CS639: Algorithmic Game Theory & Learning

Chapter 5: Learning in Games

Kirthevasan Kandasamy
UW-Madison

Outline

1. Introduction to learning in games
2. No-regret dynamics converges in zero sum games, Proof of the minimax theorem
3. No-regret dynamics converges to CE
4. Swap regret and CCE

Slides are intended as teaching aids only and do not include all material discussed in class. Students are strongly encouraged to attend lectures and take their own notes.

Ch 5.1: Introduction to learning in games

How do players choose strategies?

- A classical view is that players compute a NE and play that strategy. In practice, this is often unrealistic: players rarely solve for equilibria explicitly.
- In some settings, a trusted mediator can guide players toward a socially desirable equilibrium, but this is not always feasible in very decentralized environments (*e.g.*, congestion game, online marketplaces, ad auctions).
- In repeated games, a more realistic model is that players *adapt* based on past outcomes (*e.g.*, in a congestion game, if you see heavy traffic on one route today, you might try a different route tomorrow.).

Motivating question. When does adaptive, rational behavior converge to an equilibrium?

Convergence to equilibria via adaptive behavior

- We have already seen one adaptive process:
  - *Best-response dynamics* converge to a pure Nash equilibrium in *potential games*.
  - However, not all games are potential games.
- If best responses do not always converge, is there another form of adaptive rational behavior that does?
- **No-regret dynamics:** We study a model in which each player uses a *no-regret policy* to choose mixed strategies over repeated rounds.

Convergence to equilibria via adaptive behavior (cont'd)

In this chapter, we will see that:

1. In two-player zero-sum games, no-regret dynamics converges to a NE.
2. In general-sum games, no-regret dynamics converges to a CCE.
3. In general-sum games, *no-swap-regret dynamics* (a stronger notion) converges to a CE.

Why study no-regret dynamics?

- Provides insight into large-scale systems with many rational players (*e.g.*, online markets, ad auctions, high frequency trading), where individual players run no-regret policies to maximize their utility (*e.g.*, revenue, number of clicks).

- Leads to efficient algorithms for approximating equilibria.

No-regret dynamics

Consider an n player game with finite action spaces $\{\mathcal{A}_i\}_{i\in[n]}$. Let us assume the utilities are bounded and satisfy $u_i : \mathcal{A} \to \left[-\frac{1}{2}, \frac{1}{2}\right]$ (w.l.o.g)[1].

1) On round t each player i chooses a *mixed strategy* $p_{i,t} \in \Delta(\mathcal{A}_i)$. An action $A_{i,t}$ is sampled from $p_{i,t}$ for each player i.

2) At the end of round t, each player observes a vector of expected utilities for each action in $\mathcal{A}_i$,

$$\begin{bmatrix} u_i(1, p_{-i,t}) \\ u_i(2, p_{-i,t}) \\ \vdots \end{bmatrix} = \begin{bmatrix} \mathbb{E}_{a_{-i}\sim p_{-i,t}}[u_i(1, a_{-i})] \\ \mathbb{E}_{a_{-i}\sim p_{-i,t}}[u_i(2, a_{-i})] \\ \vdots \end{bmatrix}$$

Here, $p_{-i,t} = \bigtimes_{j\neq i} p_{j,t}$ is the distribution over the other's actions. This is possible, say, if each player observes the mixed strategies of the other players as,

$$u_i(a_i', p_{-i,t}) = \sum_{a_{-i}\in\mathcal{A}_{-i}} \left(\prod_{j\neq i} p_{j,t}(a_j)\right) u_i(a_i', a_{-i}).$$

[1]We will require the losses be bounded. The normalization to $[-1/2, 1/2]$ ensures that the losses are bounded to $[0, 1]$ when we apply online learning policies.

No-regret dynamics (cont'd)

3) From this, we can construct losses $\{\ell_{i,t}\}_t$ for each player $i \in [n]$, where $\ell_{i,t} \in [0,1]^{|\mathcal{A}_i|}$, is defined as follows,

$$\ell_{i,t}(k) = \frac{1}{2} - u_i(k, p_{-i,t}), \qquad \text{for all } k \in \mathcal{A}_i.$$

Definition (No-regret dynamics). The above interaction is called no-regret dynamics when *every* player chooses their mixed strategy $\{p_{i,t}\}_{i\in[n]}$ via some no-regret learning policy acting on the losses $\{\ell_{i,t}\}_{t=1}^{T}$.

N.B. With some additional work, this framework can be extended to situations where players observe only the realized actions of the others and not the mixed strategies $\{p_{j,t}\}_{j\neq i}$.

Player regret under no-regret dynamics

The expected loss of player i under a mixed strategy $p_{i,t}$ on round t is,

$$p_{i,t}^\top \ell_{i,t} = \sum_{k \in \mathcal{A}_i} p_{i,t}(k) \left(\frac{1}{2} - u_i(k, p_{-i,t}) \right) = \frac{1}{2} - u_i(p_{i,t}, p_{-i,t}) = \frac{1}{2} - u_i(p_t).$$

Therefore, player i's regret under a policy π_i and losses $\ell_i = \{\ell_{i,t}\}_t$ is

$$\begin{aligned} R_{i,T}(\pi_i, \ell_i) &\triangleq \sum_{t=1}^{T} p_{i,t}^\top \ell_{i,t} - \min_{p \in \Delta(\mathcal{A}_i)} \sum_{t=1}^{T} p^\top \ell_{i,t} \\ &= \sum_{t=1}^{T} \left(\frac{1}{2} - u_i(p_t) \right) - \min_{p \in \Delta(\mathcal{A}_i)} \sum_{t=1}^{T} \left(\frac{1}{2} - u_i(p, p_{-i,t}) \right) \\ &= \max_{p \in \Delta(\mathcal{A}_i)} \sum_{t=1}^{T} u_i(p, p_{-i,t}) - \sum_{t=1}^{T} u_i(p_t). \end{aligned}$$

Player regret under no-regret dynamics

We just showed,

$$R_{i,T}(\pi_i, \ell_i) = \underbrace{\sum_{t=1}^{T} p_{i,t}^\top \ell_{i,t} - \min_{p \in \Delta(\mathcal{A}_i)} \sum_{t=1}^{T} p^\top \ell_{i,t}}_{=A} = \underbrace{\max_{p \in \Delta(\mathcal{A}_i)} \sum_{t=1}^{T} u_i(p, p_{-i,t}) - \sum_{t=1}^{T} u_i(p_t)}_{=A^\star}.$$

Recalling the alternative way to write $R_{i,T}$, we can similarly show, *(try at home)*

$$R_{i,T}(\pi_i, \ell_i) = \underbrace{\sum_{t=1}^{T} p_{i,t}^\top \ell_{i,t} - \min_{k \in \mathcal{A}_i} \sum_{t=1}^{T} \ell_{i,t}(k)}_{=B} = \underbrace{\max_{k \in \mathcal{A}_i} \sum_{t=1}^{T} u_i(k, p_{-i,t}) - \sum_{t=1}^{T} u_i(p_t)}_{=B^\star}.$$

Going forward, we will use A, B when applying online learning algorithms. In particular, we know that, using ideas from Chapter 4, we can design policies π_i so that $R_{i,T}(\pi_i, \ell_i) \in o(T)$ for all players i and loss sequences $\ell_i \in [0,1]^T$. We will use $A^\star, B^\star$ when interpreting the results in the context of games.

Ch 5.2: No-regret dynamics in zero sum games

We will now study no-regret dynamics in a two player zero sum game with payoff matrix $Q \in [-1/2, 1/2]^{m \times n}$, where

$$\mathcal{A}_1 = [m], \qquad \mathcal{A}_2 = [n], \qquad u_1(i,j) = -u_2(i,j) = Q_{i,j}.$$

Recall that, $u_1(x,y) = \mathbb{E}[u_1(i,j)] = x^\top Q y$, and $u_2(x,y) = -x^\top Q y$.

Using $A^\star$ below we can write,

$$R_{1,T}(\pi_1, \ell_1) = \max_{x \in \Delta_m} \sum_{t=1}^{T} x^\top Q y_t \; - \; \sum_{t=1}^{T} x_t^\top Q y_t,$$

$$R_{2,T}(\pi_2, \ell_2) = \max_{y \in \Delta_m} \sum_{t=1}^{T} (-x_t^\top Q y) \; - \; \sum_{t=1}^{T} (-x_t^\top Q y_t) \; = \; \sum_{t=1}^{T} x_t^\top Q y_t \; - \; \min_{y \in \Delta_n} \sum_{t=1}^{T} x_t^\top Q y.$$

Recall, we showed

$$R_{i,T}(\pi_i, \ell_i) = \underbrace{\sum_{t=1}^{T} p_{i,t}^\top \ell_{i,t} - \min_{p \in \Delta(\mathcal{A}_i)} \sum_{t=1}^{T} p^\top \ell_{i,t}}_{=A} = \underbrace{\max_{p \in \Delta(\mathcal{A}_i)} \sum_{t=1}^{T} u_i(p, p_{-i,t}) \; - \; \sum_{t=1}^{T} u_i(p_t)}_{=A^\star}.$$

Minimax theorem

We will now prove the minimax theorem.

Theorem (Von Neumann's minimax theorem). In any two player zero sum game,

$$\max_{x \in \Delta_m} \min_{y \in \Delta_n} x^\top Q y = \min_{x \in \Delta_m} \max_{y \in \Delta_n} x^\top Q y$$

Proof. We have already seen that $\mathrm{LHS} \leq \mathrm{RHS}$. We will now show that $\mathrm{LHS} \geq \mathrm{RHS}$. In particular, we will show that for all $\epsilon > 0$, $\mathrm{RHS} \leq \mathrm{LHS} + \epsilon$.

Let ϵ be given. Consider the no-regret dynamics environment described above, with any no-regret policies π_1, π_2, and T rounds, where T will be chosen later (based on ϵ).

Recall: The following regret quantities, and drop the dependence of π_i, ℓ_i for simplicity,

$$R_{1,T} = \max_{x \in \Delta_m} \sum_{t=1}^{T} x^\top Q y_t - \sum_{t=1}^{T} x_t^\top Q y_t, \qquad R_{2,T} = \sum_{t=1}^{T} x_t^\top Q y_t - \min_{y \in \Delta_n} \sum_{t=1}^{T} x_t^\top Q y.$$

Let $\overline{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$ and $\overline{y} = \frac{1}{T} \sum_{t=1}^{T} y_t$ denote the time averaged mixed stragies of both players.

Minimax theorem (cont'd)

We first have,

$$\begin{aligned}\min_{y\in\Delta_n}\max_{x\in\Delta_m} x^\top Qy &\le \max_{x\in\Delta_m} x^\top Q\overline{y} \\ &= \max_{x\in\Delta_m} \frac{1}{T}\sum_{t=1}^{T} x^\top Qy_t \\ &= \frac{1}{T}R_{1,T} + \frac{1}{T}\sum_{t=1}^{T} x_t^\top Qy_t.\end{aligned}$$

You can similarly show, *(try at home)*

$$\max_{x\in\Delta_m}\min_{y\in\Delta_n} x^\top Qy \ge -\frac{1}{T}R_{2,T} + \frac{1}{T}\sum_{t=1}^{T} x_t^\top Qy_t$$

Minimax theorem (cont'd)

Recall, we just showed

$$\min_{y\in\Delta_n}\max_{x\in\Delta_m} x^\top Qy \le \frac{1}{T}R_{1,T} + \frac{1}{T}\sum_{t=1}^{T} x_t^\top Qy_t, \qquad \max_{x\in\Delta_m}\min_{y\in\Delta_n} x^\top Qy \ge -\frac{1}{T}R_{2,T} + \frac{1}{T}\sum_{t=1}^{T} x_t^\top Qy_t$$

Combining these two results, we get

$$\min_{y\in\Delta_n}\max_{x\in\Delta_m} x^\top Qy - \frac{1}{T}R_{1,T} \le \frac{1}{T}\sum_{t=1}^{T} x_t^\top Qy_t \le \max_{x\in\Delta_m}\min_{y\in\Delta_n} x^\top Qy + \frac{1}{T}R_{2,T}.$$

This implies,

$$\min_{y\in\Delta_n}\max_{x\in\Delta_m} x^\top Qy \le \max_{x\in\Delta_m}\min_{y\in\Delta_n} x^\top Qy + \frac{1}{T}\left(R_{1,T} + R_{2,T}\right)$$

As players are following no-regret policies, the last term in the RHS goes to 0 as $T \to \infty$. Hence, for any $\epsilon > 0$, we can find T large enough so the the second term of the RHS is smaller than ϵ. Therefore

$$\text{for all } \epsilon > 0, \qquad \min_{y\in\Delta_n}\max_{x\in\Delta_m} x^\top Qy \le \max_{x\in\Delta_m}\min_{y\in\Delta_n} x^\top Qy + \epsilon.$$

No-regret dynamics converges to a NE in a zero sum game

We will now show that no-regret dynamics converges to an approximate NE. Let us first recall the definition of a NE in a zero sum game.

Recall, NE in a zero sum game. A strategy profile $x^\star, y^\star$ is a NE if it satisfies,

$$\text{for all } x \in \Delta_m,\ y \in \Delta_n, \qquad x^\top Q y^\star \leq x^{\star\top} Q y^\star \leq x^{\star\top} Q y$$

We will now define an approximate NE:

Definition (Approximate NE). In a zero sum game, a strategy profile $(\overline{x}, \overline{y})$ is an ϵ-approximate NE if

$$\text{for all } x \in \Delta_m,\ y \in \Delta_n, \qquad x^\top Q \overline{y} - \epsilon \leq \overline{x}^\top Q \overline{y} \leq \overline{x}^\top Q y + \epsilon.$$

Intuitively, deviating to any other strategy will not yield more than ϵ utility for any player.

No-regret dynamics converges to an approximate NE

Claim. Suppose both players are following no-regret policies in a two-player zero sum game. Denote $\overline{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$ and $\overline{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$. After T rounds, we arrive at an ϵ_T–approximate Nash equilibrium, where

$$\epsilon_T = \frac{1}{T}\Big(R_{1,T}(\pi_1, \ell_1) + R_{2,T}(\pi_2, \ell_2)\Big) \in o(1).$$

Proof. Let us recall the regret quantities we defined,

$$R_{1,T}(\pi_1, \ell_1) \triangleq \sum_{t=1}^{T} x_t^\top \ell_{1,t} - \min_{x \in \Delta_m} \sum_{t=1}^{T} x^\top \ell_{1,t} = \max_{x \in \Delta_m} \sum_{t=1}^{T} x^\top Q y_t \; - \sum_{t=1}^{T} x_t^\top Q y_t.$$

$$R_{2,T}(\pi_2, \ell_2) \triangleq \sum_{t=1}^{T} y_t^\top \ell_{2,t} - \min_{y \in \Delta_n} \sum_{t=1}^{T} y^\top \ell_{2,t} = \sum_{t=1}^{T} x_t^\top Q y_t \; - \min_{y \in \Delta_n} \sum_{t=1}^{T} x_t^\top Q y.$$

No-regret dynamics converges to an approximate NE (cont'd)

Summing both equations and dividing by T we get,

$$\max_{x \in \Delta_m} x^\top Q\overline{y} \; - \; \min_{y \in \Delta_n} \overline{x}^\top Qy = \frac{1}{T}\left(R_{1,T}(\pi_1) + R_{2,T}(\pi_2)\right) = \epsilon_T.$$

We now have,

$$\max_{x \in \Delta_m} x^\top Q\overline{y} - \epsilon_T = \min_{y \in \Delta_n} \overline{x}^\top Qy \leq \; \overline{x}^\top Q\overline{y} \; \leq \max_{x \in \Delta_m} x^\top Q\overline{y} = \min_{y \in \Delta_n} \overline{x}^\top Qy + \epsilon_T.$$

This is precisely the definition of an ϵ_T-approximate NE. □

Recall, In a zero sum game, a strategy profile $(\overline{x}, \overline{y})$ is an ϵ-approximate NE if

$$\text{for all } x \in \Delta_m,\ y \in \Delta_n, \qquad x^\top Q\overline{y} - \epsilon \leq \overline{x}^\top Q\overline{y} \leq \overline{x}^\top Qy + \epsilon.$$

An algorithm to approximate NE in zero sum games

This suggests the following algorithm to compute an ϵ-approximate NE:

- We may use any algorithm for the experts problem (*e.g.*, Hedge) to design the policy for both players under the following losses:

$$\ell_{1,t} \in [0,1]^m, \quad \text{where,} \quad \ell_{1,t}(i) = \frac{1}{2} - u_1(i, y_t) = \frac{1}{2} - (Qy_t)_i \,.$$
$$\ell_{2,t} \in [0,1]^n, \quad \text{where,} \quad \ell_{2,t}(j) = \frac{1}{2} - u_2(x_t, j) = \frac{1}{2} + \left(Q^\top x_t\right)_j \,.$$

This assumes that the utilities are bounded, *i.e.*, $Q \in [1/2, 1/2]^{m \times n}$. If not, you will need to normalize them.

- From our previous result, we know that after T rounds, we have an ϵ_T-approximate NE, where

$$\epsilon_T = \frac{1}{T}\left(R_{1,T} + R_{2,T}\right)$$

An algorithm to approximate NE in zero sum games

- Recall, from Chapter 4, for Hedge with K actions, the regret is $2\sqrt{T\log(K)}$. This gives us,

$$\epsilon_T = \frac{1}{T}\left(R_{1,T} + R_{2,T}\right) \leq \frac{1}{T}\left(2\sqrt{T\log(m)} + 2\sqrt{T\log(n)}\right)$$
$$\leq \frac{4\sqrt{\log(\max(m,n))}}{\sqrt{T}}.$$

- To achieve an ϵ-approximate NE is is sufficient if the number of iterations T satisfies,

$$\frac{4\sqrt{\log(\max(m,n))}}{\sqrt{T}} \leq \epsilon \qquad \Longleftrightarrow \qquad T \geq \frac{16\log(\max(m,n))}{\epsilon^2}.$$

Ch 5.3: No-regret dynamics in general sum games

We will now show that in an n-player normal form game, no-regret dynamics converges to a coarse-correlated equilibrium. Let us first define an approximate CCE.

Definition (Approximate CCE). A joint distribution $s \in \Delta(\mathcal{A})$ is an ϵ-approximate CCE if for all players i, we have

$$\mathbb{E}_{a\sim s}\left[u_i(a)\right] \geq \mathbb{E}_{a\sim s}\left[u_i(a_i', a_{-i})\right] - \epsilon, \quad \text{for all } a_i' \in \mathcal{A}_i.$$

Let us also quickly recall no-regret dynamics.

Recall, no-regret dynamics. In an n player normal form game, repeated over T rounds, each player i chooses her mixed strategy $p_{i,t} \in \Delta(\mathcal{A}_i)$ via a *no-regret policy* π_i. At the end of the round, the expected payoffs $u_i(k, p_{-i,t})$ for each action $k \in \mathcal{A}_i$ are revealed to the players. Player i's regret is,

$$R_{i,T}(\pi_i, \ell_i) = \underbrace{\sum_{t=1}^{T} p_{i,t}^\top \ell_{i,t} - \min_{k\in\mathcal{A}_i} \sum_{t=1}^{T} \ell_{i,t}(k)}_{=B} = \underbrace{\max_{k\in\mathcal{A}_i} \sum_{t=1}^{T} u_i(k, p_{-i,t}) - \sum_{t=1}^{T} u_i(p_t)}_{=B^\star}.$$

No-regret dynamics converges to a CCE

Claim. Let $\epsilon_T = \max_{i \in [n]} \frac{R_{i,T}}{T}$ be the maximum average regret of any player when following their respective policies. Let $p_t = \bigtimes_{i=1}^{n} p_{i,t}$ be the joint distribution on round t. Let $\overline{p} = \frac{1}{T} \sum_{t=1}^{T} p_t$ be the time-averaged distribution of all players. Then, $\overline{p}$ is an ϵ_T–approximate CCE.

Proof. This simply requires plugging in definitions. First note that to sample an action profile from $\overline{p}$, we can simply sample some p_t from $\{p_1, p_2, \ldots, p_T\}$, and then sample an action from p_t. Hence, we can write,

$$\mathbb{E}_{a \sim \overline{p}} [u_i(a)] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{a \sim p_t} [u_i(a)] = \frac{1}{T} \sum_{t=1}^{T} u_i(p_t).$$

[1]Note that while each $p_t = p_{1,t} \times \cdots \times p_{n,t}$ is a product distribution, $\overline{p}$ may not be a product distribution. In particular, it is not equal to $\overline{p}_1 \times \cdots \times \overline{p}_n$ where $\overline{p}_i = \frac{1}{T} \sum_{t=1}^{T} p_{i,t}$.

No-regret dynamics converges to a CCE (cont'd)

Recall, $$R_{i,T}(\pi_i, \ell_i) = \underbrace{\sum_{t=1}^{T} p_{i,t}^\top \ell_{i,t} - \min_{k \in \mathcal{A}_i} \sum_{t=1}^{T} \ell_{i,t}(k)}_{=B} = \underbrace{\max_{k \in \mathcal{A}_i} \sum_{t=1}^{T} u_i(k, p_{-i,t}) - \sum_{t=1}^{T} u_i(p_t)}_{=B^\star}.$$

By a similar reasoning, we have

$$\mathbb{E}_{a \sim \overline{p}} \left[u_i(a_i', a_{-i}) \right] = \mathbb{E}_{a_{-i} \sim \overline{p}_{-i}} \left[u_i(a_i', a_{-i}) \right] \qquad \texttt{where, } \overline{p}_{-i} = \frac{1}{T} \sum_{t=1}^{T} p_{-i,t}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{a_{-i} \sim p_{-i,t}} \left[u_i(a_i', a_{-i}) \right] = \frac{1}{T} \sum_{t=1}^{T} u_i(a_i', p_{-i,t}).$$

Therefore, for any deviation a_i' for player i, we have

$$\mathbb{E}_{a \sim \overline{p}} \left[u_i(a) \right] \overset{(a)}{=} \max_{k \in \mathcal{A}_i} \frac{1}{T} \sum_{t=1}^{T} u_i(k, p_{-i,t}) - \frac{1}{T} R_{i,T} \overset{(b)}{\geq} \mathbb{E}_{a \sim \overline{p}} \left[u_i(a_i', a_{-i}) \right] - \epsilon_T.$$

This is precisely the definition of an approximate CCE. Here, (a) uses the result from the previous slide, while (b) uses the result from the previous display and the definition of ϵ_T. □

Quiz

Based on this, can you suggest an algorithm to approximate a CCE? How many iterations will it take to find an ϵ-approximate CCE?

Ch 5.4: Swap regret and correlated equilibria

We will now show that a stronger no-regret behavior called *no-swap-regret dynamics* converges to a CE. Let us first recall the experts problem from Chapter 4.

The experts problems: There are a set of K actions (experts), denoted $[K]$.
- On round t, a learner chooses a probability distribution $p_t = (p_t(1), \dots, p_t(K)) \in \Delta_K$. An action A_t is sampled from p_t and played.
- An *adversary* (environment) simultaneously (without knowledge of A_t) chooses a loss vector $\ell_t = (\ell_t(1), \dots, \ell_t(K)) \in [0,1]^K$, where $\ell_t(i)$ is the loss for action i.
- The learner incurs expected loss $\mathbb{E}[\ell_t(A_t)] = \sum_{i=1}^K p_t(i)\ell_t(i) = p_t^\top \ell_t$.
- The learner observes the entire loss vector ℓ_t, i.e the losses for *all* actions.

The (external) regret of a learner's policy π under a sequence of losses ℓ is,

$$R_T(\pi, \ell) = \sum_{t=1}^T \underbrace{p_t^\top \ell_t}_{=\mathbb{E}[\ell_t(A_t)]} - \min_{i \in [K]} \sum_{t=1}^T \ell_t(i).$$

Swap regret

We will first define *swap regret* in the context of online learning. To motivate swap regret, let us begin with the external regret as follows:

$$R_T(\pi, \ell) = \sum_{t=1}^{T} \mathbb{E}\left[\ell_t(A_t)\right] - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)\right].$$

A policy has no regret if $R_T(\pi, \ell) \in o(T)$. Equivalently, we say that a policy has no regret if for all actions $a \in [K]$, we have $R_T(\pi, \ell, a) \in o(T)$, where

$$R_T(\pi, \ell, a) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t) - \sum_{t=1}^{T} \ell_t(a)\right].$$

Swap regret (cont'd)

Instead of comparing against a fixed action, in *swap regret*, we compare against a fixed *swap function* $\sigma : [K] \to [K]$:

$$R_T^{\mathrm{sw}}(\pi, \ell, \sigma) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t) - \sum_{t=1}^{T} \ell_t(\sigma(A_t))\right]$$
$$= \sum_{t=1}^{T} p_t^\top \ell_t - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\sigma(A_t))\right]$$

Here, the swap function σ maps the action the learner took to an alternative action.

A policy satisfies *no-swap-regret* if $R_T^{\mathrm{sw}}(\pi, \ell, \sigma) \in o(T)$ for all swap functions σ.

Note that no-swap-regret is a stronger requirement than no-regret since $R_T^{\mathrm{sw}}(\pi, \ell, \sigma) = R_T(\pi, \ell, a)$ when $\sigma(a') = a$ for all $a' \in [K]$. That is, no-regret requires being competitive only against K specific swap functions.

Reduction from swap regret to external regret

It turns out, that we can construct a no-swap regret policy from no-regret policies.

Theorem (Reduction from swap regret to external regret[2]). If there is a no-regret policy, then there is a no-swap-regret policy.

Proof. We will construct our no-swap-regret policy π from K different instantiations of no-regret policies $\{\pi^{(k)}\}_{k\in[K]}$. The construction is as follows:

- On each round t, receive probability distributions $q_t^{(1)}, q_t^{(2)}, \ldots, q_t^{(K)}$ from $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(K)}$ respectively, where $q_t^{(k)} \in \Delta_K$.

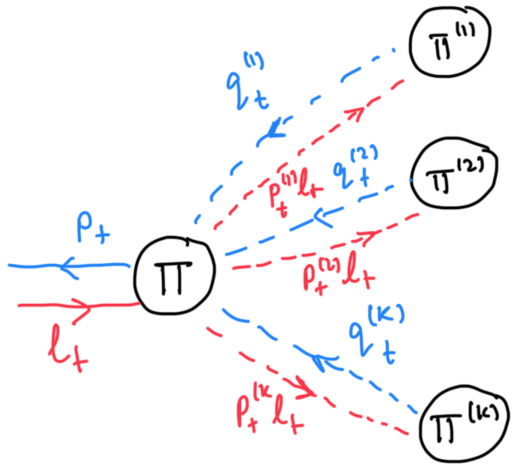
- Use $q_t^{(1)}, q_t^{(2)}, \ldots, q_t^{(K)}$ to construct a distribution p_t from which the action will be sampled on round t. We will specify how we do this shortly.

- π incurs expected loss $p_t^\top \ell_t$, and observes the loss vector ℓ_t.

- Give policy $\pi^{(k)}$ the loss vector $p_t(k)\ell_t$, which $\pi^{(k)}$ will use to compute its future distributions.

[2] **Acknowledgment:** The original proof is due to Blum and Mansour (2007). The presentation here follows Tim Roughgarden's lecture notes.

Reduction from swap regret to external regret (cont'd)



In the above construction, intuitively, the policy $\pi^{(k)}$ will be responsible for swaps from action k to other actions.

Reduction from swap regret to external regret (cont'd)

To analyze this algorithm, let us fix some swap function σ and recall the definition of the swap regret:

$$R_T^{\mathrm{sw}}(\pi, \ell, \sigma) = \sum_{t=1}^{T} p_t^\top \ell_t - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\sigma(A_t))\right]$$

We can write the second term of the RHS as,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\sigma(A_t))\right] = \sum_{t=1}^{T}\sum_{k=1}^{K} \ell_t(\sigma(k)) p_t(k).$$

Let us now consider policy $\pi^{(k)}$, which believes that actions are chosen according to $\{q_t^{(k)}\}_t$ and that the losses are $\{p_t(k)\ell_t\}_t$. Hence, its regret against a given action $a \in [K]$ is (we will drop the dependence of the regret on the losses for simplicity),

$$R_T^{(k)}(a) = \sum_{t=1}^{T} \underbrace{{q_t^{(k)}}^\top (p_t(k)\ell_t))}_{=\text{probabilities}^\top \text{losses}} - \sum_{t=1}^{T} \underbrace{p_t(k)\ell_t(a)}_{=\text{loss for action } a}.$$

Reduction from swap regret to external regret (cont'd)

We can write the first term of the RHS as,

$$\sum_{t=1}^{T} {q_t^{(k)}}^\top (p_t(k)\ell_t)) = \sum_{t=1}^{T}\sum_{j=1}^{K} p_t(k)\ell_t(j)q_t^{(k)}(j)$$

Let us write policy $\pi^{(k)}$'s regret against the action $a = \sigma(k)$, where, recall, σ is the given swap function,

$$R_T^{(k)}(\sigma(k)) = \sum_{t=1}^{T}\sum_{j=1}^{K} p_t(k)\ell_t(j)q_t^{(k)}(j) - \sum_{t=1}^{T} p_t(k)\ell_t(\sigma(k)).$$

Therefore,

$$\sum_{k=1}^{K} R_T^{(k)}(\sigma(k)) = \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{j=1}^{K} p_t(k)\ell_t(j)q_t^{(k)}(j) - \sum_{t=1}^{T}\sum_{k=1}^{K} p_t(k)\ell_t(\sigma(k)).$$

Reduction from swap regret to external regret (cont'd)

Let us put together everything we have shown so far,

$$R_T^{\mathrm{sw}}(\pi,\ell,\sigma)=\sum_{t=1}^{T}p_t^\top\ell_t-\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(\sigma(A_t))\right]=\underbrace{\sum_{t=1}^{T}\sum_{j=1}^{K}p_t(j)\ell_t(j)}_{=A}-\underbrace{\sum_{t=1}^{T}\sum_{k=1}^{K}\ell_t(\sigma(k))p_t(k)}_{=B}$$

$$\sum_{k=1}^{K}R_T^{(k)}(\sigma(k))=\underbrace{\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{j=1}^{K}p_t(k)\ell_t(j)q_t^{(k)}(j)}_{=C}-\underbrace{\sum_{t=1}^{T}\sum_{k=1}^{K}p_t(k)\ell_t(\sigma(k))}_{=B}.$$

We therefore, have,

$$\begin{aligned}R_T^{\mathrm{sw}}(\pi,\ell,\sigma)&=\sum_{k=1}^{K}R_T^{(k)}(\sigma(k))+A-C\\&=\sum_{k=1}^{K}R_T^{(k)}(\sigma(k))+\sum_{t=1}^{T}\sum_{j=1}^{K}\ell_t(j)\left(p_t(j)-\sum_{k=1}^{K}p_t(k)q_t^{(k)}(j)\right)\end{aligned}$$

Reduction from swap regret to external regret (cont'd)

We are almost done. Recall that in our construction, we use $q_t^{(1)}, q_t^{(2)}, \ldots, q_t^{(K)}$ to construct a distribution p_t from which the action will be sampled on round t. We have not yet specified how exactly we construct p_t.

Suppose we can construct p_t so that $p_t(j) = \sum_{k=1}^{K} p_t(k) q_t^{(k)}(j)$ for all $j \in [K]$. We then have,

$$R_T^{\mathrm{sw}}(\pi, \ell, \sigma) = \sum_{k=1}^{K} R_T^{(k)}(\sigma(k))$$

If $R_T^{(k)}(\sigma(k)) \leq f(T)$, then $R_T^{\mathrm{sw}}(\pi, \ell, \sigma) \leq Kf(T)$. In particular, if $R_T^{(k)}(\sigma(k)) \in o(T)$, then we also have $R_T^{\mathrm{sw}}(\pi, \ell, \sigma) \in o(T)$ and we are done!

Claim. Such a p_t can be constructed.

N.B. *We will state this without a proof, but the next slide contains some details.*

□.

Constructing p_t *(read at home)*

For the interested reader only.

The claim can be restated as follows: Given vectors $q_t^{(1)}, q_t^{(2)}, \ldots, q_t^{(K)}$, we can construct a vector p_t so that the following holds,

$$\begin{bmatrix} | \\ p_t \\ | \end{bmatrix} = \underbrace{\begin{bmatrix} | & | & & | \\ q_t^{(1)} & q_t^{(2)} & \cdots & q_t^{(K)} \\ | & | & & | \end{bmatrix}}_{=Q_t} \begin{bmatrix} | \\ p_t \\ | \end{bmatrix}$$

In other words, does Q_t have a right eigenvector with eigenvalue 1 which is also a probability vector? It is well-known that such an eigenvector exists for stochastic matrices (matrices whose columns are probability vectors).

No-swap-regret dynamics converges to a CE

We will now show that no-swap-regret dynamics converges to a CE. To define an approximate CE, we will first establish an equivalent definition for a CE.

Recall, CE (definition). A joint distribution $s \in \Delta(\mathcal{A})$ is a CE if, for all players i, we have

$$\mathbb{E}_{a\sim s}\left[u_i(a_i', a_{-i})|a_i = a_i'\right] \geq \mathbb{E}_{a\sim s}\left[u_i(a_i'', a_{-i})|a_i = a_i'\right], \quad \text{for all } a_i', a_i'' \in \mathcal{A}_i.$$

Claim. Any function $\sigma_i : \mathcal{A}_i \to \mathcal{A}_i$ which maps each action of player i is called a *swap function* for player i. A joint distribution s is a CE if and only if for all players i, we have

$$\mathbb{E}_{a\sim s}\left[u_i(a)\right] \geq \mathbb{E}_{a\sim s}\left[u_i(\sigma_i(a_i), a_{-i})\right], \quad \text{for all swap functions } \sigma_i.$$

Proof. Coming up in HW 4.

No-swap-regret dynamics converges to a CE (cont'd)

This motivates the following definition for an approximate CE.

Definition (Approximate CE). A joint distribution $s \in \Delta(\mathcal{A})$ is an ϵ-approximate correlated equilibrium if

$$\mathbb{E}_{a\sim s}\left[u_i(a)\right] \geq \mathbb{E}_{a\sim s}\left[u_i(\sigma_i(a_i), a_{-i})\right] - \epsilon, \quad \text{for all swap functions } \sigma_i.$$

Claim. Suppose the game is repeated for T rounds, which each player j choosing a mixed strategy $p_{j,t} \in \Delta(\mathcal{A}_j)$ on each round t. Let $\epsilon_T = \max_{i\in[n]} \max_{\sigma_i} \frac{1}{T} R^{\text{sw}}_{i,T}(\sigma_i)$ be the maximum swap regret of any player $i \in [n]$ for any swap function $\sigma_i : \mathcal{A}_i \to \mathcal{A}_i$. Let $\overline{p} = \frac{1}{T}\sum_{t=1}^{T} p_t$ be the time-averaged distribution of all players, where $p_t = \bigtimes_{j=1}^{n} p_{j,t}$. Then, $\overline{p}$ is an ϵ_T–approximate CE.

Proof. Coming up in HW 4.

Summary of Chapters 4, 5

Overarching goals. *(1)* Understand how strategic agents learn over time and whether this behavior converges to equilibrium concepts. *(2)* Use this to develop algorithms to approximate equilibria.

Key take-aways:

- External Regret: Measures how well a learner does compared to the best fixed action in hindsight.
- No-regret policies (e.g., Hedge) achieve zero average external regret as $T \to \infty$.
- No-regret dynamics: each player chooses their mixed strategy on each round via a no-regret policy.
  - In a two player zero-sum game, no-regret dynamics converges to a NE.
  - In a general sum game, no-regret dynamics converges to a CCE.

Summary of Chapters 4, 5 (cont'd)

Key take-aways (cont'd):

- Swap Regret: A stronger notion of regret where a learner competes against the best swap function which swaps the action taken by the player to an alternative action.
- Key reduction: If no-regret algorithms exist, then one can construct no-swap regret algorithms.
- No-swap-regret dynamics: each player chooses their mixed strategy on each round via a no-swap-regret policy.
  - In a general sum game, no-swap-regret dynamics converges to a CE.