Multi-Armed Bandits: A Novel Generic Optimal Algorithm and Applications to Networking

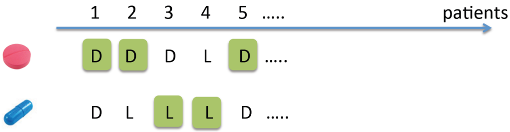
Richard Combes[1]

[1]Centrale-Supélec / L2S, France

SDN Days, 2017

Outline

A first example: sequential treatment allocation

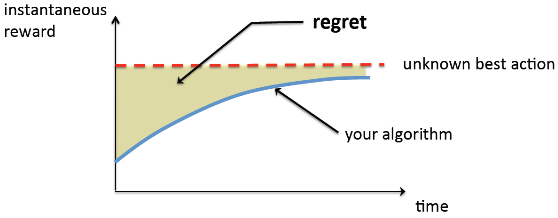


- There are T patients with the same symptoms awaiting treatment
- Two treatments exist, one is better than the other
- Based on past successes and failures which treatment should you use ?

The model

- At time n, choose action $x_n \in \mathcal{X}$, observe feedback $y_n(x_n) \in \mathcal{Y}$, and obtain reward $r_n(x_n) \in \mathbb{R}^+$.
- "Bandit feedback": rewards and feedback depend on actions (often $y_n \equiv r_n$)
- Admissible algorithm: $x_{n+1} = f_{n+1}(x_0, r_0(x_0), y_0(x_0), ..., x_n, r_n(x_n), r_n(y_n))$
- Performance metric: regret

$$R(T) = \underbrace{\max_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{n=1}^{T} r_n(x)\right]}_{\text{oracle}} - \underbrace{\mathbb{E}\left[\sum_{n=1}^{T} r_n(x_n)\right]}_{\text{your algorithm}}.$$



Bandit taxonomy: adversarial vs stochastic

Stochastic Bandit:

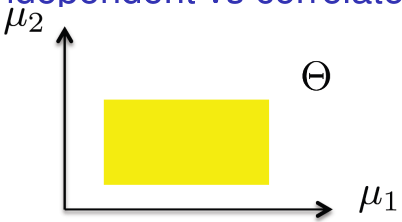
- Game against a stochastic environment
- Unknown parameters $\theta \in \Theta$
- $(r_n(x))_n$ is i.i.d with expectation θ_x

Adversarial Bandit:

- Game against a *non-adaptive* adversary
- For all x, $(r_n(x))_n$ arbitrary sequence in $\mathcal{X}$
- At time 0, the adversary "writes down $(r_n(x))_{n,x}$ in an envelope"

Engineering problems are mainly stochastic

Independent vs correlated arms



- Independent arms: $\Theta = [0,1]^K$
- Correlated arms: $\Theta \neq [0,1]^K$: choosing 1 gives information on 1 and 2

Correlation enables (sometimes much) faster learning.

Bandit taxonomy: cardinality of the set of arms

Discrete Bandits:

- $\mathcal{X} = \{1, ..., K\}$
- All arms can be sampled infinitely many times
- Regret $O(\log(T))$ (stochastic), $O(\sqrt{T})$ (adversarial)

Infinite Bandits:

- $\mathcal{X} = \mathbb{N}$, Bayesian setting (otherwise trivial)
- Explore $o(T)$ arms until a good one is found
- Regret: $O(\sqrt{T})$.

Continuous Bandits:

- $\mathcal{X} \subset \mathbb{R}^d$ convex, $x \mapsto \mu_\theta(x)$ has a *structure*
- Structures: convex, Lipschitz, linear, unimodal (quasi-convex) etc.
- Similar to derivative-free stochastic optimization
- Regret: $O(\mathbf{poly}(d)\sqrt{T})$.

Bandit taxonomy: regret minimization vs best arm identification

Sample arms and output the best arm with a given probability, similar to PAC learning

Fixed budget setting:

- T fixed, sample arms $x_1, ..., x_T$, and output $\hat{x}^T$
- Easier problem: estimation + budget allocation
- Goal: minimize $\mathbb{P}[\hat{x}^T \neq x^*]$

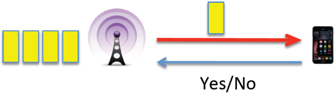
Fixed confidence setting:

- δ fixed, sample arms $x_1, ..., x_\tau$ and output $\hat{x}^\tau$
- Harder problem: estimation + budget allocation + optimal stopping (τ is a stopping time)
- Goal: minimize $\mathbb{E}[\tau]$ s.t. $\mathbb{P}[\hat{x}^\tau \neq x^*] \leq \delta$

Outline

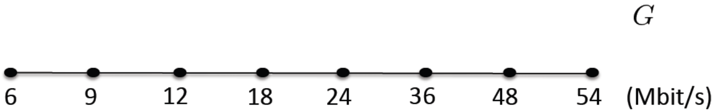
Example 1: Rate adaptation in wireless networks

- Adapting the modulation/coding scheme to the radio environment [1]



- Rates: $r_1, r_2, \ldots, r_K$
- Success probabilities: $\theta_1, \theta_2, \ldots, \theta_K$
- Throughputs: $\mu_1, \mu_2, \ldots, \mu_K$

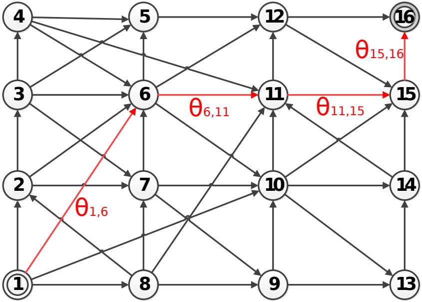
Structure: unimodality + $\theta_1 > \theta_2 > \cdots > \theta_K$.



[1] R. Combes, A. Proutiere, D. Yun, J. Ok, and Y. Yi. ”Optimal rate sampling in 802.11 systems”, IEEE INFOCOM 2014

Example 2: Shortest path routing

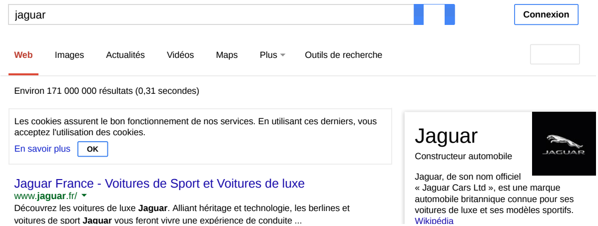
- Choose a path minimizing expected delay [2]
- Stochastic delays: $X_i(n) \sim \text{Geometric}(\theta_i)$
- Path $x \in \{0,1\}^d$, expected delay $\sum_{i=1}^d x_i/\theta_i$.
- Hop-by-hop feedback: $X_i(n)$, for $\{i : x_i(n) = 1\}$



[2] S. Talebi, Z. Zou, R. Combes, A. Proutiere, M. Johansson, "Stochastic Online Shortest Path Routing: The Value of Feedback", IEEE Trans. Automatic Control, 2017

Example 3: Learning to Rank (search engines)

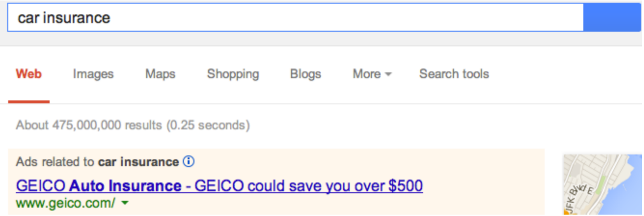
- Given a query, N relevant items, L display slots [3]
- A user is shown L items, scrolls down and selects the first relevant item
- One must show the most relevant items in the first slots.
- θ_n probability of clicking on item n (independence between items is assumed)
- Reward $r(\ell)$ if user clicks on the ℓ-th item, and 0 if the user does not click



[3] R. Combes, S. Magureanu, A. Proutiere and C. Laroche,

Example 4: Ad-display optimization

- Users are shown ads relevant to their queries[4]
- Announcers $x \in \{1, ..., K\}$, with μ_x click-through-rate and budget per unit of time c_x
- Bandit with budgets: each arm has a budget of plays
- Displayed announcer is charged per impression/click



[4] R. Combes, C. Jiang and R. Srikant, ”Bandits with Budgets: Regret Lower Bounds and Optimal Algorithms”, SIGMETRICS 2015

Outline

Optimism in the face of uncertainty

- Replace arm values by upper confidence bounds
- "Index" $b_x(n)$ such that $b_x(n) \geq \theta_x$ with high probability
- Select the arm with highest index $x_n \in \arg\max_{x \in \mathcal{X}} b_x(n)$
- Analysis idea:

$$\mathbb{E}[t_x(T)] \leq \underbrace{\sum_{n=1}^{T} \mathbb{P}[b_{x^\star}(n) \leq \theta^\star]}_{o(\log(T))} + \underbrace{\sum_{n=1}^{T} \mathbb{P}[x_n = x, b_x(n) \geq \mu^\star]}_{\text{dominant term}}.$$

$$\boxed{\text{Almost all algorithms in the literature are optimistic (sic!)}}$$

Information theory and statistics

- Distribtions P, Q with densities p and q w.r.t a measure m
- Kullback-Leibler divergence:

$$D(P||Q) = \int_x p(x) \log\left(\frac{p(x)}{q(x)}\right) m(dx),$$

- Pinsker's inequality:

$$\sqrt{\frac{D(P||Q)}{2}} \geq TV(P,Q) = \frac{1}{2}\int_x |p(x) - q(x)| m(dx).$$

- If $P, Q \sim \text{Ber}(p), \text{Ber}(q)$:

$$D(P||Q) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

- Also (Pinkser + inequality $\log(x) \leq x - 1$):

$$2(p-q)^2 \leq D(P||Q) \leq \frac{(p-q)^2}{q(1-q)}$$

The KL-divergence is ubiquitous in bandit problems

Regret Lower Bounds: general technique

- Decision x, two parameters θ, λ, with $x^\star(\lambda) = x \neq x^\star(\theta)$.
- Consider consider an algorithm with $R^\pi(T) = \log(T)$ for all parameters (unformly good):

$$\mathbb{E}_\theta[t_x(T)] = O(\log(T)) \ , \ \mathbb{E}_\lambda[t_x(T)] = T - O(\log(T)).$$

- Markov inequality:

$$\mathbb{P}_\theta[t_x(T) \geq T/2] + \mathbb{P}_\lambda[t_x(T) < T/2] \leq O(T^{-1}\log(T)).$$

- $\mathbf{1}\{t_x(T) \leq T/2\}$ is a hypothesis test, risk $O(T^{-1}\log(T))$
- Hence (Neyman-Pearson / Tsybakov):

$$\underbrace{\sum_x \mathbb{E}_\theta[t_x(T)] KL(\theta_x, \lambda_x)}_{\text{KL divergence of the observations}} \geq \log(T) - O(\log(\log(T))).$$

Concentration inequalities: Chernoff bounds

- Building indexes requires tight concentration inequalities
- Chernoff bounds: upper bound the MGF
- $X = (X_1, ..., X_n)$ independent, with mean μ, $S_n = \sum_{n'=1}^{n} X_{n'}$
- G such that $\log(\mathbb{E}[e^{\lambda(X_n-\mu)}]) \leq G(\lambda)$, $\lambda \geq 0$
- Generic technique:

$$\begin{aligned}
\mathbb{P}[S_n - n\mu \geq \delta] &= \mathbb{P}[e^{\lambda(S_n-n\mu)} \geq e^{\lambda\delta}] \\
&\leq e^{-\lambda\delta}\mathbb{E}[e^{\lambda(S_n-n\mu)}] \text{ (Markov)} \\
&= \exp(nG(\lambda) - \lambda\delta) \text{ (independence)} \\
&\leq \exp\left(-n \max_{\lambda \geq 0}\{\lambda\delta n^{-1} - G(\lambda)\}\right).
\end{aligned}$$

Concentration inequalities: Chernoff and Hoeffding's inequality

- Bounded variables: if $X_n \in [a, b]$ a.s then $\mathbb{E}[e^{\lambda(X_n-\mu)}] \le e^{\lambda^2(b-a)^2/8}$ (Hoeffding lemma)
- Hoeffding's inequality:

$$\mathbb{P}[S_n - n\mu \ge \delta] \le \exp\left(-\frac{2\delta^2}{n(b-a)^2}\right)$$

- Subgaussian variables: $\mathbb{E}[e^{\lambda(X_n-\mu)}] \le e^{\sigma^2\lambda^2/2}$, similar
- Bernoulli variables: $\mathbb{E}[e^{\lambda(X_n-\mu)}] = \mu e^{\lambda(1-\mu)} - (1-\mu)e^{-\lambda\mu}$
- Chernoff's inequality:

$$\mathbb{P}[S_n - n\mu \ge \delta] \le \exp(-nKL(\mu + \delta/n, \mu))$$

- Pinsker's inequality: Chernoff is stronger than Hoeffding.

Concentration inequalities: variable sample size and peeling

- In bandit problems, the sample size is random and depends on the samples themselves
- Intervals $\mathcal{N}_k = \{n_k, ..., n_{k+1}\}$, $\mathcal{N} = \cup_{k=1}^{K} \mathcal{N}$
- Idea: $Z_n = e^{\lambda(S_n - n\mu)}$ is a positive sub-martingale:

$$
\begin{aligned}
\mathbb{P}[\max_{n \in \mathcal{N}_k} (S_n - \mu n) \geq \delta] &= \mathbb{P}[\max_{n \in \mathcal{N}_k} Z_n \geq e^{\lambda\delta})] \\
&\leq e^{-\lambda\delta} \mathbb{E}[Z_{n_{k+1}}] \text{ (Doob's inequality)} \\
&= \exp(-\lambda\delta + n_{k+1} G(\lambda)) \\
&\leq \exp\left(-n_{k+1} \max_{\lambda \geq 0} \{\lambda\delta n_{k+1}^{-1} - G(\lambda)\} \right).
\end{aligned}
$$

- Peeling trick (Neveu): union bound over k, $n_k = (1+\alpha)^k$.

Outline

The Lai-Robbins bound

- Actions $\mathcal{X} = \{1, ..., K\}$
- Rewards $\theta = (\theta_1, ..., \theta_K) \in [0,1]^K$
- Uniformly good algorithm: $R(T) = O(\log(T))$, $\forall \theta$

Theorem (Lai '85)

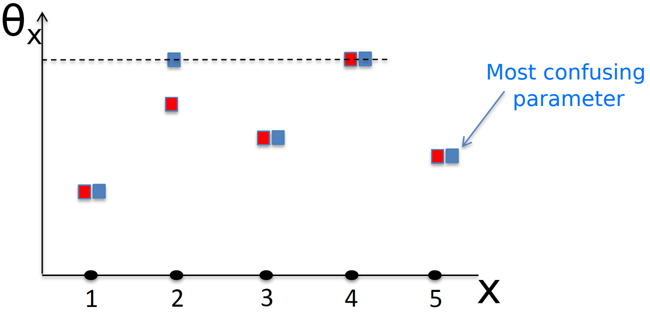
For any uniformly good algorithm, and x s.t $\theta_x < \theta^\star$ we have:

$$\lim \inf_{T \to \infty} \frac{\mathbb{E}[t_x(T)]}{\log(T)} \geq \frac{1}{KL(\theta_x, \theta^\star)}$$

- For $x \neq x^\star$, apply the generic technique with:

$$\lambda = (\theta_1, ..., \theta_{x-1}, \theta^\star + \epsilon, \theta_{x+1}, ..., \theta_K)$$

The Lai-Robbins bound



Classical bandits: algorithms

- Select the arm with highest index $x_n \in \arg\max_{x \in \mathcal{X}} b_x(n)$
- UCB algorithm (Hoeffding's ineqality):

$$b_x(n) = \underbrace{\hat{\theta}_x(n)}_{\text{empirical mean}} + \underbrace{\sqrt{\frac{2\log(n)}{t_x(n)}}}_{\text{exploration bonus}}.$$

- KL-UCB algorithm (using Garivier's inequality):

$$b_x(n) = \max\{q \le 1 : \underbrace{t_x(n)KL(\hat{\theta}_x(n), q)}_{\text{likelihood ratio}} \le \underbrace{f(n)}_{\log(\text{confidence level}^{-1})}\}.$$

with $f(n) = \log(n) + 3\log(\log(n))$.

Classical bandits: regret analysis

Theorem (Auer'02)

Under algorithm UCB, for all x s.t $\theta_x < \theta^\star$:

$$\mathbb{E}[t_x(T)] \le \frac{8\log(T)}{(\theta_x - \theta^\star)^2} + \frac{\pi^2}{6}.$$

Theorem (Garivier'11)

Under algorithm KL-UCB, for all x s.t $\theta_x < \theta^\star$ and for all $\delta < \theta^\star - \theta_x$:

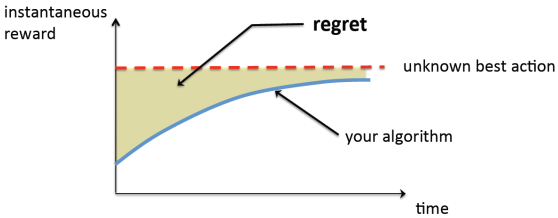
$$\mathbb{E}[t_x(T)] \le \frac{\log(T)}{KL(\theta_x + \delta, \theta^\star)} + C\log(\log(T)) + \delta^{-2}.$$

Outline

The model

- A finite set of arms $\mathcal{X}$, a parameter set Θ
- An unknown parameter $\theta \in \Theta$
- At time step n, select arm x, observe feedback $Y(n, x) \sim \nu(\theta(x))$ and receive reward $\mu(x, \theta)$
- Observations $(Y(n, x))_n$ are i.i.d. $\forall x$.
- Performance metric: regret

$$R^{\pi}(T, \theta) = T \max_{x \in \mathcal{X}} \mu(x, \theta) - \sum_{t=1}^{T} \mathbb{E}(\mu(x(t), \theta)).$$



Instances of our model

Classical bandit (Lai, 1985):

- Set of arms $\mathcal{X} = \{1, ..., |\mathcal{X}|\}$
- Parameter set $\Theta = [0, 1]^{|\mathcal{X}|}$
- Reward function: $\mu(x, \theta) = \theta(x)$
- Observations: $Y(n, x) \sim \text{Ber}(\theta(x))$
- Model for sequential treatment allocation.

Instances of our model

Linear bandit (Dani, 2008):

- Set of arms $\mathcal{X} \subset \mathbb{R}^d$ finite
- Parameter set $\theta \in \Theta$ iff $\theta(x) = \langle \phi, x \rangle, \forall x$
- Reward function: $\mu(x, \theta) = g(\theta(x))$, g link function
- Observations: $Y(n, x) \sim \mathcal{N}(\theta(x), 1)$
- Stochastic version of linear / combinatorial optimization.
- Applications: routing, channel assignment, recommender systems etc.

Instances of our model

Dueling bandits (Komiyama, 2015):

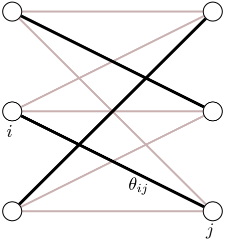
- Set of arms $\mathcal{X} = \{(i,j) \in \{1,\dots,d\}^2\}$
- Parameter set $\Theta \subset [0,1]^{d\times d}$ preference matrices
- i preferred to j w.p. $\theta(i,j) > \frac{1}{2}$, $i^\star$ Condorcet winner.
- Reward function: $\mu((i,j),\theta) = \frac{1}{2}(\theta(i^\star,i) + \theta(i^\star,j) - 1)$,
- Observations: $Y(n,x) \sim \text{Ber}(\theta(x))$
- Model for ranking using pairwise comparisons
- Applications: tournaments, learning to rank

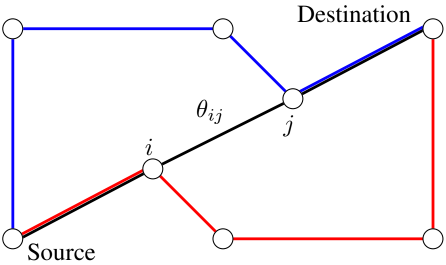
$$\begin{pmatrix} 0.5 & 0.7 & 0.9 & 0.8 \\ 0.3 & 0.5 & 0.3 & 0.1 \\ 0.1 & 0.7 & 0.5 & 0.9 \\ 0.2 & 0.9 & 0.1 & 0.5 \end{pmatrix}$$

Instances of our model

Combinatorial semi-bandits (Cesa-Bianchi, 2012):

- Set of arms $\mathcal{X} \subset \{0,1\}^d$
- Parameter set $\theta \in \Theta$ iff $\theta(x) = (\phi(1)x(1), \dots, \phi(d)x(d)), \forall x$
- Reward function: $\mu(x,\theta) = \sum_{i=1}^{d} \phi(i)x(i)$
- Observations: $Y(n,x) \in \{0,1\}^d$ with independent components and mean $\theta(x)$
- Combinatorial optimization with detailed feedback.
- Applications: routing w. link feedback, channel assignment, etc.





Regret lower bound

Theorem

Consider π a uniformly good algorithm. For any $\theta \in \Theta$, we have:

$$\lim \inf_{T \to \infty} \frac{R^{\pi}(T, \theta)}{\ln T} \geq C(\theta),$$

where $C(\theta)$ is the value of the optimization problem:

$$\underset{\eta(x) \geq 0\,,\, x \in \mathcal{X}}{\text{minimize}} \sum_{x \in \mathcal{X}} \eta(x)(\mu^{\star}(\theta) - \mu(x, \theta)) \tag{1}$$

$$\text{subject to} \sum_{x \in \mathcal{X}} \eta(x) D(\theta, \lambda, x) \geq 1\,,\ \forall \lambda \in \Lambda(\theta), \tag{2}$$

where

$$\Lambda(\theta) = \{\lambda \in \Theta : D(\theta, \lambda, x^{\star}(\theta)) = 0, x^{\star}(\theta) \neq x^{\star}(\lambda)\}.$$

The OSSB algorithm

$s(0) \leftarrow 0, N(x,1), m(x,1) \leftarrow 0$, $\forall x \in \mathcal{X}$ {Initialization}
for $t = 1, ..., T$ **do**
Compute the optimization problem (1)-(2) solution $(c(x, m(t)))_{x\in\mathcal{X}}$
where $m(t) = (m(x,t))_{x\in\mathcal{X}}$
if $N(x,t) \geq c(x, m(t))(1+\gamma) \ln\ t, \forall x$ **then**
$s(t) \leftarrow s(t-1)$
$x(t) \leftarrow x^\star(m(t))$ {Exploitation}
else
$s(t) \leftarrow s(t-1) + 1$
$\overline{X}(t) \leftarrow \arg\min_{x\in\mathcal{X}} \frac{N(x,t)}{c(x,m(t))}$
$\underline{X}(t) \leftarrow \arg\min_{x\in\mathcal{X}} N(x,t)$
if $N(\underline{X}(t), t) \leq \varepsilon s(t)$ **then**
$x(t) \leftarrow \underline{X}(t)$ {Estimation}
else
$x(t) \leftarrow \overline{X}(t)$ {Exploration}
end if
end if
{Update statistics}
Select arm $x(t)$ and observe $Y(x(t), t)$
$m(x, t+1) \leftarrow m(x,t), \forall x \neq x(t)$,
$N(x, t+1) \leftarrow N(x,t), \forall x \neq x(t)$
$m(x(t), t+1) \leftarrow \frac{Y(x(t),t) + m(x(t),t)N(x(t),t)}{N(x(t),t)+1}$
$N(x(t), t+1) \leftarrow N(x(t), t) + 1$
end for

OSSB is asymptotically optimal

We use the following natural assumptions.

- A1 Observations are either Bernoulli or Gaussian
- A2 For all x, $(\theta, \lambda) \mapsto D(x, \theta, \lambda)$ is continuous at all points where it is not infinite
- A3 For all x, the mapping $\theta \to \mu(x, \theta)$ is continuous
- A4 The solution to problem (1)-(2) is unique

Theorem

Under A1-A4, the regret of $\pi =$OSSB(ε, γ) with $\varepsilon < \frac{1}{|\mathcal{X}|}$ verifies:

$$\limsup_{T \to \infty} \frac{R^\pi(T)}{\ln T} \le C(\theta) F(\varepsilon, \gamma, \theta),$$

with $F(\varepsilon, \gamma, \theta) \to 1$ as $\varepsilon \to 0$ and $\gamma \to 0$ for all θ.

OSSB: elements of analysis

Element 1: show that in the exploitation phase, the optimal arm is selected with high probability.

Lemma

Under A1, there exists a function G such that for all $t \geq 1$:

$$\sum_{t \geq 1} \mathbb{P}\left(\sum_{x \in \mathcal{X}} N(x,t) D(m(t), \theta, x) \geq (1+\gamma) \ln t \right) \leq G(\gamma, |\mathcal{X}|).$$

Proof: Chernoff bound + Doob's maximal inequality + multi-dimensional peeling.

OSSB: elements of analysis

Element 2: show that, when θ is well estimated, so is $c(\theta)$.

Lemma

Under A1-A4, the optimal value $\theta \mapsto C(\theta)$ and the solution $\theta \mapsto c(\theta) = (c(x,\theta))_{x\in\mathcal{X}}$ are continuous at θ.

Proof: similar to Berge's theorem with an additional difficulty as the feasible set is not compact.

OSSB: elements of analysis

Element 3: show that the number of exploration / estimation rounds where θ is not well estimated is finite in expectation. Idea: after s such rounds, $N(x,t) \geq \epsilon s$ by construction.

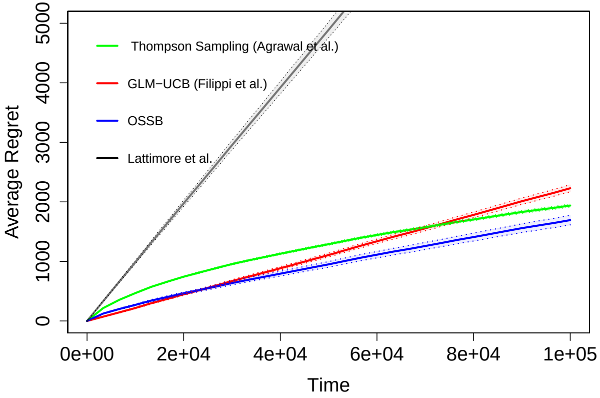
Lemma

Let $x \in \mathcal{X}$ and $\epsilon > 0$. Define $\mathcal{F}_t$ the σ-algebra generated by $(Y(x(s),s))_{1 \leq s \leq t}$. Let $\mathcal{S} \subset \mathbb{N}$ be a (random) set of rounds. Assume that there exists a sequence of (random) sets $(\mathcal{S}(s))_{s \geq 1}$ such that (i) $\mathcal{S} \subset \cup_{s \geq 1} \mathcal{S}(s)$, (ii) for all $s \geq 1$ and all $t \in \mathcal{S}(s)$, $N(x,t) \geq \epsilon s$, (iii) $|\mathcal{S}(s)| \leq 1$, and (iv) the event $t \in \mathcal{S}(s)$ is $\mathcal{F}_t$-measurable. Then for all $\delta > 0$:

$$\sum_{t \geq 1} \mathbb{P}(t \in \mathcal{S}, |m(x,t) - \theta(x)| > \delta) \leq \frac{1}{\epsilon \delta^2}.$$

Proof: Chernoff bound + Doob's optional stopping theorem.

Numerical example: linear bandits



Thank you for your attention !