Can you trust a bandit with your money?

Master M2A - Sorbonne Universite - Stochastic Optimization

Nicolas Olivain

Abstract: **Asset Allocation** is one of many problems where the opportunity cost is high. Evaluating each solution in a real setup first before making a decision can often lead to significant losses. The **stochastic optimization** framework allows for a dynamic reallocation of the resources, thus diminishing the cost of the experiment. However, this procedure is by essence uncertain, how can we be sure that it will chose wisely ? In [10] and [9], Damien Lamberton, Gilles Pagès and Pierre Tarrès investigated the asymptotic convergence and speed of a **stochastic algorithm** applied to solve a **two-armed bandit problem**. They tackle the challenge of designing an **infallible** bandit algorithm and present both convergence and speed related results. This paper was written as part of the course *Stochastic Optimization* directed by Professors A. Godichon-Baggioni and A. Guyader at Sorbonne Universite.

Key-words: Two-armed Bandits, Stochastic Algorithm, Online Optimization, Asset Allocation

1. Introduction

1.1. The bandit problem

The bandit problem is a well known problem in optimization and reinforcement learning, where limited resources must be allocated between different competing alternatives. Imagine you are in a casino, you have access to N slot machines with different expected rewards. You obviously want to maximize your profits, but you do not know which option has the highest odds in your favor and you have a finite budget. How do you spend your hard earned money ?
A trivial solution would be to split your budget evenly between each machine, but this is far from being optimal, as you could loose a lot on low-reward machines. A slightly better idea would be to test each machine evenly with for instance 25% of your budget to estimate the odds of each alternatives. Then you could go all-in with the remaining 75% on the presumed highest rewarding one. This approach is interesting as it introduces a trade-off between exploration: how much do you allocate to estimate your odds of success, and exploitation: use what you learnt from exploration to maximize your profits. If do not explore enough, you can take the wrong decision and spend your budget on a sub-optimal slot machine. On the other hand, if you explore too much, you do not have any money remaining to exploit after you've made your decision and you're back to the trivial split.
This simple example demonstrate the opportunity cost of exploration in the bandit problem. While many different strategies exist to approximate a solution to this problem, we will here consider one from the Stochastic Optimization realm: using a stochastic algorithm, we will dynamically reallocate our funds in an online fashion and try to minimize the opportunity cost.

1.2. Notations

In this paper, we consider a two-armed bandit problem applied to asset allocation on a stock exchange. We assume that we have two trading agents A and B following different strategies. As a fund manager, we have to decide the fraction of our capital that each agent will be managing to maximize our profits. We denote as X_n the fraction of the capital managed by agent A at date n. As a result agent B manages $1 - X_n$.
However we cannot just trust our two traders with the funds, we need a way to evaluate their performance. Thus we introduce the Bernoulli random variable A_n (respectively B_n) which is 1 if the trader A (respectively B) is satisfactory at date n and 0 otherwise. To match the problem described with the slot machines, we assume that A_n and B_n are stationary, meaning the probability for a trader to pass the test at a given day is independent from the past days and is always the same. We can thus introduce $\mathbb{P}(A_n) = p_A$ and $\mathbb{P}(B_n) = p_B$. We also introduce the difference $\pi = p_A - p_B$.

The problem that we consider here is symmetric in A and B, so without loss of generality, **we will consider from now on that $p_A \geq p_B$, *ie:* $\pi \geq 0$.** That is to say that agent A performs on the whole better according to our criteria than agent B.As a manager, that means we would like to allocate all of our funds to A *ie:* $X = 1$. Thus, we now introduce the natural notions of *fallibility* and *infallibility* as defined in [10].

Definition 1.1. *We say that the bandit is fallible when* $\lim_{n\to\infty} \mathbb{P}(X_n = 0) > 0$. *Meaning that there is a risk that X_n does not converge toward the optimal agent.*

Definition 1.2. *We say that the bandit is infallible when* $\lim_{n\to\infty} \mathbb{P}(X_n = 0) = 0$.

We will denote the positive step size of the algorithm as γ_n and their sum $\Gamma_n = \sum_k^n \gamma_k$. On top of that, we define C as a positive constant and ξ as a random positive constant. Both can be updated from line to line.

1.3. Preliminary results

We introduce here some usual concepts used in stochastic algorithms that will be needed in many proofs below. Given a sequence of random variables $(M_n)_{n\geq 0}$ and the associated filtration $\mathcal{F}_n$ we say that:

Definition 1.3. *If* $\mathbb{E}(M_{n+1}|\mathcal{F}_n) = M_n$ *and M_n is integrable then $(M_n)_{n\geq 0}$ is a* ***martingale****.*

Definition 1.4. *If* $\mathbb{E}(M_{n+1}|\mathcal{F}_n) > M_n$ *and* $\max(0, M_n)$ *is integrable then $(M_n)_{n\geq 0}$ is a* ***sub-martingale****.*

We now introduce two martingale convergence theorems: Doob's theorem and one of its direct consequences for sub-martingales:

Theorem 1.1. *(**Doob's theorem**) Let $(M_n)_{n\geq 0}$ be martingale, sub-martingale or super-martingale such that* $\sup_n \mathbb{E}(|M_n|) < +\infty$. *Then, $(M_n)_{n\geq 0}$ converges almost surely toward an integrable random variable M_∞.*

Theorem 1.2. *Let $(M_n)_{n\geq 0}$ be a sub-martingale majored by a constant M. Then M_n converges toward a random variable M_∞ such that* $\forall n$, $\mathbb{E}(M_\infty|\mathcal{F}_n) \geq M_n$.

2. The two-armed Bandit algorithm

In this section we derive the stochastic algorithm used to estimate X_n. Firstly we need to model a test to know if an agent if satisfactory at a given date. We model the satisfaction of agent A (respectively B) at time n with a random variable V_n^A (respectively V_n^B) over $[0, 1]$.

Definition 2.1. *We say that agent A (respectively B) is satisfactory at date n if* $V_n^A \leq p_A$ *(respectively* $V_n^B \leq p_B$*). As a result we have that* $A_n = \mathbb{1}_{V_n^A \leq p_A}$ *and* $B_n = \mathbb{1}_{V_n^B \leq p_B}$. *We consider here V_n^A and V_n^B to be uniformly distributed over* $[0, 1]$.

At each time step, we want to evaluate one of the two agent randomly to consider a reallocation of the resources. We could uniformly pick A or B for evaluation, however this would lead to higher risks. For instance, A could hold 99% of the fund, but would still have only a 50% chance of being controlled, the same as B holding only 1%. Here the control does not focus on where the risk stands, we want instead to monitor closely the trader holding a lot of assets. Thus, we are going to toss a biased coin rather than a fair one: the traders' probability of being evaluated is proportional to the fund they hold.

Definition 2.2. *Given U_n a uniform random variable over* $[0, 1]$, *we will evaluate agent A at date n if* $U_n \leq X_{n-1}$ *and agent B if* $U_n > X_{n-1}$. *We consider U_n to be independent from* (V_n^A, V_n^B).

Now that we designed the control mechanism for our two agents, we need to introduce the reward. Basically, if trader A is satisfactory at date n, we want to reward it by increasing its share of the fund at $n + 1$. On the contrary, if A is not satisfactory, we do not punish it and nothing happens. We design the reward to be proportional to the assets managed by agent B, thus we obtain that:

$$\begin{aligned} &\text{if } U_n \leq X_n \text{ and } V_n^A \leq p_A \text{ then } X_{n+1} = X_n + \gamma_{n+1}(1 - X_n) \\ &\text{if } U_n > X_n \text{ and } V_n^B \leq p_B \text{ then } 1 - X_{n+1} = 1 - X_n + \gamma_{n+1} X_n \end{aligned} \tag{1}$$

With γ_n the step-size within $]0, 1[$ such that $\sum_n \gamma_n = +\infty$ in order to nullify the impact of the initialization. We can finally derive the stochastic update mechanism of the two-armed bandit algorithm.

Definition 2.3. *Given X_0 in $]0, 1[$, we update the allocation of funds X_n at time $n + 1$ as follows:*

$$X_{n+1} = X_n + \gamma_{n+1} \left(\mathbb{1}_{\{U_{n+1} \leq X_n\} \cap A_{n+1}} (1 - X_n) - \mathbb{1}_{\{U_{n+1} > X_n\} \cap B_{n+1}} X_n \right)$$

From this definition, we can derive the associated filtration to the problem $\mathcal{F}_n = \sigma(U_{1:n}, V_{1:n}^A, V_{1:n}^B, A_{1:n}, B_{1:n})$ and the stochastic part $Z_{n+1} = (U_{n+1}, V_{n+1}^A, V_{n+1}^B)$. Let us now rewrite the algorithm in its canonical form:

$$H(X_n, Z_{n+1}) = \mathbb{1}_{\{U_{n+1} \leq X_n\} \cap A_{n+1}}(1 - X_n) - \mathbb{1}_{\{U_{n+1} > X_n\} \cap B_{n+1}} X_n$$

We can then derive $h(X_n) = \mathbb{E}(H(X_n, Z_{n+1}))$ and the martingale increment $\Delta M_{n+1} = H(X_n, Z_{n+1}) - h(X_n)$. Recalling that U_n is independent from V_n^A and V_n^B, we get that:

$$\begin{aligned} h(X_n) &= \mathbb{E}(\mathbb{1}_{\{U_{n+1} \leq X_n\} \cap A_{n+1}})(1 - X_n) - \mathbb{E}(\mathbb{1}_{\{U_{n+1} > X_n\} \cap B_{n+1}}) X_n \\ h(X_n) &= \mathbb{P}(U_{n+1} \leq X_n)\mathbb{P}(A_{n+1})(1 - X_n) - \mathbb{P}(U_{n+1} > X_n)\mathbb{P}(B_{n+1}) X_n \\ h(X_n) &= X_n p_A (1 - X_n) - (1 - X_n) p_B X_n \\ h(X_n) &= \pi X_n (1 - X_n) \end{aligned} \tag{2}$$

This let us finally rewrite X_{n+1} in a canonical stochastic algorithm from, with the deterministic h and the stochastic martingale increment ΔM_{n+1}.

$$X_{n+1} = X_n + \gamma_{n+1} \pi X_n (1 - X_n) + \gamma_{n+1} \Delta M_{n+1} \tag{3}$$

3. Convergence of the algorithm

3.1. Asymptotic behaviour

In this section we study the convergence of X_n toward X_∞ and study its properties before tackling the *infallibility* question. We will only consider the case where $p_A > p_B$, meaning that $\pi > 0$.

Lemma 3.1. *Assuming $p_A > p_B$, X_n converges toward a random variable X_∞ valued within $[0, 1]$*

Proof. First let us show that $(X_n)_{n \geq 0}$ is a sub-martingale

$$\begin{aligned} \mathbb{E}(X_{n+1} | \mathcal{F}_n) &= \mathbb{E}(X_n + \gamma_{n+1} \pi X_n (1 - X_n) + \gamma_{n+1} \Delta M_{n+1} | \mathcal{F}_n) \\ \mathbb{E}(X_{n+1} | \mathcal{F}_n) &= X_n + \gamma_{n+1} \pi X_n (1 - X_n) + \mathbb{E}(\gamma_{n+1} \Delta M_{n+1} | \mathcal{F}_n) \\ \mathbb{E}(X_{n+1} | \mathcal{F}_n) &= X_n + \gamma_{n+1} \pi X_n (1 - X_n) \\ \mathbb{E}(X_{n+1} | \mathcal{F}_n) &\geq X_n \qquad \text{as } \pi > 0 \end{aligned} \tag{4}$$

Thus $(X_n)_{n \geq 0}$ is a sub-martingale. Moreover by construction, we have that $X_n \leq 1$. We can then apply Theorem 1.2, as a result we get that X_n converges toward a $[0, 1]$ valued random variable denoted as X_∞. □

Now that we established the convergence of the algorithm, we will try to establish some more precise convergence results regarding to our objective. First we will rewrite X_n using straight forward induction as:

$$X_n = X_0 + \sum_{k=1}^{n} \pi \gamma_k X_{k-1} (1 - X_{k-1}) + \sum_{k=1}^{n} \gamma_k \Delta M_k \tag{5}$$

We then apply the expectation to both side of the equality which gives us:

$$\mathbb{E}(X_n) = X_0 + \pi \mathbb{E}\left(\sum_{k=1}^{n} \gamma_k X_{k-1} (1 - X_{k-1}) \right) + 0 \tag{6}$$

$$\mathbb{E}(X_n) - X_0 = \pi \mathbb{E}\left(\sum_{k=1}^{n} \gamma_k X_{k-1} (1 - X_{k-1}) \right) \tag{7}$$

As $\forall n$, $0 \leq X_n \leq 1$, the right side of the equation is bounded by $[-\frac{X_0}{\pi}, \frac{1 - X_0}{\pi}]$ and as such is finite when $n \to \infty$. Recalling that $\sum_n \gamma_n = +\infty$, this implies that $\lim_{n \to \infty} X_n(1 - X_n) = 0$ almost surely. Moreover, as $X_n \to X_\infty$, we get by continuity of the function $x \leftarrow x(1 - x)$ that $X_\infty(1 - X_\infty) = 0$ almost surely. This result gives us the following theorem:

Theorem 3.1. *Assuming that $p_A > p_B$, X_n converges toward a random variable X_∞ and $\mathbb{P}(X_\infty \in \{0; 1\}) = 1$*

This means that X_n asymptotically converges toward either 0 either 1, giving all of the funds to one agent. However, this theorem does not tell us if it converges toward the best one. We can now picture the problem as making sure the algorithm converges toward the *target*, and avoids the *trap*.

3.2. The equality case

Before tackling the infallibility problem in the case when $p_A > p_B$, we need to first consider the equality case. In this section we introduce important preliminaries proving the infallibility when $p_A = p_B$ under certain conditions. Lemmas 3.3 and 3.4 are **the core proof of the infallibility**, as they will generalise very well in the unequal case (see section 3.3). We begin by introducting some step related sequences.

Definition 3.1. *Given the sequence of steps $(\gamma_n)_{n\geq 1}$, we introduce $\Delta_0 = 1$, $\Delta_n = \frac{\gamma_n}{\prod_{k=1}^{n}(1-\gamma_k)}$ and $S_n = \sum_{k=1}^{n}\Delta_k$. A direct consequence of this definition is the equality $\gamma_n = \frac{\Delta_n}{S_n}$*

Lemma 3.2. $\log S_n - \sum_{k=1}^{n}\frac{\gamma_k^2}{1-\gamma_k} \leq \Gamma_n \leq \log S_n$

Proof. Note that $\Gamma_n = \sum_{k=1}^{n}\frac{\Delta_k}{S_n}$. Using comparison between sums and integral, we can get the following boundaries

$$\sum_{k=1}^{n}(1-\gamma_k)\int_{S_{k-1}}^{S_k}\frac{du}{u} = \sum_{k=1}^{n}\frac{S_{k-1}}{S_k}\int_{S_{k-1}}^{S_k}\frac{du}{u} \leq \Gamma_n \leq \int_1^{S_n}\frac{du}{u} \tag{8}$$

$$\log S_n - \sum_{k=1}^{n}\frac{\gamma_k^2}{1-\gamma_k} \leq \Gamma_n \leq \log S_n \tag{9}$$

□

Lemma 3.3. *Assuming that $p_A = p_B = 1$, if the sequence $(\Delta_n)_{n\geq 0}$ satistifes $\Delta_n = \mathcal{O}(\Gamma_n)$, then $\forall X_0 \in]0,1]$, $\mathbb{P}(X_\infty = 0) = 0$.*

Proof. Assuming $p_A = p_B$, we can rewrite the algorithm as follows:

$$\begin{aligned} S_{n+1}X_{n+1} &= S_{n+1}X_n + \Delta_{n+1}(\mathbb{1}_{U_{n+1}\leq X_n} - X_n) \\ &= S_nX_n + \Delta_{n+1}\mathbb{1}_{U_{n+1}\leq X_n} \end{aligned} \tag{10}$$

Let us define $Y_n = S_nX_n$, we will now derive an upper bound for $\mathbb{P}(X_\infty = 0|\mathcal{F}_n)$ which will hopefully converge to 0. First let us observe that we have for any positive integer p

$$\begin{aligned} \mathbb{E}\left((X_\infty - X_p)^2|\mathcal{F}_n\right) &= \mathbb{E}\left(\sum_{k=p}^{\infty}\gamma_{k+1}^2X_k(1-X_k)|\mathcal{F}_p\right) \\ &\leq \mathbb{E}\left(\sum_{k=p}^{\infty}\gamma_{k+1}^2X_k\right) \\ &= X_p\sum_{k=p}^{\infty}\gamma_{k+1}^2 \end{aligned} \tag{11}$$

Working directly on the probability, we can now derive the following upper bound:

$$\begin{aligned} \mathbb{P}(X_\infty = 0|\mathcal{F}_n) &= \frac{\mathbb{E}(\mathbb{1}_{\{X_\infty=0\}}X_p^2|\mathcal{F}_p)}{X_p^2} \leq \frac{\mathbb{E}((X_\infty - X_p)^2|\mathcal{F}_n)}{X_p^2} \\ &\leq \frac{\sum_{k=p}^{\infty}\gamma_{k+1}^2}{X_p} = S_p\frac{\sum_{k=p}^{\infty}\gamma_{k+1}^2}{Y_p} \\ &\leq C\frac{S_p}{X_p}\sum_{k\geq p+1}\frac{\Gamma_k}{S_k^2}\Delta_k \leq C\frac{S_p}{Y_p}\int_{S_p}^{\infty}\frac{\log u}{u^2}du \\ &\leq C\frac{S_p}{Y_p}^{\frac{\log S_p}{S_p}} = C\frac{\log S_p}{Y_p} \end{aligned} \tag{12}$$

We can now conclude on the upperbound using that the bounded martingale $\mathbb{P}(X_\infty = 0|\mathcal{F}_n)$ converges toward $\mathbb{1}_{\{X_\infty=0\}}$, which gives us:

$$\mathbb{P}(X_\infty = 0) = \lim_p \mathbb{E}(\mathbb{P}(X_\infty = 0|\mathcal{F}_p)) \leq C\liminf_p\frac{\log S_p}{Y_p} \tag{13}$$

Now, we need to prove that $\limsup_n \frac{Y_n}{\log S_n} = +\infty$ for the upper bound and thus the probability to go to 0. First, notice that $Y_n = x + \sum_{k=0}^{n-1} \Delta_{k+1} \mathbb{1}_{U_{k+1} \leq Y_k/S_k}$, so that for any $\lambda > 0$;

$$\limsup_n \frac{Y_n}{\log S_n} \geq \limsup_n \frac{Z_n^\lambda}{\log S_n} \quad \text{where} Z_n^\lambda = \sum_{k=0}^{n-1} \Delta_{k+1} \mathbb{1}_{U_{k+1} \leq \lambda/S_k} \tag{14}$$

We can then derive $\mathbb{E}(Z_n^\lambda)$ and $\mathrm{Var}(Z_n^\lambda)$:

$$\mathbb{E}(Z_n^\lambda) = \sum_{k=0}^{n-1} \Delta_{k+1} \min\left(1, \frac{\lambda}{S_k}\right) \tag{15}$$

$$\begin{aligned} \mathrm{Var}(Z_n^\lambda) &= \sum_{k=0}^{n-1} \Delta_{k+1}^2 \min\left(1, \frac{\lambda}{S_k}\right)\left(1 - \min\left(1, \frac{\lambda}{S_k}\right)\right) \\ &\leq C \log(S_n) \sum_{k=0}^{n-1} \Delta_{k+1} \min\left(1, \frac{\lambda}{S_k}\right) \\ &= C \log(S_n) \mathbb{E}(Z_n^\lambda) \end{aligned} \tag{16}$$

Consequently, the Bienaymé-Tchebychev inequality applied toZ_n^λ reads:

$$\mathbb{P}(|Z_n^\lambda - \mathbb{E}Z_n^\lambda| \geq \rho \mathbb{E}Z_n^\lambda) \leq C \frac{\log S_n \mathbb{E}Z_n^\lambda}{\rho^2 (\mathbb{E}Z_n^\lambda)^2} \leq C \frac{\log S_n}{\rho^2 \sum_{k=0}^{n-1} \Delta_{k+1} \min(1, \lambda/S_k)} \tag{17}$$

We can check that since $\Delta_n \min(1, \lambda/S_n) \sim \lambda \frac{\gamma_n}{1-\gamma_n} \sim \lambda \gamma_n$, we have $\lim_n \frac{\sum_{k=0}^{n-1} \Delta_{k+1} \min(1, \lambda/S_k)}{\log S_n} = \lambda$. Denoting the event $A_n^\lambda = \{Z_n^\lambda - \mathbb{E}Z_n^\lambda < \rho \mathbb{E}Z_n^\lambda\}$, for λ large enough, we have $\mathbb{P}(A_n^\lambda) \geq \frac{1}{2}$, such that $\mathbb{P}(\limsup A_n^\lambda) \geq \frac{1}{2}$. On the event $\limsup_n A_n^\lambda$, we get:

$$Z_n^\lambda \geq (1 - rho)\mathbb{E}Z_n^\lambda \geq \lambda(1 - rho) \log Sn \qquad \text{for infinitly many} n \tag{18}$$

Hence $\mathbb{P}(\limsup_n \frac{Z_n^\lambda}{\log S_n} \geq \lambda(1-\rho)) \geq \frac{1}{2}$. 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, giving us:

$$\limsup_n \frac{Z_n^\lambda}{\log S_n} \geq \lambda(1 - rho) \qquad \mathbb{P}\text{-almost surely} \tag{19}$$

This holds for every $\rho > 0$ and $\lambda > 0$, which prooves that $\limsup_n \frac{Y_n}{\log S_n} = +\infty$. This concludes that $\mathbb{P}(X_\infty = 0) = 0$: the bandit is infallible.

□

This result is key as it is our first sufficient condition for infallibility, which we will use to prove all the following theorems. We will now use it to introduce lemma 3.4, which will give a sufficient condition on γ_n instead of Δ_n, making it much easier to use, especially to prove the next theorems in section 3.3.

Lemma 3.4. *Assuming that* $0 \leq p_A = p_B \leq 1$ *and that* $\gamma_n = \mathcal{O}(\Gamma_n e^{-p_B \Gamma_n})$. *Then the two-armed bandit is infallible for any* $X_0 \in]0, 1]$.

Proof. As we have $p_A = p_B$ we can rewrite the algorithm as:

$$X_{n+1} = X_n + \gamma_{n+1} \mathbb{1}_{B_{n+1}} (\mathbb{1}_{U_{n+1} \leq X_n} - X_n) \tag{20}$$

$$X_{n+1} = X_n + \gamma_{n+1}^B (\mathbb{1}_{U_{n+1} \leq X_n} - X_n) \tag{21}$$

With the new step size $\gamma_n^B = \gamma_n \mathbb{1}_{B_n}$. We can see that here, using (γ_n, p_B) is equivalent to using $(\gamma_n^B, 1)$, meaning we can use lemma 3.3 to prove this second lemma. We define $\Delta_n^B = \frac{\gamma_n^B}{\prod_{k=1}^n (1-\gamma_k^B)}$, if $\Delta_n^B = \mathcal{O}(\Gamma_n)$, then the infallibility is established. We now introduce

$$M_n = \sum_{k=1}^{n} \log(1 - \gamma_k^B) - p_B \log(1 - \gamma_k) = \sum_{k=1}^{n} \log(1 - \gamma_k)(\mathbb{1}_{B_k} - 1) \tag{22}$$

Let us proove that M_n is a bounded martingale. Writing $\mathbb{E}(M_{n+1}|B_n)$ shows directly that M_n is indeed a martingale.

$$\mathbb{E}(M_{n+1}|B_n) = \mathbb{E}(\log(1 - \mathbb{1}_{B_{n+1}}\gamma_{n+1}) - p_B \log(1-\gamma_k)) + M_n = M_n \tag{23}$$

Now, as $u \to ue^{-p_B u}$ is non-increasing for a large enough u and integrable, we can show using $\gamma_n = \mathcal{O}(\Gamma_n e^{-p_B\Gamma_n})$ that $\sum_{k=1}^n \gamma_k^2 < +\infty$ because:

$$\sum_{k=1}^n \gamma_k^2 \leq C \sum_{k=1}^n (\Gamma_k - \Gamma_{k-1})\Gamma_k e^{-p_B\Gamma_k} \leq C \int_0^{\Gamma_n} u e^{-p_B u} du < +\infty \tag{24}$$

Using this result and the majoration $\forall x \in]0,1]$,$\log(1-x) \leq x$, it is clear that $\mathbb{E}(M_n^2) < +\infty$ and as such that M_n is a bounded martingale. Using theorem 1.2, we can say that M_n is almost surely bounded by a random variable M_∞, and by continuity we can add that the ratio

$$e^{M_n} = \frac{\prod_{k=1}^n (1-\gamma_k \mathbb{1}_{B_k})}{\prod_{k=1}^n (1-\gamma_k)^{p_B}} \qquad \text{is almost surely bounded} \tag{25}$$

As a result, we can introduce a postive random constant ξ that is $\sigma(B_n, n \geq 1)$-measurable such that

$$\Delta_n^B \leq \xi \frac{\gamma_n^B}{\prod_{k=1}^n (1-\gamma_k)^{p_B}} = \xi \gamma_n^B S_n^{p_B} \leq \xi \gamma_n S_n^{p_B} \tag{26}$$

Using the result from lemma 3.2 and $\sum_{k=1}^n \gamma_k^2 < +\infty$, we get that $S_n \leq Ce^{-\Gamma_n}$. Moreover as $\gamma_n = \mathcal{O}(\Gamma_n e^{-p_B\Gamma_n})$ it follows that:

$$\Delta_n^B \leq \xi \Gamma_n e^{-p_B\Gamma_n} (e^{\Gamma_n})^{p_B} \leq \xi \Gamma_n \tag{27}$$

With a straightforward martingale argument, it comes that $\sum_{k=1}^n \gamma_k^B \equiv p_B \sum_{k=1}^n \gamma_k$. Combining this with the inequality above, we finally obtain that $\Delta_n^B = \mathcal{O}(\sum_{k=1}^n \gamma_k^B)$, thus fulfilling the hypothesis of lemma 3.3.
As a result $\forall X_0 \in]0,1]$, $\mathbb{P}(X_\infty = 0) = 0$.

□

3.3. Infallibility

Now that we have a sufficient condition for *infallibility* in the equality case, we will extend it to the more interresting case where $p_A > p_B$, *ie* when one agent performs better than the other. First we will introduce a pathwise comparison result.

Lemma 3.5. *Let $X_0 \in [0,1]$, let (X_n) and (X'_n) be two coupled two-armed bandit algorithm built from (U_n) and (V_n), starting from $X_0 \leq X'_0$ and with parameters (p_A, p_B) and (p'_A, p_B) respectively, with $p_A \leq p'_A$. Then we get that $\forall n$, $X_n \leq X'_n$. In particular, we get the following inclusion: $\{X'_n \to 0\} \subset \{X_n \to 0\}$*

Proof. This result is shown by induction and analysing every 4 different cases. Assuming that $X_n \leq X'_n$,

- If $U_n \leq X_n \cap V_n \leq p'_A$ then $X'_{n+1} = X'_n + \gamma_{n+1}(1 - X'_n)$. Wheter X_n passes the test or not, we get that $X_{n+1} \leq X'_n + \gamma_{n+1}(1-X'_n)$. As $X_n \leq X'_n$, we have that $X_{n+1} \leq X'_{n+1}$.
- If $U_n \leq X_n \cap V_n \geq p'_A$ then $X_{n+1} = X_n \leq X'_n = X'_{n+1}$
- If $U_n \geq X_n \cap V_n \leq p_B$ then $X_{n+1} = (1-\gamma_{n+1})X_n \leq (1-\gamma_{n+1})X'_n = X'_{n+1}$
- If $U_n \geq X_n \cap V_n \geq p_B$ then $X_{n+1} = X_n \leq X'_n = X'_{n+1}$

Which demonstrates by induction that $\forall n$, $\forall n$, $0 \leq X_n \leq X'_n$. This directly implies that if $X'_n \to 0$, we get by comparison $X_n \to 0$ and thus the set inclusion $\{X'_n \to 0\} \subset \{X_n \to 0\}$. □

Finally we can introduce the theorem providing us a sufficient condition for infallibility in the general case.

Theorem 3.2. *Assuming that $p_A \geq p_B$ and that $\gamma_n = \mathcal{O}(\Gamma_n e^{-p_B\Gamma_n})$, $\forall X_0 \in]0,1]$, $\mathbb{P}(X_\infty = 0) = 0$.*

Proof. Consider two bandits X_n and X'_n using the gamma γ_n introduced above and the probabilities (p_B, p_B) and (p_A, p_B) respectively. As $p_A \geq p_B$, the Pathwise lemma 3.5 gives us that $\{X'_n \to 0\} \subset \{X_n \to 0\}$. Moreover the lemma 3.4 ensure the *infallibility* in the equality case as soon as $\gamma_n = \mathcal{O}(\Gamma_n e^{-p_B\Gamma_n})$. As such we have $\mathbb{P}(X_\infty = 0) = 0$ and as a consequence of the inclusion, $\mathbb{P}(X'_\infty = 0) = 0$, which proves the theorem for $p_A \geq p_B$. □

3.4. A pratical result: the power step

In the previous section, we derived **a general sufficient condition** to infallibility. However, in practice, the condition $\gamma_n = \mathcal{O}(\Gamma_n e^{-p_B\Gamma_n})$ is not easy to manipulate to tune hyperparameters. In this section, we will investigate a specific case we denote as *power step* and provide a necessary and sufficient condition to the *infallibility*.

Definition 3.2. *We define as power step the step sizes γ_n of the form $(\frac{C}{C+n})^\alpha$. With $\alpha \in]0,1]$ and $C > 0$.*

Theorem 3.3. *Let $\gamma_n = (\frac{C}{C+n})^\alpha$ be a power step. Then $\mathbb{P}(X_\infty = 0) = 0$ **if and only if:***

$$\alpha = 1 \quad \text{and} \quad C \leq \frac{1}{p_B}$$

Proof. We will only prove the sufficient side of this condition here and leave the necessity aside. It is clear that with $\alpha = 1$, we obtain by comparison with the harmonic series $\Gamma_n \to +\infty$ as $n \to +\infty$. Thus, according to theorem 3.2, we need to prove that $\gamma_n = \mathcal{O}(\Gamma_n e^{-p_B\Gamma_n})$. Again by comparing Γ_n to the harmonic series, we get the following equivalent:

$$\Gamma_n = C\log(n) + C' + o(1) \tag{28}$$

Consequently, the sufficient condition for X_n to be infallible according to theorem 3.2 can be derived as:

$$\gamma_n = \frac{C}{n+C} = \mathcal{O}(\log(n)n^{-Cp_b}) \tag{29}$$

Which means that we would need $\frac{n^{Cp_B-1}}{\log(n)}$ to be bounded, that is $Cp_b \leq 1$. That concludes the proof that $\alpha = 1$ and $C \leq \frac{1}{p_B}$. To prove the reciprocity of this result (*ie*: that this condition is also a necessity), one would need to prove first that $\sum_{n\geq 0}\prod_{k=1}^n(1-p_B\gamma_k) < +\infty$ implies that the algorithm is fallible and that other power step parameters would fall under this criterion. □

This result is very important in practice, as it makes the design of an asymptotically infallible bandit easy. Indeed, **one simply need to use $C = 1$ to be assured of infallibility**. In other words, the commonly used simple step size $\gamma_n = \frac{1}{n+1}$ fullfills all the hypothesis of infallibility.

4. Speed of convergence

In the previous section, we studied the asymptotic convergence of the bandit as detailled in [10]. However when deploying such an algorithm the speed of convergence is also a very important factor. With the asset allocation problem, a slow rate of convergence still presents an opportunity cost. In this section, we study the speed of convergence of the two-armed bandit as detailled in the follow-up paper [9].

4.1. Problem reformulation and first results

In order to study the rate of convergence of X_n to 1, we start by rewriting the algorithm to make the difference $1 - X_n$ explicit and define two related sequences θ_n and Y_n before introducing a couple preliminary results that will be useful to determine the speed of convergence.

$$1 - X_{n+1} = (1 - X_n)(1 - \gamma_n\pi X_n) - \gamma_n\Delta M_n \tag{30}$$

The key idea of this analysis will be to study the behaviour of $(1 - X_n)/\prod_{k=1}^n(1 - \pi\gamma_k)$ asymptotically. If we manage to control this expression asymptotically with postive constants, this will give us a speed of convergence, using some majoration of the denominator. For this sake, we introduce some useful sequences:

Definition 4.1. *We define $\theta_n = \prod_{k=1}^n(1 - \gamma_k\pi X_{k-1})$ and $Y_n = \frac{1-X_n}{\theta_n}$*

We can see from the definition that Y_n is a non-negative martingale. This comes from equation 30, when dividing both members by θ_{n+1} we obtain the relationship $Y_{n+1} = Y_n - \frac{\gamma_{n+1}}{\theta_{n+1}}\Delta M_{n+1}$ (θ_{n+1} is predictable). Using these definition and this remark, we can now introduce a first asymptotic result on $1 - X_n$.

Lemma 4.1. *On the set $\{X_\infty = 1\}$, we have for ξ a finite postive random constant and Y_∞ the limit of Y_n:*

$$\lim_{n\to+\infty}\frac{1 - X_n}{\prod_{k=1}^n(1 - \pi\gamma_k)} = \xi Y_\infty \qquad \textit{almost surely} \tag{31}$$

Proof. As a non-negative martingale, the sequence $(Y_n)_{n\in\mathbb{N}}$ has a limit Y_∞ such that $Y_\infty \geq 0$ and $\mathbb{E}(Y_\infty) < +\infty$. Recall that from theorem 3.1, $X_\infty \in \{0;1\}$, thus we have $\sum_n \gamma_{n+1} X_n(1-X_n) < +\infty$. Therefore on $\{X_\infty = 1\}$, we have $\sum_n \gamma_{n+1}(1-X_n) < +\infty$ almost surely. Let $\xi_n = \prod_{k=1}^n \frac{1-\pi\gamma_k X_{k-1}}{1-\pi\gamma_k}$. On $\{X_\infty = 1\}$ we have that:

$$\log(\xi) = \sum_{k=1}^n \log(1-\pi\gamma_k X_{k-1}) - \log(1-\pi\gamma_k) \tag{32}$$

Moreover as $\log(1-\pi\gamma_k X_{k-1}) - \log(1-\pi\gamma_k) \sim \pi\gamma_k(1-X_{k-1})$, the convergence of $\sum_n \gamma_{n+1}(1-X_n)$ gives us that $\xi_n \to \xi < +\infty$. Finally, remarking that $\xi_n Y_n = \frac{1-X_n}{\prod_{k=1}^n 1-\pi\gamma_k}$, we can conlude the proof of the lemma. □

What does this result tell us ? Notice that we have $\prod_{k=1}^n (1-\pi\gamma_k) \leq e^{-\pi\Gamma_n}$. Thus, with lemma 4.1, as ξY_∞ is finite, we get that almost surely on $\{X_\infty = 1\}$ $1 - X_n = \mathcal{O}(e^{-\pi\Gamma_n})$. If we add that $\sum_n \gamma_n^2 < +\infty$, we reach a better result as $\left(e^{\pi\Gamma_n}\prod_{k=1}^n(1-\pi\gamma_k)\right)$ converges toward a positive limit: we have for ξ' postive

$$\lim_n e^{\pi\Gamma_n}(1-X_n) = \xi' Y_\infty \tag{33}$$

This equation above gives us the speed of the algorithm **as long as** $\{Y_\infty > 0\}$. We can also notice that $\{Y_\infty = 0\} \subset \{X_\infty = 1\}$. Thus we need to refine our analysis in order to derive the speed of convergence on the set $\{Y_\infty = 0\}$. In order to introduce this new paradigm, we will derive a result which will not prove before diving into this new regime of convergence.

Lemma 4.2. *If $\sum_n \gamma_n^2 e^{\pi\Gamma_n} < +\infty$, then $(Y_n)_{n\geq 0}$ is bounded in L^2 and its limit verifies $\mathbb{E}(X_\infty Y_\infty) > 0$. Moreover on the set $\{Y_\infty = 0\}$, we have almost surely:*

$$\limsup_n \frac{Y_n}{\sum_{k\geq n}\gamma_{k+1}^2 e^{\pi\Gamma_{k+1}}} < +\infty \tag{34}$$

Secondly if $\sum_n \gamma_n^2 e^{\pi\Gamma_n} = +\infty$ and $\sup_n \gamma_n e^{\pi\Gamma_n} < +\infty$, then for every $x \in]0,1[$:

$$\{X_\infty = 1\} = \{Y_\infty = 0\} \quad \mathbb{P}\text{-almost surely} \tag{35}$$

The second result of lemma 4.2 clearly shows that in some cases, **we can experience** $\{Y_\infty = 0\}$ with a positive probability. Which means that we need to refine our analysis to that case.

4.2. The fast convergence regime

In this section, we introduce results that will allow us to tackle the case of $Y_\infty = 0$, *ie:* $1 - X_n = o(e^{-\pi\Gamma_n})$, thus converging faster than before. We call *fast* rate of convergence the case when $\sum_n 1 - X_n < +\infty$, the rationale behind this choice is made by the proof of the following theorem:

Theorem 4.1. *Assume $\sum_n \gamma_n^2 < +\infty$. On the set $\{\sum_n 1 - X_n < +\infty\}$, we have $1 - X_n \sim \xi e^{-p_A\Gamma_n}$*

Proof. Up to null events, we have that

$$\begin{aligned}\left\{\sum_n 1 - X_n < +\infty\right\} &= \left\{\sum_n \mathbb{1}_{U_n > X_n} < +\infty\right\} \\ &\subset \bigcup_{m\geq 1}\bigcap_{n\geq m}\left\{1 - X_n = (1-X_m)\prod_{k=m+1}^n (1-\mathbb{1}_{A_k}\gamma_k)\right\}\end{aligned} \tag{36}$$

Thus, we have in any case on $\{\sum_n 1 - X_n < +\infty\}$ that:

$$1 - X_n \sim \xi \prod_{k=1}^n (1-\mathbb{1}_{A_k}\gamma_k) \tag{37}$$

With ξ a positve random variable. Now, recall that for $\sum_n \gamma_n^2 < +\infty$ we have $\prod_{k=1}^n (1-\mathbb{1}_{A_k}\gamma_k) \sim \xi' e^{-p_A\Gamma_n}$. This let us conclude that on $\{\sum_n 1 - X_n < +\infty\}$, $1 - X_n \sim \xi e^{-p_A\Gamma_n}$.

□

This theorem gives us a very convienient equivalent of $1 - X_n$ when $\sum_n 1 - X_n < +\infty$: **this explicitates the fast convergence rate**. Now that we caracterized the event on which the fast rate occurs, we need to make sure that this event is not of probability zero. For this sake, we introduce theorem 4.2 that will help us characterize this occurence. Finally we will give a sufficient condition to achieve the faster rate with probabilty 1 and another one to randomly pick between the slow and the fast one in theorem 4.3.

Theorem 4.2. *We have* $\forall x \in]0,1[$, $\mathbb{P}\left(\sum_n (1 - X_n) < +\infty\right) > 0 \iff \mathbb{P}\left(\sum_{n\geq 1} \prod_{k=1}^n (1 - \mathbb{1}_{A_k}\gamma_k) < +\infty\right) > 0$ *Moreover if* $\sum_n \gamma_n^2 < +\infty$*, we have:*

$$\mathbb{P}\left(\sum_n (1 - X_n) < +\infty\right) > 0 \iff \sum_n e^{-p_A \Gamma_n} < +\infty \tag{38}$$

Proof. Let us first remark that $\{\sum_n (1 - X_n) < +\infty\} = \{\sum_n 1 - \mathbb{1}_{U_{n+1} \leq X_n}\}$. As such, we can rewrite this event as a non decreasing union of events.

$$\left\{\sum_n (1 - X_n) < +\infty\right\} = \bigcup_{n\geq 0} \bigcap_{k \geq n} \{U_{k+1} \leq X_k\} \tag{39}$$

Which consequently gives us, using the properties of the probability of a non-decreasing union of events.

$$\mathbb{P}\left(\sum_n (1 - X_n) < +\infty\right) = \lim_{n \to +\infty} \mathbb{P}\left(\bigcap_{k \geq n} \{U_{k+1} \leq X_k\}\right) \tag{40}$$

The right part of this equation will be greater that 0 if and only if for some integer $n \geq 1$ we have $\mathbb{P}(\bigcap_{k\geq n}\{U_{k+1} \leq X_k\}) > 0$. Now, working on the event itself, we get from the definition of X_n that:

$$\begin{aligned} \bigcap_{k\geq n} \{U_{k+1} \leq X_k\} &= \bigcap_{k \geq n} \left\{U_{k+1} \leq X_k \text{and} X_k = X_n \prod_{l=n+1}^{k} (1 - \mathbb{1}_{A_l}\gamma_l)\right\} \\ &= \bigcap_{k \geq n} \left\{U_{k+1} \leq X_n \prod_{l=n+1}^{k} (1 - \mathbb{1}_{A_l}\gamma_l)\right\} \end{aligned} \tag{41}$$

Now, denoting $\mathcal{B}_n$ the σ-field generated by the variable X_n and the events B_k for $k \geq n$, we get that:

$$\mathbb{P}\left(\bigcap_{k\geq n} \left\{U_{k+1} \leq X_n \prod_{l=n+1}^{k} (1 - \mathbb{1}_{A_l}\gamma_l)\right\} | \mathcal{B}_n\right) = \prod_{k=n}^{\infty} \left(1 - X_n \prod_{l=n+1}^{k} (1 - \mathbb{1}_{A_l}\gamma_l)\right) \tag{42}$$

Using a similar argument with a log as in equation 32, we get the convergence of this infinite product if and only if: $\sum_k \prod_{l=n+1}^k (1 - \mathbb{1}_{A_l}\gamma_l) < +\infty$. Which gives us $\mathbb{P}\left(\sum_{n\geq 1} \prod_{k=1}^n (1 - \mathbb{1}_{A_k}\gamma_k) < +\infty\right) > 0$. When switching the summation and product indexes and taking the expected value, we reach:

$$\mathbb{E}\left(\sum_n \prod_{k=1}^n (1 - \mathbb{1}_{A_k}\gamma_k)\right) = \sum_n \prod_{k=1}^n (1 - p_A\gamma_k) \tag{43}$$

Noticing that at the first order $1 - p_A\gamma_k \sim e^{-p_A\gamma_k}$, we get by comparison that $sum_n e^{-p_A\Gamma_n} < +\infty$ which concludes on the sufficient condition of the theorem. Finally, if $\sum_n \gamma_n^2 < +\infty$, then a result from [10] (see proof of Lemma 2), gives us the reciprocity using:

$$\prod_{k=1}^{n} \frac{1 - \mathbb{1}_{A_k}\gamma_k}{1 - p_A\gamma_k} \longrightarrow \xi > 0 \quad \text{almost surely,} \quad n \to +\infty \tag{44}$$

□

Theorem 4.3. *Let* $\epsilon_n = \frac{1}{\gamma_{n+1}} - \frac{1}{\gamma_n} - \pi$ *for* $n \geq 1$*. We denote* $\epsilon_n^+ = \max(0, \epsilon_n)$*.*

- *If* $\sum_n \gamma_n \epsilon_n^+ < +\infty$*, then* $\sum_n 1 - X_n < +\infty$ *almost surely on the set* $\{X_\infty = 1\}$
- *If* $\liminf_n \epsilon_n > 0$ *then* $\sum_n \gamma_n e^{\pi\Gamma_n} < +\infty$*, and on the event* $\{Y_\infty = 0\}$*,* $\sum_n 1 - X_n < +\infty$ *almost surely.*

4.3. Convergence speed to the target

In this section, we use all the lemmas and theorem we established above in order to derive some interesting convergence speed results of the two-armed bandit in the infallible case. We will assume that we are using a power-step with $alpha = 1$ (at least for n large enough) to ensure infallibility and make the result more pratical:

$$\gamma_n = \frac{C}{C+n}, \qquad C > 0 \tag{45}$$

Theorem 4.4. *The speed of convergence of the algorithm depend on the parameter C such that:*

- *If $\frac{1}{\pi} \leq C \leq \frac{1}{p_B}$ then $X_n \to 1$ with the **fast** rate of convergence of n^{-Cp_A}*
- *If $\frac{1}{p_A} \leq C \leq \frac{1}{\pi}$ then $X_n \to 1$ with either the **fast** rate of convergence of n^{-Cp_A} either the **slow** one $n^{-C\pi}$, both occuring with a postive probability.*
- *If $0 < C \leq \frac{1}{p_A}$ then $X_n \to 1$ with the **slow** rate of convergence of n^{-Cp_π}*

Remark It follow from theorem 4.4 that in a practical setup, finding the slow rate of convergence is easy. Indeed as detailled theorem 3.3, using $\alpha = 1$ and $C = 1$ ensure of the infallibility of the algorithm with a convergence speed of $n^{-\pi}$. However, taking advantage of the fast rate of convergence would require some parameter tuning of C, as it is dependend of the usually unknown probabilities p_A and p_B.

Proof. Firstly, if $\frac{1}{\pi} \leq C \leq \frac{1}{p_B}$, we get that $\epsilon_n = \frac{C+n+1}{C} - \frac{C+n}{C} - \pi = \frac{1}{C} - \pi < 0$. Thus it is clear that $\epsilon_n^+ = 0$. This prooves that $\sum_n \gamma_n \epsilon_n^+ < +\infty$. By using 4.3 we get that $\sum_n 1 - X_n < +\infty$. Recalling that here $\Gamma_n \sim C\log(n)$, we get that $1 - X_n \sim \xi e^{-p_A * \Gamma_n} \sim \xi n^{-p_A}$ (see Theorem 4.1).
Secondly if $\frac{1}{p_A} \leq C \leq \frac{1}{\pi}$, we have $\epsilon_n = \frac{1}{C} - pi$. Thus $\liminf_n \epsilon_n^+ > 0$. Moreover, note that $\sum_n e^{-p_A \Gamma_n} < +\infty$ because $e^{-p_A \Gamma_n} \sim n^{-Cp_A} < n^{-1}$. Using lemma 4.2, we get that $\mathbb{P}(\sum_n 1 - X_n < 0) > 0$, *ie:* $\mathbb{P}(Y_\infty = 0) > 0$. Thus, if $\{Y_\infty = 0\}$, we are back in the previous case with a speed of convergence of n^{-Cp_A}. Otherwise if $\{Y_\infty > 0\}$, we are under the regime of lemma 4.1, and we get a speed of $n^{-C\pi}$. This proves the second point, for $\frac{1}{p_A} \leq C \leq \frac{1}{\pi}$ both speed regimes are possibile with a possitive probability.
Finally if $0 < C \leq \frac{1}{p_A}$, then $\sum_n e^{-p_A \Gamma_n} = +\infty$. A similar argument with 4.2 gives us that $\mathbb{P}(\sum_n 1 - X_n < 0) = 0$ and as such that $\mathbb{P}(Y_\infty = 0) = 0$. Using lemma 4.1, we end up using the slow rate of convergence.

□

Note that similar speed results for $C > \frac{1}{p_B}$, *ie:* when X_n converges to 0, were omitted in this section. The authors of [9] managed to prove some analog results regarding the speed of convergence to the trap. These were however omitted here for the sake of simplicity.

5. Experiments

In this section, we present some numeric simulations we conducted using the two-armed bandit algorithm. The first part consists of theoritical framework describe in [10] and [9]. We use it to verify results on infallibility and speed of convergence. In the second part, we use real financial data to test the behaviour of the algorithm in a practical setup and compare it with the theory. Finally, we briefly try to extend the framework to a multi-agent setup. The notebook associated to this paper presents some more experiments that were not included here for the sake on consiceness.

5.1. Theoretical results

First, we simulate the theoritical framework detailled in section 2. The probability of success p_A and p_B are fixed before the runs. Here we chose to use $p_A = 0.7$, $p_B = 0.2$ (note that we keep $p_A > p_B$ to match the theory established above) and γ_n a power step. See the associated notebook for more details.
Figure 1 shows the influence of the value of C on the repartition of capital between the two agents (X_n and $1 - X_n$). In order to interprete de results, note that here we have $\frac{1}{p_A} \sim 1.42$, $\frac{1}{\pi} = 2$ and $\frac{1}{p_b} = 5$. We observe overall very smooth convergence, and even without comparing directly to the theoritical bound we clearly see the distinction between the slow regime when $C = 1$ and the fast one when $C = 2, 4, 5$. Interrestingly, when we use $C = 0.5$ or $C = 0.1$, even if the asymptotic convergence is assured by the theory, it is acutally slow enough that we do not see it in practice in 1000 iteration. Running the algorithm longer should see them matching their counterparts.
Figures 2 and 3 present the comparison between the oberseved speed and the theoritical one. The slow speed of convergence is illustrated with $C = 1$, whereas the fast one with $C = 3$. Both match the theory very closely.

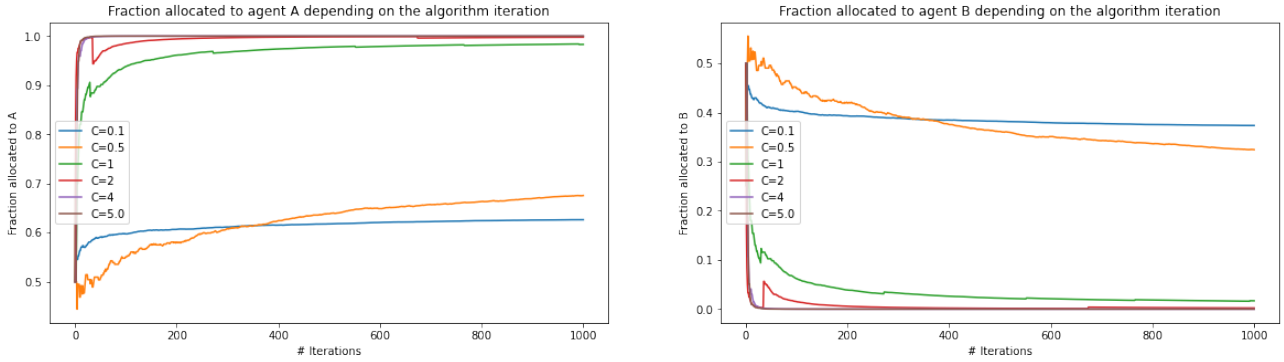

Figure 1: Influence of the value of C.

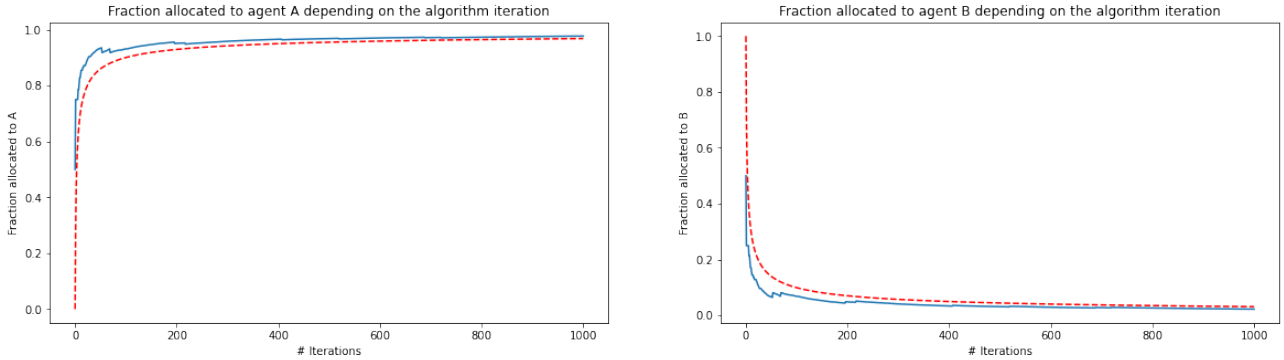

Figure 2: Slow speed regime vs theory (in dashed red).

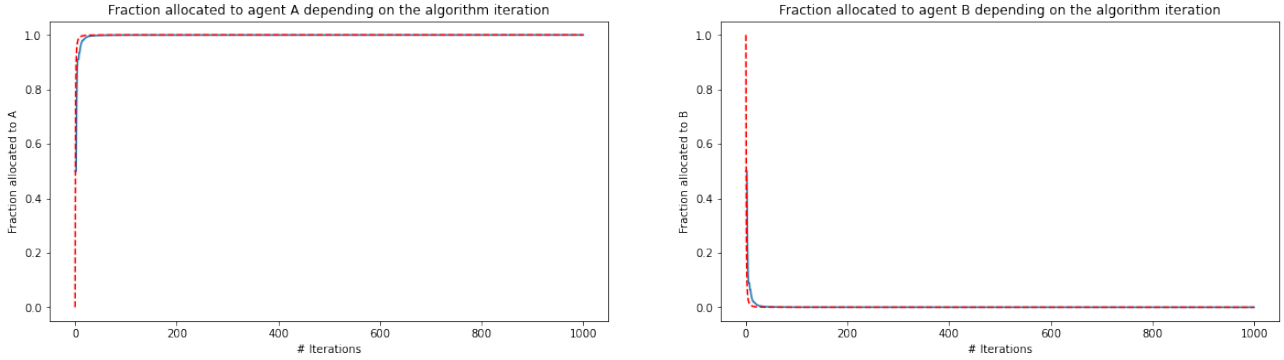

Figure 3: Fast speed regime vs theory (in dashed red).

5.2. A bandit with real data

Now that we experiemented with the theoritical setup, we try to use real data see how well does the framework holds. For this sake, we will use *Google*'s stock price between January 2011 and December 2011, resampled every two hours. We will consider two simple trading agents that will compete for capital in our imaginary hedge fund. Both will start with a 50% share. The test of satisfaction will be the increase of PnL. In other words, an agent gives us satisfaction if its PnL has increased since the last time it was tested. We implemented 3 kind of agents: one random, one relying on the expoential moving average and one on the number of consecutive price increase or decrease. Note that in practice, the test is very questionnable as well as well as the strategies employed by the agents.

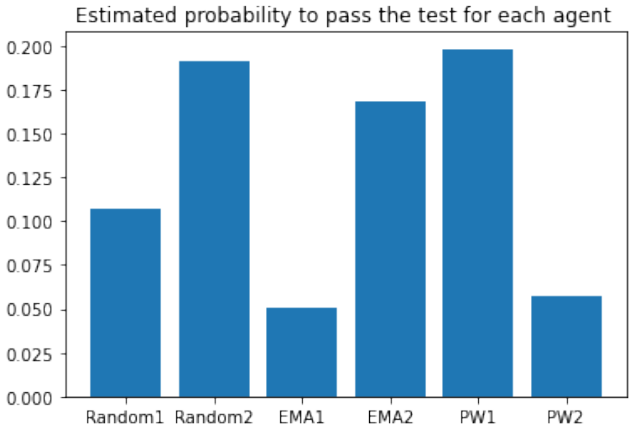

Figure 4: Estimation of each agent probability to pass the test

Figure 4 show the histogram of estimated probabilities for each agent. We created two agents of each kind with different hyper-parameters. This histogram will be useful to interprete the following results on the asset allocation when matching one agent against the other.

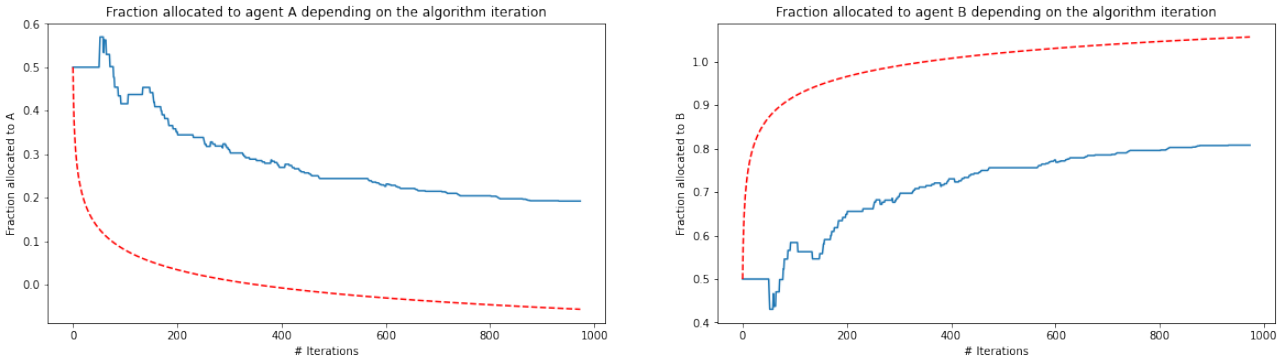

Figure 5: Agent EMA1 vs EMA2

In Figure 5, we run the two-armed bandit with the agents *EMA1* and *EMA2* (both relying on the expoential moving average with a respective α's of 0.5 and 0.9). We use $C = 4$ and as such expect a slow rate of convergence. The graph clearly shows that agent B (*EMA2*) is the one being awarded the majority of the capital, which is what we expected as here $p_B > p_A$. Moreover regarding the speed of convergence, we see that the first 200 steps are much more chaotic and slower than what the theory would expect. However, after this warmup phase, we see that the shape of the curves match fairly well. This behaviour is consistent with other similar experiments we made. Finally we can also note that 1000 steps were not enough to allocate all the ressources to agent B, as it terminates around 0.8. This could be mitigated by increase the value of C

In Figure 6, we compare the agents *PW1* and *EMA2*. These two have for particularity that p_A is very close to p_B. It is interresting to see that the bandit seems to have a hard time picking the best one. If theoritically, the setup is such that agent A should end up being favoritized, the reality is that first the bandit allocates more to agent B. However it corrects itself quickly and at the end of the simulation we see a slight trend in favour of A. Still, we are far away of the theoritical bound and it seems to be difficult to pick between two agents with similar probabilities.

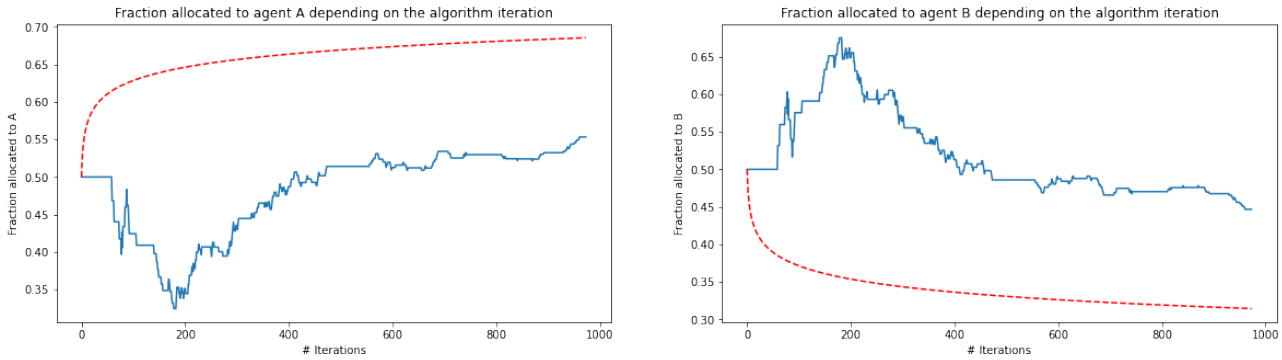

Figure 6: Agent PW1 vs EMA2

5.3. Multi-agent extension

Finally, in this section, we try to extend the framework to a multi-agent setup. For this experiement, we use 9 different *EMA* agents, we display their estimated probability of success in figure 7. Regarding the re-allocation, we keep the same idea we used in the two-armed bandit: the reward for the sucessful agent will come evenly from all the other agents, proportionally to the fraction of the capital they are currently managing.

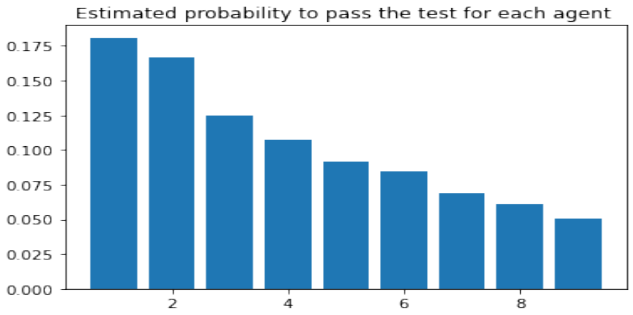

Figure 7: Estimation of each agent probability to pass the test

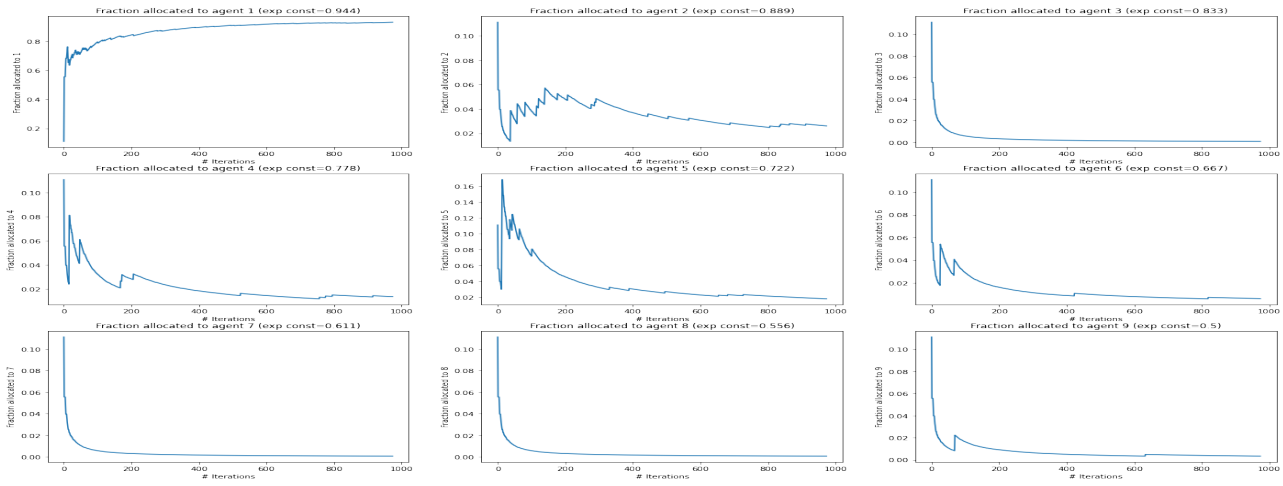

Figure 8: Fraction of the capital allocated to each agent

In figure 8 we compare the fraction allocated to each agent. Overall the framework seems to hold very well. Agent 1, with the highest probability, finishes the simulation with the highest allocation around 0.9, and most of the remaining capital is given to agent 2 with a very similar probability of sucess. Agents with a significantly lower probability have seen their allocation very quickly decrease to zero. These results are in the end very satisfying from the bandit point of view.

6. Related Work

Bandit algorithms is a well-documented topic and has many applications. Many papers tackle the issue and propose different solutions with sometimes very different approaches.

In 2004, D. Lamberton, G. Pagès and P. Tarrès investigated in [10] the convergence of the two-armed bandit. They propose a **non-penalized algorithm** and study its convergence. They also propose a parrallel between the two-armed bandit and the **Polya Urn** which was not mentionned in this paper. Next year in 2005, D. Lamberton and G. Pagès propose a second paper [9] in which they tackle the speed of convergence of this algorithm. They managed to derive some clear sufficient conditions to describe the behaviour of the algorithm. The same year, they also publish a third paper [8], proposing **a penalized bandit algorithm** and studying both its convergence and its speed.

Here we tackled the bandit problem with the **asset allocation** problem in mind, however one can find many other application to it. For instance let us reference [11], which uses bandits to **recommend advertisements**, or **wireless telecommunication** in [2]. See [3] for a more detailled sector related summary of existing applications.

Nowadays, many tackle the k-armed bandit problem and many theoritical and practical results exist to adapt to a wide range of problem. One of the most well known example of bandit algorithm is the **UCB** or **LinUCB** method [6], which uses linear rewards for each arms. These methods were later complemented with new heuristics such as **Thomson Sampling** [1]. More theoritical results can be found in G. Stoltz's work [7], [5], [4]. In [12], A. Slivkins gives a comprehensive introduction to many modern bandit techniques and some theoritical results regarding their convergence. In a broader perspective, the tradeoff between exploration and exploitation is at the heart of many **Reinforcement Learning** and **Recommander Systems** problems and algorithms.

7. Conclusion

In this paper, we derived a theoritical framework for the convergence of the two-armed bandit algorithm. We managed not only to prove its convergence, but most importantly to identify sufficient conditions to ensure its infallibility at various speed. We identified two regimes of convergence in the infallible case: a *faster* one that would need specific parameter tuning, and a *slower* one which could be reached fairly easily. One of the most insterresting result of this framework is that using the simplest power-step with $C = 1$ and $\alpha = 1$ makes the bandit infallible. Even if this comes with the slower rate of convergence, using this step requires no extra knowledge on the agents, which allows to use this algorithm to discriminate between different strategies very easily. The numerical experiments confirmed that those results seemed to transpose very well in a real setup, as well as in an extended mutli-agent framework. In the end, the two-armed bandit algorithm proposes an elegant solution to our problem, mitigating the opportunity cost by performing a dynamic online reallocation of the ressources. Please note that this paper is purely academic and holds no claim of novelty, all credits goes to the original authors.

References

[1] Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, May 2013.

[2] Stefano Boldrini, Luca De Nardis, Giuseppe Caso, Mai T. P. Le, Jocelyn Fiorina, and Maria-Gabriella Di Benedetto. muMAB: A Multi-Armed Bandit Model for Wireless Network Selection. *Algorithms*, 11(2):13, January 2018.

[3] Djallel Bouneffouf and Irina Rish. A Survey on Practical Applications of Multi-Armed and Contextual Bandits. *arXiv*, April 2019.

[4] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure Exploration for Multi-Armed Bandit Problems, June 2010. [Online; accessed 16. Apr. 2022].

[5] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvari. X-Armed Bandits. *arXiv*, January 2010.

[6] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[7] Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems. *Math. Oper. Res.*, 44(2):377–399, 2019.

[8] Damien Lamberton and Gilles Pagès. A penalized bandit algorithm. *arXiv*, October 2005.

[9] Damien Lamberton and Gilles Pagès. How fast is the bandit? *arXiv*, Oct 2005.

[10] Damien Lamberton, Gilles Pages, and Pierre Tarres. When can the two-armed bandit algorithm be trusted? *arXiv*, Jul 2004.

[11] Tyler Lu, Dávid Pál, and Martin Pál. Showing Relevant Ads via Lipschitz Context Multi-Armed Bandits. *Google Research*, 2010.

[12] Aleksandrs Slivkins. Introduction to Multi-Armed Bandits. *arXiv*, April 2019.