Computing a Nash equilibrium for repeated games

Andriy Burkov

Laval University, Canada
Motivations

- See the difference between Nash equilibria in a one-shot and a repeated normal-form game.
- Explore computation properties of finding a Nash equilibrium of a repeated game.
Plan

- Definitions
- Folk theorem
- Constructing an equilibrium
Definitions

- A repeated bimatrix game is played by two players, 1 and 2.
- Each player has a set of action choices of size $n^1$ and $n^2$.
- The game is played in rounds, with the two players simultaneously making a choice of action at each round.
- If Player 1 chooses action $1 \leq i^1 \leq n^1$ and Player 2 chooses $1 \leq i^2 \leq n^2$, they receive payoffs of $P_{i^1 i^2}^1$ and $P_{i^2 i^1}^2$, respectively.
- A strategy is a (possibly stochastic) function of the history of interactions.
Definitions (cont’d)

- Let \((p, q) \in \{(1, 2), (2, 1)\}\)

- A player \(p\)’s defensive (or maximin) strategy is defined as follows:

\[
\delta^p = \arg\max_{\pi_p} \min_{i_q} \sum_{i_p} P^p_{i_p i_q} \pi^p_{i_p}.
\]

- A strategy \(\pi^p\) is said to be individually rational if:

\[
U(\pi^p) \geq U(\delta^p)
\]
Define

$$\alpha^p = \text{argmin}_{\pi^p} \max_i \sum_{i'p} P_{iq_{i'}} \pi^p_{ip}. $$

to be an attacking strategy of player $p$.

Define

$$g^p = \sum_{ip} \sum_{iq} \delta^p_{ip} \alpha^q_{iq} $$

to be the expected payoff that player $p$ can guarantee itself by playing its defensive strategy.
Definitions (cont’d)

- Any individually rational strategy in the game defined by $A^p$ and $A^q$ yields in a nonnegative payoff for both players.
- The defensive strategy in the game defined by $A^p$ and $A^q$ has the average payoff of 0.
- The set of Nesh equilibria of the game defined by $A^p$ and $A^q$ coincides with this of $P^p$ and $P^q$.

- Thus, in the following, for simplicity we can only work with advantage matrices: this will not change the result.
Definitions (cont’d)

- When Player 1 chooses an action $i^1$ and Player 2 chooses $i^2$, the payoffs for the two players can be visualized as a point $x = (A^1_{i^1 i^2}, A^2_{i^2 i^1}) = (x^1, x^2) \in X$ in a two-dimensional space.

- The convex hull of a set of points is the set of all points that can be formed by a convex combination of points in the set.

- The set of all payoffs of players in the convex hull of $X$ is called a feasible and individually rational payoff set.
Example: Prisoner’s Dilemma
The Folk Theorem

- The Folk Theorem tells that in an infinitely repeated game and for sufficiently patient players, any feasible and individually rational payoff can be a payoff of a Nash equilibrium of this game.

- The main idea is that players can have an agreement to play a strategy bringing to each player a positive average payoff of the advantage game.
  - If one player deviates, the other one can switch to his attacking strategy forcing its opponent to obtain at most a zero average payoff for arbitrarily long time.
Which point to choose?

- The question is which point to choose in set of feasible and individually rational payoffs?
- Following Nash, Littman and Stone propose to take a point that maximizes the product of the advantages.
  - This point can be efficiently (polynomially) identified in two-player case.
Computing a NE point

- Obviously, this point is found on the outer boundary of the convex hull.
  - If not, we can always move righter or upper in the convex set, maximizing the product of advantages.
- This implies that that the point can be expressed by a weight vector \( w \) that has non-zero weight on only one or two \( x \in X \).
Computing a NE point

- Let \( x, y \in X \) be two points in \( X \). We want to find a point \( z \) on the edge between \( x \) and \( y \), \( z = w_x x + (1 - w_x) y \) for \( 0 \leq w_x \leq 1 \), such that the product of the advantages, \( z^1 z^2 \), is maximized.

- By finding a derivative of this product and solving for \( w_x \) we find:

\[
    w_x = \frac{-y^2(x^1 - y^1) - y^1(x^2 - y^2)}{2(x^2 - y^2)(x^1 - y^1)}.
\]

- If \( w_x < 0 \) or \( w_x > 1 \), then the maximum product is achieved at an endpoint.
If the weight computed for points \( x \) and \( y \) is between 0 and 1, let the weight be represented by the rational number \( \frac{r}{s} \) for integers \( 0 \leq r \leq s \).

A pair of strategies that repeats any sequence that includes the action pair \((i^1, i^2)\) a total of \( r_i \) times and \((j^1, j^2)\) a total of \( r_j \) times maximizes the product of advantages over all feasible payoffs.
The above strategies result in average advantages

\[ z = \frac{(r_i A^1_{i1j2} + r_j A^1_{j1j2}, r_i A^2_{i2i1} + r_j A^2_{j2j1})}{(r_i + r_j)}. \]

These strategies maximize the product of advantages over all strategies, so we say they represent *mutual cooperation*.
Constructing Automaton

- What if one player does not want to cooperate?
- There should be a threat of punishment integrated into the above strategy.
Constructing Automaton

Computing the number of punishment rounds:

- The average payoff of Player $q$ for cooperating is

$$z^q = \frac{r_i A_i^q_{ijp} + r_j A_j^q_{jqp}}{(r_i + r_j)}$$

- Let $d^q = \max_{x \in X} x^q$ be an upper bound on the value that $q$ can get in a single round by defecting.
Computing the number of punishment rounds (cont’d):

- The average payoff of $q$ in the advantage game for using a strategy that defects cannot be larger than
  \[
  \frac{(r_i + r_j)d^q}{r_i + r_j + a^p}
  \]

- How to set $a^p$ so that cooperation is a best response?
  - The following inequality must be satisfied:
    \[
    z^q > \frac{(r_i + r_j)d^q}{r_i + r_j + a^p}
    \]

- The above implies that
  \[
  a^p = \left\lceil \frac{(r_i + r_j)(d^q - z^q)}{z^q} \right\rceil
  \]
Conclusion

- In two player repeated game with average payoff criterion and no discounting, a Nash equilibrium can be efficiently computed (a polynomial algorithm presented).

- However, it has been recently shown by Borgs et al (2007), this result cannot be extended to three or more player case. This means that three or more player repeated game equilibrium is computationally intractable (PPAD).