Learning in Games

1 Quick Recap: Game Theory

Definition 1. *A* **normal-form (or a strategic-form) game** Γ *is defined as a triplet* $(\mathcal{N}, \mathcal{C}, \mathcal{U})$*, where*

- $\mathcal{N} = \{1, \cdots, N\}$ *is the set of* N *players (agents),*
- $\mathcal{C} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_N$ *is the strategy profile space, where* $\mathcal{C}_i$ *represents the choice set at the* i^{th} *player,*
- $\mathcal{U} = \{u_1, \cdots, u_N\}$ *is the utility profile, where* $u_i : \mathcal{C}_i \to \mathbb{R}$ *is the* i^{th} *player's utility.*

Definition 2. *Given* $\Gamma = (\mathcal{N}, \mathcal{C}, \mathcal{U})$*, a strategy profile* $(c_1, \cdots, c_N)$ *is a* **pure-strategy Nash equilibrium (PSNE)** *if*

$$u_i(c_i, \boldsymbol{c}_{-i}) \geq u_i(c_i', \boldsymbol{c}_{-i}),$$

for all $c_i' \in \mathcal{C}_i$*, for all* $i \in \mathcal{N}$*.*

Definition 3. *Given* $\Gamma = (\mathcal{N}, \mathcal{C}, \mathcal{U})$*, a mixed-strategy profile* $(\pi_1, \cdots, \pi_N)$ *is a* **mixed-strategy Nash equilibrium (MSNE)** *if*

$$u_i(\pi_i, \boldsymbol{\pi}_{-i}) \geq u_i(\pi_i', \boldsymbol{\pi}_{-i}),$$

for all $\pi_i' \in \Delta(\mathcal{C}_i)$*, for all* $i \in \mathcal{N}$*.*

Definition 4. *A* **correlated strategy** $\tau_{\mathcal{S}}$ *in* $\Gamma = (\mathcal{N}, \mathcal{C}, \mathcal{U})$ *is any probability distribution in the simplex* $\Delta(\mathcal{C}_{\mathcal{S}})$ *over a subset of players* $\mathcal{S} \subseteq \mathcal{N}$, *where* $\mathcal{C}_{\mathcal{S}} = \times_{i \in \mathcal{S}} \mathcal{C}_i$.

Definition 5. *A* **correlated equilibrium** *is a correlated strategy* $\boldsymbol{\tau} \in \Delta(\mathcal{C})$ *in a game* Γ *if*

$$U_i(\boldsymbol{\tau}) \geq \sum_{c \in \mathcal{C}_{\mathcal{S}}} \tau_{\mathcal{S}}(c) u_i(c_i', c_{-i}),$$

for all $c_i' \in \mathcal{C}_i$, *and for all* $i \in \mathcal{N}$.

What if, agents cannot guess other players' decision models (strategies and utility functions)?

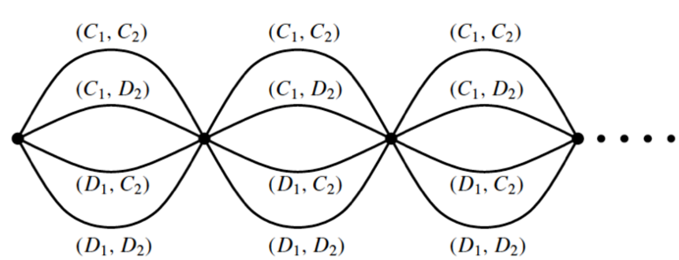
Can agents learn what NE is, if the game is played repeatedly over an infinite time horizon?

Is NE a valid solution concept in such settings?

Example: Consider the following infinitely repeated prisoner's dilemma.

Prisoner 1 \ Prisoner 2	C_2	D_2
C_1	2, 2	0, 3
D_1	3, 0	1, 1

$\Rightarrow$



This is very different from machine learning because...

- the state of the environment is dictated by multi-agent decisions $\Rightarrow$ optimality criterion is unknown!
- every agent employs an online learning algorithm, leading to a very different dynamical system as opposed to single-agent reinforcement learning.
- Involves learning mental models about other players' decisions.

In this topic, we will cover two learning paradigms:

- Fictitious Play (Best Response Dynamics)
- No-Regret Learning

However, there are many other learning paradigms. Some examples include

- Stochastic Fictitious Play
- Swap Regret Minimization
- and many more

For more details, please refer to (a non-exhaustive) list of papers posted in the course webpage (in the reading material page).

2 Fictitious Play

Can players converge to NE in one-shot games?

- Initialize beliefs about other players' strategies.
- In each turn, update beliefs according to observed actions.
- Play a best response to the other players' expected strategies (pre-play).

More formally, at player i, if W_t represents the counts of other players' strategies at time t, we have

FICTITIOUSPLAY(i, W_t)

for $j \in \mathcal{N} - \{i\}$

  for $c \in \mathcal{C}_j$

    $\hat{\pi}_{-i,t}(j, c) = \dfrac{W_t(j, c)}{\displaystyle\sum_{c' \in \mathcal{C}_j} W_t(j, c')}$ // Assessed Strategy

$c_{i,t} = \arg\max_{c \in \mathcal{C}_i} u_i(c, \hat{\boldsymbol{\pi}}_{-i,t})$ // BR to assessed strategy

return c_i

Theorem 2.1

If the empirical distribution of each player's strategies converges in fictitious play, then it converges to Nash equilibrium.

Example:

Theorem 2.2

Marginal distribution of each player' strategies in fictiticious play converges to Nash in

(i) generic payoffs in 2×2 games [Robinson-1951],

(ii) two-person zero-sum games [Miyasawa-1961].

(iii) potential games [Monderer-Shapley-1996]

(In this class, we will only cover the proof for potential games.)

Problems with Fictitious Play:

- Can players always find NE?
- If yes, which NE can/will they find?
- Need too much information – correct beliefs about all other players' strategies.
- NE is difficult to find... (will be covered later in this class – it is PPAD Hard in general.)
- Do we observe such a behavior in the real world?

3 No-Regret Learning: A Single Agent Framework

- For the sake of simplicity, consider a single agent.
- This is a classic ***online learning*** paradigm!
- Agent picks a choice $c_t \in \mathcal{C}$ at time t, and observes a cost vector $\boldsymbol{x}_t = \{x_t(c_t), x_t(-\boldsymbol{c}_t)\}$ in hindsight.
- Let $p_t = f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{t-1})$ be the probability distribution over the set of choices $\mathcal{C}$.
- Assume the agent employs an algorithm A to update p_t to effectively suit to the future cost $\boldsymbol{x}_t$.

Design a learning algorithm A against a cost function chosen by an adversary as $c_t = g(p_1, \cdots, p_t)$.

Solution: Minimize regret.

Definition 6. Internal regret *of an agent at time T for playing a strategy $\boldsymbol{c}_T = \{c_1, \cdots, c_T\}$ is*

$$R_I(\boldsymbol{c}_T) = \frac{1}{T}\sum_{t=1}^{T} x_t(c_t) - \frac{1}{T}\sum_{t=1}^{T} \min_{c \in \mathcal{C}} x_t(c). \qquad (1)$$

The second term is equivalent to playing best response in each interaction...

Definition 7. External regret *of an agent at time t for not playing a strategy $\boldsymbol{c}_T = \{c_1, \cdots, c_T\}$ is*

$$R_E(\boldsymbol{c}_T) = \sum_{t=1}^{T} x_t(c_t) - \min_{c \in \mathcal{C}} \sum_{t=1}^{T} x_t(c). \tag{2}$$

The second term here is equivalent to choosing one best response in hindsight over the past T iterations.

Definition 8. *An online learning algorithm A has* **no regret** *if, for every $\epsilon > 0$, there exists a sufficiently large time $T = T(\epsilon)$ such that, for every adversary of A, in expectation over the action realizations, we have*

$$R_E(\boldsymbol{c}_T) \leq \epsilon. \tag{3}$$

Theorem 3.1

For every set $\mathcal{C}$ of M choices and time horizon $T \geq 4 \ln M$, there is an online learning algorithm that has an expected regret of at most $2\sqrt{\ln M / T}$ for any adversary.

Corollary 1. *For any $\epsilon > 0$, there is an online learning algorithm A that produces an expected regret of at most ϵ within a time horizon $T = \frac{4 \ln M}{\epsilon^2}$.*

An example algorithm that achieves this bound:

MultiplicativeWeightUpdate

Initialize $w_1(c) = 1$ for every $c \in \mathcal{C}$

for each time step $t = 1, \cdots, T$,

use $p_t(c) = \dfrac{w_t(c)}{\sum_{c \in \mathcal{C}} w_t(c)}$ to sample from $\mathcal{C}$.

Given $\boldsymbol{x}_t$, update $w_{t+1}(c) = w_t(c) \cdot [1 - \eta x_t(c)]$

Proof of Theorem 3.1 for MultiplicativeWeightUpdate algorithm:

4 No-Regret Learning: Multi-Agent Dynamics

Given all the other agents' choices $\boldsymbol{c}_{-i,t}$, let agent i's cost (negative utilities) be denoted as $X_{i,t}(c_i, \boldsymbol{c}_{-i,t})$ at time t.
Let

$$x_{i,t}(c_i) = \mathbb{E}_{\boldsymbol{c}_{-i,t} \sim p_{-i,t}} \left[X_{i,t}(c_i, \boldsymbol{c}_{-i,t}) \right], \tag{4}$$

where $p_{-i,t} = \prod_{j \neq i} p_{j,t}$.

Assume each agent employs a no-regret algorithm A.
Then, the no-regret dynamics at an agent $i \in \mathcal{N}$ in a game $\Gamma = (\mathcal{N}, \mathcal{C}, \mathcal{X})$ is given by

NoRegretDynamics(i)

```
At each time step t = 1, ··· , T:
        Agent i independently chooses p_{i,t} using A.
        Agent i receives a cost vector x_{i,t}.
```

If every agent uses the MultiplicativeWeightUpdate algorithm, then if every agent has at most M strategies and if the costs lie between $[-x^*, x^*]$, then only $4c^{*2} \ln M/\epsilon^2$ iterations are required to drive the expected regret to at most ϵ.

Theorem 4.1

Suppose that, after T iterations of no-regret dynamics, each agent i has an expected regret of at most ϵ. Let $p_t = \prod_{i \in \mathcal{N}} p_{i,t}$ denote the outcome distribution at iteration t and $p = \frac{1}{T} \sum_{t=1}^{T} p_t$ denote the time-averaged history of these distributions. Then, p is an approximate coarse correlated equilibrium, i.e.,

$$\mathbb{E}_{\boldsymbol{c} \sim p} \left[X_i(\boldsymbol{c}) \right] \leq \mathbb{E}_{\boldsymbol{c} \sim p} \left[X_i(c', \boldsymbol{c}_{-i}) \right] + \epsilon \tag{5}$$

for every agent i and unilateral deviation $c' \in \mathcal{C}_i$.

5 Reputation Mechanisms