Close-to-opimal policies for Markovian bandits

Yan Chen

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques Appliquées
Unité de recherche : Institut National de Recherche en Informatique et en Automatique

Politiques quasi-optimales de bandits Markoviens

Close-to-opimal policies for Markovian bandits

Présentée par :

Chen YAN

Direction de thèse :

Bruno GAUJAL DIRECTEUR DE RECHERCHE, Université Grenoble Alpes	Directeur de thèse
Nicolas GAST chercheur, INRIA	Co-directeur de thèse

Rapporteurs :

DAVID ALAN GOLDBERG
Professeur associé, Cornell University
BRUNO SCHERRER
Chargé de recherche HDR, INRIA CENTRE NANCY-GRAND-EST

Thèse soutenue publiquement le **15 décembre 2022**, devant le jury composé de :

BRUNO GAUJAL Directeur de recherche, INRIA CENTRE GRENOBLE-RHONE- ALPES	Directeur de thèse
DAVID ALAN GOLDBERG Professeur associé, Cornell University	Rapporteur
BRUNO SCHERRER Chargé de recherche HDR, INRIA CENTRE NANCY-GRAND-EST	Rapporteur
KIM THANG NGUYEN Professeur des Universités, GRENOBLE INP	Examinateur
JERÔME MALICK Directeur de recherche, CNRS DELEGATION ALPES	Président
BENJAMIN LEGROS Maître de conférences, EC MANAGT NORMANDIE CAEN GROUPE LE HAVRE	Examinateur

Abstract

Bandits are one of the most basic examples of decision-making with uncertainty. A Markovian restless bandit can be seen as the following sequential allocation problem: At each decision epoch, one or several arms are activated (pulled); all arms generate an instantaneous reward that depend on their state and their activation; the state of each arm then changes in a *Markovian* fashion, based on an underlying transition matrix. Both the rewards and the probability matrices are known, and the new state is revealed to the decision maker for its next decision. The word *restless* serves to emphasize the fact that arms that are not activated can also change states, hence generalizes the simpler rested bandits. In principle, the above problem can be solved by dynamic programming, since it is a Markov decision process. The challenge that we face is the *curse of dimension*, since the size of possible states and actions grows exponentially with the number of arms of the bandit. Consequently, the focus is to design policies that solve the dilemma of computational efficiency and close-to-optimal performance.

In this thesis, we construct computationally efficient policies with provable performance bounds, that may differ depending on certain properties of the problem. In Part I , we first investigate the classical Whittle index policy (WIP) on infinite horizon problems, and prove that if it is asymptotically optimal under the global attractor assumption, then almost always it converges to the optimal value exponentially fast. The application of WIP has the additional technical assumption of indexability as a prerequisite, to get around this, we next study the LP-index policy, that is well-defined for any problem, and shares the same exponential speed of convergence as WIP under similar assumptions.

In infinite horizon, we always need the global attractor assumption for asymptotic optimality. In Part II of the thesis, we study the problem under finite horizon, so that this assumption is no-longer a concern. Instead, the LP-compatibility and the non-degeneracy are required for the asymptotic optimality and a faster convergence rate. We construct the finite horizon LP-index policy, as well as the LP-update policy, that amounts to solving new LP-index policies during the evolution of the process. This latter LP-update policy is then generalized to the broader framework of weakly coupled MDPs, together with the generalization of the non-degenerate condition. This condition also allows a more efficient implementation of the LP-update policy, as well as a faster convergence rate, if it is satisfied on the weakly coupled MDPs.

Contents

1

Introduction

The multi-armed bandit is propounded during the Second World War, and soon recognized as so difficult that it quickly became a classic, and a byword for intransigence. In fact, John Gittins had solved the problem by the late sixties, although the fact that he had done so was not generally recognized until the early eighties.

– Peter Whittle

1.1 Context: the Markovian Restless Bandits

Bandits are one of the most basic examples of decision-making with uncertainty. The word "bandit" originates from an old-fashioned name for a lever-operated slot machine in a casino. A multi-armed bandit problem can be seen as the following sequential allocation problem: At each decision epoch, one or several arms are activated (pulled) and some observable rewards are obtained. The goal is to maximize the total reward obtained by a sequence of activations, subject to some resource constraints on the activation budget.

There are at least three fundamental formalizations of the bandit problem depending on the assumed nature of the reward process: stochastic, adversarial, and Markovian. Each bandit model has its own specific playing strategies and uses distinct techniques of analysis. Roughly speaking, the stochastic bandits refer to the situation where the reward of each arm follows some unknown but stationary probability distributions, while in the adversarial scenario these distributions are non-stationary (chosen by some adversary). These two categories of bandit problems (and their generalizations) are mostly studied in the artificial intelligence and online learning literature, where the key challenge is the dilemma of exploration (acquire new knowledge) and exploitation (optimize the decisions based on existing knowledge).

The focus of this thesis is on the third kind of Markovian bandits: Each time, a subset of arms are chosen to be activated; all arms generate an instantaneous reward that depend on their state and their activation; the state of each arm then changes in a *Markovian* fashion, based on an underlying transition matrix. Both the rewards and the

probability matrices are known, and the new state is revealed to the decision maker for its next decision. We immediately realize that the Markovian bandits are Markov decision processes and in theory can always be solved by dynamic programming. The challenge that we face is the *curse of dimension*, since the numbers of possible states and actions grow exponentially with the number of arms of the bandit. Consequently, the goal of research in Markovian bandits is to design policies that solve the dilemma of computational efficiency and close-to-optimal performance.

Historically, the above Markovian multi-armed bandit problem has been solved in the particular *restful case* (non activated arms do not change their states) with one active arm at each decision epoch in Gittins [22] by the Gittins index policy, which is a greedy policy that can be computed efficiently. As we quoted at the beginning of this chapter, this problem used to have the reputation of being notoriously difficult, but once the solution given, the idea becomes quite natural. Below are two motivating examples of restful bandits taken from the book Gittins et al. [21], that can be solved by using the Gittins indices.

Example 1 (Gold Mining) A woman owns N gold mines and one gold-mining machine. Each day she must assign the machine to one of the mines. When the machine is assigned to mine n, there is a probability p_n that it extracts a proportion q_n of the gold left in the mine, and a probability $1 - p_n$ that it extracts no gold and breaks down permanently. To what sequence of mines on successive days should the machine be assigned so as to maximize the expected amount of gold mined before it breaks down?

Example 2 (Object Search) A stationary object is hidden in one of N boxes. The probability that a search of box n finds the object if it is in box n is q_n. The probability that the object is in box n is p_n, and changes by Bayes' theorem as successive boxes are searched. The cost of a single search of box n is c_n. How should the boxes be sequenced for search so as to minimize the expected cost of finding the object?

By applying an interchange argument, it can be shown that an optimal policy to the gold mining example is to allocate the machine to a mine i such that $\frac{p_i q_i x_i}{1-p_i} = \max \frac{p_n q_n x_n}{1-p_n}$, where x_n is the amount of gold remaining in mine n on a particular day. Likewise, an optimal policy for the object search problem is to search box i at each moment so that $\frac{p_i q_i}{c_i} = \max \frac{p_n q_n}{c_n}$. These expressions in the max are actually the Gittins indices for an arm (i.e. a gold mine or a box) of the restful bandits, and the Gittins index policy amounts to activating the one arm having currently the largest Gittins index.

This optimal solution proposed by Gittins, being simple and elegant, is really limited to a very special case of restful bandits. Indeed, it has been argued in Section 3.4 of Gittins et al. [21] that the optimality theorem of the Gittins index policy no longer holds, if any of the following assumptions on the model is violated: (i) The horizon needs to be infinite; (ii) The reward at time t should be discounted by γ^t, with $0 < \gamma < 1$; (iii) There can have only one arm that is subject to activation at each time. The careful readers should have noticed that these three conditions are not all satisfied in an

obvious way for the two previous examples, in order to apply the optimality theorem of Gittins index. Indeed, some extra work is needed to transform the two problems into equivalent restful bandits, so that the optimality theorem becomes applicable.

In response to these limitations, in Whittle [49] generalizes the restful bandit model in two aspects. Firstly, at each decision epoch more than one arm can be activated, and secondly, the arms that are not activated can also change states (hence the name *restless bandits*). For instance, one might imagine that in the gold mining example, the woman possesses $M > 1$ machines instead of just one machine; or in the object search example, the hidden object is non-stationary, so that it can move from one box to another during the search process. Clearly after these generalizations the problem gains a significantly large modeling power, but do we still have an index-type policy as an optimal solution to the problem under this generality?

The answer is unfortunately "no", as in Papadimitriou and Tsitsiklis [42] it has been shown that the restless bandit problem is PSPACE-hard. Nevertheless, this hardness result is not completely discouraging, as we may still hope to find an index-type policy that gives performance close to being optimal for the restless bandits. This is how the Whittle index policy (WIP) comes into the stage. This heuristic policy is introduced in Whittle [49] as a generalization of the Gittins index policy to restless bandits. It is conjectured by Whittle in the same paper that the performance of the Whittle index policy is asymptotically optimal, i.e. $V_{\text{opt}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha) \xrightarrow[N\to\infty]{} 0$, where $V_{\text{opt}}^{(N)}(\alpha)$ (resp. $V_{\text{WIP}}^{(N)}(\alpha)$) is the value of the optimal (resp. Whittle index) policy on an N-armed bandit with activated arms $M = \alpha N$ and $0 < \alpha < 1$. The conjecture is proven two years later in Weber and Weiss [48] to hold true under two additional technical assumptions: the indexability and the global attractor property. This leaves the following important questions unanswered:

(i) In cases where the asymptotic optimality holds, at which rate does the convergence occur? This question is motivated by the observation that numerous applications of WIP in practice performs well even for a small population of arms. Surprisingly, no existing result in the literature up to now has considered the convergence rate problem.

(ii) Do we still have an efficient index-type policy if the restless bandit is non-indexable? In cases where the restless bandit is indexable but the global attractor does not hold, what will happen to the asymptotic behavior of WIP, and what else can we do if WIP is no longer asymptotically optimal?

(iii) Can the index policy and the asymptotic optimality result be established in a similar manner under the finite horizon? What will remain the same and what needs to be modified?

The current thesis aims at giving answers to the questions posed above, and going beyond.

1.2 Contributions

Our contributions can be roughly divided into four parts, where Questions (i), (ii) and (iii) are discussed respectively in Section 1.2.1, 1.2.2 and 1.2.3. A future development that based on the solution to these questions is discussed in Section 1.2.4.

1.2.1 The Exponential Convergence Rate of WIP

It has been observed in numerous applications that WIP gives a very good performance. Among them we can cite wireless scheduling (Aalto et al. [1], Raghunathan et al. [45]), queuing systems (Ansell et al. [4]), crawling optimal content on the web (Avrachenkov and Borkar [7]), load-balancing (Larrnaaga et al. [32]), sensor management (Niño-Mora and Villar [36]), age of information (Hsu [28]), and the list can go on and on. As already suggested by Weiber and Weiss in their asymptotic optimality paper, one reason behind this might be that the convergence rate occurs faster than square root of N given by classical Central Limit Theorem. We give an affirmative and theoretically precise formulation to this claim: if the asymptotic optimality holds for a restless bandit, then almost always we can find two positive constants $b, c > 0$ that are independent of N, such that $V_{\text{opt}}^{(N)} - V_{\text{WIP}}^{(N)} \leq b \cdot e^{-cN}$. In other words, if WIP is asymptotically optimal, then almost always it approaches the optimal policy exponentially fast as the population of arms N grows. This confirms all the previous observations by other researchers on the excellent performance of WIP, by providing a theoretical grounding in favor of this policy, in cases that it is applicable. This part of results is detailed in Chapter 3.

1.2.2 A Unifying Linear Program Approach

In many of the successful applications which have used Whittle indices, the indexability property is established by exploiting the special structure of the restless bandit model. Unfortunately, there is no guarantee of indexability for any restless bandit problem, and in case of a non-indexable problem, the Whittle index is not well defined. This difficulty on the indexability assumption has been solved in the work Verloop [47], by using a linear program approach. More precisely, we can associate to *any* restless bandit problem a linear program, and a large collection of strict priority policies based on the solution to the linear program can be proven to be asymptotically optimal, provided that a global attractor property similar to the requirement for WIP holds true. In particular, WIP belongs to this class of asymptotically optimal policies, if the problem is indexable.

By this means we are saved from the verification of indexability, in exchange we only need to solve a linear program. Similar to the exponential convergence rate of WIP, we show that any of these LP-based policies that is asymptotically optimal actually becomes optimal exponentially fast. It is detailed in Chapter 4. This, however, does not mean that they all perform equally good. Indeed, the collection of asymptotically optimal policies based on solving the linear program is extremely large, and their performances for small values of N are observed to differ. By using the so-called LP-

index to rank states, we define the *LP-index policy* that is a good candidate among the collection of all these asymptotically optimal policies.

The global attractor assumption concerns the deterministic asymptotic behavior of WIP, and more generally is needed for any LP-based policy in infinite horizon. In practice it is almost always verified via numerical means. A counter-example that violates this assumption is given in Weber and Weiss [48], for which the limit behavior of WIP exhibits an attracting cycle, and the policy is asymptotically sub-optimal. The solution to this subtlety is by considering the restless bandit model in a finite horizon $T < \infty$, since under finite horizons we no longer need to worry about the limit behavior of the deterministic dynamical system.

1.2.3 Restless Bandits in Finite Horizon

Restless bandits in finite horizon have been studied previously in Hu and Frazier [29] and Brown and Smith [12], using the LP approach. The main idea is to construct actions at each time $0 \leq t \leq T-1$ that follow as close as possible to an optimal LP solution. A new difficulty arises in finite horizon, as the asymptotically optimal policy may no-longer be strict priorities for all time steps, and extra care needs to be taken for the tie breaking. In Brown and Smith [12] multiple tie-solving rules have been proposed, and their "Lagrange policy with optimal tie-breaking" coincides with our LP-index policy in finite horizon.

Under our more general point of view, the asymptotic optimality and the convergence rate of *any* LP-based policy is related to how close this policy can be constructed to follow an optimal LP solution. This closeness is measured by properties of the policy viewed as a map from the space of configurations of the system to the space of actions. In essence, the policy is asymptotically optimal (resp. with square root convergence rate, resp. with exponentially fast convergence rate) if the function is continuous (resp. Lipschitz continuous, resp. locally linear). Moreover, we show that these requirements on the policy function are more or less the necessary conditions for their corresponding convergence rate. This will be made precise in Chapter 5.

An important topic that (to the best of our knowledge) has remain unnoticed in the literature is the comparison between the LP-based approaches in infinite and finite horizon. In particular, if the LP-based policy in infinite horizon is not asymptotic optimal, by violation of the global attractor property, what does the policy on the corresponding finite horizon problem look like? We observe that under these situations the (finite horizon) deterministic dynamic will most probably suffer from a unstable issue, so even a tiny error is capable to propagate after several time steps into a huge deviation from the optimal trajectory. Consequently, the common LP-based policies give very poor performance on these finite horizon problems, although in theory, they still converge asymptotically to the optimal values.

To overcome this drawback, we propose the so-called LP-update policy, that solves new linear programs regularly during the T time steps, and applies new decisions based on these solutions. A naive implementation of the LP-update policy that solves

new linear programs at every time step is given in Chapter 5. After testing on various randomly generated problems, as well as the applicant screening problem that can be modeled as restless bandits, we show that the LP-update policy always outperforms the LP-index policy, and gives a significant improvement on those problems that the latter performs poorly. However, this comes at the cost of a longer computation time.

1.2.4 Generalization to Weakly Coupled MDPs

Weakly coupled MDPs can be seen as multi-armed bandits with multiple actions and multiple budget constraints, whereas the original restless bandits have only the two active and passive actions, with a single activation budget constraint. Weakly coupled MDPs under finite horizon are studied in Adelman and Mersereau [2], Astaraky and Patrick [5], Dolgov and Durfee [13], Gocgun and Ghate [24], Gocgun and Ghate [25], Meuleau et al. [34], Patrick et al. [43], as well as the two PhD thesis Hawkins [26] and Salemi Parizi [46]. In these works the authors consider the situation when the sub-MDPs are not statistically identical. Under this generality, the results rely on a Lagrange decomposition technique to solve approximately the above problem. Since the focus of this thesis is on asymptotic optimality, we shall concentrate on the special case where each sub-MDP is statistically identical, so that we can provide methods to construct policies that are not only close to being optimal, but can actually been proven theoretically to be asymptotically optimal when the population N goes to infinity.

On another branch, there exists multiple attempts in the literature to generalize the classical restless bandit model that is less general than the weakly coupled MDPs. The finite horizon multi-action single-constraint multi-armed bandit is the subject of Zayas-Cabán et al. [52] and Xiong et al. [50], while the infinite horizon analogue is considered in Hodge and Glazebrook [27] and Niño-Mora [40], with the emphasis on the generalization of Whittle's indexability and defining a similar index. The main difficulty for constructing an index-type policy on weakly coupled MDPs is that the indices, being real-values functions, are defined with respect to a single constraint.

To this end, we propose a more efficient version of the LP-update policy, considered previously in Chapter 5 for the classical restless bandits, that can be generalized in a straightforward manner to the weakly coupled MDPs. This will be detailed in Chapter 6. Our contributions are threefold: (i) we generalize the LP-update policy to weakly coupled MPDs; (ii) We propose a more efficient implementation of the naive LP-update policy considered previously in Chapter 5; (iii) We prove the convergence rate (e.g. square root, exponentially fast) of the LP-update policy under this general framework, together with the necessary conditions for these rates to hold.

1.3 Organization of the Thesis

The rest of the thesis is to make our previous informal discussion precise and rigorous. In this regard, in Chapter 2 we discuss the background, related work and main challenges. In Chapter 3 we prove the exponential convergence rate of WIP under the

non-singular and uniform global attractor assumptions. This is based on our work Gast et al. [17]. In the short Chapter 4 we discuss the LP-approach to infinite horizon problem, and show that a similar exponential convergence rate result can be proven. This completes the result of Verloop [47], and provide a good comparison to its finite horizon counter-part that we discuss next. In Chapter 5 we consider the finite horizon restless bandit problem, this is based on our work Gast et al. [18]. We provide necessary and sufficient conditions for any LP-based policy to be asymptotically optimal (resp. with square root convergence rate, resp. with exponentially fast convergence rate). The LP-index policy is defined therein, and the LP-update policy is also briefly mentioned. The (improved) LP-update policy is discussed in Chapter 6 for the more general weakly coupled Markov decision processes (MDPs). We define in the meantime the rank condition and the non-degenerate property on weakly coupled MDPs, the latter generalized its previous definition on two-action bandits. This chapter is based on our work [working paper]. Several additional numerical experiments are collected together in Chapter 7. Finally, we give a general conclusion of the thesis in Chapter 8.

The structure of the thesis is summarized in Figure 1.1, where Part I and Part II are independent from each other, and can be read separately. For convenience, we provide a table of notations, abbreviations and definitions at the end as Appendix A.

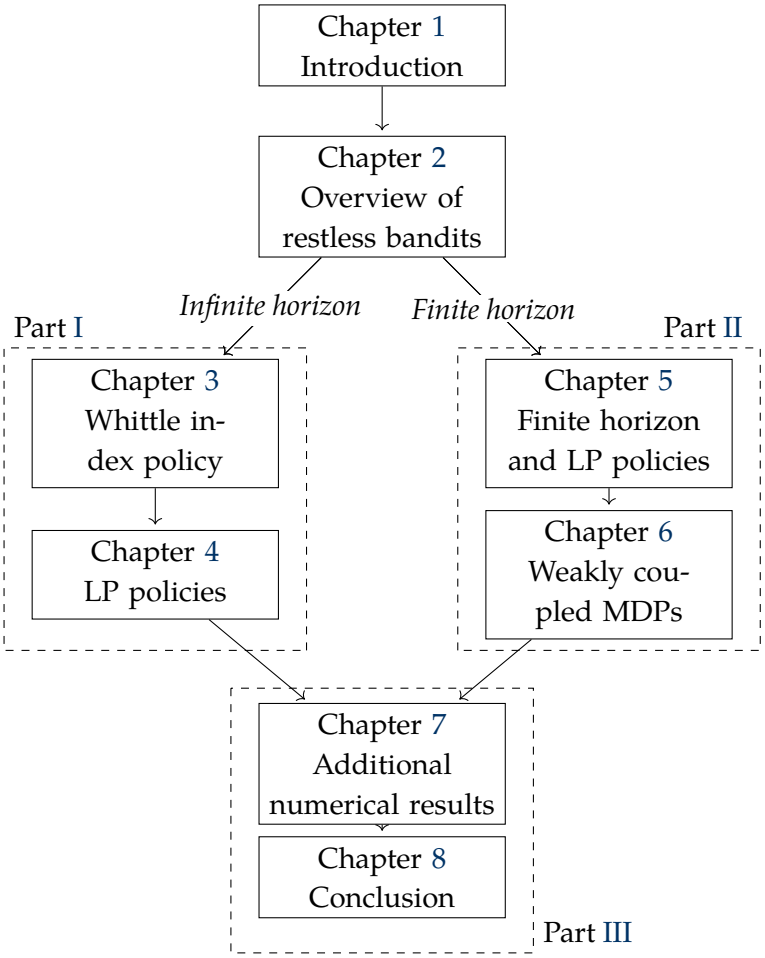


Figure 1.1 – Structure of the Thesis.

2

Background and Related Work

The starting point of the present thesis is the two classic papers Whittle [49] and Weber and Weiss [48], in which the Whittle index policy for restless bandits is defined in the first, and its asymptotic optimality (under additional assumptions) is shown in the later. In this chapter, we review the details of this line of work in Sections 2.1-2.3, which is the foundation of its subsequent developments: the linear program approach (Section 2.4), the restless bandits in finite horizon (Section 2.5), the LP-update policy and its generalization to weakly coupled MDPs (Section 2.6).

> *I can illustrate the mode of propagation of this news, when it began to propagate, by telling of an American friend of mine, a colleague of high repute, who asked an equally well-known colleague 'What would you say if you were told that the multi-armed bandit problem had been solved?' The reply was somewhat in the Johnsonian form: 'Sir, the multi-armed bandit problem is not of such a nature that it can be solved'. My friend then undertook to convince the doubter in a quarter of an hour. This is indeed a feature of John's solution: that, once explained, it carries conviction even before it is proved.*
>
> – Peter Whittle

2.1 The Restless Bandit Model

Formally, the discrete time infinite horizon restless bandit model with parameters $\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha, N\}$ is a Markov decision process defined as follows:

1. The model is composed of N statistically identical arms. Each arm evolves in a finite state space $\mathcal{S} := \{1 \dots d\}$ with an action set $\mathcal{A} = \{0, 1\}$, and the state of arm n at time t is denoted by $S_n(t) \in \{1 \dots d\}$. The state space of the whole process at time t is denoted by $\mathbf{S}(t) = \big(S_1(t), S_2(t), \dots, S_N(t)\big)$.

2. Decisions are taken at times $t \in \mathbb{N}$. At each decision epoch, a decision maker observes $\mathbf{S}(t)$ and chooses αN of the N arms to be activated, with $0 < \alpha < 1$. We set $a_n(t) = 1$ if arm n is activated at time t and $a_n(t) = 0$ otherwise. The action vector at time t is $\mathbf{a}(t) = \big(a_1(t), a_2(t), ..., a_N(t)\big)$. It satisfies $\sum_{n=1}^{N} a_n(t) = \alpha N$.

3. Arm n evolves according to Markovian laws: for all states i, j, action $a \in \mathcal{A}$ and $t \in \mathbb{N}$:
$$\mathbb{P}(S_n(t+1) = j \mid S_n(t) = i, a_n(t) = a) = P_{ij}^{a}. \tag{2.1}$$
Given $\mathbf{a}(t)$ and $\mathbf{S}(t)$, the N arms make their transitions independently.

4. For each arm that is in state i and for which action $a \in \mathcal{A}$ is taken, a reward $R_i^a \in \mathbb{R}$ is earned.

The goal of the decision maker is to compute a decision rule in order to maximize the long-term expected average reward per period. The theory of stochastic dynamic programming (e.g. Puterman [44]) shows that there exists an optimal policy which is stationary and deterministic (i.e. $\mathbf{a}(t)$ can be chosen as a time-independent deterministic function of $\mathbf{S}(t)$). Denote by Π the set of such policies, which are maps from $\mathbf{S}$ to $\mathbf{a}$. The optimization problem of the decision maker can be formalized as

$$V_{\text{opt}}^{(N)}(\alpha) := \max_{\Pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1} \frac{1}{N} \sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)}\Big] \tag{2.2}$$

$$\text{subject to} \sum_{n=1}^{N} a_n(t) = \alpha N, \text{ for all } t \in \mathbb{N}. \tag{2.3}$$

In theory, a dynamic programming approach can be used to solve Equations (2.2)-(2.3), but this approach is computationally intractable, as the numbers of possible states and actions grow exponentially with N. In fact, such problems have been proven to be PSPACE-hard in Papadimitriou and Tsitsiklis [42]. This is in big contrast with the restful bandits allowing only one active arm (i.e. $\alpha = 1/N$ and $\mathbf{P}^0 = \mathbf{I}_d$), for which the optimal solution can be obtained in $O(d^3)$ time independently of N (see for instance Niño-Mora [38] for an algorithm to compute the Gittins index). Indeed, the following two examples give hints as why it is unlikely for a simple optimal solution to exist in the multiple-activation restless case.

Example 3 ($M > 1$ vs. $M = 1$) Consider a restful bandit where each arm n is a job that takes a service time s_n to complete, and it incurs a cost c_n per unit time of delay before completion. It can be shown that the Gittins index for this problem is c_n/s_n. Now if we allow to activate at each time $M = 2$ arms, will the Gittins index policy by activating the two arms having currently the largest Gittins indices still be optimal? To be more specific, let us take $N = 3$ arms with $(c_1, s_1) = (c_2, s_2) = (1, 1)$ and $(c_3, s_3) = (2, 2)$. One might expect that any selection of two jobs for immediate service would be equally good, as the three arms share the same Gittins index. However, if we first activate job 1 and job 2, and then activate job 3 after termination of job 1, then the total cost is

$1\times1+1\times1+2\times(1+2)=8$. While if we first activate job 1 and job 3, and then activate job 2 after termination of job 1, then the cost becomes $1\times1+2\times2+1\times(1+1)=7$. Hence the second schedule does better than the first one as it uses all the available processing capacity until every job is finished. In fact, for $M=2$ with N jobs, the optimal solution amounts to finding a subset of the N jobs for which the total service time is as close as possible to $S/2$, with $S=\sum_{i=1}^{N} s_i$, and we activate concurrently the arms in these two sets. This is an alternative standard specification of the "knapsack problem", which is NP-complete.

Example 4 (Restless vs. Restful) Let $M=1$. We consider a restless bandit problem in which the transition matrices $\mathbb{P}^0$ and $\mathbb{P}^1$ for the passive and active actions are such that $\mathbb{P}^0(x,y)=\varepsilon_i\mathbb{P}^1(x,y)$ for states $y\neq x$, and $\mathbb{P}^0(x,x)=(1-\varepsilon_i)+\varepsilon_i\mathbb{P}^1(x,x)$ for some $\varepsilon_i\in[0,1], 1\leq i\leq N$. Let $\hat{\varepsilon}=\max_i \varepsilon_i$. Hence each arm may be thought to have two transition speeds (active and passive) in each state. If $\hat{\varepsilon}=0$, then the model reduces to the restful bandit problem, while on the other side of extreme, for $\hat{\varepsilon}$ close to 1, all policies are optimal. The degree of restlessness (interpreted as movement under the passive action) is easily measured through the ε_i's and the model then becomes a natural vehicle for an investigation of the quality of index policies in a restless environment. In Glazebrook et al. [23] this example is studied thoroughly and at the end a number of numeric simulations are done. They develop the so-called adaptive greedy algorithm and calculate a generalized type of index which coincides with the Gittins index in the $\hat{\varepsilon}=0$ case. The outcome of simulations could be summarized as follows: The index policy (with the index being calculated by the adaptive greedy algorithm) performs best for $\hat{\varepsilon}$ close to 0 or 1, and can be significantly sub-optimal for $\hat{\varepsilon}$ being in the mid range.

2.2 Indexability and Whittle Index

To overcome the difficulty of finding exact solutions to (2.2)-(2.3), a heuristic known as Whittle index policy (WIP) is introduced in Whittle [49]. This heuristic is obtained by computing an index ν_i for each state $i\in\mathcal{S}$, much like the Gittins index in the restful case. At a given decision epoch, WIP activates the αN arms having currently the highest indices.

The index of an arm can be computed by considering each individual arm in isolation. More precisely, for a given $\nu\in\mathbb{R}$, we define the ν-subsidized problem as the following MDP: The state space is the one of a single arm. At each time t, the decision maker chooses whether or not to activate this arm. As in the original problem, the arm evolves at time t according to (2.1). The difference lies in the passive action that is subsidized: If the arm is in state i and action 1 is taken, then as before, a reward R_i^1 is earned; if the arm is in state i and action 0 is taken, then a reward $R_i^0+\nu$ is earned. The goal of the decision maker is to maximize the long-term expected average reward per period (including passive subsidies).

For a given $\nu\in\mathbb{R}$, let us denote by $\omega(\nu)$ the set of states for which all optimal

policies of the ν-subsidized MDP are such that the passive action is optimal in these states. Naturally, for ν at the two extremities $-\infty$ and $+\infty$, the set $\omega(\nu)$ is respectively the empty set and the whole $\mathcal{S}$. If in addition $\omega(\nu)$ is monotonically increasing (in the sense of set inclusion) as ν increases, then we shall call the restless bandit *indexable*, and the Whittle index for state i will be the smallest value ν_i such that i belongs to the set $\omega(\nu_i)$. Formally:

Definition 2.2.1 (Indexability and Whittle index)**.** *A restless bandit* $(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1)$ *is* indexable *if* $\omega(\nu)$ *is increasing in* ν *(in the sense of set inclusion), namely if for all* $\nu \leq \nu'$*, we have* $\omega(\nu) \subseteq \omega(\nu')$*. In this case, the Whittle index of a state* i*, that we denote by* ν_i*, is defined as the smallest subsidy such that the passive action is optimal in this state:*

$$\nu_i := \inf_{\nu \in \mathbb{R}} \big\{ \nu \mid i \in \omega(\nu) \big\}.$$

Note that the value ν_i *is finite since the state space is finite.*

It should be emphasized that there exists restless bandit problems that are *not* indexable, in which cases the Whittle indices are not defined. Note that the Whittle index coincides with the classical definition of Gittins index for restful bandits, see Gittins et al. [21] for a proof. In this sense the Whittle index is a generalization of Gittins index for restless bandits. Since the restless bandits are in general PSPACE-hard, there is no reason that WIP should be optimal. But the relation between Whittle index and a relaxed problem (for which we discuss next) gives hope that WIP may be asymptotically optimal.

2.3 Whittle Relaxation and Asymptotic Optimality

An intuition behind the definition of Whittle index is given by considering a relaxation of the original N-armed problem (2.2) where the constraint (2.3) is replaced by $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} a_n(t) = \alpha N$. While the constraint (2.3) imposes that exactly αN arms are activated at each time step, the relaxed constraint only imposes that the time-averaged number of activated arms to be equal to αN. This gives the following optimization problem:

$$V_{\text{rel}}^{(N)}(\alpha) := \max_{\Pi} \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\sum_{t=0}^{T-1} \frac{1}{N} \sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)} \Big] \tag{2.4}$$

$$\text{subject to } \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} a_n(t) = \alpha N. \tag{2.5}$$

By using ν as a Lagrange multiplier of the constraint $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} a_n(t) = \alpha N$, the Lagrangian of the problem (2.4)-(2.5) is

$$\lim_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\sum_{t=0}^{T-1} \frac{1}{N} \sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)} \Big] + \nu \Big(\alpha - \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{n=1}^{N} a_n(t) \Big)$$

$$= \nu\alpha + \frac{1}{N}\sum_{n=1}^{N}\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1}\big(R_{S_n(t)}^{a_n(t)} - \nu a_n(t)\big)\Big]$$
$$= \frac{1}{N}\sum_{n=1}^{N}\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1}\big(R_{S_n(t)}^{a_n(t)} + \nu(1-a_n(t))\big)\Big].$$

Note that, for a fixed ν, finding a policy that maximizes the above Lagrangian can be done by solving N independent optimization problems (one for each arm in the sum of the last equation), and each problem is a ν-subsidized MDP that we described in Section 2.2. Indeed, it is not hard to show that $V_{\text{rel}}^{(N)}(\alpha) = \cdot V_{\text{rel}}^{(1)}(\alpha)$ is independent of N.

It should be clear that the constraint (2.5) is weaker than the constraint (2.3). This shows that $V_{\text{opt}}^{(N)}(\alpha) \leq V_{\text{rel}}^{(N)}(\alpha)$. Hence $V_{\text{rel}}^{(N)}(\alpha)$ is an upper bound on the value of the original optimization problem (2.2). Let us denote the long-term average expected reward of WIP to the original problem by $V_{\text{WIP}}^{(N)}(\alpha)$, i.e.

$$V_{\text{WIP}}^{(N)}(\alpha) := \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}R_{S_n(t)}^{a_n(t)}\Big],$$

where for all t, $\mathbf{a}(t)$ is chosen according to WIP.

It is then natural to expect that $V_{\text{WIP}}^{(N)}(\alpha)$ being close to $V_{\text{rel}}^{(N)}(\alpha)$ when N is large, as we expect a weaker coupling between the arms. Indeed, in Whittle [49] it is conjectured that $\lim_{N\to\infty} V_{\text{WIP}}^{(N)}(\alpha) \to V_{\text{rel}}^{(1)}(\alpha)$. We may then deduce that $\lim_{N\to\infty} V_{\text{opt}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha) = 0$, and consequently WIP is asymptotically optimal.

This conjecture is unfortunately proven to be false in Weber and Weiss [48], by providing some counterexamples in dimension $d = 4$. The good news is that with an additional assumption on the deterministic behavior of WIP, the conjecture is shown to hold true in the same paper. This deterministic dynamic of WIP can be described as follows: Let us call a *configuration vector* of an N-armed bandit at a given time step the vector representing the proportion of arms being in each state at that time. Let $\Delta^d \in \mathbb{R}^d_{\geq 0}$ be the unit d-simplex. A possible configuration vector of the system can then be represented by a point $\mathbf{m}$ in Δ^d, where m_i is the proportion of arms in state $i \in \{1 \dots d\}$. The deterministic dynamic of WIP is the map $\phi : \Delta^d \to \Delta^d$, such that given a configuration vector $\mathbf{m}$, the value $\phi_i(\mathbf{m})$ for all state i is the *expected* proportion of arms going to state i at the next time step (under WIP), knowing that the system was previously in configuration vector $\mathbf{m}$.

In order for WIP to be asymptotically optimal, it is necessary for the discrete time dynamic under the iteration of ϕ to converge towards a single fixed point. Formally, the following theorem can be proven:

Theorem 2.3.1 (Asymptotic optimality of WIP [48], modified to discrete time)**.** *Consider a discrete time recurrent restless bandit problem* $\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha\}$ *such that:*

(i) The restless bandit is indexable, so that WIP is well defined. Denote by ϕ *the map that describes the deterministic behavior of WIP in one time step, and by* Φ_t *the* t*-steps iteration of* ϕ*.*

(ii) The (unique) fixed point $\mathbf{m}^*$ *of* ϕ *is the global attractor of the discrete time dynamical system* $\Phi_{t\geq 0}(\cdot)$*: for all* $\mathbf{m} \in \Delta^d$, $\lim_{t\to\infty} \Phi_t(\mathbf{m}) = \mathbf{m}^*$.

Then $\lim_{N\to\infty} V_{\mathrm{WIP}}^{(N)}(\alpha) \to V_{\mathrm{rel}}^{(1)}(\alpha)$*, and consequently WIP is asymptotically optimal.*

To summarize, the conjecture of asymptotic optimality of WIP holds true if the restless bandit is indexable and the deterministic dynamics of WIP has a global attractor. This leaves the questions as at which rate does the convergence occur, and what to do if the two assumptions for asymptotic optimality are not satisfied, for which we provide answers in the current thesis. We may also quote the following paragraphs from Weber and Weiss [48] (with notations adapted) after their proof to the asymptotic optimality conjecture:

> ... *In fact, for the case* $d = 2$ *we have derived expressions for the equilibrium distribution of the index policy. One can give a direct proof of the truth of the conjecture. It turns out that the asymptotic difference between* $V_{\mathrm{WIP}}^{(N)}(\alpha)/N$ *and* $V_{\mathrm{rel}}^{(1)}(\alpha)$ *is even less than* $O(1/\sqrt{N})$.
>
> ... *Our impression is that counterexamples were produced for less than* 1 *in* 1000 *test problems. The size of the asymptotic sub-optimality of the index policy was no more than* 0.002% *in any example. Of course one should not place too much emphasis on results which depend on the way test problems are generated. We may be missing a class of examples for which the degree of sub-optimality is greater. A better understanding might lead to more dramatic counterexamples ... The evidence so far is that counterexamples to the conjecture are rare and that the degree of sub-optimality is very small. It appears that in most cases the index policy is a very good heuristic.*

2.4 The Linear Program Approach

As mentioned before, in many of the successful applications which have used Whittle indices, the indexability property is established by exploiting the special structure of the restless bandit model. In these cases some monotone structure can greatly simplify the task of demonstrating indexability and of recovering the Whittle indices. Indeed, as the Definition 2.2.1 of indexability is of a monotone nature, it should be expected that the models possess some kind of monotonicity in order to be indexable.

In general, it is desirable to exhibit a sufficient condition on the restless bandit for its indexability. Along this line we should cite the work of Nino-Mora [35], [37], in which the author defines the PCL-indexability (where PCL stands for "partial conservation law"), which is a stronger notion than the usual indexability. Assume the bandit is indexable, there also exists a general algorithm discussed in Niño-Mora [39], that computes the Whittle indices in cubic times in terms of the dimension d (note that this also generalizes the previous existing algorithm in Niño-Mora [38] for the computation of Gittins indices). This is later improved in Gast et al. [16] to sub-cubic times, moreover the algorithm proposed therein can also test the indexability at an additional cubic

times cost. Thanks to this algorithm that works generally for any restless bandit model, the authors run a statistical test to see if the indexability is a generic property that holds for a randomly generated model, the results are reported in Table 2 of [**?**].

In essence, the indexability of a restless bandit indeed holds true in a generic sense, meaning that if the parameters $(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1)$ are picked uniformly from an appropriate space of all possible parameters, then almost always the model is indexable (the chance increases with the dimension d and in $d = 10$ the probability is already $\geq 99.99\%$). However, if we restrict the parameter space to some particular sub-class, for instance, we require that the transition matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ to be tri-diagonal so that only the diagonal and the two off-diagonal terms can be non-zero, then the situation becomes the other way around: in $d = 10$ roughly half of the models are indexable, while in $d = 50$ the proportion decreases to only around 1.8%. This is somehow unfortunate, since any practical model for which we wish to apply WIP always has some additional structure and is never uniformly generated. In particular, the birth-and-death process falls into the tri-diagonal category, and it is not safe to claim from this statistical test that the problem is most likely to be indexable—we have to proceed by showing indexability directly from each specific model.

The aforementioned subtlety on the indexability assumption has been solved in the work Verloop [47]. The point is to realize that after the relaxation in (2.4)-(2.5), the problem can be transformed into a linear program with variables $\left\{y_{s,a}\right\}_{s\in\mathcal{S},a\in\mathcal{A}}$, for which an optimal solution $\left\{y^*_{s,a}\right\}_{s\in\mathcal{S},a\in\mathcal{A}}$ can be considered as the optimal occupation measure in stationary regime of taking action a for arms in state s. We may then partition the set of states $\mathcal{S}$ into $\mathcal{S}^+$, $\mathcal{S}^0$ and $\mathcal{S}^-$ depending on the sign of $y^*_{s,1}$ and $y^*_{s,0}$:

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

Also from an unichain assumption, there is no state such that $y^*_{s,1} = y^*_{s,0} = 0$. The claim is that any strict priority policy (which can be seen as a member in the permutation group with d elements) that assigns a priority order $\mathcal{S}^+ > \mathcal{S}^0 > \mathcal{S}^-$ is asymptotically optimal, provided that a global attractor property similar to the requirement for WIP holds true. In particular, WIP belongs to this class of asymptotically optimal policies, if the problem is indexable.

By this means we are saved from the verification of indexability, in exchange we only need to solve a linear program. However, the collection of asymptotically optimal policy based on solving the linear program may be extremely large, do they all perform equally good? If not, do we have a way to select among them a good one? We provide a possible solution in Chapter 4, by using the so-called LP-index to rank states within each of the three sets $\mathcal{S}^+$, $\mathcal{S}^0$ and $\mathcal{S}^-$. It can be shown that the LP-index of a state in $\mathcal{S}^+$ (resp. $\mathcal{S}^0$, $\mathcal{S}^-$) is positive (resp. zero, negative), hence these indices are coherent with the strict priorities of these three sets. It comes as no surprise that by adapting our proof of exponential convergence rate of WIP, one can show that all these asymptotically

optimal policies in Verloop [47] actually becomes optimal exponentially fast (modulo several technical details that we have discussed before). This, however, does not mean that they all perform equally good, as the numerical test in Chapter 5 that we run on a dimension $d = 10$ bandit with a small number of $N = 20$ arms shows: tie-solving within the three priority sets has a significant influence on performance for finite values of N, and one way that practically seems to give good performance is by using the LP-index, and we shall call the corresponding policy the *LP-index policy*.

2.5 The Finite Horizon Restless Bandit

If we consider the restless bandit model in a finite horizon $T < \infty$, we no longer need to worry about the limit behavior of the deterministic dynamical system. Consequently, both the indexability and the global attractor requirements can be avoided, by using the linear program approach under finite horizon. As a price, the policies are now time-dependent, and may take considerably more time for the computation. Indeed, in infinite horizon the linear program has roughly $|\mathcal{S}| \cdot |\mathcal{A}|$ variables with only one resource constraint, while in finite horizon we have $T \cdot |\mathcal{S}| \cdot |\mathcal{A}|$ variables with T resource constraints, one for each time. However, there are reasons other than avoiding the global attractor property that makes studying the problem under finite horizon interesting. For instance, there are real-life scenarios where the problem parameters are time-varying, or the horizon is very short (e.g. the applicant screening problem that we shall study in Chapter 5 and 6).

Restless bandits in finite horizon have been studied previously in Hu and Frazier [29] and Brown and Smith [12], using the LP approach. The main idea of their policies is to construct actions at each time t that follow as close as possible to an optimal LP solution $\left\{y^*_{s,a}(t)\right\}_{s\in\mathcal{S},a\in\mathcal{A},t\in[0,T[}$. However, a new difficulty arises as now it is possible for some time t to be such that $\mathcal{S}^0(t)$ has more than one element, where we recall that $\mathcal{S}^0(t)$ is the collection of states s such that $y^*_{s,0}(t) > 0$ and $y^*_{s,1}(t) > 0$. Indeed, it can be shown that in infinite horizon there always exists an optimal solution to the linear program for which $\left|\mathcal{S}^0\right| \leq 1$ (see Proposition 4.3.1 for a proof), so that the tie-solving using LP-index in infinite horizon actually only concerns states in $\mathcal{S}^+$ and $\mathcal{S}^-$. But this is no longer true in finite horizon. It turns out that tie-solving within states from $\mathcal{S}^0(t)$ at a time t with $\left|\mathcal{S}^0(t)\right| > 1$ is crucial for the asymptotic optimality, and the naive solution of assigning strict priorities in $\mathcal{S}^0(t)$ fails to work. In Brown and Smith [12] multiple tie-solving rules have been proposed, and their "Lagrange policy with optimal tie-breaking" is the same as our LP-index policy in finite horizon.

Under a much more general point of view, our approach in Gast et al. [18] is to treat the construction of LP-based policies as finding maps π_t for all time steps t from a configuration vector $\mathbf{M}^{(N)}(t) \in \Delta^d$ of an N-armed bandit to a decision vector $\mathbf{Y}(t)$, for which the entry $Y_{s,a}(t)$ encodes the information of the proportion of arms in state s undertaking action a at time t. If we denote by $\mathbf{m}^*(t)$ the LP-optimal configuration vector at time t, so that $m^*_s(t) = y^*_{s,0}(t) + y^*_{s,1}(t)$, then we expect these maps to satisfy $\pi_t(\mathbf{m}^*(t)) = \mathbf{y}^*(t)$, i.e. they should be LP-compatible. Recall that we mentioned

previously the main point of the LP-approach is to make the policy follow as much as possible to an optimal LP-solution. We formulate this in a very precise way, by showing that in order to have asymptotic optimality, we only need these maps π_t to be continuous and LP-compatible, and the convergence rate can be shown to be $O(1/\sqrt{N})$, if in addition these maps are Lipschitz continuous. More importantly, if these maps π_t can be constructed to be locally linear in a neighbourhood of $\mathbf{m}^*(t)$, for all time steps t, then the convergence can be shown to be exponentially fast. This is the finite horizon analogue of our exponential convergence result in infinite horizon.

To render this local linearity condition more applicable in practice, we show that it holds if and only if $\left|\mathcal{S}^0(t)\right| \geq 1$, for all time steps t. We shall call a finite horizon restless bandit that satisfies this condition *non-degenerate*. Previously in Zhang and Frazier [53], it has been shown that their LP-based policies on non-degenerate problems converge at $O(1/N)$ rate, by using a very different proof method. We improve this $O(1/N)$ rate to exponentially fast convergence, and show that the non-degenerate condition is a necessary and sufficient condition for the existence of exponential rate policy. Our proof is adapted from the infinite horizon case, where the role of "non-singular" is replaced by "non-degenerate", without the need of the global attractor assumption. To achieve this proof, some further cautions are needed for the integer rounding: a priori the values $N \cdot Y_{s,a}(t)$ are fractional numbers, with the vector $\mathbf{Y}(t)$ given by $\pi_t(\mathbf{M}^{(N)}(t))$, but we need them to be integer in order to apply the policy on an N-armed bandit. This is solved by our randomized rounding algorithm, that chooses a random feasible vector $\mathbf{Y}^{(N)}(t)$ with $N \cdot \mathbf{Y}^{(N)}(t)$ having only integer entries, while in expectation $\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right] = \mathbf{Y}(t)$.

2.6 The LP-update Policy and Its Generalization to Weakly Coupled MDPs

After a closer look into the proof of asymptotic optimality results in finite horizon, we realize that in the square root bound $O(1/\sqrt{N})$ proven for general problems, the hidden constant in the O notation that is independent of N may actually grow exponentially with T: $O(1/\sqrt{N}) = (C')^T/\sqrt{N}$ with $C' > 0$ independent of T and N. Likewise the exponential convergence rate bound $b \cdot \exp(-cN)$ for non-degenerate problems can be written more explicitly as $b' \cdot \exp(-N/c'^T)$, with $b', c' > 0$ being constants independent of both T and N. Consequently, since we are not able to prove that the constants C', c' are always possible to be chosen as smaller than 1, these bounds may become very poor if the horizon T is large.

Indeed, we have constructed a collection of tri-diagonal examples in dimension $d = 10$ with $T = 1000$, for which the infinite horizon global attractor property fails to hold numerically, and such that even arriving at $N = 10^9$, the performance of the LP-index policy is still below 80% with respect to the relaxed upper bound (after an appropriate normalization). we invite the reader to Section 7.3.2 for more details. The intuitive explanation to this phenomenon is that if the global attractor property

does not hold for the infinite horizon problem, then the (finite horizon) deterministic dynamic will most probably suffer from a unstable issue, so that it is very sensitive to small perturbation—even a tiny error is capable to propagate after several time steps into a huge deviation from the optimal trajectory. The LP-index policy that we compute from the start, being unable to take into account the large noise during the T time steps, can end up recommending a very bad choice of action.

The solution to the inadequacy of finite horizon LP-index policy in these situations is by using the so-called LP-update policy. It originates from the following nature idea: Since we are concerned about the configuration vector $\mathbf{M}^{(N)}(t)$ being deviated too much from the LP-optimal $\mathbf{m}^*(t)$, making the decision vector $\mathbf{Y}(t) = \pi_t(\mathbf{M}^{(N)}(t))$ a very bad choice, where the map π_t was computed t time steps ago, why not recompute everything? In other words, we can make decisions based on the new finite horizon restless bandit problem with initial condition $\mathbf{M}^{(N)}(t)$ and horizon $T - t$. We shall call such a replanning process as applying an *update*. The LP-based policy that applies an update at every time step will be defined as the *LP-update policy*.

The good news is we can show that the gap between the LP-update policy performance and the LP relaxation on a general finite horizon restless N-armed bandit to be bounded by $KT/\sqrt{N}$, where K is an upper-bound on the Lipschitz constants of the T functions that map an initial condition $\mathbf{m}(0)$ to the optimal LP value with horizon $1 \leq t \leq T$. Although we could not prove it, numerical evidence suggests that K is a constant independent of T. Consequently, it may be the case that the constant in the square root bound of asymptotic optimality for the LP-update policy does not grow exponentially with the horizon T, but instead grows only linearly. Indeed, the same experiment on the collection of tri-diagonal examples in dimension $d = 10$ with $T = 1000$ using the LP-update policy indicates that, arriving at $N = 100$ the performance is already above 95% with respect to the relaxed upper bound.

The obvious downside of the LP-update policy is its inefficiency, since we need to solve T linear programs instead of just one. A straightforward improvement is to only apply updates at moments for which $\mathbf{M}^{(N)}(t)$ has deviated "a lot" from the LP-optimal $\mathbf{m}^*(t)$. In Chapter 6, we shall formulate this intuitive idea in a rigorous way, in order to render the LP-update policy much more efficient. More precisely, the LP-optimal decision map π_t^*, which tells us the LP-optimal decision vector $\pi_t^*(\mathbf{M}^{(N)}(t))$ from solving the linear program with initial condition $\mathbf{M}^{(N)}(t)$ and horizon $T - t$, is actually a locally linear map in a neighbourhood of $\mathbf{m}^*(t)$, provided that a rank condition is satisfied. This rank condition is checked by verifying the invertibility of a matrix of dimension to the order of d, obtained from the parameters of the linear program. In case it holds true, the locally linear map π_t^* is then given by the inverse of this matrix. By this means, we are only obliged to solve new linear programs if the rank condition at a certain time t is not satisfied, or the decision vector $\pi_t^*(\mathbf{M}^{(N)}(t))$ with π_t^* given by the inverse matrix is not a feasible action, indicating that $\mathbf{M}^{(N)}(t)$ has deviated "a lot" from $\mathbf{m}^*(t)$.

Another big advantage of the LP-update policy compared to any other LP-based index-type policy is that it can be easily generalized to a multi-action multi-constraint framework, where "multi-action" refers to instead of having only the two passive and

active actions, we allow $A > 2$ actions for the action space $\mathcal{A}$ of each arm; and "multi-constraint" refers to having multiple resource constraints as opposed to just the single activation budget constraint αN.

The multi-action multi-constraint multi-armed bandit belongs to the more general class of problems called weakly coupled MDPs in the literature, e.g. the two PhD thesis Hawkins [26] and Salemi Parizi [46] are dedicated to this subject. Notice that there exists already multiple attempts in generalizing the classical restless bandit problem. For example, in Hodge and Glazebrook [27] and Niño-Mora [40] the infinite horizon model is generalized to the multi-action single-constraint case, with the focus of defining indexability and Whittle indices when each arm exhibit multiple actions. The finite horizon multi-action single-constraint problem has been considered in Zayas-Cabán et al. [52] and Xiong et al. [50], the main idea of their policies is to sample an action for each arm based on the occupation measure obtained from the LP solution at $t = 0$, subject to not violating the budget constraint. This approach, however, suffers from the instability issue that we mentioned before, and already in the classical two-action case gives worse performance than other more sophisticated policies (e.g. the LP-index policy). The reasons being that it does not take into account the tie-solving process, neither does it possess the local linearity for a faster convergence rate on non-degenerate problems.

Part I

Infinite Horizon

3

The Exponential Convergence Rate of WIP

It is shown in Weber and Weiss [48] that if the infinite horizon restless bandit is indexable and the associated deterministic system has a global attractor fixed point, then the Whittle index policy is asymptotically optimal in the regime where the arm population grows proportionally with the number of activation arms. In this chapter we show that, under the same conditions, this convergence rate is exponential in the arm population, unless the fixed point is *singular* (to be defined later), which almost never happens in practice. Using simulations and numerical solvers, we also investigate the singular cases, as well as how the level of singularity influences the (exponential) convergence rate. We illustrate our theorem on a Markovian fading channel model.

> I have tried to avoid long numerical computations, thereby following Riemann's postulate that proofs should be given through ideas and not voluminous computations.
>
> David Hilbert

3.1 Introduction

Despite the well-known asymptotic optimality of WIP (under some conditions) and its empirically good performance on numerous models, there is very limited research on how fast WIP becomes optimal. In this chapter we show that the convergence of the performance of WIP to the performance of an optimal policy is exponentially fast with the number N of arms, giving a theoretical explanation for the good performance of WIP in practice, even when the number of arms is small. This result holds under the same conditions as the asymptotic optimality proven in Weber and Weiss [48], namely the bandits are indexable and that the ordinary differential equation driving

the dynamics of the mean field approximation has a fixed point that is a global attractor, plus the additional conditions that the fixed point is *non-singular* (which almost always holds) and locally stable.

The proof of our main result (*i.e.* exponential convergence rate in the general case) relies on two main ingredients. The first one comes by noticing that the dynamics of the mean field approximation of the N arms under WIP, each with d states, is piecewise affine and continuous over a finite number of polytopes partitioning the configuration space (the simplex in dimension d). This piecewise linearity of the mean field approximation comes as a mixed blessing when one tries to compute the convergence rate: On the one hand the dynamics is not differentiable at the interface between the polytopes. Therefore, previous approaches based on the smoothness of the drift such as Gast et al. [15], Gast and Van Houdt [20], Ying [51] collapse here. On the other hand, when the global attracting fixed point falls into the interior of a polytope (*i.e.* it is non-singular), the dynamics in a small neighborhood around the fixed point is affine and the expected behavior of the system is relatively simple to analyze.

The second ingredient is to divide the analysis of the behavior of the stochastic system into two parts: before it enters a small neighborhood of the fixed point and after it does. The Stein's method is used to compare its behavior with its mean field approximation inside the neighborhood. Hoeffding's inequality is used to control its behavior outside the neighborhood.

Summary of contributions

In this chapter, we show that under indexability, global attraction of the fixed point of the mean field dynamics and non-singularity of this fixed point, the average performance of a stochastic Markovian bandit system under WIP converges to its mean field limit as $b \cdot \exp(-cN)$ where N is the number of arms and b, c are positive constants independent of N. Our result comes with several novelties.

- Firstly, we believe that this is the first example where an exponential convergence to a mean field limit has been obtained. This exponential rate relies crucially on the piecewise affine nature of the deterministic dynamical system, as opposed to most other mean field approximation results that prove convergence rates that are polynomial in $1/\sqrt{N}$ and for which the deterministic dynamics is smooth everywhere.

- Secondly, although a part of our proof has a large deviation flavor, our result concerns the expected behavior of the stochastic bandit and not its deviations, so that our result cannot be obtained by simply using general results on dynamical systems in the presence of random perturbations, such as the large deviation bounds presented in Section 1.5 in Kifer [31]. As for the part of our proof on concentration bounds that might have been obtained using large deviation principles, we believe that our direct proof, based on concentration inequalities, is simple enough and provides a clearer understanding of the picture.

- The contrast between singular and non-singular attractors has gone unnoticed so far. Our theoretical results (exponential convergence in the non-singular case and possibly only $O(1/\sqrt{N})$ in the singular case) are backed by numerical experiments showing that for a moderate number of arms (N ranging from 10 to 50), the relative performance of WIP w.r.t. the optimal policy can be almost perfect (less than 0.1 % difference) in the non-singular case to simply good (around 4 %) in the singular case.

Outline

The rest of the chapter is organized as follows: In Section 3.2 we introduce the restless bandit model and define the Whittle indices. We then present the main result in Section 3.3, namely exponential convergence for the performance of WIP to the optimal one in the general situation. In Section 3.4, we illustrate our results with several examples. We provide simulation and numerical estimations for the performance of WIP in different cases. In Section 3.5, we present an application of our result to the Markovian fading channel problem, where we check numerically with parameters that fall into the general case framework (non-singular global attracting fixed point). The proof of the main theorem is given in Section 3.6.

3.2 The Discrete Time Restless Bandit Model

We first describe the restless bandit model in Section 3.2.1. We then recall the definition of Whittle index in Section 3.2.2 and its relation with a linear relaxation in Section 3.2.3. This is a discrete time version of the classical continuous time model studied in Weber and Weiss [48].

3.2.1 Model description

Recall the discrete time infinite horizon restless bandit model that we defined in Chapter 2:

1. The model is composed of N statistically identical arms. Each arm evolves in a finite state space $\mathcal{S} := \{1 \dots d\}$ with an action set $\mathcal{A} = \{0, 1\}$, and the state of arm n at time t is denoted by $S_n(t) \in \{1 \dots d\}$. The state space of the whole process at time t is denoted by $\mathbf{S}(t) = \big(S_1(t), S_2(t), \dots, S_N(t)\big)$.

2. Decisions are taken at times $t \in \mathbb{N}$. At each decision epoch, a decision maker observes $\mathbf{S}(t)$ and chooses αN of the N arms to be activated, with $0 < \alpha < 1$. We set $a_n(t) = 1$ if arm n is activated at time t and $a_n(t) = 0$ otherwise. The action vector at time t is $\mathbf{a}(t) = \big(a_1(t), a_2(t), \dots, a_N(t)\big)$. It satisfies $\sum_{n=1}^{N} a_n(t) = \alpha N$.

3. Arm n evolves according to Markovian laws: for all states i, j, action $a \in \mathcal{A}$ and $t \in \mathbb{N}$:
$$\mathbb{P}(S_n(t+1) = j \mid S_n(t) = i, a_n(t) = a) = P_{ij}^{a}. \tag{3.1}$$

Given $\mathbf{a}(t)$ and $\mathbf{S}(t)$, the N arms make their transitions independently.

4. For each arm that is in state i and for which action $a \in \mathcal{A}$ is taken, a reward $R_i^a \in \mathbb{R}$ is earned.

The goal of the decision maker is to compute a decision rule in order to maximize the long-term expected average reward per period. The theory of stochastic dynamic programming (e.g. Puterman [44]) shows that there exists an optimal policy which is stationary and deterministic (i.e. $\mathbf{a}(t)$ can be chosen as a time-independent deterministic function of $\mathbf{S}(t)$). Denote by Π the set of such policies, which are maps from $\mathbf{S}$ to $\mathbf{a}$. The optimization problem of the decision maker can be formalized as

$$V_{\text{opt}}^{(N)}(\alpha) := \max_{\Pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)}\Big] \tag{3.2}$$

$$\text{subject to} \sum_{n=1}^{N} a_n(t) = \alpha N, \text{ for all } t \in \mathbb{N}. \tag{3.3}$$

We assume that α and N are such that αN is an integer (the case of non-integer values will be discussed in Section 3.4.3).

We emphasize that the symmetric arms assumption can be relaxed in a straightforward way to a finite number of classes of arms. We then need to specify the initial proportion of arms in each class, and the parameters of each class will be given separately. The transition matrices will be k-blocks matrices, if there are k classes of arms. We will study a 2 classes bandit problem in detail later in Section 3.5.

In the above formulation and in what follows, the dependence of $a_n(t)$ on $S_n(t)$ based on a policy in Π should be understood. We also assume that the parameters of the model are such that the states of the N-armed bandit form a single aperiodic closed class, regardless of the policy employed. This assumption is mostly to simplify our discussion and is also used in Weber and Weiss [48] to guarantee that neither the value of the optimization problem (3.2) nor the optimal policy depend on the initial state $\mathbf{S}(0)$ of the system at time 0. We call such a bandit an *aperiodic recurrent* bandit.

3.2.2 Indexability and Whittle index

The index of an arm can be computed by considering each individual arm in isolation. More precisely, for a given $\nu \in \mathbb{R}$, we define the ν-subsidized problem as the following MDP: The state space is the one of a single arm. At each time t, the decision maker chooses whether or not to activate this arm. As in the original problem, the arm evolves at time t according to (3.1). The difference lies in the passive action that is subsidized: If the arm is in state i and action 1 is taken, then as before, a reward R_i^1 is earned; if the arm is in state i and action 0 is taken, then a reward $R_i^0 + \nu$ is earned. The goal of the decision maker is to maximize the long-term expected average reward per period (including passive subsidies).

For a given $\nu \in \mathbb{R}$, let us denote by $\omega(\nu)$ the set of states for which all optimal policies of the ν-subsidized MDP are such that the passive action is optimal in these

states. Naturally, for ν at the two extremities $-\infty$ and $+\infty$, the set $\omega(\nu)$ is respectively the empty set and the whole $\mathcal{S}$. If in addition $\omega(\nu)$ is monotonically increasing (in the sense of set inclusion) as ν increases, then we shall call the restless bandit *indexable*, and the Whittle index for state i will be the smallest value ν_i such that i belongs to the set $\omega(\nu_i)$. Formally:

Definition 3.2.1 (Indexability and Whittle index)**.** *A restless bandit* $(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1)$ *is* indexable *if* $\omega(\nu)$ *is increasing in* ν *(in the sense of set inclusion), namely if for all* $\nu \leq \nu'$*, we have* $\omega(\nu) \subseteq \omega(\nu')$*. In this case, the Whittle index of a state* i*, that we denote by* ν_i*, is defined as the smallest subsidy such that the passive action is optimal in this state:*

$$\nu_i := \inf_{\nu \in \mathbb{R}} \big\{ \nu \mid i \in \omega(\nu) \big\}.$$

Note that the value ν_i *is finite since the state space is finite.*

3.2.3 Whittle relaxation and asymptotic optimality

An intuition behind the definition of Whittle index is given by considering a relaxation of the original N-armed problem (3.2) where the constraint (3.3) is replaced by $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} a_n(t) = \alpha N$. While the constraint (3.3) imposes that exactly αN arms are activated at each time step, the relaxed constraint only imposes that the time-averaged number of activated arms to be equal to αN. This gives the following optimization problem:

$$V_{\text{rel}}^{(N)}(\alpha) := \max_{\Pi} \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\sum_{t=0}^{T-1} \frac{1}{N} \sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)} \Big] \tag{3.4}$$

$$\text{subject to } \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} a_n(t) = \alpha N. \tag{3.5}$$

By using ν as a Lagrange multiplier of the constraint $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} a_n(t) = \alpha N$, the Lagrangian of the problem (3.4)-(3.5) is

$$\begin{aligned}
&\lim_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\sum_{t=0}^{T-1} \sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)} \Big] + \nu \Big(\alpha - \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{n=1}^{N} a_n(t) \Big) \\
&\quad = \nu\alpha + \frac{1}{N} \sum_{n=1}^{N} \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\sum_{t=0}^{T-1} \big(R_{S_n(t)}^{a_n(t)} - \nu a_n(t) \big) \Big] \\
&\quad = \frac{1}{N} \sum_{n=1}^{N} \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\sum_{t=0}^{T-1} \big(R_{S_n(t)}^{a_n(t)} + \nu(1 - a_n(t)) \big) \Big].
\end{aligned}$$

Note that, for a fixed ν, finding a policy that maximizes the above Lagrangian can be done by solving N independent optimization problems (one for each arm in the sum of the last equation), and each problem is a ν-subsidized MDP that we described in Section 2.2. Indeed, it is not hard to show that $V_{\text{rel}}^{(N)}(\alpha) = V_{\text{rel}}^{(1)}(\alpha)$ is independent of N.

It should be clear that the constraint (3.5) is weaker than the constraint (3.3). This shows that $V_{\text{opt}}^{(N)}(\alpha) \leq V_{\text{rel}}^{(N)}(\alpha)$. Hence $V_{\text{rel}}^{(N)}(\alpha)$ is an upper bound on the value of the original optimization problem (3.2). Let us denote the long-term average expected reward of WIP to the original problem by $V_{\text{WIP}}^{(N)}(\alpha)$, i.e.

$$V_{\text{WIP}}^{(N)}(\alpha) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)}\Big],$$

where for all t, $\mathbf{a}(t)$ is chosen according to WIP.

It is then natural to expect that $V_{\text{WIP}}^{(N)}(\alpha)$ being close to $V_{\text{rel}}^{(N)}(\alpha)$ when N is large, as we expect a weaker coupling between the arms. Indeed, in Whittle [49] it is conjectured that $\lim_{N\to\infty} V_{\text{WIP}}^{(N)}(\alpha) \to V_{\text{rel}}^{(1)}(\alpha)$. We may then deduce that $\lim_{N\to\infty} V_{\text{opt}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha) = 0$, and consequently WIP is asymptotically optimal.

This conjecture is unfortunately proven to be false in Weber and Weiss [48], by providing some counterexamples in dimension $d = 4$. The good news is that with an additional assumption on the deterministic behavior of WIP, the conjecture is shown to hold true in the same paper. This deterministic dynamic of WIP can be described as follows: Let us call a *configuration vector* of an N-armed bandit at a given time step the vector representing the proportion of arms being in each state at that time. Let $\Delta^d \in \mathbb{R}^d_{\geq 0}$ be the unit d-simplex. A possible configuration vector of the system can then be represented by a point $\mathbf{m}$ in Δ^d, where m_i is the proportion of arms in state $i \in \{1 \dots d\}$. The deterministic dynamic of WIP is the map $\phi : \Delta^d \to \Delta^d$, such that given a configuration vector $\mathbf{m}$, the value $\phi_i(\mathbf{m})$ for all state i is the *expected* proportion of arms going to state i at the next time step (under WIP), knowing that the system was previously in configuration vector $\mathbf{m}$.

In order for WIP to be asymptotically optimal, it is necessary for the discrete time dynamic under the iteration of ϕ to converge towards a single fixed point. Formally, the following theorem can be proven:

Theorem 3.2.2 (Asymptotic optimality of WIP [48], modified to discrete time)**.** *Consider a discrete time recurrent restless bandit problem* $\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha\}$ *such that:*

(i) The restless bandit is indexable, so that WIP is well defined. Denote by ϕ the map that describes the deterministic behavior of WIP in one time step, and by Φ_t the t-steps iteration of ϕ.

(ii) The (unique) fixed point $\mathbf{m}^*$ of ϕ is the global attractor of the discrete time dynamical system $\Phi_{t\geq 0}(\cdot)$: for all $\mathbf{m} \in \Delta^d$, $\lim_{t\to\infty} \Phi_t(\mathbf{m}) = \mathbf{m}^*$.

Then $\lim_{N\to\infty} V_{\text{WIP}}^{(N)}(\alpha) \to V_{\text{rel}}^{(1)}(\alpha)$, and consequently WIP is asymptotically optimal.

3.3 Main Results

We first show in Section 3.3.1 that, when N is large, the stochastic system governed by WIP behaves like a piecewise affine deterministic system. We then present the exponential convergence result in Section 3.3.2.

3.3.1 Piecewise affine dynamics and definition of a singular point

To avoid ambiguity in the definition of WIP, we assume that the problem is *strictly* indexable. By this, we mean that there does not exist two states that have the same Whittle index. This is mostly a technical assumption that guarantees that there is a unique[1] WIP.

Recall that the state space of a single arm is $\{1 \dots d\}$, and assume without loss of generality that the states are already sorted according to their Whittle indices in decreasing order: $\nu_1 > \nu_2 > \cdots > \nu_d$. We shall call a *configuration* of an N-armed system the vector representing the proportion of arms being in each state. Let $\Delta^d \in \mathbb{R}^d_{\geq 0}$ be the unit d-simplex, that is $\Delta^d := \{\mathbf{m} \in [0,1]^d \mid m_1 + m_2 + ... + m_d = 1\}$. A possible configuration of the system at a given time step can be represented by a point $\mathbf{m}$ in Δ^d, where m_i is the proportion of arms in state $i \in \{1 \dots d\}$.

Our result on the rate at which WIP becomes asymptotically optimal depends on the property of the iterations of a deterministic map that we define below. Denote by $\mathbf{M}^{(N)}(t)$ the N-armed system configuration at time t under WIP. The arms being time homogeneous Markov chains, we can define a map $\phi : \Delta^d \to \Delta^d$ as

$$\phi_i(\mathbf{m}) := \mathbb{E}\left[M_i^{(N)}(t+1) \mid \mathbf{M}^{(N)}(t) = \mathbf{m}\right] \tag{3.6}$$

for all $i \in \{1 \dots d\}$ and $\mathbf{m} \in \Delta^d$. It is the expected proportion of arms going to state i at time $t+1$ under WIP, knowing that the system was in configuration $\mathbf{m}$ at time t. This map has the following properties:

Lemma 3.3.1. *Assume that the bandit is indexable. Then:*

(i) The definition of ϕ does not depend on N (as long as αN is an integer) nor on t.

(ii) ϕ is a piecewise affine function, with d affine pieces, and ϕ is Lipschitz-continuous.

(iii) ϕ has a unique fixed point: there exists a unique $\mathbf{m} \in \Delta^d$ such that $\phi(\mathbf{m}) = \mathbf{m}$.

Sketch of proof. The full details of the proof are provided in Section 3.6.1. We describe the main ingredients here.

Proof of (i) and (ii) – For a given configuration $\mathbf{m} \in \Delta^d$, define $s(\mathbf{m}) \in \{1 \dots d\}$ to be the state such that $\sum_{i=1}^{s(\mathbf{m})-1} m_i \leq \alpha < \sum_{i=1}^{s(\mathbf{m})} m_i$, with the convention that $\sum_{i=1}^{0} m_i = 0$. WIP activates arms by decreasing index order. This means that when the system is in configuration $\mathbf{m}$, WIP will activate all arms that are in states 1 to $s(\mathbf{m}) - 1$, and $N(\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i)$ arms that are in state $s(\mathbf{m})$. The rest of the arms will not be activated. This means that the map ϕ satisfies:

$$\phi_j(\mathbf{m}) = \sum_{i=1}^{s(\mathbf{m})-1} m_i P^1_{ij} + (\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i) P^1_{s(\mathbf{m})j}$$

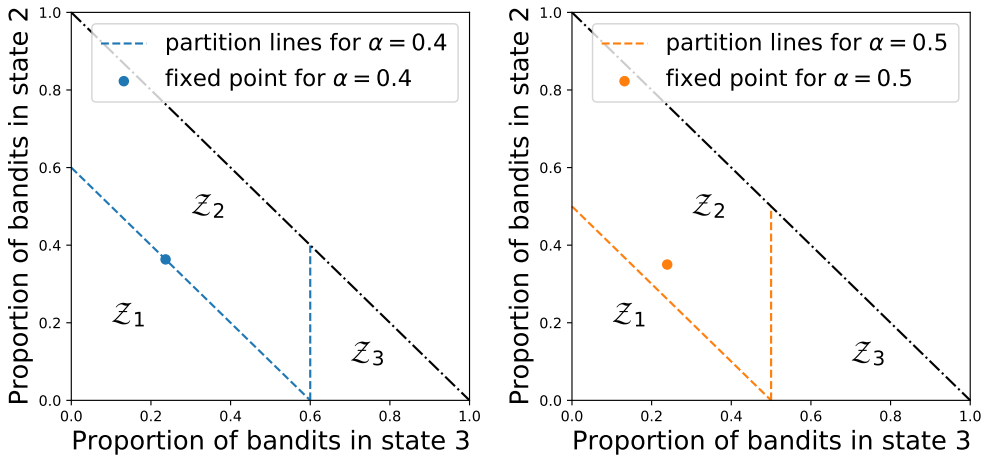
[1] If two states or more had the same index, to specify an index policy, one would need a tie-breaking rule. Our proof would work if the tie-breaking rule defines a strict order of the states.

$$+ \Big(\sum_{i=1}^{s(\mathbf{m})} m_i - \alpha\Big) P^0_{s(\mathbf{m})j} + \sum_{i=s(\mathbf{m})+1}^{d} m_i P^0_{ij}. \tag{3.7}$$

Let $\mathcal{Z}_i := \{\mathbf{m} \in \Delta^d \mid s(\mathbf{m}) = i\}$. The above expression of ϕ implies that this map is affine on each zone $\mathcal{Z}_i$, and there are d such zones. Moreover, the value of ϕ coincides on the intersection of zones, hence ϕ is continuous.

Proof of (iii) – This part of the proof is more involved, and it relies on indexability. The details are given in Section 3.6.1 where we show that indexability implies a monotonic property of ϕ that we use to obtain uniqueness. □

In what follows, we will denote by $\mathbf{m}^*$ the unique fixed point of ϕ. As we will see in Theorem 3.3.2, the rate at which WIP becomes asymptotically optimal depends on: (1) whether the iterations of ϕ converge to $\mathbf{m}^*$, (2) whether $\mathbf{m}^*$ lies strictly inside a zone $\mathcal{Z}_i$. Concerning the second property, we will call a point $\mathbf{m}$ *singular* if there exists $i \in \{1 \dots d\}$ such that $\sum_{j=1}^{i} m_j = \alpha$. Said otherwise, a fixed point is singular if it is on the boundary of two zones.



(a) Singular fixed point.

(b) Non-singular fixed point.

Figure 3.1 – An example with $d = 3$. When $\alpha = 0.4$ (Figure 3.1a) the fixed point is singular, while for $\alpha = 0.5$ (Figure 3.1b) it is not singular.

In Figure 3.1, we illustrate the notion of singular fixed point by an example in dimension $d = 3$. As $m_1 + m_2 + m_3 = 1$, the simplex Δ^3 can be represented in a 2-dimensional space as Δ^2_c, where Δ^d_c is the unit d-simplex and its interior. Our convention is that the x-coordinate of a point corresponds to m_3 (the proportion of arms in state 3), and the y-coordinate corresponds to m_2 (the proportion of arms in state 2). The colored dotted lines of Figures 3.1a and 3.1b are singular points. These lines partition the different zones $\mathcal{Z}_i$. The partition of zones, as well as the position of the unique fixed point depend on α. For this example, when $\alpha = 0.4$ (Figure 3.1a), the fixed point is singular, while for $\alpha = 0.5$ (Figure 3.1b), it is non-singular (all the other parameters in these two figures are the same).

3.3.2 Exponential convergence rate

We are now ready to state our main theorem. Assume indexability, at a given time t, WIP sorts all arms according to the Whittle indices $\nu_{S_n(t)}$ and activates the αN arms that have the highest indices. Let Φ_t be defined as the t-th iteration of the map ϕ, *i.e.* $\Phi_t : \Delta^d \to \Delta^d$ is $\Phi_0(\mathbf{m}) := \mathbf{m}$, and $\Phi_{t+1}(\mathbf{m}) := \phi\big(\Phi_t(\mathbf{m})\big)$. Recall that $\mathbf{m}^*$ is the unique fixed point of ϕ. As stated in the next theorem, the asymptotic optimality of WIP is guaranteed when $\mathbf{m}^*$ attracts all trajectories of $\Phi_{t\geq 0}(\cdot)$. In the rest of the chapter, unless otherwise specified, we use $\|\cdot\|$ to denote the $\mathcal{L}^\infty$-norm of a vector.

Theorem 3.3.2 (Exponential convergence rate theorem)**.** *Consider a discrete time recurrent restless bandit problem* $\big\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha\big\}$ *such that:*

(i) The bandit problem is indexable.

(ii) The (unique) fixed point $\mathbf{m}^*$ *of* ϕ *is not singular.*

(iii) $\mathbf{m}^*$ *is an attractor of* $\Phi_{t\geq 0}(\cdot)$*: for all* $\mathbf{m} \in \Delta^d$, $\lim_{t\to\infty} \Phi_t(\mathbf{m}) = \mathbf{m}^*$.

(iv) $\mathbf{m}^*$ *is locally stable: for all* $\varepsilon > 0$*, there exists* $\delta > 0$ *such that if* $\|\mathbf{m} - \mathbf{m}^*\| \leq \delta$*, then for all* t*:* $\|\Phi_t(\mathbf{m}) - \mathbf{m}^*\| \leq \varepsilon$.

Then there exists two constants $b, c > 0$ *that depend only on* $\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1$ *and* α*, such that for any* N *with* αN *being an integer,*

$$0 \leq V^{(N)}_{\text{rel}}(\alpha) - V^{(N)}_{\text{WIP}}(\alpha) \leq b \cdot e^{-cN}. \tag{3.8}$$

Recall that $V^{(N)}_{\text{rel}}(\alpha)$ *is the value of the relaxed problem* (3.4)-(3.5)*.*

Sketch of proof. The full details of the proof are given in Section 3.6.2. We first transform the evaluation of the performance to the analysis of the configuration of the bandit system. We then show that in stationary regime the expectation of $\mathbf{M}^{(N)}(0)$ concentrates exponentially fast on the fixed point $\mathbf{m}^*$. More precisely, there exists constants $b', c' > 0$ such that $\|\mathbb{E}[\mathbf{M}^{(N)}(0)] - \mathbf{m}^*\| \leq b' \cdot e^{-c'N}$. In order to show this:

- We first use Hoeffding's inequality in Lemma 3.6.5 to show that for any configuration $\mathbf{m}$: $\mathbb{P}\left[\|\mathbf{M}^{(N)}(1) - \phi\big(\mathbf{M}^{(N)}(0)\big)\| \geq \delta \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\right] \leq d \cdot e^{-2N\delta^2}$.
- By Lipschitz continuity of ϕ, for a time t, we apply Lemma 3.6.5 to prove Lemma 3.6.6, which bounds $\mathbb{P}\left[\|\mathbf{M}^{(N)}(t) - \Phi_t\big(\mathbf{M}^{(N)}(0)\big)\| \geq \epsilon\right]$ by a term that depends on t but decreases exponentially fast with N.
- As $\mathbf{m}^*$ is an attractor that is locally stable, this implies that when t is large enough, $\mathbf{M}^{(N)}(t)$ is within a neighborhood $\mathcal{N}$ of $\mathbf{m}^*$ with very high probability. As $\mathbf{m}^*$ is non-singular, this neighborhood can be taken to be within a zone $\mathcal{Z}_i$ on which ϕ is affine. We will choose carefully this neighborhood $\mathcal{N}$ and make sure that its choice does not depend on N. We then deduce an exponentially small upper-bound for the probability of $\mathbf{M}^{(N)}(0)$ in stationary regime being outside $\mathcal{N}$ (see Section 3.6.7), hence allows us to restrict our attention to a zone where ϕ is affine.

- The result then follows by using Stein's method on the process restricted to this affine zone, which shows that conditional on starting inside the neighborhood $\mathcal{N}$, the additive long-term distance between the large N stochastic trajectory and the deterministic trajectory is exponentially small (see Section 3.6.7).

□

We give here some comments on the assumptions of Theorem 3.3.2, their practical relevance will be discussed in Section 3.4.1. These assumptions are very similar to the ones needed to prove the asymptotic optimality of WIP in the case of continuous time bandits of Weber and Weiss [48]. The indexability property of the bandit problem is a necessary condition for WIP to be well defined and was also assumed in Weber and Weiss [48]. The non-singular condition on $\mathbf{m}^*$ is almost always satisfied (see Remark 3.3.3). The most difficult assumption to verify is point (iii) that requires $\mathbf{m}^*$ to be a global attractor. Note that in addition to being a global attractor, we also add the technical condition (iv) that $\mathbf{m}^*$ is locally stable. Actually, $\mathbf{m}^*$ being a locally stable attractor is a necessary condition in the sense that there exists examples that satisfy all assumptions of Theorem 3.3.2 except this one and for which WIP is not asymptotically optimal (see Remark 3.3.4).

Remark 3.3.3. The non-singular condition. *The key ingredient behind the proof of the exponentially fast asymptotic optimality is that the deterministic one-step dynamics ϕ of WIP is a piecewise affine continuous map inside the simplex Δ^d, with d affine pieces. In other words, the simplex Δ^d can be partitioned into d polytopes, so that ϕ is linear in each polytope and coincides on the interface of any two neighbouring polytopes. It appears that the (exponential) convergence rate is influenced by the relative position of the unique fixed point $\mathbf{m}^*$ with respect to the d polytopes, and this in turn is decided by the value of α, which we recall that αN is the activation budget of arms at each time step.*

More precisely, image d polytopes placing one next to another in a row from left to right. As α varies continuously from 0 to 1, with the other parameters $(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1)$ unchanged, the fixed point $\mathbf{m}^*(\alpha)$ that depends on α travels (continuously) from the leftmost polytope into the rightmost polytope. There will be exactly $d-1$ instances that corresponds to $d-1$ values of α such that $\mathbf{m}^*(\alpha)$ is on the interface of two neighbouring polytopes. For instance, Figure 3.1a corresponds to such a critical moment. The claim is that except for these $d-1$ values of α (that we termed as being singular*), the rest of the parameters are such that WIP becomes optimal exponentially fast (provided that the other local stability assumption that we discuss next also holds true). Moreover, the exponential rate is faster (in the sense that we can choose a larger constant $c > 0$ for the bound $b \cdot e^{-cN}$), if $\mathbf{m}^*(\alpha)$ is positioned in the middle of a polytope, rather than being close to an interface of two neighbouring polytopes. This is illustrated further in Sections 7.1.1 and 7.1.4 using an example with $d = 3$.*

The non-singularity of the fixed point $\mathbf{m}^*$ is also a necessary condition, in the sense that the following simple example satisfies all the assumptions of Theorem 3.3.2 except this one and does not satisfy (3.8). Consider the following 2 states bandit problem with $\mathbf{P}^0 = \mathbf{P}^1 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$, $\mathbf{R}^0 = (0,0)$, $\mathbf{R}^1 = (1,0)$, and $\alpha = 0.5$. The fixed point is $\mathbf{m}^* = (0.5, 0.5)$. It is singular.

It should be clear that $V_{\text{rel}}^{(1)}(\alpha) = 0.5$. *In stationary regime, the configuration* $\mathbf{M}^{(N)}$ *of the system of size* N *is distributed independently from the policy employed. Moreover, WIP will activate in priority the arms in state* 1. *This implies that the reward of WIP will be* $V_{\text{WIP}}^{(N)}(\alpha) = \mathbb{E}\left[\min(M_1^{(N)}, 0.5 \cdot N)\right]$. *As* $M_1^{(N)}$ *follows a binomial distribution of parameter* $(N, 0.5)$, *the central limit theorem shows that*

$$\lim_{N\to\infty} \sqrt{N} \cdot (V_{\text{rel}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha)) = 0.5 \cdot \mathbb{E}[\max(G, 0)] = \frac{1}{\sqrt{2\pi}},$$

where G *is a standard normal random variable.*

This example shows that, in a case where $\mathbf{m}^*$ *is singular, the convergence in* (3.8) *may occur at rate* $\Theta(1/\sqrt{N})$ *and not at exponential rate. Note on the other hand that if we take instead* $\alpha \neq 0.5$, *then* $V_{\text{WIP}}^{(N)}(\alpha)/N$ *converges to* $V_{\text{rel}}^{(1)}(\alpha) = \min(\alpha, 0.5)$ *at exponential rate, due to the fact that almost all the mass of a Gaussian distribution is concentrated around its mean value* α *(which is different from* 0.5*).*

Remark 3.3.4. The locally stable condition. *Despite this non-singular condition that holds with probability* 1 *(in an appropriate measurable space for the parameters of the restless bandit model), we also need to assume that the unique fixed point* $\mathbf{m}^*$ *is locally stable. This seems to be a natural consequence of the previous assumptions that we have already made on the model. Namely, if the fixed point is strictly inside a polytope, while in the same time being a global attractor, then it seems reasonable that the linear factor of the piecewise affine map* ϕ *in this polytope should be a stable matrix, i.e. all its eigenvalues should have modules smaller than* 1.

Unfortunately, we were unable to show this local stability as a consequence of the other assumptions, as we can not exclude the possibility of the (non-singular) fixed point being locally unstable while still attracts globally all the trajectories. Evidence suggests that this peculiar case is extremely rare, since we have not yet found numerically a single such pathologic example. On the other hand, it is indeed possible for $\mathbf{m}^*$ *to be locally stable while not being a global attractor, and we shall illustrate such an example in Section 7.1.3. Our belief is that the two assumptions of attracting globally and being stable locally are independent and can not be deduced from one to the other, although in practice they almost always hold at the same time.*

Remark 3.3.5. Cyclic and chaotic behaviors. *Although the drift* ϕ *is piecewise affine and has a unique fixed point, the long run behavior of the deterministic dynamical system* $\mathbf{m}(t+1) = \phi(\mathbf{m}(t))$ *can be cyclic or chaotic. In these cases, the fixed point is no longer a global attractor, and the performance of WIP is in general not asymptotically optimal.*

More precisely, when the dynamical system admits a cycle as a global attractor for almost every initial configuration in the simplex, then as suggested in Weber and Weiss [48], one can infer a cyclic version of Theorem 3.3.2: The performance of WIP converges to the average reward on the cycle. This average reward is in general strictly smaller than $V_{\text{rel}}^{(1)}(\alpha)$, *while* $V_{\text{opt}}^{(N)}(\alpha)/N$ *always converge to* $V_{\text{rel}}^{(1)}(\alpha)$, *regardless to the behavior of the deterministic system. Consequently, when cycles appear, the performance of WIP is asymptotically sub-optimal. This will be illustrated further via numerical examples in Sections 7.1.2 and 7.1.3.*

Remark 3.3.6. What happens when αN **is not an integer**. *The exponential convergence rate in Theorem 3.3.2 assumes that* αN *is an integer. When it is not the case, a decision maker*

cannot activate exactly αN arms at each time step. There are three natural solutions to define the model in such cases: (1) activate $\lfloor \alpha N \rfloor$ arms; (2) activate $\lceil N\alpha \rceil$ arms; (3) activates $\lfloor \alpha N \rfloor$ arms, plus one more arm being activated with probability $\alpha N - \lfloor \alpha N \rfloor$. As we further discuss in Section 3.4.3, the convergence rate in the first two solutions is much slower than in the third solution.

3.4 Numerical Experiments

In this section, we first provide statistical results to justify the conditions needed for Theorem 3.3.2, and then verify numerically the exponential convergence rate for a general 3 states restless bandit model with non-singular fixed points. We also evaluate numerically the convergence rate for a singular fixed point example. At last we investigate the situation when αN is not an integer.

3.4.1 How general is the general case?

The exponential convergence rate for the performance of WIP on a restless bandit problem is very desirable, however, several conditions have to be verified beforehand, listed in order as:

(C1) The restless bandit problem is indexable;

(C2) The unique fixed point is not singular;

(C3) The unique fixed point is a global attractor.

(C4) The unique fixed point is locally stable

Condition (C1) is mostly verified through the specific structure of the restless bandit problem and by using various techniques that are model dependent; a general method for the test of indexability is also presented in Gast et al. [16]. For Condition (C2), checking the singularity condition is straightforward, as it amounts to checking whether the sum of the first $s(\mathbf{m}*)$ coordinates of $\mathbf{m}^*$ (after the Whittle index reordering) is α. Moreover, being in an exact singular situation is improbable (for a given problem, the activation ratio α can only be singular if it satisfies an equality constraint). More generally, we also observe that the "closer" the fixed point to a singular situation, the smaller the coefficient c in Theorem 3.3.2 on the estimation of the exponential rate could be. This point will be made more precise in the next subsection.

Condition (C3) is more complicated to verify, as there is no general method to exclude cyclic or chaotic behaviors from a dynamical system. Indeed, Blondel et al. [10] shows that global properties of continuous piecewise affine functions in $\mathbb{R}^n$ is *undecidable*, as long as $n \geq 3$. At this stage, the best we can do is to verify Condition (C3) numerically, by simulating the dynamics on a large number of initial conditions over a long horizon.

As for Condition (C4), the local stability is easy to verify numerically when $\mathbf{m}^*$ is not singular: indeed, in this case the dynamical system is affine in a neighborhood of

$\mathbf{m}^*$: $\phi(\mathbf{m}) = (\mathbf{m} - \mathbf{m}^*) \cdot \mathbf{K}_{s(\mathbf{m}^*)} + \mathbf{m}^*$, where $\mathbf{K}_{s(\mathbf{m}^*)}$ is a matrix of dimension d obtained from (3.7). The dynamical system is locally stable if $\mathbf{K}_{s(m^*)}$ is a stable matrix, *i.e.* if the norm of all eigenvalues of $\mathbf{K}_{s(m^*)}$ is less than 1 [2]. If $\mathbf{K}_{s(m^*)}$ is not a stable matrix, then in most cases the fixed point will not be a global attractor and an attracting cycle will appear.

Dimension d	3	4	5	6	7
Non-indexable	653	81	5	0	0
Indexable with $\mathbf{m}^*$ not locally stable	9878	1020	82	11	0
% violating a condition of Theorem 3.3.2	0.1%	0.01%	10^{-3}%	10^{-4}%	0

Table 3.1 – Number of randomly generated instances that violate any of the conditions of Theorem 3.3.2 out of 10^7 uniformly generated restless bandit models for each dimension $d \in \{3, 4, 5, 6, 7\}$.

To give an idea of how general these conditions are, we generate a large number of discrete time restless bandit problems by choosing random parameter $(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1)$ in dimensions $d \in \{3, 4, 5, 6, 7\}$. We estimate the rarity of violations of the above conditions. More precisely, for each d, we randomly generate 10^7 instances of $(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1)$, using a uniform distribution in $[0, 1]$ for the rewards, and uniform distribution for probability vectors $\mathbf{P}_i^0$ and $\mathbf{P}_i^1$ over the simplex Δ^d. We then count the number of instances that violate conditions (C1) or (C4), the results are reported in Table 3.1. This table shows that the number of models that satisfy the conditions is more than 99.8% for $d = 3$; when $d = 7$, all generated models (among 10^7) satisfy our conditions. In our tests, what we mean by *the number of indexable instances such that* $\mathbf{m}^*$ *is not locally stable* is the number of models for which there exists $\alpha \in (0, 1)$ such that $\mathbf{m}^*$ is not locally stable. This can be done by testing each of the d matrices K_i. Note that for all these locally stable examples in Table 3.1, the corresponding $\mathbf{m}^*$ also appears to be a global attractor (numerically). However, we should point out that it is possible to construct examples for which $\mathbf{m}^*$ is locally stable while not being a global attractor. Such examples have special structures and are almost impossible to find if we generate the parameters uniformly, see Section 7.1.3.

3.4.2 The influence of how *non-singular* is a fixed point

To test how the "non-singularity" of the fixed point $\mathbf{m}^*$ affects the convergence rate, we consider the example displayed in Figure 3.1 with varying values of α in the range between 0.20 and 0.50. We emphasize that the fixed point $\mathbf{m}^* = \mathbf{m}^*(\alpha)$ is then a function of α. Numerically, these fixed points are global attractors for two reasons:

[2] Recall that ϕ is an application from Δ^d to Δ^d. This means in particular that all the rows of all matrices $\mathbf{K}_i$ sum to 1. Therefore, each of these matrices have an eigenvalue 1. When we write "the norm of all eigenvalues of $\mathbf{K}_i$ is smaller than 1", we mean 1 is an eigenvalue of $\mathbf{K}_i$ and has multiplicity one; all other eigenvalues must be of norm strictly less than 1.

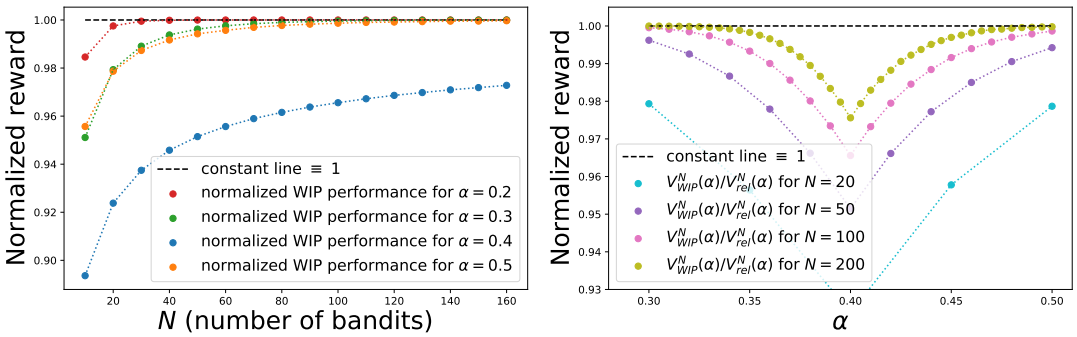
- All matrices $\mathbf{K}_i$ are locally stable because the eigenvalues of $\mathbf{K}_2$ are $\{1, -0.4\ldots, 0.08\ldots\}$ [3] while $\mathbf{K}_1 = \mathbf{P}^0$ and $\mathbf{K}_3 = \mathbf{P}^1$ are always stable matrices.

- For all tested values of α, we simulated $\Phi_t(\mathbf{m})$ from random initial points $\mathbf{m}$ and they all converge to the corresponding fixed point $\mathbf{m}^*$.

Moreover, as already shown in Figure 3.1, the fixed point $\mathbf{m}^*$ is singular when $\alpha = 0.4$, and it is non-singular for any other values of $\alpha \in [0.2, 0.5]$. This implies that all assumptions of Theorem 3.3.2 are satisfied when $\alpha \neq 0.4$. As $V_{\text{rel}}^{(N)}(\alpha)$ depends on the value of α, to make better comparisons, we consider the quantity $V_{\text{WIP}}^{(N)}(\alpha)/V_{\text{rel}}^{(N)}(\alpha)$, which is the normalized performance of WIP with respect to the relaxation upper-bound.

In Figure 3.2a, we choose four values of α as 0.2, 0.3, 0.4 and 0.5, and plot the normalized performances as a function of the number of arms N that takes values on multiples of 10. The value of $V_{\text{WIP}}^{(N)}(\alpha)$ are computed by using simulations. We repeat each simulation so that 95% confidence intervals become negligible and hence can not be seen from the pictures. In Figure 3.2b, this time we fix the value of N and plot the normalized performance as a function of α where α varies between $[0.3, 0.5]$ with a stepsize of $1/N$: $\alpha \in \{0.3, 0.3 + 1/N, 0.3 + 2/N, \ldots, 0.5\}$ (so that αN are always integers).

These two figures suggest that the convergence rate is related to how far $\mathbf{m}^*$ is away from the closest boundary of two zones (*i.e.* how non-singular it is). Here is an intuitive explanation for this phenomenon: the stochastic system in equilibrium will wander around the fixed point $\mathbf{m}^*$ that gives the optimal reward, now if $\mathbf{m}^*$ is near a boundary, it is more likely for the stochastic trajectory to jump into another neighboring polytope $\mathcal{Z}'$, in which case another affine drift applies and this may take the trajectory away from $\mathbf{m}^*$.



(a) Performance as a function of N.

(b) Performance as a function of α.

Figure 3.2 – Normalized performance of WIP for different values of α and N.

[3] We write $-0.4\ldots$ to mean a number that approaches -0.4.

3.4.3 Non integer values of αN

Our previous analysis rely on the assumption that αN is an integer. Let us briefly discuss in this subsection how to deal with non integer values of αN for the optimization problem (3.2) under (3.3). Consider the following three possible rounding procedure to replace the constraint (3.3):

- (*floor*) At each decision epoch, we activate $\lfloor \alpha N \rfloor$ arms;
- (*ceil*) At each decision epoch, we activate $\lceil \alpha N \rceil$ arms;
- (*probabilistic*) At each decision epoch, we activate $\lfloor \alpha N \rfloor$ arms, and one more arm is activated with probability $\{\alpha N\} := \alpha N - \lfloor \alpha N \rfloor$.

We denote by $V^{(N)}_{\mathrm{WIP}}(\lfloor N\alpha \rfloor/N)$, $V^{(N)}_{\mathrm{WIP}}(\lceil N\alpha \rceil/N)$ and $V^{(N)}_{\mathrm{WIP}}(\overline{\alpha})$ the reward of WIP under these three solutions. Note that these three values all coincide with our previous $V^{(N)}_{\mathrm{WIP}}(\alpha)$ if αN is an integer, but otherwise are different in general. Numerically, we discover that the average reward when always activating $\lfloor \alpha N \rfloor$ arms or always activating $\lceil \alpha N \rceil$ arms will be at distance $O(1)$ from the relaxation $V^{(N)}_{\mathrm{rel}}(\alpha)$.

Moreover, $V^{(N)}_{\mathrm{WIP}}(\overline{\alpha})$ converges at exponential rate to $V^{(N)}_{\mathrm{rel}}(\alpha)$. Here is an informal explanation: Let $\phi_{rounding}(\mathbf{m}) = \mathbb{E}\left[M^{(N)}_i(t+1) \mid \mathbf{M}^{(N)}(t) = \mathbf{m}\right]$ when any of the three *rounding* policy among *floor*, *ceil*, or *probabilistic* is used. When the rounding is probabilistic, it is not hard to show that $\phi_{probabilistic}(\mathbf{m}) = \phi(\mathbf{m})$, where $\phi(\cdot)$ is defined as in Equation (3.7) of the proof of Lemma 3.3.1. In contrast, $\phi_{floor}(\mathbf{m}) = \phi(\mathbf{m}) + O(\alpha - \lfloor \alpha N \rfloor/N)$. This shows that if the map ϕ has a unique non-singular attractor $\mathbf{m}^*$, then as N goes to infinity, the maps $\phi_{rounding}$ also have a unique non-singular attractor, that is equal to $\mathbf{m}^*$ for the *probabilistic* rounding and at distance $O(1/N)$ of $\mathbf{m}^*$ for *floor* or *ceil*. Moreover, the proof of Lemma 3.6.5 and Lemma 3.6.6 in the appendix can be adapted to obtain a concentration bound around $\phi_{rounding}$ for all policies. This guarantees an exponential convergence rate on the performance of WIP to performance on the attractor, for any of these three policies. Consequently we have $|V^{(N)}_{\mathrm{WIP}}(\overline{\alpha}) - V^{(N)}_{\mathrm{rel}}(\alpha)| \leq b \cdot e^{-CN}$, whereas $|V^{(N)}_{\mathrm{WIP}}(\lfloor N\alpha \rfloor/N) - V^{(N)}_{\mathrm{rel}}(\alpha)| = O(1/N)$.

To further illustrate these points, we consider in Figure 3.3 the same example as in Section 3.4.2, with $\alpha = 0.3$. As in Figure 5.1, the green curve represents $V^{(N)}_{\mathrm{WIP}}(\alpha)$ for N being a multiple of 10. Here, we extend this curve to all N being a multiple of 5, using the three possible rounding. The values of $V^{(N)}_{\mathrm{WIP}}(\lfloor N\alpha \rfloor/N)$, $V^{(N)}_{\mathrm{WIP}}(\overline{\alpha})$ and $V^{(N)}_{\mathrm{WIP}}(\lceil N\alpha \rceil/N)$ are plotted respectively in blue, green and red dots for $N \in \{25, 35, 45, \ldots\}$, while their values coincide for N being a multiple of 10 (which explains the zigzag of the orange and blue curves). We observe that the differences $V^{(1)}_{\mathrm{rel}}(\alpha) - V^{(N)}_{\mathrm{WIP}}(\lfloor N\alpha \rfloor/N)$ and $V^{(1)}_{\mathrm{rel}}(\alpha) - V^{(N)}_{\mathrm{WIP}}(\lceil N\alpha \rceil/N)$ converge to $\pm 0.5 \cdot (R^1_1 - R^0_1)$ when $N \to \infty$ and $\{N\alpha\} \equiv 0.5$, *i.e.* $N = 5 \cdot (2k+1)$. The behavior is quite different for the probabilistic rounding (green curve). Indeed, in this case we cannot distinguish when αN is an integer or not. This indicates that $V^{(N)}_{\mathrm{WIP}}(\overline{\alpha})$ indeed converges at exponential rate to $V^{(N)}_{\mathrm{rel}}(\alpha)$.

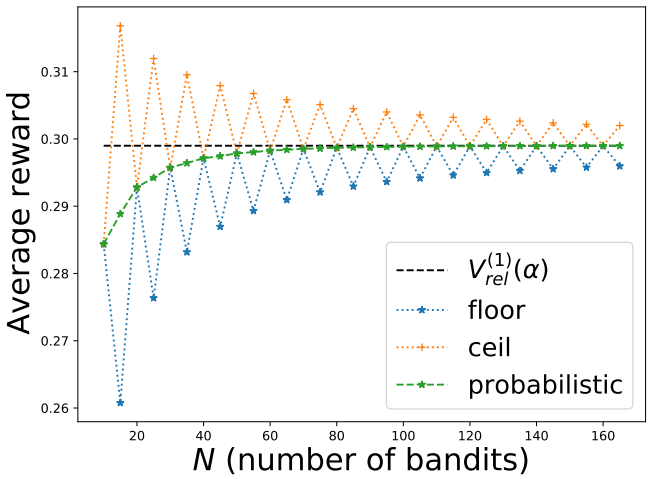


Figure 3.3 – Performance of the three policies for non integer values of αN.

3.5 Application: Markovian Fading Channels

The Markovian fading channel is a typical discrete time restless bandit model. Strictly speaking, this model has a countable infinite state space, so some approximation is needed, as we discuss later. In Ouyang et al. [41] a two-classes channel problem has been studied. By using the same scaling as here, the authors of Ouyang et al. [41] have proven the asymptotic optimality of WIP for this model, after verifying the global attractor property of the deterministic system. In this section we take a step further, evaluate numerically the convergence rate of the performance, and verify if it is exponential, as claimed in Theorem 3.3.2.

Let us first briefly review this two-class channel model (more details can be found in Ouyang et al. [41]). A Gilbert-Elliott channel is modeled as a two-states Markov chain with a bad state 0 and a good state 1. Two classes of channels are available, with the transition probability matrices for class $k \in \{1,2\}$ being $\begin{pmatrix} p_k & 1-p_k \\ r_k & 1-r_k \end{pmatrix}$, where p_k is the probability of a class k channel being in good state at time $t+1$ if it was in good state at time t, and r_k is the probability being in good state if one time step ago it was in bad state. We assume the channels are *positively correlated*, namely $p_k > r_k$ for $k = 1, 2$.

We consider a total population of N channels, a proportion β of them are from class 1. Due to limited resource, each time we can only activate a proportion α of the channels, and only a channel in good state under activation can transmit data. We assume that we can observe the state of a channel only when it is activated. Otherwise, we keep track of the state of a channel by using a belief value $b^k_{s,t}$ where $k = 1, 2$, $s = 0, 1$ and $t \geq 1$. The value $b^k_{s,t}$ is the probability for a class k channel to be in good state, provided that it was activated (hence observed) t time steps ago and was observed to be in state s. The expression of $b^k_{s,t}$ is

$$b^k_{0,t} = \frac{r_k - (p_k - r_k)^t r_k}{1 + r_k - p_k}, \; b^k_{1,t} = \frac{r_k + (1-p_k)(p_k - r_k)^t}{1 + r_k - p_k}.$$

To cast this channel model into a discrete time restless bandit problem, we treat each channel as an arm, and its state space is the whole set of possible values of $b^k_{s,t}$'s. The transition matrices $\mathbf{P}^0$, $\mathbf{P}^1$ can then be naturally written down:

$$\mathbb{P}^0(b^k_{s,t}, b^k_{s,t+1}) = 1, \qquad \mathbb{P}^1(b^k_{s,t}, b^k_{1,1}) = b^k_{s,t}, \qquad \mathbb{P}^1(b^k_{s,t}, b^k_{0,1}) = 1 - b^k_{s,t},$$

all other probabilities being 0.

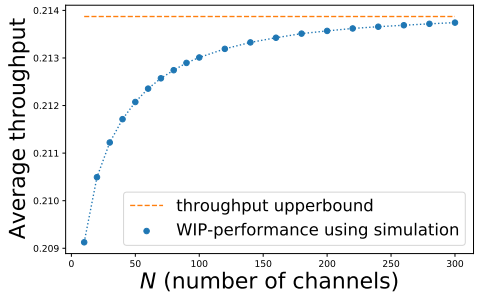
We evaluate the performance by the throughput of the system, hence we obtain a reward of 1 each time we activate a channel *and* it is in good state. Under the MDP framework, this is equivalent to assuming that state $b^k_{s,t}$ gives a reward $b^k_{s,t}$ under activation. It is shown in Ouyang et al. [41] that this problem is indexable, and that Whittle index can be calculated explicitly (via techniques due to the specific structure of the model). The index of a state $b^k_{s,t}$ is denoted by $\nu(b^k_{s,t})$ and is equal to:

$$\nu(b^k_{s,t}) = \begin{cases} \frac{(b^k_{0,t} - b^k_{0,t+1})(t+1) + b^k_{0,t+1}}{1 - p_k + (b^k_{0,t} - b^k_{0,t+1})t + b^k_{0,t+1}}, & \text{if } s = 0 \\ \frac{r_k}{(1-p_k)(1+r_k-p_k)+r_k}, & \text{otherwise.} \end{cases}$$

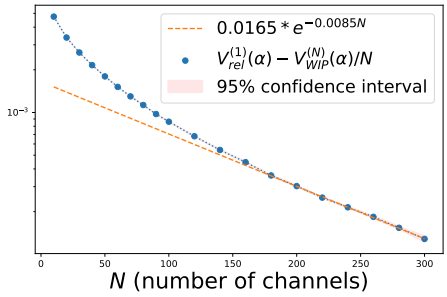
We remark that for $k = 1, 2$, the index value $\nu(b^k_{0,t})$ is an increasing function of t, and furthermore $\nu(b^k_{0,t}) \xrightarrow{t\to\infty} r_k/((1-p_k)(1+r_k-p_k)+r_k) = \nu(b^k_{1,t'})$, for any $t' \geq 1$. We shall also point out that the relative orders of the index values $\nu\left(b^k_{s,t}\right)$ between two classes $k = 1$ and $k = 2$ could be different from the orders of the belief values $b^k_{s,t}$. This indicates an interaction between classes and makes the Whittle indices for this model interesting.

The reader might have noticed that to apply Theorem 3.3.2, two assumptions are violated: first, the restless bandit model we consider here has a countable infinite state space; second, not all arms are identical (there are two classes of arms). The first point might raise some technical difficulties that we have not encountered on our previous finite state model. However, it can be shown that the states $b^k_{0,t}$ for t large are extremely rarely visited, hence using a threshold t^* and ignoring all states $b^k_{s,t}$ with $t > t^*$ (*i.e.* treating them as b^k_{s,t^*}) makes a negligible difference. Concerning the two classes of arm, we argue that having two classes of arms can be represented by a single class of arm by considering a larger state-space: the state of an arm would be $(k, b^k_{s,t})$, where k is its class and $b^k_{s,t}$ is its belief value. Compared to our model, in this new case, the arms are no longer recurrent as an arm of class k cannot become an arm of class $k' \neq k$. This implies that the quantities $V^{(N)}_{\mathrm{WIP}}(\alpha)$ and $V^{(N)}_{\mathrm{rel}}(\alpha)$ will depend on the initial condition of the system, *i.e.* on the fraction β of arms that are in class 1. Apart from that, our results apply mutatis mutandis to this case.

We can now provide some numerical results. We shall choose a parameter set that is used in Ouyang et al. [41]: $\beta = 0.6$, $\alpha = 0.3$, $(p_1, r_1) = (0.75, 0.2)$, $(p_2, r_2) = (0.8, 0.3)$. It can be shown that using these parameters, a class 2 channel that has just been activated and has been observed in good state will have the highest priority, hence should always be activated. Also a class 2 channel after 4 time steps of being idle has higher priority than a class 1 channel in any belief state. We can then characterize the fixed point $\mathbf{m}^*$



(a) WIP for two-classes channel model.



(b) The exponential convergence.

Figure 3.4 – Convergence rate for two-classes channel model.

by computing a threshold of activation of class 1 channels so that in steady-state, a proportion of $\alpha = 0.3$ of channels are activated. This gives that all class 1 channels in belief state $b^1_{0,t}$ with $t \leq 20$ will be kept idle, a fraction $0.89\ldots$ of the class 1 channels in belief state $b^1_{0,21}$ will be activated, and all class 1 channels in belief states $b^1_{0,t}$ with $t \geq 22$ will be activated. As $0.89\cdots \neq 1$, the fixed point is *not* singular.

Consequently, all conditions needed for Theorem 3.3.2 are satisfied for this model. We then use simulations to evaluate the average throughput, with N ranging from 10 to 300. We see through Figure 3.4 that a similar convergence pattern as in the 3 states model occurs, and it suggests an exponential rate convergence as claimed, with a value of the constant $c \approx 0.0085$.

3.6 Proofs of the Main Theorems

3.6.1 Proof of Lemma 3.3.1

In this section we prove Lemma 3.3.1. We first show the piecewise affine property in Lemma 3.6.1, which gives (i) and (ii). We then show the uniqueness of fixed point from a bijective property in Lemma 3.6.2, from which we conclude (iii).

Lemma 3.6.1 (Piecewise affine)**.** *ϕ is a piecewise affine continuous function, with d affine pieces.*

Proof. Let $\mathbf{m} \in \Delta^d$ be a configuration and recall $s(\mathbf{m}) \in \{1\ldots d\}$ is the state such that $\sum_{i=1}^{s(\mathbf{m})-1} m_i \leq \alpha < \sum_{i=1}^{s(\mathbf{m})} m_i$. When the system is in configuration $\mathbf{m}$ at time t, WIP will activate all arms that are in states 1 to $s(\mathbf{m}) - 1$ and not activate any arm in states $s(\mathbf{m}) + 1$ to d. Among the $Nm_{s(\mathbf{m})}$ arms in state $s(\mathbf{m})$, $N(\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i)$ of them will be activated and the rest will not be activated.

This implies that the expected number of arms in state j at time $t + 1$ will be equal to

$$\sum_{i=1}^{s(\mathbf{m})-1} Nm_i P^1_{ij} + N(\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i)P^1_{s(\mathbf{m})j} + N(\sum_{i=1}^{s(\mathbf{m})} m_i - \alpha)P^0_{s(\mathbf{m})j} + \sum_{i=s(\mathbf{m})+1}^{d} Nm_i P^0_{ij}. \quad (3.9)$$

It justifies the expression (3.7). Note that (3.7) can be reorganized to

$$\phi_j(\mathbf{m}) = \sum_{i=1}^{s(\mathbf{m})-1} m_i(P^1_{ij} - P^1_{s(\mathbf{m})j} + P^0_{s(\mathbf{m})j}) + \sum_{i=s(\mathbf{m})}^{d} m_i P^0_{ij} + \alpha(P^1_{s(\mathbf{m})j} - P^0_{s(\mathbf{m})j}).$$

Consequently $\phi(\mathbf{m}) = \mathbf{m} \cdot \mathbf{K}_{s(\mathbf{m})} + \mathbf{b}_{s(\mathbf{m})}$, where

$$\mathbf{b}_{s(\mathbf{m})} = \alpha(\mathbf{P}^1_{s(\mathbf{m})} - \mathbf{P}^0_{s(\mathbf{m})}), \text{ and } \mathbf{K}_{s(\mathbf{m})} = \begin{pmatrix} \mathbf{P}^1_1 - \mathbf{P}^1_{s(\mathbf{m})} + \mathbf{P}^0_{s(\mathbf{m})} \\ \mathbf{P}^1_2 - \mathbf{P}^1_{s(\mathbf{m})} + \mathbf{P}^0_{s(\mathbf{m})} \\ \cdots \\ \mathbf{P}^1_{s(\mathbf{m})-1} - \mathbf{P}^1_{s(\mathbf{m})} + \mathbf{P}^0_{s(\mathbf{m})} \\ \mathbf{P}^0_{s(\mathbf{m})} \\ \mathbf{P}^0_{s(\mathbf{m})+1} \\ \cdots \\ \mathbf{P}^0_d \end{pmatrix}.$$

Let $\mathcal{Z}_i := \{\mathbf{m} \in \Delta^d \mid s(\mathbf{m}) = i\}$. The above expression of ϕ implies that this map is affine on each zone $\mathcal{Z}_i$. There are d such zones with $1 \le i \le d$. It is clear from the expression that $\phi(\mathbf{m})$ is continuous on $\mathbf{m}$.

□

Lemma 3.6.2 (Bijective property)**.** *Let $\pi(s,\theta) \in \Pi$ be the policy that activates all arms in states $1,\dots,s-1$, does not activate arms in states $s+1, s+2,\dots,d$, and that activates arms in state s with probability θ. Denote by $\tilde{\alpha}(s,\theta)$ the proportion of time that the active action is taken using policy $\pi(s,\theta)$. Then, the function $(s,\theta) \mapsto \tilde{\alpha}(s,\theta)$ is a bijective map from $\{1\dots d\} \times [0,1)$ to $[0,1)$.*

Proof. The following proof is partially adapted from the proof of Weber and Weiss [48, Lemma 1]. For a given $\nu \in \mathbb{R}$, denote by $\gamma(\nu)$ the value of the subsidy-ν problem, *i.e.*

$$\gamma(\nu) := \max_{\pi\in\Pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1}\Big(R^{\pi(S(t))}_{S(t)} + \nu\big(1-\pi(S(t))\big)\Big)\Big]. \tag{3.10}$$

We defined similarly $\gamma_\pi(\nu)$ as the value under policy π for a such subsidy-ν problem. Note that for fixed π, the function $\gamma_\pi(\nu)$ is affine and increasing in ν.

By definition of indexability, $\gamma(\nu) = \max_{\pi\in\Pi}\gamma_\pi(\nu)$ is a piecewise affine, continuous and convex function of ν: it is affine on $(-\infty;\nu_d]$, on $[\nu_1;+\infty)$ and on all $[\nu_s;\nu_{s-1}]$ for $s \in \{2\dots d\}$.

Moreover, for $s \in \{2\dots d-1\}$ and $\nu \in [\nu_s;\nu_{s-1}]$, the optimal policy of (3.10) is to activate all arms up to state $s-1$. Hence,

$$\gamma(\nu) = \gamma_{\pi(s,0)}(\nu) = \gamma(\nu_{s-1}) + \big(1-\tilde{\alpha}(s,0)\big)\cdot(\nu-\nu_{s-1}).$$

Similarly, and as $\tilde{\alpha}(s+1,0) = \tilde{\alpha}(s,1)$, for $\nu \in [\nu_{s+1};\nu_s]$ we have:

$$\begin{aligned}\gamma(\nu) &= \gamma(\nu_s) + \big(1-\tilde{\alpha}(s+1,0)\big)\cdot(\nu-\nu_s)\\ &= \gamma(\nu_s) + \big(1-\tilde{\alpha}(s,1)\big)\cdot(\nu-\nu_s).\end{aligned}$$

Consequently

$$\frac{\partial \gamma}{\partial \nu}(\nu) = \begin{cases} 1 - \tilde{\alpha}(s,0), & \text{if } \nu_s < \nu < \nu_{s-1} \\ 1 - \tilde{\alpha}(s,1), & \text{if } \nu_{s+1} < \nu < \nu_s. \end{cases}$$

The convexity of $\gamma(\nu)$ implies that $1 - \tilde{\alpha}(s,0) > 1 - \tilde{\alpha}(s,1)$, hence $\tilde{\alpha}(s,1) > \tilde{\alpha}(s,0)$.

Now suppose that $\mathbf{m}^0$ and $\mathbf{m}^1$ are the equilibrium distributions of policies $\pi(s,0)$ and $\pi(s,1)$. Let $0 < \theta < 1$. The equilibrium distribution $\mathbf{m}^\theta$ induced by $\pi(s,\theta)$ is then a linear combination of $\mathbf{m}^0$ and $\mathbf{m}^1$, namely $\mathbf{m}^\theta = p \cdot \mathbf{m}^0 + (1-p) \cdot \mathbf{m}^1$, with

$$p = \frac{(1-\theta)m_s^1}{\theta m_s^0 + (1-\theta)m_s^1}.$$

Hence

$$\begin{aligned} m_s^\theta &= p m_s^0 + (1-p) m_s^1 \\ &= \frac{m_s^1 m_s^0}{\theta m_s^0 + (1-\theta) m_s^1}, \end{aligned}$$

and

$$\begin{aligned} \tilde{\alpha}(s,\theta) &= \left(\sum_{k=1}^{s-1} m_k^\theta \right) + \theta m_s^\theta \\ &= \sum_{k=1}^{s-1} \left((1-p) m_k^1 + p m_k^0 \right) + \frac{\theta \cdot m_s^1 m_s^0}{\theta m_s^0 + (1-\theta) m_s^1} \\ &= \frac{\sum_{k=1}^{s-1} \left(\theta \cdot m_s^0 m_k^1 + (1-\theta) m_s^1 m_k^0 \right) + \theta \cdot m_s^1 m_s^0}{\theta m_s^0 + (1-\theta) m_s^1}. \end{aligned}$$

Observe that $\tilde{\alpha}(s,\theta)$ is the ratio of two affine functions of θ, hence is monotone as θ ranges from 0 to 1; but as $\tilde{\alpha}(s,1) > \tilde{\alpha}(s,0)$, it is monotonically *increasing*. We hence obtain

$$1 = \tilde{\alpha}(d,1) > \tilde{\alpha}(d,0) = \tilde{\alpha}(d-1,1) > \cdots > \tilde{\alpha}(2,0) = \tilde{\alpha}(1,1) > \tilde{\alpha}(1,0) = 0,$$

which concludes the proof. □

We are now ready to finish the proof of Lemma 3.3.1(iii). Let $\mathbf{m}$ be a fixed point of the continuous map ϕ (that exists by Brouwer's fixed-point theorem). Under configuration $\mathbf{m}$, all arms that are in states from 1 to $s(\mathbf{m}) - 1$ are activated, and a fraction $\theta(\mathbf{m}) = (\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i)/m_{s(\mathbf{m})}$ of the arms that are in state $s(\mathbf{m})$ are activated. This shows that $\mathbf{m}$ also corresponds to the stationary distribution of the policy $\pi(s(\mathbf{m}), \theta(\mathbf{m}))$. The proportion of activated arms of this policy is $\tilde{\alpha}(s(\mathbf{m}), \theta(\mathbf{m})) = \alpha$. Consequently, if $\mathbf{m}'$ is another fixed point of ϕ, then $\mathbf{m}'$ would have to be the stationary distribution of some other policy of the form $\pi(s', \theta')$, with $\tilde{\alpha}(s', \theta') = \alpha$. As the function $(s,\theta) \mapsto \tilde{\alpha}(s,\theta)$ is a bijection, this implies that $s' = s(\mathbf{m})$ and $\theta' = \theta(\mathbf{m})$. Hence the fixed point of ϕ is unique.

3.6.2 Proof of Theorem 3.3.2

In this section, we explain technical details of the proof of our main result Theorem 3.3.2. In the following, we denote by $\mathcal{B}(\mathbf{m}^*, r)$ the ball centered at $\mathbf{m}^*$ with radius r.

Theorem 3.6.3. *Under the same assumptions as in Theorem 3.3.2, and assume that $\mathbf{M}^{(N)}(0)$ is already in stationary regime. Then there exists two constants $b, c > 0$ such that*

(i) $\|\mathbb{E}[\mathbf{M}^{(N)}(0)] - \mathbf{m}^*\| \leq b \cdot e^{-cN}$;

(ii) $\mathbb{P}\left[\mathbf{M}^{(N)}(0) \notin \mathcal{Z}_{s(\mathbf{m}^*)}\right] \leq b \cdot e^{-cN}$.

Let us first explain how Theorem 3.6.3 implies Theorem 3.3.2. We prove below that:

Lemma 3.6.4. *Assume that bandits are indexable, and let $\rho(\mathbf{m})$ be the instantaneous arm-averaged reward of WIP when the system is in configuration $\mathbf{m}$. Then:*

(i) ρ is piecewise affine on each of the zone $\mathcal{Z}_i$ and for all $\mathbf{m} \in \Delta^d$:

$$\rho(\mathbf{m}) = \sum_{i=1}^{s(\mathbf{m})-1} m_i R_i^1 + (\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i) R_{s(\mathbf{m})}^1 + (\sum_{i=1}^{s(\mathbf{m})} m_i - \alpha) R_{s(\mathbf{m})}^0 + \sum_{i=s(\mathbf{m})+1}^{d} m_i R_i^0. \tag{3.11}$$

(ii) $\rho(\mathbf{m}^*) = V_{\text{rel}}^{(1)}(\alpha)$.

By definition, the performance of WIP is $V_{\text{WIP}}^{(N)}(\alpha) = N \cdot \mathbb{E}\left[\rho(\mathbf{M}^{(N)}(0))\right]$. Hence from Lemma 3.6.4 we have

$$\begin{aligned} V_{\text{rel}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha) &= N \cdot V_{\text{rel}}^{(1)}(\alpha) - N \cdot \mathbb{E}\left[\rho(\mathbf{M}^{(N)}(0))\right] \\ &= N \cdot \mathbb{E}\Big[\big(\rho(\mathbf{m}^*) - \rho(\mathbf{M}^{(N)}(0))\big)\mathbb{1}_{\{\mathbf{M}^{(N)}(0) \in \mathcal{Z}_{s(\mathbf{m}^*)}\}} \\ &\qquad + \big(\rho(\mathbf{m}^*) - \rho(\mathbf{M}^{(N)}(0))\big)\mathbb{1}_{\{\mathbf{M}^{(N)}(0) \notin \mathcal{Z}_{s(\mathbf{m}^*)}\}}\Big] \end{aligned}$$

By linearity of ρ and Theorem 3.6.3(i), the first term inside the above expectation is exponentially small; by Theorem 3.6.3(ii) and since the rewards are bounded, the second term is also exponentially small.

Before proving Theorem 3.6.3, we start by proving a few technical lemmas.

3.6.3 Relation between $\mathbf{m}^*$ and $V_{\text{rel}}^{(1)}(\alpha)$ (Proof of Lemma 3.6.4)

Proof. Let $\mathbf{m} \in \Delta^d$ be a configuration and recall $s(\mathbf{m}) \in \{1 \dots d\}$ is the state such that $\sum_{i=1}^{s(\mathbf{m})-1} m_i \leq \alpha < \sum_{i=1}^{s(\mathbf{m})} m_i$. Similarly to our analysis of Lemma 3.6.1, when the system is in configuration $\mathbf{m}$, WIP will activate all arms that are in states 1 to $s(\mathbf{m}) - 1$. This

will lead an instantaneous reward of $\sum_{i=1}^{s(\mathbf{m})-1} N m_i R_i^1$. WIP will not activate arms that are in states $s(\mathbf{m})+1$ to d. This will lead an instantaneous reward of $\sum_{i=s(\mathbf{m})+1}^{d} N m_i R_i^0$. Among the $N m_{s(\mathbf{m})}$ arms in state $s(\mathbf{m})$, $N(\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i)$ of them will be activated and the rest will not be activated. This shows that $\rho(\mathbf{m})$ is given by (3.11).

For (ii), recall that $\mathbf{m}^*$ is the unique fixed point, and consider a subsidy-$\nu_{s(\mathbf{m}^*)}$ MDP, where $\nu_{s(\mathbf{m}^*)}$ is the Whittle index of state $s(\mathbf{m}^*)$. Denote by L the value of this MDP:

$$\begin{aligned} L &:= \max_{\Pi} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[R_{S_n(t)}^{a_n(t)} + (\alpha - a_n(t)) \nu_{s(\mathbf{m}^*)} \right] \\ &= \max_{\Pi} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[R_{S_n(t)}^{a_n(t)} \right] + \left(\alpha - \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[a_n(t) \right] \right) \nu_{s(\mathbf{m}^*)}. \end{aligned} \tag{3.12}$$

By definition of Whittle index, any policy of the form $\pi(s(\mathbf{m}^*), \theta)$ defined in Lemma 3.6.2 is optimal for (3.12). Moreover, if θ^* is such that $\tilde{\alpha}(s(\mathbf{m}^*), \theta^*) = \alpha$, then such a policy satisfies the constraint (3.5): $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[a_n(t)\right] = \alpha$. This shows that $L = V_{\text{rel}}^{(1)}(\alpha)$ and as all arms are identical, we have $N \cdot V_{\text{rel}}^{(1)}(\alpha) = V_{\text{rel}}^{(N)}(\alpha)$, and $\pi(s(\mathbf{m}^*), \theta^*)$ is an optimal policy for the relaxed constraint (3.5).

It remains to show that the reward of policy $\pi(s(\mathbf{m}^*), \theta^*)$ is $\rho(\mathbf{m}^*)$. This comes from the fact that the steady-state of the Markov chain induced by this policy is $\mathbf{m}^*$, and $\pi(s(\mathbf{m}^*), \theta^*)$ is such that αN arms are activated on average. Indeed, the arm-averaged reward of this policy is:

$$L = \sum_{i=1}^{s(\mathbf{m}^*)-1} m_i^* R_i^1 + \theta^* m_{s(\mathbf{m}^*)}^* R_{s(\mathbf{m}^*)}^1 + (1-\theta^*) m_{s(\mathbf{m}^*)}^* R_{s(\mathbf{m}^*)}^0 + \sum_{i=s(\mathbf{m}^*)+1}^{d} m_i^* R_i^0 \tag{3.13}$$

As the proportion of activated arms is α, we have $\sum_{i=1}^{s(\mathbf{m}^*)-1} m_i^* + \theta^* m_{s(\mathbf{m}^*)}^* = \alpha$. Hence (3.13) coincides with the expression of $\rho(\mathbf{m}^*)$ in (3.11), and $\rho(\mathbf{m}^*) = L = V_{\text{rel}}^{(1)}(\alpha)$. This concludes the proof of Lemma 3.6.4. □

3.6.4 Hoeffding's inequality (for one transition)

Lemma 3.6.5 (Hoeffding's inequality)**.** *For all $t \in \mathbb{N}$, we have*

$$\mathbf{M}^{(N)}(t+1) = \phi\left(\mathbf{M}^{(N)}(t)\right) + \mathbf{E}^{(N)}(t+1)$$

where the random vector $\mathbf{E}^{(N)}(t+1)$ is such that

$$\mathbb{E}[\mathbf{E}^{(N)}(t+1) \big| \mathbf{M}^{(N)}(t)] = \mathbf{0},$$

and for all $\delta > 0$:

$$\mathbb{P}\left[\|\mathbf{E}^{(N)}(t+1)\| \geq \delta \right] \leq d \cdot e^{-2N\delta^2}.$$

Proof. Since the N arms evolve independently, we may apply the following form of Hoeffding's inequality: Let $X_1, X_2, ..., X_N$ be N independent random variables bounded

by the interval $[0,1]$, and define the empirical mean of these variables by $\overline{X} := \frac{1}{N}(X_1 + X_2 + ... + X_N)$, then

$$\mathbb{P}\left[\overline{X} - \mathbb{E}[\overline{X}] \geq \delta\right] \leq e^{-2N\delta^2}.$$

More precisely, for a fixed $1 \leq j \leq d$, we have

$$M_j^{(N)}(t+1) = \frac{1}{N}\sum_{i=1}^{d}\sum_{k=1}^{N\cdot M_i^{(N)}(t)} \mathbb{1}_{\{U_{i,k}\leq P_{ij}(\mathbf{M}^{(N)}(t))\}}$$

where for $1 \leq i \leq d,\ 1 \leq k \leq N \cdot M_i^{(N)}(t)$, the $U_{i,k}$'s are in total N independent and identically distributed uniform $(0,1)$ random variables, and $P_{ij}(\mathbf{m})$ is the probability for an arm in state i goes to state j under WIP, when the N arms system is in configuration $\mathbf{m}$.

By definition, we have

$$\phi_j(\mathbf{M}^{(N)}(t)) = \sum_{i=1}^{d} M_i^{(N)}(t) \cdot P_{ij}(\mathbf{M}^{(N)}(t)).$$

Hence

$$\mathbb{E}\left[M_j^{(N)}(t+1)\middle|\mathbf{M}^{(N)}(t)\right] = \sum_{i=1}^{d}\frac{1}{N}\cdot N \cdot M_i^{(N)}(t)\cdot P_{ij}(\mathbf{M}^{(N)}(t)) = \phi_j(\mathbf{M}^{(N)}(t)),$$

and

$$\begin{aligned}\mathbb{P}\left[\|\mathbf{M}^{(N)}(t+1) - \phi(\mathbf{M}^{(N)}(t))\| \geq \delta\right] &= \mathbb{P}\left[\max_{1\leq j\leq d}\left|M_j^{(N)}(t+1) - \phi_j(\mathbf{M}^{(N)}(t))\right| \geq \delta\right] \\ &\leq d\cdot e^{-2N\delta^2},\end{aligned}$$

where the last inequality comes from the union bound and the above form of Hoeffding's inequality. □

3.6.5 Hoeffding's inequality (for t transitions)

Lemma 3.6.6. *There exists a positive constant K such that for all $t \in \mathbb{N}$ and for all $\delta > 0$,*

$$\mathbb{P}\left[\|\mathbf{M}^{(N)}(t+1) - \Phi_{t+1}(\mathbf{m})\| \geq (1+K+\ldots+K^t)\delta \;\middle|\; \mathbf{M}^{(N)}(0) = \mathbf{m}\right] \leq d(t+1)\cdot e^{-2N\delta^2}.$$

Proof. Since ϕ is a piecewise affine function with finite affine pieces, in particular ϕ is K-Lipschitz: there is a constant $K > 0$ such that for all $\mathbf{m}_1, \mathbf{m}_2 \in \Delta^d$:

$$\|\phi(\mathbf{m}_1) - \phi(\mathbf{m}_2)\| \leq K \cdot \|\mathbf{m}_1 - \mathbf{m}_2\|.$$

Let $t \in \mathbb{N}$ and $\mathbf{m} \in$ be fixed, we have

$$\|\mathbf{M}^{(N)}(t+1) - \Phi_{t+1}(\mathbf{m})\| \leq \|\mathbf{M}^{(N)}(t+1) - \phi(\mathbf{M}^{(N)}(t))\| + \|\phi(\mathbf{M}^{(N)}(t)) - \phi(\Phi_t(\mathbf{m}))\|$$

$$\leq \|\mathbf{M}^{(N)}(t+1) - \phi(\mathbf{M}^{(N)}(t))\| + K \cdot \|\mathbf{M}^{(N)}(t) - \Phi_t(\mathbf{m})\|.$$

By iterating the above inequality, we obtain

$$\begin{aligned}
&\|\mathbf{M}^{(N)}(t+1) - \Phi_{t+1}(\mathbf{m})\| \\
&\leq \|\mathbf{M}^{(N)}(t+1) - \phi(\mathbf{M}^{(N)}(t))\| + K \cdot \|\mathbf{M}^{(N)}(t) - \phi(\mathbf{M}^{(N)}(t-1))\| \\
&\quad + K^2 \cdot \|\mathbf{M}^{(N)}(t-1) - \Phi_{t-1}(\mathbf{m})\| \\
&\leq \sum_{s=0}^{t} K^s \cdot \|\mathbf{M}^{(N)}(t+1-s) - \phi(\mathbf{M}^{(N)}(t-s))\|,
\end{aligned}$$

where for each $0 \leq s \leq t$, we have by lemma 3.6.5: for all $\delta > 0$,

$$\mathbb{P}\left[\|\mathbf{M}^{(N)}(t+1-s) - \phi(\mathbf{M}^{(N)}(t-s))\| \geq \delta\right] \leq d \cdot e^{-2N\delta^2}.$$

Hence, using the union bound, we obtain

$$\begin{aligned}
&\mathbb{P}\left[\|\mathbf{M}^{(N)}(t+1) - \Phi_{t+1}(\mathbf{m})\| \geq (1 + K + K^2 + \ldots + K^t)\delta \,\middle|\, \mathbf{M}^{(N)}(0) = \mathbf{m}\right] \\
&\leq \mathbb{P}\left[\sum_{s=0}^{t} K^s \cdot \|\mathbf{M}^{(N)}(t+1-s) - \phi(\mathbf{M}^{(N)}(t-s))\| \geq (1 + K + K^2 + \ldots + K^t)\delta\right] \\
&\leq \mathbb{P}\left[\bigcup_{s=0}^{t} \left\{\|\mathbf{M}^{(N)}(t+1-s) - \phi(\mathbf{M}^{(N)}(t-s))\| \geq \delta\right\}\right] \\
&\leq \sum_{s=0}^{t} \mathbb{P}\left[\|\mathbf{M}^{(N)}(t+1-s) - \phi(\mathbf{M}^{(N)}(t-s))\| \geq \delta\right] \\
&\leq d(t+1) \cdot e^{-2N\delta^2},
\end{aligned}$$

and this ends the proof of Lemma 3.6.6. □

3.6.6 Exponential stability of $\mathbf{m}^*$

Lemma 3.6.7. *Under the assumptions of Theorem 3.3.2, there exists constants $b_1, b_2 > 0$ such that for all $t \geq 0$ and all $\mathbf{m} \in \Delta^d$:*

$$\|\Phi_t(\mathbf{m}) - \mathbf{m}^*\| \leq b_1 \cdot e^{-b_2 t} \cdot \|\mathbf{m} - \mathbf{m}^*\|. \tag{3.14}$$

Proof. As ϕ is locally stable, for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $\|\mathbf{m} - \mathbf{m}^*\| \leq \delta$, then for all $t \geq 0$: $\|\Phi_t(\mathbf{m}) - \mathbf{m}^*\| \leq \varepsilon$. Recall that for all $\mathbf{m} \in \mathcal{Z}_{s(\mathbf{m}^*)}$, we have $\phi(\mathbf{m}) = (\mathbf{m} - \mathbf{m}^*) \cdot \mathbf{K}_{s(\mathbf{m}^*)} + \mathbf{m}^*$. We choose $\varepsilon > 0$ so that $\mathcal{B}(\mathbf{m}^*, \varepsilon) \subset \mathcal{Z}_{s(\mathbf{m}^*)}$.

Let us now show that there exists $T > 0$ such that for all $\mathbf{m} \in \Delta^d$, $\Phi_T(\mathbf{m}) \in \mathcal{B}(\mathbf{m}^*, \varepsilon)$. We shall reason by contradiction: If this is not true, then there exists a sequence of $t \in \mathbb{N}$ that goes to infinity and a corresponding $\{\mathbf{m}_t\}_t$ such that $\|\Phi_t(\mathbf{m}_t) - \mathbf{m}^*\| \geq \varepsilon$. As Δ^d is a compact space, there exists a subsequence of $\{\mathbf{m}_t\}_t$ (denoted again as $\{\mathbf{m}_t\}_t$) that converges to an element $\bar{\mathbf{m}}$. On the other hand, as $\mathbf{m}^*$ is an attractor, there exists T_1 such that $\Phi_{T_1}(\bar{\mathbf{m}}) \in \mathcal{B}(\mathbf{m}^*, \delta/2)$. And since $\Phi_{T_1}(\cdot)$ is continuous, there exists $\eta > 0$

such that if $\|\mathbf{m} - \bar{\mathbf{m}}\| \leq \eta$, then $\|\Phi_{T_1}(\mathbf{m}) - \Phi_{T_1}(\bar{\mathbf{m}})\| \leq \delta/2$. As $\{\mathbf{m}_t\}_t$ converges to $\bar{\mathbf{m}}$, there exists T_2 such that for $t \geq T_2$, we have $\|\mathbf{m}_t - \bar{\mathbf{m}}\| \leq \eta$. Consequently for $t \geq T_2$, we have

$$\|\Phi_{T_1}(\mathbf{m}_t) - \mathbf{m}^*\| \leq \|\Phi_{T_1}(\mathbf{m}_t) - \Phi_{T_1}(\bar{\mathbf{m}})\| + \|\Phi_{T_1}(\bar{\mathbf{m}}) - \mathbf{m}^*\| \leq \delta.$$

Hence for $t \geq \max(T_1, T_2)$, by our choice of ε and δ from the local stability of ϕ, we deduce that

$$\|\Phi_t(\mathbf{m}_t) - \mathbf{m}^*\| = \|\Phi_{t-T_1}(\Phi_{T_1}(\mathbf{m}_t)) - \mathbf{m}^*\| \leq \varepsilon.$$

This gives a contradiction! Consequently, there exists T such that for all $\mathbf{m} \in \Delta^d$, $\Phi_T(\mathbf{m}) \in \mathcal{B}(\mathbf{m}^*, \varepsilon)$. This implies in particular that $\mathbf{K}_{s(\mathbf{m}^*)}$ is a stable matrix: the modules of all its eigenvalues are smaller than one. Moreover, we have for all $\mathbf{m} \in \Delta^d$ and $t \geq T$:

$$\Phi_t(\mathbf{m}) = \left(\Phi_T(\mathbf{m}) - \mathbf{m}^*\right) \cdot \mathbf{K}_{s(\mathbf{m}^*)}^{t-T} + \mathbf{m}^*.$$

As $\mathcal{Z}_{s(\mathbf{m}^*)}$ is a stable matrix, this implies that (3.14) holds for all $\mathbf{m} \in \Delta^d$. □

3.6.7 Proof of Theorem 3.6.3

We are now ready to prove the main theorem.

Proof. The proof consists of several parts.

Choice of a neighborhood $\mathcal{N}$

The fixed point $\mathbf{m}^*$ is in zone $\mathcal{Z}_{s(\mathbf{m}^*)}$ in which ϕ can be written as

$$\phi(\mathbf{m}) = (\mathbf{m} - \mathbf{m}^*) \cdot \mathbf{K}_{s(\mathbf{m}^*)} + \mathbf{m}^*.$$

As $\mathbf{m}^*$ is not singular, let $\mathcal{N}_1$ be a neighborhood of $\mathbf{m}^*$ included in $\mathcal{Z}_{s(\mathbf{m}^*)}$. Since $\mathbf{m}^*$ is locally stable, $\mathbf{K}_{s(\mathbf{m}^*)}$ is a stable matrix. We can therefore choose a smaller neighborhood $\mathcal{N}_2 \subset \mathcal{N}_1$ so that $\Phi_t(\mathcal{N}_2) \subset \mathcal{N}_1$ for all $t \geq 0$. That is, the image of $\mathcal{N}_2$ under the maps $\Phi_{t\geq 0}$ remains inside $\mathcal{N}_1$. This is possible by stability of $\mathbf{m}^*$. We next choose a neighborhood $\mathcal{N}_3 \subset \mathcal{N}_2$ and a $\delta > 0$ so that $(\phi(\mathcal{N}_3))^\delta \subset \mathcal{N}_2$, that is, the image of $\mathcal{N}_3$ under ϕ remains inside $\mathcal{N}_2$ and it is at least δ away from the boundary of $\mathcal{N}_2$. We finally fix $r > 0$ so that the intersection $\mathcal{B}(\mathbf{m}^*, r) \cap \Delta^d \subset \mathcal{N}_3$, and we choose our neighborhood $\mathcal{N}$ as

$$\mathcal{N} := \mathcal{B}(\mathbf{m}^*, r) \cap \Delta^d.$$

Note that the choice of r and δ is independent of N. From (ii) of Lemma 3.6.7, we denote furthermore by $\tilde{T} := T(r/2)$ the finite time such that for all $\mathbf{m} \in \Delta^d$, $\Phi_{\tilde{T}+1}(\mathbf{m}) \in \mathcal{B}(\mathbf{m}^*, r/2)$.

Definition and properties of the function G.

Following the generator approach used for instance in Gast et al. [19]. For $\mathbf{m} \in \Delta^d$, define $G : \Delta^d \to \mathbb{R}^d$ as

$$G(\mathbf{m}) := \sum_{t=0}^{\infty} \left(\Phi_t(\mathbf{m}) - \mathbf{m}^*\right).$$

By using Lemma 3.6.7, for all $\mathbf{m} \in \Delta^d$ we have $\|G(\mathbf{m})\| \leq \sum_{t=0}^{\infty} b_1 \cdot e^{-b_2 t} \cdot \|\mathbf{m} - \mathbf{m}^*\| < \infty$. This shows that the function G is well defined and bounded. Denote by $\overline{G} := \sup_{\mathbf{m}\in\Delta^d} \|G(\mathbf{m})\| < \infty$.

By our choice of $\mathcal{N}_2$ defined above, for all $t \geq 0$ and $\mathbf{m} \in \mathcal{N}_2$ we have:

$$\Phi_t(\mathbf{m}) = (\mathbf{m} - \mathbf{m}^*) \cdot \mathbf{K}^t_{s(\mathbf{m}^*)} + \mathbf{m}^*. \tag{3.15}$$

Hence, for all $\mathbf{m} \in \mathcal{N}_2$, we have

$$\begin{aligned} G(\mathbf{m}) &= \sum_{t=0}^{\infty} \left(\Phi_t(\mathbf{m}) - \mathbf{m}^*\right) \\ &= \sum_{t=0}^{\infty} (\mathbf{m} - \mathbf{m}^*) \cdot \mathbf{K}^t_{s(\mathbf{m}^*)} \\ &= (\mathbf{m} - \mathbf{m}^*) \cdot (\mathbf{I} - \mathbf{K}_{s(\mathbf{m}^*)})^{-1}, \end{aligned}$$

where the last equality holds because $\mathbf{K}_{s(\mathbf{m}^*)}$ is a stable matrix. Hence in $\mathcal{N}_2$, $G(\mathbf{m})$ is an *affine* function of $\mathbf{m}$.

From the definition of function G, we see that for all $\mathbf{m} \in \Delta^d$:

$$\begin{aligned} G(\mathbf{m}) - G(\phi(\mathbf{m})) &= \sum_{t=0}^{\infty} \left(\Phi_t(\mathbf{m}) - \mathbf{m}^*\right) - \sum_{t=0}^{\infty} \left(\Phi_t(\phi(\mathbf{m})) - \mathbf{m}^*\right) \\ &= \sum_{t=0}^{\infty} \left(\Phi_t(\mathbf{m}) - \mathbf{m}^*\right) - \sum_{t=1}^{\infty} \left(\Phi_t(\mathbf{m}) - \mathbf{m}^*\right) \\ &= \mathbf{m} - \mathbf{m}^*, \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}[\mathbf{M}^{(N)}(0)] - \mathbf{m}^* &= \mathbb{E}\left[G(\mathbf{M}^{(N)}(0)) - G(\phi(\mathbf{M}^{(N)}(0)))\right] \quad \text{(By the above equality)} \\ &= \mathbb{E}\left[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{M}^{(N)}(0)))\right] \quad \text{(Since } \mathbf{M}^{(N)}(0) \text{ is stationary)} \\ &= \mathbb{E}\Bigg[\mathbb{E}\left[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{m})) \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\right] \cdot \mathbb{1}_{\{\mathbf{m}\notin\mathcal{N}\}} \qquad (3.16) \\ &\qquad + \mathbb{E}\left[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{m})) \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\right] \cdot \mathbb{1}_{\{\mathbf{m}\in\mathcal{N}\}}\Bigg]. \qquad (3.17) \end{aligned}$$

In the following, we bound (3.16) and (3.17) separately.

Bound on (3.16)

As G is bounded by $\overline{G}$, we have

$$\left\| \mathbb{E}\left[\mathbb{E}\left[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{m})) \middle| \mathbf{M}^{(N)}(0) = \mathbf{m} \right] \cdot \mathbb{1}_{\{\mathbf{m} \notin \mathcal{N}\}} \right] \right\| \leq 2\overline{G} \cdot \mathbb{P}\left[\mathbf{M}^{(N)}(0) \notin \mathcal{N} \right].$$

We are left to bound $\mathbb{P}\left[\mathbf{M}^{(N)}(0) \notin \mathcal{N} \right]$. Let $u := \left(\frac{r}{2(1+K+K^2+\ldots+K^{\tilde{T}})} \right)^2$, where K is the Lipschitz constant of ϕ. We have by Lemma 3.6.6:

$$\begin{aligned}
&\mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1) - \Phi_{\tilde{T}+1}(\mathbf{m})\| \geq \frac{r}{2} \,\middle|\, \mathbf{M}^{(N)}(0) = \mathbf{m} \right] \\
&= \mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1) - \Phi_{\tilde{T}+1}(\mathbf{m})\| \geq (1+K+K^2+\ldots+K^{\tilde{T}})\sqrt{u} \,\middle|\, \mathbf{M}^{(N)}(0) = \mathbf{m} \right] \\
&\leq d(\tilde{T}+1) \cdot e^{-2uN}.
\end{aligned}$$

This shows that

$$\begin{aligned}
\mathbb{P}\left[\mathbf{M}^{(N)}(0) \notin \mathcal{N} \right] &= \mathbb{P}\left[\|\mathbf{M}^{(N)}(0) - \mathbf{m}^*\| \geq r \right] \\
&= \mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1) - \mathbf{m}^*\| \geq r \right] \quad \text{(By stationarity)} \\
&\leq \mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1) - \Phi_{\tilde{T}+1}(\mathbf{M}^{(N)}(0))\| \geq \frac{r}{2} \right] \\
&\quad + \mathbb{P}\left[\|\Phi_{\tilde{T}+1}(\mathbf{M}^{(N)}(0)) - \mathbf{m}^*\| \geq \frac{r}{2} \right] \\
&= \mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1) - \Phi_{\tilde{T}+1}(\mathbf{M}^{(N)}(0))\| \geq \frac{r}{2} \right] \\
&\leq d(\tilde{T}+1) \cdot e^{-2uN}, \qquad (3.18)
\end{aligned}$$

where the last equality comes from our choice of $\tilde{T} = T(r/2)$.

Bound on (3.17)

By Lemma 3.6.5, we have

$$\begin{aligned}
&\mathbb{E}\left[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{m})) \,\middle|\, \mathbf{M}^{(N)}(0) = \mathbf{m} \right] \cdot \mathbb{1}_{\{\mathbf{m} \in \mathcal{N}\}} \\
&= \mathbb{E}\left[G(\phi(\mathbf{m}) + \mathbf{E}^{(N)}(1)) - G(\phi(\mathbf{m})) \,\middle|\, \mathbf{M}^{(N)}(0) = \mathbf{m} \right] \cdot \mathbb{1}_{\{\mathbf{m} \in \mathcal{N}\}} \\
&= \mathbb{E}\Bigg[\left(G(\phi(\mathbf{m}) + \mathbf{E}^{(N)}(1)) - G(\phi(\mathbf{m})) \right) \cdot \mathbb{1}_{\{\|\mathbf{E}^{(N)}(1)\| < \delta\}} \\
&\qquad + \left(G(\phi(\mathbf{m}) + \mathbf{E}^{(N)}(1)) - G(\phi(\mathbf{m})) \right) \cdot \mathbb{1}_{\{\|\mathbf{E}^{(N)}(1)\| \geq \delta\}} \,\Bigg|\, \mathbf{M}^{(N)}(0) = \mathbf{m} \Bigg] \cdot \mathbb{1}_{\{\mathbf{m} \in \mathcal{N}\}}
\end{aligned}$$

By our choice of $\mathcal{N}$ and δ, for the first part of the above expectation, *i.e.* when the event $\{\|\mathbf{E}^{(N)}(1)\| < \delta\}$ occurs, $\phi(\mathbf{m}) + \mathbf{E}^{(N)}(1)$ will remain in $\mathcal{N}_2$, hence $G\left(\phi(\mathbf{m}) + \mathbf{E}^{(N)}(1)\right)$ takes the same affine form as $G(\phi(\mathbf{m}))$. Consequently

$$\mathbb{E}\left[\left(G(\phi(\mathbf{m}) + \mathbf{E}^{(N)}(1)) - G(\phi(\mathbf{m})) \right) \cdot \mathbb{1}_{\{\|\mathbf{E}^{(N)}(1)\| < \delta\}} \,\middle|\, \mathbf{M}^{(N)}(0) = \mathbf{m} \right] \cdot \mathbb{1}_{\{\mathbf{m} \in \mathcal{N}\}}$$

$$= \Big[G\big(\mathbb{E}\big[\phi(\mathbf{m}) + \mathbf{E}^{(N)}(1) \,\big|\, \mathbf{M}^{(N)}(0) = \mathbf{m}\big]\big) - G\big(\mathbb{E}\big[\phi(\mathbf{m}) \,\big|\, \mathbf{M}^{(N)}(0) = \mathbf{m}\big]\big) \Big] \mathbb{P}\big(\{\|\mathbf{E}^{(N)}(1)\| < \delta\}\big) \cdot \mathbb{1}_{\{\mathbf{m} \in \mathcal{N}\}}$$

(Thanks to the affinity of G in this case, we can interchange $\mathbb{E}$ and G)

$$= 0 \qquad \text{(By Lemma 3.6.5)}.$$

For the second part of the above expectation,

$$\begin{aligned} &\Big\| \mathbb{E}\Big[\big(G(\phi(\mathbf{m}) + \mathbf{E}^{(N)}(1)) - G(\phi(\mathbf{m}))\big) \cdot \mathbb{1}_{\{\|\mathbf{E}^{(N)}(1)\| \geq \delta\}} \Big| \mathbf{M}^{(N)}(0) = \mathbf{m} \Big] \Big\| \cdot \mathbb{1}_{\{\mathbf{m} \in \mathcal{N}\}} \\ &\leq 2\overline{G} \cdot \mathbb{P}\big(\|\mathbf{E}^{(N)}(1)\| \geq \delta\big) \\ &\leq 2d\overline{G} \cdot e^{-2N\delta^2} \quad \text{(By Lemma 3.6.5)}. \end{aligned}$$

So finally

$$\big\| \mathbb{E}\big[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{m})) \big| \mathbf{M}^{(N)}(0) = \mathbf{m}\big] \big\| \cdot \mathbb{1}_{\{\mathbf{m} \in \mathcal{N}\}} \leq 0 + 2d\overline{G} \cdot e^{-2N\delta^2} = 2d\overline{G} \cdot e^{-2N\delta^2}.$$

Conclusion of the proof

To summarize, we have obtained by (3.18):

$$\begin{aligned} \mathbb{P}\left[\mathbf{M}^{(N)}(0) \notin \mathcal{Z}_{s(\mathbf{m}^*)}\right] &\leq \mathbb{P}\left[\left(\mathbf{M}^{(N)}(0) \notin \mathcal{N}\right)\right] \\ &\leq d(\tilde{T}+1) \cdot e^{-2uN} \\ &\leq b \cdot e^{-cN}, \end{aligned}$$

and

$$\begin{aligned} \|\mathbb{E}\big[\mathbf{M}^{(N)}(0)\big] - \mathbf{m}^*\| &\leq 2d\overline{G} \cdot e^{-2N\delta^2} + 2\overline{G}(\tilde{T}+1) \cdot e^{-2uN} \\ &\leq b \cdot e^{-cN}, \end{aligned}$$

where b, c can be taken as $b := d(2\overline{G}+1)(\tilde{T}+2)$, $c := \min(\delta^2, u)$, and this concludes the proof of Theorem 3.6.3.

□

Conclusion of the Chapter

In this chapter we have proven the exponentially fast asymptotic optimality of WIP, the convergence of which being a classical result proven in Weber and Weiss [48] more than 30 years ago. Our proof not only provides a solid theoretical support for the practically observed excellent performance of WIP, but also give a better understanding as why and when WIP performs well.

One important question that we have not addressed much in this chapter is the necessity of the indexability assumption on WIP. We shall discuss in the next chapter how to completely avoid this assumption, by using a LP approach.

4

The LP Approach

In this chapter we first review a general LP approach to study the discrete time infinite horizon restless bandit model, which is proposed in Verloop [47], and incorporates the Whittle index policy we analysed in the previous chapter. The main point is that we can construct a large collection of so-called "LP-priority policies" based on the solution to the LP, that are all proven to be asymptotically optimal in Verloop [47], provided that a similar global attractor property holds (as for WIP). We then claim that this convergence actually occurs at exponential rate, provided that a similar non-singular and locally stable conditions hold (as for WIP), the later combined with the global attractor property are unified into the so-called "uniform global attractor property". Finally, we define the LP-index policy as one particular choice among the LP-priority policies, and compare it with WIP.

> Mathematics consists in proving the most obvious thing in the least obvious way.
>
> George Polya

4.1 Introduction

One potential drawback of WIP studied in the previous chapter is that it requires the technical condition of indexability on the restless bandit. Many works have been devoted to computing the indices or testing indexability, e.g. Niño-Mora [37, 39], Gast et al. [16], which makes WIP easily computable for indexable problems. Yet, we can not apply this policy if the restless bandit is non-indexable. To circumvent this weakness, another approach, based on solving linear programs, is proposed in Verloop [47], 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 existence of global

attractor), regardless of indexability. In particular, WIP is inside this set of LP-priority policies, if the restless bandit is indexable.

A related LP-approach on infinite horizon restless bandits also appears in Bertsimas and Niño-Mora [9], where the authors propose a hierarchy of N relaxations to the original PSPACE-hard problem, each corresponds to a linear program with increasing complexity, while in the meantime approaching the original problem more closely, with the N-th relaxation corresponds to the exact problem. However, the model considered therein is the infinite horizon discounted problem where arms are statistically non-identical, and no asymptotic optimality results are proven under this generality.

Summary of contributions

In this chapter, we show that the convergence claimed in Verloop [47] for the class of LP-priority policies on infinite horizon restless bandits actually occurs at exponential rate, provided that the additional non-singular (called "non-degenerate" in the LP framework) and locally stable conditions hold (as for WIP), the later combined with the global attractor property are unified into the so-called "uniform global attractor property". We then define the LP-index policy as one particular choice among the LP-priority policies, and compare it with WIP.

Outline

The rest of the chapter is organized as follows. In Section 4.2 we recall the infinite horizon restless bandit model studied previously in Chapter 3, using notations that are more adapted to the LP approach. In Section 4.3 we define the LP relaxation as well as the non-degenerate property. The exponential convergence rate theorem is stated in Section 4.4. Since its proof is very similar to the one of WIP proven in detail in Chapter 3, we shall only make comments on the minor changes. Finally, the LP-index policy is defined in Section 4.5, together with its comparison with WIP.

4.2 Model Description

An infinite horizon restless bandit model is composed of N statistically identical arms. Each arm can be considered as a Markov decision process 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. 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 $1 - \{\alpha N\}$. In other words, we use the probabilistic solution discussed in Section 3.4.3 for non-integer values of αN.

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:

$$\mathbb{P}\left[\mathbf{S}(t+1) = \mathbf{s}' \mid \mathbf{S}(t), \mathbf{A}(t), \ldots, \mathbf{S}(0), \mathbf{A}(0)\right] = \mathbb{P}\left[\mathbf{S}(t+1) = \mathbf{s}' \mid \mathbf{S}(t) = \mathbf{s}, \mathbf{A}(t) = \mathbf{a}\right] = \prod_{n=1}^{N} P^{a_n}_{s_n, s'_n}. \tag{4.1}$$

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 $\mathbf{Y}^{(N)}(t) := \left(Y_{s,1}^{(N)}(t), Y_{s,0}^{(N)}(t)\right)_{s \in \{1 \ldots d\}} \in \Delta^{2d}$, where Δ^d (and Δ^{2d}) are the d-dimensional (and $2d$-dimensional) simplex of probability vectors.

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

$$V_{\text{opt}}^{(N)}(\alpha) = \max_{\pi \in \Pi} \quad \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \Big[\sum_{t=0}^{T-1} Y_{s,a}^{(N)}(t) R_s^a \Big] \tag{4.2a}$$

$$\text{s.t.} \quad \sum_s Y_{s,1}^{(N)}(t) = \begin{cases} (\lfloor \alpha N \rfloor + 1)/N, & \text{with probability } \{\alpha N\} \\ \lfloor \alpha N \rfloor / N, & \text{otherwise.} \end{cases} \quad \forall t, \tag{4.2b}$$

$$\text{Arms follow the Markovian evolution (4.1)} \tag{4.2c}$$

Here Π is the set of Markovian stationary policies. To ease the discussion, we assume that the infinite horizon restless bandit 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.

4.3 LP Relaxation and Non-degeneracy

Similar to what we do in (3.5) for Whittle indices, we relax the constraints in (4.2b) into the following single constraint

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_s \mathbb{E}_\pi \left[Y_{s,1}^{(N)}(t) \right] = \alpha, \tag{4.3}$$

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:

$$V_{\text{rel}}^{(N)}(\alpha) = \max_{\mathbf{y} \geq \mathbf{0}} \quad \sum_{s,a} R_s^a y_{s,a} \tag{4.4a}$$

$$\text{s.t.} \quad \sum_{s} y_{s,1} = \alpha, \tag{4.4b}$$

$$y_{s,0} + y_{s,1} = \sum_{s',a} y_{s',a} P_{s's}^a \qquad \forall s, \tag{4.4c}$$

$$\sum_{s,a} y_{s,a} = 1. \tag{4.4d}$$

Denote by $\mathbf{y}^*$ an optimal solution of (4.4), and $\mathbf{m}^*$ the stationary measure, so that $m_s^* = y_{s,0}^* + y_{s,1}^*$ for all $s \in \mathcal{S}$. We define the following four sets, which together give a partition of the set of states $\mathcal{S}$.

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

Note that the unichain assumption implies that $\mathcal{S}^\emptyset$ is empty.

We say that an infinite horizon restless bandit is *non-degenerate* if there exists a solution $\mathbf{y}^*$ of (4.4) such that $\left|\mathcal{S}^0\right| \geq 1$. This notion will later be defined also in the finite horizon case. We prove that

Proposition 4.3.1. *For any infinite horizon restless bandit, the optimization problem* (4.4) *has an optimal solution* $\mathbf{y}^*$ *satisfying* $\left|\mathcal{S}^0\right| \leq 1$.

Proof. We can transform the optimization problem (4.4) into a *constraint* MDP, where the one constraint comes from (4.4b). We then apply Theorem 4.4 of Altman [3], which states that for a feasible infinite horizon MDP using the expected average cost criteria with one inequality constraint, there exists an optimal stationary policy such that the total number of randomization that it uses is at most one. Since one number of randomization corresponds exactly to one state s such that $s \in \mathcal{S}^0$, our claim follows. □

Consequently, a problem is degenerate if and only if for *any* optimal solution $\mathbf{y}^*$ of (4.4) we have $\left|\mathcal{S}^0\right| = 0$. This implies that $\sum_{s \in \mathcal{S}^+} m_s^* = \alpha$, which is the analogue of non-singularity for WIP defined in Chapter 3.

4.4 Asymptotic Optimality of LP-priority Policy with Exponential Rate

Following Definition 4.4 of Verloop [47], we define the set of LP-priorities as $\Sigma := \bigcup_{\mathbf{y}^*} \Sigma(\mathbf{y}^*)$, where $\Sigma(\mathbf{y}^*)$ is the set of permutations $\sigma = \sigma_1 \ldots \sigma_d$ 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 4.3.1, there exists $\mathbf{y}^*$ such that $\left|\mathcal{S}^0\right| \leq 1$. We shall choose this $\mathbf{y}^*$ and fix $\sigma^* \in \Sigma(\mathbf{y}^*)$. Denote by $V_{\mathrm{LP}}^{(N)}(\alpha)$ the value of the corresponding LP-priority policy. Clearly we have $V_{\mathrm{LP}}^{(N)}(\alpha) \leq V_{\mathrm{opt}}^{(N)}(\alpha) \leq V_{\mathrm{rel}}^{(N)}(\alpha)$. We wish to show the convergence of $V_{\mathrm{LP}}^{(N)}(\alpha)$ to $V_{\mathrm{rel}}^{(N)}(\alpha)$ 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

$$\Phi : \mathbf{M}^{(N)}(t) \xrightarrow[\text{policy}]{\text{LP priority}} \mathbf{Y}(t) = \mathbf{Y}^{(N)}(t) \xrightarrow[\text{Markovian transition (5.1)}]{\text{each arm follows the}} \phi(\mathbf{Y}^{(N)}(t)) \tag{4.5}$$

is a piecewise affine and continuous function from Δ^d to Δ^d, with d affine pieces, as in Lemma 3.3.1. Define the t-th iteration of maps $\Phi_{t\geq 0}(\cdot)$ as $\Phi_0(\mathbf{m}) = \mathbf{m}$, $\Phi_{t+1}(\mathbf{m}) = \Phi\big(\Phi_t(\mathbf{m})\big)$. The following Uniform Global Attractor Property captures the global attractor and locally stable assumptions in Theorem 3.3.2.

(Uniform Global Attractor Property (UGAP)) The vector $\mathbf{m}^* \in \Delta^d$ given by the optimal solution of (4.4) is a uniform global attractor of $\Phi_{t\geq 0}(\cdot)$, 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 $\|\Phi_t(\mathbf{m}) - \mathbf{m}^*\|_1 \leq \epsilon$.

The next theorem is a refinement of the asymptotic optimality result in Verloop [47] (Proposition 4.14), proving the exponential convergence rate under the additional non-degeneracy condition on the infinite horizon restless bandit.

Theorem 4.4.1. *Consider an infinite horizon restless bandit which is unichain and satisfies the UGAP. Then the LP-priority policy induced by σ^* is asymptotically optimal. Moreover, if the restless bandit is non-degenerate, then the convergence rate can be shown to be exponential.*

The proof of this theorem is similar to Theorem 3.3.2, proven in detail in the previous chapter for WIP. We briefly comment on the necessary conditions for the two theorems. Note that the latter being proved for WIP, as a preliminary, the infinite horizon restless bandit needs to be *indexable*, whereas we do not need any assumption on indexability for our result here. The non-singularity condition in Theorem 3.3.2 plays the same role as the non-degenerate condition here, and as the example of Remark 3.3.3 shows, in general this condition is necessary to ensure the exponential rate. However, this condition in infinite horizon is almost always satisfied. As we shall see later, this will not be the case for the finite horizon case.

4.5 The Infinite-Horizon LP Indices and the Whittle Indices

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:

$$\max_{\mathbf{y} \geq \mathbf{0}} \quad \sum_{s,a} (R_s^a - a\gamma^*) y_{s,a} \tag{4.6a}$$

$$\text{s.t.} \qquad y_{s,0} + y_{s,1} = \sum_{s',a} y_{s',a} P^a_{s's} \qquad \forall s, \tag{4.6b}$$

$$\sum_{s,a} y_{s,a} = 1 \tag{4.6c}$$

We transform the problem (4.6) into a MDP, with the modified rewards $\widetilde{R^a_s} := R^a_s - a\gamma^*$. The value function V^*_s for state s satisfies the Bellman equation

$$\begin{aligned} g(\gamma^*) + V^*_s &= \max_a \left\{ \widetilde{R^a_s} + \sum_{s'} V^*_{s'} \cdot P^a_{ss'} \right\} \\ &= \max \left\{ R^0_s + \sum_{s'} V^*_{s'} \cdot P^0_{ss'},\ R^1_s - \gamma^* + \sum_{s'} V^*_{s'} \cdot P^1_{ss'} \right\} \\ &= \max \left\{ Q^0_s,\ Q^1_s \right\}, \end{aligned}$$

where $g(\gamma^*)$ is the optimal value of the linear program (4.6). The LP indices for the infinite horizon restless bandit is then defined as $I_s := Q^1_s - Q^0_s$ 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 definition of Whittle indices and the concept of indexability for an infinite horizon restless bandit, given previously in Chapter 3. For each value $\gamma \in \mathbb{R}$, the value function $V_s(\gamma)$ for state s satisfies a similar Bellman equation

$$g(\gamma) + V_s(\gamma) = \max_a \left\{ R^a_s - a\gamma + \sum_{s'} V_{s'}(\gamma) \cdot P^a_{ss'} \right\}. \tag{4.7}$$

Define

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

In other words, $\mathcal{S}(\gamma)$ is the set of states for which the arg max in (4.7) is $a = 1$. The infinite horizon restless bandit 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 $\mathcal{S}(\gamma)$: $\gamma_s := \sup \left\{ \gamma \in \mathbb{R} \mid s \in \mathcal{S}(\gamma) \right\}$. 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 4.5.1. *Assume that the infinite horizon restless bandit 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 restless bandit 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 linear program (4.6) can be cast into an infinite horizon MDP denoted as X, with state space $\mathcal{S}$ and action space $\{0,1\}$. The reward in state $s \in \mathcal{S}$ under action $a \in \{0,1\}$ is $\widetilde{R_s^a} := R_s^a - a\gamma^*$. The transition probabilities are $\mathbb{P}\big(X(t+1) = y \;\big|\; X(t) = x, \text{action} = a\big) = P_{xy}^a$, for all $t \geq 0$. The theory of stochastic dynamic programming Puterman [44] shows that there exists an optimal policy which is Markovian.

So let ψ^* be such an optimal Markovian stationary policy of (4.6) formulated as a Markov decision process X, so that $\psi_{s,a}^*$ is the probability of choosing action a if $X = s$. Our previous discussion shows that

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

where $\mathbb{P}^{\psi^*}\big(X = s\big)$ refers to the probability of the process X being in state s under the policy ψ^*, with $\psi_{s,0}^* + \psi_{s,1}^* = 1$. We then deduce that

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

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 (4.6). 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 4.5.1, 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.

□

Conclusion of the Chapter

In this chapter we recall the LP approach on the infinite horizon restless bandit problem, and prove the exponentially fast asymptotic optimality of the LP-priority policies, the convergence of which being proven in Verloop [47]. Although we have successfully avoided the indexability assumption by using the LP approach, we still need a global attractor property for the convergence to hold.

We collect all known infinite horizon asymptotic results discussed previously in Table 4.1, with the claimed convergence rate and assumptions needed. A notable addition is the recent work Zhang and Frazier [54], in which by discounting, the authors succeed in constructing an asymptotically optimal policy such that no additional

assumption on the model is needed. The price to be paid is that the constant in the optimality gap depends on the discount factor, and it goes to infinity when the discount factor approached 1.

Assumptions \ Related Works	Weber and Weiss [48]	Verloop [47]	Zhang and Frazier [54]	Present Thesis
Indexabilty	✓	✗	✗	✗
Global Attractor	✓	✓	✗	✓
Local Stability	✗	✗	✗	✓
Discounting	✗	✗	✓	✗
Non-Degeneracy	✗	✗	✗	✓
Convergence Rate	$o(1)$	$o(1)$	$O(1/\sqrt{N})$	$e^{-O(N)}$

Table 4.1 – A summary of infinite horizon results appeared in different works, with the claimed convergence rate and different assumptions needed.

In the next Part II of the thesis, we shall consider the problem under a finite horizon with the LP approach, so that neither the indexability nor the global attractor assumptions are needed. Instead, we will be concerned about the *degeneracy* of the problem.

Part II

Finite Horizon

5

The LP Approach

We provide a general framework to analyse control policies for the restless Markovian bandit model under finite 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. 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. The LP-update policy is briefly mentioned, and used as comparison with the LP-index policy on the applicant screening problem, studied in the numerical section.

I think it is said that Gauss had ten different proofs for the law of quadratic reciprocity. Any good theorem should have several proofs, the more the better. For two reasons: usually, different proofs have different strengths and weaknesses, and they generalise in different directions - they are not just repetitions of each other.

– Michael Atiyah

5.1 Introduction

In this chapter we investigate the famous Markovian restless bandit problem over a finite 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 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 single budget constraint on the number of arms that can be activated at each decision epoch. The problem is considered under a finite horizon $T < \infty$.

This model arises in various domains and has numerous applications (see Zhang and Frazier [53] and the references therein for examples). Solving the (infinite horizon) problem exactly has been shown to be PSPACE-hard in Papadimitriou and Tsitsiklis [42]. 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 [49] and has been of longstanding theoretical and practical interest.

To make a comparison with the results in Part I , studying the problem under infinite horizon is theoretically interesting, but all these asymptotic optimality results mentioned therein rely on the existence of global attractor, which in most cases can only be verified numerically, and may very well not be satisfied on certain problems, as we have discussed in Chapter 3. 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 [29], 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 as in Chapter 4. Later in Zayas-Cabán et al. [52] the problem is generalized to multi-actions (instead of the two actions active and passive), and the policy proposed therein achieves an $O(\log N/\sqrt{N})$ optimality gap. In Brown and Smith [12] the same problem setting as in Hu and Frazier [29] is studied, and their policies are shown to achieve $O(1/\sqrt{N})$ optimality gap. However, as suggested by the authors of Brown and Smith [12], the convergence appears to be faster than $O(1/\sqrt{N})$ on certain problems. Indeed, later in Zhang and Frazier [53] the authors proposed a class of fluid-priority policies that incorporate the policies in the two previous works Brown and Smith [12] and Hu and Frazier [29], and show that they achieve $O(1/\sqrt{N})$ optimality gap in general, and can be improved to $O(1/N)$ if a *non-degenerate* condition holds on the restless bandit. By refining the policies, we later show that this $O(1/N)$ rate can actually be further improved to be $e^{-O(N)}$.

Summary of contributions

In this chapter, we provide a generic framework to study the relationship between restless bandit problem and the LP relaxations introduced in Hu and Frazier [29] for the finite horizon problem. 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 restless bandit. 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 5.2.3. 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 5.3.2).

b) If in addition the policy is *Lipschitz continuous*, then the asymptotic optimality occurs at rate $O(1/\sqrt{N})$.

c) If in addition the policy is *locally linear* around the LP solution, then the asymptotic optimality occurs at rate $e^{-O(N)}$.

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 restless bandit, 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 restless bandit to be *non-degenerate*, a condition already introduced in Zhang and Frazier [53] and defined in Section 5.4.1. Moreover, we provide a degenerate example in Section 5.4.3 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) we define the LP-index that refines the ranking of all states, 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 $O(1/\sqrt{N})$ rate of asymptotic optimality on this policy. We demonstrate its advantage to previous LP-based policies, both theoretically and practically.

Outline

The rest of the chapter is organized as follows: Section 5.2 defines the finite horizon restless bandit model as well as the admissible policy. Section 5.3 introduces a hierarchy of admissible policies, and prove asymptotic optimality (with convergence rate if possible) inside each of the hierarchy. Section 5.4 provides concrete constructions for the polices discussed in Section 5.3, and gives necessary and sufficient conditions for exponential convergence rate. Section 5.5 describes the LP-update policy. Section 5.6 provides numerical studies.

5.2 Model Description

We first describe the model in Section 5.2.1. We introduce the LP relaxation in Section 5.2.2. We define the admissible policy and the randomized rounding procedure in Section 5.2.3.

5.2.1 Finite horizon restless bandit

Like in infinite horizon as we discussed in Chapter 4, a finite horizon restless bandit 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 [12], Verloop [47]). 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 $P_{s,s'}^a$. Those transitions are independent among all arms: for given

states $\mathbf{s}, \mathbf{s}'$ and activation vector $\mathbf{a}$, one has:

$$\mathbb{P}\left[\mathbf{S}(t+1) = \mathbf{s}' \mid \mathbf{S}(t), \mathbf{A}(t), \ldots, \mathbf{S}(0), \mathbf{A}(0)\right] = \mathbb{P}\left[\mathbf{S}(t+1) = \mathbf{s}' \mid \mathbf{S}(t) = \mathbf{s}, \mathbf{A}(t) = \mathbf{a}\right] = \prod_{n=1}^{N} P^{a_n}_{s_n, s'_n}. \tag{5.1}$$

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 $M^{(N)}_s(t)$ the *fraction* of arms in state s at time t, and by $Y^{(N)}_{s,a}(t)$ 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 $\mathbf{Y}^{(N)}(t) := \left(Y^{(N)}_{s,1}(t), Y^{(N)}_{s,0}(t)\right)_{s \in \{1 \ldots d\}} \in \Delta^{2d}$, where Δ^d (and Δ^{2d}) are the d-dimensional (and $2d$-dimensional) simplex of probability vectors.

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

$$V^{(N)}_{\text{opt}}(\mathbf{m}(0), T) = \max_{\mathbf{Y} \geq \mathbf{0}} \quad \mathbb{E}\Big[\sum_{t=0}^{T-1} \sum_{s,a} R^a_s Y^{(N)}_{s,a}(t)\Big] \tag{5.2a}$$

$$\text{s.t.} \quad \text{Arms follow the Markovian evolution (5.1)}, \tag{5.2b}$$

$$Y^{(N)}_{s,0}(t) + Y^{(N)}_{s,1}(t) = M^{(N)}_s(t) \;\; \forall t, s, \tag{5.2c}$$

$$\sum_s Y^{(N)}_{s,1}(t) = \begin{cases} (\lfloor \alpha N \rfloor + 1)/N, & \text{with probability } \{\alpha N\} \\ \lfloor \alpha N \rfloor / N, & \text{otherwise.} \end{cases} \quad \forall t, \tag{5.2d}$$

$$M^{(N)}_s(0) = m_s(0) \;\; \forall s, \tag{5.2e}$$

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 $s \in \{1 \ldots d\}$. Note that (5.2d) represent the constraints that αN of the N arms must be activated at each time, and (5.2e) correspond to the initial condition.

5.2.2 LP relaxation

The key difficulty in the above optimization problem (5.2) is the constraint (5.2d) 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:

$$\sum_s \mathbb{E}_\pi\left[Y^{(N)}_{s,1}(t)\right] = \alpha, \;\; \forall t. \tag{5.3}$$

The key property that makes this relaxed problem simpler is that it can then be rewritten entirely by using only the variables $y_{s,a}(t) := \mathbb{E}\left[Y^{(N)}_{s,a}(t)\right]$. To see that, we will show later in Lemma 5.3.1 that the Markovian evolution (5.7) implies that

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

This implies that (5.2b) and (5.2c) can be replaced by (5.4b) 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 $V_{\text{rel}}(\mathbf{m}(0), T)$:

$$V_{\text{rel}}(\mathbf{m}(0), T) = \max_{\mathbf{y} \geq \mathbf{0}} \quad \sum_{t=0}^{T-1} \sum_{s,a} R_s^a y_{s,a}(t) \tag{5.4a}$$

$$\text{s.t.} \quad y_{s,0}(t+1) + y_{s,1}(t+1) = \sum_{s',a} y_{s',a}(t) P_{s's}^a \qquad \forall s, t, \tag{5.4b}$$

$$\sum_s y_{s,1}(t) = \alpha \qquad \forall t, \tag{5.4c}$$

$$y_{s,0}(0) + y_{s,1}(0) = m_s(0) \qquad \forall s. \tag{5.4d}$$

In the above optimization problem, the constraints (5.4c) are the relaxation of the constraints (5.2d). They impose that the expected fraction of activated arms is α at all time. The constraints (5.4b) correspond to the expected behavior of the Markovian evolution of the system. Similarly, (5.4d) correspond to the initial condition (5.2e).

Note that the optimization problem (5.4) does not depend on the arm population N. Moreover, as it is a relaxation of (5.2), it should be clear that $V_{\text{opt}}^{(N)}(\alpha) \leq V_{\text{rel}}^{(N)}(\alpha)$. Since finding an optimal policy for $V_{\text{opt}}^{(N)}(\alpha)$ is impractical, our strategy is to obtain information from optimal solutions to the linear program (5.4) to construct policies whose values converge quickly to $V_{\text{rel}}^{(N)}(\alpha)$ as N goes to infinity. As $V_{\text{opt}}^{(N)}(\alpha) \leq V_{\text{rel}}^{(N)}(\alpha)$, this will imply that they become asymptotically optimal as N goes to infinity.

5.2.3 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

$$y_{s,a} \geq 0, \quad \sum_s y_{s,1} = \alpha, \quad \text{and} \quad \sum_a y_{s,a} = m_s \quad \forall s, a. \tag{5.5}$$

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 $N y_{s,a}$ 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 $N y_{s,1}$ arms in state s *in expectation* and that works as follows:

- In a first pass, one activates $\lfloor N y_{s,1} \rfloor$ arms in state s, and we let $z_s := N y_{s,1} - \lfloor N y_{s,1} \rfloor$;
- 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, $h = \lfloor \alpha N \rfloor - \sum_s \lfloor N y_{s,1} \rfloor$ or $h = \lfloor \alpha N \rfloor + 1 - \sum_s \lfloor N y_{s,1} \rfloor$ 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 [30]. It has time complexity $O(hd \cdot \log d)$.

5.3 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 5.3.1. In Section 5.3.2, 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 5.3.3, 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 5.3.1, Theorem 5.3.2 and Theorem 5.3.3 are given respectively in Section 5.3.4, 5.3.4 and 5.3.4.

5.3.1 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:

$$\mathbf{M}^{(N)}(t) \xrightarrow[\text{policy } \pi_t(\cdot)]{\text{admissible}} \mathbf{Y}(t) \xrightarrow[\text{rounding}]{\text{randomized}} \mathbf{Y}^{(N)}(t) \xrightarrow[\text{Markovian transition (5.1)}]{\text{each arm follows the}} \mathbf{M}^{(N)}(t+1). \tag{5.6}$$

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

$$(\phi(\mathbf{y}))_s = \sum_{s',a} y_{s,a} P^a_{s',s}. \tag{5.7}$$

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 (6.21)), with an error that decreases as $O(1/\sqrt{N})$.

This observation will be used to show that a continuous admissible policy is optimal if and only if it is LP-compatible. Equation (6.20) 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 (6.22) and the fact that ϕ is linear, will be critically used in the proof of the exponential rate.

Lemma 5.3.1. *Let* $\mathbf{E}^{(N)}(t) = \mathbf{M}^{(N)}(t+1) - \phi(\mathbf{Y}^{(N)}(t))$, *where* $\phi(\cdot)$ *is given in* (5.7). *We have:*

$$\mathbb{E}\left[\mathbf{E}^{(N)}(t) \mid \mathbf{Y}^{(N)}(t)\right] = \mathbf{0}, \tag{5.8}$$

$$\mathbb{E}\left[\left\|\mathbf{E}^{(N)}(t)\right\|_1 \mid \mathbf{Y}^{(N)}(t)\right] \leq \frac{\sqrt{d}}{\sqrt{N}}, \tag{5.9}$$

$$\mathbb{P}\left[\left\|\mathbf{E}^{(N)}(t)\right\|_1 \geq \epsilon \mid \mathbf{Y}^{(N)}(t)\right] \leq 2de^{-2N\epsilon^2/d^2}. \tag{5.10}$$

A detailed proof of this result is provided in Section 5.3.4.

5.3.2 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:

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

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 (5.4), 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 $V_\pi(\mathbf{m}(0), T) = V_{\text{rel}}^{(N)}(\alpha)$.

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 $O(1/\sqrt{N})$ 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 5.4 that for all finite horizon restless bandit, there always exists a LP-compatible Lipschitz continuous policy that can be easily constructed.

Theorem 5.3.2. *Let* $\pi = \{\pi_t\}_{0\leq t\leq T-1}$ *be an admissible and continuous policy. Then:*

$$\lim_{N\to\infty} V_\pi^{(N)}(\mathbf{m}(0), T) = V_\pi(\mathbf{m}(0), T). \tag{5.11}$$

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

$$\left|V_\pi^{(N)}(\mathbf{m}(0), T) - V_\pi(\mathbf{m}(0), T)\right| \leq \frac{C}{\sqrt{N}}. \tag{5.12}$$

In particular, this implies that:

1. *If π is LP-compatible, then* $\lim_{N\to\infty} V_\pi^{(N)}(\mathbf{m}(0),T) = \lim_{N\to\infty} V_{\text{opt}}^{(N)}(\alpha) = V_{\text{rel}}^{(N)}(\alpha)$.

2. *If π is not LP compatible, then* $\limsup_{N\to\infty} V_\pi^{(N)}(\mathbf{m}(0),T) < V_{\text{rel}}^{(N)}(\alpha)$.

3. *If π is LP-compatible and Lipschitz continuous, then there exists $C' > 0$ independent of N such that*

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

Proof. (Sketch) A detailed proof is presented in Section 5.3.4. 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:

$$V_\pi^{(N)}(\mathbf{m}(0),T) - V_\pi(\mathbf{m}(0),T) = \sum_{t,a,s} R_s^a \left(\mathbb{E}\left[Y_{s,a}^{\pi,(N)}(t)\right] - y_{s,a}^\pi(t)\right). \tag{5.13}$$

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 5.3.1 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 $\mathbf{M}^{\pi,(N)}(t+1)$.

□

5.3.3 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 5.4.

We say that an LP-compatible policy $\pi = \{\pi_t\}_{0\leq t\leq T-1}$ is *locally linear* if there exists a solution $\{\mathbf{y}^*(t)\}_{0\leq t\leq T-1}$ of (5.4) such that for all $0 \leq t \leq T-1$, 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 5.3.3. *Consider a LP-compatible locally linear policy $\pi = \{\pi_t\}_{0\leq t\leq T-1}$. There exist two constants $C_1, C_2 > 0$ independent of N such that*

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

We remark that the result of exponential convergence rate in Theorem 5.3.3 is much stronger than the general square root rate given in Theorem 5.3.2. This is due to the locally linear condition. This local linearity around the optimal trajectory plays a key role in the proof of Theorem 5.3.3, as it is used in (5.18) to justify the interchange of taking expectation with applying a linear function, in order to obtain (5.19). Our later discussion in Section 5.4.3 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 (5.16), 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 [12] and Hu and Frazier [29]), it is an interesting open question to see if we can formulate an exponential convergence type result in such generic case.

5.3.4 Proof of results in Section 5.3

Proof of Lemma 5.3.1

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 (6.20). It also implies that

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

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-Schwartz, and this gives (6.21).

Equation (6.22) is a direct consequence of Hoeffding's inequality. Indeed, one has

$$\mathbb{P}\left[|E^{(N)}_s(t)| \ge \varepsilon/d \mid \mathbf{Y}^{(N)}(t)\right] \le 2e^{-N\varepsilon^2/d^2}.$$

By using the union bound, this implies that

$$\mathbb{P}\left[\left\|\mathbf{E}^{(N)}(t)\right\|_1 | \ge \varepsilon \mid \mathbf{Y}^{(N)}(t)\right] \le d\cdot\mathbb{P}\left[|E^{(N)}_s(t)| \ge \varepsilon/d \mid \mathbf{Y}^{(N)}(t)\right] \le 2de^{-N\varepsilon^2/d^2}.$$

Proof of Theorem 5.3.2

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

$$\left\|\mathbf{y}^{\pi}(t) - \mathbf{Y}^{\pi,(N)}(t)\right\|_1 \leq \left\|\pi_t(\mathbf{m}^{\pi}(t)) - \pi_t(\mathbf{M}^{\pi,(N)}(t))\right\|_1 + \left\|\pi_t(\mathbf{M}^{\pi,(N)}(t)) - \mathbf{Y}^{\pi,(N)}(t)\right\|_1 . \tag{5.14}$$

By construction of randomized rounding, $\left\|\pi_t(\mathbf{M}^{\pi,(N)}(t)) - \mathbf{Y}^{\pi,(N)}(t)\right\|_1 \leq d/N$. 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

$$\left\|\mathbf{m}^{\pi}(t+1) - \mathbf{M}^{\pi,(N)}(t+1)\right\|_1 \leq \left\|\phi(\mathbf{y}^{\pi}(t)) - \phi(\mathbf{Y}^{\pi,(N)}(t))\right\|_1 + \left\|\mathbf{E}^{(N)}(t)\right\|_1 \tag{5.15}$$

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 (5.13) implies (5.11).

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 (5.15), Lemma 5.3.1 and (5.14), we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{m}^{\pi}(t+1) - \mathbf{M}^{\pi,(N)}(t+1)\right\|_1\right] &\leq \mathbb{E}\left[\left\|\phi(\mathbf{y}^{\pi}(t)) - \phi(\mathbf{Y}^{\pi,(N)}(t))\right\|_1\right] + \mathbb{E}\left[\left\|\mathbf{E}^{(N)}(t)\right\|_1\right] \\
&\leq L\mathbb{E}\left[\left\|\mathbf{y}^{\pi}(t) - \mathbf{Y}^{\pi,(N)}(t)\right\|_1\right] + \sqrt{\frac{d}{N}} \\
&\leq L^2\mathbb{E}\left[\left\|\mathbf{m}^{\pi}(t) - \mathbf{M}^{\pi,(N)}(t)\right\|_1\right] + \frac{Ld}{N} + \sqrt{\frac{d}{N}}.
\end{aligned}$$

By a direct induction on t (which is essentially the discrete Gronwall's lemma), this implies that $\mathbb{E}\left[\left\|\mathbf{m}^{\pi}(t+1) - \mathbf{M}^{\pi,(N)}(t+1)\right\|_1\right] = O(1/\sqrt{N})$. Note however that the hidden constant in the $O(\cdot)$ grows exponentially with time t. By (5.13), this implies (5.12).

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

Proof of Theorem 5.3.3

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

$$\mathcal{E}(\delta) := \left\{\text{for all } 0 \leq t \leq T-1\text{: } \left\|\mathbf{E}^{(N)}(t) \leq \delta\right\|_1\right\},$$

where $\mathbf{E}^{(N)}(t)$ is defined as in Lemma 5.3.1, and let $\overline{\mathcal{E}(\delta)}$ be the complementary of the event $\mathcal{E}(\delta)$.

By (6.22) of Lemma 5.3.1, we have

$$\mathbb{P}\left[\overline{\mathcal{E}(\delta)}\right] \leq 2dT \cdot e^{-2N\delta^2/d^2}. \tag{5.16}$$

Assume that event $\mathcal{E}(\delta)$ holds. By definition of $\mathbf{E}^{(N)}(t)$ and (5.6), we have

$$\begin{aligned}
\left\|\mathbf{M}^{(N)}(t+1) - \mathbf{m}^*(t+1)\right\|_1 &= \left\|\phi(\mathbf{Y}^{(N)}(t)) + \mathbf{E}^{(N)}(t) - \phi(\pi_t(\mathbf{m}^*(t)))\right\|_1 \\
&\leq \left\|\phi(\mathbf{Y}^{(N)}(t)) - \phi(\mathbf{Y}(t))\right\|_1 + \left\|\phi(\mathbf{Y}(t)) - \phi(\pi_t(\mathbf{m}^*(t)))\right\|_1 + \left\|\mathbf{E}^{(N)}(t)\right\|_1 \\
&= \left\|\phi(\mathbf{Y}^{(N)}(t)) - \phi(\mathbf{Y}(t))\right\|_1 \\
&\quad + \left\|\phi(\pi_t(\mathbf{M}^{(N)}(t))) - \phi(\pi_t(\mathbf{m}^*(t)))\right\|_1 + \left\|\mathbf{E}^{(N)}(t)\right\|_1 \\
&\leq \frac{2d\ell}{N} + \ell L \cdot \left\|\mathbf{M}^{(N)}(t) - \mathbf{m}^*(t)\right\|_1 + \delta.
\end{aligned} \tag{5.17}$$

A direct induction until $t = 0$ then implies

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

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 (5.17) 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 5.4.2 for rankable finite horizon restless bandit, since then no randomized rounding is needed anywhere and $\mathbf{Y}^{(N)}(t) = \mathbf{Y}(t)$ always holds.

Consequently, we get:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - \mathbf{y}^*(t) &= \mathbb{E}\left[F_t(\mathbf{M}^{(N)}(t))\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - F_t(\mathbf{m}^*(t)) \\
&= \mathbb{E}\left[F_t\left(\phi(\mathbf{Y}^{(N)}(t-1)\mathbf{1}_{\{\mathcal{E}(\delta)\}})\right)\right] - F_t\left(\phi(\mathbf{y}^*(t-1)\right) \\
&= F_t \circ \phi\left(\mathbb{E}\left[\mathbf{Y}^{(N)}(t-1)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - \mathbf{y}^*(t-1)\right),
\end{aligned} \tag{5.18}$$

where on the last equality (5.18) 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

$$\begin{aligned}
\left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - \mathbf{y}^*(t)\right\|_1 &\leq L' \left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t-1)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - \mathbf{y}^*(t-1)\right\|_1 \\
&\leq (L')^T \left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(0)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right] - \mathbf{y}^*(0)\right\|_1.
\end{aligned} \tag{5.19}$$

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 (5.16), we have

$$\left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right] - \mathbb{E}\left[\mathbf{Y}^{(N)}(t)\mathbf{1}_{\{\mathcal{E}(\delta)\}}\right]\right\|_1 \leq 2d \cdot \mathbb{P}\left[\bar{\mathcal{E}}(\delta)\right] \leq 4d^2 T e^{-C_2 N}, \tag{5.20}$$

where $C_2 := -2\varepsilon^2/((1 + \cdots + L^{T-1})^2 d^2)$. Combining (5.19) and (5.20) 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 (5.13), all locally linear LP-compatible policies are asymptotically optimal with exponential rate, and this concludes our proof.

5.4 Existence and Construction of Policies

In this section we provide constructions of Lipschitz continuous policies and locally linear policies, defined in the previous Section 5.3. In Section 5.4.1 we define the non-degenerate and rankable restless bandit, the first being already defined for the infinite horizon problem in Chapter 4. In Section 5.4.2, 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 restless bandit is non-degenerate. We compare the non-degenerate condition with the rankable condition in Section 5.4.3. In Section 5.4.3, we construct a degenerate 2-dimensional restless bandit 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 5.4.2 and Lemma 5.4.3 are given respectively in Section 5.4.3 and 5.4.3.

5.4.1 Non-degenerate and rankable finite horizon restless bandit

Let $\{\mathbf{y}^*(t)\}_{0\le t\le T-1}$ be an optimal solution of the LP relaxed problem (5.4). 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 5.4.2 and Theorem 5.4.4, we introduce two definitions:

1. A restless bandit is *rankable* if there exists an optimal solution of (5.4) for which $\left|\mathcal{S}^0(t)\right| \le 1$ for all t. Otherwise we call this restless bandit *non-rankable*.

2. A restless bandit is *non-degenerate* if there exists an optimal solution $\{\mathbf{y}^*(t)\}_{0\le t\le T-1}$ of (5.4) for which $\left|\mathcal{S}^0(t)\right| \ge 1$ for all t. Otherwise we call this restless bandit *degenerate*. This definition coincides with the one in Zhang and Frazier [53].

Note that the non-degenerate property has already been defined for the infinite horizon problem in Chapter 4, as one of the conditions required for exponential convergence rate. The definition given above is its analogue for finite horizon problems. Recall that this non-degeneracy is like the non singularity for WIP, and hence almost always holds in infinite horizon. This will not be the case for finite horizon, as we shall illustrate around Equation (5.26) by using a simple example.

On the other hand, at first glance it appears that rankable and non-degenerate restless bandits 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 5.4.3.

Proposition 5.4.1. *Consider a restless bandit for which the LP problem* (5.4) *has a unique solution. If this restless bandit is not rankable, then it is degenerated.*

It is also possible to define rankable restless bandits in infinite horizon. However, as a consequence of Proposition 4.3.1, any infinite horizon restless bandit is rankable.

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:

$$y_{\sigma_i,1} := \pi^{\text{priority}(\sigma)}_{\sigma_i,1}(\mathbf{m}) = \min(m_{\sigma_i}, \alpha - \sum_{j=1}^{i-1} y_{\sigma_j,1}). \tag{5.21}$$

The next theorem justifies the notion of rankable restless bandit.

Theorem 5.4.2. *A restless bandit 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 5.4.3. 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 5.5.1 and Section 5.6.1). Theorem 5.4.2 shows that one has to be careful on dealing with the states in $\mathcal{S}^0(t)$: if the restless bandit is non-rankable, i.e. if $\left|\mathcal{S}^0(t)\right| > 1$ 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".

5.4.2 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 $\left|\mathcal{S}^0\right|$ 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 $|\mathcal{S}^+| + \left|\mathcal{S}^0\right|, \cdots, |\mathcal{S}^+| + 1$ 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 $\left|\mathcal{S}^0(t)\right| \le 1$, 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 restless bandit is non-degenerate.

Lemma 5.4.3. *For any finite horizon restless bandit, the water-filling policy described above is a LP-compatible Lipschitz continuous policy. Moreover, if the restless bandit is non-degenerate, i.e. if for all t, $\left|\mathcal{S}^0(t)\right| \ge 1$, then the water-filling policy is a LP-compatible locally linear policy. And if the restless bandit is degenerate, then there is no LP-compatible locally linear policy.*

The proof of Lemma 5.4.3 is postponed to Section 5.4.3. A direct consequence of this lemma, combined with Theorem 5.3.2 and Theorem 5.3.3 is that the water-filling policy is asymptotically optimal at rate at least $O(1/\sqrt{N})$.

Theorem 5.4.4. *For any finite horizon restless bandit, there exists a policy π (constructed by the water-filling procedure) and $C > 0$ such that for any N:*

$$\left|V_\pi^{(N)}(\mathbf{m}(0), T) - V_{\text{opt}}^{(N)}(\alpha)\right| \le \frac{C}{\sqrt{N}}. \tag{5.22}$$

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

$$\left|V_\pi^{(N)}(\mathbf{m}(0), T) - V_{\text{opt}}^{(N)}(\alpha)\right| \le C_1 e^{-C_2 N}. \tag{5.23}$$

Lemma 5.4.3 shows that the non-degenerate condition is necessary and sufficient for the existence of a LP-compatible locally linear policy. Theorem 5.4.4 is less precise in the sense that we only show that non-degeneracy is sufficient to obtain an exponentially asymptotically optimal policy. In Section 5.4.3, we provide an example of a restless bandit 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 (5.23) to hold.

Remark 5.4.5. *Note that the authors of Zhang and Frazier [53] 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 $\left|\mathcal{S}^0(t)\right| \leq 1$, both definition coincide and they both correspond to the same priority policy. When $\left|\mathcal{S}^0(t)\right| \geq 2$, 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 [53] 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 [53] the authors only obtain the $O(\frac{1}{N})$ convergence rate for their algorithm. We believe that this is mainly due to their rounding procedure.

5.4.3 Proof of results in Section 5.4

Proof of Proposition 5.4.1

We can actually prove the slightly more general result that claims that for any restless bandit the optimization problem (5.4) has an optimal solution $\{\mathbf{y}^*(t)\}_{0 \leq t \leq T-1}$ satisfying

$$\sum_{t=0}^{T-1} \left|\mathcal{S}^0(t)\right| \leq T. \tag{5.24}$$

Indeed, similar to our formulation of the optimization problem as a MDP that we later detail in Equation (5.33), we can transform the optimization problem (5.4) into a *constraint* MDP, where the T constraints come from (5.4c). We then apply Theorem 3.8 of Altman [3], 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 (5.24) follows.

Proposition 5.4.1 is then a direct consequence of Equation (5.24): if there exists a unique solution, it must satisfy (5.24), which by the pigeonhole principle implies that either $\left|\mathcal{S}^0(t)\right| \leq 1$ for all t (the problem is rankable) or there exists t such that $\left|\mathcal{S}^0(t)\right| = 0$ (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 (5.24) implies that one can always construct a solution such that $\left|\mathcal{S}^0(t)\right| \leq 1$ for all t (i.e. the problem rankable), or otherwise there always exists t such that $\left|\mathcal{S}^0(t)\right| = 0$ for those solutions. Yet, verifying that there are no other solutions is difficult in general.

Example 5.4.6 (A rankable and degenerate problem)**.** *Let us consider a two states restless bandit 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, $\left|\mathcal{S}^0(t)\right| = 0$ for all t. This example is rankable and is also degenerate.*

Necessary condition for exponential convergence rate

Consider a two states restless bandit 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 $0 \leq p_1, p_2, q_1, q_2 \leq 1$. Let us first establish a sufficient condition on the four parameters p_1, p_2, q_1 and q_2 so that the restless bandit is degenerate. For this simple model, solving the linear program (5.4) amounts to finding the optimal value $0 \leq \beta \leq 0.5 = \alpha$ 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 (5.4) is

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

By definition, the restless bandit is degenerate if

$$\arg\max_{0\leq\beta\leq 0.5}\left\{\beta + \min\left\{0.5,\ \beta(p_1+q_2) + (0.5-\beta)(q_1+p_2)\right\}\right\} \neq 0,\ 0.5, \tag{5.25}$$

since then $\mathcal{S}^0(0) = \{1, 2\}$. A sufficient condition for (5.25) to hold is

$$q_1 + p_2 > 1 + p_1 + q_2, \tag{5.26}$$

under which the arg max of (5.25) 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 $\left|\mathcal{S}^0(0)\right| = 2$ and $\left|\mathcal{S}^0(1)\right| = 0$.

We next consider a stochastic model with a population of N arms, where the 2-dimensional restless bandit satisfies (5.26) 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 (5.4) 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 restless bandit, and it is not due to the problem at decision epoch $t = 0$ with $\left|\mathcal{S}^0(0)\right| > 1$, but due to the fact that at $t = 1$ one has $\left|\mathcal{S}^0(1)\right| = 0$, and the optimal trajectory is on the boundary of two zones, namely $\left\{\mathbf{m} \in \Delta^d \mid \sum_{s\in\mathcal{S}^+(1)} \leq \alpha\right\}$ and $\left\{\mathbf{m} \in \Delta^d \mid \sum_{s\in\mathcal{S}^+(1)} \geq \alpha\right\}$. Note that this example implies in particular that the $O(1/\sqrt{N})$ convergence rate in Theorem 5.3.2 is tight.

Generally speaking, for a degenerate restless bandit, there exists some t for which $\left|\mathcal{S}^0(t)\right| = 0$. 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 $\left\{\mathbf{m} \in \Delta^d \mid \sum_{s\in\mathcal{S}^+(t)} \leq \alpha\right\}$ and $\left\{\mathbf{m} \in \Delta^d \mid \sum_{s\in\mathcal{S}^+(t)} \geq \alpha\right\}$. It is exactly this phenomenon that may prevent an exponentially fast convergence rate.

Proof of Theorem 5.4.2

Assume first that the restless bandit is rankable and let $\{\mathbf{y}^*(t)\}_{0\leq t\leq T-1}$ 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 (5.21), 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 (5.21). Note that if $\mathbf{m}$ is such that $0 \leq \alpha - \sum_{s\in\mathcal{S}^+(t)} m_s(t) \leq m_{s_0}$, then one has:

$$\pi_s^{\text{priority}}(\mathbf{m}) = \begin{cases} m_s & \text{if } s \in \mathcal{S}^+(t) \\ \alpha - \sum_{s\in\mathcal{S}^+(t)} m_s(t) & \text{if } s \in \mathcal{S}^0(t) \\ 0 & \text{otherwise} \end{cases} \tag{5.27}$$

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 $\left|\mathcal{S}^0(t)\right| = 1$ for all t.

Assume now that the restless bandit 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, $|\{s : \pi_{s,0}(\mathbf{m}^*(t)) > 0 \text{ and } \pi_{s,1}(\mathbf{m}^*(t)) > 0\}| \leq 1$. Hence, π cannot be LP-compatible because all solutions of (5.4) are such that there exists a time t such that $\left|\mathcal{S}^0(t)\right| \geq 2$, which is implied by the assumption that the restless bandit is non-rankable.

Proof of Lemma 5.4.3

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 $\mathcal{S}^0 := \{s_1^0, \cdots, s_{s_0}^0\}$, the next s_- states are $\mathcal{S}^- := \{s_1^-, \cdots, s_{s_-}^-\}$, and the rest $s_\emptyset$ states are $\mathcal{S}^\emptyset := \{s_1^\emptyset, \cdots, s_{s_\emptyset}^\emptyset\}$. 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:

$$\widehat{\mathcal{S}} := \Big\{s_1^+, \cdots, s_{s_+}^+, \underbrace{\overline{s}_{s_0}^0, \cdots, \overline{s}_2^0}_{\overline{\mathcal{S}^0}}, s_1^0, \underbrace{\underline{s}_2^0, \cdots, \underline{s}_{s_0}^0}_{\underline{\mathcal{S}^0}}, s_1^-, \cdots, s_{s_-}^-, s_1^\emptyset, \cdots, s_{s_\emptyset}^\emptyset\Big\}, \tag{5.28}$$

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

$$\widehat{M_s^{(N)}} := \begin{cases} M_s^{(N)}, & \text{if } s \in \mathcal{S}^+ \bigcup \mathcal{S}^- \bigcup \mathcal{S}^\emptyset \bigcup \{s_1^0\} \\ \min(M_{s_i^0}^{(N)}, y_{s_i^0,1}^*), & \text{if } s = \overline{s}_i^0 \in \overline{\mathcal{S}^0} \\ M_{s_i^0}^{(N)} - \min(M_{s_i^0}^{(N)}, y_{s_i^0,1}^*), & \text{if } s = \underline{s}_i^0 \in \underline{\mathcal{S}^0}. \end{cases} \tag{5.29}$$

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 (5.28). Let $\mathbf{Y}$ be defined as in

$$Y_s := \begin{cases} \widehat{Y}_s, & \text{if } s \in \mathcal{S}^+ \bigcup \mathcal{S}^- \bigcup \mathcal{S}^\emptyset \bigcup \{s_{s_0}^0\} \\ \widehat{Y_{\overline{s}_i^0}} + \widehat{Y_{\underline{s}_i^0}}, & \text{if } s = s_i^0 \text{ with } 1 \leq i \leq s_0 - 1. \end{cases} \tag{5.30}$$

By construction, the vector $\mathbf{Y}$ corresponds to the vector obtained by the water-filling algorithm constructed in Section 5.4.2.

Now, consider the map chain

$$(\mathbf{M}^{(N)}, \mathbf{y}^*) \xrightarrow{(5.29)} (\widehat{\mathbf{M}^{(N)}}) \xrightarrow{\text{strict priority}} \widehat{\mathbf{Y}} \xrightarrow{(5.30)} \mathbf{Y}. \tag{5.31}$$

It should be clear that (5.29) and (5.30) are Lipschitz continuous functions. As a strict priority policy is Lipschitz continuous, this shows that the water-filling policy is Lipschitz continuous.

Moreover, if $\left|\mathcal{S}^0\right| \geq 1$, then (5.29) is locally linear (and by (5.27), the strict priority policy used is also locally linear). As (5.31) is locally linear, this implies that when the restless bandit 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 (5.4). 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

$$\mathbf{0} \leq (\mathbf{m}^* + \boldsymbol{\varepsilon})\cdot\mathbf{A} = \mathbf{y}^*_{\cdot,1} + \boldsymbol{\varepsilon}\cdot\mathbf{A} \leq \mathbf{m}^* + \boldsymbol{\varepsilon}, \tag{5.32}$$

where the inequalities are considered componentwise.

Consider now a state $i \in \mathcal{S}^+$, one has $y^*_{i,1} = m^*_i$, hence (5.32) 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.

5.5 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 [12] – 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.

5.5.1 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 [53], as

well as in Definition 4.4 of Verloop [47] for the infinite horizon problem. Note that as mentioned in Section 8.1 of Verloop [47], 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 (5.4). By strong duality, there exist Lagrange multipliers $\gamma_0^*, \ldots, \gamma_{T-1}^*$ corresponding to the constraints (5.4c), such that $\{\mathbf{y}^*(t)\}_{0\le t\le T-1}$ is also an optimal solution of the following problem:

$$\max_{\mathbf{y}\ge \mathbf{0}} \quad \sum_{t=0}^{T-1}\sum_{s,a}(R_s^a - a\gamma_t^*)y_{s,a}(t) \tag{5.33a}$$

$$\text{s.t.} \quad y_{s,0}(t+1) + y_{s,1}(t+1) = \sum_{s',a} y_{s',a}(t)P_{s's}^a \qquad \forall s,t, \tag{5.33b}$$

$$y_{s,0}(0) + y_{s,1}(0) = m_s(0) \qquad \forall s. \tag{5.33c}$$

The above linear program (5.33) can be cast into a MDP denoted as $X(t)$ 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 [44] 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

$$I_s(t) := Q_{s,1}(t) - Q_{s,0}(t). \tag{5.34}$$

The *LP-index policy* is then defined as the water-filling policy, by using the values $I_s(t)$ in (5.34) 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 [12], and is the finite horizon analogue of the LP-indices for infinite horizon problems we discussed in Chapter 4.

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 [12].

Lemma 5.5.1. *The LP-indices are such that $I_s(t) \ge 0$ for all $s \in \mathcal{S}^+(t)$, $I_s(t) \le 0$ for all $s \in \mathcal{S}^-(t)$ and $I_s(t) = 0$ for all $s \in \mathcal{S}^0(t)$.*

Proof. The proof is similar to its analogue in infinite horizon stated in Proposition 4.5.1. Let ψ^* be an optimal Markovian stationary policy of (5.33) 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 $0 < \psi^*_{s,0}(t) < 1 \Rightarrow I_s(t) = 0$.

□

5.5.2 The LP-update policy

One potential drawback of the Lipschitz continuous policies with their $O(1/\sqrt{N})$ convergence rate proven in Theorem 5.3.2 is that, the constant $C > 0$ in inequality (5.22) 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 (5.4) 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 $0 \le t \le T - 1$.

Note that solving the LP problem (5.4) 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 $O(1/\sqrt{N})$, as any LP-compatible Lipschitz continuous policy does.

Theorem 5.5.2. *Let the LP-update policy be defined as above, and denote by $V^{(N)}_{\text{LP-update}}(\mathbf{m}(0), T)$ the value of LP-update policy on a restless bandit with parameter set $\{\mathbf{m}(0), T\}$. Then there exists a constant $C' > 0$ independent of N such that*

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

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

Proof. Denote by $\mathbf{y}^{t*}$ the solution of the LP (5.4) with parameter set $\left\{\mathbf{M}^{(N)}(t), T-t\right\}$ at decision epoch t. Write similarly $\mathbf{m}^{t*}$ where $m_s^{t*}(t') = y_{s,0}^{t*}(t') + y_{s,1}^{t*}(t')$ for $t \le t' \le T-1$ and $s \in \mathcal{S}$. Bellman's principle of optimality gives

$$V_{\text{rel}}(\mathbf{M}^{(N)}(t), T-t) = \sum_{s,a} y_{s,a}^{t*}(t) R_s^a + V_{\text{rel}}(\mathbf{m}^{t*}(t+1), T-(t+1)), \tag{5.35}$$

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

$$V_{\text{LP-update}}^{(N)}(\mathbf{M}^{(N)}(t), T-t) = \sum_{s,a} y_{s,a}^{t*}(t) R_s^a + \mathbb{E}\left[V_{\text{LP-update}}^{(N)}(\mathbf{M}^{(N)}(t+1), T-(t+1))\right]. \tag{5.36}$$

Denote by $Z(t) := V_{\text{LP-update}}^{(N)}(\mathbf{M}^{(N)}(t), T-t) - V_{\text{rel}}(\mathbf{M}^{(N)}(t), T-t)$ the difference between (5.35) and (5.36), one has $Z(T) = 0$ and for all $t \in \{1 \dots T-1\}$:

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

From the general theory of linear programming (see for instance Section 5.6.2 of Boyd and Vandenberghe [11]), the function $V_{\text{rel}}(\,\cdot\,, t) : \Delta^d \to \mathbb{R}$ is Lipschitz continuous with a constant denoted K_t. Let $K := \max_t K_t$. We have:

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

By Lemma 5.3.1 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 $\left\|\mathbf{Y}^{(N)}(t) - \mathbf{y}^{t*}(t)\right\|_1 \le 2d/N$ where the term $2d/N$ is caused by randomized rounding and is of order $O(\frac{1}{N})$. Recall also that $\phi(\cdot)$ is a Lipschitz function with Lipschitz constant ℓ. The dominating error hence comes from $\mathbb{E}\left[\mathbf{E}^{(N)}(t) \mid \mathbf{Y}^{(N)}(t)\right] \le c_\phi/\sqrt{N}$, where $c_\phi > 0$ is a constant independent of T and N. We therefore can bound:

$$\left|V_{\text{LP-update}}^{(N)}(\mathbf{m}(0), T) - V_{\text{rel}}^{(N)}(\alpha)\right| \le \frac{2KTc_\phi}{\sqrt{N}}. \tag{5.37}$$

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 (5.22) into a rate of $2KTc_\phi$ in (5.37), where K is an upper-bound on the Lipschitz constant $\{K_t\}_{t\ge 0}$ of the sequence of functions $\{V_{\text{rel}}(\,\cdot\,, t)\}_{t\ge 0}$. Numerical evidence suggests that the sequence $\{K_t\}_{t\ge 0}$ is in general bounded by a constant independent of T. If this is true, then the constant C' of (5.37) grows linearly with time, which is much

smaller than the exponential growth of the one in (5.22). This suggests that the LP-update policy should perform better than its non-update counterpart. We discuss the comparison between the two approaches in more details in our numerical experiments.

Later on in Chapter 6, we shall analyse the LP-update policy thoroughly under the more general weakly coupled MDPs framework. As a consequence of the results proven therein, the LP-update policy converges exponentially fast on non-degenerate problems, which is expected.

5.6 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 5.6.1. We next show the advantage of the LP-update policy to the LP-index policy on the applicant screening problem in Section 5.6.2, a model proposed in Brown and Smith [12].

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

The water-filling policy defined in Section 5.4.2 is not uniquely defined as it depends on the tie-breaking rule within $\mathcal{S}^+$ and $\mathcal{S}^-$. In Figure 5.1, we compare the two tie-breaking rules:

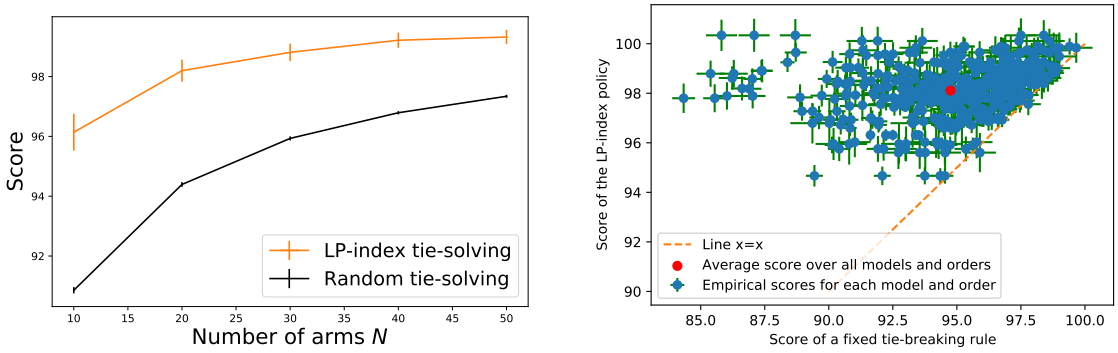
- LP-index: Give priority to the highest LP-index first, defined in Section 5.5.1;
- Random tie-solving: Ties within $\mathcal{S}^+$ and $\mathcal{S}^-$ are solved according to a random priority order that is drawn at the beginning of each simulation. The reported number for this policy is the average 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 restless bandit, recall that V_{rel} is the value of the linear program (5.4) 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:

$$\text{score}_\pi^N = \frac{V_\pi^N - V_{\text{rel}-\min}}{V_{\text{rel}} - V_{\text{rel}-\min}}. \tag{5.38}$$

The score is a number between 0 and 1 (higher being better). Theorem 5.4.4 shows that, any water-filling policy is asymptotically optimal, regardless of the tie-breaking used within $\mathcal{S}^+$ or $\mathcal{S}^-$, *i.e.* $\lim_{N\to\infty} \text{score}_\pi^N = 1$.



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

Figure 5.1 – 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.

Figure 5.1 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.

5.6.2 Case study: applicant screening problem

We discuss in this section the applicant screening problem proposed in Brown and Smith [12], 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 restless bandit 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 $\mathbb{E}\left[p_n \mid s_n\right] = a_n/(a_n + b_n)$. 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 [12], 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 5.2, the p_n are generated uniformly between 0 and 1.

- **Wrong prior** – On the right-panel of Figure 5.2, 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 this chapter, 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.

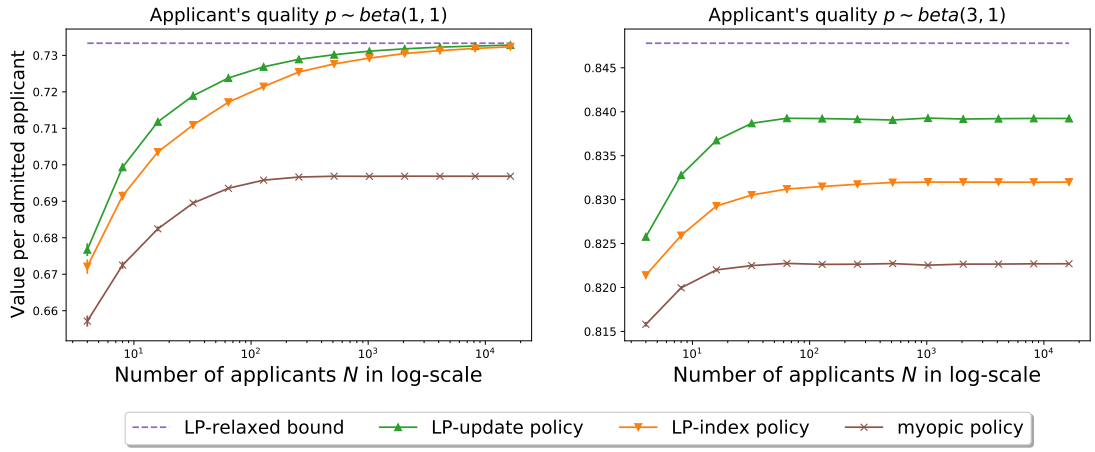


Figure 5.2 – Performance on applicant screening problem when the decision maker knows the prior distribution of p_n (left panel) or has access to a wrong prior information (right panel).

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 [12]. For this scenario, Theorem 5.4.4 and Theorem 5.5.2 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 5.2). 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 restless bandit model, as shown in the next chapter when we apply the LP-update policy on the generalized applicant screening problem.

Conclusion of the Chapter

In this chapter we have provided a general framework to construct control policies on finite horizon restless bandits, via the LP approach. It is guided by the principle that the control on the stochastic bandit should follow as much as possible the deterministically optimal one. Much like the class of LP-priority policies in infinite horizon, there is a large collection of policies that are equally good in the asymptotic regime, but may differ on performance for finite values of N. We propose the water-filling policy as a concrete construction, and its refinement, the LP-index policy, that practically gives among the best performance of all these policies.

The LP-update policy is mentioned as yet another improvement, at the exchange of efficiency, that essentially recomputes new LP-index policies during the evolution of the process. This idea will be further developed in the next chapter, via two aspects: we shall see that the LP-update policy can be generalized straightforwardly into the more general model of weakly coupled MDPs, and that resolving the LP is not always necessary in order to apply an update—we can do it in a more efficient way.

6

The LP-update Policy and its Generalization to Weakly Coupled MDPs

In the previous chapter we have briefly mentioned the LP-update policy as a possible way to improve the performance of the usual LP-index policy. As its name indicates, the LP-update policy amounts to periodically recompute a new LP-index policy. It turns out that this simple idea of applying updates is much more powerful than it might look like, as first of all it can be generalized straightforwardly into the more general framework of weakly coupled MDPs. Moreover, the number of times for updates can be reduced, after a careful analysis of the structure of the previously computed LP solutions, rendering the LP-update policy more efficient, which is the major drawback of the naive LP-update policy. We then apply the improved LP-update policy on the generalized applicant screening problem, studied in the previous chapter.

> It is my experience that proofs involving matrices can be shortened by 50% if one throws the matrices out.
>
> Emil Artin

6.1 Introduction

Markov decision processes (MDP) have proven tremendously useful as models of stochastic sequential planning problems. Dynamic programming is the principal method to address these problems under uncertainty. While generally applicable, the computational difficulty of applying classic dynamic programming algorithms to realistic problems has triggered much research into techniques to deal with large state and action spaces. One such technique is *decomposition*, for which the very large global

MDP is decomposed into N loosely dependent sub-processes, and we solve independently these N sub-problems that are exponentially smaller in size compared to the global problem. If these solutions can be pieced together effectively, and used to guide the search for a global solution that performs well, then dramatic improvements in the overall solution time can be obtained.

In this chapter we study *weakly coupled* MDPs that fall into this situation. The model originates from sequential stochastic resource allocation problems: A number of tasks must be addressed and actions consists of assigning various resource at every decision epoch to each of these tasks. We assume that each of these tasks is *additive utility independent*: the utility of achieving any collection of tasks is the sum of rewards associated with each task. In addition, we assume that each task can be viewed as an independent sub-process whose rewards and transitions are independent of the others, given a fixed action or policy. These tasks are linked only via the global resource constraints at each decision epoch, that explains the terminology "weakly coupled". Weakly coupled MDPs are widely applicable in practice, and can be used to model various scheduling, queueing, supply chain as well as health care problems.

One classical example of a weakly coupled MDP is the multi-armed restless bandit model first appeared in the famous paper Whittle [49]. In this model the bandit (global MDP) consists of N arms (sub-MDPs) that are supposed to be statistically independent, and are coupled via a single constraint: the total available resource is αN with $0 < \alpha < 1$. Two actions are possible on each arm: The passive action consumes zero unit of resource, while the active action consumes one unit of resource. By relaxing the resource constraint to require that it is only satisfied in expectation, the global N-armed bandit problem is effectively decomposed into N independent sub-problems, one for each arm. Solving these N sub-problems returns a real value index to each arm. To piece together these solutions into a feasible solution to the original global problem, we choose the active action on the αN arms having the largest indices. This is the famous Whittle index policy.

Under several additional technical assumptions, the Whittle index policy is proven to be asymptotically optimal in Weber and Weiss [48], in the sense that the gap between the performance of the Whittle index policy and the optimal policy converges to zero when N goes to infinity. Subsequently, following the same spirit, many asymptotic optimality results have been obtained for the multi-armed bandit model, under either finite or infinite horizons, with two or multiple actions available for each arm, together with a single global resource constraint. However, to the best of our knowledge, none of the asymptotic optimality results on multi-armed bandits have been generalized to weakly coupled MDPs (i.e. multi-action multi-armed bandits having multiple resource constraints). Furthermore, existing policies on weakly coupled MDPs in the literature have not been proven to be asymptotically optimal. Our work comes in to fill this gap.

Related work

There are several branches of works in the existing literature that are related to our results in this chapter. The first branch is concentrated on the study of weakly coupled MDPs under finite horizon, e.g. Adelman and Mersereau [2], Astaraky and Patrick [5], Dolgov and Durfee [13], Gocgun and Ghate [24], Gocgun and Ghate [25], Meuleau et al. [34], Patrick et al. [43], as well as the two PhD thesis Hawkins [26] and Salemi Parizi [46]. In these works the authors consider the situation when the sub-MDPs are not statistically identical. Under this generality, the results rely on a Lagrange decomposition technique to solve approximately the above problem, which works as follows: By relaxing the constraints with state independent multipliers, the N-dimensional optimization problem decouples into N one-dimensional subproblems, each involves the corresponding sub-MDP alone; the key is in choosing the value of multipliers and in transforming the optimal controls of the decoupled problems into a feasible policy to the original problem (as has been done for the simpler multi-armed bandit model), that is near optimal in practice. The focus of this chapter however, is on the special case where each sub-MDP is statistically identical, and we provide a method to construct policies that are not only close to being optimal, but can actually been proven theoretically to be asymptotically optimal when N goes to infinity. We also go further to compute the optimal convergence rate for several classes of weakly coupled MDPs.

Our results are also close to the branch of researches on asymptotic optimality results on multi-armed bandits under finite horizon as in Brown and Smith [12], Zhang and Frazier [53], Zayas-Cabán et al. [52] and Xiong et al. [50]. The finite horizon two-action single-constraint multi-armed bandit is the subject of the first two papers, the more general multi-action single-constraint multi-armed bandit is the subject of the two latter papers. All policies considered in these references are one-pass policies that involve solving a linear program only once at the very beginning. The non-degenerate property on two-action single-constraint multi-armed bandit, for which we have defined in Chapters 4 and 5 for the infinite and finite horizon problems, has also been proposed independently in Zhang and Frazier [53], together with the $O(1/N)$ rate on (finite horizon) non-degenerate models proven therein. However, none of the existing literature has ever considered the problem under the much more general weakly couple MDPs framework.

There is another related branch of works that consider the problem under infinite horizon. In Whittle [49], Weber and Weiss [48] and Verloop [47] the two-action bandit has been studied. The asymptotic optimality is proven in Weber and Weiss [48] under the additional indexability and global attractor assumptions. Later in Verloop [47] the indexability assumption has been removed. In Chapters 3 and 4 of the current thesis the optimal convergence rate of these asymptotic results are proven. In Hodge and Glazebrook [27] and Niño-Mora [40] the model is generalized to the multi-action single-constraint case. However the focus of the two latter papers is more on the generalization of indexability first proposed in Whittle [49]. For a comparison, we shall point out that when studying the problem under finite horizon as we do here,

none of the assumptions of indexability nor the global attractor property are needed for the asymptotic optimality results. On the other hand, the policies become time-dependent and the computation time may increase considerably with the horizon.

Summary of contribution

Our contribution is to generalize the asymptotic optimality results on multi-armed bandits to weakly coupled MDPs. First, we design a policy that becomes optimal as the number of components grows for general weakly coupled MDPs and then provide theoretical results on its optimal convergence rate. Note that having multiple constraints instead of one makes all the previous index-type policies not directly generalizable, since the index of a sub-process can only be defined with respect to a single constraint, and it is not obvious to incorporate multiple constraints into one real-valued meaningful index.

Nevertheless, solutions can be found by generalizing two policies of non index-type in the existing literature on bandits: the first originates from the randomized activation control policy in Zayas-Cabán et al. [52] and the occupancy-measured-reward index policy in Xiong et al. [50], that we call "the occupation measure policy" in Algorithm 3. It samples an action for each arm based on the occupation measure obtained from the solution of the relaxed problem (a linear program), subject to not violating any budget constraint. The second is the LP-update policy proposed in the previous Chapter 5, that solves new linear programs at each decision epoch, and the decision at each time is based on the solution to the linear program at that time.

However, these two generalizations do not give satisfying results: the occupation measure policy, being simple in its idea, turns out to be complicated for the theoretical analysis, and in general gives poor performance compared to other more sophisticated policies, as we show later in our case study in Section 6.5. At the very least, it is already not straightforward to show that it is a locally linear policy as defined in Chapter 5 for non-degenerate two-action bandits to ensure that it has the faster $O(1/N)$ rate (as opposed to the classical $O(1/\sqrt{N})$ rate) on this class of problems. On the other hand, the LP-update policy proposed in the previous chapter for the two-action bandits comes at the price of a much longer computation time caused by solving new linear programs at every step.

Concerning the convergence rate, which is the main topic of both Zhang and Frazier [53] and Chapter 5, the key concept is the non-degenerate property. Recall that we show in Chapter 5 that this property is a necessary and sufficient condition for two-action bandit problems to admit policies with $O(1/N)$ convergence rate, and we construct the so-called LP-index policy to achieve this rate, as well as the LP-update policy, which was supposed to even outperform the LP-index policy, but we could only prove a general $O(1/\sqrt{N})$ convergence rate in Theorem 5.5.2. The reason being that this policy is of a different nature, so the local linearity (as a consequence of non-degeneracy) can not be established in the same manner as for the LP-index policy.

In this chapter we generalize the LP-update policy defined in Chapter 5, as well

as the non-degenerate property to weakly coupled MDPs. Our policy shares the same advantage on performance as its predecessor, while in the meantime being more efficient, in the sense that it only solves new linear programs when necessary, and the algorithm knows when this necessity takes place. Remarkably, this information of necessity is given at the same moment when the algorithm checks the non-degenerate property of the problem. We then go on to provide theoretical proofs for the optimal convergence rate of the policy, which was not done in the previous chapter for two-action bandits. More precisely, we show that

1. The LP-update policy is asymptotically optimal at rate $O(1/\sqrt{N})$. Moreover, we improve this rate to $O(1/N)$ for non-degenerate weakly coupled MDPs. If the problem satisfies an additional perfect rounding condition, then the rate can be further improved to be exponentially fast along this sequence.

2. We also show that the convergence rate claimed previously, i.e. $O(1/\sqrt{N})$ in general and $O(1/N)$ for non-degenerate problems are tight.

Finally, we show a case study generalizing the applicant screening problem. This problem is first proposed in the paper of Brown and Smith [12], and has been studied again in Section 5.6.2. It is modeled as a two-action single-constraint multi-armed bandit. In this chapter we generalize the problem by allowing more actions and adding fairness constraints, so that it fits naturally into the weakly coupled MDPs framework. We show that the LP-update policy outperforms the generalized occupation measure policy, and the smaller is N, the more apparent is this advantage; whereas the LP-update policy is not much slower (roughly three to four times). This makes the latter a much better alternative for the decision maker, under a circumstance when the performance is critical.

Outline

The rest of the chapter is organized as follows: We introduce the weakly coupled MDPs model in Section 6.2. A first version of the LP-update policy, that is a direct generalization of the previous one defined in Chapter 5, is introduced in Section 6.3. The non-degenerate property and the improved version of the LP-update policy for weakly coupled MDPs is given in Section 6.4. The case study on the generalized applicant screening problem is given in Section 6.5. The proofs of the main theorems are given in Section 6.6.

6.2 Model Description

6.2.1 The weakly-coupled MDPs

We consider a finite-horizon discrete-time weakly coupled MDP composed of N statistically identical sub-MDPs, indexed by $n \in \{1 \dots N\}$. The finite state space of each

sub-MDP is the set $\mathcal{S} := \{1, 2, \dots, d\}$, and its finite action space[1] is $\mathcal{A} := \{0, 1, \dots, A\}$. The state space of the weakly coupled MDP is therefore $\mathcal{S}^N$ and the action space is a subset of $\mathcal{A}^N$. There are J types of resources, and the decision maker is allowed to use up to b_j resource of type j at each decision epoch. We assume that taking the action a_n for the component n that in state s_n uses $D_j(s_n, a_n) \geq 0$ of resource j, and that the action 0 consumes no resource: $D_j(s_n, 0) = 0$ for all $s_n \in \mathcal{S}$. Hence, the set of possible actions in state $\mathbf{s} = (s_1, s_2, \dots, s_N)$ is the set of $\mathbf{a} \in \mathcal{A}^N$ such that for all $j \in \{1 \dots J\}$: $\sum_{n=1}^{N} D_j(s_n, a_n) \leq N \cdot b_j$.

Upon choosing an action $\mathbf{a}$ in state $\mathbf{s}$, the decision maker receives a reward $\sum_{n=1}^{N} R_{s_n}^{a_n}$. We assume that the sub-processes are weakly coupled, in the sense that the N sub-MDPs are only linked through the J constraints that link the actions that can be taken in each component, *i.e.*, for a given action $\mathbf{a}$, the system transitions from a state $\mathbf{s}$ to state $\mathbf{s}' = (s'_1, s'_2, \dots, s'_N)$ with probability

$$p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) = \prod_{n=1}^{N} p(s'_n \mid s_n, a_n) = \prod_{n=1}^{N} P_{s_n, s'_n}^{a_n}, \tag{6.1}$$

where for each action a, the matrix $\mathbf{P}^a$ is a probability transition matrix of dimension $d \times d$.

This model makes a number of assumptions, that are classical in the literature (*e.g.*, Xiong et al. [50], Zayas-Cabán et al. [52]):

- We assume that all terms in $D(s, a)$ and $\mathbf{B}$ are non-negative numbers, and that $D(s, 0) = 0$. This is a natural assumption under the resource allocation context in which $a = 0$ corresponds to a *passive* action that consumes no resource. The later also implies that our resource constraint problem has at least a feasible solution by always choosing the passive action. A further remark on this point will be elaborated in Section 6.4.3.

- We assume that each sub-MDP is statistically identical. This is needed under our scaling of the arm population N. However, this assumption can be relaxed in the sense that we can incorporate the case where there is a finite number of types of sub-MDPs, by making direct sum of the state spaces of each type into a single state space.

- In our formulation of the problem, the rewards, constraints and transition probabilities do not dependent on time. This choice is only to lighten the notations: As we consider finite-horizon problem, all the results apply to the case of time-dependent parameters (it would suffice to add a dependence on t on all parameters).

Finally, we remark that this model includes the classical restless multi-armed bandit considered in Chapter 5, that corresponds to the case $\mathcal{A} = \{0, 1\}$ with $J = 1$ constraint

[1] For notational simplicity, we assume that the sub-action space of each sub-MDP n is independent of S_n. This is without loss of generality as one may forbid some actions by giving them an extremely high cost.

and $D_1(s,a) = a$. This simpler optimization problem is already PSPACE-hard (see Papadimitriou and Tsitsiklis [42]). In the rest of this chapter we focus on developing approximate solutions whose performance are provably close to optimal.

6.2.2 Symmetry simplification and population representation

Since we assume that the N tasks are statistically identical, the problem can be reformulated by keeping track of the number of sub-MDPs in each of the d states (population), and by tracking what actions is taken in which state. It will be convenient for our later consideration to normalize every quantity by dividing by N the population of the sub-MDPs.

We denote by $\mathbf{M}^{(N)}(t) = (M_s^{(N)}(t))_{s\in\mathcal{S}}$ the proportion of sub-MDPs that are in state s at decision epoch t. We denote by $\Delta^{(N),d}$ the possible values for $\mathbf{M}^{(N)}$, which is the set of vectors $\mathbf{m} \in \mathbb{R}^d$ such that $m_s \geq 0$, $\sum_{s\in\mathcal{S}} m_s = 1$ and Nm_s is an integer for all $s \in \mathcal{S}$. Let $\mathcal{U} = \mathcal{S} \times \mathcal{A}$ be the set of state-action pairs and denote $u = |\mathcal{U}|$. Upon observing $M_s^{(N)}(t)$, the action taken by decision maker is represented by $\mathbf{y} \in \mathbb{R}^u$ where $y_{s,a}$ is the proportion of sub-MDPs that are in state s and for which action a is taken. For a given $\mathbf{m} \in \Delta^{(N),d}$, we denote by $\mathcal{Y}^{(N)}(\mathbf{m})$ the set of possible actions. It is the set of possible $\mathbf{y} \in \mathcal{Y}(\mathbf{m})$ such that for all s, a, $Ny_{s,a}$ is an integer, where $\mathcal{Y}(\mathbf{m})$ is:

$$\mathcal{Y}(\mathbf{m}) := \left\{ \mathbf{y} \geq 0 \text{ such that:} \forall s \in \mathcal{S} : \sum_{a\in\mathcal{A}} y_{s,a} = m_s; D\mathbf{y} \leq \mathbf{B}; \right\}. \tag{6.2}$$

In the above definition, the first constraint guarantees that exactly one action is chosen for each arm, and the second represents the budget constraint. The budget constraints are written in matrix form by viewing $D_j(s,a)$ as a matrix whose lines are indexed by j and whose columns are indexed by (s,a). Similarly, the vector $\mathbf{B}$ is a column vector indexed by j. By definition, $\sum_{s=1}^{d} M_s^{(N)}(t) = 1$. This reduces the state space size of the MDP considerably from d^N to $\binom{N+d-1}{d-1}$, but this is still intractable when N and d are large.

6.2.3 Optimal control formulation

The decision maker's goal is to maximize the total expected reward over a finite horizon T, given an initial state configuration vector $\mathbf{m}(0) \in \Delta^{(N),d}$ of the system. 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.

This optimization problem can be written using the aggregated variables $\mathbf{Y}^{(N)}$ and $\mathbf{M}^{(N)}$ as follows.

$$V_{\text{opt}}^{(N)}(\mathbf{m}(0), T) = \max_{\mathbf{Y}} \quad \mathbb{E}\Big[\sum_{t=0}^{T-1} \sum_{(s,a)\in\mathcal{U}} R_s^a Y_{s,a}^{(N)}(t)\Big] \tag{6.3a}$$

$$\text{s.t.} \quad M_s^{(N)}(0) = m_s(0) \ \ \forall s, \tag{6.3b}$$

$$\mathbf{M}^{(N)}(t+1) \text{ follows the Markov transitions given } \mathbf{Y}^{(N)}(t)\ , \tag{6.3c}$$

$$\mathbf{Y}^{(N)}(t) \in \mathcal{Y}^{(N)}(\mathbf{M}^{(N)}(t)) \ \ \forall t \tag{6.3d}$$

6.3 The LP-update Policy for Weakly Coupled MDPs

We first introduce the linear program after relaxing the resource constraints in Section 6.3.1, which is the starting point for all policies we consider afterwards. We then define the LP-update policy with full updates and present the $O(1/\sqrt{N})$ performance guarantee that holds in all cases in Section . Some proofs of the results are postponed to Section 6.6.

6.3.1 The relaxed problem as a linear program

The main difficulty of the optimization problem is that the constraint $D\mathbf{Y}^{(N)}(t) \leq \mathbf{B}$ couples all sub-MDPs. To overcome this difficulty, a now classical approach (Brown and Smith [12], Xiong et al. [50], Zayas-Cabán et al. [52], Zhang and Frazier [53]) is to relax this constraint and consider a problem where this constraint has to be satisfied only in expectation: $D\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right] \leq \mathbf{B}$. This lead us to write a relaxed optimization problem in terms of the variables $\mathbf{y}(t) = \mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right]$. Indeed, (6.1) implies that the expectation of $\mathbf{M}^{(N)}(t+1)$ given $\mathbf{Y}^{(N)}(t)$ can be rewritten as a linear map ϕ as follows:

$$\mathbb{E}\left[M_s^{(N)}(t+1) \mid \mathbf{Y}^{(N)}(t) = \mathbf{y}\right] = (\phi(\mathbf{y}))_s := \sum_{(s',a)\in\mathcal{U}} y_{s',a} P^a_{s',s}. \tag{6.4}$$

This shows that the relaxed optimization problem is the following linear program with variables $\mathbf{y}(t) = \mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right]$:

$$V_{\text{rel}}(\mathbf{m}(0), T) = \max_{\mathbf{y} \geq 0} \quad \sum_{t=0}^{T-1} \sum_{(s,a)\in\mathcal{U}} R^a_s y_{s,a}(t) \tag{6.5a}$$

$$\text{s.t.} \quad \sum_{a\in\mathcal{A}} y_{s,a}(0) = m_s(0) \qquad \forall s, \tag{6.5b}$$

$$\sum_{a\in\mathcal{A}} y_{s,a}(t+1) = \big(\phi(\mathbf{y}(t))\big)_s \qquad \forall s, t, \tag{6.5c}$$

$$D\mathbf{y}(t) \leq \mathbf{B} \qquad \forall t, \tag{6.5d}$$

where (6.5b) corresponds to the condition on the initial state. The constraint (6.5c) corresponds to the time-evolution (6.1) plus the fact that $M_s(t+1) = \sum_a Y_{s,a}(t+1)$. Last, the constraint (6.5d) is $D\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right] \leq \mathbf{B}$ which is the relaxed version of (6.2), it also implies that $\mathbf{y}(t) \in \mathcal{Y}(\mathbf{m})$, where $m_s(t) = \sum_{a\in\mathcal{A}} y_{s,a}(t)$.

By the assumptions that $D(s,0) = 0$ and $D, \mathbf{B} \geq \mathbf{0}$, the linear program (6.5) is feasible (*e.g.* it suffices to always choose the passive action $a = 0$). In the following,

Algorithm 1: LP-update policy for weakly coupled MDPs (full updates).

Input: Time horizon T and initial configuration vector $\mathbf{m}(0)$.

for $t = 0, \dots, T-1$ **do**

  Observe the current configuration and compute $\mathbf{y}^*(t)$, which is any of the solutions of the LP (6.5) with parameters $(\mathbf{M}^{(N)}(t), T-t)$;

  Set $Y^{(N)}_{s,a}(t) = N^{-1} \left\lfloor N y^*_{s,a}(t) \right\rfloor$ for $a \neq 0$ and $Y^{(N)}_{s,0}(t) = M^{(N)}_{s,0}(t) - \sum_{a \neq 0} Y^{(N)}_{s,a}(t)$;

  Use actions $Y^{(N)}_{s,a}(t)$ over all sub-MDPs to advance to the next timestep;

end

we denote by $\mathbf{y}^*$ one of its optimal solution, and by $\mathbf{m}^*$ the sequence of vectors $\mathbf{m}^*(t)$ such that $m^*_s(t) = \sum_{a \in \mathcal{A}} y^*_{s,a}(t)$. It is the optimal state configuration vector $\mathbf{m}^*(t)$ on the relaxed problem.

6.3.2 The LP-update policy

Let us denote by $\mathbf{y}^*(t)$ be a sequence of optimal decisions for the relaxed problem and let $\mathbf{m}^*(t) := \sum_a \mathbf{y} * (t)$. To construct a policy for the system of size N, this suggests to use $\mathbf{Y}^{(N)}(t) = \mathbf{y}^*(t)$. Yet, this is in general not possible because of random fluctuations: Indeed, it is likely that $\mathbf{M}^{(N)}(t) \neq \mathbf{m}^*(t)$ which implies that, in general, $\mathbf{y}^*(t)$ is not feasible for $\mathbf{M}^{(N)}(t)$, that is

$$\text{In general: } \mathbf{y}^*(t) \notin \mathcal{Y}^{(N)}(\mathbf{M}^{(N)}(t)). \tag{6.6}$$

The classical way to solve this problem in the literature is to construct a sequence of decision rules $\pi_t : \Delta^d \to \Delta^{d(A+1)}$ such that $\pi_t(\mathbf{m}) \in \mathcal{Y}^{(N)}(\mathbf{m})$ and $\pi_t(\mathbf{m}^*(t)) = \mathbf{y}^*(t)$. This is what is used to build the randomized activation control policy in Zayas-Cabán et al. [52], the fluid-priority policies in Zhang and Frazier [53], the occupancy-measured-reward index policy in Xiong et al. [50], to name a few. In particular, it is shown in Chapter 5 that any such policy is $O(1/\sqrt{N})$ optimal if all the decision rules π_t are Lipschitz-continuous.

In this chapter, we adopt another approach, that we call the *LP-update* policy, which is a generalization of the LP-update policy of Chapter 5 introduced for two-action bandits, described as follows: at each decision epoch, we solve a new LP program starting from $\mathbf{M}^{(N)}(t)$ with horizon $T-t$. This guarantees that the newly computed $\mathbf{y}^*(t)$ is in $\mathcal{Y}(\mathbf{M}^{(N)}(t))$, by constraint (6.5b). However, this control is not necessarily feasible for the system of size N because $N y^*_{s,a}(t)$ is not necessarily an integer. The idea is then to use a rounding procedure. A naive way to do so is to use $Y^{(N)}_{s,a}(t) = N^{-1} \left\lfloor N y^*_{s,a}(t) \right\rfloor$. We will discuss an advanced rounding procedure in Section 6.4.3. All this leads to our first LP-update algorithm, that is detailed in Algorithm 1.

Similarly to what is done in Chapter 5, the next Theorem 6.3.1 shows that this algorithm is $O(1/\sqrt{N})$-optimal. The proof is an (almost) direct adaptation to the case of multi-action multi-constraint bandit.

Theorem 6.3.1. *Denote by $V^{(N)}_{\text{LP-update}}(\mathbf{m}(0), T)$ the value of the LP-update policy computed by applying Algorithm 1 and by $V_{\text{rel}}(\mathbf{m}(0), T)$ the value of the linear program (6.5).*

(i) For any weakly coupled MDPs with statistically identical arms, there exists a constant $C > 0$ such that for all N:

$$\left| V^{(N)}_{\text{LP-update}}(\mathbf{m}(0), T) - V_{\text{rel}}(\mathbf{m}(0), T) \right| \leq \frac{C}{\sqrt{N}}.$$

(ii) There exists a weakly coupled MDP with statistically identical arms and a constant $C' > 0$ such that for all N:

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

Proof. The proof of the lower bound (ii) is done in Section 6.6.3. Below, we prove (i).

We first analyze the Algorithm 1 that performs an update at each time step. At time t, the LP-update algorithm chooses a vector $\mathbf{Y}^*(t)$ that is optimal for the LP-program (6.5) with parameters $(\mathbf{M}^{(N)}(t), T - t)$, and then chooses a decision vector $\mathbf{Y}^{(N)}(t)$ such that $\|\mathbf{Y}^{(N)}(t) - \mathbf{Y}(t)\| \leq c/N$. The controller then earns $R^T\mathbf{Y}^{(N)}(t)$ and moves to the next step. Hence:

$$V^{(N)}_{\text{LP-update}}(\mathbf{M}^{(N)}(t), T - t) = \mathbb{E}\left[R^T\mathbf{Y}^{(N)}(t) + V^{(N)}_{\text{LP-update}}(\mathbf{M}^{(N)}(t+1), T - t - 1) \right] \quad (6.7)$$

Moreover, by the dynamic algorithm principle

$$V_{\text{rel}}(\mathbf{M}^{(N)}(t), T - t) = R^T\mathbf{Y}(t) + V_{\text{rel}}(\phi(\mathbf{Y}^{(N)}(t)), T - t - 1). \quad (6.8)$$

Let $Z(t) = \mathbb{E}\left[V^{(N)}_{\text{LP-update}}(\mathbf{M}^{(N)}(t), T - t) - V_{\text{rel}}(\mathbf{M}^{(N)}(t), T - t) \right]$. Combining (6.7) and (6.8) show that $Z(t) - Z(t+1)$ is equal to

$$\underbrace{R^T\mathbb{E}\left[\mathbf{Y}^{(N)}(t) - \mathbf{Y}(t) \right]}_{\text{Term A}} + \underbrace{\mathbb{E}\left[V_{\text{rel}}(\mathbf{M}^{(N)}(t+1), T - t - 1) - V_{\text{rel}}(\phi(\mathbf{Y}^{(N)}(t)), T - t - 1) \right]}_{\text{Term B}}, \quad (6.9)$$

By construction, the Term A is of order $O(1/N)$ and is equal to 0 in case of perfect rounding. Moreover, from the general theory of linear programming (see for instance Section 5.6.2 of Boyd and Vandenberghe [11]), the function $V_{\text{rel}}(\mathbf{m}, t)$ is Lipschitz continuous in $\mathbf{m}$. Denoting L_t its Lipschitz constant, the Term B is smaller than:

$$\text{(Term B)} \leq L_{T-t-1}\mathbb{E}\left[\|\mathbf{M}^{(N)}(t+1) - \phi(\mathbf{Y}^{(N)}(t))\| \right],$$

which by Lemma 6.6.1 is of order $O(1/\sqrt{N})$. This shows the first item (upper bound) of Theorem 6.3.1.

□

This first algorithm and its performance given in Theorem 6.3.1 have two important drawbacks. The first is from a computational point of view: Algorithm 1 requires to solve a new LP at each decision epoch. This is computationally expensive and actually inefficient compared to the algorithms of [12, 18, 50, 52, 53] that solve a unique LP program at decision epoch $t = 0$. The second drawback is that its performance guarantee is only $O(1/\sqrt{N})$. While this cannot be improved for general models (as shown in Theorem 6.3.1(ii)), the other algorithms can be $O(1/N)$-optimal for all problems that are *non-degenerate*. In the next section, we propose an extended definition of non-degeneracy and show how to address the two drawbacks of the current version of the LP-update policy.

6.4 Non-degenerate Problems and Improved Convergence Rate

In this section, we define what we call a non-degenerate problem, and show how it allows one to design an improved LP-update policy with selective updates that is more efficient in Section 6.4.1. As we will see, when a problem is non-degenerate, the solution to the LP starting from an initial condition $\mathbf{m} \approx \mathbf{m}^*(t)$ are locally linear. We will show that this can be used to improve both the computational efficiency of the algorithm and the rate at which the algorithm becomes asymptotically optimal. We prove in Section 6.4.2 that the new LP-update policy has a $O(1/N)$ performance guarantee for all non-degenerate problems. We discuss questions related to rounding in Section 6.4.3, which is an important step when applying the policy. Some proofs of the results are postponed to Section 6.6.

6.4.1 The non-degenerate property

We start by noting that the linear program (6.5) can be decomposed into a T-steps optimization problem, where for each step $0 \le t \le T-1$, given $\mathbf{m}(t) \in \Delta^d$, by the principle of optimality, we can write recursively

$$V_{\text{rel}}(\mathbf{m}(t), T-t) = \max_{\mathbf{y} \in \mathbb{R}^u} \quad \mathbf{R}^\top \mathbf{y} + V_{\text{rel}}(\phi(\mathbf{y}), T-t-1) \tag{6.10a}$$

$$\text{s.t.} \quad \mathbf{y} \ge \mathbf{0}, \tag{6.10b}$$

$$D\mathbf{y} \le \mathbf{B}, \tag{6.10c}$$

$$E\mathbf{y} = \mathbf{m}(t). \tag{6.10d}$$

where E is a matrix corresponding to the equality constraint $\sum_{a \in \mathcal{A}} y_{s,a}(0) = m_s(0)$.

Let $\mathbf{y}^*(t)$ be an optimal solution of the linear program (6.5) and define $\mathcal{J}^*(t)$ as the set of indices for which the constraint (6.10c) is an equality: $(D\mathbf{y}^*(t))_j = b_j$ for all $j \in \mathcal{J}^*(t)$ and $(D\mathbf{y}^*(t))_j < b_j$ for $j \in \mathcal{J} \setminus \mathcal{J}^*(t)$. We now consider the following optimization problem:

$$F_{\mathbf{y}^*}(\mathbf{m}(t), T-t) = \max_{\mathbf{y} \in \mathbb{R}^u} \quad \mathbf{R}^\top \mathbf{y} + V_{\text{rel}}(\phi(\mathbf{y}), T-t-1) \tag{6.11a}$$

$$\text{s.t.} \quad \mathbf{y} \geq \mathbf{0}, \tag{6.11b}$$
$$(D\mathbf{y})_j < b_j \qquad \forall j \notin \mathcal{J}^*(t), \tag{6.11c}$$
$$(D\mathbf{y})_j = b_j \qquad \forall j \in \mathcal{J}^*(t), \tag{6.11d}$$
$$E\mathbf{y} = \mathbf{m}(t) \tag{6.11e}$$

As (6.11) is more constrained than (6.10), we have $F_{\mathbf{y}^*}(\mathbf{m}(t), T-1) \leq V_{\text{rel}}(\mathbf{m}(t), T-t)$. Moreover, by definition when $\mathbf{m}(t) = \mathbf{m}^*(t)$, we have $F_{\mathbf{y}^*}(\mathbf{m}^*(t), T-1) = V_{\text{rel}}(\mathbf{m}^*(t), T-t)$. In what follows, we define a condition that we call non-degeneracy that guarantees that this equality is preserved in a neighbourhood of $\mathbf{m}^*(t)$.

Let $\mathcal{U}^*(t)$ as the set of indices (s,a) for which $y^*_{(s,a)} = 0$, and $\mathcal{S}^*(t)$ be the set of indices for which $m^*_s(t) > 0$, and consider the equality constraints (6.11d) and (6.11e):

$$y_{s,a} = 0 \qquad \forall (s,a) \in \mathcal{U}^*(t) \tag{6.12}$$
$$D_j(s,a) y_{s,a} = b_j \qquad \forall j \in \mathcal{J}^*(t) \tag{6.13}$$
$$E_s(s,a) y_{s,a} = m_s(t) \qquad \forall s \in \mathcal{S}^*(t) \tag{6.14}$$

The above equalities (6.12)–(6.14) can be represented by a matrix $C^*(t)$ that has $|\mathcal{J}^*(t)| + |\mathcal{S}^*(t)| + |\mathcal{U}^*(t)|$ lines and $u = |\mathcal{U}|$ columns: $C^*(t)\mathbf{y} = [\mathbf{0}; \mathbf{B}|_{\mathcal{J}^*(t)}; \mathbf{m}(t)|_{\mathcal{S}^*(t)}]^T$, where the notations $\mathbf{B}|_{\mathcal{J}^*(t)}$ and $\mathbf{m}(t)|_{\mathcal{S}^*(t)}$ indicate that the vectors are restricted to the indices $\mathcal{J}^*(t)$ or $\mathcal{S}^*(t)$.

We are now ready to define the notion of non-degeneracy on weakly coupled MDPs, that generalizes the notion previously defined on two-action single-constraint restless bandits:

Definition 6.4.1 (Non-degeneracy)**.** *An LP-problem is non-degenerate if there exists a solution $\mathbf{y}^*$ to the LP (6.5) for which at all time $t \in \{1 \dots T-1\}$, the matrix $C^*(t)$ has rank $|\mathcal{J}^*(t)| + |\mathcal{S}^*(t)| + |\mathcal{U}^*(t)|$.*

Recall that the model of weakly coupled MDPs is a generalization of the restless bandit model studied previously, for which a notion of non-degeneracy is already introduced. We show in Section 6.6 that the above definition coincides with our previous definition when we restrict to the restless bandit model. This shows that our new definition is indeed an extension to the broader class of multi-action multi-constraint bandit problems.

Remark 6.4.2. *Our general definition of degeneracy/non-degeneracy for linear programs resembles the singularly/regularly perturbed linear programs considered in Filar et al. [14] and Avrachenkov et al. [6]. A major difference is that the perturbation in the referenced works is parameterized by a real number $\varepsilon > 0$, while in our consideration it appears on the right hand side of the constraints of the LP as* any *vector $\mathbf{m} \in \Delta^d$ in a neighbourhood of $\mathbf{m}^*$ (this will become clear in the next Proposition 6.4.3), so that it can be seen as a particular case of multi-parameter deviations. As mentioned in the book Avrachenkov et al. [8]: "multi-parameter deviations are still too complex to analyze fully, and even single-parameter deviations pose significant technical challenges". We choose the vocabulary "degenerate", instead of "singular", to make it coherent with the terminology used in the previous works [18, 53].*

For a given $\mathbf{m} \in \Delta^d$ and $\varepsilon > 0$, we define the neighbourhood of $\mathbf{m}$ of size ε as $\mathcal{B}(\mathbf{m}, \varepsilon) = \{\mathbf{m}' \in \Delta^d \mid |m'_s - m_s| \leq \varepsilon \text{ and } m'_s = 0 \text{ for all } s \text{ such that } m_s = 0\}$. As we show below, having a non-degenerate problem implies that the solution of the LP problem (6.10) are locally linear in a neighbourhood of $\mathbf{m}^*$.

Proposition 6.4.3. *Assume that the problem is non-degenerate. Then, for all time t, the matrix $C^*(t)$ has a right inverse $C^+(t)$. Moreover, there exists $\varepsilon > 0$ such that:*

- *The function $\mathbf{m} \mapsto V_{\text{rel}}(\mathbf{m}, T-t)$ is linear on $\mathcal{B}(\mathbf{m}^*(t), \varepsilon)$.*
- *Choosing*

$$\mathbf{y}(\mathbf{m}) = \mathbf{y}^*(t) + C^+(t) \begin{bmatrix} \mathbf{0}|_{\mathcal{U}^*(t)} \\ \mathbf{0}|_{\mathcal{J}^*(t)} \\ (\mathbf{m} - \mathbf{m}^*(t))|_{\mathcal{S}^*(t)} \end{bmatrix} \tag{6.15}$$

is an optimal solution of (6.10) for all $\mathbf{m} \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon)$.

Proof. By standard linear algebra arguments, a $d_1 \times d_2$ matrix of rank d_1 has a right inverse. This implies that there exists a matrix $C^+(t)$ such that $C^*(t)C^+(t)$ is the identity matrix. In particular, if $\mathbf{y}(\mathbf{m})$ is defined by (6.15). then

$$C^*(t)\mathbf{y}(\mathbf{m}) = C^*(t)\mathbf{y}^*(t) + C^*(t)C^+(t) \begin{bmatrix} \mathbf{0}|_{\mathcal{U}^*(t)} \\ \mathbf{0}|_{\mathcal{J}^*(t)} \\ (\mathbf{m} - \mathbf{m}^*(t))|_{\mathcal{S}^*(t)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b}|_{\mathcal{J}^*(t)} \\ (\mathbf{m} - \mathbf{m}^*(t))|_{\mathcal{S}^*(t)} \end{bmatrix}$$

In particular, $\mathbf{y}(\mathbf{m})$ satisfies (6.12)–(6.14). This shows that there exists $\varepsilon > 0$ such that $\mathbf{y}(\mathbf{m}) \in \mathcal{Y}(\mathbf{m})$ (*i.e.*, satisfy all constraints of (6.10)) for all $\mathbf{m} \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon)$, because the constraints that are not covered by (6.12)–(6.14) are either satisfied by strict inequalities for $\mathbf{m}^*(t)$ or correspond to $m^*_s(t) = 0$.

We now prove by a backward induction on t, that for all t, there exists $\varepsilon_t > 0$ such that the function $\mathbf{m} \mapsto V_{\text{rel}}(\mathbf{m}, T-t)$ is linear on $\mathcal{B}(\mathbf{m}^*(t)\varepsilon_t)$. This is clearly true for $t = T$ for which $V_{\text{rel}}(\mathbf{m}, 0) = 0$ for all $\mathbf{m}$.

Assume now that it holds for some $t+1 \leq T$, and denote by $g(\mathbf{m}, \mathbf{y})$ the reward provided by the control $\mathbf{y}$. As shown before, for $\mathbf{m}$ close enough to $\mathbf{m}^*$, $\mathbf{y}(\mathbf{m})$ is feasible for $\mathbf{m}$. Moreover, the induction hypothesis implies that $V_{\text{rel}}(\phi(\mathbf{y}(\mathbf{m})), T-t-1)$ is locally linear in $\mathbf{m}$ for all $\mathbf{m}$ close enough to $\mathbf{m}^*(t)$. This shows that $\mathbf{m} \mapsto g(\mathbf{m}, \mathbf{y}(\mathbf{m}))$ is locally linear on $\mathcal{B}(\mathbf{m}^*(t), \varepsilon_t)$. We argue that this implies that $\mathbf{y}(\mathbf{m})$ is the optimal control for all $\mathbf{m} \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon_t)$. Indeed:

- As $V_{\text{rel}}(\mathbf{m}, T-t)$ is the solution of a linear program where $\mathbf{m}$ is a linear constraint, it is concave in $\mathbf{m}$. Moreover, by construction $\mathbf{y}(\mathbf{m}^*(t))$ provides the optimal solution for $\mathbf{m}^*(t)$.
- Let $\mathbf{m}' = 2\mathbf{m}^*(t) - \mathbf{m}$ be the symmetric of $\mathbf{m}$ with respect to $\mathbf{m}^*$. By construction $\mathbf{m}' \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon_t)$. By concavity of V_{rel}, the possible sub-optimality of $\mathbf{y}(\mathbf{m})$ and the linearity of $g(\mathbf{m}, \mathbf{y}(\mathbf{m}))$, we have:

$$V_{\text{rel}}(\mathbf{m}^*, T-t) \geq \frac{1}{2}(V_{\text{rel}}(\mathbf{m}, T-t) + V_{\text{rel}}(\mathbf{m}', T-t))$$

$$\geq \frac{1}{2}(g(\mathbf{m}', \mathbf{y}(\mathbf{m}')) + g(\mathbf{m}, \mathbf{y}(\mathbf{m})) = g(\mathbf{m}^*, \mathbf{y}(\mathbf{m}^*)) = V_{\text{rel}}(\mathbf{m}^*, T-t).$$

This shows that the inequalities must be equalities, which shows that $\mathbf{y}(\mathbf{m})$ is optimal on $\mathcal{B}(\mathbf{m}^*(t), \varepsilon_t)$.

□

6.4.2 The improved LP-update algorithm

The definition of a non-degenerate problem suggests that the original LP-update Algorithm 1 can be implemented by only recomputing the updates when necessary:

1. When at a time t the rank condition on $C^*(t)$ is not satisfied, then one cannot compute the right inverse $C^+(t)$.
2. When at a time t, the rank condition is satisfied but the stochastic trajectory has deviated too much from the optimal deterministic one, so that the suggested decision vector $\mathbf{y}^*(t) + C^+(t)(\mathbf{m}^*(t) - \mathbf{M}^{(N)}(t))$ no longer gives a feasible decision vector, *i.e.*, is not in $\mathcal{Y}(\mathbf{M}^{(N)}(t))$.

When one of this two situations occurs, the new algorithm solves a new LP with initial state $\mathbf{M}^{(N)}(t)$ over $[t, T]$. This leads to the improved LP update algorithm that is described in Algorithm 2.

As a side remark, it is not clear that testing $\mathbf{y}(t) \in \mathcal{Y}(\mathbf{M}^{(N)}(t))$ is a sufficient condition for optimality. Hence Algorithm 2 is not exactly the same as Algorithm 1, even when no update occurs in Algorithm 2. However, the square-root convergence asserted in Theorem 6.3.1 can also be proven for Algorithm 2 using the similar idea.

Also note that Algorithm 2 makes fewer updates than Algorithm 1 and is therefore computationally more efficient. As we shall see in the proof, the difference in terms of value between the two algorithms is exponentially small for non-degenerate problems. This is why, in the following theorem, we use $V^{(N)}_{\text{LP-update}}(\mathbf{m}(0), T)$ to denote interchangeably the value of the LP-update policy defined in Algorithm 1 or 2.

Theorem 6.4.4. *Denote by $V^{(N)}_{\text{LP-update}}(\mathbf{m}(0), T)$ the value of the LP-update policy defined in Algorithm 2 (or Algorithm 1), and by $V_{\text{rel}}(\mathbf{m}(0), T)$ the value of the linear program (6.5).*

- *For any non-degenerate LP-problem, there exists constant $C > 0$ such that for all N:*

$$\left| V^{(N)}_{\text{LP-update}}(\mathbf{m}(0), T) - V_{\text{rel}}(\mathbf{m}(0), T) \right| \leq \frac{C}{N}.$$

- *There exists a non-degenerate LP-problem and a constant $C' > 0$ such that for all N:*

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

The proof the Theorem 6.4.4 is done after Theorem 6.4.6, since their proofs overlap greatly.

Algorithm 2: LP-update policy for weakly coupled MDPs (selective updates).

Input: Initial configuration vector $M^N(0) = \mathbf{m}(0)$ over time span $[0, T]$.

Set $Update = TRUE$;
for $t = 0, 1, 2, \ldots, T-1$ **do**
  if $Update = FALSE$ **then**
    Set $Update := TRUE$;
    if $C^*(t)$ *has rank* $|\mathcal{J}^*(t)| + |\mathcal{S}^*(t)|$ **then**
      Set $\mathbf{y}(t) := \mathbf{y}^*(t) + C^+(t)(\mathbf{m}^*(t) - \mathbf{M}^{(N)}(t))$;
      if $\mathbf{y}(t) \in \mathcal{Y}(\mathbf{M}^{(N)}(t))$ **then**
        set $Update := FALSE$;
      end
    end
  end
  if $Update = TRUE$ **then**
    Solve LP (6.5) with initial state $M^{(N)}(t)$ over $[t, T]$. Output is $\mathbf{y}^*, \mathbf{m}^*, C^*$ over $[t, T]$;
    Set $\mathbf{y}(t) := \mathbf{y}^*(t)$;
    Set $Update := FALSE$;
  end
  Set $Y^{(N)}_{s,a}(t) := N^{-1} \left\lfloor N y_{s,a}(t) \right\rfloor$ for $a \neq 0$ and $Y^{(N)}_{s,0}(t) = M^{(N)}_{s,0}(t) - \sum_{a \neq 0} Y^{(N)}_{s,a}(t)$;
  Use actions $\mathbf{Y}^{(N)}(t)$ to advance to the next timestep;
end

6.4.3 Rounding, perfect rounding and exponential convergence rate

In the following discussion, we fix a stage $t \in \{0, 1, \ldots, T-1\}$ and omit all the dependence on t in the notations. We focus on the following rounding problem: the system is in state $\mathbf{M}^{(N)}$ and our LP-update procedure gives us a vector $\mathbf{Y} \in \mathcal{Y}(\mathbf{M}^{(N)})$. How can we compute a rational vector $\mathbf{Y}^{(N)} \in \mathcal{Y}^{(N)}(\mathbf{M}^{(N)})$ as close as possible as the original $\mathbf{Y}$. We call this problem the *rounding problem* because the difference between $\mathcal{Y}(\mathbf{m})$ and $\mathcal{Y}^{(N)}(\mathbf{m})$ is due to the fact that since the arms of a bandit can not be separated into fractional parts, the decision vector $\mathbf{Y}^{(N)}$ must be such that all terms in $N\mathbf{Y}^{(N)}$ are integer numbers. The vector $\mathbf{Y}$ being a solution to a linear program has no reason to satisfy this property. So our goal is to construct a rational vector $\mathbf{Y}^{(N)}$ that still satisfies all the budget constraints while keeping it as close as possible to a given $\mathbf{Y}$.

By assumption, all terms in $D(s, a)$ and $\mathbf{B}$ are non-negative numbers, hence an admissible solution is to use the vector $\tilde{\mathbf{Y}}$, where

$$\tilde{Y}_{s,a} = \begin{cases} N^{-1} \lfloor N Y_{s,a} \rfloor & \text{if } a \neq 0 \\ M^{(N)}_{s,0} - \sum_{a \neq 0} \tilde{Y}_{s,a} & \text{if } a = 0. \end{cases} \tag{6.16}$$

By construction, $\tilde{\mathbf{Y}} \in \mathcal{Y}^{(N)}(\mathbf{M}^{(N)})$ and $\|\tilde{\mathbf{Y}} - \mathbf{Y}\|_1 \leq \frac{dA}{N}$. Hence, this "naive" construc-

tion is used in Algorithms 1 and 2 to construct an admissible control that is at distance $O(1/N)$ from the desired $\mathbf{Y}$. We shall point out that without the presence of a passive action that consumes no resource, which guarantees a feasible $\mathbf{Y}^{(N)}$, it is not always possible to construct a decision vector $\mathbf{Y}^{(N)} \in \mathcal{Y}^{(N)}(\mathbf{M}^{(N)})$ even if there exists a feasible $\mathbf{Y} \in \mathcal{Y}(\mathbf{M}^{(N)})$.

The constructed $\mathbf{Y}^{(N)}$ might still be far from $\mathbf{Y}$ for small values of N. To obtain a $\mathbf{Y}^{(N)}$ closer to $\mathbf{Y}$, the approach developed in Gast et al. [18] is to use what is call a *randomized rounding*, that consists in using a random variable $\mathbf{Y}^{(N)}$ such that $\mathbb{E}\left[\mathbf{Y}^{(N)}\right] = \mathbf{Y}$ and $\mathbf{Y}^{(N)} \in \mathcal{Y}^{(N)}(\mathbf{M}^{(N)})$ almost surely. If such a variable exists, we say that the problem admits a *perfect rounding*:

Definition 6.4.5 (Perfect rounding)**.** *The LP-problem admits a perfect rounding for an integer N if for all $\mathbf{m} \in \Delta^{(N),d}$ and for all $\mathbf{y} \in \mathcal{Y}(\mathbf{m})$, there exists a random variable $\mathbf{Y}^{(N)}$ such that $\mathbf{Y}^{(N)} \in \mathcal{Y}^{(N)}(\mathbf{m})$ almost surely and $\mathbb{E}\left[\mathbf{Y}^{(N)}\right] = \mathbf{y}$.*

The authors of Gast et al. [18] study the single constraint case where $\mathcal{A} = \{0, 1\}$, $D(s, a) = a$ and $d = \alpha$. They show that a perfect rounding exists for all N such that αN is an integer. Another example is when the constraints (6.10b)-(6.10c)-(6.10d) form a totally unimodular matrix for all t: in that case, the solution of the LP satisfies $\mathbf{y} \in \mathcal{Y}^{(N)}(\mathbf{M}^{(N)})$ directly and no rounding is needed.

However, there are many cases where perfect rounding is impossible. In such a case, a solution improving the simple truncation (6.16), is to find $\mathbf{Y}^{(N)} \in \mathcal{Y}^{(N)}(\mathbf{M}^{(N)})$ that minimizes the distance $\|\mathbf{Y}^{(N)} - \mathbf{Y}\|_1$ while satisfying the constraints. This can be computed by solving an integer linear program (this is used in our implementation).

In the proof of Theorem 6.4.4, the largest term in the distance between $V^{(N)}_{\text{LP-update}}(\mathbf{m}(0), T)$ and $V_{\text{rel}}(\mathbf{m}(0), T)$ is governed by the error between $\mathbf{Y}^{(N)}$ and $\mathbf{Y}$. In particular, if there exists a perfect rounding, then we obtain the much faster exponential convergence rate by avoiding this rounding error. We then obtain the following convergence theorem.

Theorem 6.4.6. *Denote by $V^{(N)}_{\text{LP-update}}(\mathbf{m}(0), T)$ the value of the LP-update policy defined in Algorithm 1 or 2, and by $V_{\text{rel}}(\mathbf{m}(0), T)$ the value of the linear program (6.5). Then*

- *If the problem is non-degenerate, then there exist constants $C_1, C_2 > 0$ such that for any N for which the problem admits a perfect rounding:*

$$\left|V^{(N)}_{\text{LP-update}}(\mathbf{m}(0), T) - V_{\text{rel}}(\mathbf{m}(0), T)\right| \leq C_1 \cdot e^{-C_2 N}.$$

- *There exists a non-degenerate problem that admits a perfect rounding for an infinite number of N, and two constants C_1', C_2' such that for all such N:*

$$\left|V^{(N)}_{\text{LP-update}}(\mathbf{m}(0), T) - V_{\text{rel}}(\mathbf{m}(0), T)\right| \geq C_1' \cdot e^{-C_2' N}.$$

Proof. (Proof of Theorems 6.4.4 and 6.4.6) In what follow, we show that with very high probability, $\mathbf{M}^{(N)}(t)$ remains close to $\mathbf{m}^*(t)$ so that no update is necessarily, and the sequence of $\mathbf{Y}(t)$ computed by Algorithm 2 will always be a solution to the original

problem. To this end, we shall bound the probability of the event when there is some t for which $\mathbf{M}^{(N)}(t)$ deviates far away from $\mathbf{m}^*(t)$ from above.

By Proposition 6.4.3, there exists $\varepsilon > 0$ such that $V_{\text{rel}}(\mathbf{m}, T-t-1)$ is linear in $\mathcal{B}(\mathbf{m}^*(t), \varepsilon)$ and such that the control $\mathbf{y}(\mathbf{m})$ defined in Proposition 6.4.3 and used in Algorithm 2 is optimal when $\mathbf{M}^{(N)} \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon)$. Call $\mathcal{E}$ the event $\mathbf{M}^{(N)}(t) \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon)$. By Lemma 6.6.2, there exists $C_1, C_2 > 0$ such that the event $\mathcal{E}$ holds with probability at least $C_1 e^{-C_2 N}$. Hence, when $\mathcal{E}$ is true, Algorithm 2 behaves as Algorithm 1. This shows that (6.7) holds also for Algorithm 2, up to an $O(e^{-C_2 N})$ term due to the (exponentially small) probability that $\mathcal{E}$ does not hold. This shows that (6.9) also holds for Algorithm 2 up to an $O(e^{-C_2 N})$ term. As $V_{\text{rel}}(\mathbf{m}, T-t-1)$ is locally linear when $\mathcal{E}$ holds, (Term B) of (6.9) is smaller than $C_1 e^{-C_2 N}$.

This shows that $\mathbb{E}[Z(t)] = \sum_{t=1} R^T \mathbb{E}\left[\mathbf{Y}^{(N)}(t) - \mathbf{Y}(t)\right] + O(e^{-C_2 N})$. Hence,

- If there is not perfect rounding, $\mathbb{E}\left[\mathbf{Y}^{(N)}(t) - \mathbf{Y}(t)\right] = O(1/N)$, which gives Theorem 6.4.4.
- If we use a perfect rounding, then $\mathbb{E}\left[\mathbf{Y}^{(N)}(t) - \mathbf{Y}(t)\right] = 0$ and the convergence rate is of order $O(e^{-C_2 N})$. This gives Theorem 6.4.6.

□

Remark 6.4.7. *The authors of [52, 50] use a randomized algorithm to compute an admissible $\mathbf{Y}^{(N)}$ from $\mathbf{y}^*$ (see our description of their algorithm in Algorithm 3). Their algorithm initializes the action as all arms take the passive action 0. Then, the algorithm goes through the arms. If the nth arm is in state s, then the algorithm samples a new action a with probability $y^*_{s,a}(t)/m^*_s(t)$ and assigns this action a to the nth arms if it does not violates the budget constraints. By a central limit argument, this construction guarantees that $\mathbf{Y}^{(N)} = \mathbf{y}^*(t) + O(1/\sqrt{N})$ but not that $\mathbf{Y}^{(N)} = \mathbf{y}^* + O(1/N)$. Hence, this randomized procedure would give a $O(1/\sqrt{N})$ convergence rate and not a $O(1/N)$ convergence rate, even in the non-degenerate case.*

6.5 Case Study: Generalized Applicant Screening Problem

The applicant screening problem is proposed in Brown and Smith [12], and has been subsequently studied in Chapter 5. This problem can be modeled as a two actions single constraint restless bandit problem, on which we have compared the performance of the LP-update policy with the LP-index policy in Section 5.6.2. We study in this section a generalization of the problem allowing multiple actions while also having multiple constraints, so that the model fits naturally into our weakly coupled MDPs framework.

6.5.1 Problem description

Consider a group of N applicants applying for a job. The decision maker's goal is to hire the best possible βN applicants. The applicant n has an unknown quality level $p_n \in [0, 1]$. At each decision epoch t, the decision maker chooses for each candidate either to interview this candidate with one or two questions, or chooses not to interview

this candidate, giving the action set $\mathcal{A} = \{0, 1, 2\}$. For each interviewed applicant (i.e. $a \in \{1, 2\}$), a signal $q_n(t) \sim \text{binomial}(a, p_n)$ is returned, indicating how many among the a questions have been solved correctly by the applicant. All variables $q_n(t)$ are supposed to be independent.

Choosing an action a on an applicant consumes $D(a)$ units of resource (time, space, organization cost, etc.), for which we choose to be

$$D(s, a) = D(a) = \begin{cases} 0, & \text{if } a = 0; \\ 1, & \text{if } a = 1; \\ 1.5, & \text{if } a = 2. \end{cases} \tag{6.17}$$

The value "1.5" on action $a = 2$ is interpreted as asking a single applicant consecutively two questions consumes less resource than asking separately two applicants each with one question. At each decision epoch a total amount of αN resource is available. There is a total number of T interviewing rounds, and in the final $(T+1)$-th round, the decision maker admits βN applicants, based on the results of the interviewing rounds.

We assume that the applicants belong to two different groups, each having a population of $N/2$. And we will consider two scenarios:

- No fairness: in this scenario, there is a single budget constraint of αN for the whole population.
- Fair selection: in addition to the above constraint, the decision maker cannot spend more that γN budget on each of the single group, where $\gamma < \alpha < 2\gamma$.

Notice that the above applicant screening problem generalizes the problem studied in Brown and Smith [12] and Chapter 5 in two ways: we allow for more than one question, whereas previously we consider $\mathcal{A} = \{0, 1\}$; and we add fairness constraints, whereas previously we only have a single resource constraint that the number of interviewed candidate should not be larger than αN.

6.5.2 Modeling as weakly coupled MDPs

To cast the problem into a weakly coupled MDP as described in Section 6.2, we consider a Bayesian model in which the quality level p of an applicant from each group is generated from some beta distribution, and the decision maker's estimation on each applicant's p is updated using Bayes' rule. The state s of an applicant is hence a 2-tuple (a, b), indicating the current estimation of her quality level and is distributed according to the beta distribution $beta(a, b)$.

For the first T interviewing rounds, the action set is $\mathcal{A} = \{0, 1, 2\}$. Upon taking action 0 on an applicant, the estimation is unchanged, so the matrix $\mathbf{P}^0$ is the identity matrix. Upon taking action 1, the state is updated according to

$$(a, b) \xrightarrow{\text{action 1}} \begin{cases} (a+1, b), & \text{with probability } a/(a+b); \\ (a, b+1), & \text{with probability } b/(a+b). \end{cases}$$

This gives the matrix $\mathbf{P}^1$. Likewise, for action 2 the state is updated as

$$(a,b) \xrightarrow{\text{action 2}} \begin{cases} (a+2,b), & \text{with probability } \frac{a(1+a)}{(a+b)(1+a+b)}; \\ (a+1,b+1), & \text{with probability } \frac{2ab}{(a+b)(1+a+b)}; \\ (a,b+2), & \text{with probability } \frac{b(1+b)}{(a+b)(1+a+b)}. \end{cases}$$

This gives the matrix $\mathbf{P}^2$. The function $D(s,a)$ is given by (6.17) and is independent of the state s. The rewards are all zero during the first T rounds.

For the final $(T+1)$-th admitting round, the action set is {admit, not admit}. The reward on an admitted applicant in state (a,b) is $a/(a+b)$; the reward on a non admitted applicant is zero.

6.5.3 The occupation measure policy as a benchmark

To provide a benchmark for evaluation of the performance of the LP-update policy, we introduce in this section the occupation measure policy, which is generalized from the randomized activation control policy in Zayas-Cabán et al. [52] and the occupancy-measured-reward index policy in Xiong et al. [50]. It is a one-pass policy that solves the linear program only once at the very beginning, and constructs the occupation measure vectors $\mu^*_{s,a}(t)$ defined in (6.18) from the solution. At each decision epoch t, the budget left is initialized as $\mathbf{B} := N \cdot \mathbf{B}$, the action on each arm n is initialized as the passive 0. It then samples a new action a_n from the distribution $(\mu^*_{s_n,a}(t))_{a\in A(s_n)}$ on each arm n. If choosing action a_n instead of action 0 on arm n does not violate any of the budget constraints, then we apply action a_n on arm n, and we decrease the budget left $\mathbf{B}$; otherwise we keep action 0 on arm n and continue to sample an action on the next arm. The detailed implementation is given in Algorithm 3.

Note that this occupation measure policy is proven to be asymptotically optimal for the multi-action single-constraint multi-armed bandits in Xiong et al. [50]. Here we generalize the policy to the multi-action multi-constraint case. Its asymptotical optimality can be proved by extending the same approach.

6.5.4 Discussion on simulation results

For our simulations, we choose $\beta = 0.1$, and we consider two scenarios. The first one is such that $\alpha = 0.15$, $\gamma = 0.1$. This is the scenario where the resource is "scarce". The second scenario where the resource is "abundant" is such that $\alpha = 0.3$ and $\gamma = 0.2$ (resource being doubled). In each scenario we shall compare the effect of adding or removing the fairness constraints. Without the fairness constraints, the decision maker can distribute the total αN units of resource freely among the two groups at each interviewing round.

We assume that the decision maker's prior estimations on the two groups of applicants are respectively $beta(1,1)$ and $beta(2,2)$, so that they have the same mean but the second group has a lower variance. Throughout the horizon is fixed to $T = 10$. To make the problem more realistic, and in the same time to reduce the total number of

Algorithm 3: Occupation measure policy for weakly coupled MDPs.

Input: Time horizon T and initial configuration vector $\mathbf{m}(0)$.

Solve the linear program (6.5) with time horizon T and initial configuration vector $\mathbf{m}(0)$, obtain an optimal solution $\mathbf{y}^*$ and the corresponding $\mathbf{m}^*$;

for $t = 0, 1, \ldots, T-1$ **do**

Compute from the LP solution the occupation measure

$$\mu^*_{s,a}(t) := \begin{cases} \frac{y^*_{s,a}(t)}{m^*_s(t)}, & \text{if } m^*_s(t) > 0 \\ \mathbf{1}_{\{a=0\}}, & \text{otherwise.} \end{cases} \tag{6.18}$$

Observe the current states of the N sub-MDP's $\mathbf{s} = (s_1, s_2, \ldots, s_N)$.

Initialize $\mathbf{B} := N \cdot \mathbf{B}$, and actions on the N sub-MDP's as $A(\mathbf{s}) = (0, 0, \ldots, 0)$;

for $n = 1, 2, \ldots, N$ **do**

Sample an action a_n according to the probability vector $(\mu^*_{s_n,a}(t))_{a \in A(s_n)}$;

if $\mathbf{B} - D(s_n, a_n) \geq \mathbf{0}$ **then**

$A(s_n) := a_n$;

$\mathbf{B} := \mathbf{B} - D(s_n, a_n)$;

end

end

Apply the actions $A(\mathbf{s})$ to the N sub-MDP's ;

end

possible states of the MDP, we require in addition that no more than 10 questions can be asked on a single applicant during the 10 interviewing rounds.

The simulations are done for N ranging from 20 to $10240 = 20 \times 2^{10}$. For each value of N, we generate 1600 instances of N applicants according to the beta distributions we described previously in each scenario,. We apply the LP-update policy and occupation measure policy on each instance, with or without the fairness constraints. In each simulation the performance is evaluated based on the average quality level of the final admitted βN applicants. The results are reported in Figure 6.1.

Note that as guaranteed by Theorem 6.3.1 and Theorem 1 of Xiong et al. [50], both LP-update policy and occupation measure policy converge to the LP-relaxed bounds, as N goes to infinity. This is what we observe in Figure 6.1. The situation is however different for smaller values of N: In all cases, the LP-update policy outperforms the occupation measure policy. The smaller the value of N, the more apparent is the advantage of the LP-update policy.

We shall now discuss the effect of adding fairness constraints. We observe that in the first scarce resource scenario, the LP-relaxed performance bound becomes smaller when fairness is imposed. This should be expected since adding more constraints on a linear program can only decrease its optimal value. However, in the second scenario, when the resource is doubled, adding or removing the fairness constraints

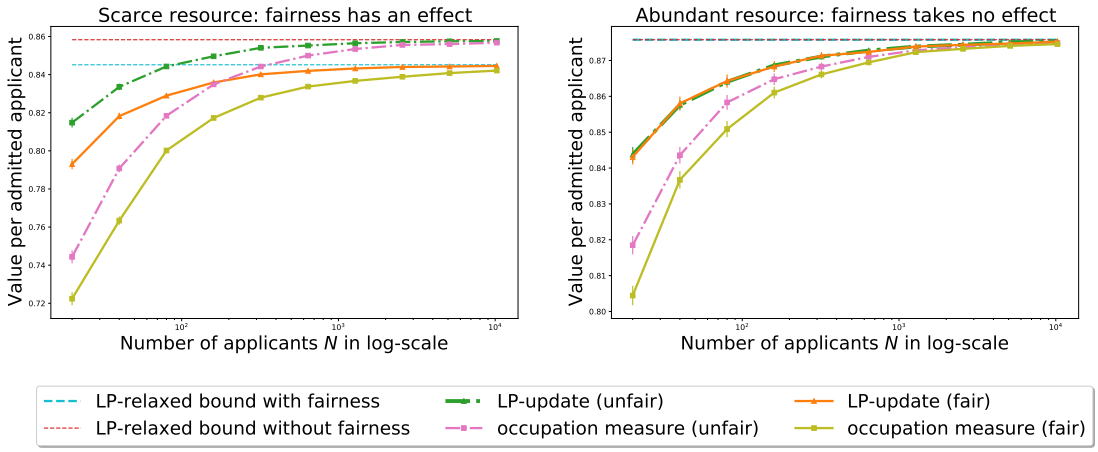


Figure 6.1 – Performance on generalized applicant screening problems when the resource is scarce (left panel) or abundant (right panel), with or without fairness constraints. Overall LP-update policy outperforms occupation measure policy in all situations, and the advantage is more apparent for small values of N. Fairness has a negative impact in the scarce resource scenario, whereas in the abundant case, it has no effect asymptotically, but can still influence the performance of the occupation measure policy.

result in the same upper-bound value, and the performances of LP-update policy are identical in the two situations. Nevertheless, these fairness constraints still play a role in both policies: for the LP-update policy we observe that the fairness constraints can be saturated when solving new linear programs by applying updates, but it turns out that this does not influence its performance at all. For the occupation measure policy, however, sampling an action is more restrictive if there are more constraints, and as a result this influences its performance, even though asymptotically they converge to the same limit.

6.5.5 Computation time analysis

Although the LP-update policy outperforms the occupation measure policy under all situations, it comes with a price of using more computations. This extra computation time is due to the fact that the LP-update policy periodically solves a new LP problem. We record the execution time for different values of N in Table 6.1, averaging over 100 runs on one problem instance under each circumstance, together with the 95% confidence interval. Here by execution time we mean the time needed to apply a policy on an instance of N applicants problem until the final phase of admission round. The program is written in Python using NumPy for data structure and PuLP for solving LPs. Note that our code is not particularly optimized and is tested on an ordinary personal laptop.

We notice that the variance of the computation time is much larger for the LP-update policy, since the number of time steps when it solves a new LP for an update varies on

		Number of applicants	$N = 20$	$N = 100$	$N = 1000$
With	fairness	Occupation measure	2.70 ± 0.01	2.72 ± 0.01	2.83 ± 0.01
		LP-update	12.69 ± 0.39	10.44 ± 0.41	7.06 ± 0.19
		(Number of updates)	(≈ 6.4)	(≈ 5.2)	(≈ 3.9)
Without	fairness	Occupation measure	2.62 ± 0.01	2.64 ± 0.01	2.75 ± 0.01
		LP-update	8.95 ± 0.31	7.93 ± 0.23	6.75 ± 0.16
		(Number of updates)	(≈ 4.5)	(≈ 3.6)	(≈ 2.8)

Table 6.1 – Computation time (in seconds) of the LP-update policy and the occupation measure policy for different values of N, as well as the number of times that the LP-update policy solves a new linear programm (the first initial one not included).

each run. Overall this variance is decreasing as N becomes large. We also remark that the LP-update policy takes less time for $N = 1000$ compared to $N = 20$. This is because for larger N the stochastic trajectory is closer to the deterministic one, consequently it is less likely to perform updates caused by a non-feasible action obtained from (6.15). Indeed, we observe from Table 6.1 that the number of updates is decreasing with N.

6.6 Proof of the Additional Results

This section is dedicated to the proofs of some technical lemmas that are used to prove Theorem 6.3.1, Theorem 6.4.4 and Theorem 6.4.6. In addition, we provide counter-examples for the lower bounds of the three theorems in Section 6.6.3. The proof for the equivalence of the two notions of non-degeneracy on restless bandits is given in Section 6.6.4.

6.6.1 One-step transition and concentration arguments

Recall that the linear function ϕ maps a decision vector $\mathbf{y}$ to a configuration vector $\phi(\mathbf{y}) = \big(\phi_1(\mathbf{y}), \dots, \phi_d(\mathbf{y})\big) \in \Delta^d$ whose sth component is

$$\phi_s(\mathbf{y}) = \sum_{s',a} y_{s',a} P^a_{s',s}. \tag{6.19}$$

This is the deterministic behavior of the Markov transition at time step t. We claim that:

Lemma 6.6.1. *Define the random vector* $\mathbf{E}^{(N)}(t) := \mathbf{M}^{(N)}(t+1) - \phi(\mathbf{Y}^{(N)}(t))$. *We have*

$$\mathbb{E}\left[\mathbf{E}^{(N)}(t) \mid \mathbf{Y}^{(N)}(t)\right] = \mathbf{0}, \tag{6.20}$$

$$\mathbb{E}\left[\left\|\mathbf{E}^{(N)}(t)\right\|_1 \mid \mathbf{Y}^{(N)}(t)\right] \le \frac{\sqrt{d}}{\sqrt{N}}, \tag{6.21}$$

$$\mathbb{P}\left[\left\|\mathbf{E}^{(N)}(t)\right\|_1 \ge \epsilon \mid \mathbf{Y}^{(N)}(t)\right] \le 2d \cdot e^{-2N\epsilon^2/d^2}. \tag{6.22}$$

Proof. For simplicity of notation, let us denote by $\mathbf{y} := \mathbf{Y}^{(N)}(t)$. There are $N y_{s,a}$ 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:

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

where $U_{s,a,i}$ are i.i.d uniform random variables in $[0,1]$. Taking expectation then gives $\mathbb{E}\left[M^{(N)}_{s'}(t+1) \mid \mathbf{Y}^{(N)}(t)\right] = \phi_{s'}(\mathbf{Y}^{(N)}(t))$, which gives (6.20). It also implies that

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

This shows that

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

where the first inequality comes from Cauchy-Schwartz, and this gives (6.21).

Equation (6.22) is a direct consequence of Hoeffding's inequality. Indeed, one has

$$\mathbb{P}\left[|E^{(N)}_s(t)| \geq \varepsilon/d \mid \mathbf{Y}^{(N)}(t)\right] \leq 2e^{-N\varepsilon^2/d^2}.$$

By using the union bound, this implies that

$$\mathbb{P}\left[\left\|\mathbf{E}^{(N)}(t)\right\|_1 | \geq \varepsilon \mid \mathbf{Y}^{(N)}(t)\right] \leq d \cdot \mathbb{P}\left[|E^{(N)}_s(t)| \geq \varepsilon/d \mid \mathbf{Y}^{(N)}(t)\right] \leq 2d \cdot e^{-N\varepsilon^2/d^2}.$$

□

6.6.2 Non-degenerate problem and concentration on a trajectory

The previous lemma can be extended to show that the stochastic system $\mathbf{M}^{(N)}(t)$ concentrates on a neighbourhood of $\mathbf{m}^*(t)$.

Lemma 6.6.2. *Assume that the problem is non-degenerate, let $\mathbf{y}^*$ be the optimal solution on the LP computed at time 0 and $\mathbf{M}^{(N)}(t)$ be the sequence of values obtained when applying Algorithm 2. Then for all $\varepsilon > 0$, there exists $C_1, C_2 > 0$ such that for all N:*

$$\mathbb{P}\left[\mathbf{M}^{(N)}(t) \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon)\right] \geq 1 - C_1 \cdot e^{-C_2 N}. \tag{6.23}$$

Moreover, if the LP has a unique solution starting from any initial point, then (6.23) also holds if $\mathbf{M}^{(N)}(t)$ is the output of Algorithm 1.

Proof. We first consider what happens when Algorithm 2 is used. We proceed by induction on t. This is clearly true for $t = 0$. Assume that this holds for some $t \geq 0$. By Proposition 6.4.3, there exists ε_t such that the control $\mathbf{y}(\mathbf{m})$ defined in Proposition 6.4.3 is optimal for all $\mathbf{m} \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon_t)$. The induction hypothesis and the continuity of $\mathbf{y}(\mathbf{m})$ therefore imply that for all $\epsilon > 0$, there exists $C_1, C_2 > 0$ such that $\|\mathbf{Y}^{(N)}(t) - \mathbf{y}^*(t)\| \leq \epsilon$ with probability at least $1 - C_1 e^{-C_2 N}$. We can write:

$$\|\mathbf{M}^{(N)}(t+1) - \mathbf{m}^*(t+1)\| \leq \|\mathbf{M}^{(N)}(t+1) - \phi(\mathbf{Y}^{(N)}(t))\| + \|\phi(\mathbf{Y}^{(N)}(t)) - \mathbf{m}^*(t+1)\|.$$

Hence, by using Lemma 6.6.1 and the union bound, for all $\epsilon > 0$, there exists $C_1', C_2' > 0$ such that $\|\mathbf{M}^{(N)}(t+1) - \mathbf{m}^*(t+1)\| \geq \epsilon$ with probability at least $1 - C_1' e^{C_2' N}$. To show (6.23), the only remaining point is to show that if s' is a state such that $m^*_{s'}(t+1) = 0$, then so is $M^{(N)}_{s'}(t+1)$. By definition of the deterministic evolution $\mathbf{m}^*(t+1) = \phi(\mathbf{y}^*(t))$ (see *e.g.*, Equation (6.19)), we have:

$$m^*_{s'}(t+1) = \sum_{a,s} y^*_{s,a}(t) P^a_{s,s'}.$$

Hence, $m^*_{s'}(t+1)$ equals 0 if for all s, a, either $y^*_{s,a}(t) = 0$ or $P^a_{s,s'} = 0$. By construction of $\mathbf{Y}^{(N)}(t)$ from $\mathbf{y}^*(t)$ $y^*_{s,a}(t) = 0$ implies that $Y^{(N)}_{s,a}(t) = 0$. This implies that $M^{(N)}_{s'}(t+1) = 0$.

To study what happens when Algorithm 1 is used, we remark that if the solution of the LP is unique, and $\mathbf{M}^{(N)}(t)$ is close enough to $\mathbf{m}^*(t)$, then the new LP solution computed by Algorithm 1 is the same $\mathbf{y}(\mathbf{M}^{(N)}(t)$ used by Algorithm 2. Hence, the proof for Algorithm 2 also applies in this case.

□

6.6.3 Proof of the lower bounds

Consider th following 2-action restless bandit with two states $\mathcal{S} = \{1, 2\}$, with parameters $\mathbf{m}(0) = (0.5, 0.5)$, $T = 2$, $A = \{0, 1\}$, $\mathbf{R}^0 = [0, 0]$, $\mathbf{R}^1 = [1, 0]$, and where the transition matrices are

$$\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \mathbf{P}^1 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}.$$

We consider that $D(s, 0) = 0$, $D(s, 1) = 1$ for any state s and we distinguish the resource constraints $b = 0.3$ and $b = 0.5$.

For any resource constraint, the solution of the LP is to choose action 1 for as many arms as possible. This gives a reward $2b$. If $b = 0.3$, the problem is non-degenerate whereas if $b = 0.5$ the problem is degenerate.

For the stochastic system with N components, this gives a reward $\frac{1}{N} \lfloor Nb \rfloor$ at time-step 0 and a reward $\min(b, \lfloor \mathbf{M}^{(N)}(1) \rfloor)$ at time 1. Since $\mathbf{M}^{(N)}(1)$ follows a binomial distribution of parameter $(N, 0.5)$, the total reward of the LP-update policy is equal to

$$\frac{1}{N} \lfloor Nb \rfloor + b + \mathbb{E}\left[\min(\left\lfloor M^{(N)}_1(1) \right\rfloor - b, 0)\right].$$

By the central limit theorem, $\lim_{N\to\infty} \sqrt{N}\mathbb{E}\left[\min(M_1^{(N)}(1) - 0.5, 0)\right] = \sqrt{2}\pi > 0$. This provides a counter-example for the lower bound of Theorem 6.3.1.

If $b = 0.3$ and N is not a multiple of 10, the problem does not admit a perfect rounding and the $2b - (\frac{1}{N}\lfloor Nb\rfloor + b + \mathbb{E}\left[\min(\left|M_1^{(N)}(1)\right| - b, 0)\right]) \geq b - \frac{1}{N}\lfloor Nb\rfloor \geq 0.1/N$. This provides a counter-example for the lower-bound of Theorem 6.4.4.

If $b = 0.3$ and N is a multiple of 10, then the problem admits a perfect rounding. In this case, classical anti-concentration arguments Matoušek and Vondrák [33] show that

$$\mathbb{P}\left[M_1^{(N)}(1) \leq 0.2\right] \geq \frac{1}{15}\exp(-16N(0.6-0.2)^2).$$

This shows that $\mathbb{E}\left[\min(\left|M_1^{(N)}(1)\right| - 0.3, 0)\right] \geq -0.1\mathbb{P}\left[M_1^{(N)}(1) \leq 0.2\right] = -\frac{1}{150}\exp(-2.56N)$. This is provides a counter-example for the lower bound of Theorem 6.4.6. □

6.6.4 Proof of the equivalence of the two notions of non-degeneracy on restless bandits

We prove that the notion of non-degeneracy as defined in Definition 6.4.1 coincides with the one given in Zhang and Frazier [53] and Gast et al. [18] for two-action bandits with a single *equality* resource constraint $b = \alpha$. To define their notion of non-degeneracy, for a given optimal solution of the LP $\mathbf{y}^*$, the authors of [18, 53] partition the state space $\mathcal{S}$ for each time t into four sub-sets:

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

In [18, 53], a problem is called non-degenerate if and only if there exists a solution $\mathbf{y}^*$ for which $\left|\mathcal{S}^0(t)\right| \geq 1$ for all t. In our definition, a problem is called non-degenerate if there exists a solution such that the corresponding matrix $C^*(t)$ satisfies some rank condition for all t. For the rest of the section, we fix a given time t and a solution $\mathbf{y}^*$ and show that $\left|\mathcal{S}^0\right| \geq 1$ if and only if C^* satisfies the rank condition.

Following our definition, let $\mathcal{U}^*$ be the set of indices (s,a) for which $y^*_{s,a} = 0$, and $\mathcal{S}^*$ be the set of states for which $m^*_s > 0$. For the two-action single-constraint case, there is a unique constraint that is satisfied with equality. Equations (6.12), (6.13) and (6.14) then becomes:

$$y_{s,a} = 0 \qquad \forall (s,a) \in \mathcal{U}^*; \tag{6.24}$$

$$\sum_s y_{s,1} = \alpha; \tag{6.25}$$

$$y_{s,0} + y_{s,1} = m_s \qquad \forall s \in \mathcal{S}^*. \tag{6.26}$$

We write them compactly in matrix form as $C^*\mathbf{y} = \left[\mathbf{0};\ \mathbf{B}|_{\mathcal{J}^*};\ \mathbf{m}(t)|_{\mathcal{S}^*}\right]^\top$.

According to our definition, the problem is non-degenerate if C^* is of rank $|\mathcal{J}^*| + |\mathcal{S}^*| + |\mathcal{U}^*|$. For the two-action single-constraint case, the partition of the set $\mathcal{S}$ shows that:

- $|\mathcal{J}^*| = 1$ (there is a single equality constraint)
- $|\mathcal{S}^*| = d - |\mathcal{S}^\emptyset|$ (each $s \notin \mathcal{S}^\emptyset$ contributes to one element of $\mathcal{S}^*$).
- $|\mathcal{U}^*| = |\mathcal{S}^+| + |\mathcal{S}^-| + 2|\mathcal{S}^\emptyset| = d + |\mathcal{S}^\emptyset| - |\mathcal{S}^0|$ (each $s \in \mathcal{S}^+ \cup \mathcal{S}^-$ contributes to one element of $\mathcal{U}^*$ and each $s \in \mathcal{S}^\emptyset$ to two).

The matrix has therefore $2d + 1 - |\mathcal{S}^0|$ rows and $|\mathcal{U}| = 2d$ columns.

Since Equations (6.24) to (6.26) are linearly independent, the matrix is therefore of rank $\min(2d + 1 - |\mathcal{S}^0|, 2d)$. This quantities equals $2d + 1 - |\mathcal{S}^0|$ if and only if $|\mathcal{S}^0| \geq 1$. This shows that, for the two-action single-constraint case studied in [18, 53], the two notions of non-degeneracy are equivalent. □

Conclusion of the Chapter

In this chapter we have constructed the LP-update policy on weakly coupled MDPs, that generalizes the restless bandit model we studied in previous chapters. A first version of the policy, described in Algorithm 1, is a direct generalization of the LP-update policy we considered in the previous chapter, and solves a new LP at each decision epoch. By investigating deeper the structure of the LPs, we realize that the optimal solution of the new LP can be easily obtained from the old ones, if a certain rank condition holds, rendering the policy more efficient, as described in Algorithm 2. This rank condition on the model is also the requirement for a faster convergence rate of the policy, and consequently is the generalization of non-degeneracy on restless bandits defined in the previous chapter.

The generalized applicant screening problem is appealing, and may be investigated further, by using different priors on the quality of the applicants, or adding more fairness constraints (e.g. race, gender, age). The weakly coupled MDPs being much broader than the restless bandits, we hope that the policies proposed in this chapter can see more interesting applications in the future.

Part III

Additional Results and Conclusion

7

Additional Numerical Experiments

This chapter is a collection of numerical experiments that extends and complements our discussions in previous chapters. These are mostly claims and observations that we were not able to formulate as mathematical theorems and prove in a rigorous way, but nevertheless can still be investigated via a numerical approach, giving insightful and practically meaningful results. We hope these results could be inspiring, and some of them be formulated as theorems and proven in the future. A summary of each subsection, as well as to which chapters it is related are given at the beginning of the sections.

> If only I had the Theorems! Then I should find the proofs easily enough.
>
> Bernhard Riemann

7.1 Extended Results from Chapter 3

This section consists of extended discussions from Chapter 3, centered around the exponential convergence rate of WIP. We estimate the optimal constant $\tilde{c}$ of Theorem 3.3.2 in Section 7.1.1. We discuss how WIP behaves when its deterministic dynamic exhibits an attracting period-2 limit cycle in Section 7.1.2. More complicated dynamics are investigated in Section 7.1.3 on a particularly constructed example. We discuss the convergence rate of WIP on singular problems in Section 7.1.4, together with a summarization of the rates under all possible scenarios of the dynamic.

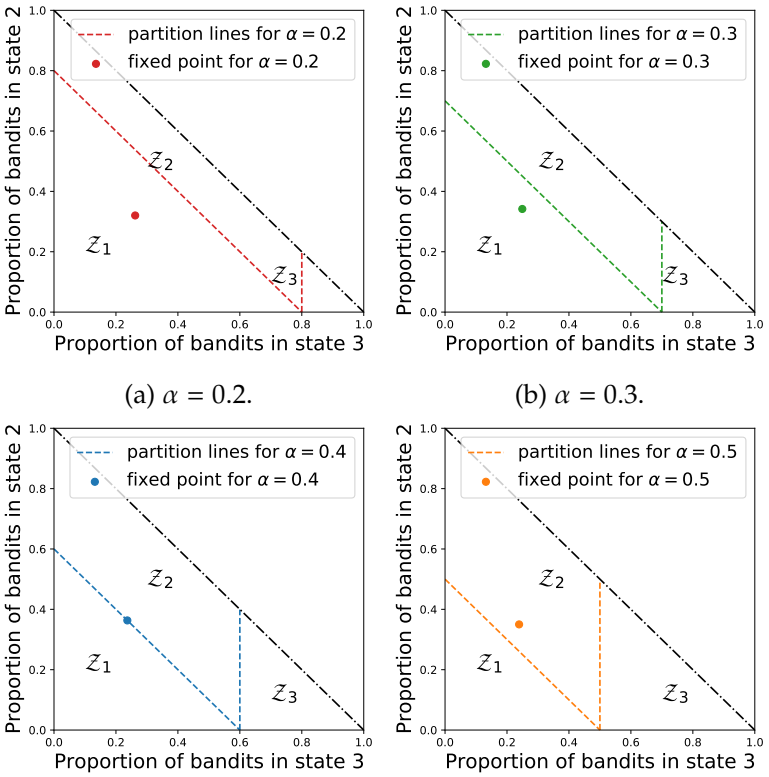
7.1.1 Finding the optimal constant in Theorem 3.3.2

Recall that Theorem 3.3.2 claims the existence of constants b and c for which $V_{\text{rel}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha) \leq b \cdot e^{-cN}$ holds true, but we do not emphasize on the optimality of the constant

c, in the sense of finding constant $\tilde{c}$ such that

$$\limsup_{N\to\infty} -\frac{1}{N}\log\left(V_{\mathrm{rel}}^{(N)}(\alpha) - V_{\mathrm{WIP}}^{(N)}(\alpha)\right) = \tilde{c}.$$

Indeed, our choice of c in the proof of Theorem 3.3.2 provided in Section 3.6.2 actually depends subtly on the given parameters, and we believe that finding $\tilde{c}$ is, if not impossible, a much more demanding task.



(a) $\alpha = 0.2$. (b) $\alpha = 0.3$.

(c) $\alpha = 0.4$. (d) $\alpha = 0.5$.

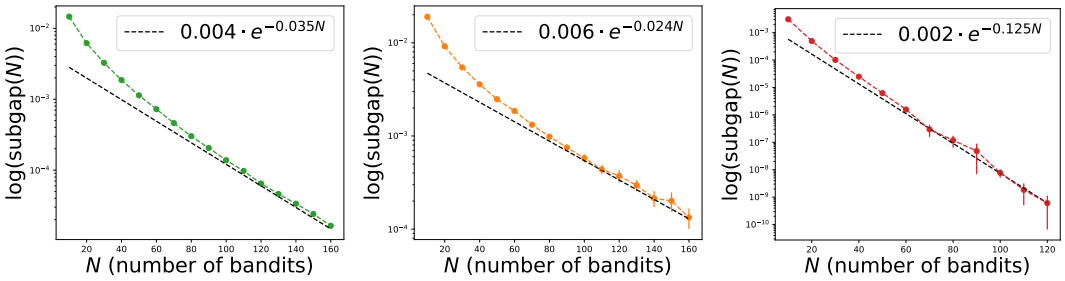
Figure 7.1 – The same model as in Figure 3.1, with $\alpha = 0.2, 0.3, 0.4, 0.5$.

Nevertheless, it is still possible to estimate numerically the value of this optimal constant $\tilde{c}$, and see how its value changes with respect to the relative position of the unique fixed point $\mathbf{m}^*(\alpha)$. Note that we write $\mathbf{m}^*(\alpha)$ to emphasize the position of the fixed point depends on α.

To this end, let us consider the quantity

$$\mathrm{subgap}(N) := V_{\mathrm{rel}}^{(1)}(\alpha) - V_{\mathrm{WIP}}^{(N)}(\alpha). \tag{7.1}$$

Theorem 3.3.2 implies that $\mathrm{subgap}(N)$ converges to 0 approximately as $b \cdot e^{-c\cdot N}$, for some constants $b, c > 0$ in non-singular cases. In Figure 7.2, we plot in log-scale the subgap (7.1) as a function of N for the same model as in Figure 3.2 and $\alpha = 0.2$, 0.3 and 0.5. For each value of α, we also plot the best-fit $\tilde{b} \cdot e^{-\tilde{c}N}$, which is a straight line in

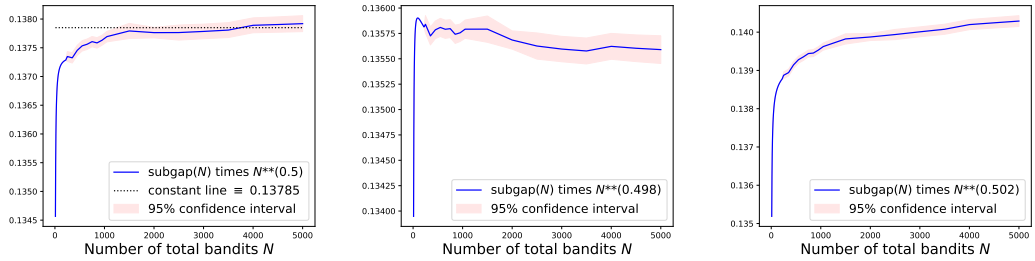


(a) $\tilde{c} \approx 0.032$ for $\alpha = 0.3$. (b) $\tilde{c} \approx 0.024$ for $\alpha = 0.5$. (c) $\tilde{c} \approx 0.125$ for $\alpha = 0.2$.

Figure 7.2 – Estimation of the optimal constant $\tilde{c}$ from Theorem 3.3.2.

log-scale. The positions of $\mathbf{m}^*(\alpha)$ are shown in Figure 7.1 for the four values of α. We see that the constant $\tilde{c}$ is around 0.03 for $\alpha = 0.3, 0.5$, and it is around 0.125 for $\alpha = 0.2$.

However, in the singular case $\alpha = 0.4$, we could not find a straight line to fit $\log\big(\text{subgap}(N)\big)$. But if we plot instead $\text{subgap}(N) \cdot \sqrt{N}$, the curve behaves like a constant. Moreover, this constant behavior is lost as soon as we plot $\text{subgap}(N) \cdot N^{\beta}$, with a power $\beta = 0.49$ or $\beta = 0.51$, as illustrated in Figure 7.3. This gives numerical evidence for an $O(\frac{1}{\sqrt{N}})$ convergence rate in this singular case, same as for the example given in Remark 3.3.3.



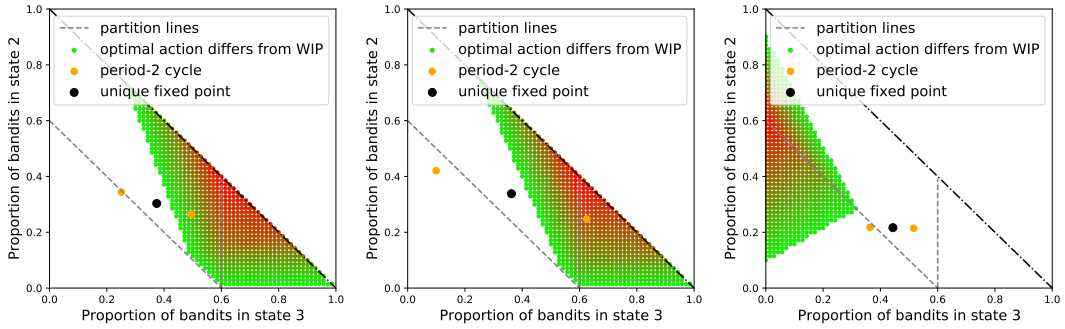
(a) $\beta = 0.5$. (b) $\beta = 0.49$. (c) $\beta = 0.51$.

Figure 7.3 – Verifying the square root convergence in the singular case $\alpha = 0.4$, by plotting $\text{subgap}(N) \cdot N^{\beta}$ with $\beta = 0.5, 0.49, 0.51$

7.1.2 WIP with an attracting period-2 cycle

In this section, we illustrate the behavior of restless bandits under WIP in three examples with $d = 3$ and $\alpha = 0.4$. For all of them, the dynamical system $\Phi_{t\geq 0}(\cdot)$ has an attracting cycle of period 2. The fixed point and the two points of the attracting cycle for each example are shown in Figure 7.4. Note that for these three examples, the matrices $\mathbf{K}_2$'s are not stable and they all have an eigenvalue smaller than -1.

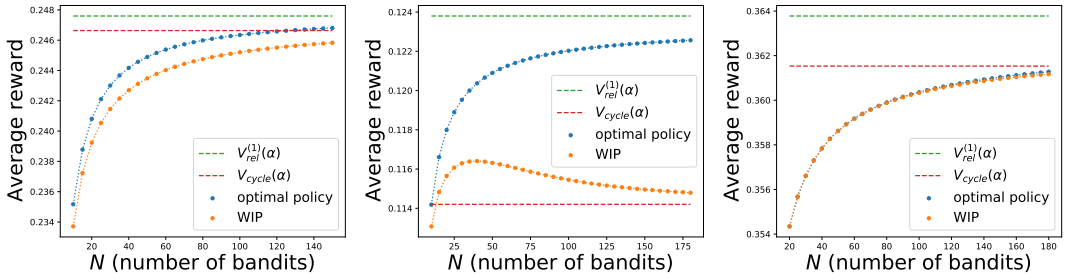
Since we are in a small dimension $d = 3$, the optimal policy in (3.2)-(3.3) can be computed directly by using a brute-de-force dynamic programming approach, provided that the arm population N is not too large. To this end, in Figure 7.4 we



(a) Example 1 (b) Example 2 (c) Example 3

Figure 7.4 – Action differences plot for three period-2 cycle examples with $d = 3$ and $N = 70$.

also highlight the configuration vectors in which the optimal policy takes a different action than WIP when the number of arms is $N = 70$. Such configuration vectors are represented as colored dots inside the triangle: starting with the greenest color, the deeper the red, the more the optimal action deviates from WIP's action on this configuration vector. The blank area means that on these configuration vectors WIP is an optimal action.



(a) Performance for Example 1 (b) Performance for Example 2 (c) Performance for Example 3

Figure 7.5 – Performance of optimal policy and WIP for three period-2 cycle examples with $d = 3$.

We then plot in Figure 7.5 the value of the optimal decision rule, $V^{(N)}_{\text{opt}}(\alpha)$, and of WIP, $V^{(N)}_{\text{WIP}}(\alpha)$, as a function of the number of arms N. We take multiples of 5 for values of N so that αN are always integers. Several comments are in order:

- As mentioned in Remark 3.3.5, $V^{(N)}_{\text{WIP}}(\alpha)$ converges to the averaged reward on the cycle, denoted here by $V_{\text{cycle}}(\alpha)$, instead of reward on the fixed point $V^{(1)}_{\text{rel}}(\alpha)$. Note that $V_{\text{cycle}}(\alpha)$ is not an upper bound on $V^{(N)}_{\text{WIP}}(\alpha)$ and sometimes, as in Example 2, $V^{(N)}_{\text{WIP}}(\alpha)$ becomes greater than $V_{\text{cycle}}(\alpha)$ for $N \approx 30$ before decreasing to this value from above.
- The quantity $V^{(N)}_{\text{opt}}(\alpha) - V^{(N)}_{\text{WIP}}(\alpha)$ converges to $V^{(1)}_{\text{rel}}(\alpha) - V_{\text{cycle}}(\alpha)$, which is strictly positive and might be relatively large, depending on the parameters.

- The gap $V^{(N)}_{\text{opt}}(\alpha) - V^{(N)}_{\text{WIP}}(\alpha)$ can be increasing with N, as in Example 2 and 3. This violates the intuition that WIP should be closer to the optimal policy as N grows. It should be contrasted with the locally stable global attractor situation, for which $V^{(N)}_{\text{opt}}(\alpha) - V^{(N)}_{\text{WIP}}(\alpha) \to 0$ exponentially fast.

Instead of a period-2 cycle, it is also possible to have more complicated shape of attracting limit cycles (of variant periods), as long as $d = 4$. This will be the case if the matrix $\mathbf{K}_{s(\mathbf{m}^*)}$ in $\phi(\mathbf{m}) = (\mathbf{m}-\mathbf{m}^*)\cdot\mathbf{K}_{s(\mathbf{m}^*)}+\mathbf{m}^*$ has a pair of conjugate complex eigenvalues (or two real eigenvalues) with norm bigger than 1. We shall next investigate such an example.

7.1.3 A peculiar example

In this section we study the following restless bandit model in dimension $d = 4$, with the parameters given by

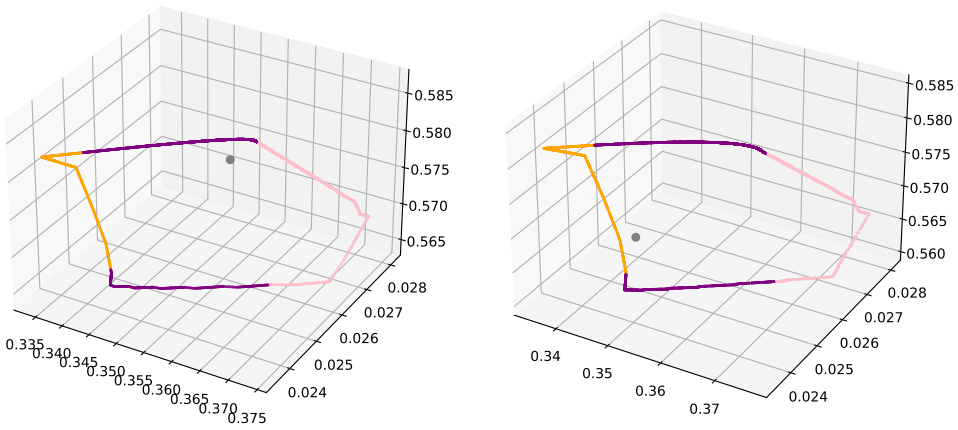
$$\mathbf{P}^0 = \begin{pmatrix} 0.23283388 & 0.28604935 & 0.15821436 & 0.32290241 \\ 0.8 & 0.1 & 0.01 & 0.09 \\ 0.01087021 & 0.01127903 & 0.96848126 & 0.00936949 \\ 0.42205252 & 0.05893614 & 0.00151789 & 0.51749346 \end{pmatrix},$$

$$\mathbf{P}^1 = \begin{pmatrix} 0.9 & 0.04 & 0.05 & 0.01 \\ 0.02 & 0.01 & 0.02 & 0.95 \\ 0.12808651 & 0.206595 & 0.17162894 & 0.49368955 \\ 0.46809243 & 0.0439124 & 0.0165773 & 0.47141787 \end{pmatrix},$$

$\mathbf{R}^1 = (0.5, 0.01, -500, -50)$ and $\mathbf{R}^0 = \mathbf{0}$. Unless otherwise specified, we shall record the fractional numbers with 8 digits of precision. The problem is indexable and the Whittle indices for the four states are $\nu_1 = 0.5$, $\nu_2 = -2.10188119$, $\nu_3 = -48.82476415$ and $\nu_4 = -56.03676124$.

The piecewise affine continuous map ϕ induced by WIP defined in Equation (3.6) has four linear pieces. Let us denote the four zones as $\mathcal{Z}_1$, $\mathcal{Z}_2$, $\mathcal{Z}_3$ and $\mathcal{Z}_4$, which is the partition of Δ^4, so that ϕ is affine in each of these zones. The interesting feature of this example is that the eigenvalues of the linear matrix factor in $\mathcal{Z}_2$ is $(0.01929854, 1.21347395, 1.03320223, 1)$, so that it has two real eigenvalues larger than 1, while the other three linear factors are stable matrices. Moreover, for α in the range $\alpha_1 < \alpha < \alpha_2$, with $\alpha_1 := 0.36428723$ and $\alpha_2 := 0.36886104$, the unique fixed point $\mathbf{m}^*(\alpha)$ of the deterministic dynamic of WIP lies strictly inside $\mathcal{Z}_2$.

In Figure 7.6 we illustrate the globally attracting limit cycles for the deterministic dynamics of WIP, with two values of α in the range $[\alpha_1, \alpha_2]$. The coordinates are the first three of the four coordinates of a point in Δ^4. These limit cycles are detected by iterating the map ϕ on randomly chosen initial points in Δ^4 after a first 1000 mixing time steps, and then we plot the next 10000 points. We repeat this process for a large enough collection of initial points, to ensure numerically that these limit cycles are indeed global attractors.



(a) $\alpha = 0.365$. (b) $\alpha = 0.368$.

Figure 7.6 – The globally attracting limit cycles for two different values of α. The three colors orange, purple and pink correspond to points respectively in $\mathcal{Z}_3$, $\mathcal{Z}_2$ and $\mathcal{Z}_1$. The grey point is the unique fixed point. In both figures it is in $\mathcal{Z}_2$, so is unstable. The first three of the four coordinates of a point in Δ^4 are presented in the 3 dimensional space of the figure.

In Figure 7.7, we do a similar experiment, this time with $\alpha = 0.3835 > \alpha_2$. Under this value of α, the unique fixed point $\mathbf{m}^*(\alpha)$ is in $\mathcal{Z}_3$ and is locally stable, since the linear factor is a stable matrix. However, it is not a global attractor. As numerically, we observe that a proportion of initial conditions are attracted to the (discrete) cycle shown in Figure 7.7. In fact, it seems that this situation occurs for any α larger than α_2 and smaller than a certain number around 0.3835.

In Figure 7.8, we simulate random trajectories of WIP with a population $N = 10^9$ of arms for this example with $\alpha = 0.3835$. There are two possible outcomes of the simulation: in the left panel the trajectory is attracted and confined to the locally stable fixed point; in the right panel the trajectory is attracted to the limit cycle shown previously in Figure 7.7.

It is then a natural question as what happens for α near the other extreme α_1. To this end, we apply a dichotomic search to compute more precisely the value of α_1, and find that it is between $\underline{\alpha_1} = 0.364287235212$ and $\overline{\alpha_1} = 0.364287235213$. Surprisingly, the unique fixed point $\mathbf{m}^*(\underline{\alpha_1})$ is a locally stable global attractor, while $\mathbf{m}^*(\overline{\alpha_1})$ is unstable and the global attractor turns into a limit cycle, with a similar shape and size to the ones shown in Figure 7.6. This phenomenon is illustrated more vividly in Figure 7.9. Since the exact value of α_1 is out of reach, it remains an interesting open problem as at the critical moment of $\alpha = \alpha_1$, whether the left panel or the right panel of Figure 7.9 will take place.

We proceed to evaluate the performance of WIP on this problem with α taking the critical value α_1 (using $\underline{\alpha_1}$ or $\overline{\alpha_1}$ makes no difference for this purpose). The simulations are done with horizon $T = 10^7$ and the arm population N ranges from 10^3 to $2 * 10^9$.

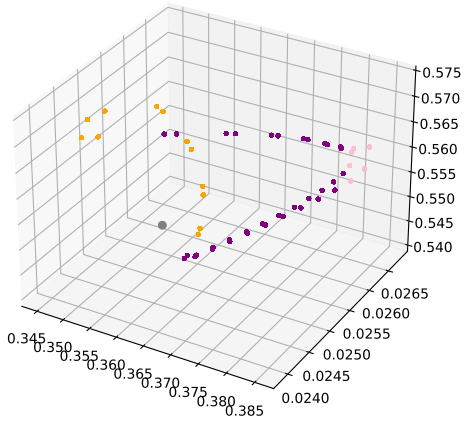
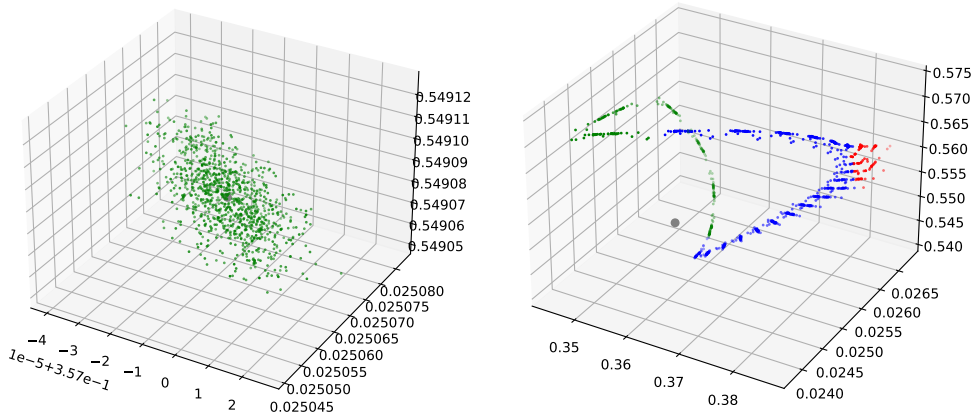
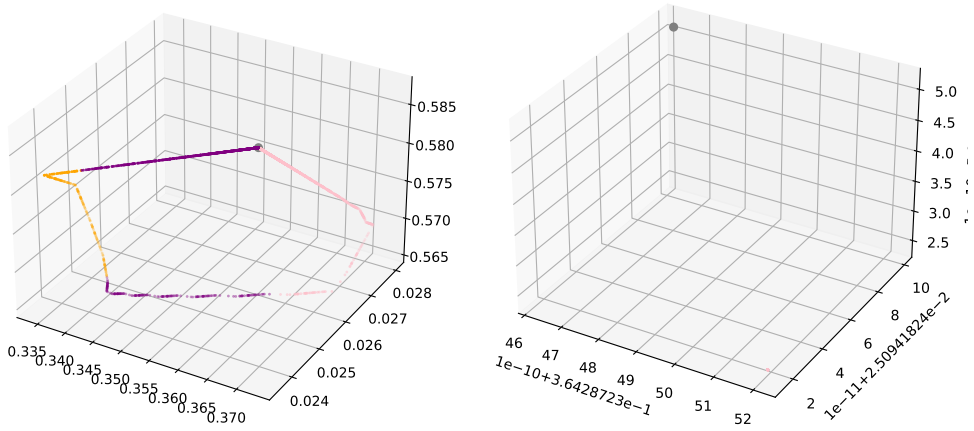


Figure 7.7 – The (not globally) attracting limit cycle for $\alpha = 0.3835 > \alpha_2 = 0.36886104$. The unique fixed point $\mathbf{m}^*(\alpha)$ is in $\mathcal{Z}_3$ and is locally stable, but is NOT a global attractor.



(a) (b)

Figure 7.8 – Two simulated trajectory of WIP with $\alpha = 0.3835$. We fix one and the same initial configuration vector with an arm population $N = 10^9$. The mixing time steps is 1000 and we plot the next 1000 points in the figure. The three colors green, blue and red correspond to points respectively in $\mathcal{Z}_3$, $\mathcal{Z}_2$ and $\mathcal{Z}_1$. There are two possible outcomes of the simulation: in the left panel the trajectory is attracted and confined to the locally stable fixed point; in the right panel the trajectory is attracted to the limit cycle shown previously in Figure 7.7.



(a) $\overline{\alpha_1} = 0.364287235213$. (b) $\underline{\alpha_1} = 0.364287235212$.

Figure 7.9 – For α near the critical value α_1, there is a "sudden jump" for the limit behavior of WIP. In the left panel with $\alpha = 0.364287235213 > \alpha_1$, the fixed point is unstable, and a globally attracting limit cycle with a shape similar to the ones in Figure 7.6 is formed. In the right panel with $\alpha = 0.364287235212 < \alpha_1$ however, the locally stable fixed point suddenly becomes the global attractor.

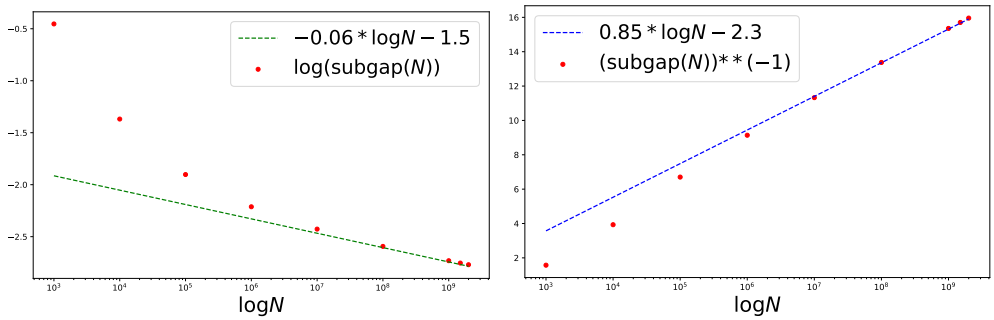
The purpose of using such large values of N is that under this critical situation, the quantity subgap(N) in Equation (7.1) converges to zero extremely slow, for this singular example.

Indeed, as shown in Figure 7.10, where we plot in $\log N$ scale, subgap(N) does not look like to converge to zero at $O(1/\sqrt{N})$ rate, as the other singular examples we studied in Remark 3.3.3 and Section 7.1.1. The most suitable exponent to fit a $O(1/N^{\beta})$ rate is around $\beta \approx 0.06$. A better fitting is found by plotting $1/\text{subgap}(N)$ as a function of $\log(N)$, shown in the right panel of Figure 7.10, which suggests that $\text{subgap}(N) = O(1/\log N)$.

To conclude, the complexity of the example studied in this section is in response to our next discussion in Section 7.1.4, which reveals the peculiarity and unpredictability of the deterministic dynamics of WIP in singular cases, not to mention its corresponding stochastic system analysis.

7.1.4 The convergence rate in singular cases

We have not succeed at proving any convergence rate result in the singular case. This seems somehow unexpected, since if the convergence rate is exponentially fast in non-singular cases, then we should at least be able to show the classical square root convergence rate in the singular case, knowing that the singular case is on the extreme of two non-singular regions. In fact, the difficulty that facing us is that in order to analyse the deterministic dynamics of WIP in the singular case, we need to understand the behavior of random products of arbitrary length of two matrices $\mathbf{K}_1$ and $\mathbf{K}_2$, that



(a) Fitting with sub square root rate. (b) Fitting with logarithmic rate.

Figure 7.10 – We take $\alpha = 0.364287235212$, which is numerically indistinguishable from the critical value α_1. It appears that the rate to which subgap(N) converges to 0 is much slower than the square root in this singular case. In the left panel we use a log-log plot, and the best-fit exponent is $\beta \approx 0.06$, implying that subgap$(N) \approx O(1/N^{0.06})$. A better fitting is found by plotting $1/\text{subgap}(N)$ as a function of $\log(N)$, shown in the right panel, which suggest that subgap$(N) = O(1/\log N)$.

corresponds to the linear factors of the two affine regions to which the singular fixed point lives. This is more or less equivalent to asking if the spectral radius of $\mathbf{K}_1$ and $\mathbf{K}_2$ is smaller, equal or larger than 1, which has been proven to be an undecidable problem in general in the literature.

Our belief is that in the singular case, even the classical $O(1/\sqrt{N})$ convergence rate does not hold without additional assumptions, indicating that it is possible to have a slower than square root convergence rate for singular problems. Indeed, as we have seen from the singular restless bandit model considered in Section 7.1.3, for which one of the two matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ is unstable (having two real eigenvalues greater than 1), while the fixed point remains a global attractor. It appears that numerically, the performance of WIP converges to the relax bound at logarithmic rate, i.e. $V^{(1)}_{\text{rel}}(\alpha) - V^{(N)}_{\text{WIP}}(\alpha)/N = \Theta(1/\log N)$. Since it can be shown that $V^{(1)}_{\text{rel}}(\alpha) - V^{(N)}_{\text{opt}}(\alpha)/N = \Theta(1/\sqrt{N})$ always holds, this indicates that WIP is susceptible to be asymptotically optimal at logarithmic rate in a singular case.

These situations are in response to the common belief that piecewise affine dynamical systems are almost as hard as non-linear systems. Along this direction we should cite Blondel et al. [10], in which it is shown that global convergence of piecewise affine continuous dynamics in $\mathbb{R}^d$ is undecidable as long as $d \geq 3$, and we only need $d \geq 2$ if we remove the continuity requirement. Interestingly, all these peculiar numerical examples that we have constructed are also in dimension 3. Of course the piecewise affine continuous dynamical systems under our consideration belong to a particular class, but the general undecidability theorem in Blondel et al. [10] indicates that it is unlikely to prove meaningful result in this direction.

To conclude this discussion, we summarize the asymptotic convergence rate of WIP according to the nature of its deterministic dynamic in Figure 7.11. Notice that

in theory, only the exponentially fast rate in the global attractor non-singular locally stable case has been proven, while the rest are empirical results supported by numerical evidence. Fortunately, the exponential convergence rate is also the case that is mostly encountered in practice.

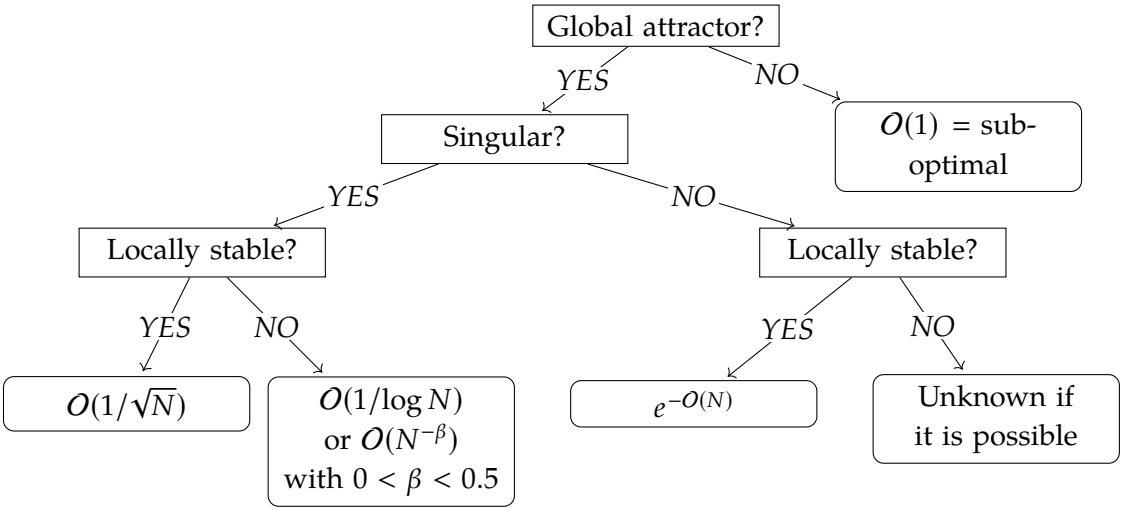


Figure 7.11 – A summary of the asymptotic convergence rate of WIP according to the nature of its deterministic dynamic. Only the exponential rate in the global attractor non-singular locally stable case is proven in a rigorous manner. Fortunately, this is also the case that is mostly encountered in practice.

7.2 Extended Results from Chapters 4 and 5

This section consists of extended discussions from Chapters 4 and 5, centered around the LP-based (non-update) policy. In Section 7.2.1, we provide a statistical test to see how likely a restless bandit model is non-degenerate in finite horizon, and satisfies the Uniform Global Attractor Property in infinite horizon, since these are the essential conditions for the corresponding LP-based policies to be asymptotically optimal at exponential rate. We investigate the time complexity for solving the LP in Section 5, under both finite and infinite horizon, as this is the most time consuming step to construct any LP-based policy.

7.2.1 How general is the general case? (continued)

Following the same spirit as the test we did in Section 3.4.1 for verifying the rarety of violating conditions in Theorem 3.3.2, in this section we provide statistics on the proportion of restless bandits that violates the UGAP for infinite horizon problems,, and the non-degenerate condition for finite horizon problems, since these are the conditions required for the corresponding LP-based policies to have an exponentially fast convergence to optimality, discussed respectively in Chapter 4 and Chapter 5.

Clearly such a statistical experiment depends on how we generate the large number of models. Denote by exp(1) an exponential distribution with parameter 1. The most general category is the following

- **(Dense model)** Each term in the transition matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ are generated using exp(1), and we normalize each line of the matrices so that the terms sum to 1.

The dense model being too vast, it is useful to restrict to some sub-categories. We propose the following:

- **(Tri-diagonal model)** Only the terms on the three main diagonals of the transition matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ are generated using exp(1), the rest of the terms are all 0. This sub-category has some practical value, as it includes birth-and-death processes, and can be applied to queueing theory, for instance.

In the finite horizon case, we obtain the following statistics with $T = 50$, shown in Table 7.1, where each percentage number is obtained from 10^5 uniformly generated samples. Note that all these non-degenerate models in Table 7.1 are also rankable, which corresponds to our discussion in Section 5.4.3. Moreover, since we are only testing the non-degeneracy using one optimal solution given by our numerical LP solver, the numbers given in Table 7.1 should be, strictly speaking, a lower-bound on the true numbers.

Scenario	Dense	Tri-diagonal
$d = 5$	89%	75%
$d = 10$	91%	60%
$d = 15$	93%	52%
$d = 20$	93%	42%

Table 7.1 – Percentage of non-degenerate finite horizon restless bandit with $T = 50$.

As for the infinite horizon case, in general it is hard to verify if a dynamical system satisfies the Uniform Global Attractor Property (UGAP) defined in Chapter 4. Hence we propose a weaker condition that is easy to check numerically:

(Stable property) The linear factor of the piecewise affine map $\Phi(\cdot)$ defined in (4.5) around $\mathbf{m}^*$ is an stable matrix, where an stable matrix is a matrix that does not have eigenvalues with norm greater than 1.

Since violation of the Stable Property implies violation of the Uniform Global Attractor Property, we shall subsequently test only the stability of an infinite horizon restless bandit (easier than UGAP). We obtain results shown in Table 7.2.

Several conclusions can be drawn from theses statistics:

- The non-degeneracy (for finite horizon restless bandit) and the stability (for infinite horizon restless bandit) are generic properties, they hold for most of the models in the dense category.

Scenario	Dense	Tri-diagonal
$d = 5$	99.9%	96.5%
$d = 10$	>99.9%	89.1%
$d = 15$	>99.9%	81.9%
$d = 20$	>99.9%	76.9%

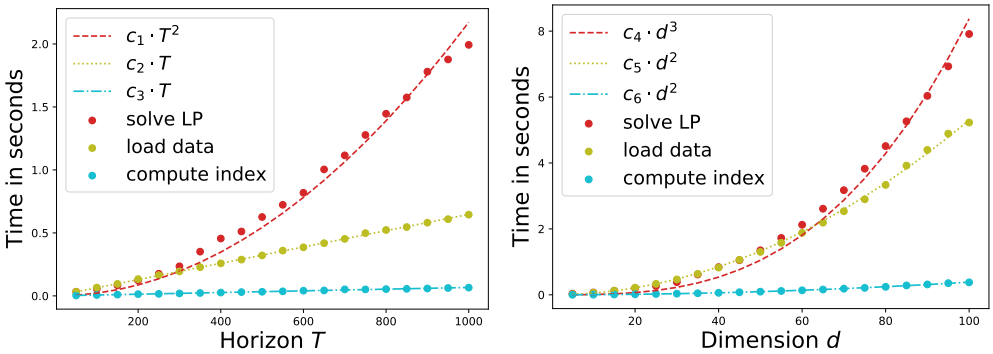
Table 7.2 – Percentage of stable infinite horizon restless bandit.

- Unfortunately, within the important tri-diagonal sub-category, the situation is the other way around. As their size d increases, the problems are more likely to be degenerate (for the finite horizon restless bandit) and unstable (for the infinite horizon restless bandit).

7.2.2 Experimental complexity of solving the LP

The most time-consuming step for the computation of any LP-based policy is to solve the linear program. In practice, the complexity to solve (4.4) and (5.4) may depend on the LP solver we use. For instance, the most common LP algorithm–the simplex method, can have exponential complexity in its worst case. To determine what is the practical complexity of solving the LP, we use the default LP solver from the `PuLP` package in Python and measure the average time to construct and solve the LP.

We first fix the dimension $d = 5$ and let T vary from 50 to 1000. For each specific value of T in this range, we sample 400 randomly generated instances and solve the corresponding LP problems and compute the finite-horizon LP indices. We record the time needed to load the data before solving the LP, the time elapsed to solve the LP, as well as the extra time required to compute the indices $I_s(t)$. The results are shown in Figure 7.12a. The constants c's are determined by minimizing the mean squared error. Similarly, we fix $T = 30$ and let d vary from 5 to 100. The results are shown in Figure 7.12b.



(a) Fix d=5 and change T

(b) Fix T=30 and change d

Figure 7.12 – Time complexity of the `PuLP` LP solver for (5.4).

These figures suggest that solving the finite horizon LP (5.4) has time complexity

$O(T^2d^3)$. Likewise, the empirical time complexity for solving the LP (4.4) of infinite horizon restless bandit is observed to be $O(d^3)$. The time to load the data is of order $O(Td^2)$, as expected. We also remark that the extra time to compute the LP indices are almost negligible, which combining with our later discussion in Section 5.6.1, suggest that it is beneficial to apply the LP-index policy, as compared to the policies in Zhang and Frazier [53] obtained after a single water-filling procedure.

7.3 Extended Results from Chapters 5 and 6

This section consists of extended discussions from Chapters 5 and 6, centered around the LP-update policy. We provide a geometric explanation as why the LP-update policy can perform much better on certain problems than the LP-index policy in Section 7.3.1. We then propose in Section 7.3.2 a set of criteria to measure the hardness of a finite horizon restless bandit, and show how the LP-update policy is capable to dealt with the hardest problems, for which any other LP-based non-update policies are incompetent.

7.3.1 Why the LP-update policy performs better?

We illustrated previously in Sections 5.6.2 and 6.5.4 that the LP-update policy outperforms the other LP-based policies on the applicant screening problem and its generalizations. On this particular problem with short horizon ($T = 5$ or $T = 10$), the better performance of the LP-update policy is explained by taking the new information into account for the up-coming decisions. In this section we take a more general (and geometric) viewpoint to explain why the LP-update policy is the winner, especially when facing the unstable problems for which the others (e.g. the LP-index policy) do wrong.

Let us first consider the following heuristic policy, that can be seen as approximating the finite horizon LP-index policy by the corresponding infinite horizon counter-part:

(LP-infinite policy) The strict priority policy by applying the LP-index policy obtained from its infinite horizon counter-part along all the T time steps [1], where the infinite horizon LP-index policy is discussed in Chapter 4.

It should be clear that the LP-infinite policy is more time efficient than the LP-index policy, and the larger the horizon the more apparent this efficiency. The reasoning for why the LP-update policy performs better then goes as follows, for which we divide into several steps. The main idea has already been mentioned in Section 2.6 for motivating the LP-update policy:

- In general, we expect the finite horizon LP-index policy to "resemble" the infinite horizon LP-index policy, when the horizon T becomes large, so that the LP-infinite policy gives nearly as good performance as the more time consuming LP-index policy, for most situations.

[1] As a reminder, we can always associate a finite horizon restless bandit with parameters $\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha, N, T, \mathbf{m}(0)\}$ its infinite horizon counter-part with parameters $\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha, N\}$, by ignoring the finite horizon and initial condition.

- This will not be the case, if the infinite horizon problem violates the Stable Property, defined in Section 7.2.1. Since the induced map Φ defined in Equation (4.5) for the priority order of the LP-infinite policy is unstable around the stationary point $\mathbf{m}^*$ (also the unique fixed point in Whittle's case), on which the highest possible stationary reward V_{rel} is given. Under such situations, the trajectory of the LP-infinite policy is mostly close to the attracting cycle, and its performance is approximately the average reward on the cycle (as we already shown in Sections 7.1.2 and 7.1.3). If this average reward turns out to be significantly smaller than V_{rel}, then the LP-infinite policy will perform poorly.

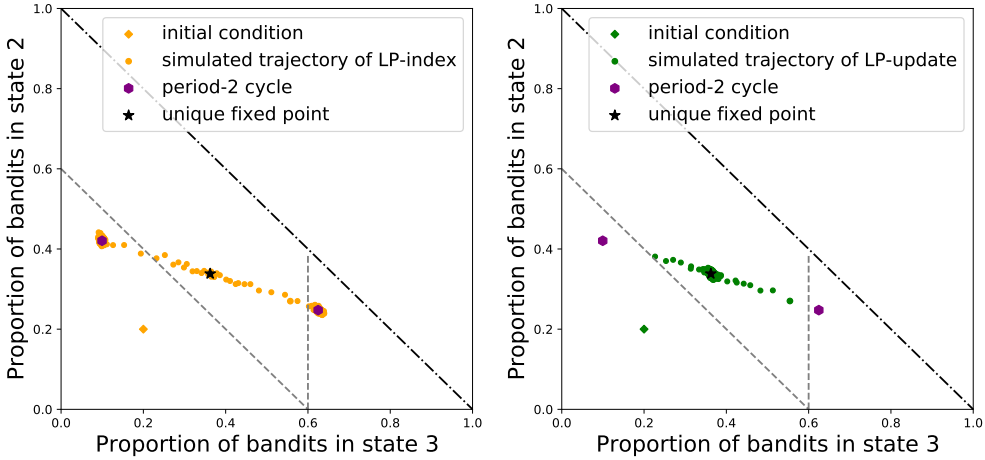
- The LP-index policy, on the other hand, is more sophisticated than the LP-infinite policy, so that under such unstable situations, it tries to avoid being attracted to the limit cycle, and keeps the trajectory close to the stationary point $\mathbf{m}^*$. This is usually done in a very subtle way, by first sending the trajectory to the stable manifold [2] of the unstable stationary point $\mathbf{m}^*$, and then apply Φ once near this stable manifold. Since along this manifold, the trajectory converges to $\mathbf{m}^*$ exponentially fast under $\Phi_{t\geq 0}$.

- However, this clever solution found by the LP-index policy only works well for the deterministic trajectory, as it is very sensitive to small perturbations. Indeed, by adding even an extremely small noise (i.e. with a very large arm population N), the stochastic trajectory by using the LP-index policy will not be able to keep staying close to $\mathbf{m}^*$, and sooner or later will be attracted to the limit cycle and stuck there for the rest of the times.

- The LP-update policy prevents this from happening, by keep sending the stochastic trajectory back to the neighbourhood of the unstable fixed point $\mathbf{m}^*$, on which the maximum stationary reward can be gained.

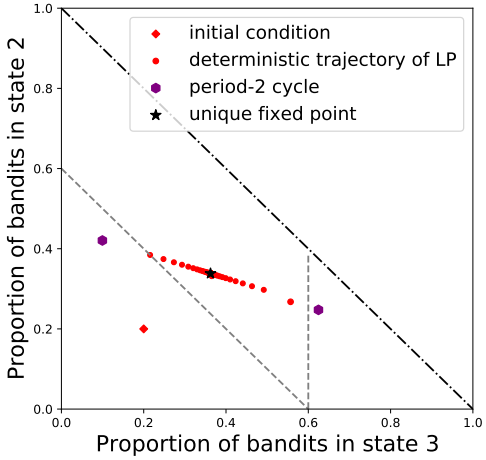
Let us use the attracting period-2 cycle example studied in Section 7.1.2 for illustration, we take the second one in Figures 7.4 and 7.5. Previously this example was considered in infinite horizon using WIP. Now we take a new look on it by studying under a long but finite horizon $T = 1000$, with an initial condition $\mathbf{m}(0) = [0.6, 0.2, 0.2]$, represented as a diamond shaped point in Figure 7.13. In Figure 7.13c, the optimal deterministic trajectory given by the LP is shown as the red points. Notice that a large proportion of the 1000 points is clustered near $\mathbf{m}^*$, only the last few of them that correspond to arriving at the end of horizon are spread around, which is caused by the effect of finite horizon. We next simulate the stochastic trajectories with an arm population $N = 10^5$ and the same initial condition, using the LP-index policy in Figure 7.13a (the orange points), and the LP-update policy in Figure 7.13b (the green points). We remark that for the LP-index policy simulation, there are two clusters of points

[2] If we write $\mathbf{V}_{\text{stable}}$ as the vector space with origin $\mathbf{m}^*$, which is spanned by the eigenvectors that correspond to the eigenvalues with norm < 1 for the linear factor of Φ in the zone $\mathcal{Z}$ of $\mathbf{m}^*$. The stable manifold of the unstable dynamic of $\Phi_{t\geq 0}$ near $\mathbf{m}^*$ is the intersection of $\mathbf{V}_{\text{stable}}$ with $\mathcal{Z}$. In general it is a polygon with codimension ≥ 1 in Δ^d.



(a) LP-index trajectory (in orange)

(b) LP-update trajectory (in green)



(c) Deterministic trajectory (in red)

Figure 7.13 – Comparing the simulated trajectories using the LP-index policy (the orange points in Figure 7.13a) and the LP-update policy (the green points in Figure 7.13b), on a unstable dimension $d = 3$ problem with an attracting period-2 limit cycle, studied previously in Section 7.1.2. The deterministically optimal trajectory is shown as the red points in Figure 7.13c.

around the period-2 attracting cycle, indicating that the trajectory has deviated from the unstable point $\mathbf{m}^*$. However, for the LP-update policy simulation, most of the points remain in a small neighbourhood of $\mathbf{m}^*$, despite the last few of them for the same reason as the deterministic trajectory. This shows the power of the LP-update policy for dealing with stability issue.

To summarize, the reason that the LP-update policy outperforms the LP-index policy (especially on unstable models with a large horizon) is that the stability issue causes the stochastic trajectory deviates fast from the optimal one, if we apply the LP-index policy that does not take into account the accumulated errors. The LP-update policy circumvents this difficulty by constantly correcting the deviation.

7.3.2 The hardest finite horizon restless bandits

The purpose of this section is twofold: firstly, we suggest a method to measure the hardness of a finite horizon restless bandit; secondly, we show how the LP-update policy is capable to dealt with the hardest problems, for which any other LP-based non-update policies are incompetent.

A natural idea to define hardness on bandit problems is that if some naive straightforward policy already gives close-to-optimal performance, then this problem should be characterized as easy. To this end, let us consider the following two heuristics for a finite horizon restless bandit:

- **(Greedy policy)** A strict priority policy based on the difference of rewards $\mathbf{R}^1$ and $\mathbf{R}^0$. The larger this difference, the higher priority is given to the corresponding state.

- **(LP-infinite policy)** A strict priority policy defined in Section 7.3.1.

It should be clear that these two heuristics are more efficient to implement, as the greedy policy requires almost nothing for computation, and from our discussion in Section 7.2.2, solving the infinite (resp. finite) horizon LP takes $O(d^3)$ (resp. $O(T^2d^3)$) time.

We next illustrate how to construct automatically a class of problems that any non-update LP-based policies are incompetent to dealt with. In our practice, we use the following rule to select hard problems, the reason should be clear from our previous discussion in Section 7.3.1:

(Hard restless bandit model) A finite horizon restless bandit is *hard*, if its infinite horizon model is unstable, and moreover the average reward on the attracting limit cycle has a score below 50, evaluated using:

$$\text{score}_\pi := \frac{V_\pi - V_{\text{rel-min}}}{V_{\text{rel}} - V_{\text{rel-min}}} \times 100, \tag{7.2}$$

where V_{rel} (resp. $V_{\text{rel-min}}$) is the optimal (resp. worst) value per-arm of the infinite horizon linear program (4.4), and π can be any policy (here it is the LP-infinite policy).

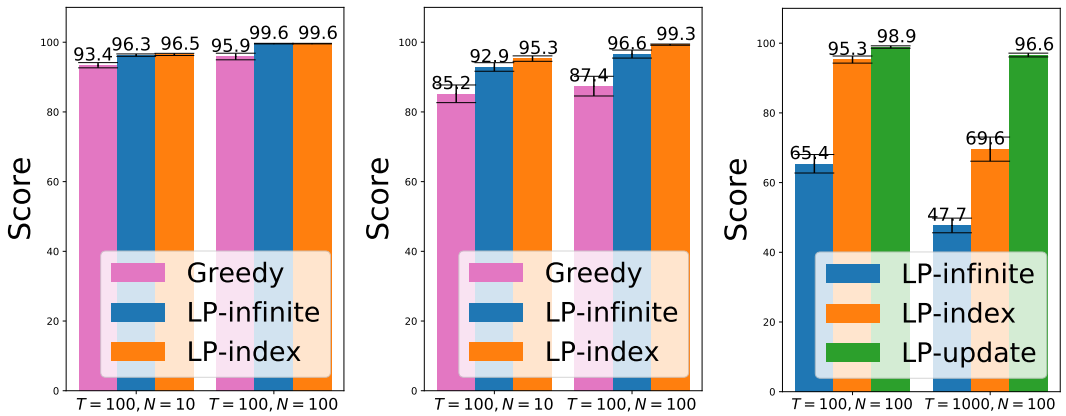
Recall from Table 7.2 that it is more probable to encounter instability in the tri-diagonal case. So it is natural to restrict to this category for finding hard models. We remark that on a test of 10^5 samples of uniformly generated tri-diagonal restless bandit models with $d = 10$, we record 1187 of them that are unstable (which also fits the statistics in Table 7.2), and 292 of them are hard, according to the above criteria.

In Figure 7.14 we study the following three scenarios, ordered in increasing difficulty. Under each scenario we choose 100 samples of restless bandit in dimension $d = 10$. We record the average score defined in (7.2) with a 95% confidence interval:

1. **(Dense)** In this scenario we choose 100 dense models, all of them being stable. We fix time horizon $T = 100$, and consider the three policies Greedy, LP-infinite and LP-index for $N = 10$ and $N = 100$. The results are shown in Figure 7.14a.

2. **(Tri-diagonal stable)** Like in the dense scenario, we consider the tri-diagonal and stable scenario, the results are shown in Figure 7.14b.

3. **(Tri-diagonal hard)** In this scenario we choose 100 tri-diagonal and hard models according to the criterion given above. We fix arm population $N = 100$, and consider the three policies LP-infinite, LP-index and LP-update for $T = 100$ and $T = 1000$. The results are shown in Figure 7.14c [3].



(a) Dense (b) Tri-diagonal stable (c) Tri-diagonal hard

Figure 7.14 – Scores of the policies on three scenarios in increasing level of hardness.

From this figure, we see that in the generic dense category, the simple greedy heuristic is already very efficient. In the tri-diagonal case, the greedy heuristic is insufficient. However, if the model is stable, then the LP-infinite policy is an ideal heuristic. The LP-index policy is also a good choice since it performs slightly better than LP-infinite, but it is much slower if the horizon T is large. When the tri-diagonal model is hard, the LP-infinite heuristic should never be considered. We can still rely

[3] Remarkably, the 100 hard models we consider here are all non-indexable in Whittle's sense (Definition 3.2.1).

on the LP-index policy if the horizon T is small,. However, when T is large, only the LP-update policy performs well.

As a remark, the scores of the LP-index policy in Figure 7.14c are simulated with an arm population $N = 100$. We see that when $T = 1000$ the average is only around 70, but the theory tells us that with a large enough N the score will eventually reach near 100. So how big the arm population should be in order to have a good score in this situation? It turns out that even with $N = 10^9$ the average score still remains at 70, so the asymptotic optimality of the LP-index policy is impractical for the hard models with a large horizon. This justifies once again the hardness of the problems, and we really need to rely on the LP-update policy, at the cost of a potentially much longer computation time.

To summarize this discussion, we propose the following characterization of the hardest finite horizon restless bandits (and more generally the weakly coupled MDPs):

A finite horizon restless bandit (weakly coupled MDPs) is among the hardest problem to solve, e.g. to obtain a near optimal policy in a reasonable computation time, if the following conditions are met:

- *Its infinite horizon problem is unstable around the stationary measure point* $\mathbf{m}^*$*, and its deterministic dynamic has an attracting limit cycle;*
- *The average reward on this cycle is low compared to the reward* V_{rel} *on the stationary measure point* $\mathbf{m}^*$ *(e.g. scores below* 50*);*
- *The finite horizon* T *is large (e.g.* $T = 1000$*);*
- *The population of arms* N *is medium (e.g.* $N = 100$*).*

The justification for the last two conditions is that, with a long horizon, solving the LP is more time consuming (recall that in Section 7.2.2 we show experimentally the time complexity for solving the LP grows quadratically with T), while we can not approximate by using the LP-infinite policy, for troubles caused by the first two conditions. In the meantime, with a medium size of arm population N, the stochastic noise is large enough so that the more sophisticated version of the LP-update policy discussed in Chapter 6, which avoids unnecessary updates by exploiting the previous LP solutions will not give much improvement, and we need to constantly solve a new LP for applying an update. On the other hand, if the arm population is small, say $N = 5$, then a direct dynamic programming approach may be favored.

8

General Conclusion and Open Questions

Physicists are concerned about what is true, while mathematicians care why it is true.
– Mr. Nobody

In this thesis we have investigated the restless bandit problem, for which the exact solution is known to be out-of-reach. We construct computationally efficient policies with provable performance bounds, that may differ depending on certain properties of the problem (e.g. singularity, degeneracy). So that we now have a better theoretical understanding, as well as a practical guide for when and why to use certain policies on a certain class of restless bandit models. We also generalize the results over the much broader framework of weakly coupled MDPs, and it is our humble wish that this may inspire more real life applications in the future.

One important theme of the thesis is the running back-and-forth between the scenarios of infinite and finite horizon. The thesis begins by investigating more closely WIP on infinite horizon restless bandit problems. For applying WIP, we always do so by checking the indexability beforehand, then computing the indices and testing numerically the global attractor property. The LP approach releases us from the verification of indexability, while the finite horizon LP approach is free from the global attractor requirement as well.

Unfortunately, the instability issue of the infinite horizon problem is inherited into his finite horizon brother, as we have illustrated in the additional numerical experiments in Chapter 7. So even we do not need the global attractor property to apply the LP-based policies on a finite horizon problem, for the seek of asymptotic optimality, still its performance is suspectable to be bad. In this sense the issue is not completely gone in finite horizon. One way that saves us out of the trouble is to apply the LP-update policy, that consists of constantly correcting the large deviations caused by the instability, at the cost of a potentially much longer computation time.

Naturally, we may then want to ask if it is possible to construct a policy in infinite horizon, that is asymptotically optimal even for unstable problems. This is an important open question already been mentioned in Verloop [47]. Necessarily such a policy can not be a strict priority policy as considered in Chapter 4. We believe that the LP-update policy in finite horizon may pave a way towards this construction, since it is shown

numerically to be immune to instability. The big challenge ahead is then to dealt with the infinite vs. finite horizon issue, as the LP-update policy can only be defined with a finite horizon and an initial condition.

One possible direction that may give insights on the theoretical aspect of this challenge is to first solve the following question: As already mentioned in Section 2.6, it is desirable to prove a result showing that the constant C' hidden in the O of Theorem 5.5.2 on asymptotic optimality of the LP-update policy does not grow exponentially with the horizon T. This can be reduced to studying the growth rate of the sensitivity constants on the initial condition of a sequence of T linear programs, that is observed numerically to be uniformly bounded in most cases, as discussed after the proof of Theorem 5.5.2. By investigating this further, we may obtain a better understanding of the link between the finite horizon and infinite horizon problems.

Appendix A

Table of Notations and Key Definitions

Notations:

N	The number of arms of the restless bandit
$\mathcal{A}$	The action space of an arm
$\mathcal{S}$	The state space of an arm
$\mathcal{U}$	The space of state-action pair (s, a) of an arm, with $s \in \mathcal{S}$ and $a \in \mathcal{A}$ a feasible action
α	The proportion of (maximal) activated arms at each time step, with $0 < \alpha < 1$
d	The number of states of an arm, or $\lvert\mathcal{S}\rvert$
T	The finite horizon
$\mathbb{P}^a$	The transition probability matrix of dimension $d \times d$ for an arm under action $a \in \mathcal{A}$
$\mathbf{R}^a$	The reward vector of dimension d for action $a \in \mathcal{A}$
Δ^d	The simplex of probability vectors of dimension d
$\mathbf{m}$	A configuration vector in Δ^d representing the proportion of arms in each state of a bandit, considered under the deterministic dynamic
$\mathbf{m}^*$	The configuration vector in Δ^d representing the stationary measure of WIP, also proven to be the unique fixed point
$\mathbf{M}$	A configuration vector in Δ^d representing the proportion of arms in each state of a bandit, considered under the stochastic dynamic
$\mathbf{M}^{(N)}$	A configuration vector in Δ^d representing the proportion of arms in each state of a stochastic N-armed bandit, emphasizing that each coordinate $M_s^{(N)}$ is an integer multiple of $1/N$
$\mathbf{m}(0)$	The initial configuration vector of the finite horizon bandit
$\mathbf{y}$	A decision vector with $y_{s,a}$ representing the proportion of arms in state s undertaking action a, for $(s, a) \in \mathcal{U}$
$\mathbf{Y}^{(N)}$	A decision vector for a stochastic N-armed bandit
$\mathbf{m}^*(t), \mathbf{y}^*(t)$	A deterministically optimal configuration (decision) vector at time t of a bandit, given by the solution to the relaxed linear program

$\mathcal{Y}(\mathbf{m})$	The set of feasible decision vectors $\mathbf{y}$, given that the configuration vector is $\mathbf{m}$
$\mathcal{Y}^{(N)}(\mathbf{m})$	The set of feasible decision vectors $\mathbf{Y}^{(N)}$ for a population N weakly coupled MDP, given that the process is in configuration vector $\mathbf{m}$
$V_{\text{rel}}^{(1)}(\alpha)$	The value per-arm of the infinite horizon relaxed problem by solving the linear program, emphasizing its dependence on α and independence on N
$V_{\pi}^{(N)}(\alpha)$	The value per-arm of a policy π on the infinite horizon restless bandit problem, emphasizing its dependence on N and α
$V_{\text{rel}}(\mathbf{m}(0), T)$	The value per-arm of the finite horizon relaxed problem by solving the linear program, emphasizing its dependence on the initial configuration vector $\mathbf{m}(0)$, and the horizon T
$V_{\pi}^{(N)}(\mathbf{m}(0), T)$	The value per-arm of a policy π on the finite horizon N-armed restless bandit problem, emphasizing its dependence on the arm population N, the initial configuration vector $\mathbf{m}(0)$, and the horizon T
$V_{\pi}(\mathbf{m}(0), T)$	The value per-arm of a policy π on the finite horizon deterministic restless bandit problem, viewing π as deterministic maps from $\mathbf{m}$ to $\mathcal{Y}(\mathbf{m})$
$\|\cdot\|$	The $\mathcal{L}^{\infty}$-norm
$\|\cdot\|_1$	The $\mathcal{L}^{1}$-norm

KEY DEFINITIONS:

Whittle index and indexability	Definitions 2.2.1 and 3.2.1
non-singularity	Section 3.3.1
non-degeneracy (for infinite horizon restless bandits)	Section 4.3
non-degeneracy (for finite horizon restless bandits)	Section 5.4.1
non-degeneracy (for weakly coupled MDPs)	Section 6.4.1
rankable (for finite horizon restless bandits)	Section 4.3
The LP-priority policy	Section 4.4
The infinite horizon LP-index policy	Section 4.5
The finite horizon LP-index policy	Section 5.5.1
The water-filling policy	Section 5.4.2
The LP-update policy for restless bandits	Section 5.5.2
The LP-update policy for weakly coupled MDPs	Sections 6.3.2 and 6.4.2
The LP-infinite policy	Section 7.3.1
The LP-compatibility	Section 5.3.2
randomized rounding	Section 5.2.3
perfect rounding	Section 6.4.3

ABBREVIATIONS:

LP	linear program
WIP	Whittle index policy
MDP	Markov decision process
UGAP	the uniform global attractor property

Bibliography

[1] Aalto S, Lassila P, Osti P (2015) Whittle index approach to size-aware scheduling with time-varying channels. *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 57–69.

[2] Adelman D, Mersereau AJ (2008) Relaxations of weakly coupled stochastic dynamic programs. *Oper. Res.* 56(3):712–727, ISSN 0030-364X.

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

[4] Ansell P, Glazebrook KD, Nino-Mora J, O'Keeffe M (2003) Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research* 57(1):21–39.

[5] Astaraky D, Patrick J (2015) A simulation based approximate dynamic programming approach to multi-class, multi-resource surgical scheduling. *European Journal of Operational Research* 245(1):309–319, ISSN 0377-2217.

[6] Avrachenkov K, Filar JA, Gaitsgory V, Stillman A (2016) Singularly perturbed linear programs and markov decision processes. *Operations Research Letters* 44(3):297–301.

[7] Avrachenkov KE, Borkar VS (2016) Whittle index policy for crawling ephemeral content. *IEEE Transactions on Control of Network Systems* 5(1):446–455.

[8] Avrachenkov KE, Filar JA, Howlett PG (2013) *Analytic perturbation theory and its applications* (SIAM).

[9] Bertsimas D, Niño-Mora J (2000) Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research* 48(1):80–90.

[10] Blondel VD, Bournez O, Koiran P, Tsitsiklis JN (2001) The stability of saturated linear dynamical systems is undecidable. *Journal of Computer and System Sciences* 62(3):442–462, ISSN 0022-0000.

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

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

[13] Dolgov DA, Durfee EH (2004) Optimal resource allocation and policy formulation in loosely-coupled Markov decision processes. *Proceedings of the Fourteenth International Conference on Automated Planning and Scheduling (ICAPS 04)*, 315–324 (Whistler, BC).

[14] Filar JA, Avrachenkov K, Altman E (1999) An asymptotic simplex method for parametric linear programming. *1999 Information, Decision and Control. Data and Information Fusion Symposium, Signal Processing and Communications Symposium and Decision and Control Symposium. Proceedings (Cat. No. 99EX251)*, 427–432 (IEEE).

[15] Gast N, Bortolussi L, Tribastone M (2018) Size Expansions of Mean Field Approximation: Transient and Steady-State Analysis. *2018 - 36th International Symposium on Computer Performance, Modeling, Measurements and Evaluation*, 1–2 (Toulouse, France).

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

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

[18] Gast N, Gaujal B, Yan C (2022) LP-based policies for restless bandits: necessary and sufficient conditions for (exponentially fast) asymptotic optimality, working paper or preprint.

[19] Gast N, Latella D, Massink M (2018) A refined mean field approximation of synchronous discrete-time population models. *Performance evaluation* 126:1–21.

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

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

[22] Gittins JC (1979) Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B* 148–177.

[23] Glazebrook K, Nino-Mora J, Ansell P (2002) Index policies for a class of discounted restless bandit problems. *Advances in Applied Probability* 34:754–774, ISSN 0001-8678.

[24] Gocgun Y, Ghate A (2010) A lagrangian approach to dynamic resource allocation. *Proceedings of the 2010 Winter Simulation Conference*, 3330–3340.

[25] Gocgun Y, Ghate A (2012) Lagrangian relaxation and constraint generation for allocation and advanced scheduling. *Computers and Operations Research* 39(10):2323–2336, ISSN 0305-0548.

[26] Hawkins JT (2003) *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. Ph.D. thesis, Massachusetts Institute of Technology.

[27] Hodge DJ, Glazebrook KD (2015) On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Advances in Applied Probability* 47(3):652–667.

[28] Hsu YP (2018) Age of information: Whittle index for scheduling stochastic arrivals.

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

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

[31] Kifer Y (1988) *Random Perturbations of Dynamical Systems*. Progress in Probability (Birkhäuser Boston), ISBN 9783764333843.

[32] Larrnaaga M, Ayesta U, Verloop IM (2016) Dynamic control of birth-and-death restless bandits: Application to resource-allocation problems. *IEEE/ACM Transactions on Networking* 24(6):3812–3825.

[33] Matoušek J, Vondrák J (2001) The probabilistic method. *Lecture Notes, Department of Applied Mathematics, Charles University, Prague* .

[34] 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.

[35] Nino-Mora J (2001) Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability* 76–98.

[36] Niño-Mora J, Villar SS (2011) Sensor scheduling for hunting elusive hiding targets via whittle's restless bandit index policy. *International Conference on NETwork Games, Control and Optimization (NetGCooP 2011)*, 1–8 (IEEE).

[37] 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.

[38] Niño-Mora J (2011) Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing* 23(2):254–267.

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

[40] Niño-Mora J (2022) Multi-gear bandits, partial conservation laws, and indexability. *Mathematics* 10(14), ISSN 2227-7390.

[41] Ouyang W, Eryilmaz A, Shroff NB (2012) Asymptotically optimal downlink scheduling over markovian fading channels. *2012 Proceedings IEEE INFOCOM*, 1224–1232 (IEEE).

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

[43] Patrick J, Puterman ML, Queyranne M (2008) Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research* 56(6):1507–1525.

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

[45] Raghunathan V, Borkar V, Cao M, Kumar PR (2008) Index policies for real-time multicast scheduling for wireless broadcast systems. *IEEE INFOCOM 2008-The 27th Conference on Computer Communications*, 1570–1578 (IEEE).

[46] Salemi Parizi M (2018) *Approximate dynamic programming for weakly coupled Markov decision processes with perfect and imperfect information*. Ph.D. thesis.

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

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

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

[50] Xiong G, Li J, Singh R (2021) Reinforcement learning for finite-horizon restless multi-armed multi-action bandits. *arXiv preprint arXiv:2109.09855* .

[51] Ying L (2017) Stein's method for mean field approximations in light and heavy traffic regimes. *POMACS* 1(1):1–27.

[52] 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)* .

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

[54] Zhang X, Frazier PI (2022) Near-optimality for infinite-horizon restless bandits with many arms. *arXiv preprint arXiv:2203.15853* .