Multi-armed bandits & repeated auctions

Vianney Perchet (with J. Weed & P. Rigollet, MIT)

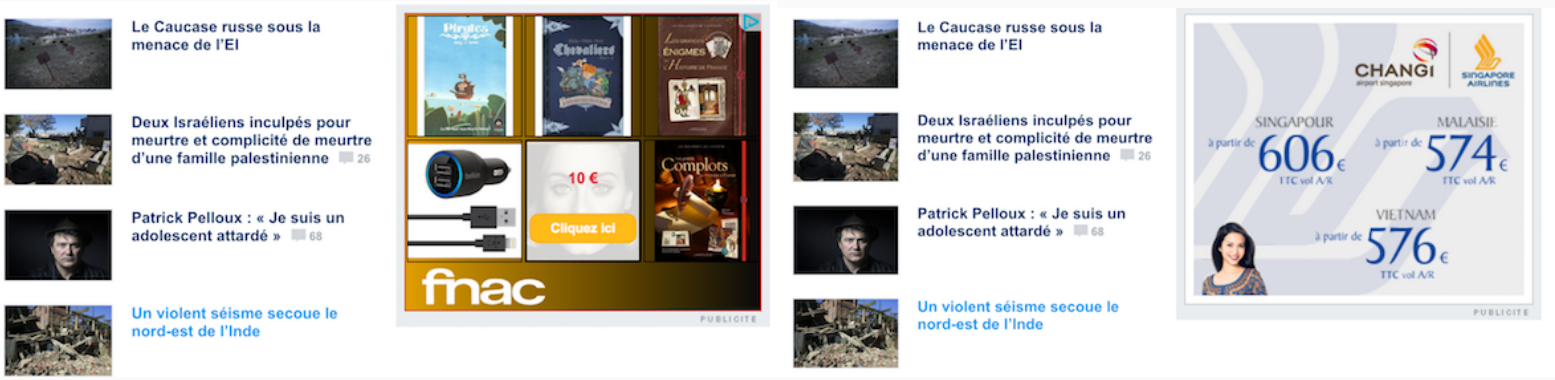
GEL 2016 Winter Workshop
Luchon, January 7 2016

CREST, ENSAE

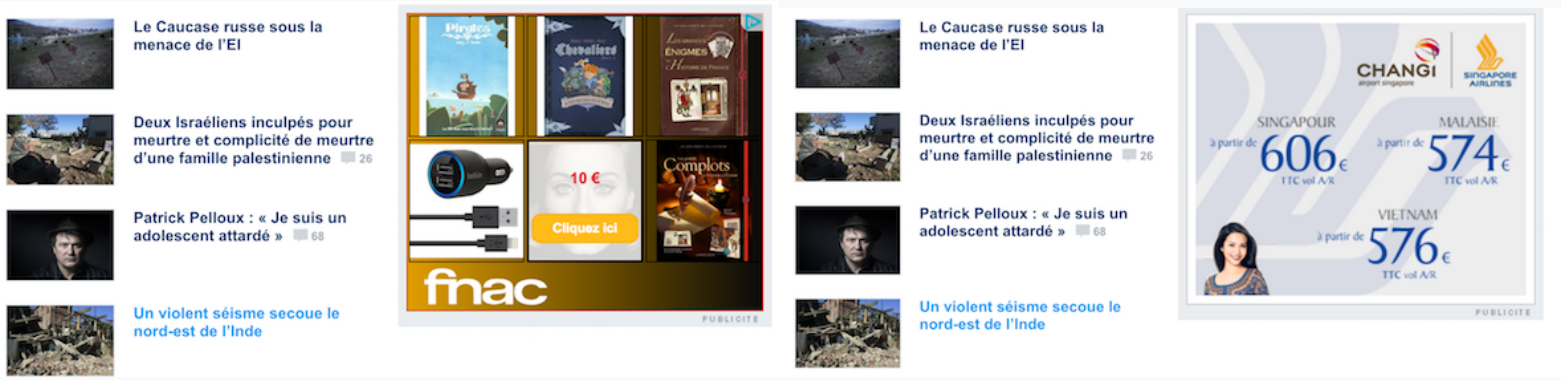
Motivations & Objectives

Example of Repeated Auctions

Example of Repeated Auctions



Example of Repeated Auctions



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 gets **paid by the client** if the user clicks on the ad

Main Problem: Repeated auctions with unknown private valuation

Learn the valuation and construct **good strategies**

Main Model

- Learning optimal reserve price [Cesa-Bianchi, Gentile, Mansour]

From the point of view of a bidder

- At round $t = 1, \ldots, T$:

  bidder bids $b_t \in [0, 1]$

  if $b_t > m_t$ (maximum other bids & reserve price)

  win good, observe value $v_t \in [0, 1]$

- Total utility: $\sum_{t=1}^{T} (v_t - m_t) \mathbb{1}\{b_t > m_t\}$
- Total **regret**:

$$\max_{b \in [0,1]} \sum_{t=1}^{T} (v_t - m_t) \mathbb{1}\{b > m_t\} - \sum_{t=1}^{T} (v_t - m_t) \mathbb{1}\{b_t > m_t\}$$

Data Assumptions - Stochastic vs Adversarial

- **Stochastic:** v_t i.i.d. $\mathbb{E}[v_t] = v \in [0, 1]$

  m_t adversarial (no assumptions on the sequence)
  m_t stochastic (i.i.d. $\mathbb{E}[m_t] = m$)

  In both cases, **expected** regret attained at v.

$$\sum_{t=1}^{T} (\mathbf{v} - m_t) \mathbb{1}\{\mathbf{v} > m_t\} - \sum_{t=1}^{T} (\mathbf{v} - m_t) \mathbb{1}\{b_t > m_t\}$$

- **Adversarial:** no assumptions at all on v_t and m_t

Tools that we will use

Variants of stochastic & adversarial multi-armed bandit

Any question on the model ?

Let’s survey some bandit literature.

First, stochastic, then adversarial.

Stochastic Multi-Armed Bandit

Two-Armed Stochastic Bandit Problems

- Two actions $i \in \{1, 2\}$, 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: $\pi_t\left(X_1^{\pi_1}, X_2^{\pi_2}, \ldots, X_{t-1}^{\pi_{t-1}}\right) \in \{1, 2\}$
- Goal: Maximize expected reward $\sum_{t=1}^{T} \mathbb{E} X_t^{\pi_t} = \sum_{t=1}^{T} \mu^{\pi_t}$
- Performance: Cumulative Regret

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

with $\Delta = |\mu^1 - \mu^2|$, the "gap" or cost of error.

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:

$$\mathbb{E}\, R_T \lesssim \frac{\log(T)}{\Delta} \wedge T\Delta$$

Worst-Case:

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

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

- 2-lines proof:

$$\pi_{t+1} \neq \star \iff \overline{X}_t^\star + \sqrt{\frac{2\log(t)}{T^\star(t)}} \leq \overline{X}_t^\sharp + \sqrt{\frac{2\log(t)}{T^\sharp(t)}}$$

$$\text{“}\Longrightarrow\text{”}\ \Delta \leq \sqrt{\frac{2\log(t)}{T^\sharp(t)}} \Longrightarrow T^\sharp(t) \lesssim \frac{\log(t)}{\Delta^2}$$

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

$$\text{Regret at stage T} \lesssim \frac{\log(T)}{\Delta^2} \times \Delta \asymp \frac{\log(T)}{\Delta}$$

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

Other Algos

- More than 2 actions, $\Delta_k = \mu^* - \mu^k$, then UCB yields

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

- 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

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

Extension

- We can plan ahead the strategy by block
  - $\log(T)$ blocks for $R_T \lesssim \frac{\log(T)}{\Delta}$
  - Only 6 blocks for $R_T \lesssim \sqrt{T}$ (as $\log\log(T) \leq 6$)
- Infinite number of actions $x \in [0, 1]$ with regularity of $x \mapsto \mu^x$
- With covariates $y \in [0, 1]$, regularity of $y \mapsto \mathbb{E}[X^i|y] = \mu^i(y)$
- More complicated feedback structures
- many, many more

Adversarial Multi-Armed Bandit

K-Armed Adversarial Bandit Problems

- K actions $i \in [K] = \{1, \dots, K\}$, outcome $X_t^i \in \mathbb{R}$ bounded in $[0, 1]$

No assumption on $X_1^i, X_2^i, \dots$

- Non-Anticipative Policy: $\pi_t\left(X_1^{\pi_1}, X_2^{\pi_2}, \dots, X_{t-1}^{\pi_{t-1}}\right) \in [K]$
- **Full Monitoring:** $\pi_t\left(X_s^i\right)_{i \in [K], s \leq t-1} \in [K]$
- Goal: Maximize reward $\sum_{t=1}^n X_t^{\pi_t}$
- Performance: Cumulative Regret

$$R_T = \max_{i \in [K]} \sum_{t=1}^T X_t^i - \sum_{t=1}^T X_t^{\pi_t}$$

Full Monitoring - EXP-algo

- Main idea: $\pi_t \sim p_t \in \Delta([K])$, more weights on best actions

$$p_t^i = \frac{e^{\eta \sum_{s=1}^{t-1} X_s^i}}{\sum_{j\in[K]} e^{\eta \sum_{s=1}^{t-1} X_s^j}}, \quad \eta \text{ is a parameter}$$

- Analysis based on potential $\Phi(Z^1, \dots, Z^K) = \frac{1}{\eta} \log\left(\sum_{i\in K} e^{\eta Z^i}\right)$

$$\nabla\Phi(Z) = \left(\frac{e^{\eta Z^i}}{\sum_{j\in[K]} e^{\eta Z^j}}\right)_i \text{ and } \nabla^2\Phi(Z) \preceq \operatorname{diag}(\eta\nabla\Phi(Z))$$

- 3-lines proof, with $Z_t = \sum_{s=1}^{t} X_s$

$$\Phi(Z_T) = \Phi(Z_{T-1}) + \left\langle \nabla\Phi(Z_{T-1}), X_T \right\rangle + \frac{1}{2} X_T^\top \nabla^2\Phi(\xi_{T-1}) X_T$$

$$\leq \Phi(Z_{T-1}) + \mathbb{E}_{p_T}[X_T^{\pi_T}] + \frac{\eta}{2} \leq \Phi(0) + \mathbb{E}[\sum_{t=1}^{T} X_t^{\pi_t}] + \frac{\eta T}{2}$$

End of analysis of EXP

- We had

$$\Phi(Z_T) \leq \Phi(0) + \mathbb{E}[\sum_{t=1}^{T} X_t^{\pi_t}] + \frac{\eta T}{2}$$

- But $\max_{i\in[K]} \sum_{t=1}^{T} X_t^i \leq \Phi(Z_T)$ and $\Phi(0) = \log(K)/\eta$, so

$$\mathbb{E}R_T = \mathbb{E}\max_{i\in[K]} \sum_{t=1}^{T} X_t^i - \sum_{t=1}^{T} X_t^{\pi_t} \leq \frac{\log(K)}{\eta} + \frac{\eta T}{2} \leq \sqrt{2\log(K)T}$$

with the choice of $\eta = \sqrt{2\log(K)/T}$

- Actually we can "improve" the bound into (**holds if** $X_t^i \leq 1$)

$$\mathbb{E}R_T = \mathbb{E}\max_{i\in[K]} \sum_{t=1}^{T} X_t^i - \sum_{t=1}^{T} X_t^{\pi_t} \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{T} \sum_{i\in[K]} p_t^i (X_t^i)^2$$

Back to bandits - EXP3 algo

- Only $X_t^{\pi_t}$ is observed, not X_t. EXP cannot be used.
- Estimate X_t by $\widehat{X}_t$

$$\widehat{X}_t^i = 1 - \left(\frac{1 - X_t^i}{p_t^i}\right) \mathbb{1}\{\pi_t = i\} \text{ and run EXP on } \widehat{X}_t$$

  - $\mathbb{E}\widehat{X}_t^i = 1 - (1 - p_t^i).0 + p_t^i \frac{1-X_t^i}{p_t^i} = X_t^i$, unbiased estimator
  - $\mathbb{E}\sum_{i\in K} p_t^i (\widehat{X}_t^i)^2 \leq 1 + \sum_{i\in[K]} p_t^i \left(\frac{1-X_t^i}{p_t^i}\right)^2 p_t^i \leq K+1$ bounded variance

- Using the "improved bound" we obtain that

$$\mathbb{E}R_T \leq \frac{\log(K)}{\eta} + \eta(K+1)T \leq 3\sqrt{\log(K)KT}$$

- Can actually be improved into $\sqrt{K\log(T)}$ which is optimal.

Enough with the classical multi-armed bandits!

Back to repeated auctions

First stochastic, then adversarial

Stochastic Repeated Auctions

Our policy: UCBid

- Round 1: bid $b_1 = 1$
- Round $t+1$ bid

$$b_{t+1} = \min\left(\bar{V}_{\omega_t} + \sqrt{\frac{3\log(t)}{2\omega_t}}, 1\right)$$

where ω_t number of auctions won.

- Our first main result

Theorem - Stochastic case

UCBid yields a regret bound of

$$\mathbb{E}R_T \leq 3 + 12\frac{\log(T)}{\Delta} \wedge 5\sqrt{T\log(T)}$$

where Δ is such that no bid m_t is in the interval $(v, v+\Delta)$

Fully stochastic case: UCBid

- If $m_t \sim \mu$ satisfies margin condition, parameter α (unknown):

Definition - margin condition

$$\forall u > 0,\ \mu\{(v, v+u)\} \leq Cu^{\alpha} \text{ for some constant } C.$$

The bigger α, the easier.

Theorem - Fully stochastic case

$$\mathbb{E}R_T \leq \begin{cases} c_1 T^{\frac{1-\alpha}{2}} \log^{\frac{1+\alpha}{2}}(T) & \text{if } \alpha < 1 \\ c_2 \log^2(T) & \text{if } \alpha = 1 \\ c_3 \log(T) & \text{if } \alpha > 1 \end{cases}$$

where the constants c_1, c_2, c_3 depend on the value α.

- Almost matching lower bound

$$\mathbb{E}R_T \geq \begin{cases} c_\alpha T^{\frac{1-\alpha}{2}} & \text{if } \alpha < 1 \\ c_\alpha \log(T) & \text{if } \alpha \geq 1 \end{cases}$$

Adversarial Repeated Auctions

Our policy: EXPTree

$$\max_{b\in[0,1]} \sum_{t=1}^{T} (v_t - m_t)\mathbb{1}\{b > m_t\} - \sum_{t=1}^{T} (v_t - m_t)\mathbb{1}\{b_t > m_t\}$$

- Main idea: Nested and weighted partitions $\mathcal{P}_t$ of $[0,1]$
  - $\mathcal{P}_t = \{[m^{(s)}, m^{(s+1)}), s = 0, \dots, t-1\}$, weights $\omega_t^{(s)} \in \mathbb{R}$
  - $m_t \in [m^{(s^*)}, m^{(s^*+1)})$: **split it** into $[m^{(s^*)}, m_t)$ and $[m_t, m^{(s^*+1)})$ and assign the same weights $\omega_t^{(s*)}$ to both.
  - After the split, weights are updated as $\omega_{t+1}^s = \omega_t^s . e^{\eta \widehat{X}_t^s}$ where $\widehat{X}_{t+1}^s$ is an unbiased estimator of a bid in the interval s.
- At round $t+1$, pick an interval $\mathcal{I}_{t+1}$ in $\mathcal{P}_{t+1}$ with proba proportional to $|\mathcal{I}_{t+1}|\omega_{t+1}$.
- Finally, **bid b_{t+1} uniform in $\mathcal{I}_{t+1}$**

Performances of EXPTree

Theorem – Upper-bound

EXPTree yields a regret bounded as

$$\mathbb{E}R_T \leq 4\sqrt{T \log(1/\Delta^\circ)}$$

with Δ° the width of interval contains the best fixed bid.

Is the dependency in Δ° necessary ? **yes**

Theorem – Lower-bound

For any algo, there exists a sequence of m_t and v_t s.t.

$$\mathbb{E}R_T \geq \frac{1}{32}\sqrt{T\lfloor \log_2(1/2\Delta^\circ)\rfloor}$$