Lecture 7

Instructor: Haipeng Luo

1 Two-player Zero-sum Games

In this lecture we explore the connection between game theory and online learning. We focus on simple two-player zero-sum games that could be represented using a matrix $G \in [0,1]^{N \times M}$, where one player (called the row player) has N possible actions and another player (called the column player) has M possible actions, and entry $G(i,j)$ represents the loss of the row player if he/she picks action i while the opponent picks action j, which is also the reward for the column player (hence zero-sum).

A classic example is the Rock-Paper-Scissors game. If we assign loss 1 for losing the game, 0 for winning and $1/2$ for a tie, then G is

$$\begin{array}{c} \\ \text{Rock} \\ \text{Paper} \\ \text{Scissors} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Rock} & \text{Paper} & \text{Scissors} \end{array} \\ \begin{pmatrix} 1/2 & 1 & 0 \\ 0 & 1/2 & 1 \\ 1 & 0 & 1/2 \end{pmatrix} \end{array}.$$

Instead of playing a fixed action (also called "pure strategy"), it often makes more sense to play a action randomly according to a distribution (called "mixed strategy"). For some mixed strategy $p \in \Delta(N)$ for the row player and some mixed strategy $q \in \Delta(M)$ for the column player, the expected loss for the row player, which is also the expected reward of the column player, is denoted by $G(p,q) = \sum_{i,j} p(i)q(j)G(i,j)$. We will also use the notation $G(i,q)$ and $G(p,j)$ to denote $\sum_j q(j)G(i,j)$ and $\sum_i p(i)G(i,j)$ respectively.

Perhaps the most important notion in game theory is the *Nash equilibrium*. A pair of mixed strategy (p,q) is called a Nash equilibrium if neither player has a incentive to change his/her strategy given that the opponent is keeping his/hers. In other words, everyone is happy about the current situation. Formally, this means that

$$G(p,q') \le G(p,q) \le G(p',q), \quad \forall p' \in \Delta(N), q' \in \Delta(M).$$

One can easily verify that for the Rock-Paper-Scissors game, playing uniformly at random for both players is a Nash Equilibrium (in fact the only one).

On the other hand, minimax solution is also a natural concept for a two-player zero-sum game. Specifically, in the worst case, playing p leads to a loss of at most $\max_q G(p,q)$ for the row player if the column player sees p before making decisions, and therefore in this sense the worst-case optimal strategy for the row player is $p^\star \in \operatorname{argmin}_p \max_q G(p,q)$, which is called the minimax strategy. Similarly, the maximin strategy for the column player is $q^\star \in \operatorname{argmax}_q \min_p G(p,q)$. Together, we call $(p^\star, q^\star)$ a minimax solution of the game.

Therefore, $\min_p \max_q G(p,q)$ and $\max_q \min_p G(p,q)$ are respectively the smallest loss and the largest reward the respective player can hope for when against an optimal opponent who plays second. How are these two values related? Intuitively, both players are playing optimally in the two expressions, but there should be no disadvantage in playing second. Therefore we should have $\min_p \max_q G(p,q) \ge \max_q \min_p G(p,q)$ (row player playing first on the left and second on the right). Indeed, this is true by a simple argument:

$$\min_p \max_q G(p,q) = \max_q G(p^\star, q) \ge G(p^\star, q^\star) \ge \min_p G(p, q^\star) = \max_q \min_p G(p,q).$$

While one may imagine that this inequality should be strict at least for some cases, the surprising fact is that the reverse inequality is also true and therefore the two values are exactly the same! In other words, if both players are playing optimally, there is no difference in playing first or second. This is the celebrated von Neumann's minimax theorem.

Theorem 1 (von Neumann's minimax theorem)**.** *For any two-player zero-sum game $G \in [0,1]^{N\times M}$, we have*

$$\min_p \max_q G(p,q) = \max_q \min_p G(p,q).$$

This single value is called the value of the game, denoted by $v(G)$. The original proof relies on a fixed-point theorem, but we will prove it in a different way *by running online learning algorithms* in the next section. For now, we discuss the connection between these different notions we have talked about so far: Nash equilibrium, minimax solution, and the value of the game.

Theorem 2. *A pair of mixed strategy (p,q) is a Nash equilibrium if and only if it is also a minimax solution. Moreover, $G(p,q)$ is the value of the game.*

Proof. Suppose (p,q) is a Nash equilibrium. By definition and optimality, we have

$$\min_{p'} \max_{q'} G(p',q') \le \max_{q'} G(p,q') = G(p,q) = \min_{p'} G(p',q) \le \max_{q'} \min_{p'} G(p',q').$$

Now by the minimax theorem, the above inequalities are actually equalities, which implies that $G(p,q) = v(G)$ and also (p,q) is a minimax solution.

Next for the other direction, if (p,q) is a minimax solution, then again by optimality and definition

$$\min_{p'} \max_{q'} G(p',q') = \max_{q'} G(p,q') \ge G(p,q) \ge \min_{p'} G(p',q) = \max_{q'} \min_{p'} G(p',q').$$

By the minimax theorem, the above is again an equality, which implies $G(p,q) = v(G)$ and (p,q) is a Nash equilibrium. □

By this theorem and the fact that minimax solutions always exist (due to compactness of the simplex), Nash equilibria also always exist.

2 Repeated Play

How should we play a game? If we know the matrix G, then playing with the minimax solutions seems to be a good strategy. However, what if G is unknown? Moreover, minimax solutions might also be too pessimistic. For example, if we play Rock-Paper-Scissors with a friend who we know prefers to play Paper for instance, then should we still play uniformly at random? In general, how do we exploit the fact that the opponent might not be optimal?

If the game is only played once, then there is little we can do. However, it is often the case that a game is repeatedly played for many times. In this case, there is hope to apply learning algorithms to learn to play well against a specific opponent. We take the row player as an example and formulated the learning model as follows: at round $t = 1, \dots, T$,

- the row player chooses mixed strategy p_t;
- the column player chooses mixed strategy q_t (which may or may not depend on p_t);
- the row player observes $G(i, q_t)$ for all $i \in [N]$.

The feedback model can be potentially extended to the more realistic case where only $G(i,j)$ is observed for some i and j drawn from p_t and q_t respectively, but for now we will stick with this easier full information feedback.

A very natural idea for the player is to make use of an expert algorithm such as Hedge, treating each available action i as an expert. Specifically, given an expert algorithm as a blackbox, p_t will be the output of this algorithm at round t, and the loss vector to pass back to the algorithm would be ℓ_t

such that $\ell_t(i) = G(i, q_t),\ \forall i$. Suppose the expert algorithm has regret bound $\mathcal{R}_T$, then it implies

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} G(p_t, q_t) &\le \min_p \frac{1}{T}\sum_{t=1}^{T} G(p, q_t) + \frac{\mathcal{R}_T}{T} \\ &= \min_p G(p, \bar{q}) + \frac{\mathcal{R}_T}{T} && (\bar{q} = \tfrac{1}{T}\textstyle\sum_{t=1}^{T} q_t) \\ &\le \max_q \min_p G(p, q) + \frac{\mathcal{R}_T}{T}.\end{aligned}$$

Therefore, if the regret is sublinear and T is large, then the average loss of the row player is very close to the value of the game, which is the smallest possible loss if against an optimal opponent. However, by using a learning algorithm instead of a minimax solution directly (if it is available), the average loss can also be much smaller in the case when the opponent is not exactly optimal (that is, when $\bar{q}$ is not close to the maximin strategy and the last inequality is loose).

However, even more interesting thing happens if the column player also uses an expert algorithm (by using the negative rewards as losses). To see this, suppose the regret bound for the column player is $\mathcal{R}'_T$:

$$\sum_{t=1}^{T} -G(p_t, q_t) - \min_q \sum_{t=1}^{T} -G(p_t, q) = \max_q \sum_{t=1}^{T} G(p_t, q) - \sum_{t=1}^{T} G(p_t, q_t) \le \mathcal{R}'_T.$$

Then we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} G(p_t, q_t) &\ge \max_q \frac{1}{T}\sum_{t=1}^{T} G(p_t, q) - \frac{\mathcal{R}'_T}{T} \\ &= \max_q G(\bar{p}, q) - \frac{\mathcal{R}'_T}{T} && (\bar{p} = \tfrac{1}{T}\textstyle\sum_{t=1}^{T} p_t) \\ &\ge \min_p \max_q G(p, q) - \frac{\mathcal{R}'_T}{T}.\end{aligned}$$

Combining the two derivations, we have

$$\min_p \max_q G(p, q) \le \frac{1}{T}\sum_{t=1}^{T} G(p_t, q_t) + \frac{\mathcal{R}'_T}{T} \le \max_q \min_p G(p, q) + \frac{\mathcal{R}_T}{T} + \frac{\mathcal{R}'_T}{T}.$$

If $\mathcal{R}_T$ and $\mathcal{R}'_T$ are sublinear, which we can indeed ensure by using for example Hedge, then the term $\frac{\mathcal{R}_T}{T} + \frac{\mathcal{R}'_T}{T}$ can be arbitrarily close to 0 as T goes to infinity. Therefore, we must have

$$\min_p \max_q G(p, q) \le \max_q \min_p G(p, q),$$

which means that we just proved the minimax theorem (recall the other direction is trivial)! This is one of the few proofs that prove a mathematical statement by running algorithms, and is taken from [Freund and Schapire, 1999].

In fact, the derivations above also tell us

$$\max_q G(\bar{p}, q) \le \min_p \max_q G(p, q) + \frac{\mathcal{R}_T}{T} + \frac{\mathcal{R}'_T}{T} \quad \text{and} \quad \max_q \min_p G(p, q) \le \min_p G(p, \bar{q}) + \frac{\mathcal{R}_T}{T} + \frac{\mathcal{R}'_T}{T},$$

which means $\bar{p}$ and $\bar{q}$ are approximately minimax solutions with error $\frac{\mathcal{R}_T}{T} + \frac{\mathcal{R}'_T}{T}$. In other words, this also provides a concrete way to calculate a minimax solution/Nash equilibrium.

3 Faster Convergence via Adaptivity

We know that the worst-case optimal regret for the expert problem is of order $\mathcal{O}(\sqrt{T})$, which means the convergence rate of the above approach is of order $\mathcal{O}(1/\sqrt{T})$. Is this the optimal rate in this specific context? The answer is no – one can in fact converge much faster using an expert algorithm

with some special adaptive property. Specifically, recall the bound we prove for Optimistic FTRL in Lecture 6 (with $m_t = \ell_{t-1}$):

$$\mathcal{R}_T \leq \frac{2+\ln N}{\eta} + \eta \sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_\infty^2 - \frac{1}{4\eta} \sum_{t=1}^{T} \|p_{t+1} - p_t\|_1^2 . \tag{1}$$

In the context of game playing, for the row player we have by Hölder's inequality (for $t \neq 1$)

$$\|\ell_t - \ell_{t-1}\|_\infty^2 = \max_i |G(i, q_t) - G(i, q_{t-1})|^2 = \max_i |\langle G(i,\cdot), q_t - q_{t-1}\rangle|^2 \leq \|q_t - q_{t-1}\|_1^2$$

where we use $G(i,\cdot)$ to denote the i-th row of G. In other words, from the row player's perspective, the stability of the environment is controlled by the stability of the column player's strategy. The exact same argument holds for the column player and therefore if both players use Optimistic FTRL with the same learning rate η, then we have

$$\mathcal{R}_T \leq \frac{2+\ln N}{\eta} + \eta + \eta \sum_{t=2}^{T} \|q_t - q_{t-1}\|_1^2 - \frac{1}{4\eta} \sum_{t=2}^{T} \|p_t - p_{t-1}\|_1^2$$

$$\mathcal{R}'_T \leq \frac{2+\ln M}{\eta} + \eta + \eta \sum_{t=2}^{T} \|p_t - p_{t-1}\|_1^2 - \frac{1}{4\eta} \sum_{t=2}^{T} \|q_t - q_{t-1}\|_1^2 .$$

Summing up the two bounds gives

$$\mathcal{R}_T + \mathcal{R}'_T \leq \frac{4+\ln(NM)}{\eta} + 2\eta + \left(\eta - \frac{1}{4\eta}\right) \sum_{t=2}^{T} \left(\|p_t - p_{t-1}\|_1^2 + \|q_t - q_{t-1}\|_1^2\right),$$

and simply setting $\eta = 1/2$ leads to

$$\mathcal{R}_T + \mathcal{R}'_T \leq 9 + 2\ln(NM),$$

which is independent of T! In other words, the average strategy $(\bar{p}, \bar{q})$ converges to the Nash equilibrium at a rate $\mathcal{O}(1/T)$ instead of $\mathcal{O}(1/\sqrt{T})$. In fact, similar results hold even if the two players do not use the exact same optimistic FTRL algorithm. The key is clearly only the special adaptive bound in the form of Eq. (1) and there are several other algorithms that enjoy similar bounds. See [Syrgkanis et al., 2015] for details.

We finally point out that even if we only look at each player's individual regret, it could still be smaller than the worst-case $\mathcal{O}(\sqrt{T})$. Take the row player as an example, suppose the column player's algorithm is stable in the sense that $\|q_t - q_{t-1}\|_1 \leq c\eta$ for some constant $c > 0$. Then the regret for the row player who uses Optimistic FTRL is

$$\mathcal{R}_T \leq \frac{2+\ln N}{\eta} + \eta + \eta \sum_{t=2}^{T} \|q_t - q_{t-1}\|_1^2 \leq \frac{2+\ln N}{\eta} + \eta + c^2 T \eta^3.$$

Setting $\eta = \left(\frac{\ln N}{T}\right)^{\frac{1}{4}}$ we have $\mathcal{R}_T = \mathcal{O}(T^{\frac{1}{4}}(\ln N)^{\frac{3}{4}})$, which is again better than $\mathcal{O}(\sqrt{T})$. The stability condition on q_t is not unfamiliar to us by now – we know that if the column player uses Hedge (with learning rate η), then stability holds with $c = 1$; if the column player uses Optimistic FTRL (with learning rate η), then stability holds with $c = 2$ (use the stability lemma of Lecture 6 to verify why this is true).

References

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems 28*, 2015.