Game Theory, Evolutionary Dynamics, and Multi-Agent Learning

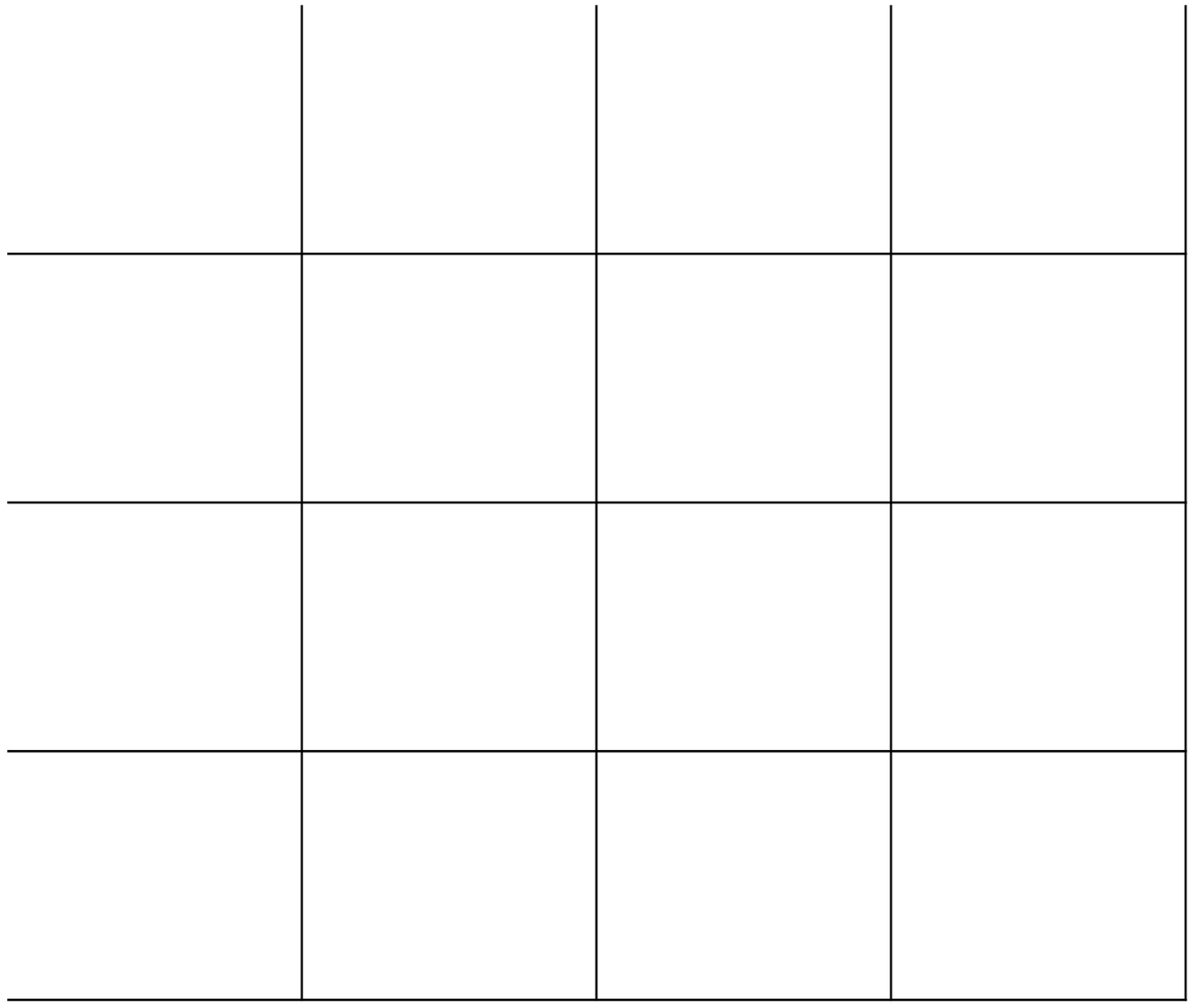
Prof. Nicola Gatti
(nicola.gatti@polimi.it)

Game theory

Game theory: basics

Normal form

- Players
- Actions
- Outcomes
- Utilities
- Strategies
- Solutions



Game theory: basics

Normal form

- Players
- Actions
- Outcomes
- Utilities
- Strategies
- Solutions

Game theory: basics

Normal form

- Players
- Actions
- Outcomes
- Utilities
- Strategies
- Solutions

		Player 2		
		Rock	**Paper**	**Scissors**
Player 1	**Rock**			
	Paper			
	Scissors			

Game theory: basics

Normal form

- Players
- Actions
- Outcomes
- Utilities
- Strategies
- Solutions

Player 1 \ Player 2	**Rock**	**Paper**	**Scissors**
Rock	Tie	Player 2 wins	Player 1 wins
Paper	Player 1 wins	Tie	Player 2 wins
Scissors	Player 2 wins	Player 1 wins	Tie

Game theory: basics

Normal form

- Players
- Actions
- Outcomes
- **Utilities**
- Strategies
- Solutions

Player 1 \ Player 2	**Rock**	**Paper**	**Scissors**
Rock	0,0	-1,1	1,-1
Paper	1,-1	0,0	-1,1
Scissors	-1,1	1,-1	0,0

Game theory: basics

Normal form

- Players
- Actions
- Outcomes
- Utilities
- **Strategies**
- Solutions

Player 2

Player 1	**Rock**	**Paper**	**Scissors**	
Rock	0,0	-1,1	1,-1	$\sigma_1(\text{Rock})$
Paper	1,-1	0,0	-1,1	$\sigma_1(\text{Paper})$
Scissors	-1,1	1,-1	0,0	$\sigma_1(\text{Scissors})$
	$\sigma_2(\text{Rock})$	$\sigma_2(\text{Paper})$	$\sigma_2(\text{Scissors})$	

Game theory: basics

Normal form

- Players
- Actions
- Outcomes
- Utilities
- Strategies
- Solutions

Player 1 \ Player 2	**Rock**	**Paper**	**Scissors**	
Rock	0,0	-1,1	1,-1	1/3
Paper	1,-1	0,0	-1,1	1/3
Scissors	-1,1	1,-1	0,0	1/3
	1/3	1/3	1/3	

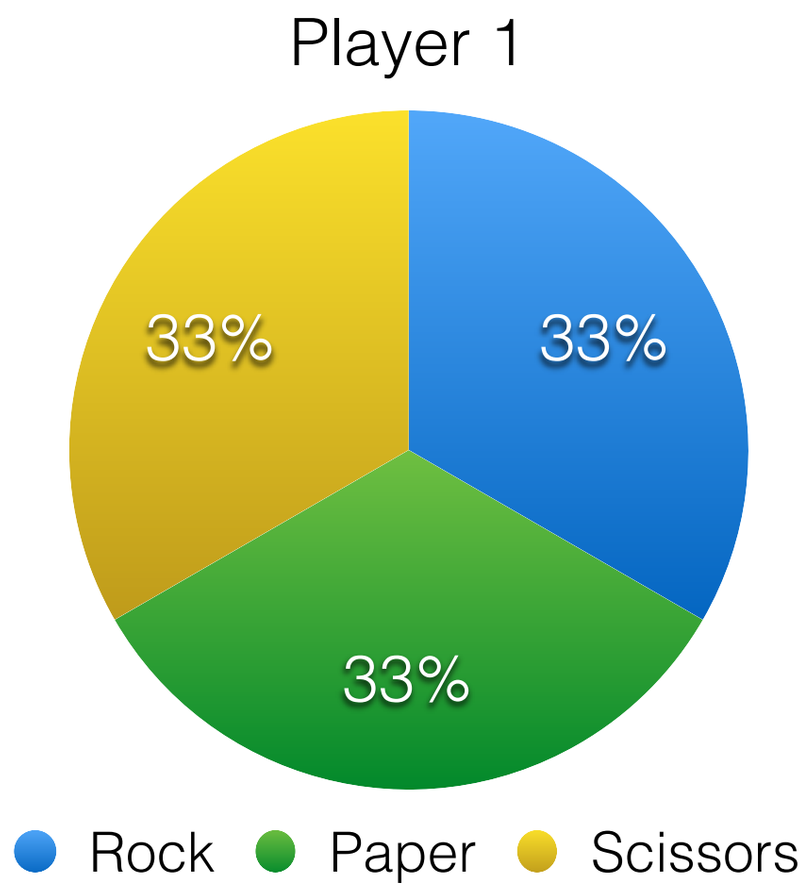
Nash Equilibrium

A strategy profile (σ_1^*, σ_2^*) is a Nash equilibrium if and if:

- $\sigma_1^* \in \arg\max_{\sigma_1} \left\{ \sigma_1 U_1 \sigma_2^* \right\}$
- $\sigma_2^* \in \arg\max_{\sigma_2} \left\{ \sigma_1^* U_2 \sigma_2 \right\}$

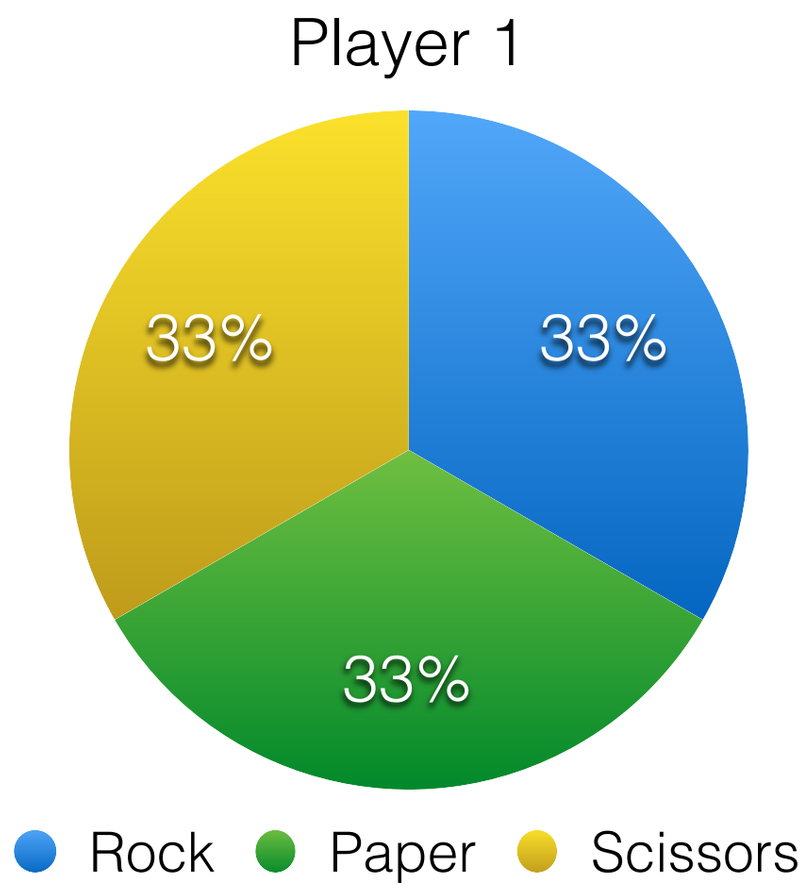
Evolutionary dynamics

An evolutionary interpretation



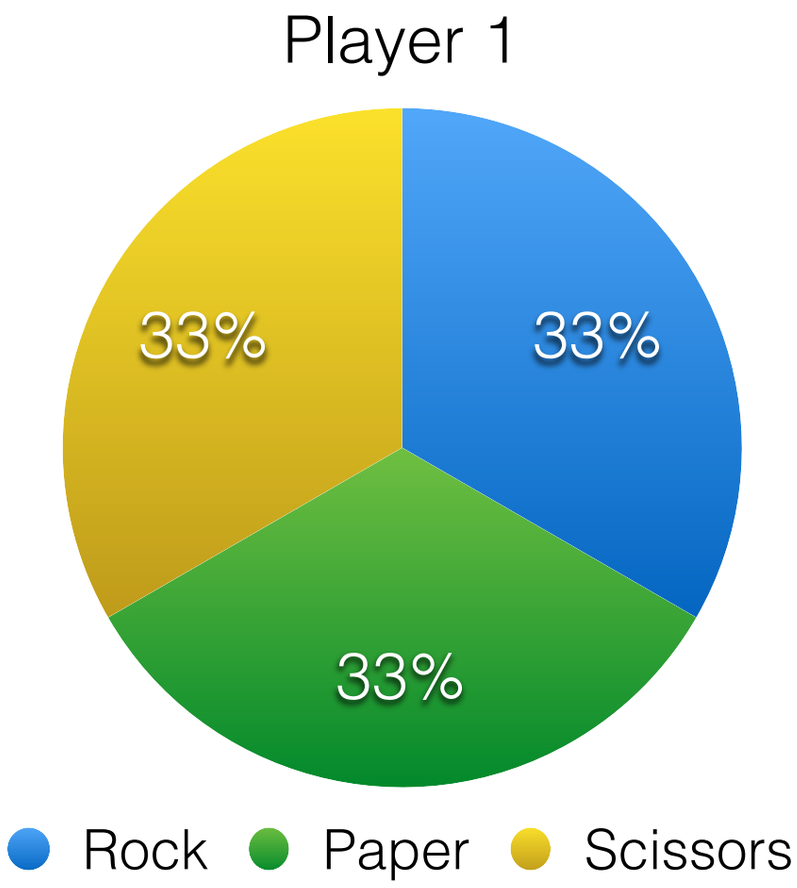
An evolutionary interpretation

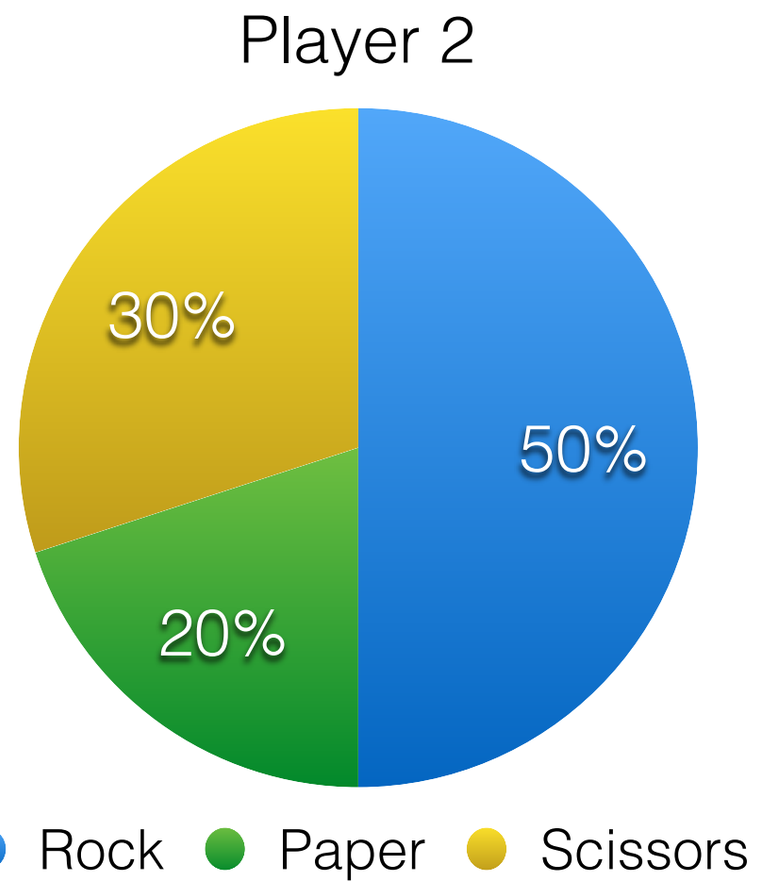
Infinite population of individuals



Bacterials

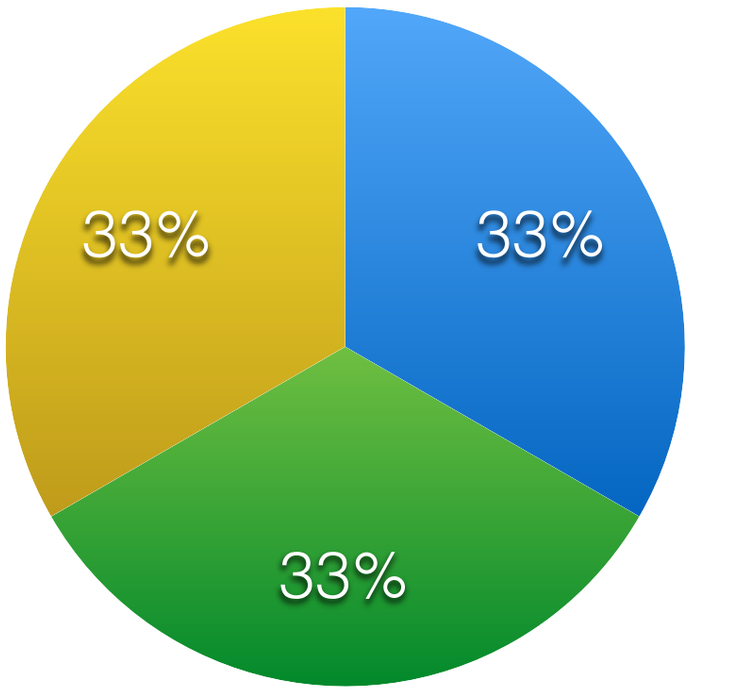
An evolutionary interpretation





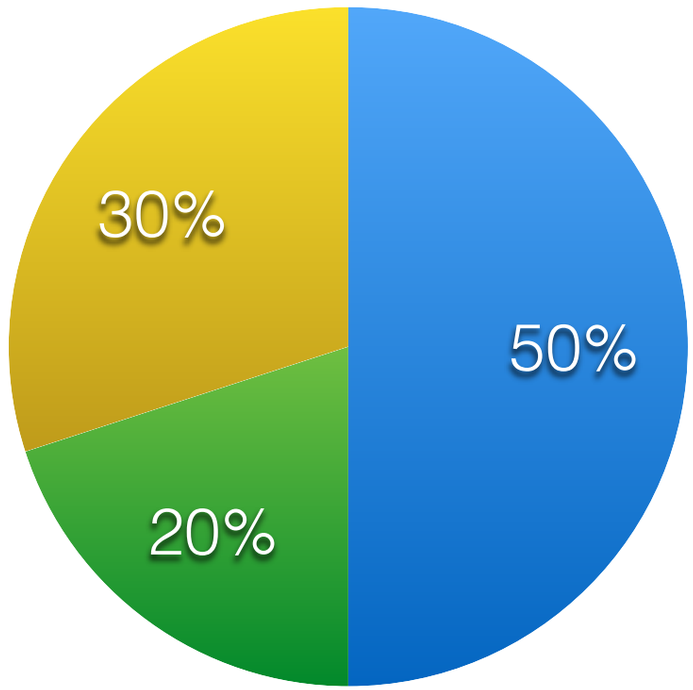
An evolutionary interpretation

Player 1



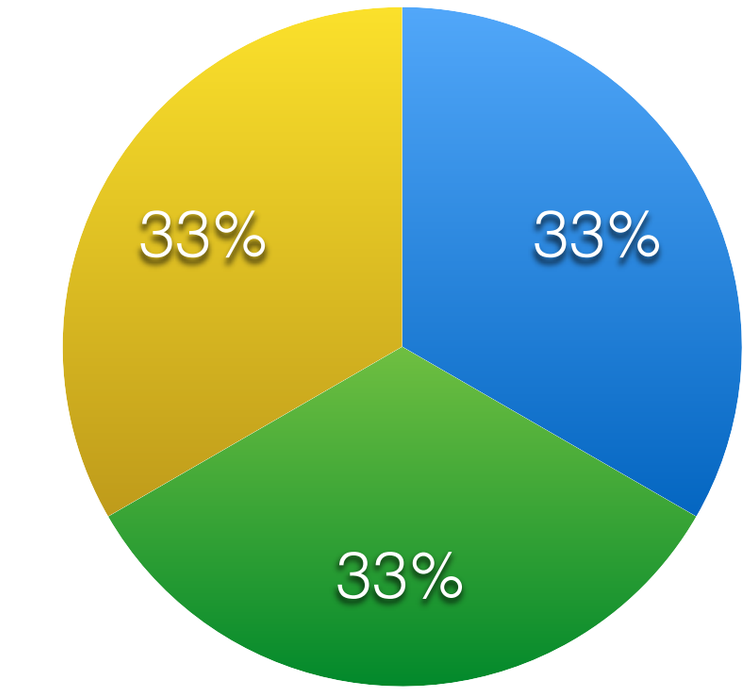
	Rock	Paper	Scissors
Rock	0,0	-1,1	1,-1
Paper	1,-1	0,0	-1,1
Scissors	-1,1	1,-1	0,0

Player 2



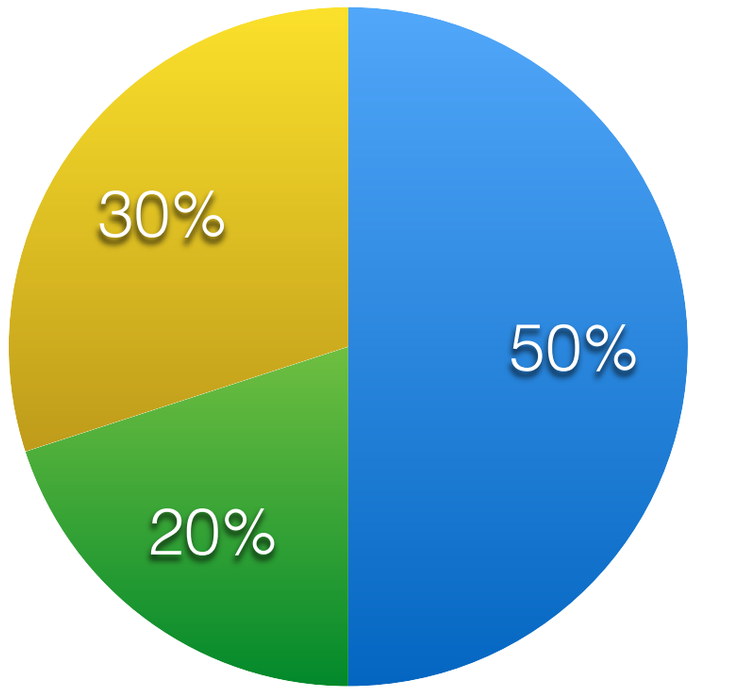
An evolutionary interpretation

Player 1



	Rock	Paper	Scissors
Rock	0,0	-1,1	1,-1
Paper	1,-1	0,0	-1,1
Scissors	-1,1	1,-1	0,0

Player 2

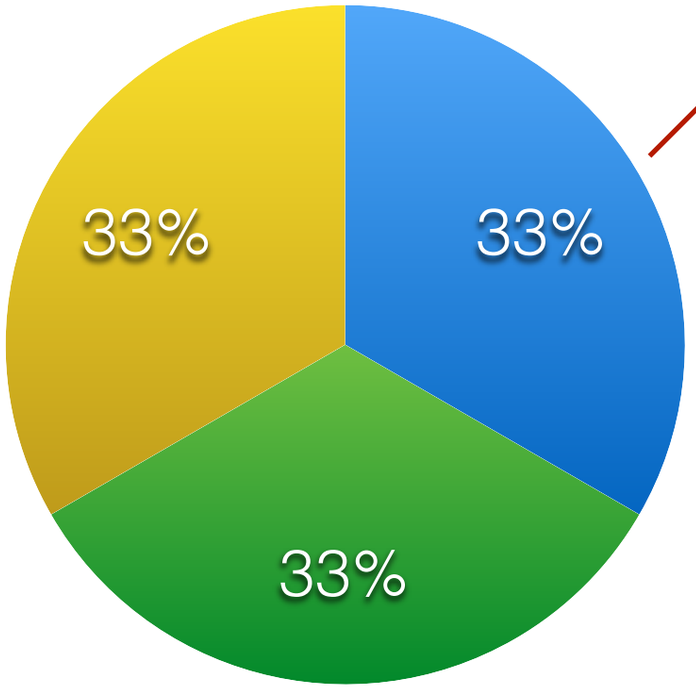


At each round, each individual of a population plays against each individual of the opponent's population weighted by the corresponding probability

An evolutionary interpretation

Player 1

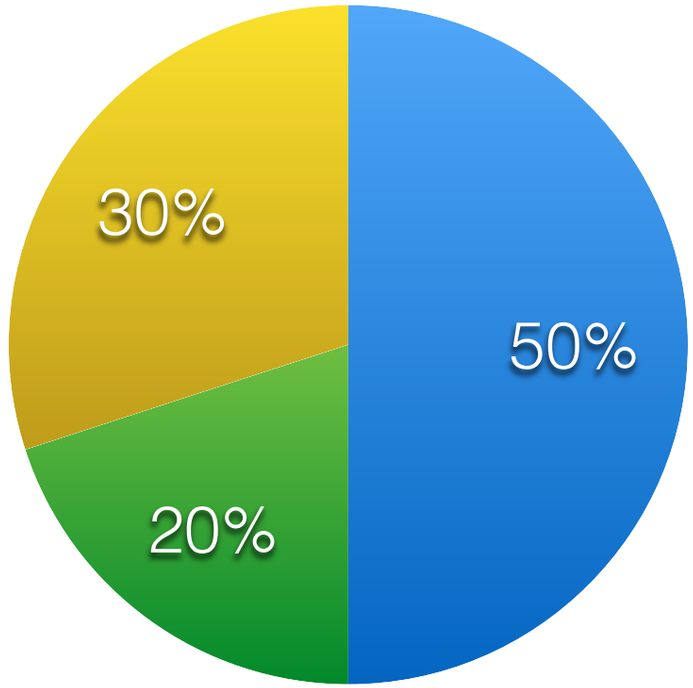
Utility = 0*1/2 - 1*1/5 + 1*3/10 = 0.1



	Rock	Paper	Scissors
Rock	0,0	-1,1	1,-1
Paper	1,-1	0,0	-1,1
Scissors	-1,1	1,-1	0,0

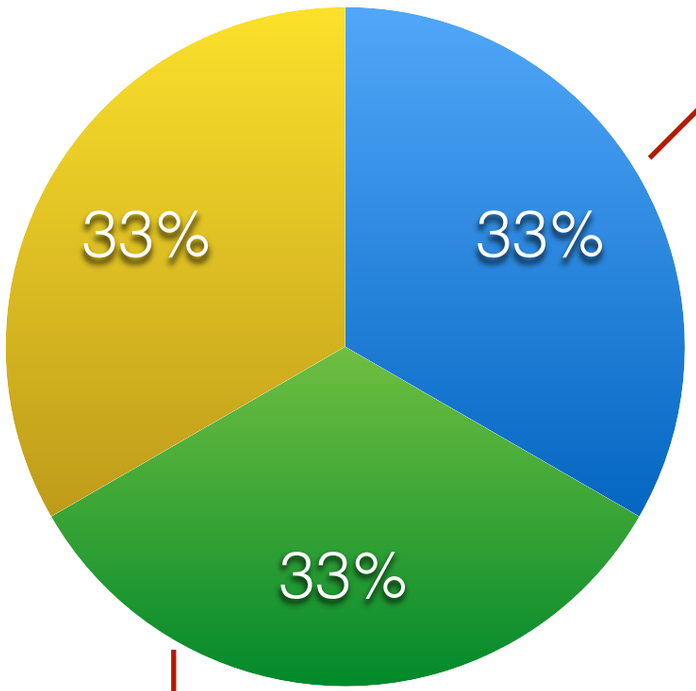
1/2 1/5 3/10

Player 2



An evolutionary interpretation

Player 1



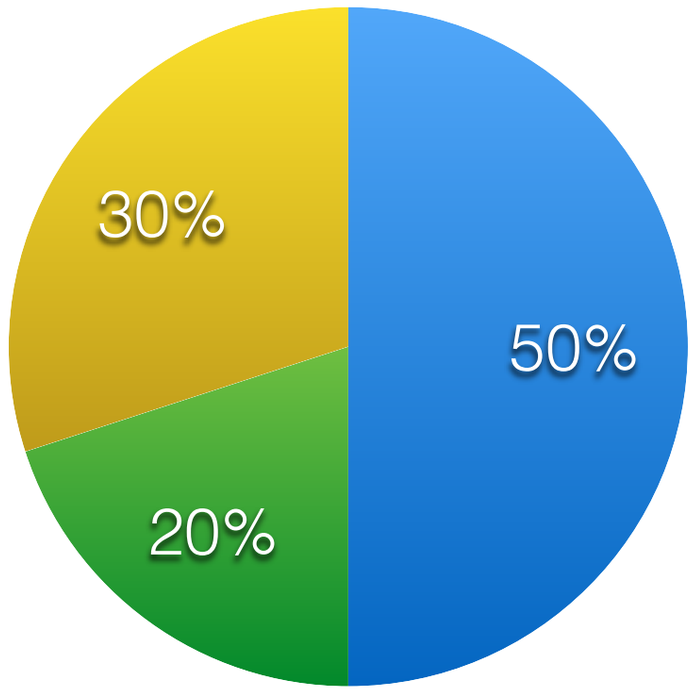
Rock Paper Scissors

Utility = 0.1

	Rock	Paper	Scissors
Rock	0,0	-1,1	1,-1
Paper	1,-1	0,0	-1,1
Scissors	-1,1	1,-1	0,0

1/2 1/5 3/10

Player 2



Rock Paper Scissors

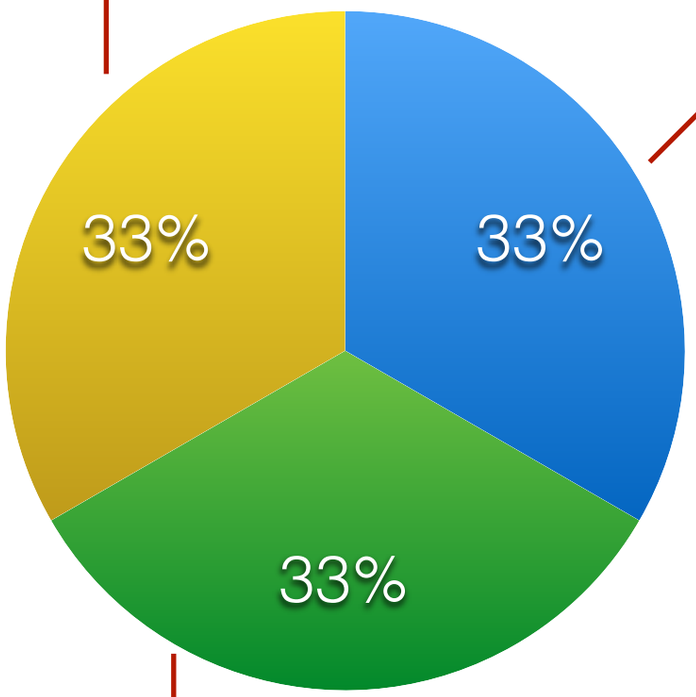
Utility = 1*1/2 - 0*1/5 - 1*3/10 = 0.2

An evolutionary interpretation

Utility = -1*1/2 - 1*1/5 - 0*3/10 = -0.3

Player 1

Utility = 0.1

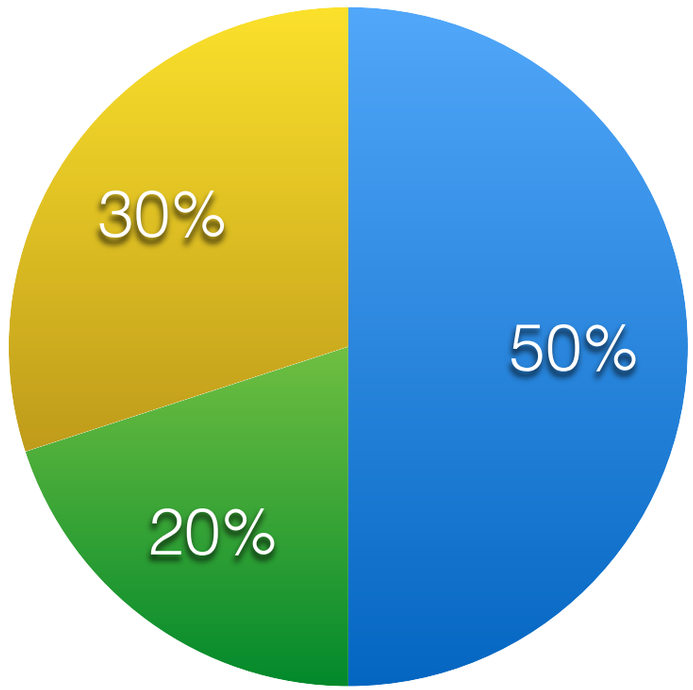


Utility = 0.2

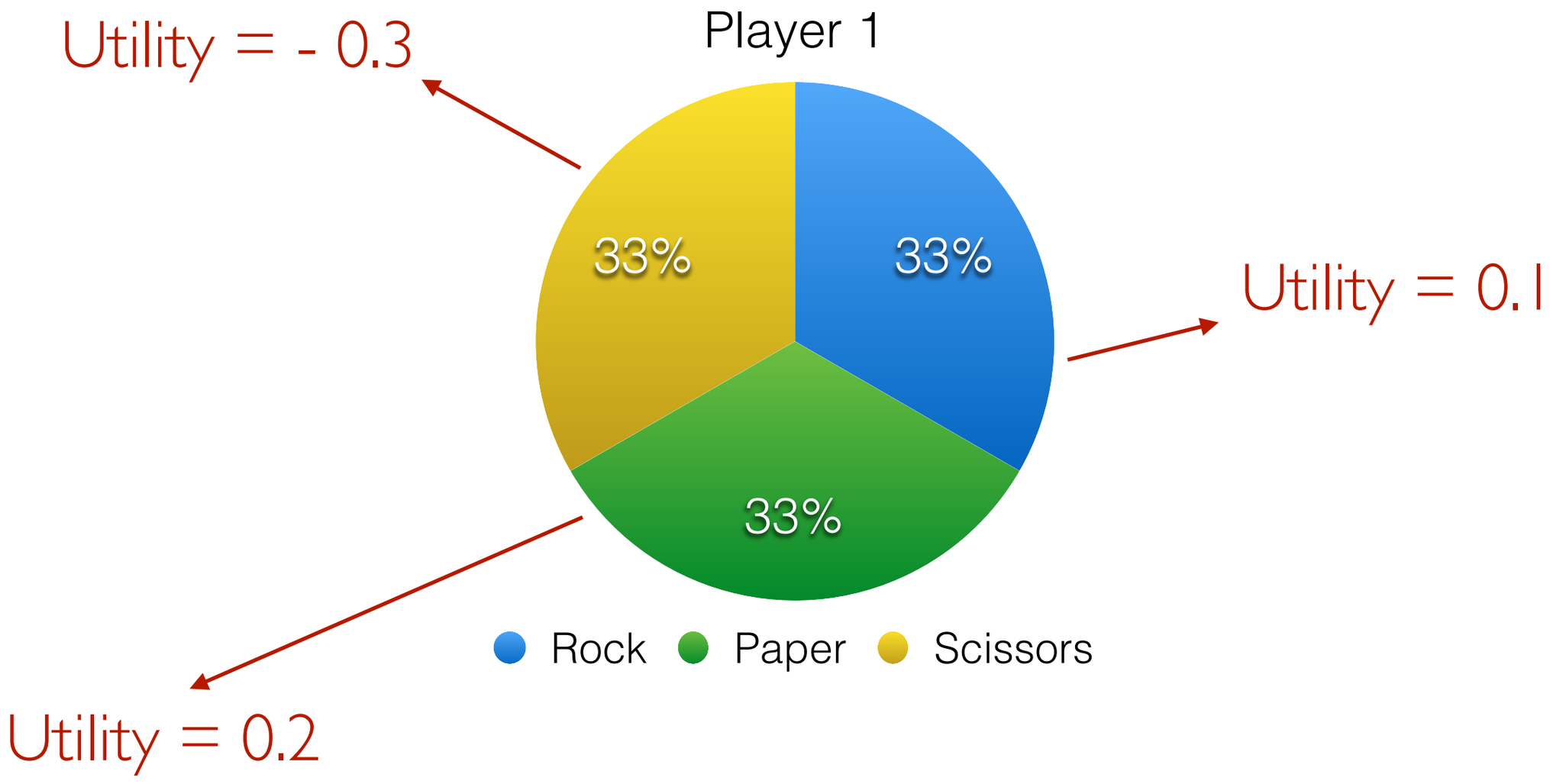
	Rock	Paper	Scissors
Rock	0,0	-1,1	1,-1
Paper	1,-1	0,0	-1,1
Scissors	-1,1	1,-1	0,0

1/2 1/5 3/10

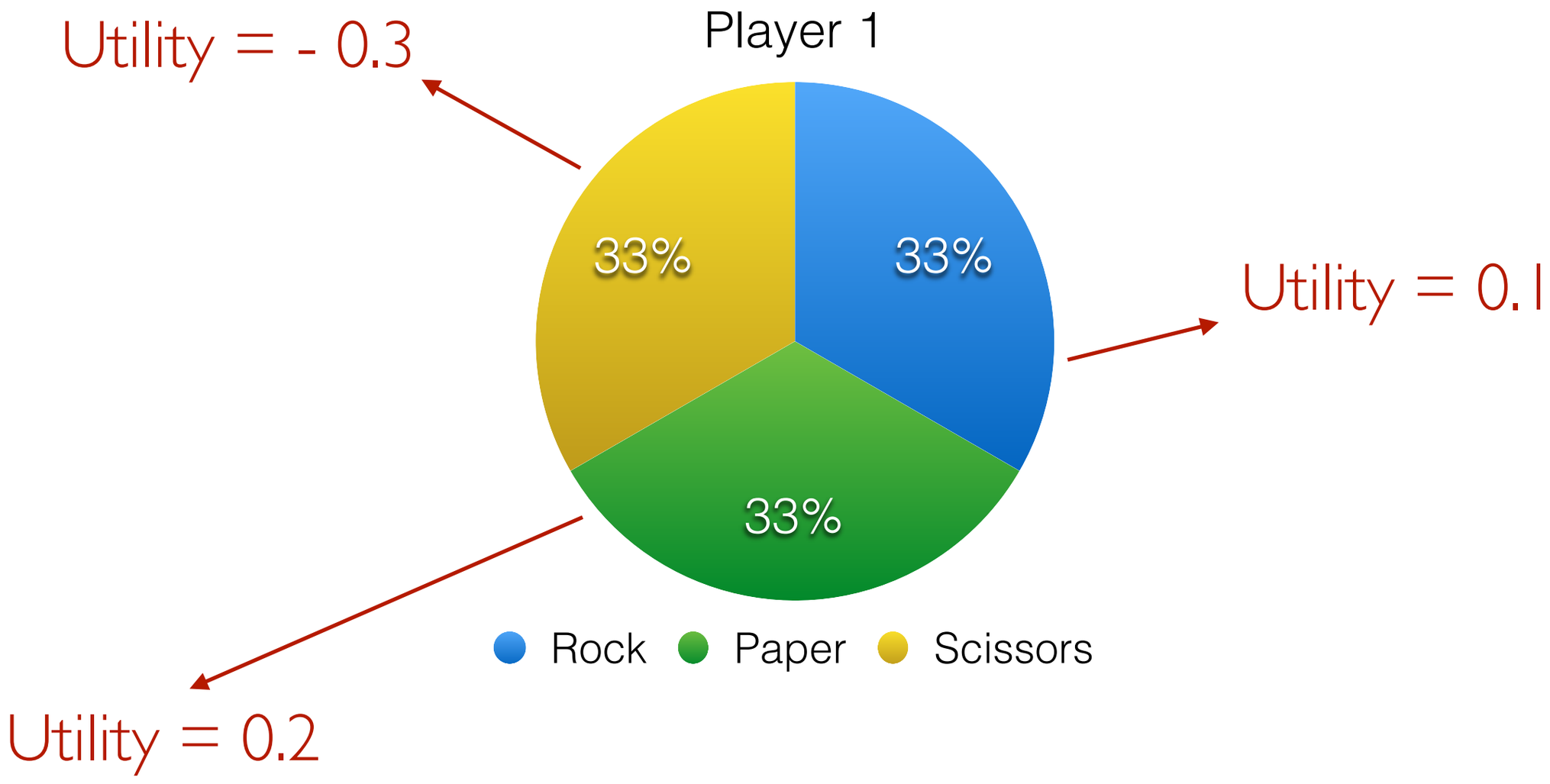
Player 2



An evolutionary interpretation

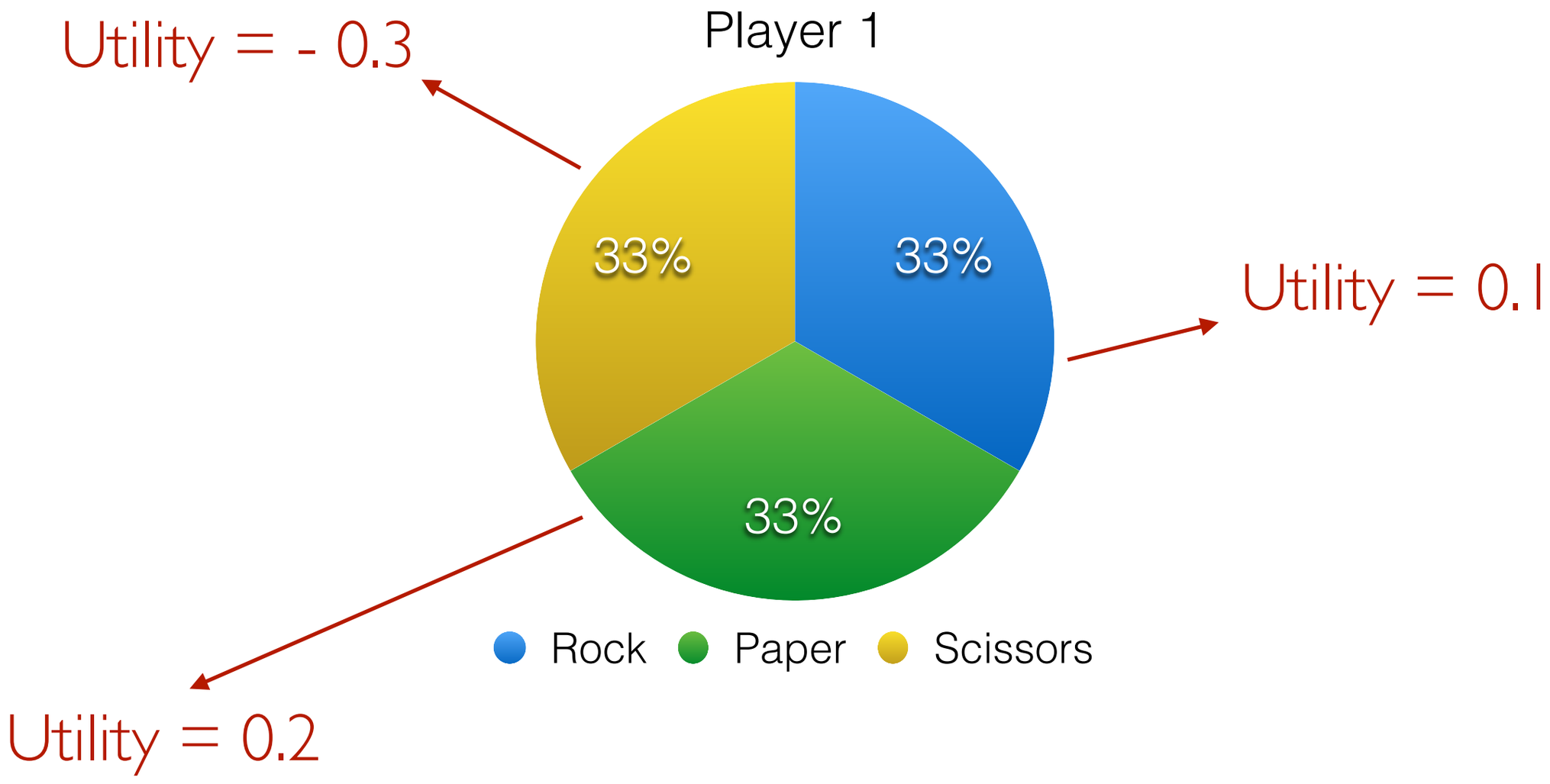


An evolutionary interpretation



The average utility is 1/3 * 0.1 + 1/3 * 0.2 - 0.3 * 1/3 = 0

An evolutionary interpretation

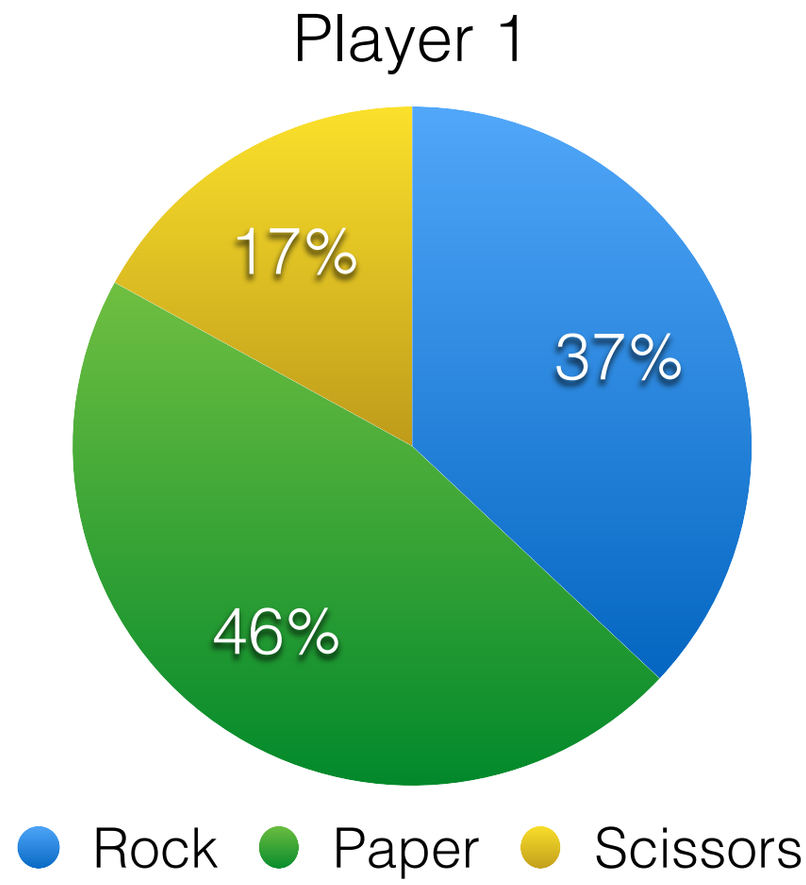


The average utility is 0

A **positive** (*utility - average utility*) leads to an **increase** of the population

A **negative** (*utility - average utility*) leads to a **decrease** of the population

An evolutionary interpretation



New population after the replication

Revision protocol

Question: how the populations change?

Replicator dynamics

$$\dot{\sigma}_1(a,t) = \sigma_1(a,t)\left(e_a U_1 \sigma_2(t) - \sigma_1(t) U_1 \sigma_2(t)\right)$$

Utility given by playing a with a probability of 1

Average population utility

Example (1)

$$\sigma_1(t) = \left[\begin{array}{ccc} \sigma_1(\mathrm{R},t) & \sigma_1(\mathrm{P},t) & \sigma_1(\mathrm{S},t) \end{array} \right]$$

$$\sigma_2(t) = \left[\begin{array}{ccc} \sigma_2(\mathrm{R},t) & \sigma_2(\mathrm{P},t) & \sigma_2(\mathrm{S},t) \end{array} \right]$$

$$U_1 = \left[\begin{array}{ccc} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{array} \right] \qquad U_2 = \left[\begin{array}{ccc} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{array} \right]$$

$$U_1(\mathrm{R},\cdot) = \left[\begin{array}{ccc} 0 & 1 & -1 \end{array} \right]$$

$$U_1(\mathrm{P},\cdot) = \left[\begin{array}{ccc} -1 & 0 & 1 \end{array} \right]$$

$$U_1(\mathrm{S},\cdot) = \left[\begin{array}{ccc} 1 & -1 & 0 \end{array} \right]$$

Example (2)

$$\dot{\sigma}_1(\mathrm{R},t) \quad = \sigma_1(\mathrm{R},t) \quad \Big(U_1(\mathrm{R},\cdot)\,\sigma_2(t) \quad -\sigma_1(t)\,U_1\,\sigma_2(t)\Big)$$

$$\dot{\sigma}_1(\mathrm{P},t) \quad = \sigma_1(\mathrm{P},t) \quad \Big(U_1(\mathrm{P},\cdot)\,\sigma_2(t) \quad -\sigma_1(t)\,U_1\,\sigma_2(t)\Big)$$

$$\dot{\sigma}_1(\mathrm{S},t) \quad = \sigma_1(\mathrm{S},t) \quad \Big(U_1(\mathrm{S},\cdot)\,\sigma_2(t) \quad -\sigma_1(t)\,U_1\,\sigma_2(t)\Big)$$

Evolutionary Stable Strategies

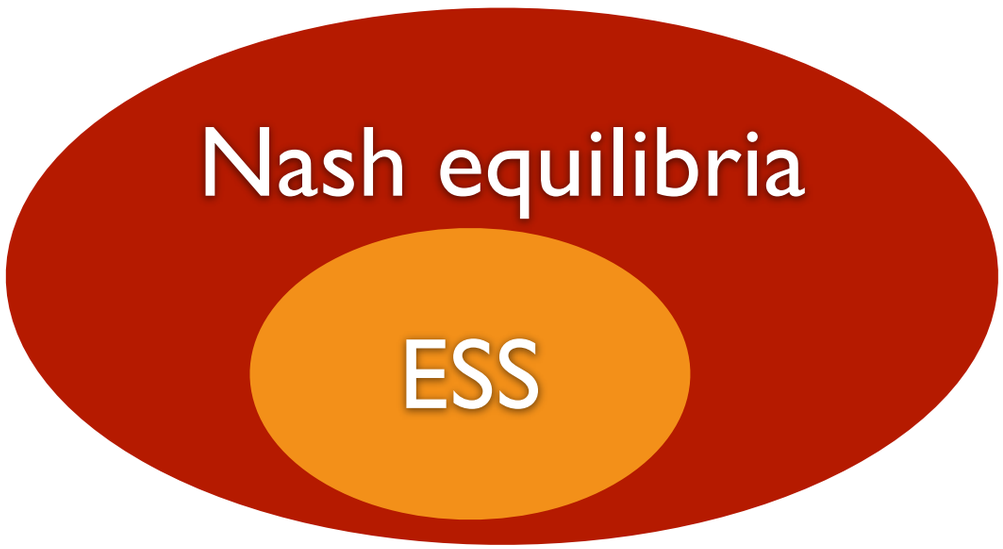
A strategy is an ESS if it is immune to invasion by mutant strategies, given that the mutants initially occupy a small fraction of population

Every ESS is an asymptotically stable fixed point of the replicator dynamics

Evolutionary Stable Strategies (1)

A strategy is an ESS if it is immune to invasion by mutant strategies, given that the mutants initially occupy a small fraction of population

Every ESS is an asymptotically stable fixed point of the replicator dynamics



While a NE always exists, an ESS may not exist

Evolutionary Stable Strategies (2)

Given a Nash equilibrium, a small perturbation could make the equilibrium unstable

Player 1 \ Player 2	**Action 1**	**Action 2**
Action 1	1, 1	0, 0
Action 2	0, 0	0, 0

Nash equilibrium

Evolutionary Stable Strategies (2)

Given a Nash equilibrium, a small perturbation could make the equilibrium unstable

Player 1 \ Player 2	Action 1	Action 2	
Action 1	1, 1	0, 0	ϵ
Action 2	0, 0	0, 0	$1-\epsilon$
	ϵ	$1-\epsilon$	

Evolutionary Stable Strategies (2)

Given a Nash equilibrium, a small perturbation could make the equilibrium unstable

The only best response is Action 1

Player 1 \ Player 2	**Action 1**	**Action 2**	
Action 1	1, 1	0, 0	ϵ
Action 2	0, 0	0, 0	$1-\epsilon$
	ϵ	$1-\epsilon$	

Evolutionary Stable Strategies (2)

Given a Nash equilibrium, a small perturbation could make the equilibrium unstable

Player 2

ESS	Action 1	Action 2	
Action 1	1, 1	0, 0	$1-\epsilon$
Action 2	0, 0	0, 0	ϵ
	$1-\epsilon$	ϵ	

Player 1

Prisoner's dilemma

		Player 2	
		Cooperate	**Defeat**
Player 1	**Cooperate**	3,3	0,5
	Defeat	5,0	1,1

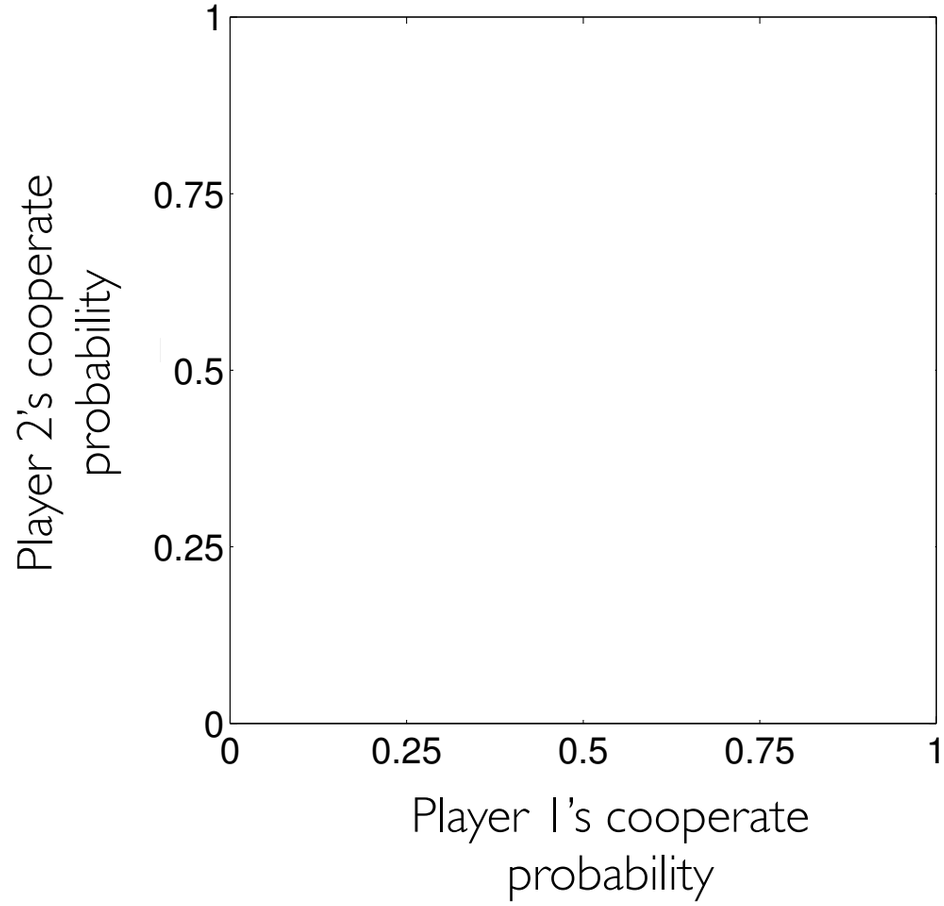
Prisoner's dilemma

Player 1 \ Player 2	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1

Prisoner's dilemma

Player 2

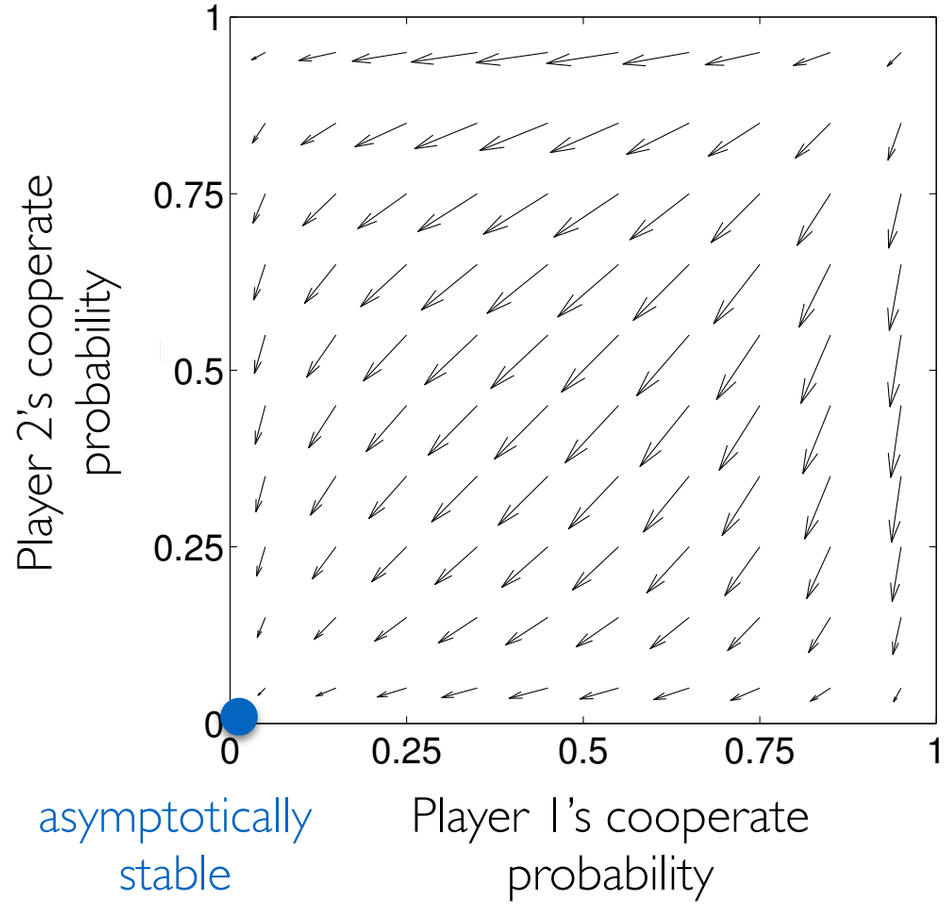
Player 1	Cooperate	Defeat
Cooperate	3,3	0,5
Defeat	5,0	1,1



Prisoner's dilemma

Player 2

Player 1	Cooperate	Defeat
Cooperate	3,3	0,5
Defeat	5,0	1,1



Stag hunt

Player 1 \ Player 2	**Stag**	**Hare**
Stag	4,4	1,3
Hare	3,1	3,3

Stag hunt

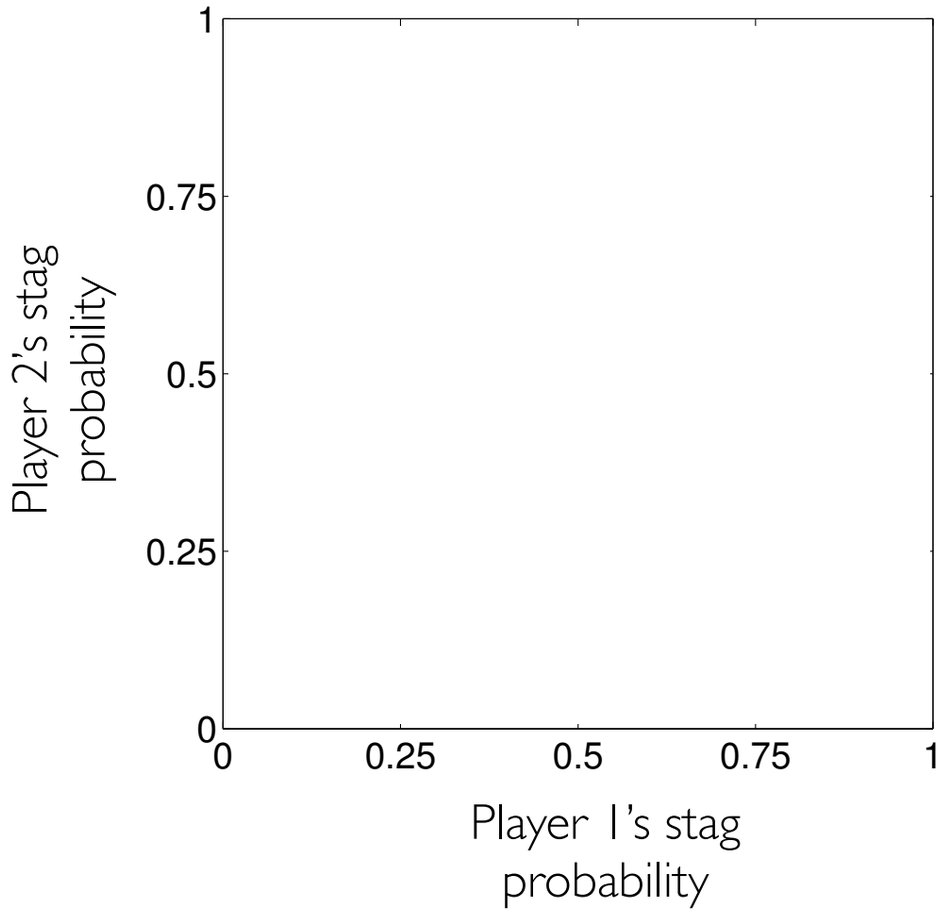
Player 2

Player 1	Stag	Hare
Stag	4,4	1,3
Hare	3,1	3,3

Stag hunt

Player 2

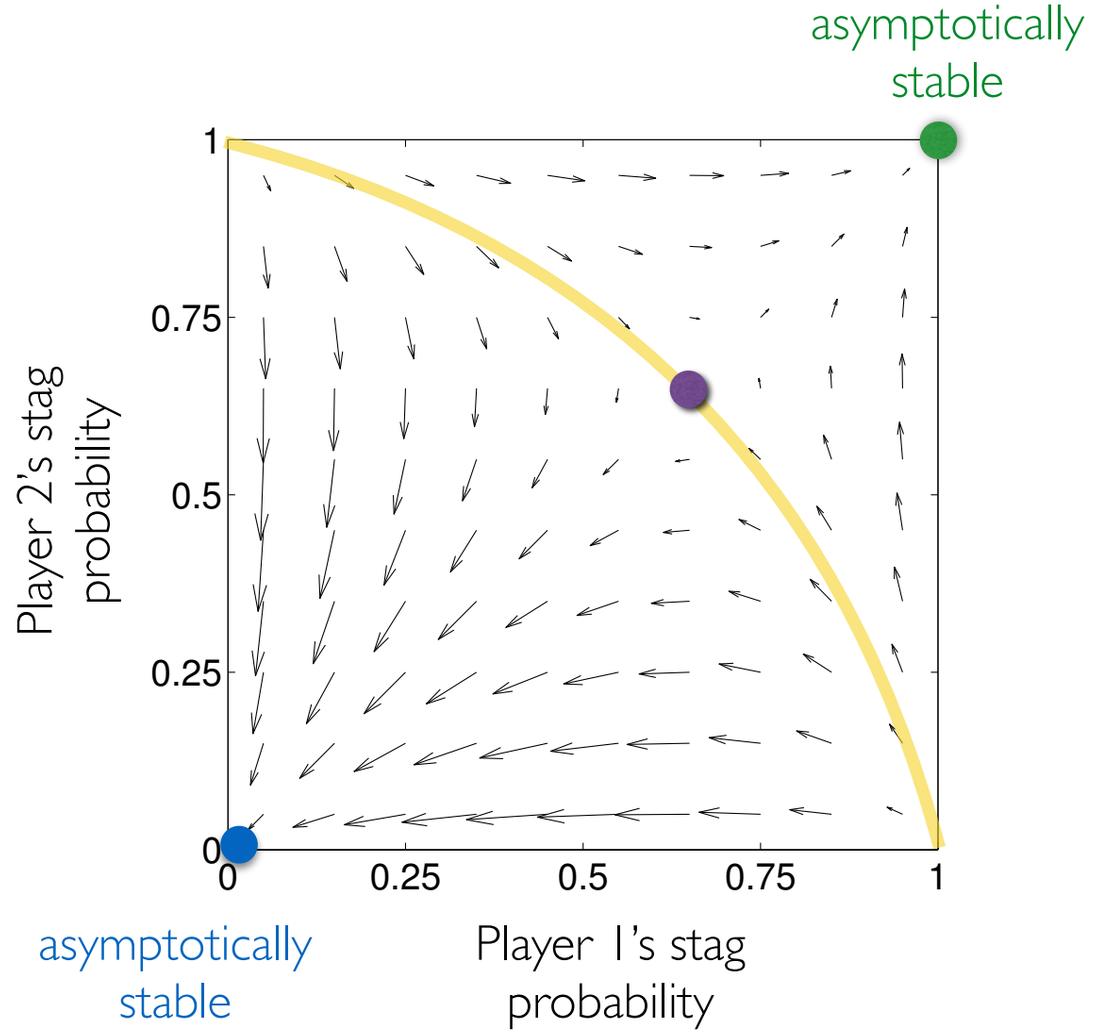
Player 1	Stag	Hare
Stag	4,4	1,3
Hare	3,1	3,3



Stag hunt

Player 2

Player 1	Stag	Hare
Stag	4,4	1,3
Hare	3,1	3,3



Matching pennies

Player 1 \ Player 2	**Head**	**Tail**
Head	0,1	1,0
Tail	1,0	0,1

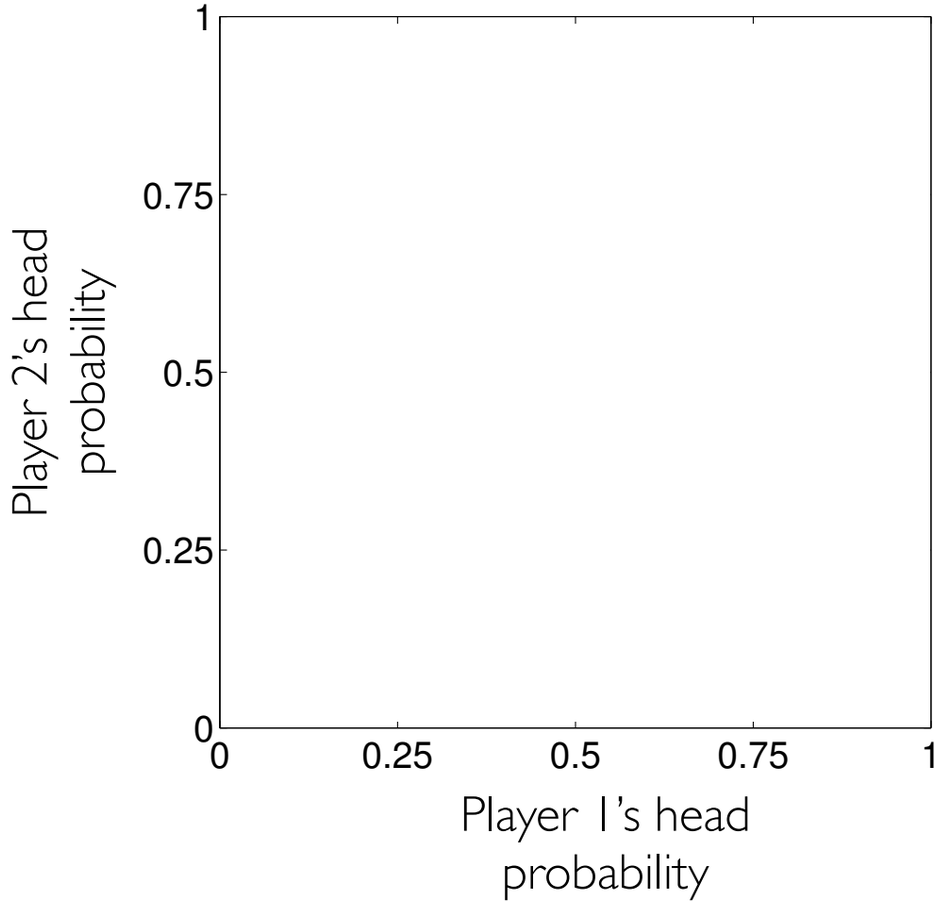
Matching pennies

Player 1 \ Player 2	Head	Tail
Head	0,1	1,0
Tail	1,0	0,1

Matching pennies

Player 2

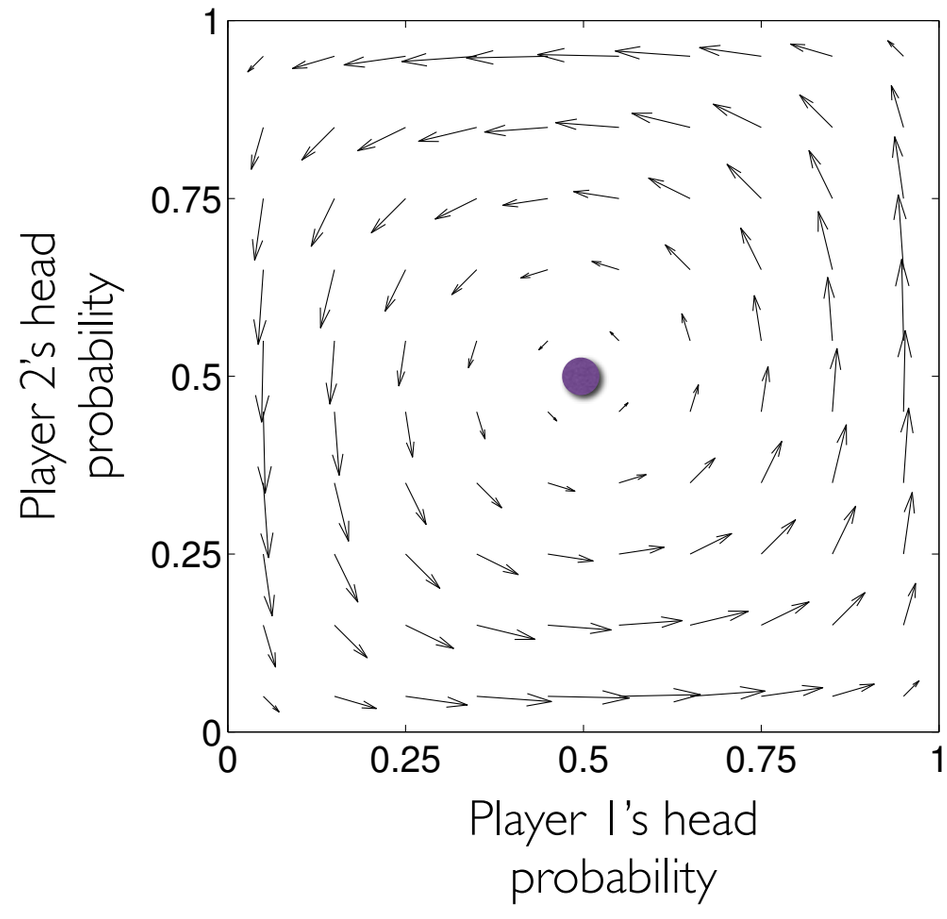
Player 1	Head	Tail
Head	0,1	1,0
Tail	1,0	0,1



Matching pennies

Player 2

Player 1	Head	Tail
Head	0,1	1,0
Tail	1,0	0,1



Starting point

$$\dot{\sigma}_1(a,t) = \sigma_1(a,t)\Big(e_a U_1 \sigma_2(t) - \sigma_1(t) U_1 \sigma_2(t)\Big)$$

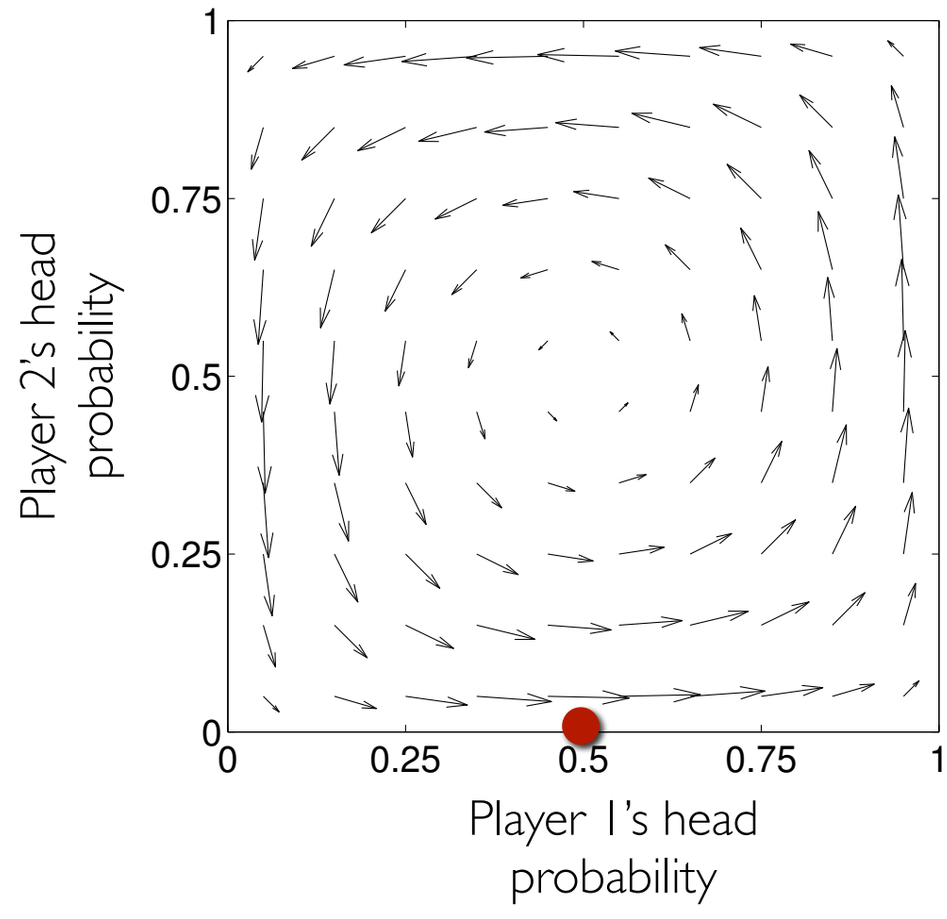
If the starting point is 0, then it will be always zero

Having a starting point with zero strategies does not make sense

Starting point

Player 2

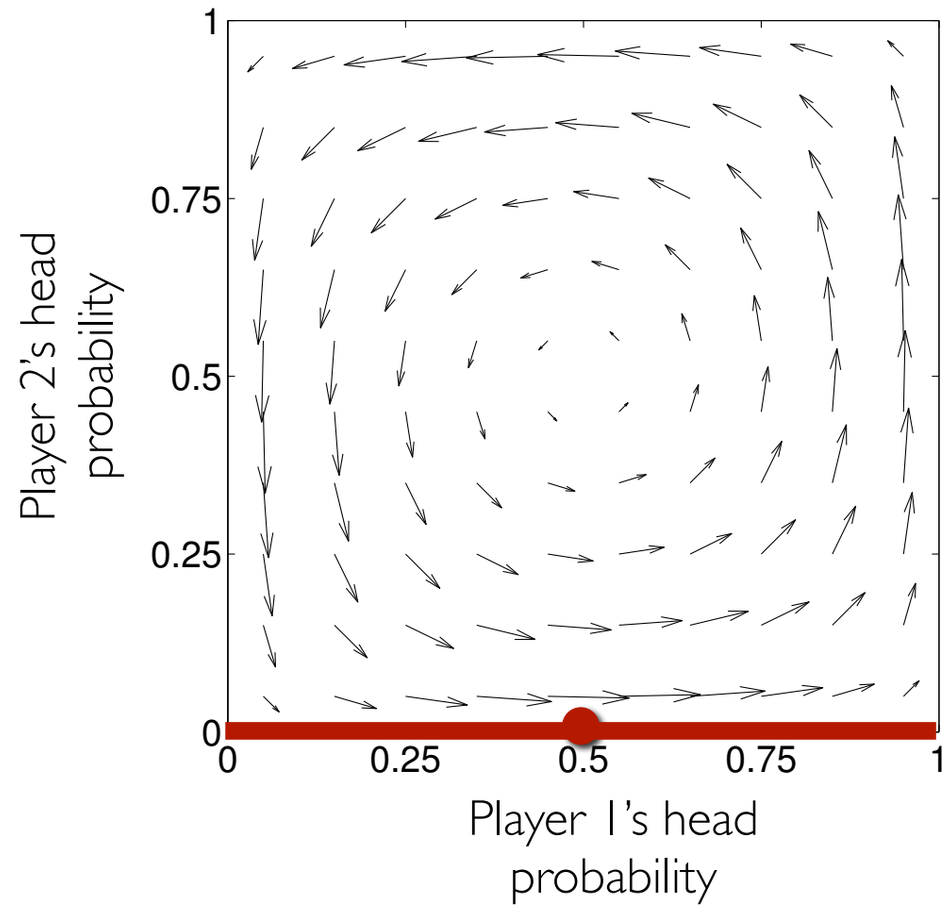
Player 1	Action 1	Action 2
Action 1	…, …	…, …
Action 2	…, …	…, …



Starting point

Player 2

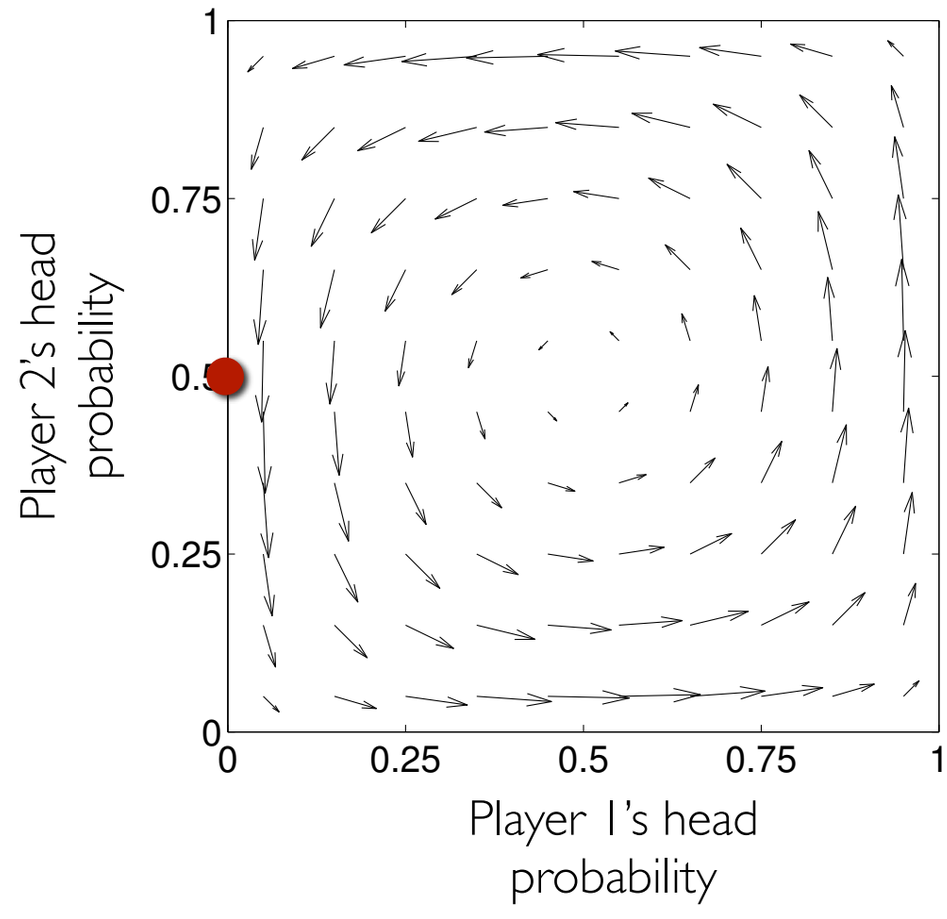
Player 1	Action 1	Action 2
Action 1	…, …	…, …
Action 2	…, …	…, …



Starting point

Player 2

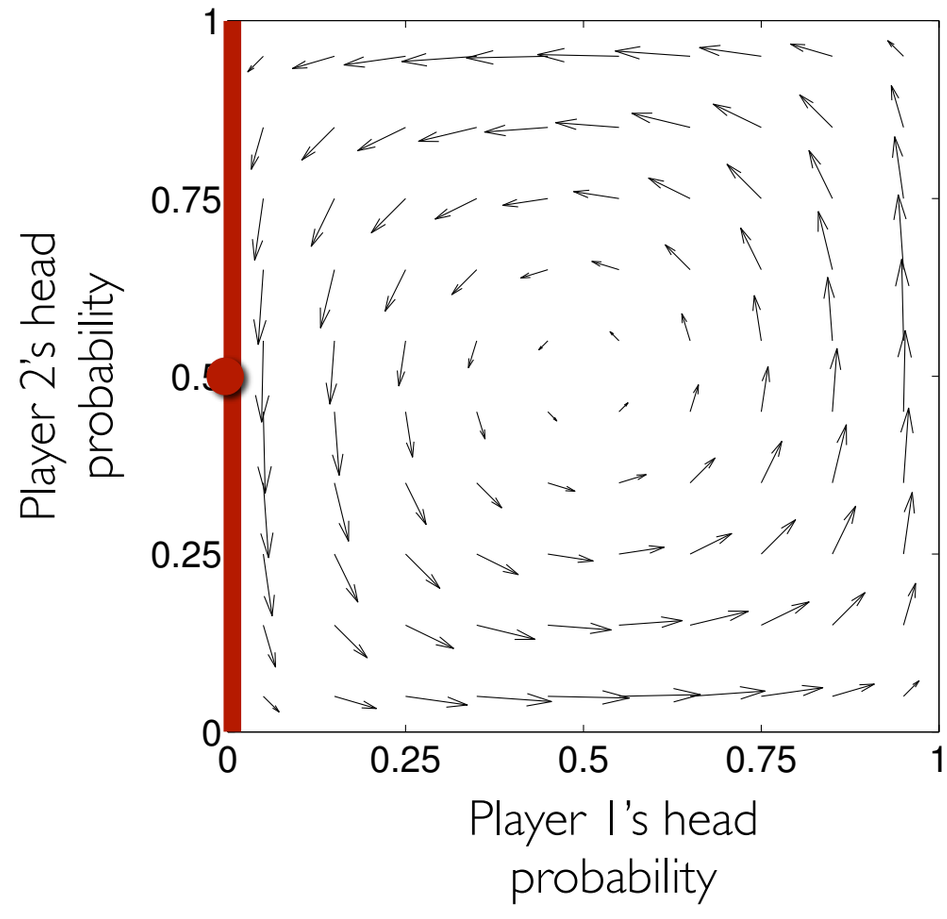
Player 1	Action 1	Action 2
Action 1	…, …	…, …
Action 2	…, …	…, …



Starting point

Player 2

Player 1	Action 1	Action 2
Action 1	…, …	…, …
Action 2	…, …	…, …



Single-population games

- There is a single population of individuals that is playing against itself
- An example: bacterials

Example

Single population

	$\sigma(\text{Type } 1)$	$\sigma(\text{Type } 2)$	$\sigma(\text{Type } 3)$
	Type 1	**Type 2**	**Type 3**
$\sigma(\text{Type } 1)$ **Type 1**	1	2	0
$\sigma(\text{Type } 2)$ **Type 2**	0	1	2
$\sigma(\text{Type } 3)$ **Type 3**	2	0	1

Single population

Convergence

Theorem. The replicator equation for a matrix game (single-population) satisfies:

1. A stable rest point is a NE

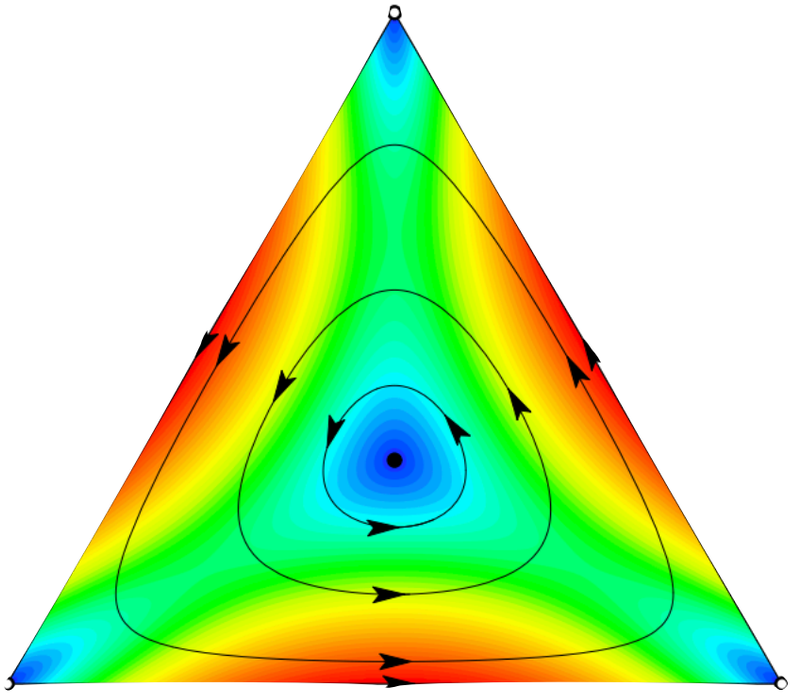
2. A convergent trajectory in the interior of the strategy space evolves to a NE

3. A strict NE is locally asymptotically stable

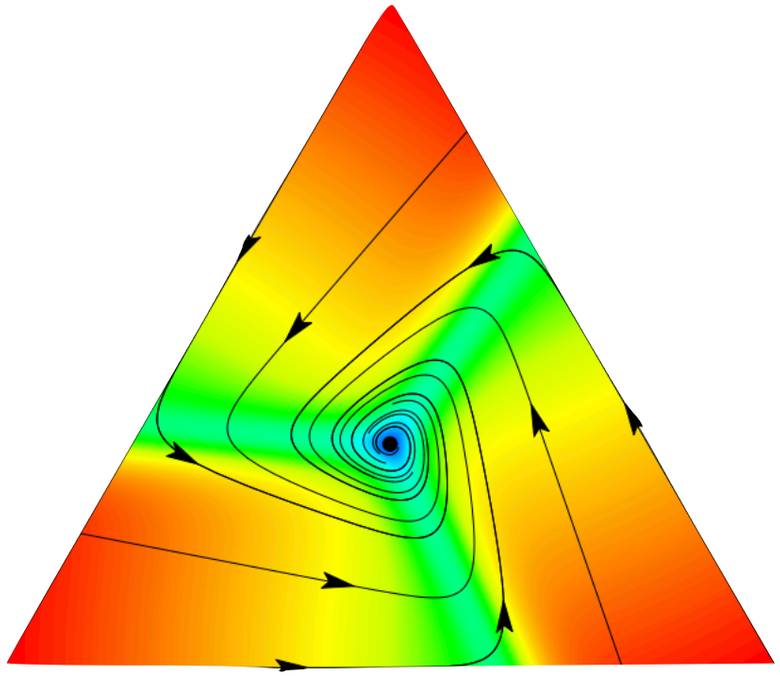
Multi-population games

- Every player is associated with a different population of individuals
- **Theorem**. (p*, q*) is a two-species ESS if and only if (p*, q*) is a strict Nash in a neighbourhood and
  - is a locally asymptotically stable rest point of the replicator equation
  - if it is in the interior, then it is a globally asymptotically stable rest point of the replicator equation

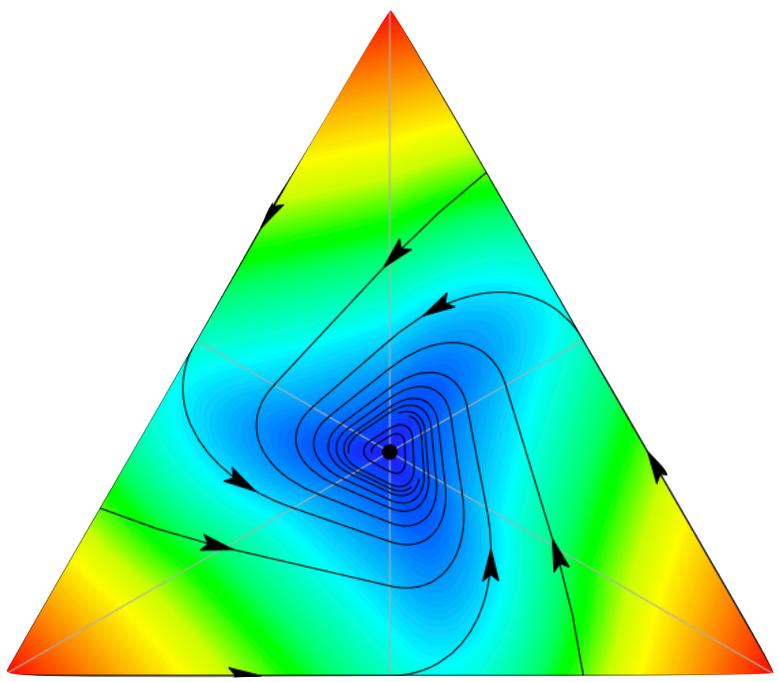
Known dynamics



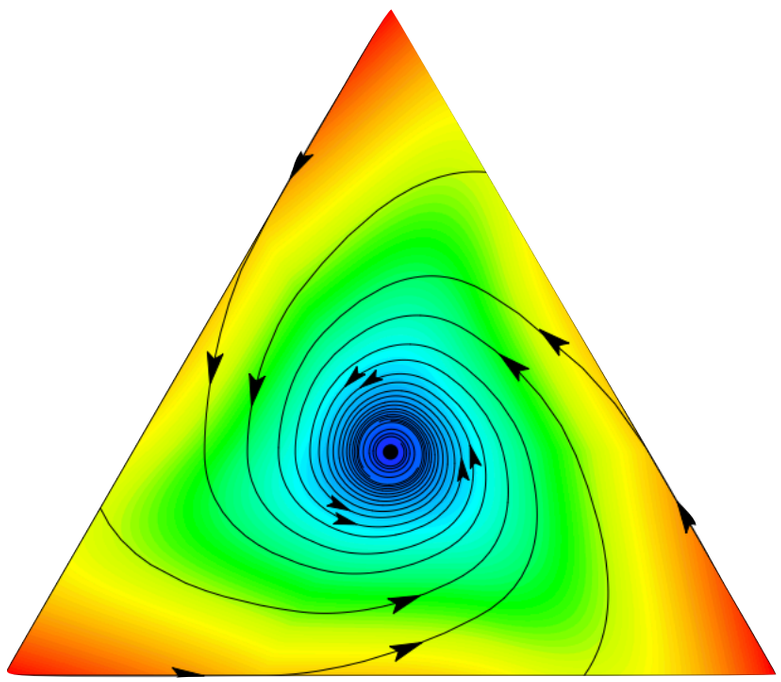
Replicator dynamics



Logit



BNN



Smith

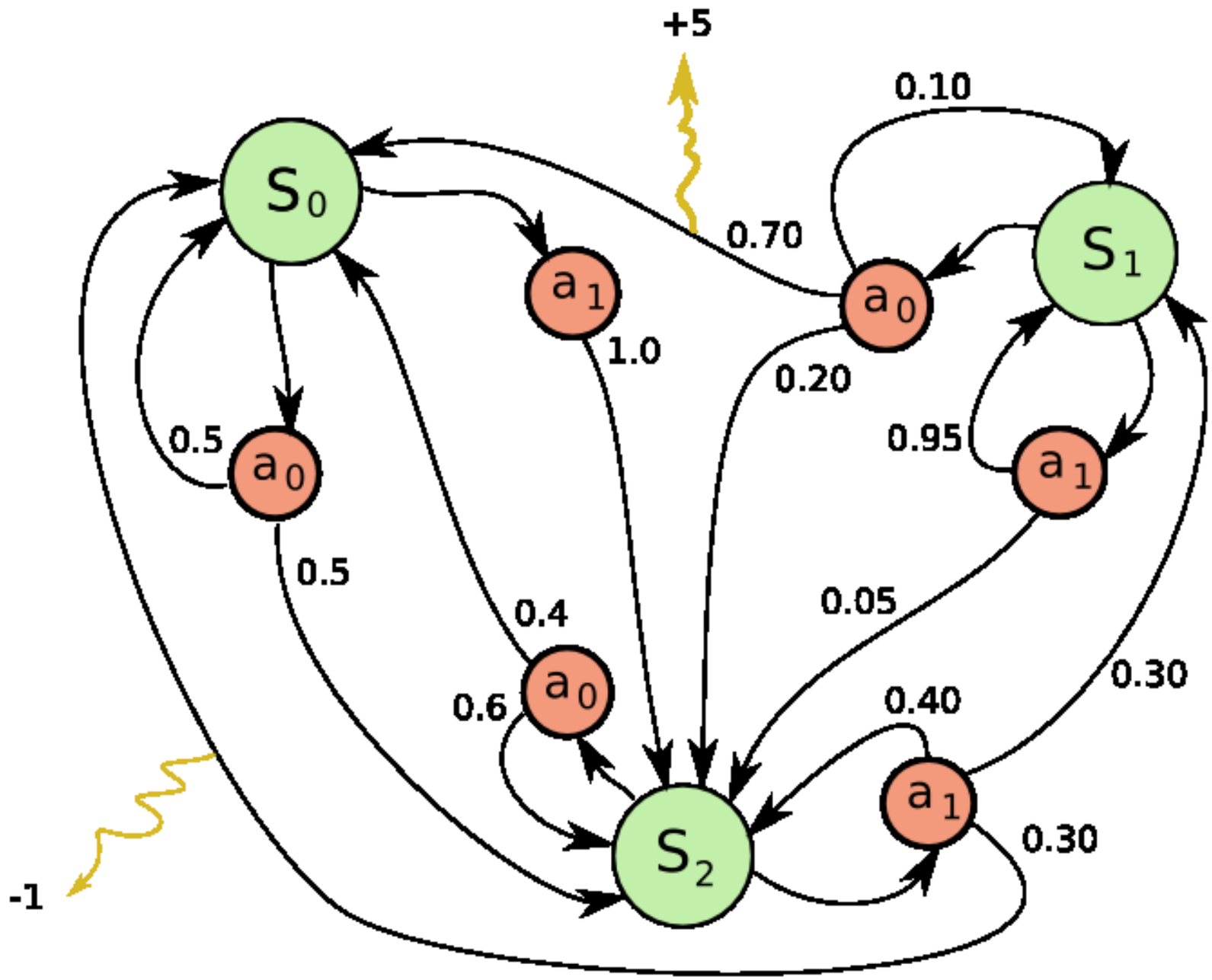
Properties

Dynamic	Family	(C)	(SD)	(NS)	(PC)
replicator	imitation	yes	yes	no	yes
logit	perturbed best response	yes	yes*	no	no
BNN	excess payoff	yes	no	yes	yes
Smith	pairwise comparison	yes	yes	yes	yes

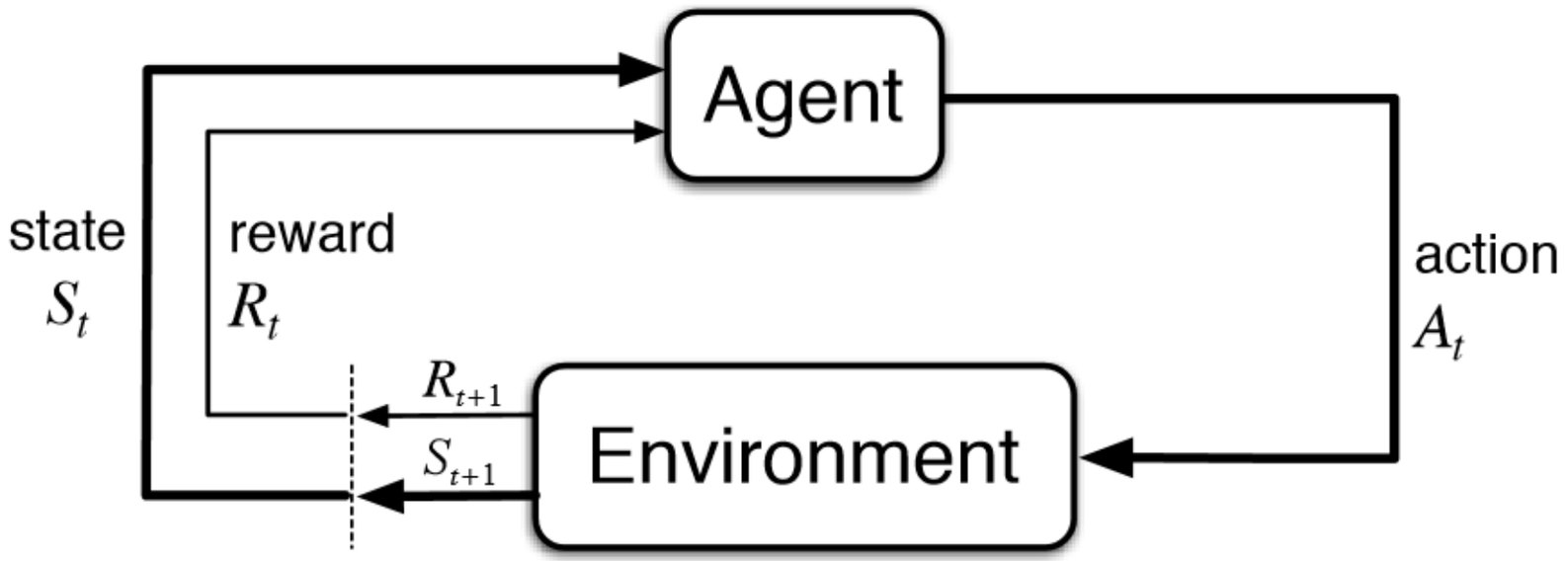
(C)	*Continuity*:	ρ is Lipschitz continuous
(SD)	*Scarcity of data*:	ρ_{ij} only depends on π_i, π_j, and x_j
(NS)	*Nash stationarity*:	$V^F(x) = \mathbf{0}$ if and only if $x \in NE(F)$
(PC)	*Positive correlation*:	$V^F(x) \neq \mathbf{0}$ implies that $V^F(x)'F(x) > 0$

Multi-agent learning

Markov decision problem



Reinforcement learning



Q-learning (I)

For every pair state/action:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s,a') - Q(s,a) \right]$$

learning rate

reward

discount factor

Example: normal-form games

Player 1
- 1 state
- 2 actions (Cooperate, Defeat)

Player 1 \ Player 2	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1

Example: normal-form games

$$Q(a) \leftarrow Q(a) + \alpha \Big(r - Q(a) \Big) \qquad \alpha = 0.2$$

$$\sigma_1(a) = \begin{cases} 1.0 & a = \text{Cooperate} \\ 0.0 & a = \text{Defeat} \end{cases}$$

$$\sigma_1(a) = \begin{cases} 0.2 & a = \text{Cooperate} \\ 0.8 & a = \text{Defeat} \end{cases}$$

		Player 2	
		Cooperate	**Defeat**
Player 1	**Cooperate**	3,3	0,5
	Defeat	5,0	1,1

Example: normal-form games

$$Q(a) \leftarrow Q(a) + \alpha\Big(r - Q(a)\Big) \qquad \alpha = 0.2$$

round	Player 2's action	Player 1's Q function
$t = 0$	—	$Q(\text{Cooperate}) = 0$
$t = 1$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.6$
$t = 2$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.48$
$t = 3$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.384$
$t = 4$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.3072$
$t = 5$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.24576$
$t = 6$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.496608$

$$\sigma_1(a) = \begin{cases} 1.0 & a = \text{Cooperate} \\ 0.0 & a = \text{Defeat} \end{cases}$$

$$\sigma_1(a) = \begin{cases} 0.2 & a = \text{Cooperate} \\ 0.8 & a = \text{Defeat} \end{cases}$$

Player 2

Player 1	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1

Example: normal-form games

$$Q(a) \leftarrow Q(a) + \alpha\Big(r - Q(a)\Big) \qquad \alpha = 0.2$$

round	Player 2's action	Player 1's Q function
$t = 0$	—	$Q(\text{Cooperate}) = 0$
$t = 1$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.6$
$t = 2$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.48$
$t = 3$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.384$
$t = 4$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.3072$
$t = 5$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.24576$
$t = 6$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.496608$

$$\sigma_1(a) = \begin{cases} 1.0 & a = \text{Cooperate} \\ 0.0 & a = \text{Defeat} \end{cases}$$

$$\sigma_1(a) = \begin{cases} 0.2 & a = \text{Cooperate} \\ 0.8 & a = \text{Defeat} \end{cases}$$

Player 2

Player 1	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1

Example: normal-form games

$$Q(a) \leftarrow Q(a) + \alpha\Big(r - Q(a)\Big) \qquad \alpha = 0.2$$

round	Player 2's action	Player 1's Q function
$t = 0$	—	$Q(\text{Cooperate}) = 0$
$t = 1$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.6$
$t = 2$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.48$
$t = 3$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.384$
$t = 4$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.3072$
$t = 5$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.24576$
$t = 6$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.496608$

$$\sigma_1(a) = \begin{cases} 1.0 & a = \text{Cooperate} \\ 0.0 & a = \text{Defeat} \end{cases}$$

$$\sigma_1(a) = \begin{cases} 0.2 & a = \text{Cooperate} \\ 0.8 & a = \text{Defeat} \end{cases}$$

Player 2

Player 1	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1

Example: normal-form games

$$Q(a) \leftarrow Q(a) + \alpha\Big(r - Q(a)\Big) \qquad \alpha = 0.2$$

round	Player 2's action	Player 1's Q function
$t = 0$	—	$Q(\text{Cooperate}) = 0$
$t = 1$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.6$
$t = 2$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.48$
$t = 3$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.384$
$t = 4$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.3072$
$t = 5$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.24576$
$t = 6$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.496608$

$$\sigma_1(a) = \begin{cases} 1.0 & a = \text{Cooperate} \\ 0.0 & a = \text{Defeat} \end{cases}$$

$$\sigma_1(a) = \begin{cases} 0.2 & a = \text{Cooperate} \\ 0.8 & a = \text{Defeat} \end{cases}$$

Player 2

Player 1	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1

Example: normal-form games

$$Q(a) \leftarrow Q(a) + \alpha\Big(r - Q(a)\Big) \qquad \alpha = 0.2$$

round	Player 2's action	Player 1's Q function
$t = 0$	—	$Q(\text{Cooperate}) = 0$
$t = 1$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.6$
$t = 2$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.48$
$t = 3$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.384$
$t = 4$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.3072$
$t = 5$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.24576$
$t = 6$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.496608$

$$\sigma_1(a) = \begin{cases} 1.0 & a = \text{Cooperate} \\ 0.0 & a = \text{Defeat} \end{cases}$$

$$\sigma_1(a) = \begin{cases} 0.2 & a = \text{Cooperate} \\ 0.8 & a = \text{Defeat} \end{cases}$$

Player 1 \ Player 2	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1

Example: normal-form games

$$Q(a) \leftarrow Q(a) + \alpha\Big(r - Q(a)\Big) \qquad \alpha = 0.2$$

round	Player 2's action	Player 1's Q function
$t = 0$	—	$Q(\text{Cooperate}) = 0$
$t = 1$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.6$
$t = 2$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.48$
$t = 3$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.384$
$t = 4$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.3072$
$t = 5$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.24576$
$t = 6$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.496608$

$$\sigma_1(a) = \begin{cases} 1.0 & a = \text{Cooperate} \\ 0.0 & a = \text{Defeat} \end{cases}$$

$$\sigma_1(a) = \begin{cases} 0.2 & a = \text{Cooperate} \\ 0.8 & a = \text{Defeat} \end{cases}$$

Player 1 \ Player 2	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1

Example: normal-form games

$$Q(a) \leftarrow Q(a) + \alpha\Big(r - Q(a)\Big) \qquad \alpha = 0.2$$

round	Player 2's action	Player 1's Q function
$t = 0$	—	$Q(\text{Cooperate}) = 0$
$t = 1$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.6$
$t = 2$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.48$
$t = 3$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.384$
$t = 4$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.3072$
$t = 5$	$a = \text{Defeat}$	$Q(\text{Cooperate}) = 0.24576$
$t = 6$	$a = \text{Cooperate}$	$Q(\text{Cooperate}) = 0.496608$

$$\sigma_1(a) = \begin{cases} 1.0 & a = \text{Cooperate} \\ 0.0 & a = \text{Defeat} \end{cases}$$

$$\sigma_1(a) = \begin{cases} 0.2 & a = \text{Cooperate} \\ 0.8 & a = \text{Defeat} \end{cases}$$

Player 2

Player 1	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1

Q-learning (2)

Softmax (a.k.a. Boltzam exploration)

$$\sigma_i(a) = \frac{\exp(Q(s,a)/\tau)}{\sum_{a'} \exp(Q(s,a')/\tau)}$$

Q-learning (2)

Softmax (a.k.a. Boltzam exploration)

$$\sigma_i(a) = \frac{\exp(Q(s,a)/\tau)}{\sum_{a'} \exp(Q(s,a')/\tau)}$$

temperature

Q-learning (2)

Softmax (a.k.a. Boltzam exploration)

$$\sigma_i(a) = \frac{\exp(Q(s,a)/\tau)}{\sum_{a'} \exp(Q(s,a')/\tau)}$$

temperature

Every action is played with strictly positive probability

The larger the temperature, the smoother the function

If the temperature is 0, we would have a best response

Example: normal-form games

$Q(\text{Cooperate})$	$Q(\text{Defeat})$	$\sigma_1(\text{Cooperate})$	$\sigma_1(\text{Cooperate})$
0	0	0.5	0.5
1	0	0.731	0.269
5	0	0.99331	0.00669
10	0	0.999955	0.000045

Example: normal-form games

$Q(\text{Cooperate})$	$Q(\text{Defeat})$	$\sigma_1(\text{Cooperate})$	$\sigma_1(\text{Cooperate})$
0	0	0.5	0.5
1	0	0.731	0.269
5	0	0.99331	0.00669
10	0	0.999955	0.000045

Example: normal-form games

$Q(\text{Cooperate})$	$Q(\text{Defeat})$	$\sigma_1(\text{Cooperate})$	$\sigma_1(\text{Cooperate})$
0	0	0.5	0.5
1	0	0.731	0.269
5	0	0.99331	0.00669
10	0	0.999955	0.000045

Example: normal-form games

$Q(\text{Cooperate})$	$Q(\text{Defeat})$	$\sigma_1(\text{Cooperate})$	$\sigma_1(\text{Cooperate})$
0	0	0.5	0.5
1	0	0.731	0.269
5	0	0.99331	0.00669
10	0	0.999955	0.000045

Self-play Q-learning dynamics

Self-play learning

Q-learning algorithm

Q-learning algorithm

Player 1 \ Player 2	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1

Learning dynamics

Assumptions:

- Time is continuous
- All the actions can be selected simultaneously

Learning dynamics

Assumptions:

- Time is continuous
- All the actions can be selected simultaneously

$$\dot{\sigma}_1(a,t) = \frac{\alpha\,\sigma_1(a,t)}{\tau}\Big(e_a U_1 \sigma_2(t) - \sigma_1(t) U_1 \sigma_2(t)\Big) - \alpha\,\sigma_1(a,t)\Big(\log(\sigma_1(a)) - \sum_{a'}\sigma_1(a')\,\log(\sigma_1(a'))\Big)$$

Learning dynamics

Assumptions:

- Time is continuous
- All the actions can be selected simultaneously

$$\dot{\sigma}_1(a,t) = \frac{\alpha\,\sigma_1(a,t)}{\tau}\Big(e_a U_1 \sigma_2(t) - \sigma_1(t) U_1 \sigma_2(t)\Big) - \alpha\,\sigma_1(a,t)\Big(\log(\sigma_1(a)) - \sum_{a'}\sigma_1(a')\,\log(\sigma_1(a'))\Big)$$

exploitation term

exploration term

Learning dynamics

Assumptions:

- Time is continuous
- All the actions can be selected simultaneously

$$\dot{\sigma}_1(a,t) = \frac{\alpha\,\sigma_1(a,t)}{\tau}\Big(e_a U_1\sigma_2(t) - \sigma_1(t)U_1\sigma_2(t)\Big) - \alpha\,\sigma_1(a,t)\Big(\log(\sigma_1(a)) - \sum_{a'}\sigma_1(a')\,\log(\sigma_1(a'))\Big)$$

exploitation term

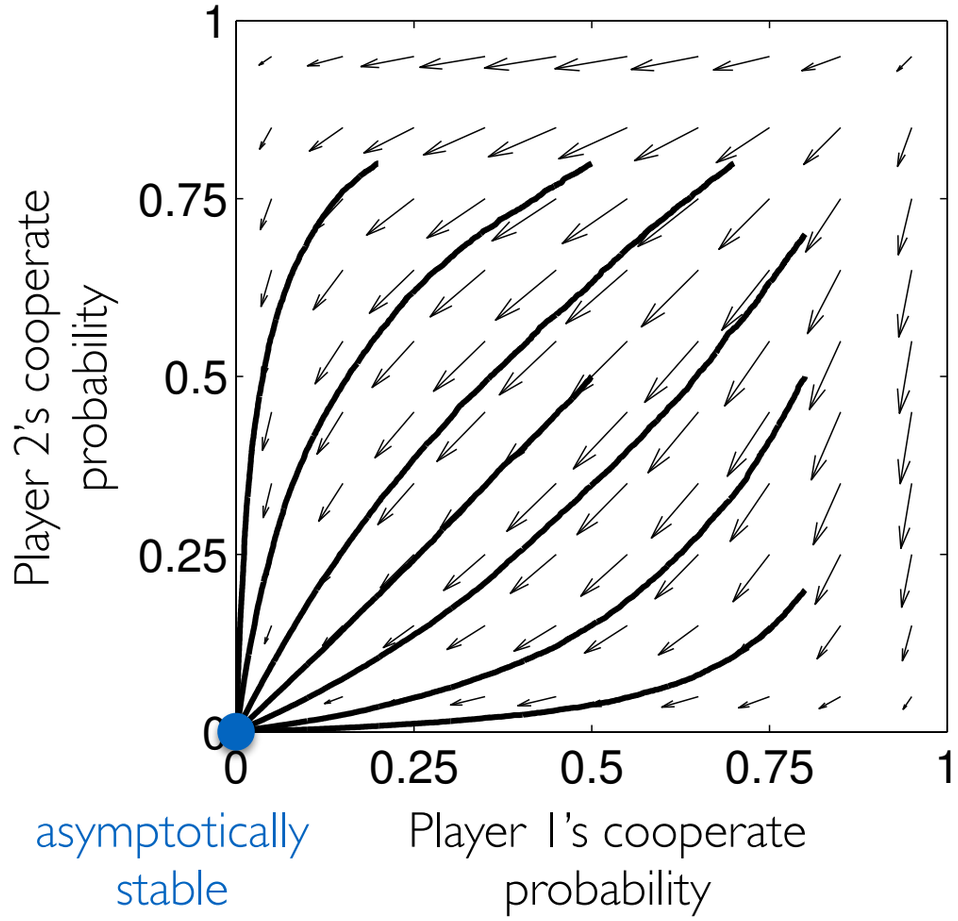
exploration term

When the temperature is 0, the Q-learning behaves as the replicator dynamics

Prisoner's dilemma

Player 2

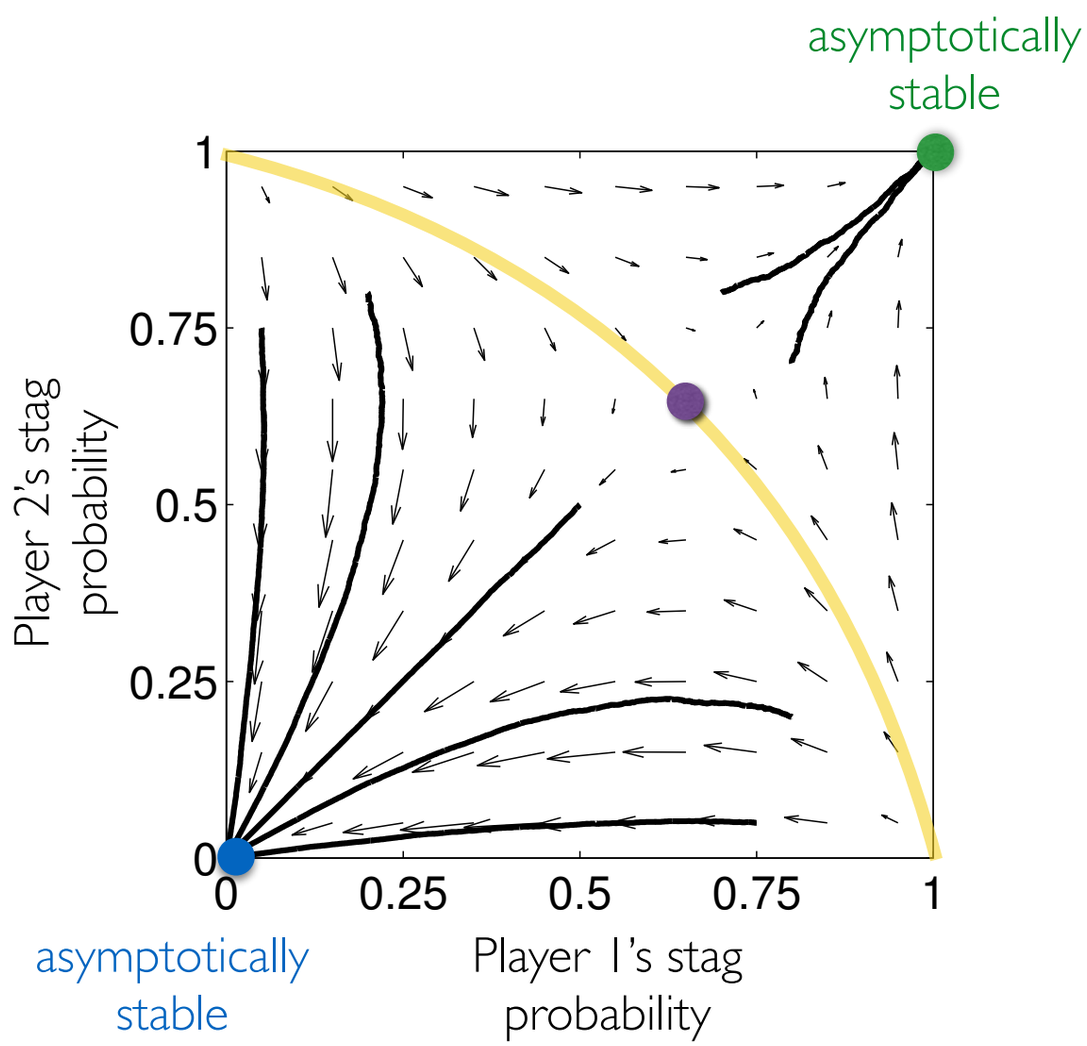
Player 1	**Cooperate**	**Defeat**
Cooperate	3,3	0,5
Defeat	5,0	1,1



Stag hunt

Player 2

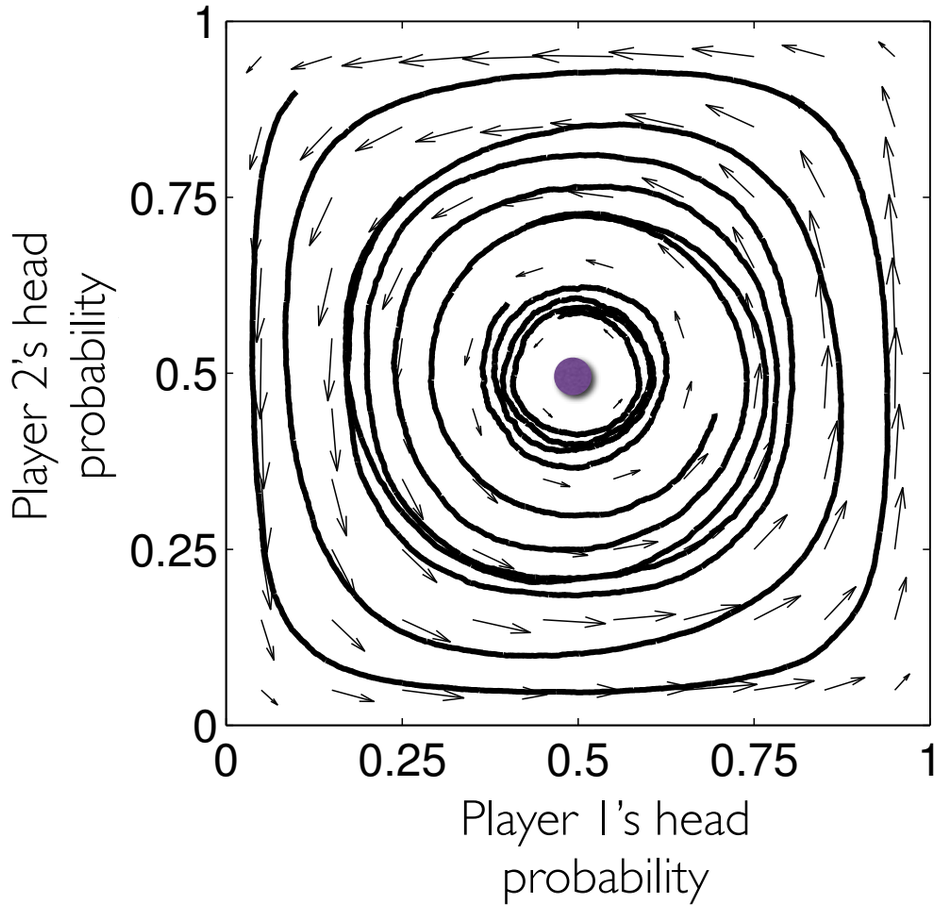
Player 1	Stag	Hare
Stag	4,4	1,3
Hare	3,1	3,3



Matching pennies

Player 2

Player 1	Head	Tail
Head	0,1	1,0
Tail	1,0	0,1



Other MAL algorithms dynamics based on replicator equation

	Algorithm	$\dot{x}$
Infinitesimal Gradient Ascent	IGA	$\alpha\partial$
IGA-Win or Learn Fast	IGA-WoLF	$\partial \cdot \begin{cases} \alpha_{min} & \text{if } V(\mathbf{x}) > V(\mathbf{x}^*) \\ \alpha_{max} & \text{otherwise} \end{cases}$
Weighted Policy Learning	WPL	$\alpha\partial \cdot \begin{cases} x & \text{if } \partial < 0 \\ (1-x) & \text{otherwise} \end{cases}$
Cross Learning	CL	$x(1-x)\ \partial$
Frequency-Adjusted Q-learning	FAQ	$\alpha x(1-x)\left[\partial \cdot \tau^{-1} - \log \frac{x}{1-x}\right]$
Regret Minimization (Polynomial Weights)	RM	$\alpha x(1-x)\ \partial \cdot \begin{cases} (1+\alpha x\partial)^{-1} & \text{if } \partial < 0 \\ (1-\alpha(1-x)\partial)^{-1} & \text{otherwise} \end{cases}$