Online Learning & Game Theory

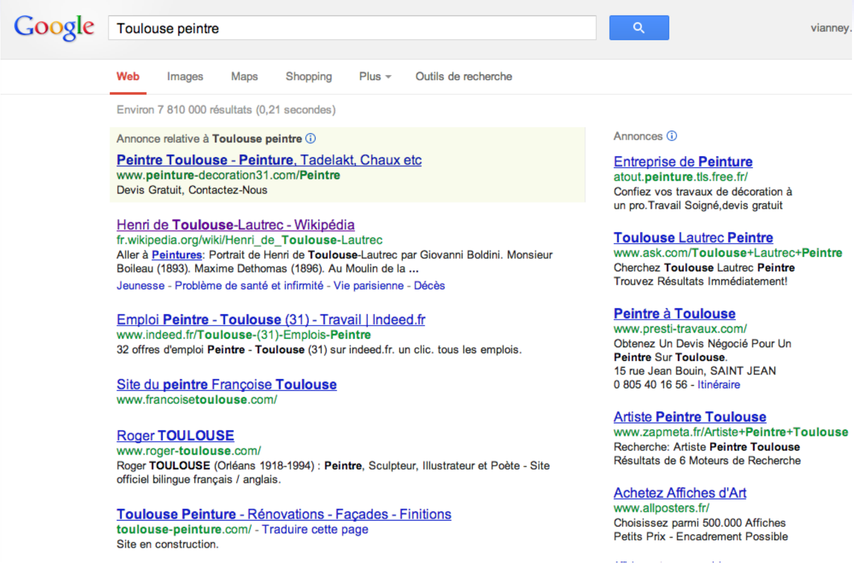
A quick overview with recent results

Vianney Perchet

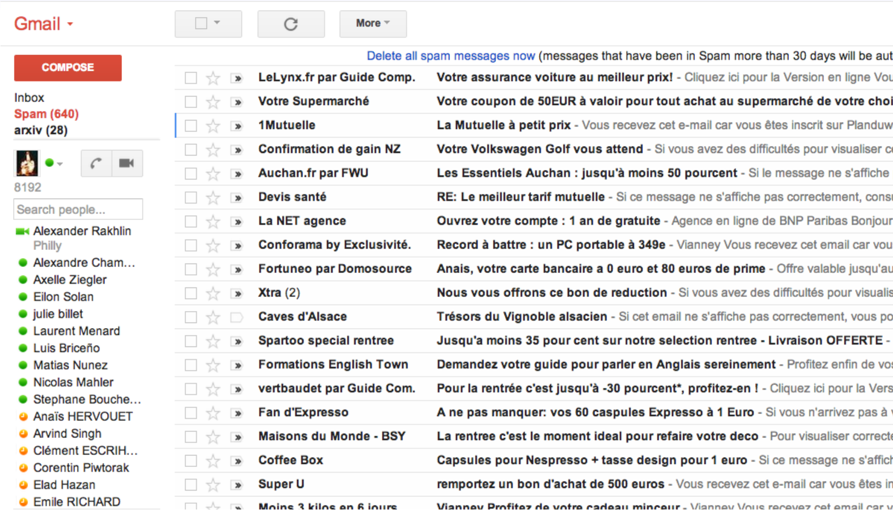
Laboratoire **P**robabilités et **M**odèles **A**léatoires
Univ. Paris-Diderot

Journées MAS 2014
27 Août 2014

Starting Examples



Starting Examples



Outline

1. First, in a stochastic environment (i.i.d. processes)
2. Then, in an adversarial environment (or individual sequences)
3. Finally, some links with game theory

First Part

Stochastic environment

Estimation of Means

$K = 2$ discrete-time proc.: $X_n^{(1)}, X_n^{(2)}$ in $[0,1]$

"The payoff of the ad $1/2$ on query n"

Estimate the means $\mu^{(1)}, \mu^{(2)}$

Hoeffding inequality: exponential decay

$\left|\overline{X}_n^{(k)} - \mu^k\right| > \varepsilon$ with proba at most $2\exp\left(-2n\varepsilon^2\right)$.

Finite number of mistakes:

$$\mathbb{E}\sum_{n\in\mathbb{N}} \mathbb{1}\left\{\left|\overline{X}_n^{(k)} - \mu^k\right| > \varepsilon\right\} \leq \frac{1}{\varepsilon^2}$$

Regret Minimization

– Choose **one** ad to display k_n. Reward: $X_n^{(k_n)}$

Maximize cumulative reward $\sum_{m=1}^{n} X_m^{(k_m)}$ or $\sum_{m=1}^{n} \mu^{(k_m)}$

Minimize Regret [Hannan'56]

$$R_n = n\mu^\star - \sum_{m=1}^{n} \mu^{(k_m)}, \quad \text{with} \quad \mu^\star = \max\{\mu^{(1)}, \mu^{(2)}\}$$

– Equivalent formulation with $\Delta = \mu^\star - \mu^k$:

$$R_n = \Delta \sum_{m=1}^{n} \mathbb{1}\{k_m \neq \star\}$$

Stochastic & Full Monitoring

- Full Monitoring: all values $X_n^{(1)}, X_n^{(2)}$ observed.
- Optimal algorithm: $k_n = \arg\max \overline{X}_n^{(k)}$:

$$\mathbb{E}R_n \leq \frac{1}{\Delta} \qquad \text{and for small } n,\ \mathbb{E}R_N \leq n\Delta$$

Bounded regret, **uniformly** in n!

- **Given** n, worst Δ is $1/\sqrt{n}$ and $\mathbb{E}R_n \leq \sqrt{n}$
- But in the examples, **only** $X_n^{(k_n)}$ is observed (bandit monitoring)!

Stochastic & Bandit Monitoring

- $\overline{X}_n^{(k)} = \frac{1}{n}\sum_{m=1}^{n} X_m^{(k)}$ **not** available, **only** $\widehat{X}_n^{(k)} = \frac{\sum_{m:k_m=k} X_m^{(k)}}{\sharp\{m : k_m = k\}}$
- with $k_n = \arg\max \widehat{X}_n^{(k)}$, $\mathbb{E}R_n = \Theta(n)$.
- **Balance exploitation** (play arg max) and **exploration** (play arg min) to get information

Upper Confidence Bound [Auer,Cesa-Bianchi,Fischer'02]

$$k_n = \arg\max \widehat{X}_n^{(k)} + \sqrt{\frac{2\log(n)}{\sharp\{m : k_m = k\}}}$$

$$\mathbb{E}R_n \leq \square \frac{\log(n)}{\Delta}$$

New policy: Explore Then Commit [P,Rigollet '13]

– Finite horizon $N \in \mathbb{N}$ given.

1) Play alternatively arm 1 and 2 as long as

$$\left|\widehat{X}_n^{(1)} - \widehat{X}_n^{(2)}\right| \leq 2\sqrt{2\frac{\log(4N/n)}{n}}$$

2) Then play for ever the best arm.

→ $\mathbb{E}\mathbb{R}_N \leq \square \frac{\log(N\Delta^2)}{\Delta}$ vs $\frac{1}{\Delta}$ with Full Info

→ Worst case $\Delta \simeq \frac{1}{\sqrt{N}}$

Full Monit & ETC: $\sqrt{N}$ vs UCB: $\sqrt{N}\log(N)$

Bandit vs Full Monitoring

Logarithmic vs bounded regret; same worst case

Bounded Regret ? [Lai,Robbins'84],[Bubeck,P,Rigollet '13]

– Without additional assumption, **No**: lower bound in $\log(n)/\Delta$

– With any given intermediate value $\mu^\sharp \in (\mu^{(1)}, \mu^{(2)})$, **yes**:

- If $\widehat{X}_n^{(1)}$ or $\widehat{X}_n^{(2)}$ above $\mu^\sharp$, then $k_n = \arg\max \widehat{X}_n^{(k)}$
- Otherwise play alternatively both arms.

$\widehat{X}_n^\star < \mu^\sharp$ on $\frac{1}{(\mu^\star - \mu^\sharp)^2}$ stages (same argument for other arm).

– If μ^* and Δ known: $\mathbb{E}R_n \leq \square \frac{1}{\Delta}$ as with Full Monit.

– If only μ^* known: $\mathbb{E}R_n \leq \square \frac{\log(1/\Delta^2)}{\Delta}$

More General Frameworks & Results

Results in worst case ("distribution independent bounds")

– Multi-armed bandit. [Auer,Cesa-Bianchi,Freund,Schapire'02],[Audibert,Bubeck'09]
$K > 2$ arms, $\mathbb{E}R_n \leq \square\sqrt{Kn}$

– Continuous bandit. [Kleinberg'08],[Bubeck,Munos,Stoltz,Szepesvari'11]
Infinite set of arms, $x \in [0,1]^d$ and $\mu(\cdot)$ Lipschitz. $\mathbb{E}R_n \leq \square n^{\frac{d+1}{d+2}}$

– Linear bandit[Dani,Hayes,Kakade'08],[Zinkevich'02],[Abernethy,Hazan,Rakhlin'08]
$x \in [0,1]^d$ and $\mu(\cdot)$ Linear. $\mathbb{E}R_n \leq \square\sqrt{n}$

– Bandit with covariates (cf Google Example) [P,Rigollet'13],[Bull'14]
Covariates $\omega \in [0,1]^d$, $\mathbb{E}[X^{(k)}|\omega] = \mu^{(k)}(\omega)$ 1-Lip. $\mathbb{E}R_n \leq \square n^{\frac{d+1}{d+2}}$

– Higher order bounds/small losses/sparsity[Hazan,Kale'10], [Gershinovitz'13], [Cappé,Garivier,Maillard,Munos,Stoltz'13], [**Gaillard**,Stoltz,van Erven'14]

$$\sqrt{n} \text{ vs } \sqrt{\sum_{m=1}^{n}\left(X_m^{(k_m)} - \mu^{(k_m)}\right)^2}, \sqrt{\sum_{m=1}^{n}\sum_{k=1}^{K} p_n^{(k)}\left(X_n^{(k)}\right)^2}$$

Second Part

Adversarial environment

What we have learned so far:

- In **worst case analysis**
  - Regret minimization in $\square \sqrt{\log(K)n}$ with full monit
  - Up to $\sqrt{K}$, learning **as fast** with bandit monit. than with full monit.
- In **distribution dependent** (not worst case)
  - Bounded regret in $\square \sum \frac{1}{\Delta_k}$
  - Additional assumption required to learn as fast in bandit monit

Adversarial World

- In the examples, data are **not i.i.d.**. Spam senders can even **adapt to spam filters**, that is:

The law of $X_{n+1}^{(k)}$ can depend on $X_1^{(1)}, \ldots, X_n^{(1)}, X_1^{(K)}, \ldots, X_n^{(K)}$ but **even on** the previous choices $k_1, \ldots, k_n$.

The environment can adapt and choose rewards strategically.

- Same def of regret (except argmax changes with time)

$$R_n = \max_k \sum_{m=1}^{n} X_m^{(k)} - \sum_{m=1}^{n} X_m^{(k_m)}$$

- Goal: a policy with sublinear regret $o(n)$ against **ANY** possible strategy of the environment (in particular any sequences $X_n^{(k)}$)

A Popular Algorithm with Full Monitoring

- With $k_n = \text{argmax}\, \overline{X}_n^{(k)}$, $\mathbb{E}R_n = \Theta(n)$. ‘
- With any **deterministic** policy, $\mathbb{E}R_n = \Theta(n)$. ‘

k with proba $\dfrac{\exp\left(\eta \sum_{m=1}^{n} X_m^{(k)}\right)}{\sum_{j=1}^{K} \exp\left(\eta \sum_{m=1}^{n} X_m^{(j)}\right)}$; temperature $\eta \simeq \sqrt{\log(K)n}$

- Regret of “exponential weights” [Auer,Cesa-Bianchi,Freund,Schapire’02]

$$\mathbb{E}R_n \leq \square \sqrt{\log(K)n}, \qquad \forall n \in \mathbb{N}$$

- Same dependency in n as worst case i.i.d., optimal in K.

Optimality and Bandit Monitoring

– **Optimality:** $\mathbb{E}R_n \geq \square\sqrt{\log(K)n}$ if $X_n^{(k)} = \pm 1$ w.p. $1/2$

$$\mathbb{E}\sum_{m=1}^{n} X_m^{(k_m)} = 0 \text{ but } \mathbb{E}\max_k \sum_{m=1}^{n} X_m^{(k)} = \square\sqrt{\log(K)n}$$

– **Bandit Monit.:** $\widetilde{X}_n^{(k)} = X_n^{(k)} \dfrac{\mathbb{1}\{k_n = k\}}{\mathbb{P}_n\{k_n = k\}}$ unbiased estim. of $X_n^{(k)}$

"Exponential weights" w.r.t. $\widetilde{X}_n^{(k)}$: $\mathbb{E}R_n \leq \square\sqrt{K\log(K)n}$

Remark: Optimal bounds are $\square\sqrt{Kn}$

Discrete/Continuous Time

– $\frac{\exp\left(\eta \sum_{m=1}^{n} X_m^{(k)}\right)}{\sum_{j=1}^{K} \exp\left(\eta \sum_{m=1}^{n} X_m^{(j)}\right)} = \nabla\Phi(V_n) := \frac{1}{\eta} \log\left(\sum_{k=1}^{K} \exp(\eta V_n^{(k)})\right)$
with $V_n^{(k)} = \sum_{m=1}^{n} X_n^{(k)} - X_m^{(k_m)}$

Deterministic continuous approx. of stochastic discrete proc.

[Benaïm,Hofbauer,Sorin'06],[Benaïm,**Faure**'13]

– $\mathbb{E}[V_{n+1}] - V_n = \left(X_{n+1}^{(k)} - \langle \nabla\Phi(V_n), X_{n+1}\rangle\right)_{k=1,\ldots,K}$

Stochastic Approx of $\dot{V} \in F(V) := \left\{U - \langle \nabla\Phi(V), U\rangle \vec{\mathbf{1}};\ U \in R^K\right\}$

– Differential inclusion with Lyapounov function $\Phi(V)$:
$\Phi(V)' = \langle \dot{V}, \Phi(V)\rangle = \left\langle U - \langle U, \nabla\Phi(V)\rangle \vec{\mathbf{1}},\ \nabla\Phi(V)\right\rangle = 0$

– $\lim R_n \leq \lim V_n = V(+\infty) = V(0) = \log(d)/\eta$

Refined Regret: Internal-Swap-

- Regret: "As well as the best constant strategy"
- Internal: "On the stages where $k_n = k$, k was the best choice" [Foster,Vohra'99]

$$R_n^{\text{int}} = \max_k \left\{ \max_j \sum_{m:k_m=k} X_m^{(j)} - X_m^{(k)} \right\}$$

- Swap: "As well as $\phi(k)$ instead of k, $\phi : [K] \to [K]$" [Blum,Mansour'07]

$$R_n^{\text{swap}} = \max_{\phi[k]\to[k]} \sum_{m=1}^{n} X_m^{(\phi(k_m))} - X_m^{(k_m)}$$

General regret

- Regret: "As well as the best constant strategy"
- General: "As well as $\xi(k_1, \ldots, k_n)$ instead of k_n, $\xi \in \Xi$" [Lehrer'02]

$$R_n^{\text{gen}} = \max_{\xi \in \Xi} \left\{ \max_j \sum_{m=1}^{n} X_m^{(\xi(k_1,\ldots,k_m))} - X_m^{(k_m)} \right\}$$

- Generalized version of "exponential weights" [P'14]

$$\mathbb{E} R_n^{\text{gen}} \leq \square \sqrt{\log(|\Xi|) n}$$

- Internal regret $\leq \square \sqrt{\log(K) n}$, Swap regret $\leq \square \sqrt{K \log(K) n}$

Third Part

Links with **Game Theory**

What we have learned in the previous section:

- In **worst case analysis**
  - Learning is **as fast** in adversarial than stochastic environment
- In the adversarial framework
  - Refined notions of regret can be minimized

Against Opponents - Game Theory

$X_n^{(k)}$ **not** arbitrary, but induced by choices of another player

- **TWO players**, simultaneous actions in $\{1, .., K\}$ and $\{1, .., L\}$
- Payoffs are defined by **two matrices** $A \in \mathbb{R}^{K \times L}$ and $B \in \mathbb{R}^{K \times L}$.
  - Player 1 picks row $k \in \{1, .., K\}$ and Player 2 column $\ell \in \{1, .., L\}$
  - Player 1 gets $A_{k,\ell}$ and Player 2 gets $B_{k,\ell}$
- Choices can be **random** $p \in \Delta([K])$ and $q \in \Delta([L])$
  - Player 1 gets $\sum_{k,\ell} p_k q_\ell A_{k,\ell} = p^T A q$; P2 gets $p^T B q$
- Online **learning**: $X_n^{(k)} = A_{k,\ell_n}$ and $Y_n^{(\ell)} = B_{k_n,\ell}$.

Assume both players minimize regret independently.

Do they "learn a solution concept" from game theory ?

Nash Equilibria

"A Nash equilibria is a situation where no player has interest to change his action" [Nash'50], [Nash'51]

- A Nash equilibria is a pair $(p^*, q^*) \in \Delta([K]) \times \Delta([L])$ such that
  - Player 1 has no interest to change given q^*:

$$(p^*)^T A q^* \geq p^T A q^*, \quad \forall p \in \Delta([K])$$

  - Player 2 has no interest to change given p^*:

$$(p^*)^T A q^* \geq (p^*)^T A q, \quad \forall q \in \Delta([L])$$

- There **always exist** Nash equilibria; generically an odd number

  [Nash'50], [Nash'51], [Shapley'74]

Are Nash Equilibria Learnable?

– Both players minimize their regret independently.

$$k_n \sim p_n \in \Delta([K]), \;\; \ell_n \sim q_n \in \Delta([L])$$

Learning Nash equilibria could mean:

- $(p_n, q_n) \in \Delta([K]) \times \Delta([L])$ cv to a NE, or to set of NE.
- $\left(\frac{1}{n}\sum_{m=1}^{n} \delta_{k_m}, \frac{1}{n}\sum_{m=1}^{n} \delta_{\ell_m}\right) \in \Delta([K]) \times \Delta([L])$ cv to a NE, or to set of NE
- $\left(\frac{1}{n}\sum_{m=1}^{n} \delta_{k_m,\ell_m}\right) \in \Delta([K] \times [L])$ cv to a NE, or to set of NE

– Nash equilibria **are not learnable** (independently): [Hart,Mas-Colell'04]
There always exists a game s.t. none of the convergence occur

– What is Learnable?
correlated eq, Minmax-Value, Potential eq [Coucheney, **Gaujal**, Mertikopolous]

Correlated Equilibria

"Players use an external device to correlate (as traffic lights); when they are told to take an action (as stop or go), it is optimal"

– A correlated equilibrium is a distribution $\pi \in \Delta([K] \times [L])$. $(k^*, \ell^*) \sim \pi$; P1 is told **secretly** to play k^*, P2 to play ℓ^*
  - if P1 plays $k^* \in [K]$, he gets $\sum_{\ell \in [L]} \pi_{k^*,\ell} A_{k^*,\ell}$. If he plays $j \in [K]$ instead, he would get $\sum_{\ell \in [L]} \pi_{k^*,\ell} A_{j,\ell}$

– $$\sum_{\ell \in [L]} \pi_{k^*,\ell} A_{k^*,\ell} \geq \sum_{\ell \in [L]} \pi_{k^*,\ell} A_{j,\ell}, \quad \text{for all } k^*, j \in [K]$$
  - Similar to no internal regret !

If both players minimize **internal regret**, empirical distribution of actions converge to the **set of correlated equilibria**. [Foster,Vohra'99]

Minmax Theory

In zero-sum games, players have **optimal** strategies

- "zero-sum": $B=-A$; P1 maximizes and P2 minimizes $p^T Aq$
- Value= $\max_{p\in\Delta([K])}\min_{q\in\Delta([L])} p^T Aq = \min_{q\in\Delta([L])}\max_{p\in\Delta([K])} p^T Aq$
- p^* optimal if $(p^*)^T Aq \geq$ Value for all $q \in \Delta([L])$.

- $R_n \leq 0 \implies \frac{1}{n}\sum_{m=1}^n X_m^{(k_m)} \geq$ Value
- $\left(\frac{1}{n}\sum_{m=1}^n \delta_{k_m}, \frac{1}{n}\sum_{m=1}^n \delta_{\ell_m}\right)$ cv to optimal strat, i.e. to NE

- NE are **fast learnable** in zero-sum game, at $O\left(\frac{1}{n}\right)$ [Harris'98]

Conclusion

- In **worst case analysis**
  - With full monitoring, learning is **as fast** in adversarial than stochastic environment
  - Up to $\sqrt{K}$, learning is **as fast** with bandit monit. than with full monit.
- In **distribution dependent** (not worst case)
  - Additional assumption required to learn as fast in bandit than in full monitoring
- In **game theoretic framework**
  - Nash equilibria are not learnable in general
  - Correlated equilibria are learnable (by minimizing internal regret)
  - In zero-sum and potential games, equilibria are learnable.

Fundamental textbook: [Cesa-Bianchi,Lugosi'06]