

# Introduction to adversarial bandit problems

Gilles Stoltz

CNRS – École normale supérieure – HEC Paris



- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

## A base one-shot game is repeated

- A decision-maker (the row player) takes actions  $I_1, I_2, \dots$  from a **finite** set  $\mathcal{X} = \{1, \dots, N\}$ .
- The opponent player (the column player) selects the outcomes  $y_1, y_2, \dots \in \mathcal{Y}$ . (The **outcome space**  $\mathcal{Y}$  is arbitrary.)
- The loss function is  $\ell : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ .

That is, at each round  $t = 1, 2, \dots$ , the opponent player chooses the **loss vector**

$$(\ell(1, y_t), \dots, \ell(N, y_t)) = \ell(\cdot, y_t)$$

and the decision-maker chooses (simultaneously) a **component**.

In the simplest setting (**full information**), both players observe and recall the action-outcome pairs  $(I_t, y_t)$ .



## Strategies for the players

For the decision-maker:

A (randomized) strategy  $\sigma$  for the decision-maker is a **sequence of functions**. The  $t$ -th of them, associates

- to the past losses  $\ell(j, y_s)$ ,  $j = 1, \dots, N$  and  $s = 1, \dots, t - 1$ ,
- a probability distribution  $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$  on the set  $\mathcal{X} = \{1, \dots, N\}$  of actions.

The played action  $I_t$  is chosen by drawing  $I_t$  according to  $\mathbf{p}_t$ .

The decision-maker aims at minimizing his cumulative loss.



## Strategies for the players

For the opponent player:

We perform a **worst-case** analysis of the decision-maker's strategy  $\sigma$  and make **no** (behavioral, stochastic) **assumption** on the opponent player's strategy  $\tau$ .

We present below strategies for the decision-maker which minimize the regret for all possible strategies of the opponent player, in an **almost sure** way.

The name "**individual sequences**" comes from this and from the fact that we fix the sequence of outcomes when we assess the quality of the decision-maker's strategy.



- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers



## Definition of the (Hannan) regret

To assess the performance of a strategy,

- we fix the realized sequence of outcomes  $y_1, y_2, \dots$
- and compare the sequence  $I_1, I_2, \dots$  of actions chosen by the decision-maker to constant sequences of pure actions  $j, j, \dots$

That is, we compare  $\widehat{L}_n = \sum_{t=1}^n \ell(I_t, y_t)$  to the  $L_{j,n} = \sum_{t=1}^n \ell(j, y_t)$ .

### Definition

The **Hannan regret**  $R_n$  is defined as the maximal difference of these cumulative payoffs,

$$\max_{j=1, \dots, N} R_{j,n} = \widehat{L}_n - \min_{j=1, \dots, N} L_{j,n} = \sum_{t=1}^n \ell(I_t, y_t) - \min_{j=1, \dots, N} \sum_{t=1}^n \ell(j, y_t)$$

## Hannan consistent strategies

The **Hannan regret** is defined as

$$\max_{j=1,\dots,N} R_{j,n} = \widehat{L}_n - \min_{j=1,\dots,N} L_{j,n} = \sum_{t=1}^n \ell(I_t, y_t) - \min_{j=1,\dots,N} \sum_{t=1}^n \ell(j, y_t)$$

### Definition

A strategy  $\sigma$  for the decision-maker is said **(Hannan) consistent** whenever

$$\limsup_{n \rightarrow \infty} \frac{R_n}{n} \leq 0 \quad \text{a.s.}$$

regardless of the strategy  $\tau$  of the opponent player.

There **exist** Hannan-consistent strategies, even many!

The earliest ones are due to **Blackwell '56** and **Hannan '57**.



- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

# A repeated zero-sum game with full information

**Parameters** (known to both players): action set  $\mathcal{X} = \{1, \dots, N\}$ , outcome set  $\mathcal{Y}$ , loss function  $\ell$

**For** each round  $t = 1, 2, \dots$ ,

- 1 the opponent player chooses the next outcome  $y_t \in \mathcal{Y}$  without revealing it;
- 2 the decision-maker chooses a probability distribution  $\mathbf{p}_t$  and draws an action  $I_t \in \{1, \dots, N\}$  according to this distribution;
- 3 the decision-maker suffers a loss  $\ell(I_t, y_t)$  and each action  $i$  suffers a loss  $\ell(i, y_t)$ ;
- 4  $y_t$ ,  $\mathbf{p}_t$ , and  $I_t$  are revealed to both players, who then know all these losses.

**Goal:** Hannan regret is to be made small,  $\widehat{L}_n - \min_{j=1, \dots, N} L_{j,n}$  a.s.



- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

## Exponential reweighting

The idea is to assign a higher probability to better-performing actions (this, in some sense, **smoothes fictitious play**).

### Exponentially weighted average predictor

$\mathbf{p}_1$  is uniform and for  $t \geq 2$ ,

$$p_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell(i, y_s)\right)}{\sum_{j=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \ell(j, y_s)\right)} = \frac{\exp(-\eta L_{i,t-1})}{\sum_{j=1}^N \exp(-\eta L_{j,t-1})}$$

where  $\eta > 0$  is a parameter to be tuned.

This strategy was introduced by Vovk '90, Littlestone and Warmuth '94. (See also Fudenberg and Levine '95, Cesa-Bianchi, Freund, Helmbold, Haussler, Schapire, and Warmuth '97, Cesa-Bianchi and Lugosi '99.)



# Exponential reweighting

## Lemma

For *all strategies*  $\tau$  of the opponent player,

$$\begin{aligned}\bar{R}_n &= \sum_{t=1}^n \sum_{i=1}^N p_{i,t} \ell(i, y_t) - \min_{j=1, \dots, N} \sum_{t=1}^n \ell(j, y_t) \\ &\leq \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{i=1}^N p_{i,t} \ell(i, y_t)^2\end{aligned}$$

## Theorem

If  $\ell$  takes values in  $[0, 1]$ , for *all strategies*  $\tau$  of the opponent player and the choice  $\eta = \sqrt{(2 \ln N)/n}$ ,

$$\bar{R}_n \leq \sqrt{2n \ln N}$$

## Consequences and comments

Via an application of the **Hoeffding-Azuma** inequality, with probability at least  $1 - \delta$ ,

$$R_n = \sum_{t=1}^n \ell(I_t, y_t) - \min_{j=1, \dots, N} \sum_{t=1}^n \ell(j, y_t) \leq \sqrt{2n \ln N} + \sqrt{\frac{n}{2} \ln \frac{1}{\delta}}$$

**Borel-Cantelli** lemma then ensures

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\sqrt{n \ln n}} \leq 1 \quad \text{a.s.}$$

The choice of  $\eta$  can be made **adaptive** in  $n$  and in the range of the payoffs by sequentially choosing a sequence  $(\eta_t)$ .

## Proof of the lemma

Recall that  $p_{i,t} = w_{i,t-1}/W_{t-1}$ , where  $W_{t-1} = w_{1,t-1} + \dots + w_{N,t-1}$ ,  $w_{i,0} = 1$ , and for  $t \geq 2$ ,

$$w_{i,t-1} = \exp(-\eta L_{i,t-1}) = \exp\left(-\eta \sum_{s=1}^{t-1} \ell(i, y_s)\right)$$

On the one hand,

$$\ln \frac{W_n}{W_0} \geq \ln \frac{\max_{j=1,\dots,N} w_{j,n}}{N} = -\eta \min_{j=1,\dots,N} L_{j,n} - \ln N$$

On the other hand, using  $e^{-x} \leq 1 - x + x/2$  for  $x \geq 0$  and  $\ln(1+u) \leq u$ , we have, for  $t = 1, \dots, n$ ,

$$\begin{aligned} \ln \frac{W_t}{W_{t-1}} &= \ln \frac{\sum_{i=1}^N e^{-\eta \ell(i, y_t)} w_{i,t-1}}{\sum_{j=1}^N w_{j,t-1}} = \ln \sum_{i=1}^N p_{i,t} e^{-\eta \ell(i, y_t)} \\ &\leq \ln \sum_{i=1}^N p_{i,t} \left(1 - \eta \ell(i, y_t) + \frac{\eta^2}{2} \ell(i, y_t)^2\right) \leq -\eta \sum_{i=1}^N p_{i,t} \ell(i, y_t) + \frac{\eta^2}{2} \sum_{i=1}^N p_{i,t} \ell(i, y_t)^2 \end{aligned}$$

Summing the upper bounds over  $t = 1, \dots, n$  and combining with the lower bound,

$$\sum_{t=1}^n \sum_{i=1}^N p_{i,t} \ell(i, y_t) - \min_{j=1,\dots,N} \sum_{t=1}^n \ell(j, y_t) \leq \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{i=1}^N p_{i,t} \ell(i, y_t)^2$$



- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

## A repeated zero-sum game with partial information

**Parameters** (known to both players): action set  $\mathcal{X} = \{1, \dots, N\}$ , outcome set  $\mathcal{Y}$ , loss function  $\ell$

**For** each round  $t = 1, 2, \dots$ ,

- 1 the opponent player chooses the next outcome  $y_t \in \mathcal{Y}$  without revealing it;
- 2 the decision-maker chooses a probability distribution  $\mathbf{p}_t$  and draws an action  $I_t \in \{1, \dots, N\}$  according to this distribution;
- 3 the decision-maker suffers a loss  $\ell(I_t, y_t)$  and each action  $i$  suffers a loss  $\ell(i, y_t)$ ;
- 4 the decision-maker **only** gets to see **his own loss**  $\ell(I_t, y_t)$  whereas  $\mathbf{p}_t$  and  $I_t$  are revealed to the opponent.

**Goal:** Hannan regret is still to be made small,  $\widehat{L}_n - \min_{j=1, \dots, N} L_{j,n}$  a.s.



## Competing with respect to $K$ strategies

As in the case of full information, the statement of the problem and the proposed forecasters **only** depend on the **values**  $\ell_{j,t} = \ell(j, y_t)$  of the losses, which can be arbitrary.

In particular, they do not depend on the underlying structure of the problem (the loss function  $\ell$  and the outcome space  $\mathcal{Y}$ ).

This therefore models the cases when the decision-maker competes with a **finite number of** base forecasting or gambling **strategies**.



- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - **Key idea: estimate the unobserved losses**
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

## Estimates of the unobserved losses

The key idea is to **estimate** the unobserved losses and to form exponentially weighted averages on these estimates.

The estimates are

$$\tilde{\ell}_{j,t} = \frac{\ell(I_t, y_t)}{p_{I_t,t}} \mathbb{I}_{\{I_t=j\}}$$

We denote by  $\mathbb{E}_t$  the **conditional expectation** at round  $t$  with respect to the information available to the decision-maker and the opponent player at the beginning of round  $t$ .

(This fixes the values of  $\mathbf{p}_t$  and  $y_t$ , only the choice of  $I_t$  according to  $\mathbf{p}_t$  involves randomness.)

## Estimates of the unobserved losses

The key idea is to **estimate** the unobserved losses and to form exponentially weighted averages on these estimates.

The estimates are

$$\tilde{\ell}_{j,t} = \frac{\ell(I_t, y_t) \mathbb{I}_{\{I_t=j\}}}{p_{I_t,t}}$$

We denote by  $\mathbb{E}_t$  the **conditional expectation** at round  $t$ .

The estimates above are (conditionally) **unbiased**: since  $I_t$  is distributed as  $\mathbf{p}_t$ ,

$$\mathbb{E}_t \left[ \tilde{\ell}_{j,t} \right] = \frac{\ell(j, y_t)}{p_{j,t}} \mathbb{E}_t \left[ \mathbb{I}_{\{I_t=j\}} \right] = \frac{\ell(j, y_t)}{p_{j,t}} p_{j,t} = \ell(j, y_t)$$

We now perform exponentially weighted averages on these unbiased estimates.

See Auer, Cesa-Bianchi, Freund, and Schapire '02.



## Exponential reweighting on estimated losses

### Exponentially weighted average predictor (version 1)

With a parameter  $\eta > 0$  to be tuned:  $\mathbf{p}_1$  is uniform and for  $t \geq 2$ ,

$$p_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{i,s}\right)}{\sum_{j=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{j,s}\right)}$$

### Theorem

If  $\ell$  takes values in  $[0, 1]$ , for **all strategies**  $\tau$  of the opponent player and the same choice  $\eta = \sqrt{(2 \ln N)/n}$ , the expectation of the regret is bounded as

$$\mathbb{E}\left[\widehat{L}_n\right] - \min_{j=1,\dots,N} \mathbb{E}\left[L_{j,n}\right] \leq \sqrt{2nN \ln N}$$

# Proof

We recall that

$$\tilde{\ell}_{j,t} = \frac{\ell(j, y_t)}{p_{j,t}} \mathbb{I}_{\{I_t=j\}}$$

Using the lemma, we first have

$$\sum_{t=1}^n \sum_{i=1}^N p_{i,t} \tilde{\ell}_{i,t} - \min_{j=1, \dots, N} \sum_{t=1}^n \tilde{\ell}_{j,t} \leq \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{i=1}^N p_{i,t} \tilde{\ell}_{i,t}^2$$

Taking the expectations of both sides and noting that

$$\sum_{i=1}^N p_{i,t} \tilde{\ell}_{i,t} = \ell(I_t, y_t)$$

and

$$\mathbb{E}_t \left[ \sum_{i=1}^N p_{i,t} \tilde{\ell}_{i,t}^2 \right] = p_{i,t} \left( \frac{\ell(i, y_t)}{p_{i,t}} \right)^2 \mathbb{E}_t \left[ \mathbb{I}_{\{I_t=i\}} \right] = \ell(i, y_t)^2$$

yields

$$\mathbb{E} \left[ \hat{L}_n \right] - \min_{j=1, \dots, N} \mathbb{E} \left[ L_{j,n} \right] \leq \frac{\ln N}{\eta} + \frac{\eta}{2} Nn$$



- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - **Exploration – exploitation trade-off**
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

## Exploration – exploitation

To get **high-probability** bounds, one needs to ensure, e.g., that all arms are pulled sufficiently and fix a common lower bound on the probabilities they are played.

### Exponentially weighted average predictor (version 2: Exp3)

With parameters  $\eta, \gamma > 0$  to be tuned:  $\mathbf{p}_1$  is uniform and for  $t \geq 2$ ,

$$p_{i,t} = (1 - \gamma) \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{i,s}\right)}{\sum_{j=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{j,s}\right)} + \frac{\gamma}{N}$$

This forecaster ensures that for properly chosen  $\eta$  and  $\gamma$  and with probability at least  $1 - \delta$ ,

$$R_n \leq \square n^{2/3} \ln \frac{1}{\delta}$$



## Sketch of proof

We start as before.

By taking into account the mixing with the uniform distribution, one gets an additional  $\gamma n$  term in the upper bound.

In addition, various quantities need to be dealt with concentration techniques, e.g., by Bernstein's inequality for martingales,

$$\sum_{t=1}^n \tilde{\ell}_{j,t} \leq \sum_{t=1}^n \ell_{j,t} + \sqrt{\sum_{t=1}^n \text{Var}_t \tilde{\ell}_{j,t} \ln \frac{1}{\delta}} + \dots$$

where the conditional variances satisfy

$$\text{Var}_t \tilde{\ell}_{j,t} \leq \mathbb{E}_t \left[ \tilde{\ell}_{j,t}^2 \right] \leq \frac{1}{\gamma/N}$$

In total, a term  $\gamma n$  has to be balanced with a  $\sqrt{n/\gamma}$  term, via the choice  $\gamma \sim n^{-1/3}$ .



## Exponential reweighting on shifted estimated payoffs

We now use the estimated **payoffs**  $\tilde{r}_{j,t} = \frac{1 - \ell(I_t, y_t)}{p_{j,t}} \mathbb{I}_{\{I_t=j\}}$

### Exponentially weighted average predictor (version 3: Exp3.P)

With parameters  $\eta, \gamma, \beta > 0$  to be tuned:  $\mathbf{p}_1$  is uniform and for  $t \geq 2$ ,

$$p_{i,t} = (1 - \gamma) \frac{\exp\left(\eta \sum_{s=1}^{t-1} \tilde{r}_{i,s} + \frac{\beta}{p_{i,s}}\right)}{\sum_{j=1}^N \exp\left(\eta \sum_{s=1}^{t-1} \tilde{r}_{j,s} + \frac{\beta}{p_{j,s}}\right)} + \frac{\gamma}{N}$$

For properly chosen parameters  $\eta, \gamma, \beta$ , and with probability at least  $1 - \delta$ , the regret of this forecaster is

$$R_n = \hat{L}_n - \min_{j=1, \dots, N} L_{j,n} \leq 6 \sqrt{nN \ln \frac{N}{\delta}} + \frac{\ln N}{2}$$

(See, e.g., Cesa-Bianchi and Lugosi '06.)

- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

## Another way of lower bounding the probabilities

### Green shift

With parameters  $\eta, \gamma, \beta > 0$  to be tuned:  $\mathbf{p}_1$  is uniform and for  $t \geq 2$ , an intermediate distribution  $\mathbf{q}_t$  is first defined as

$$q_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\tilde{\ell}_{i,s} - \frac{\beta}{\max\{p_{i,s}, \gamma\}}\right)\right)}{\sum_{j=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \left(\tilde{\ell}_{j,s} - \frac{\beta}{\max\{p_{j,s}, \gamma\}}\right)\right)}$$

where

$$\tilde{\ell}_{j,t} = \frac{\ell(I_t, y_t)}{p_{I_t,t}} \mathbb{I}_{\{I_t=j\}}$$

after which  $\mathbf{p}_t$  is defined as  $p_{i,t} = c_t q_{i,t} \mathbb{I}_{\{q_{i,t} \geq \gamma\}}$  for some normalization constant  $c_t$ .

The components of the  $\mathbf{p}_t$  are **either** equal to 0 or larger than  $\gamma$ .

Auer and Ottucsák '06 prove a similar  $\sqrt{nN \ln \frac{N}{\delta}}$  bound.



- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

## Minimax order of magnitude of the regret

Auer, Cesa-Bianchi, Freund, and Schapire '02 also show a **lower bound** for the expected regret.

### Theorem

For  $\mathcal{Y} = [0, 1]$ , there exists a loss function  $\ell : \mathbb{N} \times \mathcal{Y} \rightarrow \{0, 1\}$  such that for all  $N \geq 2$  and  $n \geq 1$ , and all strategies suited to bandit settings,

$$\sup_{y_1, \dots, y_n} \left\{ \widehat{L}_n - \min_{j=1, \dots, N} L_{j,n} \right\} \geq \frac{1}{20} \min \left\{ \sqrt{nN}, n \right\}$$

The proof relies on **Pinsker's** inequality.

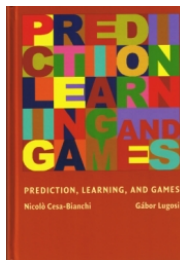
The opponent does not even need to react to our choices, he can fix a hard to predict sequence in advance.

Open question: We suspect that the minimax order is  $\sqrt{nN \ln N}$ , but this has been **open** for more 10 years now!



## More details...

For more details about what we discussed so far, see



**Prediction, Learning, and Games** by Nicolò Cesa-Bianchi and  
Gábor Lugosi

- 1 Adversarial opponent in full information
  - A repeated game
  - The regret
  - Summary of what we have seen so far
  - An efficient strategy
- 2 Adversarial opponent in a bandit problem
  - Description
  - Key idea: estimate the unobserved losses
  - Exploration – exploitation trade-off
  - Another closely related forecaster
- 3 Recent stuff and open questions
  - An open question
  - Three recent papers

## In the last three COLT editions

I checked out the programme of COLT'06, COLT'07, and COLT'08 and could only spot the following three papers in the setting of adversarial bandits.

- **The Shortest Path Problem Under Partial Monitoring** by Andras György, Tamas Linder, Gábor Lugosi, and György Ottucsák
- **High-Probability Regret Bounds for Bandit Online Linear Optimization** by Peter Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari
- **Competing in the Dark: An Efficient Algorithm for Bandit Linear Optimization** by Jacob Abernethy, Elad Hazan, and Alexander Rakhlin



## Shortest path

Here, a directed graph indicates the possible paths from  $A$  to  $B$ ; paths share some edges and one only observes the time needed on the edges composing the chosen path.

The regret can be made less than something of the order of  $\sqrt{nN}$  by the previous techniques, but since the number  $N$  of paths can be **exponential** in the number of edges

- a direct implementation is impossible,
- the bound can be bad since it involves a  $\sqrt{N}$  factor.

Gyorgy and al. deal with both issues and show an efficient (**polynomial complexity**) forecaster, whose regret bound scales essentially with  $\sqrt{n E \ln E}$ , where  $E$  is the number  $E$  of edges.



## Online linear optimization

A generalized framework considered by the two other papers involves a **convex decision set**  $\mathcal{D} \subseteq \mathbb{R}^d$  and a set  $\mathcal{L} \subset \mathbb{R}^d$  of possible loss vectors.

A **boundedness** assumption is needed: there exists  $B$  such that for all  $x \in \mathcal{D}$  and  $L \in \mathcal{L}$ , the inner products  $L \cdot x \in [0, B]$ .

The decision-maker, respectively, his opponent, sequentially chooses the decisions  $x_t \in \mathcal{D}$ , respectively, the loss vectors  $L_t \in \mathcal{L}$ .

The **regret** is then defined as

$$R_n = \sum_{t=1}^n L_t \cdot x_t - \inf_{x \in \mathcal{D}} \sum_{t=1}^n L_t \cdot x$$



## Online linear optimization: Rates

Bartlett and al. deal with the case of **high-probability** bounds in the **bandit** case.

With the help of some previous papers, one gets the following picture (up to logarithmic factors)

	Full	Bandit
UB	$\sqrt{dn}$	$d^{3/2}\sqrt{n}$
LB	$\sqrt{dn}$	$d\sqrt{n}$

Expected and high-probability Upper bounds (UB) are obtained by discretizing  $\mathcal{D}$  and applying an Exp3-type strategy.

Lower bounds (LB) follow from stochastic methods (central limit theorems).

The proposed general algorithms for the upper bounds are **not efficient** in general.



## Linear optimization: One general efficient implementation

In case of **expected** regret in a bandit setting, Abernethy et al. exhibit a general **efficient** forecaster that ensures a  $d\sqrt{n}$  bound.

Previous general efficient implementations only had a  $n^{2/3}$  rate.

The setting of online shortest path can be cast as an online linear optimization problem.

Then, one observes **only** the **total time** needed for the chosen path, and not the sequences of times taken on the edges of the path. A  $\sqrt{n}$  regret can then be ensured (in expectation).

This is to be compared to the results of Gyorgy and al. discussed earlier:  $\sqrt{n}$  if the times on the chosen edges are observed,  $n^{2/3}$  if it is the case only for the total time (both bounds in high probability).

