# Exponential Asymptotic Optimality of Whittle Index Policy.

Nicolas Gast, Bruno Gaujal, Chen Yan

# Exponential Asymptotic Optimality of Whittle Index Policy

Nicolas Gast, Bruno Gaujal and Chen Yan

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France.

Contributing authors: nicolas.gast@inria.fr; bruno.gaujal@inria.fr; chen.yan@inria.fr;

**Abstract**

We evaluate the performance of Whittle index policy for restless Markovian bandit. It is shown in Weber and Weiss (1990) that if the 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 paper 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. Our result holds for the continuous-time model of Weber and Weiss (1990) and for a discrete-time model in which all bandits make synchronous transitions. Our proof is based on the nature of the deterministic equation governing the stochastic system: We show that it is a piecewise affine continuous dynamical system inside the simplex of the empirical measure of the arms. 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.

**Keywords:** Multi-armed Bandits, Whittle Index, Asymptotic Optimality

## 1 Introduction

A multi-armed bandit (MAB) problem is a sequential allocation problem: At each decision epoch, one or several arms are activated and some observable

rewards are obtained. The goal is to maximize the total reward obtained by a sequence of activations. There are at least three fundamental formalizations of the bandit problem depending on the assumed nature of the reward process: stochastic (i.i.d.), adversarial, and Markovian. Each bandit model has its own specific playing strategies and uses distinct techniques of analysis. We focus here on the Markovian bandits (for a thorough analysis of the other two types of bandit models, see *e.g.* Lattimore and Szepesvári (2020)). Each time, a subset of arms are chosen to be activated. Each arm generates an instantaneous reward that depend on their state and their activation. The state of each activated arm then changes in a Markovian fashion, based on an underlying transition matrix (or a rate matrix in the continuous-time case). Both the reward and the new state are revealed to the decision maker for its next decision. The arms that are not activated change state according to a different transition matrix. When the underlying stochastic transition laws are assumed to be known (see Duff (1995) for a treatment of the case where the transition matrices are unknown), the optimal policy can be computed via dynamic programming, and the problem is essentially of computational nature.

The above Markovian MAB problem has been solved in the *rested case* (non activated arms do not change their states) with one active arm at each decision epoch in Gittins (1979) by the Gittins index policy, which is a greedy policy that can be computed efficiently. In Whittle (1988), Whittle generalizes the 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 (*restless bandits*), according to a different transition matrix, as mentioned before. Under these generalizations, the problem can no longer be solved by a similar efficient index-type greedy policy, and indeed it has been proven in Papadimitriou and Tsitsiklis (1999) that this problem is PSPACE-hard. In Whittle (1988), however, Whittle conjectures that, under some conditions, the so-called "Whittle index policy" (WIP) should be optimal asymptotically, *i.e.* when the number of arms goes to infinity with a fixed proportion of active arms.

This conjecture has been proven in the famous paper Weber and Weiss (1990) for the continuous-time model, under several technical conditions (further discussed in Weber and Weiss (1991)), namely when the bandits are indexable and the drift of the Markov system has a fixed point that is a global attractor. These results further reinforce the interest of Whittle index, as restless arm models have been used in a wide range of applications and Whittle index policies turn to be efficient solutions. Among them one can cite wireless scheduling Aalto et al (2015); Raghunathan et al (2008), queuing systems Ansell et al (2003), crawling optimal content on the web Avrachenkov and Borkar (2016), load-balancing Larranaga et al (2016) and sensors Niño-Mora and Villar (2011). Some partially observable Markov decision processes (POMDPs) also fall into the category of restless Markovian bandits by using a Bayesian approach to construct the transition matrices. One concrete example is the multi-channel wireless scheduling problem of Liu and Zhao (2010);

Meshram et al (2018) that we study in Section 5 of this paper. In this system, there are $N$ Gilbert-Elliott channels and the state of a channel is only observed when a transmission is scheduled on this channel. We will show in Section 5 that this example falls in our framework and we will use it to illustrate our convergence resutls.

## Contributions

Despite the well-known asymptotic optimality of WIP (under some conditions) and its empirically good performance on numerous models listed above, as well as its many extensions, there is very limited research on how fast WIP becomes optimal. In this paper 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 in the discrete-time as well as the continuous-time cases, under the same conditions as the asymptotic optimality proven in Weber and Weiss (1990), 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 condition that the fixed point is *non-singular* (which almost always holds). This last condition will be discussed in length in the rest of the paper.

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 (2018a); Gast and Van Houdt (2017); Ying (2017) 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 (in the discrete-time case) or an exponential martingale concentration inequality (in the continuous-time case) is used to control its behavior outside the neighborhood.

To be more precise, 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 (1988). 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 $\mathcal{O}(1/\sqrt{N})$ 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.

## Related work

Our work can be seen as a natural sequence of the classical paper Weber and Weiss (1990), in which we show the exponential convergence rate of the asymptotic optimality proven there, and discuss its significance and consequence. Another paper that is closely related to our work is Verloop (2016), in which the author proposes a class of priority policies using linear programming, that are all asymptotically optimal, provided that a global attractor condition holds on the model (which is also needed in our result on WIP). The advantage of the policies in Verloop (2016) is that they do not need the indexability assumption. Although the focus of our paper is on Whittle index policy, our exponential convergence rate result proven in Section 6 on the continuous-time model can be adapted to show the exponential convergence rate for the policies of Verloop (2016). This is shown in our followup paper Gast et al (2022b).

WIP and the index policies of Verloop (2016) are asymptotically optimal for the infinite horizon undiscounted problems. More recently, a series of paper have studied asymptotically optimal policies for the finite-horizon criterion. These papers use relaxation of the problem that is similar to Whittle's relaxation but adapted to finite horizon. To the best of our knowledge, these ideas first appear in Hu and Frazier (2017) in which the authors show that

for any finite-horizon Markovian bandit problem, it is possible to compute a policy whose sub-optimality gap is $\mathcal{O}(1/\sqrt{N})$. A similar result is proven for *discounted* infinite-horizon criterion in Zhang and Frazier (2022). For finite-horizon problems, a notion of non-degenerate problem is introduced in Brown and Smith (2020); Zhang and Frazier (2021) that is similar to the notion of non-singular model that we introduce in our paper. In our followup paper Gast et al (2022b), we show that for a non-degenerate problem, it is possible to compute a policy that becomes asymptotically optimal at exponential rate. We should stress that the results developed in these papers are only for *finite-horizon* or *discounted* infinite horizon models whereas our paper focussed on the *undiscounted* infinite-horizon model. This leads to distinct proof techniques and also different results:

- For finite-horizon, the LP-index policies derived in Hu and Frazier (2017); Brown and Smith (2020); Zhang and Frazier (2021); Gast et al (2022b) are always asymptotically optimal with a rate of convergence of at least $\mathcal{O}(1/\sqrt{N})$. One obtain an exponential rate of convergence only when the problem is non-degenerate. Numerical evidences provided in Gast et al (2022b) show that there exist a large fraction of problems that are degenerate.
- For infinite-horizon (that we study in this paper), WIP or the LP-priority policy studied in Verloop (2016) are only asymptotically optimal when the global attractor property is satisfied. If it is satisfied then the problem is almost always non-singular, which implies that, for almost all models, WIP becomes optimal exponentially fast when it is asymptotically optimal.

## Organization of the paper

In Section 2, we introduce the discrete-time restless bandit model: all arms change their state simultaneously in discrete time, according to transition matrix $\mathbf{P^1}$ when being activated and $\mathbf{P^0}$ when not being activated. We also define the Whittle indices and the main notations used in the paper. We then present the main result of the paper in Section 3, namely exponential convergence for the performance of WIP to the optimal one in the general situation. In Section 4, we illustrate our results with several examples. We provide simulation and numerical estimations for the performance of WIP in different cases. In Section 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). Finally, in Section 6, we extend our result to the continuous-time model (bandits are continuous-time Markov chains, and decisions are made every time when one arm changes its state). We show that exponential convergence rate also holds in the continuous-time case, and highlight the similarity and difference between the two models.

# 2 The Discrete-Time Restless Bandit Model

We first describe the restless bandit model in Section 2.1. We then recall the definition of Whittle index in Section 2.2 and its relation with a linear problem in Section 2.3. Note that in our model, all arms are synchronous. This is a discrete-time version of the classical continuous-time model studied in Weber and Weiss (1990). We discuss how to adapt our results to the latter model in Section 6.

## 2.1 Model description

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

1. The model is composed of $N$ statistically identical arms. Each arm evolves in a finite state space $\{1, 2, \ldots, d\}$ and the state of arm $n$ at time $t$ is denoted by $S_n(t) \in \{1, \ldots, d\}$. The state space of the whole process at time $t$ is denoted by $\mathbf{S}(t) = \big(S_1(t), S_2(t), \ldots, S_N(t)\big)$.
2. Decisions are taken at times $t \in \mathbb{N}$. At each decision epoch, a decision maker observes $\mathbf{S}(t)$ and chooses $\alpha N$ of the $N$ arms to be activated, where we assume that $\alpha$ and $N$ are such that $\alpha N$ is an integer (the case of non-integer values will be discussed in Section 4.3). 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 $a \in \{0, 1\}$ and $t \in \mathbb{N}$:

$$\mathbb{P}(S_n(t+1) = j \mid S_n(t) = i, a_n(t) = a) = P_{ij}^a. \tag{1}$$

   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 $a \in \{0, 1\}$ is taken, a reward $R_i^a \in \mathbb{R}$ is earned.

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 5.

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 Puterman (1994) 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 $\Pi$ 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

$$V_{\text{opt}}^{(N)}(\alpha) := \max_{\Pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1}\sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)}\Big] \tag{2}$$

$$\text{subject to} \sum_{n=1}^{N} a_n(t) = \alpha N, \text{ for all } t \in \mathbb{N}. \tag{3}$$

In the above formulation and in what follows, the dependence of $a_n(t)$ on $S_n(t)$ based on a policy in $\Pi$ should be understood. We also assume that the parameters of the model are such that the states of the $N$-arms 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 (1990) to guarantee that neither the value of the optimization problem (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.

## 2.2 Indexability and Whittle index

In theory, a dynamic programming approach can be used to solve the optimization problem (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 (1999). To overcome this difficulty, Whittle introduces in Whittle (1988) a very efficient heuristic known as Whittle index policy (WIP). This heuristic is obtained by computing an index $\nu_i$ for each state $i$. At a given decision epoch, WIP activates the $\alpha N$ arms having currently the highest indices. We describe below how these indices are defined.

The index of an arm can be computed by considering each individual arm in isolation[1]. For a given $\nu \in \mathbb{R}$, we define the $\nu$-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 (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 there exists an optimal policy of the $\nu$-subsidized MDP such that the passive action is optimal in these states. Whittle indices are defined as follows:

**Definition 1** (Indexability and Whittle index) A bandit $(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1)$ is *indexable* if $\omega(\nu)$ is increasing in $\nu$, namely if for all $\nu \le \nu'$, we have $\omega(\nu) \subseteq \omega(\nu')$.

[1] The most efficient algorithm to test indexability and compute the index can be found in Gast et al (2022a). For a given model with $d$ states, the complexity of this algorithm is $o(d^3)$.

In this case, the Whittle index of a state $i$, that we denote by $\nu_i$, 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 $\nu_i$ is finite since the state space is finite.

It should be emphasized that there exist restless bandit problems that are *not* indexable, we discuss this in more detail in Section 4.1. Note that when $\mathbf{P}^0$ is the identity matrix, bandits are rested, *i.e.* the states of the arms that are not activated do not change. In such a case, a bandit is always indexable and Whittle index coincides with the classical definition of Gittins index, see Gittins et al (2011).

## 2.3 Whittle relaxation and asymptotic optimality

An intuition behind the definition of Whittle index is given by considering a relaxation of the original $N$ arms problem (2) where the constraint (3) is replaced by $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} a_n(t) = \alpha N$. This relaxed constraint imposes the time-averaged number of activated arms to be equal to $\alpha N$. We denote by $V_{\text{rel}}^{(N)}(\alpha)$ the value of the optimal control problem. It is given by the following optimization problem:

$$V_{\text{rel}}^{(N)}(\alpha) := \max_{\Pi} \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[ \sum_{t=0}^{T-1} \sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)} \Big] \tag{4}$$

$$\text{subject to} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} a_n(t) = \alpha N. \tag{5}$$

While the original problem (2)-(3) is computationally hard to solve, the value $V_{\text{rel}}^{(N)}(\alpha)$ after the relaxation is the solution to the following linear program:

$$V_{\text{rel}}^{(N)}(\alpha) = \underset{\mathbf{x} \geq \mathbf{0}}{\text{maximize}} \quad \sum_{s,a} R_s^a x_{s,a} \tag{6a}$$

$$\text{subject to} \quad \sum_{s} x_{s,1} = \alpha \text{ and } \sum_{s,a} x_{s,a} = 1, \tag{6b}$$

$$x_{s,0} + x_{s,1} = \sum_{s',a} x_{s',a} P_{s's}^a \qquad \forall s, \tag{6c}$$

where $x_{s,a}$ is the steady state probability for an arm to be in state $s$ and for which action $a \in \{0, 1\}$ is chosen. See Verloop (2016); Gast et al (2022b) for more detailed discussion about this LP and how it is derived.

The constraint (5) is weaker than the constraint (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). In fact, the next result shows that, as the

number of arms grows, the value of the original problem converges to this value. This theorem justifies the relaxation (5) by showing that when the number of arm is large, the value of the optimization problem (2) is close to $V_{\text{rel}}^{(N)}(\alpha)$.

**Theorem 1** *Consider an aperiodic recurrent discrete-time restless bandit model with $N$ identical arms and such that the matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ are rational. Then there exists a constant $c > 0$ that depends only on $\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1$ and $\alpha$, such that for any $N$ with $\alpha N$ being an integer, we have $V_{\text{rel}}^{(N)}(\alpha) = NV_{\text{rel}}^{(1)}(\alpha)$ and*

$$0 \leq V_{\text{rel}}^{(1)}(\alpha) - \frac{V_{\text{opt}}^{(N)}(\alpha)}{N} \leq \frac{c}{\sqrt{N}}. \tag{7}$$

Note that this theorem is the analogue of Theorem 1 of Weber and Weiss (1990), that proves that $\lim_{N\to\infty} \frac{V_{\text{opt}}^{(N)}(\alpha)}{N} = V_{\text{rel}}^{(1)}(\alpha)$ for the continuous-time bandit model that we will discuss in Section 6. To the best of our knowledge, the statement of this theorem in our setting of discrete-time bandit model is new. Moreover, our result shows that the convergence is *at least* in $\mathcal{O}(1/\sqrt{N})$. For completeness, we provide a proof of Theorem 1 in Appendix A. It is an adaptation of the proof of (Weber and Weiss, 1990, Theorem 1): we use a similar coupling argument, although the coupling has to be adapted to our discrete-time setting, and we also need the additional aperiodic assumption on the model.

While Theorem 1 guarantees that the original optimization problem converges to the relaxation, it does not guarantee any result on the performance of WIP. This leaves one important question: At which speed does WIP become optimal? In the remainder of the paper we will show that, except in rare cases, when WIP is asymptotically optimal, it does so at exponential speed with the number of arms $N$. This complements Theorem 1 by proving that, under the same conditions, the convergence in (7) occurs at exponential rate.

# 3 Main Results

We first show in Section 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 3.2. Later in Section 6 we will see how to extend this result to the classical model of continuous-time bandits of Weber and Weiss (1990).

## 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 do not exist two states that have

the same Whittle index. This is mostly a technical assumption that guarantees that there is a unique[2] 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 $N$-arms 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 $\Delta^d$, 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 $N$-arms 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

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

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 2** Assume that the bandit is indexable. Then:

(i) The definition of $\phi$ does not depend on $N$ (as long as $\alpha N$ is an integer) nor on $t$.

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

(iii) $\phi$ 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 Appendix B. We only 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 $\phi$ satisfies:

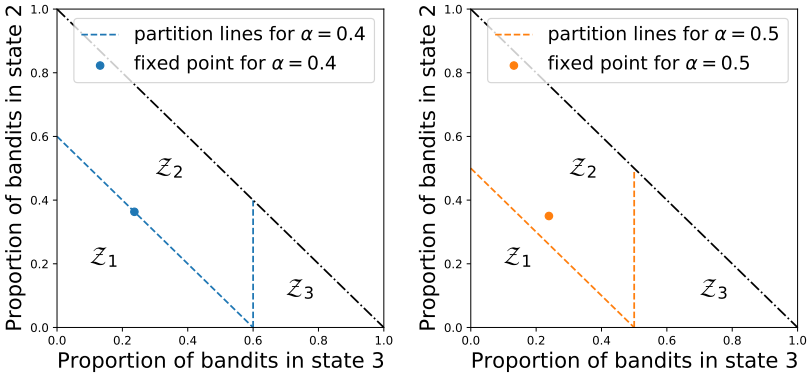
$$\begin{aligned}\phi_j(\mathbf{m}) = &\sum_{i=1}^{s(\mathbf{m})-1} m_i P^1_{ij} + (\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i) P^1_{s(\mathbf{m})j} \\ &+ (\sum_{i=1}^{s(\mathbf{m})} m_i - \alpha) P^0_{s(\mathbf{m})j} + \sum_{i=s(\mathbf{m})+1}^{d} m_i P^0_{ij}. \end{aligned} \tag{8}$$

[2] 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.

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

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

In what follows, we will denote by $\mathbf{m}^*$ the unique fixed point of $\phi$. As we will see in Theorem 3, the rate at which WIP becomes asymptotically optimal depends on: (1) whether the iterations of $\phi$ 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.

**Fig. 1**: An example with $d = 3$. When $\alpha = 0.4$ (Figure 1a) the fixed point is singular, while for $\alpha = 0.5$ (Figure 1b) it is not singular.

In Figure 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 $\Delta^3$ can be represented in a 2-dimensional space as $\Delta_c^2$, where $\Delta_c^d$ 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 1a and 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 $\alpha$. For this example, when $\alpha = 0.4$ (Figure 1a), the fixed point is singular, while for $\alpha = 0.5$ (Figure 1b), it is non-singular (all the other parameters in these two figures are the same, and are available in our Git repository).

## 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 $\alpha N$ arms that have the highest indices. We denote the long-term average expected reward of WIP as $V^{(N)}_{\text{WIP}}(\alpha)$:

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

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

Let $\Phi_t$ be defined as the $t$-th iteration of the map $\phi$, *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 $\phi$. 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 paper, unless otherwise specified, we use $\|\cdot\|$ to denote the $\mathcal{L}^\infty$-norm of a vector.

**Theorem 3** (Exponential convergence rate theorem: discrete-time case) *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 bandit problem is indexable.*

*(ii) The (unique) fixed point* $\mathbf{m}^*$ *of* $\phi$ *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* $\alpha$*, such that for any* $N$ *with* $\alpha N$ *being an integer,*

$$0 \leq V^{(N)}_{\text{rel}}(\alpha) - V^{(N)}_{\text{WIP}}(\alpha) \leq b \cdot e^{-cN}. \tag{9}$$

*Recall that* $V^{(N)}_{\text{rel}}(\alpha)$ *is the value of the relaxed problem* (4)-(5).

*Sketch of proof* The full details of the proof are given in Appendix C. 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 10 to show that for any configuration $\mathbf{m}$: $\mathbb{P}\left[\|\mathbf{M}^{(N)}(1) - \phi\big(\mathbf{M}^{(N)}(0)\big)\| \geq \delta \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\right] \leq e^{-2N\delta^2}$.
- By Lipschitz continuity of $\phi$, for a time $t$, we apply Lemma 10 to prove Lemma 11, 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 $\phi$ 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 Subsection C.4.3), hence allows us to restrict our attention to a zone where $\phi$ 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 Subsection C.4.4).

□

We give here some comments on the assumptions of Theorem 3, their practical relevance will be discussed in detail in Section 4.1. To prove that WIP is asymptotically optimal in the continuous-time case, Weber and Weiss (1990) assume two conditions: that the bandit is indexable (Assumption (i)) and that $\mathbf{m}^*$ is a global attractor (Assumption (iii)). To prove our result, we require two additional assumptions: (ii) The non-singular condition on $\mathbf{m}^*$, which is almost always satisfied (see Section 4.1); and (iv) that $\mathbf{m}^*$ is locally stable. We conjecture that condition (iii) implies condition (iv) here, but we leave this question for future work. In conclusion, our conditions are almost identical than the one needed in Weber and Weiss (1990).

Note that the most difficult assumption to verify is point (iii) that requires $\mathbf{m}^*$ to be a global attractor, as there is no general method to exclude cyclic or chaotic behaviors from a dynamical system. It is shown in Blondel et al (2001) that global properties of continuous piecewise affine functions in $\mathbb{R}^n$ are *undecidable* in general. Note that as the piecewise affine maps $\Phi$ induced by WIP form a subclass of piecewise affine functions, this does not imply that testing the global attractor property in our case is undecidable. In fact, there exist special cases for which showing the global attractor property is relatively easy. In general, this is done by finding a Lyapunov function or exhibiting some monotony property.

Assuming that a given map satisfies the global attractor property (iii) is in fact quite common in the literature (see for instance Weber and Weiss (1990) and most of the papers that use their results). Even if one cannot show mathematically that the global attractor property holds, it can still be tested numerically in an efficient way. In practice, a sufficiently large number of initial conditions suffices to approve (or disapprove) this assumption. Moreover, this condition is almost a necessary condition in the sense that there exists examples that satisfy all assumptions of Theorem 3 except this one and for which WIP is not asymptotically optimal (more comments on this in Remark 3). As already been discussed in Verloop (2016), it is a challenging question as how to design (non index-type) policies that are asymptotically optimal without the global attractor property.

**Remark 2 The singular case.** The non-singularity of the fixed point $\mathbf{m}^*$ is also necessary in the sense that the following simple example satisfies all the assumptions of Theorem 3 except this one and does not satisfy (9). 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

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

where $G$ is a standard normal random variable.

This example shows that, in a case where $\mathbf{m}^*$ is singular, the convergence in (9) 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 $\alpha$ (which is different from 0.5).

**Remark 3 Cyclic and chaotic behaviors**. Although the drift $\phi$ 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 (1990), one can infer a cyclic version of Theorem 3: 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 (from Theorem 1). Consequently, when cycles appear, the performance of WIP is asymptotically *sub-optimal*.

**Remark 4 What happens when $\alpha N$ is not an integer**. The exponential convergence rate in Theorem 3 assumes that $\alpha N$ is an integer. When it is not the case, a decision maker cannot activate exactly $\alpha N$ 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 4.3, the convergence rate in the first two solutions is much slower than in the third solution.

**Remark 5 Finding optimal constants**. Theorem 3 claims the existence of constants $b$ and $c$ for which the *inequality* (9) holds true, but we do not emphasize on the optimality of the constant $c$, in the sense of finding constant $\tilde{c}$ such that

$$\limsup_{N\to\infty} -\frac{1}{N} \log\left(V_{\text{rel}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha)\right) = \tilde{c}.$$

Our choice of $c$ in the proof of Theorem 3 provided in Appendix C actually depends subtly on the given parameters, and we believe that finding $\tilde{c}$ is, if not impossible, a much more demanding task. Nevertheless, later on in Section 4.2 we shall illustrate via numerical examples that the approximate value of $c$ is affected by the level of singularity of the fixed point, which in turn is affected by the value of $\alpha$, if all the other parameters $\{\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1\}$ are fixed.

# 4 Numerical Experiments

In this section, we first provide statistical results to justify the conditions needed for Theorem 3, 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 $\alpha N$ 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 (2022a). 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 $\alpha$. Moreover, being in an exact singular situation is improbable (for a given problem, the activation ratio $\alpha$ 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 on the estimation of the exponential rate could be. This point will be made more precise in the next subsection.

As indicated before, Condition (C3) is more complicated to verify. In our implementation, we verify numerically that Condition (C3) holds, by (1) testing if the fixed point is locally stable, and (2) 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 (8). The dynamical system is locally stable if

[3] The code and parameters to reproduce all experiments and figures of the paper are available in a Git repository https://gitlab.inria.fr/phdchenyan/code_ap2021.

$\mathbf{K}_{s(m^*)}$ is a stable matrix, *i.e.* if the norm of all[4] eigenvalues of $\mathbf{K}_{s(m^*)}$ is less than 1. 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 | 0.1% | 0.01% | $10^{-3}$% | $10^{-4}$% | 0 |

**Table 1**: Number of randomly generated instances that violate any of the conditions of Theorem 3 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 $\Delta^d$. We then count the number of instances that violate conditions (C1) or (C4), the results are reported in Table 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 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 hard to find if we generate the parameters uniformly.

## 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 1 with varying values of $\alpha$ 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 $\alpha$. Numerically, these fixed points are global attractors for two reasons:

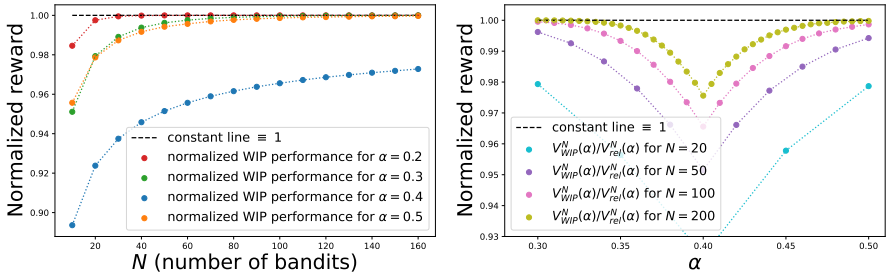
- All matrices $\mathbf{K}_i$ are locally stable because the eigenvalues of $\mathbf{K}_2$ are $\{1, -0.4\ldots, 0.08\ldots\}$ [5] while $\mathbf{K}_1 = \mathbf{P}^0$ and $\mathbf{K}_3 = \mathbf{P}^1$ are always stable matrices.

[4] Recall that $\phi$ is an application from $\Delta^d$ to $\Delta^d$. 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.

[5] In what follows, we write $-0.4\ldots$ to mean a number that approaches $-0.4$.

- For all tested values of $\alpha$, 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 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 are satisfied when $\alpha \neq 0.4$. As $V_{\text{rel}}^{(N)}(\alpha)$ depends on the value of $\alpha$, 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 2a, we choose four values of $\alpha$ 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 2b, this time we fix the value of $N$ and plot the normalized performance as a function of $\alpha$ where $\alpha$ 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 $\alpha N$ 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 $\mathbf{m}^*$.



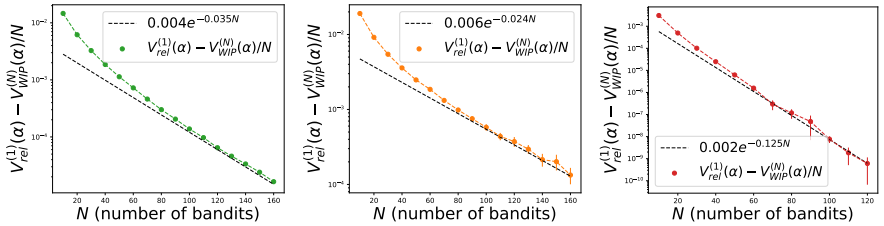

(a) Performance as a function of $N$. (b) Performance as a function of $\alpha$.

**Fig. 2**: Normalized performance of WIP for different values of $\alpha$ and $N$.

To examine more closely the convergence rate, let us consider the quantity

$$\text{subgap}(N) := V_{\text{rel}}^{(1)}(\alpha) - \frac{V_{\text{WIP}}^{(N)}(\alpha)}{N}. \tag{10}$$

Theorem 3 implies that 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 3, we plot in log-scale the subgap (10) as a function of $N$ for the same model as in Figure 2 and $\alpha = 0.2, 0.3$ and $0.5$. For each value of $\alpha$, we also plot the best-fit $b' \cdot e^{-c'N}$ which is a straight line in log-scale. The constant $c$ 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 subgap$(N) \cdot \sqrt{N}$, the curve behaves like a constant. Moreover, this constant behavior is lost [6] as soon as we plot subgap$(N) \cdot N^{\beta}$, with a power $\beta = 0.49$ or $\beta = 0.51$. This gives numerical evidence for an $\mathcal{O}(1/\sqrt{N})$ convergence rate in this singular case, same as for the example given in Remark (2). Actually, we believe that the convergence rate is $\mathcal{O}(1/\sqrt{N})$ for all singular global attractor situations, but a proof of this claim is still open to us.




(a) $c \approx 0.032$ for $\alpha = 0.3$. (b) $c \approx 0.024$ for $\alpha = 0.5$. (c) $c \approx 0.125$ for $\alpha = 0.2$.

**Fig. 3**: Estimation of the constants $c$ and $b$ from Theorem 3.

## 4.3 Non integer values of $\alpha N$

Our previous analysis rely on the assumption that $\alpha N$ is an integer. Let us briefly discuss in this subsection how to deal with non integer values of $\alpha N$ for the optimization problem (2) under (3). Consider the following three possible rounding procedure to replace the constraint (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 $V^{(N)}_{\text{WIP}}(\lfloor N\alpha \rfloor / N)$, $V^{(N)}_{\text{WIP}}(\lceil N\alpha \rceil / N)$ and $V^{(N)}_{\text{WIP}}(\overline{\alpha})$ the reward of WIP under these three solutions. Note that these three values all coincide with our previous $V^{(N)}_{\text{WIP}}(\alpha)$ if $\alpha N$ 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

[6] We refer to our Git repository for a more thorough numerical exploration of this case.

$\mathcal{O}(1)$ from the relaxation $V_{\text{rel}}^{(N)}(\alpha)$. Moreover, $V_{\text{WIP}}^{(N)}(\overline{\alpha})$ converges at exponential rate to $V_{\text{rel}}^{(N)}(\alpha)$. Here is an informal explanation: Let $\phi_{rounding}(\mathbf{m}) = \mathbb{E}\left[M_i^{(N)}(t+1) \mid \mathbf{M}^{(N)}(t) = \mathbf{m}\right]$ 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 (8) of the proof of Lemma 2. In contrast, $\phi_{floor}(\mathbf{m}) = \phi(\mathbf{m}) + \mathcal{O}(\alpha - \lfloor \alpha N \rfloor / N)$. This shows that if the map $\phi$ 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 $\mathcal{O}(1/N)$ of $\mathbf{m}^*$ for *floor* or *ceil*. Moreover, the proof of Lemma 10 and Lemma 11 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 $|V_{\text{WIP}}^{(N)}(\overline{\alpha}) - V_{\text{rel}}^{(N)}(\alpha)| \leq b \cdot e^{-CN}$, whereas $|V_{\text{WIP}}^{(N)}(\lfloor N\alpha \rfloor / N) - V_{\text{rel}}^{(N)}(\alpha)| = \mathcal{O}(1)$.

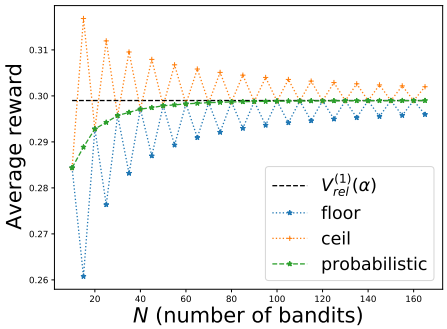



**Fig. 4**: Performance of the three policies for non integer values of $\alpha N$.

To further illustrate these points, we consider in Figure 4 the same example as in Section 4.2, with $\alpha = 0.3$. As in Figure 2, the green curve represents $V_{\text{WIP}}^{(N)}(\alpha)/N$ 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 $V_{\text{WIP}}^{(N)}(\lfloor N\alpha \rfloor / N)/N$, $V_{\text{WIP}}^{(N)}(\overline{\alpha})/N$ and $V_{\text{WIP}}^{(N)}(\lceil N\alpha \rceil / N)/N$ are plotted respectively in blue, green and red dots for $N \in \{25, 35, 45, \dots\}$, 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 $V_{\text{rel}}^{(1)}(\alpha) - V_{\text{WIP}}^{(N)}(\lfloor N\alpha \rfloor / N)$ and $V_{\text{rel}}^{(1)}(\alpha) - V_{\text{WIP}}^{(N)}(\lceil N\alpha \rceil / N)$ converge to $\pm 0.5 \cdot (R_1^1 - R_1^0)$ 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 $\alpha N$ is an integer or not. This indicates that $V_{\text{WIP}}^{(N)}(\overline{\alpha})$ indeed converges at exponential rate to $V_{\text{rel}}^{(N)}(\alpha)$.

# 5 Application: Markovian Fading Channels

The Markovian fading channel is a typical synchronous 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 (2012) a two-classes channel problem has been studied. By using the same scaling as here, the authors of Ouyang et al (2012) 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.

Let us first briefly review this two-class channel model (more details can be found in Ouyang et al (2012)). 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 $\beta$ of them are from class 1. Due to limited resource, each time we can only activate a proportion $\alpha$ 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

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

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.

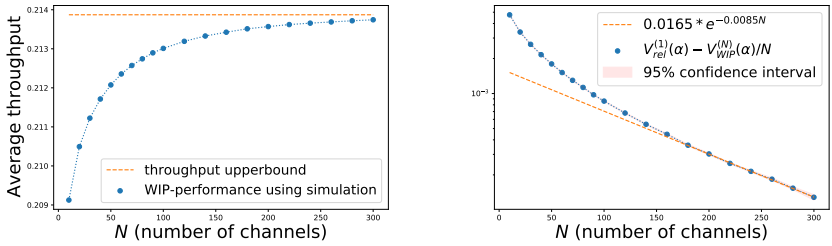
We evaluate the performance by the throughput of the system, hence we obtain a reward of 1 each time we activate a channel *and* it is in good state. Under the MDP framework, this is equivalent to assuming that state $b^k_{s,t}$ gives a reward $b^k_{s,t}$ under activation. It is shown in Ouyang et al (2012) 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 $\nu\big(b^k_{s,t}\big)$ 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, 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 $V^{(N)}_{\text{WIP}}(\alpha)$ and $V^{(N)}_{\text{rel}}(\alpha)$ will depend on the initial condition of the system, *i.e.* on the fraction $\beta$ of arms that are in class 1. Apart from that, our results apply mutatis mutandis to this case.




(a) WIP for two-classes channel model. (b) The exponential convergence.

**Fig. 5**: Convergence rate for two-classes channel model.

We can now provide some numerical results. We shall choose a parameter set that is used in Ouyang et al (2012): $\beta = 0.6$, $\alpha = 0.3$, $(p_1, r_1) = (0.75, 0.2)$,

$(p_2, r_2) = (0.8, 0.3)$. 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 $\mathbf{m}^*$ 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 $0.89\ldots \neq 1$, the fixed point is *not* singular.

Consequently, all conditions needed for Theorem 3 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 5 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 $c \approx 0.0085$.

# 6 The Continuous-Time Restless Bandit Model

Throughout the paper, we studied a *discrete-time* restless bandit problem in which all arms synchronously make a transition. In this section, we explain how to adapt the proofs done in Section 3 to the continuous-time model studied for example in Weber and Weiss (1990). We start by recalling the model of Weber and Weiss (1990) in Section 6.1. We show how the discrete-time and continuous-time models are related in Section 6.2. Finally, we state the equivalence of our main Theorem 3 in Section 6.3 for the continuous model.

## 6.1 The continuous-time bandit model

Similarly to Section 2.1, a continuous-time restless bandit problem with parameters $\left\{(\mathbf{Q}^0, \mathbf{Q}^1, \mathbf{R}^0, \mathbf{R}^1);\ \alpha, N\right\}$ is a Markov decision process defined as follows:

1. As before, the model is composed of $N$ arms each evolves in a finite state space. The state space of the process at time $t \in \mathbb{R}^+$ is the vector $\mathbf{S}(t)$.
2. In continuous time, the decision maker chooses an action $\mathbf{a}(t) \in \{0,1\}^N$. Decisions can be modified only when the process $\mathbf{S}(t)$ changes state: At each jump of the process $\mathbf{S}(t)$, the decision maker observes $\mathbf{S}(t)$ and chooses a new action vector $\mathbf{a}(t)$ that will be kept until the next jump of the process. The action vector must satisfy $\sum_{n=1}^N a_n(t) = \alpha N$ for all $t \in \mathbb{R}^+$.
3. The matrices $\mathbf{Q}^0$ and $\mathbf{Q}^1$ are $d \times d$ matrices such that $Q^a_{ij} \geq 0$ for $i \neq j$ and $\sum_j Q^a_{ij} = 0$. For two states $i \neq j$, if arm $n$ is in state $i$ at time $t$, it

jumps[7] to state $j$ at rate $Q_{ij}^{a_n(t)}$. Given $\mathbf{a}(t)$, the evolutions of the $N$ arms are independent.

4. The gain per unit time of the decision maker is $\sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)}$.

As before, the goal of the decision maker is to compute a decision rule in order to maximize the long-term average reward. Using our notation, this problem can be written as

$$V_{\text{opt}}^{(N)}(\alpha) := \max_{\Pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\int_{t=0}^{T} \sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)} dt\Big] \tag{11}$$

$$\text{subject to} \sum_{n=1}^{N} a_n(t) = \alpha N, \text{ for all } t \in \mathbb{R}_+. \tag{12}$$

This problem is the continuous-time version of the discrete-time problem (2)–(3). As before, we assume that the matrices $\mathbf{Q}^0$ and $\mathbf{Q}^1$ are such that bandit is recurrent regardless of the policy employed.

## 6.2 Whittle index, relaxation and equivalence with the discrete-time model

In this subsection, we recall briefly the definition of Whittle index and of the relaxation for the continuous-time bandit model. These definitions coincide with the ones of Weber and Weiss (1990).

As for the discrete-time case, Whittle index of continuous-time bandits is defined by considering a subsidized MDP for a single arm $n$, in which a decision maker that takes the passive action $a_n(t) = 0$ earns an extra reward $\nu$ per unit time. The definition of indexability is the same as the one in discrete-time case and the index of a state $i$, denoted by $\nu_i$, is the smallest subsidy such that the passive action is optimal for state $i$.

Similarly to the discrete-time problem, the definition of Whittle index in the continuous-time model can be justified by looking at the Lagrangian of the optimization problem (11) where the constraint (12) is replaced by the constraint (14) below. We again denote the value of this relaxed problem as $V_{\text{rel}}^{(N)}(\alpha)$.

$$V_{\text{rel}}^{(N)}(\alpha) := \max_{\Pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\int_{t=0}^{T} \sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)} dt\Big] \tag{13}$$

$$\text{subject to} \lim_{T\to\infty} \int_{0}^{T} \sum_{n=1}^{N} \frac{a_n(t)}{T} dt = \alpha N. \tag{14}$$

As we show below, when considering arms in isolation, using a discrete-time or a continuous-time model is equivalent via a standard uniformization

[7] The rates $Q_{ij}^{a^n(t)}$ are well defined because bandits evolve independently and that the probability that two arms evolve at the same time is 0.

scaling. In particular, neither the definition of Whittle index nor the value of the relaxation depend on the synchronization nature of the bandit.

**Definition 6** Let $(\mathbf{Q}^0, \mathbf{Q}^1, \mathbf{R}^0, \mathbf{R}^1)$ be the parameters of a continuous-time bandit. By a standard uniformization scaling, let $\tau := \max_i \max_a |Q^a_{ii}|$, and let $(\mathbf{P}^0, \mathbf{P}^1, \tilde{\mathbf{R}}^0, \tilde{\mathbf{R}}^1)$ be the matrices defined as follows: for all states $i \neq j$ and all action $a \in \{0, 1\}$:

$$P^a_{ij} := \frac{1}{\tau} Q^a_{ij}; \qquad P^a_{ii} := 1 - \sum_{k \neq i} P^a_{ik}; \qquad \tilde{R}^a_i := \tau R^a_i. \tag{15}$$

We call $(\mathbf{P}^0, \mathbf{P}^1, \tilde{\mathbf{R}}^0, \tilde{\mathbf{R}}^1)$ the discrete-time version of our continuous-time bandit model.

The following lemma states the equivalence of Whittle relaxation between the discrete-time and the continuous-time problems:

**Lemma 4** Let $(\mathbf{Q}^0, \mathbf{Q}^1, \mathbf{R}^0, \mathbf{R}^1)$ be a continuous-time bandit and let $(\mathbf{P}^0, \mathbf{P}^1, \tilde{\mathbf{R}}^0, \tilde{\mathbf{R}}^1)$ be its discrete-time version (15). Then:

(i) The matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ are probability matrices.

(ii) The discrete-time bandit $(\mathbf{P}^0, \mathbf{P}^1, \tilde{\mathbf{R}}^0, \tilde{\mathbf{R}}^1)$ is indexable if and only if the continuous-time bandit $(\mathbf{Q}^0, \mathbf{Q}^1, \mathbf{R}^0, \mathbf{R}^1)$ is indexable. In such a case, the indices of both bandits coincide.

(iii) The discrete-time relaxed optimization problem (4) has the same value as its continuous-time counterpart (13).

The proof of Lemma 4 is a direct consequence of uniformization: the results rely on analysis of an arm in isolation; when focus on a single arm, Bellman's equation is identical for the discrete-time and continuous-time version of the MDP.

## 6.3 Exponential convergence in the case of continuous-time model

Lemma 4 uses the fact that the Whittle relaxation is defined for arm in isolation. Hence, considering discrete-time or continuous-time bandits is equivalent. For the $N$ arms model, however, the situation is different: in the discrete-time model of Section 2 all arms change states synchronously, while in continuous-time situation, the probability that two arms make a jump at the exact same time is 0. This implies that the reward of WIP for the $N$ arms problem does depend on whether the model is synchronous or not. We denote the later by $V^{(N)}_{\mathrm{WIP-async.}}(\alpha)$.

It is shown in Weber and Weiss (1990) that the asymptotic optimality depends on the ergodic property of the solution of an ordinary differential

equation (ODE) defined in Equation (10) of Weber and Weiss (1990). Using our notation, this differential equation can be written as

$$\dot{\mathbf{m}} = \tau(\phi(\mathbf{m}) - \mathbf{m}), \tag{16}$$

where $\phi$ is defined as in Lemma 2 for a discrete-time bandit problem $(\mathbf{P}^0, \mathbf{P}^1, \tilde{\mathbf{R}}^0, \tilde{\mathbf{R}}^1)$. By applying Lemma 2 on the discrete-time problem, we see that this equation has a unique fixed point, $\mathbf{m}^*$. It is then shown in Weber and Weiss (1990) that if all the solutions of the differential equation (16) converge to $\mathbf{m}^*$, then $\lim_{N\to\infty} V^{(N)}_{\mathrm{WIP-async.}}(\alpha)/N \to V^{(1)}_{\mathrm{rel}}(\alpha)$. In the next theorem, we show that we can adapt the result of Theorem 3 to the continuous-time model.

**Theorem 5** (Exponential convergence rate theorem: continuous-time case) *Consider a continuous-time recurrent restless bandit problem* $\{(\mathbf{Q}^0, \mathbf{Q}^1, \mathbf{R}^0, \mathbf{R}^1);\ \alpha\}$ *such that:*

*(i) The bandit problem is indexable.*

*(ii) The (unique) fixed point* $\mathbf{m}^*$ *of the ODE* $\dot{\mathbf{m}} = \tau(\phi(\mathbf{m}) - \mathbf{m})$ *is not singular.*

*(iii)* $\mathbf{m}^*$ *is a global attractor of the trajectories of the ODE.*

*(iv)* $\mathbf{m}^*$ *is locally stable.*

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

$$0 \le V^{(N)}_{\mathrm{rel}}(\alpha) - V^{(N)}_{\mathrm{WIP-async.}}(\alpha) \le b \cdot e^{-cN}.$$

*Sketch of proof* The proof of this result follows the same structure as the proof of Theorem 3 but needs substantial adaptation, the full details are given in Appendix D. The main ingredients are:

- We first use a result from Darling and Norris (2008) to obtain an analogue of Hoeffding's inequality. This proves that the behavior of the $N$ arms model is close to the dynamic of the ODE (16).
- Using this and the fact that $\mathbf{m}^*$ is non-singular, we show that the stochastic system lies with high probability in a neighborhood $\mathcal{N}$ of $\mathbf{m}^*$ where $\phi$ is affine. We again use Stein's method to obtain the exponential convergence result (but this time applied to a continuous-time process).

□

This theorem is a refinement of the original asymptotic optimality result of (Weber and Weiss, 1990, Theorem 2), as it provides a bound on the rate of convergence for the performance of WIP to the optimal one. The applicability conditions are essentially similar: (Weber and Weiss, 1990, Theorem 2) also needs the assumption that $\mathbf{m}^*$ is an attractor of the ODE. We add in addition that $\mathbf{m}^*$ is locally stable and that $\mathbf{m}^*$ is not singular. Those conditions are also similar to the conditions of Theorem 3. However, we should point out that the behavior of the discrete-time dynamical system $\mathbf{m}(t+1) = \phi(\mathbf{m}(t))$

can be quite different from its continuous-time counterpart $\dot{\mathbf{m}} = \tau\big(\phi(\mathbf{m}) - \mathbf{m}\big)$: there are bandit models for which WIP is asymptotically optimal under a continuous-time model but is not for the discrete-time model. One such example is displayed in our Git repository.

# 7 Conclusion and Future Work

In this paper, we studied the performance of Whittle index policy (WIP) when there is a large number of arms. We showed that, when WIP becomes asymptotically optimal, it does so at exponential rate (unless the fixed point is singular, which barely occurs). This explains why WIP is very efficient in practice, even when the number of arms remains moderate. Our results hold for the classical model of Weber and Weiss (1990) where arms evolve asynchronously in continuous time, as well as for a synchronous discrete-time model in which all arms make their transitions simultaneously.

As for future research, we plan on investigating more closely the singular situations, as well as extending the exponential convergence rate result to those generalizations of Whittle index as in Duran and Verloop (2018); Hodge and Glazebrook (2015); Verloop (2016).

# Acknowledgements

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

# References

Aalto S, Lassila P, Osti P (2015) Whittle index approach to size-aware scheduling with time-varying channels. In: Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pp 57–69

Ansell P, Glazebrook KD, Nino-Mora J, et al (2003) Whittle's index policy for a multi-class queueing system with convex holding costs. Mathematical Methods of Operations Research 57(1):21–39

Avrachenkov KE, Borkar VS (2016) Whittle index policy for crawling ephemeral content. IEEE Transactions on Control of Network Systems 5(1):446–455

Blondel VD, Bournez O, Koiran P, et al (2001) The stability of saturated linear dynamical systems is undecidable. Journal of Computer and System Sciences 62(3):442–462

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

Darling R, Norris J (2008) Differential equation approximations for markov chains. Probab Surveys 5:37–79

Duff MO (1995) Q-learning for bandit problems. In: Proceedings of the Twelfth International Conference on International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML'95, pp 209–217

Duran S, Verloop M (2018) Asymptotic optimal control of markov-modulated restless bandits. In: International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS 2018), vol 2. ACM : Association for Computing Machinery, Irvine, US, pp 7:1–7:25

Gast N (2017) Expected Values Estimated via Mean-Field Approximation are 1/N-Accurate. In: ACM SIGMETRICS/ International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS '17, Urbana-Champaign, United States, p 26

Gast N, Van Houdt B (2017) A Refined Mean Field Approximation. Proceedings of the ACM on Measurement and Analysis of Computing Systems 1(28)

Gast N, Bortolussi L, Tribastone M (2018a) Size Expansions of Mean Field Approximation: Transient and Steady-State Analysis. In: 2018 - 36th International Symposium on Computer Performance, Modeling, Measurements and Evaluation, Toulouse, France, pp 1–2

Gast N, Latella D, Massink M (2018b) A refined mean field approximation of synchronous discrete-time population models. Performance evaluation 126:1–21

Gast N, Gaujal B, Khun K (2022a) Computing whittle (and gittins) index in subcubic time. arXiv preprint arXiv:220305207

Gast N, Gaujal B, Yan C (2022b) Lp-based policies for restless bandits: necessary and sufficient conditions for (exponentially fast) asymptotic optimality

Gittins J, Glazebrook K, Weber R (2011) Multi-armed bandit allocation indices. John Wiley & Sons

Gittins JC (1979) Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society, Series B pp 148–177

Hodge DJ, Glazebrook KD (2015) On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. Adv in Appl Probab 47(3):652–667

Hu W, Frazier P (2017) An asymptotically optimal index policy for finite-horizon restless bandits. 1707.00205

Kifer Y (1988) Random Perturbations of Dynamical Systems. Progress in Probability, Birkhäuser Boston

Kurtz TG (1978) Strong approximation theorems for density dependent markov chains. Stochastic Processes and their Applications 6(3):223–240

Larranaga 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

Lattimore T, Szepesvári C (2020) Bandit algorithms. Cambridge University Press

Liu K, Zhao Q (2010) Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. IEEE Transactions on Information Theory 56(11):5547–5567

Meshram R, Manjunath D, Gopalan A (2018) On the whittle index for restless multiarmed hidden markov bandits. IEEE Transactions on Automatic Control 63(9):3046–3053

Niño-Mora J, Villar SS (2011) Sensor scheduling for hunting elusive hiding targets via whittle's restless bandit index policy. In: International Conference on NETwork Games, Control and Optimization (NetGCooP 2011), IEEE, pp 1–8

Ouyang W, Eryilmaz A, Shroff NB (2012) Asymptotically optimal downlink scheduling over markovian fading channels. In: 2012 Proceedings IEEE INFOCOM, IEEE, pp 1224–1232

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

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

Raghunathan V, Borkar V, Cao M, et al (2008) Index policies for real-time multicast scheduling for wireless broadcast systems. In: IEEE INFOCOM 2008-The 27th Conference on Computer Communications, IEEE, pp 1570–1578

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

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

Weber RR, Weiss G (1991) Addendum to: On an index policy for restless bandits. Advances in Applied probability 23(2):429–430

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

Ying L (2017) Stein's method for mean field approximations in light and heavy traffic regimes. POMACS 1(1):1–27

Zhang X, Frazier PI (2021) Restless bandits with many arms: Beating the central limit theorem. 2107.11911

Zhang X, Frazier PI (2022) Near-optimality for infinite-horizon restless bandits with many arms. arXiv preprint arXiv:220315853

# A Proof of Theorem 1

*Proof* Let $\mathbf{m}^*$ be the fixed point of $\phi$. As $\mathbf{P}^0$, $\mathbf{P}^1$ are rational, each coordinate of $\mathbf{m}^*$ is a rational number. Let $\{N_k\}_{k\geq 0}$ be a sequence of increasing integers that goes to $\infty$, such that for all $k \geq 0$ and all $1 \leq i \leq d$, $m_i^* N_k$ and $\alpha N_k$ are integers. We then fix an $N$ from this sequence $\{N_k\}_{k\geq 0}$. Recall that $m_i N$ is the number of arms in state $i$ in configuration $\mathbf{m}$ and that $S_n(t)$ is the state of arm $n$ at time $t$. We use $\mathbf{S}(t)$ to denote the state vector of the $N$ arms system at time $t$. Let $\mathbf{S}^*$ be a state vector corresponds to configuration $\mathbf{m}^*$ with $N$ arms. This is possible as $m_i^* N$ is an integer for all $i \in \{1, \ldots, d\}$.

Note that in configuration $\mathbf{m}^*$ (*i.e.* state vector $\mathbf{S}^*$), an optimal action $\mathbf{a}^*$ under the relaxed constraint (5) will activate *exactly* $\alpha N$ arms. As $\mathbf{a}^*$ is sub-optimal compared to an optimal policy for the original $N$ arms problem (2)-(3), we have

$$\begin{aligned} V_{\text{opt}}^{(N)}(\alpha) + h(\mathbf{S}^*) &= \max_{\mathbf{a}\in\{0,1\}^N} \left\{ \sum_{n=1}^{N} R_{S_n^*}^{a_n} + \mathbb{E}_{\mathbf{a}}\Big[V\big(\mathbf{S}(1)\big) \mid \mathbf{S}(0) = \mathbf{S}^*\Big] \right\} \\ &\geq \sum_{n=1}^{N} R_{S_n^*}^{a_n^*} + \mathbb{E}_{\mathbf{a}^*}\Big[V\big(\mathbf{S}(1)\big) \mid \mathbf{S}(0) = \mathbf{S}^*\Big] \\ &= N V_{\text{rel}}^{(1)}(\alpha) + \mathbb{E}_{\mathbf{a}^*}\big[h(\mathbf{S}(1)) \mid \mathbf{S}(0) = \mathbf{S}^*\big], \end{aligned}$$

where in the above equation the function $V : \mathbf{S} \to \mathbb{R}$ is the bias of the MDP. The first line corresponds to Bellman's equation (see e.g. Equation 8.4.2 in Chapter 8 of Puterman (1994)), the second line is because $\mathbf{a}^*$ is a valid action for the $N$-arms MDP but might not be the optimal action, and the last line is because $\sum_{n=1}^{N} R_{S_n^*}^{a_n^*} = V_{\text{rel}}^{(N)}(\alpha) = N V_{\text{rel}}^{(1)}(\alpha)$.

We hence obtain

$$V_{\text{rel}}^{(1)}(\alpha) \geq \frac{V_{\text{opt}}^{(N)}(\alpha)}{N} \geq V_{\text{rel}}^{(1)}(\alpha) + \frac{\mathbb{E}_{\mathbf{a}^*} V\big(\mathbf{S}(1)\big) - h(\mathbf{S}^*)}{N}.$$

In the following, we bound $\mathbb{E}_{\mathbf{a}^*}[V\big(\mathbf{S}(1)\big) - h(\mathbf{S}^*)]$. This will be achieved in two steps.

***Step One***

We define for two state vectors $\mathbf{y}$, $\mathbf{z}$ the distance

$$\delta(\mathbf{y},\mathbf{z}) := \sum_{n=1}^{N} \mathbb{1}_{\{y_n \neq z_n\}},$$

which counts the number (among the $N$ arms) of arms that are in different states between those two vectors. Such distance satisfies the property that for all $\mathbf{y}$ and $\mathbf{z}$ such that $\delta(\mathbf{y},\mathbf{z}) = k$, we can find a sequence of state vectors $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{k-1}$ that verify $\delta(\mathbf{y},\mathbf{z}_1) = \delta(\mathbf{z}_1,\mathbf{z}_2) = ... = \delta(\mathbf{z}_{k-2},\mathbf{z}_{k-1}) = \delta(\mathbf{z}_{k-1},\mathbf{z}) = 1$. In what follows, we show that there exists $C > 0$ independent of $N$ such that for all state vectors $\mathbf{y}$ and $\mathbf{z}$, the bias function $h(.)$ satisfies:

$$|h(\mathbf{y}) - h(\mathbf{z})| \leq C \cdot \delta(\mathbf{y},\mathbf{z}).$$

In view of the above property of $\delta$, we only need to prove this for $\delta(\mathbf{y},\mathbf{z}) = 1$, *i.e.*

$$|h(\mathbf{y}) - h(\mathbf{z})| \leq C.$$

Let $\mathbf{y}$, $\mathbf{z}$ be two state vectors such that $\delta(\mathbf{y},\mathbf{z}) = 1$, and assume without loss of generality that it is arm 1 that are in different states: $y_1 \neq z_1$ and $y_n = z_n$ for $n \in \{2 \dots N\}$. We use a coupling argument as follows: We consider two trajectories of the $N$ arms system, $\mathbf{Y}$ and $\mathbf{Z}$, that start respectively in state vectors $\mathbf{Y}(0) = \mathbf{y}$ and $\mathbf{Z}(0) = \mathbf{z}$. Let $\pi^*$ be the optimal policy of the $N$ arms MDP, and suppose that we apply $\pi^*$ to the trajectory $\mathbf{Z}$. At time $t$, the action vector will be $\pi^*(\mathbf{Z}(t))$. We couple the trajectories $\mathbf{Y}$ and $\mathbf{Z}$ by applying the same action vectors $\pi^*(\mathbf{Z}(t))$ for $\mathbf{Y}$ and keeping $Y_n(t) = Z_n(t)$ for arms $n \in \{2 \dots N\}$. The $\mathbf{Z}$ trajectory follows an optimal trajectory, hence Bellman's equation is satisfied: for any $T > 0$, we have:

$$T \cdot V_{\text{opt}}^{(N)}(\alpha) + h(\mathbf{z}) = \sum_{n=1}^{N} R_{z_n}^{\pi_n^*(\mathbf{z})} + \mathbb{E}_{\pi^*}\left[ \sum_{n=1}^{N} \sum_{t=1}^{T-1} R_{Z_n(t)}^{\pi_n^*(\mathbf{Z}(t))} + V\big(\mathbf{Z}(T)\big) \mid \mathbf{Z}(0) = \mathbf{z} \right]. \tag{17}$$

Since $\mathbf{Y}$ follows a possibly sub-optimal trajectory, we have:

$$T \cdot V_{\text{opt}}^{(N)}(\alpha) + h(\mathbf{y}) \geq \sum_{n=1}^{N} R_{y_n}^{\pi_n^*(\mathbf{y})} + \mathbb{E}_{\pi^*}\left[ \sum_{n=1}^{N} \sum_{t=1}^{T-1} R_{Y_n(t)}^{\pi_n^*(\mathbf{Z}(t))} + V\big(\mathbf{Y}(T)\big) \mid \mathbf{Y}(0) = \mathbf{y} \right], \tag{18}$$

Recall that the matrices $\mathbf{P}^0, \mathbf{P}^1$ are such that a bandit is recurrent and aperiodic. This shows that the mixing time of a single arm is bounded (independently of $N$): for any policy $\pi \in \Pi$

$$\max_{i,j} \operatorname*{argmin}_{t} \left\{ \mathbb{P}_\pi \Big[ Y_1(t) = Z_1(t) \;\Big|\; Y_1(0) = i, Z_1(0) = j \Big] > 0 \right\} < \infty.$$

Because of the coupling, for $0 \leq t \leq T$ and $1 \leq n \leq N$, $Y_n(t) \neq Z_n(t)$ is only possible for $n = 1$. Furthermore, as the mixing time of an arm is bounded, for $T$ large enough, there is a positive probability, say at least $p > 0$, that $Y_1(T) = Z_1(T)$. Hence with probability smaller than $1 - p$ we have $\delta\big(\mathbf{y}(T), \mathbf{z}(T)\big) = 1$, conditional on $\mathbf{Y}(0) = \mathbf{y}$ and $\mathbf{Z}(0) = \mathbf{z}$.

Let $r := 2 \max_{1 \leq i \leq d, a \in \{0,1\}} |R_i^a|$. Subtracting (17) in (18) gives

$$\begin{aligned} |h(\mathbf{y}) - h(\mathbf{z})| &\leq T \cdot r + \left| \mathbb{E}_{\pi^*}\Big[ V\big(\mathbf{Y}(T)\big) - V\big(\mathbf{Z}(T)\big) \;\Big|\; \mathbf{Y}(0) = \mathbf{y}, \mathbf{Z}(0) = \mathbf{z} \Big] \right| \\ &\leq T \cdot r + (1-p) \max_{\mathbf{U},\mathbf{V}:\ \delta(\mathbf{U},\mathbf{V})=1} \big\{ |h(\mathbf{U}) - h(\mathbf{V})| \big\}. \end{aligned}$$

This being true for all $\mathbf{y}, \mathbf{z}$ with $\delta(\mathbf{y},\mathbf{z}) = 1$, it implies that $\max_{\mathbf{U},\mathbf{V}:\ \delta(\mathbf{U},\mathbf{V})=1} \big\{ |h(\mathbf{U}) - h(\mathbf{V})| \big\} \leq T \cdot r/p$, and we can take the constant $C := T \cdot r/p$.

***Step Two***

Recall that the state vector $\mathbf{S}^*$ corresponds to the optimal (relaxed) configuration $\mathbf{m}^*$. We now prove that

$$\mathbb{E}_{\mathbf{a}^*}[\delta(\mathbf{S}^*, \mathbf{S}(1)) \mid \mathbf{S}(0) = \mathbf{S}^*] \leq D\sqrt{N},$$

with a constant $D$ independent of $N$, where $\mathbf{S}(1)$ is the random vector conditional on $\mathbf{S}(0) = \mathbf{S}^*$ under action vector $\mathbf{a}^*$.

Indeed, let $\mathbf{x}^* := \mathbf{m}^* N$, and denote $\mathbf{X} := \mathbf{m}(1)N$ to be the random $d$-vector, with $\mathbf{m}(1)$ the random configuration corresponds to $\mathbf{S}(1)$. For each $1 \leq i \leq d$, we may write

$$X_i = (B^0_{i,1} + B^1_{i,1}) + (B^0_{i,2} + B^1_{i,2}) + ... + (B^0_{i,d} + B^1_{i,d})$$

where $B^a_{i,j} \sim Binomial(x^*_{j,a}, P^a_{ji})$ for $1 \leq j \leq d$, $a \in \{0, 1\}$; and $x^*_{j,0} + x^*_{j,1} = x^*_j$, with $x^*_{j,a}$ representing the number of arms in state $j$ taking action $a$, when optimal action vector $\mathbf{a}^*$ is applied to state vector $\mathbf{S}^*$.

By stationarity, we have

$$\mathbb{E}_{\mathbf{a}^*}(X_i) = \sum_{j=1}^{d} \sum_{a=0,1} x^*_{j,a} \cdot P^a_{ji} = x^*_i,$$

and

$$\mathrm{Var}(X_i) = \sum_{j=1}^{d} \sum_{a=0,1} x^*_{j,a} \cdot P^a_{ji}(1 - P^a_{ji}) = \mathcal{O}(N).$$

Consequently, we can bound

$$\mathbb{E}_{\mathbf{a}^*}[\delta(\mathbf{S}^*, \mathbf{S}(1))] \leq \sum_{i=1}^{d} \mathbb{E}_{\mathbf{a}^*} \left| x^*_i - X_i \right| \leq D\sqrt{N},$$

with a constant $D$ independent of $N$.

To summarize, we have

$$\mathbb{E}_{\mathbf{a}^*}\left[|h(\mathbf{S}(1)) - h(\mathbf{S}^*)|\right] \leq \mathbb{E}_{\mathbf{a}^*}\left[C \cdot \delta\big(\mathbf{S}(1), \mathbf{S}^*\big)\right] \leq CD \cdot \sqrt{N},$$

hence

$$V^{(1)}_{\mathrm{rel}}(\alpha) \geq \frac{V^{(N)}_{\mathrm{opt}}(\alpha)}{N} = V^{(1)}_{\mathrm{rel}}(\alpha) + \frac{\mathbb{E}_{\mathbf{a}^*} h(\mathbf{S}(1)) - h(\mathbf{S}^*)}{N} \geq V^{(1)}_{\mathrm{rel}}(\alpha) - \frac{CD}{\sqrt{N}}, \tag{19}$$

which implies that $V^{(N)}_{\mathrm{opt}}(\alpha)/N \to V^{(1)}_{\mathrm{rel}}(\alpha)$ when $N$ goes to $+\infty$. Moreover, from (19), the convergence rate is at least as fast as $\mathcal{O}(1/\sqrt{N})$. □

# B Proof of Lemma 2

In this appendix we prove Lemma 2. We first show the piecewise affine property in Lemma 6, which gives (i) and (ii). We then show the uniqueness of fixed point from a bijective property in Lemma 7, from which we conclude (iii).

**Lemma 6** (Piecewise affine) $\phi$ is a piecewise affine continuous function, with $d$ affine pieces.

*Proof* Let $\mathbf{m} \in \Delta^d$ be a configuration and recall $s(\mathbf{m}) \in \{1, \dots, d\}$ is the state such that $\sum_{i=1}^{s(\mathbf{m})-1} m_i \leq \alpha < \sum_{i=1}^{s(\mathbf{m})} m_i$. 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

$$\sum_{i=1}^{s(\mathbf{m})-1} Nm_i P^1_{ij} + N(\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i) P^1_{s(\mathbf{m})j} + N(\sum_{i=1}^{s(\mathbf{m})} m_i - \alpha) P^0_{s(\mathbf{m})j} + \sum_{i=s(\mathbf{m})+1}^{d} Nm_i P^0_{ij}. \tag{20}$$

It justifies the expression (8). Note that (8) 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

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

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

□

**Lemma 7** (Bijectivity) 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 $\theta$. 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, 1990, Lemma 1). For a given $\nu \in \mathbb{R}$, denote by $\gamma(\nu)$ the value of the subsidy-$\nu$ problem, *i.e.*

$$\gamma(\nu) := \max_{\pi \in \Pi} \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\Big[ \sum_{t=0}^{T-1} \Big( R^{\pi(S(t))}_{S(t)} + \nu\big(1 - \pi(S(t))\big)\Big)\Big]. \tag{21}$$

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

By definition of indexability, $\gamma(\nu) = \max_{\pi\in\Pi} \gamma_\pi(\nu)$ is a piecewise affine, continuous and convex function of $\nu$: 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 (21) 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

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

Observe that $\tilde{\alpha}(s,\theta)$ is the ratio of two affine functions of $\theta$, hence is monotone as $\theta$ 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 2(iii). Let $\mathbf{m}$ be a fixed point of the continuous map $\phi$ (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 $\phi$, 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 $\phi$ is unique.

# C Proof of Theorem 3

In this appendix, we explain technical details of the proof of our main result Theorem 3. In the following, we denote by $\mathcal{B}(\mathbf{m}^*, r)$ the ball centered at $\mathbf{m}^*$ with radius $r$.

**Theorem 8** *Under the same assumptions as in Theorem 3, 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 8 implies Theorem 3. To show this, we first prove that:

**Lemma 9** 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) $\rho$ is piecewise affine on each of the zone $\mathcal{Z}_i$ and for all $\mathbf{m} \in \Delta^d$:

$$\rho(\mathbf{m}) = \sum_{i=1}^{s(\mathbf{m})-1} m_i R_i^1 + (\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i) R_{s(\mathbf{m})}^1 + (\sum_{i=1}^{s(\mathbf{m})} m_i - \alpha) R_{s(\mathbf{m})}^0 + \sum_{i=s(\mathbf{m})+1}^{d} m_i R_i^0. \tag{22}$$

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

*Proof* Let $\mathbf{m} \in \Delta^d$ be a configuration and recall $s(\mathbf{m}) \in \{1, \ldots, d\}$ is the state such that $\sum_{i=1}^{s(\mathbf{m})-1} m_i \leq \alpha < \sum_{i=1}^{s(\mathbf{m})} m_i$. Similarly to our analysis of Lemma 6, 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 (22).

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:

$$\begin{aligned} L &:= \max_{\Pi} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[R_{S_n(t)}^{a_n(t)} + (\alpha - a_n(t))\nu_{s(\mathbf{m}^*)}\right] \\ &= \max_{\Pi} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[R_{S_n(t)}^{a_n(t)}\right] + \left(\alpha - \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[a_n(t)\right]\right) \nu_{s(\mathbf{m}^*)}. \end{aligned} \tag{23}$$

By definition of Whittle index, any policy of the form $\pi(s(\mathbf{m}^*), \theta)$ defined in Lemma 7 is optimal for (23). Moreover, if $\theta^*$ is such that $\tilde{\alpha}(s(\mathbf{m}^*), \theta^*) = \alpha$, then such a policy satisfies the constraint (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 (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 $\alpha N$ arms are activated on average. Indeed, the arm-averaged reward of this policy is:

$$L = \sum_{i=1}^{s(\mathbf{m}^*)-1} m_i^* R_i^1 + \theta^* m_{s(\mathbf{m}^*)}^* R_{s(\mathbf{m}^*)}^1 + (1-\theta^*) m_{s(\mathbf{m}^*)}^* R_{s(\mathbf{m}^*)}^0 + \sum_{i=s(\mathbf{m}^*)+1}^{d} m_i^* R_i^0 \tag{24}$$

As the proportion of activated arms is $\alpha$, we have $\sum_{i=1}^{s(\mathbf{m}^*)-1} m_i^* + \theta^* m_{s(\mathbf{m}^*)}^* = \alpha$. Hence (24) coincides with the expression of $\rho(\mathbf{m}^*)$ in (22), and $\rho(\mathbf{m}^*) = L = V_{\text{rel}}^{(1)}(\alpha)$. This concludes the proof of Lemma 9.

□

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 9 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 $\rho$ and Theorem 8(i), the first term inside the above expectation is exponentially small; by Theorem 8(ii) and since the rewards are bounded, the second term is also exponentially small.

In the rest of the section, we first prove a few technical lemma, and conclude by proving Theorem 8.

## C.1 Hoeffding's inequality (for one transition)

**Lemma 10** (Hoeffding's inequality) For all $t \in \mathbb{N}$, we have

$$\mathbf{M}^{(N)}(t+1) = \phi\big(\mathbf{M}^{(N)}(t)\big) + \boldsymbol{\epsilon}^{(N)}(t+1)$$

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

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

and for all $\delta > 0$:

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

*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

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

and

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

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

## C.2 Hoeffding's inequality (for $t$ transitions)

**Lemma 11** There exists a positive constant $K$ such that for all $t \in \mathbb{N}$ and for all $\delta > 0$,

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

*Proof* Since $\phi$ is a piecewise affine function with finite affine pieces, in particular $\phi$ 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

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

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 10: for all $\delta > 0$,

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

Hence, using the union bound, we obtain

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

and this ends the proof of Lemma 11. □

## C.3 Exponential stability of $\mathbf{m}^*$

**Lemma 12** Under the assumptions of Theorem 3, there exists constants $b_1, b_2 > 0$ such that for all $t \geq 0$ and all $\mathbf{m} \in \Delta^d$:

$$\|\Phi_t(\mathbf{m}) - \mathbf{m}^*\| \leq b_1 \cdot e^{-b_2 t} \cdot \|\mathbf{m} - \mathbf{m}^*\|. \tag{25}$$

*Proof* As $\phi$ 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 $\Delta^d$ 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 $\varepsilon$ and $\delta$ from the local stability of $\phi$, 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 (25) holds for all $\mathbf{m} \in \Delta^d$. □

## C.4 Proof of Theorem 8

We are now ready to prove the main theorem.

*Proof* The proof consists of several parts.

### C.4.1 Choice of a neighborhood $\mathcal{N}$

The fixed point $\mathbf{m}^*$ is in zone $\mathcal{Z}_{s(\mathbf{m}^*)}$ in which $\phi$ 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 $\phi$ remains inside $\mathcal{N}_2$ and it is at least to a distance $\delta$ 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 $\delta$ is independent of $N$. From $(ii)$ of Lemma 12, we denote furthermore by $\tilde{T} := T(r/2)$ the finite time such that for all $\mathbf{m} \in \Delta^d$, $\Phi_{\tilde{T}+1}(\mathbf{m}) \in \mathcal{B}(\mathbf{m}^*, r/2)$.

### C.4.2 Definition and properties of the function $G$.

Following the generator approach used for instance in Gast et al (2018b). 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 12, 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:

$$\Phi_t(\mathbf{m}) = (\mathbf{m} - \mathbf{m}^*) \cdot \mathbf{K}_{s(\mathbf{m}^*)}^t + \mathbf{m}^*. \tag{26}$$

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

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

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

$$\begin{aligned} \mathbb{E}[\mathbf{M}^{(N)}(0)] - \mathbf{m}^* &= \mathbb{E}\left[G(\mathbf{M}^{(N)}(0)) - G(\phi(\mathbf{M}^{(N)}(0)))\right] \quad \text{(By the above equality)} \\ &= \mathbb{E}\left[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{M}^{(N)}(0)))\right] \quad \text{(Since } \mathbf{M}^{(N)}(0) \text{ is stationary)} \\ &= \mathbb{E}\Big[\mathbb{E}\left[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{m})) \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\right] \cdot \mathbb{1}_{\{\mathbf{m}\notin\mathcal{N}\}} \quad (27) \\ &\quad + \mathbb{E}\left[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{m})) \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\right] \cdot \mathbb{1}_{\{\mathbf{m}\in\mathcal{N}\}}\Big]. \quad (28) \end{aligned}$$

In the following, we bound (27) and (28) separately.

### C.4.3 Bound on (27)

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

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

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 $\phi$. We have by Lemma 11:

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

This shows that

$$\begin{aligned}
\mathbb{P}\left[\mathbf{M}^{(N)}(0) \notin \mathcal{N}\right] &= \mathbb{P}\left[\|\mathbf{M}^{(N)}(0) - \mathbf{m}^*\| \geq r\right] \\
&= \mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1) - \mathbf{m}^*\| \geq r\right] \quad \text{(By stationarity)} \\
&\leq \mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1) - \Phi_{\tilde{T}+1}(\mathbf{M}^{(N)}(0))\| \geq \frac{r}{2}\right] \\
&\quad + \mathbb{P}\left[\|\Phi_{\tilde{T}+1}(\mathbf{M}^{(N)}(0)) - \mathbf{m}^*\| \geq \frac{r}{2}\right] \\
&= \mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1) - \Phi_{\tilde{T}+1}(\mathbf{M}^{(N)}(0))\| \geq \frac{r}{2}\right] \\
&\leq (\tilde{T}+1)\cdot e^{-2uN}, \qquad (29)
\end{aligned}$$

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

### C.4.4 Bound on (28)

By Lemma 10, we have

$$\begin{aligned}
&\mathbb{E}\big[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{m})) \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\big]\cdot \mathbb{1}_{\{\mathbf{m}\in\mathcal{N}\}} \\
&= \mathbb{E}\big[G(\phi(\mathbf{m}) + \boldsymbol{\epsilon}^{(N)}(1)) - G(\phi(\mathbf{m})) \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\big]\cdot \mathbb{1}_{\{\mathbf{m}\in\mathcal{N}\}} \\
&= \mathbb{E}\Big[\big(G(\phi(\mathbf{m}) + \boldsymbol{\epsilon}^{(N)}(1)) - G(\phi(\mathbf{m}))\big)\cdot \mathbb{1}_{\{\|\boldsymbol{\epsilon}^{(N)}(1)\|<\delta\}} \\
&\quad + \big(G(\phi(\mathbf{m}) + \boldsymbol{\epsilon}^{(N)}(1)) - G(\phi(\mathbf{m}))\big)\cdot \mathbb{1}_{\{\|\boldsymbol{\epsilon}^{(N)}(1)\|\geq\delta\}} \;\Big|\; \mathbf{M}^{(N)}(0) = \mathbf{m}\Big]\cdot \mathbb{1}_{\{\mathbf{m}\in\mathcal{N}\}}
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\Big[\big(G(\phi(\mathbf{m}) + \boldsymbol{\epsilon}^{(N)}(1)) - G(\phi(\mathbf{m}))\big)\cdot \mathbb{1}_{\{\|\boldsymbol{\epsilon}^{(N)}(1)\|<\delta\}} \;\Big|\; \mathbf{M}^{(N)}(0) = \mathbf{m}\Big]\cdot \mathbb{1}_{\{\mathbf{m}\in\mathcal{N}\}} \\
&= \Big[G\big(\mathbb{E}\big[\phi(\mathbf{m}) + \boldsymbol{\epsilon}^{(N)}(1) \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\big]\big) - G\big(\mathbb{E}\big[\phi(\mathbf{m}) \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\big]\big)\Big]\mathbb{P}\big(\{\|\boldsymbol{\epsilon}^{(N)}(1)\| < \delta\}\big)\cdot \\
&\quad \big(\text{Thanks to the affinity of } G \text{ in this case, we can interchange } \mathbb{E} \text{ and } G\big) \\
&= 0 \qquad \big(\text{By Lemma 10}\big).
\end{aligned}$$

For the second part of the above expectation,

$$\begin{aligned}
&\left|\left|\mathbb{E}\Big[\big(G(\phi(\mathbf{m}) + \boldsymbol{\epsilon}^{(N)}(1)) - G(\phi(\mathbf{m}))\big)\cdot \mathbb{1}_{\{\|\boldsymbol{\epsilon}^{(N)}(1)\|\geq\delta\}}\Big|\mathbf{M}^{(N)}(0) = \mathbf{m}\Big]\right|\right|\cdot \mathbb{1}_{\{\mathbf{m}\in\mathcal{N}\}} \\
&\leq 2\overline{G}\cdot \mathbb{P}\big(\|\boldsymbol{\epsilon}^{(N)}(1)\| \geq \delta\big) \\
&\leq 2\overline{G}\cdot e^{-2N\delta^2} \qquad \big(\text{By Lemma 10}\big).
\end{aligned}$$

So finally

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

### C.4.5 Conclusion of the proof

To summarize, we have obtained by (29):

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

and

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

where $b$, $c$ can be taken as $b := (2\overline{G}+1)(\tilde{T}+2)$, $c := \min(\delta^2, u)$, and this concludes the proof of Theorem 8.

□

# D Proof of Theorem 5

Recall that $\mathbf{M}^{(N)}(t)$ is the configuration of the system at time $t$, which means that $M_i^{(N)}(t)$ is the fraction of arms that are in state $i$ at time $t$. Let $\mathbf{e}_i$ be the $d$ dimensional vector that has all its component equal to 0 except the $i$th one that equals 1. The process $\mathbf{M}^{(N)}$ is a continuous-time Markov chain that jumps from a configuration $\mathbf{m}$ to a configuration $\mathbf{m} + \frac{1}{N}(\mathbf{e}_j - \mathbf{e}_i)$ when an arm jumps from state $i$ to state $j$. For $i < s(\mathbf{m})$, this occurs at rate $Nm_iQ_{ij}^1$ as all of these arms are activated. For $i > s(\mathbf{m})$, this occurs at rate $Nm_iQ_{ij}^0$ as these arms are not activated. For $i = s(\mathbf{m})$, this occurs at rate $N\big((\alpha - \sum_{k=1}^{s(\mathbf{m})-1} m_k)Q_{ij}^1 + (\sum_{k=1}^{s(\mathbf{m})} m_k - \alpha)Q_{ij}^0\big)$. Let us define:

$$\lambda_{ij}(\mathbf{m}) = \begin{cases} m_iQ_{ij}^1 & \text{if } i < s(\mathbf{m}) \\ (\alpha - \sum_{k=1}^{s(\mathbf{m})-1} m_k)Q_{ij}^1 + (\sum_{k=1}^{s(\mathbf{m})} m_k - \alpha)Q_{ij}^0 & \text{if } i = s(\mathbf{m}) \\ m_iQ_{ij}^0 & \text{if } i > s(\mathbf{m}). \end{cases}$$

The process $\mathbf{M}^{(N)}$ jumps from $\mathbf{m}$ to $\mathbf{m} + (\mathbf{e}_j - \mathbf{e}_i)/N$ at rate $N\lambda_{ij}(\mathbf{m})$. This shows that $\mathbf{M}^{(N)}$ is a density dependent population process as defined in Kurtz (1978). It is shown in Kurtz (1978) that, for any finite time $t$, the trajectories of $\mathbf{M}^{(N)}(t)$ converge to the solution of a differential equation $\dot{\mathbf{m}} = f(\mathbf{m})$ as $N$ grows, with $f(\mathbf{m}) := \sum_{i\neq j}\lambda_{ij}(\mathbf{m})(\mathbf{e}_j - \mathbf{e}_i)$. The function $f(\mathbf{m})$ is called the drift of the system. It should be clear that $f(\mathbf{m}) = \tau(\phi(\mathbf{m}) - \mathbf{m})$, where $\phi$ is defined for the discrete-time version of our continuous-time bandit problem.

For $t \geq 0$, denote by $\Phi_t\mathbf{m}$ the value at time $t$ of the solution of the differential equation that starts in $\mathbf{m}$ at time 0, it satisfies

$$\Phi_t\mathbf{m} = \mathbf{m} + \int_0^t f(\Phi_s\mathbf{m})ds.$$

Following Gast and Van Houdt (2017); Ying (2017), we denote by $L^{(N)}$ the generator of the $N$ arms system and by $\Lambda$ the generator of the differential equation. They associate to each almost-everywhere differentiable function $h$ two functions $L^{(N)}h$ and $\Lambda h$ that are defined as

$$\big(L^{(N)}h\big)(\mathbf{m}) := \sum_{i=1}^{d}\sum_{j\neq i} N\lambda_{ij}(\mathbf{m})\cdot\Big(h(\mathbf{m}+\frac{\mathbf{e}_j-\mathbf{e}_i}{N})-h(\mathbf{m})\Big),$$
$$\big(\Lambda h\big)(\mathbf{m}) := f(\mathbf{m})\cdot Dh(\mathbf{m}),$$

with $Dh$ being the differential of function $h$. The function $\Lambda h$ is defined only on points $\mathbf{m}$ for which $h(\mathbf{m})$ is differentiable. Remark that if $h(\mathbf{m})$ is an affine function in $\mathbf{m}$, *i.e.* $h(\mathbf{m}) = \mathbf{m}\cdot\mathbf{B}+\mathbf{b}$, with $\mathbf{B}$ a $d$-dimensional matrix and $\mathbf{b}$ a $d$-dimensional vector, then $\big(L^{(N)}h\big)(\mathbf{m}) = \big(\Lambda h\big)(\mathbf{m}) = f(\mathbf{m})\cdot\mathbf{B}$.

Now the analogue of Theorem 8(i) in the continuous-time case is

**Theorem 13** *Under the same assumptions as in Theorem 5, and assume that* $\mathbf{M}^{(N)}(0)$ *is already in stationary regime. Then there exists two constants* $b, c > 0$ *such that*

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

Note first that similarly, Theorem 13 implies Theorem 5.

*Proof* Define the continuous-time version of function $G$ as

$$G(\mathbf{m}) := \int_0^\infty \big(\Phi_t\mathbf{m}-\mathbf{m}^*\big)dt.$$

As for the discrete-time case, our assumptions imply that the unique fixed point is an exponentially stable attractor and a result similar to Lemma 12 can be obtained for the continuous-time case. This implies that the function $G$ is well-defined, continuous and bounded.

Recall that the function $f$ is affine in $\mathcal{Z}_{s(\mathbf{m}^*)}$: since if $\mathbf{m}\in\mathcal{Z}_{s(\mathbf{m}^*)}$, then $\phi(\mathbf{m}) = (\mathbf{m}-\mathbf{m}^*)\mathbf{K}+\mathbf{m}^*$ where $K$ is a $d\times d$ matrix, and $f(\mathbf{m}) = \tau(\phi(\mathbf{m})-\mathbf{m}) = \tau(\mathbf{m}-\mathbf{m}^*)(\mathbf{K}-\mathbf{I})$. Now suppose $\mathbf{m}\in\Delta^d$ is such that $\Phi_t\mathbf{m}$ remains inside $\mathcal{Z}_{s(\mathbf{m}^*)}$ for all $t\geq 0$, then

$$\Phi_t\mathbf{m} = (\mathbf{m}-\mathbf{m}^*)\cdot e^{t\cdot\tau(\mathbf{K}-\mathbf{I})}+\mathbf{m}^*, \text{ and } G(\mathbf{m}) = \frac{1}{\tau}(\mathbf{m}-\mathbf{m}^*)(\mathbf{K}-\mathbf{I})^{-1}.$$

So as for the discrete-time case, $G(\mathbf{m})$ is an affine function of $\mathbf{m}$, with affine factor $\mathbf{B} := \frac{1}{\tau}(\mathbf{K}-\mathbf{I})^{-1}$.

As $\mathbf{m}^*$ is non-singular, it is at a positive distance from the other zones $\mathcal{Z}_i\neq\mathcal{Z}_{s(\mathbf{m}^*)}$ and we therefore define $\delta := \min_{i\neq s(\mathbf{m}^*)} d(\mathbf{m}^*,\mathcal{Z}_i)/2 > 0$, where $d(\cdot\,,\cdot)$ is the distance under $\|\cdot\|$-norm. We then choose a neighborhood $\mathcal{N}_1 := \mathcal{B}(\mathbf{m}^*,\epsilon_1)\cap\Delta^d$ of $\mathbf{m}^*$ such that for all $t\geq 0$ and all initial condition $\mathbf{m}\in\mathcal{N}_1$, $\Phi_t(\mathbf{m})\in\mathcal{B}(\mathbf{m}^*,\delta)$. This is possible by the exponentially stable attractor property of $\mathbf{m}^*$. Following Theorem 3.2 of Gast (2017), we have

$$\mathbf{m}^* - \mathbb{E}\big[\mathbf{M}^{(N)}(0)\big] = \mathbb{E}\big[\Lambda G\big(\mathbf{M}^{(N)}(0)\big)\big]$$

$$= \mathbb{E}\big[(\Lambda - L^{(N)})G\big(\mathbf{M}^{(N)}(0)\big)\big]$$

$$= \mathbb{E}\Big[\Big((\Lambda - L^{(N)})G\big(\mathbf{M}^{(N)}(0)\big)\Big) \cdot \mathbb{1}_{\{\mathbf{M}^{(N)}(0)\in\mathcal{N}\}} \tag{30}$$

$$+ \Big((\Lambda - L^{(N)})G\big(\mathbf{M}^{(N)}(0)\big)\Big) \cdot \mathbb{1}_{\{\mathbf{M}^{(N)}(0)\notin\mathcal{N}\}}\Big], \tag{31}$$

where $\mathcal{N} := \mathcal{B}(\mathbf{m}^*, \epsilon_1/2) \cap \Delta^d$. Let $N_0 := \lceil 2/\epsilon_1 \rceil$. For $N \geq N_0$, $\mathbf{m} \in \mathcal{N}$ verifies additionally that $\Phi_t\big(\mathbf{m} + \frac{\mathbf{e}_j - \mathbf{e}_i}{N}\big) \in \mathcal{Z}_{s(\mathbf{m}^*)}$ for all $1 \leq i \neq j \leq d$ and $t \geq 0$. Hence, $G$ is locally affine and for all $m \in \mathcal{N}$ and $N \geq N_0$, we have:

$$\big(\Lambda G\big)(\mathbf{m}) = \big(L^{(N)}G\big)(\mathbf{m}) = f(\mathbf{m}) \cdot \mathbf{B}. \tag{32}$$

This shows that the first term of (30) is equal to zero.

For the second term, note that both $G$ and $\Lambda G$ are continuous functions defined on the compact region $\Delta^d$, hence they are both bounded, while $L^{(N)}G$ grows at most linearly with $N$. Hence we can choose constants $u, v > 0$ independent of $N$ such that:

$$\sup_{\mathbf{m}\in\Delta^d} \|\big(\Lambda G\big)(\mathbf{m})\| = u, \quad \sup_{\mathbf{m}\in\Delta^d} \|\big(L^{(N)}G\big)(\mathbf{m})\| \leq vN.$$

We are left to bound $\mathbb{P}\big(\mathbf{M}^{(N)}(0) \notin \mathcal{N}\big)$ exponentially from above. This could be done by using the (unnamed) proposition on page 644 of Weber and Weiss (1990). Yet, we were not able to find the paper referenced for the proof of this proposition. Hence, we provide below a direct proof of this. To achieve this, we rely on an exponential martingale concentration inequality, borrowed from Darling and Norris (2008), which in our situation can be stated as

**Lemma 14** Fix $T > 0$. Let $K$ be the Lipschitz constant of drift $f$, denote $\lambda := \max_{i,j} \lambda_{ij}$, and $c_1 := e^{-2KT}/18T$. If $\epsilon > 0$ is such that

$$1 \geq \epsilon\lambda \cdot \exp\big(\frac{\epsilon^2 e^{-KT}}{3T}\big), \tag{33}$$

then we have

$$\mathbb{P}\Big[\sup_{t\leq T} \|\mathbf{M}^{(N)}(t) - \Phi_t\mathbf{m}\| > \epsilon \Big| \mathbf{M}^{(N)}(0) = \mathbf{m}\Big] \leq 2d \cdot e^{-c_1 N\epsilon^3}. \tag{34}$$

The above lemma plays the role of Lemma 11 in discrete-time case. Note that its original form stated as Theorem 4.2 in Darling and Norris (2008) is under a more general framework, which considered a continuous-time Markov chain with countable state-space evolves in $\mathbb{R}^d$, and discussed a differential equation approximation to the trajectories of such Markov chain. As such, the right hand side of (34) has an additional term $\mathbb{P}(\Omega_0^c \cup \Omega_1^c \cup \Omega_2^c)$, with $\Omega_i^c$ being the complementary of $\Omega_i$. In our case, $\Omega_0 = \Omega_1 = \Omega$ trivially holds; while the analysis of $\Omega_2$ is more involved. However, as remarked before the statement of Theorem 4.2 in Darling and Norris (2008), if the maximum jump rate (in our case $N\lambda$) and the maximum jump size (in our case $1/N$) of the Markov chain satisfy certain inequality, which in our situation can be sated as (33), then $\Omega_2 = \Omega$. Note that the constraint (33) is satisfied as long as $\epsilon$ is sufficiently small, and consequently $\mathbb{P}(\Omega_0^c \cup \Omega_1^c \cup \Omega_2^c) = 0$.

Now let $\epsilon > 0$ be such that $\mathcal{B}(\mathbf{m}^*, 2\epsilon) \cap \Delta^d \subset \mathcal{N}$. The uniform global attractor assumption on $\mathbf{m}^*$ ensures that there exists $T > 0$ such that for all $\mathbf{m} \in \Delta^d$ and $t \geq T$: $\Phi_t\mathbf{m} \in \mathcal{B}(\mathbf{m}^*, \epsilon)$. Let such $T$ and $\epsilon$ be as in Lemma 14 that verify additionally

(33). This is possible as the right hand side of (33) converges to 0 when $\epsilon$ is small and $T$ is large.

We then have:

$$\begin{aligned}
\mathbb{P}\big[\mathbf{M}^{(N)}(0) \notin \mathcal{N}\big] &= \mathbb{P}\big[\mathbf{M}^{(N)}(T) \notin \mathcal{N}\big] \qquad \text{(By stationarity)} \\
&\leq \mathbb{P}\big[\|\mathbf{M}^{(N)}(T) - \mathbf{m}^*\| \leq 2\epsilon\big] \\
&\leq \mathbb{P}\big[\|\mathbf{M}^{(N)}(T) - \Phi_T(\mathbf{M}^{(N)}(0))\| > \epsilon\big] + \mathbb{P}\big[\|\Phi_T(\mathbf{M}^{(N)}(0)) - \mathbf{m}^*\| > \epsilon\big] \\
&= \mathbb{P}\big[\|\mathbf{M}^{(N)}(T) - \Phi_T(\mathbf{M}^{(N)}(0))\| > \epsilon\big] \qquad \big(\text{By our choice of } T\big) \\
&\leq 2d \cdot e^{-c_1 N \epsilon^3} \qquad \big(\text{We apply (34) of Lemma 14}\big).
\end{aligned}$$

So in summary, (30)-(31) gives

$$\|\mathbb{E}\big[\mathbf{M}^{(N)}(0)\big] - \mathbf{m}^*\| \leq (u + vN) \cdot 2d \cdot e^{-c_1 N \epsilon^3}. \tag{35}$$

Moreover, for any $c' > 0$ and $0 < c < c'$, $N \cdot e^{-c'N} = \mathcal{O}(e^{-cN})$, so the right hand side of (35) can be bounded by a term of the form $b \cdot e^{-cN}$. This concludes the proof of Theorem 13. □