TP3: Adversarial bandits

emilie.kaufmann@inria.fr

November 17th, 2015

In a two players zero-sum game, players A and B choose simultaneously two actions $a \in \{1...N_A\}$ and $b \in \{1..N_B\}$. The reward of player A associated to the chosen pair of action (a, b) is denoted by $R_A(a, b)$. Then one has $R_B(a, b) = -R_A(a, b)$. Such a game is represented by a matrix of gains

$$G = (R_A(i, j))_{\substack{1 \leq i \leq N_A \\ 1 \leq j \leq N_B}}$$

The game described by the matrix below can model for example two persons choosing between two possible activities (say go to the movies or to a bar). The first person (player A) prefers consensus (with a preference for activity 1) whereas the other (player B) prefers being on his own (with a slight preference for activity 2).

G =

$A\backslash B$	1	2
1	2	-1
2	0	1

Consider the following game: at each time t, player A chooses an action x_A (based on the past observed rewards), player B also plays (simultaneously, without revealing his action x_B to A), player A receives the reward $R_A(x_A, x_B)$, and player B the reward $R_B(x_A, x_B) = -R_A(x_A, x_B)$. Each of the player is playing a bandit game against an adversary.

Remark: With this particular choice of matrix G, each player is actually playing a full-information game, since when he observes the value of the reward he can identify the action of the other player and thus know what reward he would have obtained, had he chosen the other action. Thus we will implement two strategies: EWF, that uses the full information, and EXP3, that uses only the observed reward (bandit information).

EWF and EXP3 versus an oblivious adversary

We assume that B chooses in advance a sequence of actions of length n, Seq. The function

```
[Actions,Rewards]=EWFplay(n,G,eta,Seq)
```

is given. Write a function

```
[Actions,Rewards]=EXP3play(n,G,eta,beta,Seq);
```

that returns the sequence of actions chosen by A and the sequence of rewards received (both of length n) when A plays with the EXP3 strategy. (*simu.m* that draws a sample from a distribution on $\{1, \ldots, K\}$).

Question 1: For a Sequence of actions of B that you specify, present regret curves for EXP3 and EWF on the same graph. Specify the parameters chosen for each algorithm.

EXP3 versus EXP3 : Nash Equilibrium

1. Now we assume that both players are using an EXP3 strategy to select their action. Write a function

   ```
   [ActionsA,ActionsB,Rew]= EXP3vEXP3(n,eta,beta,G)
   ```

   that returns the sequences of actions chosen by each of the two players, and the rewards received by player A.

2. Illustrate that the quantities $p_{a,n} = \frac{1}{n}\sum_{t=1}^{n} \mathit{1}_{(A_t=1)}$ and $p_{b,n}$ (similarly defined) almost surely converge to some values p_a^* and p_b^*.

3. A pair (p, q) is a Nash equilibrium if when player A uses the mixed strategy defined by p (that is he plays action A with probability p, action B otherwise) and player B uses the mixed strategy defined by q, none of the players has incentive to change its strategy if the other does not. Check that (p_a^*, p_B^*) is a Nash equilibrium.

4. Show that the sequence $\frac{1}{n}\sum_{t=1}^{n} R_A(A_t, B_t)$ converges towards the value of the game.

Question 2: For a specified value of η and β, illustrate the convergence towards the Nash equilibrium and the value of the game.

Stochastic bandit or adversarial bandit?

The stochastic MAB can be viewed as an adversarial bandit problem, with an adversary drawing rewards at random from the arms' distributions. The goal of this part is to compare the EXP3 strategy on this particular adversarial problem with Thompson Sampling, one of the best strategy in the stochastic setting.

The function

```
[actions,rewards]= Thompson(n,MAB)
```

is given. Write a function

```
[actions,rewards]= EXP3Stochastic(n,eta,beta,MAB).
```

Question 3: Compare the regret of the two algorithms on easy/difficult Bernoulli bandit problems. On which type of problems does EXP3 seem interesting?