Reinforcement Learning

Based on a simple principle:

More likely to repeat an action, if it had to a positive outcome.

Reinforcement Learning

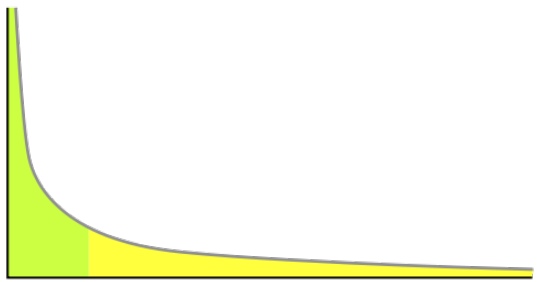
- Idea of reinforcement learning first formulated by psychologists, e.g. Bush and Mosteller (1955)
- 2 ideas from psychological learning theory
  - *law of effect*
  - *power law of practise*
- Researched by computers science and engineering (method for training AI)
- Formulated in economics by Roth/Erev (1993) and Börgers/Sarin (1997)

Reinforcement Learning: Law of Effect

- Edward Thorndike (1898): “puzzle boxes”
- Cats put into box with closed door and a release mechanism
- Thorndike measured the time it took the cats to open the door
- Cats who opened the door before, opened it faster in subsequent attempts
- **Law of effect:** Actions that produce a positive outcome are used more often in the same situation in the future

Reinforcement Learning: Power law of practise

- How long does it take to solve a problem on the first trial? On the second? The third? etc
- Statement about the functional form of the *learning curve*



- $T = aP^{-b}$
- T: Time taken, P: Trial number, a, b constants
- **Power law of practise:** Learning curve first steep, then becomes flat
- Note: Other shapes possible. Thorndike found learning curves that were flat-steep-flat

Reinforcement Learning: Erev/Roth Model

- Developed by Erev and Roth (1995, 1998) to explain experimental data
- $t = 1$ first (of many) round played
- n players with j pure strategies
- $q_{nk}(t)$ *propensity* of player n to play his strategy k in round t
  - $q_{nk}(1)$ *initial propensity*
- Updating of propensities: $q_{nk}(t+1) = q_{nk}(t) + x$
  - If strategy k was played and x was the payoff received
  - propensity unchanged, $q_{nk}(t+1) = q_{nk}(t)$, if k was not played

Reinforcement Learning: Erev/Roth Model

- From propensities to probabilities:
- $p_{nk}(t) = q_{nk}(t) / \sum q_{nj}(t)$
  - probability to play a strategy equal to its relative propensity
- Law of effect is observed: More successful strategies (higher x) are played more often
- Learning curve also steeper in early rounds:
  - $\sum q_{nj}(t)$ is an increasing function of t, so a payoff x has a bigger effect on $p_{nk}(t)$ when t is small
- Strength of initial propensities is the only parameter of the (basic) RF model
- Note that RF includes a stochastic element

Reinforcement Learning: Convergence

- Assumed: payoffs $x(k)$ bounded away from zero and infinity. All players use basic Reinforcement Learning
- Results (from Begg (2004)):
  - Each strategy is chosen infinitly often with probability 1
  - If strategy *a* strictly dominates strategy *b*, then with probability 1, the probability that the decision-maker plays strategy *a* converges to 1
  - In two-person constant-sum games, players average payoffs converge to the value of the game
- Convergence to EQ if:
  - Constant sum game + unique pure strategy equilibrium
  - 2*x*2 game + unique-mixed strategy equilibrium

- Suppose to the contrary that strategy k is chosen only a finite number of times.
- There must exist a finite time τ when k is chosen for the last time.
- To show contradiction, calculate probability that k is *not* chosen after period τ and prove that this probability is zero.

- Suppose to the contrary that strategy k is chosen only a finite number of times.
- There must exist a finite time τ when k is chosen for the last time.
- To show contradiction, calculate probability that k is *not* chosen after period τ and prove that this probability is zero.

The probability that k is chosen in period $\tau + t$, $t = 1, 2, ...$ is bounded below by

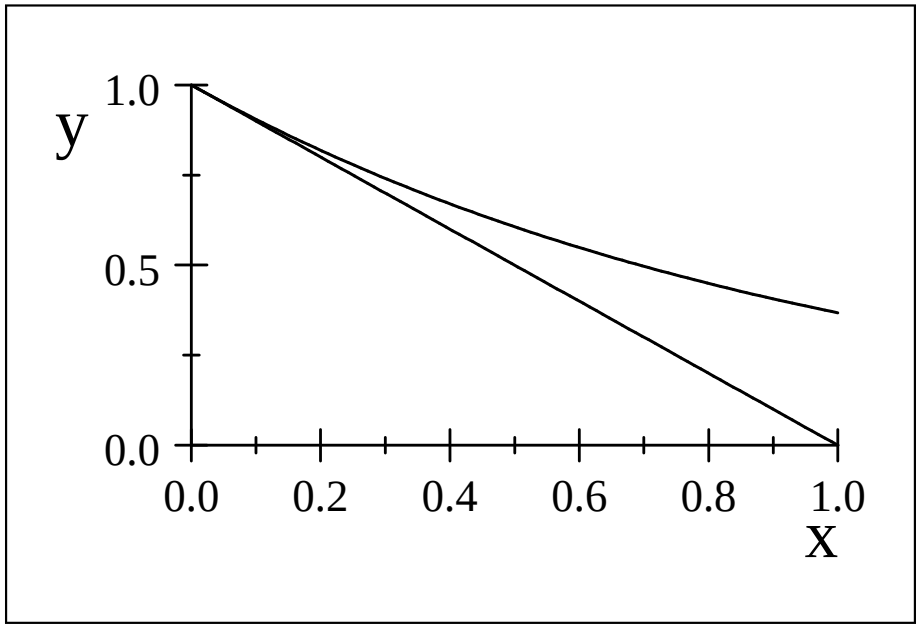
$$p_{nk}(\tau + t) \geq \frac{q_{nk}(\tau)}{\sum_j q_{nj}(\tau) + t\pi^{\max}} > \frac{1}{t} \frac{q_{nk}(\tau)}{\sum_j q_{nj}(\tau) + \pi^{\max}} =: \frac{c}{t} > 0.$$

Thus, it suffices to prove that

$$\lim_{T \to \infty} \prod_{t=1}^{T} \left(1 - \frac{c}{t}\right) = 0. \tag{1}$$

$$\lim_{T\to\infty}\prod_{t=1}^{T}\left(1-\frac{c}{t}\right)=0. \tag{1}$$

Since $1-x\leq e^{-x}, x\in[0,1]$



$$\prod_{t=1}^{T}\left(1-\frac{c}{t}\right)\leq\prod_{t=1}^{T}\exp\left(-\frac{c}{t}\right)=\exp\left(-\sum_{t=1}^{T}\frac{c}{t}\right).$$

Since the sum diverges as $T\longrightarrow\infty$, (1) follows.

Reinforcement Learning: Modifications

Extensions to the basic model, introduce more parameters (problem?)

- *Cutoff parameter*: lets probability for strategies become zero
- *Local experimentation*: reinforces strategies close to a succesful strategy
- *Forgetting*: Total propensities depreciate over time

Reinforcement Learning: Cutoff parameter

- Cutoff parameter μ: probabilities smaller than μ (and the associated propensities) are set equal to 0
- Reasoning: Probabilities that are extremely small are indistinguishable from 0
- Speeds up convergence (now in finite time)
  - Convergance of basic model can be very slow (in some simulations: not converged after 10.000s of rounds)
- Process can now converge even to strictly dominated strategies, if this strategy is chosen by chance sufficiently often

Reinforcement Learning: Local experimentation

- Also called *Generalization*, after one of the principles by B.F. Skinner
- Out of the total payoff x, only a share $1 - \epsilon$ is added to the played strategy
- The remainder ϵ is instead added to "close" strategies
- Interpretation of ϵ: Local experimentation or errors
- Requires strategies to be ordered in some meaningful way (e.g. interpreted along one dimension: price, quantities, amount given to other player, etc)

Reinforcement Learning: Forgetting

- Also called *Recency*:
  - Watson's law of recency (1930): *The response that has most recently occurred after a particular stimulus is the response most likely to be associated with that stimulus.*
- At the end of a round, each propensity is multiplied with $1 - \phi$ (were ϕ is small)
- Puts an upper bound on how big the sum of all propensities can become
- Assures that the most recent observation never becomes completely irrelevant to the overall process

Reinforcement Learning: Experimental test

- Data from 3 games:
  - Ultimatum game
  - Market game
  - Best shot game
- Data was collected for previous papers (Roth et al (1991); Prasnikar and Roth (1992)) and analysed again with Reinforcement Learning
- Each game was played 10 times, against different opponents
- Each game has a similar, one-sided subgame perfect Nash-EQ prediction

Reinforcement Learning: Ultimatum Game

- Proposer: Divides a pie of 1000 tokens; minimal step size 5 tokens
- Responder: Accepts division, or rejects. If reject, both earn 0
- Subgame perfect Nash-EQ
  - Proposer keeps 995 or 1000 tokens, offers 5 or 0
  - Responder accepts all offers/all offers but 0
- Note that any offer can be part of a non-perfect Nash-EQ

Reinforcement Learning: Market Game

- One seller: Sells 1 indivisible good
- (up to) 9 buyers: Make simultaneous offers for the good
- Seller: Accepts highest offer, or rejects
  - If seller rejects highest offer, everyone earns 0
  - If seller accepts, he earns highest offer, the buyer earns 1000 - offer, all others earn 0
  - stepsize is again 5
- Similar subgame perfect Nash-EQ: (at least 2) Buyers offer 995 or 1000, seller accepts offers of 995/1000

Reinforcement Learning: Best shot game

- 2 player investment game
- player 1 acts first by providing a quantity q_1, then player 2 sees this and provides q_2
- Payoff for both players depends on the maximum individual quantity offered $q = max\{q_1, q_2\}$
- Marginal cost function for providing quantity is increasing, marginal profit function is decreasing $\rightarrow$ if only 1 player, optimal to provide a quantity of 4
- Subgame perfect Nash-EQ: Player 1 provides 0, player 2 provides 4
- Payoff in EQ: Player 1 gets \$3.70, player 2 gets \$0.42

Reinforcement Learning: Experimental test

- In each game, subgame perfect Nash-EQ is one-sided: One player gets a lot more than the other player(s)
- What is your prediction of what happened in the 3 experiments?

Reinforcement Learning: Experimental test

- In each game, subgame perfect Nash-EQ is one-sided: One player gets a lot more than the other player(s)
- What is your prediction of what happened in the 3 experiments?
- **Ultimatum game**: First round, modal division of $(500, 500)$, last round similar (in some countries a bit higher $(600, 400)$)

Reinforcement Learning: Experimental test

- In each game, subgame perfect Nash-EQ is one-sided: One player gets a lot more than the other player(s)
- What is your prediction of what happened in the 3 experiments?
- **Ultimatum game**: First round, modal division of $(500, 500)$, last round similar (in some countries a bit higher $(600, 400)$)
- **Market game**: First round, offers dispersed, but by round 10, EQ is reached (40-70% of buyers offer EQ price)

Reinforcement Learning: Experimental test

- In each game, subgame perfect Nash-EQ is one-sided: One player gets a lot more than the other player(s)
- What is your prediction of what happened in the 3 experiments?
- **Ultimatum game**: First round, modal division of $(500, 500)$, last round similar (in some countries a bit higher $(600, 400)$)
- **Market game**: First round, offers dispersed, but by round 10, EQ is reached (40-70% of buyers offer EQ price)
- **Best shot game**: Similar to market game. Players learn quickly to not both provide positive quantities, convergance to EQ by round 10.

Reinforcement Learning: Simulations

- Can Reinforcement Learning explain the findings?
- Simulations are run on the 3 games (in a simplified version)
- Each simulation has randomly initial propensities and runs 1000 rounds
- Initial propensities are normalized such that the total sum of initial propensities (and therefore the "weight" of the initial propensities vs actual experience) is the same
- 3 models tested (all lead to mostly similar results)
  - basic model + cutoff parameter
  - basic model + local experimentation
  - basic model + local experimentation + forgetting

Reinforcement Learning: Simulations

- Results of the simulations

Reinforcement Learning: Simulations

- Results of the simulations
- **Market game**: Simulations converge rapidly to EQ

Reinforcement Learning: Simulations

- Results of the simulations
- **Market game**: Simulations converge rapidly to EQ
- **Best shot game**: Player 1 quickly places most probability weight on EQ provision of 0, player 2 moves more slowly into direction of EQ quantity (providing 4)

Reinforcement Learning: Simulations

- Results of the simulations
- **Market game**: Simulations converge rapidly to EQ
- **Best shot game**: Player 1 quickly places most probability weight on EQ provision of 0, player 2 moves more slowly into direction of EQ quantity (providing 4)
- **Ultimatum game**: Behavior has not converged to EQ after 1000 rounds
  - Even when the simulation is extended to last 1,000,000 rounds, only the model with forgetting consistently converges to EQ

Reinforcement Learning: Simulations

- What prevents convergence from happening in the ultimatum game?

Reinforcement Learning: Simulations

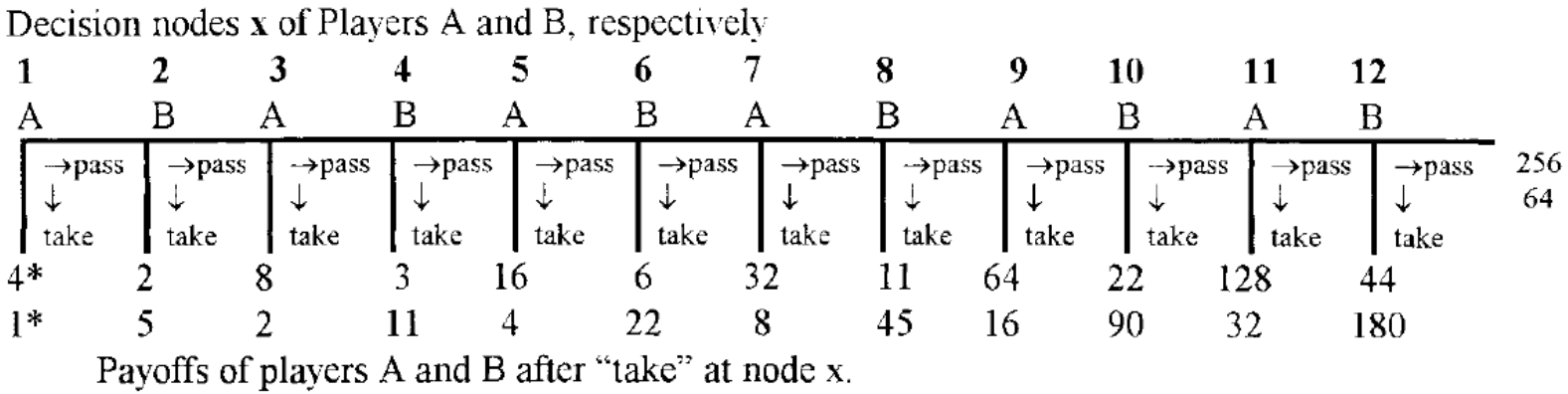
- What prevents convergence from happening in the ultimatum game?
- Player 2 only gets a very mild “penalty” for rejecting very bad offers: E.g. accepting an offer of 5 only raises the propensity of that strategy by 5 (compared to a raise of 0 for the “wrong” strategy of rejecting that offer)
- So player 2 learns only very slowly to accept bad offers
- Player 1 is faced with player 2, who rejects very bad offers for a long time. How does player 1 react to that?

Reinforcement Learning: Simulations

- What prevents convergence from happening in the ultimatum game?
- Player 2 only gets a very mild “penalty” for rejecting very bad offers: E.g. accepting an offer of 5 only raises the propensity of that strategy by 5 (compared to a raise of 0 for the “wrong” strategy of rejecting that offer)
- So player 2 learns only very slowly to accept bad offers
- Player 1 is faced with player 2, who rejects very bad offers for a long time. How does player 1 react to that?
- Strong penalty for player 1 for very bad offer that is rejected, compared to middle offer that is accepted: Rejection raises propensity by 0, accepted middle offer (e.g. of 500) raises propensity of that strategy by 500
- $\rightarrow$ Player 1 learns faster to offer fair amounts than player 2 learns to accept unfair offers $\rightarrow$ simulations stay away from EQ for a long time

Reinforcement Learning: Centipede game

- Nagel and Tang (1998)
- Use a centipede game to compare several learning theories and EQ concepts



- Presented as reduced normal form matrix
- 100 periods, random matching

Reinforcement Learning: Centipede game

Type	Model	Initial value	Parameters	Parameter values	QDM (avg.)
Static models	Equilibrium	Stage game equilibrium	—	—	1.99
	Quantal response	Individual frequency distribution	λ	0.1745	0.79
	Random	(1/7, ..., 1/7)	—	—	0.86
	Mean	Individual frequency distribution			0.54
Individualistic models	RPS	50*1/7	q	0.90	0.57
	Power-RPS	50*1/7	(r, q)	(0.58, 0.8)	0.56
	Exp.-RPS	50*1/7	(λ, q)	(0.018, 0.85)	0.60
Population model	Generalized fictitious play	(1/7, ..., 1/7)	δ	0.95	1.32

Reinforcement Learning

Some notes:

- Other papers comparing RL and FP, not all find FP worse
- ... many different ways to implement both RL and FP

Reinforcement Learning

Some notes:

- Other papers comparing RL and FP, not all find FP worse
- ... many different ways to implement both RL and FP
- Extension by Erev Roth that includes FP as a special case
- We look at EWA model instead (next week), which also includes RL and FP as special case

Reinforcement Learning

Some notes:

- Other papers comparing RL and FP, not all find FP worse
- ... many different ways to implement both RL and FP
- Extension by Erev Roth that includes FP as a special case
- We look at EWA model instead (next week), which also includes RL and FP as special case
- We also skip extensions that include negative reinforcement or a reference point

Börgers and Sarin (1997, JET)

- variant of Bush, Mosteller (1951) (aka Cross model)

Börgers and Sarin (1997, JET)

- variant of Bush, Mosteller (1951) (aka Cross model)
- no propensities; payoffs influence probablities directly (assume payoffs $0 < \pi_j < 1$)

$$
\begin{aligned}
p_j(t) &= (1 - \pi_j(t))\, p_j(t-1) + \pi_j(t) && \text{if strat. } j \text{ chosen} \\
p_j(t) &= (1 - \pi_k(t))\, p_j(t-1) && \text{if strat. } k \neq j \text{ chosen}
\end{aligned}
$$

Börgers and Sarin (1997, JET)

- variant of Bush, Mosteller (1951) (aka Cross model)
- no propensities; payoffs influence probablities directly (assume payoffs $0 < \pi_j < 1$)

$$
\begin{aligned}
p_j(t) &= (1 - \pi_j(t))\, p_j(t-1) + \pi_j(t) \quad \text{if strat. } j \text{ chosen} \\
p_j(t) &= (1 - \pi_k(t))\, p_j(t-1) \quad \text{if strat. } k \neq j \text{ chosen}
\end{aligned}
$$

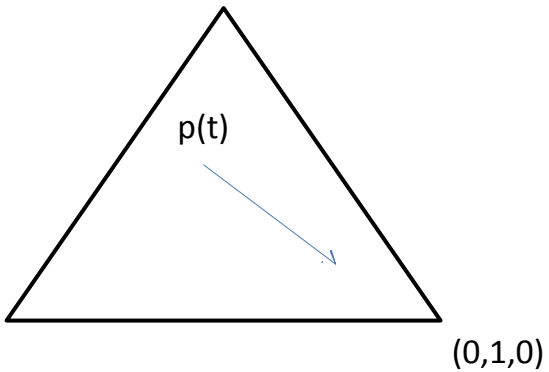
- convex combination of current prob. vector and $(0, ..., 1, ...0)$ (chosen strategy)

Börgers and Sarin (1997, JET)

- variant of Bush, Mosteller (1951) (aka Cross model)
- no propensities; payoffs influence probablities directly (assume payoffs $0 < \pi_j < 1$)

$$
\begin{aligned}
p_j(t) &= (1 - \pi_j(t))\, p_j(t-1) + \pi_j(t) \quad \text{if strat. } j \text{ chosen} \\
p_j(t) &= (1 - \pi_k(t))\, p_j(t-1) \quad \text{if strat. } k \neq j \text{ chosen}
\end{aligned}
$$

- convex combination of current prob. vector and $(0, ..., 1, ...0)$ (chosen strategy)
- e.g. if strat. 2 is played, length of vector determined by π_2



- Main result: in the cont. time limit, behaves like replicator dynamic

- Main result: in the cont. time limit, behaves like replicator dynamic
- often used as justification for replicator dynamic

- Main result: in the cont. time limit, behaves like replicator dynamic
- often used as justification for replicator dynamic
- doesn't work for Roth/Erev model

- Main result: in the cont. time limit, behaves like replicator dynamic
- often used as justification for replicator dynamic
- doesn't work for Roth/Erev model
- main differences: B-S model doesn't satisfy power law of practice

- Main result: in the cont. time limit, behaves like replicator dynamic
- often used as justification for replicator dynamic
- doesn't work for Roth/Erev model
- main differences: B-S model doesn't satisfy power law of practice
- sketch of argument:

- Main result: in the cont. time limit, behaves like replicator dynamic
- often used as justification for replicator dynamic
- doesn't work for Roth/Erev model
- main differences: B-S model doesn't satisfy power law of practice
- sketch of argument:
- Replicator dynamics:

$$\dot{p}_j = p_j \left[\pi(s_j, q) - \pi(p, q) \right]$$

(p, q relative frequencies of strat. in pop. 1 and 2, resp.)

- B-S model implies

$$\begin{aligned}
\Delta p_j(t) &= p_j(t) - p_j(t-1) \\
&= (1 - \pi_j(t))\, p_j(t-1) + \pi_j(t) - p_j(t-1) \\
&= (1 - p_j(t-1))\pi_j(t) \qquad \text{if strat. } j \text{ chosen}
\end{aligned}$$

- B-S model implies

$$\begin{aligned}\Delta p_j(t) &= p_j(t) - p_j(t-1) \\ &= (1-\pi_j(t))\, p_j(t-1) + \pi_j(t) - p_j(t-1) \\ &= (1-p_j(t-1))\pi_j(t) \qquad \text{if strat. } j \text{ chosen}\end{aligned}$$

- cont. time limit

$$\begin{aligned}\dot{p}_j &= (1-p_j)\pi_j \qquad \text{if strat. } j \text{ chosen} \\ \dot{p}_j &= -p_j\pi_k \qquad \text{if strat. } k \text{ chosen}\end{aligned}$$

- B-S model implies

$$\begin{aligned}\Delta p_j(t) &= p_j(t) - p_j(t-1) \\ &= (1-\pi_j(t))\, p_j(t-1) + \pi_j(t) - p_j(t-1) \\ &= (1-p_j(t-1))\pi_j(t) \qquad \text{if strat. } j \text{ chosen}\end{aligned}$$

- cont. time limit

$$\begin{aligned}\dot{p}_j &= (1-p_j)\pi_j \quad \text{if strat. } j \text{ chosen} \\ \dot{p}_j &= -p_j\pi_k \quad \text{if strat. } k \text{ chosen}\end{aligned}$$

- In expectation:

$$\begin{aligned}E(\dot{p}_j) &= p_j E\left[\dot{p}_j | j \text{ chosen}\right] + (1-p_j) E\left[\dot{p}_j | j \text{ not chosen}\right] \\ &= p_j E\left[(1-p_j)\pi_j | j\right] + (1-p_j) E\left[-p_j\pi_k | \neg j\right]\end{aligned}$$

$$E(\dot{p}_j) = p_j \left[(1 - p_j)\pi(s_j, q) - (1 - p_j)E\left[\pi_k | \neg j\right] \right]$$

- by the law of total prob.:

$$E\left[\pi_k | \neg j\right] = \frac{\pi(p, q) - p_j \pi(s_j, q)}{1 - p_j}$$

$$E(\dot{p}_j) = p_j \left[(1 - p_j)\pi(s_j, q) - (1 - p_j)E\left[\pi_k | \neg j\right] \right]$$

- by the law of total prob.:

$$E\left[\pi_k | \neg j\right] = \frac{\pi(p, q) - p_j \pi(s_j, q)}{1 - p_j}$$

- Hence

$$E(\dot{p}_j) = p_j \left[\pi(s_j, q) - \pi(p, q) \right]$$

$$E(\dot{p}_j) = p_j \left[(1 - p_j)\pi(s_j, q) - (1 - p_j)E\left[\pi_k | \neg j\right] \right]$$

- by the law of total prob.:

$$E\left[\pi_k | \neg j\right] = \frac{\pi(p, q) - p_j \pi(s_j, q)}{1 - p_j}$$

- Hence

$$E(\dot{p}_j) = p_j \left[\pi(s_j, q) - \pi(p, q) \right]$$

- Some more work to show that actual movement is close to replication dynamic (see B-S paper).

Reinforcement Learning: Conclusion

Positive:

- Intuitive, based on principles known from psychology
- Only 1 parameter (in the basic model)
- Predicts reasonably well
- Information/Calculation requirements very small

Negative:

- Many extensions make model adaptable, but prediction?
- Convergence can take very long, not always guaranteed. Authors suggest to study “intermediate” time frame
- Not always clear how to set the parameter of the model