Online Optimization and Bandits problems

Nicolas Gast

5 janvier 2021

What is online Optimization ?

Decisions are make sequentially but you can learn from your past decisions.

$$\text{Action} \rightsquigarrow \text{Observation} \rightsquigarrow \text{Action} \rightsquigarrow \text{Observation} \rightsquigarrow \dots$$

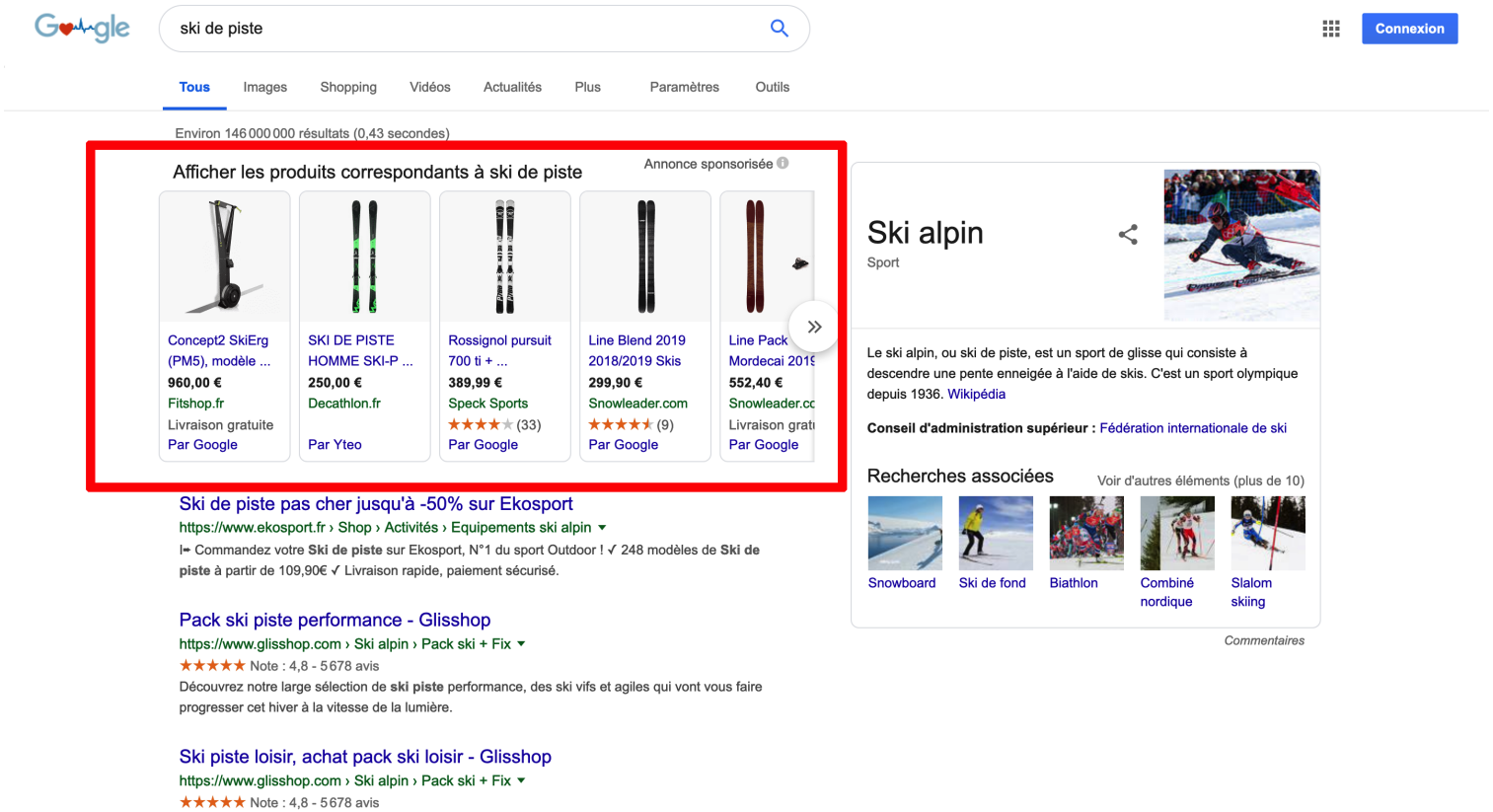
Examples from `https://en.wikipedia.org/wiki/Online_optimization` :

- K-server problem
- Job shop scheduling problem
- List update problem
- Bandit problem
- Secretary problem
- Search games
- Ski rental problem
- Linear search problem

Online Optimization is everywhere on the Internet

maximize the click-through rate

Google search

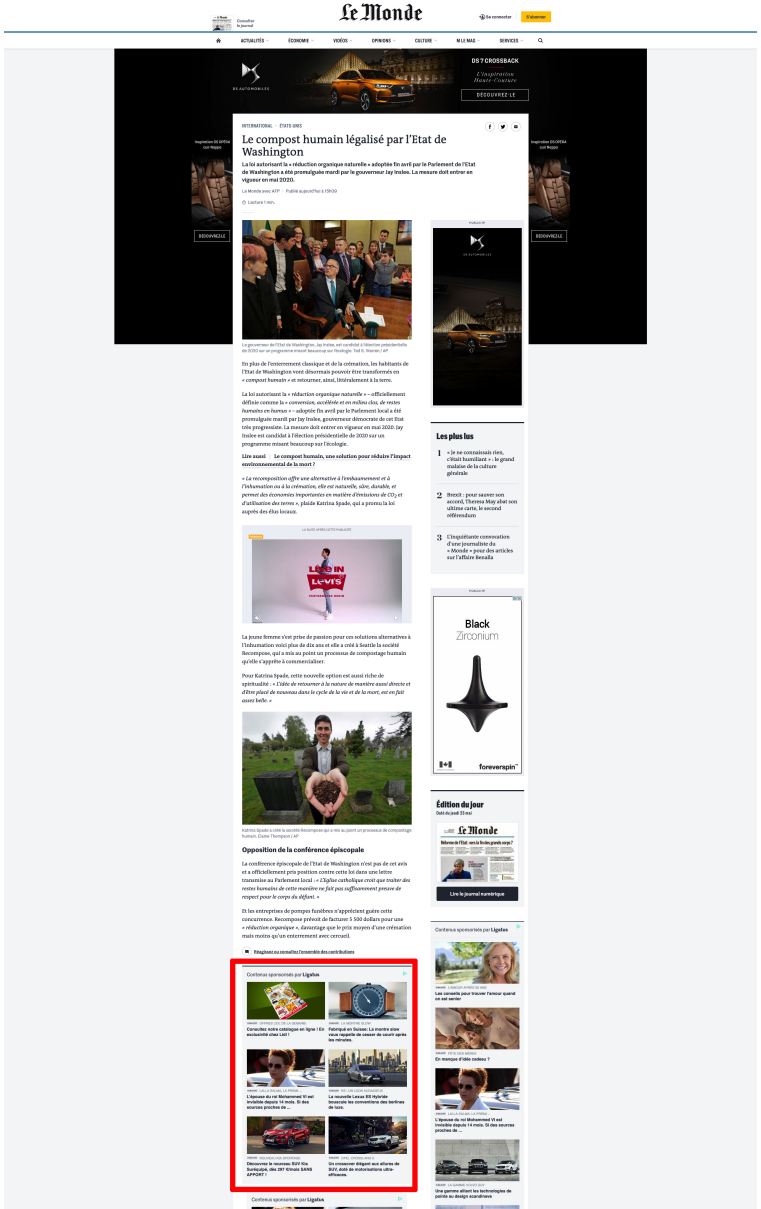


- How does google decides what items they should show you ?

Online Optimization is everywhere on the Internet

maximize the click-through rate

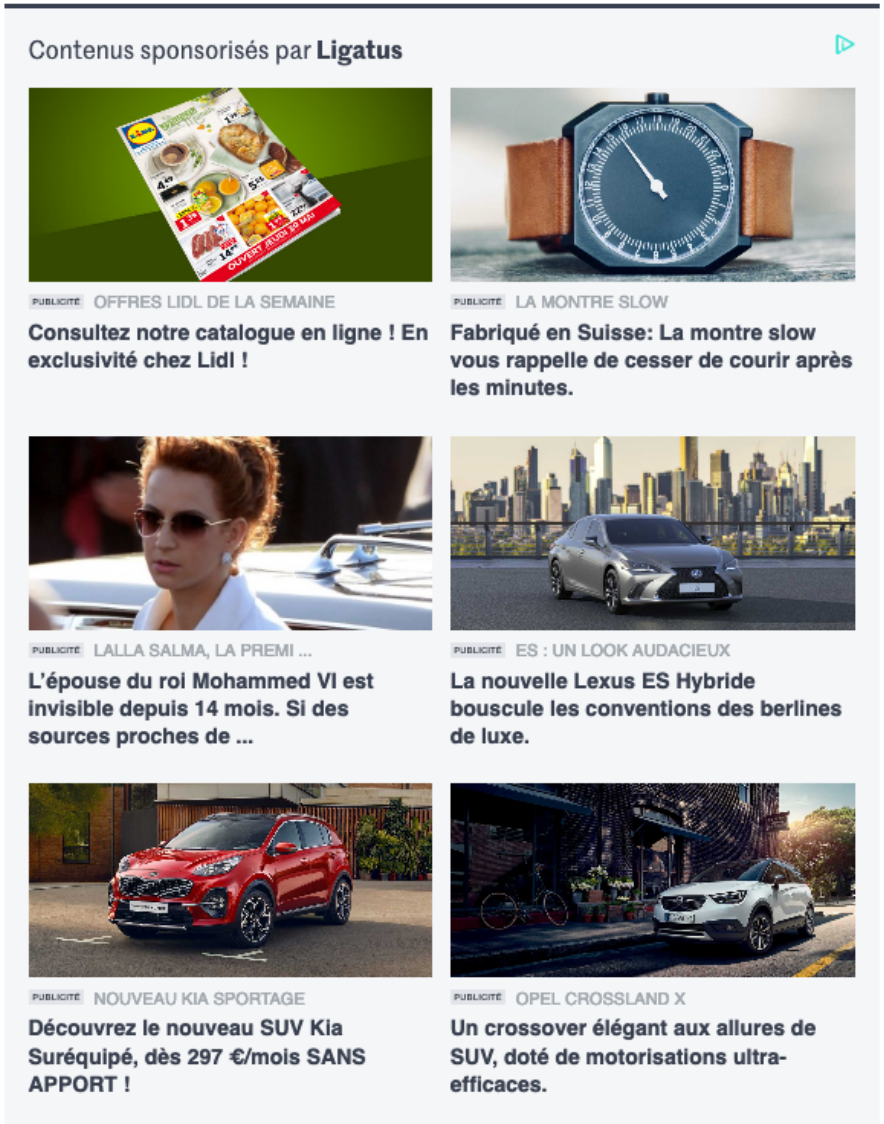
From `http://www.lemonde.fr` :



Online Optimization is everywhere on the Internet

maximize the click-through rate

From `http://www.lemonde.fr` :



- How does Ligatus/Outbrain chooses a good title ?

These problems are instances of multi-armed bandit problem

A decision-maker chooses sequential decisions.

- At time t, she chooses $I_t \in \{1 \dots n\}$.
- This gives a (random) reward R_{t,I_t}.

The decision-maker wants to maximize the sum of all rewards : $\sum_{t=1}^{T} R_{t,I_t}$.

MAB are everywhere

- Historically : sequential clinical trials (1933)
- Now on the web :
  - Ad placement
  - Price experimentation
  - Search engines
- Other examples :
  - Choosing the right expert (forecasting, scheduling)
  - Research project allocation

The mathematical formulation depends on many parameters

In this talk, I will mostly focus on stochastic, i.i.d. bandits.

There are many variants

- Knowledge :
  - Adversarial/robust (R_i are chosen by an adversary)
  - Stochastic or Markovian (R_i are random)
- Structure (e.g., combinatorial bandits)
- Observation (e.g., targeted advertising)
- Feedback

Table of Contents

Outline

The Bernoulli multi-armed bandit

At each time t, the arm i gives you a reward 1 with probability μ_i and $1-\mu_i$ with probability $1-\mu_i$:

$$\mathbb{P}(R_{t,i}=1)=\mu_i=1-\mathbb{P}(R_{t,i}=0).$$

We assume that :

- The rewards are all independent.
- The decision maker does not know μ_i.

There is a natural compromise between **exploration** and **exploitation**.

Motivation

- **Clinical trial**

μ_1 μ_2 μ_3 μ_4




  - Choose treatment I_t for patient t
  - Observe a response $X_t \in \{0, 1\}$ with $P(X_t = 1) = \mu_{I_t}$.
  - Goal : maximize the number of patients healed.

Motivation

- **Clinical trial**

μ_1 μ_2 μ_3 μ_4




  - Choose treatment I_t for patient t
  - Observe a response $X_t \in \{0, 1\}$ with $P(X_t = 1) = \mu_{I_t}$.
  - Goal : maximize the number of patients healed.

- **Online advertisement**, e.g., the choice of a title of a news article :

Title	Click proba.
"Murder victim found in adult entertainment venue"	μ_1
"Headless Body found in Topless Bar"	μ_2

  - Choose which title I_t to display to customer t
  - Observe a response $X_t \in \{0, 1\}$ (click or no click).
  - Goal : maximize your number of clicks.

Regret minimization

We define the regret of a sequence of action $\mathcal{I} = (I_1, I_2 \dots, I_T)$ as

$$\text{Regret}(\mathcal{I}) = \max_i \mathbb{E}\left[\sum_{t=1}^{T} R_{t,i}\right] - \mathbb{E}\left[\sum_{t=1}^{T} R_{t,I_t}\right]$$

Regret minimization

We define the regret of a sequence of action $\mathcal{I} = (I_1, I_2 \dots, I_T)$ as

$$\begin{aligned}\text{Regret}(\mathcal{I}) &= \max_i \mathbb{E}\left[\sum_{t=1}^{T} R_{t,i}\right] - \mathbb{E}\left[\sum_{t=1}^{T} R_{t,I_t}\right] \\ &= \mathbb{E}\left[r_* T - \sum_{t=1}^{T} R_{t,I_t}\right],\end{aligned}$$

where $r_* = \max_i \mathbb{E}[R_{t,i}] = \max_i \mathbb{E}[R_{1,i}]$.

Maximizing reward = minimizing regret.

- Goal : design strategies that have a small regret for all distribution μ.

Asymptotically optimal regret

A good policy has sub-linear regret :

$$\text{Regret}(\mathcal{I}) = o(T).$$

To achieve this, all the arms should be drawn infinitely often.

Asymptotically optimal regret

A good policy has sub-linear regret :

$$\text{Regret}(\mathcal{I}) = o(T).$$

To achieve this, all the arms should be drawn infinitely often.

Theorem (Lai and Robbins, 1985 (Asymptotically Efficient Adaptive Allocation Rules))

There exists a constant c (that depend on μ) such that any uniformly efficient[a] strategy satisfies :

$$\textit{Regret}(\mathcal{I}) \geq c \log T$$

a. Meaning $\text{Regret}(\mathcal{I}) = o(T^\alpha)$ for all μ and α.

Some ideas of policies

- **Random** – Draw each arm with probability $1/n$.
  - Exploration

Some ideas of policies

- **Random** – Draw each arm with probability $1/n$.
  - Exploration

- **Greedy :** Always choose the empirical best arm :

$$I_{t+1} = \arg\max_i \hat{\mu}_i(t)$$

  - Exploitation

Some ideas of policies

- **Random** – Draw each arm with probability $1/n$.
  - Exploration

- **Greedy :** Always choose the empirical best arm :

$$I_{t+1} = \arg\max_i \hat{\mu}_i(t)$$

  - Exploitation

- ε-**greedy** : apply "greedy" with probability $1-\varepsilon$ and "random" otherwise (each with probability ε/n)
  - Exploration and exploitation.

Some ideas of policies

- **Random** – Draw each arm with probability $1/n$.
  - Exploration
- **Greedy :** Always choose the empirical best arm :

$$I_{t+1} = \arg\max_i \hat{\mu}_i(t)$$

  - Exploitation
- ε-**greedy** : apply "greedy" with probability $1-\varepsilon$ and "random" otherwise (each with probability ε/n)
  - Exploration and exploitation.

All have linear regrets

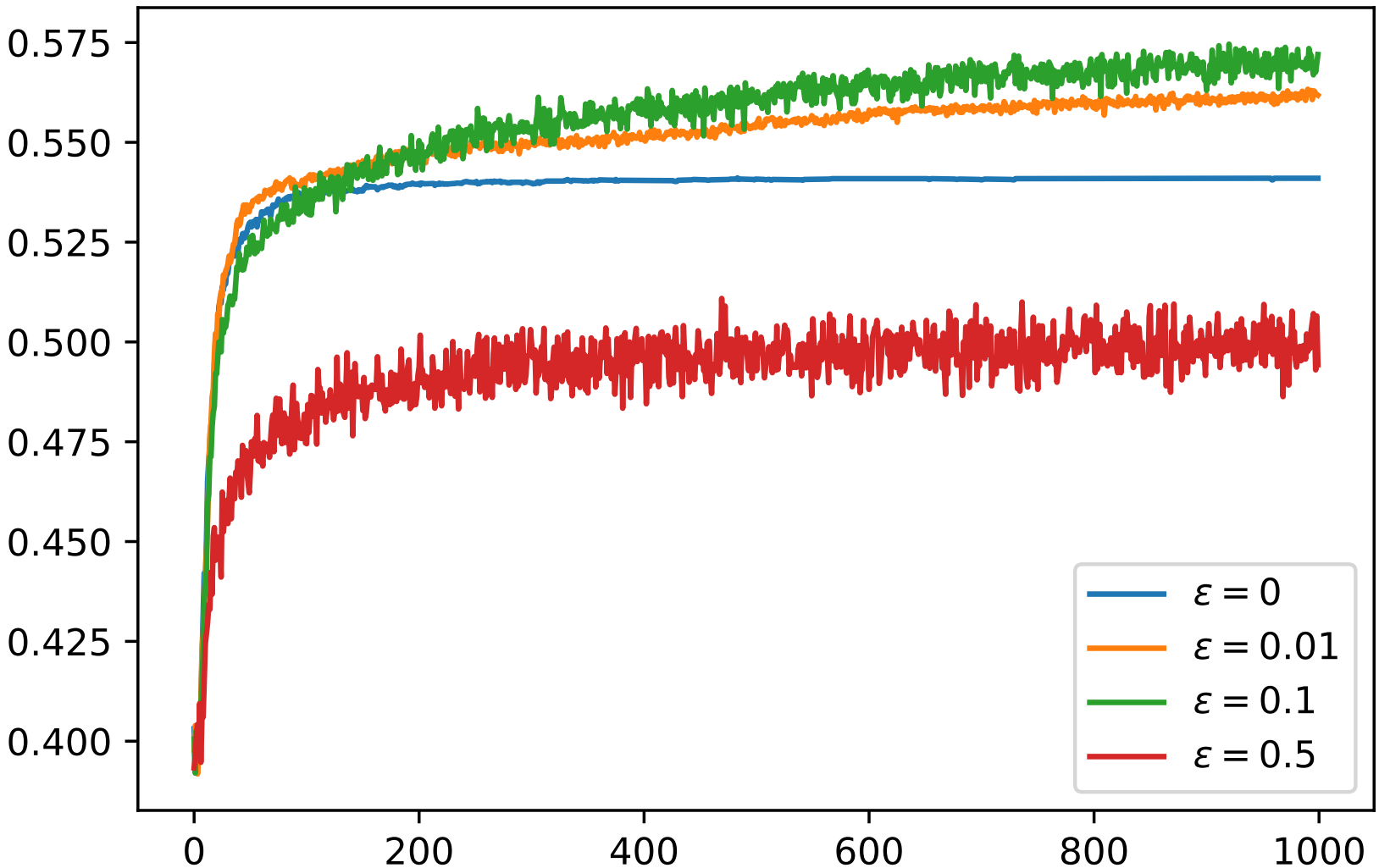
Analysis of the ϵ-greedy policy ($\epsilon > 0$)

- As $\epsilon > 0$, all arm will be chosen an infinite amount of time.
- Hence, by the law of large number : $\hat{\mu}_i(t)$ converges to the true mean as $t \to \infty$.
- Therefore, the asymptotic regret is

$$\text{Regret}(\varepsilon\text{-greedy}) = T(\sum_i (\mu_* - \mu_i))\frac{\varepsilon}{n} + o(T)$$

ε-greedy : Smaller or larger ϵ are not necessarily better

Consider 5 Bernoulli arms with success probabilities
$\mu = [0.5, 0.3, 0.6, 0.4, 0.2]$. The average reward as a function of time is :



Outline

UCB builds on Confidence Intervals

Consider a coin that gives "Head" with probability μ. Suppose that you draw a coin N times and observe K times "head". The natural estimator of μ is :

$$\hat{\mu} = \frac{K}{N}$$

UCB builds on Confidence Intervals

Consider a coin that gives "Head" with probability μ. Suppose that you draw a coin N times and observe K times "head". The natural estimator of μ is :

$$\hat{\mu} = \frac{K}{N}$$

Hoeffding inequality gives us

$$\mathbb{P}\left(\mu \geq \hat{\mu} + \sqrt{\frac{\alpha}{2N}}\right) \leq e^{-\alpha}$$

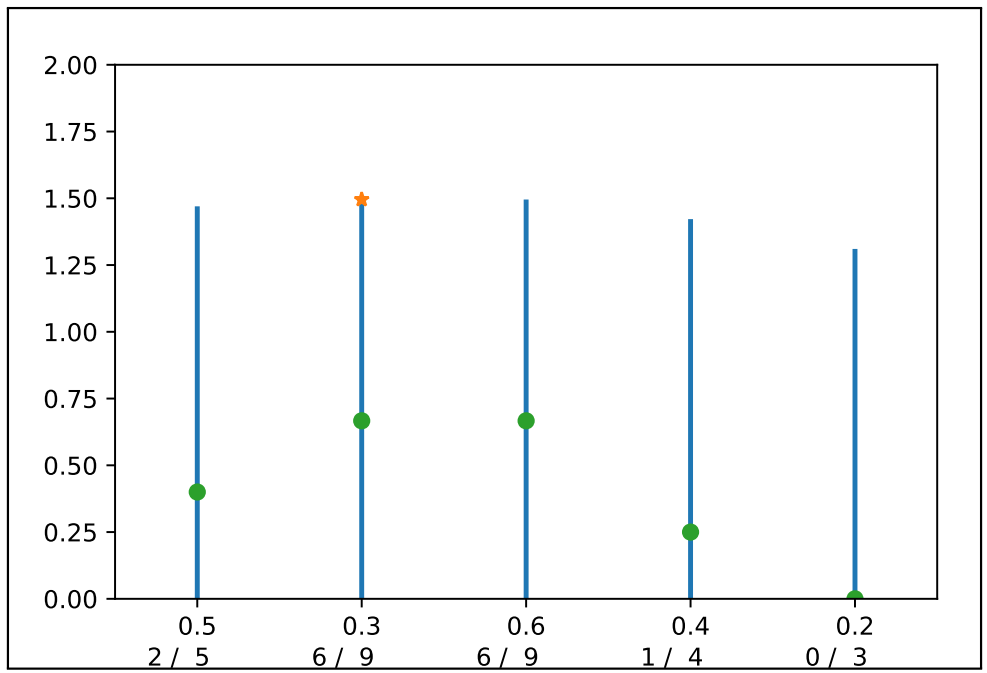
The idea of USB is to use the above bound with a growing α.

The UCB algorithm

UCB computes a confidence bound $UCB_i(t)$ such that $\mu_i(t) \leq UCB_i(t)$ with high probability. Example : $UCB1$ [Auer et al. 02] uses

$$UCB_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{\alpha \log t}{2N_i(t)}}.$$

- Choose $I_{t+1} \in UCB_i(t)$ (optimism principle).



Regret of UCB

Theorem

The algorithm UCB1 has a logarithmic regret. For all $\alpha > 2$, there exists a constant $C_\alpha > 0$ such that if a is a sub-optimal arm, then $N_i(T)$ (the number of time that this arm is chosen before T) satisfies

$$\mathbb{E}[N_i(T)] \leq \frac{2\alpha \log T}{(\mu_* - \mu_i)^2} + C_\alpha.$$

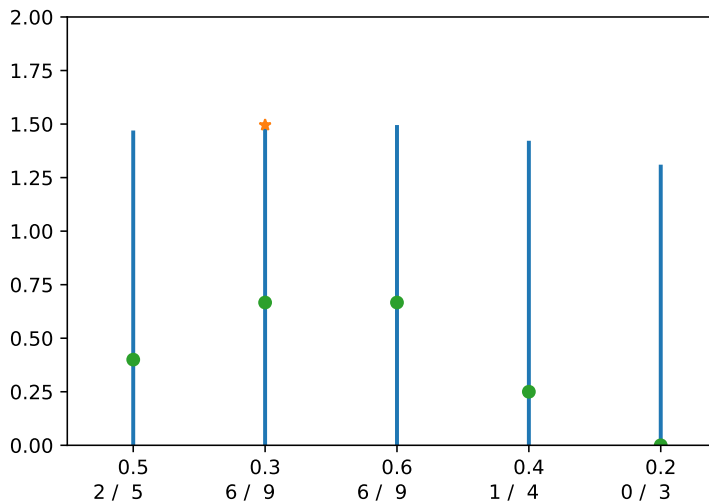
Analysis of UCB (1/3)

Using Hoeffding's inequality, one can show

$$\mathbb{P}(UCB_i(t) \leq \mu_i) \leq \frac{1}{t^{\alpha-1}}$$

$$\mathbb{P}(LCB_i(t) \geq \mu_i) \leq \frac{1}{t^{\alpha-1}},$$

where $LCB_i(t) \leq \hat{\mu}_i(t) - \sqrt{\frac{\alpha \log t}{2N_i(t)}}$.



Analysis of UCB (2/3)

Let us now consider arm 1 is optimal and arm 2 is not optimal ($\mu_2 < \mu_1$).
Let $N_2(t)$ be the number of times that arm 2 is chosen between 0 and $t-1$. Then for any stopping time τ, we have :

$$N_2(T) - N_2(\tau) = \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2\}}$$

$$= \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) \leq \mu_1\}} + \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) > \mu_1\}}$$

Analysis of UCB (2/3)

Let us now consider arm 1 is optimal and arm 2 is not optimal ($\mu_2 < \mu_1$).
Let $N_2(t)$ be the number of times that arm 2 is chosen between 0 and $t-1$. Then for any stopping time τ, we have :

$$N_2(T) - N_2(\tau) = \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2\}}$$

$$= \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) \leq \mu_1\}} + \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) > \mu_1\}}$$

$$= \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) \leq \mu_1\}} + \mathbf{1}_{\{I_t=2 \wedge UCB_2(t) > UCB_1(t) > \mu_1\}}$$

Analysis of UCB (2/3)

Let us now consider arm 1 is optimal and arm 2 is not optimal ($\mu_2 < \mu_1$).
Let $N_2(t)$ be the number of times that arm 2 is chosen between 0 and $t-1$. Then for any stopping time τ, we have :

$$N_2(T) - N_2(\tau) = \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2\}}$$

$$= \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) \leq \mu_1\}} + \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) > \mu_1\}}$$

$$= \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) \leq \mu_1\}} + \mathbf{1}_{\{I_t=2 \wedge UCB_2(t) > UCB_1(t) > \mu_1\}}$$

$$\leq \sum_{t=\tau}^{T-1} \mathbf{1}_{\{UCB_1(t) \leq \mu_1\}} + \mathbf{1}_{\{UCB_2(t) > \mu_1\}}$$

Analysis of UCB (2/3)

Let us now consider arm 1 is optimal and arm 2 is not optimal ($\mu_2 < \mu_1$).
Let $N_2(t)$ be the number of times that arm 2 is chosen between 0 and $t-1$. Then for any stopping time τ, we have :

$$N_2(T) - N_2(\tau) = \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2\}}$$

$$= \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) \leq \mu_1\}} + \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) > \mu_1\}}$$

$$= \sum_{t=\tau}^{T-1} \mathbf{1}_{\{I_t=2 \wedge UCB_1(t) \leq \mu_1\}} + \mathbf{1}_{\{I_t=2 \wedge UCB_2(t) > UCB_1(t) > \mu_1\}}$$

$$\leq \sum_{t=\tau}^{T-1} \mathbf{1}_{\{UCB_1(t) \leq \mu_1\}} + \mathbf{1}_{\{UCB_2(t) > \mu_1\}}$$

$$\leq \sum_{t=\tau}^{T-1} \mathbf{1}_{\{UCB_1(t) \leq \mu_1\}} + \mathbf{1}_{\{\mu_2 < LCB_2(t)\}} + \mathbf{1}_{\{UCB_2(t) > \mu_1 > \mu_2 > LCB_2(t)\}}$$

Analysis of UCB (3/3)

$$\sum_{t=\tau}^{T-1} \underbrace{\mathbf{1}_{\{UCB_1(t)\leq\mu_1\}} + \mathbf{1}_{\{\mu_2<LCB_2(t)\}}}_{\text{occurs with small probability (Hoeffding)}} + \underbrace{\mathbf{1}_{\{UCB_2(t)>\mu_1>\mu_2>LCB_2(t)\}}}_{\text{possible only if } N_2 \text{ is small}}$$

Indeed, for the last term :

$$\mathbf{1}_{\{LCB_2(t)<\mu_2<\mu_1<UCB_2(t)\}} \leq \mathbf{1}_{\{2\sqrt{\frac{\alpha \log t}{2N_2(t)}}>(\mu_1-\mu_2)\}} = \mathbf{1}_{\{N_2(t)\leq\frac{2\alpha \log T}{(\mu_1-\mu_2)^2}\}}$$

Analysis of UCB (3/3)

$$\sum_{t=\tau}^{T-1} \underbrace{\mathbf{1}_{\{UCB_1(t)\leq\mu_1\}} + \mathbf{1}_{\{\mu_2<LCB_2(t)\}}}_{\text{occurs with small probability (Hoeffding)}} + \underbrace{\mathbf{1}_{\{UCB_2(t)>\mu_1>\mu_2>LCB_2(t)\}}}_{\text{possible only if } N_2 \text{ is small}}$$

Indeed, for the last term :

$$\mathbf{1}_{\{LCB_2(t)<\mu_2<\mu_1<UCB_2(t)\}} \leq \mathbf{1}_{\{2\sqrt{\frac{\alpha \log t}{2N_2(t)}}>(\mu_1-\mu_2)\}} = \mathbf{1}_{\{N_2(t)\leq\frac{2\alpha \log T}{(\mu_1-\mu_2)^2}\}}$$

Let $x = \frac{2\alpha \log T}{(\mu_2-\mu_1)^2}$ and $X = \sum_{t=\tau}^{T-1} \mathbf{1}_{\{UCB_1(t)\leq\mu_1\}} + \mathbf{1}_{\{\mu_2<LCB_2(t)\}}$. We have

$$N_2(T) - N_2(\tau) \leq X + \sum_{t=\tau}^{T} \mathbf{1}_{\{N_2(t)\leq x\}}.$$

where by Hoeffding's inequality, $\mathbb{E}[X] \leq 2\sum_{t=1}^{\infty} 1/t^{\alpha-1} =: C_\alpha$.

Analysis of UCB (3/3)

$$\sum_{t=\tau}^{T-1} \underbrace{\mathbf{1}_{\{UCB_1(t)\leq\mu_1\}} + \mathbf{1}_{\{\mu_2<LCB_2(t)\}}}_{\text{occurs with small probability (Hoeffding)}} + \underbrace{\mathbf{1}_{\{UCB_2(t)>\mu_1>\mu_2>LCB_2(t)\}}}_{\text{possible only if } N_2 \text{ is small}}$$

Indeed, for the last term :

$$\mathbf{1}_{\{LCB_2(t)<\mu_2<\mu_1<UCB_2(t)\}} \leq \mathbf{1}_{\{2\sqrt{\frac{\alpha \log t}{2N_2(t)}}>(\mu_1-\mu_2)\}} = \mathbf{1}_{\{N_2(t)\leq\frac{2\alpha \log T}{(\mu_1-\mu_2)^2}\}}$$

Let $x = \frac{2\alpha \log T}{(\mu_2-\mu_1)^2}$ and $X = \sum_{t=\tau}^{T-1} \mathbf{1}_{\{UCB_1(t)\leq\mu_1\}} + \mathbf{1}_{\{\mu_2<LCB_2(t)\}}$. We have

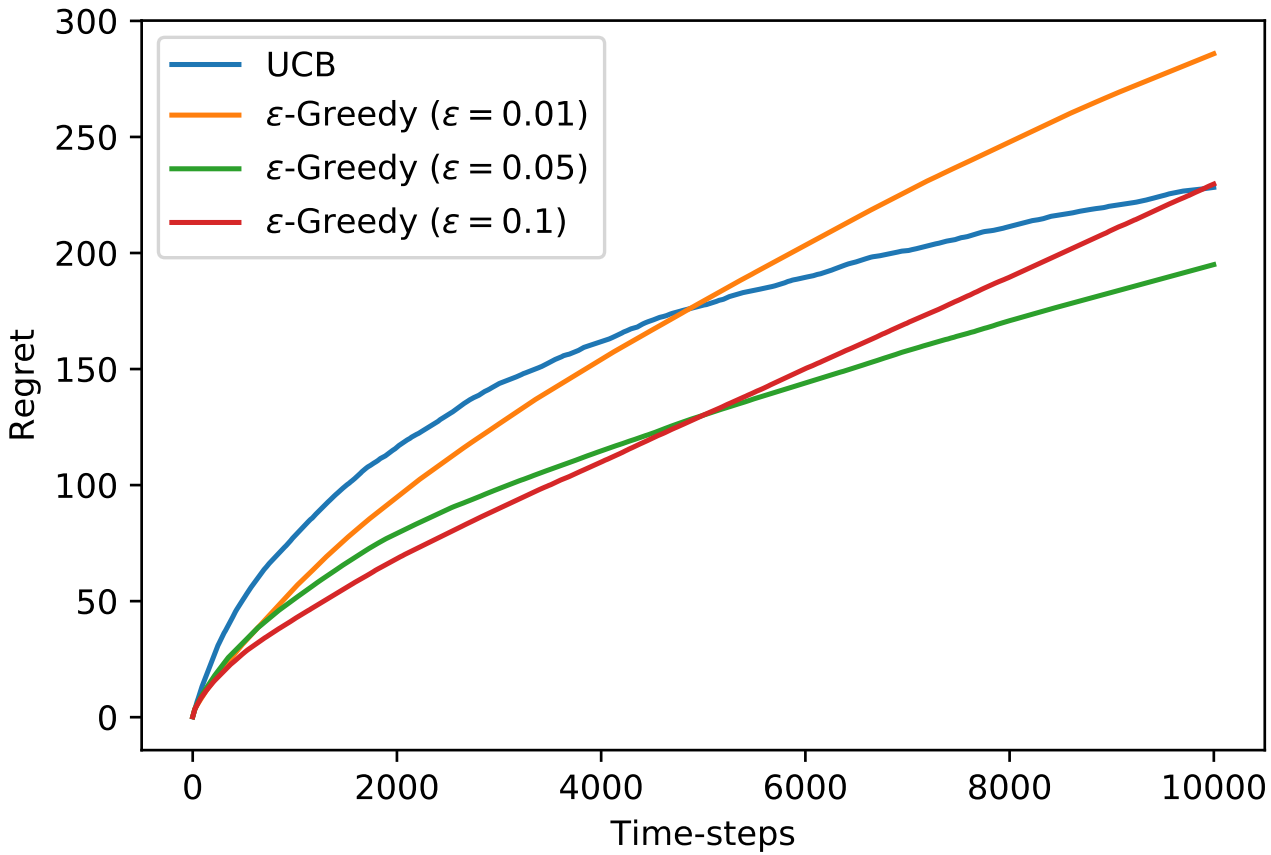
$$N_2(T) - N_2(\tau) \leq X + \sum_{t=\tau}^{T} \mathbf{1}_{\{N_2(t)\leq x\}}.$$

where by Hoeffding's inequality, $\mathbb{E}[X] \leq 2\sum_{t=1}^{\infty} 1/t^{\alpha-1} =: C_\alpha$.
By setting $\tau = \inf\{t > 0 : N_2(t) \geq x\}$, this implies that :

$$\begin{aligned}\mathbb{E}[N_2(T)] &\leq \mathbb{E}[N_2(\tau)] + \mathbb{E}[X] \\ &\leq C_\alpha + x.\end{aligned}$$

Numerical example (same parameter as before)

(Cumulative) regret as a function of the number of iterations (ϵ-greedy has a linear but small regret)

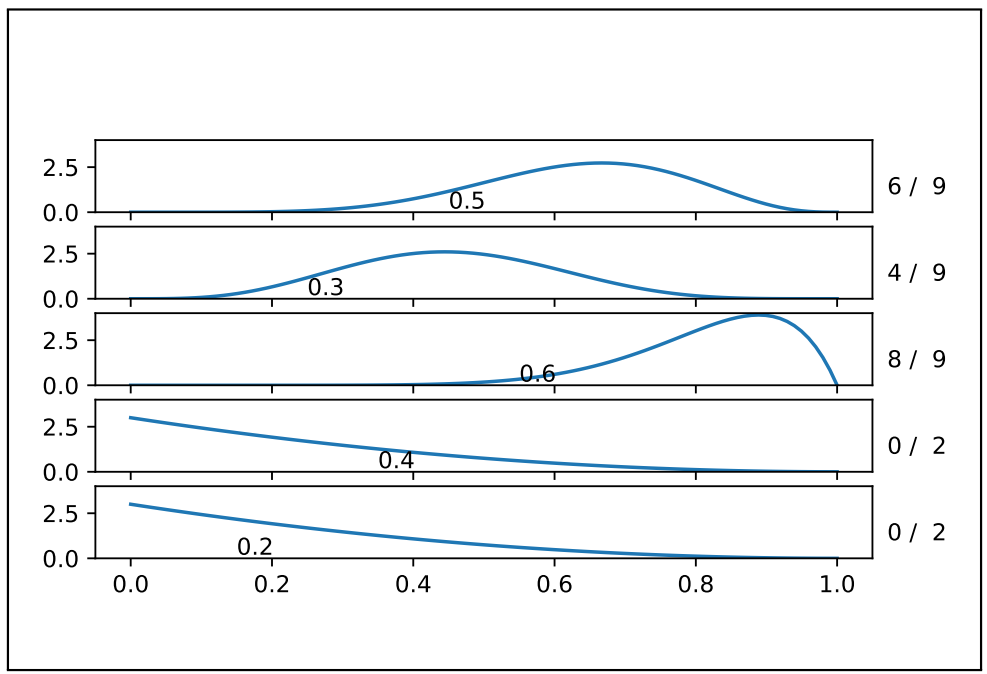


Outline

Bayesian approach : Thomson sampling

Algorithm "Thomson sampling" (1933)

- For each arm i, you maintain a probability (posterior) distribution π_i^t that represent your knowledge on μ_i.
- At each time step, you draw n independent samples $\theta_i \sim \pi_i^t$. You choose $I_{t+1} = \arg\max_i \theta_i$.
- You update your knowledge π_i^{t+1} by using Bayes rule.



Theorem (Kauffman,Korda,Munos 2012)

Thompson Sampling is asymptotically optimal[a], *i.e. it provides* $O(\log T)$ *regret with the smallest possible constant.*

a. https://arxiv.org/pdf/1205.4217.pdf

It is one of the most efficient algorithm (note : the analysis is similar to the one of UCB1).

Bayesian update

Consequence : Assume now that you start from a uniform prior on μ (i.e. you assume that μ is uniformly distributed on $[0,1]$ and that you draw n samples of X.
Then the distribution of μ conditioned on having observed p times the value 1 is a beta distribution, with probability density function $f_{p,n-p}(.)$:

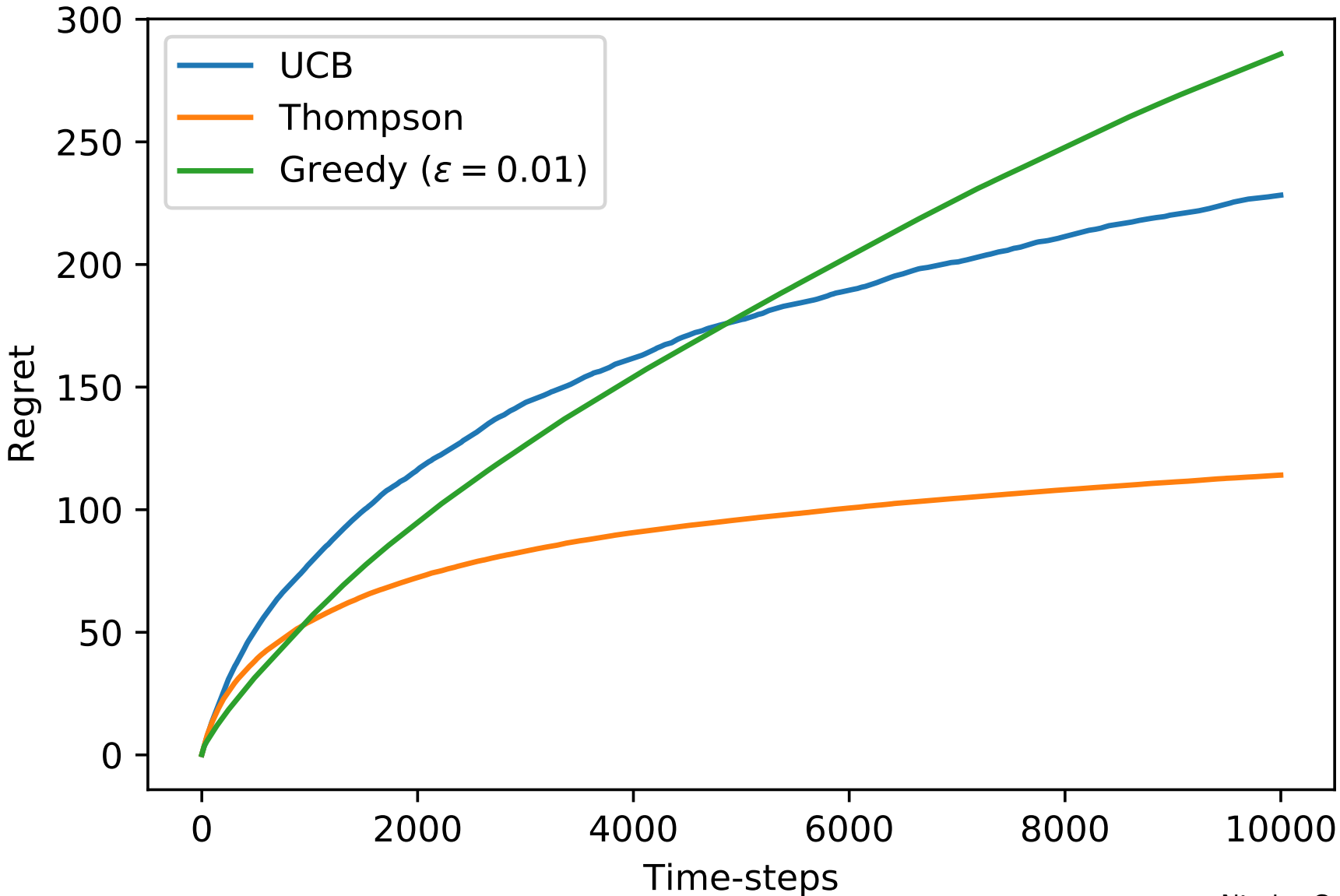
$$f_{p,n-p}(\theta) = c\theta^p(1-\theta)^{1-p},$$

where c is a constant such that the probability sums to one.

Note : can be easily adapted to more complicated prior.

Numerical example (same parameter as before)

(Cumulative) regret as a function of the number of iterations (ϵ-greedy has a linear but small regret)



Outline

The Adversarial Bandit Model

At each time step you choose an action I_t and obtain a reward R_{t,I_t}. As before, you want to minimize the expected regret :

$$\max_{i\in\{1\dots n\}} \mathbb{E}\left[\sum_{t=1}^{T} R_{t,i}\right] - \mathbb{E}\left[\sum_{t=1}^{T} R_{t,I_t}\right]$$

Now, we assume that :

- As before, you do not know R_{t,I_t}.
- Contrary to before, the distribution of R_{t,I_t} can change with time (and can depend on your past choices or rewards).

The Exp3 algorithm

Assume that you can observe all $R_{t,i}$ ("expert" case)

- Compute the accumulated score : $S_{t,i} = \sum_{k=1}^{t} R_{t,i}$
- Chooses I_{t+1} randomly such that $\mathbb{P}(I_{t+1} = i) = \frac{e^{\tau S_{t,i}}}{\sum_{j=1}^{n} e^{\tau S_{t,j}}}$.

Remarks :

- This procedures uses a softmax operator.

The Exp3 algorithm

Assume that you can observe all $R_{t,i}$ ("expert" case)

- Compute the accumulated score : $S_{t,i} = \sum_{k=1}^{t} R_{t,i}$
- Chooses I_{t+1} randomly such that $\mathbb{P}(I_{t+1} = i) = \frac{e^{\tau S_{t,i}}}{\sum_{j=1}^{n} e^{\tau S_{t,j}}}$.

Remarks :

- This procedures uses a softmax operator.
- It achieves a reward of $\frac{nT}{2}\tau + \frac{\ln n}{\tau}$
  - Choosing $\tau = 1/\sqrt{T}$ gives you a $O(\sqrt{T})$ regret.

The Exp3 algorithm

Assume that you can observe all $R_{t,i}$ ("expert" case)

- Compute the accumulated score : $S_{t,i} = \sum_{k=1}^{t} R_{t,i}$
- Chooses I_{t+1} randomly such that $\mathbb{P}(I_{t+1} = i) = \frac{e^{\tau S_{t,i}}}{\sum_{j=1}^{n} e^{\tau S_{t,j}}}$.

Remarks :

- This procedures uses a softmax operator.
- It achieves a reward of $\frac{nT}{2}\tau + \frac{\ln n}{\tau}$
  - Choosing $\tau = 1/\sqrt{T}$ gives you a $O(\sqrt{T})$ regret.
- If you can observe only R_{t,I_t}, then in the expression of $S_{t,i}$ you can replace $R_{t,i}$ by $\hat{R}_{t,i} = R_{t,I_T} \mathbf{1}_{I_t=i} / \mathbb{P}(I_t = i)$. This estimator is unbiased

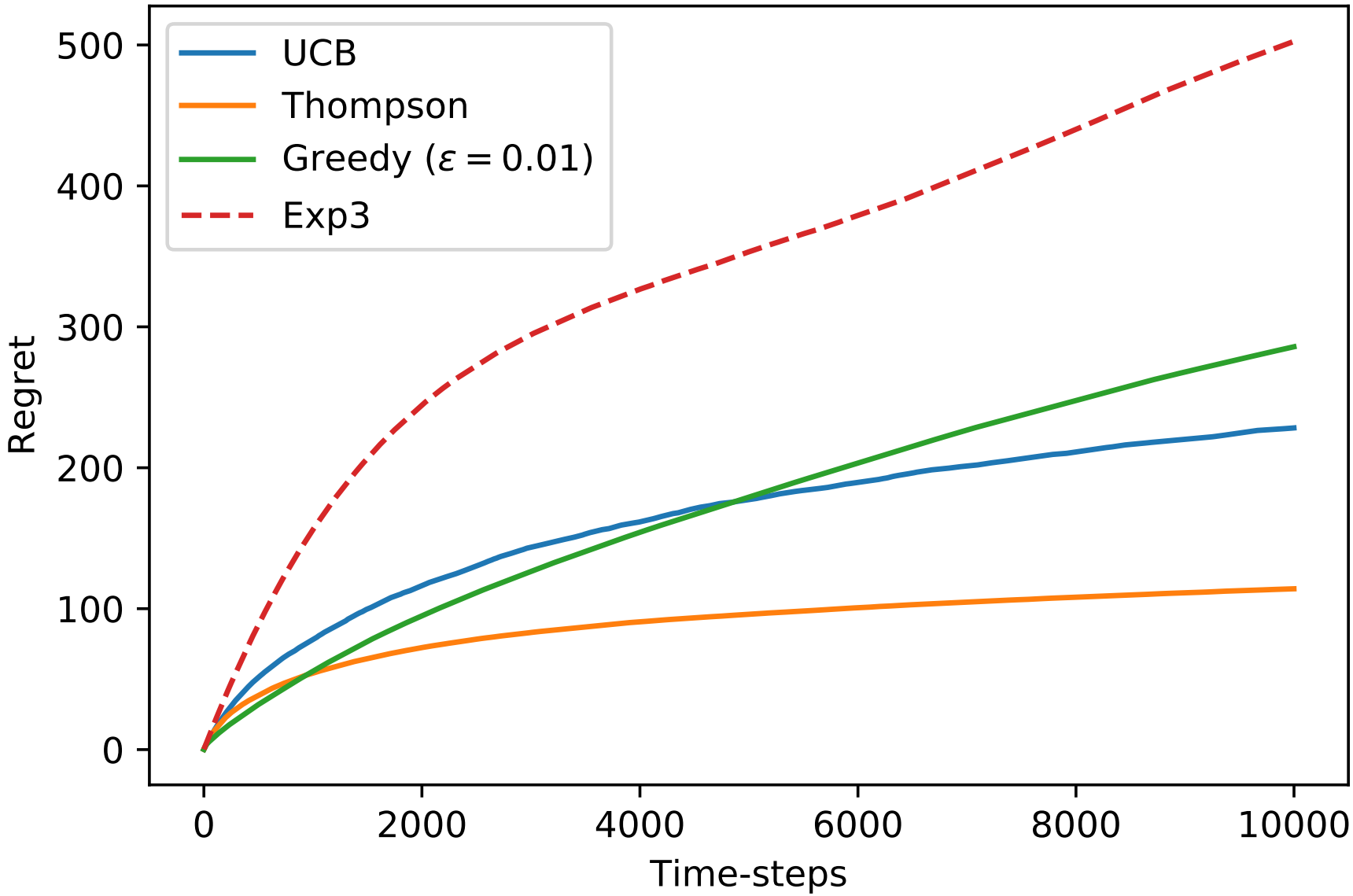
$$\mathbb{E}\left[\hat{R}_{t,i}\right] = R_{t,i}.$$

It also gives a $O(1/\sqrt{T})$ regret.

Numerical example (same parameter as before)

We look at a stochastic reward

(Cumulative) regret as a function of the number of iterations



Outline

What did we learn ?

- A way to quantify the tradeoff between exploration and exploitation.
- An algorithm that achieve this optimal tradeoff (Thompson sampling) and that uses a Bayesian framework.

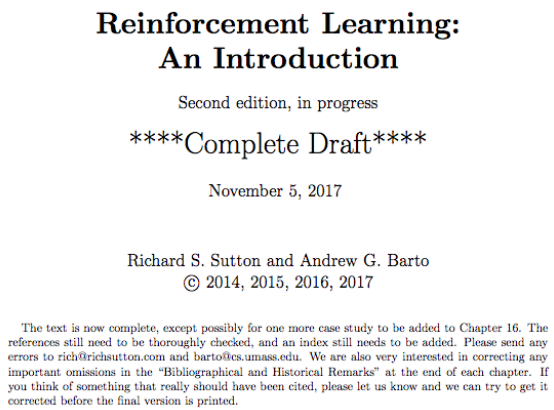
Multi-armed bandits have many variants and applications

- Online problems
- Reinforcement learning

Going further

- Markovian bandit problem
  - Index policies (Gittins and Whittle) (see Kimang and Chen)
- Combinatorial bandits
  - UCB or Exp3-like algorithm. (see Quan for application to routing)
- Continuous actions
  - Online convex optimization (see Panayotis)
- Mixing bandit problem and Markov decision processes
  - Reinforcement learning (See Bruno's talk)

Some references :



Reinforcement Learning: An Introduction

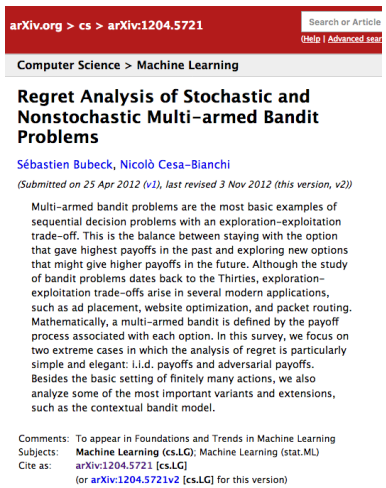
Second edition, in progress

****Complete Draft****

November 5, 2017

Richard S. Sutton and Andrew G. Barto
© 2014, 2015, 2016, 2017

The text is now complete, except possibly for one more case study to be added to Chapter 16. The references still need to be thoroughly checked, and an index still needs to be added. Please send any errors to rich@richsutton.com and barto@cs.umass.edu. We are also very interested in correcting any important omissions in the "Bibliographical and Historical Remarks" at the end of each chapter. If you think of something that really should have been cited, please let us know and we can try to get it corrected before the final version is printed.



arXiv.org > cs > arXiv:1204.5721

Search or Article

(Help | Advanced sear

Computer Science > Machine Learning

Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems

Sébastien Bubeck, Nicolò Cesa-Bianchi

(Submitted on 25 Apr 2012 (v1), last revised 3 Nov 2012 (this version, v2))

Multi-armed bandit problems are the most basic examples of sequential decision problems with an exploration-exploitation trade-off. This is the balance between staying with the option that gave highest payoffs in the past and exploring new options that might give higher payoffs in the future. Although the study of bandit problems dates back to the Thirties, exploration-exploitation trade-offs arise in several modern applications, such as ad placement, website optimization, and packet routing. Mathematically, a multi-armed bandit is defined by the payoff process associated with each option. In this survey, we focus on two extreme cases in which the analysis of regret is particularly simple and elegant: i.i.d. payoffs and adversarial payoffs. Besides the basic setting of finitely many actions, we also analyze some of the most important variants and extensions, such as the contextual bandit model.

Comments: To appear in Foundations and Trends in Machine Learning
Subjects: **Machine Learning (cs.LG)**; Machine Learning (stat.ML)
Cite as: arXiv:1204.5721 [cs.LG]
(or arXiv:1204.5721v2 [cs.LG] for this version)

Stochastic bandits Stochastic & Adversarial (Chapter 1)

+ ask Panayotis if you have questions.