Notes on (Coarse) Correlated Equilibrium and Swap Regret

Hu Fu

October 17, 2018

These notes introduces the solution concepts of correlated equilibria and coarse correalted equilibria in the context of no-regret learning dynamics played in games, and then present the black-box reduction from external to swap regret, a result due to Blum and Mansour (2007).

1 Correlated and Coarse Correlated Equilibria

We have seen that, when the two players both use no-regret learning algorithms in a zero-sum game, their time-average strategies converge to a Nash equilibrium, and this constitutes an alternative proof of von Neumann's minimax theorem. It is natural to ask what happens in games that are more general.

Consider an n-player game with action spaces $A_1, \cdots, A_n$ and each player i's utility given by $u_i(\vec{a})$ for action profile $\vec{a}$. Without loss of generality, assume $A_i - [k]$ for each i. Suppose at each round each player i uses a no-regret learning algorithm and uses (randomized) strategy s_t^i at time step t. Let $\tilde{s}_t^i$ be the time average strategy up to time step t, i.e., $\tilde{s}_t^i = \sum_{\tau=1}^t s_t^i$. (Recall that each s_t^i is a k-dimensional vector in $\Delta([k])$.)

The no-regret learning algorithm guarantess that, up to time T, for each player i and each deviation $a \in A_i$,

$$\frac{1}{T}\sum_{t=1}^{T} u_i(s_t^i, s_t^{-i}) \geq \frac{1}{T}\sum_{t=1}^{T} u_i(a, s_t^{-i}) - \epsilon(T) = u_i(a, \tilde{s}_T^{-i}) - \epsilon(T),$$

where $\epsilon(T)$ goes to 0 as T grows. We emphasize that the LHS, $\frac{1}{T}\sum_{t=1}^{T} u_i(s_t^i, s_t^{-i})$, is not equal to $u_i(\tilde{s}_T^i, \tilde{s}_T^{-i})$. If it were, then $(\tilde{s}_T^i)_i$ would constitute an approximate Nash equilibrium, which *is* the case for two-player zero-sum games. $u_i(\tilde{s}_T^i, \tilde{s}_T^{-i})$ is the expected (of player i) when all players play their randomized strategies independently, but in the expression $\frac{1}{T}\sum_{t=1}^{T} u_i(s_t^i, s_t^{-i})$, the utility is evaluated by first drawing a time step t and then all players play their time t strategy independently; therefore, overall their actions are *correlated* by the shared time step.

This suggests a solution concept where players' strategies are correlated, possibly by some external device or mediator, in which no party has a beneficial unilateral deviation.

Definition 1. A joint distribution $\boldsymbol{s} \in \Delta(\prod_i A_i)$ is a *coarse correlated Nash equilibrium* if for each player i and each deviation $a \in A_i$,

$$\mathbf{E}_{\boldsymbol{a}\sim\boldsymbol{s}}\left[u_i(a_i, \boldsymbol{a}_{-i})\right] \geq \mathbf{E}_{\boldsymbol{a}\sim\boldsymbol{s}}\left[u_i(a, \boldsymbol{a}_{-i})\right].$$

For $\epsilon > 0$, a joint distribution $\boldsymbol{s} \in \Delta(\prod_i A_i)$ is an *ϵ-approximate coarse correlated Nash equilibrium* if for each player i and each deviation $a \in A_i$,

$$\mathbf{E}_{\boldsymbol{a}\sim\boldsymbol{s}}\left[u_i(a_i, \boldsymbol{a}_{-i})\right] \geq \mathbf{E}_{\boldsymbol{a}\sim\boldsymbol{s}}\left[u_i(a, \boldsymbol{a}_{-i})\right] - \epsilon.$$

The discussion above immediately yields the following observation:

Proposition 1. *In a n-player game where $u_i \in [0,1]$, if all players use no-regret learning algorithms that guarantee an average regret $\epsilon(T)$ after time T, then the distribution D_T defined as follows constitutes an $\epsilon(T)$-approximate coarse correlated Nash equilibrium after time T: first draw t uniformly at random from $[T] = 1, \cdots, T$, then for each i draw action a_i according to s_t^i, where s_t^i is the strategy played by player i at time step t, according to the no-regret learning algorithm. As T grows, any convergent subsequence of $\{D_T\}$ (which must exist) converges to a coarse correlated Nash equilibrium.*

The qualification "coarse" refers to the fact that the deviation is rather restricted: in deliberating a possible unilateral deviation, a player considers only one fixed action no matter what the correlating device or mediator tells the player. It is natural to consider the following stronger type of deviations: whenver the correlating device tells me that I should play action a, I play some a' instead, where a' may *depend* on what a is. In other words, a deviation is a *mapping* from an action to an action, and a joint distribution over actions should be stable only when no unilaterally beneficial mapping exists. The following stronger soltuion concept captures this idea.

Definition 2. A joint distribution $\boldsymbol{s} \in \Delta(\prod_i A_i)$ is a *correlated Nash equilibrium* if for each player i and each deviation mapping $\varphi : A_i \to A_i$,

$$\mathbf{E}_{\boldsymbol{a}\sim\boldsymbol{s}}\left[u_i(a_i, \boldsymbol{a}_{-i})\right] \geq \mathbf{E}_{\boldsymbol{a}\sim\boldsymbol{s}}\left[u_i(\varphi(a_i), \boldsymbol{a}_{-i})\right].$$

Similarly, for $\epsilon > 0$, a joint distribution $\boldsymbol{s} \in \Delta(\prod_i A_i)$ is an *ϵ-approximate coarse correlated Nash equilibrium* if for each player i and each deviation mapping $\varphi : A_i \to A_i$,

$$\mathbf{E}_{\boldsymbol{a}\sim\boldsymbol{s}}\left[u_i(a_i, \boldsymbol{a}_{-i})\right] \geq \mathbf{E}_{\boldsymbol{a}\sim\boldsymbol{s}}\left[u_i(\varphi(a_i), \boldsymbol{a}_{-i})\right] - \epsilon.$$

Exercise 1. Show that every correlated Nash is a Nash, and every coarse correlated Nash is a correlated Nash.

The "diminishing regret" property of no-regret learning algorithms give rise to convergence to coarse correlated Nash, but not to the stronger solution concept of correalted Nash. It has been shown that other dynamics, which bear resemlance to no-regret learning but with some different designs, could guarantee convergence to correlated Nash (e.g. Foster and Vohra, 1993; Fudenberg and Levine, 1999; Hart and Mas-Colell, 2000). In the next section we will present a reduction due to Blum and Mansour (2007), which takes any no-regret learning algorithm as a *black-box* and uses it to design a stronger online learning algorithm whose time-average strategies converge to correlated Nash equilibria.

2 Black-Box Reduction from External Regret to Swap Regret

Suppose we run a certain online algorithms for each player in a repeated game. We'd like that after T time steps, the time-average strategies of the players constitute a $\epsilon(T)$-approximate correlated Nash equilibrium, for $\epsilon(T)$ that tends to 0 and T grows. Let's first see more clearly what property this would require from the algorithm we use.

Definition 3. In an online learning problem where the loss of action $i \in [k]$ at time step t is $\ell_t(i)$ and an algorithm is required to output a distribution s_t over actions at time step t with knowledge of $\ell_1, \cdots, \ell_{t-1}$, the *swap regret* of an algorithm after time step T is

$$\sup_{\ell_1,\cdots,\ell_T,\varphi:[k]\to[k]} \sum_{t=1}^{T}\sum_{i=1}^{k} s_t(i)[\ell_t(i) - \ell_t(\varphi(i))]. \tag{1}$$

The regret we have studied so far is called the *external regret*. Equivalently, it can be similarly defined as in (1) with φ being restricted to being a constant function.

With an argument similar to that for Proposition 1, we can see that algorithms that guarantee low swap regret converge to correlated Nash equilibria.

Proposition 2. *In a n-player game where $u_i \in [0,1]$, if all players use algorithms that guarantee an average swap regret $\epsilon(T)$ after time T, then the distribution D_T defined as follows constitutes an $\epsilon(T)$-approximate correlated Nash equilibrium after time T: first draw t uniformly at random from $[T] = 1, \cdots, T$, then for each i draw action a_i according to s_t^i, where s_t^i is the strategy played by player i at time step t, according to the algorithm. As T grows, any convergent subsequence of $\{D_T\}$ (which must exist) converges to a correlated Nash equilibrium.*

Theorem 1 (Blum and Mansour, 2007)**.** *For the expert setting with k actions, given any online learning algorithm $\mathcal{A}$ that guarantees an external regret of no more than $R(T)$ after T time steps, there is a polynomial-time algorithm which invokes $\mathcal{A}$ and guarantees a swap regret of no more than $kR(T)$ after time step T.*

This is a *black-box* reduction because the algorithm guaranteed only invokes $\mathcal{A}$ without *opening the box*, i.e., without having to know how $\mathcal{A}$ works.

Corollary 1. *If all the players in a game with k actions for each player use the algorithm in Theorem 1 by feeding the Hedge algorithm to the black-box reduction, then after time T, the joint distribution as described in Proposition 1 constitutes an $O(\frac{1}{k\sqrt{T\log k}})$-approximate correlated Nash equilibrium.*

Intuitively, we would ideally run k copies of the no-regret learning algorithm, one for each of the actions, such that whenever action i is played, we use the i-th algorithm to make sure that the regret with respect to any other fixed action is no more than $R(T)$. But this is almost self-contradictory — the algorithm corresponding to action i, by virtue of being no-regret, must output a distribution over actions, and therefore it can not be that we use that algorithm precisely when we play action i. The key idea is that we only need to guarantee that with the same probability we take action i and use the algorithm corresponding to action i. How is this possible? At each step t, each of these k no-regret learning algorithm outputs a distribution over actions; if we choose a distribution y_t over these algorithms (and therefore their outputs) such that the resulting overall distribution over the actions happens to be the same as y_t then we "happen to" be doing the right thing. We formalize the idea below.

Proof of Theorem 1. Let $\mathcal{A}_1, \cdots, \mathcal{A}_k$ be k copies of the given no-regret learning algorithm. At each time step t, we will determine on a distribution $y_t \in \Delta([k])$ and run $\mathcal{A}_i$ with probabilty $y_t(i)$; that is, if $\mathcal{A}_i$ outputs strategy $s_t^i \in \Delta([k])$, then we use strategy $z_t = \sum_i y_t(i)s_t^i$ for time step t. Then, after observing the losses $\ell_t(i)$ for each action i, we feed $\mathcal{A}_i$ with the loss vector $y_t(i)\ell_t$.

The no-regret property of algorithm $\mathcal{A}_i$ gives us that, for any $j \in [k]$,

$$\sum_{t=1}^{T} \langle s_t^i, y_t(i)\ell_t \rangle - \ell_t(j) y_t(i) = \sum_{t=1}^{T} y_t(i)[\langle s_t^i, \ell_t \rangle - \ell_t(j)] \leq R(T),$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Summing them up, we have that, for any mapping $\varphi : [k] \to [k]$,

$$kR(T) \geq \sum_{t=1}^{T} \sum_{i} y_t(i)[\langle s_t^i, \ell_t \rangle - \ell_t(\varphi(i))] = \sum_{t=1}^{T} \sum_{i=1}^{k} z_t(i)\ell_t(i) - y_t(i)\ell_t(\varphi(i)). \tag{2}$$

Recall that at time step t, we play action i with probability $z_t(i)$. Comparing (2) with (1), we see that if we could make $y_t(i) = z_t(i)$ for all i, then (2) says exactly that the swap regret of our algorithm would be bounded by $kR(T)$. Therefore y_t should be the solution to the linear system $\sum_i y_t(i) s_t^i = y_t$. Recall that y_t needs to be a probability distribution. Note that the matrix $[s_t^1, \ldots, s_t^k]$ encodes a Markov chain, with $s_t^i(j)$ representing the probability with which state i transits to state j. Viewed this way, $\sum_i y_t(i) s_t^i$ is the state distribution when we start with distribution y_t and let the chain move one step. Requiring this to be equal to y_t is to require y_t to be a stationary distribution for the Markov chain. It is well known that this always exists. □

References

Blum, A. and Mansour, Y. (2007). From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324.

Foster, D. P. and Vohra, R. V. (1993). A randomization rule for selecting forecasts. *Operations Research*, 41(4):704–709.

Fudenberg, D. and Levine, D. K. (1999). Conditional universal consistency. *Games and Economic Behavior*, 29(1):104 – 130.

Hart, S. and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150.