules are only used within $\mathcal{S}^+$ and $\mathcal{S}^-$. This implies that all policies are therefore asymptotically optimal.

In each case, we compute the average *score* of a policy on 100 randomly sampled models of dimension $d = 10$ and arm population $N \in \{10 \dots 50\}$. To generate each model, we sample the matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ as independent uniformly distributed probability matrices and the reward vectors as uniform between 0 and 1. The score is defined as follows (for ease of notation, we omit all dependence on $(\mathbf{m}(0), t)$ in this section). For a given RB, recall that V_{rel} is the value of the linear program (4) and let us denote by $V_{\text{rel}-\min}$ the value of the same linear program but where the maximization is replaced by a minimization. The value of a policy π is V_π^N. We define the score of the policy π as:

$$\text{score}_\pi^N = \frac{V_\pi^N - V_{\text{rel}-\min}}{V_{\text{rel}} - V_{\text{rel}-\min}}. \tag{44}$$



(a) Average score as a function of N (b) LP index vs arbitrary tie-solving ($N = 20$)

Figure 1 **Performance of the different tie-solving among $\mathcal{S}^+$ and $\mathcal{S}^-$: LP indices, and fixed priorities. We report the normalized score (in %) as a function of the number of arms. All policies are asymptotically equivalent but the LP-index policy performs better for all finite values of N.**

The score is a number between 0 and 1 (higher being better). Theorem 7 shows that, any water-filling policy is asymptotically optimal, regardless of the tie-breaking used within $\mathcal{S}^+$ or $\mathcal{S}^-$, *i.e.* $\lim_{N\to\infty} \text{score}_\pi^N = 1$.

Figure 1 shows that the choice of tie-solving within $\mathcal{S}^+$ and $\mathcal{S}^-$ has a significant influence on the performance of the policies. On the left figure, we plot the average score over 100 models for the LP-index policy and for 5 random orders. This figure shows that, on average, the LP-index performs much better than a random tie-solving. In the right figure, we fix $N = 20$ and for the same 100 models and 5 tie-solving rules, we plot the average score of the LP index as a function of the average score of each of the fixed tie-solving rules (this makes 500 points in total). This figure shows that the LP-index is almost always the best tie-solving rules: More precisely, among the 500 pairs of scores considered, we observe only three points that suggest that the LP-index tie-breaking rule could be beaten, and in each case the gain of this fixed order policy is much smaller than the confidence interval.

7.2. Case study: applicant screening problem

We discuss in this section the applicant screening problem proposed in Section 6.2.2 of Brown and Smith [3], and show that the LP-update policy outperforms the LP-index policy on this problem. Consider a group of N applicants applying for a job. The decision maker's goal is to hire the best possible βN applicants. Each applicant n has an unknown quality level $p_n \in [0, 1]$. At each decision epoch t, the decision maker interviews αN applicants and receives, for each interviewed candidate, a signal $d_n(t) \in \{0, 1\}$ that is distributed according to a Bernoulli distribution of parameter p_n. All variables $d_n(t)$ are supposed to be independent (given p_n).

This problem can be seen as a RB with N arms by considering a Bayesian model in which we assume that each p_n is random and distributed uniformly between 0 and 1. Each applicant (arm) is modeled by a MDP. The state s_n of this applicant is $s_n = (a_n, b_n)$ and indicates that the posterior distribution of p_n given previous observation is a beta distribution of parameters (a_n, b_b): at time 0,

$a_n = b_n = 1$. Afterwards, s_n are updated using Bayes' rule to $(a_n + d_n, b_n + 1 - d_n)$ when interviewed. An applicant's state does not change when not interviewed. The rewards are set to zero during the first $T-1$ interview periods. In the final period T, the decision maker admits βN applicants. The reward for admitting the applicant n is p_n. Note that if p_n is uniformly distributed, then $\mathbb{E}[p_n \mid s_n] = a_n/(a_n + b_n)$. The reward for those not admitted is zero.

In our numerical study, we choose the same parameters as those used in Figure 4 of Brown and Smith [3], where $\alpha = \beta = 0.25$, $T = 5$. We compute the LP-policies by assuming that the initial state of all applicants is $(1,1)$ and consider two cases:

- **Correct prior** – In the left-panel of Figure 2, the p_n are generated uniformly between 0 and 1.
- **Wrong prior** – On the right-panel of Figure 2, the p_n are generated using a distribution $beta(3,1)$, while the selection algorithm is constructed from a LP-relaxation that assumes that p_n is uniformly distributed on $[0,1]$.

The first case fits into the framework of our paper, and in particular implies the asymptotic optimality. The second case does not fall into our framework because the transition matrices that we use to construct the policies are not the correct ones. This second case corresponds to a decision maker having a wrong prior about the candidates.

As expected, the LP-index policy performance displayed in the left panel reproduces that of the Lagrange policy with optimal tie-breaking shown in Figure 4 of Brown and Smith [3]. For this scenario, Theorem 7 and Theorem 10 can be applied, and both the LP-index and the LP-update policies converge to the LP-relaxed bound. Moreover, the LP-update policy always outperforms the LP-index policy, with an advantage that is more apparent for N in the middle range. This shows the benefit of applying updates, even in this ideal scenario.

The situation is quite different when the prior of the decision maker is wrong (right panel of Figure 2). In this case, the LP-update and the LP-index policies converge to different values, that are both below the LP-relaxed bound. This is reasonable since the assumption on the p's is wrong. Here, the LP-update policy outperforms the LP-index policy by a large margin, especially when N is large. This is because by applying updates in this situation helps to correct the error due to the wrong assumption on the initial p value of each applicant. This is yet another advantage of the LP-update policy. We expect such an advantage to hold more generally on any Bayesian RB model.

8. Conclusion and future direction

In this paper we propose a general framework to study LP-based policies for RB. We show that the asymptotic behavior of these policies is closely related to properties of their corresponding deterministic maps. We also illustrate the idea of applying updates and demonstrate its advantage to any previously existing LP-based policies on finite horizon problems. We believe that as long as we can formulate the relaxed problem as a linear program, this update idea can be further applied on Weakly Coupled MDPs (see, e.g. Meuleau et al. [8]), which generalize the RB model in this paper by allowing multiple actions for each arm and multiple resource constraints. We also plan to investigate the infinite horizon problem more closely, by designing asymptotically optimal policies without the uniform global attraction property, which is an open question posed in Verloop [13].

Acknowledgements

This work is supported by the ANR project REFINO (ANR-19-CE23-0015).

References

[1] Altman E (1999) *Constrained Markov Decision Processes* (Chapman and Hall).

[2] Boyd S, Vandenberghe L (2004) *Convex Optimization* (USA: Cambridge University Press), ISBN 0521833787.

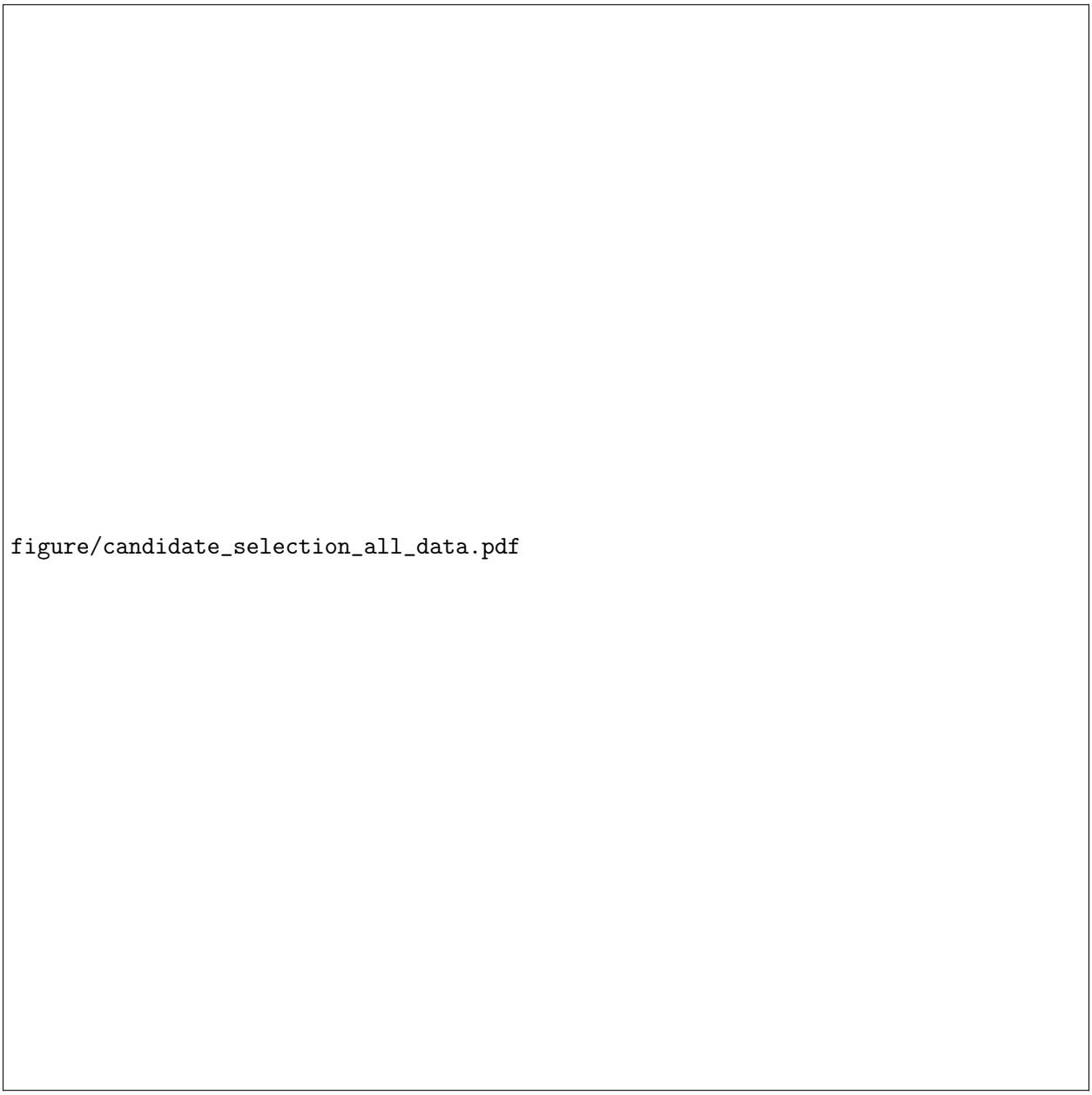


Figure 2 **Performance on applicant screening problem when the decision maker knows the prior distribution of p_n (left panel) or has access to a wrong prior information (right panel).**

[3] Brown DB, Smith JE (2020) Index policies and performance bounds for dynamic selection problems. *Manag. Sci.* 66:3029–3050.

[4] Gast N, Gaujal B, Khun K (2022) Computing whittle (and gittins) index in subcubic time. *arXiv preprint arXiv:2203.05207* .

[5] Gast N, Gaujal B, Yan C (2020) Exponential convergence rate for the asymptotic optimality of whittle index policy. *arXiv preprint arXiv:2012.09064* .

[6] Hu W, Frazier P (2017) An asymptotically optimal index policy for finite-horizon restless bandits. *arXiv preprint arXiv:1707.00205* .

[7] Ioannidis S, Yeh E (2016) Adaptive caching networks with optimality guarantees. *CoRR* abs/1604.03175.

[8] Meuleau N, Hauskrecht M, Kim KE, Peshkin L, Kaelbling LP, Dean TL, Boutilier C (1998) Solving very large weakly coupled markov decision processes. *AAAI/IAAI*, 165–172.

[9] Niño-Mora J (2007) Dynamic priority allocation via restless bandit marginal productivity indices. *TOP: An Official Journal of the Spanish Society of Statistics and Operations Research* 15:161–198.

[10] Niño-Mora J (2020) A fast-pivoting algorithm for whittle's restless bandit index. *Mathematics* 8(12), ISSN 2227-7390.

[11] Papadimitriou CH, Tsitsiklis JN (1999) The complexity of optimal queuing network control. *Math. Oper. Res* 293–305.

[12] Puterman ML (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (New York, NY, USA: John Wiley & Sons, Inc.), 1st edition.

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

[14] Weber RR, Weiss G (1990) On an index policy for restless bandits. *Journal of Applied Probability* 27(3):637–648, ISSN 00219002.

[15] Whittle P (1980) Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)* 42(2):143–149, ISSN 00359246.

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

[17] Zayas-Cabán G, Jasin S, Wang G (2017) An asymptotically optimal heuristic for general non-stationary finite-horizon restless multi-armed multi-action bandits. *Ross: Technology & Operations (Topic)* .

[18] Zhang X, Frazier PI (2021) Restless bandits with many arms: Beating the central limit theorem. *arXiv preprint arXiv:2107.11911* .