On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems

Aurélien Garivier, Eric Moulines, LTCI CNRS Telecom ParisTech

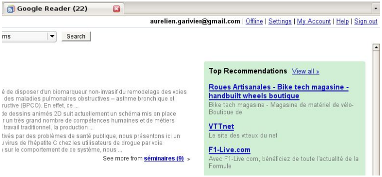
June 27, 2008

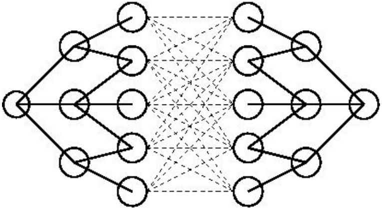
Outline

Outline

Motivating situations

- Clinical trials
- (PASCAL challenge: cf Showe-Taylor '07) Web: advertising and news feeds
- Web routing, (El Gamal, Jiang, Poor '07) Communication networks
- Economics, Auditing, Labor Market,. . .

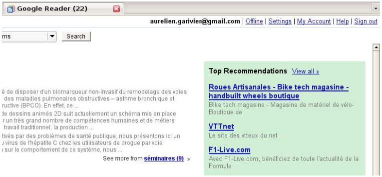


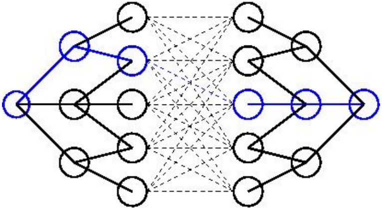


$\Longrightarrow$ Exploration versus Exploitation Dilemma

Motivating situations

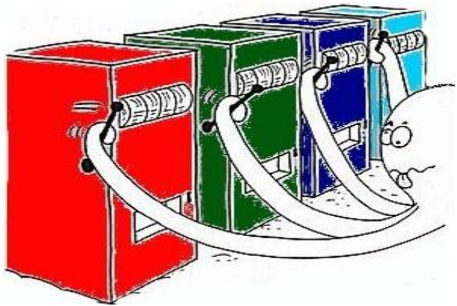
- Clinical trials
- (PASCAL challenge: cf Showe-Taylor '07) Web: advertising and news feeds
- Web routing, (El Gamal, Jiang, Poor '07) Communication networks
- Economics, Auditing, Labor Market,. . .





$\Longrightarrow$ Exploration versus Exploitation Dilemma

Idealized Problem



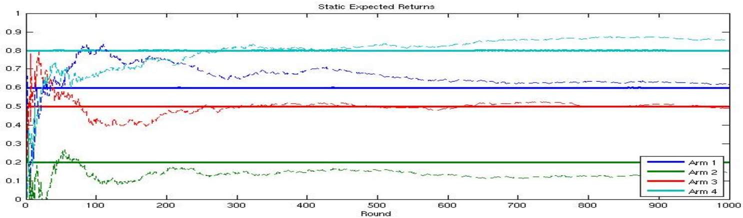
The rewards $X_t(i) \in [0, B]$ of arm i at times $t = 1, \ldots, n$ are independent with expectation $\mu_t(i)$.

At time t, a policy π:

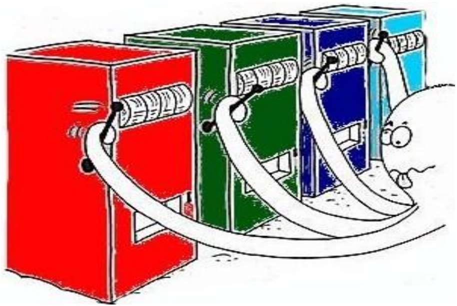
- chooses arm I_t given the past observed rewards;
- observes reward $X_t(I_t)$.

Goal: minimize expected regret

$$R_n(\pi) = \sum_{t=1..n} \mu_t(*) - \mu_t(I_t) \ .$$



Idealized Problem



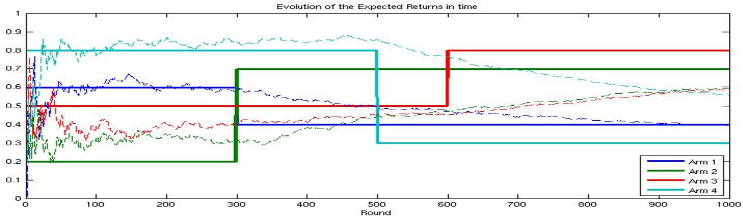
The rewards $X_t(i) \in [0, B]$ of arm i at times $t = 1, \dots, n$ are independent with expectation $\mu_t(i)$.

At time t, a policy π:

- chooses arm I_t given the past observed rewards;
- observes reward $X_t(I_t)$.

Goal: minimize expected regret

$$R_n(\pi) = \sum_{t=1..n} \mu_t(*) - \mu_t(I_t) \; .$$



The Stationary case: Methods

Classical policies:

1. **Softmax Methods** like EXP3: the arm I_t is chosen at random by the player according to some probability distribution giving more weight to arms which have so-far performed well
2. **UCB policies** arm I_t is chosen that maximizes the upper bound of a confidence interval for expected reward $\mu(i)$, which is constructed from the past observed rewards.

$$I_t = \arg\max_{1 \le i \le K} \bar{X}_t(i) + B\sqrt{\frac{\xi \log(t)}{N_t(i)}}.$$

The Stationary case: Results

1. Probabilistic setup:
   - (Lai,Robbins '85)
   $$R_n(\pi) \geq C \log n \,.$$
   - (Auer,Cesa-Bianchi,Fischer '02) rate $\log n$ reached by UCB;
   - Analysis of UCB: amounts to upper-bounding the expected number of times $\tilde{N}_t(i)$ a suboptimal arm i is played.
2. Adversarial setup:
   - (Auer, Cesa-Bianchi, Freund, Schapire '03)
   $$R_n(\pi) \geq C\sqrt{n} \,.$$
   - (Auer, Cesa-Bianchi, Freund, Schapire '03) rate reached by EXP3.
   - In a probabilistic setup, EXP3 usually has larger regret than UCB.

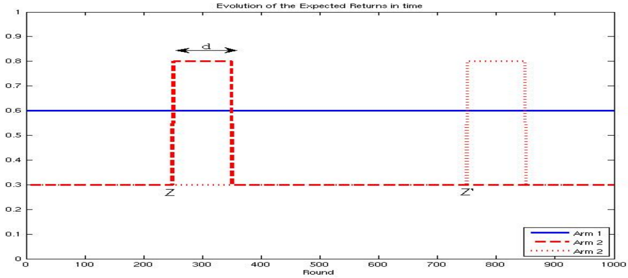
Non-stationary Policies

- Cf. results of PASCAL Exploration Vs Exploitation Challenge
- (Auer, Cesa-Bianchi, Freund, Schapire '03): EXP3.S
  - Tracking the best expert;
  - Randomized procedure working in an adversarial setup;
  - Analysis: extends EXP3
- (Szepeszvári, Koksis '06) Discounted UCB
  - Promising empirical results;
  - More difficult to analyze;
  - Problem: tuning of the discount factor?

Outline

1. The Non-stationary Bandit Problem
2. Results
   - A Lower-Bound
   - The Discounted UCB
   - The Sliding Windows UCB
3. Simulations, Conclusions and Perspectives

Setup of the Lower-bound



- The period $\{1, \ldots, T\}$ is divided into epochs of size $d \in \{1, \ldots, T\}$;
- The distribution of rewards is modified on $[Z+1, Z+d]$ (arm 2 becomes the one with highest expected reward).
- Composed game P^*:

$$\mathbb{E}^*_\pi[W] = \frac{1}{M} \sum_{Z \in d\mathbb{N}} \mathbb{E}^Z_\pi[W].$$

Lower-Bound and Consequences

- **Theorem:** For any policy π and any horizon T such that $64/(9\alpha) \leq \mathbb{E}_\pi[N_T(K)] \leq T/(4\alpha)$,

$$\mathbb{E}^*_\pi[R_T] \geq C(\mu)\frac{T}{\mathbb{E}_\pi[R_T]},$$

where $C(\mu) = \frac{32\delta(\mu(1)-\mu(K))}{27\alpha}$.

- **Corollary:** For any policy π and any positive horizon T,

$$\max\{\mathbb{E}_\pi(R_T), \mathbb{E}^*_\pi(R_T)\} \geq \sqrt{C(\mu)T} .$$

- **Remark:** as standard UCB satisfies $\mathbb{E}_\pi[N(K)] = \Theta(\log T)$,

$$\mathbb{E}^*_\pi[R_T] \quad \geq \quad c\frac{T}{\log T}.$$

Presentation of D-UCB

- Idea: give more weight to recent observations $\Longrightarrow$ *discount factor* γ
- Estimate $\mu_t(i)$ by the *discounted average*

$$\bar{X}_t(\gamma, i) = \frac{1}{N_t(\gamma, i)} \sum_{s=1}^{t} \gamma^{t-s} X_s(i) \mathbb{1}_{\{I_s = i\}}\ , \quad N_t(\gamma, i) = \sum_{s=1}^{t} \gamma^{t-s} \mathbb{1}_{\{I_s = i\}} .$$

- D-UCB policy: letting $n_t(\gamma) = \sum_{i=1}^{K} N_t(\gamma, i)$, choose

$$I_t = \arg\max_{1 \le i \le K} \bar{X}_t(\gamma, i) + 2B\sqrt{\frac{\xi \log n_t(\gamma)}{N_t(\gamma, i)}} .$$

- Compare to standard UCB:

$$I_t = \arg\max_{1 \le i \le K} \bar{X}_t(i) + B\sqrt{\frac{\xi \log(t)}{N_t(i)}} .$$

Bound on the regret

Theorem Let $\xi > 1/2$ and $\gamma \in (0,1)$. For any arm $i \in \{1,\dots,K\}$,

$$\mathbb{E}_\gamma \left[\tilde{N}_T(i)\right] \leq \mathrm{B}(\gamma) T(1-\gamma) \log \frac{1}{1-\gamma} + \mathrm{C}(\gamma) \frac{\Upsilon_T}{1-\gamma} \log \frac{1}{1-\gamma} ,$$

where

$$\mathrm{B}(\gamma) = \frac{16 B^2 \xi}{\gamma^{1/(1-\gamma)} (\Delta\mu_T(i))^2} \frac{\lceil T(1-\gamma) \rceil}{T(1-\gamma)} + \frac{2\left\lceil -\log(1-\gamma)/\log(1+4\sqrt{1-1/2\xi}) \right\rceil}{-\log(1-\gamma)\left(1-\gamma^{1/(1-\gamma)}\right)}$$

$$\to \frac{16\,\mathrm{e}\,B^2\xi}{(\Delta\mu_T(i))^2} + \frac{2}{(1-\mathrm{e}^{-1})\log\left(1+4\sqrt{1-1/2\xi}\right)}$$

and

$$\mathrm{C}(\gamma) = \frac{\gamma - 1}{\log(1-\gamma)\log\gamma} \times \log\left((1-\gamma)\xi \log n_K(\gamma)\right) \to 1 .$$

Consequences

- If horizon T and the growth rate of the number of breakpoints Υ_T are known in advance, take $\gamma = 1 - (4B)^{-1}\sqrt{\Upsilon_T/T}$:

$$\mathbb{E}_\gamma\left[\tilde{N}_T(i)\right] = O\left(\sqrt{T\Upsilon_T}\log T\right) .$$

  Assuming that $\Upsilon_T = O(T^\beta)$ for some $\beta \in [0,1)$, the regret is upper-bounded as $O\left(T^{(1+\beta)/2}\log T\right)$.

- In particular, if the number of breakpoints Υ_T is upper-bounded by Υ independently of T, taking $\gamma = 1 - (4B)^{-1}\sqrt{\Upsilon/T}$ the regret is bounded by

$$\mathbb{E}_\gamma\left[\tilde{N}_T(i)\right] = O\left(\sqrt{\Upsilon T}\log T\right) .$$

$\Longrightarrow$ D-UCB matches the lower-bound up to a factor $\log T$.

- If $\Upsilon_T \leq rT$ for a (small) positive constant r, taking $\gamma = 1 - \sqrt{r}/(4B)$ yields:

$$\mathbb{E}_\gamma\left[\tilde{N}_T(i)\right] = O\left(-T\sqrt{r}\log r\right) .$$

Insight into the analysis

$$\bar{X}_t(\gamma, i) = \mu_t(i) + \frac{\sum_{s=1}^{t} \gamma^{t-s}(\mu_s(i)-\mu_t(i))\mathbb{1}_{\{I_s=i\}}}{N_t(\gamma,i)} \quad \text{“Bias”} + \frac{\sum_{s=1}^{t} \gamma^{t-s}(X_s(i)-\mu_s(i))\mathbb{1}_{\{I_s=i\}}}{N_t(\gamma,i)} \quad \text{“Variance”}$$

- to control the bias term, abandon a few terms after each breakpoint;
- to control the variance term, new martingale bound: $\forall \eta > 0$,

$$\mathbb{P}\left(\left|\bar{X}_t(\gamma, i) - \frac{\sum_{s=1}^{t} \gamma^{t-s}\mu_s(i)\mathbb{1}_{\{I_s=i\}}}{N_t(\gamma, i)}\right| > \delta\sqrt{\frac{N_t(\gamma^2, i)}{N_t^2(\gamma, i)}}\right) \leq \left\lceil \frac{\log n_t(\gamma)}{\log(1+\eta)} \right\rceil \exp\left(-\frac{2\delta^2}{B^2}\left(1 - \frac{\eta^2}{16}\right)\right) .$$

Insight into the analysis

$$\bar{X}_t(\gamma, i) = \mu_t(i) + \frac{\sum_{s=1}^{t} \gamma^{t-s}(\mu_s(i) - \mu_t(i))\mathbb{1}_{\{I_s=i\}}}{N_t(\gamma,i)} + \frac{\sum_{s=1}^{t} \gamma^{t-s}(X_s(i) - \mu_s(i))\mathbb{1}_{\{I_s=i\}}}{N_t(\gamma,i)}$$

"Bias" "Variance"

- to control the bias term, abandon a few terms after each breakpoint;
- to control the variance term, new martingale bound:

$$\mathbb{P}\left(\left| \bar{X}_t(\gamma, i) - \frac{\sum_{s=1}^{t} \gamma^{t-s} \mu_s(i) \mathbb{1}_{\{I_s=i\}}}{N_t(\gamma, i)} \right| > \delta \sqrt{\frac{N_t(\gamma^2, i)}{N_t^2(\gamma, i)}} \right) \leq 4 \log n_t(\gamma) \exp\left(-\frac{1.99\delta^2}{B^2} \right) .$$

TELECOM ParisTech

Presentation of SW-UCB

- Idea: give weight only to recent observations $\Longrightarrow$ *sliding windows of width* τ
- Estimate $\mu_t(i)$ by the *local average*

$$\bar{X}_t(\tau, i) = \frac{1}{N_t(\tau, i)} \sum_{s=t-\tau+1}^{t} X_s(i) \mathbb{1}_{\{I_s = i\}} \,, \quad N_t(\tau, i) = \sum_{s=t-\tau+1}^{t} \mathbb{1}_{\{I_s = i\}} \,.$$

- SW-UCB policy: choose

$$I_t = \underset{1 \le i \le K}{\arg\max}\, \bar{X}_t(\tau, i) + B \sqrt{\frac{\xi \log(t \wedge \tau)}{N_t(\tau, i)}} \,.$$

- Compare to standard UCB:

$$I_t = \underset{1 \le i \le K}{\arg\max}\, \bar{X}_t(i) + B \sqrt{\frac{\xi \log(t)}{N_t(i)}}.$$

Bounds on the regret

Theorem Let $\xi > 1/2$. For any integer τ and any arm $i \in \{1, \dots, K\}$,

$$\mathbb{E}_\tau \left[\tilde{N}_T(i) \right] \leq C(\tau) \frac{T \log \tau}{\tau} + \tau \Upsilon_T + \log^2(\tau) \ ,$$

where

$$C(\tau) = \frac{4B^2\xi}{(\Delta\mu_T(i))^2} \frac{\lceil T/\tau \rceil}{T/\tau} + \frac{2}{\log \tau} \left\lceil \frac{\log(\tau)}{\log(1 + 4\sqrt{1 - (2\xi)^{-1}})} \right\rceil$$
$$\to \frac{4B^2\xi}{(\Delta\mu_T(i))^2} + \frac{2}{\log(1 + 4\sqrt{1 - (2\xi)^{-1}})} \ .$$

Consequences

- If horizon T and the growth rate of the number of breakpoints Υ_T are known in advance, take $\tau = 2B\sqrt{T\log(T)/\Upsilon_T}$:

$$\mathbb{E}_\tau\left[\tilde{N}_T(i)\right] = O\left(\sqrt{\Upsilon_T T \log T}\right) .$$

  Assuming that $\Upsilon_T = O(T^\beta)$ for some $\beta \in [0,1)$, the regret is upper-bounded as $O\left(T^{(1+\beta)/2}\sqrt{\log T}\right) \implies$ slightly better than D-UCB.

- In particular, if the number of breakpoints Υ_T is upper-bounded by Υ independently of T, taking $\tau = 2B\sqrt{T\log(T)/\Upsilon}$ the regret is bounded by

$$\mathbb{E}_\gamma\left[\tilde{N}_T(i)\right] = O\left(\sqrt{\Upsilon T \log T}\right) .$$

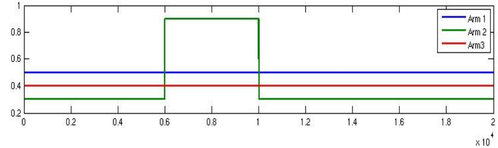
  $\implies$ SW-UCB matches the lower-bound up to a factor $\sqrt{\log T}$.

- If $\Upsilon_T \leq rT$ for a (small) positive constant r, taking $\tau = 2B\sqrt{-\log r/r}$ yields:

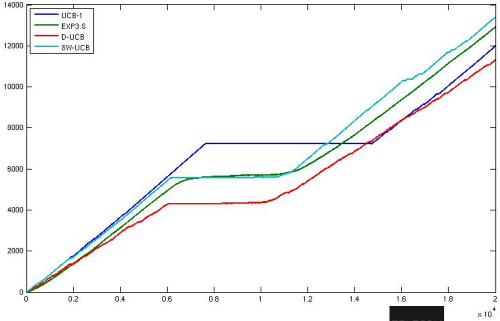
$$\mathbb{E}_\tau\left[\tilde{N}_T(i)\right] = O\left(T\sqrt{-r\log(r)}\right) .$$

Outline

Bernoulli MAB problem with two swaps

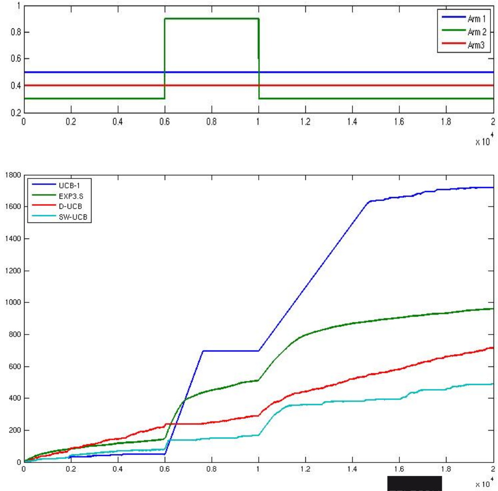


Evolution of the expected rewards



Cumulative frequency of arm 1 pulls

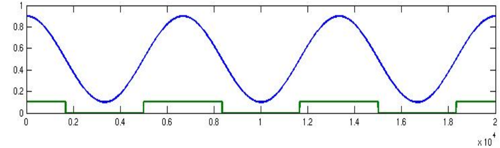
Bernoulli MAB problem with two swaps



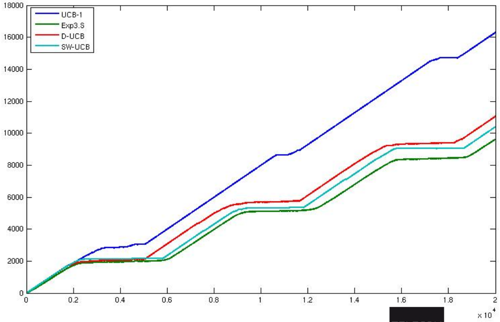
Evolution of the expected rewards

Cumulative regret

Bernoulli MAB problem with periodic rewards

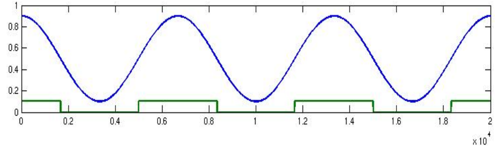


Evolution of the expected rewards

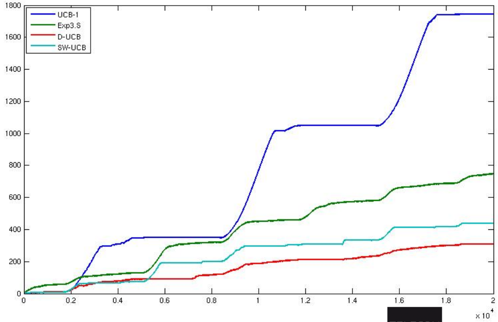


Cumulative frequency of arm 1 pulls

Bernoulli MAB problem with periodic rewards



Evolution of the expected rewards



Cumulative regret

Conclusions

- UCB methods can be efficiently adapted to face non-stationary environments;
- Interesting properties both theoretically and practically;
- No gap between stochastic and non-stochastic setups: regrets are of order $O(\sqrt{n})$;
- Other choice for the confidence interval using $N_t(\gamma^2, i)$ instead of $N_t^2(\gamma, i)$?
- Extension: data-driven choice of γ and τ;
- Generalization to smoothly-varying environments.

Conclusions

- UCB methods can be efficiently adapted to face non-stationary environments;
- Interesting properties both theoretically and practically;
- No gap between stochastic and non-stochastic setups: regrets are of order $O(\sqrt{n})$;
- Other choice for the confidence interval using $N_t(\gamma^2, i)$ instead of $N_t^2(\gamma, i)$?
- Extension: data-driven choice of γ and τ;
- Generalization to smoothly-varying environments.

Thank you for your attention!