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 Φ-Regret

- Suppose we have one player in a game, let:
  - $\mathcal{X}$ - set of available strategies.
  - x^t - strategy played at timestep t.
  - Player receives some feedback after each play through, could be a gradient of the utility or the utility itself.
  - x^{t+1} - 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: $\phi : \mathcal{X} \to \mathcal{X}$ of their strategies that when applied to the whole history of play would've yielded a strictly better utility to the player.

- *Φ-regret minimizer* - model for a decision maker that repeatedly interacts with a black-box, given a set $\mathcal{X}$ of points and a set Φ of linear transformations $\phi : \mathcal{X} \to \mathcal{X}$.

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

    - NextStrategy - regret minimizer returns an element $x^t \in \mathcal{X}$.

    - ObserveUtility (ℓ^t) - provides feedback to the minimizer from the environment, $\ell^t \ : \ \mathcal{X} \to \mathbb{R}$, based on how good the last strategy x^t was.

- *Cumulative Φ-regret* - regret minimizer's quality metric, its goal is to guarantee that Φ-regret grows asymptotically sublinearly as time T increases.

$$R^T_\Phi := \max_{\hat{\phi} \in \Phi} \left\{ \sum_{t=1}^{T} \left(\ell^t(\hat{\phi}(x^t)) - \ell^t(x^t) \right) \right\}$$

- NextStrategy and ObserveUtility alternate: the minimizer presents a new strategy x^t to the black box and receives a new utility function back ℓ^t from the black box. It then uses this feedback to formulate a new strategy x^{t+1} 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 Φ Regret

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

- Possible choices of Φ :

  - Φ = *Swap regret* - set of all mappings from $\mathcal{X}$ to $\mathcal{X}$.
    - Maximum size of Φ and therefore the highest level of hindsight rationality.

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

$$\Phi = \{\phi_{a \to b}\}_{a,b \in \Phi}$$

    where:

$$\phi_{a \to b} : x \to \begin{cases} x, & \text{if } x \neq a \\ b, & \text{if } x = a \end{cases}$$

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

  - Φ = *Trigger deviation functions*
    Φ-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*

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

    > 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 Φ-regret minimizer where Φ is chosen to be the set of constant transforms:

  $$\Phi^{\text{Const}} := \{\phi_{\hat{x}} : x \to \hat{x}\}_{\hat{x} \in \mathcal{X}}$$

  External regret:

  $$R^T := \max_{\hat{x} \in \mathcal{X}} \left\{ \sum_{t=1}^{T} \left(\ell^t(\hat{x}) - \ell^t(x^t) \right) \right\}$$

  Goal is for the cumulative regret R^T to grow sublinearly in T.

- Online linear optimization asserts sublinear regret for convex and compact domain $\mathcal{X}$, on the order $R^T = O(\sqrt{T})$.

- 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 Φ-Regret Minimization

Note: I don't understand this section

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

- **Theorem:**

  - Let $\mathcal{R}$ be a deterministic regret minimizer with external regret and every $\phi \in \Phi$ admits fixed points $\phi(x) = x \in \mathcal{X}$.

  - A Φ-regret minimizer, $\mathcal{R}_\Phi$, can be constructed from $\mathcal{R}$:

    - Each $\mathcal{R}_\Phi$ call to NextStrategy calls $\mathcal{R}$'s NextStrategy to get the next transform, ϕ^t , which is then used to compute the next strategy $x^t = \phi^t(x^t)$.

    - Each $\mathcal{R}_\Phi$ call to $\text{ObserveUtility}(\ell^t)$ constructs a linear utility function, $L^t : \phi \to \ell^t(\phi(x^t))$, that is then passed to $\mathcal{R}$'s $\text{ObserveUtility}(L^t)$.

  - This Φ-regret minimizer shares the same cumulative regret as the external regret minimizer, $R_\Phi^T = R^T$. Therefore the regret accumulated by R_Φ grows sublinearly.

- Proof:

  - $\mathcal{R}$ outputs transformations $\phi^1, \phi^2, \ldots \in \Phi$ and receives utilities $\phi \to L^1(x^1)),\ \phi \to L^2(x^2)),\ \ldots$

  - Cumulative regret for $\mathcal{R}$, then:

  $$R^T = \max_{\hat{\phi} \in \Phi} \left\{ \sum_{t=1}^{T} \left(\ell^t(\hat{\phi}(x^t)) - L^t(x^t)) \right) \right\}$$

  $L^t(x^t) = \ell^t(\phi^t(x^t))$ and $\phi^t(x^t) = x^t$, thus

  $$R^T = \max_{\hat{\phi} \in \Phi} \left\{ \sum_{t=1}^{T} \left(\ell^t(\hat{\phi}(x^t)) - \ell^t(x^t) \right) \right\}$$

  which equals R_Φ^T cumulative regret.

2. Applications of regret minimization

2.1 - Learning a Best Response Against Stochastic Opponents

- Consider the game setup:

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

  - Players $i = 1, \ldots, n-1$ play stochastically:
    At each each t they independently sample a strategy $x^{(i),t} \in \mathcal{X}^{(i)}$ from a fixed distribution.

    $x^{(i),t}$ - the strategy played by Player i at time t.

    $\bar{x}^{(i)}$ - the average strategy played by Player i.

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

  - Player n feedback function at time t, defined as
    $\ell^t := \mathcal{X}^{(n)} \ni x^{(n)} \ \to\ u^{(n)}\left(x^{(1),t}, \ldots, x^{(n-1),t}, x^{(n)}\right)$

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

$$\frac{1}{T}\sum_{t=1}^{T} x^{(n),\,t} \to \underset{\hat{x}^{(n)}\in\mathcal{X}^{(n)}}{\arg\max}\left\{u^{(n)}\left(\bar{x}^{(1)},\ \ldots,\ \bar{x}^{(n-1)},\ \hat{x}^{(n)}\right)\right\}$$

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:

$$\max_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}}\ x^{\intercal}Ay$$

  Where $\mathcal{X}$ and $\mathcal{Y}$ 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.
  - $\mathcal{X}$ - strategy space for Player 1
  - $\mathcal{Y}$ - strategy space for Player 2
  - A - payoff matrix for Player 1

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

  - Let each player be a regret minimizer, $\mathcal{R}_{\mathcal{X}}$ and $\mathcal{R}_{\mathcal{Y}}$.

  - At each time t, each minimizer outputs a strategy x^t and y^t, and receives feedback:

$$\ell^t_{\mathcal{X}} : x \to (Ay^t)^{\intercal}x \qquad \ell^t_{\mathcal{Y}} : y \to -(A^{\intercal}x^t)^{\intercal}y$$

- *Saddle point gap* (γ) - difference between the saddle points produced by the average strategies $(\hat{x},\hat{y})$ of the regret minimizers up to any time T and a given pair of strategies (x,y), used to measure convergence.

$$0 \le \gamma\,(x,y) := \left(\max_{\hat{x}\in\mathcal{X}}\{\hat{x}^{\intercal}Ay\} - x^{\intercal}Ay\right)\ +\ \left(x^{\intercal}Ay - \min_{\hat{y}\in\mathcal{Y}}\{x^{\intercal}A\hat{y}\}\right)$$

$$\gamma\,(x,y) = \max_{\hat{x}\in\mathcal{X}}\{\hat{x}^{\intercal}Ay\} - \min_{\hat{y}\in\mathcal{Y}}\{x^{\intercal}A\hat{y}\}$$

- **Theorem:** Let $R^T_{\mathcal{X}}$ and $R^T_{\mathcal{Y}}$ be the sublinear cumulative regret of the minimizers, up to time T. And, $\hat{x}^T$ and $\hat{y}^T$ be the average of the strategies. Then the saddle point gap satisfies:

$$\gamma\,(\hat{x}^T,\hat{y}^T) \le \frac{R^T_{\mathcal{X}} + R^T_{\mathcal{Y}}}{T} \to 0 \quad \text{as} \quad T \to 0$$

  (See the lecture notes for the proof.)