Regret of Narendra Shapiro Bandit Algorithms

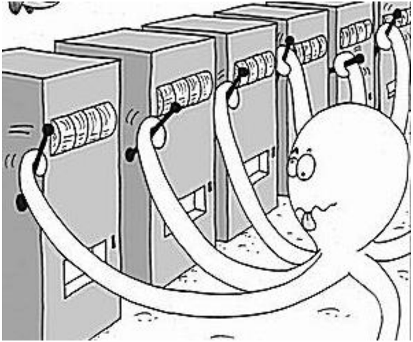
S. Gadat

Toulouse School of Economics
Joint work with F. Panloup and S. Saadane.

Inra Auzeville, September, 18 2015

I - 1 Motivations - Stochastic Bandit Games

Problem : You want to earn as much as possible in casino

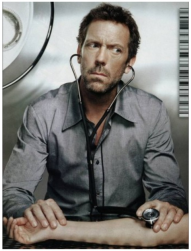


- You are in a casino and want to play with slot machines
- Each one can give you a potential gain, but these gains are not equivalent
- You *sequentially* play with one of the arms of the bandit machine

How to design a good policy to sequentially optimize the gain ?

I - 1 Motivations - Dynamic Ressource Allocation

Problem : Optimization of a sequence of clinical trials




Imagine you are a doctor :

- A sequence of patients visit you *sequentially* (one after another) for a given disease
- You choose one treatment/drug among (say) 5 availables
- The treatments are not equivalent
- You do not know where is the best drug, but you observe the effect of the prescribed treatment on each patient
- You expect to find the best drug despite some uncertainty on the effect of each treatment

How can we design a good sequence of clinical trials ?

I - 1 Motivations - Dynamic Ressource Allocation

Problem : “Fast fashion” retailer

Source : Farias & Madan, Operation Research, Vol. 9, No 2, 2011

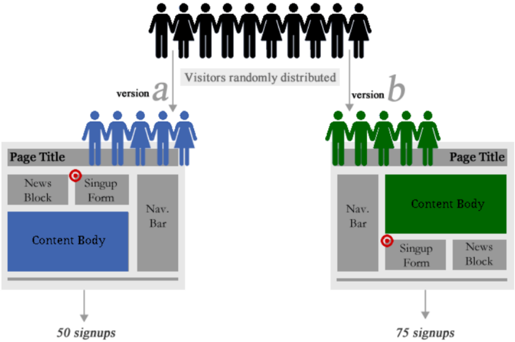
Imagine you are a firm solding clothes :

- A population of customers visit you *sequentially* (one after another) each week/day
- You observe weekly/daily sales and measure item's popularity
- You want to restock popular items and weed out unpopular ones *on-line*
- You expect to maximize your benefit while finding the best items

How can we design a good sequence of fast-fashion operations ?

I - 1 Motivations - Dynamic Ressource Allocation

Problem : "Web design"



Imagine you want to select a web page design

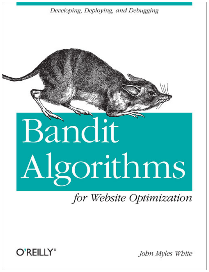
- A population of customers visit you *sequentially* (one after another)
- You randomly propose two designs a and b and measure design's popularity through the signups you obtain
- You want to propose the popular design to maximize your benefit

How can we build a good sequence of webpage propositions ?

I - 1 Motivations - Dynamic Ressource Allocation

Other motivating examples

- Pricing a product with uncertain demand to maximize revenue
- Trading (sequentially allocate a ratio of fund to the more efficient trader)
- Recommender systems :
  - advertisement
  - website optimization
  - news, blog posts



- Computer experiments
  - A code can be simulated in order to optimize a criterion
  - This simulation depends on a set of parameters
  - Simulation is costly and only few choices of parameters are possible

I - 1 Motivations - Exploration vs. Exploitation

Scientist view : Explore new ideas

Businessman view : Exploit best idea found so far

I - 2 Stochastic multi-armed bandit model

Environment :

- At your disposal : d arms with unknown parameters $\theta_1, \ldots, \theta_d$.
- For any time t, your choice is described by a variable $I_t \in \{1 \ldots, d\}$
- For anyu time t, you receive a reward, that depends on your choice I_t :

$$A_t^{I_t}$$

  For example :
  - it corresponds to the money obtained by sampling one specific slot machine in a Casino, the number of the machine is I_t.
  - it corresponds to the size of a tumor after choosing to test one drug on a patient.

Reward distribution :

- Of course, the rewards cannot be reasonnably assumed to be deterministic (otherwise I won't be there to talk about it !)
- For a fixed choice of one arm i, the rewards are i.i.d.

$$(A_t^i)_{t \geqslant 0} \sim \nu_{\theta_i}.$$

- Important assumption : the reward distributions ν_θ belong to a parametric family of probability distributions (Exponential, Poisson, . . .)

I - 2 Stochastic multi-armed bandit model

In this talk, we study the simplest case of Bernoulli rewards $\nu_p = \mathcal{B}(p)$:

- you obtain a gain of 1 with probability p
- 0 otherwise (with probability $1-p$).

What is unknown, the several probability of success : $(p_1, \ldots, p_d)$.
Without l.o.g., we assume that the first arm is the best one :

$$p_1 > \max_{2 \leqslant j \leqslant d} p_j .$$

Admissible policy :

- The agent's action follow a dynamical strategy, which is defined on-line :

$$I_t = \pi \left(A_{t-1}^{I_{t-1}} \ldots , A_1^{I_1} \right) .$$

  It means that at step t, we can use all the informations gathered from time 1 to time $t-1$ to make our decision I_t.
- The decision I_t can be driven either by
  - a deterministic function
  - a random function

  of the information from 1 to $t-1$.

Final goal : Maximize (in expectation) the cumulative rewards :

$$\mathbb{E}\left[\sum_{t=1}^{n} A_t^{I_t} \right] .$$

I - 3 Regret of Stochastic multi-armed bandit algorithms

Regret of an algorithm

Given an horizon n, we are naturally driven to minimize the expected regret R_n :

$$\mathbb{E}[R_n] = \mathbb{E} \max_{1 \leqslant j \leqslant d} \sum_{t=1}^{n} A_t^j - \mathbb{E} \sum_{t=1}^{n} A_t^{I_t} = \mathbb{E} \max_{1 \leqslant j \leqslant d} \sum_{t=1}^{n} (A_t^j - A_t^{I_t}).$$

- R_n is the maximal gain that could have been obtained minus our gain following our policy $(I_t)_{t \leqslant n}$.
- The expectation of the maximum makes the regret difficult to handle, but...

Pseudo-Regret of an algorithm

Proposition (Pseudo-regret)

If we define $\bar{R}_n := \max_{1 \leqslant j \leqslant d} \mathbb{E}\left[\sum_{t=1}^{n} (A_t^j - A_t^{I_t})\right]$, *one has*

$$\bar{R}_n \leqslant \mathbb{E}R_n \leqslant \bar{R}_n + \sqrt{\frac{n \log d}{2}}.$$

Advantage of the pseudo-regret : from a mathematical point of view, we know what arm is better than others, making $\bar{R}_n$ easier than R_n to handle.

I - 3 Regret of Stochastic multi-armed bandit algorithms

What kind of performances to expect ?

- Of course, $\mathbb{E}[R_n]$ and $\bar{R}_n$ increase with n !
- If our strategy fails to discover the best arm, it means that

$$I_t \neq 1 \qquad \text{infinitely often as } t \longrightarrow +\infty.$$

  It leads to

$$n \times (p_1 - \max_{j \geqslant 2} p_j) \lesssim \bar{R}_n,$$

  which is linear with n.
- We can expect much more better results if the strategy discovers the best arm.

- Proposition (Lower bound - (Auer, Cesa-Bianchi,Freund,Schapire 2002))

Uniformly among all policies π and among all Bernoulli distribution rewards :

$$\min_{\pi} \left\{ \sup_{2 \leqslant j \leqslant d} \max_{p_j < p_1} \mathbb{E} R_n \right\} \geqslant \frac{\sqrt{nd}}{20}.$$

This two propositions show that a strategy such that

$$\bar{R}_n \lesssim C_d \sqrt{n}$$

is a good one ($C_d \sim \sqrt{d}$).

I - 4 Roadmap

In this talk, we will :

- Briefly describe a standard *old-fashioned method*

$$X_{t+1} = X_t + \gamma_{t+1} h(X_t) + \gamma_{t+1} \Delta M_{t+1}$$

- Introduce a new one whose regret will be studied :

$$\forall n \in \mathbb{N}^* \qquad \bar{R}_n \leqslant C\sqrt{n}?$$

- Provide an asymptotic limit of this *penalized bandit* up to a correct scaling

$$\beta_n(X_n - \delta_1) \xrightarrow[n \to +\infty]{w^*} \mu$$

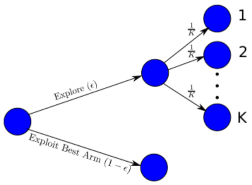
- Describe ergodic properties of the rescaled process (PDMP)

Important features of efficient algorithms :

- Fast decision from t to $t+1$ to do not slow down motion of the sequential rewards
- Adaptive with the horizon time : good strategies should no depend on the a fixed horizon time n and may be fully recursive.
- Efficient regret rate

II - 0 Some already existing method - ϵ-greedy'98

Widely used ϵ-**greedy algorithm**



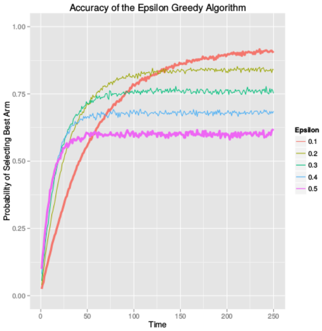
- Consider $\epsilon > 0$ and an initial guess of the ability of each arm :

$$\hat{p}_j(0) \qquad \text{is a prior information on} \qquad p_j$$

  If no information, take pick each p_j at random for example.
- Step t to $t+1$:
  - With probability $1-\epsilon$, use (one of) the best arm
  - With probability ϵ/d, pick an arm uniformly among all possibles.
  - Upgrade the estimators of the Bernoulli parameters with the empirical means $\hat{p}_j(t+1)$.
- Usually, $\epsilon = 0.1$.

II - 0 Some already existing method - ϵ-greedy'98

With 5 Bernoulli reward probabilities : $[0.1, 0.1, 0.1, 0.1, 0.9]$

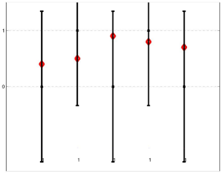


- $\epsilon = 0.1$: Businessman and
  - Learns slowly
  - Does well at the end
- $\epsilon = 0.5$: Scientist and
  - Learns quickly
  - Does not exploit at the end

Whatever ϵ is, linear regret with n.

II - 0 Some already existing method - Upper-confidence bounds'85

Popular methods that rely on the heuristic principle of optimism.



Strategy :

- Build a confidence bound around each empirical estimation of the probability of success

$$\hat{p}_i(t) \in [l_i(t); u_i(t)], \forall 1 \leqslant i \leqslant d$$

- at time t, select the arm with the highest upper confidence bound :

$$I_t = \arg\max u_i(t).$$

- Get the reward, and update the empirical estimator and the confidence bounds

$$\hat{p}_i(t+1) \in [l_i(t+1); u_i(t+1)]$$

UCB-like algorithm are shown to be optimal and satisfy

$$\limsup_{n \longrightarrow +\infty} \frac{\mathbb{E}\bar{R}_n}{\log n} \leqslant \sum_{p<p_1} \frac{1}{2(p_1 - p)}.$$

and

$$\forall (p_1, \ldots, p_d) \in [0,1]^d \qquad \bar{R}_n \lesssim \sqrt{d \log(d) n}$$

II - 1 An historical algorithm'69

The so-called Narendra-Shapiro bandit algorithm (NSa for short) defines a probability vector of $\mathcal{S}_d$

$$X_t = (X_t^1, \ldots, X_t^d) \qquad | \qquad \sum_{j=1}^{d} X_t^j = 1.$$

Idea : Use X_t to sample one arm at step t and then upgrade this probability X_t.

- In the two-armed situation with $p_2 < p_1$, denote $X_t = (x_t, 1 - x_t)$
- $X_t(1) = x_t$ is the probability to choose the first arm at step t.
- $X_t(2) = 1 - x_t$ is the probability to choose the second arm at step t.
- Upgrade formula

$$x_{t+1} = x_t + \begin{cases} \gamma_{t+1}(1 - x_t) & \text{if player 1 is selected and wins} \\ -\gamma_{t+1} x_t & \text{if player 2 is selected and wins} \\ 0 & \text{otherwise} \end{cases}$$

- Common step size :

$$\gamma_t = (1 + t/C)^{-\alpha}, \qquad \alpha \in (0,1) \qquad \text{with large enough } C.$$

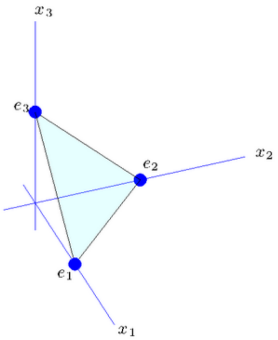
- Same idea :
  - If you win : reinforce the probability to sample I_t w.r.t. the remaining weights $(X_t^j)_{j \neq I_t}$ and decrease the probability to sample the other arms accordingly.
  - If you loose $(A_t^{I_t} = 0)$: do nothing.

II - 1 An historical algorithm'69

- Multi-armed situation, I_t : arm sampled at time t, $A_t^{I_t}$: obtained reward. Upgrade

$$\forall j \in \{1 \dots d\} \qquad X_t^j = X_{t-1}^j + \gamma_t \left[\mathbf{1}_{\{I_t = j\}} - X_{t-1}^j \right] A_t^{I_t}$$

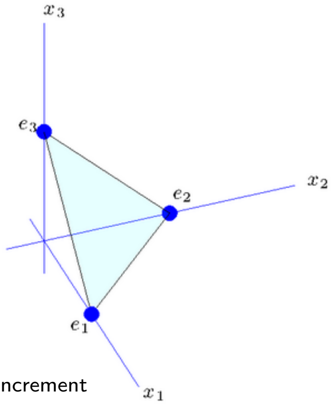
- To sum up :
  - If you win : reinforce the probability to sample I_t and decrease the probability of others.
  - If you loose ($A_t^{I_t} = 0$) : do nothing.



II - 1 An historical algorithm'69

Few words about NSa :

- Recursive stochastic algorithm
- Anytime policy
- Involves nontrivial mathematical difficulties



It can be written as mean drift + martingale increment

$$X_{t+1} = X_t + \gamma_{t+1} h(X_t) + \gamma_{t+1} \Delta M_{t+1}.$$

In the 2-armed setting ($p_2 < p_1$ and $X_t = (x_t, 1 - x_t)$), the drift on x_t is

$$h(x) = (p_1 - p_2) x (1 - x).$$

II - 1 An historical algorithm'69

Some keywords about this class of recursive algorithms ?

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

A lot is known on these Robbins-Monro (Kiefer-Wolfowitz) algorithms when :

- b is a deterministic drift and we are looking for the solution $b(x) = 0$. Standard applications : recursive quantile estimation.
- b is a gradient of a convex function U and we are looking for a minimum of U. Standard applications : Stochastic Gradient Descent (SGD).

What is known about this class of recursive algorithms ?

- Old results (Robbins, Polyak, ...) : if U is strongly convex, we can expect some non asymptotic upper bound

  $$\mathbb{E}\left[U(X_n) - \min U\right] \leqslant C\epsilon_n,$$

  where ϵ_n is a rate that should be related to the step size sequence $(\gamma_n)_{n\geqslant 1}$.
- Woodroofe'72 : Large deviation inequalities for SGD.
- Polyak averaging optimal (in the Cramer-Rao sense) of these methods.

Baseline assumption : strict convexity of U !

II - 1 An historical algorithm'69

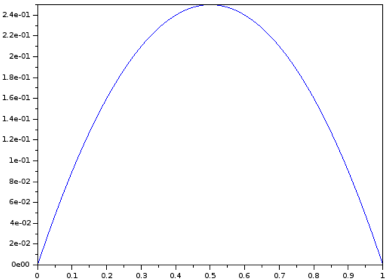
$$X_t = (x_t, 1 - x_t)$$

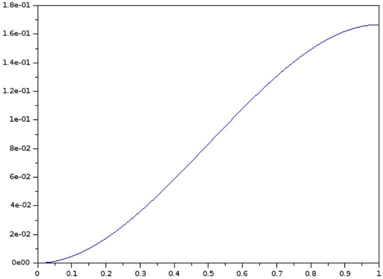
$$x_{t+1} = x_t + \gamma_{t+1} b(x_t) + \gamma_{t+1} \Delta M_{t+1}.$$

with

$$b(x) = (p_1 - p_2) x (1 - x)$$

The drift b has 2 zeros... The energy function is far from being convex !





- O.D.E. approximation $\dot{x} = h(x)$, local trap at $\{0\}$ and stable equilibrium at $\{1\}$.
- Robbins-Monro's argument : convergence to a either $\{0\}$ or $\{1\}$.
- But : the conditional variance term vanishes at 0 and 1, making impossible the use of Duflo's argument about the escape of local traps.
- Indeed, for any sequence $\gamma_t = \left(\frac{C}{t+C}\right)^{\alpha}, \qquad \alpha \in (0,1)$, the algorithm is fallible

II - 2 Improvement through penalization

- What's wrong with NSa ?
  Gittins, JRSS(B)'79 :

  Good regret properties only occur with an exploration/exploitation trade-off...
- NSa is almost a pure exploitation method : no exploration term to exit local traps.
- Main idea : Introduce a penalty term [Lamberton & Pages, EJP'09]
- In the 2-armed settings ($p_2 < p_1$ and $X_t = (x_t, 1 - x_t)$) :

$$X_{t+1} = X_t + \begin{cases} +\gamma_{t+1}(1 - X_t) & \text{if arm 1 is selected and wins} \\ -\gamma_{t+1}X_t & \text{if arm 2 is selected and wins} \\ -\rho_{t+1}\gamma_{t+1}X_t & \text{if arm 1 is selected and loses} \\ +\rho_{t+1}\gamma_{t+1}(1 - X_t) & \text{if arm 2 is selected and loses} \end{cases}$$

When one arm fails, decrease the probability to sample it.

LP'09 : Up to technical conditions on (ρ_t, γ_t) : penalized 2-armed bandit is infallible (a.s. convergence to the good target)

II - 3 Over-penalized NSa

This additional penalty term will be inefficient from the minimax regret point of view.
As a last resort : increase the penalty effect to reinforce the escape from local traps :

$$X_{t+1} = X_t + \begin{cases} +\gamma_{t+1}(1 - X_t) - \rho_{t+1}\gamma_{t+1}X_t & \text{if arm 1 is selected and wins} \\ -\gamma_{t+1}X_t + \rho_{t+1}\gamma_{t+1}(1 - X_t) & \text{if arm 2 is selected and wins} \\ -\rho_{t+1}\gamma_{t+1}X_t & \text{if arm 1 is selected and loses} \\ +\rho_{t+1}\gamma_{t+1}(1 - X_t) & \text{if arm 2 is selected and loses} \end{cases}$$

Whatever happens with the selected arm, it is penalized (escape from local traps).

A multi-armed version :

$$\begin{aligned} X_t^j &= X_{t-1}^j + \gamma_t \left[\mathbf{1}_{I_t=j} - X_{t-1}^j\right] A_t^{I_t} \\ &\quad -\gamma_t \rho_t X_{t-1}^{I_t} \left[\mathbf{1}_{I_t=j} - \frac{1 - \mathbf{1}_{I_t=j}}{d-1}\right] \end{aligned}$$

II - 3 Over-penalized NSa and infallibility

Write $X_t = X_{t-1} + \gamma_t b(X_t) + \gamma_t \rho_t \kappa(X_t) + \gamma_t \Delta M_t$. Drift :

$$b^i(x_1, \ldots, x_d) = x_i \left[(1 - x_i) p_i - \sum_{j \neq i} x_j p_j \right], \forall i \in \{1, \ldots, d\}$$

Equilibria of $\dot{X} = h(X)$: Dirac masses on each arm. Stable one : $(1, 0, \ldots, 0)$.
The Kushner-Clarck theorem → a.s. convergence towards an equilibrium (which one ?)

Theorem (Infallibility of the Over-penalized NSa)

If $p_d \leqslant p_{d-1} \leqslant \ldots \leqslant p_2 < p_1$ and $\gamma_t = \gamma_1 t^{-\alpha}$, $\rho_t = \rho_1 t^{-\beta}$, then

$$0 \leqslant \beta \leqslant \alpha \quad \textit{and} \quad \alpha + \beta \leqslant 1 \Longrightarrow \lim_{t \to +\infty} X_t = (1, 0 \ldots, 0) \quad a.s.$$

<u>Sketch of proof :</u> The penalty term induced by κ is

$$\kappa^i(x) = -x_i^2 (1 - p_i) + \frac{1}{d-1} \sum_{j \neq i} x_j^2 (1 - p_j), \forall i \in \{1, \ldots, d\}$$

If $X_\infty^1 = 0$, $\kappa^1(X_\infty) > 0$ and :

- $$\alpha \leqslant \beta \Longrightarrow \limsup \frac{\sum_t \gamma_t \Delta M_t}{\sum \gamma_t \rho_t} \geqslant 0$$
- $$\alpha + \beta \leqslant 1 \Longrightarrow \sum \gamma_t \rho_t = +\infty \Longrightarrow \sum \gamma_t \rho_t \kappa(X_t) = +\infty$$

II - 4 Non-asymptotic upper bound of the regret

We detail the picture for the two-armed over-penalized NSa

$$
\begin{aligned}
\bar{R}_n &= \max_{j\in\{1,2\}} \mathbb{E} \sum_{t=1}^{n} A_t^j - A_t^{I_t} \\
&= \mathbb{E} \sum_{t=1}^{n} \left[p_1 - (X_t^1 p_1 + (1 - X_t^1) p_2) \right] \\
&= (p_1 - p_2) \sum_{t=1}^{n} \rho_t \underbrace{\frac{1 - X_t^1}{\rho_t}}_{:= Y_t}
\end{aligned}
$$

$$X_{n+1} = X_n + \gamma_n \nabla U(X_n) + \gamma_n \Delta M_{n+1},$$

In S.A., we expect a "Central Limit Theorem" for the renormalized sequence

$$\sqrt{\gamma_n}(X_n - \arg\min U) \xrightarrow[n\to+\infty]{w^*} \mathcal{N}(0, \sigma_U^2).$$

A good news ? Be able to do the same for the sequence $(Y_t)_{t\geqslant 1}$:

$$Y_n \xrightarrow[n\to+\infty]{w^*} \mu.$$

II - 4 Non-asymptotic upper bound of the regret

We detail the picture for the two-armed over-penalized NSa

$$\begin{aligned}
\bar{R}_n &= \max_{j\in\{1,2\}} \mathbb{E} \sum_{t=1}^{n} A_t^j - A_t^{I_t} \\
&= \mathbb{E} \sum_{t=1}^{n} \left[p_1 - (X_t^1 p_1 + (1 - X_t^1)p_2)\right] \\
&= (p_1 - p_2) \sum_{t=1}^{n} \rho_t \underbrace{\frac{1 - X_t^1}{\rho_t}}_{:=Y_t}
\end{aligned}$$

If the measure μ has a finite first moment, we can expect

$$\sup_{n\geqslant 1} \mathbb{E} Y_n < \infty,$$

which implies in turn

$$\bar{R}_n \lesssim \sum_{t=1}^{n} \rho_t.$$

Find β in $\rho_t = \rho_1 t^{-\beta}$ as large as possible s.t. $\beta \leqslant \alpha, \alpha + \beta \leqslant 1$. Optimal calibration :

$$\gamma_t = \frac{\gamma_1}{\sqrt{t}} \qquad \text{and} \qquad \rho_t = \frac{\rho_1}{\sqrt{t}}.$$

II - 4 Non-asymptotic upper bound of the regret

We are turned to the random dynamical system induced by $(Y_t)_{t\geqslant 1}$. Again :

$$Y_{t+1} = Y_t + \gamma_t \varphi_t(Y_t) + \gamma_t \Delta M_{t+1}.$$

Beyond the analytic formula of φ_t, a simple picture :

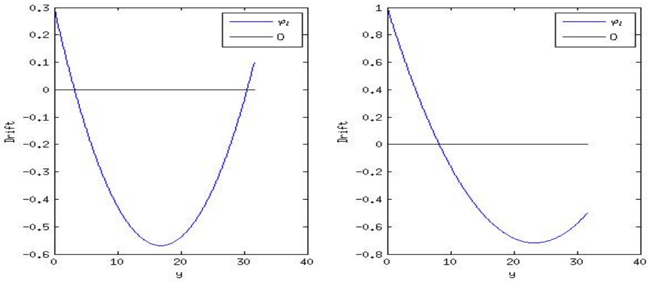


FIGURE : Drift for non penalized (left) and overpenalized (right) NSa when $y \in [0, \rho_t^{-1}]$.

To control the increments of Y_t, the right situation is much better :

Large value of Y_t are naturally decreased by φ_t

II - 4 Non-asymptotic upper bound of the regret

- Difficulty : obtaining a uniform bound over all the values $0 \leqslant p_2 < p_1 \leqslant 1$.
- Lyapunov arguments and painful computations lead to non asymptotic bound.
- Key quantity that induces the understanding of the good scaling

$$\pi = p_1 - p_2.$$

Theorem (Upper bound of the regret : 2-armed over-penalized NSa)

$$\forall n \in \mathbb{N} \qquad \sup_{p_2 < p_1} \bar{R}_n \leqslant 30\sqrt{2n}.$$

Optimal settings : $\gamma_n = \frac{9}{10\sqrt{n}}$ and $\rho_n = \frac{1}{3\sqrt{n}}$.

Sketch of proof :

Define $Z_t^{(r)} = \frac{(1-X_t)^r}{\gamma_t}$ and exhibit a mean-reverting effect for r sufficiently large

$$\mathbb{E}[Z_{t+1}^{(r)} | \mathcal{F}_t] = Z_t^{(r)} + P_{t,r}(Z_t^{(r)}).$$

- Find r such that $P_{t,r}$ is negative on $[C(\gamma_t, \pi), \gamma_t^{-1}]$ where $C(\gamma_t, \pi) = o(\gamma_t^{-1})$ and

$$\sup_{t \geqslant 0} \mathbb{E}[Z_t^{(r)}] < \infty.$$

- Exhibit a recursion between $\mathbb{E}[Z_t^{(r)}]$ and $\mathbb{E}[Z_t^{(r-1)}]$ for a result on $\sup_{t \geqslant 0} \mathbb{E}[Y_t]$

II - 4 Numerical simulations

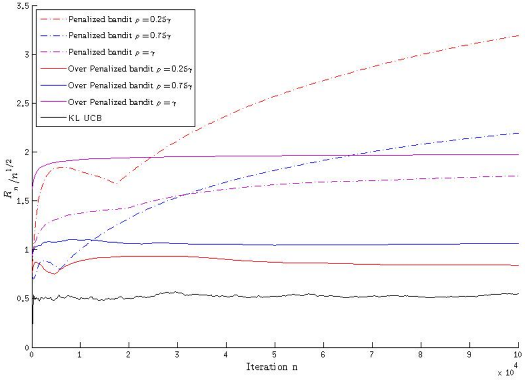


FIGURE : Evolution of $n \mapsto \sup_{(p_1,p_2)\in[0,1],p_2\leqslant p_1} \frac{\bar{R}_n}{\sqrt{n}}$ for over-penalized NSa (continuous colored line) and penalized NSa (dashed colored line) and KL UCB (black line).

- Over-penalization is important for a competitive regret
- Practical : $\bar{R}_n \leqslant \sqrt{n}$ - Theoretical : $\bar{R}_n \leqslant 30\sqrt{2n}$
- Defeated by UCB-like algorithms for the regret point of view ($\bar{R}_n \leqslant \sqrt{n}/2$)
- Much more faster than MOSS or UCB-like algorithms (1/100 of time).

II - 4 Numerical simulations

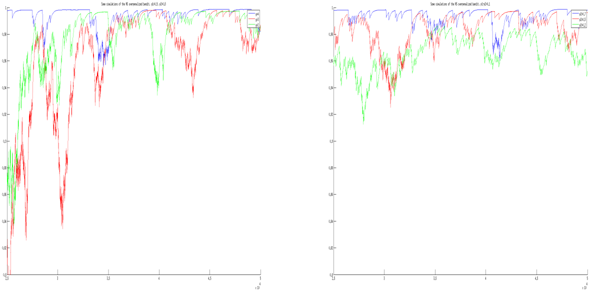


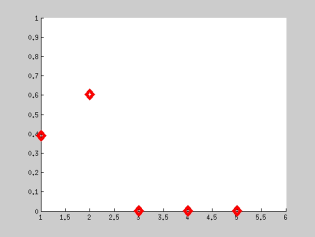
FIGURE : Evolution of the probability of Arm 1 (best one) with respect to n while $p_1 - p_2 = 0.1$. Left : ρ_1/γ_1 is varying. Right : p_2 is increasing.

Seems to behave quite particularly (maybe after a good rescaling ?)

- Some jumps randomly distributed ? (more or less frequent according to the parameters)
- Almost deterministic evolution between jumps when n is large

II - 4 Numerical simulations

Time for a short movie . . .
5 arms, $p = [0.9, 0.88, 0.8, 0.75, 0.7]$.



Let's go back to the mathematics . . .

III - 1 Rescaling

We fix $p_1 > \max(p_2, \dots, p_d)$, the "good" rescaling of what is left over by X_n^1 is

$$\bar{X}_n = \frac{(X_n^2, \dots, X_n^d)}{\rho_n}$$

Proposition

For any $f \in \mathcal{C}^2(\mathbb{R}^{d-1}, \mathbb{R})$:

$$\mathbb{E}\left[f(\bar{X}_{n+1})|\mathcal{F}_n\right] = f(\bar{X}_n) + \gamma_{n+1}\mathcal{L}_d(f)(\bar{X}_n) + o_P(\gamma_{n+1}),$$

where $\mathcal{L}_d$ is the Markov generator given by

$$\mathcal{L}_d(f)(\bar{x}) = \sum_{j=2}^{d} \underbrace{\frac{p_j}{g}\bar{x}_j}_{\text{jump rate}} \underbrace{\left[f(\bar{x} + g\mathbf{1}_j) - f(\bar{x})\right]}_{\text{jump size}} + \sum_{j=2}^{d} \underbrace{\left[\frac{1-p_1}{d-1} - p_1\bar{x}_j\right]\partial_j f(\bar{x})}_{\text{deterministic part}}.$$

- The amount of jump is low when $g = \frac{\gamma_1}{\rho_1}$ is large (seen in simulations).
- The size of jumps is large when g is large.

III - 1 Rescaling

As a tensorized process, it is enough to study the following Markov generator :

$$\mathcal{L}(f)(\bar{x}) = (a - b\bar{x})f'(\bar{x}) + cx[f(\bar{x} + g) - f(\bar{x})]$$

- Family of Piecewise Deterministic Markov Process (PDMP for short)
- Random dynamical systems with an increasing interest (encountered in many modelisation problems)
- Famous examples (among many others) :
  - Telegraph process [Kac, '74]
  - Storage models [Roberts & Tweedie,'00]
  - Randomly switched ODE [Benaïm et al.,'14] & Parrondo-like paradox
  - TCP models [Guillin, Malrieu et al.'13, Cloez & Hairer'13]

What the dynamic looks like exactly in the over-penalized NSa case ?

- Set

$$a = \frac{1-p_1}{d-1}, b = p_1, c_j = \frac{p_j}{g}, g = \frac{\gamma_1}{\rho_1}$$

- Between jumps, the evolution is deterministic and follow a differential flow

$$\dot{\phi}(\xi, t) = \left[\frac{1-p_1}{d-1} - p_1\xi\right]\partial_\xi\phi(\xi, t)$$

- Poisson jumps with an instantaneous average push of $\frac{p_j}{g}\bar{x}_j \times g$.

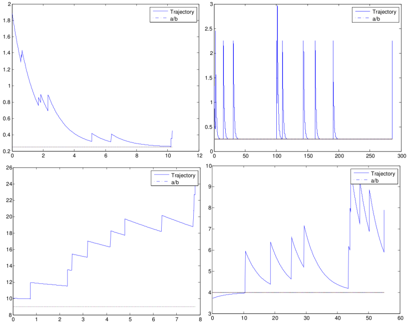
Here, the size of the jumps are deterministic.

III - 2 Trajectories of the rescaled over-penalized NSa

- $\mathcal{L}_d$ acts as a tensorized Markov generator on each coordinate.
- The problem is reduced to the study of the random dynamic system described by

$$\mathcal{L}(f)(\bar{x}) = (a - b\bar{x})f'(\bar{x}) + cx[f(\bar{x} + g) - f(\bar{x})],$$

- Examples of rescaled trajectories for several values of (a, b, c, g)



- Asymptotic direction : a/b. Bottom left : transient behaviour when $cg > b$ ... but in the bandit algorithm

$$cg - b = p_j - p_1 < 0 \quad (!)$$

III - 3 Ergodicity and Invariant measure

Ergodicity can be helpful to derive confidence bounds. It requires to obtain some mixing properties around an/the invariant measure.

$$\mathcal{L}(f)(\bar{x}) = (a - bx)f'(\bar{x}) + cx[f(\bar{x} + g) - f(\bar{x})],$$

For over-penalized NSa, the process should be studied only when $cg - b < 0$.

Proposition (Invariant measure - rescaled over-penalized NSa)

The PDMP $\bar{X}_t$ has a unique invariant measure μ supported by

$$\left[\frac{1-p_1}{p_1(d-1)}, +\infty\right[^{d-1}.$$

Sketch of proof : existence and uniqueness through a Lyapunov certificate :

$$\mathcal{L}(Id) = a - (b - cg)Id.$$

But . . . Some real difficulties :

- No explicit formula for μ . . . We are far from a standard CLT with a Gaussian distribution and even far from the simplest case of the TCP process
- Less is known about the smoothness of μ . . . Intricate situation as pointed by [Bakhtin & Hurth & Mattingly '14].

III - 4 Ergodicity and mixing rate

$\mathcal{L}$ is a non-reversible Markov operator, which is usual for this kind of kinetic models
The question : Obtaining an upper bound of the mixing rate :

$$d(L(X_t), \mu) \leqslant \epsilon(t) \longrightarrow 0 \quad \text{as} \quad t \longrightarrow +\infty.$$

- Traditional distance

$$\|L(X_t) - \mu\|_{\mathbb{L}^2(\mu)\circlearrowleft} = \sup_{f\,:\,\|f\|_{\mathbb{L}^2(\mu)}=1} \|\mathbb{E}[f(\bar{X}_t^x)] - \mu(f)\|_{\mathbb{L}_2(\mu)}$$

  Non-reversible generators : difficult to handle with the $\mathbb{L}^2$ distance, require informations on μ (Modified norms [Villani,'09], Lie brackets [Gadat & Miclo'13])
- Resort less sophisticated distances induced by trajectorial properties (instead of functional ones)
  Wasserstein distance :

$$\mathcal{W}_p(\nu_1, \nu_2) = \inf\left\{\mathbb{E}\left((X-Y)^p\right)^{\frac{1}{p}} \mid L(X) = \nu_1, L(Y) = \nu_2\right\}$$

  Total Variation distance :

$$d_{TV}(\nu_1, \nu_2) = \max_{\Omega \subset E} |\nu_1(\Omega) - \nu_2(\Omega)|$$

- Use some coupling techniques to derive quantiative bounds

III - 4 Ergodicity and mixing rate

The simple idea :

- Build a non independent coupling $(\bar{X}_t, Y_t)$ such that $\bar{X}_t$ and Y_t follow the dynamic given by $\mathcal{L}$ and $Y_0 \sim \mu$
- Try to make $\bar{X}_t$ and Y_t close to each others for the Wasserstein results

Theorem (Wasserstein ergodicity)

An explicit constant γ_p exists such that

$$\mathcal{W}_p(L(\bar{X}_t), \mu) \leqslant \gamma_p e^{-t\pi/p},$$

where $\pi = p_1 - p_2$ is the difference between the 2 probabilities of success of the 2 best arms

Optimal for $\mathcal{W}_1$. Open questions for $\mathcal{W}_p$.

- Try to make the two processes $\bar{X}_t = Y_t$ stucked rapidly for the TV results

Theorem (Total Variation ergodicity)

Some explicit constants C and α exist such that

$$d_{TV}(L(\bar{X}_t), \mu) \leqslant Ce^{-\alpha\pi t}.$$

Suspected to be far from the optimal exponents.

IV Conclusion

Statistics :

- Standard NSa Algorithm is fallible ...
- Penalized bandits are infallible
- Over-penalization : relevant for regret bounds
- Over-penalization : traduces a vanishing repelling effect on each corner of the simplex.
- Minimax result in the two-armed case :

$$\bar{R}_n \leqslant C\sqrt{2n},$$

- Much more faster than what is already existing in Bandit methods while statistically competitive (not as good as KL UCB)

Probability :

- Rescaled process as a PDMP.
- Random jumps come from the binary rewards given by each arm.
- Ergodic properties

Anecdotal :

- Used in some trading firms in « La Defense » ...

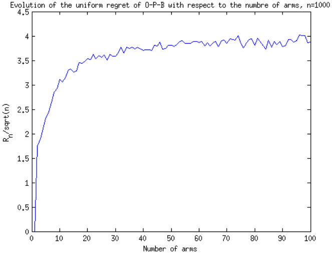
IV Conclusion

Open questions :

- Regret with d arms ? Numerical simulations lead to the conjecture

$$\bar{R}_n \leqslant C\sqrt{dn},$$

which is the known minimax rate for d-armed bandit.



Over-Penalized NSa seems to behave well ...

- What should be a generalization of Over-Penalized NSa for continuous rewards ? What is the rescaled process (suspected to be a diffusion instead of a jump process ...)
- Many challenging questions with the PDMP :
  - Spectral results and $\mathbb{L}^2$ convergence
  - Wasserstein lower bounds
  - Smoothness of the invariant measure

Thank you for your attention