Multi-agent learning

Reinforcement Learning

Gerard Vreeswijk, Intelligent Software Systems, Computer Science Department, Faculty of Sciences, Utrecht University, The Netherlands.

Monday 4th May, 2020

Reinforcement learning: motivation

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.
- A dynamic analysis:

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.
- A dynamic analysis: How do (or should) players develop their strategies and behaviour in a repeated game?

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.
- A dynamic analysis: How do (or should) players develop their strategies and behaviour in a repeated game? "Do": descriptive (economics);

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.
- A dynamic analysis: How do (or should) players develop their strategies and behaviour in a repeated game? "Do": descriptive (economics); "should": prescriptive (computer science).

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.
- A dynamic analysis: How do (or should) players develop their strategies and behaviour in a repeated game? "Do": descriptive (economics); "should": prescriptive (computer science).
- Reinforcement learning (RL) is a rudimentary learning technique.

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.
- A dynamic analysis: How do (or should) players develop their strategies and behaviour in a repeated game? “Do”: descriptive (economics); “should”: prescriptive (computer science).
- Reinforcement learning (RL) is a rudimentary learning technique.

  1. RL is stimulus-response: it plays actions with the highest past payoff.

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.
- A dynamic analysis: How do (or should) players develop their strategies and behaviour in a repeated game? "Do": descriptive (economics); "should": prescriptive (computer science).
- Reinforcement learning (RL) is a rudimentary learning technique.

  1. RL is stimulus-response: it plays actions with the highest past payoff.
  2. It is myopic: it is only interested in immediate success.

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.
- A dynamic analysis: How do (or should) players develop their strategies and behaviour in a repeated game? “Do”: descriptive (economics); “should”: prescriptive (computer science).
- Reinforcement learning (RL) is a rudimentary learning technique.
  1. RL is stimulus-response: it plays actions with the highest past payoff.
  2. It is myopic: it is only interested in immediate success.
- Reinforcement learning can be applied to learning in games.

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.
- A dynamic analysis: How do (or should) players develop their strategies and behaviour in a repeated game? "Do": descriptive (economics); "should": prescriptive (computer science).
- Reinforcement learning (RL) is a rudimentary learning technique.

  1. RL is stimulus-response: it plays actions with the highest past payoff.
  2. It is myopic: it is only interested in immediate success.

- Reinforcement learning can be applied to learning in games.
- When computer scientists mention RL, they usually mean multi-state RL,

Reinforcement learning: motivation

- Nash equilibria in repeated games is a static analysis.
- A dynamic analysis: How do (or should) players develop their strategies and behaviour in a repeated game? "Do": descriptive (economics); "should": prescriptive (computer science).
- Reinforcement learning (RL) is a rudimentary learning technique.
  1. RL is stimulus-response: it plays actions with the highest past payoff.
  2. It is myopic: it is only interested in immediate success.
- Reinforcement learning can be applied to learning in games.
- When computer scientists mention RL, they usually mean multi-state RL, but single-state RL has already interesting and theoretically important properties, especially with game theory.

Plan for today

Part I: Single-state RL in games. First half of Ch. 2 of Peyton Young (2004): "Reinforcement and Regret".

Plan for today

Part I: Single-state RL in games. First half of Ch. 2 of Peyton Young (2004): "Reinforcement and Regret".

1. By average: $\dfrac{r_1 + \cdots + r_n}{n}$.

Plan for today

Part I: Single-state RL in games. First half of Ch. 2 of Peyton Young (2004): “Reinforcement and Regret”.

1. By average: $\dfrac{r_1 + \cdots + r_n}{n}$.
2. With a discounted past : $\gamma^{n-1} r_1 + \gamma^{n-2} r_2 + \cdots + \gamma r_{n-1} + r_n$.

Plan for today

Part I: Single-state RL in games. First half of Ch. 2 of Peyton Young (2004): "Reinforcement and Regret".

1. By average: $\dfrac{r_1 + \cdots + r_n}{n}$.
2. With a discounted past : $\gamma^{n-1}r_1 + \gamma^{n-2}r_2 + \cdots + \gamma r_{n-1} + r_n$.
3. With an aspiration level (Sutton *et al.*: "reference reward").

Plan for today

Part I: Single-state RL in games. First half of Ch. 2 of Peyton Young (2004): "Reinforcement and Regret".

1. By average: $\dfrac{r_1 + \cdots + r_n}{n}$.
2. With a discounted past : $\gamma^{n-1}r_1 + \gamma^{n-2}r_2 + \cdots + \gamma r_{n-1} + r_n$.
3. With an aspiration level (Sutton *et al.*: "reference reward").

Part II: Convergence to dominant strategies. Begin of Beggs (2005): "On the Convergence of Reinforcement Learning".

	#Players	*#Actions*	*Result*
☞ Theorem 1 :	1	2	$\Pr(\text{dominant action}) = 1$
Theorem 2 :	1	≥ 2	$\Pr(\text{sub-dominant actions}) = 0$
Theorem 3 :	≥ 1	≥ 2	$\Pr(\text{dom}) = 1$, $\Pr(\text{sub-dom}) = 0$

Part I:
Single-state reinforcement learning

Part I: Single-state reinforcement learning in games

Proportional techniques: basic setup

Proportional techniques: basic setup

- There are two players: A (the protagonist) and B (the antagonist, sometimes "nature").

Proportional techniques: basic setup

- There are two players: A (the protagonist) and B (the antagonist, sometimes “nature”).
- Play proceeds in (usually an indefinite number of) rounds

$$1, \dots, t, \dots .$$

Proportional techniques: basic setup

- There are two players: A (the protagonist) and B (the antagonist, sometimes "nature").
- Play proceeds in (usually an indefinite number of) rounds

$$1, \ldots, t, \ldots.$$

- Identifiers X and Y denote finite sets of possible actions.

Proportional techniques: basic setup

- There are two players: A (the protagonist) and B (the antagonist, sometimes "nature").
- Play proceeds in (usually an indefinite number of) rounds

$$1, \dots, t, \dots.$$

- Identifiers X and Y denote finite sets of possible actions.
- Each round, t, players A and B choose actions $x \in X$ and $y \in Y$, respectively:

$$(x^1, y^1), (x^2, y^2), \dots, (x^t, y^t), \dots.$$

Proportional techniques: basic setup

- There are two players: A (the protagonist) and B (the antagonist, sometimes "nature").
- Play proceeds in (usually an indefinite number of) rounds

$$1, \ldots, t, \ldots.$$

- Identifiers X and Y denote finite sets of possible actions.
- Each round, t, players A and B choose actions $x \in X$ and $y \in Y$, respectively:

$$(x^1, y^1), (x^2, y^2), \ldots, (x^t, y^t), \ldots.$$

- A's payoff is given by a fixed non-negative function

$$u : X \times Y \to R_0^+.$$

Proportional techniques: basic setup

- There are two players: A (the protagonist) and B (the antagonist, sometimes "nature").
- Play proceeds in (usually an indefinite number of) rounds

$$1, \dots, t, \dots.$$

- Identifiers X and Y denote finite sets of possible actions.
- Each round, t, players A and B choose actions $x \in X$ and $y \in Y$, respectively:

$$(x^1, y^1), (x^2, y^2), \dots, (x^t, y^t), \dots.$$

- A's payoff is given by a fixed non-negative function

$$u : X \times Y \to R_0^+.$$

A's payoff matrix is known.

Proportional techniques: basic setup

- There are two players: A (the protagonist) and B (the antagonist, sometimes "nature").
- Play proceeds in (usually an indefinite number of) rounds

$$1, \dots, t, \dots .$$

- Identifiers X and Y denote finite sets of possible actions.
- Each round, t, players A and B choose actions $x \in X$ and $y \in Y$, respectively:

$$(x^1, y^1), (x^2, y^2), \dots, (x^t, y^t), \dots .$$

- A's payoff is given by a fixed non-negative function

$$u : X \times Y \to R_0^+.$$

A's payoff matrix is known.

- It follows that payoffs are time homogeneous, i.e.,

$$(x^s, y^s) = (x^t, y^t) \quad \Rightarrow \quad u(x^s, y^s) = u(x^t, y^t).$$

Propensity, and mixed strategy of play

Propensity, and mixed strategy of play

- Let $t \geq 0$. The propensity of A to play x at t is denoted by θ_x^t.

Propensity, and mixed strategy of play

- Let $t \geq 0$. The propensity of A to play x at t is denoted by θ_x^t.
- A simple model is cumulative payoff matching (CPM):

$$\theta_x^{t+1} = \begin{cases} \theta_x^t + u(x, y) & \text{if } x \text{ is played at round } t, \\ \theta_x^t & \text{else.} \end{cases}$$

Propensity, and mixed strategy of play

- Let $t \geq 0$. The propensity of A to play x at t is denoted by θ_x^t.
- A simple model is cumulative payoff matching (CPM):

$$\theta_x^{t+1} = \begin{cases} \theta_x^t + u(x, y) & \text{if } x \text{ is played at round } t, \\ \theta_x^t & \text{else.} \end{cases}$$

- As a vector: $\theta^{t+1} = \theta^t + u^t e^t$, where $e_x^t =_{Def} x$ is played at t ? $1 : 0$.

Propensity, and mixed strategy of play

- Let $t \geq 0$. The propensity of A to play x at t is denoted by θ_x^t.
- A simple model is cumulative payoff matching (CPM):

$$\theta_x^{t+1} = \begin{cases} \theta_x^t + u(x,y) & \text{if } x \text{ is played at round } t, \\ \theta_x^t & \text{else.} \end{cases}$$

- As a vector: $\theta^{t+1} = \theta^t + u^t e^t$, where $e_x^t =_{Def} x$ is played at t ? $1 : 0$.
- The vector of initial propensities, θ^0 is not the result of play.

Propensity, and mixed strategy of play

- Let $t \geq 0$. The propensity of A to play x at t is denoted by θ_x^t.
- A simple model is cumulative payoff matching (CPM):

$$\theta_x^{t+1} = \begin{cases} \theta_x^t + u(x,y) & \text{if } x \text{ is played at round } t, \\ \theta_x^t & \text{else.} \end{cases}$$

- As a vector: $\theta^{t+1} = \theta^t + u^t e^t$, where $e_x^t =_{Def} x$ is played at t ? $1 : 0$.
- The vector of initial propensities, θ^0 is not the result of play.
- A possible mixed strategy to play at round t is to randomise on the normalised propensity of x at t:

$$(q_x^t)_{x \in X}, \text{ where } \quad q_x^t =_{Def} \frac{\theta_x^t}{\sum_{x' \in X} \theta_{x'}^t}.$$

An example

The total accumulated payoff at round t, the sum $\sum_{x \in X} \theta_x^t$, is abbreviated by v^t.

Rounds :	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	θ^{14}	
Payoff x_1 :	1	8	3	.	.	.	7	4	.	1	.	.	.	1	.	25	θ_1^{14}
Payoff x_2 :	1	.	.	6	.	5	.	.	.	.	6	.	.	.	8	24	θ_2^{14}
Payoff x_3 :	1	.	.	.	9	.	.	.	9	.	.	2	2	.	.	23	θ_3^{14}
																72	

An example

The total accumulated payoff at round t, the sum $\sum_{x \in X} \theta_x^t$, is abbreviated by v^t.

Rounds :	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	θ^{14}	
Payoff x_1 :	1	8	3	.	.	.	7	4	.	1	.	.	.	1	.	25	θ_1^{14}
Payoff x_2 :	1	.	.	6	.	5	.	.	.	.	6	.	.	.	8	24	θ_2^{14}
Payoff x_3 :	1	.	.	.	9	.	.	.	9	.	.	2	2	.	.	23	θ_3^{14}
																72	

Remarks:

An example

The total accumulated payoff at round t, the sum $\sum_{x \in X} \theta_x^t$, is abbreviated by v^t.

Rounds :	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	θ^{14}	
Payoff x_1 :	1	8	3	.	.	.	7	4	.	1	.	.	.	1	.	25	θ_1^{14}
Payoff x_2 :	1	.	.	6	.	5	.	.	.	.	6	.	.	.	8	24	θ_2^{14}
Payoff x_3 :	1	.	.	.	9	.	.	.	9	.	.	2	2	.	.	23	θ_3^{14}
																72	

Remarks:

- Here, $v^t = 72$.

An example

The total accumulated payoff at round t, the sum $\sum_{x \in X} \theta_x^t$, is abbreviated by v^t.

Rounds :	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	θ^{14}	
Payoff x_1 :	1	8	3	.	.	.	7	4	.	1	.	.	.	1	.	25	θ_1^{14}
Payoff x_2 :	1	.	.	6	.	5	.	.	.	.	6	.	.	.	8	24	θ_2^{14}
Payoff x_3 :	1	.	.	.	9	.	.	.	9	.	.	2	2	.	.	23	θ_3^{14}
																72	

Remarks:

- Here, $v^t = 72$. Alternative formula: $v^t = \sum_{x \in X} \theta_x^0 + \sum_{s \leq t} u^t$.

An example

The total accumulated payoff at round t, the sum $\sum_{x \in X} \theta_x^t$, is abbreviated by v^t.

Rounds :	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	θ^{14}	
Payoff x_1 :	1	8	3	.	.	.	7	4	.	1	.	.	.	1	.	25	θ_1^{14}
Payoff x_2 :	1	.	.	6	.	5	.	.	.	.	6	.	.	.	8	24	θ_2^{14}
Payoff x_3 :	1	.	.	.	9	.	.	.	9	.	.	2	2	.	.	23	θ_3^{14}
																72	

Remarks:

- Here, $v^t = 72$. Alternative formula: $v^t = \sum_{x \in X} \theta_x^0 + \sum_{s \leq t} u^t$.
- It is the cumulative payoff for each action that matters, not the average payoff. (There is a difference.)

An example

The total accumulated payoff at round t, the sum $\sum_{x \in X} \theta_x^t$, is abbreviated by v^t.

Rounds :	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	θ^{14}	
Payoff x_1 :	1	8	3	.	.	.	7	4	.	1	.	.	.	1	.	25	θ_1^{14}
Payoff x_2 :	1	.	.	6	.	5	.	.	.	.	6	.	.	.	8	24	θ_2^{14}
Payoff x_3 :	1	.	.	.	9	.	.	.	9	.	.	2	2	.	.	23	θ_3^{14}
																72	

Remarks:

- Here, $v^t = 72$. Alternative formula: $v^t = \sum_{x \in X} \theta_x^0 + \sum_{s \leq t} u^t$.
- It is the cumulative payoff for each action that matters, not the average payoff. (There is a difference.)
- In this example, it is assumed that the initial propensities, θ_x^0, are one. In general, they could be anything. But $\|\theta^0\| = 0$ is forbidden.

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$\Delta q_x^t$$

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$\Delta q_x^t = q_x^t - q_x^{t-1}$$

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$\Delta q_x^t = q_x^t - q_x^{t-1} = \frac{\theta_x^t}{v^t} - \frac{\theta_x^{t-1}}{v^{t-1}}$$

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$\Delta q_x^t = q_x^t - q_x^{t-1} = \frac{\theta_x^t}{v^t} - \frac{\theta_x^{t-1}}{v^{t-1}}$$
$$= \frac{v^{t-1} \cdot \theta_x^t}{v^{t-1} \cdot v^t} - \frac{v^t \cdot \theta_x^{t-1}}{v^t \cdot v^{t-1}}$$

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$\Delta q_x^t = q_x^t - q_x^{t-1} = \frac{\theta_x^t}{v^t} - \frac{\theta_x^{t-1}}{v^{t-1}}$$
$$= \frac{v^{t-1} \cdot \theta_x^t}{v^{t-1} \cdot v^t} - \frac{v^t \cdot \theta_x^{t-1}}{v^t \cdot v^{t-1}} = \frac{v^{t-1} \cdot \theta_x^t - v^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t}$$

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$
\begin{aligned}
\Delta q_x^t = q_x^t - q_x^{t-1} &= \frac{\theta_x^t}{v^t} - \frac{\theta_x^{t-1}}{v^{t-1}} \\
&= \frac{v^{t-1} \cdot \theta_x^t}{v^{t-1} \cdot v^t} - \frac{v^t \cdot \theta_x^{t-1}}{v^t \cdot v^{t-1}} = \frac{v^{t-1} \cdot \theta_x^t - v^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{v^{t-1} \cdot (\theta_x^{t-1} + e_x^t \cdot u^t) - (v^{t-1} + u^t) \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t}
\end{aligned}
$$

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$
\begin{aligned}
\Delta q_x^t = q_x^t - q_x^{t-1} &= \frac{\theta_x^t}{v^t} - \frac{\theta_x^{t-1}}{v^{t-1}} \\
&= \frac{v^{t-1} \cdot \theta_x^t}{v^{t-1} \cdot v^t} - \frac{v^t \cdot \theta_x^{t-1}}{v^t \cdot v^{t-1}} = \frac{v^{t-1} \cdot \theta_x^t - v^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{v^{t-1} \cdot (\theta_x^{t-1} + e_x^t \cdot u^t) - (v^{t-1} + u^t) \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{\cancel{v^{t-1} \cdot \theta_x^{t-1}} + v^{t-1} \cdot e_x^t \cdot u^t - \cancel{v^{t-1} \cdot \theta_x^{t-1}} - u^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t}
\end{aligned}
$$

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$\begin{aligned}
\Delta q_x^t = q_x^t - q_x^{t-1} &= \frac{\theta_x^t}{v^t} - \frac{\theta_x^{t-1}}{v^{t-1}} \\
&= \frac{v^{t-1} \cdot \theta_x^t}{v^{t-1} \cdot v^t} - \frac{v^t \cdot \theta_x^{t-1}}{v^t \cdot v^{t-1}} = \frac{v^{t-1} \cdot \theta_x^t - v^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{v^{t-1} \cdot (\theta_x^{t-1} + e_x^t \cdot u^t) - (v^{t-1} + u^t) \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{\cancel{v^{t-1} \cdot \theta_x^{t-1}} + v^{t-1} \cdot e_x^t \cdot u^t - \cancel{v^{t-1} \cdot \theta_x^{t-1}} - u^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{v^{t-1} \cdot e_x^t \cdot u^t - u^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t}
\end{aligned}$$

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$
\begin{aligned}
\Delta q_x^t = q_x^t - q_x^{t-1} &= \frac{\theta_x^t}{v^t} - \frac{\theta_x^{t-1}}{v^{t-1}} \\
&= \frac{v^{t-1} \cdot \theta_x^t}{v^{t-1} \cdot v^t} - \frac{v^t \cdot \theta_x^{t-1}}{v^t \cdot v^{t-1}} = \frac{v^{t-1} \cdot \theta_x^t - v^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{v^{t-1} \cdot (\theta_x^{t-1} + e_x^t \cdot u^t) - (v^{t-1} + u^t) \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{\cancel{v^{t-1} \cdot \theta_x^{t-1}} + v^{t-1} \cdot e_x^t \cdot u^t - \cancel{v^{t-1} \cdot \theta_x^{t-1}} - u^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{v^{t-1} \cdot e_x^t \cdot u^t - u^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} = \frac{u^t}{v^t} \frac{v^{t-1} \cdot e_x^t - \theta_x^{t-1}}{v^{t-1}}
\end{aligned}
$$

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$
\begin{aligned}
\Delta q_x^t &= q_x^t - q_x^{t-1} = \frac{\theta_x^t}{v^t} - \frac{\theta_x^{t-1}}{v^{t-1}} \\
&= \frac{v^{t-1} \cdot \theta_x^t}{v^{t-1} \cdot v^t} - \frac{v^t \cdot \theta_x^{t-1}}{v^t \cdot v^{t-1}} = \frac{v^{t-1} \cdot \theta_x^t - v^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{v^{t-1} \cdot (\theta_x^{t-1} + e_x^t \cdot u^t) - (v^{t-1} + u^t) \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{\cancel{v^{t-1} \cdot \theta_x^{t-1}} + v^{t-1} \cdot e_x^t \cdot u^t - \cancel{v^{t-1} \cdot \theta_x^{t-1}} - u^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{v^{t-1} \cdot e_x^t \cdot u^t - u^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} = \frac{u^t}{v^t} \frac{v^{t-1} \cdot e_x^t - \theta_x^{t-1}}{v^{t-1}} \\
&= \frac{u^t}{v^t} (e_x^t - \frac{\theta_x^{t-1}}{v^{t-1}})
\end{aligned}
$$

Dynamics of the mixed strategy

We can obtain further insight in the dynamics of the process by considering the change of the mixed strategy:

$$
\begin{aligned}
\Delta q_x^t = q_x^t - q_x^{t-1} &= \frac{\theta_x^t}{v^t} - \frac{\theta_x^{t-1}}{v^{t-1}} \\
&= \frac{v^{t-1} \cdot \theta_x^t}{v^{t-1} \cdot v^t} - \frac{v^t \cdot \theta_x^{t-1}}{v^t \cdot v^{t-1}} = \frac{v^{t-1} \cdot \theta_x^t - v^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{v^{t-1} \cdot (\theta_x^{t-1} + e_x^t \cdot u^t) - (v^{t-1} + u^t) \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{\cancel{v^{t-1} \cdot \theta_x^{t-1}} + v^{t-1} \cdot e_x^t \cdot u^t - \cancel{v^{t-1} \cdot \theta_x^{t-1}} - u^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} \\
&= \frac{v^{t-1} \cdot e_x^t \cdot u^t - u^t \cdot \theta_x^{t-1}}{v^{t-1} \cdot v^t} = \frac{u^t}{v^t} \frac{v^{t-1} \cdot e_x^t - \theta_x^{t-1}}{v^{t-1}} \\
&= \frac{u^t}{v^t}\left(e_x^t - \frac{\theta_x^{t-1}}{v^{t-1}}\right) = \frac{u^t}{v^t}\left(e_x^t - q_x^{t-1}\right).
\end{aligned}
$$

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q_x^t = \frac{u^t}{v^t}(e_x^t - q_x^{t-1}).$$

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\|\Delta q^t\|$$

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\|\Delta q^t\| = \|\frac{u^t}{v^t}(e^t - q^{t-1})\|$$

---

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\|\Delta q^t\| = \|\frac{u^t}{v^t}(e^t - q^{t-1})\| = \frac{u^t}{v^t} \cdot \|e^t - q^{t-1}\|$$

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\|\Delta q^t\| = \|\frac{u^t}{v^t}(e^t - q^{t-1})\| = \frac{u^t}{v^t} \cdot \|e^t - q^{t-1}\| \leq \frac{u^t}{v^t} \cdot 2$$

---

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\begin{aligned} \|\Delta q^t\| &= \|\frac{u^t}{v^t}(e^t - q^{t-1})\| = \frac{u^t}{v^t} \cdot \|e^t - q^{t-1}\| \leq \frac{u^t}{v^t} \cdot 2 \\ &= \frac{u^t}{u_1 + \cdots + u_t} \cdot 2 \end{aligned}$$

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\begin{aligned} \|\Delta q^t\| &= \|\frac{u^t}{v^t}(e^t - q^{t-1})\| = \frac{u^t}{v^t} \cdot \|e^t - q^{t-1}\| \leq \frac{u^t}{v^t} \cdot 2 \\ &= \frac{u^t}{u_1 + \cdots + u_t} \cdot 2 \leq \frac{\max\{u^s \mid s \leq t\}}{t \cdot \min\{u^s \mid s \leq t\}} \cdot 2 \end{aligned}$$

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\begin{aligned}\|\Delta q^t\| &= \|\frac{u^t}{v^t}(e^t - q^{t-1})\| = \frac{u^t}{v^t} \cdot \|e^t - q^{t-1}\| \leq \frac{u^t}{v^t} \cdot 2 \\ &= \frac{u^t}{u_1 + \cdots + u_t} \cdot 2 \leq \frac{\max\{u^s \mid s \leq t\}}{t \cdot \min\{u^s \mid s \leq t\}} \cdot 2 = \frac{2}{t}.\end{aligned}$$

---

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \ldots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\begin{aligned}\|\Delta q^t\| = \|\frac{u^t}{v^t}(e^t - q^{t-1})\| = \frac{u^t}{v^t}\cdot \|e^t - q^{t-1}\| &\leq \frac{u^t}{v^t}\cdot 2 \\ = \frac{u^t}{u_1 + \cdots + u_t}\cdot 2 \leq \frac{\max\{u^s \mid s \leq t\}}{t\cdot \min\{u^s \mid s \leq t\}}\cdot 2 &= \frac{2}{t}.\end{aligned}$$

So $\lim_{t\to\infty} \|\Delta q^t\| = 0$

---

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\begin{aligned} \|\Delta q^t\| &= \|\frac{u^t}{v^t}(e^t - q^{t-1})\| = \frac{u^t}{v^t}\cdot \|e^t - q^{t-1}\| \leq \frac{u^t}{v^t}\cdot 2 \\ &= \frac{u^t}{u_1 + \cdots + u_t}\cdot 2 \leq \frac{\max\{u^s \mid s \leq t\}}{t\cdot \min\{u^s \mid s \leq t\}}\cdot 2 = \frac{2}{t}. \end{aligned}$$

So $\lim_{t\to\infty} \|\Delta q^t\| = 0 \not\Rightarrow$ convergence![1]

[1]Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\begin{aligned} \|\Delta q^t\| &= \|\frac{u^t}{v^t}(e^t - q^{t-1})\| = \frac{u^t}{v^t}\cdot \|e^t - q^{t-1}\| \leq \frac{u^t}{v^t}\cdot 2 \\ &= \frac{u^t}{u_1 + \cdots + u_t}\cdot 2 \leq \frac{\max\{u^s \mid s \leq t\}}{t\cdot \min\{u^s \mid s \leq t\}}\cdot 2 = \frac{2}{t}. \end{aligned}$$

So $\lim_{t\to\infty} \|\Delta q^t\| = 0 \not\Rightarrow$ convergence![1] Does q^t converge?

[1]Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\begin{aligned}\|\Delta q^t\| &= \|\frac{u^t}{v^t}(e^t - q^{t-1})\| = \frac{u^t}{v^t} \cdot \|e^t - q^{t-1}\| \leq \frac{u^t}{v^t} \cdot 2 \\ &= \frac{u^t}{u_1 + \cdots + u_t} \cdot 2 \leq \frac{\max\{u^s \mid s \leq t\}}{t \cdot \min\{u^s \mid s \leq t\}} \cdot 2 = \frac{2}{t}.\end{aligned}$$

So $\lim_{t\to\infty} \|\Delta q^t\| = 0 \not\Rightarrow$ convergence![1] Does q^t converge? If so, to a Pareto optimal strategy?

---

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$.

Dynamics of the mixed strategy: convergence

So the dynamics of the mixed strategy in round t is given by

$$\Delta q^t = \frac{u^t}{v^t}(e^t - q^{t-1}).$$

On coordinate x it is

$$\Delta q^t_x = \frac{u^t}{v^t}(e^t_x - q^{t-1}_x).$$

We have:

$$\begin{aligned} \|\Delta q^t\| &= \|\frac{u^t}{v^t}(e^t - q^{t-1})\| = \frac{u^t}{v^t}\cdot \|e^t - q^{t-1}\| \leq \frac{u^t}{v^t}\cdot 2 \\ &= \frac{u^t}{u_1 + \cdots + u_t}\cdot 2 \leq \frac{\max\{u^s \mid s \leq t\}}{t \cdot \min\{u^s \mid s \leq t\}}\cdot 2 = \frac{2}{t}. \end{aligned}$$

So $\lim_{t\to\infty} \|\Delta q^t\| = 0 \not\Rightarrow$ convergence![1] Does q^t converge? If so, to a Pareto optimal strategy? Cf. (Beggs, 2005).

[1] Think of the non-converging series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \ldots$.

Abstraction of past payoffs t^p

In 1991 and 1993, B. Arthur proposed the following update formula:

$$\Delta q^t = \frac{u^t}{Ct^p + u^t}(e^t - q^{t-1})$$

Abstraction of past payoffs t^p

In 1991 and 1993, B. Arthur proposed the following update formula:

$$\Delta q^t = \frac{u^t}{Ct^p + u^t}(e^t - q^{t-1})$$

Consequently,

$$\|\Delta q^t\| \leq \frac{1}{Ct^p}.$$

Abstraction of past payoffs t^p

In 1991 and 1993, B. Arthur proposed the following update formula:

$$\Delta q^t = \frac{u^t}{Ct^p + u^t}(e^t - q^{t-1})$$

Consequently,

$$\|\Delta q^t\| \leq \frac{1}{Ct^p}.$$

- If $p > 1$, then late payoffs are not longer that important.

Abstraction of past payoffs t^p

In 1991 and 1993, B. Arthur proposed the following update formula:

$$\Delta q^t = \frac{u^t}{Ct^p + u^t}(e^t - q^{t-1})$$

Consequently,

$$\|\Delta q^t\| \leq \frac{1}{Ct^p}.$$

- If $p > 1$, then late payoffs are not longer that important.
- If $p = 2$, then there is convergence

Abstraction of past payoffs t^p

In 1991 and 1993, B. Arthur proposed the following update formula:

$$\Delta q^t = \frac{u^t}{Ct^p + u^t}(e^t - q^{t-1})$$

Consequently,

$$\|\Delta q^t\| \leq \frac{1}{Ct^p}.$$

- If $p > 1$, then late payoffs are not longer that important.
- If $p = 2$, then there is convergence $\Rightarrow$ may lock into sub-optimal play.

Abstraction of past payoffs t^p

In 1991 and 1993, B. Arthur proposed the following update formula:

$$\Delta q^t = \frac{u^t}{Ct^p + u^t}(e^t - q^{t-1})$$

Consequently,

$$\|\Delta q^t\| \leq \frac{1}{Ct^p}.$$

- If $p > 1$, then late payoffs are not longer that important.
- If $p = 2$, then there is convergence $\Rightarrow$ may lock into sub-optimal play.
- In related research, where the value of p is determined through psychological experiments, it is estimated that $p < 1$.

Abstraction of past payoffs t^p

In 1991 and 1993, B. Arthur proposed the following update formula:

$$\Delta q^t = \frac{u^t}{Ct^p + u^t}(e^t - q^{t-1})$$

Consequently,

$$\|\Delta q^t\| \leq \frac{1}{Ct^p}.$$

- If $p > 1$, then late payoffs are not longer that important.
- If $p = 2$, then there is convergence $\Rightarrow$ may lock into sub-optimal play.
- In related research, where the value of p is determined through psychological experiments, it is estimated that $p < 1$.

B. Arthur (1993): "On Designing Economic Agents that Behave Like Human Agents". In: *Journal of Evolutionary Economy* **3**, pp 1-22.

Designing Economic Agents (Arthur, 1991)

chooses one of N possible actions at each time that have random payoffs or profits drawn from a stationary distribution that is unknown in advance. This would be the case, for example, where a firm, government agency, or research department is faced each period with a choice among N alternative pricing schemes, or policy options, or research projects, each with consequences that are poorly understood at the outset and that vary from "trial" to "trial". The agent chooses one alternative at each time, observes its consequence or payoff, and over time updates his choice as a result. What makes this iterated choice problem interesting is the tension between *exploitation* of high-payoff actions that have been undertaken many times and are therefore well understood, and *exploration* of seldom-tried actions that potentially may have higher average payoff.

The classic multi-arm-bandit version of this problem is to design a learning algorithm or automaton that maximizes some criterion—such as expected average payoff. Our problem is different. It is to design a learning algorithm or learning automaton that can be tuned to choose actions in this iterated choice situation the way humans

action. That is, it sets $p_t = S_t / C_t$.

2) Chooses one action from the set according to the probabilities p_t and triggers that action.

3) Observes the payoff received and updates strengths by adding the chosen actions's j's payoff to action j's strength. That is, where action j is chosen, it sets the strengths to $S_t + \beta_t$ where $\beta_t = \Phi(j)e_j$; (e_j is the jth unit vector).

4) Renormalizes the strengths to sum to a value from a prechosen time sequence. In this case, it renormalizes strengths to sum to $C_t = Ct^{\nu}$.

This last step allows us to set the rate and deceleration of the learning via the parameters C and ν that are fixed in advance. The rate of learning, it turns out, is proportional to $1/(Ct^{\nu})$. Parameters C and ν thus define a two-parameter family of algorithms that can be used to calibrate the automaton.

The algorithm has a simple behavioral interpretation (at least when $\nu = 0$). The strength vector summarizes the current confidence the agent or automaton has learned to associate with actions 1 through N. Confidence associated with an action increases according to the (random) payoff it brings in when taken. The automaton chooses its ac-

Decaying past payoffs

In 1995, Erev and Roth proposed the following update formula:

$$\theta^{t+1} = \lambda \theta^t + u^t \cdot e^t,$$

where $0 \leq \lambda \leq 1$

Decaying past payoffs

In 1995, Erev and Roth proposed the following update formula:

$$\theta^{t+1} = \lambda\theta^{t} + u^{t} \cdot e^{t},$$

where $0 \leq \lambda \leq 1$ determines a decay of previous propensities.

Decaying past payoffs

In 1995, Erev and Roth proposed the following update formula:

$$\theta^{t+1} = \lambda\theta^t + u^t \cdot e^t,$$

where $0 \leq \lambda \leq 1$ determines a decay of previous propensities. It can be shown:

$$\Delta q^t = \frac{u^t}{\sum_{s \leq t} \lambda^{t-s} \cdot u^s} (e^t - q^{t-1})$$

Decaying past payoffs

In 1995, Erev and Roth proposed the following update formula:

$$\theta^{t+1} = \lambda\theta^t + u^t \cdot e^t,$$

where $0 \leq \lambda \leq 1$ determines a decay of previous propensities. It can be shown:

$$\Delta q^t = \frac{u^t}{\sum_{s \leq t} \lambda^{t-s} \cdot u^s}(e^t - q^{t-1})$$

Since

$$\Big(\sum_{s \leq t} \lambda^{t-s}\Big) \min\{u^s \mid s \leq t\}$$

Decaying past payoffs

In 1995, Erev and Roth proposed the following update formula:

$$\theta^{t+1} = \lambda\theta^t + u^t \cdot e^t,$$

where $0 \leq \lambda \leq 1$ determines a decay of previous propensities. It can be shown:

$$\Delta q^t = \frac{u^t}{\sum_{s \leq t} \lambda^{t-s} \cdot u^s} (e^t - q^{t-1})$$

Since

$$\Big(\sum_{s \leq t} \lambda^{t-s}\Big) \min\{u^s \mid s \leq t\} \;\leq\; \sum_{s \leq t} \lambda^{t-s} \cdot u^s$$

Decaying past payoffs

In 1995, Erev and Roth proposed the following update formula:

$$\theta^{t+1} = \lambda\theta^t + u^t \cdot e^t,$$

where $0 \leq \lambda \leq 1$ determines a decay of previous propensities. It can be shown:

$$\Delta q^t = \frac{u^t}{\sum_{s \leq t} \lambda^{t-s} \cdot u^s}(e^t - q^{t-1})$$

Since

$$\Big(\sum_{s\leq t} \lambda^{t-s}\Big) \min\{u^s \mid s \leq t\} \;\leq\; \sum_{s \leq t} \lambda^{t-s} \cdot u^s \;\leq\; \Big(\sum_{s \leq t} \lambda^{t-s}\Big) \max\{u^s \mid s \leq t\}$$

Decaying past payoffs

In 1995, Erev and Roth proposed the following update formula:

$$\theta^{t+1} = \lambda\theta^t + u^t \cdot e^t,$$

where $0 \leq \lambda \leq 1$ determines a decay of previous propensities. It can be shown:

$$\Delta q^t = \frac{u^t}{\sum_{s \leq t} \lambda^{t-s} \cdot u^s}(e^t - q^{t-1})$$

Since

$$\Big(\sum_{s \leq t} \lambda^{t-s}\Big) \min\{u^s \mid s \leq t\} \;\leq\; \sum_{s \leq t} \lambda^{t-s} \cdot u^s \;\leq\; \Big(\sum_{s \leq t} \lambda^{t-s}\Big) \max\{u^s \mid s \leq t\}$$

and since

$$\sum_{s \leq t} \lambda^{t-s} = 1 + \lambda + \lambda^2 + \cdots + \lambda^t = \frac{1 - \lambda^{t+1}}{1 - \lambda}$$

for $\lambda \neq 1$,

Decaying past payoffs

In 1995, Erev and Roth proposed the following update formula:

$$\theta^{t+1} = \lambda\theta^t + u^t \cdot e^t,$$

where $0 \leq \lambda \leq 1$ determines a decay of previous propensities. It can be shown:

$$\Delta q^t = \frac{u^t}{\sum_{s \leq t} \lambda^{t-s} \cdot u^s}(e^t - q^{t-1})$$

Since

$$\left(\sum_{s \leq t} \lambda^{t-s}\right) \min\{u^s \mid s \leq t\} \;\leq\; \sum_{s \leq t} \lambda^{t-s} \cdot u^s \;\leq\; \left(\sum_{s \leq t} \lambda^{t-s}\right) \max\{u^s \mid s \leq t\}$$

and since

$$\sum_{s \leq t} \lambda^{t-s} = 1 + \lambda + \lambda^2 + \cdots + \lambda^t = \frac{1 - \lambda^{t+1}}{1 - \lambda}$$

for $\lambda \neq 1$, the mixed strategy tends to change at a rate $\sim 1 - \lambda$.

Past payoffs with an aspiration level

Suppose an aspiration level $a^t \in R$ at every round

Past payoffs with an aspiration level

Suppose an aspiration level $a^t \in R$ at every round:

$$\begin{aligned} u_x^t > a^t \quad &\Rightarrow \quad \text{positively reinforce action } x \\ u_x^t < a^t \quad &\Rightarrow \quad \text{negatively reinforce action } x \end{aligned}$$

Past payoffs with an aspiration level

Suppose an aspiration level $a^t \in R$ at every round:

$$\begin{aligned} u_x^t > a^t \quad &\Rightarrow \quad \text{positively reinforce action } x \\ u_x^t < a^t \quad &\Rightarrow \quad \text{negatively reinforce action } x \end{aligned}$$

Correspondingly, the mixed strategy evolves according to

$$\Delta q^t = (u^t - a^t)(e^t - q^{t-1}).$$

Past payoffs with an aspiration level

Suppose an aspiration level $a^t \in R$ at every round:

$$u_x^t > a^t \quad \Rightarrow \quad \text{positively reinforce action } x$$
$$u_x^t < a^t \quad \Rightarrow \quad \text{negatively reinforce action } x$$

Correspondingly, the mixed strategy evolves according to

$$\Delta q^t = (u^t - a^t)(e^t - q^{t-1}).$$

Typical definitions for aspiration:

Past payoffs with an aspiration level

Suppose an aspiration level $a^t \in R$ at every round:

$$\begin{aligned} u_x^t > a^t \quad &\Rightarrow \quad \text{positively reinforce action } x \\ u_x^t < a^t \quad &\Rightarrow \quad \text{negatively reinforce action } x \end{aligned}$$

Correspondingly, the mixed strategy evolves according to

$$\Delta q^t = (u^t - a^t)(e^t - q^{t-1}).$$

Typical definitions for aspiration:

- **Average past payoffs.** $a^t =_{Def} v^t / t$. A.k.a. reinforcement comparison method (Sutton *et al.*, 1998), or satisficing play (Stimpson *et al.*, 2001).

Past payoffs with an aspiration level

Suppose an aspiration level $a^t \in R$ at every round:

$$\begin{aligned} u_x^t > a^t \quad &\Rightarrow \quad \text{positively reinforce action } x \\ u_x^t < a^t \quad &\Rightarrow \quad \text{negatively reinforce action } x \end{aligned}$$

Correspondingly, the mixed strategy evolves according to

$$\Delta q^t = (u^t - a^t)(e^t - q^{t-1}).$$

Typical definitions for aspiration:

- **Average past payoffs.** $a^t =_{Def} v^t / t$. A.k.a. reinforcement comparison method (Sutton *et al.*, 1998), or satisficing play (Stimpson *et al.*, 2001).
- **Discounted past payoffs.** $a^t =_{Def} \sum_{s \leq t} \lambda^{t-s} \cdot u^s$ (Erev & Roth, 1995).

Past payoffs with an aspiration level

Suppose an aspiration level $a^t \in R$ at every round:

$$\begin{aligned} u_x^t > a^t \quad &\Rightarrow \quad \text{positively reinforce action } x \\ u_x^t < a^t \quad &\Rightarrow \quad \text{negatively reinforce action } x \end{aligned}$$

Correspondingly, the mixed strategy evolves according to

$$\Delta q^t = (u^t - a^t)(e^t - q^{t-1}).$$

Typical definitions for aspiration:

- **Average past payoffs.** $a^t =_{Def} v^t / t$. A.k.a. reinforcement comparison method (Sutton *et al.*, 1998), or satisficing play (Stimpson *et al.*, 2001).
- **Discounted past payoffs.** $a^t =_{Def} \sum_{s \leq t} \lambda^{t-s} \cdot u^s$ (Erev & Roth, 1995).

Börgers and Sarin (2000). "Naïve Reinforcement Learning with Endogeneous Aspirations" in: *Int. Economic Review* **41**, pp. 921-950.

Adequacy of reinforcement learning

Adequacy of reinforcement learning

Does reinforcement learning lead to optimal behaviour against B?

Adequacy of reinforcement learning

Does reinforcement learning lead to optimal behaviour against B?

If A and B would both converge to optimal behaviour, i.e., to a best response, this would yield a Nash equilibrium.

Adequacy of reinforcement learning

Does reinforcement learning lead to optimal behaviour against B?

If A and B would both converge to optimal behaviour, i.e., to a best response, this would yield a Nash equilibrium.

Turns out to be too demanding.

Adequacy of reinforcement learning

Does reinforcement learning lead to optimal behaviour against B?

If A and B would both converge to optimal behaviour, i.e., to a best response, this would yield a Nash equilibrium.

Turns out to be too demanding. Less demanding:

Does reinforcement learning converge to optimal behaviour when B is stationary (perhaps with noise)?

Adequacy of reinforcement learning

Does reinforcement learning lead to optimal behaviour against B?

If A and B would both converge to optimal behaviour, i.e., to a best response, this would yield a Nash equilibrium.

Turns out to be too demanding. Less demanding:

Does reinforcement learning converge to optimal behaviour when B is stationary (perhaps with noise)?

- A history is a finite sequence of actions $\xi^t : (x_1, y_1), \dots, (x_t, y_t)$.

Adequacy of reinforcement learning

Does reinforcement learning lead to optimal behaviour against B?

If A and B would both converge to optimal behaviour, i.e., to a best response, this would yield a Nash equilibrium.

Turns out to be too demanding. Less demanding:

Does reinforcement learning converge to optimal behaviour when B is stationary (perhaps with noise)?

- A history is a finite sequence of actions $\xi^t : (x_1, y_1), \dots, (x_t, y_t)$.
- A strategy for A is a function $g : H \to \Delta(X)$ that maps histories to probability distributions over X. Let $q_{t+1} =_{Def} g(\xi^t)$.

Optimality against stationary opponents

- Assume that B plays a fixed probability distribution $q^* \in \Delta(Y)$.

Optimality against stationary opponents

- Assume that B plays a fixed probability distribution $q^* \in \Delta(Y)$.
- The combination of θ^0, g and q^*
  yields a realisation $\omega = (x_1, y_1), \dots, (x_t, y_t), \dots$.

Optimality against stationary opponents

- Assume that B plays a fixed probability distribution $q^* \in \Delta(Y)$.
- The combination of θ^0, g and q^*
  yields a realisation $\omega = (x_1, y_1), \dots, (x_t, y_t), \dots$.
- Define $B(q^*) =_{Def} \{ x \in X \mid x \text{ is a best response to } q^* \}$.

Optimality against stationary opponents

- Assume that B plays a fixed probability distribution $q^* \in \Delta(Y)$.
- The combination of θ^0, g and q^* yields a realisation $\omega = (x_1, y_1), \dots, (x_t, y_t), \dots$.
- Define $B(q^*) =_{Def} \{ x \in X \mid x \text{ is a best response to } q^* \}$.

Definition. A strategy g is called optimal against q^* if, with probability one,

$$\text{for all } x \notin B(q^*) : \lim_{t \to \infty} q_x^t = 0.$$

Optimality against stationary opponents

- Assume that B plays a fixed probability distribution $q^* \in \Delta(Y)$.
- The combination of θ^0, g and q^* yields a realisation $\omega = (x_1, y_1), \dots, (x_t, y_t), \dots$.
- Define $B(q^*) =_{Def} \{\, x \in X \mid x \text{ is a best response to } q^* \,\}$.

Definition. A strategy g is called optimal against q^* if, with probability one,

$$\text{for all } x \notin B(q^*) : \lim_{t\to\infty} q_x^t = 0.$$

Theorem (Beggs, 2005). Given finite action sets X and Y, cumulative payoff matching on X is optimal against every stationary distribution q^* on Y.

Optimality against stationary opponents

- Assume that B plays a fixed probability distribution $q^* \in \Delta(Y)$.
- The combination of θ^0, g and q^* yields a realisation $\omega = (x_1, y_1), \dots, (x_t, y_t), \dots$.
- Define $B(q^*) =_{Def} \{ x \in X \mid x \text{ is a best response to } q^* \}$.

Definition. A strategy g is called optimal against q^* if, with probability one,

$$\text{for all } x \notin B(q^*) : \lim_{t \to \infty} q_x^t = 0.$$

Theorem (Beggs, 2005). Given finite action sets X and Y, cumulative payoff matching on X is optimal against every stationary distribution q^* on Y.

Peyton Young (2004, p. 17): "Its proof is actually quite involved (...)".

Part II: Convergence to dominant strategies

Alan Beggs, Economics professor, Wadham College, Oxfo

The learning model

Single-state cumulative payoff matching (Erev & Roth, 1995).

The learning model

Single-state cumulative payoff matching (Erev & Roth, 1995).

As usual:

$$A_i(n) = \begin{cases} A_i(n-1) + \pi_i(n) & \text{if action } i \text{ is chosen at round } n, \\ A_i(n-1) & \text{else.} \end{cases}$$

The learning model

Single-state cumulative payoff matching (Erev & Roth, 1995).

As usual:

$$A_i(n) = \begin{cases} A_i(n-1) + \pi_i(n) & \text{if action } i \text{ is chosen at round } n, \\ A_i(n-1) & \text{else.} \end{cases}$$

As usual:

$$\Pr_i(n+1) = \frac{A_i(n)}{\sum_{j=1}^{m} A_j(n)}.$$

The learning model

Single-state cumulative payoff matching (Erev & Roth, 1995).

As usual:

$$A_i(n) = \begin{cases} A_i(n-1) + \pi_i(n) & \text{if action } i \text{ is chosen at round } n, \\ A_i(n-1) & \text{else.} \end{cases}$$

As usual:

$$\mathrm{Pr}_i(n+1) = \frac{A_i(n)}{\sum_{j=1}^{m} A_j(n)}.$$

The following two assumptions are made:

The learning model

Single-state cumulative payoff matching (Erev & Roth, 1995).

As usual:

$$A_i(n) = \begin{cases} A_i(n-1) + \pi_i(n) & \text{if action } i \text{ is chosen at round } n, \\ A_i(n-1) & \text{else.} \end{cases}$$

As usual:

$$\Pr_i(n+1) = \frac{A_i(n)}{\sum_{j=1}^{m} A_j(n)}.$$

The following two assumptions are made:

1. All past, current and future payoffs $\pi_i(n)$ are bounded away from zero and bounded from above.

The learning model

Single-state cumulative payoff matching (Erev & Roth, 1995).

As usual:

$$A_i(n) = \begin{cases} A_i(n-1) + \pi_i(n) & \text{if action } i \text{ is chosen at round } n, \\ A_i(n-1) & \text{else.} \end{cases}$$

As usual:

$$\Pr_i(n+1) = \frac{A_i(n)}{\sum_{j=1}^{m} A_j(n)}.$$

The following two assumptions are made:

1. All past, current and future payoffs $\pi_i(n)$ are bounded away from zero and bounded from above. So there are $0 < k_1 \leq k_2$ such that all payoffs are in $[k_1, k_2]$.

The learning model

Single-state cumulative payoff matching (Erev & Roth, 1995).

As usual:

$$A_i(n) = \begin{cases} A_i(n-1) + \pi_i(n) & \text{if action } i \text{ is chosen at round } n, \\ A_i(n-1) & \text{else.} \end{cases}$$

As usual:

$$\Pr_i(n+1) = \frac{A_i(n)}{\sum_{j=1}^{m} A_j(n)}.$$

The following two assumptions are made:

1. All past, current and future payoffs $\pi_i(n)$ are bounded away from zero and bounded from above. So there are $0 < k_1 \leq k_2$ such that all payoffs are in $[k_1, k_2]$.
2. The initial $A_i(0)$ are strictly positive.

Choice of actions

Lemma 1. *Each action is chosen infinitely often a.s.*

Choice of actions

Lemma 1. *Each action is chosen infinitely often a.s.*

Proof.

Choice of actions

Lemma 1. *Each action is chosen infinitely often a.s.*

Proof. From the above assumptions it follows that

$$\Pr_i(n+1) = \frac{A_i(n)}{\sum_{j=1}^{m} A_j(n)} \geq \frac{A_i(0)}{A_i(0) + nk_2}.$$

Choice of actions

Lemma 1. *Each action is chosen infinitely often a.s.*

Proof. From the above assumptions it follows that

$$\mathrm{Pr}_i(n+1) = \frac{A_i(n)}{\sum_{j=1}^{m} A_j(n)} \geq \frac{A_i(0)}{A_i(0) + nk_2}.$$

(Which is like worst case for i: as if i was never chosen and all previous n rounds actions $\neq i$ received the maximum possible payoff.)

Choice of actions

Lemma 1. *Each action is chosen infinitely often a.s.*

Proof. From the above assumptions it follows that

$$\Pr_i(n+1) = \frac{A_i(n)}{\sum_{j=1}^{m} A_j(n)} \geq \frac{A_i(0)}{A_i(0) + nk_2}.$$

(Which is like worst case for i: as if i was never chosen and all previous n rounds actions $\neq i$ received the maximum possible payoff.)

Apply the so-called conditional Borel-Cantelli lemma:[2]

Choice of actions

Lemma 1. *Each action is chosen infinitely often a.s.*

Proof. From the above assumptions it follows that

$$\Pr_i(n+1) = \frac{A_i(n)}{\sum_{j=1}^{m} A_j(n)} \geq \frac{A_i(0)}{A_i(0) + nk_2}.$$

(Which is like worst case for i: as if i was never chosen and all previous n rounds actions $\neq i$ received the maximum possible payoff.)

Apply the so-called conditional Borel-Cantelli lemma:[2] if $\{E_n\}_n$ are events, and

$$\sum_{n=1}^{\infty} \Pr(E_n \mid X_{n-1}, \dots, X_1)$$

is unbounded, then an infinite number of E_n's occur a.s. $\square$

[2] A.k.a. the *second Borel-Cantelli lemma*, or the *Borel-Cantelli-Lévy lemma* (Shiryaev, p. 518).

Unboundedness of propensities, and convergence

Lemma 2. *For each i, A_i tends to infinity with probability one.*

Unboundedness of propensities, and convergence

Lemma 2. *For each* i, A_i *tends to infinity with probability one.*

Proof.

Unboundedness of propensities, and convergence

Lemma 2. *For each i, A_i tends to infinity with probability one.*

Proof. For each i, action i is chosen infinitely often with probability one.

Unboundedness of propensities, and convergence

Lemma 2. *For each i, A_i tends to infinity with probability one.*

Proof. For each i, action i is chosen infinitely often with probability one. Since payoff per round is bounded from below by k_1,

Unboundedness of propensities, and convergence

Lemma 2. *For each i, A_i tends to infinity with probability one.*

Proof. For each i, action i is chosen infinitely often with probability one. Since payoff per round is bounded from below by k_1, we have $\sum_{j=1}^{\infty} k_1 \leq A_i$, where j runs over rounds where i is chosen. $\square$

Unboundedness of propensities, and convergence

Lemma 2. *For each i, A_i tends to infinity with probability one.*

Proof. For each i, action i is chosen infinitely often with probability one. Since payoff per round is bounded from below by k_1, we have $\sum_{j=1}^{\infty} k_1 \leq A_i$, where j runs over rounds where i is chosen. $\square$

Now Lemma 1 + Lemma 2 + martingale theory suffice to claim convergence as follows.

Unboundedness of propensities, and convergence

Lemma 2. *For each i, A_i tends to infinity with probability one.*

Proof. For each i, action i is chosen infinitely often with probability one. Since payoff per round is bounded from below by k_1, we have $\sum_{j=1}^{\infty} k_1 \leq A_i$, where j runs over rounds where i is chosen. □

Now Lemma 1 + Lemma 2 + martingale theory suffice to claim convergence as follows. Suppose there are only two possible actions: a_1 and a_2.

Unboundedness of propensities, and convergence

Lemma 2. *For each i, A_i tends to infinity with probability one.*

Proof. For each i, action i is chosen infinitely often with probability one. Since payoff per round is bounded from below by k_1, we have $\sum_{j=1}^{\infty} k_1 \leq A_i$, where j runs over rounds where i is chosen. □

Now Lemma 1 + Lemma 2 + martingale theory suffice to claim convergence as follows. Suppose there are only two possible actions: a_1 and a_2. The expression

$$E[\,\pi(a_i)\mid \text{history}\,]$$

denotes the expected payoff of action a_i, given history of play.

Unboundedness of propensities, and convergence

Lemma 2. *For each i, A_i tends to infinity with probability one.*

Proof. For each i, action i is chosen infinitely often with probability one. Since payoff per round is bounded from below by k_1, we have $\sum_{j=1}^{\infty} k_1 \leq A_i$, where j runs over rounds where i is chosen. $\square$

Now Lemma 1 + Lemma 2 + martingale theory suffice to claim convergence as follows. Suppose there are only two possible actions: a_1 and a_2. The expression

$$E[\,\pi(a_i) \mid \text{history}\,]$$

denotes the expected payoff of action a_i, given history of play.

Theorem 1. *If a_1 is dominant, i.e., if, for all histories h*

$$E[\,\pi(a_1) \mid h\,] > \gamma E[\,\pi(a_2) \mid h\,]$$

for some fixed $\gamma > 1$

Unboundedness of propensities, and convergence

Lemma 2. *For each i, A_i tends to infinity with probability one.*

Proof. For each i, action i is chosen infinitely often with probability one. Since payoff per round is bounded from below by k_1, we have $\sum_{j=1}^{\infty} k_1 \leq A_i$, where j runs over rounds where i is chosen. □

Now Lemma 1 + Lemma 2 + martingale theory suffice to claim convergence as follows. Suppose there are only two possible actions: a_1 and a_2. The expression

$$E[\,\pi(a_i) \mid \text{history}\,]$$

denotes the expected payoff of action a_i, given history of play.

Theorem 1. *If a_1 is dominant, i.e., if, for all histories h*

$$E[\,\pi(a_1) \mid h\,] > \gamma E[\,\pi(a_2) \mid h\,]$$

for some fixed $\gamma > 1$, then the probability that a_1 will be played converges to one.

Convergence to dominant action: proof in a nutshell

Convergence to dominant action: proof in a nutshell

- If a_1 is dominant (like in the above inequality), the objective is to show that

$$\frac{A_2}{A_1}(n) \rightsquigarrow 0, \text{ a.s.}$$

Convergence to dominant action: proof in a nutshell

- If a_1 is dominant (like in the above inequality), the objective is to show that

$$\frac{A_2}{A_1}(n) \rightsquigarrow 0, \text{ a.s.}$$

- To this end, it suffices to show that

$$\frac{A_2^{\epsilon}}{A_1}(n) \rightsquigarrow C, \text{ a.s.} \tag{1}$$

for some C and for some $1 < \epsilon < \gamma$ (which is possible, since $\gamma > 1$).

Convergence to dominant action: proof in a nutshell

- If a_1 is dominant (like in the above inequality), the objective is to show that

$$\frac{A_2}{A_1}(n) \rightsquigarrow 0, \text{ a.s.}$$

- To this end, it suffices to show that

$$\frac{A_2^{\epsilon}}{A_1}(n) \rightsquigarrow C, \text{ a.s.} \tag{1}$$

for some C and for some $1 < \epsilon < \gamma$ (which is possible, since $\gamma > 1$).

For, in such case,

$$\lim_{n\to\infty} \frac{A_2}{A_1}(n) = \lim_{n\to\infty} \frac{A_2^{\epsilon}}{A_1} \cdot \frac{1}{A_2^{\epsilon-1}} = \lim_{n\to\infty} \frac{A_2^{\epsilon}}{A_1} \cdot \lim_{n\to\infty} \frac{1}{A_2^{\epsilon-1}} = C \cdot 0.$$

Convergence to dominant action: proof in a nutshell

- If a_1 is dominant (like in the above inequality), the objective is to show that

$$\frac{A_2}{A_1}(n) \rightsquigarrow 0, \text{ a.s.}$$

- To this end, it suffices to show that

$$\frac{A_2^{\epsilon}}{A_1}(n) \rightsquigarrow C, \text{ a.s.} \qquad (1)$$

for some C and for some $1 < \epsilon < \gamma$ (which is possible, since $\gamma > 1$).

For, in such case,

$$\lim_{n\to\infty} \frac{A_2}{A_1}(n) = \lim_{n\to\infty} \frac{A_2^{\epsilon}}{A_1} \cdot \frac{1}{A_2^{\epsilon-1}} = \lim_{n\to\infty} \frac{A_2^{\epsilon}}{A_1} \cdot \lim_{n\to\infty} \frac{1}{A_2^{\epsilon-1}} = C \cdot 0.$$

- To this end, Beggs shows that, for some $n \geq N$, and for all $1 < \epsilon < \gamma$, (1) is a so-called non-negative super-martingale.

Convergence to dominant action: proof in a nutshell

- If a_1 is dominant (like in the above inequality), the objective is to show that

$$\frac{A_2}{A_1}(n) \rightsquigarrow 0, \text{ a.s.}$$

- To this end, it suffices to show that

$$\frac{A_2^{\epsilon}}{A_1}(n) \rightsquigarrow C, \text{ a.s.} \tag{1}$$

for some C and for some $1 < \epsilon < \gamma$ (which is possible, since $\gamma > 1$).

For, in such case,

$$\lim_{n\to\infty} \frac{A_2}{A_1}(n) = \lim_{n\to\infty} \frac{A_2^{\epsilon}}{A_1} \cdot \frac{1}{A_2^{\epsilon-1}} = \lim_{n\to\infty} \frac{A_2^{\epsilon}}{A_1} \cdot \lim_{n\to\infty} \frac{1}{A_2^{\epsilon-1}} = C \cdot 0.$$

- To this end, Beggs shows that, for some $n \geq N$, and for all $1 < \epsilon < \gamma$, (1) is a so-called non-negative super-martingale.

  Why?

Convergence to dominant action: proof in a nutshell

- If a_1 is dominant (like in the above inequality), the objective is to show that

$$\frac{A_2}{A_1}(n) \rightsquigarrow 0, \text{ a.s.}$$

- To this end, it suffices to show that

$$\frac{A_2^{\epsilon}}{A_1}(n) \rightsquigarrow C, \text{ a.s.} \tag{1}$$

for some C and for some $1 < \epsilon < \gamma$ (which is possible, since $\gamma > 1$).

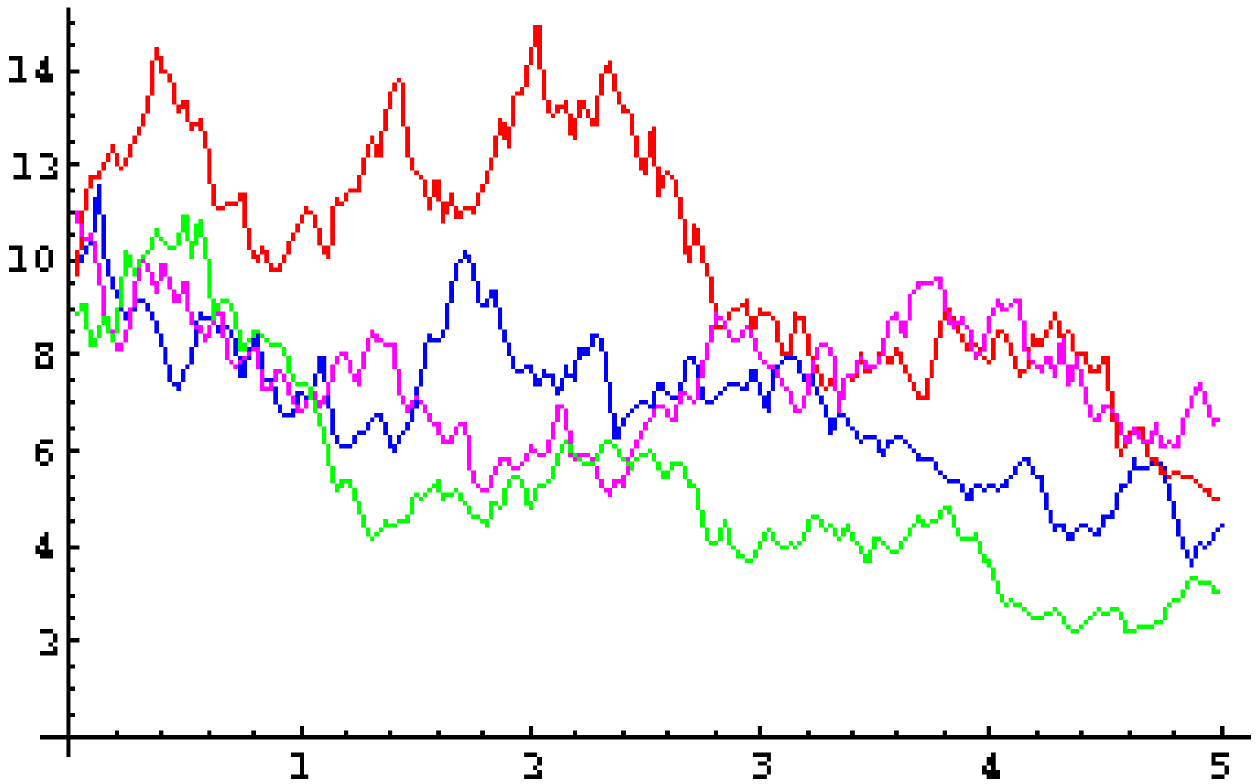
For, in such case,

$$\lim_{n\to\infty} \frac{A_2}{A_1}(n) = \lim_{n\to\infty} \frac{A_2^{\epsilon}}{A_1} \cdot \frac{1}{A_2^{\epsilon-1}} = \lim_{n\to\infty} \frac{A_2^{\epsilon}}{A_1} \cdot \lim_{n\to\infty} \frac{1}{A_2^{\epsilon-1}} = C \cdot 0.$$

- To this end, Beggs shows that, for some $n \geq N$, and for all $1 < \epsilon < \gamma$, (1) is a so-called non-negative super-martingale.

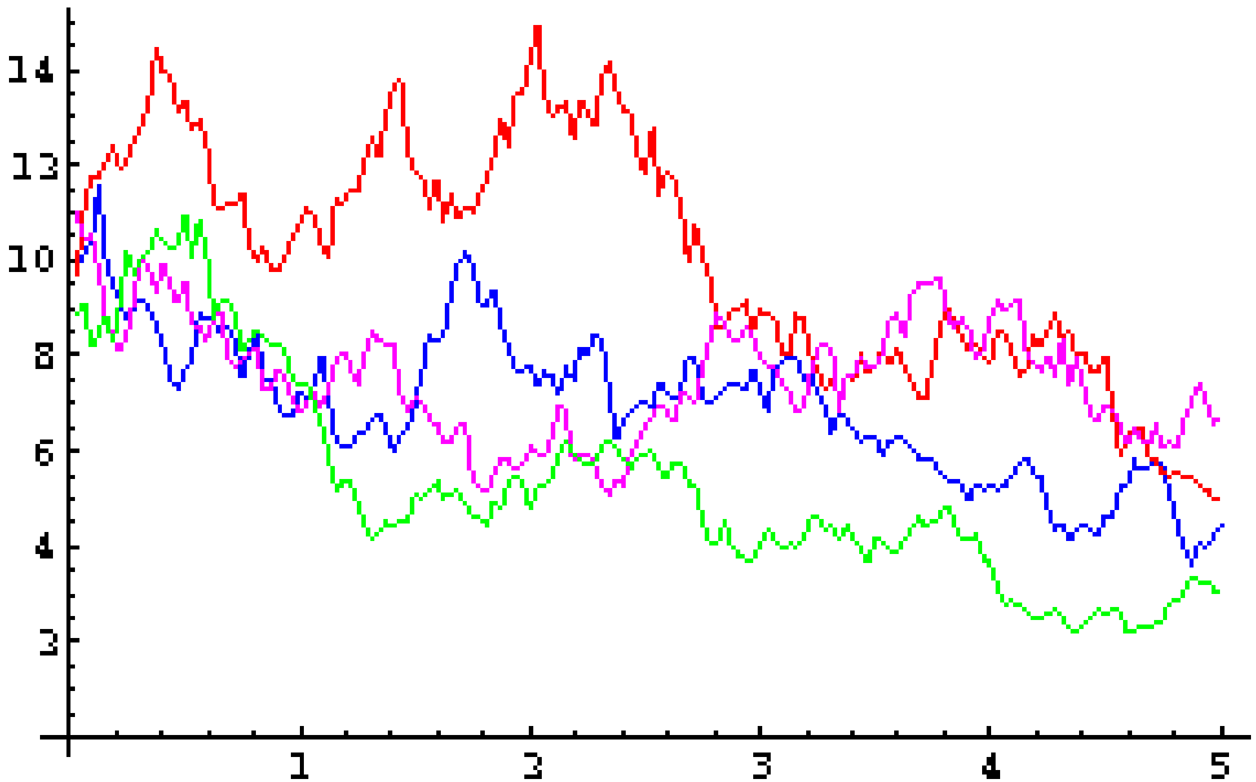
  Why? For it is known that every non-negative super-martingale converges to a finite limit C a.s.

Super-martingale (informal idea)



A super-martingale embodies the concept of one's capital in an unfair gambling game, for example roulette.

Super-martingale (informal idea)



A super-martingale embodies the concept of one's capital in an unfair gambling game, for example roulette. (Recall pockets "0" and "00".)

Super-martingale (formal definition)

A super-martingale is a *stochastic process* in which the conditional expectation of the next value, given the current and preceding values, is less than or equal to the current value:

$$E[\, Z_{n+1} \mid Z_1, \ldots, Z_n \,] \leq Z_n$$

Super-martingale (formal definition)

A super-martingale is a *stochastic process* in which the conditional expectation of the next value, given the current and preceding values, is less than or equal to the current value:

$$E[\, Z_{n+1} \mid Z_1, \dots, Z_n \,] \leq Z_n$$

1. **Expectations decrease.** Taking expectations on both sides yields $E[\, Z_{n+1} \,] \leq E[\, Z_n \,]$.

Super-martingale (formal definition)

A super-martingale is a *stochastic process* in which the conditional expectation of the next value, given the current and preceding values, is less than or equal to the current value:

$$E[\, Z_{n+1} \mid Z_1, \dots, Z_n \,] \leq Z_n$$

1. **Expectations decrease.** Taking expectations on both sides yields $E[\, Z_{n+1} \,] \leq E[\, Z_n \,]$. (Tower property of expectation: $E[E[X|Y]] = E[X]$.)

Super-martingale (formal definition)

A super-martingale is a *stochastic process* in which the conditional expectation of the next value, given the current and preceding values, is less than or equal to the current value:

$$E[\, Z_{n+1} \mid Z_1, \dots, Z_n \,] \leq Z_n$$

1. **Expectations decrease.** Taking expectations on both sides yields $E[\, Z_{n+1} \,] \leq E[\, Z_n \,]$. (Tower property of expectation: $E[E[X|Y]] = E[X]$.)
2. **Expectations converge.** From (1) and the monotone convergence theorem[3] it follows that the *expectations* of a non-negative super-martingale converge to a limit L somewhere in $[0, E[\, Z_1 \,]]$.

Super-martingale (formal definition)

A super-martingale is a *stochastic process* in which the conditional expectation of the next value, given the current and preceding values, is less than or equal to the current value:

$$E[\, Z_{n+1} \mid Z_1, \dots, Z_n \,] \leq Z_n$$

1. **Expectations decrease.** Taking expectations on both sides yields $E[\, Z_{n+1} \,] \leq E[\, Z_n \,]$. (Tower property of expectation: $E[E[X|Y]] = E[X]$.)

2. **Expectations converge.** From (1) and the monotone convergence theorem[3] it follows that the *expectations* of a non-negative super-martingale converge to a limit L somewhere in $[0, E[\, Z_1 \,]]$.

3. Doob's Martingale Convergence Theorem:

Super-martingale (formal definition)

A super-martingale is a *stochastic process* in which the conditional expectation of the next value, given the current and preceding values, is less than or equal to the current value:

$$E[\, Z_{n+1} \mid Z_1, \dots, Z_n \,] \leq Z_n$$

1. **Expectations decrease.** Taking expectations on both sides yields $E[\, Z_{n+1} \,] \leq E[\, Z_n \,]$. (Tower property of expectation: $E[E[X|Y]] = E[X]$.)
2. **Expectations converge.** From (1) and the monotone convergence theorem[3] it follows that the *expectations* of a non-negative super-martingale converge to a limit L somewhere in $[0, E[\, Z_1 \,]]$.
3. Doob's Martingale Convergence Theorem: **values themselves converge a.s.** as well.

Super-martingale (formal definition)

A super-martingale is a *stochastic process* in which the conditional expectation of the next value, given the current and preceding values, is less than or equal to the current value:

$$E[\, Z_{n+1} \mid Z_1, \dots, Z_n \,] \leq Z_n$$

1. **Expectations decrease.** Taking expectations on both sides yields $E[\, Z_{n+1} \,] \leq E[\, Z_n \,]$. (Tower property of expectation: $E[E[X|Y]] = E[X]$.)

2. **Expectations converge.** From (1) and the monotone convergence theorem[3] it follows that the *expectations* of a non-negative super-martingale converge to a limit L somewhere in $[0, E[\, Z_1 \,]]$.

3. Doob's Martingale Convergence Theorem: **values themselves converge a.s.** as well. I.e., let $\{Z_n\}_n$ be a super-martingale, such that $E[\,|Z_n|\,]$ is bounded.

Super-martingale (formal definition)

A super-martingale is a *stochastic process* in which the conditional expectation of the next value, given the current and preceding values, is less than or equal to the current value:

$$E[\, Z_{n+1} \mid Z_1, \dots, Z_n \,] \leq Z_n$$

1. **Expectations decrease.** Taking expectations on both sides yields $E[\, Z_{n+1} \,] \leq E[\, Z_n \,]$. (Tower property of expectation: $E[E[X|Y]] = E[X]$.)
2. **Expectations converge.** From (1) and the monotone convergence theorem[3] it follows that the *expectations* of a non-negative super-martingale converge to a limit L somewhere in $[0, E[\, Z_1 \,]]$.
3. Doob's Martingale Convergence Theorem: **values themselves converge a.s.** as well. I.e., let $\{Z_n\}_n$ be a super-martingale, such that $E[\, |Z_n| \,]$ is bounded. Then $\lim_{n\to\infty} Z_n$ exists a.s. and is finite.

[3] Ordinary mathematics.

To show that A_2^ϵ / A_1 is a non-neg super-martingale

$$E\left[\Delta\frac{A_2^\epsilon}{A_1}(n+1) \mid \text{history}\right]$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

$$E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid \text{history}\right] =$$

$$E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 1, \text{history}\right]$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

$$E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid \text{history}\right] =$$

$$E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 1, \text{history}\right] \Pr(1 \mid \text{history})$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

$$
\begin{aligned}
E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid \text{history}\right] = \\
& E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 1, \text{history}\right] \Pr(1 \mid \text{history}) \\
& \quad + \ E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 2, \text{history}\right]
\end{aligned}
$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

$$E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid \text{history}\right] =$$

$$E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 1, \text{history}\right] \Pr(1 \mid \text{history})$$

$$+\ E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 2, \text{history}\right] \Pr(2 \mid \text{history})$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

$$E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid \text{history}\right] =$$

$$E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 1, \text{history}\right] \Pr(1 \mid \text{history})$$

$$+ \; E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 2, \text{history}\right] \Pr(2 \mid \text{history})$$

$$= E\left[\frac{A_2^\epsilon(n)}{A_1(n) + \pi_1(n+1)} - \frac{A_2^\epsilon(n)}{A_1(n)}\right]$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

$$E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid \text{history}\right] =$$

$$E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 1, \text{history}\right] \Pr(1 \mid \text{history})$$

$$+ \; E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 2, \text{history}\right] \Pr(2 \mid \text{history})$$

$$= E\left[\frac{A_2^\epsilon(n)}{A_1(n) + \pi_1(n+1)} - \frac{A_2^\epsilon(n)}{A_1(n)}\right] \frac{A_1(n)}{A_1(n) + A_2(n)}$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

$$E\left[\Delta\frac{A_2^\epsilon}{A_1}(n+1) \mid \text{history}\right] =$$

$$E\left[\Delta\frac{A_2^\epsilon}{A_1}(n+1) \mid 1, \text{history}\right] \Pr(1 \mid \text{history})$$

$$+\ E\left[\Delta\frac{A_2^\epsilon}{A_1}(n+1) \mid 2, \text{history}\right] \Pr(2 \mid \text{history})$$

$$= E\left[\frac{A_2^\epsilon(n)}{A_1(n)+\pi_1(n+1)} - \frac{A_2^\epsilon(n)}{A_1(n)}\right] \frac{A_1(n)}{A_1(n)+A_2(n)}$$

$$+\ E\left[\frac{(A_2(n)+\pi_2(n+1))^\epsilon}{A_1(n)} - \frac{A_2^\epsilon(n)}{A_1(n)}\right]$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

$$
\begin{aligned}
E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid \text{history}\right] = {} & \\
& E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 1, \text{history}\right] \Pr(1 \mid \text{history}) \\
& \quad + \; E\left[\Delta \frac{A_2^\epsilon}{A_1}(n+1) \mid 2, \text{history}\right] \Pr(2 \mid \text{history})
\end{aligned}
$$

$$
\begin{aligned}
= E & \left[\frac{A_2^\epsilon(n)}{A_1(n) + \pi_1(n+1)} - \frac{A_2^\epsilon(n)}{A_1(n)}\right] \frac{A_1(n)}{A_1(n) + A_2(n)} \\
& + \; E\left[\frac{(A_2(n) + \pi_2(n+1))^\epsilon}{A_1(n)} - \frac{A_2^\epsilon(n)}{A_1(n)}\right] \frac{A_2(n)}{A_1(n) + A_2(n)}
\end{aligned}
$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

$$
\begin{aligned}
E\left[\Delta\frac{A_2^\epsilon}{A_1}(n+1) \mid \text{history}\right] = \\
& E\left[\Delta\frac{A_2^\epsilon}{A_1}(n+1) \mid 1, \text{history}\right] \Pr(1 \mid \text{history}) \\
& + \; E\left[\Delta\frac{A_2^\epsilon}{A_1}(n+1) \mid 2, \text{history}\right] \Pr(2 \mid \text{history})
\end{aligned}
$$

$$
\begin{aligned}
= E & \left[\frac{A_2^\epsilon(n)}{A_1(n)+\pi_1(n+1)} - \frac{A_2^\epsilon(n)}{A_1(n)}\right] \frac{A_1(n)}{A_1(n)+A_2(n)} \\
& + \; E\left[\frac{(A_2(n)+\pi_2(n+1))^\epsilon}{A_1(n)} - \frac{A_2^\epsilon(n)}{A_1(n)}\right] \frac{A_2(n)}{A_1(n)+A_2(n)} \\
& \leq \dots
\end{aligned}
$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Taylor expansion for, say, $n = 4$:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x) + \underbrace{\frac{h^4}{4!}f''''(x+\theta h)}_{\text{Lagrange remainder}}$$

for some $\theta \in (0, 1)$.

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Taylor expansion for, say, $n = 4$:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x) + \underbrace{\frac{h^4}{4!}f''''(x+\theta h)}_{\text{Lagrange remainder}}$$

for some $\theta \in (0, 1)$.

Applied to $f(x) = x^{-1}$ and $n = 2$ we obtain

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Taylor expansion for, say, $n = 4$:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x) + \underbrace{\frac{h^4}{4!}f''''(x+\theta h)}_{\text{Lagrange remainder}}$$

for some $\theta \in (0, 1)$.

Applied to $f(x) = x^{-1}$ and $n = 2$ we obtain

$$\begin{aligned}
(x+h)^{-1} &= x^{-1} + h(-x^{-2}) + \frac{h^2}{2!}(2(x+\theta h)^{-3}) \\
&= x^{-1} - hx^{-2} + h^2(x+\theta h)^{-3} \\
&= \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{(x+\theta h)^3}.
\end{aligned}$$

To show that A_2^{ϵ}/A_1 is a non-neg super-martingale

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Taylor expansion applied to $f(x) = x^{-1}$ and $n = 2$ yields

$$\begin{aligned}(x+h)^{-1} &= x^{-1} - hx^{-2} + h^2(x+\theta h)^{-3} \\ &= \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{(x+\theta h)^3}\end{aligned}$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Taylor expansion applied to $f(x) = x^{-1}$ and $n = 2$ yields

$$\begin{aligned}(x+h)^{-1} &= x^{-1} - hx^{-2} + h^2(x+\theta h)^{-3} \\ &= \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{(x+\theta h)^3}\end{aligned}$$

For non-negative x and h we have $x^3 \leq (x+\theta h)^3$ so that

$$\leq \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{x^3}.$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Taylor expansion applied to $f(x) = x^{-1}$ and $n = 2$ yields

$$\begin{aligned}(x+h)^{-1} &= x^{-1} - hx^{-2} + h^2(x+\theta h)^{-3} \\ &= \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{(x+\theta h)^3}\end{aligned}$$

For non-negative x and h we have $x^3 \leq (x+\theta h)^3$ so that

$$\leq \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{x^3}.$$

This first inequality puts an upper bound on the left-hand side:

$$\frac{1}{A_1(n) + \pi_1(n+1)}$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Taylor expansion applied to $f(x) = x^{-1}$ and $n = 2$ yields

$$\begin{aligned}(x+h)^{-1} &= x^{-1} - hx^{-2} + h^2(x+\theta h)^{-3} \\ &= \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{(x+\theta h)^3}\end{aligned}$$

For non-negative x and h we have $x^3 \leq (x+\theta h)^3$ so that

$$\leq \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{x^3}.$$

This first inequality puts an upper bound on the left-hand side:

$$\frac{1}{A_1(n) + \pi_1(n+1)} \leq$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Taylor expansion applied to $f(x) = x^{-1}$ and $n = 2$ yields

$$\begin{aligned}(x+h)^{-1} &= x^{-1} - hx^{-2} + h^2(x+\theta h)^{-3} \\ &= \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{(x+\theta h)^3}\end{aligned}$$

For non-negative x and h we have $x^3 \leq (x+\theta h)^3$ so that

$$\leq \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{x^3}.$$

This first inequality puts an upper bound on the left-hand side:

$$\frac{1}{A_1(n) + \pi_1(n+1)} \leq \frac{1}{A_1(n)}$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Taylor expansion applied to $f(x) = x^{-1}$ and $n = 2$ yields

$$\begin{aligned}(x+h)^{-1} &= x^{-1} - hx^{-2} + h^2(x+\theta h)^{-3} \\ &= \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{(x+\theta h)^3}\end{aligned}$$

For non-negative x and h we have $x^3 \leq (x+\theta h)^3$ so that

$$\leq \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{x^3}.$$

This first inequality puts an upper bound on the left-hand side:

$$\frac{1}{A_1(n) + \pi_1(n+1)} \leq \frac{1}{A_1(n)} - \frac{\pi_1(n+1)}{A_1^2(n)}$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Taylor expansion applied to $f(x) = x^{-1}$ and $n = 2$ yields

$$\begin{aligned}(x+h)^{-1} &= x^{-1} - hx^{-2} + h^2(x+\theta h)^{-3} \\ &= \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{(x+\theta h)^3}\end{aligned}$$

For non-negative x and h we have $x^3 \leq (x+\theta h)^3$ so that

$$\leq \frac{1}{x} - \frac{h}{x^2} + \frac{h^2}{x^3}.$$

This first inequality puts an upper bound on the left-hand side:

$$\frac{1}{A_1(n) + \pi_1(n+1)} \leq \frac{1}{A_1(n)} - \frac{\pi_1(n+1)}{A_1^2(n)} + \frac{\pi_1^2(n+1)}{A_1^3(n)}.$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Similarly, applying Taylor expansion to $f(x) = (x+h)^\epsilon$ with $n = 2$ yields

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Similarly, applying Taylor expansion to $f(x) = (x+h)^\epsilon$ with $n = 2$ yields

$$(x+h)^\epsilon = x^\epsilon + h\epsilon x^{\epsilon-1} + h^2(\epsilon-1)\epsilon(x+\theta h)^{\epsilon-2}.$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Similarly, applying Taylor expansion to $f(x) = (x+h)^\epsilon$ with $n = 2$ yields

$$(x+h)^\epsilon = x^\epsilon + h\epsilon x^{\epsilon-1} + h^2(\epsilon-1)\epsilon(x+\theta h)^{\epsilon-2}.$$

For non-negative x, $h \in [k_1, k_2]$, and $\epsilon > 1$, we have

$$(\epsilon-1)(x+\theta h)^{\epsilon-2}$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Similarly, applying Taylor expansion to $f(x) = (x+h)^\epsilon$ with $n = 2$ yields

$$(x+h)^\epsilon = x^\epsilon + h\epsilon x^{\epsilon-1} + h^2(\epsilon-1)\epsilon(x+\theta h)^{\epsilon-2}.$$

For non-negative x, $h \in [k_1, k_2]$, and $\epsilon > 1$, we have

$$(\epsilon - 1)(x+\theta h)^{\epsilon-2} \leq (\epsilon-1)(x+k_2)^{\epsilon-2}$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Similarly, applying Taylor expansion to $f(x) = (x+h)^\epsilon$ with $n = 2$ yields

$$(x+h)^\epsilon = x^\epsilon + h\epsilon x^{\epsilon-1} + h^2(\epsilon-1)\epsilon(x+\theta h)^{\epsilon-2}.$$

For non-negative x, $h \in [k_1, k_2]$, and $\epsilon > 1$, we have

$$(\epsilon-1)(x+\theta h)^{\epsilon-2} \leq (\epsilon-1)(x+k_2)^{\epsilon-2} \leq Cx^{\epsilon-2}$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Similarly, applying Taylor expansion to $f(x) = (x+h)^\epsilon$ with $n = 2$ yields

$$(x+h)^\epsilon = x^\epsilon + h\epsilon x^{\epsilon-1} + h^2(\epsilon-1)\epsilon(x+\theta h)^{\epsilon-2}.$$

For non-negative x, $h \in [k_1, k_2]$, and $\epsilon > 1$, we have

$$(\epsilon-1)(x+\theta h)^{\epsilon-2} \leq (\epsilon-1)(x+k_2)^{\epsilon-2} \leq Cx^{\epsilon-2}$$

for some constant C

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Similarly, applying Taylor expansion to $f(x) = (x+h)^\epsilon$ with $n = 2$ yields

$$(x+h)^\epsilon = x^\epsilon + h\epsilon x^{\epsilon-1} + h^2(\epsilon-1)\epsilon(x+\theta h)^{\epsilon-2}.$$

For non-negative x, $h \in [k_1, k_2]$, and $\epsilon > 1$, we have

$$(\epsilon-1)(x+\theta h)^{\epsilon-2} \leq (\epsilon-1)(x+k_2)^{\epsilon-2} \leq Cx^{\epsilon-2}$$

for some constant C, so that

$$(x+h)^\epsilon \leq x^\epsilon + h\epsilon x^{\epsilon-1} + h^2 C\epsilon x^{\epsilon-2}.$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Similarly, applying Taylor expansion to $f(x) = (x+h)^\epsilon$ with $n = 2$ yields

$$(x+h)^\epsilon = x^\epsilon + h\epsilon x^{\epsilon-1} + h^2(\epsilon-1)\epsilon(x+\theta h)^{\epsilon-2}.$$

For non-negative x, $h \in [k_1, k_2]$, and $\epsilon > 1$, we have

$$(\epsilon-1)(x+\theta h)^{\epsilon-2} \leq (\epsilon-1)(x+k_2)^{\epsilon-2} \leq Cx^{\epsilon-2}$$

for some constant C, so that

$$(x+h)^\epsilon \leq x^\epsilon + h\epsilon x^{\epsilon-1} + h^2 C\epsilon x^{\epsilon-2}.$$

This second inequality puts an upper bound on the left-hand side:

$$(A_2(n) + \pi_2(n+1))^\epsilon$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Similarly, applying Taylor expansion to $f(x) = (x+h)^\epsilon$ with $n = 2$ yields

$$(x+h)^\epsilon = x^\epsilon + h\epsilon x^{\epsilon-1} + h^2(\epsilon-1)\epsilon(x+\theta h)^{\epsilon-2}.$$

For non-negative x, $h \in [k_1, k_2]$, and $\epsilon > 1$, we have

$$(\epsilon-1)(x+\theta h)^{\epsilon-2} \leq (\epsilon-1)(x+k_2)^{\epsilon-2} \leq Cx^{\epsilon-2}$$

for some constant C, so that

$$(x+h)^\epsilon \leq x^\epsilon + h\epsilon x^{\epsilon-1} + h^2 C\epsilon x^{\epsilon-2}.$$

This second inequality puts an upper bound on the left-hand side:

$$(A_2(n) + \pi_2(n+1))^\epsilon \leq A_2^\epsilon(n)$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

Similarly, applying Taylor expansion to $f(x) = (x+h)^\epsilon$ with $n = 2$ yields

$$(x+h)^\epsilon = x^\epsilon + h\epsilon x^{\epsilon-1} + h^2(\epsilon-1)\epsilon(x+\theta h)^{\epsilon-2}.$$

For non-negative x, $h \in [k_1, k_2]$, and $\epsilon > 1$, we have

$$(\epsilon-1)(x+\theta h)^{\epsilon-2} \leq (\epsilon-1)(x+k_2)^{\epsilon-2} \leq Cx^{\epsilon-2}$$

for some constant C, so that

$$(x+h)^\epsilon \leq x^\epsilon + h\epsilon x^{\epsilon-1} + h^2 C\epsilon x^{\epsilon-2}.$$

This second inequality puts an upper bound on the left-hand side:

$$(A_2(n) + \pi_2(n+1))^\epsilon \leq A_2^\epsilon(n) + \ \ldots \ + \ etc.$$

(Take $x = A_2(n)$ and $h = \pi_2(n+1)$.)

To show that A_2^{ϵ}/A_1 is a non-neg super-martingale

To show that A_2^ϵ / A_1 is a non-neg super-martingale

- Using $E[aX + b] = aE[X] + b$ and factoring out common terms, we obtain

To show that A_2^ϵ / A_1 is a non-neg super-martingale

- Using $E[aX + b] = aE[X] + b$ and factoring out common terms, we obtain

$$\cdots \leq \frac{A_1}{A_1 + A_2} \frac{A_2^\epsilon}{A_1^2}(n) \left[-E[\pi_1(n+1)] + c_1 \frac{E[\pi_1(n+1)^2]}{A_1(n)} \right] + \frac{1}{A_1 + A_2} \frac{\epsilon A_2^\epsilon}{A_1}(n) \left[E[\pi_2(n+1)] + c_2 \frac{E[\pi_2(n+1)^2]}{A_2(n)} \right].$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

- Using $E[aX + b] = aE[X] + b$ and factoring out common terms, we obtain

$$\cdots \leq \frac{A_1}{A_1 + A_2} \frac{A_2^\epsilon}{A_1^2}(n) \left[-E[\pi_1(n+1)] + c_1 \frac{E[\pi_1(n+1)^2]}{A_1(n)} \right] + \frac{1}{A_1 + A_2} \frac{\epsilon A_2^\epsilon}{A_1}(n) \left[E[\pi_2(n+1)] + c_2 \frac{E[\pi_2(n+1)^2]}{A_2(n)} \right].$$

- Because payoffs are bounded, $E[\pi_1(\dots)] > \gamma E[\pi_2(\dots)]$, $1 - \gamma < \epsilon - \gamma < 0$, constants K_1, K_2, $K_3 > 0$ can be found such that

$$\cdots \leq \frac{A_2^\epsilon}{A_1(A_1 + A_2)} \left(K_1(\epsilon - \gamma) + \frac{K_2}{A_1} + \frac{K_3}{A_2} \right)(n) .$$

To show that A_2^ϵ / A_1 is a non-neg super-martingale

- Using $E[aX + b] = aE[X] + b$ and factoring out common terms, we obtain

$$\cdots \leq \frac{A_1}{A_1 + A_2} \frac{A_2^\epsilon}{A_1^2}(n) \left[-E[\pi_1(n+1)] + c_1 \frac{E[\pi_1(n+1)^2]}{A_1(n)} \right] + \frac{1}{A_1 + A_2} \frac{\epsilon A_2^\epsilon}{A_1}(n) \left[E[\pi_2(n+1)] + c_2 \frac{E[\pi_2(n+1)^2]}{A_2(n)} \right].$$

- Because payoffs are bounded, $E[\pi_1(\dots)] > \gamma E[\pi_2(\dots)]$, $1 - \gamma < \epsilon - \gamma < 0$, constants K_1, K_2, $K_3 > 0$ can be found such that

$$\cdots \leq \frac{A_2^\epsilon}{A_1(A_1 + A_2)} \left(K_1(\epsilon - \gamma) + \frac{K_2}{A_1} + \frac{K_3}{A_2} \right)(n) .$$

- For $\epsilon \in (1, \gamma)$ and for n large enough, this expression is non-positive. $\square$

Generalisation of Begg's Theorem 1

Let there be $m \geq 2$ alternative actions, $a_1, \ldots, a_m$ (rather than $m = 2$).

Generalisation of Begg's Theorem 1

Let there be $m \geq 2$ alternative actions, $a_1, \ldots, a_m$ (rather than $m = 2$).

Theorem 2. *If the expected payoff (conditional on the history) of* a_i *dominates the expected payoff (conditional on the history) of* a_j*, for all* $j \neq i$*, then the probability that* a_j *will be played converges to zero, for all* $j \neq i$.

Generalisation of Begg's Theorem 1

Let there be $m \geq 2$ alternative actions, $a_1, \ldots, a_m$ (rather than $m = 2$).

Theorem 2. *If the expected payoff (conditional on the history) of a_i dominates the expected payoff (conditional on the history) of a_j, for all $j \neq i$, then the probability that a_j will be played converges to zero, for all $j \neq i$.*

Applied to games:

Generalisation of Begg's Theorem 1

Let there be $m \geq 2$ alternative actions, $a_1, \ldots, a_m$ (rather than $m = 2$).

Theorem 2. *If the expected payoff (conditional on the history) of a_i dominates the expected payoff (conditional on the history) of a_j, for all $j \neq i$, then the probability that a_j will be played converges to zero, for all $j \neq i$.*

Applied to games:

Theorem 3. *In a game with finitely many actions and players, if a player learns according the ER scheme then,*

Generalisation of Begg's Theorem 1

Let there be $m \geq 2$ alternative actions, $a_1, \dots, a_m$ (rather than $m = 2$).

Theorem 2. *If the expected payoff (conditional on the history) of a_i dominates the expected payoff (conditional on the history) of a_j, for all $j \neq i$, then the probability that a_j will be played converges to zero, for all $j \neq i$.*

Applied to games:

Theorem 3. *In a game with finitely many actions and players, if a player learns according the ER scheme then,*

1. *With probability 1, the probability and empirical frequency that he plays any action that is strictly dominated by another pure strategy converges to zero.*

Generalisation of Begg's Theorem 1

Let there be $m \geq 2$ alternative actions, $a_1, \ldots, a_m$ (rather than $m = 2$).

Theorem 2. *If the expected payoff (conditional on the history) of a_i dominates the expected payoff (conditional on the history) of a_j, for all $j \neq i$, then the probability that a_j will be played converges to zero, for all $j \neq i$.*

Applied to games:

Theorem 3. *In a game with finitely many actions and players, if a player learns according the ER scheme then,*

1. *With probability 1, the probability and empirical frequency that he plays any action that is strictly dominated by another pure strategy converges to zero.*

2. *Hence if he has a strictly dominant strategy, with probability 1, the probability and empirical frequency with which he plays that action converges to 1.*

Generalisation of Begg's Theorem 1

Let there be $m \geq 2$ alternative actions, $a_1, \ldots, a_m$ (rather than $m = 2$).

Theorem 2. *If the expected payoff (conditional on the history) of a_i dominates the expected payoff (conditional on the history) of a_j, for all $j \neq i$, then the probability that a_j will be played converges to zero, for all $j \neq i$.*

Applied to games:

Theorem 3. *In a game with finitely many actions and players, if a player learns according the ER scheme then,*

1. *With probability 1, the probability and empirical frequency that he plays any action that is strictly dominated by another pure strategy converges to zero.*
2. *Hence if he has a strictly dominant strategy, with probability 1, the probability and empirical frequency with which he plays that action converges to 1.*

(Beggs, 2005).

Summary

- There are several rules for reinforcement learning on single states.
- Sheer convergence is often easy to prove.
- Proving convergence to best actions in a stationary environment is more difficult.
- Convergence to best actions in non-stationary environments, e.g., convergence to dominant actions, or best responses in self-play, is state-of-the art research.

What next?

What next?

- **No-regret learning**: this is a generalisation of reinforcement learning

  No-regret $=_{Def}$ play those actions that would have been successful in the past.

What next?

- **No-regret learning**: this is a generalisation of reinforcement learning

  No-regret $=_{Def}$ play those actions that would have been successful in the past.

- **Similarities** with reinforcement learning:

What next?

- **No-regret learning**: this is a generalisation of reinforcement learning

  No-regret $=_{Def}$ play those actions that would have been successful in the past.

- **Similarities** with reinforcement learning:

  1. Driven by past payoffs.

What next?

- **No-regret learning**: this is a generalisation of reinforcement learning

  No-regret $=_{Def}$ play those actions that would have been successful in the past.

- **Similarities** with reinforcement learning:

  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.

What next?

- **No-regret learning**: this is a generalisation of reinforcement learning

  No-regret $=_{Def}$ play those actions that would have been successful in the past.

- **Similarities** with reinforcement learning:

  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Myopic.

What next?

- **No-regret learning**: this is a generalisation of reinforcement learning

  No-regret $=_{Def}$ play those actions that would have been successful in the past.

- **Similarities** with reinforcement learning:

  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Myopic.

- **Differences**:

What next?

- **No-regret learning**: this is a generalisation of reinforcement learning

  No-regret $=_{Def}$ play those actions that would have been successful in the past.

- **Similarities** with reinforcement learning:

  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Myopic.

- **Differences**:

  1. No-regret learning als learns from hypothetical payoffs.

What next?

- **No-regret learning**: this is a generalisation of reinforcement learning

  No-regret $=_{Def}$ play those actions that would have been successful in the past.

- **Similarities** with reinforcement learning:

  1. Driven by past payoffs.
  2. Not interested in (the behaviour of) the opponent.
  3. Myopic.

- **Differences**:

  1. No-regret learning als learns from hypothetical payoffs.
  2. It is more easy to obtain results regarding performance.