3. Regret minimization and hindsight rationality

Link

---

0 - Introduction

- How can we 'learn' from repeatedly playing a game?
- How does this *dynamic* concept of learning relate to the *static* concept of game-theoretic equilibria?

1 - Hindsight rationality and -Regret

- Suppose we have one player in a game, let:
  - - set of available strategies.
  - - strategy played at timestep .
  - Player receives some feedback after each play through, could be a gradient of the utility or the utility itself.
  - - some 'better' strategy it has formulated based on the feedback.

- *Hindsight rationality* - a player has "learnt" to play the game when looking back at the history of play, they cannot think of any transformation: of their strategies that when applied to the whole history of play would've yielded a strictly better utility to the player.

- - model for a decision maker that repeatedly interacts with a black-box, given a set of points and a set of linear transformations .

  - At each time , the regret minimizer interacts with the environment through two operations:

* - regret minimizer returns an element .

* () - provides feedback to the minimizer from the environment, , based on how good the last strategy was.

- *Cumulative -regret* - regret minimizer's quality metric, its goal is to guarantee that -regret grows asymptotically sublinearly as time increases.

- and alternate: the minimizer presents a new strategy to the black box and receives a new utility function back from the black box. It then uses this feedback to formulate a new strategy and so on.

- The regret minimizer's decision making is online in the sense that its strategy depends on its past strategies and environmental feedback (observed utility functions).

1.1 - Some Notable Choices for the Set of Transformations Regret

- What set of transitions should the agent consider for its ? The size of defines the agent's rationality.

- Possible choices of :

  - = *Swap regret* - set of all mappings from to .

    * Maximum size of and therefore the highest level of hindsight rationality.

  - = *Internal regret* - set of all single-point deviations. Each strategy maps directly to a different strategy.

where:

> When all agents in a multiplayer general-sum game use internal or swap regret, their empirical frequency of play converges to a *correlated equilibrium.*

- = *Trigger deviation functions*
  -regret is efficiently bounded with a polynomial dependence on the size of the game tree.

  > When all agents in a multiplayer general-sum game use trigger deviation functions, their empirical frequency of play converges to an *extensive-form correlated equilibrium*

- = *External regret* (constant transforms) - requires that the player not regret substituting all of the strategies they played with the same strategy .

  > When all agents in a multiplayer general-sum game use external regret, their empirical frequency of play converges to a *coarse correlated equilibrium.*

  > When all agents in a two-player zero-sum game use external regret, their average strategies converge to *Nash equilibrium.*

Note: My understanding is that internal regret is the regret associated with choosing an action at a single decision point; whereas, external regret is the regret associated with entire strategies.

1.2 - A Very Important Special Case: Regret Minimization

- *Regret minimizer* (external regret minimizer) - special case of the -regret minimizer where is chosen to be the set of constant transforms:
  External regret:

  Goal is for the cumulative regret to grow sublinearly in .

- Online linear optimization asserts sublinear regret for convex and compact domain , on the order .

- External regret minimization guarantees:
  - Nash equilibrium in two-player zero-sum games.

    - Coarse correlated equilibrium in multiplayer general-sum games
    - Best responses to static stochastic opponents in multiplayer general-sum games.
    - etc.

1.3 From Regret Minimization to -Regret Minimization

Note: I don't understand this section

- There exits a construction such that -regret minimization reduces to regret minimization. [2008-Gordon]

- **Theorem:**

    - Let be a deterministic regret minimizer with external regret and every admits fixed points .

    - A -regret minimizer, , can be constructed from :

        * Each call to calls 's to get the next transform, , which is then used to compute the next strategy .

        * Each call to constructs a linear utility function, , that is then passed to 's .

    - This -regret minimizer shares the same cumulative regret as the external regret minimizer, . Therefore the regret accumulated by grows sublinearly.

- Proof:

  - outputs transformations and receives utilities

  - Cumulative regret for , then:

and , thus

which equals cumulative regret.

2. Applications of regret minimization 2.1 - Learning a Best Response Against Stochastic Opponents

- Consider the game setup:

  - Playing a repeated -player general-sum game with multilinear utilities.

  - Players play stochastically:
    At each each they independently sample a strategy from a fixed distribution.

    - the strategy played by Player at time .

    - the average strategy played by Player .

  - Player is the learning agent, it picks strategies according to an algorithm that guarantees sublinear external regret.

  - Player feedback function at time , defined as

- The average strategy played by Player converges to a best response:

2.2 - Self-Play Convergence to Bilinear Saddle Points
(such as Nash equilibrium in a two-player zero-sum game)

- Regret minimization can be used to converge to bilinear saddle points.

- Bilinear saddle points are solutions of the form:

Where and are convex and compact sets.

- This type of optimization problem is pervasive in game-theory.
 - *Example:* Computation of Nash equilibria in two-player zero-sum games.
 - - strategy space for Player 1
 - - strategy space for Player 2
 - - payoff matrix for Player 1

- Use self play between regret minimizers to converge to bilinear saddle-points:

  – Let each player be a regret minimizer, and .

  – At each time , each minimizer outputs a strategy and , and receives feedback:

- *Saddle point gap* () - difference between the saddle points produced by the average strategies of the regret minimizers up to any time and a given pair of strategies , used to measure convergence.

- **Theorem:** Let and be the sublinear cumulative regret of the minimizers, up to time . And, and be the average of the strategies. Then the saddle point gap satisfies:
 (See the lecture notes for the proof.)