

Minimization of regret and convergence to sets of equilibria

Gilles Stoltz

CNRS – École normale supérieure – HEC Paris



- 1 The model of individual sequences
 - A repeated game against Nature
 - Hannan regret
 - Explicit strategies and convergence rates
- 2 Convergence to correlated equilibria
- 3 Bandit games

- 1 The model of individual sequences
 - A repeated game against Nature
 - Hannan regret
 - Explicit strategies and convergence rates
- 2 Convergence to correlated equilibria
- 3 Bandit games

A base one-shot game is repeated

- A decision-maker (the row player) takes actions l_1, l_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 payoff function is $r : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$.

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**.

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



Strategies for the players

For the decision-maker:

A (randomized) strategy σ 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 $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$ 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 maximizing his cumulative payoff.

Strategies for the players

For the opponent player:

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

We present below strategies for the decision-maker which minimize the regret for all possible strategies of the opponent player, in a $\mathbb{P}_{\sigma, \tau}$ -**a.s.** 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 The model of individual sequences
 - A repeated game against Nature
 - **Hannan regret**
 - Explicit strategies and convergence rates
- 2 Convergence to correlated equilibria
- 3 Bandit games

Definition of 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 l_1, l_2, \dots of actions chosen by the decision-maker to constant sequences of pure actions j, j, \dots

That is, we compare $\hat{X}_n = \sum_{t=1}^n r(l_t, y_t)$ 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,

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

Hannan consistent strategies

The **Hannan regret** is defined as

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

Definition

A strategy σ for the decision-maker is said **Hannan consistent** (or universally consistent) whenever

$$\limsup_{n \rightarrow \infty} \frac{R_n}{n} \leq 0 \quad \mathbb{P}_{\sigma, \tau}\text{-a.s.}$$

regardless of the strategy τ of the opponent player.

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

See, among others, those of **Blackwell '56**, **Hannan '57**, Fudenberg and Levine '95.



Properties of Hannan-consistent strategies

We consider a finite zero-sum game (when \mathcal{X} and \mathcal{Y} are finite sets and the opponent has payoff function $-r$).

If the decision-maker uses a Hannan-consistent strategy, then his **average payoff** is at least the **value v** of the base one-shot game,

$$\liminf_{n \rightarrow \infty} \frac{\widehat{X}_n}{n} \geq v = \min_{\mathbf{q}} \max_{\mathbf{p}} r(\mathbf{p}, \mathbf{q}) .$$

The proof denotes by $\bar{\mathbf{q}}_n = \frac{1}{n} \sum_{t=1}^n \delta_{y_t}$ the empirical distribution of plays of the opponent player,

$$\begin{aligned} \liminf \frac{\widehat{X}_n}{n} &\geq \liminf \max_{j=1, \dots, N} \frac{X_{j,n}}{n} = \liminf \max_{j=1, \dots, N} r(j, \bar{\mathbf{q}}_n) \\ &= \liminf \max_{\mathbf{p}} r(\mathbf{p}, \bar{\mathbf{q}}_n) \geq \min_{\mathbf{q}} \max_{\mathbf{p}} r(\mathbf{p}, \mathbf{q}) = v . \end{aligned}$$

If **both** players use such strategies, then $\widehat{X}_n/n \rightarrow v$.

Properties of Hannan-consistent strategies

As a **consequence**, if both players use Hannan-consistent strategies σ and τ , then the product of the marginal distributions of plays converges to the set of minimax distributions (**same** guarantee as for **fictitious play**).

Indeed, denoting by

$$\bar{\mathbf{p}}_n = \frac{1}{n} \sum_{t=1}^n \delta_{I_t} \quad \text{and} \quad \bar{\mathbf{q}}_n = \frac{1}{n} \sum_{t=1}^n \delta_{J_t}$$

the empirical distributions of plays, we get from the previous slide (and by symmetry)

$$\lim_{n \rightarrow \infty} \max_{\mathbf{p}} r(\mathbf{p}, \bar{\mathbf{q}}_n) = \lim_{n \rightarrow \infty} \min_{\mathbf{q}} r(\bar{\mathbf{p}}_n, \mathbf{q}) = v$$

and thus

$$r(\bar{\mathbf{p}}_n, \bar{\mathbf{q}}_n) \longrightarrow v .$$

Properties of Hannan-consistent strategies

What about the empirical frequencies of **pairs of actions**?

We denote

$$\bar{\pi}_n = \frac{1}{n} \sum_{t=1}^n \delta_{(I_t, Y_t)} .$$

$(\bar{\pi}_n)$ does **not** converge to the set of **minimax** distributions, **not** even to the larger set of **correlated** equilibria.

But it converges to the even **larger** so-called **Hannan set**

$$\mathcal{H} = \left\{ \begin{array}{l} \pi : \quad \forall k, k', \quad \sum_{i,y} \pi(i,y)r(i,y) \geq r(k, \pi^2) \\ \text{and} \quad - \sum_{i,y} \pi(i,y)r(i,y) \geq -r(\pi^1, k') \end{array} \right\}$$

where π^1 and π^2 denote the first and second marginal distributions of the joint probability distribution π over $\mathcal{X} \times \mathcal{Y}$.



- 1 The model of individual sequences
 - A repeated game against Nature
 - Hannan regret
 - Explicit strategies and convergence rates
- 2 Convergence to correlated equilibria
- 3 Bandit games

Aim: Non-asymptotic statements

We want not only to know that $(\bar{\mathbf{p}}_n, \bar{\mathbf{q}}_n)$ is asymptotically a minimax equilibrium or that $\bar{\pi}_n$ belongs asymptotically to the Hannan set.

We want statements of the form:

- $(\bar{\mathbf{p}}_n, \bar{\mathbf{q}}_n)$ is an ε_n -minimax equilibrium,
- or $\bar{\pi}_n$ is in an ε_n -neighbourhood of the Hannan set,

for some ε_n .

We will actually have **high-probability statements**:

With probability at least $1 - \delta$, either of the previous statements is true with ε_n of the form $\square \sqrt{n \ln(N/\delta)}$.

To do so, we bound the regret with high probability.



Regret vs. expected regret

Recall that we want to minimize the **Hannan regret**

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

and that l_t is drawn at random according to \mathbf{p}_t .

Denote by \mathbb{E}_t the **conditional expectation** at round t ,

$$\mathbb{E}_t[r(l_t, y_t)] = \sum_{i=1,\dots,N} p_{i,t} r(i, y_t) = r(\mathbf{p}_t, y_t)$$

By **martingale convergence** (r is bounded),

$$\bar{X}_n - \hat{X}_n = \sum_{t=1,\dots,n} r(\mathbf{p}_t, y_t) - \sum_{t=1,\dots,n} r(l_t, y_t) = o_{\mathbb{P}}(n)$$

We may thus focus on the **expected regret** \bar{R}_n ,

$$\bar{R}_n = \max_{j=1,\dots,N} X_{j,n} - \bar{X}_n = \max_{j=1,\dots,N} \sum_{t=1,\dots,n} r(j, y_t) - \sum_{t=1,\dots,n} r(\mathbf{p}_t, y_t)$$



Regret vs. expected regret

Since we are interested in convergence rates, let's be more specific.

r takes values in $[0, 1]$, thus **Hoeffding–Azuma** inequality ensures that with probability at least $1 - \delta$,

$$\Delta_n = \bar{X}_n - \hat{X}_n = \sum_{t=1, \dots, n} r(\mathbf{p}_t, y_t) - \sum_{t=1, \dots, n} r(l_t, y_t) \leq \sqrt{\frac{n}{2} \ln \frac{1}{\delta}}$$

As a consequence (as $R_n = \bar{R}_n + \Delta_n$), with probability $1 - \delta$,

$$\begin{aligned} & \max_{j=1, \dots, N} \sum_{t=1, \dots, n} r(j, y_t) - \sum_{t=1, \dots, n} r(l_t, y_t) \\ & \leq \left(\max_{j=1, \dots, N} \sum_{t=1, \dots, n} r(j, y_t) - \sum_{t=1, \dots, n} r(\mathbf{p}_t, y_t) \right) + \left(\sum_{t=1, \dots, n} r(\mathbf{p}_t, y_t) - \sum_{t=1, \dots, n} r(l_t, y_t) \right) \\ & \leq \left(\max_{j=1, \dots, N} \sum_{t=1, \dots, n} r(j, y_t) - \sum_{t=1, \dots, n} r(\mathbf{p}_t, y_t) \right) + \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} \end{aligned}$$



Regret vs. expected regret

We will see in a minute that this leads to the upper bound, holding with probability at least $1 - \delta$,

$$R_n \leq \sqrt{\frac{n}{2} \ln N} + \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} .$$

An application of the **Borel-Cantelli** lemma then yields (via the choice, e.g., of $\delta_n = 1/n^2$)

$$\limsup \frac{R_n}{\sqrt{n \ln n}} < +\infty \quad \text{and thus,} \quad \limsup \frac{R_n}{n} \leq 0 \quad \text{a.s.}$$



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

p_1 is uniform and for $t \geq 2$,

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

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

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

Exponential reweighting

An important assumption is that the payoff function takes bounded values, $r : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$.

Theorem

For *all strategies* τ of the opponent player, the expected regret 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 $\eta = \sqrt{8 \ln N / n}$.

We now **prove** this bound, and discuss later the **tuning of η** .

Note: It can be shown to be **optimal**.



Hoeffding's lemma

In the **proof**, the following lemma, due to Hoeffding, will be the key step:

Lemma

A bounded random variable X , $0 \leq X \leq B$, satisfies, for all $s > 0$,

$$s\mathbb{E}[X] \leq \log \mathbb{E} \left[e^{sX} \right] \leq s\mathbb{E}[X] + \frac{s^2}{8} B^2$$

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 $t \geq 2$,

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

On the one hand,

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

On the other hand, for $t = 1, \dots, n$,

$$\begin{aligned} \ln \frac{W_t}{W_{t-1}} &= \ln \frac{\sum_{i=1}^N e^{\eta r(i, y_t)} w_{i,t-1}}{\sum_{j=1}^N w_{j,t-1}} = \ln \sum_{i=1}^N p_{i,t} e^{\eta r(i, y_t)} \\ &\leq \eta \left(\sum_{i=1}^N p_{i,t} r(i, y_t) \right) + \frac{\eta^2}{8} = \eta r(\mathbf{p}_t, y_t) + \frac{\eta^2}{8} \end{aligned}$$

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

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



Tuning of η

We can use a “doubling trick” or use all information available by allowing η to depend on time,

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

where $\eta_t = \sqrt{8 \ln N / (t - 1)}$.

Auer, Cesa-Bianchi, and Gentile '02 show that the (expected) regret of this strategy is less than $1 + 2\sqrt{(n/2) \ln N}$.

These tuning issues can also be overcome by choosing a different **potential function**.



Extension to convex payoffs

Assume now that the game is given by $\mathcal{X} = \mathcal{S}$ the simplex of probability distributions in \mathbb{R}^N and a payoff function $r : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ **concave** in the first argument.

The Hannan regret is

$$\begin{aligned} \max_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^n r(\mathbf{x}, y_t) - \sum_{t=1}^n r(\mathbf{p}_t, y_t) &\leq \max_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^n \nabla r(\mathbf{p}_t, y_t) \cdot (\mathbf{x} - \mathbf{p}_t) \\ &\leq \max_{j=1, \dots, N} \sum_{t=1}^n (\nabla r(\mathbf{p}_t, y_t))_j - \sum_{t=1}^n \nabla r(\mathbf{p}_t, y_t) \cdot \mathbf{p}_t \end{aligned}$$

Thus, defining \mathbf{p}_t as an exponentially weighted average (over the sums of the **components of the sub-gradients**) leads to a regret less than $\sqrt{n \ln N}$.



- 1 The model of individual sequences
 - A repeated game against Nature
 - Hannan regret
 - Explicit strategies and convergence rates
- 2 Convergence to correlated equilibria
- 3 Bandit games

Definition of correlated equilibria

We consider the **general case** of 2-players finite games, with payoff functions r_1 and r_2 .

The set of **correlated equilibria** is formed by the following joint distributions,

$$\mathcal{C} = \left\{ \begin{array}{l} \pi : \quad \forall \varphi, \varphi', \quad \sum_{i,y} \pi(i,y)r_1(i,y) \geq \sum_{i,y} \pi(i,y)r_1(\varphi(i), y) \\ \quad \text{and} \quad \sum_{i,y} \pi(i,y)r_2(i,y) \geq \sum_{i,y} \pi(i,y)r_2(i, \varphi'(y)) \end{array} \right\}$$

where φ , resp. φ' , is any function that maps the action space of the decision-maker, resp. of the opponent player, into itself.

Note: The results for the Hannan set also held in this general framework, but not the ones indicating convergence to the set of minimax equilibria.

Definition of internal regret

In the definition of correlated equilibria, it is enough to consider all functions φ and φ' that differ from identity only at one point.

Following this intuition, we define the $j \rightarrow k$ modification of a given strategy as the strategy playing k whenever the given strategy plays j or k and playing according to it otherwise.

The cumulative payoff of this strategy is denoted by $\widehat{X}_n^{j \rightarrow k}$ (for the decision-maker).

Definition

The **internal regret** of a given strategy of the decision-maker is defined by

$$R_n^{\text{int}} = \max_{j \neq k} \widehat{X}_n^{j \rightarrow k} - \widehat{X}_n .$$

Convergence to the set of correlated equilibria

If **both** players minimize their **internal regrets**, then the empirical frequencies of pairs of actions

$$\bar{\pi}_n = \frac{1}{n} \sum_{t=1}^n \delta_{(I_t, Y_t)}$$

are each ε_n -**correlated equilibria**.

In particular, the sequence $(\bar{\pi}_n)$ converges to the **set \mathcal{C}** of correlated equilibria.

The ε_n are given by the maximum of the two internal regrets (decision-maker, opponent player) at round n .

As we show on the next slide, typically, $\varepsilon_n = \square \sqrt{n \ln(N/\delta)}$ with high probability $1 - \delta$.



A simple trick to minimize internal regret

The trick is described in Stoltz and Lugosi '05. A related trick was found independently by Blum and Mansour '07.

We simply use the exponentially weighted average predictor on the **pseudo-actions** $j \rightarrow k$ and solve a **fixed-point** equation.

More precisely, we choose at rounds $t \geq 2$, the distribution \mathbf{p}_t such that

$$\mathbf{p}_t = \frac{\exp(\eta \widehat{X}_{t-1}^{k \rightarrow j})}{\sum_{k' \neq j'} \exp(\eta \widehat{X}_{t-1}^{k' \rightarrow j'})} \mathbf{p}_t^{k \rightarrow j}.$$

Since there are $N(N-1)$ pseudo-predictors, the regret can be deduced from the usual one:

with probability at least $1 - \delta$, the internal regret is less than

$$R_n^{\text{int}} = \sqrt{\frac{n}{2} \ln N(N-1)} + 2\sqrt{\frac{n}{2} \ln \frac{1}{\delta}}.$$

- 1 The model of individual sequences
 - A repeated game against Nature
 - Hannan regret
 - Explicit strategies and convergence rates
- 2 Convergence to correlated equilibria
- 3 Bandit games

A repeated game against Nature, with partial monitoring

Parameters (known to both players): number N of actions, outcome set \mathcal{Y} , payoff function r

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

- 1 the environment chooses the next outcome $y_t \in \mathcal{Y}$ without revealing it;
- 2 the forecaster chooses a probability distribution \mathbf{p}_t and draws an action $I_t \in \{1, \dots, N\}$ according to this distribution;
- 3 the forecaster receives reward $r(I_t, y_t)$ and each action i gets reward $r(i, y_t)$;
- 4 **only his own reward $r(I_t, y_t)$** is revealed to the forecaster.

Hannan regret is to be made small, $\max_{j=1, \dots, N} X_{j,n} - \widehat{X}_n = o(n)$ a.s.

Estimates of the unobserved payoffs

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

The estimates are

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

We still 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 payoffs

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

The estimates are

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

We still 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 ,

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

We now perform exponentially weighted averages on these unbiased estimates. (Well, almost...)

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

Exponential reweighting on estimated payoffs

We use the estimated payoffs $\tilde{r}_{j,t} = \frac{r(I_t, y_t)}{p_{I_t,t}} \mathbb{1}_{[I_t=j]}$

Exponentially weighted average predictor

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

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

This forecaster has a \sqrt{n} regret in expectation and a $O(n^{2/3})$ regret with high probability.

The mixing is needed for high probability bounds to bound from below the $p_{i,t}$, which in turn, **bounds** from above the **conditional variances** of the estimators $\tilde{r}_{j,t}$.

Exponential reweighting on shifted estimated payoffs

We use the estimated payoffs $\tilde{r}_{j,t} = \frac{r(l_t, y_t)}{p_{l_t,t}} \mathbb{I}_{[l_t=j]}$

Exponentially weighted average predictor

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

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

For properly chosen parameters η, γ, β , and with probability at least $1 - \delta$, the regret of this forecaster is

$$\max_{j=1, \dots, N} X_{j,n} - \hat{X}_n \leq 6 \sqrt{nN \ln \frac{N}{\delta}} + \frac{\ln N}{2}$$

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 payoff function $r : \mathbb{N} \times \mathcal{Y} \rightarrow \{0, 1\}$ such that for all $N \geq 2$ and $n \geq 1$, and all strategies suited to bandit settings,

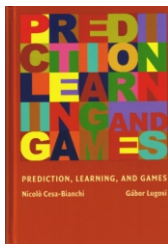
$$\max_{j=1,\dots,N} X_{j,n} - \bar{X}_n \geq \frac{1}{20} \min \left\{ \sqrt{nN}, n \right\}$$

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

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

Conclusion

For more details, see



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