Advanced Course in Machine Learning Spring 2010

Games Against Nature

Handouts are jointly prepared by Shie Mannor and Shai Shalev-Shwartz

In the previous lectures we talked about experts in different setups and analyzed the regret of the algorithm by comparing its performance to the performance of the best *fixed* experts (and later the best shifting expert). In this lecture we consider the game theory connection and present games against Nature. Along the way, we present one of the most common tools to analyze prediction problems: approachability theory.

The setup in today's lecture is that of full information. The next lecture will be devoted to the partial information setup. We start from a more general model for the game and then show how to apply it to different online learning setups.

1 The Model

The model is comprised of a single player playing against "Nature." The game is repeated in time, and at stage t the decision maker has to choose an action $a_t \in A$ and Nature chooses (simultaneously) an action $b_t \in B$. As a result the decision maker obtains a reward $r_t \approx R(a_t, b_t)$ (that is, the reward can be stochastic: we will only need finite second moments). The game continues ad infinitum. We let the average reward be denoted by

$$\hat{r}_t = \frac{1}{t}\sum_{\tau=1}^{t} r_\tau.$$

Note: There is no reward for Nature, therefore this is not a game in the standard sense of the word (or, one can say this is a zero-sum game).

The decision maker keeps track of the rewards and of Nature's actions. We consider the empirical frequency of Nature's actions as:

$$q_t(b) = \frac{1}{t}\sum_{\tau=1}^{t} 1\{b_t = b\}$$

and note that $q_t \in \Delta(B)$, the set of distributions over B.

1.1 The stationary case

If Nature is stationary (i.e., the actions are generated from an IID source q^*) then:

$$q_t \to q^* \qquad \text{a.s.}$$

(In fact, we have exponentially fast convergence: $\Pr(\|q_t - q^*\| > \epsilon) \le C\exp(-C' t\epsilon^2)$.) In that case, one can hope to obtain a reward as high as the best response reward:

$$r^*(q) = \max_p \sum_{a,b} q(a)p(b)r(a,b) = \max_a \sum_b q(b)r(a,b).$$

By obtaining we mean:

$$\hat{r}_t \to r^*(q^*) \qquad \text{a.s.}$$

Here is a simple fictitious play algorithm that obtains that:

1. Observe b_t and for an estimate:
$$q_t = \frac{1}{t}\sum_{\tau=1}^{t} 1\{b_t = b\}.$$
2. Play $a_t \in \arg\max r(a, q_t)$.

This algorithm is also based on the celebrated certainty equivalence scheme.

Theorem 1 *The Fictitious Play algorithm satisfies that $\hat{r}_t \to r^*(q^*)$ a.s.*

But what happens if Nature is *not* stationary?

1.2 Arbitrary source

Suppose now that the sequence $b_1, b_2, \ldots$ is generated by an arbitrary process. Arbitrary here means not necessarily stochastic. Clearly, we cannot assume that q_t converges. Our objective of having the average reward converge to $r^*(q^*)$ is not well defined anymore since q^* may not exist.

We can define the average regret as:
$$R_t = r^*(q_t) - \hat{r}_t.$$
This is a *random variable*. Randomness is determined by randomness in the algorithm.

The basic questions is therefore: Can we find an algorithm such that
$$\limsup R_t \leq 0 \qquad \text{a.s. ?}$$
If such an algorithm exists we call it 0-regret (we will later call such an algorithm 0 external regret, but this is sufficient for now). This is, of course, the same notion from the previous two lectures where we consider the average regret as opposed to the cumulative regret.

Nature models.

1. Oblivious. Nature writes down the sequence of $b_1, b_2, \ldots$ at time 0 (not disclosing them).
2. Non-oblivious. Nature is adversarial and it tries to maximize the regret. Nature may even be aware of any randomization the decision maker does (but not the value of private coin tosses).

Observations:

1. An non-oblivious opponent is a very strong model: it encompasses a worst case view on disturbances in many systems and it generalizes play against an adversary.
2. Fictitious play would fail since randomization is needed. Fictitious play is called here "follow the leader" (FtL).
3. If the leader does not change (asymptotically), FtL does have 0 regret.

More interestingly, as long as there are not many switches, FtL "works." More precisely, we say that FtL switches from action a to a' at time t if $a_{t-1} = a$ and $a_t = a'$. We let the number of switches be N_t. We say that FtL exhibits infrequent switches along a history if for every $\epsilon > 0$ there exists T such that $N_t/t < \epsilon$ for all $t \geq T$.

Theorem 2 *If FtL exhibits infrequent switches along a history it satisfies $\limsup_{t\to\infty} R_t \leq 0$ along that history.*

Proof: Home exercise. (Note that we do not use almost sure quantifiers since clearly FtL is not optimal for every history.)

1.3 A generalized notion of regret

In general, regret can be defined as the difference between the obtained (cumulative reward) and the reward that would have been obtained by the best strategy in a reference set. That is:

$$R_t = \sup_{\text{strategy } \sigma} r(\sigma, \text{history}) - \hat{r}_t,$$

where $r(\sigma, \text{history})$ is an estimate of the average reward if playing σ. This is not always well defined or achievable. In the example above, the set of strategies is simply the set of stationary strategies. One can easily think of other sets of strategies such as the set of strategies that depend on the last observation from Nature. In that case: the set of strategies is identified with $p_t \approx p(a|b_{t-1}) \in \Delta(A)^{|B|}$ and the reward as a function of history is defined as:

$$r(\sigma, \text{history}) = \frac{1}{t} \sum_{\tau=1}^{t} \sum_{a} p(a|b_{t-1}) r(a, b_t),$$

where b_0 is defined is one of the members of B. We observe that this comparison class is richer than the comparison class we considered above which can be identified with $p(a) \in \Delta(A)$. We will show later that there is an asymptotical 0-regret strategy against this particular comparison class.

2 Blackwell's Approachability

We now introduce a useful tool in the analysis of repeated games against Nature called Blackwell's approachability theory.

Let us define a vector-valued two-player game. We call the players P1 and P2 to distinguish them from the decision makers above.

We consider a two player vector-valued repeated game where both P1 and P2 choose actions as before from finite sets A and B. The reward is now a k-dimensional vector, $m(a, b) \in \mathbb{R}^k$. As before, the stage game reward is $m_t \approx m(a_t, b_t)$ (the reward can be a random vector).

The average reward is

$$\hat{m}_t = \frac{1}{t} \sum_{\tau=1}^{t} m_t.$$

P1's task is to approach a *target set* T, namely to ensure convergence of the average reward vector to this set irrespectively of P2's actions. Formally, let $T \subset \mathbb{R}^k$ denote the target set. In the following, d is the Euclidean distance in $\mathbb{R}^k$. The set-to-point distance between a point x and a set T is $d(x, T) = \inf_{y \in T} d(x, y)$. (We let $P_{\pi,\sigma}$ denote the probability measure when P1 plays the policy π and P2 plays policy σ.)

Definition 1 *A policy π^* of P1* approaches *a set $T \subset \mathbb{R}^k$ if*

$$\lim_{n \to \infty} d(\hat{m}_n, T) = 0 \quad P_{\pi^*,\sigma}\text{-a.s., for every } \sigma \in \Sigma\,.$$

A policy $\sigma^* \in \Sigma$ of P2 excludes *a set T if for some $\delta > 0$,*

$$\liminf d(\hat{m}_n, T) > \delta \quad P_{\pi,\sigma^*}\text{-a.s. for every } \pi \in \Pi\,,$$

The policy π^* (σ^*) will be called an approaching (excluding) policy for P1 (P2). A set is approachable if there exists an approaching policy. Noting that approaching a set and its topological closure are the same, we shall henceforth suppose that the set T is closed.

The notion of approachability and excludability assumes uniformity with respect to time (and the strategy of P2 (approachability) or P1 (excludability).

2.1 The projected game

Let u be a unit vector in the reward space $\mathbb{R}^k$. We often consider the *projected game in direction* u as the zero-sum game with the same dynamics as above, and *scalar* rewards $r_n = m_n \cdot u$. Here "$\cdot$" stands for the standard inner product in $\mathbb{R}^k$. Denote this game by $\Gamma(u)$.

2.2 The Basic Approachability Results

For any $x \notin T$, denote by C_x a closest point in T to x, and let u_x be the unit vector in the direction of $C_x - x$, which points from x to the goal set T. The following theorem requires, geometrically, that there exists a (mixed) action $p(x)$ such that the set of all possible (vector-valued) expected rewards is on the other side of the hyperplane supported by C_x in direction u_x.

Theorem 3 *Assume that for every point $x \notin T$ there exists a strategy $p(x)$ such that:*

$$(m(p(x), q) - C_x) \cdot u_x \geq 0\,, \quad \forall q \in \Delta(B)\,. \tag{1}$$

Then T is approachable by P1. An approaching policy is given as follows: If $\hat{m}_n \notin T$, play $p(\hat{m}_n)$, otherwise, play arbitrarily.

Proof Let $y_n = C_{\hat{m}_n}$ and denote by F_n the filtration generated by the history up to time n. We further let $d_n = \|\hat{m}_n - y_n\|$. We want to prove that $d_n \to 0$ a.s.. We have that:

$$\begin{aligned}
\mathbb{E}(d_{n+1}^2|F_n) &= \mathbb{E}(\|\hat{m}_{n+1} - y_{n+1}\|^2 \,\Big|F_n) \\
&\leq \mathbb{E}(\|\hat{m}_{n+1} - y_n\|^2 \,\Big|F_n) \\
&= \mathbb{E}(\|\hat{m}_{n+1} - \hat{m}_n + \hat{m}_n - y_n\|^2 \,\Big|F_n) \\
&= \|\hat{m}_n - y_n\|^2 + \mathbb{E}(\|\hat{m}_{n+1} - \hat{m}_n\|^2|F_n) + 2\mathbb{E}((\hat{m}_n - y_n) \cdot (\hat{m}_{n+1} - \hat{m}_n)|F_n).
\end{aligned}$$

Now, since $\hat{m}_{n+1} - \hat{m}_n = m_{n+1}/(n+1) - \hat{m}_n/(n+1)$ we have that:

$$\mathbb{E}(d_{n+1}^2|F_n) \leq d_n^2 + \frac{C}{n^2} + 2\mathbb{E}((\hat{m}_n - y_n) \cdot (\hat{m}_{n+1} - \hat{m}_n)|F_n).$$

Expanding the last term we obtain:

$$\begin{aligned}
(\hat{m}_n - y_n) \cdot (\hat{m}_{n+1} - \hat{m}_n) &= (\hat{m}_n - y_n) \cdot (m_{n+1}/(n+1) - \hat{m}_n/(n+1)) \\
&= (\hat{m}_n - y_n) \cdot (y_n/n + 1 - \hat{m}_n/(n+1) + m_{n+1}/(n+1) - y_n/(n+1)) \\
&= -d_n^2/(n+1) + \frac{1}{n+1}(\hat{m}_n - y_n) \cdot (m_{n+1}/(n+1) - y_n/(n+1))
\end{aligned}$$

Now, the expected value of the last term is negative so we obtain:

$$\mathbb{E}(d_{n+1}^2|F_n) \leq (1 - \frac{2}{n+1})d_n^2 + \frac{c}{n^2}.$$

It follows by Lemma 1 that $d_n \to 0$ almost surely. □

Remarks:

1. **Convergence Rates.** The convergence rate of the above policy is $O(\sqrt{T})$ and is *independent* of the dimension. The only dependence kicks in through the magnitude of the randomness (the second moment, to be exact).

2. **Complexity.** There are two distinct elements to computing an approaching strategy as in Theorem 3. The first is finding the closest point C_x and the second is solving the projected game. Solving the projected 0-sum game can be easily done using linear programming (or other methods) with polynomial dependence on the number of actions of both players. Finding C_x, however, can be in general a very hard problem as finding the closest point in a non-convex set is NP-hard. There are, however, some easy instances such as the case where T is convex and described in some compact form. In fact, it is enough to assume that a convex T has a separation oracle (i.e., we can query in polytime if a point belongs to T or not).

3. **Is a set approachable?** In general, it is NP-hard even to determine if a point is approachable where hardness here is measured in the dimension (if the dimension is fixed it is not hard to decide if a point is approachable).

4. **The game theory connection.** The above result generalizes the celebrated min-max theorem. To observe that, take a one dimensional problem. In that case the approachable set is the segment $[v, \infty]$.

For *convex* target sets, the condition of the last theorem turns out to be both sufficient and necessary. Moreover, this condition may be expressed in a simpler form, which may be considered as a generalization of the minimax theorem for scalar games. Given a stationary policy $q \in \Delta(B)$ for P2, let $\Phi(A, q) \triangleq \mathrm{co}(\{m(p, q)\}_{p \in \Delta(A)})$, where co is the convex hull operator. The Euclidean unit sphere in $\mathbb{R}^k$ is denoted by $\mathbb{B}^k$. The following theorem is characterizes convex approachable sets in an elegant way.

Theorem 4 *Let T be a closed convex set in $\mathbb{R}^k$.*

(i) T is approachable if and only if $\Phi(A, q) \cap T \neq \emptyset$ for every stationary policy $q \in \Delta(B)$.

(ii) If T is not approachable then it is excludable by P2. In fact, any stationary policy q that violates (i) is an excluding policy.

(iii) T is approachable if and only if $\mathrm{val}\,\Gamma(u) \geq \inf_{m \in T} u \cdot m$ for every $u \in \mathbb{B}^k$, where val *is the value of the (scalar) 0-sum game.*

Condition (i) in Theorem 4 is sometimes very easy to check, as we see below.

3 Back to regret

We are now ready to use approachability for proving we can minimize the regret.

Consider the following vector-valued game. When the decision maker plays a and Nature plays b and a reward r_t is obtained the vector-valued reward is $m_t = (r_t, e_b)$ where e_b is a vector of zeros except for the b-th entry which is one. It holds that:

$$\hat{m}_t = (\hat{r}_t, q_t).$$

Now, define the following target set $T \subseteq \mathbb{R} \times \Delta(B)$:

$$T = \{(r, q) : r \geq r^*(q), q \in \Delta(B)\}.$$

We claim that T is convex. Indeed, it follows that $r^*(q)$ is convex as a maximum of linear functions. The set T is convex as the epigraph of a convex function.

We now claim that T is approachable. By Theorem 4, a necessary and sufficient condition is that $\Phi(A, q) \cap T \neq \emptyset$ for every q. Fix some q and let $p^* \in \Delta(A)$ be a member of the argmax of r, that is: $p^* \in \arg\max r(p, q)$. But this is easy to show since $m(p^*, q) \in \Phi(A, q)$ and $m(p^*, q) \in T$. This means that by using approachability we have that $d(\hat{m}_t, T) \to 0$.

What is left is to argue that approaching T implies that $\hat{r}_t - r^*(q_t) \leq 0$ asymptotically. This holds since r^* is a uniformly continuous function (it is convex, continuous and on a compact domain).

We have thus proved:

Theorem 5 *There exists a strategy that guarantees that*

$$\limsup \hat{r}_t - r^*(q_t) \geq 0 \qquad \textit{a.s.}$$

In fact, we have proved that the convergence rate is $O(\sqrt{T})$.

We now return to the problem where we considered generalized regret. We claim a 0-regret strategy does exist. Indeed, consider the target set of the form:

$$T = \{(r, \pi) \in \mathbb{R} \times \Delta(B^2) : \; r \geq \max_{p \in \Delta(A)^B} \sum_{b,b' \in B} \pi(b, b') p(a|b) r(a, b')\},$$

where we identify p with a conditional probability of choosing an action given the past observation (note that it suffices to choose a pure action). It is easy to see that T is convex as an epigraph of a convex function. Now, we need to define the game: when P1 chooses a, P2 chooses b' and the previous action chosen by P2 was b the reward is a vector whose entries are $r(a, b')$ in the first coordinate and the remaining coordinates are zero except for one at the $b \times B + b'$ coordinate. It remains an easy exercise to show that the set T is approachable. (We note that a slight extension of approachability is needed: see "The Empirical Bayes Envelope and Regret Minimization in Competitive Markov Decision Processes." MOR 28(1):327-345, S. Mannor and N. Shimkin.)

4 Calibration

The definition of calibration and a very easy proof using approachability is provided in the attached note.

A Appendix

Lemma 1 *Assume e_t is a non-negative random variable, measurable according to the sigma algebra F_t ($F_t \subset F_{t+1}$) and that*

$$\mathbb{E}(e_{t+1}|F_t) \leq (1 - d_t)e_t + cd_t^2. \tag{2}$$

Further assume that $\sum_{t=1}^{\infty} d_t = \infty$, $d_t \geq 0$, and that $d_t \to 0$. Then $e_t \to 0$ P-a.s.

Proof First note that by taking the expectation of Eq. (2) we get:

$$\mathbb{E}e_{t+1} \leq (1 - d_t)\mathbb{E}e_t + cd_t^2.$$

According to Bertsekas and Tsitsiklis (Neuro-dynamic programming, page 117) it follows that $\mathbb{E}e_t \to 0$. Since e_t is non-negative it suffices to show that e_t converges. Fix $\epsilon > 0$, let

$$V_t^\epsilon \stackrel{\Delta}{=} \max\{\epsilon, e_t\}.$$

Since $d_t \to 0$ there exists $T(\epsilon)$ such that $cd_t < \epsilon$ for $t > T$. Restrict attention to $t > T(\epsilon)$. If $e_t < \epsilon$ then

$$\mathbb{E}(V_{t+1}^\epsilon|F_t) \leq (1 - d_t)\epsilon + cd_t^2 \leq \epsilon \leq V_t^\epsilon.$$

If $e_t > \epsilon$ we have:

$$\mathbb{E}(V_{t+1}^\epsilon|F_t) \leq (1 - d_t)e_t + d_t e_t \leq V_t^\epsilon.$$

V_t^ϵ is a super-martingale, by a standard convergence argument we get $V_t^\epsilon \to V_\infty^\epsilon$.

By definition $V_t^\epsilon \geq \epsilon$ and therefore $\mathbb{E}V_t^\epsilon \geq \epsilon$. Since $\mathbb{E}\left[\max(X, Y)\right] \leq \mathbb{E}X + \mathbb{E}Y$ it follows that $\mathbb{E}V_t^\epsilon \leq \mathbb{E}e_t + \epsilon$. So that $\mathbb{E}V_\infty^\epsilon = \epsilon$. Now we have a positive random variable, with expectation ϵ which is above ϵ with probability 1. It follows that $V_\infty^\epsilon = \epsilon$.

To summarize, we have shown that for every $\epsilon > 0$ with probability 1:

$$\limsup_{t \to \infty} e_t \leq \limsup_{t \to \infty} V_t^\epsilon = \lim_{t \to \infty} V_t^\epsilon = \epsilon\,.$$

Since ϵ is arbitrary and e_t non-negative it follows that $e_t \to 0$ almost surely.