MIT 6.7980 — Topics in Multiagent Learning Thu, Sep 24, 2026

Lecture 4

Foundations of learning in games

Instructor: Prof. Gabriele Farina (gfarina@mit.edu)

With this lecture 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.

L4.1 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 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 A for the player, so $X = \Delta(A)$. At each time $t = 1, 2, ...$, the player will play some strategy $x^{(t)} \in X$, receive some form of feedback, and will incorporate that feedback to formulate a "better" strategy $x^{(t+1)} \in X$ 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. However, for the purposes of the abstract model we are building today, let's not make any assumptions about how the feedback is assigned; we will strive to build algorithms that perform competitively under *any* feedback—even adversarial one.

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 L4.1 (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 $\phi : X \to X$ 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 at the following formalization.

Definition L4.2 (Φ-regret minimizer)**.** Given the convex and compact strategy set X and a set Φ of linear transformations $\phi : X \to X$, a *Φ-regret minimizer for the set 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()` takes no input, and has the effect that the regret minimizer will output an element $x^{(t)} \in X$.
- `ObserveUtility`$\big(u^{(t)}\big)$ provides the environment's feedback to the regret minimizer, in the form of a linear utility function $u^{(t)} : X \to \mathbb{R}$ that evaluates how good the last-output point $x^{(t)}$ was. The utility function can depend adversarially on the outputs $x^{(1)}, ..., x^{(t)}$.

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

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

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.

L4.1.1 Notable choices of transformations Φ

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 $X = \Delta(A)$.

- Φ = 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.
- Φ = set of all "probability mass transport" on X, defined as $\Phi = \{\phi_{a\to b}\}_{a,b\in A}$, where

$$\big(\phi_{a\to b}(x)\big)_s := \begin{cases} 0 & \text{if } \; s = a \quad \text{(remove mass from } \; a \; \text{...)} \\ x_b + x_a & \text{if } \; s = b \quad \text{(... and give it to } \; b\text{)} \\ x_s & \text{otherwise.} \end{cases}$$

  This is known as *internal regret*.

Theorem L4.3 (Informal; formal version in Theorem L4.11)**.** 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 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 L4.4 (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.

Theorem L4.5 (Informal; formal version in Theorem L4.11)**.** 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 L4.6 (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.

L4.1.2 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 L4.7 (Regret minimizer)**.** Let X be a set. An *external regret minimizer for X* —or simply *"regret minimizer for X"*—is a Φ^{const}-regret minimizer for the special set of *constant* transformations

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

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

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

Again, the goal for a regret minimizer is to ensure its cumulative regret $\text{Reg}^{(T)}$ grows sublinearly in T.

An important result asserts the existence of algorithms that guarantee sublinear regret for any convex and compact domain X, typically of the order $\text{Reg}^{(T)} = O\big(\sqrt{T}\big)$ asymptotically.

As we will show below, 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 discussed, regret minimization is *one* instantiation of Φ-regret minimization—and perhaps the smallest sensible instantiation. Then, clearly, coming up with a regret minimizer for a set X cannot be harder than the problem of coming up with a Φ-regret minimizer for 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 discuss more about this in Supplementary reading S6.

L4.2 Applications of regret minimization

To establish regret minimization as a meaningful abstraction for learning in games, we check that regret minimizing and Φ-regret minimizing dynamics indeed lead to the expected behavior in common scenarios.

Definition L4.8 (Canonical learning setup)**.** In the cases that we will mention, a recurring idea will be to consider the setup in which all players $i \in [n]$ play according to the outputs $x_i^{(t)}$ of a Φ-regret minimizer. At each iteration, the utility function that each player i observes from the environment is the utility function u_i of that player, evaluated in the strategies played by all players, that is,

$$u^{(t)} : \Delta(A_i) \to \mathbb{R} \qquad\qquad u^{(t)}(x_i) := u_i\big(x_i, x_{-i}^{(t)}\big).$$

Given its importance, we give to this natural setup the name of *canonical learning setup*.

L4.2.1 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 $x_i^{(t)} \in X_i$ 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:

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

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,

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

(You should try to prove this!)

L4.2.2 Learning a Nash equilibrium in two-player zero-sum games

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

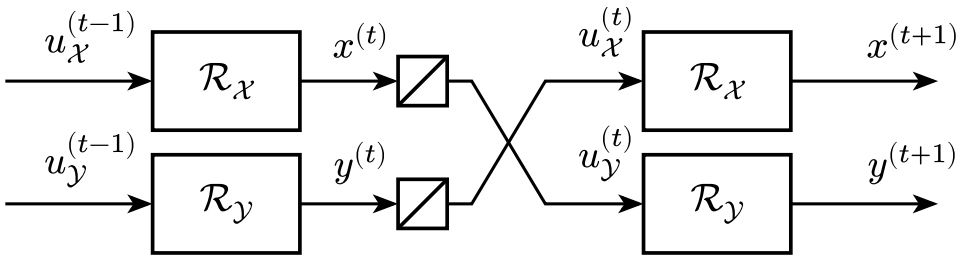
$$\max_{x \in X} \min_{y \in Y} x^\top U y, \tag{1}$$

where X and Y are convex compact sets and U 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 X and 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 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, R_X and R_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 $u_X^{(t)}, u_Y^{(t)}$ defined as

$$u_X^{(t)} : x \mapsto \big(U y^{(t)}\big)^\top x, \qquad u_Y^{(t)} : y \mapsto -\big(U^\top x^{(t)}\big)^\top y.$$

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*

$$0 \leq \gamma(x,y) \coloneqq \Big(\max_{\hat{x}\in X}\{\hat{x}^\top U y\} - x^\top U y\Big) + \Big(x^\top U y - \min_{\hat{y}\in Y}\{x^\top U \hat{y}\}\Big) = \max_{\hat{x}\in X}\{\hat{x}^\top U y\} - \min_{\hat{y}\in Y}\{x^\top U \hat{y}\}.$$

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

Theorem L4.9. Consider the self-play setup summarized in the figure above, where R_X and R_Y are regret minimizers for the sets X and Y, respectively. Let $\mathrm{Reg}_X^{(T)}$ and $\mathrm{Reg}_Y^{(T)}$ be the (sublinear) regret cumulated by R_X and R_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

$$\gamma\big(\overline{x}^{(T)}, \overline{y}^{(T)}\big) = \frac{\mathrm{Reg}_X^{(T)} + \mathrm{Reg}_Y^{(T)}}{T} \to 0 \qquad \text{as } \; T \to \infty.$$

Proof. By definition of regret,

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

Letting $T \to \infty$ and using the sublinearity of regret, we obtain the statement. □

L4.2.3 Proof of the minimax theorem

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

Theorem L4.10 (Minimax theorem)**.** Let X and Y be convex compact sets, and let U be a matrix. Suppose that a regret minimizer R_X for set X guaranteeing sublinear regret no matter the sequence of utilities can be constructed. Then,

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

Proof. One direction of the equality, specifically

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

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

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

$$y^{(t)} \in \arg\min_{y\in Y} \left(x^{(t)}\right)^\top U y.$$

The utility function observed by R_X at each time t is set to the linear function

$$u_X^{(t)}(x) = x^\top U y^{(t)}.$$

Letting $\overline{x}^{(T)} \in X$ and $\overline{y}^{(T)} \in Y$ 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

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

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

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

Combining the expressions, we obtain

$$\max_{x\in X}\min_{y\in Y} x^\top U y \ge \min_{y\in Y}\max_{x\in X} x^\top U y - \frac{\mathrm{Reg}_X^{(T)}}{T}.$$

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

L4.2.4 Learning (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 L4.11 (Formal version of Theorems L4.3 and L4.5)**.** 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 $(a_1, ..., a_n)$ independendently 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

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

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

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

as

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

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. $\square$

Note that Theorem L4.11 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 Φ-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.