(Close to) Optimal Policies for Finite Horizon Restless Bandits

Nicolas Gast, Bruno Gaujal, Chen Yan

(Close to) Optimal Policies for Finite Horizon Restless Bandits

Nicolas Gast, Bruno Gaujal, Chen Yan
Univ. Grenoble ALpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
{nicolas.gast,bruno.gaujal,chen.yan}@inria.fr

Abstract

Most restless Markovian bandits problems in infinite horizon can be solved quasi-optimally: the famous Whittle index policy is known to become asymptotically optimal exponentially fast as the number of arms grows, at least under certain conditions (including having a so-called indexable problem). For restless Markovian bandits problems in finite horizons no such optimal policy is known. In this paper, we define a new policy, based on a linear program relaxation of the finite horizon problem (called the LP-filling policy), that is asymptotically optimal under no condition. Furthermore we show that for *regular* problems (defined in the paper) the LP-filling policy becomes an index policy (called the LP-regular policy) and becomes optimal *exponentially* fast in the number of arms. We also introduce the LP-update policy that significantly improves the performance compared to the LP-filling policy for large time horizons. We provide numerical studies that show the prevalence of our LP-policies over previous solutions.

1 Introduction

In a *restless Markovian bandit*, a decision maker faces N arms and chooses which αN arms of those N to activate at each decision epoch. Each arm possesses an internal state whose evolution is Markovian and depends on whether this arm is activated or not. This forms a Markov decision process that possesses a special structure. Restless Markovian bandits problems have been shown to be PSPACE-hard in [15]. The classical *Whittle index policy* has been introduced for infinite horizon bandits problems in [18]. It scales well with the number of arms. It is in general suboptimal but is proven in [17] to be asymptotically optimal as N goes to infinity, under several technical assumptions.

Whittle index policy has been used in numerous domains where restless Markovian bandits are natural modeling tools, and it performs well in all of them, even for a moderate number of arms. Among these applications in the existing literature we may cite: wireless fading channels [14]; charging vehicles [19]; queue and birth-and-death processes [1], [11]; medical treatments [2] or age of information [9]. The work in [6] has provided a theoretical grounding for the excellent performance of Whittle index policy, by proving that the aforementioned asymptotic optimality claimed in [17] occurs *exponentially* fast in N when the model is *non-singular*.

In this paper we consider the same classical model as in [17], but under a discrete time *finite horizon* T. Finite horizon multi-armed bandits have drawn extensive research attention for a long time, because of their practical and theoretical interest, see [3]. In these cases, Whittle indexes cannot be defined properly and no asymptotically optimal policy is known. The difficulty to compute a policy for restless bandits is that the decision maker has to activate exactly αN arms at each decision epoch. This creates dependencies among arms. A key idea that motivated the original definition of Whittle index in [18] is to relax this constraint by assuming that the time average activation of arms is αN. Our starting point is to use a tighter relaxation that is adapted to the finite horizon case, namely, we consider a problem in which the expected number of activated arm should be αN at

Preprint. Under review.

each decision epoch. This relaxation transforms the original finite horizon optimization problem into a linear program. By using the structure of the solutions of this linear program, we define several control policies. The goal of this paper is to provide a thorough study of these policies. Our main contributions are:

1. Similarly to the exponential convergence rate theorem claimed in [6] for the infinite horizon problem under the additional assumption of non-singularity, we encounter a similar regularity issue here for the finite horizon problem (see Definition 1), but with a completely different meaning for being regular. We show in Theorem 1 that under the additional assumption of being *regular* (in the sense we define in this paper), the LP-regular policy becomes optimal *exponentially* fast while classical convergence rates for such approximations, based on central limit arguments, is usually in the square root.
2. We show that the more general LP-filling policy is asymptotically optimal in Theorem 2. One advantage of this asymptotic optimality result compared to its analogue in the infinite horizon scenario is that we do not need *any* technical assumptions on the model, whereas in [17] the asymptotic optimality is true only when the model is *indexable* (so that Whittle index is well defined) and the deterministic dynamics has a *global attractor*. Both assumptions (indexability and global attraction) are extremely hard to verify.
3. Finally, we propose the LP-update policy for the finite horizon problem, which solves new linear programs based on the stochastic trajectory at each point in time. We show in Theorem 3 that the LP-update policy is also asymptotically optimal. We then provide numerical examples showing that the LP-update policy can significantly improve the performance compared to the LP-filling policy, notably when the time horizon T is large and when the deterministic dynamics has stability issue. This shows the benefit of using feedback updates, under situations where the performance is critical in spite of the increase in computation time.

1.1 Related work

A priority policy based on the solution of a linear program has already been proposed in [16], in which the author studies the *continuous time* infinite horizon restless bandits using time average reward criterion. Two key differences are: (a) we consider the finite horizon problem; (b) we are using dual variables of the linear program solution to define a *non-priority* policy which is asymptotically optimal in all generality, whereas [16] defines a set of *priority* policies and shows that under some stability assumptions they are all asymptotically optimal.

Another paper with similar ideas as ours is [4], in which the authors consider the discrete time infinite horizon discounted restless bandits, and define a primal-dual heuristic index policy based on the linear program solution. However, they do not prove any asymptotic optimality result.

Gittins index policy ([7]) was originally designed for infinite horizon discounted *rested* bandits, where passive arms remain frozen. This model can be regarded as a special case of the restless bandit models we considered in this paper. In [13] the author proposes an algorithm computing a finite horizon version of Gittins index that performs well in practice. We compare the performance of the finite horizon Gittins index policy with our LP-filling policy, and show that, while both of them have similar computational complexities, the LP policy outperforms Gittins even for moderate numbers of arms. In fact, here Gittins is not asymptotically optimal contrary to the LP policy.

2 Model description

A discrete time finite horizon restless Markovian bandit model with parameter set $\left\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha, N, T, \mathbf{m}\right\}$ is a Markov decision process (MDP) defined as follows. The model is composed of N statistically identical arms[1]. Each arm evolves in a finite state space $\{1 \dots d\}$ and the state of the nth arm at time t is denoted by $S_n(t) \in \{1 \dots d\}$. The state space of all the arms at time t is denoted by $\mathbf{S}(t) = \big(S_1(t), \dots, S_N(t)\big)$. Decisions are taken at times $t \in \{0, \dots, T-1\}$. At each decision epoch, a decision maker observes $\mathbf{S}(t)$ and chooses a fraction αN of the N arms to be activated, where we assume that α and N are such that αN is an integer.

[1]The case with several arm types can be handled by aggregating all types into one and using a single block diagonal transition matrix with one block per type. See also the example in Section 6.

We write $A_n(t) = 1$ if arm n is activated at time t and $A_n(t) = 0$ otherwise. The action vector at time t is $\mathbf{A}(t) = \big(A_1(t), \dots, A_N(t)\big)$. It must satisfy $\sum_{n=1}^N A_n(t) = \alpha N$. For each arm that is in state s and whose action is a, the decision maker earns an immediate reward R_s^a. We assume that $|R_s^a| \leq 1$. Given $S_n(t) = s$ and $A_n(t) = a$, the arm n makes a Markovian transition to a state s' with probability $P_{s,s'}^a$. Those transitions are independent among all arms: for given states $\mathbf{s}, \mathbf{s}'$ and activation vector $\mathbf{a}$, one has:

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

For a given initial condition $\mathbf{s}(0)$, the *finite horizon restless bandit (FHRB)* problem can be written as

$$V_{\text{opt}}^{(N)}(\mathbf{s}(0), T) = \max_{\Pi} \quad \mathbb{E}\Big[\frac{1}{N} \sum_{t=0}^{T-1} \sum_{n=1}^N R_{S_n(t)}^{A_n(t)}\Big] \tag{2a}$$

$$\text{s.t.} \quad \sum_{n=1}^N A_n(t) = \alpha N, \text{ for all } t \in \{0, \dots, T-1\}, \tag{2b}$$

$$\text{Arms follow the Markovian evolution (1),} \tag{2c}$$

$$S_n(0) = s_n(0) \text{ for all } n \in \{1 \dots N\}. \tag{2d}$$

Here $\Pi = \{(\pi_0, \dots, \pi_{T-1})\}$ is the set of eligible decision policies, with $\pi_t : \mathbf{S}(t) \to \mathbf{A}(t)$ the decision rule that the decision maker uses at time t, which chooses the action vector $\mathbf{A}(t)$ when the state is $\mathbf{S}(t)$.

The key difficulty in the above optimization problem is the constraint Equation (2b) that couples the evolution of all arms. In the following, we construct a decision rule based on a relaxation of this constraint. The goal of the paper is to demonstrate that this construction is very efficient.

3 LP-based Indexes

By construction, the n arms are exchangeable, which means that the arms state vector $\mathbf{S}(t)$ at time t can be replaced by its empirical measure, $\mathbf{M}(t) \in \Delta^d$ where $M_s(t)$ is the fraction of arms in state s at time t and Δ^d is the d-dimensional simplex. In particular, for the initial state, now denoted $\mathbf{m}$, we set $m_s := \frac{1}{N} \sum_{n=1}^N \mathbf{1}_{\{s_n(0)=s\}}$ for all $s \in \{1 \dots d\}$. As for the decision at time t, the decision maker observes the vector $\mathbf{M}(t)$ and chooses $\mathbf{Y}(t) := \big(Y_{s,0}(t), Y_{s,1}(t)\big)_{s \in \{1 \dots d\}}$ where $Y_{s,a}(t)$ is the fraction of arms that are in state s at time t for which decision $a \in \{0, 1\}$ is taken (note that by construction $\sum_a Y_{s,a}(t) = M_s(t)$ and $\sum_s Y_{s,1}(t) = \alpha$).

To construct our heuristics, we consider the optimization problem (2) where we replace (2b) by the relaxed constraint $\sum_{n=1}^N \mathbb{E}\big[A_n(t)\big] = \alpha N$. Since the cost and the constraints only depend on the average number of bandits in each state, the states and the activations can be replaced by expectations. Let $m_s(t) := \mathbb{E}\left[M_s(t)\right]$ and $y_{s,a}(t) := \mathbb{E}\left[Y_{s,a}(t)\right]$ for all states s, actions a and time-steps t. We denote by $V_{\text{rel}}(\mathbf{m}, T)$ the value of this relaxed optimization problem:

$$V_{\text{rel}}(\mathbf{m}, T) = \max_{\mathbf{y} \geq \mathbf{0}} \quad \sum_{t=0}^{T-1} \sum_{a,s} R_s^a y_{s,a}(t) \tag{3a}$$

$$\text{s.t.} \quad \sum_s y_{s,1}(t) = \alpha \qquad \forall t, \tag{3b}$$

$$y_{s,0}(t+1) + y_{s,1}(t+1) = \sum_{s',a} y_{s',a}(t) P_{s's}^a \qquad \forall s, t, \tag{3c}$$

$$y_{s,0}(0) + y_{s,1}(0) = m_s \qquad \forall s. \tag{3d}$$

In the above optimization problem, the constraints (3b) are the relaxation of the constraints (2b). They impose that the expected fraction of activated arms is α at all time. The constraints (3c) are the analog of (2c) written as linear constraints. They correspond to the Markovian evolution of the system. Similarly, (3d) correspond to the initial condition (2d).

Note that the optimization problem (3) does not depend on N anymore. Moreover, as it is a relaxation of (2), it should be clear that $V_{\text{opt}}^{(N)}(\mathbf{m}, T) \leq V_{\text{rel}}(\mathbf{m}, T)$. Yet, an optimal policy for (3) is usually not a valid control policy for (2) because the constraint (2b) imposes that the number of activated arms is *exactly* αN at time t (and not just in expectation). The key contribution of the paper is to use the relaxation (3) to construct an efficient policy for (2). For that, we will use the Lagrangian multipliers that correspond to the critical constraints (3b).

By construction, the LP problem (3) has at least one solution that we denote by $\mathbf{y}^*$. We also denote by $\mathbf{m}^*$ the corresponding state vector, where $m_s^*(t) = y_{s,0}^* + y_{s,1}^*$. Hence, by strong duality, there exists Lagrange multipliers $\gamma_0, \ldots, \gamma_{T-1}$ corresponding to the constraints (3b), such that $\mathbf{y}^*$ is also a solution to the problem

$$V_{\text{rel}}(\mathbf{m}, T) = \max_{\mathbf{y} \geq \mathbf{0}} \quad \sum_{t=0}^{T-1} \sum_{a,s} (R_s^a + a\gamma_t) y_{s,a}(t) \tag{4a}$$

$$\text{s.t.} \quad y_{s,0}(t+1) + y_{s,1}(t+1) = \sum_{s',a} y_{s',a}(t) P_{s's}^a \qquad \forall s, t, \tag{4b}$$

$$y_{s,0}(0) + y_{s,1}(0) = m_s \qquad \forall s. \tag{4c}$$

One can see the optimization problem 4 as a linear formulation of the following Markov decision problem: Let X be a Markov reward process with state space $\{1 \ldots d\}$ and action space $\{0, 1\}$. The reward in state $s \in \{1 \ldots d\}$ under action $a \in \{0, 1\}$ is $R_s^a + a\gamma_t$. The transition probabilities are $\mathbb{P}\big(X(t+1) = y \;\big|\; X(t) = x, \text{action} = a\big) = P_{xy}^a$. The initial condition is $X(0) \sim \mathbf{m}$, by interpreting $\mathbf{m}$ as a probability vector, and the time horizon is T.

Let us denote by $Q_{s,a}(t)$ the Q-value of the state-action pair (s, a) at time t for this MDP. We define the LP-index of the state s at time t as

$$I_s(t) := Q_{s,1}(t) - Q_{s,0}(t). \tag{5}$$

A positive $I_s(t)$ means that an arm in state s should be activated, while a negative $I_s(t)$ means that an arm in state s should not be activated. For arms in states s such that $I_s(t) = 0$, both actions have the same merit and they can be activated or not for problem (4). This suggests that a policy that activates *(i)* all the arms with positive index, *(ii)* no arms with negative index and *(iii)* some arms with a null index to reach the right number of activated arms, should perform well for the original problem (2). This also suggests that states with a null index should play a critical role. This is corroborated in the next section by introducing the notion of regularity that concerns precisely states with a null index.

4 Exponential convergence rate of the LP-regular policy

In this section, we focus on finite horizon restless bandits (FHRB) that are *regular*: Any FHRB $\big\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha, T, \mathbf{m}\big\}$ is regular if, for all time t, the number of states such that $I_s(t) = 0$ is exactly one and that this state satisfies $y_{s,0}^*(t) > 0$ and $y_{s,1}^*(t) > 0$. We define a decision rule for regular FHRBs that we call the LP-regular policy as follows:

Definition 1 (The LP-regular policy)**.** *At each decision epoch $0 \leq t \leq T-1$, the LP-regular policy enumerates the arms by decreasing order of index values $I_s(t)$ and activates the αN arms having the largest indices. Ties are resolved by using a fixed and predetermined priority among states.*

We denote the value of this policy by $V_{\text{LP}-\text{reg}}^{(N)}(\mathbf{m}, T)$. As the LP-regular policy is a valid policy for our original problem, it should be clear that

$$V_{\text{LP}-\text{reg}}^{(N)}(\mathbf{m}, T) \leq V_{\text{opt}}^{(N)}(\mathbf{m}, T) \leq V_{\text{rel}}(\mathbf{m}, T).$$

What we show below is that, if a FHRB is regular, then $V_{\text{LP}-\text{reg}}^{(N)}(\mathbf{m}, T)$ converges to $V_{\text{rel}}(\mathbf{m}, T)$ at exponential speed, as the number of arms N goes to infinity. This implies that the LP-regular policy becomes optimal exponentially fast for regular problems. This is quite different than convergence rate that one usually gets in this kind of approximations using the mean behavior. In general the convergence rate is in $\mathcal{O}(1/\sqrt{N})$ (as for our next Theorems 2 and 3) and is obtained using central limit theorem approaches.

Theorem 1. *Consider a* regular *FHRB* $\left\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha, N, T, \mathbf{m}\right\}$. *Then there exist* $C_1, C_2 > 0$ *that do not depend on* N *such that for all* N *with* αN *being an integer we have:*

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

Sketch of proof. A full proof of the theorem is given in Appendix A.3. We describe below the main ingredients of this proof. To emphasize the dependence on N, we denote by $M_s^{(N)}(t)$ the fraction of arms that are in state s at time t when following the LP-regular policy defined above, and by $Y_{s,a}^{(N)}(t)$ the fraction of arms that are in state s at time t and for which decision a is taken.

The first step of the proof is to decompose the evolution of the process $(\mathbf{M}^{(N)}(t))_{t\geq 0}$ in two stages. The first stage is the decision stage. We show that there exists a continuous piecewise affine deterministic map $\pi^t_{LP-reg} : \Delta^d \to \Delta^{2d}$ such that $\pi^t_{LP-reg}(\mathbf{M}^{(N)}(t)) := \mathbf{Y}^{(N)}(t)$. This stage is then followed by a Markovian transition stage where each arm makes an independent transition. Let $\phi : \Delta^{2d} \to \Delta^d$ be the map that is defined as $(\phi(\mathbf{y}))_s = \sum_{a=0}^{1} \sum_{s'=1}^{d} y_{s,a} P^a_{s,s'}$. By using the Markov property and Hoeffding's inequality, we show that $\mathbb{E}\left[\mathbf{M}^{(N)}(t+1)) \mid \mathbf{Y}^{(N)}(t)\right] = \phi(\mathbf{Y}^{(N)}(t))$ and that $\mathbb{P}\left(\left\|\mathbf{M}^{(N)}(t+1) - \phi\left(\mathbf{Y}^{(N)}(t)\right)\right\|_1 \geq \varepsilon \mid \mathbf{Y}^{(N)}(t)\right) \leq e^{-2N\varepsilon^2}$. This shows that the evolution of $\mathbf{M}^{(N)}(t)$ and $\mathbf{Y}^{(N)}(t)$ are essentially deterministic as N goes to infinity and converge in probability to some deterministic values $\mathbf{m}^\infty(t)$ and $\mathbf{y}^\infty(t)$ that satisfy $\mathbf{m}^\infty(t+1) = \phi(\mathbf{m}^\infty(t)))$ and $\mathbf{m}^\infty(t+1) = \pi^t_{LP-reg}(\mathbf{y}^\infty(t))$.

The second step of the proof is to show that the solution of the LP problem (3), $\mathbf{y}^*$ and the corresponding $\mathbf{m}^*$ satisfies $\mathbf{m}^*(t+1) = \phi(\mathbf{y}^*(t))$ and $\mathbf{y}^*(t+1) = \pi^t_{LP-reg}(\mathbf{m}^*(t))$. The former corresponds to constraint (3c). For a regular problem, the latter is guaranteed by the definition of the LP-regular policy. This implies that $\mathbf{m}^\infty = \mathbf{m}^*$ and $\mathbf{y}^\infty = \mathbf{y}^*$. Lastly, when the model is regular, $\mathbf{m}^*(t)$ lies strictly inside a linear region of $\pi^t_{LP-reg}(\cdot)$. We prove that this implies $\left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t) \mid \mathbf{M}^{(N)}(0)\right] - \mathbf{y}^*(t)\right\|_1 \leq e^{-CN}$ for some constant C independent of N. □

The above result guarantees that for regular models, the LP-regular policy becomes optimal exponentially fast. We should emphasize that in the above proof, we use the regular condition twice. The first time is when we show that the limit $\mathbf{y}^\infty$ is a solution of the LP problem (3). In fact, when there are two states $s \neq s'$ such that $I_s(t) = I_{s'}(t) = 0$, there exist many ways to deal with ties and most of them will not provide an asymptotically optimal policy. We provide a 2 states FHRB example in Appendix B for which the optimal LP solution activates a bit of both state 1 and state 2 at time 1:

$$y^*_{1,0}(1) \approx 0.12, \quad y^*_{1,1}(1) \approx 0.04, \quad y^*_{2,0}(1) \approx 0.63, \quad y^*_{2,1}(1) \approx 0.21. \tag{6}$$

In fact, we show numerically that for this example, giving a higher priority to state 1 or to state 2 are asymptotically suboptimal. To be asymptotically optimal, a decision rule should activate a proportion of arms close to the ones given in (6). The next Section 5 provides a generic way to deal with this situation. The second time when we use regularity is to show that $\mathbf{m}^*(t)$ lies strictly inside a linear region of $\pi^t_{LP-reg}(\cdot)$. This is essential to obtain the exponential convergence rate. When this does not hold, one cannot really hope better than an $\mathcal{O}(\frac{1}{\sqrt{N}})$ convergence rate.

5 Asymptotic optimality of the LP policies in the general case

Unfortunately, the LP-regular policy that solves ties by a fixed priority order is not always optimal when the model is not regular. In fact, the example that we provide in Appendix B is such that there exists no fixed priority order that is asymptotically optimal. In this section we propose two solutions to construct an LP-policy that is asymptotically optimal as the number of arms grows. The first one is actually a generalization of the LP-regular policy to deal with ties when there are at least two states $s \neq s'$ such that $I_s(t) = I_{s'}(t) = 0$. It uses the idea of "water filling". The second solution applies updates based on the stochastic trajectory. It is computationally more expensive but has a better performance in practice.

5.1 The LP-filling policy using water filling

For regular cases, the LP-regular policy is asymptotically optimal essentially because the map $\pi^t_{LP-reg} : \mathbf{M}^{(N)}(t) \mapsto \mathbf{Y}^{(N)}(t)$ is continuous and satisfies $\mathbf{y}^*(t) = \pi^t_{LP-reg}(\mathbf{m}^*(t))$. When the

problem is not regular, defining a continuous map $\pi^t(\cdot)$ such that $\mathbf{y}^*(t) = \pi^t(\mathbf{m}^*(t))$ cannot be done by a policy as simple as a fixed priority order. A first idea to circumvent this difficulty would be to activate exactly $Ny^*_{s,1}(t)$ arms for each state s but this would probably be impossible since $Ny^*_{s,1}(t)$ might be larger than the stochastic quantity $M_s^{(N)}(t)$ and might not be an integer. Hence, we introduce a new policy called the LP-filling policy, whose value is denoted by $V^{(N)}_{\mathrm{LP-fill}}(\mathbf{m}, T)$.

Definition 2 (The LP-filling policy)**.** *At each decision epoch $0 \leq t \leq T-1$, we activate αN arms according to the following water filling rules (for each rule, we choose arms in decreasing order of index $I_s(t)$, ties being resolved by a fixed priority order):*

- *Activate the arms with a positive index, up to a total number of αN if there are enough such arms.*
- *If the number of active arms has not reached αN, activate arms such that $I_s(t) = 0$ and activate at most $\lfloor Ny^*_{s,1}(t) \rfloor$ of them.*
- *Complete the set of αN active arms if necessary with remaining arms (with null or negative indexes $I_s(t)$).*

For all time steps t such that at most one state s is such that $I_s(t) = 0$, this policy is simply a priority policy and coincides with the LP-regular policy defined in Section 4 for regular models.

To illustrate how the LP-filling policy works, consider a FHRB example with $N = 10$ arms, $\alpha = 0.3$, and $d = 4$ states for which for a given t, $I_1(t) > I_2(t) = I_3(t) = 0 > I_4(t)$ and $y^*_{\cdot,1}(t) = (0.1, 0.1, 0.1, 0)$. In such a case, arms in state 1 are activated in priority because $I_1(t) > 0$. Then, we activate some arms in state 2 or 3 while respecting the constraint that we should not activate more than one arm in state 2 or 3 (here we assume that ties are resolved by giving higher priority to state 2 than state 3). If this is not enough, we complete with the remaining arms in states 2, 3 and 4 (in this order). This would give the following activations:

State $\mathbf{M}^{(N)}(t)$	Decisions $\mathbf{Y}^{(N)}(t)$	Reason
$(0.4, 0.2, 0.2, 0.2)$	$(0.3, 0, 0, 0)$	Priority to state 1.
$(0.2, 0.3, 0.2, 0.3)$	$(0.2, 0.1, 0, 0)$	Priority to state 1 then at most one arm in state 2.
$(0.1, 0.3, 0.3, 0.3)$	$(0.1, 0.1, 0.1, 0)$	Priority to state 1 then at most one arm in state 2 or 3.
$(0, 0.2, 0.2, 0.6)$	$(0, 0.2, 0.1, 0)$	Activate at most one arm in state 2 or 3, then complete.

As stated in the next theorem, the LP-filling policy is asymptotically optimal for any FHRB. Recall that, when the problem is regular, the LP-filling policy coincides with LP-regular which becomes optimal at an exponential rate. In non-regular cases the proved convergence rate is much slower as it is only $\mathcal{O}(\frac{1}{\sqrt{N}})$. This is due to the fact that for non-regular problems $\mathbf{m}^*(t)$ lies on the boundary between two linear pieces of $\phi_t(\cdot)$ for some t. In any neighborhood of $\mathbf{m}^*(t)$, the map $\phi_t(\cdot)$ is not linear.

Theorem 2. *For any FHRB $\left\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha, N, T, \mathbf{m}\right\}$, there exists $C_3 > 0$ that does not depend on N such that for all N with αN being an integer we have:*

$$0 \leq V_{\mathrm{rel}}(\mathbf{m}, T) - V^{(N)}_{\mathrm{LP-fill}}(\mathbf{m}, T) \leq \frac{C_3}{\sqrt{N}}.$$

Consequently $\lim_{N\to\infty} V^{(N)}_{\mathrm{LP-fill}}(\mathbf{m}, T) = V^{(N)}_{\mathrm{opt}}(\mathbf{m}, T)$.

Sketch of proof. A full proof of the theorem is given in Appendix A.4. We first show that for all $0 \leq t \leq T-1$, there exists a decision rule $\pi^{t,N}_{LP-fill}(\cdot) : \Delta^d \to \Delta^{2d}$ that is continuous and piecewise affine with finitely many affine pieces, and such that the decision rule induced by the water filling is at distance $2d/N$ of $\pi^{t,N}_{LP-fill}(\cdot)$. We then combine those facts together with $\|\mathbf{M}(t+1) - \phi(\mathbf{Y}(t))\|_1 = \mathcal{O}(\frac{1}{\sqrt{N}})$ to obtain the result. □

5.2 The LP-update policy

As mentioned earlier, a first difficulty when trying to apply the control $\mathbf{y}^*$ given by the LP problem (3) to the original problem (2) is that in general $y^*_{s,0}(t) + y^*_{s,1}(t) \neq M_s^{(N)}(t)$. Hence, one cannot always activate exactly $Ny^*_{s,1}(t)$ arms in state s at time t. This problem disappears at time 0 because

$\mathbf{m}^*(0) = \mathbf{m} = \mathbf{M}^{(N)}(0)$ (constraints (2d) and (3d)). This suggests to solve the LP problem at each stage starting from $\mathbf{M}^{(N)}(t)$ with horizon $T-t$, and to activate exactly $Ny^{t*}_{s,1}(0)$ arms that are in state s, where $\mathbf{y}^{t*}$ is the LP solution at this stage. Yet, a second difficulty is that $Ny^{t*}_{s,1}(0)$ might not be an integer. To solve this second problem, we propose to use a randomized rounding algorithm inspired from [10]. This leads to the definition of the LP-update policy, whose value is denoted by $V^{(N)}_{\mathrm{LP-up}}(\mathbf{m}, T)$:

Definition 3 (The LP-update policy)**.** *At a given decision epoch t, we solve the problem* (3) *with initial condition $\mathbf{M}^{(N)}(t)$ and horizon $T-t$. Denoting $\mathbf{y}^{t*}$ the optimal solution of this LP problem at stage t, we then use the* randomized procedure *detailed in Appendix A.5.1 to activate $Ny^{t*}_{s,1}(0)$ arms in state s in expectation.*

We claim that the LP-update policy is also asymptotically optimal:

Theorem 3. *For FHRB $\left\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha, N, T, \mathbf{m}\right\}$, there exists $C_4 > 0$ that does not depend on N such that for all N with αN being an integer we have:*

$$0 \leq V_{\mathrm{rel}}(\mathbf{m}, T) - V^{(N)}_{\mathrm{LP-up}}(\mathbf{m}, T) \leq \frac{C_4}{\sqrt{N}}.$$

Proof. A full proof of the theorem is given in Appendix A.5. We show that the function $V_{\mathrm{rel}}(\,\cdot\,, t) : \Delta^d \to \mathbb{R}$ is Lipschitz-continuous. We then define a sequence $\{x(t)\}_{0 \leq t \leq T}$ for $0 \leq t \leq T$ as

$$Z(t) := V_{\mathrm{rel}}(\mathbf{M}^{(N)}(t), T-t) - V^{(N)}_{\mathrm{LP-up}}(\mathbf{M}^{(N)}(t), T-t).$$

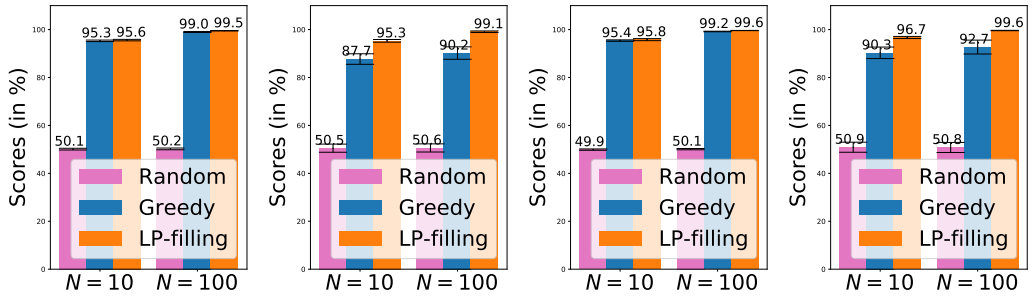
Note that $Z(0) = V_{\mathrm{rel}}(\mathbf{m}, T) - V^{(N)}_{\mathrm{LP-up}}(\mathbf{m}, T)$ and $Z(T) = 0$. We then apply the Bellman's principle of optimality, as well as a triangle inequality, to show that the difference between $Z(t)$ and $Z(t+1)$ can be upper bounded by $\mathcal{O}(\frac{1}{\sqrt{N}})$. □

Previous policies, LP-regular and LP-filling take a decision at time t by assuming that $\mathbf{M}(t) \approx \mathbf{m}^*$, which may become false exponentially fast with t. Hence it may be possible that for a large T, an extremely large N is required for the convergence in Theorems 1 or 2 to be apparent. Meanwhile, LP-update "corrects" its decisions at each time t according to the actual output from $t-1$. This can significantly improve the convergence. Of course, applying the LP-update policy requires to solve an LP problem at each time step, which is computationally expensive. In most cases, reported in Figure 1, the LP-filling policy has very good performance. However, when the dynamics of the system is complex or sensitive to perturbations, LP-update becomes much better than LP-filling and is worth the extra computations. Such examples are reported in Appendix C.3.

6 Numerical Experiments

In order to assess the performance of the LP-based policies, we randomly generate multi-armed bandits models of different types. The results are summarized in Figure 1, and they differ significantly depending on the examples. In all examples, the reward of passive arms is $\mathbf{R}^0 = \mathbf{0}$. We compare the performance of the two variants of the LP policy (LP-filling and LP-update) with two baseline policies: the random policy which at each time step chooses αN arms uniformly at random, and the greedy policy which activates the αN arms having the largest rewards $\mathbf{R}^1$. We also consider rested arms, for which the finite horizon Gittins index policy (called the Gittins policy for short) introduced in [13] is believed to be a very efficient heuristic and has a proven regret in a Bayesian setting [12].

For each considered model, we normalize the performance between 0 and 100, where 100 is the value of the LP relaxation (3) (which is an upper bound on the performance of any policy) and 0 is a lower bound that is obtained by using a minimizer instead of a maximizer in (3). In each case, we evaluate the performance of certain policy by taking average over 100 sets of parameters, and report the 95% confidence interval. We provide more details on our experimental setups in Appendix C.1. To solve problem (3) and compute the indices, we use the default LP solver from the `PuLP` package in Python. Its empirical time complexity turns out to be $\mathcal{O}(T^2 d^3)$. All our simulation tasks can be ran in parallel, and the most time demanding ones are the different arms cases in Figure 1, and evaluating the LP-update policy in Figure 2. Both of them take no more than half an hour to complete on an eight cores CPU. More details can be found in Appendix C.2.



(a) Dense identical arms (b) Sparse identical arms (c) Dense different arms (d) Sparse different arms

Figure 1: Performance of the different heuristics on randomly generated FHRB.

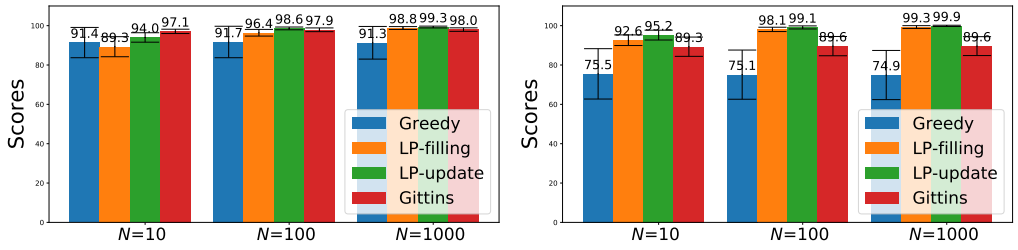
Scenario 1: Dense models. In Figure 1a and 1c, we consider *dense* models. We generate transition matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ by generating matrices of numbers between 0 and 1 and normalizing each line so that it sums to 1. We choose $d = 10$, $T = 50$, $\alpha = 0.5$ and $N = 10$ or $N = 100$. In this scenario, the greedy policy provides a performance that is close to optimal (above 95), which means that such a randomly generated example is easy to control. This leaves a little margin of improvement to the LP-filling policy, but it still performs slightly better, even with a moderate number of arms.

Scenario 2: Sparse models. In Figures 1b and 1d, we consider *sparse* models where each line of the transition matrices $\mathbf{P}^0$ and $\mathbf{P}^1$ only has *two* non-zero terms, and their positions are random and uniformly picked. In Figure 1b, we choose $d = 10$ and $N = 10$ or $N = 100$ identical arms each having the same sparse transition matrices and we choose $\alpha = 0.5$. Here, the LP-filling policy outperforms the greedy policy to a greater extent. We have also encountered sparse models for which the asymptotic optimality of the LP-filling policy occurs extremely slowly with respect to the arm population N. We discuss in Appendix C.3 how the LP-update policy can greatly improve the situation of those difficult cases.

Scenario 3: Different arms Our model assumes that all arms have the same transition matrices but does not assume that such matrices are irreducible. Hence, it is possible to consider non-identical arms by considering block-diagonal matrices and an initial condition that puts arms in different states. This is what we do to obtain Figure 1c and Figure 1d, where we consider 10 different arms with either one (for $N = 10$) or ten (for $N = 100$) copies of each arm. The performance in the case of different arms is qualitatively similar to the one of identical arms: the LP-filling policy performs very well, it improves over greedy only for sparse bandits.

Scenario 4: Rested bandits models Rested bandits can be viewed as a particular case of restless bandits for which $\mathbf{P}^0$ is the identity matrix. The Gittins index policy is a well-known policy that is optimal for *discounted infinite horizon* rested bandit problems with one arm activation (see Section 3.4 of [8]). A finite horizon Gittins index is introduced in [13], that can be computed in $\mathcal{O}(T^2d^3)$ (similarly to the empirical complexity of our LP-indices). In Figures 2a and 2b we consider sparse *rested* bandits and we compare our LP policies with Gittins (and with Greedy as a baseline). Figure 2a is with $\alpha = 0.1$, $d = 10$, $T = 50$ and $N = 10, 100, 1000$ respectively, and we choose $\alpha = 0.5$ in Figure 2b. For finite horizons, the Gittins policy is known to perform very well. It also has regret guarantees for Bayesian bandits [12]. Yet, it is in general not optimal. The results reported in 2a and 2b show indeed that Gittins performs well. Yet, for large values of N, Gittins is outperformed by the LP policies. In fact, these figures suggest that Gittins is not asymptotically optimal here because its performance does not seem to increase with N. This is in contrast with LP-filling and LP-update, which are asymptotically optimal and outperform the Gittins policy, as soon as the activation proportion α is not too small and the arm population N is of moderate size. Note that for $N = 10$ and $\alpha = 0.1$, Greedy performs very well. However, the confidence intervals are quite loose for Greedy, indicating that Greedy outperforms LP-filling for some models but performs poorly for others. Note that for all those models, LP-update performs better than LP-filling but only by a small margin. In fact, for most (rested and restless) models LP-filling and LP-update perform almost identically but there exist some models for which LP-update largely outperforms LP-filling. We study

this in more detail in Appendix C.3 where we consider sparse bandit models for which the LP-filling policy does not provide a good performance, even for $N = 1000$. In all these examples, LP-update improves dramatically the situation.



(a) $\alpha = 0.1$ (b) $\alpha = 0.5$

Figure 2: Performance of the different heuristics on randomly generated *rested* bandit models.

Regularity We also did statistical tests to see if models tend to be regular or not. We choose $\alpha = 0.5$, $T = 50$, identical arms, and report the proportions in Table 1. According to our experiments, a dense model tends to be regular (more than 95% of the time) and this figure does not seem vary much with the dimension. On the contrary, for sparse models, the proportion of regular models is much smaller and seems to decrease with the dimension. For rested models, the proportion is even smaller: among 1000 randomly generated examples, they are all irregular. Note that the regularity condition is only necessary to obtain the *exponential* convergence rate, and the LP-filling policy is *always* asymptotically optimal.

Scenario	Dense restless models	Sparse restless models	Rested models
$d = 10$	96.8%	24.8%	0%
$d = 15$	98.7%	11.2%	0%
$d = 20$	98.5%	6.0%	0%

Table 1: Percentage of regular models among 1000 randomly generated parameter sets.

7 Conclusion and future work

In this paper we introduce new policies to solve the finite horizon restless bandit problem. These policies are the first to be shown asymptotically optimal when the number of arms grows. The convergence is even exponentially fast when the bandit problem is regular. These theoretical properties are backed by numerical experiments that show the superiority of our LP policies over previously proposed heuristics in several scenarios.

Actually, LP-based policies can also be defined for the infinite horizon restless bandit problem. As for future work, we plan to develop such policies that are asymptotically optimal when the number of arms grows, and that can replace the Whittle index policy when the later cannot be properly defined, *i.e.* when the problem is not indexable, or when the problem is indexable but does not have the global attractor property.

References

[1] PS Ansell, Kevin D Glazebrook, José Nino-Mora, and M O'Keeffe. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research*, 57(1):21–39, 2003.

[2] Turgay Ayer, Can Zhang, Anthony Bonifonte, Anne C. Spaulding, and Jagpreet Chhatwal. Prioritizing Hepatitis C Treatment in U.S. Prisons. *Operations Research*, 67(3):853–873, May 2019.

[3] Donald A. Berry and Bert Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Springer, October 1985.

[4] Dimitris Bertsimas and José Niño-Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic, 1997.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, USA, 2004.

[6] Nicolas Gast, Bruno Gaujal, and Chen Yan. Exponential convergence rate for the asymptotic optimality of whittle index policy, 2020.

[7] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, pages 148–177, 1979.

[8] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[9] Yu-Pin Hsu. Age of information: Whittle index for scheduling stochastic arrivals, 2018.

[10] Stratis Ioannidis and Edmund Yeh. Adaptive caching networks with optimality guarantees. *CoRR*, abs/1604.03175, 2016.

[11] Maialen Larrnaaga, Urtzi Ayesta, and Ina Maria Verloop. Dynamic control of birth-and-death restless bandits: Application to resource-allocation problems. *IEEE/ACM Transactions on Networking*, 24(6):3812–3825, 2016.

[12] Tor Lattimore. Regret analysis of the finite-horizon gittins index strategy for multi-armed bandits. In *Conference on Learning Theory*, pages 1214–1245. PMLR, 2016.

[13] José Niño-Mora. Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing*, 23(2):254–267, 2011.

[14] Wenzhuo Ouyang, Atilla Eryilmaz, and Ness B Shroff. Asymptotically optimal downlink scheduling over markovian fading channels. In *2012 Proceedings IEEE INFOCOM*, pages 1224–1232. IEEE, 2012.

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

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

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

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

[19] Zhe Yu, Yunjian Xu, and Lang Tong. Large scale charging of electric vehicles: A multi-armed bandit approach. *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] This work is a theoretical work. We clearly explain under which conditions our results hold or not in the introduction and around the theorems.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is theoretical. We have read the seven points in Potential Negative Societal Impacts section on the website and tried to think of other applications but could not find any that has negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the discussion at the beginning of Section 6, as well as Section C.1.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 1 and Figure 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] All our simulation tasks can be ran in parallel, and the most time demanding ones are the different arms cases in Figure 1, and evaluating the LP-update policy in Figure 2. Both of them take no more than half an hour to complete on an eight cores CPU.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of the main theorems

A.1 Notations

We use the indices $a \in \{0,1\}$ for actions, $s, s' \in \{1 \dots d\}$ for states and $t \in \{0 \dots T-1\}$ for time steps (also called decision epochs). For conciseness, sums over a, s or t are implicitly sums over the whole range of the corresponding values (for instance, $\sum_{a,t} y_{s,a}(t)$ denotes $\sum_{a=0}^{1} \sum_{t=0}^{T-1} y_{s,a}(t)$, etc.)

A bold letter (e.g., $\mathbf{y}$, $\mathbf{m}$) denotes a vector whereas a normal letter (e.g., $y_{s,a}(t)$, $m_s(t)$) denotes a scalar. The bold letter $\mathbf{m}$ always denotes a state vector (that live in $\Delta^d \subset \mathbb{R}^d$) where as $\mathbf{y} = (\mathbf{y}_{\cdot,0}, \mathbf{y}_{\cdot,1})$ denotes a state-action vector (that live in $\Delta^{2d} \subset \mathbb{R}^{2d}$). For a vector $\mathbf{m} \in \mathbb{R}^d$, we denote by $\|\mathbf{m}\|_1 = \sum_s |m_s|$ the L_1 norm of $\mathbf{m}$. For a vector $\mathbf{y} \in \mathbb{R}^{2d}$, we denote by $\|\mathbf{y}\|_1 = \sum_{s,a} |y_{s,a}|$ the L_1 norm of $\mathbf{y}$. Apart from matrices, capital letters (e.g., $\mathbf{Y}^{(N)}$, $\mathbf{M}^{(N)}$) denotes random variables whereas small letters denote deterministic values (e.g., $\mathbf{y}$, $\mathbf{m}$). $\mathbb{E}\left[\mathbf{M}^{(N)}(t)\right]$ denotes the expectation and $\mathrm{var}\left[\mathbf{M}^{(N)}(t)\right]$ denotes the variance.

We recall that $\mathbf{y}^*$ denotes an optimal solution of the LP (3) and $\mathbf{m}^*$ is defined by $m_s^*(t) = \sum_a y_{s,a}^*(t)$.

A.2 Two technical lemmas

We first start the proofs by two technical lemmas that will be used in the proofs of all theorems. The first one emphasizes the relationship between the LP-index $I_s(\cdot)$ and the structure of the optimal decisions $\mathbf{y}^*$. The second one shows that the state vector $\mathbf{M}^{(N)}(t+1)$ is essentially a linear and deterministic function of $\mathbf{Y}^{(N)}(t)$.

A.2.1 Structure of the optimal problem

Lemma 4. *A vector $\mathbf{y}^*$ is a solution of the LP problem* (3) *if and only if the three following conditions hold:*

1. $y_{s,1}^*(t) = 0$ *for all s such that $I_s(t) < 0$,*
2. $y_{s,0}^*(t) = 0$ *for all s such that $I_s(t) > 0$,*
3. *It satisfies the structural constraints $y_{s,a}^*(t) \geq 0$,* (3b)*,* (3c) *and* (3d)*.*

Proof. Recall that the optimization (4) is

$$\begin{aligned} \max_{\mathbf{y} \geq \mathbf{0}} \quad & \sum_{t=0}^{T-1} \sum_{a,s} (R_s^a + a\gamma_t) y_{s,a}(t) \\ \text{s.t.} \quad & y_{s,0}(t+1) + y_{s,1}(t+1) = \sum_{s',a} y_{s',a}(t) P_{s's}^a \qquad \forall s, t, \\ & y_{s,0}(0) + y_{s,1}(0) = m_s \qquad \forall s. \end{aligned} \tag{7}$$

This problem can be formulated as the following Markov decision problem (MDP): Let X be a MDP with state space $\{1 \dots d\}$ and action space $\{0,1\}$. The reward in state $s \in \{1 \dots d\}$ under action $a \in \{0,1\}$ is $R_s^a + a\gamma_t$. The transition probabilities are $\mathbb{P}\big(X(t+1) = y \;\big|\; X(t) = x, \text{action} = a\big) = P_{xy}^a$. The initial condition is $X(0) \sim \mathbf{m}$, by interpreting $\mathbf{m}$ as a probability vector, and the time horizon is T. A policy for this MDP is a sequence $\psi_0, \dots, \psi_{T-1}$ where $\psi_t : \{1 \dots d\} \to \{0,1\}$ is the decision rule at time t and $\psi_t(s)$ is the probability that the action taken in state s is a. Using this interpretation, the LP problem (4) can be reformulated as

$$\begin{aligned} \max_{\mathbf{0} \le \psi \le \mathbf{1}} \quad & \sum_{t=0}^{T-1} \sum_{s} x_s(t) \left((R_s^1 + \gamma_t)\psi_t(s) + R_s^0(1-\psi_t(s)) \right) \\ \text{s.t.} \quad & x_s(t+1) = \sum_{s'} x'_s(t) \left(P_{s's}^1 \psi_t(s) + P_{s's}^0 (1-\psi_t(s)) \right) \qquad \forall s, t, \\ & x_s(0) = m_s \qquad \forall s. \end{aligned} \tag{8}$$

The equivalence between problem (7) and (8) comes by setting $y_{s,1}(t) = x_s(t)\psi_t(s)$ and $y_{s,0}(t) = x_s(t)(1-\psi_t(s))$.

Recall that $I_s(t) = Q_t(s,1) - Q_t(s,0)$ where $Q_t(s,a)$ are the Q-values of this MDP. By definition of Q-values, a policy ψ is optimal if and only if $\psi_s(t) = 0$ for all s such that $I_s(t) < 0$ and $\psi_s(t) = 1$ for all s such that $I_s(t) > 0$. Hence, a vector $\mathbf{y}^*$ is a solution of the LP problem (3) if and only if it is a solution to (4) and satisfies (3b). This is equivalent to the statement of the lemma. □

A.2.2 Deterministic approximation of the N-armed bandit

Recall that the function $\phi : \Delta^{2d} \to \Delta^d$ maps a vector $\mathbf{y} \in \Delta^d$ to a vector $\phi(\mathbf{y}) = ((\phi(\mathbf{y}))_1 \dots (\phi(\mathbf{y}))_d) \in \Delta^d$ whose sth component is

$$(\phi(\mathbf{y}))_s = \sum_{s',a} y_{s,a} P_{s',s}^a.$$

This induces the following lemma.

Lemma 5. *Let* $\mathbf{E}^{(N)}(t) := \mathbf{M}^{(N)}(t+1) - \phi\left(\mathbf{Y}^{(N)}(t)\right)$. *Then, we have*

$$\mathbb{E}\left[\mathbf{E}^{(N)}(t) \mid \mathbf{Y}^{(N)}(t)\right] = \mathbf{0}, \tag{9}$$

$$\mathbb{E}\left[\left\|\mathbf{E}^{(N)}(t)\right\|_1\right] \le \frac{\sqrt{d}}{\sqrt{N}}, \tag{10}$$

$$\mathbb{P}\left(\left\|\mathbf{E}^{(N)}(t)\right\|_1 \ge \epsilon\right) \le 2de^{-2N\epsilon^2/d^2}. \tag{11}$$

Proof. For notational convenience, let us denote by $\mathbf{y} := \mathbf{Y}(t)$. There are $Ny_{s,a}$ arms in state s and whose action is a. Each of these arms makes a transition to state s' with probability $P_{s,s'}^a$. This shows that

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

where the variables $U_{s,a,i}$ are i.i.d uniform random variable and where the function $\mathbf{1}_E$ is a random variable that equals 1 if the event E is true and 0 otherwise.

As a result, $\mathbb{E}\left[M_{s'}^{(N)}(t+1)\right] = (\phi(\mathbf{Y}^{(N)}(t)))_s$. Moreover,

$$\mathbb{E}\left[|E_{s'}^{(N)}(t+1)|^2\right] \le \text{var}\left[M_{s'}^{(N)}(t+1)\right] = \frac{1}{N^2} \sum_{s,a} N y_{s,a} P_{s,s'}^a (1 - P_{s,s'}^a) \le \frac{\sum_{s,a} y_{s,a} P_{s,s'}^a}{N}.$$

This shows that

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

where the first inequality comes from Cauchy-Schwarz.

Equation (11) is an almost direct consequence of Hoeffding's inequality. Indeed, by Hoeffding's inequality, one has $\mathbb{P}\left(|E_s^{(N)}(t+1)| \ge \varepsilon/d\right) \le 2e^{-N\varepsilon^2/d^2}$. By using the union bound, this implies that $\mathbb{P}\left(\left\|\mathbf{E}^{(N)}(t+1)\right\|_1 | \ge \varepsilon\right) \le d \cdot \mathbb{P}\left(|E_s^{(N)}(t+1)| \ge \varepsilon/d\right) \le 2de^{-N\varepsilon^2/d^2}$. □

A.3 Proof of Theorem 1

By definition, $V_{\rm rel}(\mathbf{m}, T) = \sum_{t,a,s} y^*_{s,a}(t) R^a_s$ where $\mathbf{y}^*$ is the solution of the LP (3). Moreover, as $Y^{(N)}_{s,0}(t)$ is the fraction of arms that are in state s and not activated at time t, and $Y^{(N)}_{s,1}(t)$ is the fraction of arms that are in state s and activated at time t, the reward of the stochastic system is $V^{(N)}_{\rm LP-reg}(\mathbf{m}, T) = \sum_{t,a,s} \mathbb{E}\left[Y^{(N)}_{s,a}(t) R^a_s\right]$. We show below that, for regular models, there exists constants C_1 and $C_2 > 0$ such that $\left\| \mathbb{E}\left[Y^{(N)}_{s,a}(t)\right] - y^*_{s,a}(t) \right\|_1 \leq C_1 e^{-C_2 N}/T$, which implies the theorem since $|R^a_s| \leq 1$ for all s, a.

Denote by σ_t the permutation of the states $\{1 \dots d\}$ that corresponds to the strict priority used by the LP-regular policy (this policy is such that $\sigma_t(1)$ has the largest index and $\sigma_t(d)$ has the smallest index). Let $s_t(\mathbf{m})$ be a state defined as

$$s_t(\mathbf{m}) := k \in \{1 \dots d\} \text{ such that } \sum_{i=1}^{k-1} m_{\sigma_t(i)} \leq \alpha < \sum_{i=1}^{k} m_{\sigma_t(i)}. \tag{12}$$

The LP-regular policy is a strict priority policy. Hence, by construction LP-regular will activate all arms that are in states $\sigma_t(1), \dots, \sigma_t(s_t(\mathbf{m}) - 1)$. It will activate some of the arms in state $s_t(\mathbf{m})$ in order to satisfy the constraint that the number of activated arms is exactly αN. It will activate no arms that are in states $\sigma_t(s_t(\mathbf{m}) + 1), \dots, \sigma_t(d)$. Therefore, we define $\pi^t_{LP-reg}(\mathbf{m}) := \mathbf{y} \in \Delta^{2d}$ as

$$(\pi^t_{LP-reg}(\mathbf{m}))_{s,1} = \begin{cases} m_s & \text{if } s \in \{\sigma_t(1), \dots, \sigma_t(s_t(\mathbf{m}) - 1)\}; \\ \alpha - \sum_{i=1}^{s_t(\mathbf{m})-1} m_{\sigma_t(i)} & \text{if } s = s_t(\mathbf{m}); \\ 0 & \text{if } s \in \{\sigma_t(s_t(\mathbf{m}) + 1), \dots, \sigma_t(d)\}; \end{cases}$$

$$\begin{aligned} (\pi^t_{LP-reg}(\mathbf{m}))_{s,0} &= m_s - (\pi^t_{LP-reg}(\mathbf{m}))_{s,1} \\ &= \begin{cases} 0 & \text{if } s \in \{\sigma_t(1), \dots, \sigma_t(s_t(\mathbf{m}) - 1)\}; \\ \sum_{i=1}^{s_t(\mathbf{m})} m_{\sigma_t(i)} - \alpha & \text{if } s = s_t(\mathbf{m}); \\ m_s & \text{if } s \in \{\sigma_t(s_t(\mathbf{m}) + 1), \dots, \sigma_t(d)\}; \end{cases} \end{aligned}$$

The function $\pi^t_{LP-reg}(\cdot)$ is a piecewise linear continuous function. Hence, it is Lipschitz-continuous. Moreover, by construction, for the stochastic N-arms system, one has $\mathbf{Y}^{(N)}(t) = \pi^t_{LP-reg}(\mathbf{M}^{(N)}(t))$.

As the problem is regular, for all t, there exists a unique state such that $I_s(t) = 0$. By Lemma 4, at time $t = 0$, this state must be $s_t(\mathbf{m}^*(0))$ because $I_s(0) > 0$ implies that $y^*_{s,1}(0) = m^*_s$ and $I_s(0) < 0$ implies that $y^*_{s,1}(0) = 0$. This implies that for $t = 0$, one has $\mathbf{y}^*(0) = \pi^t_{LP-reg}(\mathbf{m}^*(0))$. Moreover, by (3c), one has $\mathbf{m}^*(1) = \phi(\mathbf{y}^*(0))$. Hence, by induction on t, it holds that $y^*(t) = \pi^t_{LP-reg}(\mathbf{m}^*(t))$ for all $t \in \{0 \dots T-1\}$.

As $y^*_{s,1}(t) > 0$ and $y^*_{s,0}(t) > 0$ for $s = s_t(\mathbf{m}^*(t))$, by definition of $s_t(\mathbf{m})$ in (12), there exists ε_t such that for all $\mathbf{m} \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon_t)$ one has $s_t(\mathbf{m}) = s_t(\mathbf{m}^*(t))$, where $\mathcal{B}(\mathbf{m}^*(t), \varepsilon_t) := \{\mathbf{m} : \|\mathbf{m} - \mathbf{m}^*(t)\|_1 \leq \varepsilon_t\}$ is the ball centered at $\mathbf{m}^*(t)$ of radius ε_t. Taking $\varepsilon := \max_t \varepsilon_t$, this implies that:

$$\text{There exists } \varepsilon > 0 \text{ such that, for all } t, \text{ the function } \pi^t_{LP-reg} \text{ is linear on } \mathcal{B}(\mathbf{m}^*(t), \varepsilon). \tag{13}$$

Let $\delta > 0$ whose value will be determined later, and let $\mathcal{E}(\delta)$ be the event "for all $t \in \{0 \dots T-1\}$: $\|\mathbf{E}^{(N)}(t) \leq \delta\|$", where $\mathbf{E}^{(N)}(t) := \mathbf{M}^{(N)}(t+1) - \phi\left(\mathbf{Y}^{(N)}(t)\right)$ is as in Lemma 5. Let $\bar{\mathcal{E}}(\delta)$ be the complementary of the event $\mathcal{E}(\delta)$. Let L_t be the Lipschitz constant of the map $\phi \circ \pi^t_{LP-reg}$ and $L := \max_t L_t$. Assume that $\mathcal{E}(\delta)$ holds, we have:

$$\begin{aligned} \|\mathbf{M}^{(N)}(t+1) - \mathbf{m}^*(t+1)\| &= \|\phi(\pi^t_{LP-reg}(\mathbf{M}^{(N)}(t))) + \mathbf{E}^{(N)}(t) - \phi(\pi^t_{LP-reg}(\mathbf{m}^*(t)))\| \\ &\leq \|\phi(\pi^t_{LP-reg}(\mathbf{M}^{(N)}(t))) - \phi(\pi^t_{LP-reg}(\mathbf{m}^*(t)))\| + \delta \\ &\leq L \|\mathbf{M}(t) - \mathbf{m}^*(t)\|_1 + \delta \\ &\leq (1 + \dots + L^t)\delta, \end{aligned}$$

where the last inequality is a direct induction until $t = 0$ (and holds because $\mathbf{M}^{(N)}(0) = \mathbf{m}^*(0) = \mathbf{m}$).

Let $\varepsilon > 0$ be as in 13 above, and let $F_t : \Delta^d \to \mathbb{R}^{2d}$ be a linear function such that $\pi^t_{LP-reg}(\mathbf{m}) = F_t(\mathbf{m})$ for $\mathbf{m} \in \mathcal{B}(\mathbf{m}^*(t), \varepsilon)$. Take $\delta := \varepsilon/(1 + \cdots + L^{T-1})$. The above equation implies that when $\mathcal{E}(\delta)$ is true, one has $\pi^t_{LP-reg}(\mathbf{M}^{(N)}(t)) = F_t(\mathbf{M}^{(N)}(t))$ holds for all t. Hence,

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

where we used the linearity of $F_t \circ \phi$.

This implies that

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

where L' is an upper bound on the Lipschitz constants of maps $F_t \circ \phi$ for $0 \le t \le T-1$.

By Lemma 5, Equation (11), $\mathbb{P}\left(\mathcal{E}(\delta)\right) \ge 1 - T\mathbb{P}\left(\left\|\mathbf{E}^{(N)}(t)\right\|_1 \ge \epsilon\right) \ge 1 - 2Td e^{-2N\delta^2/d^2}$. Hence, taking $C_2 := -2\varepsilon^2/((1 + \cdots + L^{T-1})^2 d^2)$, we have

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

This concludes the proof by using $C_1 = 4T^2d^2(1 + (L')^T)$.

A.4 Proof of Theorem 2

The most laborious part of the proof is to translate the LP-filling policy give in Definition 2 into a map $\pi^t_{LP-fill}(\cdot)$. As in the proof of Theorem 1, let σ_t be the permutation of the states $\{1 \dots d\}$ that corresponds to the priority used at time t and let $s_t(\mathbf{m})$ be a state defined as in Equation (12). We also define $\mathcal{S}^+(t) := \{s \mid I_s(t) > 0\}$, $\mathcal{S}^0(t) := \{s \mid I_s(t) = 0\}$ and $\mathcal{S}^-(t) := \{s \mid I_s(t) < 0\}$. We distinguish three cases:

- If $s_t(\mathbf{m}) \in \mathcal{S}^+(t)$, this means that we will only activate states s such that $I_s(t) > 0$. In this case $\pi^t_{LP-fill}(\cdot)$ coincides with $\pi^t_{LP-reg}(\cdot)$.
- If $s_t(\mathbf{m}) \in \mathcal{S}^-(t)$, this means that we will activate all states s such as $I_s(t) \ge 0$ and some states s such that $I_s(t) < 0$. In this case $\pi^t_{LP-fill}(\cdot)$ also coincides with $\pi^t_{LP-reg}(\cdot)$.
- If $s_t(\mathbf{m}) \in \mathcal{S}^0(t)$, we will activate all states such that $I_s(t) > 0$ and a fraction $\beta := \alpha - \sum_{s:I_s(t)>0} \mathbf{M}^{(N)}_s(t) \ge 0$ states such that $I_s(t) = 0$. We detail this below.

Denote by

$$\widetilde{M}^{(N)}_s(t) := \begin{cases} M^{(N)}_s(t) & \text{for } s \in \mathcal{S}^+(t) \\ \min\left\{M^{(N)}_s(t), \frac{\lfloor y^*_{s,1}(t)N \rfloor}{N}\right\} & \text{for all } s \in \mathcal{S}^0(t)\end{cases}$$

By construction $N\widetilde{M}^{(N)}_s(t) \in \mathbb{N}$. We distinguish two subcases:

1. If $\sum_{s\in\mathcal{S}^0(t)} \widetilde{M}^{(N)}_s(t) \ge \beta$, by construction of $\pi^t_{LP-fill}(\cdot)$, this means that we will not activate more than $\widetilde{M}^{(N)}_s(t)$ fraction of arms in state $s \in \mathcal{S}^0(t)$. Let k be such that

$$\sum_{i=1}^{k-1} \widetilde{M}^{(N)}_{\sigma_t(i)} \le \alpha < \sum_{i=1}^{k} \widetilde{M}^{(N)}_{\sigma_t(i)}.$$

The activation vector $Y^{(N)}_{\cdot,1}(t)$ in this case is

- $Y^{(N)}_{s,1}(t) = \widetilde{M}^{(N)}_s(t)$, for $s = \sigma_t(i)$ with $i \leq k-1$ (we activate exactly $\widetilde{M}^{(N)}_s(t)$ of such arms);
- $Y^{(N)}_{\sigma_t(k),1}(t) = \alpha - \sum_{i=1}^{k-1} \widetilde{M}^{(N)}_{\sigma_t(i)}$ (we complete to activate a fraction exactly α of arms);
- $Y^{(N)}_{s,1}(t) = 0$, for $s = \sigma_t(i)$ with $i \geq k+1$.

2. If $\sum_{s\in\mathcal{S}^0(t)} \widetilde{M}^{(N)}_s(t) \geq \beta$, we need to activate more than $\widetilde{M}^{(N)}_s(t)$ arms in some of the states $s \in \mathcal{S}^0(t)$. Let

$$\gamma := \alpha - \sum_{s\in\mathcal{S}^+(t)} M^{(N)}_s(t) - \sum_{s\in\mathcal{S}^0(t)} \widetilde{M}^{(N)}_s(t) \geq 0,$$

and define $1 \leq k \leq |\mathcal{S}^0(t)|$ such that

$$\sum_{i=|\mathcal{S}^+(t)|+1}^{|\mathcal{S}^+(t)|+k-1} \left(M^{(N)}_{\sigma_t(i)}(t) - \widetilde{M}^{(N)}_{\sigma_t(i)}(t)\right) \leq \gamma < \sum_{i=|\mathcal{S}^+(t)|+1}^{|\mathcal{S}^+(t)|+k} \left(M^{(N)}_{\sigma_t(i)}(t) - \widetilde{M}^{(N)}_{\sigma_t(i)}(t)\right).$$

The values $\mathbf{Y}^{(N)}_{\cdot,1}(t)$ in this case are then given by

- $Y^{(N)}_{s,1}(t) = M^{(N)}_s(t)$, for $s = \sigma_t(i)$ with $1 \leq i \leq |\mathcal{S}^+(t)| + k - 1$ (we activate all of them);
- for $\sigma_t(|\mathcal{S}^+(t)| + k)$,

$$Y^{(N)}_{\sigma_t(|\mathcal{S}^+(t)|+k),1}(t) = \widetilde{M}^{(N)}_{\sigma_t(|\mathcal{S}^+(t)|+k)}(t) + \alpha - \sum_{i=|\mathcal{S}^+(t)|+1}^{|\mathcal{S}^+(t)|+k-1} \left(M^{(N)}_{\sigma_t(i)}(t) - \widetilde{M}^{(N)}_{\sigma_t(i)}(t)\right),$$

- $Y^{(N)}_{s,1}(t) = \widetilde{M}^{(N)}_s(t)$, for $s = \sigma_t(i)$ with $|\mathcal{S}^+(t)| + k + 1 \leq i \leq |\mathcal{S}^+(t)| + |\mathcal{S}^0(t)|$;
- $Y^{(N)}_{s,1}(t) = 0$, for $s \in \mathcal{S}^-(t)$.

This defines a map $\pi^{t,N}_{LP-fill}(\cdot)$ that depends on N because of the use of the integer part in the expressions $\lfloor y^*_{s,1}(t)N \rfloor / N$. This map is defined only for vectors $\mathbf{M}^{(N)}(t)$ such that every coordinate is an integer multiple of $1/N$. By abuse of notation, we define $\pi^{t,\infty}_{LP-fill}$ as the limit of this map when N goes to infinity. As the only difference between the two maps is the rounding, we have $\left\|\pi^{t,\infty}_{LP-fill}(\mathbf{m}) - \pi^{t,N}_{LP-fill}(\mathbf{m})\right\|_1 \leq 2d/N$ for all $\mathbf{m} \in \Delta^d$. Moreover, $\pi^{t,\infty}_{LP-fill}(\cdot)$ is Lipschitz continuous since it is piecewise linear and continuous.

By construction, $\mathbf{Y}^{(N)}(t) = \pi^{t,N}_{LP-fill}(\mathbf{M}^{(N)}(t))$. Moreover, if $\mathbf{m}^*$is the optimal trajectory given by the LP solution, then by Lemma 4, $\pi^{t,\infty}_{LP-fill}(\cdot)$ activates a fraction $m^*_s(t)$ in state $s \in \mathcal{S}^+(t)$ and a fraction $y^*_{s,1}(t)$ in state $s \in \mathcal{S}^0(t)$. Hence, for the LP solution: $\mathbf{y}^*(t) = \pi^{t,\infty}_{LP-fill}(\mathbf{m}^*(t))$.

Denote by K_t the Lipschitz constant of the map $\pi^{t,\infty}_{LP-fill}(\cdot)$, and let $K := \max_t K_t$, also let $K' := Kl$ where l is the Lipschitz constant of the map $\phi(\cdot)$. We have

$$\begin{aligned}
&\left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right] - \mathbf{y}^*(t)\right\|_1 \\
&= \left\|\mathbb{E}\left[\pi^{t,N}_{LP-fill}(\mathbf{M}^{(N)}(t))\right] - \pi^{t,\infty}_{LP-fill}(\mathbf{m}^*(t))\right\|_1 \\
&\leq \left\|\mathbb{E}\left[\pi^{t,N}_{LP-fill}(\mathbf{M}^{(N)}(t))\right] - \mathbb{E}\left[\pi^{t,\infty}_{LP-fill}(\mathbf{M}^{(N)}(t))\right]\right\|_1 + \left\|\mathbb{E}\left[\pi^{t,\infty}_{LP-fill}(\mathbf{M}^{(N)}(t))\right] - \pi^{t,\infty}_{LP-fill}(\mathbf{m}^*(t))\right\|_1 \\
&\leq \frac{2d}{N} + K'\left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t-1)\right] - \mathbf{y}^*(t-1)\right\|_1 + \frac{K\sqrt{d}}{\sqrt{N}}.
\end{aligned}$$

An easy induction then implies that $\left\|\mathbb{E}\left[\mathbf{Y}^{(N)}(t)\right] - \mathbf{y}^*(t)\right\|_1 \leq C'/\sqrt{N}$, with $C' > 0$ a constant independent of N.

A.5 Randomized rounding and proof of Theorem 3

A.5.1 Randomized rounding

We first explain how to use a *randomized procedure* to activate exactly αN arms while activating $Ny_{s,1}^{t*}(0)$ arms in state s in expectation. Note that by construction, $\sum_s Ny_{s,1}^{t*}(0) = \alpha N \in \mathbb{N}$ and $\mathbf{y}^{t*}(0) \leq \mathbf{M}^{(N)}(t)$. Hence, in a first pass, one can activate $\left\lfloor Ny_{s,1}^{t*}(0) \right\rfloor$ arms in state s. Let $z_s := Ny_{s,1}^{t*}(0) - \left\lfloor Ny_{s,1}^{t*}(0) \right\rfloor$. In a second pass, one activates an extra $ZN_s \in \{0,1\}$ arms in state s such that in expectation $\mathbb{E}[ZN_s] = z_s \in [0,1)$ and $\sum_s ZN_s = N\alpha - \sum_s \left\lfloor Ny_{s,1}^{t*}(0) \right\rfloor$.

This problem can be formalized as follows:

(Randomized rounding) *Consider d values $z_1 \dots z_d$ such that $z_s \in [0,1)$ and $\sum_s z_s = h \in \mathbb{N}$. Define $\mathbf{V} := \left\{ \mathbf{v} \in \{0,1\}^d \mid \sum_{s=1}^d v_s = \sum_s z_s = h \in \mathbb{N} \right\}$. The problem is to find a distribution μ among $\mathbf{V}$ such that if $\mathbf{v} \sim \mathbf{V}$, then $v_s \sim \mathbf{Bernoulli}(z_s)$ for $1 \leq s \leq d$.*

An efficient algorithm to solve this problem is presented in Section 5.2.3 of [10]. It has complexity $\mathcal{O}(hd \log d)$. The support of the distribution μ given by this algorithm is at most d, hence the algorithm is also space efficient.

A.5.2 Proof of Theorem 3

The value of the LP-update policy starting in $\mathbf{M}^{(N)}(t)$ at time t satisfies:

$$\begin{aligned} V_{\mathrm{LP-up}}^{(N)}(\mathbf{M}^{(N)}(t), T-t) &= \mathbb{E}\left[\sum_{s,a} Y_{s,a}^{(N)}(t) R_s^a + V_{\mathrm{LP-up}}^{(N)}(\mathbf{M}^{(N)}(t+1), T-(t+1)) \right] \\ &= \sum_{s,a} y_{s,a}^{t*}(0) R_s^a + \mathbb{E}\left[V_{\mathrm{LP-up}}^{(N)}(\mathbf{M}^{(N)}(t+1), T-(t+1)) \right], \end{aligned} \tag{14}$$

where $\mathbf{y}^{t*}$ is the solution of the LP problem with initial condition $\mathbf{M}^{(N)}(t)$ and horizon $T-t$.

Moreover, by Bellman's principle of optimality, we have:

$$V_{\mathrm{rel}}(\mathbf{M}^{(N)}(t), T-t) = \sum_{s,a} y_{s,a}^{t*}(0) R_s^a + V_{\mathrm{rel}}(\mathbf{m}^{t*}(1), T-(t+1)). \tag{15}$$

Denoting by $Z(t) := V_{\mathrm{LP-up}}^{(N)}(\mathbf{M}^{(N)}(t), T-t) - V_{\mathrm{rel}}(\mathbf{M}^{(N)}(t), T-t)$ and subtracting (15) to (14), we get:

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

By the general theory of linear programming, the function $V_{\mathrm{rel}}(\,\cdot\,, t) : \Delta^d \longrightarrow \mathbb{R}$ is Lipschitz continuous with a constant denoted ℓ_t (see for instance Section 5.6.2 of [5]). Denote also by $\ell := \max_t \ell_t$. We have:

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

Defining $\mathbf{E}^{(N)}(t)$ and ϕ as in Lemma 13, we have

$$\begin{aligned} \mathbf{M}^{(N)}(t+1) &= \phi(\mathbf{Y}^{(N)}(t)) + \mathbf{E}^{(N)}(t) \\ \mathbf{m}^{t*}(1) &= \phi(\mathbf{y}^{t*}(0)) \end{aligned}$$

Moreover, by construction $\left\| \mathbf{Y}^{(N)}(t) - \mathbf{y}^{t*}(0) \right\|_1 \leq 2d/N$ (the rounding error affects at most one arm in each state for each action). As ϕ is Lipschitz continuous (because it is linear) and as $\mathbb{E}\left[\left\| \mathbf{E}^{(N)}(t) \right\|_1 \right] = O(1/\sqrt{N})$, there exists a constant C' independent of N such that $\mathbb{E}\left[\left\| \mathbf{M}^{(N)}(t+1) - \mathbf{m}^{t*}(1) \right\|_1 \right] \leq C'/\sqrt{N}$.

This concludes the proof by using $C_4 = 2d\ell T C'$.

B Non-optimality of LP-regular policy for non-regular models

In this appendix we provide a numerical study on a 2 state ($d = 2$) 3 step ($T = 3$) restless bandit model that is not regular. It serves to illustrate that the LP-regular policy as defined in Section 4 is not asymptotically optimal for non-regular models. We report floating numbers with 3 digits of precision.

The parameters of the FHRB are as follows:

$$\mathbf{P}^1 = \begin{pmatrix} 0.2 & 0.8 \\ 0.95 & 0.05 \end{pmatrix}, \mathbf{P}^0 = \begin{pmatrix} 0.6 & 0.4 \\ 0.15 & 0.85 \end{pmatrix}, \tag{16}$$

$$\mathbf{R}^1 = (0.6, 0.2), \mathbf{R}^0 = (0, 0), \mathbf{m} = (0.5, 0.5), \alpha = 0.25. \tag{17}$$

The solution of the LP relaxation (3) gives

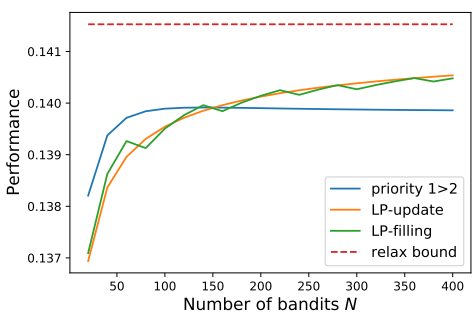
$$y_{1,0}(0) = 0.25,\ y_{1,1}(0) = 0.25,\ y_{2,0}(0) = 0.5,\ y_{2,1}(0) = 0;$$

$$y_{1,0}(1) \approx 0.12,\ y_{1,1}(1) \approx 0.04,\ y_{2,0}(1) \approx 0.63,\ y_{2,1}(1) \approx 0.21;$$

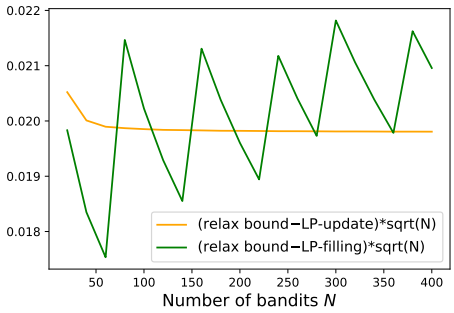
$$y_{1,0}(2) = 0,\ y_{1,1}(2) = 0.25,\ y_{2,0}(2) = 0.75,\ y_{2,1}(2) = 0.$$

The solution also gives $I_1(1) = I_2(1) = 0$, so the model is not regular at $t = 1$.

Since the system has only two states, the LP-regular policy is determined by an arbitrary order between the two states at time 1. If we choose "state 2 > state 1", the value is below 0.137 (not shown in Figure 3a). The best value is obtained by choosing "state 1 > state 2" at time one for LP-regular. We have computed the value of this priority policy for populations of arms N ranging from 20 to 400, using dynamic programming. The result is shown as the blue curve in Figure 3a. Also shown in this figure are the values of the LP-filling policy given in Definition 2 by water filling (the green curve), and the value of the LP-update policy given in Definition 3 (the orange curve), also computed exactly by dynamic programming.



(a) $d = 2$, $T = 3$ singular example.



(b) Verify the $\mathcal{O}(\frac{1}{\sqrt{N}})$ convergence rate.

Figure 3: Transient behavior of the values of three policies, LP-regular (aka priority), LP-filling and LP-update as N grows, for the small non-regular bandit problem given in (16) and (17).

Observe that numerically, the LP-filling policy and the LP-update policy become asymptotically optimal, whereas the LP-regular policy does not. This indicates that solving ties of $I_s(t)$ by using a single predetermined order is not working, and some more sophisticated considerations, taking $\mathbf{m}$ into account, are needed.

In Figure 3b, we plot the quantity $\left|V_{\mathrm{rel}}(\mathbf{m}, T) - V^{(N)}_{\mathrm{LP}-\star}(\mathbf{m}, T)\right| \sqrt{N}$ as a function of N for both LP-filling and LP-update to check if the convergence speed is indeed of the order $1/\sqrt{N}$, as claimed in Theorems 2 and 3. For LP-update, this quantity quickly stays close to 0.0198, and the oscillating effect of LP-filling is due to integer rounding, introducing an error term in $\mathcal{O}(1/N)$.

C Detailed Numerical Experiments

C.1 Experimental methodology and choice of performance measures

In our numerical studies in Section 6, we evaluate the performance of several policies on various parameter sets $\left\{(\mathbf{P}^0, \mathbf{P}^1, \mathbf{R}^0, \mathbf{R}^1); \alpha, N, T, \mathbf{m}\right\}$ via the following method: for each policy, we simulate 80 stochastic trajectories and compute the average reward. This is called the *simulated performance*. We also calculate the relaxed upper bound obtained by solving (3), as well as a relaxed lower bound where we replace the maximization of (3) by a minimization:

$$V_{\text{lower}}(\mathbf{m}, T) = \min_{\mathbf{y} \geq \mathbf{0}} \quad \sum_{t=0}^{T-1} \sum_{a,s} R_s^a y_{s,a}(t) \tag{18a}$$

$$\text{s.t.} \quad \sum_s y_{s,1}(t) = \alpha \qquad \forall t, \tag{18b}$$

$$y_{s,0}(t+1) + y_{s,1}(t+1) = \sum_{s',a} y_{s',a}(t) P_{s's}^a \qquad \forall s, t, \tag{18c}$$

$$y_{s,0}(0) + y_{s,1}(0) = m_s \qquad \forall s. \tag{18d}$$

We then define the *score* as a measure of performance of this policy on one parameter set:

$$\text{score} := \frac{\text{simulated performance} - V_{\text{lower}}(\mathbf{m}, T)}{V_{\text{rel}}(\mathbf{m}, T) - V_{\text{lower}}(\mathbf{m}, T)} \times 100$$

The score is a number between 0 and 100 and measures the relative performance w.r.t. the upper bound $V_{\text{rel}}(\mathbf{m}, T)$. Note that the perfect score 100 is not attainable for a finite number of bandits since we measure the relative performance w.r.t. the strict upper bound $V_{\text{rel}}(\mathbf{m}, T)$.

All numbers reported in the paper are taken by averaging the "score" of K randomly generated models (where $K \in \{10, 100\}$ depending on the time taken to compute a value of "score"). The confidence intervals that we report are $\pm 2\bar{\sigma}_{\text{score}}/\sqrt{K-1}$, where $\bar{\sigma}_{\text{score}}$ is the empirical standard deviations of those K scores.

All programming is done in Python, using `numpy` to generate random variables and to run the simulations, and using `PuLP` to solve the linear programs. All code is available in the supplementary material. All tasks were run in parallel on an Intel Corei7–8706G CPU. We estimate the total simulation time to be of the order of 10 hours on our machine. In Table 2 is shown the order of times needed to complete the simulations of one parameter set in Figure 1(for random, greedy and LP-filling policies) and Figure 2(for greedy, LP-filling, LP-update and Gittins policies). Note that:

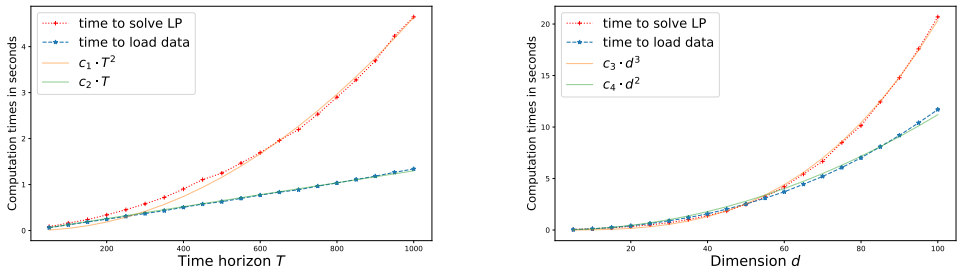
- The time is much larger for the "different models" than for the "identical models" because those models are of higher dimension.
- The time is also much larger for the rested models because for those models we implement Gittins and LP-update, that both take time.

Scenarios	Time for one parameter set
Identical arms – Figure 1a and 1b	5 seconds
Different arms – Figure 1c and 1d	4 minutes
Rested model (LP-update and Gittins) – Figure 2	7 minutes

Table 2: Average time to compute 80 simulation runs used to obtain the average "score" of one parameter set for each of the models on our machine. This includes the time to solve the upper and lower bounds.

C.2 Experimental complexity for solving the LP

The complexity to solve (3) may depend on the LP solver we use. For instance, the most common LP algorithm–the simplex method, can have exponential complexity in its worst case. To determine what is the practical complexity of the LP, we use the default LP solver from the `PuLP` package in Python and measure the time to construct and solve the LP.



(a) Fix d=5 and change T

(b) Fix T=30 and change d

Figure 4: Time complexity of the PuLP LP solver for (3).

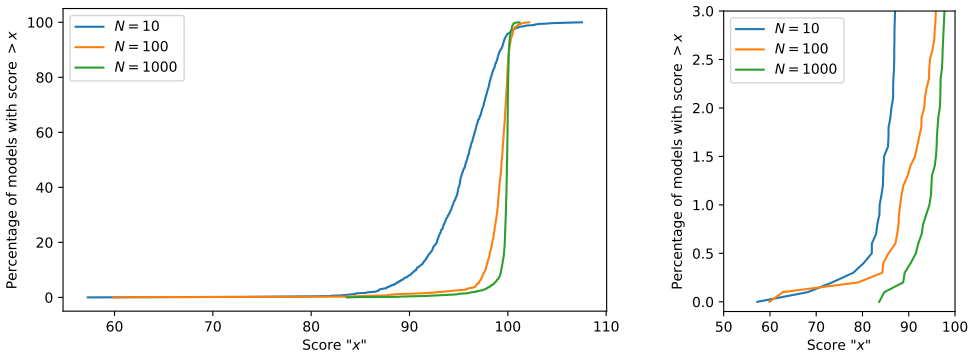
We first fix d=5 and let T vary from 50 to 1000. For each specific value of T in this range, we solve 1600 samples of random uniformly generated LP problems given in (3). We then record the average time elapsed to solve the LP, as well as the average time needed to load the data before solving the LP. The results are shown in Figure 4a. The constants c_1 and c_2 are determined by minimizing the mean squared error. Similarly, we fix T=30 and let d vary from 5 to 100. The results are shown in Figure 4b. This figure suggests that the empirical time complexity for solving the LP are to the order of $\mathcal{O}(T^2d^3)$.

For comparison, the algorithm proposed in [13] to compute the finite horizon Gittins index also has time complexity $\mathcal{O}(T^2d^3)$. The model considered in [13] is restful and can be seen as a special case of our restless bandit model here by taking $\mathbf{P}^0 = I_d$ (passive arms remain frozen). We also compare in Section 6 the performance of this finite horizon Gittins index policy with our LP policies for restful bandits.

C.3 The LP-update policy on "worst-case" models

As reported in Section 6, the LP-filling policy performs in general very well for all tested models, and its score converges quickly to 100 as N grows. Yet, there are some examples for which LP-filling does not perform that well. We study such examples in this section.

To obtain a model for which LP-filling does not perform well, we randomly generated 1000 sparse models and compute the score for $N = 10$, $N = 100$ and $N = 1000$. We report in Figure 5 the empirical CDFs of the scores. Note that some models have a score slightly above 100. This is due to randomness in the simulation (for better readability, the confidence intervals are not shown in this figure). Among these 1000 generated models, most of them have a score of 98 and above (for $N = 1000$) and the minimal score (for $N = 1000$) is 83.

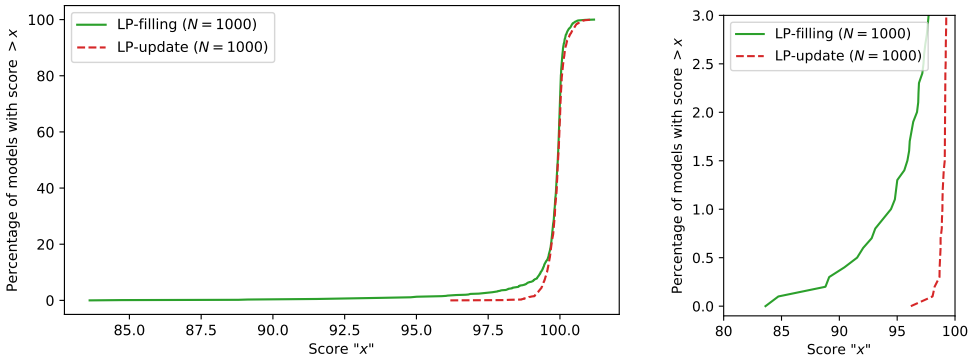


(a) Empirical CDF

(b) Zoom on the 3% worst performing examples.

Figure 5: Emprical CDF of the scores of the LP-filling policies for sparse models.

We now zoom on the case for $N = 1000$ for which LP-filling performs very well in the vast majority of the cases but only around 80 for worst cases. In Figure 6, we compare the performance of the LP-filling with the one of the LP-update policy. We observe that if LP-update performs similarly to LP-filling in the vast majority of cases, it largely improves the tail of the distribution: out of the same 1000 random sparse models, the worst performance is now around 96 (compared to 83 for the LP-filling).



(a) Empirical CDF

(b) Zoom on the 3% worst performing examples.

Figure 6: Comparison of the empirical CDF of the scores of the LP-update policy versus the LP-filling policy (for $N = 1000$).

To understand what makes the difference between the LP-update and LP-filling, in the rest of this section, we generate more models until we find a few of them for which the performance for $N = 1000$ was below 80 and study one of them in details. For this model, we compare the performance of the LP-filling and LP-update are reported in Table 3 for $N = 20, 1000, 10000, 100000$. Note that we also tested with other models that all have the same qualitative behavior.

Our first remark is that the LP-filling policy converges extremely slowly. Our second remark is that the LP-update policy greatly improves the performance, even for an arm population as small as $N = 20$. Following our discussion in Section 6, the LP-filling policy (based on the solution of a single LP) will apply the same priority order for a large proportion of the T time steps. If the deterministic system induced by this priority order is very *sensitive to perturbation*, then the stochastic trajectory is susceptible to deviate quickly from the optimal trajectory, and this can lead to significant sub-optimality even for very large N. Using LP-update (solving a new LP at each time step) corrects these deviations and constantly forces the stochastic trajectory to stay close to the optimal one, which explains the big advantage of LP-update to LP-filling on these sensitive models.

Policy	random	greedy	LP-filling	LP-update
$N = 20$	44.0	59.2	66.6	94.3
$N = 1000$	43.9	59.5	72.5	99.4
$N = 10000$	43.4	59.5	78.8	99.9
$N = 100000$	43.9	59.5	83.6	99.9

Table 3: Scores on one "worst-case" sparse model with N = 20, 1000, 10000, 100000.

The explanation for the relatively bad performance of LP-filling for this example comes from the stability properties of the deterministic function $\pi^{t,N}_{LP-fill}(\cdot)$. For this example, when starting from the initial condition

$$\mathbf{m} = [0.00217, 0.17684, 0.07073, 0.12528, 0.12958, 0.06513, 0.14282, 0.15356, 0.07572, 0.05817],$$

the deterministic optimal trajectory stays very close to the point

$$\mathbf{m}_{unstable} := [0.02853411, 0.0000000, 0.18785833, 0.49351031, 0.0681266, \\ 0.08889696, 0.02025134, 0.03834716, 0.02701284, 0.04746235] \in \Delta^{10},$$

and the LP-filling policy will use the priorities $\sigma := (4, 8, 9, 3, 5, 7, 2, 6, 1, 10)$ from time step 6 until time step 97.

The main problem is that applying a fixed priority order σ induces a map $\phi \circ \pi^{t,N}_{LP-fill}(\cdot)$ that is unstable around $\mathbf{m}_{unstable}$ (because it is locally linear and has an eigenvalue $1.16107056 > 1$). Consequently, the stochastic trajectory will in general does not stay close to $\mathbf{m}_{unstable}$ but will go closer to a stable point:

$$\mathbf{m}_{stable} := [0.00278525, 0.0000000, 0.09017394, 0.0040191, 0.25828372, \\ 0.42918881, 0.03318102, 0.0049398, 0.01724258, 0.16018578] \in \Delta^{10}.$$

The reward $\mathbf{m}_{stable}$ is much smaller than the one of $\mathbf{m}_{unstable}$, which explains why LP-filling performs poorly. On the other hand, when applying the LP-update policy, it will change dynamically the priorities in order to adapt to the stochastic trajectories. This will keep the stochastic trajectory close to $\mathbf{m}_{unstable}$ and will therefore provide a higher reward.