Recommender Systems (and applications)

R. Gaudel

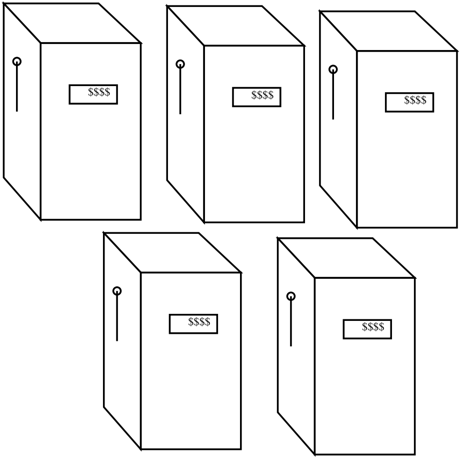
[1] ENSAI, CREST

October 2019

École nationale de la statistique et de l'analyse de l'information

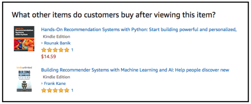
Part II

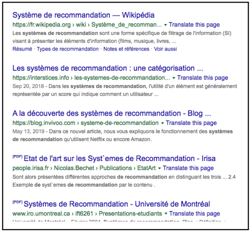
Bandits Theory

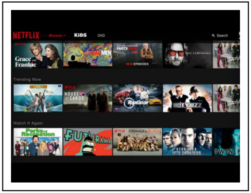


Last Time in a Nutshell

- A Zoo of Recommender Systems





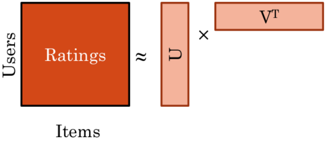


- Focus on Collaborative Filtering

Objets

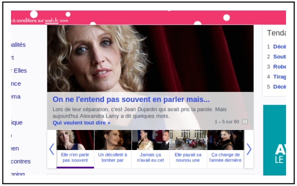
Utilisateurs

...Sequel...	TAO	QARMA
?	2	4
?	2	5
5	?	?



Today Focus: Content-based recommendation

- Example: News Recommendation

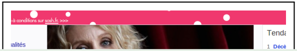


- Data
  - News' features $\mathbf{x}$: text, date, category...
  - Log of previous recommendations: $(\mathbf{x}^{(1)}, \mathit{click})$, $(\mathbf{x}^{(2)}, \mathit{click})$, $(\mathbf{x}^{(3)}, \mathit{noClick})$, $(\mathbf{x}^{(4)}, \mathit{click})$...

- Model
  - Logistic Regression...
    - ★ $\hat{z} = \mathbf{P}(z = 1|\mathbf{x}) \stackrel{def}{=} \hat{f}_{\mathbf{w}^T,b}(\mathbf{x}) = \sigma(\mathbf{x}\mathbf{w}^T + b)$
    - ★ with $\sigma(\dot{z}) = \frac{1}{1+e^{-\dot{z}}}$ (sigmoid)

Today Focus: Content-based recommendation

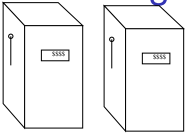
- Example: News Recommendation



- Problem solved. What else ?
- Data not at all independent !
  - Data result from past recommendations
  - Past recommendations results from a model
  - Model learned from data
  - Data result from past recommendations
  - ...
  - $\Longrightarrow$ Exploration / Exploitation trade-off

- Data
  - New
  - Log $(\mathbf{x}^{(3)}, noClick)$,
    $(\mathbf{x}^{(4)}$
- Model
  - Logistic Regression...
    - ★ $\hat{z} = \mathbf{P}(z = 1|\mathbf{x}) \overset{def}{=} \hat{f}_{\mathbf{w}^T,b}(\mathbf{x}) = \sigma(\mathbf{x}\mathbf{w}^T + b)$
    - ★ with $\sigma(\dot{z}) = \frac{1}{1+e^{-\dot{z}}}$ (sigmoid)

Oversimplified Example: Facing 2 Options

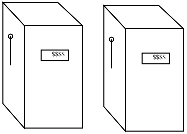


	blue option	**brown option**
nb +1:	3	60
nb 0:	7	40

Which arm to play ?

30 remaining trials
Obj: maximize total gain

Oversimplified Example: Facing 2 Options



	blue option	**brown option**
nb +1:	3	60
nb 0:	7	40
true mean:	0.7	0.6

Which arm to play ?
Play right arm
(better empirical average & higher confidence)

Is it really the best option ?
NO !
You should also explore
(from time to time)

Outline

Stochastic Multi-Armed Bandit

- Game

Learner (you, your program)	Action: pull arm i_t $\longrightarrow$ $\longleftarrow$ Reward: r_t	Environment (user, client)

- Parameters
  - K: nb arms (previously known as options) — known
  - ν_i: reward distribution of arm i — unknown
  - μ_i: $\mathbb{E}[\nu_i]$ — unknown
  - μ^*: $max_{i=1,\dots,K}\mu_i$ — unknown
  - Δ_i: $\mu^* - \mu_i$ — unknown
- Setting
  - At each time-step t
    - Choose arm i_t (to draw)
    - Get reward $r_t \sim \nu_i$
- Objective
  - Find a strategy to choose $i_1, \dots, i_T$ in order to

$$\text{maximize} \sum_{t=1}^{T} r_t$$

Outline

A/B Testing

- Context
  - Choose between option A and option B ($K = 2$ arms)
- Solution
  1. Apply both options
     - Up to time t / up to budget m
     - With random assignment
     - Log efficiency of each assignment
  2. Choose the best option
     - Given the logs
     - Using statistical test
     - Conclusion: A > B A < B A ? B
  3. Apply the winning option
     - Up to time T
- Aka. Explore Then Commit (ETC) in Bandit community

Notations

- Denote $T_{i,t-1}$ the number of trial of option i from time-step 1 to $t-1$

$$T_{i,t-1} = \sum_{s=1}^{t-1} 1_{i_s=i}$$

- Denote $\hat{\mu}_{i,t-1}$ the empirical mean reward when choosing option i from time-step 1 to $t-1$:

$$\hat{\mu}_{i,t-1} = \frac{1}{T_{i,t-1}} \sum_{s=1}^{t-1} 1_{i_s=i} r_s$$

A/B Testing Strategy

- A/B testing strategy
  - Try each of the $K = 2$ available options m/K times
  - Go with the winner for the remaining rounds

A/B testing at time-step t

- $T_{i,t-1} = \sum_{s=1}^{t-1} 1_{i_s=i}$
- $\hat{\mu}_{i,t-1} = \frac{\sum_{s=1}^{t-1} 1_{i_s=i} r_s}{T_{i,t-1}}$
- Choose option

$$i_t = \begin{cases} A, & \text{if } t \leqslant m \text{ and } (t \bmod 2 = 0) \\ B, & \text{if } t \leqslant m \text{ and } (t \bmod 2 = 1) \\ \underset{i \in \{A, B\}}{\operatorname{argmax}} \ \hat{\mu}_{i,m} & \text{if } t > m \end{cases}$$

ENSAI

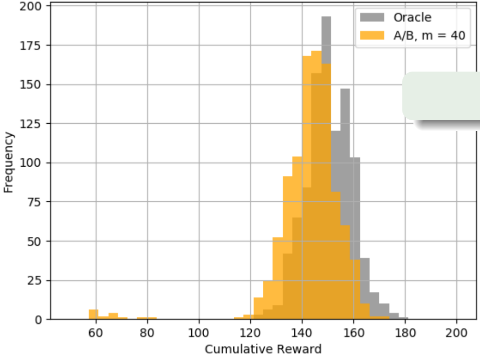
Empirical Analysis

- Specific environment
  - $r_t | i_t = A \sim \mathcal{N}(0.5, 1)$
  - $r_t | i_t = B \sim \mathcal{N}(0.2, 1)$ $\qquad (\Delta_B = 0.3)$
  - $T = 300$
  - $1,000$ "games"
- Questions
  - Sum of rewards $\sum_{t=1}^{T} r_t$ (mean value, distribution)
  - $T_{A,T}$ (mean value, distribution)

Sum of Rewards

$r_t | i_t = A \sim \mathcal{N}(0.5, 1)$ $T = 300$
$r_t | i_t = B \sim \mathcal{N}(0.2, 1)$ $1,000$ games

- Mean value
  - Oracle: 150
  - A/B ($m = 40$): 143



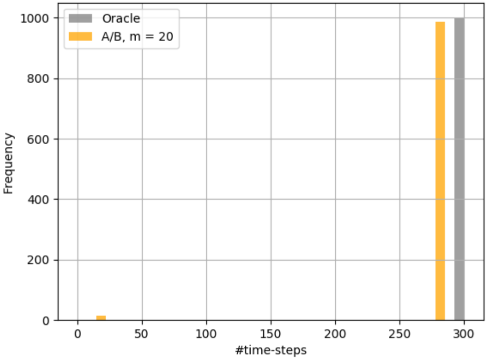
Lets explain

Histogram of Sum of rewards at time-step $T = 300$

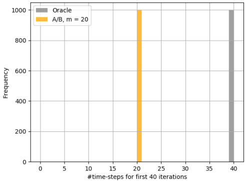
Number of Trials of Best Option

$r_t | i_t = A \sim \mathcal{N}(0.5, 1)$ $T = 300$
$r_t | i_t = B \sim \mathcal{N}(0.2, 1)$ $1,000$ games

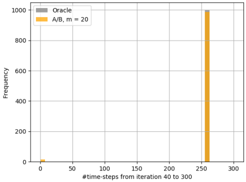
- Mean value
  - Oracle: 300
  - A/B ($m = 40$): 276



Histogram of number of trials of best option at time-step $T = 300$



On 40 first trials



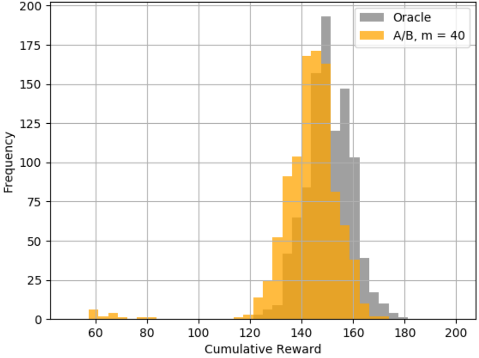
On remaining trials

Sum of Rewards

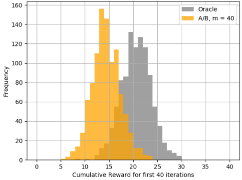
$r_t | i_t = A \sim \mathcal{N}(0.5, 1)$ $\quad T = 300$
$r_t | i_t = B \sim \mathcal{N}(0.2, 1)$ $\quad 1,000$ games

Question

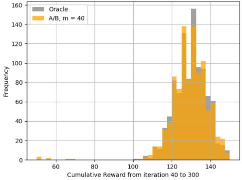
- Consequence if m is small ?
- Consequence if m is big ?



Histogram of Sum of rewards at time-step $T = 300$



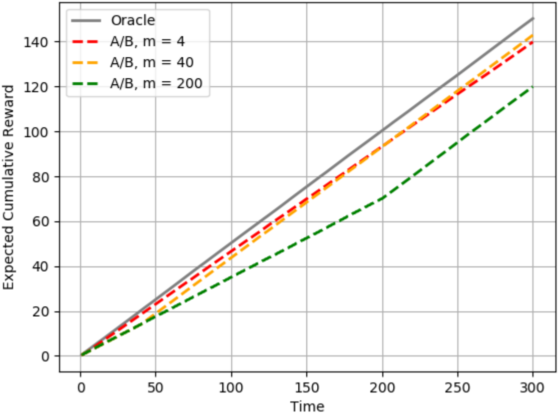
On 40 first trials



On remaining trials

Sum of Rewards

$r_t | i_t = A \sim \mathcal{N}(0.5, 1)$ $T = 300$
$r_t | i_t = B \sim \mathcal{N}(0.2, 1)$ $1,000$ games



Best value for m ?

Averaged sum of rewards from time-step $T = 0$ to $T = 300$

Theoretical Analysis

- Specific environment
  - $r_t | i_t = A \sim \mathcal{N}(\mu, 1)$
  - $r_t | i_t = B \sim \mathcal{N}(\mu - \Delta, 1)$, $\Delta > 0$
  - T trials
  - Each option tried $m/2$ times
- Questions
  - Sum of rewards $\sum_{t=1}^{T} r_t$ (mean value, distribution)
  - $\Longrightarrow T_{A,T}$ distribution
    - ★ $\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = T\mu - \mathbb{E}[T_{B,T}]\Delta$

      from $\sum_{t=1}^{T} \left(r_t.1_{i_t=A} + r_t.1_{i_t=B}\right)$ and $\mathbb{E}\left[r_t.1_{i_t=A} | i_t = i\right] = 1_{i_t=i}\mathbb{E}\left[r_t | i_t = i\right]$
    - ★ $T_{A,T}$ has only two possible outcomes: $m/2$ or $m/2 + (T - m)$

Questions

- Lemma
  - Upper-bound of the probability to identify option B as the best after m trials
- "Theorem"
  - Lower-Bound on expected sum of rewards at time T
- Corollary
  - Optimal value for m

Theoretical Analysis

- Specific environment
  - $r_t | i_t = A \sim \mathcal{N}(\mu, 1)$
  - $r_t | i_t = B \sim \mathcal{N}(\mu - \Delta, 1)$, $\Delta > 0$
  - T trials
  - Each option tried $m/2$ times

Usefull

- Let $z \sim \mathcal{N}(\mu, \sigma)$

$$\mathbb{P}\left(\frac{z}{\sigma} > s\right) \leqslant \exp\left(\frac{s^2}{2}\right)$$

Lemma

- Probability to identify option B as the best after m trials

$$\mathbb{P}\left(\underset{i \in \{A, B\}}{\operatorname{argmax}} \ \hat{\mu}_{i,m} = B\right) = \mathbb{P}\left(\hat{\mu}_{B,m} > \hat{\mu}_{A,m}\right)$$

$$= \mathbb{P}\left(\frac{2}{m}\sum_{t=1}^{m/2}(x_t + \mu - \Delta) > \frac{2}{m}\sum_{t=1}^{m/2}(y_t + \mu)\right)$$

$$= \mathbb{P}\left(\frac{\sum_{t=1}^{m/2} x_t - \sum_{t=1}^{m/2} y_t}{\sqrt{m}} > \frac{\sqrt{m}}{2}\Delta\right) \leqslant \exp\left(-\frac{m\Delta^2}{8}\right)$$

with $x_t \sim \mathcal{N}(0, 1)$, $y_t \sim \mathcal{N}(0, 1)$

Theoretical Analysis

- Specific environment
  - $r_t | i_t = A \sim \mathcal{N}(\mu, 1)$
  - $r_t | i_t = B \sim \mathcal{N}(\mu - \Delta, 1)$, Δ
  - T trials
  - Each option tried $m/2$ tim

Usefull

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = T\mu - \mathbb{E}[T_{B,T}]\Delta$$

$$\mathbb{P}\left(\underset{i \in \{A, B\}}{\operatorname{argmax}} \; \hat{\mu}_{i,m} = B\right) \leqslant \exp\left(-\frac{m\Delta^2}{8}\right)$$

Theorem

- Expected cumulative reward (to lower-bound)

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = T\mu - \frac{m}{2}\Delta - (T - m)\Delta\mathbb{P}\left(\underset{i \in \{A, B\}}{\operatorname{argmax}} \; \hat{\mu}_{i,m} = B\right)$$

$$\geqslant T\mu - \frac{m}{2}\Delta - (T - m)\Delta \exp\left(-\frac{m\Delta^2}{8}\right)$$

Theoretical Analysis

- Specific environment
  - $r_t | i_t = A \sim \mathcal{N}(\mu, 1)$
  - $r_t | i_t = B \sim \mathcal{N}(\mu - \Delta, 1)$, $\Delta > 0$
  - T trials
  - Each option tried $m/2$ times

Corollary

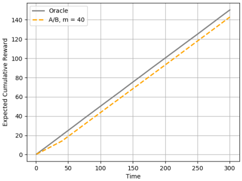
- Recall

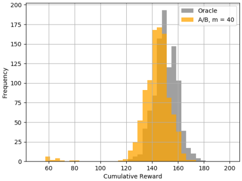
$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \geqslant T\mu - \frac{m}{2}\Delta - (T-m)\Delta \exp\left(-\frac{m\Delta^2}{8}\right)$$

- Best value for m (with T large enough)

$$m = \frac{8}{\Delta^2}\log\left(\frac{T\Delta^2}{4}\right) \quad \longrightarrow \quad \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \geqslant T\mu - \Delta - \frac{4}{\Delta}\left(1 + \log\left(\frac{T\Delta^2}{4}\right)\right)$$

Take-Home Message





- Recap
  - Explore then Commit (to te good or to the bad)
  - Optimal m of the order $\frac{1}{\Delta^2} \log(T.\Delta)$
- Remarks
  - Optimal m depends on Δ
    - Do you know Δ?
    - What about more than 2 options?
  - Optimal m increases with T
    - Consequence if T unknown?

Outline

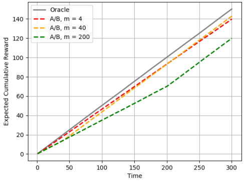
Regret

- Bandit true objective : find a strategy to choose $i_1, \ldots, i_T$ in order to

$$\text{minimize } R_T \overset{def}{=} T.\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[T_{i,T}\right] \quad \text{(a.k.a (pseudo-)regret)}$$

as a replacement for maximize $\sum_{t=1}^{T} r_t$

- Any "interesting" algorithm "converges": $\frac{1}{T}\sum_{t=1}^{T} r_t \xrightarrow[T\to\infty]{} \mu^*$
  - Equivalent to $\frac{R_T}{T} \xrightarrow[T\to\infty]{} 0$, $R_T = o(T)$
  - aka. Zero-regret learner / vanishing regret / sublinear regret
- Remaining question: at which speed ?
- Standard settings
  - $R_T = O(\sqrt{T})$ $\quad \frac{R_T}{T} = O(\frac{1}{\sqrt{T}})$
  - $R_T = O(\log(T))$ $\quad \frac{R_T}{T} = O(\frac{\log(T)}{T})$



A/B Testing Regret Bounds

- Specific environment
  - $K = 2$
  - $\nu_0 = \mathcal{N}(\mu, 1)$
  - $\nu_1 = \mathcal{N}(\mu - \Delta, 1)$, $\Delta > 0$
  - Horizon: T
  - Each option tried m times
- (Cheating) instance-dependent bound ($m = \frac{8}{\Delta^2} \log\left(\frac{T\Delta}{4\sqrt{\pi}}\right)$)

$$R_T \leqslant \frac{4}{\Delta}\left(1 + \log\left(\frac{T\Delta}{4\sqrt{\pi}}\right)\right)$$

$$R_T = O\left(\frac{1}{\Delta}\log(T)\right)$$

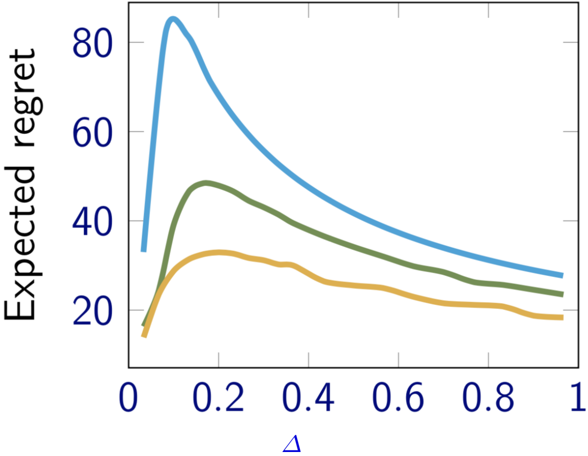
- Worst-case bound

$$R_T = O\left(\sqrt{T}\right)$$

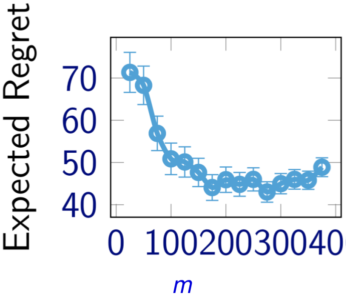
- Not anytime: best m depends on T

Experimental Analysis

$K = 2$ $\nu_0 = \mathcal{N}(0,1)$
$10,000$ games $\nu_1 = \mathcal{N}(-\Delta,1)$



$T = 1,000$



$T = 2,000$, $\Delta = 0.1$

blue: some upper-bound given m "optimal"
green / yellow: m given by theoretical analysis

Lower Bound on the Regret

Lower Bound on the Regret

For any policy that has sub-polynomial regret for all 1-subgaussian distribution (i.e., $R_T = o(T^p)$ for all p > 0 and all $\nu_1, \ldots, \nu_K$), for any set of distributions $\nu_1, \ldots, \nu_K$,

$$\liminf_{T \to +\infty} \frac{R_T}{\log(T)} \geq \sum_{i:\Delta_i > 0} \frac{2}{\Delta_i}$$

- Corollary
  - $R_T = \Omega(\log(T)) \implies$ Standard algorithms are optimal (up to a constant)
  - Explore at least $\frac{2\log(T)}{\Delta_i}$ times arm $i \implies$ Never-ending exploration

Outline

ε_n-greedy

- Spread A/B Testing exploration along time
- K arms

ε_n-greedy at time-step t

- $T_{i,t-1} = \sum_{s=1}^{t-1} 1_{i_s=i}$
- $\hat{\mu}_{i,t-1} = \frac{\sum_{s=1}^{t-1} 1_{i_s=i} r_s}{T_{i,t-1}}$
- $\varepsilon_t = cK/d^2t$, with c and d two parameters
- Pull the arm

$$i_t = \begin{cases} \text{argmax}_i \ \hat{\mu}_{i,t-1}, & \text{with prob. } (1-\varepsilon_t) \\ i, & \text{with prob. } \varepsilon_t/K \end{cases}$$

ε_n-greedy Regret

$$i_t = \begin{cases} \text{argmax}_i \ \hat{\mu}_{i,t-1}, & \text{with prob. } (1 - \varepsilon_t) \\ i, & \text{with prob. } \varepsilon_t / K \end{cases}$$

Question

- Expected number of time arm i is drawn due to exploration rule ?

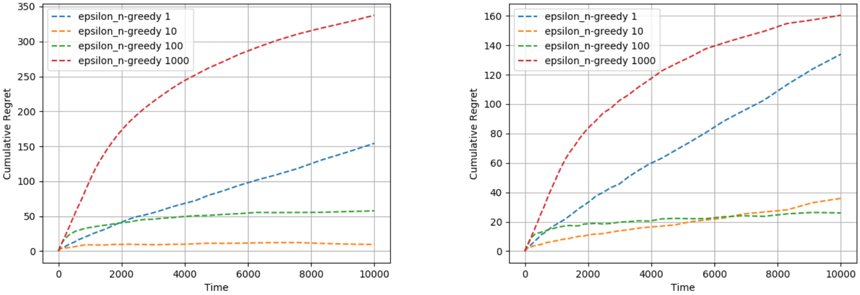
Bound on the regret of ε_n-greedy

If $0 < d < min\{\Delta_i : \Delta_i \neq 0, i \in 1, \ldots, K\}$, ν_i support is included in $[0, 1]$, and $c > 5$, and arms $i_1, \ldots, i_T$ are chosen by ε_n-greedy strategy,

$$R_T \leq \frac{K}{d^2} \ln T + o(\ln T)$$

Experimental Analysis

$K = 2$ $\nu_0 = Ber(0.9)\nu_1 = Ber(0.9\Delta)$
$T = 10,000$ 1000 replications



$\Delta = 0.2$ $\Delta = 0.1$

- c/d^2 to be tuned ...
- ... application per application

Outline

UCB1

Upper Confidence Bound (Auer et. al, 2002)

UCB1 at time-step t

- $T_{i,t-1} = \sum_{s=1}^{t-1} 1_{i_s=i}$
- $\hat{\mu}_{i,t-1} = \frac{\sum_{s=1}^{t-1} 1_{i_s=i} r_s}{T_{i,t-1}}$
- Pull the arm

$$i_t = \underset{i}{\operatorname{argmax}} \; \hat{\mu}_{i,t-1} + \sqrt{\frac{2 \ln t}{T_{i,t-1}}}$$

Be Optimist in the Face of Uncertainty

- From where comes $UCB(i,t) = \hat{\mu}_{i,t-1} + \sqrt{\frac{2\ln t}{T_{i,t-1}}}$?
- Remark: if $X_1, X_2, \ldots, X_n$ are independent and σ-subgaussian with mean μ and $\hat{\mu} = \frac{\sum_{s=1}^{n} X_s}{n}$, then for any $\varepsilon \geqslant 0$

$$\mathbb{P}\left(\hat{\mu} \leqslant \mu - \varepsilon\right) \leqslant \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

  - What is $UCB(i,t)$?

$$\mathbb{P}\left(\mu \geqslant UCB(i,t)\right) \leqslant \frac{\exp\left(-1/\sigma^2\right)}{t}$$

- Graphics / Demo
- Remark: $T_{i,t-1}$ depends on values r_t, so the analysis is (partially) wrong.

Application

- Recall: $UCB(i,t) = \hat{\mu}_{i,t-1} + \sqrt{\frac{2 \ln t}{T_{i,t-1}}}$
- For following situations (nb of wins / nb of trials)
  1. Which arm a "greedy" strategy would pull ?
  2. Which arm would you pull ?
  3. Which arm UCB1 would pull ?

arm 1	arm 2	arm 3
3 / 13	60 / 160	6 / 20
9 / 10	8 / 10	7 / 10
18 / 20	8 / 10	7 / 10

- An arm has been pulled T times with empirical mean $\hat{\mu}$. At which iteration, its UCB1 value will be greater than $\hat{\mu} + \delta$?

Regret

Problem-Dependent Bound for UCB1

If ν_i support is included in $[0, 1]$ and arms $i_1, \ldots, i_T$ are chosen by UCB1 strategy,

$$R_T \leq \sum_{i:\Delta_i>0} \frac{8}{\Delta_i} \ln T + O(1)$$

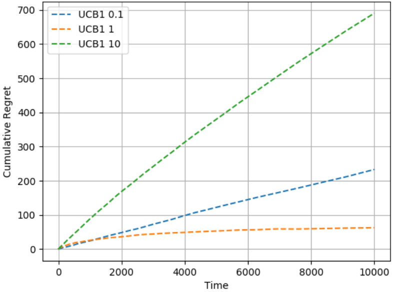
Worst-Case Bound for UCB1

If ν_i support is included in $[0, 1]$ and arms $i_1, \ldots, i_T$ are chosen by UCB1 strategy,

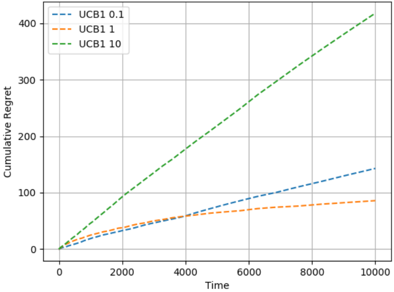
$$R_T \leq 8\sqrt{TK \ln T} + O(1)$$

Experimental Analysis

$K = 2 \quad \nu_0 = Ber(0.9)\nu_1 = Ber(0.9\Delta)$
$T = 10,000 \quad$ 1000 replications



$\Delta = 0.2$

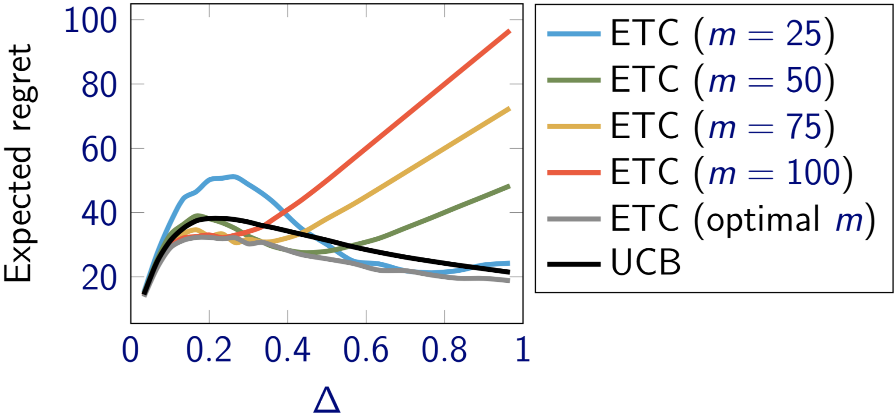


$\Delta = 0.1$

- Good behavior ...
- ... with standard parameters

Experimental Analysis

$K = 2 \qquad \nu_0 = \mathcal{N}(0, 1)$
$T = 1,000 \qquad \nu_1 = \mathcal{N}(-\Delta, 1)$



- UCB almost the best while no parameter to tune
- Results similar with ε_n-greedy
- (So simple to implement)

UCB: a Huge Family

- UCB2
- UCB-V (learn the variance)
- KL-UCB (almost optimal for Bernouilli distributions)
- AO-UCB (asymptotically optimal on 1-subgaussian distributions)
- ...

Outline

Thompson Sampling (for Bernoulli distributions)

a.k.a. Probability Matching, Bayesian Bandits

- Assumption: ν_i are Bernoulli distributions

Thompson Sampling at time-step t

- Let $\tilde{\mu}_i$ be a sample from $Beta(S_{i,t} + 1, F_{i,t} + 1)$ for each arm i
- Pull the arm

$$i_t = \underset{i}{\text{argmax}}\ \tilde{\mu}_i$$

- Get reward r_t
- $(\tilde{r} \sim Bernouilli(r_t))$
- If $r == 1$, $\quad S_{i,t} \leftarrow S_{i,t} + 1 \quad$ else $\quad F_{i,t} \leftarrow F_{i,t} + 1$

- Exercise: ν is Bernoulli distribution of parameter μ, with a uniform prior on μ. After T trials, you did collect S successes and F fails. What's the posterior distribution for μ ?

Apply Bayesian Framework

- Generative model

$$\mu_i \sim \mathit{Uniform}([0,1]) \qquad \forall i \tag{1}$$
$$r_t \mid i_t \sim \mathit{Bernouilli}\left(\mu_{i_t}\right) \tag{2}$$

- A posteriori distribution on μ_i (after T trials: S successes and F fails)

$$\Pr\left(\mu_i \mid S, F\right) \propto \mu_i^S (1-\mu_i)^F$$

Corresponds to distribution $\mathit{Beta}(S+1, F+1)$

Regret

Bound on the regret of Thompson Sampling

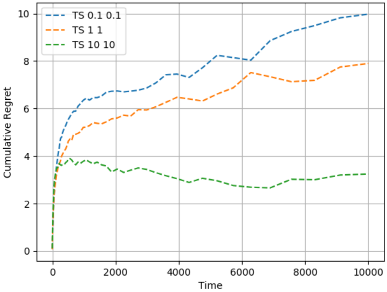
If ν_i support is included in $[0, 1]$, and arms $i_1, \ldots, i_T$ are chosen by Thompson Sampling strategy,

$$R_T \leq O\left(\left(\sum_{i:\Delta_i>0} \frac{1}{\Delta_i^2}\right)^2 \ln T\right)$$

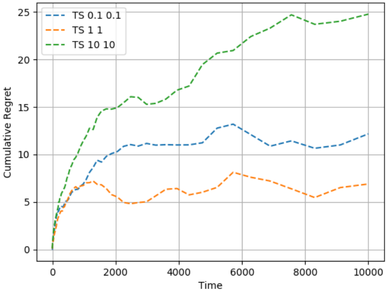
Experimental Analysis

$K = 2$ $\quad \nu_0 = Ber(0.9) \nu_1 = Ber(0.9\Delta)$

$T = 10,000$ $\quad$ 1000 replications



$\Delta = 0.2$



$\Delta = 0.1$

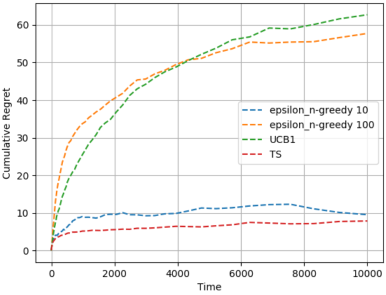
- Good behavior ...
- ... with a large range of parameters

Outline

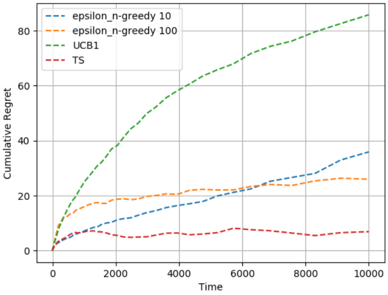
Experimental Analysis

$K = 2$ $\quad \nu_0 = Ber(0.9) \nu_1 = Ber(0.9\Delta)$

$T = 10,000$ $\quad$ 1000 replications



$\Delta = 0.2$



$\Delta = 0.1$

- GG to Thompson Sampling !

(Some) Known regret Bounds

	bound on the regret
lower bound	$\liminf_{T\to+\infty} \frac{R_T}{\ln T} \geq \sum_{i:\Delta_i>0} \frac{\Delta_i}{\mathrm{kl}(\mu_i,\mu*)}$
ε_n-greedy	$\frac{K}{d^2} \ln T + o(\ln T)$
UCB1	$\sum_{i:\Delta_i>0} \frac{8}{\Delta_i} \ln T + O(1)$
Thompson Sampling	$O\left(\left(\sum_{i:\Delta_i>0} \frac{1}{\Delta_i^2}\right)^2 \ln T\right)$
KL-UCB (Bernouilli arms)	$\alpha \sum_{i:\Delta_i>0} \frac{\Delta_i}{\mathrm{kl}(\mu_i,\mu*)} \ln T + O(1)$

Respective Strengths

- ε_n-greedy
  - Easy to apply to more tricky learning systems
  - Random
  - $\mathbb{P}$(pull arm i at time-step t) > 0
- UCB1
  - Easy to tune
- Thompson Sampling
  - Easy to apply to more tricky learning systems (if you're Bayesian)
  - Random
  - $\mathbb{P}$(pull arm i at time-step t) > 0

Applications

- MCTS (search in trees)
  - Go-playing Artificial Intelligences (up to AlphaGo)
  - General game (artificial) players
- $\boxed{\text{A/B Testing}}$
- Big names (Google and Co.), at least ε_n-greedy

Outline

A/B Testing vs. Anytime Bandits

- Drawback of anytime approaches: maintain all he options
  - Code choosing the arm
  - Data Storage
  - Clients affectation
  - ...

A/B Testing may be less expensive
$\Rightarrow$ use dedicated bandits (simple regret)

Simple Regret

Rational

- Focus on minimizing / controlling $p_e = \mathbb{P}(\text{selected arm} \neq \text{best arm})$
- ... with m as small as possible
- Typically: fix p_e and adapt m to data

Adaptive Approaches

- Spread exploration budget non-uniformly
- Examples:
  - Successive Reject: stop exploration as soon as possible ... arm by arm
  - k-best arms identification[a]

[a] Emilie Kaufmann, Olivier Cappé, Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. The Journal of Machine Learning Research, Volume 17 Issue 1, 2016

Example: Successive Reject

Successive Reject Algorithm

- $A = \{1, \ldots, K\}$
- $n_0 = 0$
- $\forall k \in \{1, \ldots, K-1\}, n_k = \frac{\alpha}{K+1-k}$
  // Exploration Phase
- For k in $1, \ldots, K-1$
  - Select $n_k - n_{k-1}$ times each arm in A
  - Identify the worst arm i in A
  - Remove i: $A \leftarrow A \setminus \{i\}$

  // Explotation Phase
- For remaining t
  - Select the only remaining arm

Theoretical Analysis of Successive Reject

Probability of Failure of Successive Reject

For some constant c,

$$\mathbb{P}(\text{selected arm} \neq \text{best arm}) \leq K^2 \exp\left(-c \frac{n}{\log(K) \sum_{i:\Delta_i>0} \frac{1}{\Delta_i^2}}\right)$$

Outline

Context

- A huge/infinite number of arms
  - How to manage it ?
- Examples
  - News
  - Advertisement
  - Songs
  - Youtube videos
- $\Rightarrow$ Put arms in a metric space
  - Neighbor arms have similar reward distribution

Contextual Bandit

- Parameters
  - $\mathcal{X} \subseteq \mathbb{R}^d$: set of arms — known
  - $\theta^* \in \mathbb{R}^d$: parameters — unknown
- Setting
  - At each time-step t
    - ★ choose arm x_t (to draw)
    - ★ get reward $r_t \sim \nu_{x_t}$ s.t. $\mathbb{E}[r_t] = \langle x_t, \theta^* \rangle$
- Objective
  - Find a strategy to choose $x_1, \dots, x_T$ in order to

$$\text{minimize } R_T = T . \max_{x \in \mathcal{X}} \langle x, \theta^* \rangle - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \qquad \text{(a.k.a (pseudo-)regret)}$$

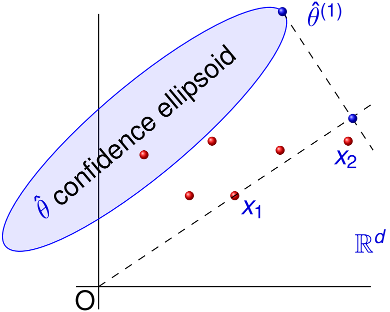
- Question: estimator of θ ?

ENSAI

OFUL

Optimism in the Face of Uncertainty
Linear Bandit Algorithm (Abbasi-Yadkori et. al, 2011)

- Optimism in face of uncertainty strategy on the estimator of θ^*



OFUL at time-step t

- Let $\mathbf{V}_t = \lambda.I + \sum_{s=1}^{t} x_s.x_s^T$
- Denote $\hat{\theta}$ the (regularized) least square estimator of θ
- Let $C_t = \left\{\tilde{\theta} \in \mathbb{R}^d : \|\hat{\theta} - \tilde{\theta}\|_{\mathbf{V}_t} \leq R\sqrt{d \ln\left(\frac{1+tL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}s\right\}$
- Pull the arm

$$x_t = \underset{(x,\tilde{\theta})\in(\mathcal{X},C_t)}{\text{argmax}} \langle x, \tilde{\theta}\rangle$$

Regret Bound

Bound on the regret of OFUL

Under conditions on distributions and bounds on expected rewards, if arms $x_1, \ldots, x_T$ are chosen by OFUL strategy, with probability at least $1 - \delta$

$$R_T \leq O\left(\sqrt{dT \ln T}\sqrt{\ln\frac{1}{\delta} + \ln T}\right)$$

GLM-UCB

- Extension to Generalized Linear Model (includes Logistic Regression)
  - Sarah Filippi, Olivier Cappé, Aurélien Garivier, Csaba Szepesvári . Parametric Bandits: The Generalized Linear Case. NIPS'10.

Outline

Context

- Recommend based on **the identity** of the object
  - Done: Multi-Armed Bandits
- Recommend based on **the features** of the object
  - Done: GLM-Bandit
- Recommend based on **the identity** of the object **and the user**
  - To be done now

Collaborative Filtering in the Bandit Setting

$$
\begin{tikzcd}
\text{known} & \begin{pmatrix}1 & . & . & 3\\ . & . & 2 & .\\ 2 & 1 & . & .\end{pmatrix} \arrow[d, "MC"'] & \begin{pmatrix}1 & . & . & 3\\ . & . & 2 & .\\ 2 & 1 & . & 4\end{pmatrix} \arrow[d, "MC"'] & \begin{pmatrix}1 & . & . & 3\\ . & . & 2 & 3\\ 2 & 1 & . & 4\end{pmatrix} \arrow[d, "MC"'] & \cdots \\
\text{inferred} & \begin{pmatrix}1\,1\,2\,3\\ 1\,1\,2\,3\\ \boxed{2\,1\,2\,\mathbf{3}}\end{pmatrix} \arrow[ur, "R+F"] & \begin{pmatrix}1\,1\,2\,3\\ \boxed{1\,1\,2\,\mathbf{4}}\\ 2\,1\,2\,4\end{pmatrix} \arrow[ur, "R+F"] & \begin{pmatrix}1\,2\,2\,3\\ \boxed{\mathbf{2}\,1\,2\,3}\\ 2\,1\,2\,4\end{pmatrix} \arrow[ur, "R+F"] &
\end{tikzcd}
$$

MC : matrix completion
R+F : recommendation + feedback

Collaborative Filtering in the Bandit Setting

$$\begin{tikzcd}
\text{known} & \begin{pmatrix} 1 & . & . & 3 \\ . & . & 2 & . \\ 2 & 1 & . & . \end{pmatrix} \arrow[d, "\text{MC}"'] & & \begin{pmatrix} 1 & . & . & 3 \\ . & . & 2 & . \\ 2 & 1 & . & 4 \end{pmatrix} \arrow[d, "\text{MC}"'] & & \begin{pmatrix} 1 & . & . & 3 \\ . & . & 2 & 3 \\ 2 & 1 & . & 4 \end{pmatrix} \arrow[d, "\text{MC}"'] & & \cdots \\
\text{inferred} & \begin{pmatrix} 1 & 1 & 2 & 3 \\ 1 & 1 & 2 & 3 \\ \boxed{2\;1\;2\;3} \end{pmatrix} \arrow[urr, "\text{R+F}"] & & \begin{pmatrix} 1 & 1 & 2 & 3 \\ \boxed{1\;1\;2\;4} \\ 2 & 1 & 2 & 4 \end{pmatrix} \arrow[urr, "\text{R+F}"] & & \begin{pmatrix} 1 & 2 & 2 & 3 \\ \boxed{2\;1\;2\;3} \\ 2 & 1 & 2 & 4 \end{pmatrix} \arrow[urr, "\text{R+F}"] & &
\end{tikzcd}$$

MC : mat

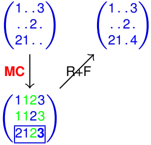
R+F : rec

A sequence of recommendations

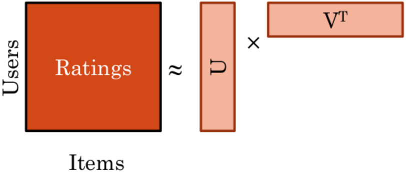
$\Longrightarrow$ Exploration-Exploitation dilemma

Approach 1: SeALS

(F. Guillou & R. Gaudel & P. Preux, 2016)



Matrix Completion
ALS-WR



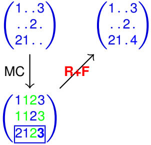
$$\underset{\mathbf{U},\mathbf{V}}{\text{argmin}} \sum_{(i,j)\in\mathcal{S}} \left(\mathbf{R}_{i,j} - \mathbf{U}_i \mathbf{V}_j^T\right)^2 + \lambda \left(\sum_i \#\mathcal{J}(i) ||\mathbf{U}_i||^2 + \sum_j \#\mathcal{I}(j) ||\mathbf{V}_j||^2 \right)$$

Algorithm: alternate

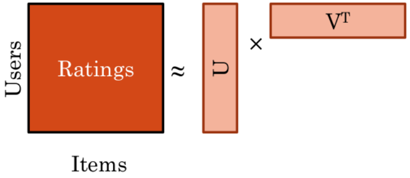
1. Fix **U** and solve remaining least square problem
2. Fix **V** and solve remaining least square problem

Approach 1: SeALS

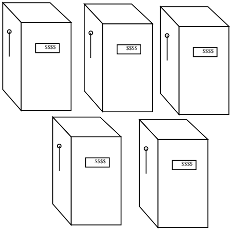
(F. Guillou & R. Gaudel & P. Preux, 2016)



Matrix Completion
ALS-WR

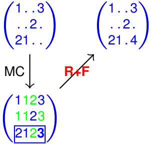


Recommendation
ε_n**-greedy**

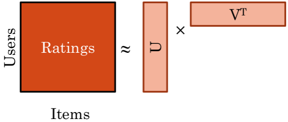


Approach 2: BeWARE

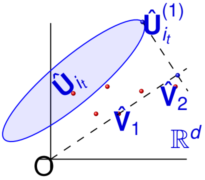
(J. Mary & R. Gaudel & P. Preux, 2015)



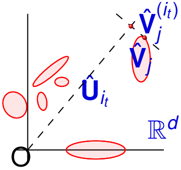
Matrix Completion
ALS-WR



Recommendation
LinUCB (two flavors)



Confidence interval on users



Confidence interval on items

PTS

(Jaya Kawale, Hung Bui, Branislav Kveton, Long Tran Thanh, Sanjay Chawla. Efficient Thompson Sampling for Online Matrix-Factorization Recommendation. NIPS'2015)

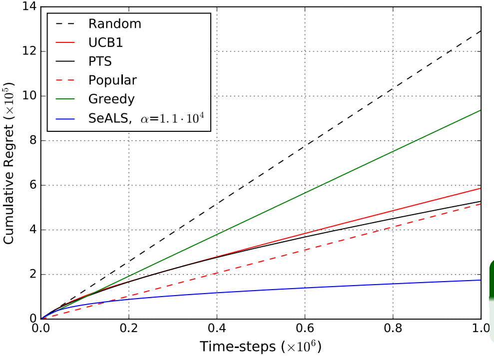
- Apply Thompson Sampling strategy to model

$$\mathbf{U}_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2 I_K)$$

$$\mathbf{V}_j \overset{iid}{\sim} \mathcal{N}(0, \sigma_v^2 I_K)$$

$$r_{i,j} | U, V \overset{iid}{\sim} \mathcal{N}(\mathbf{U}_i \mathbf{V}_j^T, \sigma^2)$$

Experimental Results on MovieLens-1M



Cumulative regret vs. time-step

- MovieLens 1M
  - 6,040 × 3,706
- Setting
  - Start with empty matrix
  - Perform 10^6 recom. 1 by 1
  - Store cumulative regret

Conclusion

- Exploration helps !

Outline

Context

- The environment is adversarial (no more "random iid")
  - Why ?
  - Which constraints remains ?
  - How to manage it ?
- Why ?
  - Why not ? Can we trust the "iid" assumption ?
  - What about shifting probabilities ?
  - Small regret, even in the worst case
- Which constraints remain ?
  - Potential values for r_t chosen in advance : $X_1, \dots, X_T \in \mathbb{R}^K$
  - At each time-step t
    - Learner chose action i_t
    - Learner gets reward $r_t = X_{t,i_t}$
  - Environment do not react to actions.

Regret

- Regret

$$R_T = \mathbb{E}\left[max_{i=1,\dots,K} \sum_{t=1}^{T} X_{t,i} - \sum_{t=1}^{T} X_{t,i_t} \right]$$

- Worst-case regret

$$R_T^* = \sup_{X_1,\dots,X_T \in \mathbb{R}^K} R_T(X_1, \dots, X_T)$$

- Some important questions
  - Does there exists a strategy s.t. $R_T^* = o(n)$? (Yes)
  - How small can we make R_T^*? $(O(\sqrt{Kn}))$
  - Let see Exp3 which achieves that worst-case regret

ENSAI

Exp3

Exponential-weight algorithm for Exploration and Exploitation

- Assumption: $X_1, \ldots, X_T \in [0,1]^K$

Exp3 t

- $\forall i, P_{ti} \leftarrow \dfrac{\exp(\eta S_{t-1,i})}{\sum_{j=1}^{K} \exp(\eta S_{t-1,j})}$
- Sample $i_t \sim P_t$
- Get reward r_t
- $\forall i, S_{t,i} \leftarrow S_{t-1,i} + 1 - \frac{\mathbb{1}_{i_t=i}(1-r_t)}{P_{ti}}$

- Rational $\mathbb{E}\left[1 - \frac{\mathbb{1}_{i_t=i}(1-r_t)}{P_{ti}}\right] = X_{ti}$

Regret Bound

Bound on the regret of Exp3

Let $X_1, \ldots, X_T \in [0,1]^K$, $\eta = \log(K)/(2TK)$, the expected regret of Exp3 satisfies

$$R_T \leqslant \sqrt{2TKlog(K)}.$$

- Remark: back to worst-case bound of iid setting.

Outline

Conclusion

- Context: choose the best option
- Optimality: requires to balance exploration and exploitation
- Strategies
  - A/B Testing (not anytime $\Rightarrow$ not optimal)
  - Dozens of "better" solutions

- Do we care about optimality ?
  - A/B Testing for strategies (with simple regret algorithms)
    - $\Theta(T)$ loss (aka. regret)
  - Basic algorithms for products, Movies, ads ... (when advanced prediction models)
    - $O(\sqrt{T})$ loss
    - Each option to maintain
  - "Advanced" algorithms for products, Movies, ads ... (when simple prediction models)
    - $O(\log(T))$ loss
    - Each option to maintain

And More

- Adapt bandit to specific setting
  - Time varying best arm
  - Restriction on what to serve
  - Cannot serve same arm twice (movie, song) $\Rightarrow$ probabilistic algorithms
  - Adversarial context
  - Delayed feedback / update on nights
  - Baseline arm
  - ...
- Take a look at
  - Bandit algorithms for Website optimization. John Myles White. O'Reilly Media.
  - Sebastien Bubeck's blog and tutorials
  - Tor Lattimore and Csaba Szepesvári's online book and tutorials
  - Bubeck and Cesa-Bianchi's book
  - ...