CSCI699: Topics in Learning & Game Theory

Lecture 6

Lecturer: Shaddin Dughmi *Scribes: Omkar Thakoor & Umang Gupta*

1 No regret learning in zero-sum games

Definition 1. *A 2-player game of complete information is said to be zero-sum if* $U_2(a_1,\ a_2) = -U_1(a_1,\ a_2) \forall$ *action profiles* $(a_1,\ a_2)$

Such a game can be described by an $n \times m$ matrix A, where n, m are the number of actions of P_1, P_2 resp. and we have,

$$A_{ij} = U_1(i,j) \qquad (\Rightarrow U_2(i,j) = -A_{ij}) \tag{1}$$

Example We can represent the well-known Rock-Paper-Scissor game with the following matrix (Table 2 shows the general-sum game representation of this game.)

$$\begin{array}{c} \\ R \\ P \\ S \end{array} \begin{array}{c} \begin{array}{ccc} R & P & S \end{array} \\ \left[\begin{array}{ccc} 0 & -1 & +1 \\ +1 & 0 & -1 \\ -1 & +1 & 0 \end{array}\right] \end{array}$$

Next, let $x \in \Delta_n$, $y \in \Delta_m$ be mixed strategies of P_1, P_2. Then, P_1 plays his action i with probability x_i, and P_2 plays his action j with probability y_j, and thus P_1 gets a payoff A_{ij} with probability $x_i y_j$. Hence, his expected payoff can be written as,

$$U_1(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j A_{ij} = x^T A y \qquad (=> U_2(x,y) = -x^T A y)$$

Recall that, by definition, (x,y) form a Nash Equilibrium iff

$$\begin{aligned} & U_1(x,y) \geq U_1(x',y)\ \forall x' \quad \& \quad U_2(x,y) \geq U_2(x,y')\ \forall y' \\ \Rightarrow \quad & x^T A y \geq x'^T A y\ \forall x' \quad \& \quad x^T A y \leq x^T A y'\ \forall y' \end{aligned}$$

Equivalently, we can reduce it to,

$$x^T A y \geq (Ay)_i\ \forall i \in [n] \quad \& \quad x^T A y \leq (x^T A)_j\ \forall j \in [m]$$

Definition 2. *The maximin value of A is*

$$\max_x \min_y x^T A y = \max_x \min_{j \in [m]} (x^T A)_j$$

Further, we call any $x^* \in \text{argmax}_x \min_y x^T A y$ as P_1's maximin strategy.

Similarly, we also define the following:

Definition 3. *The minimax value of A is*

$$\min_y \max_x x^T A y = \min_y \max_{i \in [n]} (Ay)_i$$

Further, we call any $y^* \in \text{argmin}_x \max_y x^T A y$ as P_2's minimax strategy.

An interesting question that arises, is, "Which is better, moving first or second?". To answer this, we first show the following result, which is also known as the weak 2nd mover's advantage.

Theorem 1. $maximin(A) \leq minimax(A)$

Proof.

$$\begin{aligned}
maximin(A) &= \max_x \min_y x^T A y \\
&= \min_y x^{*T} A y && \text{(By definition of } x^* \text{ (Def. 2))} \\
&\leq x^{*T} A y^* && \text{(By definition of min)} \\
&\leq \max_x x^T A y^* && \text{(By definition of max)} \\
&= \min_y \max_x x^T A y && \text{(By definition of } y^* \text{ (Def. 3))} \\
&= minimax(A)
\end{aligned}$$

□

Next, we prove, that the two values are in fact equal.

Theorem 2 (Minimax Thm)**.** $maximin(A) = minimax(A)$

Before proceeding with the proof of the theorem, note the following consequence of the theorem.

Corollary 3. *$(x^*,\ y^*)$ is a Nash Equilibrium of the simultaneous game.*

Proof. Using Theorem 2, it follows that every inequality in the proof of Theorem 1 is, in fact, an equality. Hence,

$$x^{*T}Ay^* = \max_x x^T Ay^* \quad \& \quad x^{*T}Ay^* = \min_y x^{*T}Ay$$

Thus, x^*, y^* are best responses to each other, making $(x^*,\ y^*)$ a Nash Equilibrium by definition. □

Next, we present the proof of Theorem 2.

Proof. (Minimax Thm)
Consider the following setting.
- Assume that the two players play repeatedly T times where T is large.
- At each step $t = 1, \ldots, T$, they play simultaneously: P_1, P_2 choose x^t, y^t respectively. P_1 sees all the previous iterations before choosing x^t, but not see y^t. Similarly, P_2 sees all the previous iterations before choosing y^t, but not see x^t.
- Utility of P_1 at time t is $x^t Ay^t$, which is also the loss of P_2 at time t.
- Let $U_i^t = (Ay^t)_i$ be the utility of action i for P_1 at time t. Similarly, let $C_j^t = (x^t A)_j$ be the cost of action j for P_2 at time t.
- Each player faces an online learning problem: P_1 must select x^t based only on $U_i^1, \ldots, U_i^{t-1}\ \forall i$. similarly, P_2 must select y^t based only on $C_j^1, \ldots, C_j^{t-1}\ \forall j$.
- Assume each uses a no-external regret algorithm (e.g. Multiplicative Weights).
- Let $\delta(T)$ be bound on external regret.
- Let $v = \frac{1}{T}\sum_{t=1}^{T} x^t Ay^t$

Then, we have,

$$\begin{aligned} v &\geq \max_i \frac{1}{T}\sum_{t=1}^{T}(Ay^t)_i - \delta(T) && \text{(since no-ext regret)} \\ &= \max_i \left(A\frac{\sum_{t=1}^{T} y^t}{T}\right)_i - \delta(T) \\ &\geq \min_y \max_i (Ay)_i - \delta(T) \\ &= minimax(A) - \delta(T) \end{aligned}$$

Similarly, from P_2's perspective, we can get,

$$v \leq maximin(A) + \delta(T)$$

Combining the two inequalities,

$$\begin{aligned} minimax(A) &\leq maximin(A) + 2\delta(T) \\ \Rightarrow minimax(A) &= maximin(A) && \text{(since } \delta(T) \to 0 \text{ as } T \to \infty) \end{aligned}$$

□

As a result of this, the following immediately follows by definition:

Corollary 4. $\left(\sum_{t=1}^{T} \frac{x^t}{T}, \sum_{t=1}^{T} \frac{y^t}{T}\right)$ *is* δ*-NE.*

2 No regret learning in general-sum games

In this section, we discuss no regret learning in general-sum games. We talk about equilibrium criterion — Correlated equilibrium & Coarse correlated equilibrium and we talk about their existence in general sum games. We introduce a new regret measure/benchmark called *swap regret* and an algorithm that has vanishing swap regret.

Recall that the *Games of Complete Information* have :

- N players, $n = 1, \ldots, n$
- Action set A_i for each player i; $A = A_1 \times A_2 \ldots \times A_n$
- utility function — $u_i : A \to [-1, 1]$ for player i; $u_i(a_1, a_2 \ldots, a_n)$ is player's utility when players play $A = (a_1, \ldots, a_n)$

Definition 4. *Correlated equilibrium is a distribution* χ *over* A *such that* $\forall i, \forall j, j' \in A_i$

$$E_{a\sim\chi}[u_i(a)|a_i = j] \geq E_{a\sim\chi}[u_i(j', a_{-i})|a_i = j]$$

In other words, if player i *is recommended an action* j*, he should choose to move* j*. This is similar to saying that* $\forall i, \forall s : A_i \to A_i$ *(swapping function)*

$$E_{a\sim\chi}[u_i(a)] \geq E_{a\sim\chi}[u_i(s(a_i), a_{-i})]$$

Note that knowing his own recommended action player can infer something about other player's move yet he is better off playing the recommended action.

For example, consider the chicken-dare game in which players can choose to chicken-out or dare. Pay-offs are as mentioned in table 2. If players are recommended action profiles (C,D), (D,C) & (C,C) with equal probability, it is a Correlated equilibrium . None of the players would want to deviate from recommended strategy in this case. Consider player 1 if it is recommended to chicken out, it knows other player is going to dare and chicken out with equal probability and so player 1's pay off is better off not deviating. If player 1 is recommended dare, it knows other player is going to chicken out and hence won't deviate.

	Chicken	Dare
Chicken	0, 0	-2, 1
Dare	1, -2	-10, -10

Table 1: Pay-off in Chicken Dare Game

Definition 5. *Coarse correlated equilibrium is a distribution χ such that $\forall i, j$*

$$E_{a\sim\chi}[u_i(a)] \geq E_{a\sim\chi}[u_i(a_j, a_{-i})]$$

Above definition of Coarse correlated equilibrium states that the agent is not going to make any profit by deviating to some other constant/fixed action. Note the subtle difference in definition of Coarse correlated equilibrium from that of Correlated equilibrium . Note that Correlated equilibrium uses a smart swap function. Therefore, Coarse correlated equilibrium is a weaker equilibrium criteria as there is no correlation between swap and the recommended action in general as compared to Correlated equilibrium . Another way to look at it is to compute Correlated equilibrium , we assume players have more knowledge and hence can respond better in which case it becomes difficult to make them obey.

$$DominantStrategy \subseteq NashEq. \subseteq CorrelatedEq. \subseteq CoarseCorrelatedEq.$$

For example, consider the rock-paper-scissor game and the pay off matrix given as in table 2. Choosing all the actions equally likely except the diagonal once is a Coarse correlated equilibrium . To understand why, suppose the player 1 is playing rock, player 2 will respond with either paper or scissor with equal probability, but if he starts responding with paper higher possibility he will incur more loss when player 1 plays scissor (recall player 1 play rock as well scissor with equal probability). Hence, player 2 is not better off playing other strategies.

This is not correlated equilibrium strategy. If player 2 is instructed to play paper. He knows that other player is playing either rock or scissor and player2's average pay off is 0. He can change his move to rock and improve pay off to 1/2 as against to 0 if he plays recommended action. In this case, player2 could exploit knowledge of action recommended to him to improve pay off.

	Rock	Paper	Scissor
Rock	0,0	-1,1	1,-1
Paper	1,-1	-0,0	-1,1
Scissor	-1,-1	1,-1	0,0

Table 2: Pay-off in rock, paper, scissor Game

Definition 6. *δ-correlated equilibrium (or approximate Correlated equilibrium) is a distribution χ over A such that $\forall i, \forall j, j' \in A_i$ & $\delta > 0$*

$$E_{a\sim\chi}[u_i(a)|a_i = j] \geq E_{a\sim\chi}[u_i(j', a_{-i})|a_i = j] - \delta$$

In other words, if player i is recommended action j, he is only better off deviating from his action by a small quantity δ. This is similar to saying that $\forall i, \forall s : A_i \to A_i$ (swapping function)

$$E_{a\sim\chi}[u_i(a)] \geq E_{a\sim\chi}[u_i(s(a_i), a_{-i})] - \delta$$

Definition 7. *δ-coarse correlated equilibrium (or approximate Coarse correlated equilibrium) is a distribution χ such that $\delta > 0$ & $\forall i, j$*

$$E_{a\sim\chi}[u_i(a)] \geq E_{a\sim\chi}[u_i(a_j, a_{-i})] - \delta$$

Note that in both δ-correlated equilibrium & δ-coarse correlated equilibrium inequalities can slack by an arbitrary δ from the Correlated equilibrium & Coarse correlated equilibrium criterion.

2.1 Existence of δ-coarse correlated equilibrium

Theorem 5. *Fix a game of complete information, suppose that player plays the game repeatedly T times & each player uses a vanishing external regret algorithm. Then, the time averaged mixed strategy profiles form an approximate Coarse correlated equilibrium*

Formally,
Let $P_i^t \in \mathcal{D}(A_i)$ be player i's mixed strategy at time t.
Let $\chi^t = P_1^t \times \dots P_n^t$ &
Let $\bar{\chi} = \frac{1}{T}\sum_{i=1}^{T} x^t$ be the time averaged joint action profile distribution,
Then $\bar{\chi}$ is a δ-coarse correlated equilibrium where $\delta(T) \to 0$ as $T \to \infty$

Note that $\bar{\chi}$ can be realized in two steps by sampling time t from uniform $\{1, T\}$, then sampling from χ^t

Proof. Fix player i & assume it uses a vanishing external regret algorithm to choose his action. Vanishing external regret implies that switching to fixed action j in hindsight cannot give more that $\delta(T)$ per time step & on an average $\delta(T) \to 0$ as $T \to \infty$. Recall, Multiplicative weights is one such vanishing external regret algorithm. For a vanishing external regret algorithm, we know that at time T,

$$\frac{1}{T}\sum_t E_{a^t\sim\chi^t}[u_i(a^t)] \geq \frac{1}{T}\sum_t E_{a^t\sim\chi^t}[u_i(a_j^t, a_{-i}^t)] - \delta(T) \quad \text{for any } j$$

Since expectations are linear so we reconsider the above equations as picking some t at random & get action from it. Therefore the above eq. can be approximately written using $\bar{\chi}$ as follows

$$E_{a\sim\bar{\chi}}[u_i(a)] \geq E_{a\sim\bar{\chi}}[u_i(a_j, a_{-i})] - \delta(T)$$

Above relation clearly implies that δ-coarse correlated equilibrium exists in above scenario and it can be achieved if agents use a vanishing external regret algorithm. □

Corollary 6. *δ-coarse correlated equilibrium exists for every $\delta > 0$ and every finite game. Moreover, a vanishing external regret learning agent dynamics arrive at such δ-coarse correlated equilibrium*

2.2 Swap Regret

Definition 8. *Fix an online learning environment, where an adversary chooses cost function at each time , $C_1 \dots C_T : A \to [-1, 1]$. An online learning algorithm which plays a mixed strategy $P_t \in \mathcal{D}(A)$ at time t has swap regret as follows*

$$\textit{swap-regret}_{alg}^t = \frac{1}{T}\Big(\sum_{t=1}^{T} E(C^t(j)) - \min_{s:A\to A}\sum_{t=1}^{T} E(C^t(s(j)))\Big)$$

s is the swap function.

Thus, we say that an algorithm has vanishing swap regret (or no swap regret) if

$$\textit{swap-regret}_{alg}^T \xrightarrow{T\to\infty} 0 \quad \forall \textit{adversaries}, \forall C_t(a)$$

Note that swap-regret criteria is a stronger criteria than the external regret criteria in that swap function is smart, i.e. it can look at the player's action and make the swap depending on the action as against choosing a fixed action in external regret criteria.

The main idea is that :

- Swap-regret benchmark moves around algorithm.
- It imposes some type of local optimality.
- It has local modification of action.

2.3 Existence of δ-correlated equilibrium

Theorem 7. *Fix a game of complete information, suppose that player play the game repeatedly T times & each player uses a vanishing swap regret algorithm. Then, the time averaged mixed strategy profiles forms an approximate Correlated equilibrium .*

Formally,
Let $P_i^t \in \mathcal{D}(A_i)$ *be player i's mixed strategy at time t*
Let $\chi^t = P_1^t \times \dots P_n^t$ *&*
Let $\bar{\chi} = \frac{1}{T}\sum_{i=1}^{T} x^t$ *be the time averaged joint action profile distribution*

Then, $\bar{\chi}$ *is a* δ*-correlated equilibrium where* $\delta(T) \to 0$ *as* $T \to \infty$

Note that $\bar{x}$ can be realized in two steps by sampling uniform $\{1, T\}$ and then sampling from x^t

Proof. Fix player i and use a vanishing swap regret algorithm to choose action. No swap regret implies that swapping the action with a action j in hindsight cannot give more that $\delta(T)$ per time step and on an average $\delta(T) \to 0$ as $T \to \infty$. For a vanishing swap regret algorithm we know that at time T

$$\frac{1}{T}\sum_t E_{a^t \sim \chi^t}[u_i(a^t)] \geq \frac{1}{T}\sum_t E_{a^t \sim \chi^t}[u_i(s(a_i^t), a_{-i}^t)] - \delta(T)$$

where, $s : A \to A$ is swap function

Since expectations are linear so we consider the above equation as picking some t at random & get action from it. There the above eq. can be approximately written using $\bar{\chi}$ as follows

$$E_{a \sim \bar{\chi}}[u_i(a)] \geq E_{a \sim \bar{\chi}}[u_i(a_j, a_{-i})] - \delta(T)$$

Above eq. implies that δ-coarse correlated equilibrium exists in above scenario and it can be achieved if agents use a no swap regret algorithm. □

Corollary 8. δ*-correlated equilibrium exists for every* $\delta > 0$ *and every finite game. Moreover, a no swap regret learning agent dynamics arrive at such* δ*-correlated equilibrium*

So far, we have shown that given an algorithm with vanishing swap regret, δ-correlated equilibrium always exists. Next we try to establish existence of a no swap regret algorithm

2.4 Existence of a no swap regret algorithm

Theorem 9. *In any online learning setting with finitely many actions A (at max n), there is a reduction from no swap regret algorithm to a no external regret algorithm.*

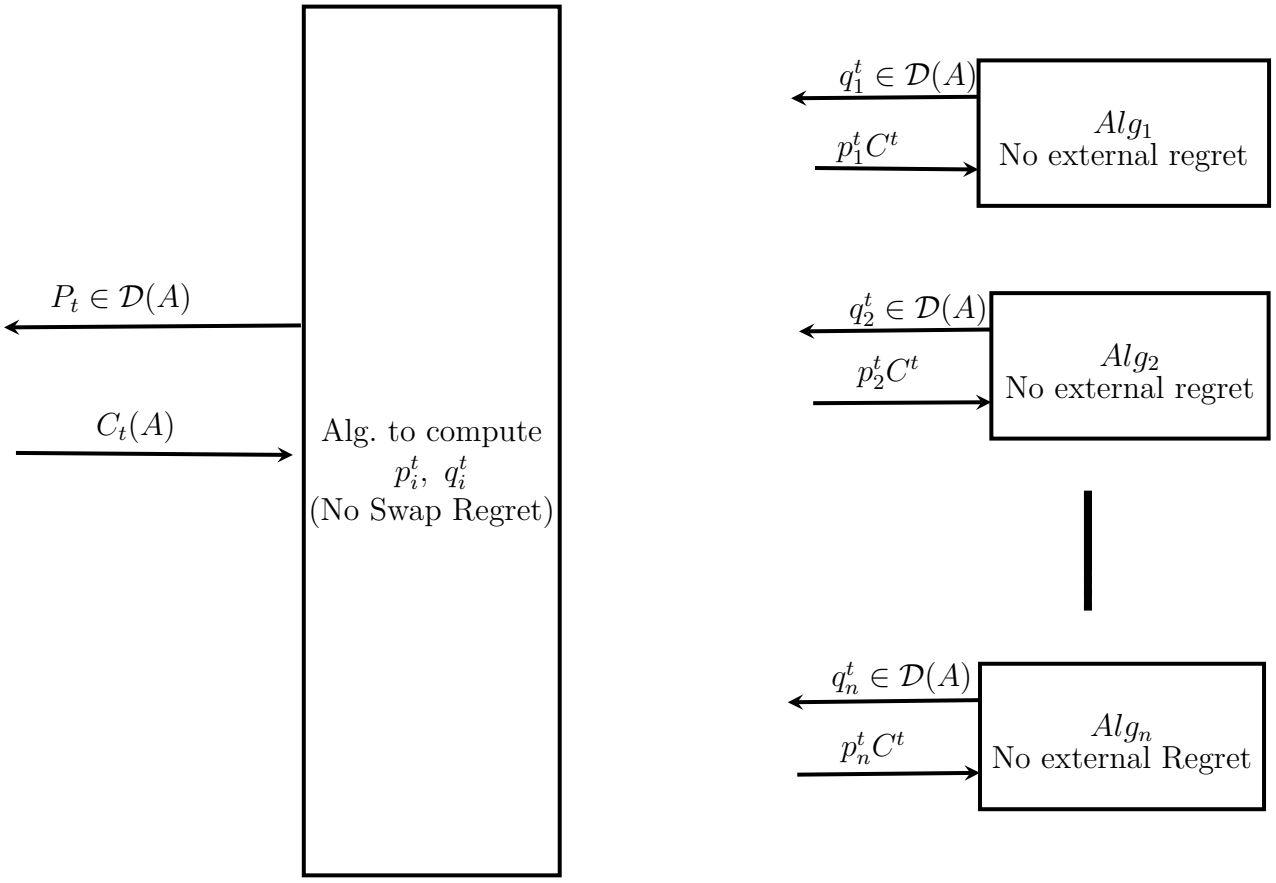

Figure 1: No Swap Regret Algorithm

Proof. The algorithm for the above is as follows:

- Instantiate n copies of vanishing external regret algorithm.
- Each instance of algorithm gets some proportion of cast.
- This is intuitively interpreted in following way — First instance of no external regret algorithm is incharge for swapping action 1 to some other fixed action, and similarly for other instances. See fig. 1

Now, for each time step $t = 1 \dots T$

- Receive $q_1^t \dots q_n^t \in \mathcal{D}(A)$ from $Alg_1 \dots Alg_n$
- Aggregate $q_1^t \dots q_n^t$ to get P^t
- Play according to P^t
- Receive $C^t(A) \in [-1, 1]$
- Compute and give $p_i^t C^t$ to each of $Alg_1 \dots Alg_n$

To achieve no swap regret we need for any $s : A \to A$ for

$$\frac{1}{T}\sum_T E_{a\sim P_t}[C^t(i)] \leq \frac{1}{T}\sum_T E_{a\sim P_t}[C^t(s(i))] + \delta(T)$$

$$\frac{1}{T}\sum_T \sum_i P_i^t C^t(i) \leq \frac{1}{T}\sum_T \sum_i P_i^t C^t(s(i)) + \delta(T) \tag{2}$$

We know $Alg_1 \dots Alg_n$ have vanishing external regret for each Alg_j
Let,

$$k = \operatorname*{argmin}_j \frac{1}{T}\sum_T c^t(j)$$
$$c^t = p_j^t C^t$$

$$\frac{1}{T}\sum_T E_{i\sim q_j^t}[c^t(i)] \leq \frac{1}{T}\sum_T E_{i\sim q_j^t}[c^t(k)] + \delta'(T)$$

$$\frac{1}{T}\sum_T \sum_i q_{ji}^t p_j C^t(i) \leq \frac{1}{T}\sum_T p_j C^t(k) + \delta'(T) \qquad \forall\ j = 1 \dots n \tag{3}$$

We need to prove eq. 2. Since eq. 3 holds for any k, therefore it should also hold for any $s : A \to A$ as it will map to one of the k. We can write,

$$\frac{1}{T}\sum_{T}\sum_{i} q^t_{ji} p_j C^t(i) \leq \frac{1}{T}\sum_{T} p_j C^t(s(i)) + \delta'(T) \qquad \forall\ j = 1 \dots n$$

Adding all eq. from $1 \dots n$.

$$\frac{1}{T}\sum_{T}\sum_{i}\sum_{j} q^t_{ji} p_j C^t(i) \leq \frac{1}{T}\sum_{T}\sum_{j} p_j C^t(s(i)) + \delta'(T) \qquad \forall\ j = 1 \dots n \tag{4}$$

By eq. 4 & eq. 2, we can conclude that algorithm should have $\forall t$

$$\sum_{j} q_{ji} p^t_j = P^t_i = p^t_j \qquad \forall t$$

Assume, P^t_i, p^t_j are vectors and q_{ji} is a matrix of transition probabilities from j to i. We want

$$p^t Q^t = P^t = P^t Q^t \tag{5}$$

We know that — *Every markov chain Q has a stationary dist. $\mathcal{P}$. Moreover such a distribution can be computed efficiently in $\mathcal{O}(\text{No. of States} = n)$.* Therefore, eq. 5 has a solution.

Therefore, we can create a no swap regret algorithm from n instance of no external regret algorithm. This proves the existence of no swap regret algorithm and hence δ-correlated equilibrium

□