When can the two-armed bandit algorithm be trusted ?

DAMIEN LAMBERTON [*] GILLES PAGÈS [†] PIERRE TARRÈS [‡]

October 29, 2002

Abstract

We investigate the asymptotic behaviour of the so-called two-armed bandit algorithm. It is an example of stochastic approximation procedure whose associated ODE has both a repulsive and an attractive equilibrium, at which the procedure is noiseless. We show that if the gain parameter is constant or goes to 0 not too fast, the algorithm does fall in the noiseless repulsive equilibrium with positive probability whereas it always converges to its natural attractive target when the gain parameter goes to zero at some appropriate rates depending on the parameters of the model. We also elucidate the behaviour of the constant step algorithm when the step goes to 0. Finally, we highlight the connection between the algorithm and the Polya urn. An application to asset allocation is briefly described.

Key words: Two-armed bandit algorithm, Stochastic Approximation, learning automata, Polya urn, asset allocation.

2001 AMS classification: 62L20, secondary 93C40, 91E40, 68T05, 91B32 .

Introduction

The aim of this paper is to deeply investigate the asymptotic behaviour of the so-called two-armed bandit algorithm. This stochastic approximation procedure is widely known in the fields of mathematical psychology and learning automata (see [13] and [15]). Our own motivations are both theoretical and practical as it will be seen further on. Let us first introduce the algorithm itself in a financial context, namely as an adaptive optimal asset allocation model.

[*]Équipe d'analyse et de mathématiques appliquées, ESA 8050,Univ. Marne-la-Valle, Cité Descartes, 5, Bld Descartes, Champs-sur-Marne, F-77454 Marne-la-Vallée Cedex 2, France. dlamb@math.univ-mlv.fr

[†]Laboratoire de Probabilités et Modélisation aléatoire, UMR 7599, Université Paris 6, case 188, 4, pl. Jussieu, F-75252 Paris Cedex 5. gpa@ccr.jussieu.fr

[‡]Laboratoire de Statistique et Probabilités, UMR CNRS C5583, Université Paul Sabatier, 118, route de Narbonne F-31062 Toulouse Cedex 4. tarres@math.ups-tlse.fr

Imagine a fund managed by only two traders, say A and B: every day each of them is in charge of a percentage of the fund, which may vary from day to day. The few wealthy investors (the shareholders) who created the fund wish ideally to allocate the whole fund to the most efficient trader but of course they do not know who he is. They simultaneously want to make some advantage of the performances of the best trader as soon as possible. This means to devise a periodic re-allocation procedure of the fund to the traders based on their (daily or monthly) performances. On the other hand this procedure should be not too "upsetting" to the traders in order to preserve their motivation and self-confidence: one way is to enhance reward rather than punishment. Taking all these specifications into account suggests to proceed as follows: let X_n be the fraction of the fund managed by trader A during day n, the fraction $1 - X_n$ being managed by trader B. Every day, one trader is chosen *at random* and his performances of the day are evaluated. Assume it is A for a moment. If they are considered as outstanding, trader A is rewarded by an extra-allocation for day $n+1$ of $\gamma_{n+1}\%$ of the fraction managed by the trader B during day n (whatever the performances of trader B are since he was not checked). So trader A will manage $X_n + \gamma_{n+1}(1-X_n)\,\%$ of the fund during the day $n+1$. If his performances are not high enough to deserve a reward, nothing happens. The same procedure is applied to trader B when he is checked: if B has outstanding performances, he is allocated an extra $\gamma_{n+1}\%$ of the share managed by A during day n so that, during day $n+1$, the share managed by A will be reduced to $X_{n+1} = X_n - \gamma_{n+1}X_n$ (whatever his performances on day n were). One models the daily performance evaluations of A and B by two sequences of events $(A_n)_{n\geq 1}$ and $(B_n)_{n\geq 1}$ respectively: $A_n = \{A\text{'s performances on day } n \text{ are outstanding}\}$, $B_n = \{B\text{'s performances on day } n \text{ are outstanding}\}$.

A natural policy for the investors of the fund is to reduce the risks induced by this strategy by controlling the largest possible part (in average) of the whole fund. So tossing up for the checked trader with a fair coin is not appropriate. What seems more efficient is to use for the daily toss a biased (virtual) coin so that the probability for trader A or B to be checked at the end of day n is equal to the share of the fund they managed that day, namely X_n and $1 - X_n$ respectively. This virtual coin can be tossed by generating on a computer some i.i.d. random numbers U_n, $n \geq 1$ and by setting

$$\{A \text{ is checked at then end of day } n\} = \{U_{n+1} \leq X_n\}$$

$$\{B \text{ is checked at then end of day } n\} = \{U_{n+1} > X_n\}.$$

All this leads to the following dynamics for X_n: for every $n \geq 1$,

$$\begin{aligned} X_{n+1} &= X_n + \gamma_{n+1}\big((1 - X_n)\mathbf{1}_{\{U_{n+1}\leq X_n\}\cap A_{n+1}} - X_n\mathbf{1}_{\{U_{n+1}>X_n\}\cap B_{n+1}}\big) \qquad (1)\\ X_0 &= x \in [0,1] \end{aligned}$$

where $(\gamma_n)_{n\geq 1}$ is the sequence of gain parameters (or *steps*) satisfying

$$\forall\, n \in \mathbb{N}^*,\ \gamma_n \in (0,1) \quad \text{and} \quad \Gamma_n := \gamma_1 + \cdots + \gamma_n \to +\infty \quad \text{as} \quad n \to +\infty. \qquad (2)$$

(Note this includes the constant step setting $\gamma_n = \gamma \in (0,1)$). The fact that γ_n lies in $(0,1)$ is induced by the modelling (it is a percentage). On the other hand the fact that Γ_n

goes to infinity is a necessary condition to "forget" the starting value: if $\lim_n \Gamma_n < +\infty$, X_n would still converge *a.s.* toward a random variable X_∞, but one could no show that X_∞ takes its values in $\{0,1\}$.

This recursive random procedure was first introduced by Norman in mathematical psychology (1968, see [13]) and then, independently, by Shapiro and Narendra in the engineering literature as a linear learning automata (1969, see [15]). In this field it is known as the Linear Reward-Inaction (L_{R-I}) scheme (see the survey [10] and the book [11] by Narendra about learning automata theory). In both cases, only the constant step setting is considered. The application to optimal adaptive asset allocation in a financial context has been developed in [12].

The algorithm (1) is often mentioned in the literature about stochastic approximation and recursive stochastic algorithms, this time mainly in its decreasing step version (see [5]), as the *two-armed-bandit*. In fact, from a mathematical point of view, it is one of the simplest example of a stochastic approximation algorithm having a "noiseless trap". We will come back further on this property which was another motivation for investigating this algorithm.

The sequence $(U_n)_{n\geq 1}$ and the events A_n, B_n, $n \geq 1$ are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We will make some further assumptions on the events A_n and B_n, namely that the sequence

$$(\mathbf{1}_{A_n}, \mathbf{1}_{B_n})_{n\geq 1} \quad \text{is} \quad \text{i.i.d.}$$

This assumption corresponds to a "stationary" situation: the traders' daily performances are supposed to be independent and "statistically invariant" *i.e.* identically distributed: so one sets

$$\mathbb{P}(A_1) = p_{_A} \quad \text{and} \quad \mathbb{P}(B_1) = p_{_B}.$$

Of course, the owners of the fund do not know whether $p_{_A} > p_B$ or $p_{_A} < p_B$. Finally, one assumes that the sequences

$$(U_n)_{n\geq 1} \quad \text{and} \quad (\mathbf{1}_{A_n}, \mathbf{1}_{B_n})_{n\geq 1} \quad \text{are independent}$$

i.e. the daily tosses are in no way influenced by the respective past (and future) performances of A and B except for the shares respectively managed that day.

To elucidate the *a.s.* asymptotic behaviour of this allocation procedure, one could call upon classical Stochastic Approximation methods like the so-called Ordinary Differential Equation (*ODE*) method. It consists in comparing the asymptotic behaviour of the algorithm $(X_n)_{n\geq 1}$ with that of the related $ODE \equiv \dot{x} = \pi h(x)$ with $\pi := p_{_A} - p_{_B}$ and $\pi\, h(x) := \frac{1}{\gamma_{n+1}} \mathbb{E}(X_{n+1} - X_n \,|\, X_n = x) = \pi\, x(1-x)$ is the *mean* function of the algorithm (see section 2 below). One readily checks that this *ODE* admits two equilibria, 0 and 1, and that, when $p_{_A} > p_{_B}$, its flow $\Phi(x,t) = \frac{x}{(1-x)e^{-\pi t} + x}$ uniformly converges on compact sets of $(0,1]$ toward 1 as $t \to \infty$: the equilibrium 1 is stable with an attraction interval $(0,1]$; on the other hand 0 is repulsive (0 is then called a *trap* for the algorithm). Thus, the celebrated "conditional convergence" Theorem due to Kushner & Clark in [8] says that, under technical assumptions fulfilled here, almost every path of the algorithm that visits infinitely many times a compact subset of the attracting interval of a stable equilibrium will converge toward it. Applying that to a path of the two-armed bandit algorithm shows that *if it does not converge to* 0, then it necessarily visits infinitely often the compact interval $[\varepsilon, 1]$ for some $\varepsilon > 0$ and hence converges toward 1.

In some way it is not really surprising that this approach fails since stability is a second order property whereas the ODE method is based on a first order approximation. Recent sophisticated first order approaches like [2] cannot be more efficient for the same reason.

There is wide literature in Stochastic Approximation about traps and how not to fall into them (see [9, 14, 3, 16, 17, 6]). They all rely on the fact that, if the noise is exciting enough at a repulsive equilibrium x^* then *a.s.*, the algorithm will not converge to it. By "exciting enough" one means that a conditional variance term at x^* is positive. But the main feature of the two-armed-bandit algorithm is that its two equilibria (0 and 1) lie at the boundary of its state space $[0, 1]$, so the above conditional variance term is necessarily identically 0 at the repulsive equilibrium $x^* = 0$ (and at $x^* = 1$ as well). So, the behaviour of the two-armed-bandit algorithm cannot be solved using these approaches.

As far as we know, from a mathematical point of view, the asymptotic behaviour of the algorithm has not been elucidated in the literature. The present paper derives from results obtained independently by the third author in [17] and the other two authors.

Heuristics, probably suggested by the behaviour of the mean algorithm, seems to consider that the procedure described above works well in practice.

So, it is interesting for both theoretical and practical motivations to analyze the behaviour of the two-armed-bandit algorithm, that is

– Is it possible to choose the gain parameter sequence so that the algorithm *a.s.* never fails?

– Conversely, does the algorithm "fall in its noiseless trap 0" for some sequences of gain parameters?

This leads to introduce the following terminology when $0 \le p_B < p_A \le 1$. (Inverting the rôle played by A and B solves the case $0 \le p_A < p_B \le 1$.) The two-armed-bandit algorithm is

– *fallible* when starting from $x \in (0, 1)$ if $\mathbb{P}_x(X_n \to 0) > 0$,

– *a.s. infallible* if $\mathbb{P}_x(X_n \to 0) = 0$ for every $x \in (0, 1)$.

Although not directly interested by the critical case $p_A = p_B$, we will deeply investigate it since it is key to solve the general case thanks to a comparison result.

The paper is organized as follows. In Section 1 is stated the main theoretical result of the paper, namely Theorem 1, concerning the convergence and the fallibility of the algorithm. Two corollaries show its consequences on usual parametrized families of steps. Section 2 is devoted to some elementary although important facts on which relies the proof of Theorem 1, Proposition 2 on one hand and the comparison result stated in Proposition 3 on the other hand. Section 3 is mainly devoted to the proof of Theorem 1: Subsection 3.1

solves item (b) and Subsection 3.3 solves item (c). Subsection 3.2 has a particular status: it is a kind of bridge between Subsections 3.1 and 3.3: we focus on the special case where the step γ_n is constant which is the historical setting considered by those who devised the procedure. It is shown in Theorem 2 that the (positive) probability of failure for the algorithm with constant step γ goes to 0 as γ goes to zero. Some bounds are displayed, the optimality of which are not known to us. Section 4 makes a connection between regular Polya urns and the two-armed bandit algorithm: we show that the two-armed bandit algorithm can be seen as a generalized Polya urn. Thus, we retrieve partially the infallibility results of Theorem 1 using standard methods of proof for the Polya urns like the "moment method" and the log-method. In the martingale case ($p_A = p_B$) these approaches yield some more information about the distribution of the *a.s.* limit X_∞ of X_n. The last section contains some provisional remarks and additional results.

Note that, except for the notations and the elementary facts contained in Section 2, other subsections are self-contained and can be read independently.

NOTATIONS: • The letter C will denote a positive real constant that may change from line to line.

• The letter ξ will denote a *random* positive real constant that may change from line to line.

• Let $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ be two sequences of positive real numbers. The symbol $a_n \asymp b_n$ is for $a_n = O(b_n)$ and $b_n = O(a_n)$ whereas the symbol $a_n \sim b_n$ means $\lim_n a_n/b_n = 1$.

1 The main result

Theorem 1 (a) *A.s.* CONVERGENCE: • *If* $0 < p_B < p_A \leq 1$ *and* $x \in (0,1)$, $(X_n)_{n\geq 0}$ *is a bounded* $\mathcal{F}$*-sub-martingale, hence* $\mathbb{P}_x$*-a.s. converging toward a random variable* X_∞*. The random variable* X_∞ *takes values in* $\{0,1\}$ *and*

$$\mathbb{P}_x(X_\infty = 1) = x + \pi \sum_{n\geq 0} \gamma_{n+1}\mathbb{E}_x(h(X_n)) > x + \pi\,\gamma_1 x(1-x) > x.$$

(If $p_B = 0$ *and* $p_A > 0$*, then,* X_n *is nondecreasing and converges toward* 1*.)*

• *If* $0 < p_B = p_A \leq 1$ *and* $x \in (0,1)$*, then* $(X_n)_{n\geq 0}$ *is a bounded* $\mathcal{F}$*-martingale* $\mathbb{P}_x$*-a.s. converging toward a random variable* X_∞*.*

Moreover,
if $\sum_{n\geq 0} \gamma_{n+1}^2 = +\infty$*,* X_∞ *is* $\{0,1\}$*-valued with distribution* $\mathcal{B}ernoulli(x)$*.*

(If $p_B = p_A = 0$*, then* $X_n = x$*,* $\mathbb{P}_x$*-a.s. for every* $n \geq 0$*.)*

• *If* $x \in \{0,1\}$*, then* $X_n = x$*,* $\mathbb{P}_x$*-a.s. for every* $n \geq 0$*.*

(b) CONVERGENCE TO 0 WITH POSITIVE PROBABILITY: *If*

$$\sum_{n\geq 0} \prod_{k=1}^{n} (1 - p_B\gamma_k) < +\infty \tag{3}$$

then, for every $x \in [0,1)$

$$\mathbb{P}_x(X_\infty = 0) > 0.$$

In particular,

– *if* $0 < p_B < p_A \leq 1$, *then, for every* $x \in (0,1)$,
the two-armed-bandit algorithm starting from x is fallible.

– *if* $0 < p_B = p_A \leq 1$ *and* $\sum_n \gamma_n^2 < +\infty$, *then, for every* $x \in (0,1)$,

$$\mathbb{P}_x(X_\infty = 0),\ \mathbb{P}_x(X_\infty = 1) \ \ \textit{and} \ \ \mathbb{P}_x(X_\infty \in (0,1)) > 0.$$

(c) Convergence to a nonzero value: *Assume* $0 \leq p_B \leq p_A \leq 1$ *and*

$$\gamma_n = O\left(\Gamma_n e^{-p_B \Gamma_n}\right). \tag{4}$$

Then, for every $x \in (0,1]$,

$$\mathbb{P}_x\left(X_\infty = 0\right) = 0.$$

In particular,

– *if* $0 \leq p_B < p_A \leq 1$
then, for every $x \in (0,1]$,

$$X_\infty = 1 \quad \mathbb{P}_x\text{-}a.s. \ \ i.e. \text{ the algorithm is } a.s. \text{ infallible,}$$

– *when* $0 \leq p_B = p_A \leq 1$
then, for every $x \in (0,1)$,

$$X_\infty \in (0,1) \ \ \mathbb{P}_x\text{-}a.s.$$

Proof: This theorem follows from Propositions 2, 4 and 5 below. These propositions can be seen as steps of the proofs of the theorem. $\diamond$

Corollary 1 *(*Fallibility*) Let* $p_B \in (0,1]$.
(a) Constant Step: *If the step* $\gamma_n := \gamma \in (0,1)$, *the two-armed-bandit algorithm does converge toward* 0 *with positive probability. Namely,*

$$\forall\, x \in (0,1), \quad \mathbb{P}_x(X_\infty = 0) \geq (1-x)^{\frac{1}{p_B \gamma}} > 0.$$

(b) Power Step (I): *One considers the family of "power"steps* $\gamma_n := \left(\frac{C}{n+C}\right)^\alpha$, $0 < \alpha \leq 1$, $n \geq 1$. *These step sequences satisfy (2) and Assumption (3) reads*

$$(0 < \alpha < 1) \qquad \textit{or} \qquad (\alpha = 1 \ \ \textit{and} \ \ C > 1/p_B).$$

(c) *In particular, if* $0 < p_B < p_A \leq 1$, *the two-armed bandit algorithm is fallible starting from any* $x \in [0,1)$ *for the step sequences specified in the above items* (a) *and* (b).

Proof: (b) is a straightforward consequence of Proposition 4.
(a) The lower bound for $\mathbb{P}_x(X_\infty = 1)$ follows from Theorem 1(a). That for $\mathbb{P}_x(X_\infty = 0)$ needs further care. It relies on Inequality (8) established in the proof of Proposition 4: setting $\gamma_n = \gamma \in (0,1)$, it reads

$$\mathbb{P}_x(X_\infty = 0) \geq \mathbb{E}_x\left(\prod_{n\geq 1}\left(1 - x\prod_{k=1}^{n-1}(1-\mathbf{1}_{B_k}\gamma_k)\right)\right).$$

Then the computations can easily be carried on: Jensen Inequality yields

$$\mathbb{P}_x(X_\infty = 0) \quad \geq \exp\left(\textstyle\sum_{n\geq 1}\mathbb{E}_x\log(1-xZ_n)\right) \quad \text{with} \quad Z_n := \prod_{k=1}^{n-1}(1-\gamma\mathbf{1}_{B_k}).$$

$$\begin{aligned}
\text{Now} \qquad\qquad \mathbb{E}_x\left(\log(1-xZ_n)\right) &= -\sum_{m\geq 1}\frac{x^m}{m}\,\mathbb{E}(Z_n^m)\\
&= -\sum_{m\geq 1}\frac{x^m}{m}\left((1-\gamma)^m p_B + 1 - p_B\right)^{n-1}\\
\text{so that} \qquad \sum_{n\geq 1}\mathbb{E}_x\left(\log(1-xZ_n)\right) &= -\frac{1}{p_B}\sum_{m\geq 1}\frac{x^m}{m(1-(1-\gamma)^m)}\\
&\geq \frac{1}{p_B\gamma}\log(1-x).
\end{aligned}$$

Finally, for every $x \in [0,1]$,

$$\mathbb{P}_x(X_\infty = 0) \geq (1-x)^{\frac{1}{p_B\gamma}}. \qquad \diamond \tag{5}$$

Corollary 2 *(*Infallibility*) Let $0 < p_B \leq p_A \leq 1$.*
(a) Power step (II): *If $\gamma_n := (\frac{C}{n+C})^\alpha$ with*

$$\alpha = 1 \qquad \textit{and} \qquad C \leq \frac{1}{p_B}$$

then Assumptions (2) and (4) are fulfilled.
(b) Power step (III): *Set $\gamma_n := \frac{\Delta_n}{1+\Delta_1+\cdots+\Delta_n}$ where $(\Delta_n)_{n\geq 1}$ is a sequence of positive real numbers satisfying $\Delta_n \sim C\,n^{1/p_B-1}\log^\alpha n$. Then*

$$\gamma_n \sim \frac{1}{p_B n} \qquad \textit{and Assumption (4) reads} \qquad \alpha \leq \frac{1}{p_B}.$$

(c) In particular, if $0 < p_B < p_A \leq 1$, the two-armed bandit algorithm is a.s. infallible for the step sequences specified in the above items (a) and (b).

PROOF: (a) On easily checks that if $0 < \alpha < 1$ the assumption cannot be fulfilled. If $\alpha = 1$, $\Gamma_n = C \log n + C' + o(1)$. Consequently, Assumption (4) reads $1/n = O\left(\log(n) n^{-C\,p_B}\right)$, *i.e.* $C\,p_B \leq 1$.

(b) One has $S_n := \Delta_1 + \cdots + \Delta_n \sim C\,p_B n^{1/p_B} \log^{\alpha} n$ so that $\gamma_n \sim \frac{1}{p_B}\frac{1}{n}$. Furthermore $S_n \asymp e^{\Gamma_n}$ (see the preliminary remark below Lemma 1 in Subsection 3.3 for more details, especially Inequality (14)). Then Assumption (4) reads $\gamma_n = O(\Gamma_n S_n^{-p_B})$. Finally, notice that $\Gamma_n S_n^{-p_B} = O\left(\dfrac{\log^{1-\alpha p_B} n}{n}\right)$. $\diamondsuit$

Remarks: • It is to be noticed that the critical case for the steps which is not entirely elucidated by the above results lies "inside" $\gamma_n \sim \frac{1}{p_B n}$.

• The asymptotics of the constant step setting when the step γ goes to 0 is elucidated in Theorem 2 of subsection 3.2.

2 Some elementary facts

The random innovation at time n is clearly $\varepsilon_n := (U_n, \mathbf{1}_{A_n}, \mathbf{1}_{B_n})$ (the ε_n's are i.i.d.). Let $\mathcal{F}_n := \sigma(\varepsilon_1, \ldots, \varepsilon_n)$ denote its natural filtration. It follows from Equation (1) that $(X_n)_{n\geq 1}$ is obviously a $[0,1]$-valued $\underline{\mathcal{F}}$-Markov chain (homogeneous if $\gamma_n = \gamma$). For notational convenience, we will denote by $\mathbb{P}_x$ the distribution of the whole sequence $(X_n)_{n\geq 0}$ starting at $x \in [0,1]$. One also derives from Equation (1) some straightforward properties of the algorithm.

Proposition 1 *For every $x \in (0,1)$ and every $n \geq 1$, $X_n \in (0,1)$. On the other hand, both states 0 and 1 are absorbing i.e., if $x \in \{0,1\}$, $X_n = x$ for every $n \geq 1$, $\mathbb{P}_x$-a.s.*

Viewed as a stochastic approximation procedure, its canonical form reads

$$\begin{aligned} X_n &= X_{n-1} + \gamma_n \pi\, h(X_{n-1}) + \gamma_n \Delta M_n \qquad (6)\\ \text{where } \pi &:= p_A - p_B, \qquad h(x) := x(1-x)\\ \text{and } \Delta M_n &:= (1 - X_{n-1})\mathbf{1}_{\{U_n \leq X_{n-1}\}\cap A_n} - X_{n-1}\mathbf{1}_{\{U_n > X_{n-1}\}\cap B_n} - \pi\, h(X_{n-1}). \end{aligned}$$

is a $\underline{\mathcal{F}}$-martingale increment.

Remark: The *mean* algorithm associated with $(X_n)_{n\geq 1}$ is the deterministic recursive procedure defined by

$$x_{n+1} = x_n + \gamma_{n+1}\pi\, h(x_n),\; x_0 \in [0,1].$$

It can be solved very easily: when $\pi = p_A - p_B > 0$ and $x_0 \in (0,1]$, the sequence x_n is $[0,1]$-valued and nondecreasing, hence converging toward x_∞. Since the series $\sum_n \gamma_n h(x_{n-1}) < +\infty$ whereas $\sum_n \gamma_n = +\infty$, it is obvious that $h(x_\infty) = 0$. Hence $x_\infty = 1$ since $x_\infty \geq x_0 > 0$. So, the mean algorithm never fails in pointing out the best trader since it asymptotically assigns the whole fund to manage to A when $p_A > p_B$ (and to B when

$p_B > p_A$). Unfortunately, it needs to know *a priori* who is the best trader *i.e.* whether $p_A > p_B$ or $p_B > p_A$.

Similarly, Equation (6) shows that $(X_n)_{n\geq 1}$ is a bounded sub-martingale and one derives (see Proposition 2 below) that then $\mathbb{P}_x$-*a.s.* X_n converges toward a $\{0,1\}$-valued random variable X_∞. But this time, there is no straightforward argument showing that the procedure always points out the best trader *e.g.* $X_\infty = 1$ $\mathbb{P}_x$-*a.s.* when $p_A > p_B$ and $x \in (0,1]$. The next proposition yields some first answers about the behaviour of the algorithm.

Proposition 2 (*a*) SUB-MARTINGALE CASE: *If* $0 < p_B < p_A \leq 1$ *and* $x \in (0,1)$, $(X_n)_{n\geq 0}$ *is a bounded* $\mathcal{F}$*-sub-martingale, hence* $\mathbb{P}_x$*-a.s. converging toward a random variable* X_∞*, taking values in* $\{0,1\}$ *and*

$$\mathbb{P}_x(X_n \to 1) = x + \pi \sum_{n\geq 0} \gamma_{n+1}\mathbb{E}_x(h(X_n)) > x + \pi\,\gamma_1 x(1-x) > x.$$

(If $p_B = 0$ *and* $p_A > 0$*, then,* X_n *is nondecreasing and converges toward* 1*.)*

(*b*) MARTINGALE CASE: *If* $0 < p_B = p_A \leq 1$ *and* $x \in (0,1)$*, then* $(X_n)_{n\geq 0}$ *is a bounded* $\mathcal{F}$*-martingale* $\mathbb{P}_x$*-a.s. converging toward a random variable* X_∞*. Moreover,*

– *if* $\sum_{n\geq 0} \gamma_{n+1}^2 = +\infty$*, then* X_∞ *is* $\{0,1\}$*-valued with Bernoulli distribution* $\mathcal{B}(x)$*,*

– *if* $\sum_{n\geq 0} \gamma_{n+1}^2 < +\infty$*, then* X_∞ *is* $[0,1]$*-valued and satisfies* $\mathbb{P}_x\left(X_\infty \in (0,1)\right) > 0$*.*

(If $p_B = p_A = 0$*, then* $X_n = x$*,* $\mathbb{P}_x$*-a.s.)*

Proof: (*a*) $(X_n)_{n\geq 0}$ is obviously a bounded $\mathcal{F}$-sub-martingale. Furthermore its *a.s.* limit, say X_∞, satisfies

$$\begin{aligned}\mathbb{E}_x(X_\infty) &= \lim_n \mathbb{E}_x(X_n) = x + \lim_n \mathbb{E}_x\left(\sum_{k=0}^{n-1} \gamma_{k+1}\pi\, h(X_k) + \gamma_{k+1}\Delta M_{k+1}\right)\\ &= x + \pi \sum_{n\geq 0}\gamma_{n+1}\mathbb{E}_x h(X_n).\end{aligned}$$

Hence $\sum_{n\geq 0} \gamma_{n+1}\mathbb{E}_x(h(X_n)) < +\infty$ since $\pi > 0$ and consequently $\sum_{n\geq 0}\gamma_{n+1}h(X_n) < +\infty$ *a.s.* which in turn implies $\liminf_n h(X_n) = 0$ since h is nonnegative and $\sum_{n\geq 0}\gamma_{n+1} = +\infty$. It follows that $h(X_\infty) = 0$ so that $X_\infty \in \{0,1\}$ *a.s.*. Finally, $\mathbb{P}_x(X_\infty = 1) = \mathbb{E}_x(X_\infty)$.

(*b*) $(X_n)_{n\geq 0}$ is obviously a bounded $\mathcal{F}$-martingale. When $p_A = p_B$, an elementary computation shows that

$$\mathbb{E}_x(X_n(1-X_n)) = (1 - p_A\gamma_n^2)\mathbb{E}_x\left(X_{n-1}(1-X_{n-1})\right) = x(1-x)\prod_{k=1}^{n}(1 - p_A\gamma_k^2)$$

so that

$$\mathbb{E}_x(X_\infty(1-X_\infty)) = x(1-x)\prod_{n\geq 1}(1-p_A\gamma_n^2).$$

The announced result follows since the infinite product converges toward a nonzero limit iff $\sum_n \gamma_n^2 < +\infty$. $\diamondsuit$

One may specify without loss of generality the definition of events A_n and B_n: these two events never interact so only the marginal distributions of A_1 and B_1 are involved in the distribution of the whole sequence $(X_n)_{n\geq 0}$. So, one set

$$A_n := \{V_n \leq p_A\} \qquad \text{and} \qquad B_n := \{V_n \leq p_B\} \tag{7}$$

where $(U_n)_{n\geq 1}$, $(V_n)_{n\geq 1}$ are two independent i.i.d. $U([0,1])$-distributed sequences.

Then, this "coupled" algorithm is pathwise monotonous as a function of p_A, the parameter p_B being fixed. This is established in the proposition below.

Proposition 3 *(Pathwise comparison result) Let $x \in (0,1)$. Let (X_n) and (X'_n) denote two two-armed-bandit algorithms "built" from the sequences (U_n) and (V_n), starting from $x \leq x'$ and associated to the parameters (p_B, p_A) and (p_B, p'_A) respectively with $p_A \leq p'_A$. Then for every $n \in \mathbb{N}$,*

$$X_n \leq X'_n.$$

In particular

$$\{X'_n \to 0\} \subset \{X_n \to 0\}.$$

Proof: The result follows from what happens between time 0 and 1. One inspects the four possible cases

– On $\{U_1 \leq x'\} \cap \{V_1 \leq p'_A\}$, $X'_1 = x' + \gamma_1(1-x')$ and $X_1 \leq x + \gamma_1(1-x) \leq X'_1$.

– On $\{U_1 \leq x'\} \cap \{V_1 > p'_A\}$, $X'_1 = x'$ and $X_1 \leq x \leq x'$.

– On $\{U_1 > x'\} \cap \{V_1 \leq p_B\}$, $X'_1 = (1-\gamma_1)x'$ and $X_1 = (1-\gamma_1)x \leq X'_1$.

– On $\{U_1 > x'\} \cap \{V_1 > p_B\}$, $X'_1 = x' \geq x = X_1$. $\diamondsuit$

Remark: One checks that, when $p_A \geq p_B$, the trajectories of the general form of the algorithm are nondecreasing as a function of their starting value.

In particular, *the function $x \mapsto \mathbb{P}_x(X_\infty = 1)$ is nondecreasing.*

3 When does the two-armed-bandit algorithm fail ?

3.1 Quite often. . .

Proposition 4 *If $\displaystyle\sum_{n\geq 0}\prod_{k=1}^{n}(1-p_B\gamma_k) < +\infty$ then*

$$\mathbb{P}_x(X_n \to 0) > 0 \textit{ for every } x \in [0,1).$$

Proof: One considers the event

$$D_\infty := \left\{ U_n > x \prod_{k=1}^{n-1} (1 - \gamma_k \mathbf{1}_{B_k}) \text{ for every } n \geq 1 \right\}$$

where $\prod_{k=1}^{0} = 1$. One checks by induction that, on D_∞, $X_n = x \prod_{k=1}^{n} (1 - \gamma_k \mathbf{1}_{B_k})$. The algorithm is nonincreasing toward its limit X_∞. Hence

$$\mathbb{E}_x \left(\mathbf{1}_{D_\infty} X_\infty \right) \leq \lim_n \mathbb{E}_x \left(\mathbf{1}_{D_\infty} X_n \right) \leq \lim_n \mathbb{E}_x \left(x \prod_{k=1}^{n} (1 - \gamma_k \mathbf{1}_{B_k}) \right) = x \prod_{n \geq 1} (1 - p_{_B} \gamma_n) = 0$$

since $p_{_B} > 0$, so that $X_\infty = 0$ sur D_∞. On the other hand, $(U_n)_{n \geq 1}$ being i.i.d., uniformly distributed and independent of the sequence $(B_n)_{n \geq 1}$,

$$\mathbb{P}_x(D_\infty) = \mathbb{E}_x \left(\mathbb{P}_x(D_\infty / \sigma(B_n,\, n \geq 1)) \right) = \mathbb{E}_x \left(\prod_{n \geq 1} \left(1 - x \prod_{k=1}^{n-1} (1 - \mathbf{1}_{B_k} \gamma_k) \right) \right).$$

Consequently

$$\mathbb{P}_x(X_\infty = 0) \geq \mathbb{E}_x \left(\prod_{n \geq 1} \left(1 - x \prod_{k=1}^{n-1} (1 - \mathbf{1}_{B_k} \gamma_k) \right) \right). \tag{8}$$

Now, the events B_k are independent, hence

$$\mathbb{E}_x \left(\sum_{n \geq 1} \prod_{k=1}^{n-1} (1 - \mathbf{1}_{B_k} \gamma_k) \right) = \sum_{n \geq 1} \prod_{k=1}^{n-1} (1 - p_{_B} \gamma_k) < +\infty$$

so that $\sum_{n \geq 1} \prod_{k=1}^{n-1} (1 - \mathbf{1}_{B_k} \gamma_k) < +\infty$ $\mathbb{P}_x$-*a.s.* Consequently the infinite product $\prod_{n \geq 1} \left(1 - x \prod_{k=1}^{n-1} (1 - \mathbf{1}_{B_k} \gamma_k) \right)$ converges toward a $\mathbb{P}_x$-*a.s.* positive random variable. Hence, $\mathbb{P}_x(D_\infty) > 0$. $\diamondsuit$

3.2 (Especially with constant step although...)

The "failure" result obtained in Corollary 1(b) for the algorithm with "slowly" decreasing step is to be compared with the asymptotics of its behaviour with constant step. We know from Corollary 1(a) that, if $0 < p_{_B} \leq p_{_A} \leq 1$, and $\gamma_n = \gamma \in (0, 1)$, then, for every $x \in (0, 1]$, the algorithm with step γ (denoted $(X_n^\gamma)_{n \geq 0}$ in this paragraph) does fail with positive probability: namely, it converges $\mathbb{P}_x$-*a.s.* toward a $\{0, 1\}$-valued random variable X_∞^γ satisfying $\mathbb{P}(X_\infty^\gamma = 0) > 0$. The following theorem shows however that the probability of failure goes to 0 as $\gamma \to 0$.

The fallibility of the algorithm with constant step and this property is known by specialists in Learning Automata theory (see the discussion in chap. 5 in [11]), although not clearly established mathematically in full generality.

Theorem 2 *Assume that* $0 \le p_{_B} < p_{_A} \le 1$ *and* $\gamma_n = \gamma \in (0,1)$. *Then, for every* $x \in (0,1]$,

$$\mathbb{P}_x(X^\gamma_\infty = 1) \ge 1 - \frac{2p_{_A}\gamma}{\pi(1-\gamma)^2}\left(\frac{1}{x} - 1\right) \quad \textit{hence} \quad \lim_{\gamma \to 0_+} \mathbb{P}_x(X^\gamma_\infty = 1) = 1.$$

Proof: We focus on the function

$$\psi_\gamma(x) := \mathbb{E}_x\left(\sum_{n\ge 0} h(X_n)\right) = \sum_{n \ge 0} P^n_\gamma h(x)$$

where $P_\gamma(x, dy)$ denotes the Markov transition of the two-armed-bandit algorithm with constant step $\gamma \in (0,1)$. A straightforward computation shows that

$$P_\gamma(f)(x) = p_{_A}\, x\, f(x + \gamma(1-x)) + p_{_B}(1-x) f(x(1-\gamma)) + (1 - p_{_A} x - p_{_B}(1-x)) f(x).$$

At this stage, it is convenient to introduce the operator Q_γ formally defined by

$$Q_\gamma(g)(x) := \frac{P_\gamma(gh)}{h}.$$

It reads on functions $g : [0,1] \to \mathbb{R}$

$$\begin{aligned} Q_\gamma(g)(x) &= (1-\gamma)(p_{_A}(x + \gamma(1-x))g(x+\gamma(1-x)) + p_{_B}(1 - x(1-\gamma))g(x(1-\gamma))) \\ &\quad + (1 - p_{_A} x - p_{_B}(1-x))g(x). \end{aligned}$$

It is clear that ψ_γ satisfies $\psi_\gamma = h\,\chi_\gamma$ where $\chi_\gamma := \sum_{n\ge 0} Q^n_\gamma(\mathbf{1})$. One shows by successive inductions and a little elementary Calculus that χ_γ and $x \mapsto \frac{1}{\pi\,\gamma\,x} - \chi_\gamma(x)$ are absolutely decreasing functions (an infinitely differentiable function f is *absolutely decreasing* on $(0,1)$ if its successive derivatives $f^{(n)}$ satisfy $(-1)^n f^{(n)} \ge 0$ for every $n \ge 0$). On one hand, one derives that ψ_γ is indefinitely differentiable on $(0,1]$ and, on the other hand that

$$0 \le \chi_\gamma(x) \le \frac{1}{\pi\,\gamma\,x}, \quad |\chi'_\gamma(x)| \le \frac{1}{\pi\,\gamma\,x^2} \quad \text{and} \quad 0 \le \chi''_\gamma(x) \le \frac{2}{\pi\,\gamma\,x^3}. \tag{9}$$

Then,it follows from the definition of ψ_γ that

$$\psi_\gamma - P_\gamma \psi_\gamma = h \qquad \text{and} \qquad \mathbb{P}_x(X^\gamma_\infty = 1) = x + \pi\,\gamma\,\psi_\gamma(x). \tag{10}$$

The first of these two identities reads

$$\begin{aligned} p_{_A} x(\psi_\gamma(x) - \psi_\gamma(x + \gamma(1-x))) + p_{_B}(1-x)(\psi_\gamma(x) - \psi_\gamma((1-\gamma)x)) &= x(1-x) \\ \textit{i.e.} \quad p_{_A}\, x \int_\gamma^0 \psi'_\gamma(x + t(1-x))(1-x)dt + p_{_B}(1-x)\int_\gamma^0 \psi'_\gamma(x - tx)(-x)dt &= x(1-x). \end{aligned}$$

Hence, simplifying by $x(1-x)$ for every $x \in (0,1)$, yields

$$p_{_A} \int_0^\gamma \psi'_\gamma(x + t(1-x))\,dt - p_{_B}\int_0^\gamma \psi'_\gamma(x - tx)\,dt = -1.$$

Now,
$$\begin{aligned}\psi'_\gamma(x+t(1-x)) &= \psi'_\gamma(x)+\int_0^t \psi''_\gamma(x+s(1-x))(1-x)\,ds\\ \psi'_\gamma(x-tx) &= \psi'_\gamma(x)+\int_0^t \psi''_\gamma(x(1-s))(-x)\,ds\end{aligned}$$

so that

$$\pi\,\gamma\psi'_\gamma(x)+p_{_A}(1-x)\int_0^\gamma\int_0^t \psi''_\gamma(x+s(1-x))\,ds\,dt+p_{_B}\,x\int_0^\gamma\int_0^t \psi''_\gamma(x-sx)\,ds\,dt=-1.$$

Combining a rewriting of this identity with an obvious inequality leads to

$$\begin{aligned}-\psi'_\gamma(x) &\geq \frac{1}{\pi\,\gamma}\left(1+\int_0^\gamma\int_0^t\Big(p_{_A}(1-x)\,\psi''_\gamma(x+s(1-x))+p_{_B}\,x\,\psi''_\gamma(x(1-s))\Big)\,ds\,dt\right)\\ &\geq \frac{1}{\pi\,\gamma}-\frac{p_{_A}(1-x)+p_{_B}\,x}{\pi\gamma}\frac{\gamma^2}{2}\sup_{u>x(1-\gamma)}|\psi''_\gamma(u)|.\end{aligned}$$

Plugging Inequalities (9) in the equality $\psi''_\gamma=\chi''_\gamma\,h+2\,\chi'_\gamma\,h'+\chi_\gamma\,h''$ yields

$$|\psi''_\gamma(x)|\leq\frac{2(1-x)}{\pi\gamma x^2}+\frac{2|1-2x|}{\pi\gamma x^2}+\frac{2}{\pi\gamma x}\leq\frac{4}{\pi\gamma x^2}$$

and consequently

$$-\psi'_\gamma(x)\geq\frac{1}{\pi\,\gamma}-\frac{p_{_A}-\pi\,x}{\pi\gamma}\frac{2\gamma^2}{\pi\gamma(1-\gamma)^2x^2}.$$

Now $\psi_\gamma(y)=\psi_\gamma(1)+\displaystyle\int_y^1(-\psi'_\gamma(u))du=\int_y^1(-\psi'_\gamma(u))du$ ($\psi_\gamma(1)=0$), hence

$$\psi_\gamma(x)\geq\frac{1-x}{\pi\,\gamma}+\frac{2p_{_A}}{\pi^2(1-\gamma)^2}\left(1-\frac{1}{x}\right).$$

Finally, one comes to

$$\mathbb{P}_\gamma(X_n^\gamma\to 1)\geq x+1-x-\frac{2p_{_A}\gamma}{\pi(1-\gamma)^2}\left(\frac{1}{x}-1\right).\quad\diamond$$

3.3 But not always !

Proposition 5 *Assume $0\leq p_{_B}\leq p_{_A}\leq 1$ and Assumption (4). Then, for every $x\in(0,1]$,*

$$\mathbb{P}_x\left(X_\infty=0\right)=0. \tag{11}$$

Remark: As soon as $p_{_B}>0$, Assumption (4) implies $\sum_{n\geq1}\gamma_n^2<+\infty$: for $n\geq1$,

$$\begin{aligned}\sum_{k=1}^n\gamma_k^2 &\leq C\sum_{k=1}^n(\Gamma_k-\Gamma_{k-1})\Gamma_k e^{-p_B\Gamma_k}\\ &\leq C\int_0^{\Gamma_n}ue^{-p_Bu}du\leq C\int_0^{+\infty}ue^{-p_Bu}du<+\infty\end{aligned}$$

since $u\mapsto ue^{-p_Bu}$ is nonincreasing for u large enough.

Lemma 1 *(Case $p_{_A} = p_{_B} = 1$) Assume $p_{_A} = p_{_B} = 1$ and set $\Delta_0 = 1$ and $\Delta_n = \dfrac{\gamma_n}{\prod_{k=1}^n (1-\gamma_k)}$, for $n \geq 1$. If the sequence $(\Delta_n)_{n\in\mathbb{N}}$ satisfies*

$$\Delta_n = O(\Gamma_n) \tag{12}$$

then, for every $x \in (0,1]$, $\mathbb{P}_x(X_\infty = 0) = 0$.

Preliminary remark: With the notations of the Lemma, we have

$$\gamma_n = \frac{\Delta_n}{S_n}, \quad \text{with} \quad S_n = \sum_{k=0}^{n} \Delta_k. \tag{13}$$

The partial sums S_n and Γ_n satisfy for every $n \geq 1$,

$$\log S_n - \sum_{k=1}^{n} \frac{\gamma_k^2}{1-\gamma_k} \leq \Gamma_n \leq \log S_n. \tag{14}$$

This follows from the easy comparisons (with $S_0 = \Delta_0 = 1$),

$$\Gamma_n = \sum_{k=1}^{n} \frac{\Delta_k}{S_k} \begin{cases} \leq & \displaystyle\int_1^{S_n} \frac{du}{u} = \log S_n \\ \geq & \displaystyle\sum_{k=1}^{n} \frac{S_{k-1}}{S_k} \int_{S_{k-1}}^{S_k} \frac{du}{S_{k-1}} \geq \sum_{k=1}^{n} (1-\gamma_k) \int_{S_{k-1}}^{S_k} \frac{du}{u} \geq \log S_n - \sum_{k=1}^{n} \frac{\gamma_k^2}{1-\gamma_k}. \end{cases}$$

Hence $S_n \to \infty$ as $n \to \infty$ since $\Gamma_n \to \infty$ and $\gamma_n \leq C\Gamma_n/S_n \leq C\,\Gamma_n e^{-\Gamma_n}$. Consequently, the former remark applies with $p_{_B} = 1$ and shows that

$$\sum_{n\geq 1} \gamma_n^2 < +\infty.$$

Finally, Assumption (12) implies

$$\Gamma_n \sim \log S_n \quad \text{and} \quad S_n \asymp e^{\Gamma_n} \quad \text{as } n \to \infty. \tag{15}$$

Proof of Lemma 1: The algorithm can now be rewritten as follows (we assume $p_{_A} = p_{_B} = 1$).

$$S_{n+1}X_{n+1} = S_{n+1}X_n + \Delta_{n+1}\left(\mathbf{1}_{\{U_{n+1}\leq X_n\}} - X_n\right).$$

Hence,

$$S_{n+1}X_{n+1} = S_n X_n + \Delta_{n+1}\mathbf{1}_{\{U_{n+1}\leq X_n\}}.$$

Let $Y_n := S_n X_n$. We will first prove that $\lim_n Y_n = +\infty$, *a.s.* Since the sequence $(\Delta_n/\Gamma_n)_{n\geq 1}$ is bounded and $\mathbb{E}\left(\mathbf{1}_{\{U_{n+1}\leq X_n\}} \mid \mathcal{F}_n\right) = X_n$, we have (see, *e.g.*, Theorem 2.7.33 in [4])

$$\left\{\sum_{n=0}^{\infty} \frac{\Delta_{n+1}}{\Gamma_{n+1}} \mathbf{1}_{\{U_{n+1}\leq X_n\}} = \infty\right\} = \left\{\sum_{n=0}^{\infty} \frac{\Delta_{n+1}}{\Gamma_{n+1}} X_n = \infty\right\} \quad a.s.$$

But $$\Delta_{n+1}X_n \;\ge\; \Delta_{n+1}x\prod_{k=1}^{n}(1-\gamma_k) \ge \Delta_{n+1}x\prod_{k=1}^{n+1}(1-\gamma_k) = \gamma_{n+1}\,x$$

so that
$$\sum_{n=0}^{\infty}\frac{\Delta_{n+1}}{\Gamma_{n+1}}X_n \ge x\sum_{n=0}^{\infty}\frac{\gamma_{n+1}}{\Gamma_{n+1}} = +\infty$$

since $\Gamma_n \to \infty$ and $\gamma_n \to 0$ as $n \to \infty$. Consequently the nondecreasing sequence $(Y_n)_{n\ge 0}$ satisfies

$$\limsup_n Y_n = \limsup_n \sum_{k=0}^{n-1}\Delta_{k+1}\mathbf{1}_{\{U_{k+1}\le X_k\}} \ge \gamma_1\sum_{n\ge 0}\frac{\Delta_{n+1}}{\Gamma_{n+1}}\mathbf{1}_{\{U_{n+1}\le X_k\}} = +\infty.$$

Next, we prove that $\limsup_n \dfrac{Y_n}{\log S_n} = +\infty\ a.s.$ One may write $Y_n = x + \sum_{k=0}^{n-1}\Delta_{k+1}\mathbf{1}_{\{U_{k+1}\le \frac{Y_k}{S_k}\}}$ so that for any $\lambda > 0$,

$$\limsup_n \frac{Y_n}{\log S_n} \ge \limsup_n \frac{Z_n^\lambda}{\log S_n}$$

where $$Z_n^\lambda = \sum_{k=0}^{n-1}\Delta_{k+1}\mathbf{1}_{\{U_{k+1}\le \frac{\lambda}{S_k}\}}.$$

We have $$\mathbb{E}\,Z_n^\lambda = \sum_{k=0}^{n-1}\Delta_{k+1}\min(1,\frac{\lambda}{S_k})$$

and
$$\begin{aligned}
\mathrm{Var}(Z_n^\lambda) &= \sum_{k=0}^{n-1}\Delta_{k+1}^2\min(1,\frac{\lambda}{S_k})\Big(1-\min(1,\frac{\lambda}{S_k})\Big)\\
&\le C\,\log S_n\sum_{k=0}^{n-1}\Delta_{k+1}\min(1,\frac{\lambda}{S_k})\\
&= C\,\log S_n\,\mathbb{E}\,Z_n^\lambda.
\end{aligned}$$

Consequently,

$$\mathbb{P}(|Z_n^\lambda - \mathbb{E}\,Z_n^\lambda| \ge \rho\,\mathbb{E}\,Z_n^\lambda) \le C\,\frac{\log S_n\,\mathbb{E}\,Z_n^\lambda}{\rho^2(\mathbb{E}\,Z_n^\lambda)^2} \le C\,\frac{\log S_n}{\rho^2\sum_{k=0}^{n-1}\Delta_{k+1}\min(1,\lambda/S_k)}.$$

One checks that $\lim_n \dfrac{\sum_{k=0}^{n-1}\Delta_{k+1}\min(1,\lambda/S_k)}{\log S_n} = \lambda$ since $\Delta_n\min(1,\lambda/S_{n-1}) \sim \lambda\dfrac{\gamma_n}{1-\gamma_n} \sim \lambda\,\gamma_n$.

Let $A_n^\lambda = \{|Z_n^\lambda - \mathbb{E}\,Z_n^\lambda| < \rho\,\mathbb{E}\,Z_n^\lambda\}$. For λ large enough, $\mathbb{P}(A_n^\lambda) \ge 1/2$, so that $\mathbb{P}(\limsup_n A_n^\lambda) \ge 1/2$. Now, on the event $\limsup_n A_n^\lambda$,

$$Z_n^\lambda \ge (1-\rho)\mathbb{E}\,Z_n^\lambda \ge \lambda(1-\rho)\log S_n \qquad \text{for infinitely many } n.$$

Hence $\mathbb{P}\left(\limsup_n \frac{Z_n^\lambda}{\log S_n} \geq \lambda(1-\rho)\right) \geq 1/2$. But the random variable $\limsup_n \frac{Z_n^\lambda}{\log S_n}$ lies in the asymptotic σ-field of the i.i.d. random variables U_n's, hence

$$\limsup_n \frac{Z_n^\lambda}{\log S_n} \geq \lambda(1-\rho) \qquad \mathbb{P}_x\text{-}a.s.$$

This holds for every $\rho > 0$ and $\lambda > 0$ so that $\limsup_n \frac{Y_n}{\log S_n} = +\infty \; \mathbb{P}_x$-$a.s.$.

On the other hand, for any positive integer p,

$$\begin{aligned}
\mathbb{E}\left((X_\infty - X_p)^2 \mid \mathcal{F}_p\right) &= \mathbb{E}\left(\sum_{k=p}^\infty \gamma_{k+1}^2 X_k(1-X_k) \mid \mathcal{F}_p\right)\\
&\leq \mathbb{E}\left(\sum_{k=p}^\infty \gamma_{k+1}^2 X_k \mid \mathcal{F}_p\right) = X_p \sum_{k=p}^\infty \gamma_{k+1}^2.
\end{aligned}$$

Now observe that

$$\begin{aligned}
\mathbb{P}\left(X_\infty = 0 \mid \mathcal{F}_p\right) &= \frac{\mathbb{E}\left(\mathbf{1}_{\{X_\infty=0\}} X_p^2 \mid \mathcal{F}_p\right)}{X_p^2} \leq \frac{\mathbb{E}\left((X_\infty - X_p)^2 \mid \mathcal{F}_p\right)}{X_p^2}\\
&\leq \frac{\sum_{k=p}^\infty \gamma_{k+1}^2}{X_p} = S_p \frac{\sum_{k=p}^\infty \gamma_{k+1}^2}{Y_p}\\
&\leq C\,\frac{S_p}{Y_p} \sum_{k\geq p+1} \frac{\Gamma_k}{S_k^2}\Delta_k \leq C\,\frac{S_p}{Y_p}\int_{S_p}^{+\infty} \frac{\log u}{u^2}du\\
&\leq C\,\frac{S_p}{Y_p} \times \frac{\log S_p}{S_p} = C\,\frac{\log S_p}{Y_p}.
\end{aligned}$$

One concludes by noting that the bounded martingale $\mathbb{P}_x(X_\infty = 0 \mid \mathcal{F}_p)$ $\mathbb{P}_x$-$a.s.$ converges to $\mathbf{1}_{\{X_\infty=0\}}$ so that

$$\mathbb{P}\left(X_\infty = 0\right) = \lim_p \mathbb{E}\left(\mathbb{P}\left(X_\infty = 0 \mid \mathcal{F}_p\right)\right) \leq C\, \liminf_p \frac{\log S_p}{Y_p} = 0. \quad \diamond$$

Lemma 2 *Assume* $0 < p_A = p_B < 1$ *and Assumption (4). Then, for every* $x \in (0,1]$,

$$\mathbb{P}_x\left(X_\infty = 0\right) = 0. \tag{16}$$

Proof of Lemma 2: In that case, the algorithm can be written as follows (see the specification of the algorithm in (7)).

$$X_{n+1} = X_n + \gamma_{n+1}\mathbf{1}_{B_{n+1}}\left(\mathbf{1}_{\{U_{n+1}\leq X_n\}} - X_n\right), \quad \text{with} \quad B_{n+1} = \{V_{n+1} \leq p_B\}.$$

By conditioning on the σ-field generated by the events B_n, $n \geq 1$, we easily deduce from Lemma 1 that if the sequences $(\Delta_n^B)_{n\in\mathbb{N}}$, defined by $\Delta_n^B = \gamma_n \mathbf{1}_{B_n}/\prod_{k=1}^n (1-\gamma_k \mathbf{1}_{B_k})$ and $\gamma_n^B = \gamma_n \mathbf{1}_{B_n}$ satisfy assumption (12) in Lemma 1, the announced result is proved.

Now, for $n \geq 1$, define

$$M_n = \sum_{k=1}^n \log\left(1-\gamma_k \mathbf{1}_{B_k}\right) - p_B \log(1-\gamma_k) = \sum_{k=1}^n \log(1-\gamma_k)\left(\mathbf{1}_{B_k} - p_B\right).$$

The sequence (M_n) is a martingale and $\sup_n \mathbb{E}\, M_n^2 < +\infty$ because $\sum_n \gamma_n^2 < \infty$ (as follows from the remark below Proposition 5). Therefore, the ratio

$$\frac{\prod_{k=1}^n (1-\gamma_k)^{p_B}}{\prod_{k=1}^n (1-\gamma_k \mathbf{1}_{B_k})} \quad \text{is } a.s. \text{ bounded.}$$

Consequently, there exists a $\sigma(B_n, n \geq 1)$-measurable random positive constant ξ such that, $\mathbb{P}_x$-$a.s.$,

$$\Delta_n^B \leq \xi \frac{\gamma_n^B}{\prod_{k=1}^n (1-\gamma_k)^{p_B}} = \xi\, \gamma_n^B S_n^{p_B} \leq \xi\, \gamma_n S_n^{p_B}.$$

Inequality (14) and $\sum_{n\geq 1} \gamma_n^2 < +\infty$ imply that $S_n \leq C e^{\Gamma_n}$. In turn, Assumption (4) yields

$$\Delta_n^B \leq \xi\, \Gamma_n e^{-p_B \Gamma_n} (e^{\Gamma_n})^{p_B} \leq \xi\, \Gamma_n.$$

Now, a straightforward martingale argument shows that, $a.s.$,

$$\sum_{k=1}^n \gamma_k^B \sim p_B \sum_{k=1}^n \gamma_k$$

so that $\Delta_n^B = O(\gamma_1^B + \cdots + \gamma_n^B)$ and the proposition follows from Lemma 1. $\diamond$

Proof of Proposition 5: Using Proposition 3 (pathwise comparison result), one may assume without loss of generality that $p_A = p_B > 0$. Lemma 2 completes the proof. $\diamond$

4 The two-armed-bandit algorithm as a generalized Polya urn

In this section, we propose two proofs of Proposition 5 – in some partial cases – in which the martingale case is based on methods directly inspired by the Polya urn.

4.1 Short background on Polya urn

Assume that an urn contains at time 0, r red balls and b black balls. At every time n one draws at random a ball from the urn and then puts it back in the urn with another ball of the same colour. Then, at every time n, the urn contains (once the new ball has been put in the urn) exactly $r+b+n$ balls. Let B_n denote the number of black balls inside the urn at time n, let $X_n := \frac{B_n}{r+b+n}$ denote the proportion of black balls at time n and $Y_n := \frac{B_n}{r+b}$. One models the drawings using a sequence $(U_n)_{n\geq 1}$ of i.i.d. random variables uniformly distributed over $[0,1]$ as follows: if $U_{n+1} \leq X_n$ the ball drawn at time $n+1$ is black, otherwise it is red. Then, these sequences satisfy respectively

$$
\begin{aligned}
B_0 &:= b \quad \text{and} \quad B_{n+1} = B_n + \mathbf{1}_{\{U_{n+1}\leq X_n\}},\\
Y_0 &:= \frac{b}{r+b} \quad \text{and} \quad Y_{n+1} = Y_n + \frac{1}{r+b}\mathbf{1}_{\{U_{n+1}\leq X_n\}},\\
X_0 &:= \frac{b}{r+b} \quad \text{and} \quad X_{n+1} = X_n + \frac{1}{r+b+n+1}(\mathbf{1}_{\{U_{n+1}\leq X_n\}} - X_n).
\end{aligned}
$$

Consequently, the regular Polya urn appears as a special case of the two-armed-bandit algorithm (in the martingale setting $p_A = p_B = 1$) corresponding to a rational starting value $X_0 = \frac{b}{r+b}$ and a step $\gamma_n := \frac{1}{r+b+n}$ *i.e.* $\Delta_n = \frac{1}{r+b}$.

This suggests to try extending some classical methods of proof devised for the Polya urn to solve the martingale case of the two-armed-bandit algorithm, with the hope, in some cases, to get more accurate results, *e.g.* concerning the distribution of the limit X_∞.

4.2 The moment approach

Following another classical method devised to solve the Polya urn (see *e.g.* [1]), it is possible to obtain some moment estimates for the limiting distribution of the X_n's in the martingale case $p_A = p_B = 1$. When $\Delta_n = \Delta > 0$, this limiting distribution is even explicit.

Proposition 6 *Assume that $p_A = p_B = 1$ and that the sequence $(\Delta_n)_{n\geq 1}$ is nonincreasing (Δ_1 may be greater than $\Delta_0 = 1$).*

(a) For every $x \in [0,1]$ and for every integer $m \geq 1$,

$$
\mathbb{E}_x(X_\infty^{m+1}) \leq \prod_{k=0}^{m}\left(1 - \frac{1-x}{S_k}\right) \quad \textit{and} \quad \mathbb{E}_x((1-X_\infty)^{m+1}) \leq \prod_{k=0}^{m}\left(1 - \frac{x}{S_k}\right). \tag{17}
$$

In particular, for every $x \in (0,1)$,

$$
\mathbb{P}_x(X_\infty = 1) \leq \inf_m \mathbb{E}_x(X_\infty^m) = 0 \quad \textit{and} \quad \mathbb{P}_x(X_\infty = 0) \leq \inf_m \mathbb{E}_x((1-X_\infty)^m) = 0
$$

since $\displaystyle\sum_{k\geq 1}\frac{1}{S_k} \geq \sum_{k\geq 1}\frac{1}{1+k\Delta_1} = +\infty.$

(b) *If, moreover* $\Delta_n = \Delta > 0,\ n \geq 1$ *(and* $\Delta_0 = 1$*), then* $\gamma_n = \dfrac{\Delta}{n\,\Delta + 1}$ *and*

$$X_\infty \stackrel{\mathcal{L}}{\sim} \beta\left(\frac{x}{\Delta}; \frac{1-x}{\Delta}\right).$$

Remark: Item (b) is a classical result about Polya urn.

Proof: (a) One uses the notation $Y_n = S_n X_n$ of Lemma 1 and sets

$$Z_n^{(m)} := \frac{Y_n}{S_n} \frac{Y_n + \Delta_{n+1}}{S_{n+1}} \times \cdots \times \frac{Y_n + \Delta_{n+1} + \cdots + \Delta_{n+m}}{S_{n+m}}.$$

$$\begin{aligned}
\mathbb{E}_x(Z_{n+1}^{(m)}/\mathcal{F}_n) &= \mathbb{E}_x(Z_{n+1}^{(m)} \mathbf{1}_{\{U_{n+1} > X_n\}}/\mathcal{F}_n) + \mathbb{E}_x(Z_{n+1}^{(m)} \mathbf{1}_{\{U_{n+1} \leq X_n\}}/\mathcal{F}_n) \\
&= \frac{Y_n}{S_{n+1}} \frac{Y_n + \Delta_{n+2}}{S_{n+2}} \times \cdots \times \frac{Y_n + \Delta_{n+2} + \cdots + \Delta_{n+m+1}}{S_{n+m+1}} (1 - X_n) \\
&\quad + \frac{Y_n + \Delta_{n+1}}{S_{n+1}} \frac{Y_n + \Delta_{n+1} + \Delta_{n+2}}{S_{n+2}} \times \cdots \times \frac{Y_n + \Delta_{n+1} + \cdots + \Delta_{n+m+1}}{S_{n+m+1}} X_n
\end{aligned}$$

Recall that $X_n = \dfrac{Y_n}{S_n}$ and $1 - X_n = \dfrac{S_n - Y_n}{S_n}$. Hence

$$\begin{aligned}
\mathbb{E}_x(Z_{n+1}^{(m)}/\mathcal{F}_n) &= \frac{S_n - Y_n}{S_{n+m+1}} \times \underbrace{\frac{Y_n}{S_n} \frac{Y_n + \Delta_{n+2}}{S_{n+1}} \times \cdots \times \frac{Y_n + \Delta_{n+2} + \cdots + \Delta_{n+m+1}}{S_{n+m}}}_{\leq Z_n^{(m)} \text{ since } \Delta_{n+i+1} \leq \Delta_{n+i}} \\
&\quad + \frac{Y_n + \Delta_{n+1} + \cdots + \Delta_{n+m+1}}{S_{n+m+1}} \times \underbrace{\frac{Y_n}{S_n} \frac{Y_n + \Delta_{n+1}}{S_{n+1}} \times \cdots \times \frac{Y_n + \Delta_{n+1} + \cdots + \Delta_{n+m}}{S_{n+m}}}_{Z_n^{(m)}} \\
&\leq Z_n^{(m)}.
\end{aligned}$$

The sequence $(Z_n^{(m)})_{n\geq 0}$ is then a super-martingale. On the other hand $Z_n^{(m)}$ obviously converges toward X_∞^{m+1} since, for every $k \leq m$, $\frac{\Delta_{n+k}}{S_{n+m}} \leq \gamma_{n+m} \to 0$ as $n \to +\infty$ because the sequence $(\Delta_n)_{n\geq 1}$ is bounded and $S_n \uparrow +\infty$. Consequently, for instance *via* Fatou's Lemma,

$$\begin{aligned}
\forall\, m \in \mathbb{N}, \quad \mathbb{E}_x(X_\infty^{m+1}) \leq Z_0^{(m)} &= x \prod_{k=1}^{m} \frac{x + \Delta_1 + \cdots + \Delta_k}{1 + \Delta_1 + \cdots + \Delta_k} = x \prod_{k=1}^{m} \frac{x - 1 + S_k}{S_k} \\
&= \prod_{k=0}^{m} \left(1 - \frac{1-x}{S_k}\right).
\end{aligned}$$

One proceeds symmetrically with $\widetilde{X}_n := 1 - X_n$ and $\widetilde{Y}_n = S_n - Y_n$ to establish the moment inequalities concerning $1 - X_\infty$.

Hence, the Lebesgue dominated convergence Theorem implies that when $m \to \infty$,

$$\mathbb{P}_x(X_\infty = 0) = \lim_m \mathbb{E}_x((1-X_\infty)^{m+1}) \le \prod_{k\ge 0}\left(1-\frac{x}{S_k}\right) = 0$$

since $\displaystyle\sum_{n\ge 0}\frac{1}{S_n} \ge \sum_{n\ge 0}\frac{1}{1+n\Delta_1} = +\infty$.

(b) When $\Delta_n = \Delta$, $n \ge 1$, the same proof shows that $(Z_n^{(m)})_{n\ge 0}$ is a martingale, hence, for every $m \ge 0$,

$$\mathbb{E}_x(X_\infty^{m+1}) = \prod_{k=0}^{m}\left(1-\frac{1-x}{1+k\Delta}\right) = \prod_{k=0}^{m}\frac{x/\Delta + k}{(1-x)/\Delta + x/\Delta + k}$$

hence X_∞ has the moments of a $\beta\,(x/\Delta;(1-x)/\Delta)$ distribution. Both distributions have compact support hence they are equal. $\diamond$

The above result can be extended to the general martingale case.

Corollary 3 *If $p_A = p_B \in (0,1]$ and the sequence $(\Delta_n)_{n\ge 1}$ is nonincreasing (Δ_1 may be greater than $\Delta_0 = 1$). Then, for every $x \in (0,1)$ $\mathbb{P}_x(X_\infty = 1) = \mathbb{P}_x(X_\infty = 0) = 0$ and, for every $m \in \mathbb{N}$,*

$$\mathbb{E}_x(X_\infty^{m+1}) \le \prod_{k=0}^{m}\left(1-(1-x)\prod_{\ell=1}^{k}(1-\gamma_\ell)\right) \tag{18}$$

$$\mathbb{E}_x((1-X_\infty)^{m+1}) \le \prod_{k=0}^{m}\left(1-x\prod_{\ell=1}^{k}(1-\gamma_\ell)\right). \tag{19}$$

Proof: Dealing with the case $p := p_A = p_B < 1$ still needs conditioning with respect to the σ-algebra generated by the events B_n. However we will proceed slightly differently than in the proof of Proposition 5. We introduce the successive stopping times

$$\forall\, \omega \in \Omega, \quad \tau_0^B(\omega) := 0, \quad \tau_n^B(\omega) := \min\{k > \tau_{n-1}^B \,/\, \omega \in B_k\},\ n \ge 1.$$

The τ_n's are $\mathbb{P}_x$-*a.s.* finite iff $p_B > 0$. Then, set $\widetilde{\gamma}_n := \gamma_{\tau_n^B}$ and $\widetilde{\Delta}_n$ and $\widetilde{S}_n := 1 + \widetilde{\Delta}_1 + \cdots + \widetilde{\Delta}_n$, $n \ge 1$, like in Lemma 1 so that $\widetilde{\gamma}_n = \widetilde{\Delta}_n/\widetilde{S}_n$. One checks that

$$\frac{\widetilde{\Delta}_{n+1}}{\widetilde{\Delta}_n} = \frac{\Delta_{\tau_{n+1}^B}}{\Delta_{\tau_n^B}}\prod_{k=\tau_n^B+1}^{\tau_{n+1}^B-1}(1-\gamma_k) \le \frac{\Delta_{\tau_{n+1}^B}}{\Delta_{\tau_n^B}},$$

so that $\widetilde{\Delta}_n$ is nonincreasing as long as Δ_n is. It follows from Proposition 6 and obvious equalities that, for every $m \ge 0$,

$$\mathbb{E}_x(X_\infty^{m+1}) \le \mathbb{E}_x\left(\prod_{k=0}^{m}\left(1-\frac{1-x}{\widetilde{S}_k}\right)\right) \quad \text{and} \quad \mathbb{E}_x((1-X_\infty)^{m+1}) \le \mathbb{E}_x\left(\prod_{k=0}^{m}\left(1-\frac{x}{\widetilde{S}_k}\right)\right).$$

Now $\widetilde{S}_n \leq 1 + n\widetilde{\Delta}_1$, $n \geq 1$. Then, proceeding as in the proof of Proposition 6(a) yields

$$\mathbb{P}_x(X_\infty = 0) = \lim_m \mathbb{E}_x((1 - X_\infty)^{m+1}) \leq \prod_{k \geq 0} \left(1 - \frac{x}{1 + k\Delta_1}\right) = 0.$$

One gets similarly that $\mathbb{P}_x(X_\infty = 1) = 0$. The moment bounds follow from the easy fact that $\widetilde{\Delta}_n \leq \Delta_n$ so that $\widetilde{S}_n \leq S_n = (\prod_{k=1}^n (1 - \gamma_k))^{-1}$. $\diamond$

Remark: The above bounds (18) and (19) do not involve $p_{_B}$. In fact they can be improved by replacing γ_n by $\gamma_{\tau_n^B}$ in their right hand and taking the expectation with respect to $\mathbb{E}_x$. Then, one may use that $\tau_n^B - \tau_{n-1}^B$ is i.i.d. with Geometric distribution $G(p_{_B})$.

4.3 The *log*-martingale approach

The classical *log*-martingale approach to Polya urn yields a new proof of Lemma 1 when the sequence $(\Delta_n)_{n \geq 1}$ is bounded. We use the same notations as in the original lemma.

Proposition 7 *Assume $p_{_A} = p_{_B} = 1$ and $(\Delta_n)_{n \geq 1}$ is bounded.*
(a) Then, there exists a martingale $(M_n)_{n \geq 1}$ with bounded increments such that, for every $x \in (0, 1)$,

$$\sup_n \left| \log\left(\frac{X_n}{1 - X_n}\right) - M_n \right| < +\infty \quad \mathbb{P}_x\text{-}a.s.$$

(b) Consequently, for every $x \in (0, 1)$, $\mathbb{P}_x(X_\infty = 0) = \mathbb{P}_x(X_\infty = 1) = 0$.

Proof: (a) Set $Z_n := \log\left(\dfrac{X_n}{1 - X_n}\right) = \log\left(\dfrac{Y_n}{S_n - Y_n}\right) = \log Y_n - \log(S_n - Y_n)$.

$$\begin{aligned}\left|\log Y_n - \left(\log x + \sum_{k=1}^n \frac{\Delta Y_k}{Y_{k-1}}\right)\right| &\leq \sum_{k=1}^n \left|\log\left(\frac{Y_k}{Y_{k-1}}\right) - \frac{\Delta Y_k}{Y_{k-1}}\right| \leq \frac{1}{2} \sum_{k=1}^n \left(\frac{\Delta Y_k}{Y_{k-1}}\right)^2 \\ &\leq \frac{\sup_n \Delta_n}{2} \sum_{k=1}^n \frac{\Delta Y_k}{Y_{k-1}^2} \leq \frac{c^2 \sup_n \Delta_n}{2} \sum_{k \geq 1} \frac{\Delta Y_k}{Y_k^2} < +\infty\end{aligned}$$

where $c := 1 + \sup_n \Delta_n / x$ satisfies $Y_k / Y_{k-1} \leq 1 + \Delta_k / Y_{k-1} \leq c$. Similarly

$$\left|\log(S_n - Y_n) - \left(\log(1 - x) + \sum_{k=1}^n \frac{\Delta_k - \Delta Y_k}{S_{k-1} - Y_{k-1}}\right)\right| \leq \frac{\sup_n \Delta_n}{2} \sum_{k \geq 1} \frac{(\Delta_k - \Delta Y_k)}{(S_{k-1} - Y_{k-1})^2} < +\infty.$$

Combining these inequalities yields

$$\sup_n \left| Z_n - \sum_{k=1}^n \left(\frac{\Delta Y_k}{Y_{k-1}} - \frac{\Delta_k - \Delta Y_k}{S_{k-1} - Y_{k-1}}\right)\right| < +\infty \quad \mathbb{P}_x\text{-}a.s.$$

Now $M_n := \sum_{k=1}^{n} \left(\frac{\Delta Y_k}{Y_{k-1}} - \frac{\Delta_k - \Delta Y_k}{S_{k-1} - Y_{k-1}}\right)$ is a martingale since

$$\begin{aligned}
\mathbb{E}_x(\Delta M_n/\mathcal{F}_{n-1}) &= \mathbb{E}_x(\Delta Y_n/\mathcal{F}_{n-1})\frac{1}{Y_{n-1}} - (\Delta_n - \mathbb{E}_x(\Delta Y_n/\mathcal{F}_{n-1}))\frac{1}{S_{n-1} - Y_{n-1}} \\
&= \frac{X_{n-1}\Delta_n}{Y_{n-1}} - \frac{(1 - X_{n-1})\Delta_n}{S_{n-1} - Y_{n-1}} = \frac{\Delta_n}{S_{n-1}} - \frac{\Delta_n}{S_{n-1}} = 0.
\end{aligned}$$

Furthermore its increments are bounded. As a matter of fact ΔY_k and $\Delta_k - \Delta Y_k$ are never simultaneously zero and are upper-bounded by Δ_k ; on the other hand Y_{k-1} and $S_{k-1} - Y_{k-1}$ are lower bounded by x and $1 - x$ respectively, hence, for every $k \geq 1$,

$$|\Delta M_k| \leq \frac{\Delta_k}{Y_{k-1} \wedge (S_{k-1} - Y_{k-1})} \leq \frac{\|\Delta\|_\infty}{x \wedge (1 - x)}.$$

(b) The Law of Iterated Logarithm for martingales with bounded increments says that, on the event $\{< M >_\infty = +\infty\}$, the martingale (M_n) satisfies $\liminf_n M_n = -\infty$ and $\limsup_n M_n = +\infty$ *a.s.* Meanwhile Z_n converges toward $\log\left(\frac{X_\infty}{1 - X_\infty}\right) \in \overline{\mathbb{R}}$ *a.s.* The difference of these two quantities remaining bounded, it follows that $< M >_\infty < +\infty$ $\mathbb{P}_x$-*a.s.* Hence, the martingale M_n converges toward a finite limit and, consequently, $X_\infty \in (0, 1)$ $\mathbb{P}_x$-*a.s.* $\diamondsuit$

Remark: The above assumption does not embody the Power step (III) setting of Corollary 2(b) ($p_A = p_B = 1$ and) $\Delta_n \sim C \log n$ *i.e.* the closest case to the critical case that we can get.

The extension to general p_A and p_B, $0 < p_B \leq p_A$, in that framework consists in proving that the sequence $(\Delta_n^B)_{n\geq 1}$ is *a.s.* bounded. One shows using martingale methods of Lemma 2 that this leads to the condition $\gamma_n = O\left(e^{-p_B \Gamma_n}\right)$ which is, as expected, more stringent than Assumption (4).

5 Additional results

• Regularity of $x \mapsto \mathbb{P}_x(X_\infty = 1)$ when $p_A > p_B$

One can obtain some regularity results for the function $x \mapsto \mathbb{P}_x(X_\infty = 1)$ as soon as $p_A \neq p_B$ (keep in mind that in that setting, X_∞ is $\{0, 1\}$-valued). Namely

Proposition 8 *If $p_A > p_B$, the function $x \mapsto \mathbb{P}_x(X_\infty = 1)$ is nondecreasing and analytic on* $(0, 1]$.

Proof: The only point to establish is analyticity. One starts from the second equality in (10) and the tools developed in the proof of Proposition 2. We also adopt the same notations. Indeed, function χ is analytic on $(0, 1]$ since it is an absolutely decreasing function. Then, ψ_γ is analytic as well and consequently so is $x \mapsto \mathbb{P}_x(X_n \to 1)$. $\diamondsuit$

• About the distribution of X_∞

When $0 < p_A = p_B \leq 1$ and $\sum_{n\geq 1} \gamma_n^2 < +\infty$, the conditional distribution of X_∞ given $\{X_\infty \neq 0, 1\}$ is continuous. This follows from Theorem 3.IV.13 in [5].

• Still open questions...

The main open question is of course to find a necessary and sufficient condition for the algorithm to be *a.s.* infallible. For example, when $p_A = p_B = 1$, Assumption (3) ensuring that the algorithm fails with a positive probability is easier to express using the partial sums S_n of the Δ_n's by

$$\sum_{n\geq 1} \frac{1}{S_n} < +\infty. \tag{20}$$

If $\Delta_n = \log n \, \log_2^\beta n$, Assumption (20) is equivalent to $\beta > 1$ whereas $\Delta_n = O(\Gamma_n)$ in Lemma 1 reads $\beta = 0$. So we are facing a $\log\log$ problem.

Furthermore, it follows from the Borel-Cantelli Lemma for independent events that

$$\limsup_n \frac{Y_n}{\Delta_n} \geq \limsup_n \mathbf{1}_{\{U_n \geq x/S_{n-1}\}} = 1 \quad \mathbb{P}_x\text{-}a.s.$$

when $\sum_{n\geq 1} 1/S_n = +\infty$.

It is to be noticed that, when $\Delta_n = \log n \, \log_2^\beta n$ for some $\beta \in (0,1)$, this straightforwardly implies that $\limsup_n Y_n / \log S_n = +\infty$ (which was the key step of Lemma 1). Unfortunately, for such sequences Δ_n, Lemma 1 only implies that, $\mathbb{P}_x(X_\infty = 0) = 0$ iff $\limsup_n \frac{Y_n}{\Delta_n} > 1$ $\mathbb{P}_x$-*a.s.*

• A last remark:

Let $\sigma_n := \min\{k > \sigma_{n-1} \,/\, U_k < X_{k-1}\}$ and $\sigma_0 := 0$ denote the increasing break times, Assumption (20) is equivalent to $\mathbb{P}_x(\sigma_1 = +\infty) > 0$ for every $x \in (0,1)$.

Otherwise, all the σ_n's are $\mathbb{P}_x$-*a.s.* finite for every $x \in (0,1)$: this follows from the expression of $\mathbb{P}_x(\sigma_1 \geq k)$ and from the Markov property.

Acknowledgement: The authors are grateful to Jean-Claude Fort for stimulating discussions.

References

[1] P. Baldi, L. Mazliak, P. Priouret (2000), *Martingales et chaînes de Markov*, Hermann, 2è édition, Paris, 222p. or (2002) *Martingales and Markov chains*, Chapman & Hall/CRC, London, 192p.

[2] M. Benaïm, M.W. Hirsch (1996), Asymptotic pseudo-trajectories, chain recurrent flows and stochastic approximations, *J. Dynamics and Diff. Equations*, **8**, n^0 1, pp. 141-176.

[3] O. Brandière, M. Duflo (1996), Les algorithmes stochastiques contournent-ils les pièges ?, *Ann. Inst. H. Poincaré Probab. Statist.*, **32**, pp. 395-427.

[4] D. Dacunha-Castelle, M. Duflo (1983), *Probabilités et Statistiques, 2. Problèmes à temps mobile*, Masson, Paris.

[5] M. Duflo (1996), *Algorithmes stochastiques*, coll. Mathématiques & Applications, **23**, Springer-Verlag, Berlin, 319p.

[6] J.C. Fort, G. Pagès (2002), Decreasing step stochastic algorithms: *a.s.* behaviour of weighted empirical measures, *Monte Carlo Methods and Applications*, **8**, $n^0 3$, pp. 221-320.

[7] P. Hall, C.C. Heyde (1980), *Martingale Limit Theory and its Application*, Academic Press, New York, 1980, 308p.

[8] H.J. Kushner, D.S. Clark (1978), Stochastic Approximation for Constrained and Unconstrained Systems, *Applied Math. Science Series*, **26**, Springer.

[9] V.A. Lazarev (1992), Convergence of stochastic approximation procedures in the case of a regression equation with several roots, (transl. from) *Problemy Pederachi Informatsii*, vol. **28**, $\mathrm{n}^0 1$.

[10] K.S. Narendra, M.A.L. Thathachar (1974) Learning Automata - A survey, *IEEE Trans. Systems, Man., Cybernetics*, S.M.C-4, pp. 323-334.

[11] K.S. Narendra, M.A.L. Thathachar (1989) *Learning Automata - An introduction*, Prentice Hall, Englewood Cliffs, NJ, 476p.

[12] M. Niang (1999), *Algorithme de Narendra et application à l'allocation d'actifs*, Rapport de stage de DEA, Olympia Capital Management et Univ. Marne-la-vallée (France), 51p.

[13] M.F. Norman (1968), On linear Models with Two Absorbing Barriers, *J. of Mathematical Psychlogy*, **5**, pp.225-241.

[14] R. Pemantle (1990), Nonconvergence to unstable points in urn models and stochastic approximations, *Annals of Probability*, **18**, n^0 2, pp. 698-712.

[15] I.J. Shapiro, K.S. Narendra, Use of Stochastic Automata for Parameter Self-Optimization with Multi-Modal Perfomance Criteria, *IEEE Trans. Syst. Sci. and Cybern.*, SSC-5, pp.352-360.

[16] P. Tarrès, Pièges répulsifs, *C.R.A.S. Acad. Sc. de Paris*, Série I, **330**, pp.125-130, 2000.

[17] P. Tarrès, *Algorithmes stochastiques et marches aléatoires renforcées*, Thèse de l'ENS Cachan (France), novembre 2001.