In this lecture we use Blackwell's approachability theorem to formulate both external and internal regret minimizing algorithms. Our study is based primarily on the algorithms presented by Hart and Mas-Colell [6, 7]; see also [3] for a summary.

Throughout the lecture we consider a finite two-player game, where each player i has a finite pure action set A_i; let $A = \prod_i A_i$, and let $A_{-i} = \prod_{j \neq i} A_j$. We let a_i denote a pure action for player i, and let $s_i \in \Delta(A_i)$ denote a mixed action for player i. We will typically view s_i as a vector in $\mathbb{R}^{A_i}$, with $s_i(a_i)$ equal to the probability that player i places on a_i. We let $\Pi_i(\boldsymbol{a})$ denote the payoff to player i when the composite pure action vector is $\boldsymbol{a}$, and by an abuse of notation also let $\Pi_i(\boldsymbol{s})$ denote the expected payoff to player i when the composite mixed action vector is $\boldsymbol{s}$.

The game is played repeatedly by the players. We let $h^T = (\boldsymbol{a}^0, \ldots, \boldsymbol{a}^{T-1})$ denote the history up to time T. The *external regret* of player i against action s_i after history h^T is:

$$ER_i(h^T; s_i) = \sum_{t=0}^{T-1} \Pi_i(s_i, \boldsymbol{a}_{-i}^t) - \Pi_i(a_i^t, \boldsymbol{a}_{-i}^t).$$

The *internal regret* of player i of action a_i against action a_i' after history h^T is:

$$IR_i(h^T; a_i, a_i') = \sum_{t=0}^{T-1} \mathcal{I}\{a_i^t = a_i\} \left(\Pi_i(a_i', \boldsymbol{a}_{-i}^t) - \Pi_i(a_i, \boldsymbol{a}_{-i}^t)\right).$$

We let $p_i^T \in \Delta(A_i)$ denote the marginal empirical distribution of player i's play up to time T:

$$p_i^T(a_i) = \frac{1}{T} \sum_{t=0}^{T-1} \mathcal{I}\{a_i^t = a_i\}.$$

1 External Regret Minimization

Recall that a strategy for player 1 is external regret minimizing, or Hannan consistent, if regardless of the (possibly history-dependent) strategy of player 2, there holds:

$$\limsup_{T \to \infty} \max_{a_1 \in A_1} \frac{1}{T} ER_1(h^T; a_1) \leq 0.$$

To translate from external regret minimization to the Blackwell approachability setting, we define a game with vector-valued payoffs where the payoff vector to player 1 is negative regret. Formally, define $\hat{\Pi} : A_1 \times A_2 \to \mathbb{R}^{A_1}$ by:

$$\hat{\Pi}(a_1, a_2)(a_1') = \Pi_1(a_1, a_2) - \Pi_1(a_1', a_2).$$

Thus $\hat{\Pi}$ measures the improvement in player 1's payoff by playing a_1 instead of a_1'. The key observation we require is the following:

$$ER_1(h^T; a_1) = -\sum_{t=0}^{T-1} \hat{\Pi}(a_1^t, a_2^t)(a_1).$$

Thus, in the notation of Lecture 13, we have:

$$\hat{\Pi}^T(a_1) = -\frac{1}{T} ER_1(h^T; a_1).$$

Hannan consistency is equivalent to requiring that for all $a_1 \in A_1$, there holds:

$$\liminf_{T\to\infty} \hat{\Pi}^T(a_1) \geq 0.$$

We thus conclude: *there exist Hannan consistent algorithms for player 1 if and only if the nonnegative orthant $S = \{\boldsymbol{u} : u(a_1) \geq 0, a_1 \in A_1\}$ is approachable for player 1 in the zero-sum game with vector-valued payoffs $\hat{\mathbf{\Pi}}$.*

In one direction, we have already established the existence of Hannan consistent algorithms for player 1 (e.g., the multiplicative weights algorithm), so the nonnegative orthant must be approachable. Further, Blackwell's approachability theorem then ensures that any halfspace containing the orthant is also approachable.

More interesting, however, is the use of approachability to construct a Hannan consistent algorithm for player 1. Our approach will be to build the strategy suggested in the proof of Blackwell's theorem (see Lecture 13), by "mixing" optimal strategies that arise from approachability of all the halfspaces containing S. We start by first finding an optimal strategy for the scalar zero-sum game induced by any halfspace containing S. Without loss of generality, we restrict attention to halfspaces of the form:

$$H = \{\boldsymbol{u} : \boldsymbol{V} \cdot \boldsymbol{u} \geq 0\},$$

where the vector $\boldsymbol{V}$ is nonzero, and has all nonnegative components. Such a halfspace has the property that its tangent hyperplane is also tangent to S. (Clearly approachability of all such halfspaces implies approachability of any halfspace containing S.)

To ensure approachability of H, we must find a mixed action s_1 for player 1 such that:

$$\min_{a_2 \in A_2} \boldsymbol{V} \cdot \hat{\mathbf{\Pi}}(s_1, a_2) \geq 0. \tag{1}$$

(Recall that this is the *Blackwell condition*.) From the definition of $\hat{\mathbf{\Pi}}$, the preceding relation holds if and only if, for each $a_2 \in A_2$:

$$\Pi_1(s_1, a_2) \left(\sum_{a_1 \in A_1} V(a_1) \right) \geq \sum_{a_1 \in A_1} V(a_1) \Pi_1(a_1, a_2).$$

If we choose:

$$s_1(a_1) = \frac{V(a_1)}{\sum_{a_1' \in A_1} V(a_1')}, \tag{2}$$

then (1) holds with equality for all $a_2 \in A_2$. (Note the denominator is positive since $\boldsymbol{V} \neq 0$.)

We now use this construction to build the strategy suggested in the proof of Blackwell's approachability theorem. The idea is to project $\hat{\mathbf{\Pi}}^{T-1}$ onto the nonnegative orthant, and then play the optimal action s_1 for the resulting halfspace. Note that $P_S(\hat{\mathbf{\Pi}}^{T-1})(a_1) = [\hat{\Pi}^{T-1}(a_1)]^+$, so:

$$P_S(\hat{\mathbf{\Pi}}^{T-1})(a_1) - \hat{\mathbf{\Pi}}^{T-1}(a_1) = \left[\frac{1}{T} ER_1(h^T; a_1)\right]^+.$$

Thus Blackwell's strategy is as follows. At time 0, player 1 plays any mixed action. At time T, if $\hat{\mathbf{\Pi}}^{T-1} \in S$—i.e., if $ER_1(h^T; a_1) \leq 0$ for all a_1—then player 1 can play according to any mixed action. If $\hat{\mathbf{\Pi}}^{T-1} \notin S$, then player 1 plays the following mixed action s_1^T:

$$s_1^T(a_1) = \frac{\left[ER_1(h^T; a_1)\right]^+}{\sum_{a_1' \in A_1} \left[ER_1(h^T; a_1')\right]^+}, \quad a_1 \in A_1.$$

The preceding expression follows from (2). From the proof of the approachability theorem, we conclude that this strategy for player 1 ensures the average vector payoff approaches the nonnegative orthant; in other words, this is a Hannan consistent algorithm for player 1.

We make two remarks on this algorithm:

1. Notice that the mixed action s_1^T depends on more than just the empirical distribution of player 2's action—it also depends on the past history of *player 1's* play. Thus the algorithm just constructed is *not* a variant of fictitious play.

2. It is possible to provide a finite time bound on the regret of this algorithm, in the spirit of the bounds proven in Lecture 11 for the multiplicative weights algorithm. In particular, it is possible to show that if player 1 uses this algorithm, then:

   $$\mathbb{E}[\max_{a_1 \in A_1} ER_1(h^T; a_1)] \leq O(\sqrt{T|A_1|}).$$

   (See [3] for details.)

2 Internal Regret Minimization

We now consider the same approach as the previous section, but for *internal* regret minimization. Recall that a strategy for player 1 is *internal regret minimizing* if regardless of the (possibly history-dependent) strategy of player 2, there holds for all $a_1, a_1' \in A_1$

$$\limsup_{T \to \infty} \frac{1}{T} IR_1(h^T; a_1, a_1') \leq 0.$$

By analogy with the preceding section, we consider a vector-valued payoff $\hat{\Pi} : A_1 \times A_2 \to \mathbb{R}^{A_1 \times A_1}$ defined as:

$$\hat{\Pi}(a_1, a_2)(a_1', a_1'') = \mathcal{I}\{a_1 = a_1'\}(\Pi(a_1, a_2) - \Pi(a_1'', a_2)).$$

It then follows that:

$$\hat{\Pi}^T(a_1, a_1') = -\frac{1}{T} IR_1(h^T; a_1, a_1').$$

Thus internal regret minimization is equivalent to the requirement that for all $a_1, a_1' \in A_1$, there holds:

$$\liminf_{T\to\infty} \hat{\Pi}^T(a_1, a_1') \geq 0.$$

We conclude that *there exist internal regret minimizing algorithms for player 1 if and only if the nonnegative orthant* $S = \{\boldsymbol{u} : u(a_1, a_1') \geq 0, a_1, a_1' \in A_1\}$ *is approachable for player 1 in the zero-sum game with vector-valued payoffs* $\hat{\boldsymbol{\Pi}}$.

As in the preceding section, we use the strategy constructed in the approachability theorem to present an internal regret minimizing algorithm for player 1. We start by considering approachability of halfspaces of the form:

$$H = \{\boldsymbol{u} : \boldsymbol{V} \cdot \boldsymbol{u} \geq 0\},$$

where $\boldsymbol{V}$ is nonzero, and all components of $\boldsymbol{V}$ are nonnegative. We wish to find a mixed action s_1 for player 1 such that:

$$\min_{a_2 \in A_2} \boldsymbol{V} \cdot \hat{\boldsymbol{\Pi}}(s_1, a_2) \geq 0.$$

Fix $a_2 \in A_2$. Then the preceding expression is equivalent to:

$$\sum_{a_1', a_1'' \in A_1} V(a_1', a_1'') \sum_{a_1 \in A_1} s_1(a_1) \mathcal{I}\{a_1 = a_1'\}(\Pi(a_1, a_2) - \Pi(a_1'', a_2)) \geq 0.$$

Simplifying, the preceding expression is equivalent to the requirement that:

$$\sum_{a_1, a_1' \in A_1} V(a_1, a_1') s_1(a_1) \Pi(a_1, a_2) - \sum_{a_1, a_1' \in A_1} V(a_1', a_1) s_1(a_1') \Pi(a_1, a_2) \geq 0.$$

The preceding expression will hold with equality for all $a_2 \in A_2$ as long as for all $a_1 \in A_1$, there holds:

$$\sum_{a_1' \in A_1} V(a_1, a_1') s_1(a_1) - V(a_1', a_1) s_1(a_1') = 0. \tag{3}$$

Thus approachability of H has been reduced to determining whether there exists a mixed strategy s_1 such that (3) holds. Define the $A_1 \times A_1$ matrix $\boldsymbol{Q}$ as:

$$Q(a_1, a_1') = V(a_1, a_1'), \text{ if } a_1 \neq a_1'; \;\; Q(a_1, a_1) = -\sum_{a_1' \neq a_1} V(a_1, a_1'). \tag{4}$$

Then $\boldsymbol{Q}$ is the rate matrix of a continuous time Markov chain on the finite state space A_1, and such a chain must have at least one invariant distribution, i.e., a distribution s_1 such that $s_1 \boldsymbol{Q} = 0$; such an invariant distribution is also a mixed action satisfying (3). This establishes that H is approachable.

As in the preceding section, we can use this construction together with the strategy of the proof of the approachability theorem to give an internal regret minimizing strategy for player 1. At time

T, we project the average payoff $\hat{\mathbf{\Pi}}^{T-1}$ onto the nonnegative orthant S, and use the optimal strategy suggested by the resulting halfspace. We have:

$$P_S(\hat{\mathbf{\Pi}}^{T-1})(a_1, a_1') - \hat{\mathbf{\Pi}}^{T-1}(a_1, a_1') = \left[\frac{1}{T} IR_1(h^T; a_1, a_1')\right]^+.$$

Thus when $\hat{\mathbf{\Pi}}^{T-1} \notin S$, player 1 plays a mixed action s_1^T that is an invariant distribution for the continuous time Markov chain with rate matrix Q defined as in (4), with:

$$V(a_1, a_1') = \left[\frac{1}{T} IR_1(h^T; a_1, a_1')\right]^+.$$

The approachability theorem then implies that if player 1 plays using this algorithm, the average payoff approaches the nonnegative orthant; in other words, this algorithm is internal regret minimizing. (A similar eigenvector calculation is used by Blum and Mansour to show that any external regret minimizing algorithm can be efficiently "converted" into an internal regret minimizing algorithm; see [1].)

We conclude by reinterpreting the algorithm via a slightly different presentation. Choose a constant $\mu > \sup_{a_1, a_1'} |Q(a_1, a_1')|$, and define:

$$\boldsymbol{P} = \boldsymbol{I} + \boldsymbol{Q}/\mu,$$

where $\boldsymbol{I}$ is the identity matrix. Then $\boldsymbol{P}$ is a stochastic matrix, i.e., all its entries are nonnegative and all its rows sum to one; the former follows by choice of μ, and the latter since all rows of $\boldsymbol{Q}$ sum to zero. Thus $\boldsymbol{P}$ is the transition matrix of a discrete time Markov chain on the finite state space A_1, and further, a mixed action s_1 is an invariant distribution s_1 for this chain if and only if it is an invariant distribution for the continuous time Markov chain with rate matrix $\boldsymbol{Q}$.

If we write the components of $\boldsymbol{P}$ explicitly in terms of the internal regrets, we find:

$$P(a_1, a_1') = \frac{1}{\mu}\left[\frac{1}{T} IR_1(h^T; a_1, a_1')\right]^+, \text{ if } a_1' \neq a_1;$$
$$P(a_1, a_1) = 1 - \sum_{a_1' \neq a_1} P(a_1, a_1').$$

Hart and Mas-Colell view the preceding transition probabilities as a specification for repeated play [6]. In particular, they consider an algorithm for player 1 where $s_1^T(a_1') = P(a_1^{T-1}, a_1')$. Considering the expression above for $P(a_1, a_1')$, we see that this algorithm involves increasing weight on pure actions for which internal regret is high *against the most recently played pure action*. Hart and Mas-Colell refer to this algorithm as "regret matching." (Note that this is *not* the algorithm constructed via Blackwell approachability above, where player 1 plays according to the stationary distribution of the matrix $\boldsymbol{P}$.) Hart and Mas-Colell observe that while regret matching is not internal regret minimizing, if all players play according to the regret matching strategy, then the resulting joint distribution of play converges to the set of correlated equilibria. This is an elegant result in the theory of learning in games, because of the simplicity of regret matching.

Indeed, regret matching might be considered the simplest of the algorithms for which convergence to correlated equilibria is guaranteed. Note, however, that it requires the strong assumption that all players are using the same algorithm.

Additional remarks:

1. As before, observe that the internal regret minimizing algorithm constructed above also has the property that it depends on the entire past history of both players' actions (through the regret vector).

2. It is possible to find finite time bounds for internal regret minimizing algorithms as well. The best of these bounds are $\sqrt{2}$ higher than the corresponding bounds for their external regret minimizing counterparts; thus, for example, the best achievable bound on internal regret (in the general setting) at time T is $\sqrt{T \log |A_1|}$. (Informally, this inflation occurs because the set of "experts" we are checking against is of size $|A_1|^2$, rather than size $|A_1|$; see [3] for details.)

3. Clearly, internal regret minimization requires a more sophisticated algorithmic procedure than external regret minimization; in particular, computing a stationary distribution typically requires an eigenvector calculation.

3 Potential-Based Approachability

We conclude by briefly surveying a generalization of approachability that turns out to be quite powerful; the approach we present here is studied in more detail by Hart and Mas-Colell [7] and Cesa-Bianchi and Lugosi [2].

For definiteness, we fix attention on the external regret minimization setting, though the same constructions can also be applied for internal regret minimization. We define $\hat{\mathbf{\Pi}}$ as in Section 1, and again let S be the nonnegative orthant. Suppose there exists a real-valued function $\Phi(\hat{\mathbf{\Pi}})$ with the following properties:

1. There exists a monotonically increasing, concave function ϕ such that $\Phi(\hat{\mathbf{\Pi}}) = \sum_{a_1 \in A_1} \phi(\hat{\Pi}(a_1))$.

2. For $\hat{\mathbf{\Pi}} \notin S$, there exists a mixed action s_1 such that:

$$\nabla \Phi(\hat{\mathbf{\Pi}}) \cdot \hat{\mathbf{\Pi}}(s_1, a_2) \geq 0, \quad \text{for all } a_2 \in A_2. \tag{5}$$

The idea in these assumptions is that Φ measures the quality of the payoff vector $\hat{\mathbf{\Pi}}$. The first condition ensures that Φ resembles a sum of "utility functions" in each element of $\hat{\mathbf{\Pi}}$. The second condition is a generalization of the Blackwell condition. It requires that if $\hat{\mathbf{\Pi}} \notin S$, then there exists a mixed action of player 1 that guarantees that the resulting payoff to player 1 lies on the same side of the subspace defined by the normal $\nabla \Phi(\hat{\mathbf{\Pi}})$ as the nonnegative orthant. It is straightforward to show that under conditions on the Hessian of Φ, one can recreate a strategy similar to the proof of Blackwell's approachability theorem to ensure that the average payoff converges to the set S almost surely [7, 2].

One example of a potential is provided by:

$$\phi(x) = \begin{cases} -\frac{1}{2}x^2, & \text{if } x \leq 0; \\ 0, & \text{otherwise.} \end{cases}$$

It is straightforward to check that using this potential, we obtain:

$$\nabla\Phi(\hat{\mathbf{\Pi}}^{T-1})(a_1) = \left[\frac{1}{T}ER_1(h^T; a_1)\right]^+.$$

Thus the generalized Blackwell condition (5) is equivalent to the standard Blackwell condition (1), with $V(a_1) = [(1/T)ER_1(h^T; a_1)]^+$.

Another example is provided by considering $\phi(x) = -e^{-x/\varepsilon}$, for some $\varepsilon > 0$. In this case:

$$\nabla\Phi(\hat{\mathbf{\Pi}}^{T-1})(a_1) = \frac{1}{\varepsilon}e^{(1/T)ER_1(h^T;a_1)/\varepsilon}.$$

Following the analysis of Section 1, we see that a mixed action s_1^T satisfying (5) is given by:

$$s_1^T(a_1) = \frac{e^{(1/T)ER_1(h^T;a_1)/\varepsilon}}{\sum_{a_1' \in A_1} e^{(1/T)ER_1(h^T;a_1)/\varepsilon}} = \frac{e^{\Pi_1(a_1, p_2^{T-1})/\varepsilon}}{\sum_{a_1' \in A_1} e^{\Pi_1(a_1', p_2^{T-1})/\varepsilon}}.$$

The last equality follows by multiplying top and bottom by $\exp((1/T)\sum_{t=0}^{T-1}\Pi_1(a_1^t, a_2^t))$.

Thus using the exponential potential, we recover the logistic fictitious play of Fudenberg and Levine [5], or equivalently, the multiplicative weights algorithm of Freund and Schapire [4]. We conclude that the multiplicative weights algorithm can be recovered as a special case of algorithms arising via Blackwell approachability. Note, however, that general stochastic fictitious play algorithms will not emerge as special cases of algorithms constructed via the Blackwell condition; to see this, note that stochastic fictitious play algorithms only involve responses to the empirical distribution of the opponent, while algorithms constructed via approachability generally involve the entire past history of both players' actions (as discussed above).

References

[1] A. Blum and Y. Mansour. From external to internal regret. In *Proceedings of COLT*, pages 621–636, 2005.

[2] N. Cesa-Bianchi and G. Lugosi. Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51:239–261, 2003.

[3] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, United Kingdom, 2004.

[4] Y. Freund and R. Schapire. Adaptive game playing with multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.

[5] D. Fudenberk and D. Levine. Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19:1065–1089, 1995.

[6] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

[7] S. Hart and A. Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98:26–54, 2001.