Topics on strategic learning

Sylvain Sorin

IMJ-PRG
Université P. et M. Curie - Paris 6
IMJ-PRG
sylvain.sorin@imj-prg.fr

Stochastic Methods in Game Theory
National University of Singapore
Institute for Mathematical Sciences
November 16, 2015

Topics on strategic learning II:

Complements and extensions

1. Conditional expectation

Recall that the total regret at stage n that the player wants to control is of the form:

$$\sum_{m=1}^{n} U_m^k - \omega_m, \quad k \in K$$

where $\omega_m = U_m^{k_m}$ is the random payoff at stage m.
Let $x_m \in \Delta(K)$ be the strategy of the player at stage m, then

$$\mathsf{E}(\omega_m | h_{m-1}) = \langle U_m, x_m \rangle$$

so that $\omega_m - \langle U_m, x_m \rangle$ is a bounded martingale difference.
Hoeffding-Azuma's concentration inequality for a process $\{Z_n\}$ of martingale differences with $|Z_n| \leq L$ states that:

$$\mathsf{P}\{|\bar{Z}_n| \geq \varepsilon\} \leq 2\exp(-\frac{n\,\varepsilon^2}{2L^2})$$

Hence the average difference between the payoff and its conditional expectation is controlled.

Thus we will study quantities of the form:

$$\sum_{m=1}^{n} U_m^k - \langle U_m, x_m \rangle, \quad k \in K.$$

or equivalently, because of the linearity:

$$\sum_{m=1}^{n} \langle U_m, x \rangle - \langle U_m, x_m \rangle, \quad x \in \Delta(K).$$

Similarly the internal no-regret condition becomes:

$$\sum_{m=1}^{n} x_m^i [U_m^j - U_m^i] \leq o(n), \quad \forall i,j \in K.$$

2. Procedures in law

Assume that the actual move k_n is not observed and define a pseudo-process $\tilde{R}$ defined through the conditional expected regret:

$$R_n = U_n - \omega_n \mathbf{1}, \qquad \tilde{R}_n = U_n - \langle U_n, x_n \rangle \mathbf{1}$$

and introduce the associated strategy $\tilde{\sigma}$.
Then consistency holds both for the pseudo and the realized processes under $\tilde{\sigma}$.

2. Procedures in law

Assume that the actual move k_n is not observed and define a pseudo-process $\tilde{R}$ defined through the conditional expected regret:

$$R_n = U_n - \omega_n \mathbf{1}, \qquad \tilde{R}_n = U_n - \langle U_n, x_n \rangle \mathbf{1}$$

and introduce the associated strategy $\tilde{\sigma}$.
Then consistency holds both for the pseudo and the realized processes under $\tilde{\sigma}$.

3. Experts and generalized consistency

Experts

External consistency can be considered as a robustness property of σ facing a given finite family of “external” experts using procedures $\phi \in \Phi$:

$$\lim \frac{1}{n}[\sum_{m=0}^{n} \langle \phi_m - x_m, U_m \rangle]^+ = 0, \quad \forall \phi \in \Phi.$$

The typical case corresponds to a constant choice : $\phi = k$ and $\Phi = K$.

In general “k” will be the (random) move of expert k, that the player follows with probability x_m^k at stage m.

U_m^k is then the payoff to expert k at stage m.

Internal consistency corresponds to experts adjusting their behavior to the one of the predictor.

From external to internal consistency

Stoltz and Lugosi (2005)
Consider a family $\psi^{ij}, (i,j) \in K \times K$ of experts and θ an algorithm that satisfies external consistency with respect to this family.
Define σ inductively as follows.
Given some element $p \in \Delta(K)$, let $p(ij)$ be the vector obtained by adding p^i to the j^{th} component of p.
Let $q_{n+1}(p)$ be the distribution induced by θ at stage $n+1$ given the history h_n and the behavior $\psi^{ij}(h_n) = p(ij)$ of the experts.
Assume that the map $p \mapsto q_{n+1}(p)$ is continuous and let $\bar{p}_{n+1}$ be a fixed point which defines $\sigma(h_n) = x_{n+1}$.

The fact that σ is an incarnation of θ implies that it performs well facing any ψ^{ij} hence

$$[\sum_{m=0}^{n}\langle\psi_m^{ij}-x_m, U_m\rangle] \leq o(n), \qquad \forall i,j$$

which is

$$[\sum_{m=0}^{n}\langle\bar{p}(ij)_m-\bar{p}_m, U_m\rangle] \leq o(n), \qquad \forall i,j$$

hence

$$[\sum_{m=0}^{n}\bar{p}_m^i(U_m^j-U_m^i)] \leq o(n), \qquad \forall i,j$$

and this is the internal consistency condition.

Blum and Mansour (2007)
Consider K parallel algorithms $\{\phi[k]\}$ having no external regret, that generates each a (row) vector $q[k] \in \Delta(K)$ then define σ by the invariant measure p with

$$p = pQ.$$

Given the outcome $U \in \mathbb{R}^K$, add $p^k U$ to the entry of algorithm $\phi[k]$. Expressing the fact that $\phi[k]$ satisfies no external regret gives, at stage m, for all $j \in K$

$$[\sum_{m=0}^{n} p_m^k U_m^j - \langle q[k]_m, p_m^k U_m \rangle] \leq o(n)$$

Note that $\sum_k \langle q[k]_m, p_m^k U_m \rangle = \sum_k \langle p_m^k q[k]_m, U_m \rangle = \langle p_m, U_m \rangle$, hence by summing over k, for any function $M : K \mapsto K$, corresponding to a perturbation of σ with $j = M(k)$ the difference between the performances of σ_M and σ will satisfy as well

$$[\sum_{m=0}^{n} \sum_{k} p_m^k U_m^{M(k)} - \langle p_m, U_m \rangle] \leq o(n).$$

This is the internal consistency for “swap experts”.

Large range

Blum and Mansour (2007); Cesa-Bianchi and Lugosi (2006); Lehrer (2003).

Consider an even larger set of experts that are allowed (in addition to be adapted to the past history) to choose their actions and to be active as a function of the choice of the predictor.

Explicitly every expert $s \in S$ (finite) is characterized, at stage m, conditional to the past, by :

- a choice function $f^s_m : K \to K$
- an activity function $\tau^s_m : K \to [0,1]$.

Given a predictor ϕ which prediction at stage m has a law p_m the regret facing s is :

$$r^s_m = \sum_k p^k_m \tau^s_m(k)[U^{f^s_m(k)}_m - U^k_m]$$

We assume that the functions f^s, τ^s are known by the predictor. Then there exists a consistent procedure.

4. Bandit framework

This is the case where given the move k and the vector U the only information to the predictor is the realization $\omega = U^k$ (the vector U is not announced).
Define the pseudo regret vector at each stage n by:

$$\hat{U}^k_n = \frac{\omega_n}{\sigma^k_n} \mathbf{1}_{\{k_n = k\}}$$

and note that it is an unbiased estimator of the true regret.
To keep the outcome bounded one may have to perturb the strategy and bt same asymptotic properties hold.
(Auer, Cesa-Bianchi, Freund, Shapire, 2002)

For recent advances, see Bubeck and Cesa-Bianchi (2012), chapter 5.

5.Link with on-line convex optimization

Cesa-Bianchi N. and G. Lugosi (2006) Chapter 11
Bubek S. (2011) Introduction to online optimization.
Hazan E. (2009) The convex optimization approach to regret minimization.
Rakhlin A. (2009) Lecture notes on online learning.
Shalev-Shwartz S. (2011) Online learning and online convex optimization, *Foundations and Trends in Machine Learning*, **4**, 107-194.

Recall that we are interested in upper bounds for quantities of the form:

$$\sum_{m=1}^{n} [\langle U_m, x\rangle - \langle U_m, x_m\rangle], \quad x \in \Delta(K).$$

where $\{U_m\}$ is an unknown process and $\{x_m\}$ the adapted procedure.

One can extend the framework to the case where: - X is a convex compact subset of some euclidean set $\mathbb{R}^d$ and
- $\{f_m\}$ is a collection of L-Lipschitz convex functions (that corresponds to a cost) defined on X.

The quantity to control is now:

$$A_n = \sum_{m=1}^{n} [f_m(x_m) - f_m(x)], \quad x \in X.$$

By convexity:

$$f_m(x) - f_m(x_m) \geq \langle \nabla f_m(x_m), x - x_m \rangle$$

so that:

$$A_n \leq \sum_{m=1}^{n} \langle -\nabla f_m(x_m), x - x_m \rangle$$

which corresponds to the previous case with $U_m = -\nabla f_m(x_m)$

Recall that the basic gradient equation for a (fixed) convex function:

$$\dot{x}_t = -\nabla f(x_t)$$

leads to two discrete time algorithms, given a sequence of step sizes $\{a_n\}$.
Prox (proximal algorithm) corresponds to x_{n+1} minimizing

$$f(x) + \frac{1}{2a_n}\|x - x_n\|^2$$

which is

$$x_{n+1} - x_n = -a_n \nabla f(x_{n+1})$$

Euler is given by

$$x_{n+1} - x_n = -a_n \nabla f(x_n)$$

Note that Euler corresponds to Prox applied to the linearization of f near x_n: $f(x)$ is replaced by $f(x_n) + \langle \nabla f(x_n), x - x_n \rangle$

When considering the problem with constraints

$$\min f(x);\ x \in X$$

the previous equation corresponds to the projected gradient (Polyack)

$$x_{n+1} = \Pi_X(x_n - a_n \nabla f(x_n))$$

where Π_X is the orthogonal projection operator.
This amounts to replace the penalization $\|x - x_n\|^2$ by $\|x - x_n\|^2 \chi_X$ where χ_X is the indicator function of X.
More generally one can use a penalization generated by a distance given by a Bregman function (Beck and Teboulle (2003)) $B_h(x,y) = h(x) - h(y) - \langle x - y, \nabla h(y) \rangle$.
Then x_{n+1} minimizes/

$$\langle x, \nabla f(x_n) \rangle + \frac{1}{a_n} B_h(x, x_n)$$

so that: $\nabla f(x_n) + \frac{1}{a_n}[\nabla h(x_{n+1}) - \nabla h(x_n)] = 0$
(mirror descent) leading either to projection (if ∇h is defined on all X) or interior methods (if $\|\nabla h(x_t)\| \to +\infty$ whenever $x_t \to x \in \partial X$).

A first algorithm, due to Zinkevich (2003), extends the projected gradient to our framework by changing the function to minimize at each step:

$$x_{n+1} = \Pi_X(x_n - a_n \nabla f_n(x_n)).$$

Define $g_n = \nabla f_n(x_n)$ and $y_{n+1} = x_n - a_n g_n$ so that for any $x \in X$:

$$\|y_{n+1} - x\|^2 = \|x_n - x\|^2 - 2a_n \langle x_n - x, g_n \rangle + a_n^2 \|g_n\|^2$$

Since $x_{n+1} = \Pi_X(y_{n+1})$:

$$\|x_{n+1} - x\|^2 \leq \|x_n - x\|^2 - 2a_n \langle x_n - x, g_n \rangle + a_n^2 \|g_n\|^2$$

which gives:

$$\langle x_n - x, g_n \rangle \leq \frac{1}{2a_n} [\|x_n - x\|^2 - \|x_{n+1} - x\|^2] + \frac{a_n}{2} L^2.$$

By summing we obtain:

$$A_N = \sum_{n=1}^{N} \langle x_n - x, g_n \rangle \leq \frac{1}{2a_1} \|x_1 - x\|^2 - \frac{1}{2a_N} \|x_{N+1} - x\|^2$$

$$+ \sum_{n=2}^{N} [\frac{1}{2a_n} - \frac{1}{2a_{n-1}}] \|x_n - x\|^2 + \sum_{n=1}^{N} \frac{a_n}{2} L^2$$

hence with $m(X)$ being the diameter of X:

$$A_N \leq m(X)^2 + \frac{L^2}{2} \sum_{n=1}^{N} a_n$$

and the choice of $a_n = n^{-1/2}$ leads to:

$$A_N \leq \sqrt{N}(m(X)^2/2 + L^2).$$

Alternatively a constant step size a would give:

$$A_N \leq (m(X)^2/2 + L^2)[\frac{1}{a} + aN]$$

hence the choice $a[N] = N^{-1/2}$. Use then the “doubling trick”.

This algorithm was of the form $x_{n+1} = T_n(x_n, g_n)$.

Other procedures are based on an operator of the kind:

$$x = \text{ argmin}[\eta\langle x, y\rangle + \rho(x)]$$

where ρ is a bounded smooth regularization function adapted to X, η is a positive parameter, and y will be the state variable. We will write $x = S(y;\eta)$. Note that one has:

$$x = \text{ argmax}[\langle x, -\eta y\rangle - \rho(x)]$$

so that if ρ is convex l.s.c., $x \in \nabla\rho^*(-\eta y)$, where ρ^* is the Fenchel conjugate of ρ.
The procedure will be typically:
$x_{n+1} = S(y_n;\eta_n)$ and
$y_{n+1} = y_n - \nabla f_n(x_n)$.

Example 1

$X = \Delta(K)$, $\rho(x) = \sum x^k \log(x^k)$ (entropy) so that $x = S(y; \eta)$ is given by

$$x^k \div \exp(-\eta y^k)$$

which corresponds to the logit map.

$x_{n+1} = S(y_n; \eta)$ is the exponential weight algorithm:

$$x^k_{n+1} = \frac{\exp[-\sum_{m=1}^n \nabla f_m^k(x_m)]}{\sum_{\ell \in K} \exp[-\sum_{m=1}^n \nabla f_m^\ell(x_m)]}$$

and it satisfies:

$$A_N \leq L[\frac{1}{\eta} + \eta N]$$

so that again $\eta = N^{-1/2}$ will give a rate of convergence of $N^{-1/2}$.

Example 2

$\eta_n = \frac{1}{\varepsilon n}$, $x_{n+1} = S(y_n; \eta_n)$ is an ε smooth payoff best reply to $\bar{y}_n$, that remains bounded (smooth fictitious play).

The recursive equation is now:

$$x_{n+1} = S(\bar{y}_n; \varepsilon^{-1})$$

$$\bar{y}_{n+1} - \bar{y}_n = \frac{1}{n+1}[-\nabla f_n(x_n) - \bar{y}_n].$$

6. Imperfect monitoring

The model

Consider a finite zero-sum two person repeated game defined by G from $I \times J$ to $\mathbb{R}$. In addition there is a finite signal set S and a map M from $I \times J$ to $\Delta(S)$.
At each stage n, given a profile of moves (i_n, j_n), a signal s_n with law $M(i_n, j_n)$ is sent to player 1 and this is his only information.
Player 2 is Nature and knows the all history.
Given $y \in Y = \Delta(J)$, let $M(i,y) = \sum_j y^j M(i,j)$ be the linear extension and denote by $m(y) \in \Delta(S)^I = \{\ M(i,y), i \in I\}$ be the "flag" induced by y.
This is the maximal information that player 1 can obtain if player 2 uses y i.i.d..
This model appears in repeated games (Mertens, Sorin and Zamir, 1994) and the analysis of external regret in this framework was done by Rustichini (1999).

Given a n-stage play the average flag is $\bar{\mu}_n$, where $\mu_r = m(j_r)$ (hence also $m(\bar{y}_n)$) and the evaluation of player 1 is $d(\bar{\mu}_n)$ where:

$$d(\mu) = \max_{x \in \Delta(I)} \min_{y \in \Delta(J); m(y) = \mu} G(x,y).$$

Note that in general best replies are not pure.
The external regret is then $r_n = d(\bar{\mu}_n) - \bar{G}_n$.

Internal regret

Cesa-Bianchi, Lugosi and Stoltz (2006), Lehrer and Solan (2007), Lugosi, Mannor and Stoltz (2008), Perchet (2009)
To specify a notion of internal consistency we use the regularity of the model to define for each $\varepsilon > 0$ a finite discretization $(\mu[\ell], x[\ell]; \ell \in L)$ such that there exists $\delta > 0, \eta > 0$ with:
- the set of flags is covered by balls $B(\mu(\ell), \delta)$
- for any $\mu \in B(\mu(\ell), \delta)$ and $x \in B(x(\ell), \eta)$, x is a $\varepsilon-$best reply to μ for the evaluation $d(\mu)$.

One can now introduce the vector of internal regret.
Let $A_n[\ell]$ be the set of stages before n where player 1 uses $x[\ell]$ and $N_n[\ell]$ its cardinality. $\bar{\mu}_n[\ell]$ resp. $\bar{G}_n[\ell]$, are the corresponding average flag resp. payoff. Then:

$$R_n[\ell] = d(\bar{\mu}_n[\ell]) - \bar{G}_n[\ell], \qquad \ell \in L$$

and define $\varepsilon-$internal consistency as:

$$\limsup_{n \to +\infty} \frac{N_n[\ell]}{n} [R_n[\ell] - \varepsilon]^+ \to 0, \qquad \forall \ell \in L.$$

The main result in this framework is the existence of $\varepsilon-$internal consistent strategies.

Perchet (2009)
Assume first that player 1 is informed of the vector of signals (indexed by I) at each. He can use a calibrated strategy associated to L such that at a stage of type ℓ he “predicts ” $\mu[\ell]$ and plays $x[\ell]$. Then asymptotically on these stages (if their frequency is large enough) his prediction will be correct, his average moves closed to $x[\ell]$ hence the average regret $R_n[\ell]$ small.
To reduce to the previous situation one constructs an estimator of the flag via a pertubation of the strategy (like in the bandit framework).

Perchet V. (2011b)
Using a specfic discretization trough Laguerre diagrams allows to get a speed of convergence of $O(n^{-1/3})$ which is optimal (compared to $\varepsilon + O(n^{-1/2})$).

Approachability

Perchet V. (2011a)
The framework is as above except that G is now from $I \times J$ to $\mathbb{R}^d$. $G(x,y)$ is the multilinear extension to $X = \Delta(I) \times Y = \Delta(J)$.
Let

$$P(x,\mu) = \{G(x,y); m(y) = \mu, y \in Y\} \subset \mathbb{R}^d$$

be the set of payoffs compatible with the strategy $x \in X$ and the flag μ.

Proposition

A closed convex set $C \subset \mathbb{R}^d$ is approachable (by player 1) if and only if

$$\forall \mu \in m(Y), \exists x \in X \qquad \textit{such that} \qquad P(x,\mu) \subset C.$$

Note that this is exactly Blackwell's condition in the full monitoring case.

Proof

Assume that the condition holds.
Then for each $\varepsilon > 0$ one constructs as above a finite family $\{\mu[\ell], x[\ell], \ell \in L\}$ with $P(x[\ell], \mu[\ell]) \subset C$.
A calibrated strategy associated to this set L, such that $x[\ell]$ is played when $\mu[\ell]$ is predicted, will induce on average, on stages of type ℓ, a payoff near C.
One uses the convexity of C to deduce approchability.
Finally if there exists a signal μ_0 such that

$$\forall x \in X, \exists y \in Y, m(y) = \mu_0 \qquad \text{and} \qquad G(x,y) \notin C$$

one can assume $d(G(x,y), C) \geq \delta > 0$ by compactness.

Given σ strategy of player 1 in the n-stage game let

$$z_n = \mathsf{E}_{\sigma,\mu_0}\,[\bar{i}_n]$$

be the expectation of the average move of player 1 facing signals with distribution μ_0 at each stage. Then, let τ be $y(z_n)$ i.i.d. and by convexity

$$\mathsf{E}_{\sigma,\tau}[d(\bar{g}_n, C)] \geq d(G(z_n, y(z_n)), C) \geq \delta > 0$$

■

In addition there are convex sets that are neither approachable nor excludable.
Extensions to games with payoff correspondence: Mannor, Perchet and Stoltz (2014).