Lecture 14:
Learning in Games:

Part I: Intro and Concept

Yi, Yung (이융)
KAIST, Electrical Engineering
http://lanada.kaist.ac.kr
yiyung@kaist.edu

Intro: Learning in Games

Learning and Dynamics in Games

- Tension: Even though strategic form games model "one shot" interactions
  - NE: better motivated as the outcome of a dynamic process
  - Unclear how to interpret mixed strategies and Bernoulli payoffs as "one shot".
- Resolution
  - Define interactive processes that lead to NE
- Sorry for the notations
  - Some notations are used with different mathematical symbols
  - Please understand that making slides takes a lot of time

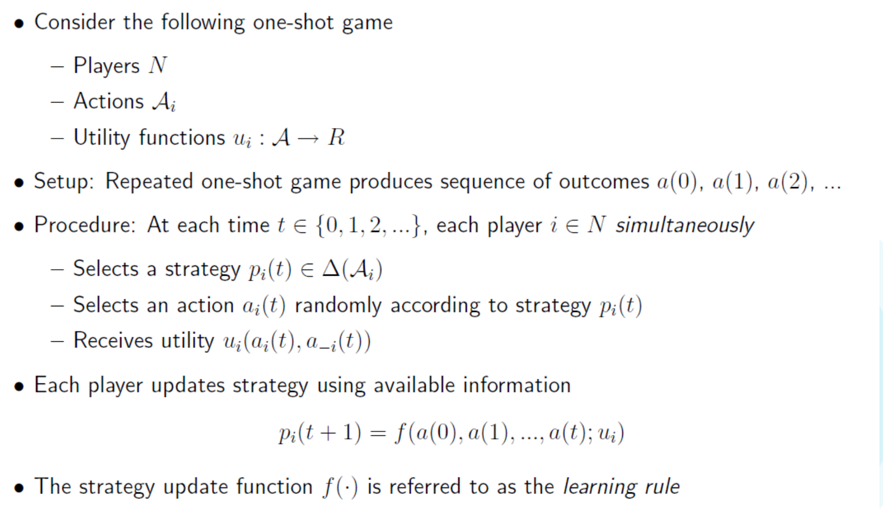
Games with Complete/Incomplete Information

- Complete Information
  - Every player knows the number of players and the strategy sets of all the players and their utility functions.
  - Every player knows that every player is rational.
- Incomplete Information
  - A player knows some information about himself, but has partial (or no) information about the others.
- Learning for Incomplete Information
  - Investigate ways in which players can optimize their own utility while simultaneously learning from experience or observations.

Taxonomy: Learning Algorithms

- Fully distributed (or Uncoupled) learning algorithm
  - A player does not use information about the other players.
  - Builds his strategies and updates them by using own-actions and own-utilities.
- Partially distributed learning algorithm
  - A player implements his updating rule after receiving some data about others.
  - The amount and the kind of data may depend on each algorithm (We will discuss this later)

Framework: Learning in Games (1)

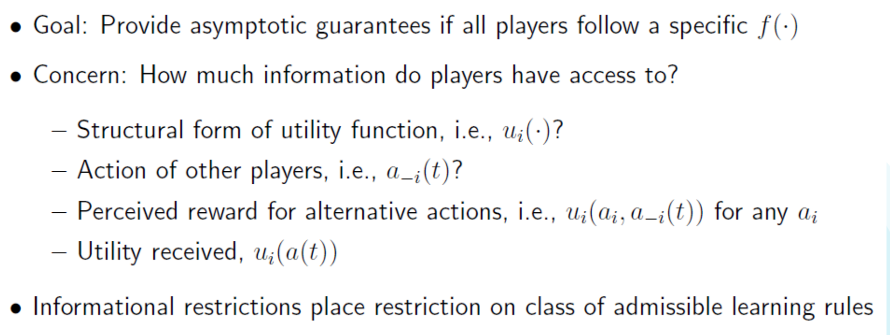


- Consider the following one-shot game
  - Players N
  - Actions $\mathcal{A}_i$
  - Utility functions $u_i : \mathcal{A} \rightarrow R$
- Setup: Repeated one-shot game produces sequence of outcomes $a(0)$, $a(1)$, $a(2)$, ...
- Procedure: At each time $t \in \{0, 1, 2, ...\}$, each player $i \in N$ *simultaneously*
  - Selects a strategy $p_i(t) \in \Delta(\mathcal{A}_i)$
  - Selects an action $a_i(t)$ randomly according to strategy $p_i(t)$
  - Receives utility $u_i(a_i(t), a_{-i}(t))$
- Each player updates strategy using available information

$$p_i(t+1) = f(a(0), a(1), ..., a(t); u_i)$$

- The strategy update function $f(\cdot)$ is referred to as the *learning rule*

Framework: Learning in Games (2)



- Goal: Provide asymptotic guarantees if all players follow a specific $f(\cdot)$
- Concern: How much information do players have access to?
  - Structural form of utility function, i.e., $u_i(\cdot)$?
  - Action of other players, i.e., $a_{-i}(t)$?
  - Perceived reward for alternative actions, i.e., $u_i(a_i, a_{-i}(t))$ for any a_i
  - Utility received, $u_i(a(t))$
- Informational restrictions place restriction on class of admissible learning rules

Some Concepts

- Better Reply Graph
- Finite Improvement Property
- Weakly Acyclic Game
- Correlated Equilibrium

Better Reply Graph G*

- Define the following **better reply** graph
  - Nodes are joint actions, $a \in \mathcal{A}$
  - Edges are unilateral better replies
- Details:
  - Let $a = (a_i, a_{-i})$ and $a' = (a_i', a_{-i})$ be two distinct nodes.
  - There is an edge from a to a' if

$$U_i(a_i', a_{-i}) > U_i(a_i, a_{-i})$$

i.e., deviating agent i experienced an improvement

- Example: Stag hunt

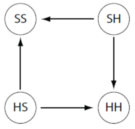
	S	H
S	$3, 3$	$0, 1$
H	$1, 0$	$1, 1$

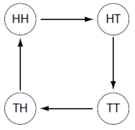
  - Why isn't there an edge from (H, H) to (S, S)?
  - What are the dead end nodes?

- Example: Matching pennies

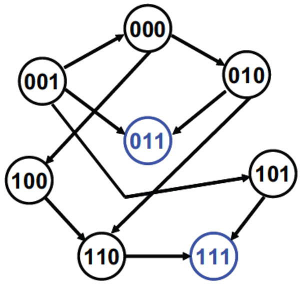
	S	H
S	$1, -1$	$-1, 1$
H	$-1, 1$	$1, -1$

Why are there cycles?





FIP (Finite Improvement Property)



- A game has the **finite improvement property** if *every* path in the better reply graph leads to an NE.
- See above illustration: 3 players, 2 moves each.

Potential game and FIP (1)

- The finite improvement property is difficulty to verify.
- Alternative: Potential games.
- A game is a **potential game** if there exists a potential function

$$\phi : \mathcal{A} \to \mathbf{R}$$

with the following property. Let (a_i, a_{-i}) and (a_i', a_{-i}) be two distinct joint actions.

$$U_i(a_i', a_{-i}) - U_i(a_i, a_{-i}) = \phi(a_i', a_{-i}) - \phi(a_i, a_{-i})$$

- In words: Unilateral changes in any agent's payoff quantitatively equals changes in a potential function. Player's utility functions are *aligned* with the potential function.
- FACT: Any potential game has the finite improvement property.
- How? A unilateral better reply increases the potential function. From finiteness, it cannot perpetually increases.

Potential Game and FIP (2)

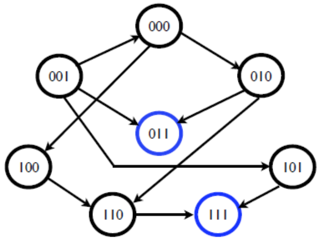
- Broader definition: **Generalized ordinal potential game**

$$U_1(a_i', a_{-i}) - U_i(a_i, a_{-i}) > 0$$
$$\Downarrow$$
$$\phi(a_i', a_{-i}) - \phi(a_i, a_{-i}) > 0$$

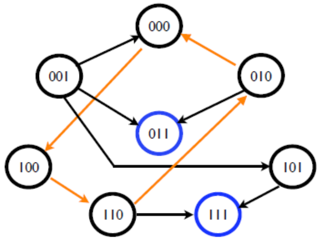
- FACT: Any generalized ordinal potential game has the finite improvement property.

Weakly Acyclic Game (1)

- Weakly acyclic games provide a generalization of potential games.
- A game is a **weakly acyclic game** if there exists a better reply path from any action profile to a pure Nash equilibrium.



Potential Game



Weakly Acyclic Game

- FACT: A weakly acyclic game does not necessarily have the finite improvement property.

Weakly Acyclic Game (2)

- Alternative defintion: A game is a **weakly acyclic game** if there exists a potential function

$$\phi : \mathcal{A} \to \mathbf{R}$$

with the following property: For any action profile a that is *not* a Nash equilibrium, there exist a player i with an action a_i' such that

$$U_i(a_i', a_{-i}) - U_i(a_i, a_{-i}) > 0 \quad \text{and} \quad \phi(a_i', a_{-i}) - \phi(a_i, a_{-i}) > 0$$

- In words: At least one agent's utility function is *aligned* with the potential function. (weaker form of alignment)

Summary

Correlated Strategies

- In a Nash equilibrium, players choose strategies (or randomize over strategies) independently.
- For games with multiple Nash equilibria, one may want to allow for randomizations between Nash equilibria by some form of communication prior to the play of the game.

Example Consider the Battle of the Sexes game:

	Ballet	Football
Ballet	1, 4	0, 0
Football	0, 0	4, 1

Suppose that the players flip a coin and go to the Ballet if the coin is Heads, and to the Football game if the coin is tails, i.e., they randomize between two pure strategy Nash equilibria, resulting in a payoff of $(5/2, 5/2)$ that is not a Nash equilibrium payoff.

Traffic Intersection Game (1)

Consider a game where two cars arrive at an intersection simultaneously. Row player (player 1) has the option to play U or D, and the column player (player 2) has the option to play L or R with payoffs as follows.

	L	R
U	5, 1	0, 0
D	4, 4	1, 5

- There are two pure strategy Nash equilibria: (U, L) and (D, R).
- To find the mixed strategy Nash equilibria, assume player 1 plays U with probability p and player 2 plays L with probability q. Using the mixed equilibrium characterization, we have

$$5q = 4q + (1-q) \Rightarrow q = \frac{1}{2}$$
$$5p = 4p + (1-p) \Rightarrow p = \frac{1}{2}$$

- This implies that there is a unique mixed strategy equilibrium with expected payoff (5/2,5/2).

Traffic Intersection Game (2)

- Assume that there is a publicly observable random variable (such as a fair coin) such that with probability 1/2 (Head), player 1 plays U and player 2 plays L, and with probability 1/2 (Tail), player 1 plays D and player 2 plays R.
- The expected payoff for this play of the game increases to (3,3).
- We show that no player has an incentive to deviate from the "recommendation" of the coin.
- If player 1 sees a Head, he believes that player 2 will play L, and therefore playing U is his best response (similar argument when he sees a Tail).
- Similarly, if player 2 sees a Head, he believes that player 1 will play U, and therefore playing L is his best response (similar argument when he sees a Tail).
- When the recommendation of the coin is part of a Nash equilibrium, no player has an incentive to deviate

---

Traffic Intersection Game (3)

- With a publicly observable random variable, we can get any payoff vector in the convex hull of Nash equilibrium payoffs.
  - Note that the convex hull of a finite number of vectors $x_1, \dots, x_k$ is given by

$$\text{conv}(\{x_1, \dots, x_k\}) = \{x \mid x = \sum_{i=1}^{k} \lambda_i x_i,\ \lambda_i \geq 0,\ \sum_{i=1}^{k} \lambda_i = 1\}$$

- The coin flip is one way of communication prior to the play.
- A more general form of communication is to find a trusted mediator who can perform general randomizations.
- Consider next a more elaborate signalling scheme.
- Suppose the players find a mediator who chooses $x \in \{1, 2, 3\}$ with equal probability 1/3. She then sends the following messages:
  - If $x = 1$, player 1 plays U, player 2 plays L.
  - If $x = 2$, player 1 plays D, player 2 plays L.
  - If $x = 3$, player 1 plays D, player 2 plays R.

Traffic Intersection Game (4)

- We show that no player has an incentive to deviate from the "recommendation" of the mediator:
  - If player 1 gets the recommendation U, he believes player 2 will play L, so his best response is to play U.
  - If player 1 gets the recommendation D, he believes player 2 will play L, R with equal probability, so playing D is a best response.
  - If player 2 gets the recommendation L, he believes player 1 will play U, D with equal probability, so playing L is a best response.
  - If player 2 gets the recommendation R, he believes player 1 will play D, so his best response is to play R.
- Thus the players will follow the mediator's recommendations.
- With the mediator, the expected payoffs are $(10/3, 10/3)$, strictly higher than what the players could get by randomizing between Nash equilibria.

Correlated Equilibrium (1)

- The preceding examples lead us to the notions of correlated strategies and "correlated equilibrium".
- Let $\Delta(S)$ denote the set of probability measures over the set S.
- Let R be a random variable taking values in $S = \Pi_{i=1}^{n} S_i$ distributed according to π.
  - An instantiation of R is a pure strategy profile and the i^{th} component of the instantiation will be called the recommendation to player i.
  - Given such a recommendation, player i can use conditional probability to form a posteriori beliefs about the recommendations given to the other players.

Correlated Equilibrium (2)

- $G = (N, \{S_i\}_{i \in N}, \{u_i\}_{i \in N})$

- Denote S = $\prod S_i$ and let q be a probability distribution on S

- Denote the probability of s ∈ S by q(s).

- The prob. dist. q is a correlated equilibrium if, for every i ∈ N, every s_i, $s'_i \in S_i$

$$\sum_{s_{-i} \in S_{-i}} q(s_i, s_{-i}) u_i(s_i, s_{-i}) \geq \sum_{s_{-i} \in S_{-i}} q(s_i, s_{-i}) u_i(s'_i, s_{-i})$$

If we divide both sides by $q(s_i)$ (= $\sum_{s_{-i} \in S_{-i}} q(s_i, s_{-i})$)

$$\underbrace{\sum_{s_{-i} \in S_{-i}} q(s_{-i}|s_i) u_i(s_i, s_{-i})}_{\text{i'th conditional expected payoff from playing } s_i} \geq \underbrace{\sum_{s_{-i} \in S_{-i}} q(s_{-i}|s_i) u_i(s'_i, s_{-i})}_{\text{i'th conditional expected payoff from playing } s'_i}$$

Interpretation of Correlated Equilibrium

$$\sum_{s_{-i} \in S_{-i}} q(s_{-i}|s_i) u_i(s_i, s_{-i}) \geq \sum_{s_{-i} \in S_{-i}} q(s_{-i}|s_i) u_i(s'_i, s_{-i})$$

- Consider the following 2 stage procedure

1. Recommendation : an action-tuple s ∈ S is drawn via the q.

Each player i is told only his part of the outcome (i.e., s_i).

2. Switch : each player is given the chance to switch to an alternative action $s'_i \neq s_i$

- If q is a correlated equilibrium, i's conditional expected payoff from playing alternative action $s'_i \neq s_i$ is no higher than playing drawn action s_i
- If q is a product measure, i.e. the play of different player is independent, correlated equilibrium is equal to Nash eq.

Lecture 15:
Learning in Games:

Part II: Partially Distributed

Yi, Yung (이융)
KAIST, Electrical Engineering
http://lanada.kaist.ac.kr
yiyung@kaist.edu

Contents

- Environment
- 1. Best Response Dynamics
- 2. Fictitious Play Based Learning
- 3. Logit Equilibrium Learning (Logic Dynamics)
  (smoothed fictitious play)

Environment of 3 algorithms

- Every player knows his utility function and can observe at each stage the actions played by others
- They don't know the other's utility function.
  ⇒ Don't know equilibria of game
- Continuous/discrete action space (BRD, FPL)
  Only discrete action space (LOGIT)

How to define update time?

- Discrete time
  - Synchronous: Each player updates simultaneously
  - Asynchronous: At each time, only one player updates.
    - For fair update chances, it is often assumed "sequential"
- Continuous time
  - Often, assume that each player has its own Poisson clock with, say, unite rate
  - No simultaneous update
  - Long-term fair update chances
- You can use your own update timing assumption, depending on the target applications

---

1. Best Response Dynamics

Best Response Dynamics (BRD)

- Procedure (Asynchronous/Sequential Version)

1. Starting state, say a(0)=(a_1(0),a_2(0),…,a_k(0))
2. Player i updates his action to his best response to a_{-i}(0).
3. Player j updates his action to his best response to the new action profile which only one action has been updated.
4. Another player updates his action to his best response to the new action profile which two actions have been updated. And so on.

- There also exists a synchronous BRD

$$a_i(t+1) = B_i(a_{-i}(t))$$

Synchronous BRD with or without inertia

- Synchronous BRD

$$a_i(t+1) = B_i(a_{-i}(t))$$

  - Often, experience oscillations

- Synchronous BRD with inertia

$$a_i(t+1) = \begin{cases} a_i(t) & \text{with probability } \epsilon \\ \arg\max_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i}(t)) & \text{with probability } 1-\epsilon \end{cases}$$

  - Emulates "asynchronous" BRD
  - Helps in avoiding oscillations

Convergence of BRD

- Not many general results on the convergence of BRD
- Often, ad-hoc proof has to be done
- Convergence for potential games
- Young (2004)
  - In weakly acyclic games,
  - A-BRD converges with probability one to a pure Nash equilibrium
  - As discussed, weakly acyclic games is the superset of potential games
    - WAG is just a sufficient condition for the convergence
- Need to check the convergence of BRD for the corresponding applications

---

2. Fictitious Play

Fictitious Play (mixed strategy)

- A learning rule is of the form

$$p_i(t+1) = f(a(1), a(2), ..., a(t); U_i)$$

- Define empirical frequencies $q_i(t)$ as follows:

How many a_i are played until t

$$q_i^{a_i}(t) = \frac{1}{t}\sum_{\tau=1}^{t} I\{a_i(\tau) = a_i\}$$

- Fictitious play: Each player best responds to empirical frequencies

$$a_i(t+1) \in \arg\max_{a_i^* \in \mathcal{A}_i} \sum_{a_{-i} \in \mathcal{A}_{-i}} \left(u_i(a_i^*, a_{-i}) \boxed{\prod_{j \neq i} q_j^{a_j}(t)} \right)$$

Empirical distribution of a_{-i}

Difference from BRD?

---

3. Logit Learning or Logit Dynamics

Logit Learning

- Logit (or Boltzmann-Gibbs) learning can be interpreted as a variant of fictitious play or variant of best-response dynamics

- A modification of fictitious play
- Players' responses are smoothed by small random trembles.
- Let $x_{-i}(t) \in \Delta(A_{-i})$ be i's forecast for the opponents' behavior at time t
  - $x_{-i}(t)$ can be: (i) (ii)
- i chooses his action a_i with probability

$$q^i(a_i | x_{-i}(t)) = \frac{e^{\frac{1}{\gamma_i} u_i(a_i, x_{-i}(t))}}{\sum_{a_i' \in A_i} e^{\frac{1}{\gamma_i} u_i(a_i', x_{-i}(t))}}$$

Logistic function (or Bolzmann-Gibbs Distribution)

Where $q^i(a_i)$ is the prob. that i chooses action a_i

Logit Learning (cont'd)

- $q^i(a_i | x_{-i}(t)) = \frac{e^{\frac{1}{\gamma_i} u_i(a_i, x_{-i}(t))}}{\sum_{a_i' \in A_i} e^{\frac{1}{\gamma_i} u_i(a_i', x_{-i}(t))}}$

- If $\gamma_i > 0$ is close to 0,
  - the learning rule closely approximates a best response.
- If γ_i becomes larger,
  - the learning tends to a uniform distribution.

  ⇒ γ_i : level of rationality of player i

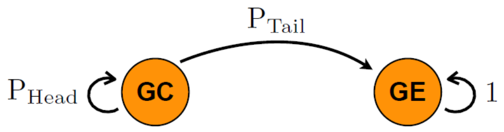
Summary

Convergence of BRD

- Young (2004)
  - In weakly acyclic games,
  - A-BRD converges with probability one to a pure Nash equilibrium
  - As discussed, weakly acyclic games is the superset of potential games
    - WAG is just a sufficient condition for the convergence
- Things to do
  - We have to see some papers which use BRD and prove that it converges to pure NE.
    - Case-by-case
  - Are there such papers? Need to google

- Almost sure

- Recall: St. Petersburg Paradox
  - Player repeatedly flips coing
  - Game terminates when player first flips Tail
  - Payoff is 2^x where x is the time where player first flips Tail
- Does the game end "surely" or "almost surely"?
- Inspect: We can represent the game as a *Markov chain*



- Enumerate all possible game trajectories that result in end
  - $z_1 = \{T\}$ with probability $\mathrm{P}_{\mathrm{Tail}}$
  - $z_2 = \{H, T\}$ with probability $(\mathrm{P}_{\mathrm{Head}})\,\mathrm{P}_{\mathrm{Tail}}$
  - $z_3 = \{H, H, T\}$ with probability $(\mathrm{P}_{\mathrm{Head}})^2\,\mathrm{P}_{\mathrm{Tail}}$
  - $z_k = \{H, ..., H, T\}$ with probability $(\mathrm{P}_{\mathrm{Head}})^{k-1}\,\mathrm{P}_{\mathrm{Tail}}$
- Probability that game ends

$$\sum_{k=1}^{\infty} \mathbf{Prob}[z_k] = \mathrm{P}_{\mathrm{Tail}} \sum_{k=0}^{\infty} (\mathrm{P}_{\mathrm{Head}})^k = 1$$

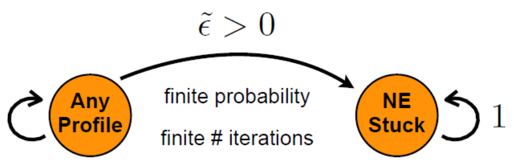
- The game ends almost surely... not surely.. Why?
- Consider the sequence $\{T, T, T, ...\}$
- The sequence does not result in end game. Probability of sequence is 0.
- Almost surely $=$ surely for all practical purposes

- Proof of BR

- Fact: Cournot dynamics with inertia converges almost surely to a NE in potential games

$$a_i(t+1) = \begin{cases} a_i(t) & \text{with probability } \epsilon \\ \arg\min_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i}(t)) & \text{with probability } 1-\epsilon \end{cases}$$

- Proof possesses similar idea to previous page



- What is the probability of always avoiding path for k consecutive timesteps

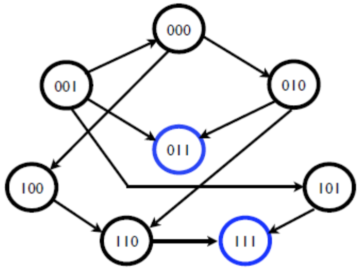
$$(1-\tilde{\epsilon})^k \to 0$$

- Can just lower bound probability from any action profile a and results still holds

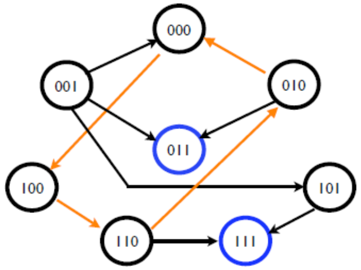
$$P_a \geq \tilde{\epsilon} > 0$$

- Direction of proof: Find $\tilde{\epsilon}$ and finite $\#$ iterations such that statement is true
- Inspection: Suppose $a(t) = a^{\mathrm{ne}}$. Then $a(\tau) = a^{\mathrm{ne}}$ for all times $\tau \geq t$
- Revised direction: Find $\tilde{\epsilon}$ and finite $\#$ iterations such that a NE is played once

- Key insight: Use better reply graph to find $\tilde{\epsilon}$ and bound on # iterations



Potential Game



Weakly Acyclic Game

- The following sequence of actions occurs with positive probability $\tilde{\epsilon} > 0$:
  - All but 1 player uses inertia, and non-inertia player does best reply
  - Repeat until NE
  - Once NE reached – stuck
- What is an upper bound on the number of iterations in above sequence?

Summary of BRD (Conjecture)

- Asynchronous (or sequential) BRD
  - Converges to NE (for a special class of the game mentioned earlier)
- Synchronous
  - BRD with inertia: Similarly converges to NE
  - Pure BRD: May not converge due to possible oscillations (even in potential games?)
- Basic idea on convergence
  - Asynchronous or inertia: 동시에 update를 해서, oscillation이 나지 않을 가능성 확보

Fictitious Play (pure strategy)

$$a_i(t+1) \in \arg\max_{a_i^* \in \mathcal{A}_i} \sum_{a_{-i} \in \mathcal{A}_{-i}} \left(u_i(a_i^*, a_{-i}) \prod_{j \neq i} q_j^{a_j}(t) \right)$$

- 상대편마다, 특정 action을 얼마나 많이 play했는가에 대한 2차원 matrix를 유지하여, 그것을 가지고 자신의 action을 선택

Fictitious Play (mixed strategy)

- A learning rule is of the form

$$p_i(t+1) = f(a(1), a(2), ..., a(t); U_i)$$

- Define empirical frequencies $q_i(t)$ as follows:

How many a_i are played until t

$$q_i^{a_i}(t) = \frac{1}{t} \sum_{\tau=1}^{t} I\{a_i(\tau) = a_i\}$$

- Fictitious play: Each player best responds to empirical frequencies

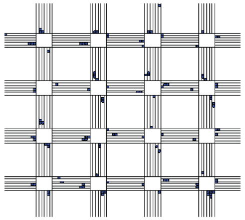
$$p_i(t+1) \in \arg\max_{p_i \in \Delta(\mathcal{A}_i)} u_i(p_i, q_{-i}(t))$$

where

$$u_i(p_i, q_{-i}(t)) = \sum_{a \in \mathcal{A}} u_i(a_1, a_2, ..., a_n) p_i^{a_i} \prod_{j \neq i} q_j^{a_j}(t)$$

- FP facts: Beliefs (i.e., empirical frequencies) converge to NE for
  - For 2-player games with 2 moves per player
  - Zero sum games with arbitrary moves per player
  - **Potential games**

- App of FP



- FP requirements for distributed routing:
  - Track empirical frequencies $q_j(t)$
  - Compute best response (focus on just pure strategies)

$$a_i(t+1) \in \arg\max_{a_i^* \in \mathcal{A}_i} \sum_{a_{-i} \in \mathcal{A}_{-i}} \left(u_i(a_i^*, a_{-i}) \prod_{j \neq i} q_j^{a_j}(t) \right)$$

- Question: How big is $\mathcal{A}_{-i}$?
  - 2 moves, 2 players → 4 joint actions
  - 2 moves, 3 players → 8 joint actions
  - 2 moves, n players → 2^n joint actions
  - m moves, n players → m^n joint actions
- Problem #1: Computing best response computationally prohibitive
- Problem #2: Observing action profiles prohibitive
- Engineering perspective: Develop simple algorithms with similar guarantees

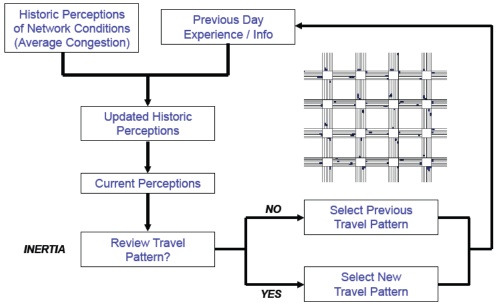
---

Joint Strategy vs. Individual Strategy

- 관점
  - Player i의 입장에서 상대방의 strategy를 상대방마다 추적을 기록을 할 것인가, 아니면, 상대방 전체를 하나로 보고 할 것인가?
- Individual
  - Fictitious play of the earlier slides
  - Sometimes, called Brown's FP (1951)
- Joint strategy
  - 다음장에서 볼 것.
- 2개가 차이가 있기는 한 것인가? (곧 볼 것)

Models of Behavior

- Actual drivers do not select driving pattern by FP
- Consider the following driver behavioral model


Historic Perceptions of Network Conditions (Average Congestion)
Previous Day Experience / Info
Updated Historic Perceptions
Current Perceptions
INERTIA
Review Travel Pattern?
NO
YES
Select Previous Travel Pattern
Select New Travel Pattern

- Description of model
  - Let $a_i(t)$ be the route chosen by player i at day t
  - Let $V_i^{a_i}(t)$ be the average congestion on route a_i up to day t

$$V_i^{a_i}(t) = \frac{1}{t}\sum_{\tau=1}^{t} u_i(a_i, a_{-i}(t)) = \frac{1}{t}\Big(u_i(a_i, a_{-i}(t)) + (t-1)V_i^{a_i}(t-1)\Big)$$

  - Decision rule day $t+1$:

$$a_i(t+1) = \begin{cases} a_i(t) & \text{with probability } \epsilon \\ \arg\min_{a_i \in \mathcal{A}_i} V_i^{a_i}(t) & \text{with probability } 1-\epsilon \end{cases}$$

  where $\epsilon \in (0,1)$ is referred to as the player's inertia.
- Think of model as algorithm. Asymptotic guarantees?

Joint Strategy FP

- Previous algorithm similar to fictitious play
- Define $z_{-i}(t)$ as the empirical frequency of other player's joint moves up to time t

$$z_{-i}^{a_{-i}}(t) = \frac{1}{t}\sum_{\tau=1}^{t} I\{a_{-i}(\tau) = a_{-i}\}$$

  Note that $z_{-i}^{a_{-i}(t)} \neq \prod_{j \neq i} q_j^{a_j(t)}$
- Key: Player views all other players as a single player with action set $\mathcal{A}_{-i}$
- Presumption: Each player presumes all other players playing collectively according to joint strategy $z_{-i}(t)$
- Is this tractable?

$$\begin{aligned} u_i(a_i, z_{-i}(t)) &= \sum_{a_{-i} \in \mathcal{A}_i} u_i(a_i, a_{-i}) z_{-i}^{a_{-i}}(t) \\ &= \frac{1}{t}\sum_{\tau=1}^{t} u_i(a_i, a_{-i}(\tau)) \\ &= V_i^{a_i}(t) \end{aligned}$$

- JSFP w/ inertia restated:

$$a_i(t+1) = \begin{cases} a_i(t) & \text{with probability } \epsilon \\ \arg\max_{a_i \in \mathcal{A}_i} u_i(a_i, z_{-i}(t)) & \text{with probability } 1-\epsilon \end{cases}$$

- **Fact:** JSFP w/ inertia converges almost surely to pure NE in (generic) potential games

Generalized JSFP

$$\beta_{-i}^{a_{-i}}(t) = \frac{\left(z_{-i}^{a_{-i}}(t)\right)^{\gamma_i}}{\sum_{a'_{-i}\in\mathcal{A}_{-i}}\left(z_{-i}^{a'_{-i}}(t)\right)^{\gamma_i}}$$

- If \gamma_i =1, then just a JSFP.
- We are going to look at another variant of JSFP.

Logit Learning (cont'd)

- $q^i(a_i|x_{-i}(t)) = \frac{e^{\frac{1}{\gamma_i}u_i\ (a_i\ ,x_{-i}(t))}}{\sum_{a_i'\in A_i} e^{\frac{1}{\gamma_i}u_i\ (a_i',x_{-i}(t))}}$
- If $\gamma_i > 0$ is close to 0,
  - – the learning rule closely approximates a best response.
- If γ_i becomes larger,
  - – the learning tends to a uniform distribution.
  - $\Rightarrow$ γ_i : level of rationality of player i
- **Logit equilibrium**

- the limit of the logit learning procedure if it converges.
- ε-Nash equilibrium

Logit Equilibirum

- Let q_j^i be the prob. that player i chooses j'th action.

Given a distribution q^i on $\Delta(A_i)$, the amount of information conveyed by q^i : $-\sum_j q_j^i \ln q_j^i$ (Shannon entropy function)

- Player i's actual (or modified) utility U$_i$

$$U_i(q^i, x_{-i}) = \underbrace{u_i(q^i, x_{-i})}_{\text{Current payoff}} - \gamma_i \underbrace{\sum_j q_j^i \ln q_j^i}_{\text{Information}}$$

- Optimal (or maximizer) q^i is given by the Logistic function
  - – Please check!

Convergence of Logit Learning

- A potential game is a case the logit learning converges.
- The convergence time to a η-Nash equilibrium under the logit learning is of order $K \log\log K + \log\frac{1}{\eta}$

Lecture 16:
Learning in Games:

Part III: Fully Distributed

Yi, Yung (이융)
KAIST, Electrical Engineering
http://lanada.kaist.ac.kr
yiyung@kaist.edu

Taxonomy: Learning Algorithms

- Fully distributed (or Uncoupled) learning algorithm
  - A player does not use information about the other players.
  - Builds his strategies and updates them by using own-actions and own-utilities.
  - Can we even do something with this small information? Maybe
- Partially distributed learning algorithm
  - A player implements his updating rule after receiving some data about others.
  - The amount and the kind of data may depend on each algorithm (We will discuss this later)

Contents

- Trial and error learning
- Regret matching based learning
- Reinforcement learning
- Boltzmann-Gibbs learning

Contents

- Trial and error learning
- Regret matching based learning
- Reinforcement learning
- Boltzmann-Gibbs learning

Trial and Error Learning (TEL)

- Also, called Learning by Experimenting
- Environment
  - All players do not know their utility functions, only the value of the function at given stage is known.
- Procedure

1. At time t,
   - Player i is in state $s_{i,t} = (a_{i,t}, u_{i,t})$
2. At t+1,
   - Before choosing an action, each player i does experiment with probability $\varepsilon_i \in (0,1)$
   - If he does not experiment, $a_{i,t+1} = a_{i,t}$
   - Otherwise, he plays $a_i' \in A_i$ drawn uniformly at random. Only if received utility $u_{i,t+1}$ is greater than old $u_{i,t}$ he adopts the new state $s_{i,t+1} = (a_i', u_{i,t+1})$

Convergence of TEL (Young 2009)

- Let G_A be the set of all K-player games on the finite action space A that has at least one pure Nash eq.

- [Theorem] If all players use interactive trial and learning and the experimentation rate ε is sufficiently small, then for games in G_A a Nash equilibrium is played at least 1- ε proportion of the time.

Contents

- Trial and error learning

- Regret matching based learning

- Reinforcement learning

- Boltzmann-Gibbs learning

Regret-Matching (Young 2000)

- Single-agent learning before multi-agent learning in game
  - Setup:
    - Two players: Player 1 vs. Nature
    - Actions set: $\mathcal{A}_1$ and $\mathcal{A}_N$
    - Payoffs: $u : \mathcal{A}_1 \times \mathcal{A}_N \rightarrow R$

		Nature	
		Rain	No Rain
P_1	Umbrella	1	0
	No umbrella	0	1

Player 1's Payoff

- Player repeatedly interacts with nature
  - Player's action day t: $a_1(t)$
  - Nature's action day t: $a_N(t)$
  - Payoff day t: $u(a_1(t), a_N(t))$
- Goal: Implement strategy that provides guarantees with regard to average performance?

Regret

- Challenge: Hard to predict what nature is going to do
- Thoughts: Can a player optimize “what if” scenarios?
- Definition: Player’s average payoff at day t

$$\bar{u}(t) = \frac{1}{t}\sum_{\tau=1}^{t} u(a_1(\tau), a_N(\tau))$$

- Definition: Player’s perceived average payoff at day t if committed to fixed action and nature was unchanged

$$\bar{v}^{a_1}(t) = \frac{1}{t}\sum_{\tau=1}^{t} u(a_1, a_N(\tau))$$

- Definition: Player’s *regret* at day t for not having used action a_1

$$\bar{R}^{a_1}(t) = \bar{v}^{a_1}(t) - \bar{u}(t)$$

If this value is large, then I have to think “oh I have to choose a_1 more from now on”

- Example:

Day	1	2	3	4	5	6	…
Player’s Decision	NU	U	NU	U	NU	NU	…
Nature’s Decision	R	NR	R	R	NR	R	…
Payoff	0	0	0	1	1	0	…

  - $\bar{u}(6)$?
  - $\bar{v}^{U}(6)$?
  - $\bar{v}^{NU}(6)$?
  - $\bar{R}^{U}(6)$?
  - $\bar{R}^{NU}(6)$?

Regret-Matching Learning

- Positive regret = Player could have done something better in hindsight
- Q: Is it possible to make positive regret vanish asymptotically “irrespective” of nature?
- Consider the strategy *Regret Matching*: At day t play strategy $p(t) \in \Delta(\mathcal{A}_1)$

$$\begin{aligned} p^{U}(t+1) &= \frac{[\bar{R}^{U}(t)]_+}{[\bar{R}^{U}(t)]_+ + [\bar{R}^{NU}(t)]_+} \\ p^{NU}(t+1) &= \frac{[\bar{R}^{NU}(t)]_+}{[\bar{R}^{U}(t)]_+ + [\bar{R}^{NU}(t)]_+} \end{aligned}$$

- Notation: $[\cdot]_+$ is projection to positive orthant, i.e., $[x]_+ = \max\{x, 0\}$
- Strategy generalizes to more than two actions
- **Fact:** Positive regret asymptotically vanishes irrespective of nature

$$\begin{aligned} [\bar{R}^{U}(t)]_+ &\to 0 \\ [\bar{R}^{NU}(t)]_+ &\to 0 \end{aligned}$$

A randomized strategy with more probability to the strategy with larger regret

---

- Example revisited:

Day	1	2	3	4	5	6	…
Player’s Decision	NU	U	NU	U	NU	NU	…
Nature’s Decision	R	NR	R	R	NR	R	…
Payoff	0	0	0	1	1	0	…

  - Regret matching strategy day 2?
  - Regret matching strategy day 3?
  - Regret matching strategy day 4?
  - Regret matching strategy day 5?
  - Regret matching strategy day 6?
- Why does positive regret vanish?

More formally

- Learning rule $p_i(t+1) = f(a(0), a(1), \dots, a(t); u_i)$
  - Consider the learning rule $f(\cdot)$ where

$$p_i^{a_i}(t+1) = \frac{[\bar{R}_i^{a_i}(t)]_+}{\sum_{\tilde{a}_i \in \mathcal{A}_i} [\bar{R}_i^{\tilde{a}_i}(t)]_+}$$

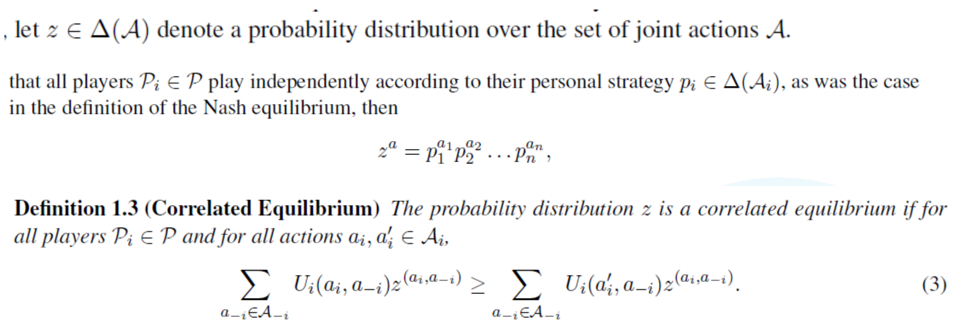
    - $p_i^{a_i}(t+1)$ = Probability player i plays action a_i at time $t+1$
    - $\bar{R}_i^{a_i}(t)$ = Regret of player i for action a_i at time t
  - Fact: Max regret of all players goes to 0 (think of other players as "nature")

$$[\bar{R}_i^{a_i}(t)]_+ \to 0$$

  - Result restated: The behavior converges to a "no-regret" point

---

Correlated Equilibrium (Remind)



, let $z \in \Delta(\mathcal{A})$ denote a probability distribution over the set of joint actions $\mathcal{A}$.

that all players $\mathcal{P}_i \in \mathcal{P}$ play independently according to their personal strategy $p_i \in \Delta(\mathcal{A}_i)$, as was the case in the definition of the Nash equilibrium, then

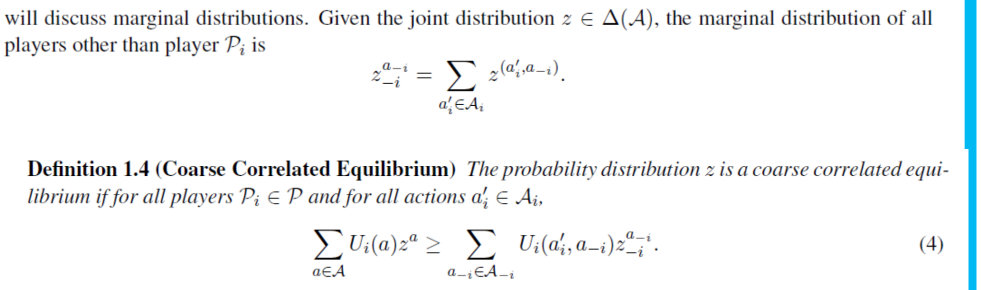
$$z^a = p_1^{a_1} p_2^{a_2} \dots p_n^{a_n},$$

Definition 1.3 (Correlated Equilibrium) *The probability distribution z is a correlated equilibrium if for all players $\mathcal{P}_i \in \mathcal{P}$ and for all actions $a_i, a_i' \in \mathcal{A}_i$,*

$$\sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i, a_{-i}) z^{(a_i, a_{-i})} \geq \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i', a_{-i}) z^{(a_i, a_{-i})}. \quad (3)$$

- How to interpret?
  - Joint action a is randomly chosen by Nature according to z
  - Each player is informed of his action a_i
  - Each player has a chance to change his action a_i
  - Each player i's conditional expected payoff for action a_i is at least as good as for other action a_i'

Coarse Correlated Equilibrium



will discuss marginal distributions. Given the joint distribution $z \in \Delta(\mathcal{A})$, the marginal distribution of all players other than player $\mathcal{P}_i$ is

$$z_{-i}^{a_{-i}} = \sum_{a_i' \in \mathcal{A}_i} z^{(a_i', a_{-i})}.$$

Definition 1.4 (Coarse Correlated Equilibrium) *The probability distribution z is a coarse correlated equilibrium if for all players $\mathcal{P}_i \in \mathcal{P}$ and for all actions $a_i' \in \mathcal{A}_i$,*

$$\sum_{a \in \mathcal{A}} U_i(a) z^a \geq \sum_{a_{-i} \in \mathcal{A}_{-i}} U_i(a_i', a_{-i}) z_{-i}^{a_{-i}}. \quad (4)$$

- How to interpret?
  - Before the joint action a is drawn, each player i is given the chance to opt out, in which case she chooses any action beforehand
  - If he does not opt out, then follow Nature's suggestion
  - No player choose to opt out given that all other players opt to stay in

---

No-regret point? NE or what?

- Rewrite regret in terms of empirical frequency $z(t) \in \Delta(\mathcal{A})$

$$\bar{u}_i(t) = \frac{1}{t}\sum_{\tau=1}^{t} u_i(a(\tau)) = u_i(z(t))$$

$$\bar{v}_i^{a_i}(t) = \frac{1}{t}\sum_{\tau=1}^{t} u_i(a_i, a_{-i}(t)) = u_i(a_i, z_{-i}(t))$$

$$\bar{R}_i^{a_i}(t) = \bar{v}_i^{a_i}(t) - \bar{u}_i(t) = u_i(a_i, z_{-i}(t)) - u_i(z(t))$$

- Characteristic of no-regret point

$$\bar{R}_i^{a_i}(t) \leq 0 \quad \Leftrightarrow \quad u_i(a_i, z_{-i}(t)) \leq u_i(z(t))$$

- No-regret point restated: For any player i and action a_i

$$u_i(a_i, z_{-i}(t)) \leq u_i(z(t))$$

- No-regret point = Coarse correlated equilibrium
- **Theorem:** If all players follow the regret matching strategy then the empirical frequency converges to the set of coarse correlated equilibrium.

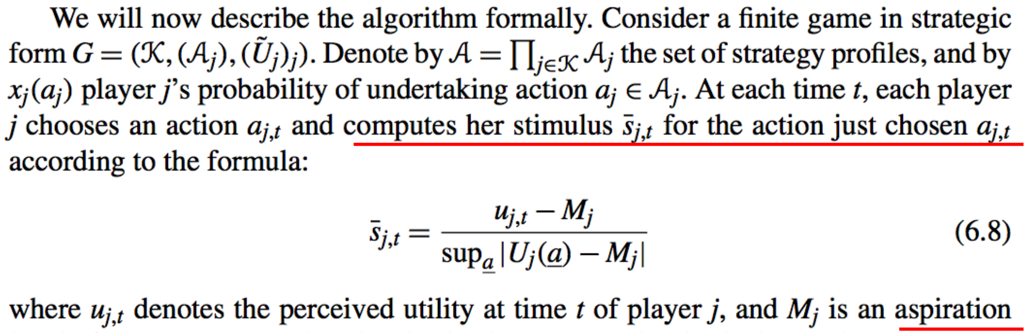
Contents

- Trial and error learning
- Regret matching based learning
- Reinforcement learning
- Boltzmann-Gibbs learning

Introduction for Reinforcement Learning

- Basic idea: Reinforcement (재강화?)
  - The higher the payoff from taking an action in the past, the more likely it will be taken in the future.
- Originally, developed for a single-agent situation (self-learning)
  - A classical machine learning topic
    - Supervised Learning
    - Unsupervised Learning
    - Reinforcement Learning
- Player only knows his past choices and perceived utilities.
- Using this information, each player tries to learn a strategy that maximizes expected future reward.

Bush & Mosteller Algorithm (1955) (1)



We will now describe the algorithm formally. Consider a finite game in strategic form $G = (\mathcal{K}, (\mathcal{A}_j), (\tilde{U}_j)_j)$. Denote by $\mathcal{A} = \prod_{j \in \mathcal{K}} \mathcal{A}_j$ the set of strategy profiles, and by $x_j(a_j)$ player j's probability of undertaking action $a_j \in \mathcal{A}_j$. At each time t, each player j chooses an action $a_{j,t}$ and computes her stimulus $\bar{s}_{j,t}$ for the action just chosen $a_{j,t}$ according to the formula:

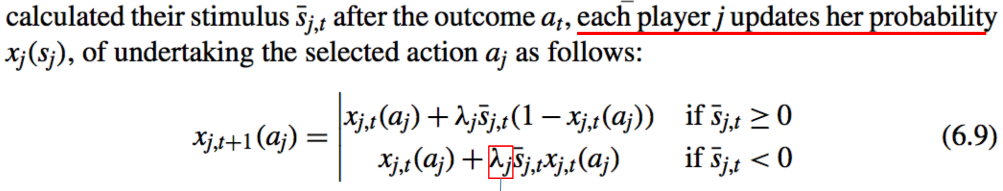
$$\bar{s}_{j,t} = \frac{u_{j,t} - M_j}{\sup_{\underline{a}} |U_j(\underline{a}) - M_j|} \qquad (6.8)$$

where $u_{j,t}$ denotes the perceived utility at time t of player j, and M_j is an aspiration level of player j. Hence the stimulus is always a number in the interval $[-1,1]$. Note

Stimulus: How am I satisfied about a particular action (measured by my aspiration level)?

Positive→ above aspiration, Negative→ below aspiration

Bush & Mosteller Algorithm (2)



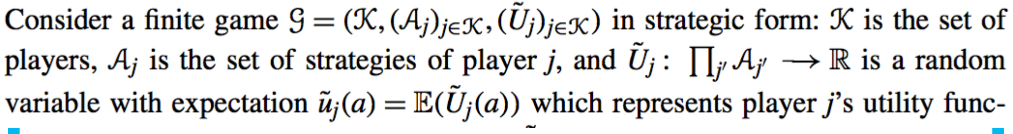
that player j is assumed to know the denominator $\sup_{\underline{a}} |U_j(\underline{a}) - M_j|$. Secondly, having calculated their stimulus $\bar{s}_{j,t}$ after the outcome a_t, each player j updates her probability $x_j(s_j)$, of undertaking the selected action a_j as follows:

$$x_{j,t+1}(a_j) = \left| \begin{array}{ll} x_{j,t}(a_j) + \lambda_j \bar{s}_{j,t}(1 - x_{j,t}(a_j)) & \text{if } \bar{s}_{j,t} \geq 0 \\ x_{j,t}(a_j) + \boxed{\lambda_j} \bar{s}_{j,t} x_{j,t}(a_j) & \text{if } \bar{s}_{j,t} < 0 \end{array} \right. \qquad (6.9)$$

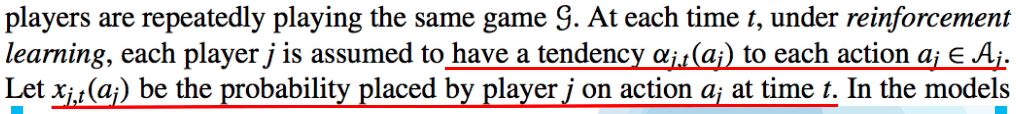
Learning rate

Larger stimulus → larger probability

Arthur's Algorithm (1993) (1)

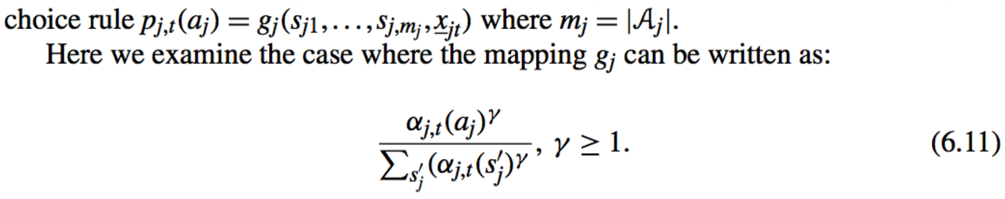


Consider a finite game $\mathcal{G} = (\mathcal{K}, (\mathcal{A}_j)_{j\in\mathcal{K}}, (\tilde{U}_j)_{j\in\mathcal{K}})$ in strategic form: $\mathcal{K}$ is the set of players, $\mathcal{A}_j$ is the set of strategies of player j, and $\tilde{U}_j : \prod_{j'} \mathcal{A}_{j'} \longrightarrow \mathbb{R}$ is a random variable with expectation $\tilde{u}_j(a) = \mathbb{E}(\tilde{U}_j(a))$ which represents player j's utility func-



players are repeatedly playing the same game $\mathcal{G}$. At each time t, under *reinforcement learning*, each player j is assumed to have a tendency $\alpha_{j,t}(a_j)$ to each action $a_j \in \mathcal{A}_j$. Let $x_{j,t}(a_j)$ be the probability placed by player j on action a_j at time t. In the models

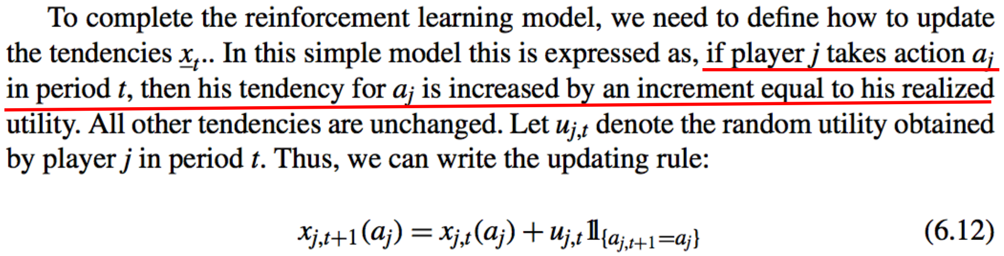
Arthur's Algorithm (1993) (2)



of reinforcement learning, we consider that these probabilities are determined by the choice rule $p_{j,t}(a_j) = g_j(s_{j1}, \ldots, s_{j,m_j}, \underline{x}_{jt})$ where $m_j = |\mathcal{A}_j|$.

Here we examine the case where the mapping g_j can be written as:

$$\frac{\alpha_{j,t}(a_j)^\gamma}{\sum_{s'_j}(\alpha_{j,t}(s'_j))^\gamma}, \gamma \geq 1. \tag{6.11}$$



To complete the reinforcement learning model, we need to define how to update the tendencies $\underline{x}_t$.. In this simple model this is expressed as, if player j takes action a_j in period t, then his tendency for a_j is increased by an increment equal to his realized utility. All other tendencies are unchanged. Let $u_{j,t}$ denote the random utility obtained by player j in period t. Thus, we can write the updating rule:

$$x_{j,t+1}(a_j) = x_{j,t}(a_j) + u_{j,t} \mathbb{1}_{\{a_{j,t+1}=a_j\}} \tag{6.12}$$

Contents

- Trial and error learning
- Regret matching based learning
- Reinforcement learning
- Boltzmann-Gibbs learning

Boltzmann-Gibbs Learning (Partially Distributed)

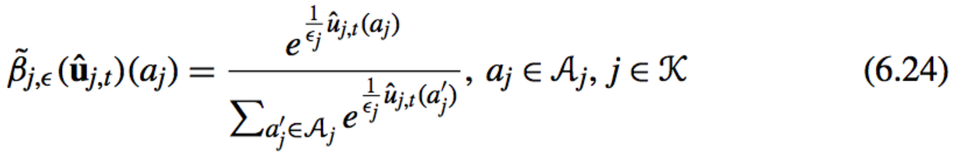
- Recall:

Compare the above with:

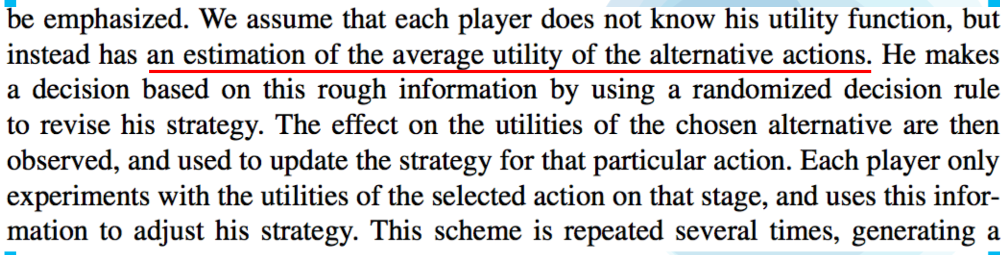
$$q^i(a_i|x_{-i}(t)) = \frac{e^{\frac{1}{\gamma_i}u_i(a_i,x_{-i}(t))}}{\sum_{a'_i \in A_i} e^{\frac{1}{\gamma_i}u_i(a'_i,x_{-i}(t))}}$$

Logistic function
(or Bolzmann-Gibbs
Distribution)

Boltzmann-Gibbs Learning (Fully Distributed)



$$\tilde{\beta}_{j,\epsilon}(\hat{\mathbf{u}}_{j,t})(a_j) = \frac{e^{\frac{1}{\epsilon_j}\hat{u}_{j,t}(a_j)}}{\sum_{a'_j \in \mathcal{A}_j} e^{\frac{1}{\epsilon_j}\hat{u}_{j,t}(a'_j)}}, \; a_j \in \mathcal{A}_j, \; j \in \mathcal{K} \qquad (6.24)$$



be emphasized. We assume that each player does not know his utility function, but instead has an estimation of the average utility of the alternative actions. He makes a decision based on this rough information by using a randomized decision rule to revise his strategy. The effect on the utilities of the chosen alternative are then observed, and used to update the strategy for that particular action. Each player only experiments with the utilities of the selected action on that stage, and uses this information to adjust his strategy. This scheme is repeated several times, generating a