Exponential Convergence Rate for the Asymptotic Optimality of Whittle Index Policy

Nicolas Gast, Bruno Gaujal, Chen Yan

Submitted to the Annals of Applied Probability

EXPONENTIAL CONVERGENCE RATE FOR THE ASYMPTOTIC OPTIMALITY OF WHITTLE INDEX POLICY

BY NICOLAS GAST[1], BRUNO GAUJAL[1] AND CHEN YAN[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France, nicolas.gast@inria.fr; bruno.gaujal@inria.fr; chen.yan@univ-grenoble-alpes.fr.*

We evaluate the performance of Whittle index policy for restless Markovian bandits, when the number of bandits grows. It is proven in [30] that this performance is asymptotically optimal if the bandits are indexable and the associated deterministic system has a global attractor fixed point. In this paper we show that, under the same conditions, the convergence rate is exponential in the number of bandits, unless the fixed point is *singular* (to be defined later). 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 bandits. Using simulations and numerical solvers, we also investigate the cases where the conditions for the exponential rate theorem are violated, notably when attracting limit cycles appear, or when the fixed point is singular. We illustrate our theorem on a Markovian fading channel model, which has been well studied in the literature. Finally, we extend our synchronous model results to the asynchronous model.

1. Introduction. A multi-armed bandit (MAB) problem is a sequential allocation problem: At each decision epoch, one or several bandits 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, 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.* [4, 20]). Each time, a subset of bandits are chosen to be activated. All bandits generate an instantaneous reward that depend on their state and their activation. The state of each activated bandit then changes in a Markovian fashion, based on an underlying transition matrix (or a rate matrix in the asynchronous case). Both the reward and the new state are revealed to the decision maker for its next decision. The bandits that are not activated change state according to a different transition matrix. When the underlying stochastic transition laws are assumed to be known (see [6] 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 *restful case* (non activated bandits do not change their states) with one active bandit at each decision epoch in [13] by the Gittins index policy, which is a greedy policy that can be computed efficiently. In [32] Whittle generalizes the model in two aspects. Firstly, at each decision epoch more than one bandit can be activated, and secondly, the bandits that are not activated can also change states (*restless bandits*), according to a different stochastic 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 [26] that this problem is PSPACE-hard.

MSC2020 subject classifications: Primary 90C40; secondary 37H12, 60F10, 68M20.
Keywords and phrases: Multi-armed Bandits, Whittle Index, Asymptotic Optimality.

In [32], however, Whittle conjectures that, under some conditions, the so-called "Whittle index policy" (WIP) should be optimal asymptotically, *i.e.* when the number of bandits goes to infinity with a fixed proportion of active bandits.

This conjecture has been proven in the famous paper [30] for the asynchronous model, under several technical conditions (further discussed in [31]), 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 bandit 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 [1, 15, 28], queuing systems [2], crawling optimal content on the web [3], load-balancing [19] and sensors [24]. In particular, partially observable Markov decision process (POMDP) falls 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 [21, 22]. 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. This example will be further analyzed in Section 5 to illustrate our result.

There are several extensions to the restless bandit model. In [14], individual bandits have multiple levels of activation (instead of only two) and are subject to an overall resource constraint; in [7], the environment in which lives the bandits also changes along time; in [29], the author considers the larger set of all priority policies (to which WIP belongs) and aims at finding best policies among them. Of all those generalizations, some similar index policies have been proposed and asymptotic optimality results have been proved therein. However, to the best of our knowledge, nobody has considered the question of evaluating the convergence rate of the performance of the Whittle index policy to the optimal one.

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 bandits, giving a theoretical explanation for the good performance of WIP in practice, even when the number of bandits is small. This result holds in the synchronous as well as the asynchronous cases, under the same conditions as the asymptotic optimality proven in [30], 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*. This last condition will be discussed in length in the rest of the paper. The exponential convergence rate extends to the multi-action case developed in [14]. This extension is direct (the piece-wise affine structure of the dynamics is preserved in the multi-action case) and will not be discussed further in 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 bandits, each with d states, under WIP 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 [9, 17, 11, 33] 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 synchronous case) or an exponential martingale concentration inequality (in the asynchronous 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 bandits 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 bandits 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 [16]. 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 what is actually going on.
- 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 in $1/\sqrt{N}$ in the singular case) are backed by numerical experiments showing that for a moderate number of bandits (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.
- Finally, we also investigate the behavior of the restless bandits under WIP when the fixed point of the deterministic dynamical system is *not* a global attractor. In the synchronous case, the system can become periodic, with a stable periodic cycle of variant length (the fixed point being unstable). In such cases, the performance of WIP converges to the average performance over the cycle and in general is not asymptotically optimal.

Organization of the paper. In Section 2, we introduce the synchronous restless bandit model: all bandits 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: singular, non-singular, cyclic. 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 classical asynchronous bandit model (bandits are continuous-time Markov chains, and decisions are made every time when one bandit changes its state). We show that exponential convergence rate also holds in the asynchronous case, and highlight the differences between the synchronous and asynchronous models.

2. The synchronous 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 relaxation in Section 2.3. Note that in our model, all bandits are synchronous. This is a discrete-time version of the classical continuous-time model studied in [30]. We will discuss an extension of our results to the latter model in Section 6.

2.1. *Model description.* The synchronous 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 bandits. Each bandit evolves in an identical finite state space $\{1 \dots d\}$ and the state of bandit n at time t is denoted by $S_n(t) \in \{1 \dots d\}$. The state space of the whole process is denoted at time t by $\mathbf{S}(t) = \big(S_1(t), S_2(t), ..., S_N(t)\big)$.
2. Decisions are taken at times $t \in \mathbb{N}$. At each decision epoch, a decision maker observes $\mathbf{S}(t)$ and chooses αN of the N bandits to be activated, where we assume that α and N are such that αN is an integer. We set $a_n(t) = 1$ if bandit 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. Bandit 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 $\mathbf{a}(t)$, the N bandits make their transitions independently.
4. If bandit n is in state i and the decision maker takes action $a \in \{0, 1\}$, a bounded reward $R_i^a \in \mathbb{R}$ is earned.

The goal of the decision maker is to compute a decision rule in order to maximize the long-term expected average reward. The theory of stochastic dynamic programming [27] shows that there exists an optimal policy which is Markovian and stationary (*i.e.* $\mathbf{a}(t)$ can be chosen as a time-independent function of $\mathbf{S}(t)$). Denote by Π the set of such Markovian stationary policies, the optimization problem of the decision maker can be formalized as

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

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

In the rest of the paper, we assume that matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ are such that the states form a single aperiodic closed class, regardless of the policy employed. This assumption was also used in [30] and guarantees 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 unichain* bandit.

2.2. *Indexability and Whittle index.* In theory, a dynamic programming approach can be used to solve Equations (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 [26]. To overcome this difficulty, Whittle introduces in [32] a very efficient heuristic known as Whittle index policy (WIP). This heuristic is obtained by computing an index ν_i for each state i. At a given decision epoch, WIP activates the αN bandits having currently the highest indices. We describe below how these indices are defined.

The index of a bandit can be computed by considering each individual bandit in isolation. For a given $\nu \in \mathbb{R}$, we define the subsidy-ν problem as the following MDP. The state space is the one of a single bandit. At each time t, the decision maker chooses whether or not to activate this bandit. As in the original problem, the bandit evolves at time t according to (1). The difference lies in the passive action that is subsidized: If the bandit is in state i and action 1 is taken, then as before, a reward R_i^1 is earned; if the bandit 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 total expected reward (including passive subsidies) earned over an infinite horizon. 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 ν-subsidized MDP such that the passive action is optimal in these states. Whittle indices are defined as follows:

DEFINITION 2.1 (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 ν, namely if for all $\nu \leq \nu'$, we have $\omega(\nu) \subseteq \omega(\nu')$. In this case, the Whittle index of a state i, that we denote by ν_i, is defined as the smallest subsidy such that the passive action is optimal in this state:

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

It should be emphasized that there exist restless bandit problems that are *not* indexable, although this is relatively rare (we discuss this in more detail in Section 4.1). Note that when $\mathbf{P}^0$ is the identity matrix, bandits are restful, *i.e.* the states of the bandits 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 [12]. An algorithm to test indexability and calculate the index, as well as a geometric interpretation can be found in [23]. For given parameters $(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1)$, the complexity of this algorithm is $\mathcal{O}(2^d)$. However, if a restless bandit is known to be indexable beforehand, then its indices can be computed by a greedy algorithm in time $\mathcal{O}(d^3)$.

2.3. *Whittle relaxation and the asymptotic optimality theorem.* An intuition behind the definition of Whittle index is that it is related to a relaxation of the original N bandits problem (2) where the constraint (3) is replaced by $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} a_n(t) = \alpha N$. While the constraint (3) imposes that exactly αN bandits are activated at each time step, the relaxed constraint only imposes the time-averaged number of activated bandits to be equal to αN:

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

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

By using ν as a Lagrange multiplier of the constraint $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} a_n(t) = \alpha N$, the Lagrangian of the problem (4)-(5) is

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

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

Note that, for a fixed ν, finding a policy that maximizes the above Lagrangian can be done by solving N independent optimization problems (one for each bandit), and each problem is a ν-subsidized MDP.

It should be clear that 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 bandits grows[1], the value of the original problem converges to this value:

THEOREM 2.1. Consider a synchronous restless bandit model with N identical aperiodic unichain bandits and such that the matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ are rational. Then $V_{\text{rel}}^{(N)}(\alpha) = NV_{\text{rel}}^{(1)}(\alpha)$ and

$$\limsup_{N\to\infty} \sqrt{N}\left(\frac{V_{\text{opt}}^{(N)}(\alpha)}{N} - V_{\text{rel}}^{(1)}(\alpha)\right) < \infty. \tag{6}$$

The above theorem justifies the relaxation (5) by showing that when the number of bandit is large, the value of the optimization problem (2) is close to $V_{\text{rel}}^{(N)}(\alpha)$: $\lim_{N\to\infty} \frac{V_{\text{opt}}^{(N)}(\alpha)}{N} = V_{\text{rel}}^{(1)}(\alpha)$. In Theorem 1 of [30], the result $\lim_{N\to\infty} \frac{V_{\text{opt}}^{(N)}(\alpha)}{N} = V_{\text{rel}}^{(1)}(\alpha)$ was proved for the asynchronous bandit model that we will discuss in Section 6. To the best of our knowledge, the statement of this theorem in our setting of synchronous 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 2.1 in Appendix A. It is an adaptation of the proof of [30, Theorem 1]: we use a similar coupling argument, although the coupling has to be adapted to our synchronous setting, and we also need the additional aperiodic assumption on the model.

While Theorem 2.1 guarantees that the original optimization problem converges to the relaxation, it does not guarantee any result on the performance of WIP. This leaves two important questions: *(1) Is indexability a sufficient condition for WIP to be asymptotically optimal (for our synchronous bandit model)? (2) If WIP is asymptotically optimal, then, at which speed does it become optimal?* In the remainder of the paper we will see that, similarly to the asynchronous bandit model, there exist examples for which WIP is *not* asymptotically optimal. We will also exhibit sufficient conditions to guarantee asymptotic optimality that is similar to the conditions of [30]. Our main result concerns the rate of convergence. We will show that, except in rare cases, when WIP is asymptotically optimal, it does so at exponential speed with the number of bandits N. This complements Theorem 2.1 by proving that, under the same conditions, the convergence in (6) 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 asynchronous bandits of [30].

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

[1] In the rest of the paper, unless otherwise specified, we restrict our attention to the values of N such that αN is an integer. The notation $\lim_{N\to\infty} f(N) = \ell$ is to be understood as $\lim_{k\to\infty} f(kN_0) = \ell$ where N_0 is the smallest positive integer such that $\alpha N_0 \in \mathbb{N}$. We will discuss non-integer values of αN in Section 4.4.

there does not exist two states that have the same Whittle index. This is mostly a technical assumption that guarantees that there is a unique[2] WIP.

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

LEMMA 3.1. Assume that the bandits are unichain and indexable. The map ϕ satisfies:

(i) The definition of ϕ does not depend on N (as long as αN is an integer) nor on t.
(ii) ϕ is a piecewise affine function, with d affine pieces, and ϕ is Lipschitz-continuous.
(iii) ϕ has a unique fixed point, *i.e.* 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 bandits by decreasing index order. This means that when the system is in configuration $\mathbf{m}$, WIP will activate all bandits that are in states 1 to $s(\mathbf{m}) - 1$, and $N(\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i)$ bandits that are in state $s(\mathbf{m})$. The other bandits will not be activated. This implies that the map ϕ is defined as:

(7)
$$\phi_j(\mathbf{m}) = \sum_{i=1}^{s(\mathbf{m})-1} m_i P^1_{ij} + (\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i) P^1_{s(\mathbf{m})j} + (\sum_{i=1}^{s(\mathbf{m})} m_i - \alpha) P^0_{s(\mathbf{m})j} + \sum_{i=s(\mathbf{m})+1}^{d} m_i P^0_{ij}.$$

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

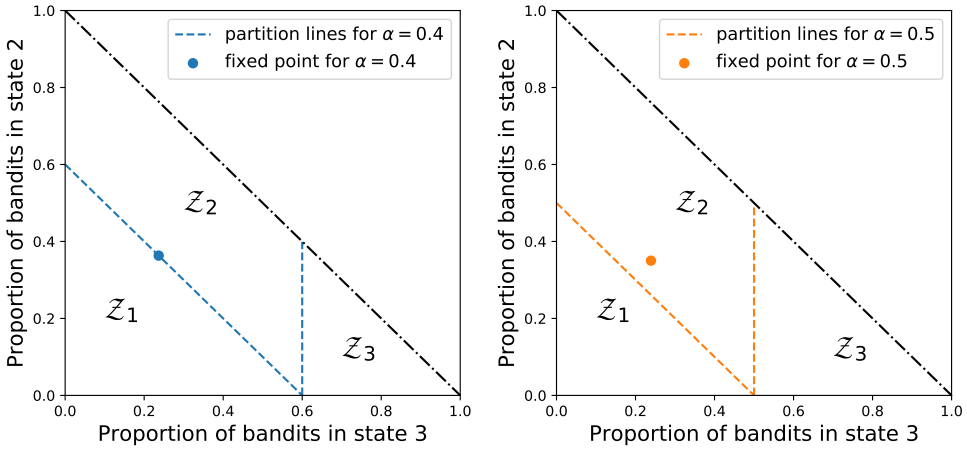
Proof of (iii) – This part of the proof is more involved, and it relies on indexability. By (7), for each $i \in \{1 \dots d\}$, there exist a matrix $\mathbf{K}_i$ and a vector $\mathbf{b}_i$ such that for all $\mathbf{m} \in \mathcal{Z}_i$, we have:

(8)
$$\phi(\mathbf{m}) = \mathbf{m} \cdot \mathbf{K}_i + \mathbf{b}_i.$$

[2] If two states or more had the same index, to specify an index policy, one would need a tie-breaking rule. Our proof could be easily adapted if either the tie-breaking rule defines a strict order of the states or if ties are broken at random.

The assumption that each bandit is unichain implies that the linear equation $\phi(\mathbf{m}) = \mathbf{m}$ has a unique solution for each couple $(\mathbf{K}_i, \mathbf{b}_i)$. Hence, ϕ has at most one fixed point inside each zone $\mathcal{Z}_i$. To conclude, we prove in Appendix B that the indexability of bandits implies a monotonic property of ϕ that we use to obtain uniqueness. □

In what follows, we will denote by $\mathbf{m}^*$ the unique fixed point of ϕ. As we will see in Theorem 3.2, the rate at which WIP becomes asymptotically optimal depends on (1) whether the iterations of ϕ converge to $\mathbf{m}^*$ and (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 Δ^3 can be represented in a 2-dimensional space as Δ_c^2, where Δ_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 bandits in state 3), and the y-coordinate corresponds to m_2 (the proportion of bandits in state 2). The colored dotted lines of Figures 1a and 1b correspond to the set of singular points. These lines partition the different zones $\mathcal{Z}_i$. The partition of zones, as well as the position of the unique fixed point depend on α. For this example, when $\alpha = 0.4$ (Figure 1a), the fixed point is singular, while for $\alpha = 0.5$ (Figure 1b), it is non-singular.

3.2. *Exponential convergence rate.* We are now ready to state our main theorem. Assume that bandits are indexable, at a given time t, WIP sorts all bandits according to the Whittle indices $\nu_{S_n(t)}$ and activates the αN bandits that have the highest indices. We denote the long-term average expected reward of WIP as $V_{\mathrm{WIP}}^{(N)}(\alpha)$:

$$V_{\mathrm{WIP}}^{(N)}(\alpha) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1}\sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)}\Big], \qquad \text{where for all } t,\ \mathbf{a}(t) \text{ is chosen according to WIP.}$$

Let Φ_t be defined as the t-th iteration of the map ϕ, *i.e.* $\Phi_t : \Delta^d \to \Delta^d$ is $\Phi_0(\mathbf{m}) := \mathbf{m}$, and $\Phi_{t+1}(\mathbf{m}) := \phi\big(\Phi_t(\mathbf{m})\big)$. Recall that $\mathbf{m}^*$ is the unique fixed point of ϕ. As stated

in the next theorem, the asymptotic optimality of WIP is guaranteed when $\mathbf{m}^*$ attracts all trajectories of $\Phi_{t\geq 0}(\cdot)$. In the rest of the paper, unless otherwise specified, we use $\|\cdot\|$ to denote the $\mathcal{L}^\infty$-norm of a vector.

THEOREM 3.2 (Exponential convergence rate theorem-synchronous case). Consider a synchronous restless bandit problem $\left\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha\right\}$ such that:

(i) Bandits are unichain and indexable.
(ii) $\mathbf{m}^*$ is a uniform global attractor of $\Phi_{t\geq 0}(\cdot)$, *i.e.* for all $\epsilon > 0$, there exists $T(\epsilon) > 0$ such that for all $t \geq T(\epsilon)$ and all $\mathbf{m} \in \Delta^d$, one has $\|\Phi_t(\mathbf{m}) - \mathbf{m}^*\| \leq \epsilon$.
(iii) $\mathbf{m}^*$ is not singular.

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

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

Recall that $V_{\text{rel}}^{(N)}(\alpha)$ is the value of the relaxed problem (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 bandits 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 $\|\mathbb{E}[\mathbf{M}^{(N)}(0)] - \mathbf{m}^*\| \leq b' \cdot e^{-cN}$. In order to show this:

- We first use Hoeffding's inequality in Lemma C.3 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 ϕ, for a time t, we apply Lemma C.3 to prove Lemma C.4, 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.
- We combine this with the uniformly global attractor property to show that when t is large enough, $\mathbf{M}^{(N)}(t)$ is within a neighborhood $\mathcal{N}$ of $\mathbf{m}^*$ with very high probability. As $\mathbf{m}^*$ is non-singular, this neighborhood can be taken to be within a zone $\mathcal{Z}_i$ on which ϕ is affine. We will choose carefully this neighborhood $\mathcal{N}$ and make sure that its choice does not depend on N. We then deduce an exponentially small upper-bound for the probability of $\mathbf{M}^N(0)$ in stationary regime being outside $\mathcal{N}$ (see Subsection C.5.3), hence allows us to restrict our attention to a zone where ϕ is affine.
- The result then follows by using Stein's method on the process restricted to this affine zone, which shows that conditional on starting inside the neighborhood $\mathcal{N}$, the additive long-term distance between the large N stochastic trajectory and the deterministic trajectory is exponentially small (see Subsection C.5.4).

□

Here are some comments on the assumptions of Theorem 3.2. These assumptions are very similar to the ones needed to prove the asymptotic optimality of WIP in the case of asynchronous bandits of [30]. The indexability and unichain property of the bandit problem is a necessary condition for WIP to be well defined and was also assumed in [30]. The most difficult assumption to verify is point *(ii)* that requires $\mathbf{m}^*$ to be a global attractor. Note that our assumption *(ii)* is slightly stronger than the one used in [30] as we assume that $\mathbf{m}^*$ attracts the trajectories *uniformly* in the initial condition. We use this to obtain our rate of convergence. We will see in Section 4.3 that this condition is necessary in the sense that there exists examples that satisfy all assumptions of Theorem 3.2 except this one and for

which WIP is not asymptotically optimal (more comments on this in Remark 3.2). In [30] the authors have proven that this is not possible when $d = 3$ for the asynchronous case, and they give an example of dimension $d = 4$ in [31] for which the deterministic differential equation has an attracting limit cycle. Their proof of the impossibility in dimension 3 relies on Bendixson's Criteria to exclude limit cycles and Poincaré-Bendixson Theorem to exclude chaotic behaviors. The synchronous situation is different: by randomly generating examples, we are able to find attracting limit cycles (of period 2) for $d = 3$ and more complicated shape of cycles for $d = 4$ (see Section 4).

REMARK 3.1. **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 3.2 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 $V^{(1)}_{\text{rel}}(\alpha) = 0.5$. 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 bandits in state 1. This implies that the reward of WIP will be $V^{(N)}_{\text{WIP}}(\alpha) = \mathbb{E}\left[\min(M^{(N)}_1, 0.5 \cdot N)\right]$. As $M^{(N)}_1$ follows a binomial distribution of parameter $(N, 0.5)$, the central limit theorem shows that

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

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 $\mathcal{O}(1/\sqrt{N})$ and not at exponential rate. Note that if we take instead $\alpha \neq 0.5$, $V^{(N)}_{\text{WIP}}(\alpha)/N$ converges to $V^{(1)}_{\text{rel}}(\alpha) = \min(\alpha, 0.5)$ at exponential rate due to the fact that almost all the mass of a Gaussian distribution is concentrated around its mean value α (which is different from 0.5).

REMARK 3.2. **Cyclic and chaotic behaviors**. Although the drift ϕ is piecewise affine and has a unique fixed point, the long run behavior of the deterministic dynamical system $\mathbf{m}(t+1) = \phi(\mathbf{m}(t))$ can be cyclic or chaotic. In these cases, the fixed point is no longer a global attractor, and the performance of WIP is in general not asymptotically optimal.

More precisely, when the dynamical system admits a cycle as a global attractor for almost every initial configuration in the simplex, then as suggested in [30], one can infer a cyclic version of Theorem 2.1: The performance of WIP converges to the average reward on the cycle. This average reward is in general strictly smaller than $V^{(1)}_{\text{rel}}(\alpha)$, while $V^{(N)}_{\text{opt}}(\alpha)/N$ always converge to $V^{(1)}_{\text{rel}}(\alpha)$, regardless to the behavior of the deterministic system (from Theorem 2.1). Consequently, when cycles appear, the performance of WIP is asymptotically *sub-optimal*. See also Sections 4.3.

REMARK 3.3. **What happens when αN is not an integer**. The exponential convergence rate in Theorem 3.2 assumes that αN is an integer. When it is not the case, a decision maker cannot activate exactly αN bandits at each time step. There are three natural solutions to define the model in such cases: (1) activate $\lfloor \alpha N \rfloor$ bandits; (2) activate $\lceil N\alpha \rceil$ bandits; (3) activates $\lfloor \alpha N \rfloor$ bandits, plus one more bandit being activated with probability $\alpha N - \lfloor \alpha N \rfloor$. As we further discuss in Section 4.4, the convergence rate in the first two solutions is much slower than in the third solution.

REMARK 3.4. **Finding optimal constants**. Theorem 3.2 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 3.2 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 α, if all the other parameters $\left\{\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1\right\}$ are fixed.

4. Numerical Experiments. In this section, we first provide statistical results to justify the conditions needed for Theorem 3.2, and then verify numerically the exponential convergence rate for a general 3 states restless bandit model with non-singular fixed points. We also evaluate numerically the convergence rate for a singular fixed point example. We then investigate the situation when attracting limit cycle appears, as well as the solutions for αN not being integers. To ensure reproducibility[3], all parameters used in our numerical experiments are provided in Appendix E.

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 uniform global attractor.

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 [23]. As its running time is exponential in the number of states, it is only applicable when this number is small (say $d < 10$). For Condition (C2), being in an exact singular situation is improbable (for a given problem, the activation ratio α can only be singular if it satisfies an equality constraint). More generally, we will also check that the closer the fixed point to a singular situation, the smaller the coefficient c in the estimation of the exponential rate could be. This point will be made more precise in the next subsection.

Condition (C3) is more complicated to verify as there is no general method to exclude cyclic or chaotic behaviors from a dynamical system. A necessary condition for (C3) to hold is that the unique fixed point is locally stable. Numerically, this is easy to verify 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 defined as in (8). The dynamical system is locally stable if $\mathbf{K}_{s(m^*)}$ is a stable matrix, *i.e.* if the norm of all eigenvalues of $\mathbf{K}_{s(m^*)}$ is less than 1 [4]. If $\mathbf{K}_{s(m^*)}$ is not a stable matrix, the fixed point will not be a global attractor and an attracting cycle will appear in most cases, see Sections 4.3.

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

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

Dimension d	3	4	5	6	7
Nb. of non-indexable instances	653	81	5	0	0
Nb. of indexable instances such that $\mathbf{m}^*$ is not locally stable	9878	1020	82	11	0
% of examples violating a condition of Theorem 3.2	0.1%	0.01%	10^{-3}%	10^{-4}%	0.

TABLE 1

Number of randomly generated instances that violate any of the conditions of Theorem 3.2 out of 10^7 *uniformly generated restless bandit models for each dimension* $d \in \{3,4,5,6,7\}$.

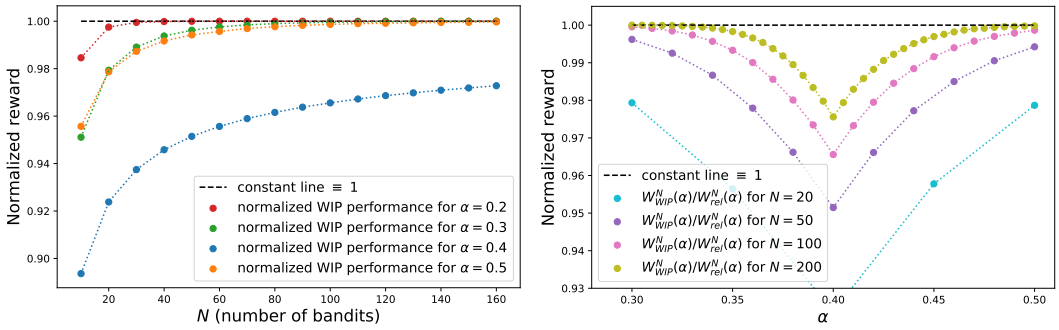
To check how general are these conditions, we generate a large number of synchronous restless bandit problems by choosing random parameter $(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1)$ in dimensions $d \in \{3,4,5,6,7\}$. We estimate the rarity of violations of the above conditions. More precisely, for each d, we randomly generate 10^7 instances of $(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1)$, using a uniform distribution in $[0,1]$ for the rewards, and uniform distribution for probability vectors $\mathbf{P}_i^0$ and $\mathbf{P}_i^1$ over the simplex Δ^d. We then count the number of instances that violate conditions (C1) or (C3), 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 (where K_i is defined as in (8)). To test numerically if $\mathbf{m}^*$ is a global attractor, we check if $\lim_{t\to\infty} \Phi_t(\mathbf{m}) = \mathbf{m}^*$ on 10^5 randomly generated initial conditions $\mathbf{m}$. Note that we exclude $d=2$ from this table: it can be shown that when $d=2$, all models are indexable and the fixed point $\mathbf{m}^*$ is always a globally stable attractor. Concerning indexability, a similar statistical test was already presented in [23, (Table 2, left table, $\beta=1$)]. Our results are essentially equivalent (note however that [23] does not study the stability of $\mathbf{m}^*$ but only tests indexability).

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 α in the range between 0.20 and 0.50. Although we have no formal proof, we believe that the fixed points are uniform global attractors for two reasons:

- All matrices $\mathbf{K}_i$ are locally stable because the eigenvalues of $\mathbf{K}_2$ are $\{1, -0.4\ldots, 0.08\ldots\}$ while $\mathbf{K}_1$ and $\mathbf{K}_3 = \mathbf{K}_d$ are always stable matrices.
- For all tested values of α, we simulated $\Phi_t(\mathbf{m})$ from random initial points $\mathbf{m}$ and they all converge to the corresponding fixed point $\mathbf{m}^*$.

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 3.2 are satisfied when $\alpha \neq 0.4$. As $V_{\text{rel}}^{(N)}(\alpha)$ depends on the value of α, to make better comparisons, we consider the quantity $V_{\text{WIP}}^{(N)}(\alpha)/V_{\text{rel}}^{(N)}(\alpha)$, which is the normalized performance of WIP with respect to the relaxation upper-bound. In Figure 2a, we choose four values of α as 0.2, 0.3, 0.4 and 0.5, and plot the normalized performances as a function of the number of bandits N that takes values on multiples of 10, so that αN are always integers. The value of $V_{\text{WIP}}^{(N)}(\alpha)$ are cumputed 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 α where α varies between $[0.3, 0.5]$ with a stepsize of $1/N$: $\alpha \in \{0.3, 0.3+1/N, 0.3+2/N, \ldots, 0.5\}$ (so that again αN are always integers). These two figures suggest that the convergence rate is related to how far $\mathbf{m}^*$ is away from the closest boundary of two zones (*i.e.* how non-singular it is). Here is an intuitive explanation for this phenomenon: the stochastic system in equilibrium

will wander around the fixed point $\mathbf{m}^*$ that gives the optimal reward, now if $\mathbf{m}^*$ is near a boundary, it is more likely for the stochastic trajectory to jump into another neighboring polytope $\mathcal{Z}'$, in which case another affine drift applies and this may take the trajectory away from $\mathbf{m}^*$.



(a) Varying N for $\alpha = 0.2, 0.3, 0.4$ and 0.5. (b) Varying α for $N = \in \{20, 50, 100, 200\}$.

Fig 2: Normalized performance of WIP.

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

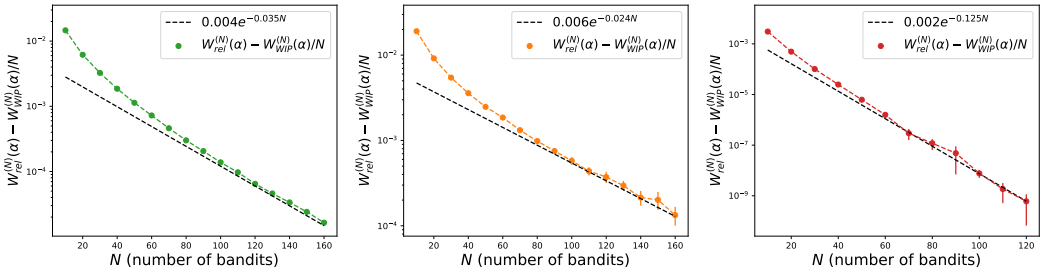
$$\text{subgap}(N) := V^{(1)}_{\text{rel}}(\alpha) - \frac{V^{(N)}_{\text{WIP}}(\alpha)}{N}. \tag{10}$$

Theorem 3.2 implies that subgap(N) converges to 0 approximately as $b \cdot e^{-c \cdot N}$, for some constants $b, c > 0$. 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 α, we also plot the best-fit $b' e^{-c' N}$ 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[5] 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 $\mathcal{O}(\frac{1}{\sqrt{N}})$ convergence rate in this singular case, same as for the example given in Remark (3.1). Actually, we believe that the convergence rate is $\mathcal{O}(\frac{1}{\sqrt{N}})$ for all singular global attractor situations, but a proof of this claim is still open to us.

4.3. *WIP is suboptimal in cases of cycle: examples of an attracting period-2 cycle with* $d = 3$. In the synchronous case, it is possible to have an attracting periodic cycle for the dynamics $\Phi_{t \geq 0}(\cdot)$ as soon as $d = 3$. Several new features that we have not encountered in the previous uniform global attracting fixed point situation will appear. To motivate our discussion, recall that under the assumptions of Theorem 3.2, the quantity $V^{(N)}_{\text{WIP}}(\alpha)/N$ converges to $V^{(1)}_{\text{rel}}(\alpha)$ exponentially fast. As $N \cdot V^{(1)}_{\text{rel}}(\alpha) \geq V^{(N)}_{\text{opt}}(\alpha) \geq V^{(N)}_{\text{WIP}}(\alpha)$, this guarantees that WIP becomes asymptotically optimal *at least* exponentially fast in N. Such nice feature is no longer true, however, when the dynamical system has a periodic cycle.

In all of the randomly generated examples presented in Table 1, only very few examples do not have a unique locally stable attractor. In all of those counter-examples, the limiting

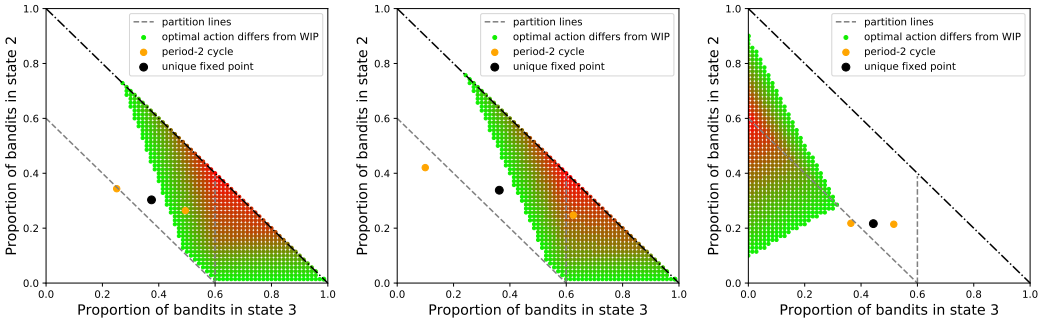
[5] We refer to our git repository for a more thorough numerical exploration of this case.



(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: Fitting subgap(N) with a line in log-scale.

behavior is a cycle of period 2. In this subsection, we illustrate the behavior of restless bandits under WIP in three examples with $d = 3$ and $\alpha = 0.4$. For all of them, the dynamical system $\Phi_{t\geq 0}(\cdot)$ has an attracting cycle of period 2. The fixed point and the two points of the attracting cycle for each example are shown in Figure 4. Note that for these three examples, the matrices $\mathbf{K}_2$'s are not stable and they all have an eigenvalue smaller than -1.



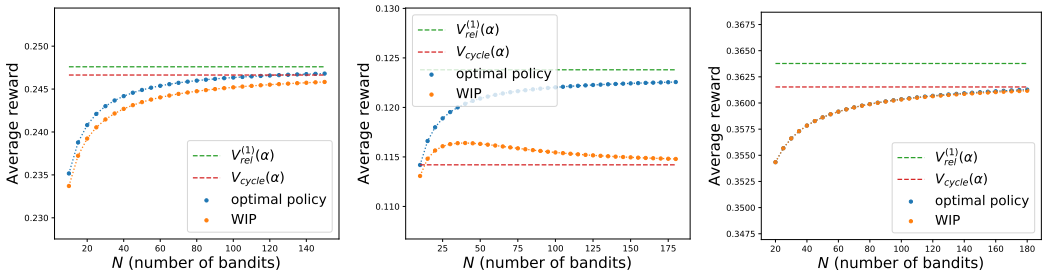
(a) Example 1 (b) Example 2 (c) Example 3

Fig 4: Action differences plot for three period-2 cycle examples with $d = 3$ and $N = 70$.

In Figure 4, we also highlight the configurations in which the optimal policy takes a different action than WIP when the number of bandits is $N = 70$. Such configurations are represented as colored dots, starting with the greenest color, the deeper the red, the more the optimal action deviates from WIP's action on this configuration. The blank area means that on these configurations WIP is an optimal action.

We then plot in Figure 5 the value of the optimal decision rule, $V_{\text{opt}}^{(N)}(\alpha)/N$, and of WIP, $V_{\text{WIP}}^{(N)}(\alpha)/N$, as a function of the number of bandits N. We take multiples of 5 for values of N so that αN are always integers. Several comments are in order:

- As mentioned in Remark 3.2, $V_{\text{WIP}}^{(N)}(\alpha)/N$ converges to the averaged reward on the cycle, denoted here by $V_{\text{cycle}}(\alpha)$, instead of reward on the fixed point $V_{\text{rel}}^{(1)}(\alpha)$. Note that $V_{\text{cycle}}(\alpha)$ is not an upper bound on $V_{\text{WIP}}^{(N)}(\alpha)/N$ and sometimes, as in Example 2, $V_{\text{WIP}}^{(N)}(\alpha)/N$ becomes greater than $V_{\text{cycle}}(\alpha)$ for $N \approx 30$ before decreasing to this value from above.
- The quantity $\left(V_{\text{opt}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha)\right)/N$ converges to $V_{\text{rel}}^{(1)}(\alpha) - V_{\text{cycle}}(\alpha)$, which is strictly positive and might be relatively large, depending on the parameters.



(a) Performance for Example 1. (b) Performance for Example 2. (c) Performance for Example 3.

Fig 5: Performance of optimal policy and WIP for three period-2 cycle examples with $d = 3$.

- The gap $\left(V_{\text{opt}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha)\right)/N$ can be increasing with N, as in Example 2 and 3. This violates the intuition that WIP should be closer to the optimal policy as N grows. It should be contrasted with the uniform global attractor situation for which $\left(V_{\text{opt}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha)\right)/N \to 0$ exponentially fast.

As a final remark, we would like to point out that it is possible to have more complicated shape of attracting limit cycles (of variant periods), as long as $d = 4$, and an unstable matrix $\mathbf{K}$ can have either a pair of conjugate complex eigenvalues or two real eigenvalues with norm bigger than 1.

4.4. *Non integer values of* αN. Our previous analysis rely on the assumption that αN is an integer. Let us briefly discuss in this subsection how to deal with non integer values of αN for the optimization problem (2) under (3). Consider the following three possible modifications for constraint (3):

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

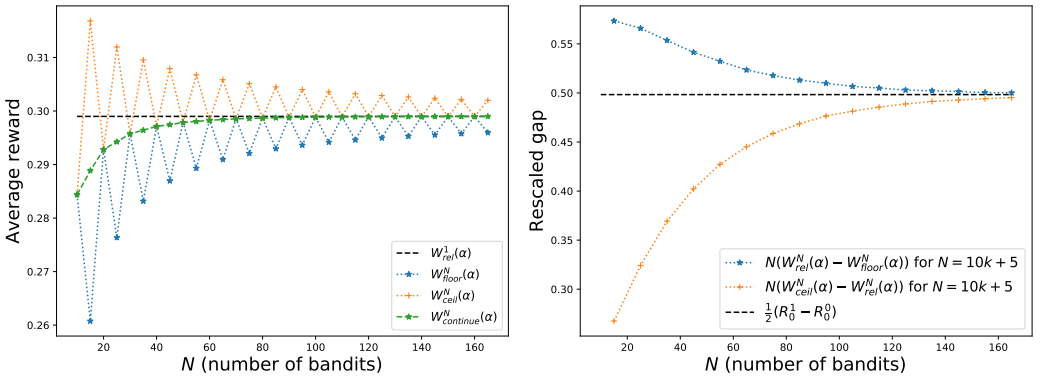
We denote by $V_{\text{WIP}}^{(N)}(\lfloor N\alpha \rfloor/N)$, $V_{\text{WIP}}^{(N)}(\lceil N\alpha \rceil/N)$ and $V_{\text{WIP}}^{(N)}(\overline{\alpha})$ the reward of WIP under these three solutions. Note that these three values all coincide with our previous $V_{\text{WIP}}^{(N)}(\alpha)$ if αN is an integer, but otherwise are different in general. We claim that under the assumptions of Theorem 3.2, the average reward when always activating $\lfloor \alpha N \rfloor$ bandits or always activating $\lceil \alpha N \rceil$ bandits will be at distance $\mathcal{O}(1/N)$ from the relaxation $V_{\text{rel}}^{(N)}(\alpha)$:

(11)
$$\liminf \Delta R \cdot \{\alpha N\} \leq \lim_{N\to\infty} V_{\text{rel}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\lfloor N\alpha \rfloor/N) \leq \limsup \Delta R \cdot \{\alpha N\},$$

(12)
$$\liminf -\Delta R \cdot (1 - \{\alpha N\}) \leq \lim_{N\to\infty} V_{\text{rel}}^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\lceil N\alpha \rceil/N) \leq \limsup -\Delta R \cdot (1 - \{\alpha N\}),$$

where $\Delta R := R^1_{s(\mathbf{m}^*)} - R^0_{s(\mathbf{m}^*)}$. On the contrary, $V_{\text{WIP}}^{(N)}(\overline{\alpha})$ converges at exponential rate to $V_{\text{rel}}^{(N)}(\alpha)$. Indeed, recall the definition of $\phi(\mathbf{m})$, especially equation (7). In Lemma 3.1 we showed that ϕ does not depend on N as long as αN is an integer. Now the modification *(continue)* is actually a natural extension (and is the only one) on the definition of ϕ to all integers N, while keeping the property that its definition does not depend on N for a given α,

in contrast with *(floor)* and *(ceil)*. Consequently a Hoeffding's equality holds in this extension for Φ_t on all N (see Lemma C.3 and Lemma C.4 in the appendix), regardless whether αN is an integer, and it gives an exponential bound in N for the deviation. The rest of the proof then works exactly the same as in our original proof for Theorem 3.2 under this extension.



(a) Performance of the three solutions (b) $\mathcal{O}(1/N)$ convergence rate for $floor$ and $ceil$

Fig 6: Illustration for non integer values of αN on our previous example with $d = 3$.

To illustrate these points, we consider in Figure 6 the same example as in Section 4.2, with $\alpha = 0.3$. As in Figure 2, the green curve represents $V^{(N)}_{\mathrm{WIP}}(\alpha)/N$ for N being a multiple of 10. Here, we extend this curve to all N being a multiple of 5, using the three modifications. The values of $V^{(N)}_{\mathrm{WIP}}(\lfloor N\alpha \rfloor/N)/N$, $V^{(N)}_{\mathrm{WIP}}(\overline{\alpha})/N$ and $V^{(N)}_{\mathrm{WIP}}(\lceil N\alpha \rceil/N)/N$ 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. Now as claimed in (11) and (12), the differences *gap-floor* $V^{(N)}_{\mathrm{rel}}(\alpha) - V^{(N)}_{\mathrm{WIP}}(\lfloor N\alpha \rfloor/N)$ and *gap-ceil* $V^{(N)}_{\mathrm{rel}}(\alpha) - V^{(N)}_{\mathrm{WIP}}(\lceil N\alpha \rceil/N)$ converge to $\pm 0.5 \cdot (R^1_1 - R^0_1)$ when $N \to \infty$ and $\{N\alpha\} \equiv 0.5$, *i.e.* $N = 5 \cdot (2k+1)$. This is indeed verified in Figure 6b. However, the green dots in Figure 6a have their positions along the green curve that is originally plotted only for $N = 10k$. This indicates that $V^{(N)}_{\mathrm{WIP}}(\overline{\alpha})$ converges at exponential rate to $V^{(N)}_{\mathrm{rel}}(\alpha)$.

5. Illustration: Markovian fading channels. The Markovian fading channel is a typical synchronous restless bandit model. In [25] a two-classes channel problem has been studied. By using the same scaling as here, the authors of [25] 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 3.2.

Let us first briefly review this two-classes channel model, more details could be referred from [25]. A Gilbert-Elliott channel is modeled as a two-state Markov chain with a bad state 0 and a good state 1. Two classes of channels are available, with the transition probability matrix for class $k \in \{1, 2\}$ being $\begin{pmatrix} p_k & 1-p_k \\ r_k & 1-r_k \end{pmatrix}$, where p_k is the probability of a class k channel being in good state at time $t+1$ if it was in good state at time t, and r_k is the probability being in good state if one time step ago it was in bad state. We assume the channels are *positively correlated*, namely $p_k > r_k$ for $k = 1, 2$.

We consider a total population of N channels, a proportion β of them are from class 1. Due to limited resource, each time we can only activate a proportion α of the channels, and only a

channel in good state under activation can transmit data. We assume that we can observe the state of a channel only when it is activated. Otherwise, we keep track of the state of a channel by using a belief value $b_{s,t}^k$ where $k = 1, 2$, $s = 0, 1$ and $t \geq 1$. The value $b_{s,t}^k$ 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 $b_{s,t}^k$ is

$$b_{0,t}^k = \frac{r_k - (p_k - r_k)^t r_k}{1 + r_k - p_k},\ b_{1,t}^k = \frac{r_k + (1 - p_k)(p_k - r_k)^l}{1 + r_k - p_k}.$$

To cast this channel model into a synchronous restless bandit problem, we treat each channel as a bandit, and its state space is the whole set of possible values of $b_{s,t}^k$'s. The transition matrices $\mathbf{P}^0$, $\mathbf{P}^1$ can then be naturally written down, *e.g.*

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

all other probabilities being 0.

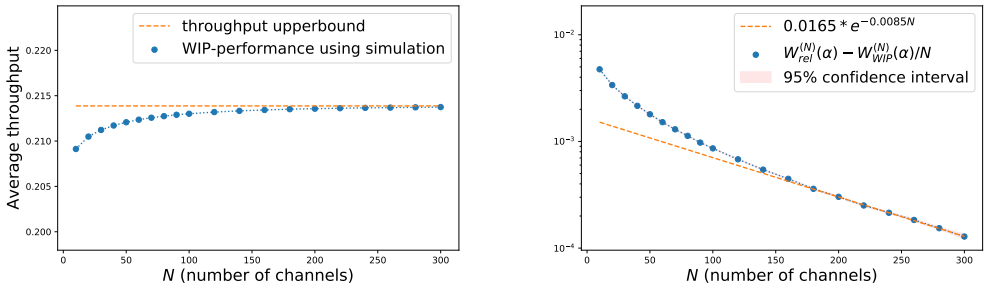
We evaluate the performance by the throughput of the system, hence we obtain a reward of 1 each time we activate a channel and it is in good state. Under the MDP framework, this is equivalent to assuming that state $b_{s,t}^k$ gives a reward $b_{s,t}^k$ under activation. It is shown in [25] that this 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 $b_{s,t}^k$ is denoted by $\nu(b_{s,t}^k)$ and is equal to:

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

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

The reader might have noticed that to apply Theorem 3.2, two assumptions are violated: first, the restless bandit model we consider here has a countable infinite state space; second, not all bandits are identical (there are two classes of bandits). 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 $b_{0,t}^k$ for t large are extremely rarely visited, hence using a threshold t^* and ignoring all states $b_{s,t}^k$ with $t > t^*$ (*i.e.* treating them as b_{s,t^*}^k) makes a negligible difference. Concerning the two classes of bandit, we argue that having two classes of bandits can be represented by a single class of bandit by considering a larger state-space: the state of a bandit would be $(k, b_{s,t}^k)$, where k is its class and $b_{s,t}^k$ is its belief value. Compared to our model, in this new case, the bandits are no longer unichain as a bandit of class k cannot become a bandit of class $k' \neq k$. This implies that the quantities $V_{\mathrm{WIP}}^{(N)}(\alpha)$ and $V_{\mathrm{rel}}^{(N)}(\alpha)$ will depend on the initial condition of the system, *i.e.* on the fraction β of bandits that are in class 1. Apart from that, our results apply mutatis mutandis to this case.

We can now provide some numerical results. We shall choose a parameter set that is used in [25]: $\beta = 0.6$, $\alpha = 0.3$, $(p_1, r_1) = (0.75, 0.2)$, $(p_2, r_2) = (0.8, 0.3)$. 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.



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

Fig 7: Convergence rate for two-classes channel model.

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 3.2 are satisfied for this model. We then use simulations to evaluate the average throughput, with N ranging from 10 to 300. We see through Figure 7 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. Extension to the asynchronous model. Throughout the paper, we studied a *synchronous* restless bandit problem in which all bandits synchronously make a transition. In this section, we explain how to adapt the proofs done in Section 3 to the asynchronous model studied for example in [30]. We start by recalling the model of [30] in Section 6.1. We show how the synchronous and asynchronous models are related in Section 6.2. Finally, we state the equivalent of our main Theorem 3.2 in Section 6.3.

6.1. *The asynchronous bandit model.* Similarly to Section 2.1, an asynchronous restless bandit problem with parameters $\{(\mathbf{Q}^0, \mathbf{Q}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha, N\}$ is a Markov decision process defined as follows:

1. As before, the model is composed of N bandits 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. Bandit n evolves as a continuous-time Markov chain of kernel $\mathbf{Q}^{a_n(t)}$, *i.e.* $\mathbf{Q}^0$ 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 $i \neq j$, bandit n jumps from state i to state j at rate $Q^{a_n(t)}_{ij}$. Given $\mathbf{a}(t)$, the evolutions of the N bandits are independent.
4. The gain per unit time of the decision maker is $\sum_{n=1}^N R^{a_n(t)}_{S_n(t)}$.

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

$$V^{(N)}_{\text{opt}}(\alpha) := \sup_{\mathbf{a}\in\Pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\int_{t=0}^{T} \sum_{n=1}^{N} R^{a_n(t)}_{S_n(t)} dt\Big] \tag{13}$$

$$\text{subject to} \sum_{n=1}^{N} a_n(t) = \alpha N, \text{ for all } t \in \mathbb{R}_+. \tag{14}$$

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

6.2. *Whittle index, relaxation and equivalence with the synchronous model.* In this subsection, we recall briefly the definition of Whittle index and of the relaxation for the asynchronous bandit model. These definitions coincide with the ones of [30].

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

Similarly to the synchronous problem, the definition of Whittle index in the asynchronous model can be justified by looking at the Lagrangian of the optimization problem (13) where the constraint (14) is replaced by the constraint (16) below. We again denote the value of this relaxed problem as $V_{\text{rel}}^{(N)}(\alpha)$.

$$V_{\text{rel}}^{(N)}(\alpha) := \sup_{\mathbf{a}\in\Pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\int_{t=0}^{T} \sum_{n=1}^{N} R_{S_n(t)}^{a_n(t)} dt\Big] \tag{15}$$

$$\text{subject to} \lim_{T\to\infty} \int_0^T \sum_{n=1}^{N} \frac{a_n(t)}{T} dt = \alpha N. \tag{16}$$

As we show below, when considering bandits in isolation, using a synchronous or an asynchronous model is equivalent. In particular, neither the definition of Whittle index nor the value of the relaxation depend on the synchronization nature of a bandit.

DEFINITION 6.1. Let $(\mathbf{Q}^0, \mathbf{Q}^1, \mathbf{R}^0, \mathbf{R}^1)$ be the parameters of an asynchronous bandit. Let $\tau := \max_i \max_a |Q_{ii}^a|$. 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 $a \in \{0, 1\}$:

$$P_{ij}^a := \frac{1}{\tau} Q_{ij}^a; \qquad P_{ii}^a := 1 - \sum_{k\neq i} P_{ik}^a; \qquad \tilde{R}_i^a := \tau R_i^a. \tag{17}$$

We call $(\mathbf{P}^0, \mathbf{P}^1, \tilde{\mathbf{R}}^0, \tilde{\mathbf{R}}^1)$ the synchronous version of our asynchronous bandit model.

The following lemma states the equivalence of Whittle relaxation between the synchronous and the asynchronous problems:

LEMMA 6.1. Let $(\mathbf{Q}^0, \mathbf{Q}^1, \mathbf{R}^0, \mathbf{R}^1)$ be an asynchronous bandit and let $(\mathbf{P}^0, \mathbf{P}^1, \tilde{\mathbf{R}}^0, \tilde{\mathbf{R}}^1)$ be its synchronous version (17). Then:

(i) The matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ are probability matrices.
(ii) The synchronous bandit $(\mathbf{P}^0, \mathbf{P}^1, \tilde{\mathbf{R}}^0, \tilde{\mathbf{R}}^1)$ is indexable if and only if the asynchronous 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 synchronous relaxed optimization problem (4) has the same value as its asynchronous counterpart (15).

The proof of Lemma 6.1 is a direct consequence of uniformization: the results rely on analysis of a bandit in isolation; when focus on a single bandit, Bellman's equation is identical for the synchronous and asynchronous version of the MDP.

6.3. *Exponential convergence in the case of asynchronous model.* Lemma 6.1 uses the fact that the Whittle relaxation is defined for bandit in isolation. Hence, considering synchronous or asynchronous bandits is equivalent. For the N bandits model, however, the situation is different: in the synchronous model of Section 2 all bandits change states at the same time, while in asynchronous situation, the probability that two bandits makes a jump at the exact same time is 0. This implies that the reward of WIP for the N bandits 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 [30] that the asymptotic optimality depends on the ergodic property of the solution of an ordinary differential equation (ODE) defined in Equation (10) of [30]. Using our notation, this differential equation can be written as

$$\dot{\mathbf{m}} = \tau(\phi(\mathbf{m}) - \mathbf{m}), \tag{18}$$

where ϕ is defined as in Lemma 3.1 for a synchronous bandit problem $(\mathbf{P}^0, \mathbf{P}^1, \tilde{\mathbf{R}}^0, \tilde{\mathbf{R}}^1)$. By Lemma 3.1, this equation has a unique fixed point, $\mathbf{m}^*$. It is shown in [30] that if all the solutions of the differential equation (18) 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 extend the result of Theorem 3.2 to the asynchronous model.

THEOREM 6.2 (Exponential convergence rate theorem-asynchronous case). Consider an asynchronous restless bandit problem $\{(\mathbf{Q}^0, \mathbf{Q}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha\}$ such that:

(i) Bandits are unichain and indexable.
(ii) The unique fixed point $\mathbf{m}^*$ of the ODE $\dot{\mathbf{m}} = \tau(\phi(\mathbf{m}) - \mathbf{m})$ is a uniform global attractor of the trajectories of the ODE.
(iii) $\mathbf{m}^*$ is not singular.

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 α, such that for any N with αN being an integer,

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

SKETCH OF PROOF. The proof of this result follows the same structure as the proof of Theorem 3.2 but needs substantial adaptation, the full details are given in Appendix D. The main ingredients are:

- We first use a result from [5] to obtain an analogue of Hoeffding's inequality. This proves that the behavior of the N bandits model is close to the dynamic of the ODE (18).
- 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 ϕ 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 [30, 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: [30, Theorem 2] also needs the assumption that $\mathbf{m}^*$ is an attractor of the ODE. We add in addition that the attractivity is uniform in the initial points and that $\mathbf{m}^*$ is not singular. Those conditions are also similar to the conditions of Theorem 3.2. It would be tempting to say that an asynchronous bandit satisfies the assumptions of Theorem 6.2 if and only if its synchronous version satisfies the assumptions of Theorem 3.2. Yet, this is true for the indexability and the non-singularity of

the fixed point, but does not hold in general, since 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 $\dot{\mathbf{m}} = \tau\big(\phi(\mathbf{m}) - \mathbf{m}\big)$. Consider for instance the synchronous model of Section 4.3: the discrete-time dynamical system has an attracting cycle of period 2; in the asynchronous version of this model, $\mathbf{m}^*$ is a uniform global attractor. [6] This indicates that there are bandit models for which WIP is asymptotically optimal under an asynchronous scaling but is not for the synchronous scaling.

7. Conclusion and future work. In this paper, we studied the performance of Whittle index policy (WIP) when there is a large number of bandits. We showed that, when WIP becomes asymptotically optimal, it does so at exponential rate (unless the fixed point is singular, which occurs with probability 0). This explains why WIP is very efficient in practice, even when the number of bandits remains moderate. Our results hold for the classical model of [30] where bandits evolve asynchronously as well as for a synchronous model in which all bandits make their transitions simultaneously. Yet, while Whittle indices are the same for both models, we provided examples for which WIP is asymptotically optimal for the asynchronous model but is not for the synchronous model.

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 [7, 14, 29].

REFERENCES

[1] AALTO, S., LASSILA, P. and 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* 57–69.

[2] ANSELL, P., GLAZEBROOK, K. D., NINO-MORA, J. and O'KEEFFE, M. (2003). Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research* **57** 21–39.

[3] AVRACHENKOV, K. E. and BORKAR, V. S. (2016). Whittle index policy for crawling ephemeral content. *IEEE Transactions on Control of Network Systems* **5** 446–455.

[4] BUBECK, S. and CESA-BIANCHI, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *CoRR* **abs/1204.5721**.

[5] DARLING, R. W. R. and NORRIS, J. R. (2008). Differential equation approximations for Markov chains. *Probab. Surveys* **5** 37–79.

[6] DUFF, M. O. (1995). Q-learning for Bandit Problems. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*. *ICML'95* 209–217. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[7] DURAN, S. and VERLOOP, M. (2018). Asymptotic optimal control of Markov-modulated restless bandits. In *International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS 2018)* **2** 7:1–7:25. ACM : Association for Computing Machinery, Irvine, US.

[8] 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 . ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems , SIGMETRICS '17* **1** 26.

[9] GAST, N., BORTOLUSSI, L. and TRIBASTONE, M. (2018). Size Expansions of Mean Field Approximation: Transient and Steady-State Analysis. In *2018 - 36th International Symposium on Computer Performance, Modeling, Measurements and Evaluation* 1-2.

[10] GAST, N., LATELLA, D. and MASSINK, M. (2018). A refined mean field approximation of synchronous discrete-time population models. *Performance evaluation* **126** 1–21.

[6] From (18) we infer that, if $\lambda_{\text{sync.}}$ is an eigenvalue of an affine factor $\mathbf{K}$ and $\lambda_{\text{async.}}$ is its asynchronous counterpart, then $\tau \cdot (\lambda_{\text{sync.}} - 1) = \lambda_{\text{async.}}$. So $\lambda_{\text{sync.}} < -1$ for the examples in Section 4.3 implies $\Re(\lambda_{\text{async.}}) < 0$ for the asynchronous models, and $\lambda_{\text{async.}}$ is a stable eigenvalue (for the continuous-time ODE).

[11] GAST, N. and VAN HOUDT, B. (2017). A Refined Mean Field Approximation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* **1**.
[12] GITTINS, J., GLAZEBROOK, K. and WEBER, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.
[13] GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B* 148–177.
[14] HODGE, D. J. and GLAZEBROOK, K. D. (2015). On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Adv. in Appl. Probab.* **47** 652–667.
[15] HSU, Y.-P. (2018). Age of information: Whittle index for scheduling stochastic arrivals. In *2018 IEEE International Symposium on Information Theory (ISIT)* 2634–2638. IEEE.
[16] KIFER, Y. (1988). *Random Perturbations of Dynamical Systems*. *Progress in Probability*. Birkhäuser Boston.
[17] KOLOKOLTSOV, V. N., LI, J. and YANG, W. (2011). Mean Field Games and Nonlinear Markov Processes Technical Report, ArXiv e-prints.
[18] KURTZ, T. G. (1978). Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and their Applications* **6** 223–240.
[19] LARRNAAGA, M., AYESTA, U. and VERLOOP, I. M. (2016). Dynamic control of birth-and-death restless bandits: Application to resource-allocation problems. *IEEE/ACM Transactions on Networking* **24** 3812–3825.
[20] LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit algorithms*. Cambridge University Press.
[21] LIU, K. and ZHAO, Q. (2010). Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory* **56** 5547–5567.
[22] MESHRAM, R., MANJUNATH, D. and GOPALAN, A. (2018). On the Whittle index for restless multiarmed hidden Markov bandits. *IEEE Transactions on Automatic Control* **63** 3046–3053.
[23] NIÑO-MORA, J. (2007). Dynamic priority allocation via restless bandit marginal productivity indices. *Top* **15** 161–198.
[24] NIÑO-MORA, J. and VILLAR, S. S. (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)* 1–8. IEEE.
[25] OUYANG, W., ERYILMAZ, A. and SHROFF, N. B. (2012). Asymptotically optimal downlink scheduling over Markovian fading channels. In *2012 Proceedings IEEE INFOCOM* 1224–1232. IEEE.
[26] PAPADIMITRIOU, C. H. and TSITSIKLIS, J. N. (1999). The Complexity Of Optimal Queuing Network Control. *Math. Oper. Res* 293–305.
[27] PUTERMAN, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st ed. John Wiley & Sons, Inc., New York, NY, USA.
[28] RAGHUNATHAN, V., BORKAR, V., CAO, M. and KUMAR, P. R. (2008). Index policies for real-time multicast scheduling for wireless broadcast systems. In *IEEE INFOCOM 2008-The 27th Conference on Computer Communications* 1570–1578. IEEE.
[29] VERLOOP, M. (2016). Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *Annals of Applied Probability* **26** 1947–1995.
[30] WEBER, R. R. and WEISS, G. (1990). On an Index Policy for Restless Bandits. *Journal of Applied Probability* **27** 637–648.
[31] WEBER, R. R. and WEISS, G. (1991). Addendum to 'On an index policy for restless bandits'. *Advances in Applied probability* **23** 429–430.
[32] WHITTLE, P. (1988). Restless Bandits: activity allocation in a changing world. *Journal of Applied Probability* **25A** 287–298.
[33] YING, L. (2017). Stein's method for mean field approximations in light and heavy traffic regimes. *POMACS* **1** 1-27.

APPENDIX A: PROOF OF THEOREM 2.1

PROOF. Let $\mathbf{m}^*$ be the fixed point of ϕ. As $\mathbf{P}^0$, $\mathbf{P}^1$ are rational, each coordinate of $\mathbf{m}^*$ is a rational number. Note that the proof could be adapted to non rational $\mathbf{P}^0, \mathbf{P}^1$ by using a continuity argument.

Let $\{N_k\}_{k\geq 0}$ be a sequence of increasing integers that goes to ∞, such that for all $k \geq 0$ and all $1 \leq i \leq d$, $m_i^* N_k$ and αN_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 bandits in state i in configuration $\mathbf{m}$ and that $S_n(t)$ is the state of bandit n at time t. We use $\mathbf{S}(t)$ to denote the state vector of the N bandits

system at time t. Let $\mathbf{S}^*$ be a state vector corresponds to configuration $\mathbf{m}^*$ with N bandits. This is possible as $m_i^* N$ is an integer for all $i \in \{1 \dots d\}$.

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* αN bandits. As $\mathbf{a}^*$ is sub-optimal compared to an optimal policy for the original N bandit problem (2)-(3), we have

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

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 [27, Equation (8.4.2) in Chapter 8]), the second line is because $\mathbf{a}^*$ is a valid action for the N-bandits 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

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

In the following, we bound $\mathbb{E}_{\mathbf{a}^*}[V\big(\mathbf{S}(1)\big) - V(\mathbf{S}^*)]$. 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 bandits) of bandits 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 $\mathbf{z}$,

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

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

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

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 bandit 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 bandits system, $\mathbf{Y}$ and $\mathbf{Z}$, that start respectively in state vectors $\mathbf{Y}(0) = \mathbf{y}$ and $\mathbf{Z}(0) = \mathbf{z}$. Let π^* be the optimal policy of the N bandits MDP, and suppose that we apply π^* 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 bandits $n \in \{2 \dots N\}$. The $\mathbf{Z}$ trajectory follows an optimal trajectory, hence Bellman's equation is satisfied: for any $T > 0$, we have:

$$T \cdot V_{\text{opt}}^{(N)}(\alpha) + V(\mathbf{z}) = \sum_{n=1}^N R_{z_n}^{\pi_n^*(\mathbf{z})} + \mathbb{E}_{\pi^*} \left[\sum_{n=1}^N \sum_{t=1}^{T-1} R_{Z_n(t)}^{\pi_n^*(\mathbf{Z}(t))} + V\big(\mathbf{Z}(T)\big) \;\Big|\; \mathbf{Z}(0) = \mathbf{z} \right]. \tag{19}$$

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

$$(20)\quad T\cdot V_{\rm opt}^{(N)}(\alpha)+V(\mathbf{y})\geq\sum_{n=1}^{N}R_{y_n}^{\pi_n^*(\mathbf{y})}+\mathbb{E}_{\pi^*}\left[\sum_{n=1}^{N}\sum_{t=1}^{T-1}R_{Y_n(t)}^{\pi_n^*(\mathbf{Z}(t))}+V\big(\mathbf{Y}(T)\big)\;\Big|\;\mathbf{Y}(0)=\mathbf{y}\right],$$

Recall that the matrices $\mathbf{P}^0,\mathbf{P}^1$ are such that a bandit is unichain and aperiodic. This shows that the mixing time of a single bandit 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 a bandit 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 (19) in (20) gives

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

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

Step Two. 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

$$X_i=(B_{i,1}^0+B_{i,1}^1)+(B_{i,2}^0+B_{i,2}^1)+\ldots+(B_{i,d}^0+B_{i,d}^1)$$

where $B_{i,j}^a\sim Binomial(x_{j,a}^*,P_{ji}^a)$ 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 bandits in state j taking action a, when optimal action vector $\mathbf{a}^*$ is applied to state vector $\mathbf{S}^*$.

By stationarity, we have

$$\mathbb{E}_{\mathbf{a}^*}(X_i)=\sum_{j=1}^{d}\sum_{a=0,1}x_{j,a}^*\cdot P_{ji}^a=x_i^*,$$

and

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

Consequently, we can bound

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

with a constant D independent of N.

To summarize, we have

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

hence

$$V^{(1)}_{\text{rel}}(\alpha) \geq \frac{V^{(N)}_{\text{opt}}(\alpha)}{N} = V^{(1)}_{\text{rel}}(\alpha) + \frac{\mathbb{E}_{\mathbf{a}^*} V(\mathbf{S}(1)) - V(\mathbf{S}^*)}{N} \geq V^{(1)}_{\text{rel}}(\alpha) - \frac{CD}{\sqrt{N}}, \tag{21}$$

which implies that $V^{(N)}_{\text{opt}}(\alpha)/N \to V^{(1)}_{\text{rel}}(\alpha)$ when N goes to $+\infty$. Moreover, from (21), the convergence rate is at least as fast as $\mathcal{O}(1/\sqrt{N})$. □

APPENDIX B: PROOF OF LEMMA 3.1

In this appendix we prove Lemma 3.1. We first show the piecewise affine property in Lemma B.1, which gives (i) and (ii). We then show the uniqueness of fixed point from a bijective property in Lemma B.2, from which we conclude (iii).

LEMMA B.1 (Piecewise affine). ϕ is a piecewise affine continuous function, with d affine pieces.

PROOF. Let $\mathbf{m} \in \Delta^d$ be a configuration and recall $s(\mathbf{m}) \in \{1 \dots d\}$ 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 bandits that are in states 1 to $s(\mathbf{m}) - 1$ and not activate any bandit in states $s(\mathbf{m}) + 1$ to d. Among the $Nm_{s(\mathbf{m})}$ bandits 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 bandits in state j at time $t+1$ will be equal to

$$\sum_{i=1}^{s(\mathbf{m})-1} Nm_i P^1_{ij} + N(\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i) P^1_{s(\mathbf{m})j} + N(\sum_{i=1}^{s(\mathbf{m})} m_i - \alpha) P^0_{s(\mathbf{m})j} + \sum_{i=s(\mathbf{m})+1}^{d} Nm_i P^0_{ij}. \tag{22}$$

It justifies the expression (7). Note that (7) can be reorganized to

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

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

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

Let $\mathcal{Z}_i := \{\mathbf{m} \in \Delta^d \mid s(\mathbf{m}) = i\}$. The above expression of ϕ 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 B.2 (Bijectivity). Let $\pi(s,\theta) \in \Pi$ be the policy that activates all bandits in states $1,\dots,s-1$, does not activate bandits in states $s+1, s+2, \dots, d$, and that activates bandits in state s with probability θ. Denote by $\tilde{\alpha}(s,\theta)$ the proportion of time that the active action is taken using policy $\pi(s,\theta)$. Then, the function $(s,\theta) \mapsto \tilde{\alpha}(s,\theta)$ is a bijective map from $\{1\dots d\} \times [0,1)$ to $[0,1)$.

PROOF. The following proof is partially adapted from the proof of [30, Lemma 1]. For a given $\nu \in \mathbb{R}$, denote by $\gamma(\nu)$ the value of the subsidy-ν problem, *i.e.*

$$\gamma(\nu) := \sup_{\pi\in\Pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1}\Big(R_{S(t)}^{\pi(S(t))} + \nu\big(1-\pi(S(t))\big)\Big)\Big]. \tag{23}$$

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

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

Moreover, for $s \in \{2\dots d-1\}$ and $\nu \in [\nu_s;\nu_{s-1}]$, the optimal policy of (23) is to activate all bandits 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

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

and

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

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

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

which concludes the proof. □

We are now ready to finish the proof of Lemma 3.1(iii). Let $\mathbf{m}$ be a fixed point of ϕ. Under configuration $\mathbf{m}$, all bandits 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 bandits 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 bandits of this policy is $\tilde{\alpha}(s(\mathbf{m}),\theta(\mathbf{m})) = \alpha$. Consequently, if $\mathbf{m}'$ is another fixed point of ϕ, then $\mathbf{m}'$ would have to be the stationary distribution of some other policy of the form $\pi(s',\theta')$, with $\tilde{\alpha}(s',\theta') = \alpha$. As the function $(s,\theta) \mapsto \tilde{\alpha}(s,\theta)$ is a bijection, this implies that $s' = s(\mathbf{m})$ and $\theta' = \theta(\mathbf{m})$. Hence the fixed point of ϕ is unique.

APPENDIX C: PROOF OF THEOREM 3.2

In this appendix, we explain technical details of the proof of our main result Theorem 3.2. In the following, we denote by $\mathcal{B}(\mathbf{m}^*, r)$ the ball centered at $\mathbf{m}^*$ with radius r.

THEOREM C.1. Under the same assumptions as in Theorem 3.2, and assume that $\mathbf{M}^{(N)}(0)$ is already in stationary regime. Then there exists two constants $b, c > 0$ such that

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

Let us first explain how Theorem C.1 implies Theorem 3.2. We prove below that:

LEMMA C.2. Assume that bandits are indexable and unichain, and let $\rho(\mathbf{m})$ be the instantaneous bandit-averaged reward of WIP when the system is in configuration $\mathbf{m}$. Then:

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

(24)
$$\rho(\mathbf{m}) = \sum_{i=1}^{s(\mathbf{m})-1} m_i R_i^1 + (\alpha - \sum_{i=1}^{s(\mathbf{m})-1} m_i) R_{s(\mathbf{m})}^1 + (\sum_{i=1}^{s(\mathbf{m})} m_i - \alpha) R_{s(\mathbf{m})}^0 + \sum_{i=s(\mathbf{m})+1}^{d} m_i R_i^0.$$

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

By definition, the performance of WIP is $V_{\text{WIP}}^{(N)}(\alpha) = N \cdot \mathbb{E}\left[\rho(\mathbf{M}^{(N)}(0))\right]$. Hence from Lemma C.2 we have

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

By linearity of ρ and Theorem C.1(i), the first term inside the above expectation is exponentially small; by Theorem C.1(ii) and since the rewards are bounded, the second term is also exponentially small.

Before proving Theorem C.1, we start by proving a few technical lemmas.

C.1. Relation between $\mathbf{m}^*$ and $V_{\text{rel}}^{(1)}(\alpha)$ (Proof of Lemma C.2).

PROOF. Let $\mathbf{m} \in \Delta^d$ be a configuration and recall $s(\mathbf{m}) \in \{1 \dots d\}$ is the state such that $\sum_{i=1}^{s(\mathbf{m})-1} m_i \leq \alpha < \sum_{i=1}^{s(\mathbf{m})} m_i$. Similarly to our analysis of Lemma B.1, when the system is in configuration $\mathbf{m}$, WIP will activate all bandits 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 bandits 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 $N m_{s(\mathbf{m})}$ bandits 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 (24).

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:

$$\begin{aligned} L &:= \max_{\mathbf{a}\in\Pi} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[R_{S_n(t)}^{a_n(t)} + (\alpha - a_n(t)) \nu_{s(\mathbf{m}^*)} \right] \\ &= \max_{\mathbf{a}\in\Pi} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[R_{S_n(t)}^{a_n(t)} \right] + \left(\alpha - \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[a_n(t) \right] \right) \nu_{s(\mathbf{m}^*)}. \end{aligned} \tag{25}$$

By definition of Whittle index, any policy of the form $\pi(s(\mathbf{m}^*), \theta)$ defined in Lemma B.2 is optimal for (25). Moreover, if θ^* 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 bandits 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 αN bandits are activated on average. Indeed, the bandit-averaged reward of this policy is:

$$L = \sum_{i=1}^{s(\mathbf{m}^*)-1} m_i^* R_i^1 + \theta^* m_{s(\mathbf{m}^*)}^* R_{s(\mathbf{m}^*)}^1 + (1-\theta^*) m_{s(\mathbf{m}^*)}^* R_{s(\mathbf{m}^*)}^0 + \sum_{i=s(\mathbf{m}^*)+1}^{d} m_i^* R_i^0 \tag{26}$$

As the proportion of activated bandits is α, we have $\sum_{i=1}^{s(\mathbf{m}^*)-1} m_i^* + \theta^* m_{s(\mathbf{m}^*)}^* = \alpha$. Hence (26) coincides with the expression of $\rho(\mathbf{m}^*)$ in (24), and $\rho(\mathbf{m}^*) = L = V_{\text{rel}}^{(1)}(\alpha)$. This concludes the proof of Lemma C.2.

□

C.2. Hoeffding's inequality (for one transition).

LEMMA C.3 (Hoeffding's inequality). For all $t \in \mathbb{N}$, we have

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

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

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

and for all $\delta > 0$:

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

PROOF. Since the N bandits 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 a bandit in state i goes to state j under WIP, when the N bandits 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

$$\begin{aligned}\mathbb{P}\left[\|\mathbf{M}^{(N)}(t+1) - \phi(\mathbf{M}^{(N)}(t))\| \geq \delta\right] &= \mathbb{P}\left[\max_{1\leq j\leq d}\left|M_j^{(N)}(t+1) - \phi_j(\mathbf{M}^{(N)}(t))\right| \geq \delta\right] \\ &\leq e^{-2N\delta^2} \quad \text{(By the above form of Hoeffding's inequality).}\end{aligned}$$

□

C.3. Hoeffding's inequality (for t transitions).

LEMMA C.4. There exists a positive constant K such that for all $t \in \mathbb{N}$ and for all $\delta > 0$,

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

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

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

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

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

By iterating the above inequality, we obtain

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

where for each $0 \leq s \leq t$, we have by lemma C.3: 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 C.4. □

C.4. Exponential stability of $\mathbf{m}^*$.

LEMMA C.5. Under the assumptions of Theorem 3.2:

(i) The matrix $K_{s(\mathbf{m}^*)}$ is a stable matrix.
(ii) There exist constants $b_1, b_2 > 0$ such that for all $t \geq 0$ and all $\mathbf{m} \in \Delta^d$:

$$\|\Phi_t(\mathbf{m}) - \mathbf{m}^*\| \leq b_1 \cdot e^{-b_2 t} \cdot \|\mathbf{m} - \mathbf{m}^*\|. \tag{27}$$

PROOF. As $\mathbf{m}^*$ is non-singular and is a uniform global attractor, there exists $T > 0$ such that for all $\mathbf{m} \in \Delta^d$ and $t \geq T$: $\Phi_t(\mathbf{m}) \in \mathcal{Z}_{s(\mathbf{m}^*)}$. Recall that for all $\mathbf{m} \in \mathcal{Z}_{s(\mathbf{m}^*)}$, we have $\phi(\mathbf{m}) = (\mathbf{m} - \mathbf{m}^*) \cdot \mathbf{K}_{s(\mathbf{m}^*)} + \mathbf{m}^*$. This shows that for all $\mathbf{m} \in \Delta^d$ and $t \geq T$:

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

It implies in particular that $\mathbf{K}_{s(\mathbf{m}^*)}$ is a stable matrix, *i.e.* the norm of all of its eigenvalues is smaller than 1. Consequently (27) is true for all $\mathbf{m} \in \mathcal{Z}_{s(\mathbf{m}^*)}$. As $\mathcal{Z}_{s(\mathbf{m}^*)}$ is reached in a finite time T from any initial condition $\mathbf{m}$, this implies (27) for all $\mathbf{m} \in \Delta^d$. □

C.5. Proof of Theorem C.1.

We are now ready to prove the main theorem.

PROOF. The proof consists of several parts.

C.5.1. *Choice of a neighborhood $\mathcal{N}$.* The fixed point $\mathbf{m}^*$ is in zone $\mathcal{Z}_{s(\mathbf{m}^*)}$ in which ϕ can be written as

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

As $\mathbf{m}^*$ is not singular, let $\mathcal{N}_1$ be a neighborhood of $\mathbf{m}^*$ included in $\mathcal{Z}_{s(\mathbf{m}^*)}$. Since $\mathbf{m}^*$ is a uniformly global attractor, $\mathbf{K}_{s(\mathbf{m}^*)}$ is a stable matrix. We can therefore choose a smaller neighborhood $\mathcal{N}_2 \subset \mathcal{N}_1$ so that $\Phi_t(\mathcal{N}_2) \subset \mathcal{N}_1$ for all $t \geq 0$. That is, the image of $\mathcal{N}_2$ under the maps $\Phi_{t\geq 0}$ remains inside $\mathcal{N}_1$. This is possible by stability of $\mathbf{m}^*$. We next choose a

neighborhood $\mathcal{N}_3 \subset \mathcal{N}_2$ and a $\delta > 0$ so that $(\phi(\mathcal{N}_3))^\delta \subset \mathcal{N}_2$, that is, the image of $\mathcal{N}_3$ under ϕ remains inside $\mathcal{N}_2$ and it is at least δ away from the boundary of $\mathcal{N}_2$. We finally fix $r > 0$ so that the intersection $\mathcal{B}(\mathbf{m}^*, r) \cap \Delta^d \subset \mathcal{N}_3$, and we choose our neighborhood $\mathcal{N}$ as

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

Note that the choice of r and δ is independent of N. From (ii) of Lemma C.5, 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.5.2. *Definition and properties of the function G.*. Following the generator approach used for instance in [10]. For $\mathbf{m} \in \Delta^d$, define $G : \Delta^d \to \mathbb{R}^d$ as

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

By using Lemma C.5(ii), for all $\mathbf{m} \in \Delta^d$ we have $\|G(\mathbf{m})\| \le \sum_{t=0}^{\infty} b_1 \cdot e^{-b_2 t} \cdot \|\mathbf{m} - \mathbf{m}^*\| < \infty$. 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 $t \ge 0$ and $\mathbf{m} \in \mathcal{N}_2$ we have:

$$\Phi_t(\mathbf{m}) = (\mathbf{m} - \mathbf{m}^*) \cdot \mathbf{K}^t_{s(\mathbf{m}^*)} + \mathbf{m}^*. \tag{28}$$

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

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

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

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

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

Hence

$$\begin{aligned} \mathbb{E}[\mathbf{M}^{(N)}(0)] - \mathbf{m}^* &= \mathbb{E}\big[G(\mathbf{M}^{(N)}(0)) - G(\phi(\mathbf{M}^{(N)}(0)))\big] \quad \text{(By the above equality)} \\ &= \mathbb{E}\big[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{M}^{(N)}(0)))\big] \quad \text{(Since } \mathbf{M}^{(N)}(0) \text{ is stationary)} \\ &= \mathbb{E}\Big[\mathbb{E}\big[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{m})) \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\big] \cdot \mathbb{1}_{\{\mathbf{m} \notin \mathcal{N}\}} \qquad (29) \\ &\quad + \mathbb{E}\big[G(\mathbf{M}^{(N)}(1)) - G(\phi(\mathbf{m})) \mid \mathbf{M}^{(N)}(0) = \mathbf{m}\big] \cdot \mathbb{1}_{\{\mathbf{m} \in \mathcal{N}\}}\Big]. \qquad (30) \end{aligned}$$

In the following, we bound (29) and (30) separately.

C.5.3. *Bound on* (29). As G is bounded by $\overline{G}$, we have

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

We are left to bound $\mathbb{P}\left[\mathbf{M}^{(N)}(0)\notin\mathcal{N}\right]$. Let $u := \left(\frac{r}{2(1+K+K^2+\ldots+K^{\tilde{T}})}\right)^2$, where K is the Lipschitz constant of ϕ. We have by Lemma C.4:

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

This shows that

$$\begin{aligned}
\mathbb{P}\left[\mathbf{M}^{(N)}(0)\notin\mathcal{N}\right] &= \mathbb{P}\left[\|\mathbf{M}^{(N)}(0)-\mathbf{m}^*\|\geq r\right]\\
&= \mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1)-\mathbf{m}^*\|\geq r\right] \quad \text{(By stationarity)}\\
&\leq \mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1) - \Phi_{\tilde{T}+1}(\mathbf{m})\| \geq \frac{r}{2} \;\middle|\; \mathbf{M}^{(N)}(0)=\mathbf{m}\right] + \mathbb{P}\left[\|\Phi_{\tilde{T}+1}(\mathbf{m})-\mathbf{m}^*\|\geq\frac{r}{2}\right]\\
&\quad (\text{with } \mathbf{m} \text{ being an arbitrary element of } \Delta^d)\\
&= \mathbb{P}\left[\|\mathbf{M}^{(N)}(\tilde{T}+1) - \Phi_{\tilde{T}+1}(\mathbf{m})\| \geq \frac{r}{2} \;\middle|\; \mathbf{M}^{(N)}(0)=\mathbf{m}\right] \quad \text{(By our choice of } \tilde{T}=T(r/2))\\
&\leq (\tilde{T}+1)\cdot e^{-2uN}
\end{aligned} \tag{31}$$

C.5.4. *Bound on* (30). By Lemma C.3, we have

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

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

$$\begin{aligned}
&\mathbb{E}\Big[\big(G(\phi(\mathbf{m})+\mathbf{E}^{(N)}(1)) - G(\phi(\mathbf{m}))\big)\cdot\mathbb{1}_{\{\|\mathbf{E}^{(N)}(1)\|<\delta\}} \;\Big|\; \mathbf{M}^{(N)}(0)=\mathbf{m}\Big]\cdot\mathbb{1}_{\{\mathbf{m}\in\mathcal{N}\}}\\
&= \Big[G\big(\mathbb{E}\big[\phi(\mathbf{m})+\mathbf{E}^{(N)}(1) \;\big|\; \mathbf{M}^{(N)}(0)=\mathbf{m}\big]\big) - G\big(\mathbb{E}\big[\phi(\mathbf{m}) \;\big|\; \mathbf{M}^{(N)}(0)=\mathbf{m}\big]\big)\Big]\mathbb{P}\big(\{\|\mathbf{E}^{(N)}(1)\|<\delta\}\big)\cdot\mathbb{1}_{\{\mathbf{m}\in\mathcal{N}\}}\\
&\big(\text{Thanks to the affinity of } G \text{ in this case, we can interchange } \mathbb{E} \text{ and } G\big)\\
&= 0 \qquad \big(\text{By Lemma C.3}\big).
\end{aligned}$$

For the second part of the above expectation,

$$\begin{aligned}
&\left|\left|\mathbb{E}\left[\big(G(\phi(\mathbf{m})+\mathbf{E}^{(N)}(1))-G(\phi(\mathbf{m}))\big)\cdot \mathbb{1}_{\{\|\mathbf{E}^{(N)}(1)\|\geq\delta\}}\middle|\mathbf{M}^{(N)}(0)=\mathbf{m}\right]\right|\right|\cdot \mathbb{1}_{\{\mathbf{m}\in\mathcal{N}\}}\\
&\leq\ 2\overline{G}\cdot\mathbb{P}\big(\|\mathbf{E}^{(N)}(1)\|\geq\delta\big)\\
&\leq\ 2\overline{G}\cdot e^{-2N\delta^2}\quad\big(\text{By Lemma C.3}\big).
\end{aligned}$$

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.5.5. *Conclusion of the proof.* To summarize, we have obtained by (31):

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

and

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

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 C.1.

□

APPENDIX D: PROOF OF THEOREM 6.2

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 bandits 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 ith 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 a bandit jumps from state i to state j. For $i<s(\mathbf{m})$, this occurs at rate $Nm_iQ_{ij}^1$ as all of these bandits are activated. For $i>s(\mathbf{m})$, this occurs at rate $Nm_iQ_{ij}^0$ as these bandits 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 [18]. It is shown in [18] 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 ϕ is defined for the synchronous version of our asynchronous 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 [11, 33], we denote by $L^{(N)}$ the generator of the N bandits system and by Λ the generator of the differential equation. They associate to each almost-everywhere differentiable function h two functions $L^{(N)}h$ and Λh 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 Λh 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 C.1(i) in the asynchronous case is

THEOREM D.1. Under the same assumptions as in Theorem 6.2, 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 D.1 implies Theorem 6.2.

PROOF. Define the asynchronous version of function G as

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

As for the synchronous case, our assumptions imply that the unique fixed point is an exponentially stable attractor and a result similar to Lemma C.5 can be obtained for the asynchronous 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 $\phi(\mathbf{m}) = (\mathbf{m} - \mathbf{m}^*)\mathbf{K} + \mathbf{m}^*$ with the linear matrix K as in (8), 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 synchronous 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 [8], we have

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

$$= \mathbb{E}\Big[\Big((\Lambda - L^{(N)})G\big(\mathbf{M}^{(N)}(0)\big)\Big) \cdot \mathbb{1}_{\{\mathbf{M}^{(N)}(0)\in\mathcal{N}\}} \tag{32}$$

$$+ \Big((\Lambda - L^{(N)})G\big(\mathbf{M}^{(N)}(0)\big)\Big) \cdot \mathbb{1}_{\{\mathbf{M}^{(N)}(0)\notin\mathcal{N}\}}\Big], \tag{33}$$

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:

$$\big(\Lambda G\big)(\mathbf{m}) = \big(L^{(N)}G\big)(\mathbf{m}) = f(\mathbf{m}) \cdot \mathbf{B}. \tag{34}$$

This shows that the first term of (32) is equal to zero.

For the second term, note that both G and ΛG are continuous functions defined on the compact region Δ^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 [30]. 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 [5], which in our situation can be stated as

LEMMA D.2. 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

$$1 \geq \epsilon\lambda \cdot \exp\big(\frac{\epsilon^2 e^{-KT}}{3T}\big), \tag{35}$$

then we have

$$\mathbb{P}\Big[\sup_{t\leq T} \|\mathbf{M}^{(N)}(t) - \Phi_t\mathbf{m}\| > \epsilon \,\Big|\, \mathbf{M}^{(N)}(0) = \mathbf{m}\Big] \leq 2d \cdot e^{-c_1 N\epsilon^3}. \tag{36}$$

The above lemma plays the role of Lemma C.4 in synchronous case. Note that its original form stated as Theorem 4.2 in [5] 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 (36) has an additional term $\mathbb{P}(\Omega_0^c \cup \Omega_1^c \cup \Omega_2^c)$, with Ω_i^c being the complementary of Ω_i. In our case, $\Omega_0 = \Omega_1 = \Omega$ trivially holds; while the analysis of Ω_2 is more involved. However, as remarked before the statement of Theorem 4.2 in [5], 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 (35), then $\Omega_2 = \Omega$. Note that the constraint (35) is satisfied as long as ϵ 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 ϵ be as in Lemma D.2 that verify additionally (35). This is possible as the right hand side of (35) converges to 0 when ϵ is small and T is large.

We then have:

$$\begin{aligned}\mathbb{P}\big[\mathbf{M}^{(N)}(0) \notin \mathcal{N}\big] &= \mathbb{P}\big[\mathbf{M}^{(N)}(T) \notin \mathcal{N}\big] \qquad \text{(By stationarity)} \\ &\leq \mathbb{P}\big[\|\mathbf{M}^{(N)}(T) - \mathbf{m}^*\| \leq 2\epsilon\big]\end{aligned}$$

$$\leq \mathbb{P}\big[\|\mathbf{M}^{(N)}(T) - \Phi_T(\mathbf{M}^{(N)}(0))\| > \epsilon\big] \; + \; \mathbb{P}\big[\|\Phi_T(\mathbf{M}^{(N)}(0)) - \mathbf{m}^*\| > \epsilon\big]$$
$$= \mathbb{P}\big[\|\mathbf{M}^{(N)}(T) - \Phi_T(\mathbf{M}^{(N)}(0))\| > \epsilon\big] \qquad \big(\text{By our choice of } T\big)$$
$$\leq 2d \cdot e^{-c_1 N \epsilon^3} \qquad \big(\text{We apply (36) of Lemma D.2}\big).$$

So in summary, (32)-(33) gives

$$\|\mathbb{E}\big[\mathbf{M}^{(N)}(0)\big] - \mathbf{m}^*\| \leq (u + vN) \cdot 2d \cdot e^{-c_1 N \epsilon^3}. \tag{37}$$

Moreover, for any $c' > 0$ and $0 < c < c'$, $N \cdot e^{-c'N} = \mathcal{O}(e^{-cN})$, so the right hand side of (37) can be bounded by a term of the form $b \cdot e^{-cN}$. This concludes the proof of Theorem D.1. □

APPENDIX E: PARAMETERS USED IN THE NUMERICAL EXPERIMENTS.

For completeness, in this appendix we provide the parameters used in our numerical experiments in Section 4. The numbers are recorded in 8 digits of precision.

E.1. Parameters for the example of Section 4.2. The parameters for the example with $d = 3$ in Section 4.2 are

$$\mathbf{P}^0 = \begin{pmatrix} 0.30368587 & 0.25184515 & 0.44446898 \\ 0.40839084 & 0.41334941 & 0.17825975 \\ 0.66146205 & 0.1840829 & 0.15445505 \end{pmatrix}, \; \mathbf{P}^1 = \begin{pmatrix} 0.23763148 & 0.42381178 & 0.33855674 \\ 0.54401527 & 0.27028947 & 0.18569526 \\ 0.06938943 & 0.38776507 & 0.54284550 \end{pmatrix},$$

$\mathbf{R}^1 = (0.99663977, 0.22770951, 0.17300611)$ and $\mathbf{R}^0 = \mathbf{0}$. This example has also been used in our discussion at the end of Section 3.1 as well as in Section 4.4.

E.2. Parameters for the examples of Section 4.3. The parameters for the three period-2 cyclic examples in Section 4.3 are:

Example 1

$$\mathbf{P}^0 = \begin{pmatrix} 0.5214073 & 0.40392496 & 0.07466774 \\ 0.0158415 & 0.21455666 & 0.76960184 \\ 0.53722329 & 0.37651148 & 0.08626522 \end{pmatrix}, \; \mathbf{P}^1 = \begin{pmatrix} 0.24639364 & 0.23402385 & 0.51958251 \\ 0.49681581 & 0.49509821 & 0.00808597 \\ 0.37826553 & 0.15469252 & 0.46704195 \end{pmatrix},$$

$\mathbf{R}^1 = (0.72232506, 0.18844869, 0.25752477)$ and $\mathbf{R}^0 = \mathbf{0}$.

Example 2

$$\mathbf{P}^0 = \begin{pmatrix} 0.02232142 & 0.10229283 & 0.87538575 \\ 0.03426605 & 0.17175704 & 0.79397691 \\ 0.52324756 & 0.45523298 & 0.02151947 \end{pmatrix}, \; \mathbf{P}^1 = \begin{pmatrix} 0.14874601 & 0.30435809 & 0.54689589 \\ 0.56845754 & 0.41117331 & 0.02036915 \\ 0.25265570 & 0.27310439 & 0.47423991 \end{pmatrix},$$

$\mathbf{R}^1 = (0.37401552, 0.11740814, 0.07866135)$ and $\mathbf{R}^0 = \mathbf{0}$.

Example 3

$$\mathbf{P}^0 = \begin{pmatrix} 0.47819592 & 0.02090623 & 0.50089785 \\ 0.08063373 & 0.15456935 & 0.76479692 \\ 0.66552514 & 0.08481946 & 0.24965540 \end{pmatrix}, \; \mathbf{P}^1 = \begin{pmatrix} 0.00279465 & 0.37327924 & 0.62392611 \\ 0.51582335 & 0.46333908 & 0.02083756 \\ 0.41875202 & 0.17776712 & 0.40348086 \end{pmatrix},$$

$\mathbf{R}^1 = (0.97658608, 0.53014109, 0.40394919)$ and $\mathbf{R}^0 = \mathbf{0}$.

APPENDIX: ACKNOWLEDGEMENTS

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