

# Strategies for Minimizing Regret under Imperfect Monitoring

Gilles Stoltz

CNRS – École normale supérieure – HEC Paris



This is a joint work with **Gábor Lugosi** (ICREA and Universitat Pompeu Fabra, Barcelona) and **Shie Mannor** (McGill University, Montreal), published by *Mathematics of Operations Research*.

The story: The **key** step was made **in Montreal** in April 2006, when Gábor was on sabbatical at McGill and I visited Shie and him to solve in a constructive way the most general form of a prediction problem under imperfect monitoring.

Shie (with Nahum Shimkin) on one side, and Gábor and I (with Nicolò Cesa-Bianchi) on the other side had solved some special cases independently before that.



A base finite game is repeated.

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

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

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

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

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



The strategies for the players are as follows.

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 payoffs  $r(j, y_s)$ ,  $j = 1, \dots, N$  and  $s = 1, \dots, t - 1$ ,
- a probability distribution  $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$  over 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 maximizing his cumulative payoff.



The strategies for the players are as follows.

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 a quantity called **regret** for **all possible** strategies of the opponent, in a 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.



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  $\hat{X}_n = \sum_{t=1}^n r(I_t, y_t)$  to the  $X_{j,n} = \sum_{t=1}^n r(j, y_t)$ .

### Definition

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

$$\max_{j=1, \dots, N} R_{j,n} = \max_{j=1, \dots, N} X_{j,n} - \hat{X}_n = \max_{j=1, \dots, N} \sum_{t=1}^n r(j, y_t) - \sum_{t=1}^n r(I_t, y_t)$$

The **Hannan regret** is defined as

$$R_n = \max_{j=1,\dots,N} R_{j,n} = \max_{j=1,\dots,N} X_{j,n} - \widehat{X}_n = \max_{j=1,\dots,N} \sum_{t=1,\dots,n} r(j, y_t) - \sum_{t=1,\dots,n} r(l_t, y_t)$$

### Definition

A strategy  $\sigma$  for the decision-maker is said **Hannan-consistent** (or universally 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!

See, for the earliest ones, those of **Blackwell '56** and **Hannan '57**.

I now recall one simple such Hannan-consistent strategy.



The natural idea is to assign a **higher probability** to **better-performing** actions.

### Exponentially weighted average predictor

$\mathbf{p}_1$  is uniform and for  $t \geq 2$ ,  $\mathbf{p}_t$  is defined by its components,

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

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

This strategy was introduced by Vovk '90, Littlestone and Warmuth '94.



## Theorem

For *all strategies*  $\tau$  of the opponent player, the expected regret of this strategy is bounded as

$$\max_{j=1,\dots,N} \sum_{t=1,\dots,n} r(j, y_t) - \sum_{t=1,\dots,n} r(\mathbf{p}_t, y_t) \leq \frac{\ln N}{\eta} + \frac{\eta n}{8} = \sqrt{\frac{n}{2} \ln N}$$

with  $\eta = \sqrt{8 \ln N / n}$ .

(The proof is short and relies on Hoeffding's lemma.)

Thus, by Hoeffding–Azuma inequality, with probability  $1 - \delta$ , the **true regret**  $R_n \leq \square \sqrt{n \ln(N/\delta)}$ .

The Borel–Cantelli lemma then ensures that with probability 1, one has  $R_n = o(n)$ .



In the general case, the feedback received after choosing the action  $I_t$  may contain much **less information** than the value of the outcome  $y_t$ .

The formulation of this general problem is as follows.

A base **finite game** is to be repeated.

It is parameterized by

- the strategy sets  $\mathcal{X} = \{1, \dots, N\}$  and  $\mathcal{Y} = \{1, \dots, M\}$  for the decision-maker and the opponent player,
- a payoff function  $r : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$  for the decision-maker,
- a (finite) set  $\mathcal{S}$  of possible **signals**,  $\Delta(\mathcal{S})$  denoting the set of probability distributions on  $\mathcal{S}$ ,
- a **feedback** function  $H : \mathcal{X} \times \mathcal{Y} \rightarrow \Delta(\mathcal{S})$ .



The repeated game is played as follows.

*Parameters* (known to both players): number  $N$  of actions, number  $M$  of outcomes, payoff function  $r$ , random feedback function  $H$

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

- the opponent player chooses the next outcome  $y_t \in \{1, \dots, M\}$  without revealing it;
- the decision-maker chooses a probability distribution  $\mathbf{p}_t$  and draws an action  $I_t \in \{1, \dots, N\}$  according to this distribution;
- the forecaster receives a reward  $r(I_t, y_t)$  and each action  $i$  gets a reward  $r(i, y_t)$ , where **none** of these values **is revealed** to the forecaster;
- **only a feedback**  $s_t$  drawn at random according to  $H(I_t, y_t)$  is revealed to the decision-maker.



**Example:** Dynamic pricing.

A vendor sells **T-shirts** on the Internet and chooses prices in

$$\mathcal{X} = \{9.90, 14.90, 19.90, 24.90, 29.90, \dots, 99.90\}$$

To the  $t$ -th customer, she/he offers the T-shirt at a **price**  $l_t$ .

Customers connect one by one to his web site.

Each of them has in mind a **maximal price**  $y_t \in \mathcal{X} = \mathcal{Y}$  she/he is willing to pay – but does not tell it to the vendor.

When  $y_t \geq l_t$ , the product is bought and the vendor suffers a loss of earnings  $r(l_t, y_t) = l_t - y_t$ .

Otherwise, no deal takes place and the loss equals a fixed  $c$  (accounting for all her/his charges),  $r(l_t, y_t) = -c$ .

The feedback is thus  $H(l_t, y_t) = \delta_{\mathbb{I}_{\{y_t \geq l_t\}}}$ .

We want the **same average payoff** as the the **best constant price**.



Full monitoring: the outcomes are revealed,  $H(i, j) = \delta_j$

Bandit games: the obtained payoffs are revealed,  $H(i, j) = \delta_{r(i, j)}$

Noisy binary observations:  $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ , the matrix representations of the feedback and payoff functions are

$$\begin{bmatrix} (1 - \varepsilon_0, \varepsilon_0) & (\varepsilon_1, 1 - \varepsilon_1) \\ (1 - \varepsilon_0, \varepsilon_0) & (\varepsilon_1, 1 - \varepsilon_1) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Label-efficient prediction: 3 actions, 2 outcomes, the matrix representations of the feedback and payoff functions are

$$\begin{bmatrix} \delta_a & \delta_b \\ \delta_a & \delta_a \\ \delta_a & \delta_a \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A **trade-off** needs to be done between getting **information** (action 1) and getting **rewards** (actions 2 and 3).



A case study: Take the following feedback and payoff matrices,

$$H = \begin{bmatrix} (1, 0) & (1/2, 1/2) & (0, 1) \\ (1, 0) & (1/2, 1/2) & (0, 1) \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 1 \end{bmatrix}$$

If the **averaged observed distribution of signals** is close to  $\Delta = (1/2, 1/2)$ , the decision-maker does not know whether it followed

- from a constant choice of outcome 2,  $\mathbf{q}_2 = (0, 1, 0)$ ;
- from outcomes 1 and 3 played equally often,  $\mathbf{q}_{13} = (1/2, 0, 1/2)$ ;
- or any possible mixing of both,  $\mathbf{q} = \alpha \mathbf{q}_{13} + (1 - \alpha) \mathbf{q}_2$ .

Action 2 is the optimal one against  $\mathbf{q}_2$  (average payoff of 3 vs. 0).

One should play action 1 against  $\mathbf{q}_{13}$  (payoff 2.5 vs. 0.5).

In the **mixing case**, the best action **depends** on the mixing parameter  $\alpha$ .



A case study: Take the following feedback and payoff matrices,

$$H = \begin{bmatrix} (1, 0) & (1/2, 1/2) & (0, 1) \\ (1, 0) & (1/2, 1/2) & (0, 1) \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 1 \end{bmatrix}$$

If the **averaged observed distribution of signals** is close to  $\Delta = (1/2, 1/2)$ , the decision-maker does not know whether it followed

- from a constant choice of outcome 2,  $\mathbf{q}_2 = (0, 1, 0)$ ;
- from outcomes 1 and 3 played equally often,  $\mathbf{q}_{13} = (1/2, 0, 1/2)$ ;
- or any possible mixing of both,  $\mathbf{q} = \alpha \mathbf{q}_{13} + (1 - \alpha) \mathbf{q}_2$ .

Action 1:	$5\alpha/2$
Action 2:	$3(1 - \alpha) + \alpha/2 = 3 - 5\alpha/2$

The worst best payoff is obtained by equating the two payoffs. We should think that the **opponent** was indeed **playing this worst**



## Summary:

We start the formalization of the previous result as follows, by first extending linearly  $r$ .

The **best payoff** the decision-maker can get **for sure given**  $\Delta = (1/2, 1/2)$  equals

$$\begin{aligned} \min_{\alpha} \max \{ r(\delta_1, \mathbf{q}), r(\delta_2, \mathbf{q}) \} &= \min_{\mathbf{q}: H(\mathbf{q})=\Delta} \max_{\mathbf{p}} r(\mathbf{p}, \mathbf{q}) \\ &= \max_{\mathbf{p}} \min_{\mathbf{q}: H(\mathbf{q})=\Delta} r(\mathbf{p}, \mathbf{q}) \end{aligned}$$

by application of the minmax theorem.

The decision-maker should play any element of

$$\operatorname{argmax}_{\mathbf{p}} \min_{\mathbf{q}: H(\mathbf{q})=\Delta} r(\mathbf{p}, \mathbf{q})$$

(these are also **equalizers**).



In **full monitoring**, the aim of the decision-maker is to have a cumulative reward

$$\sum_{t=1}^n r(I_t, y_t)$$

close to (i.e., within  $o(n)$  of) the cumulative reward of the best **fixed action**,

$$\max_{j=1, \dots, N} \sum_{t=1}^n r(j, y_t) = n \max_{\mathbf{p}} r(\mathbf{p}, \bar{\mathbf{q}}_n)$$

where  $\bar{\mathbf{q}}_n$  is the empirical distribution of  $y_1, \dots, y_n$ .

Here, this quantity can in general **not be achieved**.

The reasonable goal is the optimal ex-post payoff

$$n \max_{\mathbf{p}} \min_{\mathbf{q}: H(\mathbf{q})=H(\bar{\mathbf{q}}_n)} r(\mathbf{p}, \mathbf{q})$$



We now give a precise formalization and, again, **extend linearly**  $r$  and  $H$ .

For probability distributions  $\mathbf{p}$  and  $\mathbf{q}$  on  $\{1, \dots, N\}$  and  $\{1, \dots, M\}$ , we define

$$r(\mathbf{p}, \mathbf{q}) = \sum_{i,j} p_i q_j r(i, j)$$

$$\text{and } H(\cdot, \mathbf{q}) = \sum_{j=1, \dots, N} q_j \begin{bmatrix} H(1, j) \\ H(2, j) \\ \vdots \\ H(N, j) \end{bmatrix} \in (\Delta(\mathcal{S}))^N$$

Denote by  $\mathcal{F}$  the set of the  $\Delta$  that may be written as  $H(\cdot, \mathbf{q})$  for some  $\mathbf{q}$ .

The **target function**  $\rho$  is defined as

$$\rho(\mathbf{p}, \Delta) = \min_{\mathbf{q}: H(\cdot, \mathbf{q}) = \Delta} r(\mathbf{p}, \mathbf{q})$$

Recall that  $\bar{\mathbf{q}}_n$  the empirical distribution of  $y_1, \dots, y_n$ .

The previous slides explained why, even with the knowledge of  $\hat{\Delta}_n = H(\cdot, \bar{\mathbf{q}}_n)$ , we cannot hope to do better than  $\max_{\mathbf{p}} \rho(\mathbf{p}, H(\cdot, \bar{\mathbf{q}}_n))$ . The latter is the **optimal ex-post** payoff.

The following theorem shows that it is achievable even in a sequential fashion.

### Theorem (Rustichini '99)

*The decision-maker has a strategy  $\sigma$  such that for all strategies  $\tau$  of Nature,*

$$\limsup_{n \rightarrow \infty} \max_{\mathbf{p}} \rho(\mathbf{p}, H(\cdot, \bar{\mathbf{q}}_n)) - \frac{1}{n} \sum_{t=1, \dots, n} r(I_t, y_t) \leq 0 \quad \text{a.s.}$$



## Definition

The **Rustichini regret**  $R_n$  under partial monitoring is defined as

$$R_n = n \max_{\mathbf{p}} \rho(\mathbf{p}, H(\cdot, \bar{\mathbf{q}}_n)) - \sum_{t=1, \dots, n} r(I_t, y_t)$$

The strategies  $\sigma$  such that  $R_n = o(n)$  a.s. against all strategies  $\tau$  are said **Rustichini consistent**.

Rustichini's proof relies on an approachability theorem for a continuum of types (see Mertens, Sorin, and Zamir '94). It is neither **constructive** nor indicates **convergence rates**.

We deal with both issues.



Two special cases had been dealt with so far,

- the case when the feedback depends **only on the outcome**, see Mannor and Shimkin '03;
- the **Hannan-consistent** case when

$$n \max_{\mathbf{p}} \rho(\mathbf{p}, H(\cdot, \bar{\mathbf{q}}_n)) = \max_{j=1, \dots, N} \sum_{t=1, \dots, n} r(j, y_t) ;$$

this corresponds, for instance, to **dynamic pricing**, multi-armed bandits, label-efficient prediction, noisy binary prediction (provided that  $\varepsilon_1 + \varepsilon_0 \neq 1$ );

see Piccolboni and Schindelhauer '01, Cesa-Bianchi, Lugosi, and Stoltz '06, where a **necessary and sufficient condition** for Hannan consistency is proposed and a  $O(n^{2/3})$  rate for the regret of an explicit forecaster is exhibited and proved to be optimal.



Our new and complete solution relies on **several layers** of techniques, namely,

- the use of a lazy strategy that **groups rounds** together, to be able to **estimate** the original **feedback** distributions,
- which in turns allows **estimation** of (a pessimistic lower bounds on) the unobserved **payoffs**;
- as well as some classical **exploration–exploitation** trade-off and **linearized** upper bounds on the quantities at hand by using sub-gradients of concave functions.

We now give some more details... very few, in view of the remaining time...



The forecaster only sees the signals  $s_t \sim H(I_t, y_t)$  and aims at **reconstructing**  $H(\cdot, \bar{\mathbf{q}}_n) = (H(\cdot, y_1) + \dots + H(\cdot, y_n))/n$ .

The key is to find an unbiased estimate for each  $H(i, y_t)$ . For instance,

$$\hat{h}_{i,t} = \frac{\delta_{s_t}}{p_{i,t}} \mathbb{I}_{\{I_t=i\}}$$

is such that

$$\mathbb{E}_t [\hat{h}_{i,t}] = \frac{1}{p_{i,t}} \mathbb{E}_t [\delta_{s_t} \mathbb{I}_{\{I_t=i\}}] = \frac{1}{p_{i,t}} \mathbb{E}_t [H(I_t, y_t) \mathbb{I}_{\{I_t=i\}}] = \frac{1}{p_{i,t}} p_{i,t} H(i, y_t) = H(i, y_t)$$

Thus, for all  $m$  large enough and  $b$ , with  $\Pi$  projection onto  $\mathcal{F}$

$$\hat{\Delta}^b = \Pi \left( \frac{1}{m} \sum_{t=bm+1}^{(b+1)m} [\hat{h}_{i,t}]_{i=1, \dots, N} \right) \quad \text{estimates} \quad \Delta^b = \frac{1}{m} \sum_{t=bm+1}^{(b+1)m} H(\cdot, y_t)$$



We continue by indicating some analytical properties of  $\rho$ , where we recall that for  $\Delta \in \mathcal{F}$ ,

$$\rho(\mathbf{p}, \Delta) = \min_{\mathbf{q} : H(\cdot, \mathbf{q}) = \Delta} r(\mathbf{p}, \mathbf{q})$$

Thus,  $\rho$  is **concave** in its first argument and **convex** in the second argument.

In addition, it can be shown that  $\rho$  is **uniformly Lipschitz** in its second argument.

The forecaster is as follows (and can be implemented efficiently).



*Parameters:* Integer  $m \geq 1$ , real numbers  $\eta, \gamma > 0$

*Initialization:*  $\mathbf{w}^0 = (1, \dots, 1)$

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

- 1 if  $bm + 1 \leq t < (b + 1)m$  for some integer  $b$ , choose the distribution  $\mathbf{p}_t = \mathbf{p}^b = (1 - \gamma)\tilde{\mathbf{p}}^b + \gamma\mathbf{u}$ , where  $\tilde{\mathbf{p}}^b$  is defined component-wise as

$$\tilde{p}_i^b = \frac{w_i^b}{\sum_{j=1}^N w_j^b}$$

and  $\mathbf{u}$  denotes the uniform distribution,  $\mathbf{u} = (1/N, \dots, 1/N)$ ;

- 2 draw an action  $I_t$  from  $\{1, \dots, N\}$  according to it;
- 3 if  $t = (b + 1)m$  for some integer  $b$ , perform the update

$$w_i^{b+1} = w_i^b e^{\eta(\nabla\rho(\mathbf{p}^b, \hat{\Delta}^b))_i} \quad \text{for each } i = 1, \dots, N,$$

where for all  $\Delta \in \mathcal{F}$ ,  $\nabla\rho(\cdot, \Delta)$  is a sub-gradient of  $\rho(\cdot, \Delta)$  and  $\hat{\Delta}^b$  is defined in the previous slides.



## Theorem (Lugosi, Mannor, and Stoltz '08)

*The regret*

$$R_n = n \max_{\mathbf{p}} \rho(\mathbf{p}, H(\cdot, \bar{\mathbf{q}}_n)) - \sum_{t=1, \dots, n} r(I_t, y_t)$$

*is bounded with overwhelming probability by*

- $O(n^{4/5})$  in the most general case,
- $O(n^{3/4})$  in the case of random signals depending on outcome only,
- $O(n^{2/3})$  in the case of deterministic signals,
- $O(n^{1/2})$  in the case of deterministic signals depending on outcome only.

Cesa-Bianchi, Lugosi, and Stoltz '06 show with label-efficient prediction that the  $n^{2/3}$  rate is optimal (and that the  $n^{1/2}$  rate is optimal as well has been known for a decade now).