Strategies for prediction under imperfect monitoring

Gábor Lugosi

ICREA and Pompeu Fabra University, Barcleona, Spain

joint work with

Shie Mannor (Technion, McGill)

Gilles Stoltz (ENS, Paris),

randomized prediction

Prediction game:

number $\mathbf{N}$ of actions,
cardinality $\mathbf{M}$ of outcome space,
reward function $\mathbf{r : \{1, \ldots, N\} \times \{1, \ldots, M\} \longrightarrow [0, 1]}$,
number $\mathbf{n}$ of game rounds.

For each round $\mathbf{t = 1, 2, \ldots, n,}$

(1) the environment chooses the next outcome $\mathbf{y_t}$;

(2) the forecaster chooses $\mathbf{p_t}$ and determines the random action $\mathbf{I_t}$, distributed according to $\mathbf{p_t}$;

(3) the environment reveals $\mathbf{y_t}$;

(4) the forecaster receives a reward $\mathbf{r(I_t, y_t)}$.

regret

The goal of the forecaster is to minimize the regret

$$\max_{i=1,\ldots,N} \frac{1}{n} \sum_{t=1}^{n} r(i, J_t) - \frac{1}{n} \sum_{t=1}^{n} r(I_t, J_t)$$

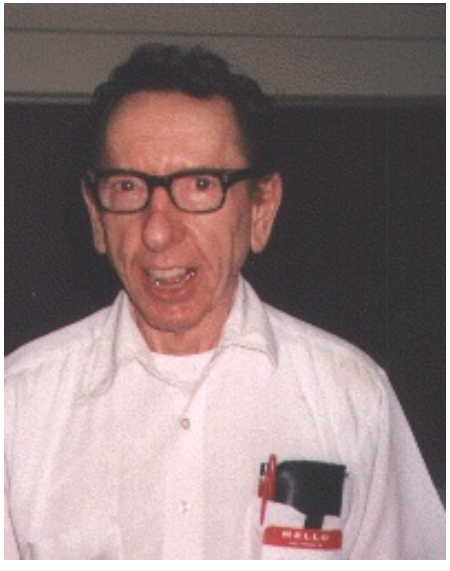
Hannan (1957) and Blackwell (1956) showed that the forecaster has a strategy such that the regret goes to zero almost surely for all strategies of the environment.

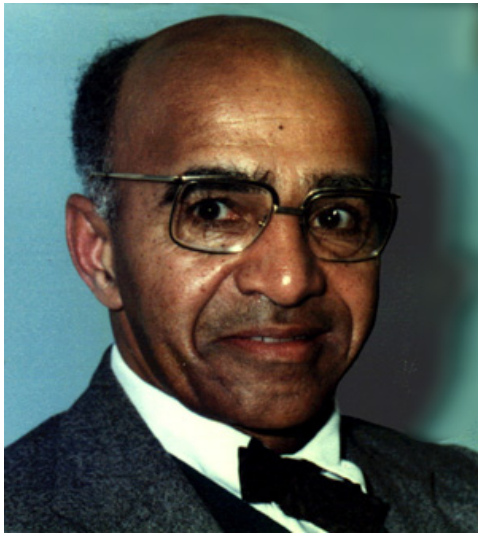
regret

The goal of the forecaster is to minimize the regret

$$\max_{i=1,\ldots,N} \frac{1}{n} \sum_{t=1}^{n} r(i, J_t) - \frac{1}{n} \sum_{t=1}^{n} r(I_t, J_t)$$

Hannan (1957) and Blackwell (1956) showed that the forecaster has a strategy such that the regret goes to zero almost surely for all strategies of the environment.





basic ideas

If $\mathbf{r(p,j) = \sum_{i=1}^{N} p_i r(i,j)}$ then by martingale convergence, with probability at least $\mathbf{1-\delta}$,

$$\mathbf{\frac{1}{n}\sum_{t=1}^{n} r(I_t, J_t) \geq \frac{1}{n}\sum_{t=1}^{n} r(p_t, J_t) - \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}} \ .$$

It suffices to study $\mathbf{(1/n)\sum_{t=1}^{n} r(p_t, J_t)}$.

exponential weights

The exponentially weighted average forecaster selects action I_t with probability

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

where $\eta > 0$. It is well known that

$$\max_{i=1,\ldots,N} \frac{1}{n} \sum_{t=1}^{n} r(i, J_t) - \frac{1}{n} \sum_{t=1}^{n} r(p_t, J_t) \leq \frac{\ln N}{n\eta} + \frac{\eta}{8} .$$

With the choice $\eta = \sqrt{8 \ln N / n}$ the upper bound becomes $\sqrt{\ln N / (2n)}$.

Multi-armed bandits

The forecaster only observes $r(I_t, J_t)$ but not $r(i, J_t)$ for $i \neq I_t$.
Trick: estimate $r(i, J_t)$ by

$$\widetilde{r}(i, J_t) = \frac{r(I_t, J_t)\mathbb{I}_{[I_t=i]}}{p_{I_t,t}}$$

This is an unbiased estimate:

$$\mathbb{E}_t\widetilde{r}(i, J_t) = r(i, J_t)$$

Use the estimated losses to define exponential weights. One gets

$$\mathbb{E}\frac{1}{n}\left(\max_{i\leq N}\sum_{t=1}^{n} r(i, J_t) - \sum_{t=1}^{n} r(p_t, J_t)\right) = O\left(\sqrt{\frac{N \ln N}{n}}\right) ,$$

Auer, Cesa-Bianchi, Freund, and Schapire (2002).

imperfect monitoring

$\mathbf{S}$ is a finite set of signals.

Feedback matrix: $\mathbf{H} : \{1, \ldots, N\} \times \{1, \ldots, M\} \longrightarrow \mathcal{P}(S)$.

For each round $\mathbf{t = 1, 2 \ldots, n}$,

1. the environment chooses the next outcome $J_t \in \{1, \ldots, M\}$ without revealing it;
2. the forecaster chooses p_t and draws an action $I_t \in \{1, \ldots, N\}$ according to it;
3. the forecaster receives reward $r(I_t, J_t)$ and each action i gets reward $r(i, J_t)$, none of these values is revealed to the forecaster;
4. a feedback s_t drawn at random according to $H(I_t, J_t)$ is revealed to the forecaster.

target

Define

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

$$\mathbf{H(\cdot, q) = (H(1,q), \ldots, H(N,q))}$$

where $\mathbf{H(i,q) = \sum_j q_j H(i,j)}$.

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

$\mathcal{F}$ is the set of "observable" vectors of signal distributions $\mathbf{\Delta}$.
The key quantity is

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

ρ is concave in $\mathbf{p}$ and convex in $\mathbf{\Delta}$.

rustichini's theorem

The value of the base one-shot game is

$$\max_{p} \min_{q} r(p, q) = \max_{p} \min_{\Delta \in \mathcal{F}} \rho(p, \Delta)$$

If $\bar{q}_n$ is the empirical distribution of $y_1, \ldots, y_n$, even with the knowledge of $H(\cdot, \bar{q}_n)$ we cannot hope to do better than $\max_p \rho(p, H(\cdot, \bar{q}_n))$.

Rustichini (1999) proved that there exists a strategy such that for all strategies of the opponent, almost surely,

$$\limsup_{n \to \infty} \left(\max_{p} \rho\left(p, H(\cdot, \bar{q}_n)\right) - \frac{1}{n} \sum_{t=1,\ldots,n} r(I_t, y_t) \right) \leq 0$$

rustichini's theorem

Rustichini's proof relies on an approachability theorem for a continuum of types (Mertens, Sorin, and Zamir, 1994).

It is non-constructive.

It does not imply any convergence rate.

We construct efficiently computable strategies that guarantee fast rates of convergence.

deterministic feedback, depending on the outcome

Consider first the special case $H(I_t, J_t) = h(J_t)$, deterministic. Clearly,

$$r(p, j) \geq \rho(p, \delta_{h(j)}) \ .$$

We introduce a forecaster motivated by the gradient-based strategies (see Cesa-Bianchi and Lugosi, 2006, Section 2.5).

The forecaster uses a sub-gradient of $\rho(\cdot, \delta_{h(J_t)})$. $I_t = i$ with probability

$$p_{i,t} = \frac{e^{\eta \sum_{s=1}^{t-1} \left(\widetilde{r}(p_s, \delta_{h(J_s)})\right)_i}}{\sum_{j=1}^{N} e^{\eta \sum_{s=1}^{t-1} \left(\widetilde{r}(p_s, \delta_{h(J_s)})\right)_j}},$$

where $\left(\widetilde{r}(p_s, \delta_{h(J_s)})\right)_i$ is the i-th component of a sub-gradient $\widetilde{r}(p_s, \delta_{h(J_s)}) \in \nabla \rho(p_s, \delta_{h(J_s)})$ of the concave function $\rho(\cdot, \delta_{h(J_s)})$.

analysis

$$
\begin{aligned}
& n\rho(p, H(\bar{q}_n)) - \sum_{t=1}^{n} r(I_t, J_t) \\
\approx \quad & n\rho(p, H(\bar{q}_n)) - \sum_{t=1}^{n} r(p_t, J_t) \\
\leq \quad & n\rho(p, H(\bar{q}_n)) - \sum_{t=1}^{n} \rho(p_t, \delta_{h(J_t)}) \\
\leq \quad & \sum_{t=1}^{n} \left(\rho(p, \delta_{h(J_t)}) - \rho(p_t, \delta_{h(J_t)})\right)
\end{aligned}
$$

(by convexity of ρ in the second argument)

analysis

$$
\begin{aligned}
&\leq \sum_{t=1}^{n} \tilde{r}(p_t, \delta_{h(J_t)}) \cdot (p - p_t) \\
&\quad \text{(by concavity of } \rho \text{ in the first argument)} \\
&\leq \frac{\ln N}{\eta} + \frac{nK^2\eta}{2} \\
&= O(\sqrt{n \ln N})
\end{aligned}
$$

random feedback, depending on the outcome

We still assume $H(I_t, J_t) = H(J_t)$, but $H(j)$ is a distribution over signals.

At time t the forecaster observes s_t drawn from $H(J_t)$.

$$H(\bar{q}_n) = \frac{1}{n} \sum_{t=1}^{n} H(J_t)$$

needs to be estimated.

Idea: a lazy strategy. Group together $m \ll n$ time rounds and use the same mixed strategy.

In the $\mathbf{b}$-th group one can calculate

$$\frac{1}{\mathbf{m}} \sum_{\mathbf{t=bm+1}}^{\mathbf{(b+1)m}} \delta_{\mathbf{S_t}}$$

and project it to the set $\mathcal{F}$ of feasible distributions:

$$\widehat{\boldsymbol{\Delta}}^{\mathbf{b}} = \boldsymbol{\Pi} \left(\frac{1}{\mathbf{m}} \sum_{\mathbf{t=bm+1}}^{\mathbf{(b+1)m}} \delta_{\mathbf{S_t}} \right) .$$

If

$$\boldsymbol{\Delta}^{\mathbf{b}} = \frac{1}{\mathbf{m}} \sum_{\mathbf{t=bm+1}}^{\mathbf{(b+1)m}} \mathbf{H}(\mathbf{J_t}) \quad (\in \mathcal{F})$$

then $\widehat{\boldsymbol{\Delta}}^{\mathbf{b}} \approx \boldsymbol{\Delta}^{\mathbf{b}}$ by (vector valued) martingale convergence.

the strategy

For each round $\mathbf{t = 1, 2, \ldots}$

1. If $\mathbf{bm + 1 \le t < (b + 1)m}$ for some integer $\mathbf{b}$, choose the distribution $\mathbf{p_t = p^b}$ given by

$$\mathbf{p_{k,t} = p_k^b = \frac{w_k^b}{\sum_{j=1}^N w_j^b}}$$

   and draw an action $\mathbf{I_t}$ from $\mathbf{\{1, \ldots, N\}}$ according to it;

2. if $\mathbf{t = (b + 1)m}$ for some integer $\mathbf{b}$, perform the update

$$\mathbf{w_k^{b+1} = w_k^b \, e^{\eta \left(\widetilde{r}\left(p^b, \widehat{\Delta}^b\right)\right)_k}} \qquad \text{for each } \mathbf{k = 1, \ldots, N},$$

   where for all $\mathbf{\Delta}$, $\mathbf{\widetilde{r}(\cdot, \Delta)}$ is a sub-gradient of $\boldsymbol{\rho(\cdot, \Delta)}$.

performance

By optimizing parameters $m \approx \sqrt{n}$ and $\eta \approx n^{-1/4}\sqrt{\ln N}$, we get the regret bound

$$\max_p \rho\left(p, H(\cdot, \bar{q}_n)\right) - \frac{1}{n} \sum_{t=1,\ldots,n} r(I_t, y_t)$$
$$= O(n^{-1/4}\sqrt{\ln(nN/\delta)})$$

which holds with probability $> 1 - \delta$.

the general case

Now the random feedback depends on the action–outcome pairs: $\mathbf{H(I_t, J_t)}$.

Again, we need to estimate the (unobserved) $\mathbf{H(\cdot, \overline{q}_n)}$. Let

$$\widehat{\mathbf{h}}_{\mathbf{i,t}} = \frac{\delta_{\mathbf{s_t}}}{\mathbf{p_{i,t}}} \mathbb{I}_{\mathbf{[I_t = i]}} \ .$$

Then $\widehat{\mathbf{h}}_{\mathbf{i,t}}$ is conditionally unbiased:

$$\mathbb{E}_{\mathbf{t}} \left[\widehat{\mathbf{h}}_{\mathbf{i,t}} \right] = \frac{\mathbf{1}}{\mathbf{p_{i,t}}} \mathbb{E}_{\mathbf{t}} \left[\delta_{\mathbf{s_t}} \mathbb{I}_{\mathbf{[I_t = i]}} \right] = \mathbf{H(i, J_t)} \ .$$

the general case

Define

$$\widehat{\Delta}^b = \Pi\left(\frac{1}{m}\sum_{t=bm+1}^{(b+1)m}\left[\widehat{h}_{i,t}\right]_{i=1,\ldots,N}\right)$$

This will be close to

$$\Delta^b = \frac{1}{m}\sum_{t=bm+1}^{(b+1)m} H(\cdot, J_t)$$

provided $p_{i,t}$ is not too small.

the strategy

For each round $\mathbf{t = 1, 2, \ldots}$

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

$$\widetilde{p}_k^b = \frac{w_k^b}{\sum_{j=1}^N w_j^b}$$

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

2. draw an action $\mathbf{I_t}$ from $\mathbf{\{1, \ldots, N\}}$ according to it;

3. if $\mathbf{t = (b + 1)m}$ for some integer $\mathbf{b}$, perform the update

$$w_k^{b+1} = w_k^b \, e^{\eta \left(\widetilde{r}\left(p^b, \widehat{\Delta}^b\right)\right)_k} \qquad \text{for each } k = 1, \ldots, N,$$

   where for all $\mathbf{\Delta \in \mathcal{F}}$, $\widetilde{\mathbf{r}}\mathbf{(\cdot, \Delta)}$ is a sub-gradient of $\boldsymbol{\rho}\mathbf{(\cdot, \Delta)}$.

performance

By choosing $m \approx n^{3/5}$, $\gamma \approx n^{-1/5}$, and $\eta \approx n^{-1/5}\sqrt{\ln N}$, we obtain

$$\max_p \rho\left(p, H(\cdot, \bar{q}_n)\right) - \frac{1}{n} \sum_{t=1,\ldots,n} r(l_t, y_t) = O(n^{-1/5} N \sqrt{\ln(nN/\delta)})$$

which holds with probability $> 1 - \delta$.

For deterministic feedback this can be improved to $O(n^{-1/3} N^{2/3} \sqrt{\ln(1/\delta)})$.

In the deterministic case the rates (as a function of n) cannot be improved.

remarks

The strategies involve computation of $\mathbf{l_2}$ projections to a convex set and computation of (sub)gradients of piecewise linear concave functions.

These can be done in time polynomial in $\mathbf{N}$ and $|\mathbf{S}|$.

Are the obtained rates optimal in the case of random feedback?

hannan consistency

Our strategies are *Hannan consistent* whenever

$$\max_{\mathbf{p}} \rho(\mathbf{p}, \mathbf{H}(\cdot, \mathbf{q})) = \max_{\mathbf{i}=1,\ldots,\mathbf{N}} \mathbf{r}(\mathbf{i}, \mathbf{q}) \ .$$

Piccolboni and Schindelhauer (2001), and Cesa-Bianchi, Lugosi, and Stoltz (2006) investigate this special case for deterministic feedback.

For example, Hannan consistency is possible if there is a matrix $\mathbf{K}$ such that $\mathbf{R} = \mathbf{KH}$.

We have a different sufficient condition (independent of the rewards):
If $\mathbf{H}$ doesn't have identical columns, then for all $\mathbf{q}$,

$$\max_{\mathbf{p}} \rho(\mathbf{p}, \mathbf{H}(\cdot, \mathbf{q})) = \max_{\mathbf{i}=1,\ldots,\mathbf{N}} \mathbf{r}(\mathbf{i}, \mathbf{q}) \ .$$