CS 6840 Algorithmic Game Theory Oct 23, 2024

Lecture 24: Danger of No Regret Learning

Instructor: Eva Tardos *Scribe: Jiaqi Wang*

1 Stackelberg Game

1.1 Motivating Example

Consider a two-player game where the row player chooses between U and D, and the column player chooses between L and R. The payoff matrix is shown below, with (a, b) representing the utilities for the row and column players, respectively.

	L	R
U	(1, 1)	(3, 0)
D	(0, 0)	(2, 1)

Table 1: Payoff matrix for a two player game

Note that U is the dominant strategy for row player. Therefore, row player always chooses U. Given this, the column player will choose L to maximize his payoff. This implies that (U, L) is the unique Nash, where row player has payoff 1.

We may ask whether the row player can do better by *announcing his choice first*?

For instance, if he announces D, the column player will choose R to maximizes his payoff. (D, R) gives row player payoff 2, making him better off than the Nash.

Even better, if the row player announces a mixed strategy - choosing U with probability $\frac{1}{2} - \epsilon$, and D with probability $\frac{1}{2} + \epsilon$. So long as $\epsilon > 0$, choosing R is the best response, as it maximizes the expected payoff (Choosing L gives the column player an expected payoff: $u_2(U, L) * (\frac{1}{2} - \epsilon) + u_2(D, L) * 0 = \frac{1}{2} - \epsilon$, while choosing R gives him: $u_2(U, R) * (\frac{1}{2} + \epsilon) + u_2(D, R) * 0 = \frac{1}{2} + \epsilon$). For small ϵ, the row player gets a better payoff $2.5 - \epsilon$.

Remark. One may ask whether the row player can be better off announcing D while choosing U instead. In the case of playing against a no regret learner, the column player will learn if the row player cheats. Therefore, we assume that the row player must stay truthful to his announced strategy.

1.2 Stackelberg Equilibrium

In a multi-player setting, the Stackelberg game involves one leader and multiple followers. The leader (assumed to be player i) selects a mix strategy α_i over $s \in S_i$. Observing this, other players play Nash of the resulting game.

More formally, we consider a 2 player game with player 1 being the leader and player 2 being the follower. Let S_1, S_2 (discrete) be their corresponding strategy sets.

The first player chooses a mixed strategy α over S_1, with $\alpha_s \geq 0, \forall s \in S_1$, and $\sum_{s \in S_1} a_s = 1$.

Then the second player chooses his best response z^* to α by maximizing his expected payoff.

$$z^*(\alpha) = \arg\max_{z \in S_2} u_2(\alpha, z) = \arg\max_{z \in S_2} \sum_s \alpha_s u_2(s, z)$$

The utility of the first player is $u_1(\alpha, z^*(\alpha)) = \sum_{s \in S_1} \alpha_s u_1(s, z^*)$. And he will choose the mixed strategy α to maximize this payoff

$$\alpha^* = \arg\max_{\alpha} u_1(\alpha, z^*(\alpha)) = \arg\max_{\alpha} \sum_{s \in S_1} \alpha_s u_1(s, z^*)$$

Definition 1. The *Stackelberg equilibrium* of a game is a pair of strategies α^*, a mixed strategy over S_1, and $z^* \in S_2$ that maximizes $u_1(\alpha, z^*(\alpha))$ under the constraint that $z^*(\alpha) = \arg\max_{z \in S_2} u_2(\alpha, z)$ is a *best response* to α. We call the value $u_1(\alpha, b)$ the *Stackelberg value* of the game.

Question: Would the Stackelberg game has the same payoff as a Nash in a zero sum game? Alternatively speaking, could we do better than announcing a NE in zero sum game?

Lemma 1. *In a 2 person zero sum game, Stackelberg value = Nash value.*

Proof. Let A be the payoff matrix, and x, y be the mixed strategies of two players respectively. Player 1 seeks to $\max_x x^T A y$, and player 2 seeks to $\min_y x^T A y$.

Minmax theorem states that

$$\max_x \min_y x^T A y = \min_y \max_x x^T A y = \text{Nash value}$$

Notice that $\max_x \min_y x^T A y$ fits the Stackelberg game where player 1 announces its strategy x first, and player 2 chooses his best respond y. Therefore, Stackelberg value $= \max_x \min_y x^T A y =$ Nash value. i.e., by announcing your strategy first, the best payoff the leader can achieve is the Nash payoff.

Remark. This result does not hold for general games. One reason it holds here is that two players have the common payoff function A.

■

2 Playing against a No Regret Learner

Knowing that one of the players (the learner) is playing a no-regret learning strategy, what is the optimal gameplay for the other player (the leader)?

2.1 No worse than Stackelberg Value

Theorem 1. *Let V be the Stackelberg value of the game. If the learner is playing a no-regret learning algorithm, and its best response in the Stackelberg game is unique, then the leader can guarantee at least $VT - o(T)$ utility.*

This theorem says that the first player can achieve a value close to Stackelberg value. To see how can this happen, we will give the following qualatitive argument.

Let (α^*, z^*) be the Stackelberg equilibrium. Suppose that you choose the mixed strategy α^*, with $\alpha_s \geq 0$, so that $z^*(\alpha^*) = \arg\min_{z \in S_2} \sum_{s \in S_1} \alpha_s u_2(s, z)$ is unique. Then the second best strategy for the learner, $\hat{z} = \arg\min_{z \in S_2 \setminus \{z^*\}} \sum_{s \in S_1} \alpha_s u_2(s, z)$ satisfies the supoptimality gap $u_2(\alpha^*, z^*) - u_2(\alpha^*, \hat{z}) \geq \kappa$.

The no-regret learner has an expected payoff of $T \sum_{s \in S_1} \alpha_s u_2(s, z)$ for each strategy $z \in S_2$ over T period. Instead of maximizing this expected value, there are two sources of errors: (1) By the law of large numbers, the difference between the mean and expected payoff vanishes as $T \to \infty$. (2) The error from no-regret learning is $o(T)$, which also goes to 0 as $T \to \infty$.

When T is large enough,

$$(\text{error in expectation} + \text{error in regret}) \ < T\kappa$$

Here κ is the gap the gap between the learner's payoff from the best response z^* and the second best response $\hat{z}$. In particular, average error is below κ after $o(T)$ time.

The no regret condition implies that the learner must play z^* for all but a small portion (*i.e.* $o(T)$) of rounds from 0 to T. Hence, the average payoff of the leader is almost the same as the Stackelberg payoff.

2.2 Example: Leader Outperforming Stackelberg

We may ask whether the leader can exploit the no regret learner and do better than Stackelberg value? The answer is yes.

Consider the following example. The game consists of the row player (leader), and the column player (learner), and has the following payoff matrix.

	L	M	R
U	$(0, \epsilon)$	$(-2, -1)$	$(-2, 0)$
D	$(0, -1)$	$(-2, 1)$	$(2, 0)$

Table 2: Payoff matrix in a two player game

First we will show that both NE value and Stackelberg value are 0.

Observe that (U, L) and $((\frac{1}{2}U, \frac{1}{2}D), R)$ are Nash. And no other Nash can do better. All Nash gives the row player payoff 0.

For Stackelberg game, to get the value > 0, the row player wants to choose D. But once he plays D with more than $\frac{1}{2}$ probability, the column player chooses M. Therefore, the best payoff the row player (leader) gets is 0.

However, playing against a *mean-based* no regret learner (choosing the strategy that minimizes the accumulative cost in the past), we can hope to do better.

Consider the following strategy: suppose the overall time is T.

Row player: play U during 0 to $\frac{T}{2}$ rounds, and then switch to D and play it for x times.

For the column player, now the historical payoff of each strategies are:

1. $L : \epsilon\frac{T}{2} - x$
2. $M = -\frac{T}{2} + x$

3. $R = 0$

After $x > \epsilon\frac{T}{2}$ (playing D for x additional rounds), $Payoff(R) = 0 \geq Payoff(L), Payoff(M)$, the column player will choose R in $\frac{T}{2}(1+\epsilon)\cdots T$ rounds. This gives the row player a total payoff

$$\sim 0 \cdot \frac{T}{2} + 2 \cdot \frac{T}{2}(1-\epsilon) = T(1-\epsilon) > 0$$

This shows that we can take advantage of the no regret learner and achieve a strictly better payoff than either Nash or Stackelberg game.

In the next class, we'll explore how Google leveraged mean-based no-regret learning in auctions to build its fortune. We'll also discuss strategies to modify the algorithm to defend against this approach.