Regret Minimization under Partial Monitoring

Nicolò Cesa-Bianchi

Università degli Studi di Milano

joint work with
Gábor Lugosi and Gilles Stoltz

Playing a repeated zero-sum game

Known loss matrix with entries in $[0,1]$

	1	$\cdots$	M
1	$\ell(1,1)$	$\cdots$	$\ell(1,M)$
$\vdots$	$\vdots$	$\ell(I_t,y_t)$	$\vdots$
N	$\ell(N,1)$	$\cdots$	$\ell(N,M)$

For $t = 1,2,\ldots$

- Row player (forecaster) chooses distribution p_t over $\{1,\ldots,N\}$
- Column player (adversary) chooses action $y_t \in \{1,\ldots,M\}$
- Row player draws $I_t \in \{1,\ldots,N\}$ according to p_t

Regret and Hannan consistency

Play at round t may depend on past plays (I_s, y_s), $s < t$

Regret

$$R_n = \frac{1}{n}\sum_{t=1}^{n} \ell(I_t, y_t) - \min_{k=1,\dots,N} \frac{1}{n}\sum_{t=1}^{n} \ell(k, y_t)$$

Forecaster is Hannan consistent if

$$\limsup_{n\to\infty} R_n = 0 \qquad \text{with probability 1}$$

irrespective to what adversary does

Game with full information

After drawing I_t the forecaster observes the adversary's play y_t

	1	$\cdots$	y_t	$\cdots$	M
1	$\ell(1,1)$	$\cdots$	$\ell(1,y_t)$	$\cdots$	$\ell(1,M)$
$\vdots$			$\vdots$		
I_t	$\vdots$		$\ell(I_t,y_t)$		$\vdots$
$\vdots$			$\vdots$		
N	$\ell(N,1)$	$\cdots$	$\ell(N,y_t)$	$\cdots$	$\ell(N,M)$

Regret vanishes at rate $\sqrt{\frac{\ln N}{n}}$

Nonstochastic bandits

After drawing I_t the forecaster observes his own loss $\ell(I_t, y_t)$

	1	$\cdots$	y_t	$\cdots$	M
1	$\ell(1,1)$		$\cdots$		$\ell(1,M)$
$\vdots$					
I_t	$\vdots$		$\ell(I_t, y_t)$		$\vdots$
$\vdots$					
N	$\ell(N,1)$		$\cdots$		$\ell(N,M)$

Regret vanishes at rate $\sqrt{\dfrac{N \ln N}{n}}$

Partial monitoring

$$\begin{array}{|ccc|}\hline \ell(1,1) & \cdots & \ell(1,M) \\ \vdots & \ell(I_t,y_t) & \vdots \\ \ell(N,1) & \cdots & \ell(N,M) \\ \hline \end{array}$$

Loss matrix L

$$\begin{array}{|ccc|}\hline h(1,1) & \cdots & h(1,M) \\ \vdots & h(I_t,y_t) & \vdots \\ h(N,1) & \cdots & h(N,M) \\ \hline \end{array}$$

Feedback matrix H

- After drawing I_t the forecaster observes a signal $h(I_t, y_t)$
- For $H \equiv L$ this reduces to nonstochastic bandits

Dynamic pricing

	1	2	3	4	5
1	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
2	c	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$
3	c	c	0	$\frac{1}{4}$	$\frac{1}{2}$
4	c	c	c	0	$\frac{1}{4}$
5	c	c	c	c	0

Loss matrix H

	1	2	3	4	5
1	1	1	1	1	1
2	0	1	1	1	1
3	0	0	1	1	1
4	0	0	0	1	1
5	0	0	0	0	1

Feedback matrix H

- Forecaster's action is the price at which a product sold online is offered to t-th customer
- Adversary's action is maximum price at which t-th customer is willing to buy the product
- Feedback is 1 for SOLD and 0 for NOT SOLD

Previous work

- Repeated games: [Hannan, 1956] [Blackwell, 1956] "Prediction with Expert Advice" (computer science)
- Nonstochastic bandits: [Baños, 1968] [Megiddo, 1980] [Auer, C-B, Freund and Schapire, 2002]
- Partial monitoring: [Mertens, Sorin, and Zamir, 1994] [Rustichini, 1999] [Piccolboni and Schindelhauer, 2001]

Partial monitoring

- Rustichini establishes existence of Hannan consistent strategies (even for stochastic signals)
- Piccolboni and Schindelhauer give general conditions for convergence of expected regret
- This work: explicit algorithms with optimal rates for actual regret (Hannan consistency)

Upper bound

[Piccolboni and Schindelhauer, 2001] [C-B, Lugosi, and Stoltz, 2005]

Recall rate for nonstochastic bandits: $\sqrt{(N \ln N)/n}$

Theorem

If a partial monitoring game (L, H) *satisfies* $L = K\,H$ *for some matrix* K*, then there exists a forecaster whose regret is at most*

$$c \left(\frac{N^2 \ln N}{n} \right)^{1/3} \qquad \textit{w.h.p.}$$

$\longrightarrow$ Hannan consistency for the dynamic pricing problem

Dependence on M?

Proof ideas

- Exponential weighting scheme

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

- Pseudo-loss

$$\widehat{\ell}(i, y_t) = \frac{k(i, I_t)\, h(I_t, y_t)}{p_{I_t,t}}$$

- Since $L = K\,H$

$$\mathbb{E}\left[\widehat{\ell}(i, y_t) \,\middle|\, I_1, \dots, I_{t-1}\right] = \sum_{j=1}^{N} \frac{k(i, j)\, h(j, y_t)}{p_{j,t}} \times p_{j,t} = \ell(i, y_t)$$

- Forecaster's distribution

$$\mathbb{P}(I_t = i) = (1-\gamma)\frac{w_{i,t-1}}{\sum_{j=1}^{N} w_{j,t-1}} + \frac{\gamma}{N}$$

The revealing action game [Helmbold, Littlestone, and Long, 2000]

	0	1
0	0	1
1	1	0
2	1	1

Loss matrix L

	0	1
0	a	a
1	a	a
2	b	c

Feedback matrix H

Theorem

If a forecaster plays the revealing action at most m times, then its regret is at least $c_1 \frac{m}{n} + c_2 \frac{1}{\sqrt{m}}$ *for some* $y_1, \dots, y_n$

This construction can be generalized to obtain $\left(\frac{\ln N}{n}\right)^{1/3}$

Gap theorem [Piccolboni and Schindelhauer, 2001]

In any partial monitoring problem,

- either the regret is $\Omega(1)$ for all forecasters
- or there exists a forecaster whose regret is $O\left(n^{-1/3}\right)$