New Perspectives for Multi-Armed Bandits and Their Applications

Vianney Perchet

Workshop Learning & Statistics
IHES, January 19 2017

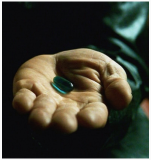
CMLA, ENS Paris-Saclay

Motivations & Objectives

Classical Examples of Bandits Problems

- Size of data: n patients with some proba of getting cured
- Choose one of two treatments to prescribe

or



- Patients **cured** or **dead**

1) **Inference:** Find the best treatment between the red and blue

2) **Cumul:** Save as many patients as possible

Classical Examples of Bandits Problems

- Size of data: *n* banners with some proba of click
- Choose one of two ads to display

or

- Banner **clicked** or **ignored**

1) **Inference:** Find the best ad between the red and blue
2) **Cumul:** Get as many clicks as possible

Classical Examples of Bandits Problems

- Size of data: *n* auctions with some expected revenue
- Choose one of two strategies(bid/opt out) to follow

or

- Auction **won** or **lost**

1) **Inference:** Find the best strategy between the red and blue

2) **Cumul:** Win as many profitable auctions as possible

Classical Examples of Bandits Problems

- Size of data: n mails with some proba of spam
- Choose one of two actions: spam or ham

or

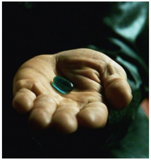
- Mail **correctly** or **incorrectly** classified

1) **Inference:** Find the best strategy between the red and blue
2) **Cumul:** Minimize number of errors as possible

Classical Examples of Bandits Problems

- Size of data: n patients with some proba of getting cured
- Choose one of two treatments to prescribe

or

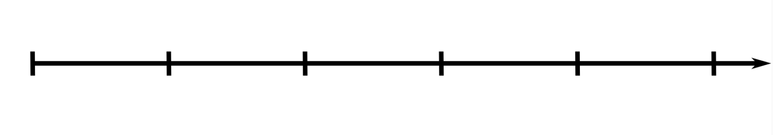


- Patients **cured** ♡ or **dead** ☠

1) **Inference:** Find the best treatment between the red and blue

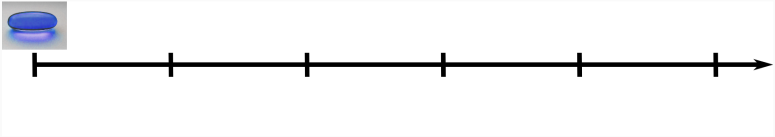
2) **Cumul:** Save as many patients as possible

Two-Armed Bandit



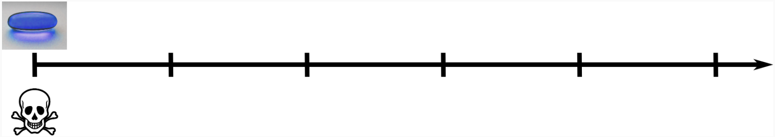
- Patients arrive and are treated sequentially.

Two-Armed Bandit



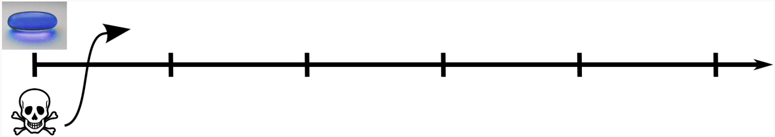
- Patients arrive and are treated sequentially.

Two-Armed Bandit



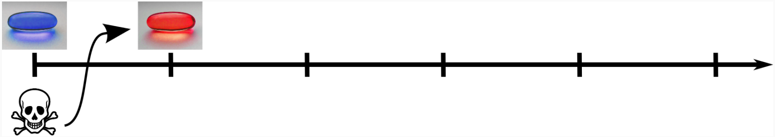
- Patients arrive and are treated sequentially.

Two-Armed Bandit



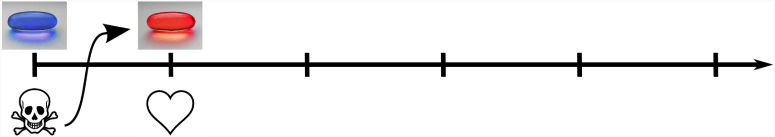
- Patients arrive and are treated sequentially.

Two-Armed Bandit



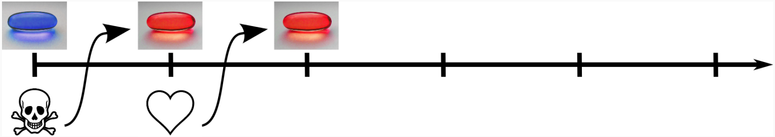
- Patients arrive and are treated sequentially.

Two-Armed Bandit



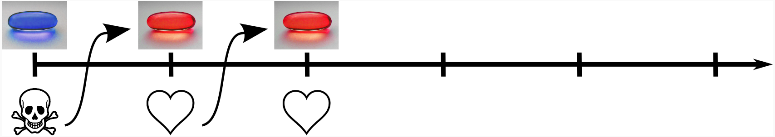
- Patients arrive and are treated sequentially.

Two-Armed Bandit



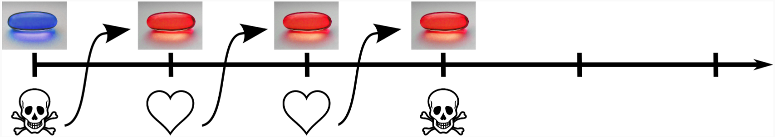
- Patients arrive and are treated sequentially.

Two-Armed Bandit



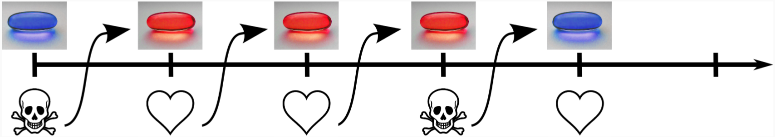
- Patients arrive and are treated sequentially.

Two-Armed Bandit



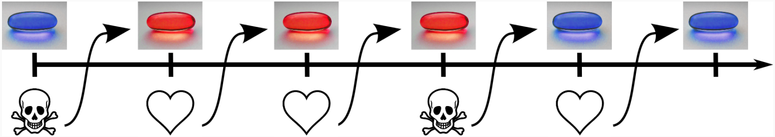
- Patients arrive and are treated sequentially.

Two-Armed Bandit



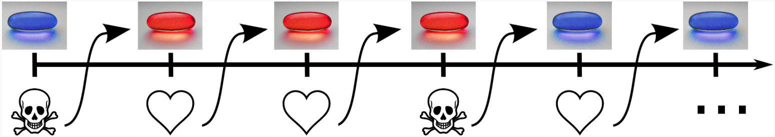
- Patients arrive and are treated sequentially.

Two-Armed Bandit



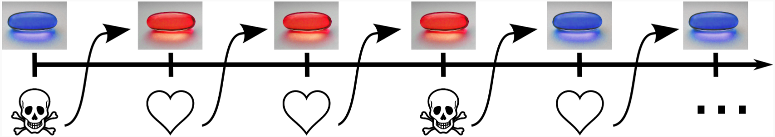
- Patients arrive and are treated sequentially.

Two-Armed Bandit



- Patients arrive and are treated sequentially.

Two-Armed Bandit



- Patients arrive and are treated sequentially.
- Save as many as possible.

A bit of theory

Stochastic Multi-Armed Bandit

K-Armed Stochastic Bandit Problems

- K actions $i \in \{1, \ldots, K\}$, outcome $X_t^i \in \mathbb{R}$ (sub-)Gaussian, bounded

$$X_1^i, X_2^i, \ldots, \sim \mathcal{N}(\mu^i, 1) \quad \text{i.i.d.}$$

- Non-Anticipative Policy: $\pi_t\left(X_1^{\pi_1}, X_2^{\pi_2}, \ldots, X_{t-1}^{\pi_{t-1}}\right) \in \{1, \ldots, K\}$
- Goal: Maximize expected reward $\sum_{t=1}^{T} \mathbb{E} X_t^{\pi_t} = \sum_{t=1}^{T} \mu^{\pi_t}$
- Performance: Cumulative Regret

$$R_T = \max_{i \in \{1,2\}} \sum_{t=1}^{T} \mu^i - \sum_{t=1}^{T} \mu^{\pi_t} = \Delta_i \sum_{t=1}^{T} \mathbb{1}\{\pi_t = i \neq \star\}$$

with $\Delta_i = \mu^\star - \mu^i$, the “gap” or cost of error i.

Most Famous algorithm [Auer, Cesa-Bianchi, Fisher, ’02]

- UCB - “Upper Confidence Bound”

$$\pi_{t+1} = \arg\max_i \left\{ \overline{X}_t^i + \sqrt{\frac{2\log(t)}{T^i(t)}} \right\},$$

where $T^i(t) = \sum_{t=1}^t \mathbb{1}\{\pi_t = i\}$ and $\overline{X}_t^i = \frac{1}{T_t^i} \sum_{s: i_s = i} X_s^i$.

Regret:

$$\mathbb{E}\, R_T \lesssim \sum_k \frac{\log(T)}{\Delta_k}$$

Worst-Case:

$$\mathbb{E}\, R_T \lesssim \sup_\Delta K \frac{\log(T)}{\Delta} \wedge T\Delta$$
$$\simeq \sqrt{KT\log(T)}$$

Ideas of proof $\pi_{t+1} = \arg\max_i \left\{ \overline{X}^i_t + \sqrt{\frac{2\log(t)}{T^i(t)}} \right\}$

- 2-lines proof:

$$\pi_{t+1} = i \neq \star \iff \overline{X}^\star_t + \sqrt{\frac{2\log(t)}{T^\star(t)}} \leq \overline{X}^i_t + \sqrt{\frac{2\log(t)}{T^i(t)}}$$

$$\text{``}\Longrightarrow\text{''}\, \Delta_i \leq \sqrt{\frac{2\log(t)}{T^i(t)}} \Longrightarrow T^i(t) \lesssim \frac{\log(t)}{\Delta_i^2}$$

- Number of mistakes grows as $\frac{\log(t)}{\Delta_i^2}$; each mistake costs Δ_i.

$$\text{Regret at stage T} \lesssim \sum_i \frac{\log(T)}{\Delta_i^2} \times \Delta_i \approx \sum_i \frac{\log(T)}{\Delta_i}$$

- “ $\Longrightarrow$ ” actually happens with overwhelming proba
- “optimal”: no algo can always have a regret smaller than $\sum_i \frac{\log(T)}{\Delta_i}$

Other Algos

- Other algo, ETC [Perchet,Rigollet], pulls in round robin then eliminates

$$R_T \lesssim \sum_k \frac{\log(T\Delta^k)}{\Delta^k}, \text{ worst case } R_T \leq \sqrt{T\log(K)K}$$

- Other algo, MOSS [Audibert, Bubeck], variants of UCB

$$R_T \lesssim K\frac{\log(T\Delta^{\min}/K)}{\Delta^{\min}}, \text{ worst case } R_T \leq \sqrt{TK}$$

- Infinite number of actions $x \in [0,1]^d$ with $\Delta(x)$ 1 Lipschitz. Discretize + UCB gives

$$R_T \lesssim T\varepsilon + \sqrt{\frac{T}{\varepsilon}} \leq T^{2/3}$$

Very interesting....

useful ?

no...

Here is a list of reasons

On the basic assumptions

1. Stochastic: Data are not iid, patients are different

   ill-posedness, feature selection/model selection
2. Different Timing: several actions for one reward

   pomdp, learn trade bias/variance
3. Delays: Rewards not received instantaneously

   grouping, evaluations
4. Combinatorial: Several decisions at each stage

   combinatorial optimization, cascading
5. Non-linearity: concave gain, diminishing returns, etc

Investigating (past/present/futur) them

Patients are different

- We assumed (implicitly ?) that all patients/users are identical
- Treatments efficiency 9proba of clicks) depend on age, gender...
- Those covariates or **contexts** are observed/known **before** taking the decision of blue/red pill

  The decision (and regret...) should ultimately depend on it

General Model of Contextual Bandits

- **Covariates:** $\omega_t \in \Omega = [0,1]^d$, i.i.d., law μ (equivalent to) λ

  The cookies of a user, the medical history, etc.
- **Decisions:** $\pi_t \in \{1, .., K\}$

  The decision can (should) depend on the context ω_t
- **Reward:** $X_t^k \in [0,1] \sim \nu^k(\omega_t)$, $\mathbb{E}[X^k|\omega] = \mu^k(\omega)$

  The expected reward of action k depend on the context ω
- **Objectives:** Find the best decision given the request

  Minimize regret $R_T := \sum_{t=1}^{T} \mu^{\pi^\star(\omega_t)}(\omega_t) - \mu^{\pi_t}(\omega_t)$

Regularity assumptions

1. Smoothness of the pb: Every μ^k is β-hölder, with $\beta \in (0, 1]$:

$$\exists\, L > 0,\ \forall\, \omega, \omega' \in \mathcal{X},\ \|\mu(\omega) - \mu(\omega')\| \ \leq\ L\|\omega - \omega'\|^{\beta}$$

2. Complexity of the pb: (α-margin condition) $\exists C_0 > 0$,

$$\mathbb{P}_X \left[0 < \left|\mu^1(\omega) - \mu^2(\omega)\right| < \delta\right] \ \leq\ C_0 \delta^{\alpha}$$

Regularity assumptions

1. **Smoothness of the pb:** Every μ^k is β-hölder, with $\beta \in (0, 1]$:

$$\exists\, L > 0,\ \forall\, \omega, \omega' \in \mathcal{X},\ \|\mu(\omega) - \mu(\omega')\| \ \leq\ L\|\omega - \omega'\|^{\beta}$$

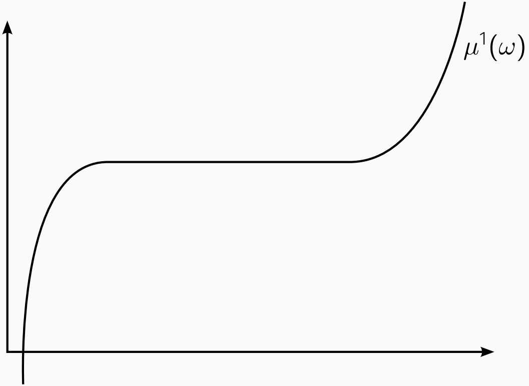
2. **Complexity of the pb:** (α-margin condition) $\exists C_0 > 0$,

$$\mathbb{P}_X\left[0 < \left|\mu^\star(\omega) - \mu^\sharp(\omega)\right| < \delta\right] \ \leq\ C_0 \delta^{\alpha}$$

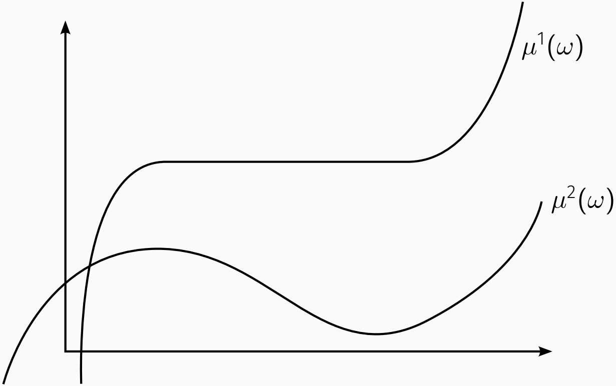
where $\mu^\star(\omega) = \max_k \mu^k(\omega)$ is the maximal μ^k and
$\mu^\sharp(\omega) = \max\left\{\mu^k(\omega)\, s.t.\, \mu^k(\omega) < \mu^\star(\omega)\right\}$ is the second max.

With $K > 2$: $\mu^\star$ is β-Hölder but $\mu^\sharp$ is not continuous.

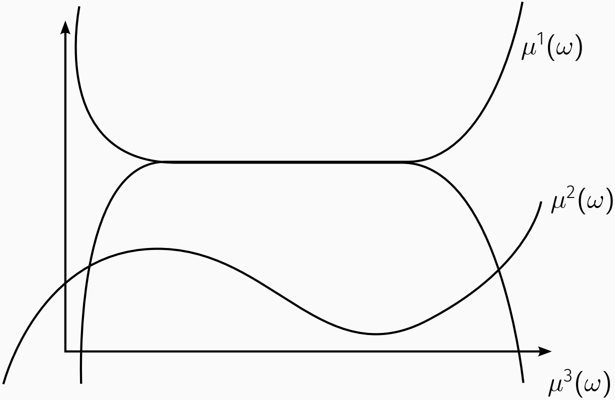
Regularity: an easy example (α big)



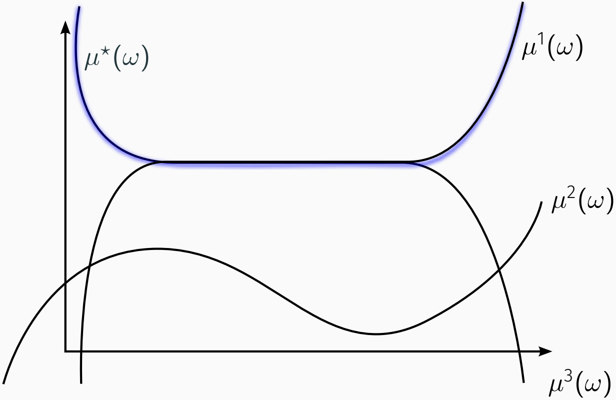
Regularity: an easy example (α big)



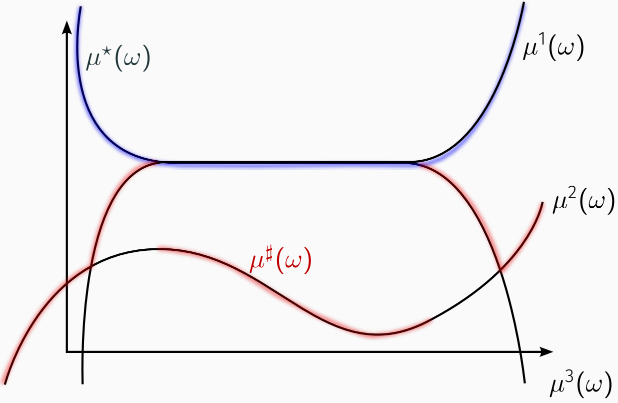
Regularity: an easy example (α big)



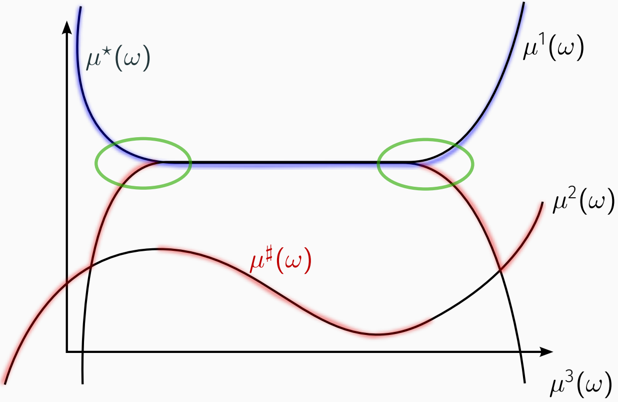
Regularity: an easy example (α big)



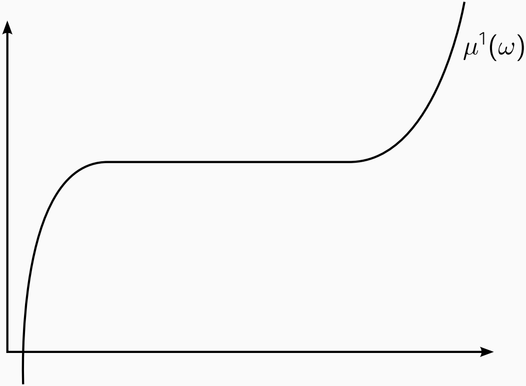
Regularity: an easy example (α big)



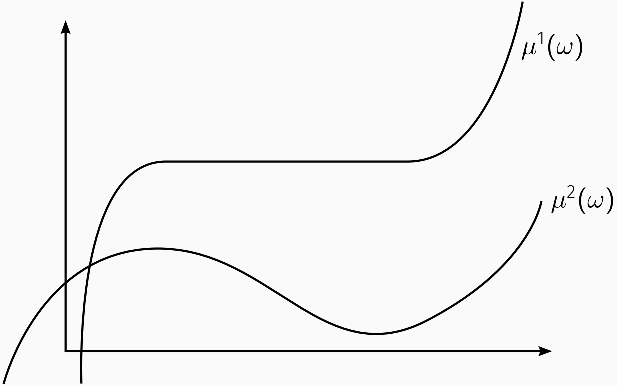
Regularity: an easy example (α big)



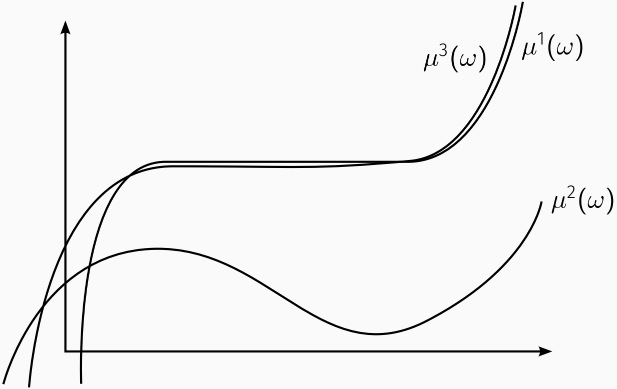
Regularity: a hard example (α small)



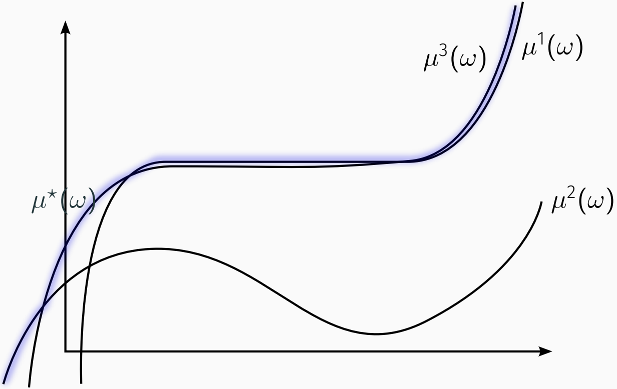
Regularity: a hard example (α small)



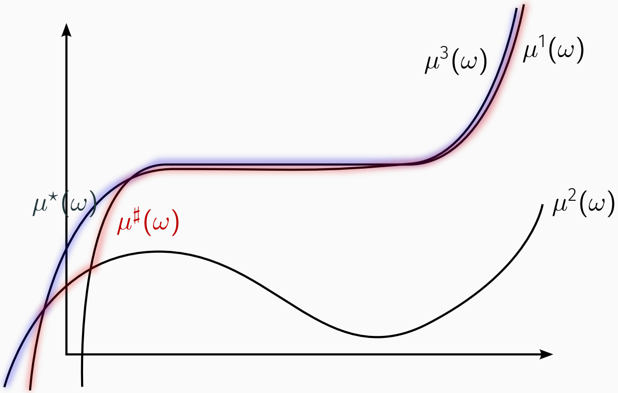
Regularity: a hard example (α small)



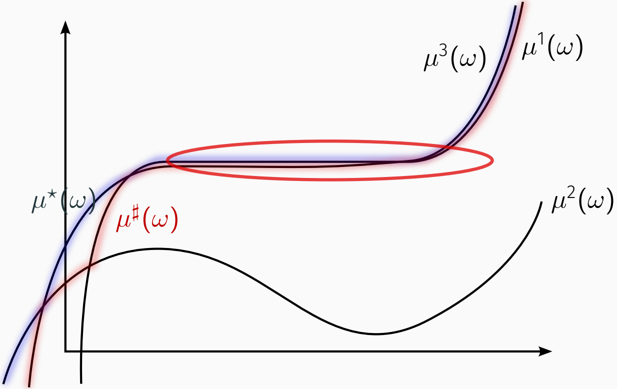
Regularity: a hard example (α small)



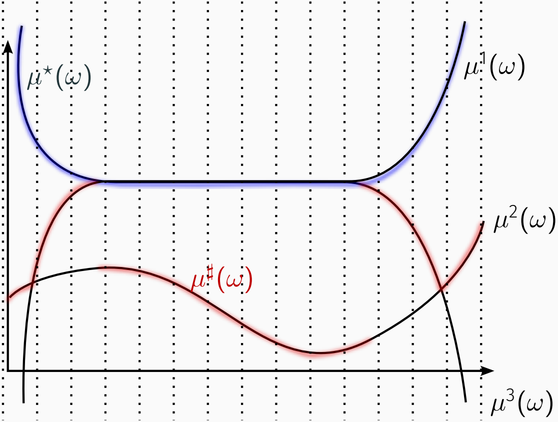
Regularity: a hard example (α small)



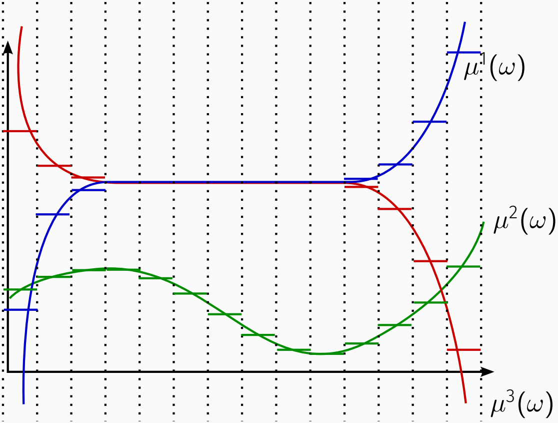
Regularity: a hard example (α small)



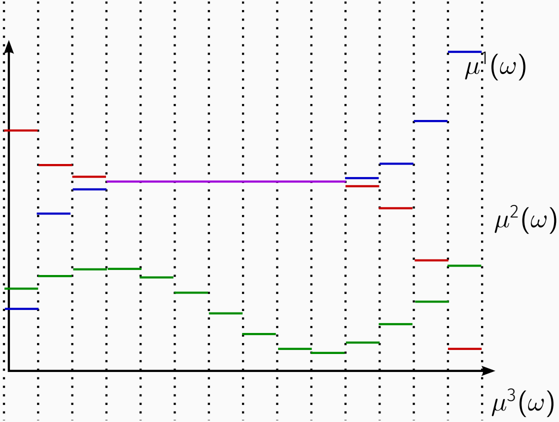
Binned policy



Binned policy



Binned policy



Binned Successive Elimination (BSE)

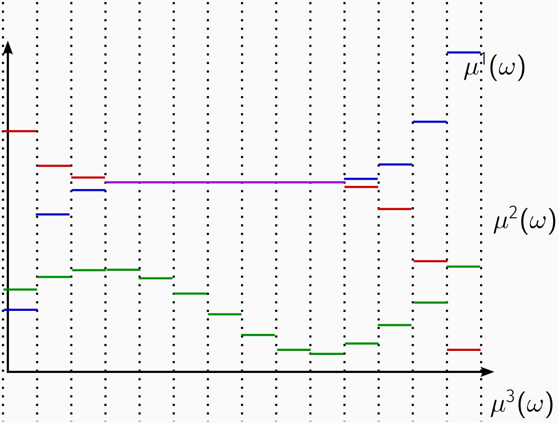
Theorem [P. and Rigollet ('13)]

If $\alpha < 1$, $\mathbb{E}[R_T(\mathrm{BSE})] \lesssim T\left(\frac{K\log(K)}{T}\right)^{\frac{\beta(1+\alpha)}{2\beta+d}}$, bin side $\left(\frac{K\log(K)}{T}\right)^{\frac{1}{2\beta+d}}$.

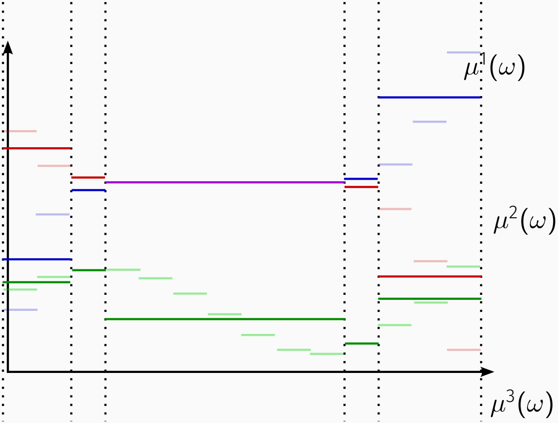
For $K = 2$, matches lower bound: minimax optimal w.r.t. T.

- Same bound with full monit [Audibert and Tsybakov, '07]
- No $\log(T)$: difficulty of nonparametric estimation washes away the effects of exploration/exploitation.
- $\alpha < 1$: cannot attain fast rates for easy problems.
- Adaptive partitioning !

Suboptimality of (BSE) for $\alpha \geq 1$



Suboptimality of (BSE) for $\alpha \geq 1$



Adaptive BSE (ABSE)

Theorem [P. and Rigollet ('13)]

For all α, $\mathbb{E}[R_T(\mathrm{ABSE})] \lesssim T\left(\frac{K\log(K)}{T}\right)^{\frac{\beta(1+\alpha)}{2\beta+d}}$.

For $K = 2$, matches lower bound: minimax optimal w.r.t. T.

- Same bound than (BSE) even for easy problems $\alpha \geq 1$.

This is not the solution

1. dimensions dependent bound: $T^{1-\frac{\beta}{2\beta+d}}$

   $d = +\infty$ **and** $\beta = 0$, lots of contexts, no regularity

   Online selection of models ?

   Ill-posed pb $\mu(\cdot)$ not β-holder

   Estimation/Approx errors

   Performance = Approx Error + Regret(β, d, T)

2. Non-stationarity of **arms**: Value are not i.i.d., evolve with time. Ex. ads for movies.

   Cumulative objectives clearly not the solution.

   Discount ? How, why, at which speeds ?

3. Non-stationarity of **sets** of arms:

   Arms arrive and disappears

   How incorporate a new arm ? which index ?

This was really not the solution

1. Non-stationarity of **sets** of arms:

   Arms arrive and disappears

   How incorporate a new arm ? which index ?

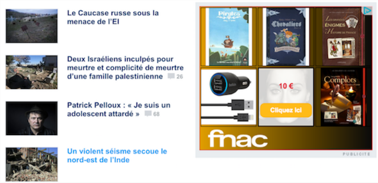
2. Contexts (covariates) are not in $\mathbb{R}^d$

   Rather descriptions, texts, id, images...How to embed ?

   training set is influenced by algorithms...

Different Timing

Example of Repeated Auctions

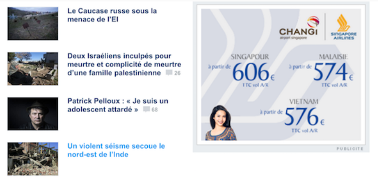


Le Caucase russe sous la menace de l'EI

Deux Israéliens inculpés pour meurtre et complicité de meurtre d'une famille palestinienne 26

Patrick Pelloux : « Je suis un adolescent attardé » 68

Un violent séisme secoue le nord-est de l'Inde



Le Caucase russe sous la menace de l'EI

Deux Israéliens inculpés pour meurtre et complicité de meurtre d'une famille palestinienne 26

Patrick Pelloux : « Je suis un adolescent attardé » 68

Un violent séisme secoue le nord-est de l'Inde

CHANGI
airport singapore
SINGAPORE AIRLINES
SINGAPOUR à partir de 606€ TTC vol A/R
MALAISIE à partir de 574€ TTC vol A/R
VIETNAM à partir de 576€ TTC vol A/R
PUBLICITÉ

Ad slot sold by lemonde.fr. **2nd-price auctions**

- Several (marketing) companies places bids
- Highest bid wins (...), say criteo, **pays to lemonde** 2nd bid (...)
- criteo chooses ad of a client, fnac or singapore airlines
- criteo **paid by the client** if the user clicks on the ad

Main Problem: Repeated auctions with unknown private valuation

Learn valuations, find which ad to display & **good strategies**

Repeated auctions

1. Can be modeled as a bandit pb with **Extra Structure**
2. Actually, Criteo (Google, Facebook) paid if **the user buys something after the click**

Needs several ”costly” auctions to seal a deal

Auctions lost can also help to seal deal (competitor displays ad for free)

Optimal strategy in repeated auctions, **learn it** ? (POMDP ?)

Reward timing per user,
decision timing by opportunities

Other examples - repeated A/B tests

- Companies test new technologies (algo, hardware, etc.) before putting in productions. Sequences of AB tests

  Timing of Decisions: each day, continue, stop or validate the current AB test

  Timing of Rewards: Total improvements of implemented techno.
- The longer AB test are, the more confident (reduces variance) but less and less implementation

Online tradeoff risks/performances

Delays

Rewards are not observed immediately

- **Clinical trials**: have to wait 6 months to see results.

  A trial length is 3 year : 6 phases

  Regret is still $\sqrt{T}$
- **Marketing** (ad displays), only see if users buy

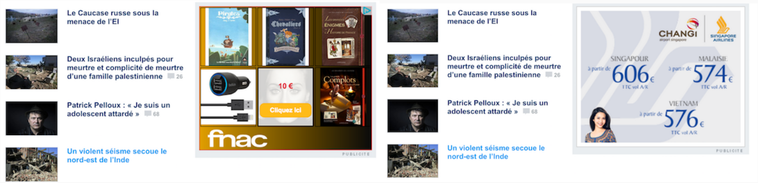
  No feedback is either no sale (forever) or no sale **yet**

  Build estimators with censured/missing data

  Feasible with iid data... but they are not!

Combinatorial Structure

Large Decision spaces



Le Caucase russe sous la menace de l'EI

Deux Israéliens inculpés pour meurtre et complicité de meurtre d'une famille palestinienne 26

Patrick Pelloux : « Je suis un adolescent attardé » 68

Un violent séisme secoue le nord-est de l'Inde

Le Caucase russe sous la menace de l'EI

Deux Israéliens inculpés pour meurtre et complicité de meurtre d'une famille palestinienne 26

Patrick Pelloux : « Je suis un adolescent attardé » 68

Un violent séisme secoue le nord-est de l'Inde

- Choose not to display 1 ad, but 4, 6, 10...
- Paid if sales after click (even if unrelated)

  Lots of correlations (between products, positions, colors/style of banner, **time**, etc.)

  Some products are seen, other are not (carrousels...)
- Too many possibilities of (almost) equal performances

  Compete with the best $R_T \leq \sqrt{KT}$

  but at least top 5%, $R_T \leq \sqrt{\log(K)\frac{1}{5\%}T}$??

Bandit theory is quite neat

To be ”applied”, or relevant, need LOTS of work

Anybody is welcome to join & collaborate!

Model selection, Feature extractions, Missing Data, Censured Data,

Combinatorial Optimization, New techniques estimators..