MIT 6.S890 — Topics in Multiagent Learning Tue, Sep 17th 2024

Lecture 4
Learning in games: Foundations

Instructor: Prof. Gabriele Farina (✉ gfarina@mit.edu)[⋆]

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? Multiple answers are correct. However, today we focus on a powerful answer through the concept of *hindsight rationality*.

Take the point of view of one player in a game, and denote with $\mathcal{X}$ 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 U_1, so $\mathcal{X} = \Delta^A$. At each time $t = 1, 2, ...$, the player will play some strategy $x^{(t)} \in \mathcal{X}$, receive some form of feedback, and will incorporate that feedback to formulate a "better" strategy $x^{(t+1)} \in \mathcal{X}$ 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 x, they would have been strictly better by consistently playing different strategy x' instead. Can we really say that the player has "learnt" how to play? Perhaps not. This concept goes under the name of *hindsight rationality*:

Definition 1.1 (Hindsight rationality, informal). The player has "learnt" to play the game if 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.

We have thus arrived to the following formalization.

Definition 1.2 (Φ-regret minimizer). Given the strategy set $\mathcal{X}$ 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 $x^{(t)} \in \mathcal{X}$;
- `ObserveUtility`$\left(u^{(t)}\right)$ 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 $x^{(t)}$ was. The utility function can depend adversarially on the outputs $x^{(1)}, ..., x^{(t)}$ if the regret minimizer is deterministic (*i.e.*, does not use randomness internally[1]).

[⋆] These notes are class material that has not undergone formal peer review. The TAs and I are grateful for any reports of typos.

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

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

The goal for a Φ-regret minimizer is to guarantee that its Φ-regret grows asymptotically sublinearly as time T increases, no matter the sequence of utility functions $u^{(t)}$.

Calls to `NextStrategy` and `ObserveUtility` keep alternating to each other: first, the regret minimizer will output a point $x^{(1)}$, then it will received feedback $u^{(1)}$ from the environment, then it will output a new point $x^{(2)}$, 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 $x^{(1)}, ..., x^{(t-1)}$ and corresponding observed utility functions $u^{(1)}, ..., u^{(t-1)}$, 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 $\mathcal{X} = \Delta^A$.

- $\Phi =$ set of *all* stochastic matrices, mapping $\Delta^A \to \Delta^A$. This notion of Φ-regret is known under the name *swap regret*. This notion is related to convergence to the set of correlated equilibria.
- $\Phi =$ set of all "probability mass transport" on $\mathcal{X}$, defined as

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

  This is known as *internal regret.*

  > **Theorem 1.1** (Informal; formal version in Theorem 2.3). 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.

  In sequential games, the above concept extends to $\Phi =$ a particular set of linear transformations called *trigger deviation functiona.* 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.

- $\Phi =$ 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 $\hat{x} \in \Delta^A$. 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.

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

Theorem 1.3 (Informal; formal version in Theorem 2.3). 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.

Corollary 1.1 (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 An 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

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

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

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

Once again, the goal for a regret minimizer is to have its cumulative regret Reg^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 $\text{Reg}^T = O\left(\sqrt{T}\right)$ 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.

■ **Teaser: 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. We will talk more about it in Lecture 8.

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

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

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:

$$\boldsymbol{u}^{(t)} := \mathcal{X}_n \ni x_n \mapsto u_n\Big(x_1^{(t)}, ..., x_{n-1}^{(t)}, x_n\Big).$$

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

$$\frac{1}{T}\sum_{t=1}^{T} x_n^{(t)} \quad \xrightarrow{\text{a.s.}} \quad \arg\max_{\hat{x}_n \in \mathcal{X}_n}\{u_n(\overline{x}_1, ..., \overline{x}_{n-1}, \hat{x}_n)\}.$$

(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

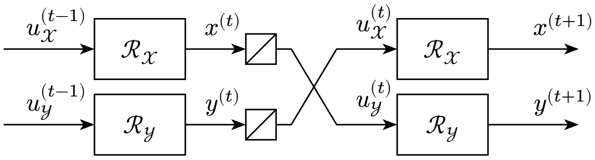
$$\max_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} x^\top \mathrm{U}_1 y, \tag{1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are convex compact sets and U_1 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 (1) 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 U_1 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 $x^{(t)}$ and $y^{(t)}$, respectively. Then, they receive feedback $u_\mathcal{X}^{(t)}, u_\mathcal{Y}^{(t)}$ defined as

$$u_\mathcal{X}^{(t)} : x \mapsto \big(\mathrm{U}_1 y^{(t)}\big)^\top x, \qquad u_\mathcal{Y}^{(t)} : y \mapsto -\Big({\mathrm{U}_1}^\top x^{(t)}\Big)^\top y.$$

We can summarize the process pictorially as follows.



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 (1), where convergence is measured via the *saddle point gap*

$$0 \le \gamma(x,y) := \left(\max_{\hat{x}\in\mathcal{X}}\{\hat{x}^\top \mathrm{U}_1 y\} - x^\top \mathrm{U}_1 y\right) + \left(x^\top \mathrm{U}_1 y - \min_{\hat{y}\in\mathcal{Y}}\{x^\top \mathrm{U}_1 \hat{y}\}\right) = \max_{\hat{x}\in\mathcal{X}}\{\hat{x}^\top \mathrm{U}_1 y\} - \min_{\hat{x}\in\mathcal{X}}\{x^\top \mathrm{U}_1 \hat{y}\}.$$

A point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ has zero saddle point gap if and only if it is a solution to (1).

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

$$\gamma\Big(\overline{x}^{(T)}, \overline{y}^{(T)}\Big) \le \frac{\text{Reg}_{\mathcal{X}}^{(T)} + \text{Reg}_{\mathcal{Y}}^{(T)}}{T} \to 0 \qquad \text{as } T \to \infty.$$

Proof. By definition of regret,

$$\begin{aligned}
\frac{\text{Reg}_{\mathcal{X}}^{(T)} + \text{Reg}_{\mathcal{Y}}^{(T)}}{T} &= \frac{1}{T}\max_{\hat{x}\in\mathcal{X}}\left\{\sum_{t=1}^{T} u_{\mathcal{X}}^{(t)}(\hat{x})\right\} - \frac{1}{T}\sum_{t=1}^{T} u_{\mathcal{X}}^{(t)}(x^{t}) + \frac{1}{T}\max_{\hat{y}\in\mathcal{Y}}\left\{\sum_{t=1}^{T} u_{\mathcal{Y}}^{(t)}(\hat{y})\right\} - \frac{1}{T}\sum_{t=1}^{T} u_{\mathcal{Y}}^{(t)}(y^{t}) \\
&= \frac{1}{T}\max_{\hat{x}\in\mathcal{X}}\left\{\sum_{t=1}^{T} u_{\mathcal{X}}^{(t)}(\hat{x})\right\} + \frac{1}{T}\max_{\hat{y}\in\mathcal{Y}}\left\{\sum_{t=1}^{T} u_{\mathcal{Y}}^{(t)}(\hat{y})\right\} \qquad \Big(\text{since } u_{\mathcal{X}(x^{(t)})}^{(t)} + u_{\mathcal{Y}(y^{(t)})}^{(t)} = 0\Big) \\
&= \frac{1}{T}\max_{\hat{x}\in\mathcal{X}}\left\{\sum_{t=1}^{T} \hat{x}^{\top}\mathrm{U}_1 y^{(t)}\right\} + \frac{1}{T}\max_{\hat{y}\in\mathcal{Y}}\left\{\sum_{t=1}^{T} -\big(x^{(t)}\big)^{\top}\mathrm{U}_1\hat{y}\right\} \\
&= \max_{\hat{x}\in\mathcal{X}}\Big\{\hat{x}^{\top}\mathrm{U}_1\overline{y}^{(T)}\Big\} - \min_{\hat{y}\in\mathcal{Y}}\Big\{\Big(\overline{x}^{(T)}\Big)^{\top}\mathrm{U}_1\hat{y}\Big\} = \gamma\Big(\overline{x}^{(T)}, \overline{y}^{(T)}\Big).
\end{aligned}$$

□

2.3 Proof of the minimax theorem

The very existence of regret minimizers is a powerful enough fact to imply the minimax theorem!

Theorem 2.2 (Minimax theorem). Let $\mathcal{X}$ and $\mathcal{Y}$ be convex compact sets, and let U_1 be a matrix. Suppose that a regret minimizer $\mathcal{R}_{\mathcal{X}}$ for set $\mathcal{X}$ guaranteeing sublinear regret no matter the sequence of utilities can be constructed. Then,

$$\max_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}} x^{\top}\mathrm{U}_1 y = \min_{y\in\mathcal{Y}}\max_{x\in\mathcal{X}} x^{\top}\mathrm{U}_1 y.$$

Proof. One direction of the equality, specifically

$$\max_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}} x^{\top}\mathrm{U}_1 y \le \min_{y\in\mathcal{Y}}\max_{x\in\mathcal{X}} x^{\top}\mathrm{U}_1 y,$$

follows from definition (this is often called *weak duality*).

To show the reverse inequality, we will interpret the bilinear saddle point $\min_{\{y\in\mathcal{Y}\}}\max_{\{x\in\mathcal{X}\}} x^{\top}\mathrm{U}_1 y$ as a repeated game. At each time t, we will let a regret minimizer $\mathcal{R}_{\mathcal{X}}$ pick actions $x^{(t)} \in \mathcal{X}$, whereas we will always assume that $y^{(t)} \in \mathcal{Y}$ is chosen by the environment to best respond to $x^{(t)}$, that is,

$$y^{(t)} \in \arg\min_{y\in\mathcal{Y}} \big(x^{(t)}\big)^{\top}\mathrm{U}_1 y.$$

The utility function observed by $\mathcal{R}_{\mathcal{X}}$ at each time t is set to the linear function

$$u_{\mathcal{X}}^{(t)}(x) = x^{\top}\mathrm{U}_1 y^{(t)}.$$

Letting $\overline{x}^{(T)} \in \mathcal{X}$ and $\overline{y}^{(T)} \in \mathcal{Y}$ be the average strategies output up to time T, that is,

$$\overline{x}^{(T)} := \frac{1}{T}\sum_{t=1}^{T} x^{(t)} \qquad \overline{y}^{(T)} := \frac{1}{T}\sum_{t=1}^{T} y^{(t)},$$

then we have

$$\max_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}} x^\top \mathrm{U}_1 y \geq \min_{y\in\mathcal{Y}}\left\{\left(\overline{x}^{(T)}\right)^\top \mathrm{U}_1 y\right\} = \frac{1}{T}\min_{y\in\mathcal{Y}}\sum_{t=1}^{T}\left(x^{(t)}\right)^\top \mathrm{U}_1 y \geq \frac{1}{T}\sum_{t=1}^{T}\left(x^{(t)}\right)^\top \mathrm{U}_1 y^{(t)}.$$

The important insight is that the right-hand side can be related to the regret incurred on $\mathcal{X}$: by definition,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(x^{(t)}\right)^\top \mathrm{U}_1 y^{(t)} &= -\frac{\mathrm{Reg}_{\mathcal{X}}^{(T)}}{T} + \frac{1}{T}\max_{x\in\mathcal{X}}\left\{\sum_{t=1}^{T} x^T \mathrm{U}_1 y^{(t)}\right\} \\ &= -\frac{\mathrm{Reg}_{\mathcal{X}}^{(T)}}{T} + \max_{x\in\mathcal{X}} x^T \mathrm{U}_1 \overline{y}^{(T)} \\ &\geq -\frac{\mathrm{Reg}_{\mathcal{X}}^{(T)}}{T} + \min_{y\in\mathcal{Y}}\max_{x\in\mathcal{X}} x^T \mathrm{U}_1 y\end{aligned}$$

Combining the expressions, we obtain

$$\max_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}} x^\top \mathrm{U}_1 y \geq \min_{y\in\mathcal{Y}}\max_{x\in\mathcal{X}} x^\top \mathrm{U}_1 y - \frac{\mathrm{Reg}_{\mathcal{X}}^{(T)}}{T}.$$

Letting $T \to \infty$ proves the result. □

2.4 Convergence to the set of correlated and coarse-correlated equilibria

The previous result is in fact a direct corollary of the more general connection between Φ-regret minimization and the set of coarse-correlated equilibria in multiplayer general-sum games. We present a general form of this connection in the next theorem.

Theorem 2.3 (Formal version of Theorems 1.1 and 1.3). Let $x_1^{(t)}, ..., x_n^{(t)}$ the strategies played by the players at any time t, and let $\Phi\text{-Reg}_i^{(t)}$ 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 $\mu^{(T)}$ that selects a time $\bar{t}$ uniformly at random from the set $\{1, ..., T\}$, and selects actions $(a_1, ..., a_n)$ independently according to the $x_i^{(\bar{t})}$, that is,

$$\mu^{(T)} := \frac{1}{T}\sum_{t=1}^{T} x_1^{(t)} \otimes ... \otimes x_n^{(t)}.$$

This distribution satisfies the inequality

$$\max_{\phi\in\Phi} \mathbb{E}_{a\sim\mu^{(T)}}[u_i(\phi(a_i), a_{-i}) - u_i(a_i, a_{-i})] \leq \frac{\Phi\text{-Reg}_i^{(T)}}{T}.$$

Proof. Pick an arbitrary $\phi \in \Phi$. With the usual slight abuse of notation, we will denote with $\phi(a)$, where a is an action, as the strategy returned by ϕ when evaluated in the *deterministic* strategy that places all the mass on a. Expanding the specific structure of $\mu^{(T)}$, we can decompose the expectation

$$\mathbb{E}_{a\sim\mu^{(T)}}[u_i(\phi(a_i), a_{-i}) - u_i(a_i, a_{-i})]$$

as

$$
\begin{aligned}
&\mathbb{E}_{a\sim\mu^{(T)}}[u_i(\phi(a_i),a_{-i})-u_i(a_i,a_{-i})]\\
&\qquad=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{a\sim x_1^{(t)}\otimes\ldots\otimes x_n^{(t)}}[(u_i(\phi(a_i),a_{-i})-u_i(a_i,a_{-i}))]\\
&\qquad=\frac{1}{T}\sum_{t=1}^{T}\left(u_i\left(\mathbb{E}_{a_i\sim x_i^{(t)}}\phi(a_i),\mathbb{E}_{a_{-i}\sim\otimes x_{-i}^{(t)}}a_{-i}\right)-u_i\left(\mathbb{E}_{a_i\sim x_i^{(t)}}a_i,\mathbb{E}_{a_{-i}\sim\otimes x_{-i}^{(t)}}a_{-i}\right)\right)\\
&\qquad=\frac{1}{T}\sum_{t=1}^{T}\left(u_i\left(\phi\left(\mathbb{E}_{a_i\sim x_i^{(t)}}a_i\right),\mathbb{E}_{a_{-i}\sim\otimes x_{-i}^{(t)}}a_{-i}\right)-u_i\left(\mathbb{E}_{a_i\sim x_i^{(t)}}a_i,\mathbb{E}_{a_{-i}\sim\otimes x_{-i}^{(t)}}a_{-i}\right)\right)\\
&\qquad=\frac{1}{T}\sum_{t=1}^{T}\left(u_i\left(\phi\left(x_i^{(t)}\right),x_{-i}^{(t)}\right)-u_i\left(x_i^{(t)},x_{-i}^{(t)}\right)\right)
\end{aligned}
$$

where the second equality follows by linearity of ϕ and u_i. Taking now a maximum over $\phi \in \Phi$, and recognizing the definition of Φ-regret on the right-hand side, we obtain the desired inequality. □

Note that Theorem 2.3 holds for any set Φ. The approximate equilibria found this way are sometimes called approximate Φ-equilibria. In the special cases of $\Phi =$ all constant transformations, it is clear that the previous result implies convergence to the set of coarse correlated equilibria. For correlated equilibria, we need to convince ourselves that any arbitrary mapping $A \to A$ can be represented via a stochastic matrix. This is indeed the case, by constructing the matrix whose columns indicate what action is assigned to each action in A by the mapping. (You should convince yourself!) Finally, for the case of $\Phi =$ all probability mass transportations, it is enough to note that the Phi regret of any stochastic matrix transformations is at most $|A|$ times larger than the worst possible regret of a probability mass transportation between two actions.