Regret bounds for Narendra-Shapiro bandit algorithms

Sébastien Gadat

Toulouse School of Economics, UMR 5604
Université Toulouse I Capitole, France
e-mail: sebastien.gadat@math.univ-toulouse.fr

Fabien Panloup

Institut de Mathématiques de Toulouse, UMR 5219
118, route de Narbonne F-31062 Toulouse Cedex 9, France
e-mail: fabien.panloup@math.univ-toulouse.fr

Sofiane Saadane

Institut de Mathématiques de Toulouse, UMR 5219
118, route de Narbonne F-31062 Toulouse Cedex 9, France
e-mail: sofiane.saadane@math.univ-toulouse.fr

Abstract: Narendra-Shapiro (NS) algorithms are bandit-type algorithms that have been introduced in the sixties (with a view to applications in Psychology or learning automata), whose convergence has been intensively studied in the stochastic algorithm literature. In this paper, we adress the following question: are the Narendra-Shapiro (NS) bandit algorithms competitive from a *regret* point of view? In our main result, we show that some competitive bounds can be obtained for such algorithms in their penalized version (introduced in [20]). More precisely, up to an over-penalization modification, the pseudo-regret $\bar{R}_n$ related to the penalized two-armed bandit algorithm is uniformly bounded by $C\sqrt{n}$ (where C is made explicit in the paper). We also generalize existing convergence and rates of convergence results to the multi-armed case of the over-penalized bandit algorithm, including the convergence toward the invariant measure of a Piecewise Deterministic Markov Process (PDMP) after a suitable renormalization. Finally, ergodic properties of this PDMP are given in the multi-armed case.

Keywords and phrases: Regret, Stochastic Bandit Algorithms, Piecewise Deterministic Markov Processes.

1. Introduction

1.1. Generalities and objectives

The so-called Narendra-Shapiro (NS) algorithm is a bandit-type algorithm which has been introduced in [23] and developed by [22] as a linear learning automata. Such algorithms have been primarily considered by the probabilistic community

*Authors are indebted to Sébastien Gerchinovitz and Aurélien Garivier for numerous motivating discussions on the subject.

as an interesting benchmark of stochastic algorithm. An almost complete historical overview on recursive markovian methods may be found in the seminal contributions of [12] or [16].

As a growing field of interest, many different bandit procedures have been developed recently and it is necessary to briefly recall what kind of d-armed bandit algorithm we will study in the sequel. Let us consider a slot machine with arms $A_1, \ldots, A_d$. When playing arm A_i, the probability of success is $p_i \in (0,1)$ (we get 1 coin with probability p_i and none with probability $1-p_i$). Of course, a gambler is not aware of the value of $p = (p_i)_{1 \leq i \leq d}$ and he wants to design a strategy to locate the optimal arm (for instance A_1 if we assume that $p_1 > p_j, \forall j \neq 1$). The Linear Reward Inaction (LRI) procedure defines a random Markov sequence of $\mathbb{R}^d$, which will be denoted $(X_n)_{n \geq 1}$ and X_n^i represents at step n the probability to select the arm A_i. A common expected task of bandit algorithms is to determine which arm is the most profitable. More precisely, an algorithm is an infallible bandit algorithm if it converges (at least in a probabilistic sense) toward the best arm as the number of attempts tends to infinity.

As mentioned before, stochastic multi-armed bandits are known in the field of stochastic approximation as a very simple example of recursive approximation scheme where the limit behaviour cannot be trivially described through the ODE methods mainly as a result of the presence of several noiseless traps. For instance, results of [24] cannot be conveniently applied in this context. Important advances on the understanding of NS-bandit algorithms have been obtained in [21] where an explicit dissection of the possible limiting behaviour of this algorithm is found and necessary and sufficient conditions of infallibility are given (see also [19] for a study on the convergence rates of bandit algorithms).

More recently, stochastic bandit algorithms have received an outstanding growing attention of statisticians and machine learning researchers, certainly owing to its widespread range of applications, for instance in game theory, statistics, signal processing, clinical trials or finance and its relevancy for Big Data problems. A pioneering work of [14] introduced the problem of the optimization of the expected reward of the algorithm, by using forwards induction policy as it is the case by the Markov algorithms quoted above. In most of the applications reminded above, the goal to maximize the rewards of the forecaster (and therefore its cumulative gain) appears quite naturally. Of course, the optimal strategy that would always select the arm A_{i^*} that maximises the probability of success, is not realistic since the probabilities $(p_1, \ldots, p_d)$ are unknown. In order to have a good estimate of the efficiency of the algorithm, one may first study the asymptotic rate of convergence of the algorithm when $n \longrightarrow +\infty$. But for practical purpose with a finite number of observations (in the framework of clinical trials for instance), one may prefer to focus on the so-called *regret* of the algorithm. It is defined in [18] as the loss, after a finite number of attempts, between the chosen policy and the best possible one, which consists in testing the best arm A_{i^*}. Hence, a natural way to assess the performance of a policy is to compute its *regret* and a good algorithm aims to minimize its expected regret after n steps (see Subsection 2.2 for a precise definition). There exists a

recent literature on statistical works that aim to optimize sequencial allocation procedures in several contexts: for online-regression (see [9]), adaptive routing ([15]) or many other applications. A complete state of the art may be found in the book [10] and the references therein. It is well known that the optimality of these several policies rely on an exploration-exploitation tradeoff as pointed in [3] and some recent sophisticated statistical procedures use confidence bounds to define suitable optimal ([3], [2], [8]).

In some sense, the penalized bandit algorithm developed in [20] also relies on an exploration-exploitation paradigm (see Subsection 2.4: the exploitation is given by the recursive reinforcement learning and the exploration relies on the penalization term. Note that when the penalization term is omitted, the NS-algorithm does not possess a natural exploration term, which implies some real difficulties to ensure infallibility. For NS-algorithms, convergence and asymptotic rates have been deeply studied in [21] and [19] for the *crude* two-armed NS-algorithm and in [20] for the penalized (two-armed) NS-algorithm (see below for background on crude and penalized NS-algorithms). But, to the best of our knowledge, there is no result about the regret in this context.

Thus, we propose in this paper to adress the following questions: are NS-algorithms competitive from a regret viewpoint and in the case of positive answer, what are the associated upper-bounds ?

Due to a lack of robustness (or more precisely to a too much constraining condition of infallibity), it seems that a crude NS-algorithm can not be competitive from a regret viewpoint. This is the purpose of Section 2.1 and 2.2 and 2.3 where we recall the definition of crude NS-algorithms and some basics on the regret.

Then, we mainly focus on the penalized NS-algorithms whose definition is given in Section 2.4. In this section, we also define *over-penalized* NS-algorithms which are slightly more penalized algorithms than the previous ones. From a theoretical point of view, this slight modification accentuates the robustness of the algorithm, which plays a non negligible role for the establishment of uniform upper-bounds for the pseudo-regret. It can also be mentioned that, by construction, these algorithms generate "*anytime*" bandit policies, *i.e.* that do not depend on the horizon, which may be a very important feature of bandit algorithms for practical applications.

Section 3 is devoted to the main results: in Theorem 3.2, we establish an upper-bound of the *pseudo-regret* $\bar{R}_n$ of the over-penalized algorithm in the two-armed. This upper-bound appears to be uniformly minimax with respect to all available bandit policies (see [2] for a complete state of the art on the regret in various frameworks, including our). A result is also given for the "original" penalized algorithm but the bound is not completely uniform.

In this section, we also extend some existing convergence and rate of convergence results of the algorithm to the multi-armed case. In the "critical" case (see below for details), the normalized algorithm converges in distribution toward a PDMP (Piecewise Deterministic Markov Process). In the two-armed case, we

also provide a sharp study of its convergence to equilibrium by using a method developed in [5].

The rest of the paper is devoted to the proofs of the main results: Section 4 presents the proofs of the regret analysis, Section 5 establishes the weak limit of the rescaled multi-armed bandit algorithm. At last, Section 6 gathers all the proofs of the ergodicity rates.

2. General settings and definitions

2.1. Crude NS-algorithm

Before going further, let us recall the mechanism of the NS-algorithms for multi-armed bandit problems with d arms. We first define d independent infinite sequences of i.i.d. Bernoulli random variables $A^i = (A_n^i)_{n\geq 1}$ where $A_n^i \sim \mathcal{B}(p_i)$ where $n \geq 1$ and $i \in [\![1; d]\!]$. Each A_n^i represents the test of the arm i at step n, its success occurs with the probability $p_i = \mathbb{P}(A_n^i = 1)$ and the optimal arm corresponds to the highest value of p_i. In the sequel, we will assume that only one arm is optimal (*i.e.* reaches the maximal value $\max_{1\leq i\leq d} p_i$) and without loss of generality, we define A_1 as the optimal one.

Given any real vector $p = (p_1, \ldots, p_d) \in (0,1)^d$, the crude NS-algorithm is a LRI algorithm that defines recursively a sequence of discrete probability measures $(X_n)_{n\geq 1}$ on $\{1, \ldots, d\}$. This sequence $(X_n)_{n\geq 1}$ is a Markov chain and this procedure is recalled in Algorithm 1.

Algorithm 1: Multi-armed NS algorithm

Input: $p = (p_1, \ldots, p_d) \in (0,1)^d$, Step size sequences $(\gamma_n)_{n\geq 1}$ of $(0,1)^{\mathbb{N}}$
Initialization: Initialize the probability vector $X_0 = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right) \in \mathcal{S}_d$.
for $n = 1$ *to* N **do**

- Sample $I_n \in \{1, \ldots, d\}$ with respect to the distribution X_{n-1}.
- Compute the reward of the chosen arm I_n, denoted $A_n^{I_n} \in \{0, 1\}$.
- Update the probability distribution as follows:

$$X_n^j = X_{n-1}^j + \gamma_n \left[\delta_{I_n}(j) - X_{n-1}^j\right] A_n^{I_n} \tag{1}$$

where $\delta_k(j)$ is equal to 1 if $k = j$ and 0 otherwise.

end

The behaviour of (1) is rather simple: the arm I_n is selected at step n with the current distribution X_n and is evaluated. If the arm fails, the weights are kept unchanged. In case of success, the weight of the arm I_n is increased by $\gamma_n \times (1 - X_n(I_n))$, where $1 - X_n(I_n)$ represents the remaining weight spread on the $d-1$ other arms. In the meantime, other weights of the other arms are decreased by a factor $(1 - \gamma_n)$.
In the two-armed case, the convergence of Algorithm 1 has been deeply studied in [21] and [19]. Without recalling these results, let us remark that :

- On the one hand, some non adaptive sequences of $(\gamma_n)_{n\geq 1}$ may lead to a very fast convergence to the target but in that case, a strong dependence between $(\gamma_n)_{n\geq 1}$ and the probabilities of success p_1 and p_2 is required (*i.e.* Algorithm 1 is not adaptive here). In particular, without any assumptions on (p_1, p_2) and $(\gamma_n)_{n\geq 1}$, the algorithm is fallible when its rate is fast: there exists some values of p_1 and p_2 such that there is a positive probability that the NS algorithm selects the wrong arm.
- On the other hand, the crude NS-algorithm may be infallible without conditions on p_1 and p_2 but in this case, the rate of convergence is very bad. According to the results of [19], the best (always) infallible case is obtained for $\gamma_n = 1/n$ and the associated rate is about $n^{-(p_1-p_2)}$.

As we will see below, these remarks will imply that the crude NS-algorithm can not be competitive from a regret point of view.

2.2. Regret of multi-armed bandit procedures

In bandit games, one considers some predictions where at each stage t between step 1 and step n, a forecaster chooses an arm I_n, receives a reward $A_n^{I_n}$ and then can use this information to choose the next arm at step $n+1$. As introduced in the seminal work of [25], the rewards are sampled independently from a fixed product distribution at each step t and a natural way to assess the performance of the policy is to compute its regret with respect to the best action. Using our notation A_n^i introduced in paragraph 2.1, the regret R_n is the random variable defined as

$$R_n := \max_{1\leq j\leq d} \sum_{k=1}^{n} \left[A_k^j - A_k^{I_k} \right].$$

where $(A_k^j)_{k,j}$ is a sequence of independent random variables. A good strategy corresponds to a selection procedure that minimizes the expected regret $\mathbb{E}R_n$, optimal ones are referred to as minimax strategies. A policy that yields an expected regret of the order $\inf \sup \mathbb{E}R_n$ (where the supremum is considered over all parameters $p \in (0,1)^d$ and the infimum is considered over all possible strategies for this game) are thus minimax. For a general overview on minimax algorithms for several kind of sequential games, we refer to [2] and [7]. The former expected regret cannot be easy handled and is generally replaced in statistical analysis by the pseudo-regret defined as

$$\bar{R}_n := \max_{1\leq j\leq d} \mathbb{E} \sum_{k=1}^{n} \left[A_k^j - A_k^{I_k} \right].$$

Next result describes how these two quantities are related.

Proposition 2.1. *Consider a stochastic bandit problem with Bernoulli rewards described by the vector p. We have (uniformly in p)*

$$0 \leq \mathbb{E}R_n - \bar{R}_n \leq \sqrt{\frac{n \log d}{2}}.$$

Furthermore, for every integers n and d and for any (admissible) strategy,

$$\mathbb{E}[R_n] \geq \frac{1}{20}\sqrt{nd} \quad \text{(uniformly in } p\text{)}.$$

We refer to Proposition 34 of [2] for a detailed proof of the first property and to Theorem 5.1 of [4] for the second one. As mentioned in the result, the bounds are uniform in p. The rate orders are strongly different if a dependence in p is authorized.

This result shows in particular that it is reasonable to focus on upper-bounds for the pseudo-regret in order to study those of the "true" regret. Note that such bounds exist in the literature. For example, it is shown in [1] that the MOSS policy leads to a uniform upper-bound of order $49\sqrt{nd}$ for the pseudo-regret.

2.3. Crude NS algorithm from a regret point of view

We now discuss on the regret of Algorithm 1. One can instantaneously remark that the infallibility of the procedure (convergence in probability to the good target) is in fact a necessary condition for the efficiency of the algorithm according to the regret point of view. In other words, when the algorithm is fallible, a moment thought leads to the conclusion that the growth of the regret with n is linear and is absolutely not competitive (minimax uniform rate is of the order $\sqrt{nd}$ up to multiplicative constants). Taken together, this last remark and paragraph 2.1 show that the crude NS Algorithm 1 cannot be convenient to minimize the regret. We are thus naturally driven to focus on the penalized NS-algorithm introduced in [20], which has some better adaptive convergence properties.

2.4. Penalized and Over-penalized NS-algorithm

The main difference between the crude NS-algorithm and its penalized version introduced by [20] relies on the exploitation of the failure of the selected arms. Algorithm 1 only uses the sequence of successes to update the probability distribution X_n over the d arms: the value of X_n is modified (and increased) *iff* $A_n^{I_n} = 1$. In the opposite, the penalized NS-algorithm of [20] also uses the information carried by the failure of the arm I_n and can decrease the value of the probability of the selected arm. Note that the penalized procedure of [20] is proposed only in the case of $d = 2$ arms. In that case, $x_n := X_n(1)$ (resp. $1 - x_n$) is the probability to choose the arm 1 (resp. the arm 2) at time t. The update formula of $(x_n)_{n\geq 1}$ is

$$\begin{aligned} x_n \quad = \quad & x_{n-1} + \gamma_n \left[\delta_{I_n}(1) - x_{n-1}\right] A_n^{I_n} \\ & -\gamma_n \rho_n \left[x_{n-1}\delta_{I_n}(1) - (1 - x_{n-1})\delta_{I_n}(2)\right] (1 - A_n^{I_n}) \end{aligned} \tag{2}$$

In case of success of the selected arm I_n, this algorithm mimics the crude NS-algorithm (increase of $X_n(I_n)$ by $\gamma_n \times (1 - X_n(I_n))$ and decrease of other weights

by a factor $(1-\gamma_n)$). But in case of failure, the weight of the selected arm is now decreased by a factor $(1-\gamma_n\rho_n)$ (whereas the probability of drawing the other arm is increased of the corresponding quantity).

We will see that this algorithm possesses some good properties from a regret point of view. However, we will also find that in some cases (when p_1 and p_2 are too close to 1), the penalty becomes too small to derive (theoretically) some uniform upper-bounds for the pseudo-regret. More precisely, if one assumes that $p_1 > p_2$, the capacity of $(X_n)_{n\geq 1}$ to get out of a neighborhood of 0 decreases with p_2 and can not be controlled uniformly. This is why we introduce in this paper a slightly "over-penalized" algorithm in order to bypass this lack of uniformity. For a given $\sigma \in [0,1]$, $(x_n^\sigma)_{n\geq 0}$, is the algorithm defined by:

$$\begin{aligned} x_n^\sigma &= x_{n-1}^\sigma + \gamma_n \left[\delta_{I_n}(1) - x_{n-1}^\sigma\right] A_n^{I_n} \\ &\quad -\gamma_n\rho_n \left[x_{n-1}^\sigma \delta_{I_n}(1) - (1-x_{n-1}^\sigma)\delta_{I_n}(2)\right] \left(1 - A_n^{I_n} B_n^\sigma\right) \end{aligned} \tag{3}$$

where $(B_n^\sigma)_n$ is a sequence of i.i.d. random variables with a Bernoulli distribution $\mathcal{B}(\sigma)$, independent of $(A_n^j)_{n,j}$ and such that for all $n \in \mathbb{N}$, B_n^σ and I_n are also independent. Writing

$$1 - A_n^{I_n} B_n^\sigma = 1 - A_n^{I_n} + A_n^{I_n}(1 - B_n^\sigma),$$

one should understand the over-penalization as (slightly) penalizing a successful arm with a probability σ. The case $\sigma = 1$ corresponds to the penalized bandit introduced in [20] described by Equation (2) whereas when $\sigma = 0$, the arm is always penalized when it plays. More precisely, this modification means that the increment of x_n^σ is slightly weaker than in the previous case.
It should be mention that this overpenalization is an exploration term, which plays a central role in the exploration-exploitation tradeoff and is a cornerstone for efficient bandit algorithms (see [18] for example). Hence, one should understand this overpenalization as a completion of the exploration ability of the penalized bandit: a success is randomly penalized to escape of local traps.

In the multidimensional case, the algorithm is given by the following procedure:

Algorithm 2: Multi-armed σ-Over-Penalized NS-algorithm

Input: $p \in (0,1)^d$, $\sigma \in (0,1)$, step size sequences $(\gamma_n)_{n\geq 1}$ and $(\rho_n)_{n\geq 1}$of $(0,1)^{\mathbb{N}}$

Initialization: Initialize the probability vector $X_0 = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right) \in \mathcal{S}_d$.

for $n = 1$ *to* N **do**

- Sample $I_n \in \{1, \ldots, d\}$ with respect to the distribution X_{n-1}.
- Compute the reward of the chosen arm I_n, denoted $A_n^{I_n} \in \{0,1\}$.
- Update the probability distribution according to

$$\begin{aligned} X_n^j &= X_{n-1}^j + \gamma_n \left[\delta_{I_n}(j) - X_{n-1}^j\right] A_n^{I_n} \\ &\quad -\gamma_n \rho_n X_{n-1}^{I_n}(1 - A_n^{I_n} B_n^{\sigma}) \left[\delta_{I_n}(j) - \frac{1-\delta_{I_n}(j)}{d-1}\right] \end{aligned} \tag{4}$$

end

Conversely to the two-armed case, we have here to choose how to distribute the penalty to the other arms. The (natural) choice made above is to divide it fairly, *i.e.* to spread it uniformly over the other arms. Note that alternative algorithms (not studied here) could be considered. The penalty could be distributed proportionally to the probabilities of drawing the other arms for example.

Once again, we will say that the algorithm is over-penalized if $\sigma < 1$. When $\sigma = 1$, the algorithm is a multi-armed version of (2) and is referred as the multi-armed penalized NS-algorithm.

Before stating the main results, one wants to understand what performances could be reached by the penalized NS-algorithms from a regret point of view. For the sake of simplicity, we choose to focus on the two-armed penalized NS-algorithm $(x_n)_n$ given by (2) and we recall (in a slightly less general form) the convergence results of Proposition 3, Theorems 3 and 4 of [20].

Theorem 2.1 (Lamberton & Pages, [20])**.** *Let $0 \leq p_2 < p_1 \leq 1$ and $\gamma_n = \gamma_1 n^{-\alpha}$ and $\rho_n = \rho_1 n^{-\beta}$ with $(\alpha, \beta) \in (0, +\infty)$ and $(\gamma_1, \rho_1) \in (0,1)^2$. Let $(x_n)_n$ be the algorithm given by (2).*

i) If $0 < \beta \leq \alpha$ and $\alpha + \beta \leq 1$, the penalized two-armed bandit is infallible.

ii) If furthermore $0 < \beta < \alpha$ and $\alpha + \beta < 1$, then $\dfrac{1-x_n}{\rho_n} \longrightarrow \dfrac{1-p_1}{p_2 - p_1}$ a.s.

iii) If $\alpha = \beta \leq 1/2$ and $g = \gamma_1/\rho_1$: $\dfrac{1-x_n}{\rho_n} \xrightarrow{\mathcal{L}} \mu$, where μ is the stationary distribution of the PDMP whose generator $\mathcal{L}$ acts on $\mathcal{C}_c^1(\mathbb{R}_+)$ as

$$\forall f \in \mathcal{C}_c^1(\mathbb{R}_+) \qquad \mathcal{L}(f)(y) = p_2 y \frac{f(y+g) - f(y)}{g} + (1 - p_1 - p_1 y) f'(y).$$

Now, let us stress that in the two-armed case,

$$\begin{aligned}\bar{R}_n &= p_1 n - \mathbb{E}\left(\sum_{k=1}^{n} A_k^{I_k}\right) \\ &= p_1 n - \mathbb{E}\left(\sum_{k=1}^{n} x_k p_1 + (1-x_k)p_2\right) = (p_1-p_2)\mathbb{E}\left(\sum_{k=1}^{n}(1-x_k)\right) \\ &= (p_1-p_2)\sum_{k=1}^{n}\rho_k \mathbb{E}\left(\frac{1-x_k}{\rho_k}\right). \end{aligned} \tag{5}$$

In particular,

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left[\frac{1-x_n}{\rho_n}\right] < +\infty \implies \bar{R}_n \leq C(p_1-p_2)\sum_{k=1}^{n}\rho_k, \tag{6}$$

where C is a constant that may depend on p_1 and p_2. In order to minimize the rate of increase of the pseudo-regret, one thus have to minimize the sequence $(\rho_n)_{n\in\mathbb{N}}$. By Theorem 2.1, it seems that the potential optimal choice corresponds to the one of statement (iii). Indeed, the infaillibility occurs only when $\alpha \geq \beta$ and $\alpha + \beta \leq 1$ and Equation (5) suggests that β should be chosen as large as possible to minimize the r.h.s. of (6), leading to $\alpha = \beta = 1/2$.
In this case, Equation (6) makes us think that $\bar{R}_n$ is of order $\sqrt{n}$, which is (owing to Proposition 2.1 and the comments below) the "good" order of convergence from a uniform minimax point of view. Thus, in the two-armed case, it seems that, in order to obtain a competitive upper-bound, it "remains to replace" the convergence in distribution of Theorem 2.1(iii) by a L^1-boundedness of the sequence $((1-X_n)/\rho_n)_{n\geq 1}$ and to show that the associated bound leads to uniform bounds for the pseudo-regret in terms of (p_1, p_2).

In the other cases where the algorithm is infallible, *i.e.* in Part (ii) of Theorem 2.1), the conditions imply that $\rho_n = n^{-\beta}$ with $\beta < 1/2$. In this case, the algorithm can not be competitive from a regret viewpoint. This is why in the sequel, we will only focus on the case

$$\gamma_n = \frac{\gamma_1}{\sqrt{n}} \quad \text{and} \quad \rho_n = \frac{\rho_1}{\sqrt{n}}.$$

It is important to point out that the penalized and over-penalized bandit algorithms are "anytime" policies, meaning that these algorithms are completely recursive and does not require the knowledge of the stopping horizon n. These policies do not require the use of a "doubling trick" (see [10] Section 2.3 for an example of making an "anytime" strategy from a dependent horizon one). At last, these bandit algorithms are very light and simple to handle from a numerical point of view.

3. Main Results

3.1. Regret of the over-penalized two-armed bandit

In this part, we establish some uniform upper-bounds for the two-armed σ-over-penalized NS-algorithm. Our main result is Theorem 3.2. Before stating it, we choose to state a new result when $\sigma = 1$, *i.e.* for the "original" penalized NS-algorithm introduced in [20].

Theorem 3.1. *Let $(X_n)_{n\geq 0}$ be the two-armed penalized bandit defined by (2) with $(\gamma_1, \rho_1) \in (0,1)^2$. Then, for every $\delta \in (0,1)$, there exists $C_\delta > 0$ such that*

$$\forall n \in \mathbb{N}, \qquad \sup_{(p_1,p_2)\in[0,1],p_2\leq p_1\wedge(1-\delta)} \bar{R}_n \leq C_\delta \sqrt{n}.$$

Remark 3.1. *The pseudo-regret of the original penalized-NS algorithm is not completely uniform. From a theoretical point of view, there is a lack of penalty when p_2 is too large, which in turns generates a lack of mean-reverting effect for the sequence $((1-X_n)/\rho_n)_{n\geq 1}$ when X_n is close to 0. Note that this constraint also appears numerically (see Figure 1 below). Following carefully the proof of the result, the constant C_δ could be made explicit.*

This explains the interest of the over-penalization, illustrated by the next result.

Theorem 3.2. *Let $(X_n)_{n\geq 0}$ be the two-armed σ-overpenalized bandit defined by (3) with $\sigma \in [0,1)$ and $(\gamma_1, \rho_1) \in (0,1)^2$.. Then,*

(a) There exists $C_\sigma(\gamma_1, \rho_1)$ such that

$$\forall n \in \mathbb{N}, \qquad \sup_{(p_1,p_2)\in[0,1],p_2<p_1} \bar{R}_n \leq C_\sigma(\gamma_1,\rho_1)\sqrt{n}.$$

(b) Furthermore, the choice $\sigma = 0$, $\gamma_n = 2.63\rho_n = 0.89/\sqrt{n}$ yields

$$\forall n \in \mathbb{N}, \qquad \sup_{(p_1,p_2)\in[0,1],p_2<p_1} \bar{R}_n \leq 31.1\sqrt{2n}. \tag{7}$$

Remark 3.2. *Once again, for every $\sigma > 0$, C_σ could be made explicit in terms of γ_1 and ρ_1. The second bound is obtained by an optimization of this quantity.*

In Figure 1, we draw a numerical approximation of $n \mapsto \sup_{p_2<p_1} \dfrac{R_n}{\sqrt{n}}$ for the penalized and over-penalized algorithms. First, remark that the blue curve indicates that the upper bound "44" in Theorem 3.2 is not sharp and the over-penalized algorithms satisfy a uniform upper-bound of the order $0.9 \times \sqrt{n}$. This bound is obtained with $\sigma = 0$, and $\gamma_n = \frac{1}{\sqrt{4+n}} = 4\rho_n$ (red line in Figure 1), suggesting that the rewards should *always* be over-penalized with $\rho_n = \frac{\gamma_n}{4}$. At last, the leading constant is always lower than 9/10 and converges to a limiting value when $n \longrightarrow +\infty$.
As compared to our theoretical results, it seems that the boundedness of $R_n/\sqrt{n}$ also holds uniformly over $(p_1 > p_2)$ for the standardly penalized one. However,

this numerical results proves that the over-penalization introduced in the NS-bandit algorithm is not only a theoretical tool but also leads to competitive numerical bounds.
At last, we observe that the statistical performance of the KL UCB algorithm (see *e.g.* [8] and references therein) are slightly better. Our simulations [1] suggest that the regret of the KL UCB algorithm satisfies the uniform bound $R_n \leq \sqrt{n}/2$, but it is worth noticing that the over-penalized bandit algorithm is much more faster than the initial UCB algorithm (phenomenon increased when compared to KL UCB).

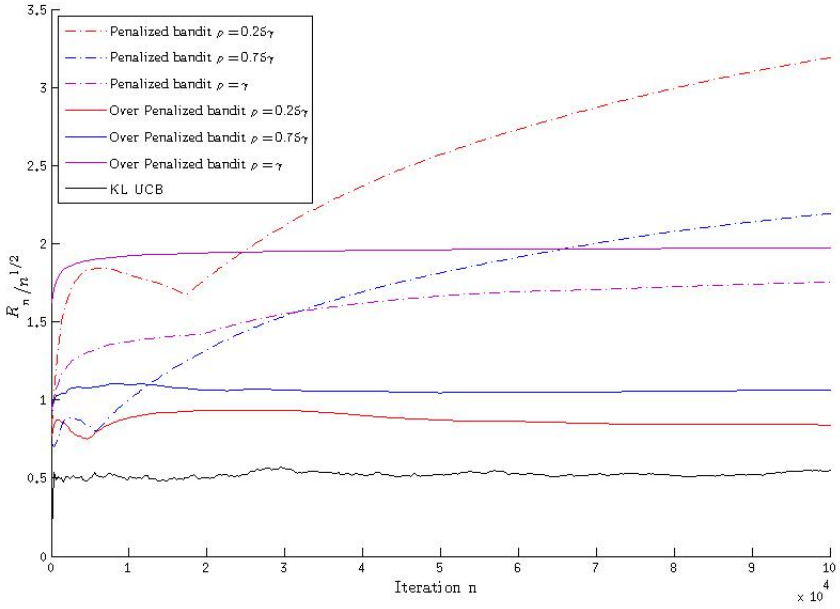

FIG 1. *Evolution of* $n \mapsto \sup_{(p_1,p_2)\in[0,1],p_2\leq p_1} \frac{\bar{R}_n}{\sqrt{n}}$ *for over-penalized algorithms (continuous colored line) and penalized algorithms (dashed colored line) and KL UCB (black line).*

3.2. Convergence of the multi-armed over-penalized bandit

We start this section by an extension of Theorem 2.1 of [20] to the (over) penalized algorithm in the multi-armed situation. The first result describes the pointwise convergence of the over-penalized algorithm.

[1] Performed with the Matlab package available on the website http://mloss.org/software/view/415/

Proposition 3.1 (Convergence of the multi-armed over-penalized bandit)**.** *Consider* $p_d \leq \ldots \leq p_2 < p_1$ *and* $\gamma_n = \gamma_1 n^{-\alpha}, \rho_n = \rho_1 n^{-\beta}$ *with* $(\alpha, \beta) \in (0, +\infty)$ *and* $(\gamma_1, \rho_1) \in (0,1)^2$*. Algorithm 2 with* $\sigma \in (0,1]$ *satisfies*

i) If $0 < \beta \leq \alpha$ *and* $\alpha + \beta \leq 1$*, then* $\lim_{n\to+\infty} X_n = (1, 0, \ldots, 0)$ *a.s.*
ii) If furthermore $0 < \beta < \alpha$ *and* $\alpha + \beta < 1$*, then*

$$\forall i \in \{2, \ldots, d\}, \quad \frac{X_n^i}{\rho_n} \longrightarrow \frac{1 - \sigma p_1}{(d-1)(p_1 - p_i)} \qquad a.s.$$

Proposition 3.2 provides a sharp description of the weak convergence of the *normalized* algorithm if we consider $Y_{n,j} = \frac{X_{n,j}}{\rho_n}$. It establishes that $(Y_{n,.})_{n\geq 0}$ converges toward a dynamic of a *Piecewise Deterministic Markov Process* (PDMP for short in what follows).

Proposition 3.2 (Weak convergence of the normalized algorithm)**.** *Under the assumptions of Proposition 3.1, if* $\alpha = \beta \leq 1/2$ *and* $g = \gamma_1/\rho_1$*, then*

$$\frac{1}{\rho_n}(X_{n,2}, \ldots, X_{n,d}) \xrightarrow{\mathcal{L}} \mu_{d,g},$$

where μ_d *is the (unique) stationary distribution of the Markov process whose generator* $\mathcal{L}_d$ *acts on compactly supported functions* f *of* $\mathcal{C}^1((\mathbb{R}_+)^{d-1})$ *as follows:*

$$\begin{aligned} \mathcal{L}_d f(y_2, ..., y_d) &= \sum_{i=2,\ldots,d} \frac{p_i y_i}{g}(f(y_2, ..., y_i + g, ..., y_d) - f(y_2, ..., y_i, ...y_d)) \\ &+ \sum_{i=2,\ldots,d} (\frac{1 - \sigma p_1}{d-1} - p_1 y_i)\partial_i f(y_2, ..., y_d). \end{aligned} \tag{8}$$

3.3. Ergodicity of the limiting process

In this section, we focus on the long behavior of the limiting Markov process which appears (after normalization) in Proposition 3.2. As mentioned before, this process is a PDMP and its long time behavior can be sharply studied with some arguments in the spirit of [5]. We also learned the existence of a close study in the PhD thesis of Florian Bouguet (personal communication). Such properties are stated for both the one-dimensional and the multidimensional cases.

3.3.1. One-dimensional case

Setting

$$a = 1 - p_1, b = p_1, g = \frac{\gamma_1}{\rho_1}, c = \frac{p_2}{g},$$

the generator $\mathcal{L}$ given by Proposition 3.2 may be written for any $\mathcal{C}^1$-function $f : \mathbb{R}_+^* \to \mathbb{R}$ as

$$\mathcal{L}f(x) = \underbrace{(a - bx)f'(x)}_{\text{deterministic part}} + \underbrace{cx}_{\text{jump rate}} \underbrace{(f(x+g) - f(x))}_{\text{jump size}} \tag{9}$$

In what follows, we will assume that a, b, c and g are positive numbers. On the one hand, the deterministic flow that guides the PDMP between the jumps is given by

$$\begin{cases} \partial_t \phi(x,t) &= (a - bx)\partial_x \phi(x,t) \\ \phi(x,0) &= x \in \mathbb{R}_+^* \end{cases}$$

so that

$$\phi(x,t) = \frac{a}{b} + \left(x - \frac{a}{b}\right) e^{-bt}.$$

Hence, if $x > \frac{a}{b}$ (resp. $x < \frac{a}{b}$), $t \mapsto \phi(x,t)$ decreases (resp. increases) and converges exponentially fast to $\frac{a}{b}$.

On the other hand, the PDMP possesses some positive jumps occuring with a Poisson intensity "$c.x$", whose size is deterministic and equals to g.
From the finiteness and positivity of g, it is easy to show that for every positive starting point, the process is $a.s.$ well-defined on $\mathbb{R}_+$, positive and does not explode in finite time. The fact that the size of the jumps is deterministic is less important and what follows could be easily generalized to a random size g (under adapted integrability assumptions). In Figure 2 below, some paths of the process are represented with different values of the parameters.

3.3.2. Convergence results

As pointed in Figure 2, the long-time behavior of the process certainly depends on the relation between the mean-reverting effect generated by "$-bx$" and the frequency and size of the jumps.

Invariant measure The process admits a unique invariant distribution if $b - cg > 0$. Actually, the existence is ensured by the fact that $V(x) = x$ is a Lyapunov function for the process:

$$\forall x \in \mathbb{R}_+^*, \quad \mathcal{L}V(x) = a - (b - cg)x = a - (b - cg)V(x)$$

Among other arguments, the uniqueness is ensured by Theorem 3.3 (the convergence in Wasserstein distance of the process toward the invariant distribution implies in particular its uniqueness). We denote it μ_∞ in what follows. It could be also shown that $\mathrm{Supp}(\mu_\infty) = (a/b, +\infty)$, that the process is strongly ergodic on $(a/b, +\infty)$ (see *e.g.* [13] for background) and that if $b - cg > 0$, the process explodes when $t \to +\infty$ (this case corresponds to the bottom left of Figure 2). Finally, let us remark that for the limiting PDMP of the bandit algorithm,

$$b - cg = p_1 - p_2 = \pi$$

and thus, the ergodicity condition coincides with the positivity of π.

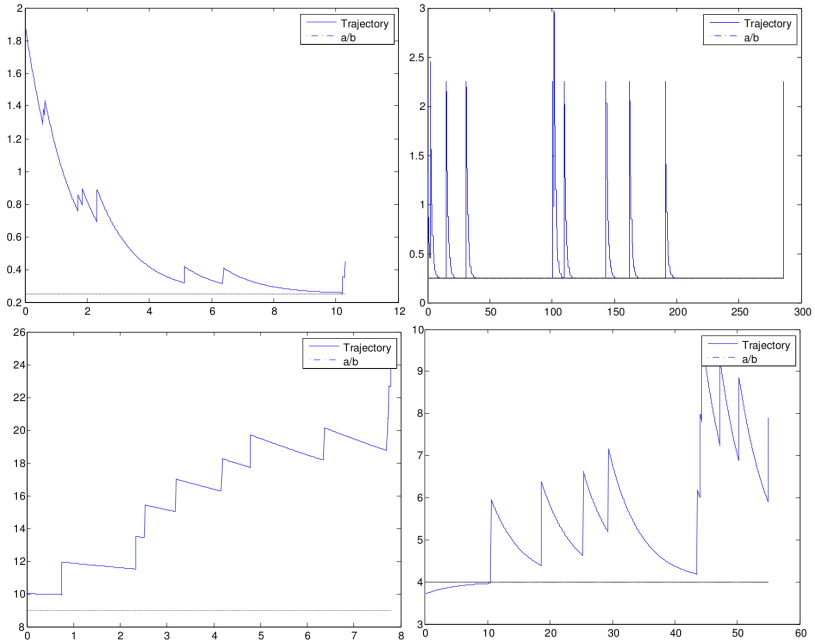

FIG 2. *Exact simulation of trajectories of a process driven by (9) when* $g = 0.1, a = 0.2, b = 0.8, c = 0.2$ *(top left)* $g = 2, a = 0.2, b = 0.8, c = 0.1$ *(top right),* $g = 2, a = 0.9, b = 0.9, c = 0.15$ *(bottom left) and* $g = 2, a = 0.8, b = 0.2, c = 0.05$ *(bottom right).*

Wasserstein results We aim to obtain rates of convergence results for the PDMP toward μ_∞ for two distances, namely the Wasserstein distance and the total variation distance. There exist rather different ways to obtain such results using coupling arguments or PDEs. Here, we use coupling techniques following the work of [6] and [11]. Before stating our results, let us recall that the p-Wasserstein distance is defined for any probability measures μ and ν on $\mathbb{R}^d$ by

$$\mathcal{W}_p(\mu,\nu) = \inf\left\{\mathbb{E}\left((X-Y)^p\right))^{\frac{1}{p}} \mid \mathcal{L}(X) = \mu, \mathcal{L}(Y) = \nu\right\}.$$

Denote by μ_0 the initial distribution of the PDMP and μ_t its law at time t, we can now state the main result on the PDMP associated to the rescaled two-armed overpenalized bandit, which is driven by (9).

Theorem 3.3 (One dimensional PDMP)**.** *Let* $p \geq 1$ *and denote for every* $t \geq 0$ $\mu_t := \mathcal{L}(X_t^{\mu_0})$ *where* $(X_t^{\mu_0})$ *is a Markov process driven by* (9) *with initial distribution* μ_0 *(with support included in* $\mathbb{R}_+^*$*). If* $p = 1$*, we have*

$$\left|\int x(\mu_0 - \mu_\infty)(dx)\right| e^{-\pi t} \leq \mathcal{W}_1(\mu_t, \mu_\infty) \leq \mathcal{W}_1(\mu_0, \mu_\infty)e^{-t\pi}$$

and if $p > 1$, a constant C_p exists such that

$$\mathcal{W}_p(\mu_t, \mu_\infty) \quad \leq \quad \gamma_p e^{-\frac{t\pi}{p}}.$$

where $(\gamma_p)_{p\geq 1}$ satisfies the recursion $\gamma_p^p = \gamma_{p-1}^{p-1}[pa + (1+g)^p]$.

Remark 3.3. *If $p = 1$, the lower and upper bounds imply the optimality of the rate obtained in the exponential. For $p > 1$, the optimality of the exponent $e^{-\pi t/p}$ is still an open question.*

We know give a corollary for the limiting process which appears in Proposition 3.2.

Corollary 3.1 (Multi-dimensional PDMP)**.** *Let $(Y_t)_{t\geq 0}$ be the PDMP driven by (8) with initial distribution $\mu_0 \in (\mathbb{R}_+^*)^d$ Then, the conclusions of Theorem 3.3 hold with $\pi = p_1 - p_2$.*

The proof of this result is almost obvious due to the "tensorized" form of the generator $\mathcal{L}_d$. Actually, for every starting point $y = (y_2, \ldots, y_d)$, all the coordinates $(Y_t^i)_{t\geq 0}$ are independent one-dimensional PDMPs with generator $\mathcal{L}$ defined by (9) with parameters

$$a_i = \frac{1 - \sigma p_1}{d-1}, \quad b_i = p_1 \quad \text{and} \quad c_i = p_i/g. \tag{10}$$

The result then follows easily from Theorem 3.3 with a global rate given by $\min\{b_i - c_i g, i = 2, \ldots, d\} = p_1 - p_2$. The details are left to the reader.

3.4. Total variation results

When some bounds are available for the Wasserstein distance, a classical way to deduce an upper bound of the total variation is to build a two-step coupling. In a first step, use a Wasserstein coupling to bring the paths sufficiently close (with a probability controlled by the Wasserstein bound). In a second step, we use a total variation coupling to try to stick the paths with a high probability. In our case, the jump size is deterministic and sticking the paths implies a non trivial coupling of the jump times. Some of the ideas to obtain the results below are in the spirit of [6], who follows this strategy for the TCP process.

Theorem 3.4. *Let μ_0 be a starting distribution with moments of any order. Then, for every $\varepsilon > 0$, there exists $C_\varepsilon > 0$ such that*

$$\|\mu_0 P_t - \mu_\infty P_t\|_{TV} \leq C_\varepsilon e^{-(\alpha\pi - \varepsilon)t} \quad \text{with } \alpha = \frac{1}{2 + \frac{b\pi}{ac}}.$$

Once again, this result can be extended to the multi-armed case.

Corollary 3.2. *Let $(Y_t)_{t\geq 0}$ be the PDMP driven by (8) with initial distribution $\mu_0 \in (\mathbb{R}_+^*)^d$ Then, the conclusions of Theorem 3.4 hold with $\alpha\pi$ replaced by*

$$\sum_{i=2}^{d} \frac{1}{2 + \frac{b_i \pi_i}{a_i c_i}} \pi_i$$

where $\pi_i = p_1 - p_i$ *and* a_i, b_i *and* c_i *are defined by* (10).

The proof of this result is based on the remark which follows Corollary 3.1. Owing to the "tensorization" property, the probability for coupling all the coordinates before time t is essentially the product of the probabilities of coupling of each coordinate. Once again, the details are left to the reader.

4. Proof of the regret bound (Theorems 3.1 and 3.2)

This section is devoted to the study of the regret of the penalized two-armed bandit procedure described in Algorithm 2. We will mainly focus on the proof of the explicit bound given in Theorem 3.2(b) and we will give the main ideas for the proofs of Theorems 3.1 and 3.1(a).

4.1. Notations

In order to lighten the notations, X_n^1 will be summarized by X_n, so that $X_n^2 = 1 - X_n$.
The proofs are then strongly based on a sharp study of the behavior of the (positive) sequence $(Y_n)_{n\geq 1}$ defined by

$$\forall n \geq 1 \qquad Y_n = \frac{1 - X_n}{\gamma_n}. \tag{11}$$

According to the important remark after Equation (5), we will consider in the sequel the following sequences $(\gamma_n)_{n\geq 1}$ and $(\rho_n)_{n\geq 1}$:

$$\forall n \geq 1, \quad \gamma_n = \frac{\gamma_1}{\sqrt{n}} \quad \text{and} \quad \rho_n = \frac{\rho_1}{\sqrt{n}} = \tilde{\rho}_1 \gamma_n \quad \text{and} \quad \tilde{\rho}_1 = \frac{\rho_1}{\gamma_1},$$

where γ_1 and ρ_1 are constants in $(0, 1)$ that will be precised later. In the meantime, we also define

$$\pi = p_1 - p_2 \in (0, 1).$$

With this setting, the pseudo-regret is

$$\bar{R}_n = \pi \sum_{n=1}^{n} \gamma_n \mathbb{E}[Y_n].$$

Remark here that we have substituted the division by ρ_n in (5) by a normalization with γ_n. This will be easier to handle in the sequel. The main question is now to obtain a convenient upper bound for $\mathbb{E}[Y_n]$. More precisely, remark that

$$\forall n_0 \in \mathbb{N} \quad \forall n \leq n_0 - 1, \quad \bar{R}_n \leq \pi n \leq \pi \sqrt{n_0 - 1} \sqrt{n},$$

and conversely for every $n \geq n_0$,

$$\begin{aligned}\frac{\bar{R}_n}{\sqrt{n}} &\leq \pi\sqrt{n_0-1}+\pi\sup_{n\geq n_0}\mathbb{E}[Y_n]\frac{1}{\sqrt{n}}\sum_{n=n_0}^{n}\frac{\gamma_1}{\sqrt{k}}\\ &\leq \pi\left(\sqrt{n_0-1}+2\gamma_1\sup_{n\geq n_0}\mathbb{E}[Y_n]\right). \qquad (12)\end{aligned}$$

Thus it is enough to derive an upper bound of $\mathbb{E}[Y_n]$ after an iteration n_0 that can be of the order $1/\pi^2$. In particular, the "suitable" choice of n_0 will strongly depend on the value of π.

4.2. Evolution of $(Y_n)_{n\geq 1}$

Random dynamical system In order to understand the mechanism and difficulties of the penalized procedure, let us first roughly describe the behavior of the sequences $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$. According to the definition of Algorithm 2 in the two-armed case, a careful inspection of (2) leads to

$$\begin{aligned}\mathbb{E}\left[X_{n+1}|X_n\right] &= X_n+\gamma_{n+1}X_n(1-X_n)\left[p_1-p_2\right]\\ &\quad +\gamma_{n+1}\rho_{n+1}\left[(1-X_n)^2(1-\sigma p_2)-X_n^2(1-\sigma p_1)\right].\end{aligned}$$

One can remark that the drift term may be splitted in two parts, the main part is the usual drift of Narendro-Shapira bandit algorithms described by h:

$$h(x)=[p_1-p_2]x(1-x). \qquad (13)$$

The second term comes from the penalization procedure and depends on σ:

$$\kappa_\sigma(x)=(1-\sigma p_2)(1-x)^2-(1-\sigma p_1)x^2. \qquad (14)$$

As a consequence, we can write the evolution of $(X_n)_{n\geq 0}$ as follows:

$$1-X_{n+1}=1-X_n-\gamma_{n+1}\left[h(X_n)+\rho_{n+1}\kappa_\sigma(X_n)+\Delta M_{n+1}\right], \qquad (15)$$

where ΔM_{n+1} is a Martingale increment. From the equation above, we easily derive that

$$\forall n\geq 1,\quad Y_{n+1}=Y_n\left(1+\gamma_n(\epsilon_n-\pi X_n)\right)-\rho_{n+1}\kappa(X_n)+\Delta M_{n+1}$$

where

$$\epsilon_n=\frac{1}{\gamma_{n+1}}-\frac{1}{\gamma_n}=\frac{1}{\gamma_1}\left(\sqrt{n+1}-\sqrt{n}\right)\leq\frac{1}{2\gamma_1\sqrt{n}}=\frac{\gamma_n}{2\gamma_1^2}. \qquad (16)$$

It follows that the increments of $(Y_n)_{n\geq 1}$ are given by

$$\Delta Y_{n+1}:=Y_{n+1}-Y_n=\gamma_n\varphi_n(Y_n)-\Delta M_{n+1}$$

where the drift function φ_n acting on the sequence $(Y_n)_{n\geq 1}$ is defined as

$$\varphi_n(y) = \underbrace{y \times [\epsilon_n + \pi(\gamma_n y - 1)]}_{:=\varphi_n^1(y)} - \underbrace{\frac{\rho_{n+1}}{\gamma_n}\kappa_\sigma(1-\gamma_n y)}_{:=\varphi_n^2(y)}.$$

To better understand the underlying effects of the dynamical system, recall that the definition of the sequence $(Y_n)_{n\geq 1}$ implies that $Y_n \in [0, {\gamma_n}^{-1}]$ with ${\gamma_n}^{-1} \sim Cn^{1/2}$. We aim to obtain a uniform bound (over n) of $\mathbb{E}[Y_n]$, it is thus important to understand the behaviour of the drift φ_n over $[0, {\gamma_n}^{-1}]$.

Crude algorithm In order to get an upper bound for $(\mathbb{E}(Y_n))_{n\geq 1}$, one generally aims to exploit the behaviour of the drift term $y \longmapsto \varphi_n(y)$ and needs to establish a mean reverting property out of a compact set (here for large values of y). When dealing with the crude bandit algorithm (*i.e.* when $\rho_1 = 0$ and described by Algorithm 1), the drift is reduced to φ_n^1, which does not produce a sufficient mean reverting effect: $\varphi_n^1(y)$ is negative *iff*

$$\epsilon_n - \pi(1-\gamma_n y) < 0 \Longleftrightarrow y \leq {\gamma_n}^{-1} - \frac{\epsilon_n}{\pi\gamma_n} \Longleftrightarrow x \geq \frac{\epsilon_n}{\pi}.$$

When x is close to 0 (in some sense depending on n, π and γ_1), φ_n^1 becomes repulsive: the fact that the crude bandit algorithm does not always converge to the good target can be understood as a consequence of this remark.

Penalized algorithm When the drift φ_n contains a non null penalty, the second term $-\varphi_n^2$ may help the dynamics to be not repulsive when x is close to 0, *i.e.* when y is larger than $1/\gamma_n$. It can be checked that $\kappa_\sigma(0) = 1 - \sigma p_2$ and

$$\lim_{n\to+\infty} \varphi_n\left({\gamma_n}^{-1}\right) = \frac{1}{2\gamma_1^2} - \frac{\gamma_1}{\rho_1}(1-\sigma p_2).$$

This quantity is negative under the condition

$$1 - \sigma p_2 \geq \frac{\rho_1}{2\gamma_1^3}. \tag{17}$$

This last assumption can be easily fulfilled if one chooses a suitable triple $(\sigma, \rho_1, \gamma_1)$ and such a choice can be *independent* on p_2. However, this property can be false if one chooses $\sigma = 1$ and $\rho_1/(2\gamma_1^3) > 1 - p_2$. Furthermore, note that the mean reverting effect guaranteed by κ_σ is very weak (and in some sense too weak to obtain a uniform bound for $\mathbb{E}(Y_n)$ since $\varphi_n(\gamma_n^{-1}) = \mathcal{O}(1)$).
Finally, note that symmetrically, φ_n may become positive in the neighborhood of $y = 0$. We show in Figure 3 and Figure 4 the behaviour of φ_n in the two "opposite" cases when $1 - \sigma p_2 < \rho_1/(2\gamma_1^3)$ or when $1 - \sigma p_2 > \rho_1/(2\gamma_1^3)$. When σ is large (*i.e.* $\sigma = 1$ in Figure 3), the mean reverting effect of φ_n^2 may not be " strong enough" to neutralize φ_n^1. This is not yet the case when σ is chosen small enough (with respect to ρ_1 and γ_1) as pointed in Figure 4. In particular, $\sigma < 1 - \frac{\rho_1}{2\gamma_1^3}$ is convenient for *any* value of p_2.

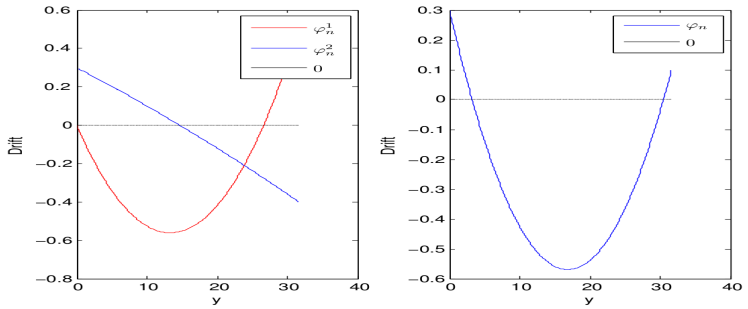

FIG 3. *Drift decomposition (left) and global (right) when* $y \in [0, \frac{1}{\gamma_n}]$ *with* $\gamma_1 = \rho_1 = 1$, $p_1 = 0.7$,$p_2 = 0.6$, $\sigma = 1$.

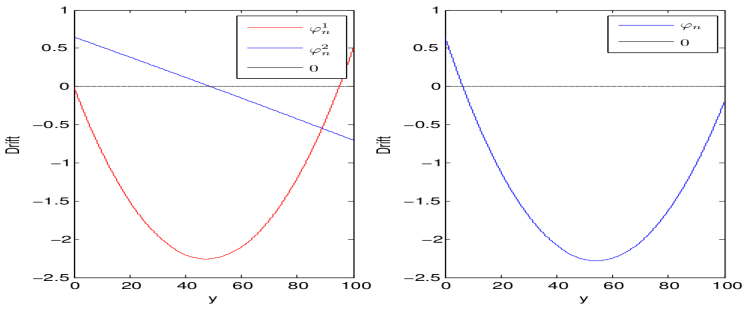

FIG 4. *Drift decomposition (left) and global (right) when* $y \in [0, \frac{1}{\gamma_n}]$ *with* $\gamma_1 = \rho_1 = 1$, $p_1 = 0.7$,$p_2 = 0.6$, $\sigma = 0.5$.

4.3. Increase of exponent

In order to overcome the difficulties previously described, the first natural idea is to introduce some stopping times relative to the 3 areas defined by φ_n in the spirit of [20]. Unfortunately, this approach remained unsuccessful, our concurrent idea is to introduce the nicer sequence $(Z_n^{(r)})_{n\geq 0}$:

$$\forall n \geq 1 \qquad Z_n^{(r)} = \frac{(1 - X_n)^r}{\gamma_n}. \tag{18}$$

Our strategy is first to establish a link between $\mathbb{E}[Z_n^{(r)}]$ and $\mathbb{E}[Z_n^{(r+1)}]$ and second to obtain a uniform control of $\mathbb{E}[Z_n^{(r)}]$ for a properly chosen integer r.

Let us define the bounded function on $[0, 1]$:

$$\forall \gamma \in [0, 1] \qquad h_r(\gamma) = \frac{(1 + \gamma)^r - 1 - r\gamma}{r\gamma^2}. \tag{19}$$

The first key element is given by the next proposition.

Proposition 4.1. *Let $r \in \mathbb{N}^*$, $\gamma_1 \in (0,1)$ and $0 < \epsilon \leq \epsilon_0 = \frac{1}{3}$, and set*

$$n_0(\epsilon, \pi, \gamma_1) := \left\lfloor \frac{1}{4\epsilon^2 \gamma_1^2 \pi^2} \right\rfloor + 1. \tag{20}$$

Then, if $2\epsilon\gamma_1^2(r-\varepsilon) \leq 1$,

$$\sup_{n \geq n_0} \mathbb{E} Z_n^{(r)} \leq \mathbb{E} Z_{n_0}^{(r)} + \frac{r}{\pi(r-\epsilon)} \left[\tilde{\rho}_1 + h_r(\gamma_{n_0}) + \pi \sup_{n \geq n_0} Z_n^{(r+1)} \right].$$

In particular, for $r = 1, 2$, the previous inequality holds for every $\gamma_1 \in (0,1)$ and $\varepsilon \in (0, 1/3]$.

Remark 4.1. *Note that the conditions on the parameters are not really intrinsic. The result is written in view to the establishment of the explicit constant given in Theorem 3.2 (b) for which we use the increase of exponent for $r = 1, 2$. In particular, in the proofs of Theorems 3.1 and 3.2 (a), this property will be used in a slightly different way.*

Proof For any integer $r > 0$ and $n \geq 0$, the binomial formula applied to (15) leads to

$$\begin{aligned} (1 - X_{n+1})^r &= (1 - X_n - \Delta X_{n+1})^r \\ &= (1 - X_n)^r - r(1 - X_n)^{r-1} \Delta X_{n+1} \\ &\quad + \sum_{j=0}^{r-2} \binom{r}{j} (1 - X_n)^j (-\Delta X_{n+1})^{r-j}. \end{aligned}$$

where $\sum_{\emptyset} = 0$ and $\Delta X_{n+1} = X_{n+1} - X_n = \gamma_{n+1}[h(X_n) + \rho_{n+1}\kappa_\sigma(X_n) + \Delta M_{n+1}]$. From the definition of h given in (13), we get

$$(1-x)^{r-1}[h(x) + \rho_{n+1}\kappa_\sigma(x)] = \pi x (1-x)^r + \rho_{n+1}\kappa_\sigma(x)(1-x)^{r-1}.$$

If we define now

$$\beta_n^{(r)} = -r\rho_{n+1}(1-X_n)^{r-1}\kappa_\sigma(X_n) + \frac{1}{\gamma_{n+1}} \sum_{j=0}^{r-2} \binom{r}{j} (1-X_n)^j (-\Delta X_{n+1})^{r-j}, \tag{21}$$

we can then conclude using (18) that

$$\begin{aligned} & Z_{n+1}^{(r)} \\ &= Z_n^{(r)} \frac{\gamma_n}{\gamma_{n+1}} - \gamma_n r \pi X_n Z_n^{(r)} + \beta_n^{(r)} - r(1-X_n)^{r-1} \Delta M_{n+1} \\ &= Z_n^{(r)} \left(1 + \gamma_n \left[\frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n} - r\pi X_n \right] \right) + \beta_n^{(r)} - r(1-X_n)^{r-1} \Delta M_{n+1} \\ &= Z_n^{(r)} \left(1 + \gamma_n \left[\epsilon_n - r\pi X_n \right] \right) + \beta_n^{(r)} - r(1-X_n)^{r-1} \Delta M_{n+1} \\ &= Z_n^{(r)} \left(1 + \gamma_n \left[\epsilon_n - r\pi \right] \right) + r\pi\gamma_n (1-X_n) Z_n^{(r)} + \beta_n^{(r)} - r(1-X_n)^{r-1} \Delta M_{n+1} \\ &= Z_n^{(r)} \left(1 + \gamma_n \left[\epsilon_n - r\pi \right] \right) + r\pi\gamma_n Z_n^{(r+1)} + \beta_n^{(r)} - r(1-X_n)^{r-1} \Delta M_{n+1}. \end{aligned} \tag{22}$$

The formulation above is important: it exhibits a contraction of $(1+\gamma_n[\epsilon_n - r\pi])$ on $Z_n^{(r)}$ that can be used jointly with an upper bound of $Z_n^{(r+1)}$ and a simple majorization of $\beta_n^{(r)}$. In this view, we study (21): $|\Delta X_{n+1}| \leq \gamma_{n+1}$ *a.s.* and (14) yields $|\kappa_\sigma(x)| \leq (1-\sigma p_2)$. Now, with h_r given in (19), we get

$$\beta_n^{(r)} \leq r\tilde{\rho}_1\gamma_{n+1} + \sum_{j=0}^{r-2}\binom{r}{j}(\gamma_{n+1})^{r-j-1} \leq r\left(\tilde{\rho}_1 + h_r(\gamma_{n+1})\right)\gamma_{n+1}.$$

For any $\epsilon \in (0,1)$, we can see in (22) that the contraction coefficient can be useful as soon as n is large enough. More precisely, using (16), we see that

$$\epsilon_n \leq \epsilon \Longleftarrow n \geq n_0(\epsilon,\pi,\gamma_1) := \left\lfloor \frac{1}{4\epsilon^2\gamma_1^2\pi^2} \right\rfloor + 1.$$

Then, for every $n \geq n_0(\epsilon,\pi,\gamma_1)$,

$$1+\gamma_n[\epsilon_n - r\pi] \leq 1 - \alpha_r\gamma_n \quad \text{with} \quad \alpha_r = \pi(r-\varepsilon).$$

In the sequel, we will omit the dependence of n_0 in (ϵ,π,γ_1) and will just use the notation n_0. Also remark that under the condition $2\epsilon\gamma_1^2(r-\varepsilon) \leq 1$, we have $\alpha_r\gamma_j < 1$ for every $\pi \in (0,1)$ and for every $j \geq n_0$ (one can in particular check that $2\epsilon\gamma_1^2(r-\varepsilon) \leq 1$ is true for every $\epsilon \in (0,1/3)$ and $\gamma_1 \in (0,1)$ if $r = 1,2$). Thus, by a simple recursion based on (22), one obtains for every $n \geq n_0+1$,

$$\mathbb{E}(Z_n^{(r)}) \leq \mathbb{E}(Z_{n_0}^{(r)})\prod_{j=n_0}^{n-1}(1-\alpha_r\gamma_j) + \sum_{j=n_0}^{n-1}\left(r\pi\gamma_j\mathbb{E}(Z_j^{(r+1)}) + \beta_j^{(r)}\right)\prod_{l=j}^{n-1}(1-\alpha_r\gamma_l)$$

If we call $\Theta_r = r\left(\pi \sup_{j\geq n_0}\mathbb{E}(Z_j^{(r+1)}) + \tilde{\rho}_1 + h_r(\gamma_j)\right)$, an iteration of the previous inequality yields:

$$\mathbb{E}(Z_n^{(r)}) \leq \mathbb{E}(Z_{n_0}^{(r)}) + \Theta_r\sum_{j=n_0}^{n-1}\gamma_j\prod_{l=j}^{n-1}(1-\alpha_r\gamma_l).$$

We aim to apply Lemma A.1 (deferred to the appendix section) to the last term. It is possible as soon as

$$n_0 \geq \frac{1}{(\alpha_r\gamma_1)^2}.$$

This last condition is fulfilled for any $r \geq 1$ when $\frac{1}{4\epsilon^2\gamma_1^2\pi^2} \geq \frac{1}{(1-\epsilon)^2\pi^2\gamma_1^2}$, *i.e.* when $\epsilon \leq 1/3$.
Then, by Lemma A.1, one deduces that $\forall\epsilon \leq 1/3$ and $\forall n \geq n_0$:

$$\sup_{n\geq n_0}\mathbb{E}Z_n^{(r)} \leq \mathbb{E}Z_{n_0}^{(r)} + \frac{r}{\pi(r-\epsilon)}\left[\tilde{\rho}_1 + h_r(\gamma_{n_0}) + \pi\sup_{n\geq n_0}Z_n^{(r+1)}\right].$$

□

From the last proposition and a recursive argument, we can now deduce the following key observations.

Corollary 4.1. *Assume that $\epsilon \in (0, 1/3)$, $\gamma_1 \in (0, 1)$ and that n_0 is defined in (20). Then,*

$$\sup_{n \geq n_0} \mathbb{E}Y_n \leq \mathbb{E}[Z_{n_0}^{(1)}] + \frac{\mathbb{E}[Z_{n_0}^{(2)}]}{1-\epsilon} + \frac{1}{\pi(1-\epsilon)} \left[\tilde{\rho}_1 + \frac{\tilde{\rho}_1}{1-\epsilon/2} + \frac{1}{2(1-\epsilon/2)} \right] + \frac{\sup_{n \geq n_0} \mathbb{E}Z_n^{(3)}}{(1-\epsilon)(1-\epsilon/2)}. \tag{23}$$

Remark 4.2. *Once again, this property is established in view to Theorem 3.2 (b). For Theorems 3.1 and 3.2 (a) with $\sigma \in (0, 1)$, we will need to use it for large values of r (see the end of this section for details).*

4.4. Bound for $(\mathbb{E}(Z_n^{(3)})_{n \geq n_0}$

As seen in Corollary 4.1, our next task is to bound $\mathbb{E}(Z_n^{(3)})$ for $n \geq n_0$ to obtain a tractable application of Equation (23). Such a bound is reached through a careful inspection of the increments $\Delta Z_{n+1}^{(3)} := Z_{n+1}^{(3)} - Z_n^{(3)}$.

Lemma 4.1 (Decomposition of $Z_n^{(3)}$)**.** *For every $n \geq 1$,*

$$\mathbb{E}[\Delta Z_{n+1}^{(3)} | \mathcal{F}_n] = \gamma_{n+1}(1 - X_n) P_n(X_n) + \Delta R_n,$$

where for every $n \in \mathbb{N}$, P_n is a polynomial function defined by

$$P_n(x) = \frac{(1-x)^2}{\gamma_{n+1}}(\varepsilon_n - 3\pi x) - 3\tilde{\rho}_1(1-x)\kappa_\sigma(x) + 3\left(x(1-x)^2 p_1 + x^2(1-x)p_2\right) + \gamma_{n+1}\left(-x(1-x)^2 p_1 + x^3 p_2\right), \tag{24}$$

and if γ_1 and n_0 satisfy the assumptions of Proposition 4.1, then

$$\forall n \geq n_0, \quad \Delta R_n \leq (1 - \sigma p_2)\left[3\gamma_{n+1}\rho_{n+1}^2 + \gamma_{n+1}^2 \rho_{n+1}^3\right].$$

Remark 4.3.

- *The keypoint is that $\gamma_k = \gamma_1 k^{-1/2}$ and thus the series $\sum_{n \geq 1} \Delta R_k$ is uniformly bounded whatever π is. This will be enough to obtain a competitive upper bound of the regret. With the choice of n_0 given in (20), a careful inspection of Lemma 4.1 leads to*

$$\sum_{k \geq n_0} \Delta R_k \leq 12\gamma_1^4 \tilde{\rho}_1^2 \epsilon\pi + \frac{16}{3}\gamma_1^8 \tilde{\rho}_1^3 \epsilon^3 \pi^3. \tag{25}$$

- *As in Remark 4.2, let us notice that for Theorems 3.1 and 3.2 (a) with $\sigma \in (0, 1)$, we will need to use such a development with some larger values of r (see the end of this section for details).*

Proof. We use again Equation (22) and deduce that

$$Z_{n+1}^{(3)} - Z_n^{(3)} = (1-X_n)^3(\epsilon_n - 3\pi X_n)) - 3\tilde{\rho}_1\gamma_{n+1}(1-X_n)^2\kappa_\sigma(X_n) \tag{26}$$

$$+\frac{1}{\gamma_{n+1}}\sum_{j=0}^{1}\binom{3}{j}(1-X_n)^j(-\Delta X_{n+1})^{3-j} - 3(1-X_n)^2\Delta M_{n+1} \tag{27}$$

First, remark that terms in Equation (26) are associated to the first two terms in the definition of P_n introduced in (24) up to a multiplication by $(1-X_n)\gamma_{n+1}$. Second, we can easily compute the expectations involved in the sum of Equation (27) since the events are all disjointed. On the one hand, when $j=1$ we have

$$\begin{aligned}\frac{1}{\gamma_{n+1}}\mathbb{E}[(-\Delta X_{n+1})^2|\mathcal{F}_n] \;&=\; \gamma_{n+1}\sigma\left(p_1X_n(1-X_n)^2 + p_2(1-X_n)X_n^2\right)\\ &+\gamma_{n+1}(1-\sigma)\left(p_1X_n(1-X_n-\rho_{n+1}X_n)^2 + p_2(1-X_n)(X_n-\rho_{n+1}(1-X_n))^2\right)\\ &+\gamma_{n+1}\rho_{n+1}^2\left[X_n^3(1-p_1)+(1-X_n)^3(1-p_2)\right].\end{aligned}$$

Further computations yield:

$$\begin{aligned}\frac{1}{\gamma_{n+1}}\mathbb{E}[(-\Delta X_{n+1})^2|\mathcal{F}_n] \;&=\; \gamma_{n+1}X_n\left(\sigma p_1(1-X_n)^2 + \sigma p_2X_n^2\right) + \Delta A_n^{(1)}\\ &\quad+\underbrace{\gamma_{n+1}\rho_{n+1}^2\left[X_n^3(1-\sigma p_1)+(1-X_n)^3(1-\sigma p_2)\right]}_{:=\Delta R_n^{(1)}}\end{aligned}$$

with $\Delta A_n^{(1)} = -2\rho_{n+1}\gamma_{n+1}X_n(1-X_n)(1-\sigma)(X_np_1+(1-X_n)p_2)$. On the other hand, we can also compute the term when $j=0$:

$$\begin{aligned}\frac{1}{\gamma_{n+1}}\mathbb{E}[(-\Delta X_{n+1})^3|\mathcal{F}_n] \;&=\; \gamma_{n+1}^2X_n(1-X_n)\left(p_2X_n^2 - p_1(1-X_n)^2\right)\\ &\quad+\Delta A_n^{(2)} + \underbrace{\gamma_{n+1}^2\rho_{n+1}^3\left[X_n^4(1-\sigma p_1)-(1-X_n)^4(1-\sigma p_2)\right]}_{:=\Delta R_n^{(2)}}\end{aligned}$$

with $\Delta A_n^{(2)} \le 3\gamma_{n+1}^2\rho_{n+1}(1-\sigma)X_n(1-X_n)^2\left(\pi X_n + \rho_{n+1}(1-X_n)p_2\right)$. Set $\Delta R_n^{(3)} = (1-X_n)\Delta A_n^{(1)} + \Delta A_n^{(2)}$ and $\Delta R_n := 3(1-X_n)\Delta R_n^{(1)} + \Delta R_n^{(2)}$. Plugging the previous controls into (27) yields

$$\mathbb{E}[\Delta Z_{n+1}^{(3)}|\mathcal{F}_n] \le \gamma_{n+1}(1-X_n)P_n(X_n) + \Delta R_n. \tag{28}$$

Note that $\Delta R_n^{(1)}$ can be upper bounded as follows:

$$3(1-X_n)\Delta R_n^{(1)} \le 3\gamma_{n+1}\rho_{n+1}^2(1-\sigma p_2)\max_{0\le t\le 1}\left[\frac{1-\sigma p_1}{1-\sigma p_2}t^3(1-t)+(1-t)^4\right].$$

Since $1-\sigma p_1 \le 1-\sigma p_2$, a short functional study shows that $at^3(1-t)+(1-t)^4$ when $a\in(0,1)$ reaches its maximal value for $t=0$. It leads to

$$3(1-X_n)\Delta R_n^{(1)} \le 3\gamma_{n+1}\rho_{n+1}^2(1-\sigma p_2).$$

For $\Delta R_n^{(2)}$, we have $\Delta R_n^{(2)} \leq \gamma_{n+1}^2 \rho_{n+1}^3 (1-\sigma p_2) \max_{0\leq t\leq 1} \left[\frac{1-\sigma p_1}{1-\sigma p_2} t^4 - (1-t)^4\right]$, which involves an increasing function of t. Thus, we have

$$\Delta R_n^{(2)} \leq \gamma_{n+1}^2 \rho_{n+1}^3 (1-\sigma p_1) \leq \gamma_{n+1}^2 \rho_{n+1}^3 (1-\sigma p_2).$$

Finally, if γ_1 and n_0 satisfy the assumptions of Proposition 4.1, then for every $n \geq n_0$, $\gamma_n \leq 2/3$ and it follows that $\Delta R_n^{(3)} \leq 0$. The result follows according to Equations (28). □

In order to bound $\sup_{n\geq n_0} \mathbb{E}(Z_n^{(3)})$, we have now to precisely study the polynomial function P_n and exhibit a mean reverting effect on its dynamics.

Proposition 4.2. *Let $\epsilon \in (0, \frac{1}{3})$, $\tilde{\rho}_1 \leq \frac{227}{232}$ and $\frac{1}{3\sqrt{2}(1-\sigma)\tilde{\rho}_1} \leq \gamma_1^2 \leq \frac{3}{2(1+\tilde{\rho}_1)}$. Then*

i) The polynomial P_n given by (24) is negative on $[0, 1 - \frac{2(1+\tilde{\rho}_1)}{\pi}]$.

ii) $Z_n^{(3)}$ satisfies

$$\begin{aligned}\sup_{n\geq n_0} \mathbb{E}Z_n^{(3)} \quad &\leq \quad \mathbb{E}Z_{n_0}^{(3)} + 12\gamma_1^4 \tilde{\rho}_1^2 \epsilon\pi + \frac{16}{3}\gamma_1^8 \tilde{\rho}_1^3 \epsilon^3 \pi^3 \\ &\quad + \frac{8\gamma_1^4 \epsilon(1+\tilde{\rho}_1)\left[1 + (1+\tilde{\rho}_1)[2 + 6\tilde{\rho}_1 + 12\gamma_1^2\epsilon]\right]}{\pi}.\end{aligned}$$

Remark 4.4. *The above result is given under some technical conditions which will lead to a sharp explicit bound. Nevertheless, the reader has to keep in mind that in view of the condition on σ, the "universal" bound on $(\mathbb{E}(Z_n^{(3)}))_{n\geq n_0}$ is only accessible when $\sigma < 1$, i.e. in the over-penalized case. When $\sigma = 1$, some bounds will be atteignable only if p_2 is not too large (see (17) for a similar statement when $r = 1$) and in order to alleviate the constraint on p_2, one will need to take a larger exponent than $r = 3$ (see Subsection 4.5 for details).*

Proof. We first provide the proof of *i*). The function P_n introduced in (24) is a third degree polynomial and for $n \geq n_0$:

$$\begin{aligned}P_n(0) \quad &= \quad \frac{\epsilon_n}{\gamma_{n+1}} - 3\tilde{\rho}_1 \kappa_\sigma(0) \\ &\leq \quad \frac{\gamma_n}{2\gamma_1^2 \gamma_{n+1}} - 3\tilde{\rho}_1(1-\sigma p_2) \\ &\leq \quad \frac{\sqrt{1+n_0^{-1}}}{2\gamma_1^2} - 3\tilde{\rho}_1(1-\sigma p_2)\end{aligned}$$

Since $p_2 < 1$, this last quantity is negative if one has

$$\rho_1 \gamma_1 \geq \frac{1}{3\sqrt{2}(1-\sigma)}. \tag{29}$$

In a same way, we can check that $P_n(1) = \gamma_{n+1} p_2 > 0$ and thus P_n has one root in the interval $(0,1)$. A careful inspection on the leading coefficient (denoted $a_n x^3$) of P_n in (24) shows that

$$a_n = \left[3(1+\sigma\tilde{\rho}_1) - \frac{3}{\gamma_{n+1}} - \gamma_{n+1}\right]\pi.$$

The leading coefficient a_n is negative as soon as $3(1+\sigma\tilde{\rho}_1) \leq \frac{3}{\gamma_{n+1}}$. Again, the choice of n_0 in (20) shows that this last condition is fulfilled as soon as

$$\frac{1}{\epsilon} \geq 2\gamma_1\pi(\gamma_1 + \sigma\rho_1). \tag{30}$$

But remark that we have assumed $\epsilon \in (0, 1/3]$ so that $\frac{1}{\epsilon} \geq 3$. As a consequence, (29) and (30) are satisfied as soon as $(\gamma_1, \tilde{\rho}_1)$ satisfies

$$\frac{1}{3\sqrt{2}(1-\sigma)\tilde{\rho}_1} \leq \gamma_1^2 \leq \frac{3}{2(1+\tilde{\rho}_1)}$$

Hence, if (29) and (30) hold, P_n possesses one root in $(-\infty, 0)$ and another one in $(1, +\infty)$. Consequently, P_n has a unique root in $(0,1)$. We consider now:

$$\xi_n = \frac{2(1+\tilde{\rho}_1)}{\pi}\gamma_{n+1} := \xi\gamma_{n+1}.$$

We compute that:

$$\begin{aligned} &P_n(1-\xi_n) \\ &= \frac{\xi_n^2}{\gamma_{n+1}}(\epsilon_n - 3\pi(1-\xi_n)) - 3\tilde{\rho}_1\xi_n\left[(1-\sigma p_2)\xi_n^2 - (1-\sigma p_1)(1-\xi_n)^2\right] \\ &\quad +3\left[(1-\xi_n)\xi_n^2 p_1 + (1-\xi_n)^2\xi_n p_2\right] + \gamma_{n+1}\left[(1-\xi_n)^3 p_2 - \xi_n^2(1-\xi_n)\right]. \end{aligned}$$

Hence, replacing ξ_n by $\xi\gamma_{n+1}$ and simplifying by γ_{n+1}, we see that $P_n(1-\xi_n)$ is negative when

$$\overbrace{\frac{\xi^2\epsilon_n}{(1-\xi_n)} + 3\tilde{\rho}_1(1-\sigma p_1)(1-\xi_n)\xi + 3p_1\gamma_{n+1}\xi^2 + 3p_2(1-\xi_n)\xi + p_2(1-\xi_n)^2}^{:=A_n(\xi)}$$
$$\leq \underbrace{3\pi\xi^2 + \frac{3\tilde{\rho}_1\gamma_{n+1}^2\xi^3(1-\sigma p_2)}{1-\xi_n} + \gamma_{n+1}^2\xi^2}_{:=B_n(\xi)}.$$

From (16), we know that $\epsilon_n \leq \frac{\gamma_{n+1}}{2\gamma_1^2}$, and $1-\xi_n \leq 1$ thus

$$A_n(\xi) \leq \xi^2\gamma_{n+1}\left(\frac{1}{2\gamma_1^2(1-\xi_n)} + 3p_1\right) + 3\xi\left(\tilde{\rho}_1 + 1\right) + 1$$

In the meantime, we will use the simple lower bound $B_n(\xi) \geq 3\pi\xi^2$. We can check that $1-\xi_n = 1 - \frac{2(1+\tilde{\rho}_1)\gamma_{n+1}}{\pi} \geq 1 - 4\epsilon(1+\tilde{\rho}_1)\gamma_1^2$ since $\gamma_{n_0} \leq 2\epsilon\gamma_1^2\pi$. Thus

$$\begin{aligned} & A_n\left(\frac{2(1+\tilde{\rho}_1)}{\pi}\right) \\ & \leq \frac{4(1+\tilde{\rho}_1)^2}{\pi^2}\gamma_{n+1}\left[3p_1 + \frac{1}{2\gamma_1^2\left[1-4\epsilon(1+\tilde{\rho}_1)\gamma_1^2\right]}\right] + \frac{6(1+\tilde{\rho}_1)^2}{\pi} + 1 \\ & \leq \frac{(1+\tilde{\rho}_1)^2}{\pi}\left[24\epsilon\gamma_1^2 p_1 + \frac{4\epsilon}{1-4\epsilon(1+\tilde{\rho}_1)\gamma_1^2} + 7\right] \end{aligned}$$

and

$$B_n\left(\frac{2(1+\tilde{\rho}_1)}{\pi}\right) \geq \frac{12(1+\tilde{\rho}_1)^2}{\pi}.$$

As a consequence, $P_n(1-\xi_n)$ is negative if one has

$$5 \geq 24\epsilon\gamma_1^2 + \frac{4\epsilon}{1-4\epsilon(1+\tilde{\rho}_1)\gamma_1^2}$$

From the constraint on γ_1, another computation shows that the above condition is fulfilled when $\epsilon^2\frac{128(1+\tilde{\rho}_1)}{3} - \epsilon[84+40(1+\tilde{\rho}_1)] + 45 \geq 0$. We then remark that all values of ϵ in $(0, \frac{1}{3}]$ can be conveniently used when $\tilde{\rho}_1 \leq \frac{227}{232}$.

To obtain *ii*), the main idea is to use the sharp estimation of the sign of P_n on $[0,1]$ and obtain an upper bound of $\mathbb{E}Z_n^{(3)}$: note that

$$\begin{aligned} & \sup_{0\leq t\leq 1} \gamma_{n+1}(1-t)P_n(t) \\ & = \gamma_{n+1}\sup_{1-\xi_n\leq t\leq 1}(1-t)P_n(t) \\ & = \gamma_{n+1}\sup_{1-\xi_n\leq t\leq 1}\Big\{(1-t)^3\left[\epsilon_n - 3\pi t\right] - 3\tilde{\rho}_1(1-t)^2\kappa_\sigma(t) \\ & \quad +3\left[t(1-t)^3p_1 + t^2(1-t)^2p_2\right] + \gamma_{n+1}\left[-t(1-t)^3p_1 + t^3(1-t)p_2\right]\Big\} \end{aligned}$$

Now, we have seen in the proof of *i*) that $t \in [1-\xi_n, 1] \Longrightarrow \epsilon_n \leq 3\pi t$. Hence, using $\kappa_\sigma(t) \geq -(1-\sigma p_1)t^2$, we have

$$\begin{aligned} & \sup_{0\leq t\leq 1} \gamma_{n+1}(1-t)P_n(t) \\ & \leq \gamma_{n+1}\left[3\tilde{\rho}_1(1-\sigma p_1)\xi_n^2 + 3p_1\xi_n^3 + p_2\xi_n^2 + \gamma_{n+1}\xi_n\right] \\ & \leq \frac{C_1(\tilde{\rho}_1,p_1,p_2,\sigma)}{\pi^2}\gamma_{n+1}^3 + \frac{C_2(\tilde{\rho}_1,p_1)}{\pi^3}\gamma_{n+1}^4 \end{aligned}$$

with $C_1(\tilde{\rho}_1,p_1,p_2,\sigma) = (1+\tilde{\rho}_1)\left(12\tilde{\rho}_1(1+\tilde{\rho}_1)(1-\sigma p_1) + 4p_2(1+\tilde{\rho}_1) + 2\pi\right)$ and $C_2(\tilde{\rho}_1,p_1) = 24p_1(1+\tilde{\rho}_1)^3$ shortenned in C_1 and C_2 in what follows. We apply

Lemma 4.1 to upper bound $\sup_{n\geq n_0}\mathbb{E}Z_n^{(3)}$:

$$\begin{aligned}
&\sup_{n\geq n_0}\mathbb{E}Z_n^{(3)}\\
&\quad\leq\ \mathbb{E}Z_{n_0}^{(3)}+\sup_{n\geq n_0}\mathbb{E}\sum_{k=n_0}^{n}\Delta Z_{n+1}^{(3)}\\
&\quad\leq\ \mathbb{E}Z_{n_0}^{(3)}+\sup_{n\geq n_0}\mathbb{E}\left[\sum_{k=n_0}^{n}\gamma_{k+1}(1-X_k)P_k(X_k)+\Delta R_k\right]\\
&\quad\leq\ \mathbb{E}Z_{n_0}^{(3)}+\frac{C_1}{\pi^2}\sum_{k=n_0}^{\infty}\gamma_{n+1}^3+\frac{C_2}{\pi^3}\sum_{k=n_0}^{\infty}\gamma_{n+1}^4+\sum_{k=n_0}^{\infty}\mathbb{E}\Delta R_k
\end{aligned}$$

Using a simple comparison argument with the integrals $\int_{n_0}^{\infty}t^{-\alpha}dt$, we get

$$\sum_{k=n_0}^{\infty}\gamma_{n+1}^3\leq 2\gamma_1^3 n_0^{-1/2}\leq 4\gamma_1^4\epsilon\pi\qquad\text{and}\qquad\sum_{k=n_0}^{\infty}\gamma_{n+1}^4\leq\gamma_1^4 n_0^{-1}\leq 4\gamma_1^6\epsilon^2\pi^2.$$

We then deduce that

$$\sup_{n\geq n_0}\mathbb{E}Z_n^{(3)}\leq\mathbb{E}Z_{n_0}^{(3)}+\frac{4\gamma_1^4\epsilon C_1}{\pi}+\frac{4\gamma_1^6\epsilon^2 C_2}{\pi}+\sum_{k=n_0}^{\infty}\Delta R_k.$$

The result follows using now (25). □

Explicit bound We can now conclude the proof of Theorem 3.2.

Proof of Theorem 3.2 (b). We consider the extreme over-penalized case obtained with $\sigma=0.$ and use a power increment until $r=3$. Recall that $n_0:=n_0(\epsilon,\pi,\gamma_1)$ is defined by (20). In particular, $\sqrt{n_0-1}\leq(2\epsilon\gamma_1\pi)^{-1}$ and for $i=1,2,3$, $\pi\mathbb{E}[Z_{n_0}^{(i)}]\leq(2\epsilon\gamma_1^2)^{-1}+(\gamma_1)^{-1}$. Gathering the results of Proposition 4.2 *ii*) and Corollary 4.1 and plugging them into (12), a series of computations yields:

$$\frac{\sup_{p_1\geq p_2}\bar{R}_n}{\sqrt{n}}\leq T_1(\gamma_1,\tilde{\rho}_1,\epsilon)+\frac{2\gamma_1}{(1-\epsilon)(1-\epsilon/2)}T_2(\gamma_1,\tilde{\rho}_1,\epsilon),$$

where

$$\begin{aligned}
T_1(\gamma_1,\tilde{\rho}_1,\epsilon)=\frac{1}{2\epsilon\gamma_1}+\left(\frac{1}{\epsilon\gamma_1}+2\right)\left(1+\frac{1}{1-\epsilon}+\frac{1}{(1-\epsilon)(1-\epsilon/2)}\right)\\
+2\rho_1\left(\frac{1}{1-\epsilon}+\frac{1}{(1-\epsilon)(1-\epsilon/2)}\right)+\frac{\gamma_1}{(1-\epsilon)(1-\epsilon/2)},
\end{aligned}$$

and

$$\begin{aligned}
&T_2(\gamma_1,\tilde{\rho}_1,\epsilon)\\
&\quad=\ \gamma_1^4\left[8\epsilon(1+\tilde{\rho}_1)\left(1+(1+\tilde{\rho}_1)(2+6\tilde{\rho}_1+12\gamma_1^2\epsilon)\right)+12\tilde{\rho}_1^2\epsilon+\frac{16}{3}\gamma_1^4\tilde{\rho}_1^3\epsilon^3\right].
\end{aligned}$$

Theorem 3.2(b) follows by minimizing $(\gamma_1, \tilde{\rho}_1, \epsilon) \longmapsto T_1(\gamma_1, \tilde{\rho}_1, \epsilon) + T_2(\gamma_1, \tilde{\rho}_1, \epsilon)$ up to the constraints

$$\epsilon \leq 1/3,\ \frac{1}{3\sqrt{2}\tilde{\rho}_1} \leq \gamma_1^2 \leq \frac{3}{2(1+\tilde{\rho}_1)},\ \tilde{\rho}_1 \leq 227/232.$$

The "best" upper bound was obtained by setting $\gamma_1 = 0.89, \tilde{\rho}_1 = 0.38, \epsilon = 1/9$, leading to the regret upper bound

$$\bar{R}_n \leq 44\sqrt{n}.$$

□

4.5. Regret of the penalized bandit of [20]

Sketch of proof of Theorem 3.1 and 3.2 (a). We prove these results together. We thus consider $\gamma_1 \in (0,1)$, $\rho_1 \in (0,1)$ and $\sigma \in [0,1]$. A variant of Proposition 4.1 about the increase of exponent is still valid. First, one can remark that if one sets $\varepsilon_r = r - 1/2$ (so that $\alpha_r = \pi/2$), then Lemma A.1 can be applied with $\tilde{n} \geq (\frac{\pi}{2}\gamma_1)^{-2}$. Thus, one sets $n_0(\lambda) := \lfloor \frac{\lambda^2}{\pi^2} \rfloor + 1$ with $\lambda \geq 2\gamma_1^{-1}$. After a simple adaptation of the proof of Proposition 4.1, one deduces that for every $r \geq 1$,

$$\sup_{n \geq n_0(\lambda)} \mathbb{E}Z_n^{(r)} \leq \mathbb{E}Z_{n_0}^{(r)} + \frac{2r}{\pi}\left[\tilde{\rho}_1 + h_r(\gamma_{n_0(\lambda)}) + \pi \sup_{n \geq n_0(\lambda)} Z_n^{(r+1)}\right].$$

By an iteration, it follows by using the fact that $\pi\mathbb{E}[Z_{n_0(\lambda)}^{(i)}] \leq \pi\gamma_{n_0(\lambda)}^{-1} \leq \gamma_1^{-1}(\lambda+1)$ that for every $r \geq 1$, some constants $C_r^1(\lambda)$ and $C_r^2(\lambda)$ exist (depending only on σ, γ_1 and ρ_1) such that,

$$\sup_{n \geq n_0(\lambda)} \pi\mathbb{E}[Y_n] \leq C_r^1(\lambda) + C_r^2(\lambda)\pi \sup_{n \geq n_0(\lambda)} \mathbb{E}Z_n^{(r+1)}. \tag{31}$$

It remains to upper bound $\sup_{n \geq n_0(\lambda)} \mathbb{E}Z_n^{(r)}$ for r large enough. Once again, a simple adaptation of the proof of Lemma 4.1 yields for $r \geq 3$:

$$\mathbb{E}[\Delta Z_{n+1}^{(r)} | \mathcal{F}_n] = \gamma_{n+1}(1 - X_n)^{r-1} P_n^{(r)}(X_n) + \Delta R_n^{(r)}.$$

with

$$\begin{aligned} &P_n^{(r)}(x) \\ &= \frac{(1-x)^2}{\gamma_{n+1}}(\varepsilon_n - r\pi x) - r\tilde{\rho}_1(1-x)\kappa_\sigma(x) + \binom{r}{r-2}\left(x(1-x)^2 p_1 + x^2(1-x)p_2\right) \\ &+ \gamma_{n+1}\binom{r}{r-3}\left(-x(1-x)^2 p_1 + x^3 p_2\right) \end{aligned} \tag{32}$$

and $\Delta R_n^{(r)} \leq C_r \gamma_{n+1}^3$ (where C_r does not depend on π). We want to prove that $P_n^{(r)}$ is negative on $[0, 1-\xi_n]$ with $\xi_n = \xi\gamma_{n+1} \in (0,1)$ where ξ is a constant

to calibrate. We follow the lines of the proof of Proposition 4.2 but we can use some rougher arguments since we do not search for explicit constants. First, $P_n^{(r)}(0) = \frac{\varepsilon_n}{\gamma_{n+1}} - r\tilde{\rho}_1\kappa_\sigma(0)$, so that

$$P_n^{(r)}(0) < 0 \quad \Longleftrightarrow \quad \gamma_1\rho_1 > \frac{1}{r\sqrt{2}(1-\sigma p_2)}.$$

On the one hand, for every $\sigma < 1$, one can find a r sufficiently large for which this condition holds. On the other hand, when $\sigma = 1$ (case which corresponds to Theorem 3.1), we need now to assume that there exists $\delta > 0$ such that $p_2 < 1-\delta$ (in this case, the condition is satisfied if $r > (\gamma_1\rho_1\sqrt{2}\delta)^{-1}$). For such a r, one remarks that the leading coefficient $a_n^{(r)}$ (related to x^3) is

$$a_n^{(r)} = \left(-\frac{r}{\gamma_{n+1}} + \binom{r}{r-2} + r\sigma\tilde{\rho}_1 - \gamma_{n+1}\right)\pi.$$

One deduces that $a_n^{(r)}$ is negative for every $n \geq n_1^\sigma$ where

$$n_1^\sigma := \left\lfloor \gamma_1^2 \left(\frac{r-1}{2} + \sigma\tilde{\rho}_1\right)^2 \right\rfloor.$$

Assume that $\lambda \geq \sqrt{n_1^\sigma}$ in order to get $n_0(\lambda) \geq n_1^\sigma$. Since $P_n^{(r)}(1) = \gamma_{n+1}p_2 > 0$ and $\deg(P_n^{(r)}) = 3$, it follows that $P_n^{(r)}$ has exactly one root in $(0,1)$ for every $n \geq n_0$ and that $P_n^{(r)}$ is negative on $[0, 1-\xi_n]$ as soon as $P_n^{(r)}(1-\xi_n) < 0$. Let n be such that $\xi\gamma_{n+1} \leq 1/2$. Then, some rough estimations yield that $P_n^{(r)}(1-\xi_n)$ is negative if

$$\frac{r\pi}{2}\xi^2 - c_r\xi - 1 > 0,$$

where c_r is a constant which does not depend on π. One then checks that there exists another constant η_r such that the previous property is fulfilled is $\xi \geq \eta_r/\pi$. Then, $P_n^{(r)}(1 - \frac{\eta_r}{\pi}\gamma_{n+1}) < 0$ is negative as soon as $\xi\gamma_{n+1} < 1/2$. This is true for every $n \geq n_0(\lambda)$ as soon as $\lambda \geq 2\gamma_1\eta_r$. We can conclude from what preceeds that there exist $r \geq 3$ and $\lambda > 0$ such that for every $n \geq n_0(\lambda)$, for every $(p_1, p_2) \in [0,1]^2$, such that $p_1 > p_2$ (resp. $p_1 > p_2$ and $p_2 < 1-\delta$) if $\sigma < 1$ (resp. if $\sigma = 1$)

$$\mathbb{E}[\Delta Z_{n+1}^{(r)}] \leq \gamma_{n+1} \sup_{t \in [1-\frac{\eta_r}{\pi}\gamma_{n+1}, 1]} (1-t)P_n^{(r)}(t) + C_r\gamma_{n+1}^3.$$

Using that $\gamma_{n+1} \leq \pi/\lambda$ if $n \geq n_0(\lambda)$, a constant C_λ exists such that on

$$\forall t \in [1 - \frac{\eta_r}{\pi}\gamma_{n+1}, 1] \qquad P_n^{(r)}(t) \leq C_\lambda\gamma_{n+1}/\pi.$$

Under the previous conditions, we deduce

$$\sup_\pi \left(\pi \sup_{n \geq n_0(\lambda)} \mathbb{E}Z_n^{(r)}\right) \leq \sup_\pi \left(\pi \sum_{n \geq n_0} C\gamma_{n+1}^3\left(\pi^{-2} + \pi^{-1}\right)\right) < +\infty.$$

The result follows by plugging this inequality into (31). □

5. Almost sure and weak limit of the over-penalized bandit

We provide here the proofs of Propositions 3.1 and 3.2. For the sake of simplicity, we restrict our study to $\sigma = 1$ (always over-penalization of the bandit), and the argument can be adapted for any values of $\sigma \in (0, 1]$.

5.1. A.s. convergence of the multi-armed bandit (Proposition 3.1)

Recall first that $X_n = (X_{n,1}, ..., X_{n,d})$, the multi-armed penalized bandit (4) permits to define for $i \in \{2, ..., d\}$,

$$X_{n+1,i} = X_{n,i} + \gamma_{n+1} h_i(X_n) + \gamma_{n+1}\rho_{n+1}\kappa_i(X_n) + \gamma_{n+1}\Delta M_{n+1,i},$$

where the main part of the drift h_i is defined as

$$h_i(x_1, ..., x_d) = (1 - x_i)x_i p_i - x_i \sum_{j \neq i} x_j p_j,$$

and the penalty drift is

$$\kappa_i(x_1, ..., x_d) = -x_i^2(1 - p_i) + \frac{1}{d-1} \sum_{j \neq i} x_j^2 (1 - p_j).$$

Hence, the martingale increment is simply obtained as

$$\begin{aligned} \Delta M_{n+1,i} &= ((1 - X_{n,i}) 1_{V_{n+1,i}, A_{n+1,i}} - X_{n,i} \sum_{j \neq i} 1_{V_{n+1,j}, A_{n+1,j}} - h_i(X_n)) \\ - & \quad \rho_{n+1}(X_{n,i} 1_{V_{n+1,i}, A^c_{n+1,i}} - \frac{1}{d-1} \sum_{j \neq i} X_{n,j} 1_{V_{n+1,j}, A^c_{n+1,j}} + \kappa_i(X_n)) \end{aligned}$$

Proof of Proposition 3.1. We start by (i) and identify the stationary point of the ODE method. The ODE $\dot{x} = h(x)$ possesses a finite number of equilibria that can be easily identified. Indeed we begin by solving the equation $h_1(x) = 0$. Since

$$h_1(x) = x_1 \sum_{i=2}^{d} x_j (p_1 - p_j) \geq 0,$$

we either have $x_1 = 1$ and $x_2 = ... = x_d = 0$ or $x_1 = 0$.
Then, the equation $h_2(x) = 0$ with $x_1 = 0$ may be reduced to

$$x_2 \sum_{i=3}^{d} x_j (p_2 - p_j) \geq 0.$$

The same argument leads to $x_2 = 1$ or $x_2 = 0$ and a straightforward recursion shows that the equilibria of the ODE are $(\delta^i)_{1 \leq i \leq d}$, with $(\delta^i)_{1 \leq i \leq d}$ defined as

$$\delta^i_i = 1 \quad \text{and} \quad \delta^i_j = 0 \quad \forall j \neq i.$$

Let us emphasize that to discriminate among these equilibria, it is not possible to use the second derivative criterium that relies on $\left(\frac{\partial h_i}{\partial x_j}\right)_{i,j}$ to decide their stability. Instead, it is possible to check that δ^1 fulfills the Lyapunov certificate with the function $V(x) = (x_2^2 + \ldots + x_d^2)$. If we denote $h = (h_1, \ldots, h_d)$, one has

$$\langle \nabla V(x), h(x) \rangle = \sum_{j=2}^{d} x_j^2 \sum_{k \neq j} x_k (p_j - p_k).$$

Considering x in a closed neighborhood of δ^1 defined as $x_j \leq \epsilon/d, \forall j \geq 2$ (implying that $x_1 > 1 - \epsilon$), we see that

$$\begin{aligned} \langle \nabla V(x), h(x) \rangle &= x_1 \sum_{j=2}^{d} x_j^2 (p_j - p_1) + \sum_{k=2}^{d} x_k^2 \sum_{j \neq k, j>1} x_j (p_k - p_j) \\ &\leq -(1-\epsilon)(p_1 - p_2) \sum_{j=2}^{d} x_j^2 + \epsilon \sum_{j=2}^{d} x_j^2, \end{aligned}$$

and the term above is negative as soon as ϵ is chosen such that

$$\epsilon \leq \frac{1}{p_1 - p_2 + 1}.$$

Oppositely, the other equilibria $(\delta^j), j \neq 1$ are unstable: this can be easily deduced from the unstability of the two-armed bandit by testing the first arm versus the arm j.

The martingale increment $\Delta M_{n+1,i}$ being uniformly bounded, we can apply the Kushner-Clark theorem (see *e.g.* [17]) and conclude that $(X_{n,i})_{n \geq 0}$ either converges to 1 or 0 a.s. As a consequence, it is also true that $(X_n)_{n \geq 0}$ converges a.s. We now make this limit explicit and show that $(X_n)_{n \geq 0}$ converges toward $(1, ..., 0)$ a.s. We start by noticing that $h_1(x) = x_1 \sum_{j \geq 2} x_j (p - 1 - p_j) \geq 0$, which implies that

$$X_{n,1} \geq X_{0,1} + \sum_{j=1}^{n-1} \gamma_j \rho_j \kappa_{j-1,1}(X_{j-1}) + \sum_{j=1}^{n-1} \gamma_j \Delta M_j. \tag{33}$$

The martingale increment ΔM_j is bounded and a large enough C exists such that $\Delta M_j \leq \sqrt{C}$. This implies that

$$\left\| \sum_{j=1}^{n-1} \gamma_j \Delta M_j \right\|_{L^2}^2 \leq C \sum_{j=1}^{n-1} \gamma_j^2 \leq C \sup_{j \in \mathbb{N}} \left(\frac{\gamma_j}{\rho_j} \right) \sum_{j=1}^{n-1} \gamma_j \rho_j.$$

Since $\sum \rho_j \gamma_j = +\infty$, we can deduce that

$$\lim_{n\longrightarrow+\infty} \frac{\mathbb{E}\left[\sum_{j=1}^{n} \gamma_j \Delta M_j\right]^2}{\sum_{j=1}^{n-1} \gamma_j \rho_j} = 0 \qquad \text{so that} \qquad \limsup_{n\to\infty} \frac{\sum_{j=1}^{n-1} \gamma_j \Delta M_j}{\sum_{j=1}^{n-1} \gamma_j \rho_j} \geq 0.$$

Consider now an event $\omega \in \{X_{\infty,1} = 0\}$, we have

$$\lim_{n\longrightarrow+\infty} \kappa_1(X_n(\omega)) = \frac{1}{d-1} \sum_{k\geq 2} (1-p_k) X_{\infty,k}(\omega)^2,$$

and according to the Toeplitz Lemma we deduce that

$$\lim_{n\to\infty} \frac{\sum_{j=1}^{n-1} \gamma_j \rho_j \kappa_1(X_{j-1}(\omega))}{\sum_{j=1}^{n-1} \gamma_j \rho_j} = \frac{1}{d-1} \sum_{k\geq 2} (1-p_k) X_{\infty,k}(\omega)^2 > 0.$$

Putting together this last remark with Equation (33) leads to the conclusion

$$\limsup_{n\to\infty} \frac{X_{n,1}(\omega)}{\sum_{j=1}^{n-1} \gamma_j \rho_j} > 0.$$

We obtain a contradiction with the boundedness of $(X_n)_{n\geq 1}$ and conclude that $\mathbb{P}(X_{\infty,1} = 0) = 0$. For (ii), we refer to [20] since the arguments here are similar. □

5.2. *Weak convergence of the normalized bandit (Proposition 3.2)*

The proof of the weak convergence follows the lines of [20]. The idea is to prove the tightness of the pseudo-trajectories associated to the normalized sequence and then to show that any weak limit of this sequence is a solution of the martingale problem $(\mathcal{L}, \mathcal{C}^1_K(\mathbb{R}_+, (\mathbb{R}_+)^{d-1})$ where $\mathcal{L}$ is the infinitesimal generator defined in Proposition 3.2. Then, proving that uniqueness holds for the solutions of the martingale problem and for the invariant distribution, the convergence follows. Here, we choose to only detail the key step of the characterization of the limit. The rest of the proof can be obtained by a simple generalization of that of [20].

Proposition 5.1. *Let f be a continuously differentiable function with compact support in $\mathbb{R}_+^{d-1}$. We have*

$$\mathbb{E}(f(Y_{n+1,2}, ..., Y_{n+1,d}) - f(Y_{n,2}, ..., Y_{n,d}) | \mathcal{F}_n) = \gamma_{n+1} \mathcal{L}_d f(Y_{n,2}, ..., Y_{n,d}) + o_P(1),$$

where $\mathcal{L}_d$ is the PDMP generator defined in (8) and $\mathcal{F}_n = \sigma(Y_k, k \leq n)$.

Proof. Since the proof does not depend on σ, we assume that $\sigma = 1$ for the sake of clarity. We first give an alternative expression for the variables $Y_{n,i}$ for $i \geq 2$.

$$Y_{n+1,i} = Y_{n,i} + \gamma_{n+1}\left(\frac{1-p_1}{d-1} - (p_1 - p_i)Y_{n,i}\right) + \gamma_{n+1}C_{n,i} - g\Delta M_{n+1,i},$$

where $C_{n,i} = (\kappa_i(X_n) - \frac{1-p_1}{d-1}) + Y_{n,i}(p_1 - p_i + (\epsilon_n + \frac{\rho_n}{\rho_{n+1}}(p_i - \sum_{j\neq i} X_{n,j}p_j))) = o_P(1)$ since $(\epsilon_n)_{n\geq 0}$ converges 0 and $(X_{n,i})_{n\geq 0}$ converges to 0 in probability for $i \geq 2$. We rewrite this as follows

$$Y_{n+1,i} = Y_{n,i} + \gamma_{n+1}\left(\frac{1-p_1}{d-1} - (p_1 - p_i)Y_{n,i} + C_{n,i}\right) + G_{n,i} + g\Delta\tilde{M}_{n+1,i},$$

where $G_{n,i} = g(1-X_{n,i})(1_{V_{n+1,i},A_{n+1,i}} - X_{n,i}p_i)$ and $\Delta\tilde{M}_{n+1,i} = \Delta M_{n+1,i} - G_{n,i}$. We consider a function $f \in \mathcal{C}^1(\mathbb{R}_+^{d-1})$ with a compact support.

$$f(Y_{n+1}) - f(Y_n) = \sum_{i=2}^{d} f(Y_{n+1,2}, ...Y_{n+1,i}, ...Y_{n+1,d}) - f(Y_{n,2}, ...Y_{n,i}, ...Y_{n,d}).$$

We will use the following notation $F_i(Y_k) = f(Y_{n,2}, ...Y_{k,i}, ...Y_{n,d})$. It means that the first $i-1$ variables are $(Y_{n,2}, Y_{n,3}, ...)$ and the $d-i$ last ones are $(Y_{n+1,i+1}, Y_{n+1,i+2}, ..., Y_{n+1,d})$. We have

$$F_i(Y_{n+1,i}) - F_i(Y_{n,i}) = F_i(Y_{n+1,i}) - F_i(\overline{Y}_{n,i}) + F_i(\overline{Y}_{n,i}) - F_i(Y_{n,i}),$$

where

$$\widetilde{Y}_{n,i} = Y_{n,i} + \gamma_{n+1}\left(\frac{1-p_1}{d-1} - (p_1 - p_i)Y_{n,i} + C_{n,i}\right),$$

and

$$\overline{Y}_{n,i} = \widetilde{Y}_{n,i} + G_{n,i}.$$

We begin by writing

$$F_i(Y_{n+1,i}) - F_i(\overline{Y}_{n,i}) = \partial_i F_i(\widetilde{Y}_{n,i})\Delta\widetilde{M}_{n+1,i} + \gamma_{n+1}V_{n+1,i},$$

where the first order Taylor approximation formula yields

$$\exists\, \theta \in [0,1]: \qquad V_{n+1,i} = \left[F_i(\widetilde{Y}_{n,i} + \theta\Delta\widetilde{M}_{n+1,i}) - F_i(\widetilde{Y}_{n,i})\right]\Delta\widetilde{M}_{n+1,i}.$$

As a consequence, $V_{n+1,i} = o_P(1)$ and we are now going to prove that

$$\mathbb{P} - \lim_{n\to\infty} \mathbb{E}\left(\frac{F_i(Y_{n+1}) - F_i(Y_n) - \gamma_{n+1}\mathcal{A}_i F_i(Y_n)}{\gamma_{n+1}} | \mathcal{F}_n\right) = 0,$$

where

$$\begin{aligned}\mathcal{A}_i f(Y_2, ..., Y_d) &= \frac{p_i Y_i}{g}(f(Y_2, ..., Y_i + g, ..., Y_d) - f(Y_2, ..., Y_i, ...Y_d)) \\ &+ \left(\frac{1-p_1}{d-1} - p_1 Y_i\right)\partial_i f(Y_2, ..., Y_d).\end{aligned}$$

We compute

$$\begin{aligned}\mathbb{E}(F_i(\overline{Y}_{n,i}))|\mathcal{F}_{n,i}) &= p_i X_{n,i} F_i(\widetilde{Y}_{n,i} + g(1-X_{n,i})(1-p_i X_{n,i})) \\ &+ (1-gp_i X_{n,i}) F_i(\widetilde{Y}_{n,i} - gp_i X_{n,i}(1-X_{n,i}))).\end{aligned}$$

Let us decompose the r.h.s. of the above equation in two parts, we denote

$$F_{n,i} = p_i X_{n,i}(F_i(\widetilde{Y}_{n,i} + g(1-X_{n,i})(1-p_i X_{n,i})) - F_i(Y_{n,i})), \tag{34}$$

and

$$G_{n,i} = (1-gp_i X_{n,i})(F_i(\tilde{Y}_{n,i} - gp_i X_{n,i}(1-X_{n,i}))) - F_i(Y_{n,i})). \tag{35}$$

Note that (34) is the jump part of the PDMP and (35) the deterministic one. If $i \geq 2$, $(X_{n,i})_{n\geq 1}$ converges to 0 in probability and $\rho_n {\gamma_{n+1}}^{-1} = g + o(\rho_n)$, thus:

$$\begin{aligned}{\gamma_{n+1}}^{-1} F_{n,i} &= {\gamma_{n+1}}^{-1} \rho_n p_i Y_{n,i}(F_i(Y_{n,i} + g + o_P(1)) - F_i(Y_{n,i})) \\ &= \frac{p_i Y_{n,i}}{g}(1+o(\rho_n))\left[F_i(Y_{n,i} + g + o_P(1)) - F_i(Y_{n,i})\right].\end{aligned}$$

As a consequence, the asymptotic behaviour of (34) is given by

$$\mathbb{P} - \lim_{n\to\infty} \left(\frac{F_{n,i}}{\gamma_{n+1}} - p_i Y_{n,i} \frac{F_i(Y_{n,i}+g) - F_i(Y_{n,i})}{g} \right) = 0.$$

We now study (35) and compute

$$\begin{aligned}\widetilde{Y}_{n,i} - gX_{n,i}(1-p_i X_{n,i}) &= Y_{n,i} + \gamma_{n+1}\left(\frac{1-p_1}{d-1} - p_1 Y_{n,i}\right) \\ &+ \gamma_{n+1} p_i Y_{n,i} - gp_i X_{n,i}(1-X_{n,i}) + \gamma_{n+1} C_{n,i} \\ &= Y_{n,i} + \gamma_{n+1}\left(\frac{1-p_1}{d-1} - p_1 Y_{n,i}\right) \\ &+ \underbrace{\gamma_{n+1} p_i Y_{n,i} - g\rho_n p_i Y_{n,i}(1-X_{n,i}) + \gamma_{n+1} C_{n,i}}_{:=\gamma_{n+1}\widetilde{C}_{n,i}},\end{aligned}$$

where we used $g\rho_n = \gamma_n$. Since $\widetilde{C}_{n,i}$ converges to 0 in probability, we obtain:

$$\begin{aligned}&\gamma_{n+1}^{-1} G_{n,i} \\ &= \gamma_{n+1}^{-1}(1+o(\rho_n))\left(F_i(Y_{n,i} + \gamma_{n+1}\left[\frac{1-p_1}{d-1} - p_1 Y_{n,i}\right] + \gamma_{n+1}\tilde{C}_{n,i}) - F_i(Y_{n,i})\right) \\ &= \left[\frac{1-p_1}{d-1} - p_1 Y_{n,i}\right] \frac{\left(F_i(Y_{n,i} + \gamma_{n+1}\left[\frac{1-p_1}{d-1} - p_1 Y_{n,i}\right] + \gamma_{n+1}\tilde{C}_{n,i}) - F_i(Y_{n,i})\right)}{\gamma_{n+1}\left[\frac{1-p_1}{d-1} - p_1 Y_{n,i}\right]} \\ &+ o_P(1).\end{aligned}$$

We finally obtain that the limiting behaviour of (35):

$$\mathbb{P} - \lim_{n\to\infty} \left(\frac{G_{n,i}}{\gamma_{n+1}} - \left(\frac{1-p_1}{d-1} - p_1 Y_i\right) \partial_i F_i(Y_{n,i}) \right) = 0.$$

This ends the proof of the proposition. □

6. Ergodicity of the PDMP

From now on, the variable $(X_t)_{t\geq 0}$ will refer to a trajectory of the PDMP associated to the normalized (over)-penalized bandit and bear no relation with the multi-armed Bandit sequence $(X_n)_{n\geq 1}$.

6.1. Wasserstein results

We begin the study of the ergodicity of the PDMP whose infinitesimal generator is (9) with some computations of the moments of the process.

Lemma 6.1. *Let $(X_t)_{t\geq 0}$ be a Markov process, whose generator $\mathcal{L}$ is defined by (9). If $\pi := b - cg > 0$, then $\sup \mathbb{E}[(X_t^x)^p] \leq C(1+|x|^p)$. In particular, the invariant distribution π has moments of any order and*

$$\forall t \geq 0 \qquad \mathbb{E}(X_t) = \frac{a}{\pi} + \left(\mathbb{E}(X_0) - \frac{a}{\pi}\right) e^{-t\pi}$$

Proof. Let us define $f_p(x) = x^p$, we have

$$\begin{aligned} \mathcal{L}f_p(x) &= p(a-bx)x^{p-1} + cx((x+g)^p - x^p) \\ &= -p\pi f_p(x) + paf_{p-1}(x) + c\sum_{k=0}^{p-2} C_p^k g^{p-k} f_{k+1}(x), \end{aligned} \tag{36}$$

where we adopt the convention $\Sigma_\emptyset = 0$. If we define now $\alpha_p(t) = \mathbb{E}(X_t^p)$, the previous relation shows that α_p satisfies the ODE for any integer $p \geq 1$ defined by

$$\alpha_p(t)' + p\pi\alpha_p(t) = pa\alpha_{p-1}(t) + c\sum_{k=0}^{p-2} C_p^k g^{p-k}\alpha_{k+1}(t).$$

For example, with $p = 1$ we have $\alpha_1'(t) = -\pi\alpha_1(t) + a$, which implies that

$$\alpha_1(t) = \frac{a}{\pi} + \left(\mathbb{E}(X_0) - \frac{a}{\pi}\right) e^{-t\pi}.$$

The control of the moments of order $p > 1$ then follows from a recursion. □

6.1.1. Rescaled two-armed bandit & Theorem 3.3

In what follows, we will exploit Equation (36) to obtain a suitable upper bound of the Wasserstein distance $\mathcal{W}_p$ between the law of X_t and the invariant measure μ_∞ of the PDMP. For this purpose, we remark that the generator (9) possesses the stochastic monotonicity property: *i.e.* there exists a coupling (X, Y) starting from (x, y) (with $x > y$) such that $X_t \geq Y_t$ for any $t \geq 0$. The increase of the jump rate (with respect to the position) and the positivity of the jumps are of first importance for this property. Such a coupling could be built as follows: we only allow simultaneous jumps of both components or a single jump of the

highest one (see ([6]) for a similar procedure). The generator of this coupling (X, Y) starting from (x, y) with $x > y$ is given by:

$$\begin{aligned}\mathcal{L}_{\mathcal{W}} f(x,y) &= (a-bx)\partial_x f(x,y) + (a-by)\partial_y f(x,y) \\ &+cy\left(f(x+g,y+g)-f(x,y)\right) + c(x-y)\left(f(x+g,y)-f(x,y)\right) \quad (37)\end{aligned}$$

with a symmetric expression when $y > x$. We now prove the main result.

Proof of theorem 3.3. Let μ_0 be a probability on $\mathbb{R}_+^*$ and denote by μ_∞ the invariant distribution of the PDMP. Set

$$\mathcal{C}_t = \{\nu \in \mathcal{P}(\mathbb{R}^2), \nu(dx \times \mathbb{R}_+) = \mu_t(dx), \nu(\mathbb{R}_+ \times dy) = \mu_\infty(dy)\}.$$

For any $\nu \in \mathcal{C}$, let $(X_t, Y_t)_{t\geq 0}$ denote the Markov process driven by (37) starting from ν. From the definition of $\mathcal{W}_p$ and the stationary of (Y_t), we have for any t:

$$\mathcal{W}_p(\mu_t, \mu_\infty) \leq \inf\{\nu \in \mathcal{C}_0, \left(\int_{\mathbb{R}_+^2} \mathbb{E}[|X_t^x - Y_t^y|^p]\nu(dx,dy)\right)^{\frac{1}{p}}.$$

At the price of a potential exchange of the coordinates, we can now work with some deterministic starting points x and y such that $x > y > 0$. Owing to the monotonicity of $\mathcal{L}_{\mathcal{W}}$, we thus have for any $p \geq 1$

$$\mathbb{E}(|X_t^x - Y_t^y|)^p = \mathbb{E}(X_t^x - Y_t^y)^p.$$

Assume now that $p \in \mathbb{N}^*$, we remark that $\mathcal{L}_{\mathcal{W}}$ acts on $(x, y) \mapsto (x-y)^p$ as

$$\mathcal{L}_{\mathcal{W}}(x-y)^p = -p\pi(x-y)^p + pa(x-y)^{p-1} + c\sum_{k=0}^{p-2} C_p^k g^{p-k}(x-y)^{k+1}.$$

Setting $\beta_p(t) = \mathbb{E}\,|X_t^x - Y_t^y|^p$, it is then immediate to check that

$$\dot{\beta}_p(t) + \pi p \beta_p(t) = \left(pa\beta_{p-1}(t) + c\sum_{k=0}^{p-2} C_p^k g^{p-k}\beta_{k+1}(t)\right). \quad (38)$$

When $p = 1$, (38) implies that $\beta_1(t) = \beta_1(0)e^{-\pi t} \Rightarrow \mathbb{E}[X_t^x - Y_t^y] = (x-y)e^{-\pi t}$, so that

$$\mathcal{W}_1(\mu_t, \mu_\infty) \leq \mathcal{W}_1(\mu_0, \mu_\infty)e^{-t\pi}.$$

For the lower-bound, one uses that

$$\mathcal{W}_1(\mu_t, \mu_\infty) \geq \inf\left\{\nu_t \in \mathcal{C}_t, \left|\int (x-y)\nu_t(dx,dy)\right|\right\} = \left|\mathbb{E}[X_t^{\mu_0}] - \mathbb{E}[Y_t^{\mu_\infty}]\right|,$$

which implies that

$$\mathcal{W}_1(\mu_t, \mu_\infty) \geq \left|\int \mathbb{E}[X_t^x - Y_t^y]\mu_0(dx)\mu_\infty(dy)\right| = \left|\int (x-y)\mu_0(dx)\mu_\infty(dy)\right| e^{-\pi t}.$$

The lower-bound follows.

Now, let us consider the case $p > 1$ (with $p \in \mathbb{N}$). For $p = 2$, we have

$$\left(\beta_2(t)e^{2\pi t}\right)' e^{-2\pi t} = (2a + cg^2)\beta_1(0)e^{-\pi t},$$

and an integration leads to $\beta_2(t)e^{2\pi t} - \beta_2(0) = \frac{2a+cg^2}{\pi}\beta_1(0)[e^{\pi t} - 1]$. As a consequence,

$$\beta_2(t) \leq e^{-2\pi t}\beta_2(0) + \frac{2a + cg^2}{\pi}\beta_1(0)e^{-\pi t}.$$

Using the inequalities $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ and $\beta_2 \geq \mathcal{W}_2^2$, we thus deduce that

$$\mathcal{W}_2(\mu_t, \mu_\infty) \leq \mathcal{W}_2(\mu_0, \mu_\infty)e^{-\pi t} + \sqrt{\frac{2a + cg^2}{\pi}}\sqrt{\mathcal{W}_1(\mu_0, \mu_\infty)}e^{-\frac{\pi t}{2}}.$$

The result follows when $p = 2$ by setting

$$\gamma_2 := \mathcal{W}_2(\mu_0, \mu_\infty) + \sqrt{\frac{2a + cg^2}{\pi}}\sqrt{\mathcal{W}_1(\mu_0, \mu_\infty)}.$$

A recursive argument based on (38) shows that a constant γ_p exists that only depends on μ_0 and μ_∞ such that

$$W_p(\mu_t, \mu_\infty) \leq \gamma_p e^{-\frac{\pi}{p}t}.$$

□

6.2. Proof of Total variation results

As mentioned before, the idea is to wait that the paths get close (with a probability controlled by the Wasserstein bound) and then to try to stick them (with high probability). Since the jump size is deterministic, sticking the paths implies a non trivial coupling of the jump times which is described in the lemma below. We begin by establishing the next useful lemma.

Lemma 6.2. *Let $\varepsilon > 0$ and $t \geq \frac{1}{b}\ln(1+\varepsilon)$. There exists a coupling $(X_t, Y_t)_{t\geq 0}$ of paths driven by (9) such that on $A_{x_0,\varepsilon}$*

$$\mathbb{P}(X_t = Y_t, t \geq s) \geq \left(1 - \frac{c}{b}x_0\varepsilon - e^{-\frac{a}{b}cs} - \frac{c\varepsilon}{b}\right)\max(0, 1 - \frac{c}{b}\varepsilon(x_0 + g)),$$

where $A_{x_0,\varepsilon} = \left\{(x,y)|\,\frac{a}{b} < x \leq x_0, 0 < x - y \leq \varepsilon\right\}$.

Proof Let $\varepsilon > 0$ and $(x, y) \in A_{x_0,\varepsilon}$ (in particular, $x > y$). Denote by T_1^x and T_1^y the first jumps of (X_t^x) and (X_t^y) respectively and by T_2^x the second jump of (X_t^x). One remarks that

$$\mathbb{P}(X_t = Y_t, t \geq s) \geq \mathbb{P}(X^x_{T_1^y} = X^y_{T_1^y}, T_1^y \leq s).$$

We aim to build a coupling that leads to a sharp lower-bound of the r.h.s. For this purpose, remark that if $T_1^x < T_1^y < T_2^x$, the triple (T_1^x, T_1^y, T_2^x) satisfies

$$X^y_{T_1^y} = X^x_{T_1^y} \iff \frac{a}{b} + \left(y - \frac{a}{b}\right) e^{-bT_1^y} + g = \frac{a}{b} + \left(X^x_{T_1^x} - \frac{a}{b}\right) e^{-b(T_1^y - T_1^x)}.$$

Using that $X^x_{T_1^x} = \frac{a}{b} + (x - \frac{a}{b})e^{-T_1^x} + g$ and defining $\psi(t) = \frac{1}{b}\ln\left(e^{bt} + \frac{x-y}{g}\right)$, we can verify that $X^y_{T_1^y} = X^x_{T_1^y} \leq s$ and $T_1^x < T_1^y < T_2^x$ as soon as

$$T_1^y = \psi(T_1^x) \leq s \quad \text{and} \quad T_2^x \geq \psi(T_1^x),$$

since $\psi(t) \geq t$. We are naturally encouraged to consider $S_1^{x,s} = \psi(T_1^x)1_{\{\psi(T_1^x)\leq s\}}$ and it is well known that the law of (T_1^x, T_1^y) can be described through the maximal coupling :

$$T_1^y = \Theta U + (1-\Theta)V_y, \quad \psi(T_1^x) = \Theta U + (1-\Theta)V_x,$$

where V_x, V_y, Θ and U are independent, $U \sim \frac{\mathbb{P}_{T_1^y} \wedge \mathbb{P}_{\psi(T_1^x)}}{\|\mathbb{P}_{S_1^{x,s}} \wedge \mathbb{P}_{T_1^y}\|_{TV}}$ and $\Theta \sim \mathcal{B}(p)$ where $p = \|\mathbb{P}_{S_1^{x,s}} \wedge \mathbb{P}_{T_1^y}\|_{TV}$. With this coupling, if $q(t,z) = \mathbb{P}(T_1^z \geq \psi(t) - t)$, the Strong Markov property yields

$$\mathbb{P}(T_2^x - T_1^x | (T_1^x, T_1^y)) = \mathbb{P}(T_2^x \geq \psi(T_1^x) | T_1^x) = q(T_1^x, X^x_{T_1}).$$

Since, $z \mapsto q(t,z)$ is increasing and $x > a/b$ (from the assumption on $A_{x_0,\epsilon}$), we deduce that $X^x_{T_1} \leq x + g$ and it follows that

$$\mathbb{P}(T_2^x > T_1^y | (T_1^x, T_1^y)) \geq q(t, x+g) \geq q(0, x+g).$$

using that $t \mapsto \psi(t) - t$ is a non-decreasing function. As a consequence, one obtains that with this coupling,

$$\mathbb{P}(X^x_{T_1^y} = X^y_{T_1^y}, T_1^y \leq s) \geq q(0, x+g)\mathbb{P}(\Theta = 1) = q(0, x+g)\|\mathbb{P}_{S_1^{x,s}} \wedge \mathbb{P}_{T_1^y}\|_{TV}. \tag{39}$$

It remains to find a lower bound of the total variation distance involved in the r.h.s. of the above inequality . Recall that

$$\|\mathbb{P}_{S_1^{x,s}} \wedge \mathbb{P}_{T_1^y}\|_{TV} = \int_0^{+\infty} f_y(t) \wedge g_{x,s}(t)dt,$$

where f_y and $g_{x,s}$ denote respectively the densities of T_1^y and $S_1^{x,s}$. We have

$$\forall t > 0, \quad f_y(t) = c\phi(y,t)e^{-\int_0^t c\phi(y,u)du} \quad \text{with} \quad \phi(y,t) = \frac{a}{b} + (y - \frac{a}{b})e^{-bt},$$

and a change of variable yields

$$\forall t > 0, \quad g_x(t) = f_x(\psi^{-1}(t))(\psi^{-1})'(t)1_{\{\psi(0)\leq t\leq s\}}. \tag{40}$$

On the one hand, since $(x,y) \in A_{x_0,\epsilon}$, we can check that

$$\forall t \geq 0, \quad \phi(x,t) - \varepsilon e^{-bt} \leq \phi(y,t) \leq \phi(x,t),$$

and then we conclude that

$$\forall t > 0, \quad f_y(t) \geq f_x(t) - \varepsilon e^{-bt}.$$

One the other hand, remark that

$$\forall t > \psi(0), \quad \psi^{-1}(t) = \frac{1}{b} \ln \left(e^{bt} - \frac{x-y}{g} \right) \leq t \quad \text{and} \quad (\psi^{-1})'(t) = \frac{e^{bt}}{e^{bt} - \frac{x-y}{g}} \geq 1,$$

and we deduce from (40) that $\forall t \in [\psi(0), s]$:

$$g_x(t) \geq c\phi(x, \psi^{-1}(t)) e^{-\int_0^t c\phi(x,s)ds} \geq c\phi(x,t) e^{-\int_0^t c\phi(x,s)ds} = f_x(t).$$

Note that we used that $t \mapsto \phi(x,t)$ is decreasing since $x > a/b$. Thus,

$$\left(\mathbb{P}_{T_1^y} \wedge \mathbb{P}_{S_1^x} \right)(dt) \geq h(t)dt \quad \text{with} \quad h(t) = (f_x(t) - \varepsilon e^{-bt}) 1_{\psi(0) \leq t \leq s} dt.$$

As a consequence,

$$\| \mathbb{P}_{S_1^{x,s}} \wedge \mathbb{P}_{T_1^y} \|_{TV} \geq e^{-\int_0^{\psi(0)} c\phi(x,u)du} - e^{-\int_0^s c\phi(x,u)du} - \frac{\varepsilon}{b}.$$

Checking that $\psi(0) \leq \varepsilon/b$ and that $\forall t \geq 0$, $a/b \leq \phi(x,t) \leq x \leq x_0$, we deduce that

$$\| \mathbb{P}_{S_1^{x,s}} \wedge \mathbb{P}_{T_1^y} \|_{TV} \geq e^{-\frac{cx_0\varepsilon}{b}} - e^{-\frac{a}{b}cs} - \frac{\varepsilon}{b} \geq 1 - \frac{cx_0\varepsilon}{b} - e^{-\frac{a}{b}cs} - \frac{\varepsilon}{b},$$

where in the second line, we used $e^{-u} \geq 1-u$ for $u \geq 0$. To conclude the proof, it remains to plug this inequality into (39) and to remark that

$$q(0, x+g) \geq q(0, x_0+g) = e^{-\int_0^{\psi(0)} c\phi(x_0+g,s)ds} \geq 1 - c\psi(0)(x_0+g) \geq 1 - \frac{c}{b}\varepsilon(x_0+g).$$

□

We provide now the proof of the ergodicity w.r.t. the total variation distance.

Proof of Theorem 3.4. For any starting distribution μ_0,

$$\|\mu_0 P_t - \mu_\infty\|_{TV} \leq \int \|\delta_x P_t - \delta_y P_t\|_{TV} \mu_0(dy) \mu_\infty(dx). \tag{41}$$

The idea is to use the Wasserstein coupling during a time t_1 and then to try to stick the paths on the interval $[t_1, t]$ using Lemma 6.2. Consider $A_{x_0,\varepsilon}$ defined in lemma 6.2 and the alternative set $A^*_{x_0,\varepsilon} = \{(x,y), a/b < y < x_0, 0 < y - x \leq \varepsilon\}$. Set $B_{x_0,\varepsilon} = A_{x_0,\varepsilon} \cup A^*_{x_0,\varepsilon}$, we have

$$1 - \|\delta_x P_t - \delta_y P_t\|_{TV} \geq \mathbb{P}(X_t^x = Y_t^y | (X_{t_1}^x, Y_{t_1}^y) \in B_{x_0,\varepsilon}) \mathbb{P}((X_{t_1}^x, Y_{t_1}^y) \in B_{x_0,\varepsilon}). \tag{42}$$

Since the Wasserstein coupling preserves the order and that $x > a/b$ $\mu_\infty(dx)$-$a.s.$, one remarks that $\mu_\infty(dx)$-$a.s.$,

$$(X^x_{t_1}, Y^y_{t_1}) \in B_{x_0,\varepsilon} \Longleftrightarrow \begin{cases} X^x_{t_1} - X^y_{t_1} \leq \varepsilon \text{ and } X^x_{t_1} \leq x_0 & \text{if } x \geq y \\ X^y_{t_1} - X^x_{t_1} \leq \varepsilon \text{ and } X^y_{t_1} \leq x_0 & \text{if } x < y. \end{cases}$$

It follows that for every $p > 0$, $\mu_\infty(dx)$ almost surely:

$$\begin{aligned} \mathbb{P}((X^x_{t_1}, Y^y_{t_1}) \in B^c_{x_0,\varepsilon}) &\leq \mathbb{P}(|X^x_{t_1} - X^y_{t_1}| > \varepsilon) + \mathbb{P}(X^x_{t_1} > x_0) + \mathbb{P}(X^y_{t_1} > x_0) \\ &\leq \frac{1}{\varepsilon}\mathbb{E}[|X^x_{t_1} - X^y_{t_1}|] + \frac{1}{x_0^p}\left(\mathbb{E}[(X^x_{t_1})^p] + \mathbb{E}[(X^y_{t_1})^p]\right). \end{aligned}$$

By Theorem 3.3 and Lemma 6.1, a constant C_p exists such that C_p depends on p, μ_0 and μ_∞ but not on t_1 and satisfies:

$$\int \mathbb{P}((X^x_{t_1}, Y^y_{t_1}) \in B^c_{x_0,\varepsilon})\mu_0(dy)\mu_\infty(dx) \leq \frac{\mathcal{W}_1(\mu_0,\mu_\infty)}{\varepsilon}e^{-\pi t_1} + \frac{C_p}{x_0^p}.$$

Finally, Lemma 6.2 leads to

$$\begin{aligned} &\mathbb{P}(X^x_t = Y^y_t | (X^x_{t_1}, Y^y_{t_1}) \in B_{x_0,\varepsilon}) \\ &\geq \left(1 - \frac{c}{b}x_0\varepsilon - e^{-\frac{a}{b}c(t-t_1)} - \frac{c\varepsilon}{b}\right)\left\{0 \vee 1 - \frac{c}{b}\varepsilon(x_0+g)\right\} \end{aligned}$$

so that plugging the previous inequalities into (42) and (41), one deduces that for every $p > 1$, a constant $\tilde{C}_p$ exists such that for every $t \geq 0$, for every x_0 and ε such that $x_0\varepsilon \leq b/2c$ (with $x_0 > 1$ and $\varepsilon \in (0,1)$),

$$\|\mu_0 P_t - \mu_\infty\|_{TV} \leq \tilde{C}_p\left(x_0\varepsilon + e^{-\frac{a}{b}c(t-t_1)} + \varepsilon + \frac{1}{\varepsilon}e^{-\pi t_1} + \frac{1}{x_0^p}\right).$$

If we try to optimize the above bound, we set $t_1 = \delta t$, $x_0 = C_1 e^{\alpha t}$, $\varepsilon = C_2 e^{-\beta t}$ with $\delta \in (0,1)$ and $\beta > \alpha > 0$ and deduce that a constant $\check{C}_p$ exists such that

$$\|\mu_0 P_t - \mu_\infty\|_{TV} \leq \check{C}_p \exp\left(-t\left\{\beta - \alpha \wedge \frac{ca}{b}(1-\delta) \wedge \delta\pi - \beta \wedge \alpha p\right\}\right).$$

We can choose p as large as we want (μ_0 has moments of any order) and thus α arbitrarily small, the result then follows using an optimization on (β, δ). □

Appendix A: Technical result for the pseudo-regret upper bound

Lemma A.1. *Let $\alpha > 0$, $\gamma_1 \in (0,1)$ and $\tilde{n} \in \mathbb{N}$ such that $\alpha\gamma_{\tilde{n}} < 1$ and $\tilde{n} \geq 1/(\alpha\gamma_1)^2$). We have*

$$\forall n \geq \tilde{n} \qquad \sum_{j=\tilde{n}}^{n-1} \gamma_j \prod_{l=j}^{n-1}(1 - \alpha\gamma_l) \leq \frac{1}{\alpha}$$

Proof Let $j \geq \tilde{n}$. By the inequality $\ln(1+x) \geq x$ for $x > -1$, we have

$$\prod_{l=j}^{n-1}(1-\alpha\gamma_l) = \exp\left(\ln\sum_{l=j}^{n-1}(1-\alpha\gamma_l)\right) \leq \exp\left(-\sum_{l=j}^{n-1}\alpha\gamma_l\right)$$

Using that $x \mapsto 1/\sqrt{x}$ is decreasing,

$$\sum_{l=j}^{n-1}\gamma_l = \sum_{l=j}^{n-1}\frac{\gamma_1}{\sqrt{l}} \geq \gamma_1\sum_{l=j}^{n-1}\int_l^{l+1}\frac{1}{\sqrt{x}}dx = \gamma_1\int_j^n\frac{1}{\sqrt{x}}dx = 2\gamma_1(\sqrt{n}-\sqrt{j})$$

so that

$$\sum_{j=n_0}^{n-1}\gamma_j\prod_{l=j}^{n-1}(1-\alpha\gamma_l) \quad \leq \quad \gamma_1 e^{-2\alpha\gamma_1\sqrt{n}}\sum_{j=n_0}^{n-1}\frac{e^{2\alpha\gamma_1\sqrt{j}}}{\sqrt{j}}.$$

Checking that $x \mapsto \frac{1}{\sqrt{x}}e^{\alpha\gamma_1\sqrt{x}}$ is non-decreasing on $[\frac{1}{(\alpha\gamma_1)^2},\infty)$ one deduces that for any $j \geq n_0$,

$$\sum_{j=n_0}^{n-1}\frac{1}{\sqrt{j}}e^{2\alpha\gamma_1\sqrt{j}} \leq \int_{n_0}^{n}\frac{1}{\sqrt{x}}e^{2\alpha\gamma_1\sqrt{x}}dx \leq \frac{1}{\alpha\gamma_1}.$$

The lemma follows.

□

References

[1] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, pages 2635–2686, 2009.

[2] Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010.

[3] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397422, 2002.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32:4877, 2002.

[5] Jean-Baptiste Bardet, Alejandra Christen, Arnaud Guillin, Florent Malrieu, and Pierre-André Zitt. Total variation estimates for the TCP process. *Electronic Journal of Probability*, 18:1–21, 2013.

[6] Jean-Baptiste Bardet, Alejandra Christen, Arnaud Guillin, Florent Malrieu, and Pierre-André Zitt. Total variation estimates for the TCP process. *Electron. J. Probab.*, 18:no. 10, 21, 2013.

[7] S. Bubeck and N. Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*, volume 5 of *Foundations and Trends in Machine Learning*. 2012.
[8] O. Cappé, A. Garivier, O. Maillard, R. Munos, and G. Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41:1516–1 541, 2013.
[9] N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59:392411, 1999.
[10] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 3 Market place, Woodstock, Oxfordshire, 2006.
[11] Bertrand Cloez. Wasserstein decay of one dimensional jump-diffusions. *preprint*, 2012.
[12] M. Duflo. *Random Iterative Models*. Classics in Mathematics. Springer-Verlag, New-York, 1997.
[13] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. Characterization and convergence.
[14] C. Gittins and H. Robbins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41:148–177, 1979.
[15] A. György and G. Ottucska. Adaptive routing using expert advice. *Computer Journal-Oxford*, 49:180189, 2006.
[16] H. Kushner and G. Yin. *Stochastic approximation and recursive algorithms and applications, Second edition*. Applications of Mathematics, 35. Stochastic Modelling and Applied Probability. Springer-Verlag, New-York, 2003.
[17] H.J. Kushner and D.S. Clark. *Stochastic approximation for constrained and unconstrained systems*. Springer-Verlag, Berlin, 1978.
[18] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.
[19] Damien Lamberton and Gilles Pagès. How fast is the bandit? *Stochastic Analysis and Applications*, 26:603–623, 2008.
[20] Damien Lamberton and Gilles Pagès. A penalized bandit algorithm. *Electronic Journal of Probability*, 13:341–373, 2008.
[21] Damien Lamberton, Gilles Pagès, and Pierre Tarrès. When can the two-armed bandit algorithm be trusted ? *Annals of Applied Probability*, 14:1424–1454, 2004.
[22] K. S. Narendra and I. J. Shapiro. Use of stochastic automata for parameter self-optimization with multi-modal perfomance criteria. *IEEE Trans. Syst. Sci. Cybern.*, 5:352–360, 1969.
[23] M.F. Norman. On linear models with absorbing barriers. *J. Math. Psych.*, 5:225–241, 1968.
[24] Robin Pemantle. Non-convergence to unstable points in urn models and stochastic approximations. *Annals of Probability*, 18:698–712, 1990.
[25] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.