MIT 6.S890 — Topics in Multiagent Learning (F23) Mon, September 18th 2021

Lecture 4

Learning in games: Foundations

Instructor: Gabriele Farina*

With this class we begin to explore what it means to "learn" in a game, and how that "learning", which is intrinsically a *dynamic* and *local* (per-player) concept, relates to the much more *static* and *global* concept of game-theoretic equilibrium.

1 Hindsight rationality and Φ-regret

What does it mean to "learn" in games? The answer to this question is delicate. The history of learning in games historically spanned several subfields. A powerful definition for what "learning in games" means is through the concept of *hindsight rationality*.

Take the point of view of one player in a game, and let $\mathcal{X}$ be their set of available strategies. In normal-form games, we have seen that a strategy is just a distribution over the set of available actions S, so $\mathcal{X} = \Delta^S$. At each time $t = 1, 2, \dots$, the player will play some strategy $\boldsymbol{x}^{(t)} \in \mathcal{X}$, receive some form of feedback, and will incorporate that feedback to formulate a "better" strategy $\boldsymbol{x}^{(t+1)} \in \mathcal{X}$ for the next repetition of the game. A typical (and natural) choice of "feedback" is just the utility of the player, given what all the other agents played.

Now suppose that the game is played infinite times, and looking back at what was played by the player we realize that every single time the player played a certain strategy $\boldsymbol{x}$, they would have been strictly better by consistently playing different strategy $\boldsymbol{x}'$ instead. Can we really say that the player has "learnt" how to play? Perhaps not.

That leads to the idea of *hindsight rationality*:

> **Definition 1.1** (Hindsight rationality, informal). The 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 at the whole history of play would have given strictly better utility to the player.

This leads to the following definition.

> **Definition 1.2** (Φ-regret minimizer). Given a set $\mathcal{X}$ of points and a set Φ of linear transformations $\phi : \mathcal{X} \to \mathcal{X}$, a Φ*-regret minimizer for the set* $\mathcal{X}$ is a model for a decision maker that repeatedly interacts with a black-box environment. At each time t, the regret minimizer interacts with the environment through two operations:
>
> - NextStrategy has the effect that the regret minimizer will output an element $\boldsymbol{x}^{(t)} \in \mathcal{X}$;
> - ObserveUtility($u^{(t)}$) provides the environment's feedback to the regret minimizer, in the form of a linear utility function $u^{(t)} : \mathcal{X} \to \mathbb{R}$ that evaluates how good the last-output point $\boldsymbol{x}^{(t)}$ was. The

*MIT EECS. gfarina@mit.edu.

utility function can depend adversarially on the outputs $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(t)}$ if the regret minimizer is deterministic (i.e., does not use randomness internally[a]).

Its quality metric is its cumulative Φ-*regret*, defined as the quantity

$$R_\Phi^{(T)} := \max_{\hat{\phi}\in\Phi}\left\{\sum_{t=1}^{T}\left(u^{(t)}(\hat{\phi}(\boldsymbol{x}^{(t)})) - u^{(t)}(\boldsymbol{x}^{(t)})\right)\right\}, \tag{1}$$

The goal for a Φ-regret minimizer is to guarantee that its Φ-regret grows asymptotically sublinearly as time T increases.

[a] When randomness is involved, the utility function cannot depend adversarially on $\boldsymbol{x}^{(t)}$ or guaranteeing sublinear regret would be impossible. Rather, $u^{(t)}$ must be conditionally independent on $\boldsymbol{x}^{(t)}$, given all past random outcomes.

Calls to NextStrategy and ObserveUtility keep alternating to each other: first, the regret minimizer will output a point $\boldsymbol{x}^{(1)}$, then it will received feedback $u^{(1)}$ from the environment, then it will output a new point $\boldsymbol{x}^{(2)}$, and so on. The decision making encoded by the regret minimizer is *online*, in the sense that at each time t, the output of the regret minimizer can depend on the prior outputs $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(t-1)}$ and corresponding observed utility functions $u^{(1)}, \ldots, u^{(t-1)}$, but no information about future utilities is available.

1.1 Some notable choices for the set of transformations Φ considered

The size of the set of transformations Φ considered by the player defines a natural notion of how "rational" the agent is. There are several choices of interest for Φ for a normal-form strategy space $\mathcal{X} = \Delta^S$.

1. Φ = set of *all* stochastic matrices, mapping $\Delta^S \to \Delta^S$. This notion of Φ-regret is known under the name *swap regret*.

2. Φ = set of all "probability mass transport" on $\mathcal{X}$, defined as

$$\Phi = \{\phi_{a\to b}\}_{a,b\in S}, \qquad \text{where} \quad \phi_{a\to b}(\boldsymbol{x})[s] := \begin{cases} 0 & \text{if } s = a \quad \text{(remove mass from } a\text{...)} \\ \boldsymbol{x}[b] + \boldsymbol{x}[a] & \text{if } s = b \quad \text{(... and give it to } b\text{)} \\ \boldsymbol{x}[s] & \text{otherwise.} \end{cases}$$

   This is known as *internal regret*.

   Theorem 1.1 (Informal, formal version in Appendix B)**.** When all agents in a multiplayer general-sum normal-form game play so that their internal or swap regret grows sublinearly, their average correlated distribution of play converges to the set of *correlated equilibria* of the game.

3. Sneak peak: in sequential games, the above concept extends to Φ = a particular set of linear transformations called *trigger deviation functions*. It is known that in this case the Φ-regret can be efficiently bounded with a polynomial dependence on the size of the game tree. The reason why this choice of deviation functions is important is given by the following fact.

   Theorem 1.2 (Informal)**.** When all agents in a multiplayer general-sum extensive-form game play so that their Φ-regret relative to trigger deviation functions grows sublinearly, their average correlated distribution of play converges to the set of *extensive-form correlated equilibria* of the game.

4. Φ = constant transformations. In this case, we are only requiring that the player not regret substituting *all* of the strategies they played with the *same* strategy $\hat{\boldsymbol{x}} \in \Delta^S$. Φ-regret according to this set of transformations Φ is usually called *external* regret, or more simply just *regret.* While this seems like an extremely restricted notion of rationality, it actually turns out to be already extremely powerful. We will spend the rest of this class to see why.

Theorem 1.3 (Informal)**.** When all agents in a multiplayer general-sum normal-form game play so that their external regret grows sublinearly, their average correlated distribution of play converges to the set of *coarse correlated equilibrium* of the game.

Theorem 1.4 (Informal)**.** When all agents in a two-player zero-sum normal-form game play so that their external regret grows sublinearly, their average strategies converge to the set of *Nash equilibria* of the game.

1.2 A very important special case: regret minimization

The special case where Φ is chosen to be the set of constant transformations is so important that it warrants its own special definition and notation.

Definition 1.3 (Regret minimizer)**.** Let $\mathcal{X}$ be a set. An *external regret minimizer for* $\mathcal{X}$—or simply *"regret minimizer for* $\mathcal{X}$*"*—is a Φ^{const}-regret minimizer for the special set of *constant* transformations

$$\Phi^{\text{const}} := \{\phi_{\hat{\boldsymbol{x}}} : \boldsymbol{x} \mapsto \hat{\boldsymbol{x}}\}_{\hat{\boldsymbol{x}} \in \mathcal{X}}.$$

Its corresponding Φ^{const}-regret is called "*external regret*" or simply "*regret*", and it is indicated with the symbol

$$R^{(T)} := \max_{\hat{\boldsymbol{x}} \in \mathcal{X}} \left\{ \sum_{t=1}^{T} \left(u^{(t)}(\hat{\boldsymbol{x}}) - u^{(t)}(\boldsymbol{x}^{(t)}) \right) \right\}. \tag{2}$$

Once again, the goal for a regret minimizer is to have its cumulative regret R^T grow sublinearly in T.

An important result in the subfield of *online linear optimization* asserts the existence of algorithms that guarantee sublinear regret for any convex and compact domain $\mathcal{X}$, typically of the order $R^T = O(\sqrt{T})$ asymptotically.

As it turns out, external regret minimization alone is enough to guarantee convergence to Nash equilibrium in two-player zero-sum games, to coarse correlated equilibrium in multiplayer general-sum games, to best responses to static stochastic opponents in multiplayer general-sum games, and much more. Before we delve into those aspects, however, we first show another important property of regret minimization: general Φ-regret minimization can reduced to it, in a precise sense.

1.3 From regret minimization to Φ-regret minimization

As we have seen, regret minimization is a very narrow instantiation of Φ-regret minimization—perhaps the smallest sensible instantiation. Then, clearly, the problem of coming up with a regret minimizer for a set $\mathcal{X}$ cannot be harder than the problem of coming up with a Φ-regret minimizer for $\mathcal{X}$ for richer sets of transformation functions Φ. It might then seem surprising that there exists a construction that reduces Φ-regret minimization to regret minimization.

More precisely, a result by Gordon et al. [2008] gives a way to construct a Φ-regret minimizer for $\mathcal{X}$ starting from any regret minimizer *for the set of functions* Φ. The result goes as follows.

Theorem 1.5 (Gordon et al. [2008]). Let $\mathcal{R}$ be a deterministic regret minimizer for the set of transformations Φ whose (external) cumulative regret $R^{(T)}$ grows sublinearly in T, and assume that every $\phi \in \Phi$ admits a fixed point $\phi(\boldsymbol{x}) = \boldsymbol{x} \in \mathcal{X}$. Then, a Φ-regret minimizer $\mathcal{R}_\Phi$ can be constructed starting from $\mathcal{R}$ as follows:

- Each call to $\mathcal{R}_\Phi$.NextStrategy first calls NextStrategy on $\mathcal{R}$ to obtain the next transformation $\phi^{(t)}$. Then, a fixed point $\boldsymbol{x}^{(t)} = \phi^{(t)}(\boldsymbol{x}^{(t)})$ is computed and output.
- Each call to $\mathcal{R}_\Phi$.ObserveUtility$(u^{(t)})$ with linear utility function $u^{(t)}$ constructs the linear utility function $L^{(t)} : \phi \mapsto u^{(t)}(\phi(\boldsymbol{x}^{(t)}))$, where $\boldsymbol{x}^{(t)}$ is the last-output strategy, and passes it to $\mathcal{R}$ by calling $\mathcal{R}$.ObserveUtility$(L^{(t)})$.[b]

Furthermore, the Φ-regret $R_\Phi^{(T)}$ cumulated up to time T by $\mathcal{R}_\Phi$ we have just defined is exactly equal to the (external) cumulative regret $R^{(T)}$ cumulated by $\mathcal{R}$:

$$R_\Phi^{(T)} = R^{(T)} \qquad \forall\, T = 1, 2, \dots.$$

So, because the regret cumulated by $\mathcal{R}$ grows sublinearly by hypothesis, then so does the Φ-regret cumulated by $\mathcal{R}_\Phi$.

[b] On the surface, it might look like $L^{(t)}$ is independent on the last output $\phi^{(t)}$ of the regret minimizer $\mathcal{R}$, and thus, that it trivially satisfies the requirements of Definition 1.3. However, that is not true: x^t is a fixed point of ϕ^t, and since x^t enters into the definition of L^t, if $\mathcal{R}$ picks ϕ^t randomly, it might very well be that L^t is not conditionally independent on ϕ^t. We sidestep this issue by requiring that $\mathcal{R}$ is deterministic (cf. Footnote *a*).

Proof. The proof of correctness of the above construction is deceptively simple. Since $\mathcal{R}$ outputs transformations $\phi^{(1)}, \phi^{(2)}, \dots \in \Phi$ and receives utilities $\phi \mapsto u^{(1)}(\phi(\boldsymbol{x}^{(1)})), \phi \mapsto u^{(2)}(\phi(\boldsymbol{x}^{(2)})), \dots$, its cumulative regret $R^{(T)}$ is by definition

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

Now, since by construction $\boldsymbol{x}^{(t)}$ is a fixed point of $\phi^{(t)}$, $\phi^{(t)}(\boldsymbol{x}^{(t)}) = \boldsymbol{x}^{(t)}$, and therefore we can write

$$R^{(T)} = \max_{\hat\phi \in \Phi} \left\{ \sum_{t=1}^{T} \left(u^{(t)}(\hat\phi(\boldsymbol{x}^{(t)})) - u^{(t)}(\boldsymbol{x}^{(t)}) \right) \right\}, \tag{3}$$

where the right-hand side is exactly the cumulative Φ-regret $R_\Phi^{(T)}$ incurred by $\mathcal{R}_\Phi$, as defined in (2). $\square$

2 Applications of regret minimization

In order to establish regret minimization as a meaningful abstraction for learning in games, we must check that regret minimizing and Φ-regret minimizing dynamics indeed lead to "interesting" or expected behavior in common situations.

2.1 Learning a best response against stochastic opponents

As a first smoke test, let's verify that over time a regret minimizer would learn how to best respond to static, stochastic opponents. Specifically, consider this scenario. We are playing a repeated n-player general-sum game with multilinear utilities (this captures normal-form game and extensive-form games alike), where

Players $i = 1, \dots, n-1$ play stochastically, that is, at each t they independently sample a strategy $\boldsymbol{x}_i^{(t)} \in \mathcal{X}_i$ from the same fixed distribution (which is unknown to any other player). Formally, this means that

$$\mathbb{E}[\boldsymbol{x}_i^{(t)}] = \bar{\boldsymbol{x}}_i \qquad \forall\, i = 1, \dots, n-1, \quad t = 1, 2, \dots.$$

Player n, on the other hand, is learning in the game, picking strategies according to some algorithm that guarantees sublinear external regret, where the feedback observed by Player n at each time t is their own linear utility function:

$$\boldsymbol{u}^{(t)} \coloneqq \mathcal{X}_n \ni \boldsymbol{x}_n \mapsto u_n(\boldsymbol{x}_1^{(t)}, \dots, \boldsymbol{x}_{n-1}^{(t)}, \boldsymbol{x}_n).$$

Then, the average of the strategies played by Player n converges almost surely to a best response to $\bar{\boldsymbol{x}}_1, \dots, \bar{\boldsymbol{x}}_{n-1}$, that is,

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_n^{(t)} \quad \xrightarrow{\text{a.s.}} \quad \operatorname*{arg\,max}_{\hat{\boldsymbol{x}}_n \in \mathcal{X}_n} \Big\{ u_n(\bar{\boldsymbol{x}}_1, \dots, \bar{\boldsymbol{x}}_{n-1}, \hat{\boldsymbol{x}}_n) \Big\}.$$

(You should try to prove this!)

2.2 Self-play convergence to bilinear saddle points (such as a Nash equilibrium in a two-player zero-sum game)

It turns out that regret minimization can be used to converge to bilinear saddle points, that is solutions to problems of the form

$$\max_{\boldsymbol{x} \in \mathcal{X}} \min_{\boldsymbol{y} \in \mathcal{Y}} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{y}, \tag{4}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are convex compact sets and $\boldsymbol{A}$ is a matrix. These types of optimization problems are pervasive in game-theory. The canonical prototype of bilinear saddle point problem is the computation of Nash equilibria in two-player zero-sum games (either normal-form or extensive-form). There, a Nash equilibrium is the solution to (4) where $\mathcal{X}$ and $\mathcal{Y}$ are the strategy spaces of Player 1 and Player 2 respectively (probability simplexes for normal-form games or sequence-form polytopes for extensive-form games), and $\boldsymbol{A}$ is the payoff matrix for Player 1. Other examples include social-welfare-maximizing correlated equilibria and optimal strategies in two-team zero-sum adversarial team games.

The idea behind using regret minimization to converge to bilinear saddle-point problems is to use *self play*. We instantiate two regret minimization algorithms, $\mathcal{R}_\mathcal{X}$ and $\mathcal{R}_\mathcal{Y}$, for the domains of the maximization and minimization problem, respectively. At each time t the two regret minimizers output strategies $\boldsymbol{x}^{(t)}$ and $\boldsymbol{y}^{(t)}$, respectively. Then, they receive feedback $u_\mathcal{X}^{(t)}, u_\mathcal{Y}^{(t)}$ defined as

$$u_\mathcal{X}^{(t)} : \boldsymbol{x} \mapsto (\boldsymbol{A}\boldsymbol{y}^{(t)})^\top \boldsymbol{x}, \qquad u_\mathcal{Y}^{(t)} : \boldsymbol{y} \mapsto -(\boldsymbol{A}^\top \boldsymbol{x}^{(t)})^\top \boldsymbol{y}.$$

We summarize the process pictorially in Figure 1.

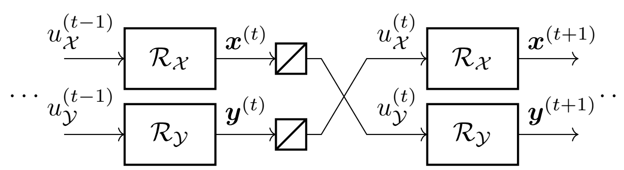

Figure 1: The flow of strategies and utilities in regret minimization for games. The symbol ⧄ denotes computation/construction of the utility function.

A well known folk theorem establish that the pair of average strategies produced by the regret minimizers up to any time T converges to a saddle point of (4), where convergence is measured via the *saddle point gap*

$$0 \le \gamma(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \left(\max_{\hat{\boldsymbol{x}} \in \mathcal{X}}\{\hat{\boldsymbol{x}}^\top \boldsymbol{A}\boldsymbol{y}\} - \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y}\right) + \left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y} - \min_{\hat{\boldsymbol{y}} \in \mathcal{Y}}\{\boldsymbol{x}^\top \boldsymbol{A}\hat{\boldsymbol{y}}\}\right) = \max_{\hat{\boldsymbol{x}} \in \mathcal{X}}\{\hat{\boldsymbol{x}}^\top \boldsymbol{A}\boldsymbol{y}\} - \min_{\hat{\boldsymbol{x}} \in \mathcal{X}}\{\boldsymbol{x}^\top \boldsymbol{A}\hat{\boldsymbol{y}}\}.$$

A point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$ has zero saddle point gap if and only if it is a solution to (4).

Theorem 2.1. Consider the self-play setup summarized in Figure 1, where $\mathcal{R}_\mathcal{X}$ and $\mathcal{R}_\mathcal{Y}$ are regret minimizers for the sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $R_\mathcal{X}^{(T)}$ and $R_\mathcal{Y}^{(T)}$ be the (sublinear) regret cumulated by $\mathcal{R}_\mathcal{X}$ and $\mathcal{R}_\mathcal{Y}$, respectively, up to time T, and let $\bar{\boldsymbol{x}}^{(T)}$ and $\bar{\boldsymbol{y}}^{(T)}$ denote the average of the strategies produced up to time T. Then, the saddle point gap $\gamma(\bar{\boldsymbol{x}}^{(T)}, \bar{\boldsymbol{y}}^{(T)})$ of $(\bar{\boldsymbol{x}}^{(T)}, \bar{\boldsymbol{y}}^{(T)})$ satisfies

$$\gamma(\bar{\boldsymbol{x}}^{(T)}, \bar{\boldsymbol{y}}^{(T)}) \le \frac{R_\mathcal{X}^{(T)} + R_\mathcal{Y}^{(T)}}{T} \to 0 \qquad \text{as } T \to \infty.$$

Proof. By definition of regret,

$$\begin{aligned}
\frac{R_\mathcal{X}^{(T)} + R_\mathcal{Y}^{(T)}}{T} &= \frac{1}{T}\max_{\hat{\boldsymbol{x}} \in \mathcal{X}}\left\{\sum_{t=1}^T u_\mathcal{X}^{(t)}(\hat{\boldsymbol{x}})\right\} - \frac{1}{T}\sum_{t=1}^T u_\mathcal{X}^{(t)}(\boldsymbol{x}^t) + \frac{1}{T}\max_{\hat{\boldsymbol{y}} \in \mathcal{Y}}\left\{\sum_{t=1}^T u_\mathcal{Y}^{(t)}(\hat{\boldsymbol{y}})\right\} - \frac{1}{T}\sum_{t=1}^T u_\mathcal{Y}^{(t)}(\boldsymbol{y}^t) \\
&= \frac{1}{T}\max_{\hat{\boldsymbol{x}} \in \mathcal{X}}\left\{\sum_{t=1}^T u_\mathcal{X}^{(t)}(\hat{\boldsymbol{x}})\right\} + \frac{1}{T}\max_{\hat{\boldsymbol{y}} \in \mathcal{Y}}\left\{\sum_{t=1}^T u_\mathcal{Y}^{(t)}(\hat{\boldsymbol{y}})\right\} \qquad (\text{since } u_\mathcal{X}^{(t)}(\boldsymbol{x}^{(t)}) + u_\mathcal{Y}^{(t)}(\boldsymbol{y}^{(t)}) = 0) \\
&= \frac{1}{T}\max_{\hat{\boldsymbol{x}} \in \mathcal{X}}\left\{\sum_{t=1}^T \hat{\boldsymbol{x}}^\top \boldsymbol{A}\boldsymbol{y}^{(t)}\right\} + \frac{1}{T}\max_{\hat{\boldsymbol{y}} \in \mathcal{Y}}\left\{\sum_{t=1}^T -(\boldsymbol{x}^{(t)})^\top \boldsymbol{A}\hat{\boldsymbol{y}}\right\} \\
&= \max_{\hat{\boldsymbol{x}} \in \mathcal{X}}\left\{\hat{\boldsymbol{x}}^\top \boldsymbol{A}\bar{\boldsymbol{y}}^{(T)}\right\} - \min_{\hat{\boldsymbol{y}} \in \mathcal{Y}}\left\{(\bar{\boldsymbol{x}}^{(T)})^\top \boldsymbol{A}\hat{\boldsymbol{y}}\right\} = \gamma(\bar{\boldsymbol{x}}^{(T)}, \bar{\boldsymbol{y}}^{(T)}).
\end{aligned}$$

□

A Appendix: Recap of notation from past classes

Definition A.1. A finite n-player normal-form game is described by:
- a set of actions (aka. pure strategies) for each player i, denoted S_i;
- a utility/payoff function for each player i: $u_i : \times_j S_j \to \mathbb{R}$
  (Remark: this can be thought of as a tensor).

Definition A.2. A *randomized/mixed strategy* for player i is any $\boldsymbol{x}_i \in \Delta^{S_i}$.

Definition A.3. Player i's *expected utility* is

$$u_i(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \mathop{\mathbb{E}}_{\substack{s_1 \sim \boldsymbol{x}_1 \\ \cdots \\ s_n \sim \boldsymbol{x}_n}} u_i(s_1, \ldots, s_n) = \sum_{s_1 \in S_1} \cdots \sum_{s_n \in S_n} \boldsymbol{x}_1[s_1] \cdots \boldsymbol{x}_n[s_n] \cdot u_i(s_1, \ldots, s_n).$$

Definition A.4 (Correlated equilibrium)**.** An *correlated equilibrium* is a joint distribution $D(s_1, \ldots, s_n)$ such that for every player i, and every pair of pure strategies a and b such that s_i is sampled with nonzero probability,

$$\mathop{\mathbb{E}}_{s_{-i}\sim D(\cdot|s_i=a)} u_i(a, s_{-i}) \geq \mathop{\mathbb{E}}_{s_{-i}\sim D(\cdot|s_i=a)} u_i(b, s_{-i}),$$

or equivalently (by multiplying by $\mathbb{P}_D[s_i = a]$),

$$\mathop{\mathbb{E}}_{s\sim D}\Big[\big(u_i(b, s_{-i}) - u_i(a, s_{-i})\big) \cdot \mathbb{1}[s_i = a]\Big] \leq 0.$$

Definition A.5 (Approximate correlated equilibrium)**.** D is an ϵ-approximate correlated equilibrium if

$$\mathop{\mathbb{E}}_{s\sim D}\Big[\big(u_i(b, s_{-i}) - u_i(a, s_{-i})\big) \cdot \mathbb{1}[s_i = a]\Big] \leq \epsilon.$$

B Appendix: Proof of Theorem 1.1

Theorem B.1 (Formal version of Theorem 1.1)**.** Let $\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_n^{(t)}$ the strategies played by the players at any time t, and let $\mathrm{IntReg}_i^{(t)}$ denote the internal regret incurred by Player i up to time t. Consider now the average correlated distribution of play up to any time T, that is, the distribution $D^{(T)}$ that selects a time $\bar{t}$ uniformly at random from the set $\{1, \ldots, T\}$, and selects actions $(s_1, \ldots, s_n)$ independendently according to the $\boldsymbol{x}_i^{(\bar{t})}$. This distribution is an $\epsilon^{(T)}$-correlated equilibrium, where

$$\epsilon^{(T)} := \max_{i\in\{1,\ldots,n\}} \frac{\mathrm{IntReg}_i^{(T)}}{T}$$

and IntReg denotes the internal regret of player i.

Proof. Pick any player i and strategies $a, b \in S_i$. The maximum benefit that player i can obtain by deviating from action a to action b is given by,

$$\mathop{\mathbb{E}}_{s\sim D^{(T)}}\Big[\big(u_i(b, s_{-i}) - u_i(a, s_{-i})\big) \cdot \mathbb{1}[s_i = a]\Big].$$

Expanding the specific structure of $D^{(T)}$, we can decompose the expectation as

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{s\sim \boldsymbol{x}_1^{(t)}\otimes\cdots\otimes\boldsymbol{x}_n^{(t)}}\Big[\big(u_i(b, s_{-i}) - u_i(a, s_{-i})\big) \cdot \mathbb{1}[s_i = a]\Big] \\
&\qquad = \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{s_i\sim \boldsymbol{x}_i^{(t)}} \mathop{\mathbb{E}}_{s_{-i}\sim \otimes \boldsymbol{x}_{-i}^{(t)}}\Big[\big(u_i(b, s_{-i}) - u_i(a, s_{-i})\big) \cdot \mathbb{1}[s_i = a]\Big] \\
&\qquad = \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{s_i\sim \boldsymbol{x}_i^{(t)}}\Big[\big(u_i(b, \boldsymbol{x}_{-i}^{(t)}) - u_i(a, \boldsymbol{x}_{-i}^{(t)})\big) \cdot \mathbb{1}[s_i = a]\Big]
\end{aligned}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \Big(\big(u_i(b, \boldsymbol{x}_{-i}^{(t)}) - u_i(a, \boldsymbol{x}_{-i}^{(t)})\big) \cdot \boldsymbol{x}_i^{(t)}[a] \Big).$$

At this point, the proof is concluded if we can show that the argument in the square brackets is upper bounded by $\mathrm{IntReg}_i^{(T)}$. This is easy by considering the deviation $\phi_{a \to b}$ defined above:

$$\phi_{a \to b}(\boldsymbol{x})[s] := \begin{cases} 0 & \text{if } s = a \quad \text{(remove mass from } a\text{...)} \\ \boldsymbol{x}[b] + \boldsymbol{x}[a] & \text{if } s = b \quad \text{(... and give it to } b\text{)} \\ \boldsymbol{x}[s] & \text{otherwise.} \end{cases}$$

In particular, by definition of internal regret,

$$\begin{aligned} \mathrm{IntReg}_i^{(T)} &\geq \sum_{t=1}^{T} \Big(u_i(\phi_{a \to b}(\boldsymbol{x}_i^{(t)}), \boldsymbol{x}_{-i}^{(t)}) - u_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{x}_{-i}^{(t)}) \Big) \\ &= \sum_{t=1}^{T} \sum_{s_i \sim \boldsymbol{x}_i^{(t)}} \big(\phi_{a \to b}(\boldsymbol{x}_i^{(t)})[s_i] - \boldsymbol{x}_i^{(t)}[s_i]\big) \cdot u_i(s_i, \boldsymbol{x}_{-i}^{(t)}) \\ &= \sum_{t=1}^{T} \Big(- \boldsymbol{x}_i^{(t)}[a] \cdot u_i(a, \boldsymbol{x}_{-i}^{(t)}) + \boldsymbol{x}_i[a] \cdot u_i(b, \boldsymbol{x}_{-i}^{(t)}) \Big) \\ &= \sum_{t=1}^{T} \Big(\big(u_i(b, \boldsymbol{x}_{-i}^{(t)}) - u_i(a, \boldsymbol{x}_{-i}^{(t)})\big) \cdot \boldsymbol{x}_i^{(t)}[a] \Big). \end{aligned}$$

So, the maximum benefit of any deviation of player i is bounded above by $\frac{1}{T}\mathrm{IntReg}_i^{(T)}$. It follows that the maximum deviation across any player is at most $\epsilon^{(T)}$ as defined in the statement. $\square$

References

Geoffrey J Gordon, Amy Greenwald, and Casey Marks. No-regret learning in convex games. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 360–367. ACM, 2008.