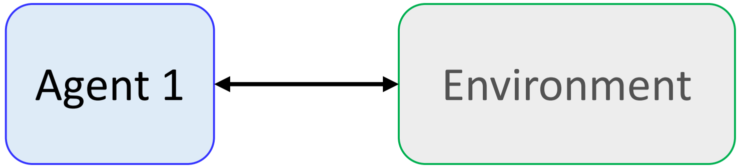
5. Learning in Repeated Games

Introduction

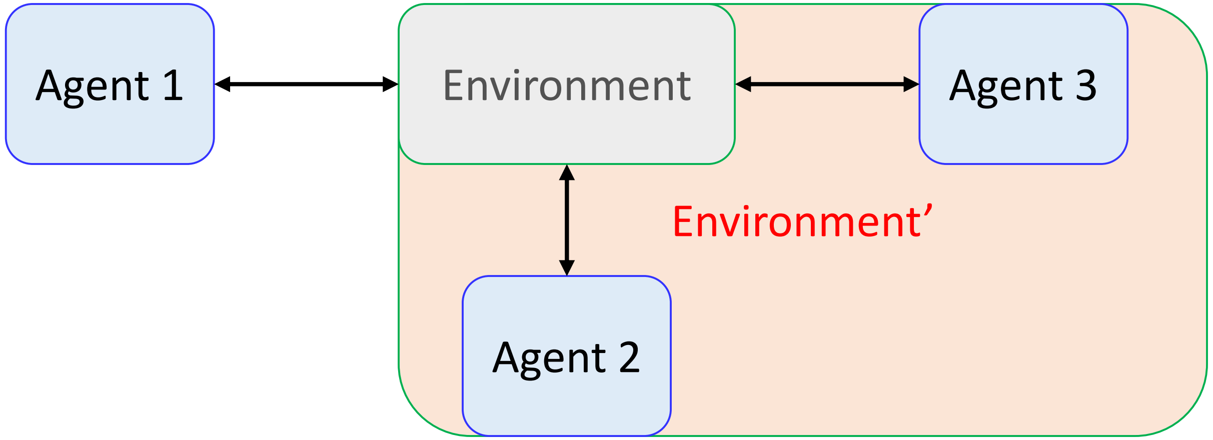
- We concentrate on techniques drawn primarily from two disciplines
  - Artificial intelligence
  - game theory

The integrations between learning and teaching



- Most work in **artificial intelligence** concerns the learning performed by an individual agent
  - The goal is to design an agent that learns to function successfully in an environment that is
    - unknown
    - (potentially) changing as the agent is learning

The integrations between learning and teaching



- **Multiagent setting** adds additional complexities
  - Environment contains other agents
    - Environment is changing as other agents are learning
    - Environment is changing depending on other agents’ actions
    - ➢ *The learning of the other agents will be impacted by the learning performed by our protagonist*

The simultaneous learning of the agents means that every learning rule leads to a dynamical system

The integrations between learning and teaching

	L	R
T	$1, 0$	$3, 2$
B	$2, 1$	$4, 0$

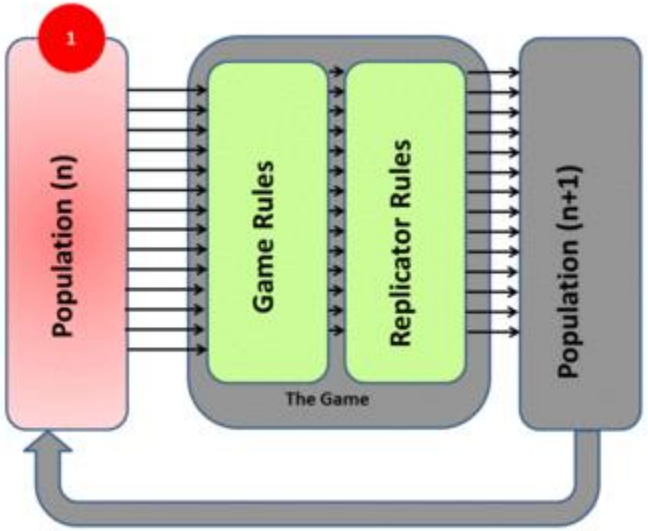
Stackelberg game

- Player 1 (the row player) has a dominant strategy, namely B.
- (B, L) is the unique Nash equilibrium of the game
  - ✓ If player 1 were to play B repeatedly, it is reasonable to expect that player 2 would always respond with L.
- What will happen if player 1 chooses to play T?
  - Then, player 2's best response would be R, yielding player 1 a payoff of 3 (>2 for Nash)
- In a single-stage game it would be hard for player 1 to convince player 2 that he (player 1) will play T, since it is a strictly dominated strategy.1
- However, in a repeated-game setting, player 1 could repeatedly play T; presumably, after a while player 2, if he has any sense at all, would get the message and start responding with R.

Player 1's role is a teacher, trying to modifying player 2's strategy

What constitutes learning?

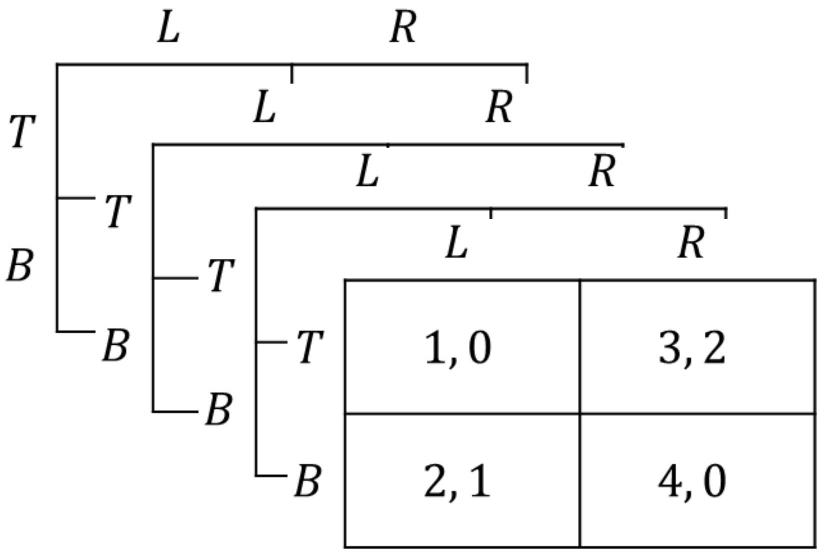
1. Evolutionary game



- Evolutionary game is for modeling large populations
  - Largely inspired by evolutionary biology
  - Consist of a large number of players, who repeatedly paly a given game among themselves

What constitutes learning?

2. Repeated (Matrix) game



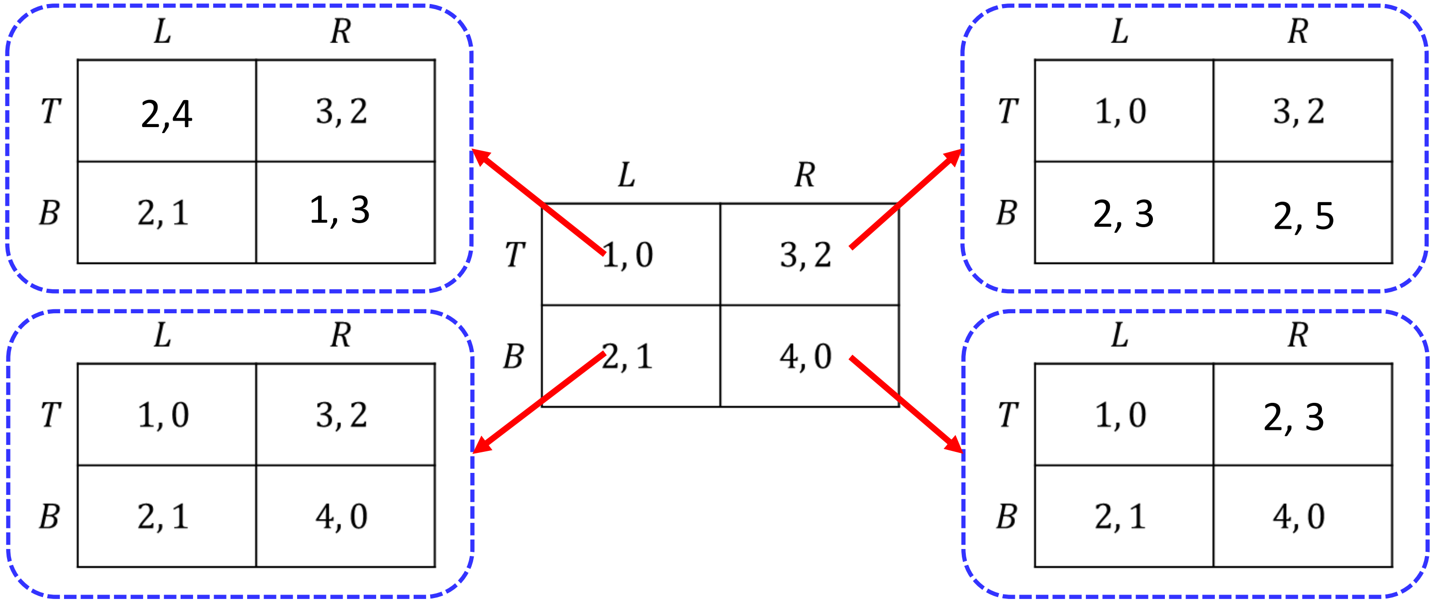
	L	R
T	$1, 0$	$3, 2$
B	$2, 1$	$4, 0$

- Repeated game is used for learning setting because
  - Each time the same players are involved
  - Each time the same game is played

- The experience so far → a strategy for selecting future action
  - Tit-for-Tat and Triger strategies in repeated PD game
  - A more general strategies can be obtained

What constitutes learning?

3. Stochastic game



	L	R
T	2,4	3, 2
B	2, 1	1, 3

	L	R
T	1, 0	3, 2
B	2, 1	4, 0

	L	R
T	1, 0	3, 2
B	2, 1	4, 0

	L	R
T	1, 0	3, 2
B	2, 3	2, 5

	L	R
T	1, 0	2, 3
B	2, 1	4, 0

- Stochastic game is a moral general setting where learning is taking place
  - The game transits to another game depending on the joint actions by agents
  - Same players and same actions sets are used through games
- Most of the techniques discussed in the context of repeated games are applicable more generally to stochastic games
  - ✓ specific results obtained for repeated games do not always generalize.

What constitutes learning?

Additional aspects for repeated and stochastic games

	Opponent's strategy unknown
Game is known	Need to learn only opponents' strategies
Game is unknown	Need to learn both the payoffs and opponents' strategies (e.g., Multiagents reinforcement learning)

Evolutionarily games and replicator dynamics

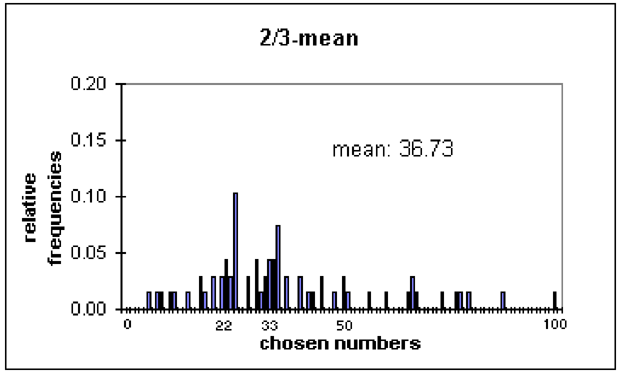
K-beauty contest

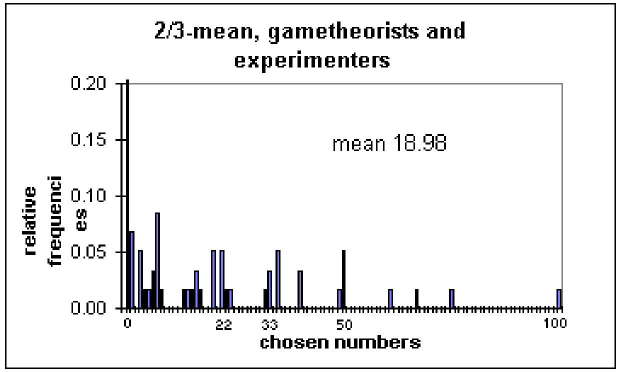
- **The rules of the basic beauty-contest game.**
  - Each person of ***N-players*** is asked to choose a (***real*** or ***integer***) number from the ***interval 0 to 100***.
  - The winner is the person whose choice is closest to ***p*** times the ***mean*** of the choices of all players (where p is, for example, ***2/3***).
  - The winner gets a ***fixed*** prize of ***$20***. In case of a tie the prize is ***split*** amongst those who tie.
  - The same game may be repeated ***several periods.*** Subjects ***are informed*** of the ***mean, 2/3 mean*** and ***all choices*** after each period

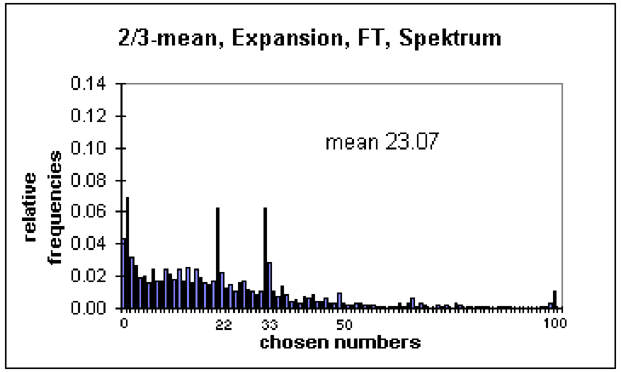
What would you play?
What is a Nash equilibrium?

K-beauty contest

- **The rules of the basic beauty-contest game.**
  - Each person of ***N-players*** is asked to choose a (***real*** or ***integer***) number from the ***interval 0 to 100***.
  - The winner is the person whose choice is closest to ***p*** times the ***mean*** of the choices of all players (where p is, for example, ***2/3***).
  - The winner gets a ***fixed*** prize of ***$20***. In case of a tie the prize is ***split*** amongst those who tie.
  - The same game may be repeated ***several periods.*** Subjects ***are informed*** of the ***mean, 2/3 mean*** and ***all choices*** after each period







Nagel (1995)

Introduction

- Most models of economic behavior are based on the assumption of rationality of economic agents and common knowledge of rationality.
- Do people play Nash equilibrium?
- In the context of the k-beauty game, we saw that even very smart game theorists do not play the unique Nash equilibrium (or the unique strategy profile surviving iterated elimination of strictly dominated strategies).
- Why?
  - Either because in new situations, it is often quite complex to work out what is “best”.
  - Or more likely, because, again in new situations, individuals are uncertain about how others will play the game.
- If we played the k-beauty game several more times, behavior would have approached or in fact reached the Nash equilibrium prediction.

Introduction

- This reasoning suggests the following:
- Perhaps people behave using simple rules of thumb; these are somewhat "myopic," in the sense that they do not involve full computation of optimal strategies for others and for oneself.
- But they are also "flexible" rules of thumb in the sense that they adapt and respond to situations, including to the (actual) behavior of other players.
- What are the implications of this type of adaptive behavior?
- Two different and complementary approaches:
  - evolutionary game theory
  - Learning in games.

Evolution and game theory

- The theory of evolution goes back to Darwin's classic, The Origins of Species (and to Wallace)
- Darwin focused mostly on evolution and adaptation of an organism to the environment in which it was situated. But in The Descent of Man, in the context of sexual selection, he anticipated many of the ideas of evolutionarily game theory.
- evolutionary game theory was introduced by John Maynard Smith in Evolution and the Theory of Games, and in his seminal papers, Maynard Smith (1972) "Game Theory and the Evolution of Fighting" and Maynard Smith and Price (1973) "The Logic of Animal Conflict".
- The theory was formulated for understanding the behavior of animals in game-theoretic situations (to a game theorist, all situations). But it can equally well be applied to modeling "myopic behavior" for more complex organisms—such as humans.

Evolution and game theory

- In its simplest form the story goes like this: each organism is born programmed to play a particular strategy.
- The game is the game of life—with payoffs given as fitness (i.e., expected number of off springs). If the organism is successful, it has greater fitness and more offspring, also programmed to play in the same way. If it is unsuccessful, it likely dies without offspring.
- Mutations imply that some of these offspring will randomly play any one of the feasible strategies.

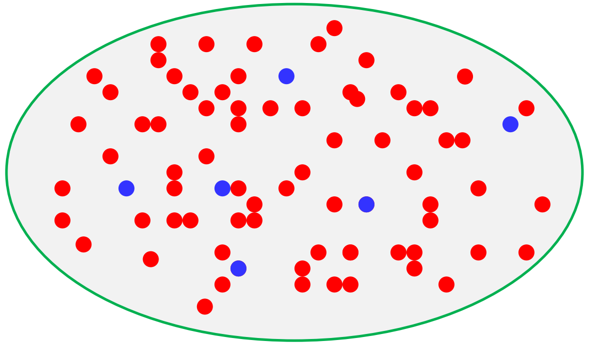
Two approaches to evolutionary Game Theory

- There are two approaches to evolutionarily game theory
  - The evolutionarily stable strategy:
    - Static property of the game
    - Equilibrium concept to analysis evolutionarily game
  - The evolutionary dynamics:
    - Dynamic property of the game
    - An explicit model of the process by which the frequency of strategies change in the population

The evolutionarily stable strategy:

The setting

Large population of agents



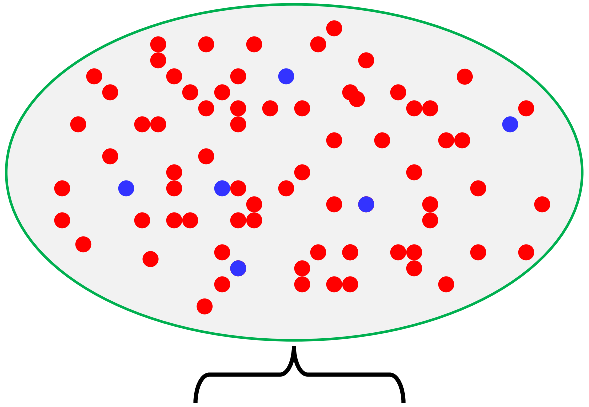
- Consider a large population of agents (organisms, animals, humans).
- Agents are designed to play a certain way
  - Incumbents (다수) play I
  - Mutants (돌연변이) play M (a small fraction)

- At each instant, each agent is randomly matched with one other agent from the population, whose distribution is modeled by the mixed strategy

- **Which strategy is evolutionally stable or not?**

The setting

Large population of agents



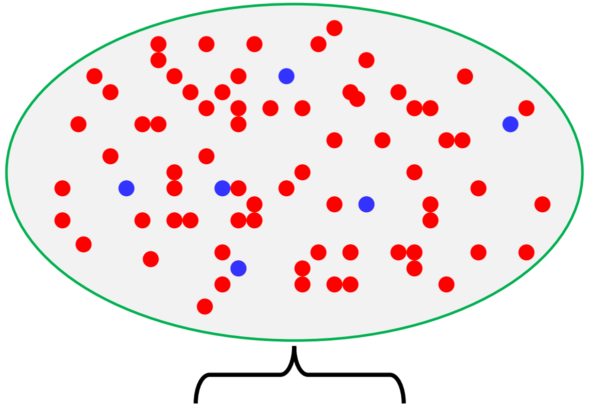
	$1-\epsilon$ I	ϵ M
● I	$3, 3$	$0, 4$
● M	$4, 0$	$1, 1$

- At each instant, each agent is randomly matched with one other agent from the population, and they play a symmetric strategic form game.
- The population is fictitiously represented by the column player
- The population's distribution is modeled by the mixed strategy $(1-\epsilon, \epsilon)$
  - Mutant: a small portion of ϵ of mutants plays M
  - Incumbent: the rest of the population $(1-\epsilon)$ play I

- The interaction between each pair of organisms can be modeled as a two player sym. game $G(N, A, u)$ with a common set A of actions and a payoff function u.
  - $u(a, a')$ is the expected reward, representing fitness
  - The fitness can indicates expected number of offspring

The setting

Large population of agents

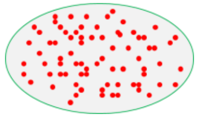


$$\begin{array}{c|c|c|} & 1-\epsilon & \epsilon \\ & I & M \\ \hline \bullet\ I & 3,3 & 0,4 \\ \hline \bullet\ M & 4,0 & 1,1 \\ \hline \end{array}$$

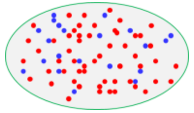
$$u(I, Population) = u(I, I)(1-\epsilon) + u(I, M)\epsilon$$

$$u(M, Population) = u(M, I)(1-\epsilon) + u(M, M)\epsilon$$

- Strategies with higher payoffs expand and those with lower payoffs contract
  - If $u(I, Population) > u(M, Population)$: Mutants will distinguish



  - If $u(M, Population) > u(I, Population)$: Mutants will expand



Motivating example

Populations

	$1-\epsilon$	ϵ
	C	D
C	$3, 3$	$0, 4$
D	$4, 0$	$1, 1$

<Prisoner’s dilemma as an evolutionarily game>

- Case 1: Incumbent (playing C) vc. the population:
  - C vc. $[(1-\epsilon)C+\epsilon D]$
  - The expected payoff : $(1-\epsilon)3+\epsilon 0 = 3(1-\epsilon)$

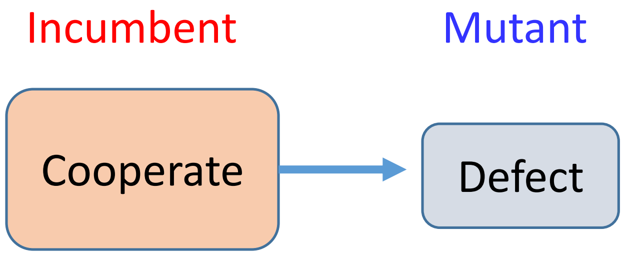
- Case 2: Mutant (playing D) vc. the population:
  - D vc. $[(1-\epsilon)C+\epsilon D]$
  - The expected payoff : $(1-\epsilon)4+\epsilon 1 = 4(1-\epsilon)+\epsilon$

- We need to compare the expected payoffs of incumbents and mutants when involved in random matchings with other individuals
  - Since $4(1-\epsilon)+\epsilon > 3(1-\epsilon)$, mutant performs better than the incumbent
    - ➢ Cooperation (Incumbent) is not a an evolutionarily stable strategy

Evolutionarily stable strategy – Intuition

A strictly dominated strategy is not an evolutionarily stable

$$\begin{array}{c|c|c|}
 & \multicolumn{2}{c}{\text{Populations}} \\
 & 1-\epsilon & \epsilon \\
 & C & D \\
\hline
C & 3,3 & 0,4 \\
 & \wedge & \wedge \\
D & 4,0 & 1,1 \\
\hline
\end{array}$$



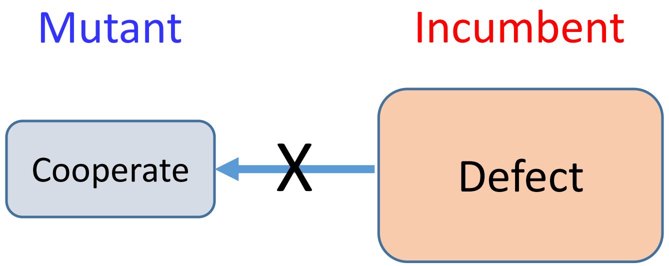
- **Cooperation,** which is as strictly dominated strategy, is not evolutionarily stable

Evolutionarily stable strategy – Intuition

A strictly dominated strategy is not an evolutionarily stable

Populations

	ϵ	$1-\epsilon$
	C	D
C	$3, 3$	$0, 4$
D	$4, 0$	$1, 1$



Mutant

Incumbent

Cooperate

X

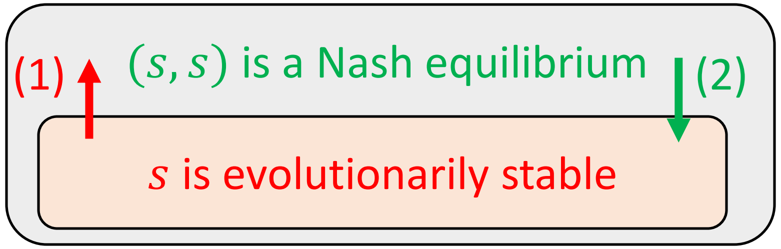
Defect

- Case 1: Mutant (playing C) vc. the population:
  - C vc. $[\epsilon C + (1-\epsilon)D]$
  - The expected payoff : $(1-\epsilon)0 + \epsilon 3 = 3\epsilon$
- Case 2: Incumbent (playing D) vc. the population:
  - D vc. $[\epsilon C + (1-\epsilon)D]$
  - The expected payoff : $\epsilon 4 + (1-\epsilon)1 = (1-\epsilon) + 4\epsilon$

The incumbent is more successful than the mutant on random matchings.

- Implies that mutations from D tend to extinguish

Evolutionarily stable strategy – Intuition



(1)

(s, s) is a Nash equilibrium

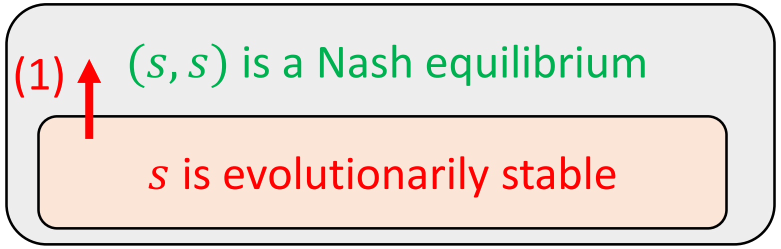
(2)

s is evolutionarily stable

Evolutionarily stable strategy – Intuition

From evolutionarily stable strategy to symmetric Nash equilibrium

$$\begin{array}{c|c|c|}
 & \epsilon & 1-\epsilon \\
 & b & c \\
\hline
b & 0,0 & 1,1 \\
\hline
c & 1,1 & 0,0 \\
\hline
\end{array}$$



(1) (s, s) is a Nash equilibrium

s is evolutionarily stable

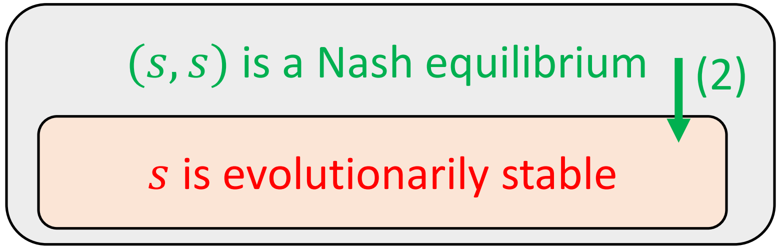
- Case 1: Mutant (playing b) vc. the population:
  - b vc. $[\epsilon b + (1-\epsilon)c]$
  - The expected payoff : $\epsilon 0 + (1-\epsilon)1 = (1-\epsilon)$
- Case 2: Incumbent (playing c) vc. the population:
  - c vc.$[\epsilon b + (1-\epsilon)c]$
  - The expected payoff : $\epsilon 1 + (1-\epsilon)0 = \epsilon$
- Mutant (playing b) performs better than the incumbent (playing c)
  - ➢ c is not a symmetric NE, because turning to b is strictly profitable
  - ➢ **Strategy c is not an evolutionarily stable strategy**

- If a strategy s is not a symmetric NE, then it is not a evolutionarily stable
- If s is evolutionarily stable $\Rightarrow$ (s, s) is Nash equilibrium

Evolutionarily stable strategy – Intuition

A Nash equilibrium strategy is not necessarily an evolutionarily stable strategy

$$\begin{array}{c|c|c|}
 & \epsilon & 1-\epsilon \\
 & a & b \\
\hline
a & \boxed{1,1} & 0,0 \\
\hline
b & 0,0 & \boxed{0,0} \\
\hline
\end{array}$$



- Case 1: Mutant (playing a) vc. the population:
    - a vc. $[\epsilon a + (1-\epsilon)b]$
    - The expected payoff : $\epsilon 1 + (1-\epsilon)1 = \epsilon$
- Case 2: Incumbent (playing b) vc. the population:
    - b vc.$[\epsilon a + (1-\epsilon)b]$
    - The expected payoff : $\epsilon 0 + (1-\epsilon)0 = 0$
- Mutant (playing a) performs better than the incumbent (playing b)
    - ➢ **Strategy b is not an evolutionarily stable strategy**
    - ➢ This is true despite the symmetric profile (b, b) being a Nash equilibrium

Formal definition (Biology)

- Evolutionarily stable strategies have also been studied by **evolutionarily Biologist**

Definition (evolutionarily stable strategy (ESS))

Given a symmetric two-player normal-form game $G = (N, S, u)$, a strategy $s^* \in S$ is an evolutionarily stable strategy if there exists $\bar{\epsilon} > 0$ such that for any $s \neq s^*$and for any $\epsilon < \bar{\epsilon}$, we have

$$\underbrace{u(s^*, \overset{\text{population}}{\epsilon s + (1-\epsilon)s^*})}_{\substack{\text{Payoff to incumbent } s^* \\ \text{(다수)}}} > \underbrace{u(s, \overset{\text{population}}{\epsilon s + (1-\epsilon)s^*})}_{\substack{\text{Payoff to mutant } s \\ \text{(소수)}}}$$

- Due to the property of expectation, the above equation can be written as

$$\epsilon u(s^*, s) + (1-\epsilon)u(s^*, s^*) > \epsilon u(s, s) + (1-\epsilon)u(s, s^*)$$

- Two interpretations:
    - We can state that the incumbents perform better than the mutants on random matchings
    - The strategy s^* cannot be invaded by s

Formal definition (Economics)

- Evolutionarily stable strategies have also been studied by **Economists**

Definition (evolutionarily stable strategy (ESS))

A mixed strategy $s^* \in S$ is evolutionarily stable if for any $s \in S$ the following two conditions hold:

(a) $u(s^*, s^*) \geq u(s, s^*)$

(b) if $u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

s^*: Incumbent strategy

s: Mutant strategy

- Condition (a) states essentially that the symmetric profile (s^*, s^*) is a Nash equilibrium
  - $u(s^*, s^*) > u(s, s^*)$: a strict Nash equilibrium
  - $u(s^*, s^*) \geq u(s, s^*)$: not-a-strict Nash equilibrium

- Condition (b) says that if the symmetric profile(s^*, s^*) is not a strict Nash equilibrium, then the mutant must perform poorly when playing against another mutant.

Equivalence of the Two definitions

Theorem

The two definitions of evolutionarily stable strategies are equivalent.

Proof: (First definition → second definition)

$$u(s^*, \epsilon s + (1-\epsilon)s^*) > u(s, \epsilon s + (1-\epsilon)s^*)$$

First definition

$$\Downarrow$$

(a) $u(s^*, s^*) \geq u(s, s^*)$

(b) if $u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

second definition

- Since the first definition holds for any $\epsilon < \bar{\epsilon}$, as $\epsilon \to 0$,

$$u(s^*, \epsilon s + (1-\epsilon)s^*) > u(s, \epsilon s + (1-\epsilon)s^*) \text{ implies}$$

$$u(s^*, s^*) \geq u(s, s^*)$$

Thus establishing part 1 of the second definition

Equivalence of the Two definitions

Theorem

The two definitions of evolutionarily stable strategies are equivalent.

Proof: (First definition → second definition)

- To establish part 2, suppose that $u(s^*, s^*) = u(s, s^*)$. Recall that u is linear in its arguments (since it is expected utility), so definition 1, $u(s^*, \epsilon s + (1-\epsilon)s^*) > u(s, \epsilon s + (1-\epsilon)s^*)$, can be written as

$$\epsilon u(s^*, s) + (1-\epsilon)u(s^*, s^*) > \epsilon u(s, s) + (1-\epsilon)u(s, s^*)$$

- Since $u(s^*, s^*) = u(s, s^*)$, this is equivalent to

$$\epsilon u(s^*, s) + \cancel{(1-\epsilon)u(s^*, s^*)} > \epsilon u(s, s) + \cancel{(1-\epsilon)u(s, s^*)}$$

$$\Rightarrow \epsilon u(s^*, s) > \epsilon u(s, s)$$

Since $\epsilon > 0$, part 2 of the second definition, $u(s^*, s) > u(s, s)$, follows

Equivalence of the Two definitions

Theorem

The two definitions of evolutionarily stable strategies are equivalent.

Proof: (Second definition → First definition)

$$u(s^*, \epsilon s + (1-\epsilon)s^*) > u(s, \epsilon s + (1-\epsilon)s^*)$$

First definition

$$\Uparrow$$

(a) $u(s^*, s^*) \geq u(s, s^*)$

(b) $if\ u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

second definition

Equivalence of the Two definitions

Theorem

The two definitions of evolutionarily stable strategies are equivalent.

Proof: (Second definition → First definition)

$$u(s^*, \epsilon s + (1-\epsilon)s^*) > u(s, \epsilon s + (1-\epsilon)s^*)$$

First definition

$$\uparrow$$

(a) $u(s^*, s^*) \geq u(s, s^*)$

(b) $if\ u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

second definition

- We have two possibilities in $u(s^*, s^*) \geq u(s, s^*)$
  - $u(s^*, s^*) > u(s, s^*)$
  - $u(s^*, s^*) = u(s, s^*)$

Equivalence of the Two definitions

Theorem

The two definitions of evolutionarily stable strategies are equivalent.

Proof: (Second definition → First definition)

$$u(s^*, \epsilon s + (1-\epsilon)s^*) > u(s, \epsilon s + (1-\epsilon)s^*)$$

First definition

$$\Uparrow$$

(a) $u(s^*, s^*) \geq u(s, s^*)$

(b) $if\ u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

second definition

- If $u(s^*, s^*) > u(s, s^*)$, then the condition in the first definition

$$u(s^*, \epsilon s + (1-\epsilon)s^*) > u(s, \epsilon s + (1-\epsilon)s^*)$$

is satisfied for $\epsilon = 0$, and hence for sufficiently small ϵ as well

Equivalence of the Two definitions

Theorem

The two definitions of evolutionarily stable strategies are equivalent.

Proof: (Second definition → First definition)

$$u(s^*, \epsilon s + (1-\epsilon)s^*) > u(s, \epsilon s + (1-\epsilon)s^*)$$

First definition

$$\Uparrow$$

(a) $u(s^*, s^*) \geq u(s, s^*)$

(b) $if\ u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

second definition

- If $u(s^*, s^*) = u(s, s^*)$, then the second definition implies

$$u(s^*, s) > u(s, s)$$

$$\Rightarrow \epsilon u(s^*, s) > \epsilon u(s, s)$$

Since $\epsilon > 0$

$$\Rightarrow \epsilon u(s^*, s) + (1-\epsilon)u(s^*, s^*) > \epsilon u(s, s) + (1-\epsilon)u(s, s^*)$$

Since $(1-\epsilon)u(s^*, s^*) = (1-\epsilon)u(s, s^*)$

then the condition in first definition holds.

evolutionarily Stability and Nash Equilibrium

Definition (Evolutionarily stable strategy (ESS))

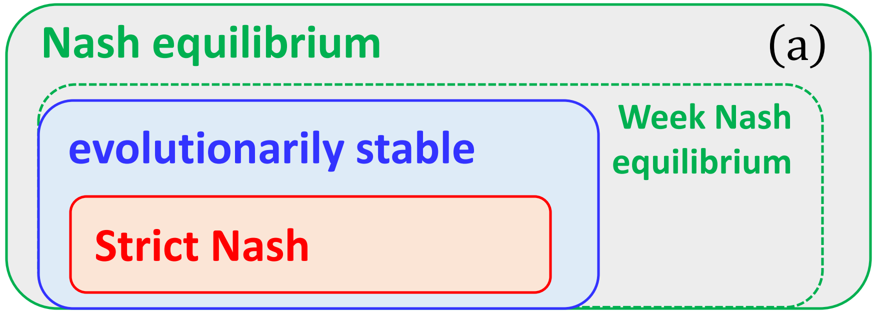
A mixed strategy $s^* \in S$ is evolutionarily stable if for any $s \in S$ the following two conditions hold:

(a) $u(s^*, s^*) \geq u(s, s^*)$
(b) if $u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

Theorem

- A strict (symmetric) Nash equilibrium of a symmetric game is an evolutionarily stable strategy.
- An evolutionarily stable strategy is a Nash equilibrium

- Their converses are not true, however.



A non-strict Nash equilibrium can be an evolutionarily stable strategy

	a	b
a	$1, 1$	$1, 1$
b	$1, 1$	$0, 0$

- We show that a non strict Nash equilibrium can be evolutionarily stable

  s^* is *evolutionariliy stable* if $u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

- The symmetric profile (a, a) is a symmetric Nash equilibrium, but not a strict one
  - $u_1(a, a) = u_1(b, a) = 1$

- Strategy a is evolutionarily stable or not?

  ➢ Stable because

$$u(s^*, s^*) = u(s, s^*), \text{ then } u(s^*, s) > u(s, s)$$

$$\underset{1}{u_1(a, a)} = \underset{1}{u_1(b, a)}, \text{ then } \underset{1}{u_1(a, b)} > \underset{0}{u_1(b, b)}$$

Evolution of social convention

	L	R
L	$2, 2$	$0, 0$
R	$0, 0$	$1, 1$

- This example deals with a game that presents **multiple** evolutionarily stable strategies

- The profiles (L, L) and (R, R) are two strict Nash equilibrium solutions
  - Both L and R are involuntarily stable strategies
  - These strategies need not be equally good

Battle of the sexes

	L	R
L	$0, 0$	$2, 1$
R	$1, 2$	$0, 0$

- There exist non symmetric pure Nash equilibrium solutions
  - means that we have no monomorphic (단일형) population
- There may exist evolutionarily stable mixed strategies
  - Strategies correspond to genes
  - Mixed strategies correspond to a polymorphic (다형성) population
  - In the parlance of evolutionarily biology, this is a population with multiple genes
- The solution $(s^*, s^*) = \left[\left(\frac{2}{3}, \frac{1}{3}\right), \left(\frac{2}{3}, \frac{1}{3}\right)\right]$ is a symmetric mixed-strategy Nash equilibrium, which yields a polymorphic population
- The mixed strategy s^* is ESS

  Mixed NE is always week Nash
  - ➢ s^* is *evolutionariliy stable* if $u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$
  - ➢ $u(s^*, L) = \frac{1}{3} > u(L, L) = 0, \qquad u(s^*, R) = \frac{4}{3} > u(R, R) = 0,$

    L and R are supports for any s

Monomorphic and Polymorphic Evolutionarily Stability

$vs.$

	$Hawk$	$Dove$
$Hawk$	$\left(\frac{V-C}{2},\frac{V-C}{2}\right)$	$(V,0)$
$Dove$	$(0,V)$	$\left(\frac{V}{2},\frac{V}{2}\right)$

- We develop a game characterized by monomorphic evolutionarily stable strategies
- V is the prize of victory and C is the cost of fight
- If $V > C$, then the game can be assimilated to the Prisoner's dilemma
  - Exist a unique strict Nash equilibrium (Hawk, Hawk)
  - Consequently, Hawk is an evolutionarily stable strategy
- evolutionarily interpretation:
  - ➢ all individuals will end up selecting an aggressive behavior, namely strategy Hawk

Polymorphic and Polymorphic Evolutionarily Stability

$vs.$

	$Hawk$	$Dove$
$Hawk$	$\left(\frac{V-C}{2},\frac{V-C}{2}\right)$	$(V,0)$
$Dove$	$(0,V)$	$\left(\frac{V}{2},\frac{V}{2}\right)$

- What will happen if $V < C$?
- Then the profiles $(Hawk, Dove)$ and $(Dove, Hawk)$ are two non symmetric NEs
- We also have a mixed Nash equilibrium, which is given by $(s^*, s^*) = \left[\left(\frac{V}{C}, 1-\frac{V}{C}\right), \left(\frac{V}{C}, 1-\frac{V}{C}\right)\right]$
- Is this mixed strategy $s^* = \left(\frac{V}{C}, 1-\frac{V}{C}\right)$ Evolutionarily stable?

Polymorphic and Polymorphic Evolutionarily Stability

	$Hawk$	$Dove$
$Hawk$	$\left(\frac{V-C}{2}, \frac{V-C}{2}\right)$	$(V, 0)$
$Dove$	$(0, V)$	$\left(\frac{V}{2}, \frac{V}{2}\right)$

- What will happen if $V < C$?
- Then the profiles $(Hawk, Dove)$ and $(Dove, Hawk)$ are two non symmetric NEs
- We also have a mixed Nash equilibrium, which is given by $(s^*, s^*) = \left[\left(\frac{V}{C}, 1-\frac{V}{C}\right), \left(\frac{V}{C}, 1-\frac{V}{C}\right)\right]$
- Is this mixed strategy $s^* = \left(\frac{V}{C}, 1-\frac{V}{C}\right)$ Evolutionarily stable?
- The mixed strategy s^* is ESS
  - s^* is *evolutionariliy stable* if $u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

$$u(s^*, s^*) = u(D, s^*), u(s^*, s^*) = u(D, s^*) \longrightarrow \begin{array}{l} u(s^*, D) > u(D, D) \\ u(s^*, H) > u(H, H) \end{array}$$

- As V increases, more players playing Hawk are in the evolutionarily stable strategy

Nash equilibrium does not imply ESS

- Consider the modified rock-paper-scissors game:

	Rook	Paper	Scissors
Rook	γ, γ	-1, 1	1, -1
Paper	1, -1	γ, γ	-1, 1
Scissors	-1, 1	1, -1	γ, γ

- Here $0 \leq \gamma < 1$ (If $\gamma = 0$, this is the standard rock-paper-scissors game)

- For all such γ, there is a unique mixed strategy equilibrium $s^* = (1/3, 1/3, 1/3)$, with expected payoff $u(s^*, s^*) = \frac{\gamma}{3}$.

- But for $\gamma > 0$, this is not ESS. For example, $s = R$ would invade, since

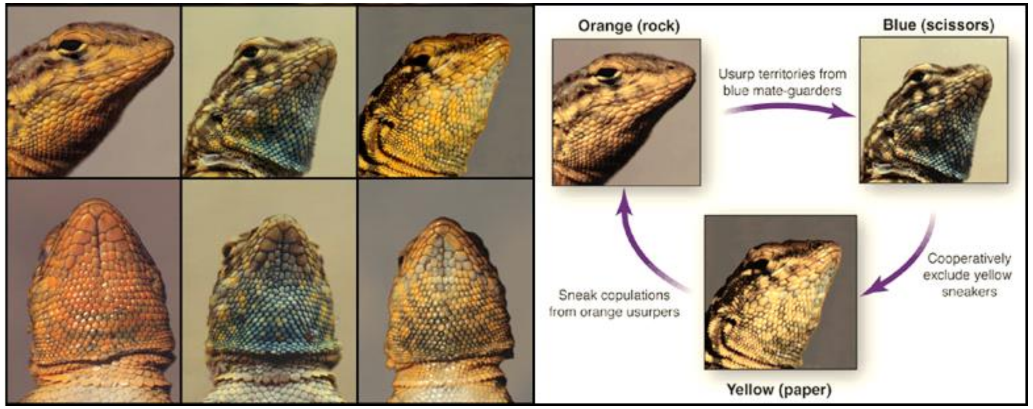
$$u(s^*, s) = \gamma \times \frac{1}{3} - 1 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{\gamma}{3} < u(s, s) = \gamma$$

Under ESS: if $u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

Hold when one player is plying a mixed Nash: $u(s^*, s^*) = u(R, s^*) = u(P, s^*) = \mathrm{u}(S, s^*)$

- This also shows that ESS doesn't necessarily exist

Do Animas play games?

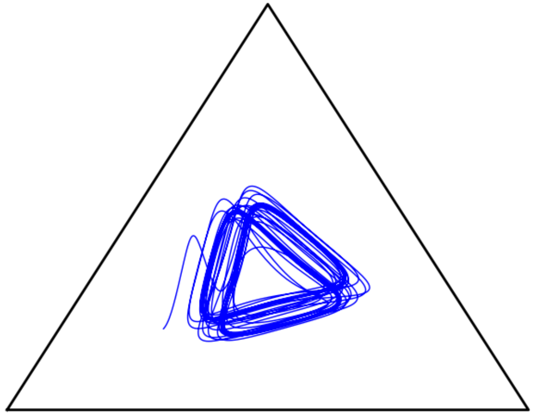

Orange (rock)
Blue (scissors)
Usurp territories from blue mate-guarders
Cooperatively exclude yellow sneakers
Sneak copulations from orange usurpers
Yellow (paper)

- Side-blotched lizards seem to play a version of the Hawk-Dove game. Three productive strategies for male lizards with distinct throat colors (that are genetically determined):
  - orange color: very aggressive and defend large territories;
  - blue color: less aggressive defense smaller territories;
  - yellow color: not aggressive, opportunistic mating behavior.
- Tails seem to be as follows:
  - when all are orange, yellow does well; when all are yellow, blue does well; and when all are blue, orange does well.
- This is similar to the modified rock-paper-scissors pattern, and in nature, it seems that there are fluctuations in composition of male colorings as we should expect on the basis that the game does not have any evolutionarily stable strategies.

The Replicator Dynamics

Dynamics

- The discussion of “dynamics” so far was largely heuristic.
- Are there actual dynamics of populations resulting from “game-theoretic” interactions that lead to evolutionarily stable strategies?
- Question at the intersection of game theory and population dynamics.
- The answer to this question is yes, and here we will discuss the simplest example, replicator dynamics.
- Throughout, we continue to focus on symmetric games.



Replicator Dynamics (Derivation)

- The replicator dynamics describes the evolution of the strategies under the assumption that the players are reactive to the environment
  - The players observe the population behavior and adopt their best-response strategies

- The replicator dynamics provides also an opportunity to look into asymptotic stability and to analyze the link with the evolutionarily stable strategies

Replicator Dynamics (Derivation)

- The evolutionary models encountered in the previous chapter do not consider any explicit dynamics
- Replicator dynamics model how the players change dynamically their strategies
- Let us enumerate the possible actions using an index $s = 1, 2, \dots, K$.
- Let us indicate with x_s the percentage of the population playing a strategy s
  - Obviously, it must hold that the sum of the percentages over the whole action space must sum up to one, i.e.,

$$\sum_{s=1}^{K} x_s = 1$$

  - Denote $x = (x_1, \dots, x_K)$ the population distribution
    - corresponds to a polymorphic strategy (evolutionary biology perspective)
    - corresponds to a mixed strategy (game theoretic perspective)

Replicator Dynamics (Derivation)

- Define the fitness at time t for playing a generic strategy s against a population playing x

$$u(s, x(t)) = \sum_{s'=1}^{K} u(s, s')x_{s'}(t)$$

$$= \mathbf{e}_s^T U x(t)$$

Pure action

e.x., $\mathbf{e}_3 = [0,0,1]^T$

- Define the average fitness (payoff of population) at time t resulting from a population distribution given by $x(t)$, which can be computed as

$$\bar{u}(x(t)) = \sum_{s=1}^{K} x_s(t)u(s, x(t))$$

- Using the definition of fitness $u(s, x(t))$, we can express $\bar{u}(x(t))$ as

$$\bar{u}(x(t)) = \sum_{s=1}^{K} x_s(t)u(s, x(t)) = \sum_{s=1}^{K} x_s(t) \sum_{s'=1}^{K} u(s, s')x_{s'}(t) = x(t)^T U x(t)$$

with $U_{ss'} = u(s, s')$

Replicator Dynamics (Derivation)

- Let $n_s(t)$ the number of individual playing the strategy s
- Let $N(t)$ the total number of individual
- $x_s(t) = n_s(t)/N(t)$
- $x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_s(t) \\ \vdots \\ x_K(t) \end{bmatrix} = \begin{bmatrix} n_1(t)/N(t) \\ \vdots \\ n_s(t)/N(t) \\ \vdots \\ n_K(t)/N(t) \end{bmatrix}$
- We assume that during the small time interval τ, only a τ fraction of the population takes part in pairwise competition, that is playing games as:

$$n_s(t+\tau) = (1-\tau)n_s(t) + \tau n_s(t)u(s, x(t)) \tag{1}$$

$$N(t+\tau) = (1-\tau)N(t) + \tau N(t)\bar{u}(x(t)) \tag{2}$$

Replicator Dynamics (Derivation)

$$n_s(t+\tau) = (1-\tau)n_s(t) + \tau n_s(t)u(s,x(t)) \quad (1)$$

$$N(t+\tau) = (1-\tau)N(t) + \tau N(t)\bar{u}(x(t)) \quad (2)$$

$$\text{LHS of } \frac{(1)}{(2)} = \frac{n_s(t+\tau)}{N(t+\tau)} = x_s(t+\tau)$$

$$\text{RHS of } \frac{(1)}{(2)} = \frac{(1-\tau)n_s(t) + \tau n_s(t)u(s,x(t))}{(1-\tau)N(t) + \tau N(t)\bar{u}(x(t))}$$

$$= \frac{(1-\tau)x_s(t)\cancel{N(t)} + \tau x_s(t)\cancel{N(t)}u(s,x(t))}{(1-\tau)\cancel{N(t)} + \tau \cancel{N(t)}\bar{u}(x(t))} \qquad \because x_s(t) = n_s(t)/N(t)$$

$$= \frac{x_s(t)\{(1-\tau) + \tau\bar{u}(x(t))\} - x_s(t)\tau\bar{u}\big(x(t)\big) + \tau x_s(t)u(s,x(t))}{(1-\tau) + \tau\bar{u}(x(t))}$$

$$= x_s(t) + \frac{-x_s(t)\tau\bar{u}\big(x(t)\big) + \tau x_s(t)u(s,x(t))}{(1-\tau) + \tau\bar{u}(x(t))}$$

$$= x_s(t) + \tau x_s(t)\frac{\big[u(s,x(t)) - \bar{u}\big(x(t)\big)\big]}{(1-\tau) + \tau\bar{u}(x(t))}$$

$$\text{LHS} = \text{RHS} \Rightarrow x_s(t+\tau) - x_s(t) = \tau x_s(t)\frac{[u(s,x(t)) - \bar{u}(x(t))]}{(1-\tau) + \tau\bar{u}(x(t))}$$

Replicator Dynamics (Derivation)

$$\Rightarrow x_s\ (t+\tau) - x_s(t) = \tau x_s(t)\,\frac{[u(s,x(t)) - \bar{u}(x(t))]}{(1-\tau)+\tau\bar{u}(x(t))}$$

$$\Rightarrow \frac{x_s(t+\tau) - x_s(t)}{\tau} = x_s(t)\,\frac{[u(s,x(t)) - \bar{u}(x(t))]}{(1-\tau)+\tau\bar{u}(x(t))}$$

$$\Rightarrow \lim_{\tau\to 0}\frac{x_s(t+\tau) - x_s(t)}{\tau} = \lim_{\tau\to 0} x_s(t)\,\frac{[u(s,x(t)) - \bar{u}(x(t))]}{(1-\tau)+\tau\bar{u}(x(t))}$$

$$\Rightarrow \dot{x}_s(t) = x_s(t)[u(s,x(t)) - \bar{u}(x(t))]$$

Replicator Dynamics

Definition (Continuous-time replicator dynamics)

The continuous-time version of the dynamics takes the form

$$\dot{x}_s(t) = x_s(t)[u(s, x(t)) - \bar{u}(x(t))]$$

which is called continuous replicator

Definition (Discrete-time Replicator dynamics)

For each $s = 1,2, \dots, K$ and for all t and τ, the replicator dynamic is given by

$$x_s(t + \tau) - x_s(t) = \tau x_s(t)[u(s, x(t)) - \bar{u}(x(t))]$$

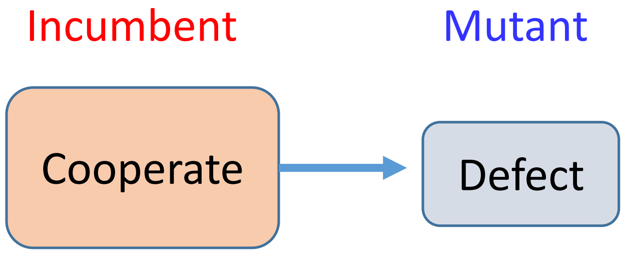
- The greater the fitness of a strategy relative to the average fitness, the greater its relative increase in the population.

Replicator dynamics example: Prisoner's Dilemma

Populations

	$1-\epsilon$	ϵ
	C	D
C	$3, 3$	$0, 4$
D	$4, 0$	$1, 1$

<Prisoner's dilemma as an evolutionarily game>

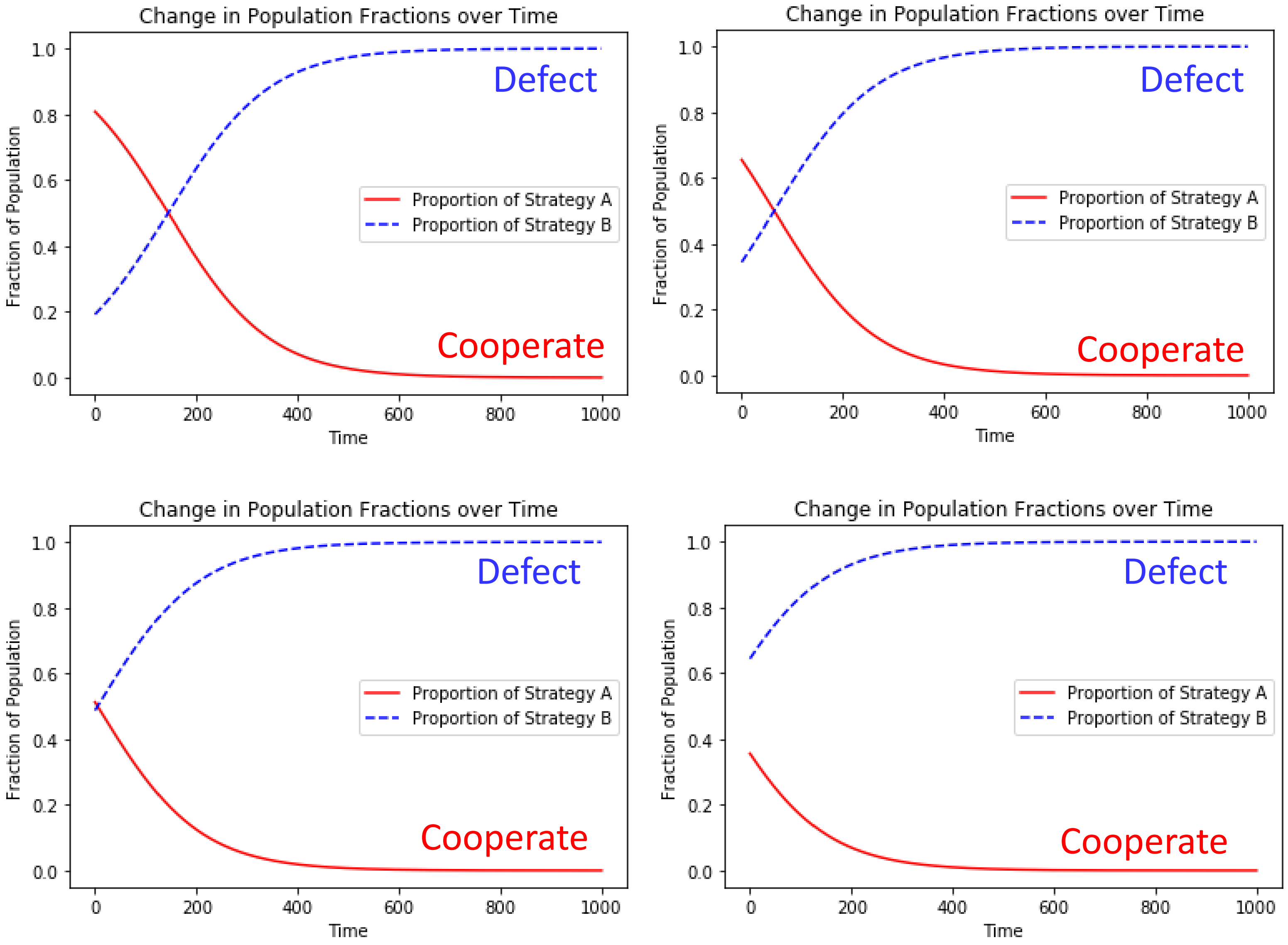


- Case 1: Incumbent (playing C) vc. the population:
  - C vc. $[(1-\epsilon)C+\epsilon D]$
  - The expected payoff : $(1-\epsilon)3+\epsilon 0 = 3(1-\epsilon)$

- Case 2: Mutant (playing D) vc. the population:
  - D vc. $[(1-\epsilon)C+\epsilon D]$
  - The expected payoff : $(1-\epsilon)4+\epsilon 1 = 4(1-\epsilon)+\epsilon$

- We need to compare the expected payoffs of incumbents and mutants when involved in random matchings with other individuals
  - Since $4(1-\epsilon)+\epsilon > 3(1-\epsilon)$, mutant performs better than the incumbent
    - ➢ Cooperation (Incumbent) is not a an evolutionarily stable strategy

Replicator dynamics example: Prisoner's Dilemma



- Regardless of the initial percentages, it always converges to Nash Equilibrium strategy

Replicator dynamics example: Rock Paper Scissor game

- Consider the modified rock-paper-scissors game:

	Rook	Paper	Scissors
Rook	γ, γ	-1, 1	1, -1
Paper	1, -1	γ, γ	-1, 1
Scissors	-1, 1	1, -1	γ, γ

- Here $0 \leq \gamma < 1$ (If $\gamma = 1$, this is the standard rock-paper-scissors game)
- For all such γ, there is a unique mixed strategy equilibrium $s^* = (1/3, 1/3, 1/3)$, with expected payoff $u(s^*, s^*) = \frac{\gamma}{3}$.
- But for $\gamma > 0$, this is not ESS. For example, $s = R$ would invade, since

$$u(s^*, s) = \gamma \times \frac{1}{3} - 1 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{\gamma}{3} < u(s, s) = \gamma$$

Under ESS: if $u(s^*, s^*) = u(s, s^*)$, then $u(s^*, s) > u(s, s)$

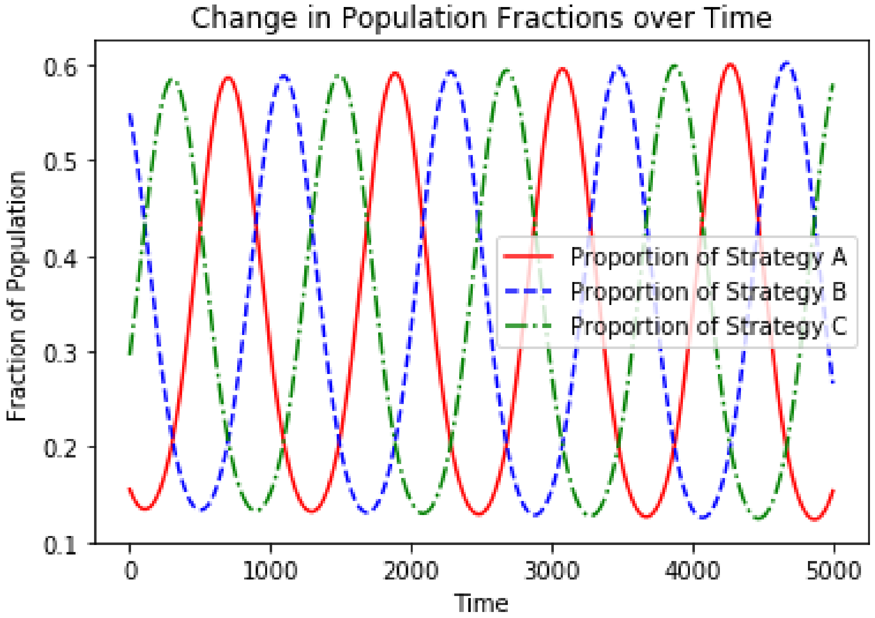
Hold when one player is plying a mixed Nash: $u(s^*, s^*) = u(R, s^*) = u(P, s^*) = \mathrm{u}(S, s^*)$

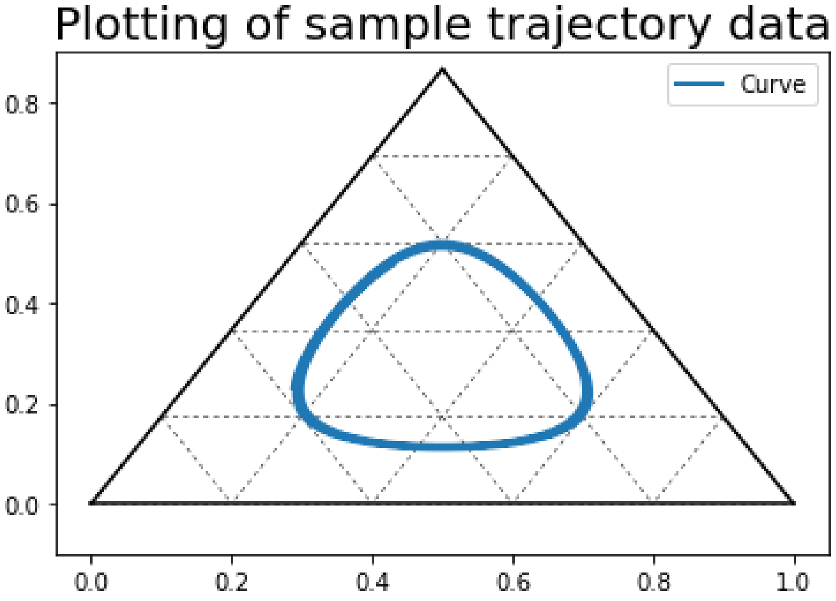
- This also shows that ESS doesn't necessarily exist

Replicator dynamics example: Rock Paper Scissor game

	Rook	Paper	Scissors
Rook	γ, γ	-1, 1	1, -1
Paper	1, -1	γ, γ	-1, 1
Scissors	-1, 1	1, -1	γ, γ

$\gamma = 0$

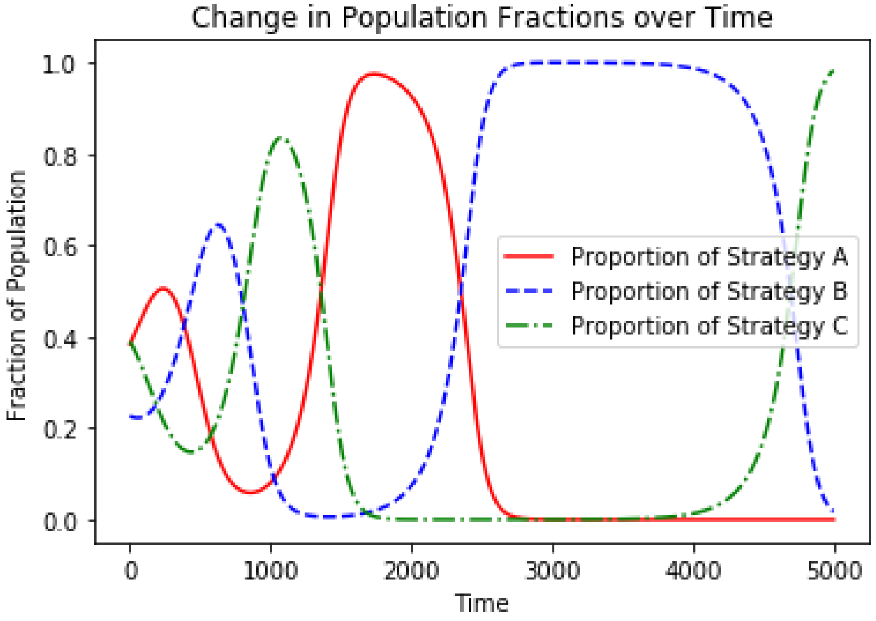


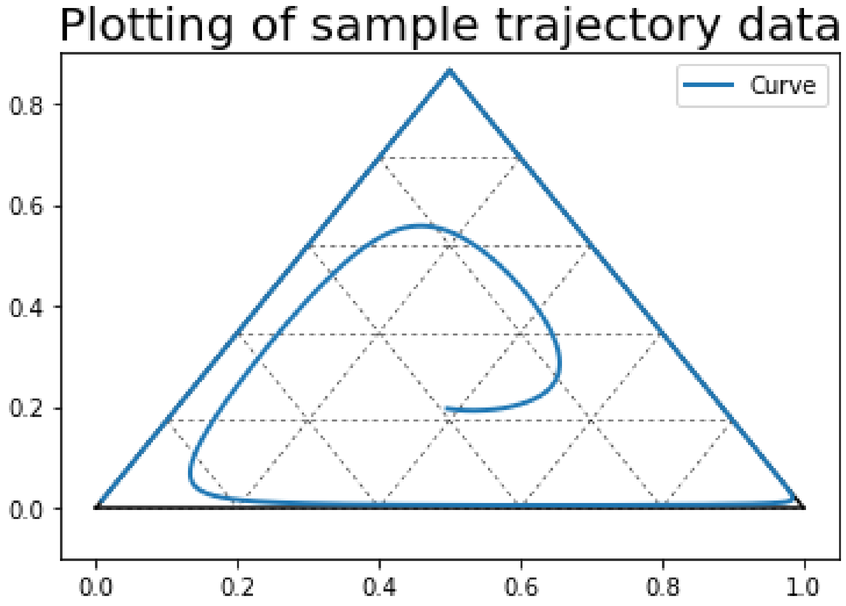


Replicator dynamics example: Rock Paper Scissor game

	Rook	Paper	Scissors
Rook	γ, γ	-1, 1	1, -1
Paper	1, -1	γ, γ	-1, 1
Scissors	-1, 1	1, -1	γ, γ

$\gamma = 0.1$

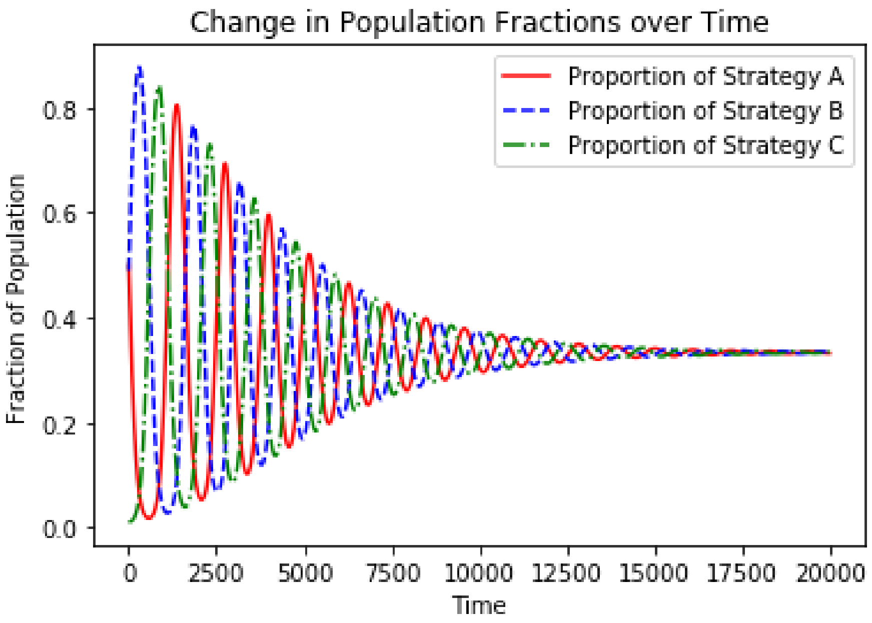


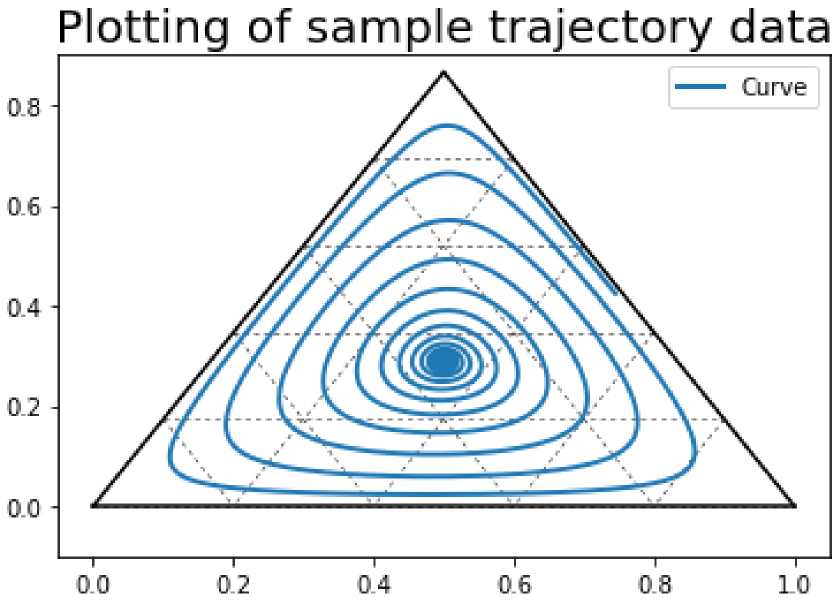


Replicator dynamics example: Rock Paper Scissor game

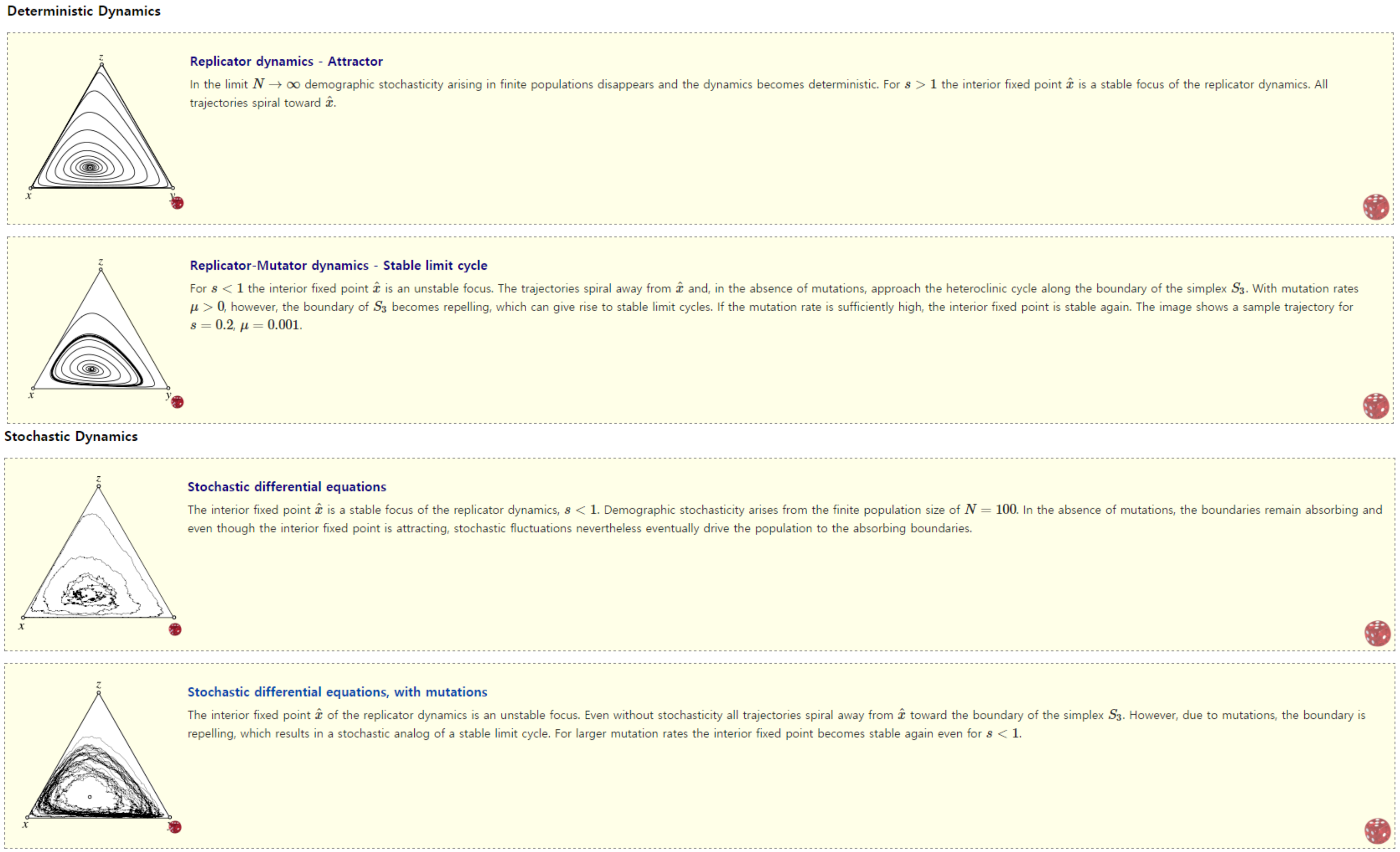
	Rook	Paper	Scissors
Rook	γ, γ	-1, 1	1, -1
Paper	1, -1	γ, γ	-1, 1
Scissors	-1, 1	1, -1	γ, γ

$\gamma = -0.1$





Further population dynamics



Deterministic Dynamics

Replicator dynamics - Attractor

In the limit $N \to \infty$ demographic stochasticity arising in finite populations disappears and the dynamics becomes deterministic. For $s > 1$ the interior fixed point $\hat{x}$ is a stable focus of the replicator dynamics. All trajectories spiral toward $\hat{x}$.

Replicator-Mutator dynamics - Stable limit cycle

For $s < 1$ the interior fixed point $\hat{x}$ is an unstable focus. The trajectories spiral away from $\hat{x}$ and, in the absence of mutations, approach the heteroclinic cycle along the boundary of the simplex S_3. With mutation rates $\mu > 0$, however, the boundary of S_3 becomes repelling, which can give rise to stable limit cycles. If the mutation rate is sufficiently high, the interior fixed point is stable again. The image shows a sample trajectory for $s = 0.2$, $\mu = 0.001$.

Stochastic Dynamics

Stochastic differential equations

The interior fixed point $\hat{x}$ is a stable focus of the replicator dynamics, $s < 1$. Demographic stochasticity arises from the finite population size of $N = 100$. In the absence of mutations, the boundaries remain absorbing and even though the interior fixed point is attracting, stochastic fluctuations nevertheless eventually drive the population to the absorbing boundaries.

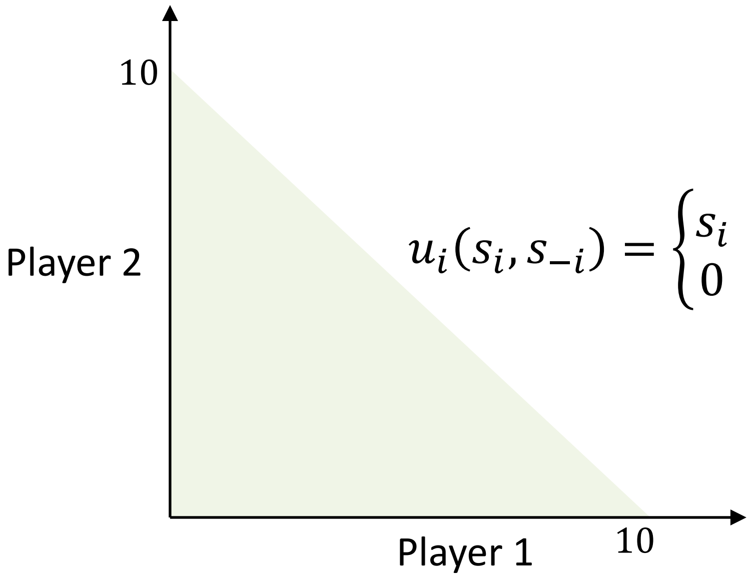
Stochastic differential equations, with mutations

The interior fixed point $\hat{x}$ of the replicator dynamics is an unstable focus. Even without stochasticity all trajectories spiral away from $\hat{x}$ toward the boundary of the simplex S_3. However, due to mutations, the boundary is repelling, which results in a stochastic analog of a stable limit cycle. For larger mutation rates the interior fixed point becomes stable again even for $s < 1$.

http://wiki.evoludo.org/index.php?title=Stochastic_dynamics_in_finite_populations

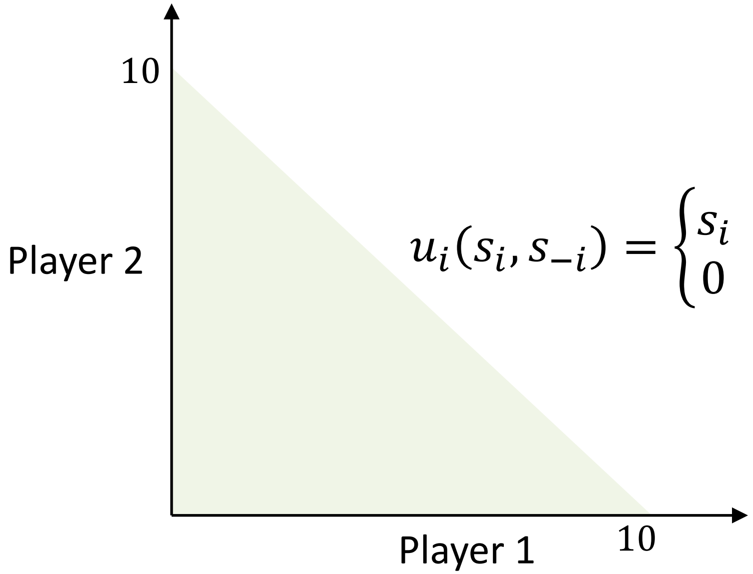
Replicator dynamics example: A sense of fairness

- One natural game to use for investigating the evolution of fairness is *divide-the-cake*
  - Suppose that two individuals are presented with a resource of size C by a third party.
  - A strategy for a player, in this game, consists of an amount of cake that he would like, thus $a_1 \in [0, C]$
  - If the sum of strategies for each player is less than or equal to C, each player receives the amount he asked for. However, if the sum of strategies exceeds C, players receive nothing
  - The feasible set of the game is as follow



$$u_i(s_i, s_{-i}) = \begin{cases} s_i & \text{if } s_i + s_{-i} \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

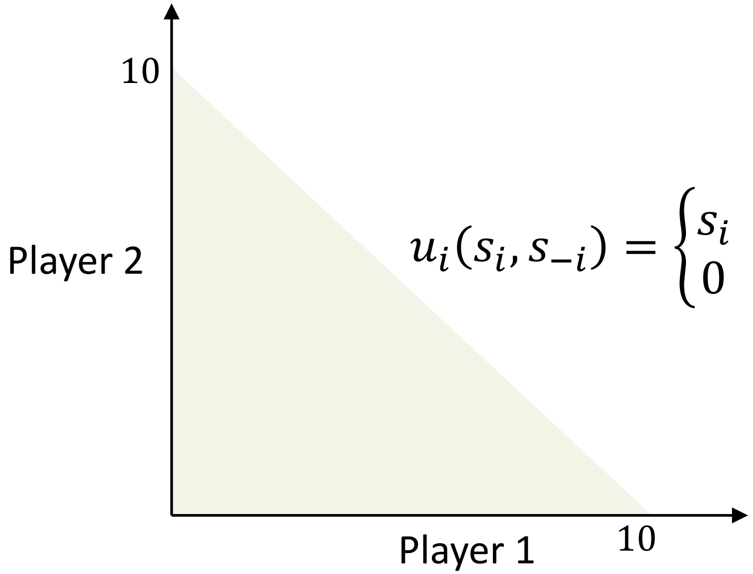
Replicator dynamics example: A sense of fairness



$$u_i(s_i, s_{-i}) = \begin{cases} s_i & \text{if } s_i + s_{-i} \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

- What is the Nash equilibrium?

Replicator dynamics example: A sense of fairness



$$u_i(s_i, s_{-i}) = \begin{cases} s_i & \text{if } s_i + s_{-i} \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

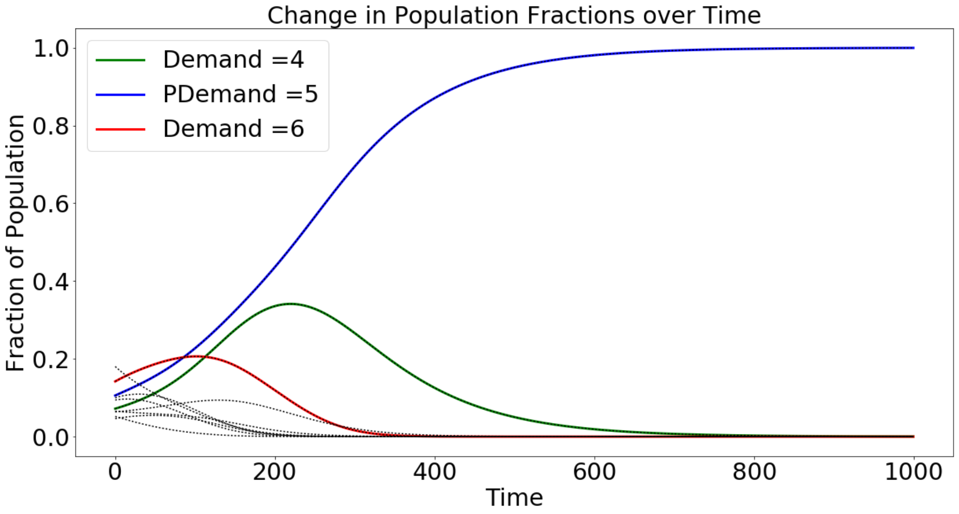
- What is the Nash equilibrium?
  - There are an infinite number of Nash equilibria for this game
    - Player 1 asks for $p \in [0, C]$ of the cake
    - Player 2 asks for $C - p$
  - Therefore, the equal split is only one of infinitely many Nash equilibria

Replicator dynamics example: A sense of fairness

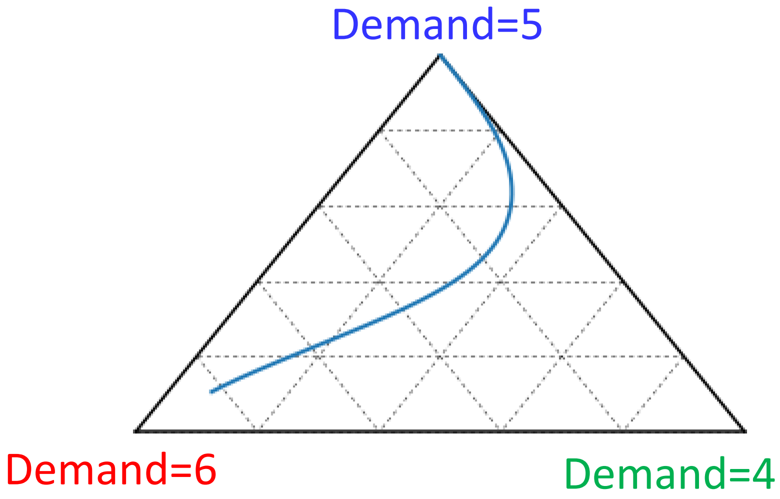
- Let’s model this game using Replicator dynamics
  - Let’s assume that the cake is divided into 10 equally sized slices
  - Each player’s strategy conforms to one of the following 11 possible types:
    - Demand 0, Demand 1, …, Demand 10
  - For the replicator dynamics, the state of the population is represented by a vector $[p_0, p_1, \dots, p_{10}]$, where each p_i denotes the frequency of the strategy “Demand i slices” in the population

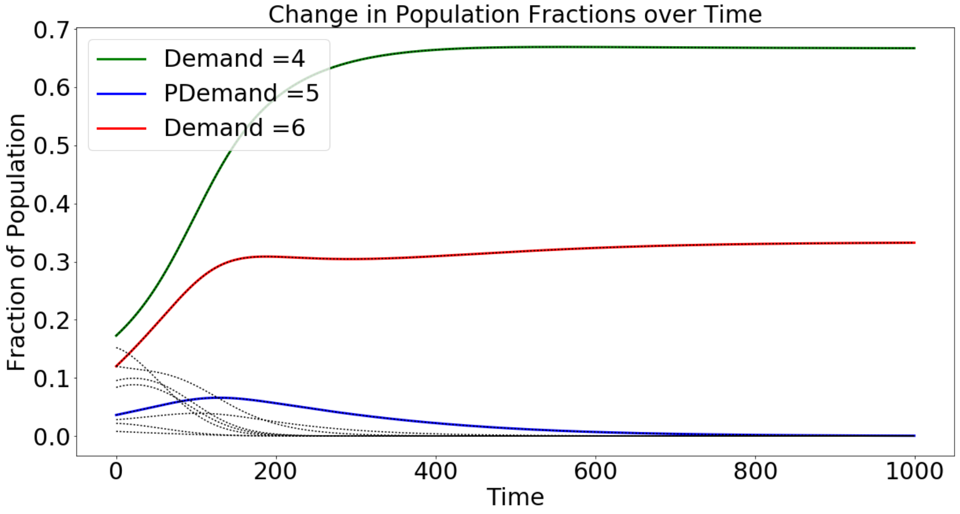
- The replicator dynamics allows us to model how the distribution of strategies in the population changes over time, beginning from a particular initial condition

Replicator dynamics example: A sense of fairness

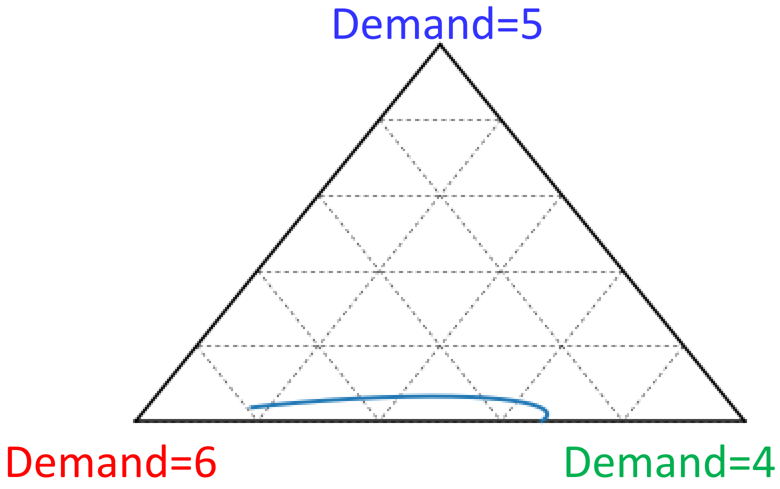


Evolution to equal division





Evolution to unequal division



Replicator dynamics example: A sense of fairness

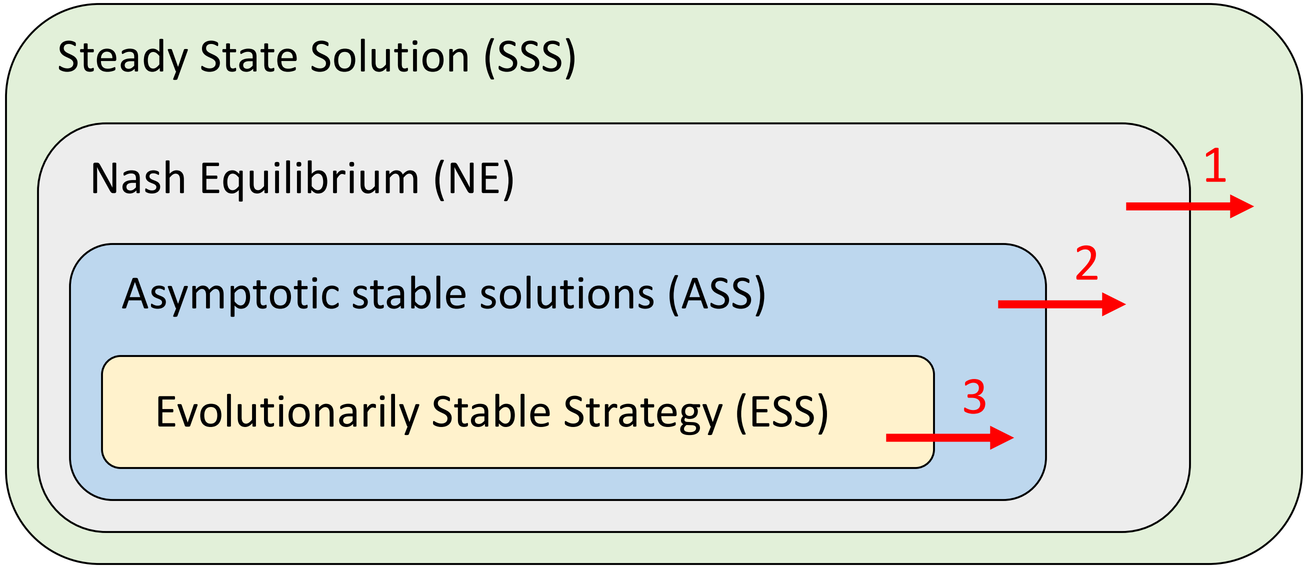
- In a population of bounded rational agents who modify their behaviors in a manner described by the replicator dynamics, fair division one, although not the only, evolutionary outcome.
- Evolves differently depending on the initial distribution
- The tendency of fair division to emerge can be measured by determining the size of the basin of attraction of the state where everyone in the population uses the strategy Demand 5 slices
- Skyrms (1996) measured this size using MC method, 62%

Stationarity, equilibria, and asymptotic stability

- Given a vector of distribution x^*, is such a vector a stationary state?
  - In systems theory, a stationary state is a state for which the first-order derivative is null, namely $\dot{x}^*(t) = 0$.

- Is a given vector of distribution x^* asymptotically stable?
  - There exists a neighborhood of x^* such that any trajectory starting from any x_0 in this neighborhood is such that the continuous replicator dynamics provides a trajectory that converges to x^*

Stationarity, equilibria, and asymptotic stability


Steady State Solution (SSS)
Nash Equilibrium (NE)
Asymptotic stable solutions (ASS)
Evolutionarily Stable Strategy (ESS)
1
2
3

Theorem

If x^* is a Nash equilibrium, then it is a steady state.

Theorem

If x^* is asymptotically stable, then it is a Nash equilibrium.

Theorem

If x^* is Evolutionarily stable, then it is a asymptotically stable.

Stationarity, equilibria, and asymptotic stability

Theorem

If x^* is a Nash equilibrium, then it is a stationary state.

Proof:

- Assume x^* is a Nash equilibrium. Consequently, x^* must also be a best-response to itself:

$$\text{(1)} \quad u(s, x^*) - \bar{u}(x^*) \leq 0 \text{ for all } s$$
$$\text{(2)} \quad u(s, x^*) - \bar{u}(x^*) = 0 \text{ for all } s \text{ in the support of } x^*$$

(1) is true : $u(s, x^*) \leq \sum_{s=1}^{K} x_s^* u(s, x^*) \; = \bar{u}(x^*)$ because $x^* = BR(x^*)$

(2) is true due to indifference principle under Nash mixed strategy

- ➢ when $x_s^*(t) \neq 0, u(s, x^*) - \bar{u}(x^*) = 0$

- The two conditions (1) and (2) guarantees for any s, either $u(s, x^*) - \bar{u}(x^*) = 0$ or $x_s^*(t) = 0$, which makes

$$\dot{x}_s(t) = x_s(t) \frac{[u(s, x(t)) - \bar{u}(x(t))]}{\bar{u}(x(t))} = 0$$

Stationarity, equilibria, and asymptotic stability

Theorem

If x^* is asymptotically stable, then it is a Nash equilibrium.

- The proof is immediate if x^* corresponds to a pure strategy(monomorphic population).
- In the case where x^* corresponds to a mixed strategy Nash equilibrium, the proof is also straightforward but long. The basic idea is that equation (Continuous replicator) implies that we are moving in the direction of "better replies"—relative to the average. If this process converges, then there must not exist any more (any other) strict better replies, and thus we must be at a Nash equilibrium.
- The converse is again not true.

Stationarity, equilibria, and asymptotic stability

Theorem

If x^* is asymptotically stable, then it is a Nash equilibrium.

- Example showing the converse is not true.

	A	B
A	$1, 1$	$0, 0$
B	$0, 0$	$\boxed{0, 0}$

- Here $(B, B)$is a Nash equilibrium, but clearly it is not asymptotically stable, since B is weakly dominated, and thus any perturbation away from (B, B)will start a process in which the fraction of agents playing A steadily increases.

- Take $x(t) = (\epsilon, 1 - \epsilon)$; then
  - $x_A = \epsilon, u\big(A, x(t)\big) - \bar{u}\big(x(t)\big) = \epsilon - \epsilon^2 > 0$
  - Playing A returns a payoff higher than the average payoff computed over the population, and therefore the percentage of people playing A increases

Stationarity, equilibria, and asymptotic stability

Theorem

If x^* is Evolutionarily stable, then it is a asymptotically stable.

- The proof is again somewhat delicate, but intuitively straightforward. The first definition of ESS states that for small enough perturbations, the evolutionarily stable strategy is a strict best response. This essentially implies that in the neighborhood of the ESS x^*, x^* will do better than any other strategy x, and thus according to (Continuous replicator), the fraction of those playing x^* should increase, thus implying asymptotic stability.

- This is the key result that justifies the focus on Evolutionary stable strategies

Examples

Stag-hunt game

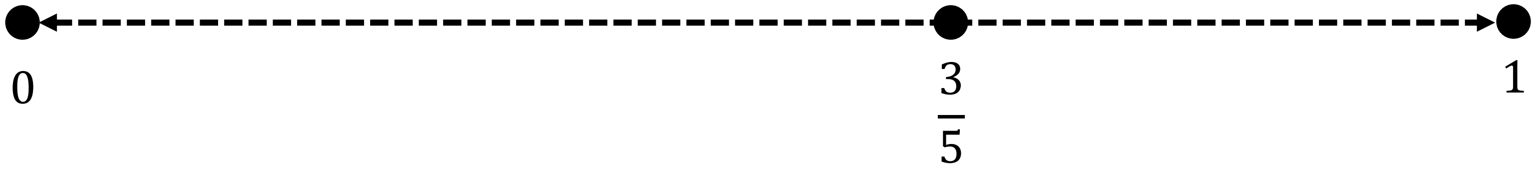
		S	H
$x_S(t)$	S	$5, 5$	$0, 3$
$1 - x_S(t)$	H	$3, 0$	$3, 3$

$$x(t) = \begin{bmatrix} x_s(t) \\ 1 - x_s(t) \end{bmatrix}$$

$$\dot{x}_s(t) = x_s(t)[u(x_s(t), x(t)) - \bar{u}(x(t))]$$

$$= x_s(t)\left([1, 0]\begin{bmatrix} 5 & 0 \\ 3 & 3 \end{bmatrix}\begin{bmatrix} x_s(t) \\ 1 - x_s(t) \end{bmatrix} - [x_s(t), 1 - x_s(t)]\begin{bmatrix} 5 & 0 \\ 3 & 3 \end{bmatrix}\begin{bmatrix} x_s(t) \\ 1 - x_s(t) \end{bmatrix}\right)$$

$$= x_s(t)\big(1 - x_s(t)\big)(5x_s(t) - 3)$$



Examples

Hawk-Dove game

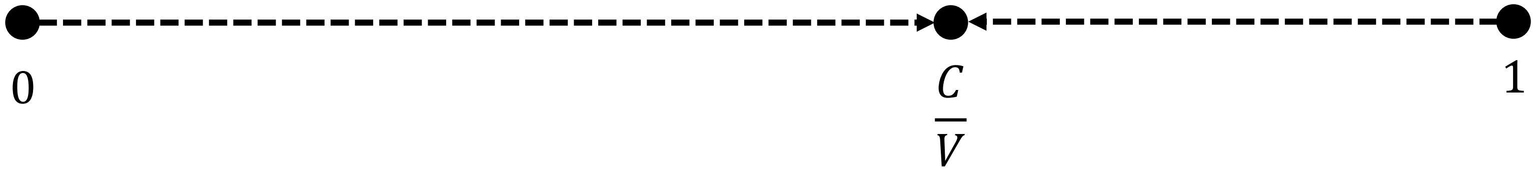
		H	D
$x_H(t)$	H	$\left(\frac{V-C}{2},\frac{V-C}{2}\right)$	$(V,0)$
$1-x_H(t)$	D	$(0,V)$	$\left(\frac{V}{2},\frac{V}{2}\right)$

$$x(t)=\begin{bmatrix} x_H(t) \\ 1-x_H(t) \end{bmatrix}$$

$$\dot{x}_H(t)=x_H(t)[u(x_H(t),x(t))-\bar{u}(x(t))]$$

$$=x_H(t)\left([1,0]\begin{bmatrix} \frac{(V-C)}{2} & V \\ 0 & \frac{V}{2} \end{bmatrix}\begin{bmatrix} x_H(t) \\ 1-x_H(t) \end{bmatrix}-[x_H(t),1-x_H(t)]\begin{bmatrix} \frac{(V-C)}{2} & V \\ 0 & \frac{V}{2} \end{bmatrix}\begin{bmatrix} x_H(t) \\ 1-x_H(t) \end{bmatrix}\right)$$

$$=-x_H(t)\big(1-x_H(t)\big)\left(x_H(t)-\frac{C}{V}\right)$$



Examples

Prisoner’s Dilemma game

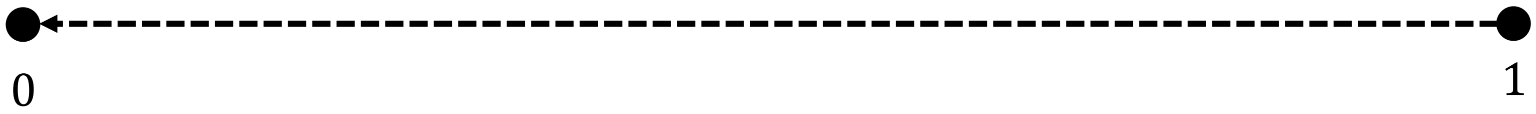
		C	D
$x_C(t)$	C	3, 3	0, 5
$1 - x_C(t)$	D	5, 0	1, 1

$$x(t) = \begin{bmatrix} x_C(t) \\ 1 - x_C(t) \end{bmatrix}$$

$$\dot{x}_C(t) = x_C(t)[u(x_C(t), x(t)) - \bar{u}(x(t))]$$

$$= x_C(t)\left([1, 0]\begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix}\begin{bmatrix} x_C(t) \\ 1 - x_C(t) \end{bmatrix} - [x_C(t), 1 - x_C(t)]\begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix}\begin{bmatrix} x_C(t) \\ 1 - x_C(t) \end{bmatrix}\right)$$

$$= -x_H(t)\big(1 - x_H(t)\big)(x_H(t) + 1)$$



Why Evolutionary Game Theory?

- A numerous researchers hope evolutionary game theory will provide tools for addressing a number of deficiencies in the traditional theory of games

(1) The equilibrium selection problem
  - If one restrict to select only pure strategies, some games lose solutions
  - When there are multiple NEs, various refinement concepts exist

(2) The problem of hype rational agents
  - Numerous experiment results show rationality assumptions does not describe the behavior of real human subjects.
  - Evolutionary game theory can better describe and predict the choices of human subjects

(3) The lack of a dynamical theory in the traditional theory of games
  - Most game theoretic equilibrium concepts are static
    - Even extensive form game, traditional game theory represents an individual's strategy as a specification of what choice that individual would make at each information set in the game (selection can be made prior the game)

Evolution vs. Learning

- Evolution is a good model for fully myopic behavior. But even when individuals follow rules of thumb, they are not fully myopic.

- Moreover, in evolution, the time scales are long. We need “mutations,” which are random and, almost by definition, rare.

- It would be better that one player observes his opponents' behavior, learns from these observations, and makes the best move in response to what he has learned

- In most (human) game-theoretic situations, even if individuals are not fully rational, they can imitate more successful strategies quickly, and learn the behavior of their opponents and best respond to those.

- This suggests a related but distinct approach to dynamic game-theoretic behavior, which is taken in the literature on *learning in games*.
  - Note that this is different from Bayesian game-theoretic learning, which focuses on learning the game structure but not strategies

Fictitious play

Introduction

- Most economic theory relies on equilibrium analysis based on Nash equilibrium or its refinements.
- The traditional explanation for when and why equilibrium arises is that it results from analysis and introspection by the players in a situation with all common knowledge on
  - the rules of the game
  - the rationality of the players
  - the payoff functions of players are all common knowledge.
- In this lecture, we develop an alternative explanation why equilibrium arises as the long-run outcome of a process

Fictitious play

- One of the earliest learning rules, introduced in Brown (1951), is the fictitious play.
- The most compelling interpretation of fictitious play is as a “belief-based” learning rule
  - ➢ players form beliefs about opponent play (from the entire history of past play) and behave rationally with respect to these beliefs.

Fictitious play-Setup

- We focus on a two player strategic form game $G = (\{1,2\}, S, u)$
- The players play this game at times $t = 1,2, \ldots$
- The stage payoff of player i is again given by $u_i(a_i, a_{-i})$ (for the pure strategy profile (a_i, a_{-i}))
- For $t = 1,2, \ldots$ and $i = 1, 2$, define the function $\eta_i^t \colon A_{-i} \to \mathbb{N}$
    - $\eta_i^t(a_{-i})$ is the number of times player i has observed the action a_{-i} before time t. Let $\eta_i^0(a_{-i})$ represent a starting point (or fictitious past)
- For example, consider a two player game, with $A_2 = \{U, D\}$.
    - $\eta_1^0(U) = 3$ and $\eta_1^0(D) = 5$
    - player 2 plays U, U, D in the first three periods
    - then, $\eta_1^3(U) = 3 + 2 = 5$ and $\eta_1^3(D) = 5 + 1 = 6$

The basic idea

- The basic idea of fictitious play is that each player assumes that his opponent is using a stationary mixed strategy, and updates his beliefs about this stationary mixed strategies at each step.

- Players choose actions in each period (or stage) to maximize that period's expected payoff given their prediction of the distribution of opponent's actions, which they form according to:

$$\mu_i^t(a_{-i}) = \frac{\eta_i^t(a_{-i})}{\sum_{a'_{-i} \in A_{-i}} \eta_i^t\left(a'_{-i}\right)}$$

- Player i forecasts player $-i$'s strategy at time t to be the empirical frequency distribution of the past play

Factious paly model of learning

- Given player $i's$ belief/forecast about his opponents play, he chooses his action at time t to maximize his payoff, i.e.,

$$a_i^t \in \underset{a_i \in A_i}{\operatorname{argmax}}\, u_i\left(a_i, \mu_i^t\right)$$

Remarks:

- Even though fictitious play is “belief based,” it is also myopic, because players are trying to maximize current payoff without considering their future payoffs.
- Perhaps more importantly, they are also not learning the “true model” generating the empirical frequencies (that is, how their opponent is actually playing the game).
- In this model, every player plays a pure best response to opponents’ empirical distributions.
- Not a unique rule due to multiple best responses. Traditional analysis assumes player chooses any of the pure best responses.

Example

- Consider the fictitious play of the following game:

	L	R
U	$4, 4$	$1, 1$
D	$5, 1$	$2, 2$

- $\eta_i^t(s_{-i})$ is the number of times player i has observed the action s_{-i} before time t
- $\mu_i^t(s_{-i})$ is player i's forecast on player $-i$'s strategy at time t

- Note that this game is dominant solvable (D is a strictly dominant strategy for the row player), and the unique $NE(D, R)$.
- Assume $\eta_1^0 = (3, 0)$ and $\eta_2^0 = (1, 2.5)$. Then fictitious play proceeds as follows:

$t = 0$

$$\eta_1^0 = \begin{bmatrix} \eta_1^0(a_2 = L) \\ \eta_1^0(a_2 = R) \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \qquad \mu_1^0 = \begin{bmatrix} \mu_1^0(a_2 = L) \\ \mu_1^0(a_2 = R) \end{bmatrix} = \begin{bmatrix} 3/3 \\ 0 \end{bmatrix}$$

$$\eta_2^0 = \begin{bmatrix} \eta_2^0(a_1 = U) \\ \eta_2^0(a_1 = D) \end{bmatrix} = \begin{bmatrix} 1 \\ 2.5 \end{bmatrix} \qquad \mu_2^0 = \begin{bmatrix} \mu_2^0(a_1 = U) \\ \mu_2^0(a_1 = D) \end{bmatrix} = \begin{bmatrix} 1/3.5 \\ 2.5/3.5 \end{bmatrix}$$

Example

- Consider the fictitious play of the following game:

	L	R
U	$4,4$	$1,1$
D	$5,1$	$2,2$

- $\eta_i^t(s_{-i})$ is the number of times player i has observed the action s_{-i} before time t
- $\mu_i^t(s_{-i})$ is player i's forecast on player $-i$'s strategy at time t

- Note that this game is dominant solvable (D is a strictly dominant strategy for the row player), and the unique $NE(D,R)$.
- Assume $\eta_1^0 = (3,0)$ and $\eta_2^0 = (1,2.5)$. Then fictitious play proceeds as follows:

$t = 0$

$$\eta_1^0 = \begin{bmatrix} \eta_1^0(a_2 = L) \\ \eta_1^0(a_2 = R) \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \qquad \mu_1^0 = \begin{bmatrix} \mu_1^0(a_2 = L) \\ \mu_1^0(a_2 = R) \end{bmatrix} = \begin{bmatrix} 3/3 \\ 0 \end{bmatrix}$$

- player 1 thinks player 2 will play L more often, thus $a_1^0 = D$

$$\eta_2^0 = \begin{bmatrix} \eta_2^0(a_1 = U) \\ \eta_2^0(a_1 = D) \end{bmatrix} = \begin{bmatrix} 1 \\ 2.5 \end{bmatrix} \qquad \mu_2^0 = \begin{bmatrix} \mu_2^0(a_1 = U) \\ \mu_2^0(a_1 = D) \end{bmatrix} = \begin{bmatrix} 1/3.5 \\ 2.5/3.5 \end{bmatrix}$$

- player 2 thinks player 1 will play D more often, thus $a_2^0 = R$

Example

- Consider the fictitious play of the following game:

	L	R
U	$4, 4$	$1, 1$
D	$5, 1$	$2, 2$

- $\eta_i^t(s_{-i})$ is the number of times player i has observed the action s_{-i} before time t
- $\mu_i^t(s_{-i})$ is player i's forecast on player $-i$'s strategy at time t

- Note that this game is dominant solvable (D is a strictly dominant strategy for the row player), and the unique $NE(D, R)$.
- Assume $\eta_1^0 = (3, 0)$ and $\eta_2^0 = (1, 2.5)$. Then fictitious play proceeds as follows:

$t = 0$

$$\eta_1^0 = \begin{bmatrix}\eta_1^0(a_2 = L)\\ \eta_1^0(a_2 = R)\end{bmatrix} = \begin{bmatrix}3\\ 0\end{bmatrix} \qquad \mu_1^0 = \begin{bmatrix}\mu_1^0(a_2 = L)\\ \mu_1^0(a_2 = R)\end{bmatrix} = \begin{bmatrix}3/3\\ 0\end{bmatrix}$$

- player 1 thinks player 2 will play L more often, thus $a_1^0 = D$

$$\eta_2^0 = \begin{bmatrix}\eta_2^0(a_1 = U)\\ \eta_2^0(a_1 = D)\end{bmatrix} = \begin{bmatrix}1\\ 2.5\end{bmatrix} \qquad \mu_2^0 = \begin{bmatrix}\mu_2^0(a_1 = U)\\ \mu_2^0(a_1 = D)\end{bmatrix} = \begin{bmatrix}1/3.5\\ 2.5/3.5\end{bmatrix}$$

- player 2 thinks player 1 will play D more often, thus $a_2^0 = R$

Example

- Consider the fictitious play of the following game:

	L	R
U	$4, 4$	$1, 1$
D	$5, 1$	$2, 2$

- $\eta_i^t(s_{-i})$ is the number of times player i has observed the action s_{-i} before time t
- $\mu_i^t(s_{-i})$ is player i's forecast on player $-i$'s strategy at time t

- Note that this game is dominant solvable (D is a strictly dominant strategy for the row player), and the unique $NE(D, R)$.
- Assume $\eta_1^0 = (3, 0)$ and $\eta_2^0 = (1, 2.5)$. Then fictitious play proceeds as follows:

$t = 1$

$$\eta_1^1 = \begin{bmatrix} \eta_1^1(a_2 = L) \\ \eta_1^1(a_2 = R) \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \qquad \mu_1^1 = \begin{bmatrix} \mu_1^1(a_2 = L) \\ \mu_1^1(a_2 = R) \end{bmatrix} = \begin{bmatrix} 3/4 \\ 1/4 \end{bmatrix}$$

- player 1 thinks player 2 will play L more often, thus $a_1^1 = D$

$$\eta_2^1 = \begin{bmatrix} \eta_2^1(a_1 = U) \\ \eta_2^1(a_1 = D) \end{bmatrix} = \begin{bmatrix} 1 \\ 3.5 \end{bmatrix} \qquad \mu_2^1 = \begin{bmatrix} \mu_2^1(a_1 = U) \\ \mu_2^1(a_1 = D) \end{bmatrix} = \begin{bmatrix} 1/4.5 \\ 3.5/4.5 \end{bmatrix}$$

- player 2 thinks player 1 will play D more often, thus $a_2^1 = R$

Example

- Since D is a dominant strategy for the row player, he always plays D, and μ_2^t converges to (0, 1) with probability 1.

- Therefore, player 2 will end up playing R.

- The remarkable feature of the fictitious play is that players don’t have to know anything about their opponent’s payoff. They only form beliefs about how their opponents will play.

Convergence of Fictitious play to pure strategies

- Let $\{a^t\}$ be a sequence of strategy profiles generated by fictitious play (FP).
- Let us now study the asymptotic behavior of the sequence $\{s^t\}$, i.e., the convergence properties of the sequence $\{a^t\}$ as $t \rightarrow \infty$
- We first define the notion of convergence to pure strategies.

Definition

The sequence $\{a^t\}$ converges to a if there exists T such that $a^t = a$ for all $t \geq T$

Theorem

Let $\{a^t\}$ be a sequence of strategy profiles generated by fictitious play.

1) If $\{a^t\}$ converges to $\bar{a}$, then $\bar{a}$ is a pure strategy *Nash equilibrium*
2) Suppose that for some t, $a^t = a^*$, where a^* is a *strict Nash equilibrium*. Then $a^\tau = a^*$ for all $\tau > t$.

Convergence of Fictitious play to pure strategies

- Part 1 is straightforward (Asymptotically stable strategy is Nash equilibrium)
- Consider part 2
- Let $a^t = a^*$. We will show that $a^{t+1} = a^*$.
- Note that for all $a_{-i} \in A_{-i}$

$$\mu_i^{t+1}(a_{-i}) = (1-\alpha)\mu_i^t(a_{-i}) + \alpha s_{-i}^t(a_{-i}), \text{ with } \quad s_{-i}^t(a_{-i}) = \begin{cases} 1 & \text{if } a_{-i} = a_{-i}^* \\ 0 & \text{otherwise} \end{cases}$$

  - $\mu_i^t(a_{-i})$ is player i's belief on player $-i$'s strategy at time t
    - player i believes player $-i$ will select action a_{-i} with a probability $\mu_i^t(a_{-i})$
  - $s_{-i}^t(a_{-i})$ is the probability that player $-i$ actually select action a_{-i}^*
  - $\alpha = \frac{1}{\left[\sum_{a_{-i}} \eta_i^t(a_{-i}) + 1\right]}$

- Regard μ_i^{t+1} and s_{-i} are strategies, i.e., probability distribution on the possible actions, i.e., $\mu_i^{t+1}, s_{-i} \in \Delta(A_{-i})$

$$\mu_i^{t+1} = (1-\alpha)\mu_i^t + \alpha s_{-i}^t$$

Convergence of Fictitious play to pure strategies

$$\mu_i^{t+1} = (1-\alpha)\mu_i^t + \alpha s_{-i}^t$$

- Therefore, by the linearity of the expected utility, we have for all $a_i \in A_i$,

$$u_i\left(a_i,\ \mu_i^{t+1}\right) = (1-\alpha)u_i\left(a_i,\ \mu_i^t\right) + \alpha u_i\left(a_i,\ s_{-i}^t\right)$$

- Since a_i^* maximizes both terms

$$a_i^* = a_i^t \in \operatorname*{argmax}_{a_i \in A_i} u_i\left(a_i, \mu_i^t\right) \qquad \because \text{ assumption } a^t = a^*$$

$$a_i^* = BR\left(s_{-i}^t\right) = \operatorname*{argmax}_{a_i} u_i\left(a_i,\ s_{-i}^t\right)$$

$$= \operatorname*{argmax}_{a_i} u_i\left(a_i,\ a_{-i}^*\right) \qquad s_{-i}^t(a_{-i}) = \begin{cases} 1 & \text{if } a_{-i} = a_{-i}^* \\ 0 & \text{otherwise} \end{cases}$$

- it follows a_i^* will be played at $t+1$

$$a_i^* = \operatorname*{argmax}_{a} u_i\left(a, \mu_i^{t+1}\right)$$

- Thus

$$a_i^{t+1} = a_i^t = a_i^*$$

Convergence of Fictitious play to mixed strategies

- The preceding notion of convergence only applies to pure strategies. We next provide an alternative notion of convergence, i.e., convergence of empirical distributions or beliefs.
    - Converged in pure strategy profiles

      $$(A,B) \to (B,A) \to \cdots \to (A,B) \to (A,B) \to (A,B) \to (A,B) \to (A,B) \to (A,B)$$

    - Converged in mixed strategy profiles in the time-average sense

      $$(A,B) \to (B,A) \to \cdots \to (A,B) \to (B,A) \to (A,B) \to (B,A) \to (A,B) \to (B,B)$$

      Player 1: $(A{:}\,1/2 \quad \mathrm{B}{:}\,1/2)$ Player 2: $(A{:}\,1/2 \quad \mathrm{B}{:}\,1/2)$

Definition

The sequence $\{a^t\}$ converges to $\sigma \in S$ in the time-average sense if for all i and for all $a_i \in A_i$, we have

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T-1} I\{a_i^t = a_i\}}{T} = \sigma(a_i)$$

- In other words, $\mu^T_{-i}(a_i)$ converges $\sigma_i(a_i)$ as $T \to \infty$

Convergence in Matching Pennies: An example

- Example illustrates convergence of the fictitious play sequence in the time-average sense.

	$Heads$	$Tails$
$Heads$	$1,-1$	$-1,1$
$Tails$	$-1,1$	$1,-1$

Time	η_1^t	η_2^t	Play
0	$(0,0)$	$(0,2)$	(H,H)
1	$(1,0)$	$(1,2)$	(H,H)
2	$(2,0)$	$(2,2)$	(H,T)
3	$(2,1)$	$(3,2)$	(H,T)
4	$(2,2)$	$(4,2)$	(T,T)
5	$(2,3)$	$(4,3)$	(T,T)
6	$\ldots$	$\ldots$	(T,H)

- In this example, play continues as a deterministic cycle.
- The time average converges to the unique Nash equilibrium,

$$\big((1/2,1/2),(1/2,1/2)\big)$$

More general convergence result

Theorem

Suppose a fictitious play sequence $\{a^t\}$ converges to σ in the time-average sense. Then σ is a Nash equilibrium.

Proof:

- Suppose a^t converges to σ in the time-average sense
- Suppose, to obtain a **contradiction**, that σ is not a Nash equilibrium
- Then there exist some i, a_i, $a_i' \in A_i$ with $\sigma_i(a_i) > 0$ such that

$$u_i(a_i', \sigma_{-i}) > u_i(a_i, \sigma_{-i})$$

Note that if σ is Nash equilibrium for all a_i, $a_i' \in A_i$ with $\sigma_i(a_i) > 0$ the following is satisfied

$$u_i(a_i', \sigma_{-i}) \leq u_i(a_i, \sigma_{-i}) = u_i(\sigma_i, \sigma_{-i})$$

because a_i is included to the support for $\sigma, \text{i. e.}, \sigma_i(a_i) > 0$

More general convergence result

- Choose $\epsilon > 0$ such that

$$\epsilon < \frac{1}{2}[u_i(a_i', \sigma_{-i}) - u_i(a_i, \sigma_{-i})] \qquad (1)$$

- Choose T sufficiently large that for all $t \geq T$, we have

$$\left|\mu_i^T(a_{-i}) - \sigma_{-i}(a_{-i})\right| < \frac{\epsilon}{\max\limits_{a \in A} u_i(a)} \text{ for all } a_{-i} \qquad (2)$$

which is possible $\mu_i^t \to \sigma_{-i}$ by assumption

- Then, for any $t \geq T$, we have

$$\begin{aligned} u_i(a_i, \mu_i^t) &= \sum\nolimits_{a_{-i}} u_i(a_i, a_{-i})\mu_i^t(a_{-i}) \\ &\leq \sum\nolimits_{a_{-i}} u_i(a_i, a_{-i})\sigma_{-i}(a_{-i}) + \epsilon && \because (2) \\ &< \sum\nolimits_{a_{-i}} u_i(a_i', a_{-i})\sigma_{-i}(a_{-i}) - \epsilon && \because (1) \\ &\leq \sum\nolimits_{a_{-i}} u_i(a_i', a_{-i})\mu_i^t(a_{-i}) = u_i\left(a_i', \mu_i^t\right) && \because (2) \end{aligned}$$

- This shows that after sufficiently large t, a_i is never played, implying that as $t \to \infty, \mu_{-i}^t(a_i) \to 0$.
- But this contradicts the fact that $\sigma_i(a_i) > 0$,completing the proof.

Example: The Anti-Coordination game

- The theorem gives sufficient conditions for the empirical distribution of the players' action to convergence to a mixed-strategy equilibrium
- However, it does not make any claims about the distribution of the particular outcomes (payoffs that each player can have)
- Consider the following *Anti-Coordination game*

	A	B
A	$0, 0$	$1, 1$
B	$1, 1$	$0, 0$

- What are the Nash equilibriums?

$$(A, A), (B, B), \left(A\colon\frac{1}{2}, B\colon\frac{1}{2}\right)$$

Example: The Anti-Coordination game

Round	1's action	2's action	1's belief	2's belief
0			$(1, 0.5)$	$(1, 0.5)$
1	B	B	$(1, 1.5)$	$(1, 1.5)$
2	A	A	$(2, 1.5)$	$(2, 1.5)$
3	B	B	$(2, 2.5)$	$(2, 2.5)$
4	A	A	$(3, 2.5)$	$(3, 2.5)$
5	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- The strategy of each player converges to the mixed strategy $(0.5, 0.5)$, which is the mixed strategy Nash equilibrium
- However, the payoff received by each player is 0, since the players never hit the outcomes with positive payoff
- Thus, although the empirical distribution of the strategies converges to the mixed strategy Nash equilibrium, **the players may not receive the expected payoff of the Nash equilibrium**.

Example: Shapley’s Almost-Rock-Paper-Scissors game

- The empirical distributions of players actions need not converge at all.
- Consider the following rock-paper-scissors game proposed by Shapley

	Rook	Paper	Scissors
Rook	$0, 0$	0, 1	1, 0
Paper	1, 0	$0, 0$	0, 1
Scissors	0, 1	1, 0	$0, 0$

- The unique Nash equilibrium of this game is for each player to play the mixed strategy is $(1/3, 1/3, 1/3)$

Example: Shapley's Almost-Rock-Paper-Scissors game

Round	1's action	2's action	1's belief	2's belief
0			$(0, 0, 0.5)$	$(0, 0.5, 0)$
1	R	S	$(0, 0, 1.5)$	$(1, 0.5, 0)$
2	R	P	$(0, 1, 1.5)$	$(2, 0.5, 0)$
3	R	P	$(0, 2, 1.5)$	$(3, 0.5, 0)$
4	S	P	$(0, 3, 1.5)$	$(3, 0.5, 1)$
5	S	P	$(0, 3, 2.5)$	$(3, 1.5, 1)$
⋮	⋮	⋮	⋮	⋮

- The empirical play of this game never converges to any fixed distribution

When empirical distribution converges?

Theorem

Each of the following is a sufficient condition for the empirical frequencies of play to converge in fictions play:

- The game is zero sum:
- The game is solvable by iterated elimination of strictly dominated strategies;
- The game is potential game
- The game is $2 \times n$ and has generic payoffs

Summary

- Fictitious play is very sensitive to the players' initial beliefs
- Fictitious play is somewhat paradoxical in that each agent assumes a stationary policy of the opponent, yet, no agent plays a stationary policy except when the process happens to converge to one
- It is simple to state and gives rise to nontrivial properties
- Because players only are thinking about their opponent's actions, they are not playing attention to whether they are actually been doing well.

Extension:

- How to define fictitious play if one has continuous action space?

Stochastic approximation approaches

Introduction

- We will focus on the algorithms that have been used for learning how to choose the optimal actions when agents are playing matrix games repeatedly

- We going to discuss the learning algorithms that looks similar to gradient ascent algorithm

$$w' \leftarrow w + \eta \frac{\partial L(w)}{\partial w}$$

- Centralized learning
  - Infinitesimal gradient ascent (IGA)
    - ➢ Win or Learn Fast (WoLF) + IGA = WoLF-IGA

- Decentralized learning
  - Policy Hill Climbing (PHC)
    - ➢ Win or Learn Fast (WoLF) + PHC = WoLF-PHC
  - Linear reward-inaction (L_{RI}) algorithm
  - Lagging anchor algorithm
  - L_{RI} −lagging anchor algorithm

Infinitesimal gradient ascent (IGA)

- Used in in relatively simple two-action/two-player general-sum games

	β action 1	$1-\beta$ action 2
α action 1	(r_{11}, c_{11})	(r_{12}, c_{12})
$1-\alpha$ action 2	(r_{21}, c_{21})	(r_{22}, c_{22})

- Payoff matrix for the row player

$$R_r = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$$

- Payoff matrix for the column player

$$R_c = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

- $P(a_r = 1) = \alpha, \quad P(a_r = 2) = 1 - \alpha$
- $P(a_c = 1) = \beta, \ \ P(a_c = 2) = 1 - \beta$

Infinitesimal gradient ascent (IGA)

- Define the expected payoff to each player
  - $V_r(\alpha, \beta) = \alpha\beta r_{11} + \alpha(1-\beta)r_{12} + (1-\alpha)\beta r_{21} + (1-\alpha)(1-\beta)r_{22}$
    $= u_r\alpha\beta + \alpha(r_{12} - r_{22}) + \beta(r_{21} - r_{22}) + r_{22}$
    with $u_r = r_{11} - r_{12} - r_{21} + r_{22}$
  - $V_c(\alpha, \beta) = \alpha\beta c_{11} + \alpha(1-\beta)c_{12} + (1-\alpha)\beta c_{21} + (1-\alpha)(1-\beta)c_{22}$
    $= u_c\alpha\beta + \alpha(c_{12} - c_{22}) + \beta(c_{21} - c_{22}) + c_{22}$
    with $u_c = c_{11} - c_{12} - c_{21} + c_{22}$

- We can compute the gradient of the payoff function with respect to the strategy as

$$\frac{\partial V_r(\alpha, \beta)}{\partial \alpha} = \beta u_r + (r_{12} - r_{12})$$

$$\frac{\partial V_c(\alpha, \beta)}{\partial \beta} = \alpha u_c + (c_{12} - c_{12})$$

Infinitesimal gradient ascent (IGA)

- The gradient ascent (GA) algorithm then becomes

$$\alpha_{k+1} = \alpha_k + \eta \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \alpha_k} = \alpha_k + \eta[\beta_k u_r + (r_{12} - r_{12})\ ]$$

$$\beta_{k+1} = \beta_k + \eta \frac{\partial V_c(\alpha_k, \beta_k)}{\partial \beta_k} = \beta_k + \eta[\alpha_k u_c + (c_{12} - c_{12})\ ]$$

Infinitesimal gradient ascent (IGA)

Theorem

If both players follow infinitesimal gradient ascent (IGA), where $\eta \rightarrow 0$, then their strategies will converge to a Nash equilibrium, or the average payoffs over time will converge in the limit of the expected payoffs of a Nash equilibrium

- To implement IGA, one needs to know
  - ✓ the payoff matrix in advance (his own)
  - ✓ the current strategy of other agent
- Difficult to select parameters, i.e., step size decaying rate
  - With poorly selected learning rate, the strategy oscillates between 0 and 1
- Not a practical algorithm

WoLF-IGA

- The WoLF-IGA algorithm, introduced by Bowling and Veloso
- Allows the player to update its strategy based on the current gradient and a **variable learning rate**.
- The **value of the learning rate** is
  - Smaller when the player is wining, and
  - It is larger when the player is losing

- Used in in relatively simple two-action/two-player general-sum games

	β action 1	$1-\beta$ action 2
α action 1	(r_{11}, c_{11})	(r_{12}, c_{12})
$1-\alpha$ action 2	(r_{21}, c_{21})	(r_{22}, c_{22})

WoLF-IGA

- The updating rules of the WoLF-IGA algorithm as follows:

$$\alpha_{k+1} = \alpha_k + \eta l_{r,k} \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \alpha_k}$$

$$\beta_{k+1} = \beta_k + \eta l_{c,k} \frac{\partial V_c(\alpha_k, \beta_k)}{\partial \beta_k}$$

- The varying learning rates, $l_{r,k}$ and $l_{c,k}$, can be computed as

$$l_{r,k} = \begin{cases} l_{min}, & \text{if } V_r(\alpha_k, \beta_k) > V_r(\alpha^*, \beta_k) \\ l_{max}, & \text{otherwise} \end{cases}$$

Current strategy α_kl s doing well

$$l_{c,k} = \begin{cases} l_{min}, & \text{if } V_c(\alpha_k, \beta_k) > V_r(\alpha_k, \beta^*) \\ l_{max}, & \text{otherwise} \end{cases}$$

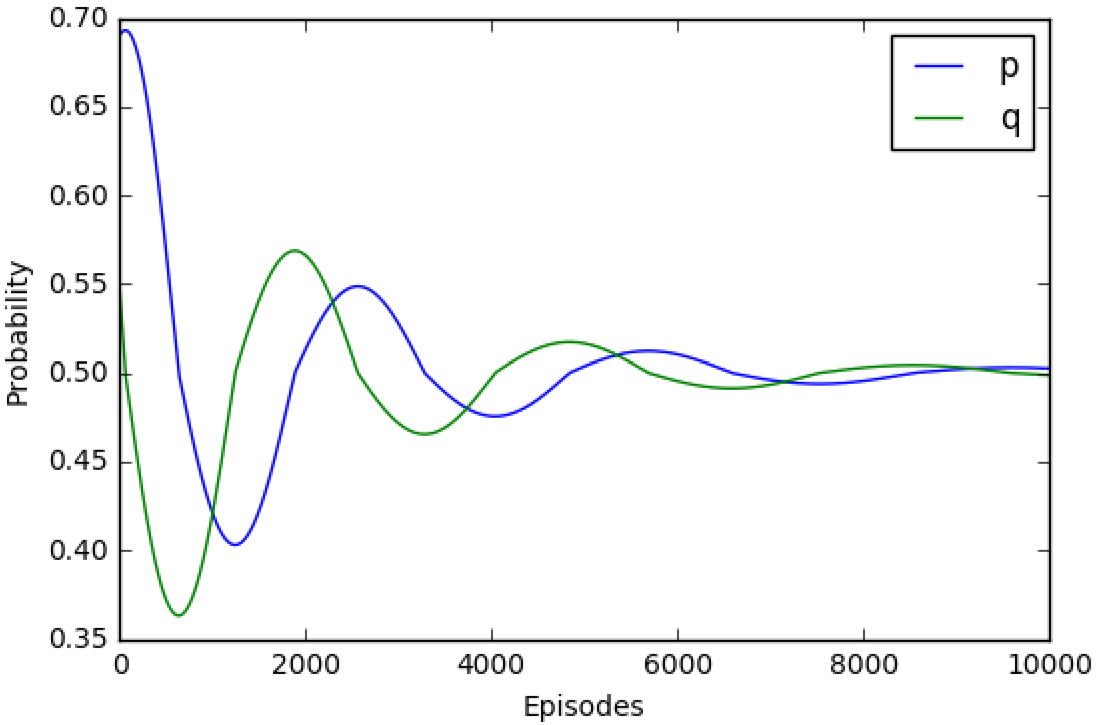
Current strategy β_k is doing well

where (α^*, β^*) is Nash equilibrium

- In a two-player two-action matrix game if each player uses the WoLF-IGA algorithm with $l_{max} > l_{min}$, the players' strategy converges to an NE as the step size $\eta \to 0$.
- It requires the knowledge of $V_r(\alpha^*, \beta_k)$ and $V_r(\alpha_k, \beta^*)$ to compute varying learning rate
  - ➢ To do this, each agent needs to know its own payoff matrix and other's current strategy
- WoLF-IGA algorithm is only available for the two actions

WoLF-IGA

	$Heads$ q	$Tails$ $1-q$
$Heads$ p	$1,-1$	$-1,1$
$Tails$ $1-p$	$-1,1$	$1,-1$



Decentralized Learning in Matrix Games

- Decentralized learning means that there is no central learning strategy for all of the agents. Instead, each agent learns its own strategy.
  - Used when an agent has “incomplete information”
  - The agent **does not know**
    - Its own reward function
    - The other players strategies
    - Other players’ reward function
  - The agent **only knows**
    - its own action and
    - the received reward at each time step

Policy Hill Climbing (PHC)

- Policy Hill Climbing (PHC) algorithm is a more practical version of the gradient descent algorithm

- PHC is based on Q-learning algorithm

- PHC is a rational algorithm
  - Converge to the optimal mixed strategies if the other players are not learning and are therefore playing stationary strategies
  - However, if the other players are learning (non-stationary), PHC may not converge to a stationary policy

- PHC algorithm learns mixed strategies

- PHC does not require much of information. It does not need to know
  - Agent’s payoff matrix
  - Agent’s recent actions executed
  - Other agents’ current strategies (we don’t need to know opponent’s strategies!)

Policy Hill Climbing (PHC)

```
Algorithm  Policy hill − climbing (PHC) algorithm for agent $i$
Initialize
        learning rate $\alpha \in (0,1], \delta \in (0,1]$
        discunt factor $\gamma \in (0,1)$
        exploration rate $\epsilon$
        $Q_i(a_i) \leftarrow 0$ and $\pi_i(a_i) \leftarrow \frac{1}{|A_i|}$ $\forall a_i \in A_i$
  Repeat
    (a) select an action $a_i$ according to the straegy $\pi(a_i)$ with some exploration rate $\epsilon$
    (b) observe the immediate reward $r_i$
    (c) update $Q$ values:
            $Q_i(a_i) = (1-\alpha)Q_i(a_i) + \alpha\left(r_i + \gamma \max_{a_i'} Q_i(a_i')\right)$
    (d) Update the strategy $\pi_i(a_i)$ and constrain it to a legal probability distribution
            $\pi_i(a_i) = \pi_i(a_i) + \begin{cases} \delta & if\ a_i = \max_{a_i'} Q_i(a_i') \\ -\frac{\delta}{|A_i|-1} & otherwise \end{cases}$
```

WoLF-PHC

- The WoLF-PHC algorithm is an extension of the PHC algorithm
- It uses the mechanism of win-or-learn-fast (WoLF) so that the PHC algorithm converges to an NE in self-play
- The algorithm has two different learning rates:
  - δ_w when the algorithm is wining
  - δ_l when the algorithm is losing
  - The difference between the average strategy and the current strategy is used as a criterion to decide when the algorithm wins or lose.
- The learning rate δ_l for losing is larger than the learning rate δ_w for winning
  - When a player is losing, it learns faster then when wining
  - Allow the player to
    - adapt quickly to the changes in the strategies of the other player when it is doing more poorly than expected
    - learns cautiously when it is doing better than expected, which also gives the other players the time to adopt to the player's strategy changes

WoLF-PHC

- The WoLF-PHC algorithm is an extension of the PHC algorithm
  - WoLF(win-or-learn-fast) allows variable learning rate→ faster convergence

WoLF

(c) update Q values:

$$Q_i(a_i) = (1-\alpha)Q_i(a_i) + \alpha\left(r_i + \gamma \max_{a_i'} Q_i(a_i')\right)$$

update estimate of average policy $\bar{\pi}(s,a')$:

$$C \leftarrow C+1$$

$$\forall a' \in A_i, \qquad \bar{\pi}_i(a') \leftarrow \pi_i(a') + \frac{1}{C}\left(\pi_i(a') - \bar{\pi}_i(a')\right)$$

(d) Update the strategy $\pi_i(s,a_i)$ and constrain it to a legal probability distribution

$$\pi_i(a_i) = \pi_i(s,a_i) + \begin{cases} \delta & \text{if } a_i = \max_{a_i'} Q_i(a_i') \\ -\dfrac{\delta}{|A_i|-1} & \text{otherwise} \end{cases}$$

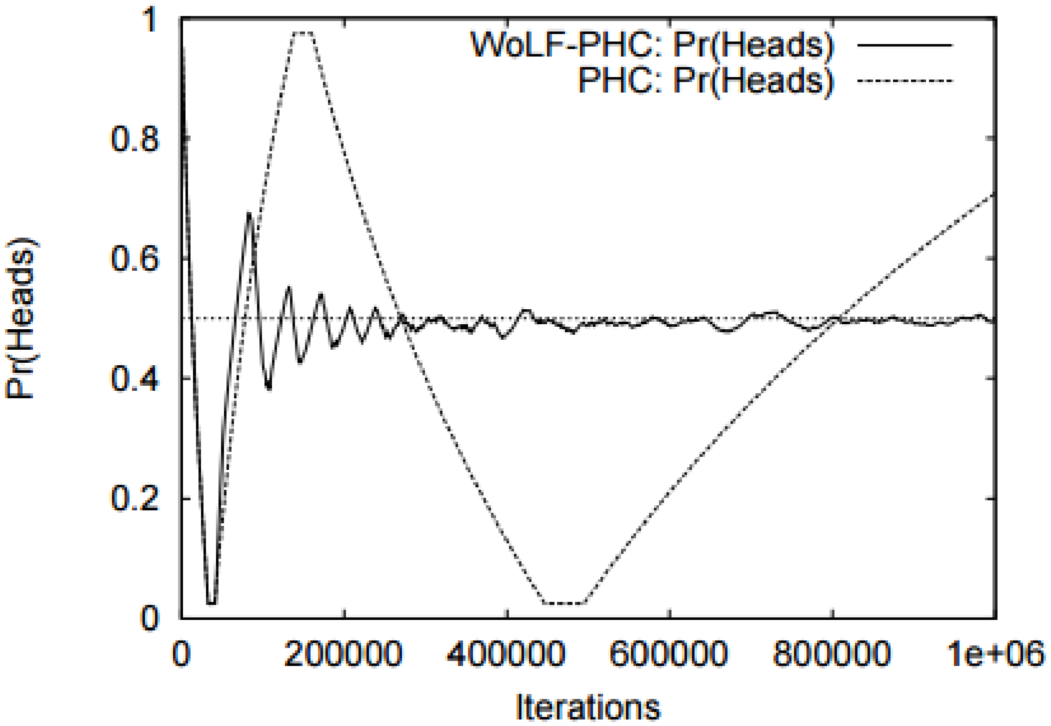
where

$$\delta = \begin{cases} \delta_w & \text{if } \sum_{a_i} \pi_i(a_i)Q_i(a_i) > \sum_{a_i} \bar{\pi}_i(a_i)Q_i(a_i) \\ \delta_l & \text{otherwise} \end{cases}$$

- Average strategy can be computed in running average sense
- The difference between the average strategy and the current strategy is used as a criterion to decide when the algorithm wins or lose.

WoLF-PHC

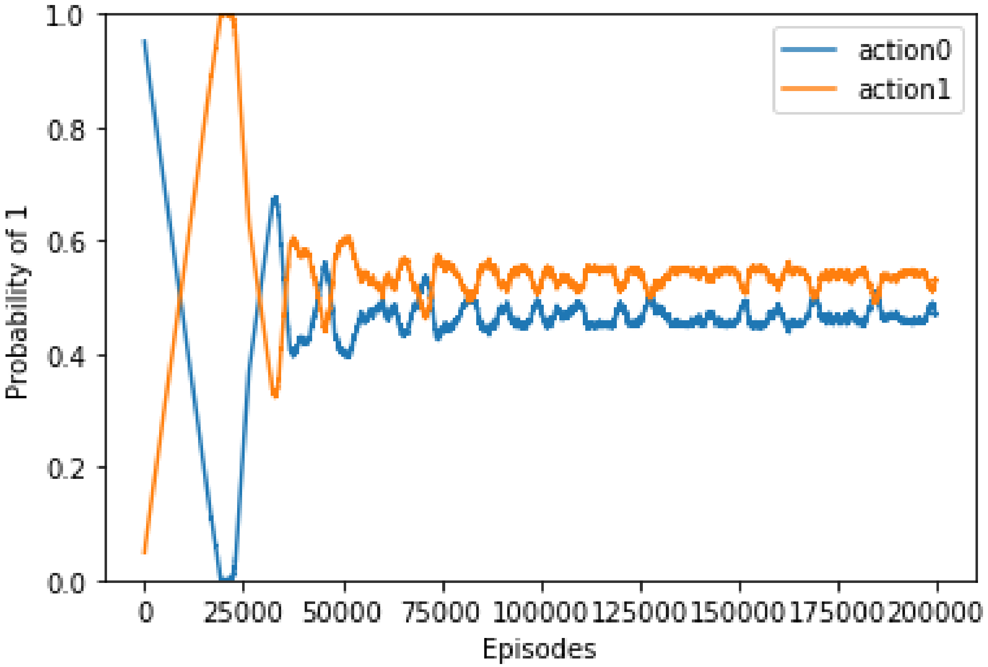
	$Heads$	$Tails$
$Heads$	$1, -1$	$-1, 1$
$Tails$	$-1, 1$	$1, -1$



- The algorithm oscillate about the NE as expected by the theory because both players are learning
- It takes may iterations to converge about the 50% equilibrium point
- Choosing all the parameters is difficult

WoLF-PHC

	Heads	*Tails*
Heads	$1,-1$	$-1,1$
Tails	$-1,1$	$1,-1$



- The algorithm oscillate about the NE as expected by the theory because both players are learning
- It takes may iterations to converge about the 50% equilibrium point
- Choosing all the parameters is difficult

WoLF-PHC

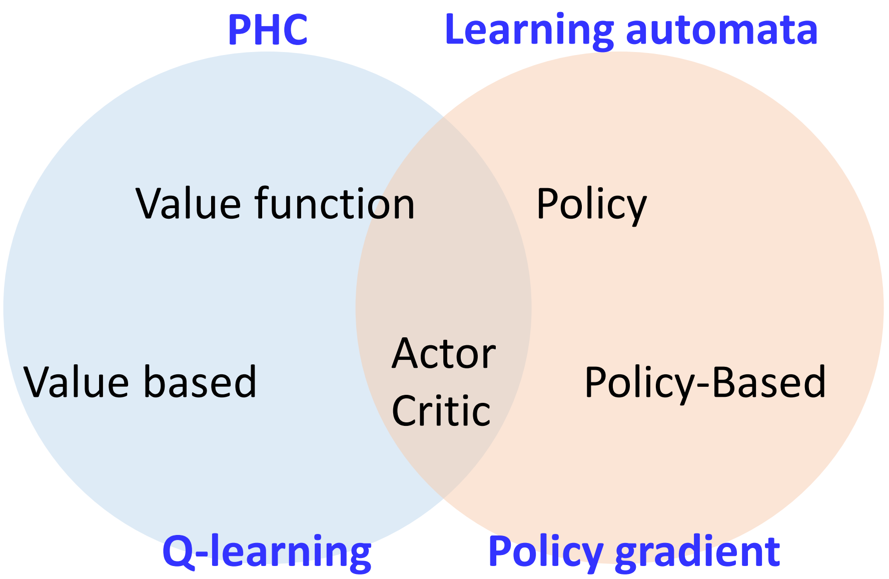
- The WoLF-PHC algorithm exhibits the property of convergences as it makes the player converge to one of its NEs.

- The algorithm is also a rational learning algorithm because
    - It makes the player converge to its optimal strategy when its opponent plays a stationary strategy

- WoLF-PHC algorithm is widely applied to a variety of stochastic games.

Learning Automata

- Learning automation is a learning unit for adaptive decision making in an unknown environment.
- The objective of learning automation is to learning the optimal action or strategy by updating its action probability distribution based on the environment response.
- The learning automata approach is a completely decentralized learning algorithm because each learner only considers its action and the received reward from the environment and ignores any information from other agents such as the actions taken by other agents.
- Two typical learning algorithms based on learning automata
  - Linear reward-inaction (L_{RI})
  - Linear reward-penalty (L_{RP})
- It directly learn the policy



Linear reward-inaction (L_{RI}) algorithm

- The linear reward-inaction (L_{RI}) algorithm for player $i (i = 1, \dots, n)$ is defined as follows:

$$\begin{cases} P_c^i(k+1) = p_c^i(k) + \eta r^i(k)\left(1 - p_c^i(k)\right) & \text{if } a_c \text{ is the current action at } k \\ P_j^i(k+1) = p_j^i(k) - \eta r^i(k) p_j^i(k) & \text{for all } a_j^i \neq a_c^i \text{ is the current action at } k \end{cases}$$

  - $p_c^i(k)$ is the probability for player i to select action $a_c^i (c = 1, \dots, m)$ at time k
  - $0 < \eta < 1$ is the learning parameter
  - $r^i(k)$ response of the environment given player i's action a_c^i at k
  - $\sum_j P_j^i(k+1) = \sum_j P_j^i(k)$
  - Increase the probability of selecting the chosen action, while reducing probability of selecting other actions.

- For example, assume a_1^i is selected among$\left(a_1^i, a_2^i, a_3^i, a_4^i\right)$ is the current action at k

$$\begin{cases} P_1^i(k+1) = p_1^i(k) + \eta r^i(k)\left(1 - p_1^i(k)\right) \quad \Rightarrow p_1^i(k) + \eta r^i(k)\left(p_2^i(k) + p_3^i(k) + p_4^i(k)\right) \\ P_2^i(k+1) = p_2^i(k) - \eta r^i(k) p_2^i(k) \\ P_3^i(k+1) = p_3^i(k) - \eta r^i(k) p_3^i(k) \\ P_4^i(k+1) = p_4^i(k) - \eta r^i(k) p_4^i(k) \end{cases}$$

$$\sum_j P_j^i(k+1) = \sum_j P_j^i(k) + \eta r^i(k)\Big(1 - \underbrace{\left(p_1^i(k) + p_1^i(k) + p_1^i(k) + p_1^i(k)\right)}_{1}\Big) = \sum_j P_j^i(k)$$

Linear reward-inaction (L_{RI}) algorithm

- The linear reward-inaction (L_{RI}) algorithm for player $i(i = 1, \dots, n)$ is defined as follows:

$$\begin{cases} P_c^i(k+1) = p_c^i(k) + \eta r^i(k)\left(1 - p_c^i(k)\right) & \text{if } a_c \text{ is the current action at } k \\ P_j^i(k+1) = p_j^i(k) - \eta r^i(k) p_j^i(k) & \text{for all } a_j^i \neq a_c^i \text{ is the current action at } k \end{cases}$$

  - $p_c^i(k)$ is the probability for player i to select action $a_c^i(c = 1, \dots, m)$ at time k
  - $0 < \eta < 1$ is the learning parameter
  - $r^i(k)$ response of the environment given player i's action a_c^i at k
  - $\sum_j P_j^i(k+1) = \sum_j P_j^i(k)$
  - Increase the probability of selecting the chosen action, while reducing probability of selecting other actions.

- In a matrix game with n players, if each player uses the L_{RI} algorithm,
  - Convergence to NE under the assumption that the game only has strict NEs in pure strategies (S. Lakshmivarahan, 1981)

Linear reward-penalty (L_{RP}) algorithm

- The linear reward-penalty (L_{RP}) algorithm for player $i(i = 1, \dots, n)$ is defined as follows:

$$\begin{cases} P_c^i(k+1) = p_c^i(k) + \eta_1 r^i(k)\left(1 - p_c^i(k)\right) - \eta_2[1 - r^i(k)]p_c^i(k) & \text{if } a_c \text{ is the current action} \\ P_j^i(k+1) = p_j^i(k) - \eta_1 r^i(k) p_j^i(k) + \eta_2[1 - r^i(k)]\left[\dfrac{1}{m-1} - p_j^i(k)\right] & \text{for all } a_j^i \neq a_c^i \end{cases}$$

  - $p_c^i(k)$ is the probability for player i to select action $a_c^i(c = 1, \dots, m)$ at time k
  - $0 < \eta_1, \eta_2 < 1$ is the learning parameters
  - $r^i(k)$ response of the environment given player i's action a_c^i at k
  - $\sum_j P_j^i(k+1) = \sum_j P_j^i(k)$

- For example, assume a_1^i is chosen among$\left(a_1^i, a_2^i, a_3^i, a_4^i\right)$ is the current action at k

$$\begin{cases} P_1^i(k+1) = p_1^i(k) + \eta_1 r^i(k)\left(1 - p_1^i(k)\right) & - \eta_2[1 - r^i(k)]p_1^i(k) \\ P_2^i(k+1) = p_2^i(k) - \eta_1 r^i(k) p_2^i(k) & + \eta_2[1 - r^i(k)][1/3 - p_2^i(k)] \\ P_3^i(k+1) = p_3^i(k) - \eta_1 r^i(k) p_3^i(k) & + \eta_2[1 - r^i(k)][1/3 - p_3^i(k)] \\ P_4^i(k+1) = p_4^i(k) - \eta_1 r^i(k) p_4^i(k) & + \eta_2[1 - r^i(k)][1/3 - p_4^i(k)] \end{cases}$$

If $r^i(k)$ is large, give a larger reward

If $r^i(k)$ is small, give a larger penalty

Linear reward-penalty (L_{RP}) algorithm

- The linear reward-penalty (L_{RP}) algorithm for player $i (i = 1, \dots, n)$ is defined as follows:

$$\begin{cases} P_c^i(k+1) = p_c^i(k) + \eta_1 r^i(k)\left(1 - p_c^i(k)\right) - \eta_2[1 - r^i(k)]p_c^i(k) & \text{if } a_c \text{ is the current action} \\ P_j^i(k+1) = p_j^i(k) - \eta_1 r^i(k) p_j^i(k) + \eta_2[1 - r^i(k)]\left[\frac{1}{m-1} - p_j^i(k)\right] & \text{for all } a_j^i \neq a_c^i \end{cases}$$

  - $p_c^i(k)$ is the probability for player i to select action $a_c^i (c = 1, \dots, m)$ at time k
  - $0 < \eta_1, \eta_2 < 1$ is the learning parameters
  - $r^i(k)$ response of the environment given player i's action a_c^i at k
  - $\sum_j P_j^i(k+1) = \sum_j P_j^i(k)$

- For example, assume a_1^i is chosen among$\left(a_1^i, a_2^i, a_3^i, a_4^i\right)$ is the current action at k

$$\begin{cases} P_1^i(k+1) = p_1^i(k) + \eta_1 r^i(k)\left(1 - p_1^i(k)\right) & - \eta_2[1 - r^i(k)]p_1^i(k) \\ P_2^i(k+1) = p_2^i(k) - \eta_1 r^i(k) p_2^i(k) & + \eta_2[1 - r^i(k)][1/3 - p_2^i(k)] \\ P_3^i(k+1) = p_3^i(k) - \eta_1 r^i(k) p_3^i(k) & + \eta_2[1 - r^i(k)][1/3 - p_3^i(k)] \\ P_4^i(k+1) = p_4^i(k) - \eta_1 r^i(k) p_4^i(k) & + \eta_2[1 - r^i(k)][1/3 - p_4^i(k)] \end{cases}$$

$$\begin{aligned} \sum\nolimits_j P_j^i(k+1) = \sum\nolimits_j P_j^i(k) &+ \eta_1 r^i(k)\left(1 - \left(p_1^i(k) + p_1^i(k) + p_1^i(k) + p_1^i(k)\right)\right) \\ &+ \eta_2[1 - r^i(k)]\left(1/3 + 1/3 + 1/3 - \left(p_1^i(k) + p_1^i(k) + p_1^i(k) + p_1^i(k)\right)\right) \\ = \sum\nolimits_j P_j^i(k) & \end{aligned}$$

Linear reward-penalty (L_{RP}) algorithm

- The linear reward-penalty (L_{RP}) algorithm for player $i (i = 1, \ldots, n)$ is defined as follows:

$$\begin{cases} P_c^i(k+1) = p_c^i(k) + \eta_1 r^i(k)\left(1 - p_c^i(k)\right) - \eta_2[1 - r^i(k)]p_c^i(k) & \text{if } a_c \text{ is the current action} \\ P_j^i(k+1) = p_j^i(k) - \eta_1 r^i(k) p_j^i(k) + \eta_2[1 - r^i(k)]\left[\dfrac{1}{m-1} - p_j^i(k)\right] & \text{for all } a_j^i \neq a_c^i \end{cases}$$

  - $p_c^i(k)$ is the probability for player i to select action $a_c^i (c = 1, \ldots, m)$ at time k
  - $0 < \eta_1, \eta_2 < 1$ is the learning parameters
  - $r^i(k)$ response of the environment given player i's action a_c^i at k
  - $\sum_j P_j^i(k+1) = \sum_j P_j^i(k)$

- fIn a two-player zero-sum matrix game, if each player uses the L_{RP} and chooses $\eta_2 < \eta_1$, then the expected value of the fully mixed strategies for both players can be made arbitrarily close to an NE (S. Lakshmivarahan, 1982)

- This means that the L_{RP} algorithm can guarantee the convergence to an NE in the sense of expected value but not the players' strategy itself.

Comparison

	Algorithms			
	Centralized	Decentralized		
Applicability	WoLF-IGA	WoLF-PHC	L_{RI}	L_{RP}
Allowable actions	Two actions	No limit	No limit	No limit
Convergence	Pure NE & Mixed NE	Pure NE & Mixed NE	Pure NE	Fully Mixed NE