Simple models of reinforcement learning...

Matteo Marsili
Abdus Salam ICTP, Trieste, Italy

Take home message

- Modeling of competition in large populations (ecology, traffic, financial markets, etc)
- Evolution population = individual learning
- Individuals are not small, even in large populations
- Symmetry of Nash equilibria (specialization):
  Specialization requires minimal form of rationality (account for your impact)
  Specialization expected in crowded environments
- Warning: lots of maths, very little ecology...

The setting

- N agents $\underset{\text{constraints}}{\overset{\text{preferences}}{\rightleftarrows}}$ P resources

```
E.g. financial markets (speculators and information patterns),
     commodity markets (buyers and sellers),
     production economies (firms and goods' markets),
     urban traffic (vehicles and streets),
     internet traffic (surfers and Internet),
     ecosystems (species and resources), ...
```

adaptive agents = preferences + bounded rationality
noise = experimentation, environmental volatility, ...

- Question:

$$\text{collective behavior} = F\left(\text{structure of preferences},\ \text{degree of rationality},\ \ldots\right)$$

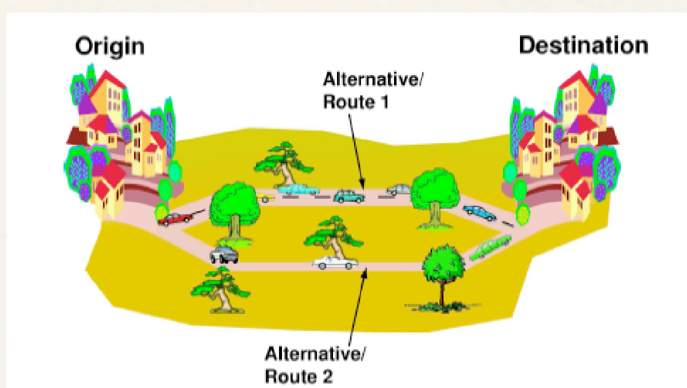
Order - disorder transition:
symmetric (mixed strat.)
asymmetric (pure strat.)

Identical/heterogeneous
agents and/or resources

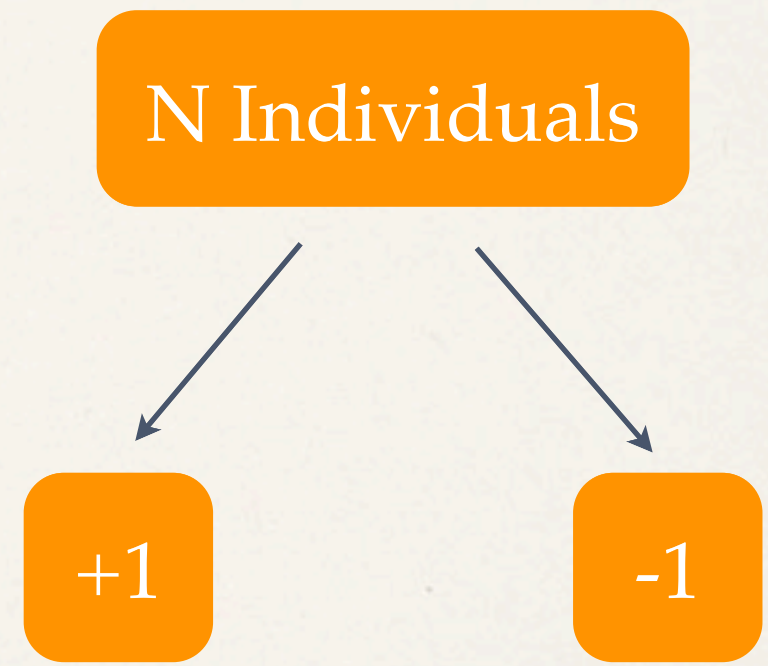
Learning, memory,...

Two equivalent resources

- One species, two resources
- Drivers, two roads (D. Helbing)



- Traders, two stocks
- Game theory:
  Payoff matrix (2x2 game)
  Pairwise random matching (Nx2 game)



	+1	-1
+1	0	1
-1	1	0

Modeling competition: Evolution

- A population of N_s individuals is associated to each strategy $s=\pm 1$

- Fitness = payoff in game with random opponent
  $dN_s = u_s N_s dt$ $\quad$ $u_s = 1 - n_s$ $\quad$ $n_s = N_s/N$

- … replicator dynamics… $\quad$ Lyapunov function … $\Rightarrow n_s \to 1/2$

$$\begin{aligned} \dot{n}_s &= n_s \left[u_s - \sum_{s'} u_{s'} n_{s'} \right] \\ \dot{n}_+ &= n_+(1-n_+)(1-2n_+) \end{aligned}$$

$$\begin{aligned} H &= \frac{1}{2}(1-2n_+)^2 \\ \dot{H} &= \frac{dH}{dn_+}\dot{n}_+ \\ &= -2n_+(1-n_+)(1-2n_+)^2 \leq 0 \end{aligned}$$

- J. Maynard-Smith *Evolution and the theory of games*, Cambridge (1982)
  J. W. Weibull, *Evolutionary game theory*, MIT (1995) (imitation ~ evolution)

Modeling competition: Learning

- Reinforcement learning
- N (fixed) individuals
- Each individual attach a score $U_{s,i}$ to each resource:
  Prior beliefs: $U_{s,i}(0)=0$
  Reward resource depending on payoff: $U_{s,i}(t+dt) = U_{s,i}(t) + 1-n_s$
  Choose resource with highest score: $P\{s_i(t)=s\} \sim \exp[\Gamma U_{s,i}(t)], \quad \Gamma>0$
- ... $P\{s_i(t)=s\}$ follows same dynamics as $n_s = N_s/N$ in replicator dynamics
  $\Rightarrow n_s \rightarrow 1/2$ (individuals learn to flip coins!)
- Fudenberg, Levine *The theory of learning in games* MIT (1998)
  Rustichini, Games and Econ. Behav., **29**, 244-273 (1999).

Back to game theory

- N=2
- Symmetric Nash equilibrium:
  P{s=+1}=1/2 (mixed strategy)
- Asymmetric Nash equilibrium:
  s_1=+1, s_2=-1 or s_1=+1, s_2=-1 (pure strategies)
- N individuals (random matching): $u_s = 1 - n_s$
  1 symmetric Nash equilibrium: P{s_i=+1}=1/2 $\forall i$
  Exponentially many asymmetric Nash equilibria!

	+1	-1
+1	0	1
-1	1	0

Questions

- Efficiency:
  symmetric NE: sqrt(N) individuals make the wrong choice!
  asymmetric NE: at most one in worse resource (but can't do better)

- Why do individual fail to learn the optimal NE?
  Account for yourself when learning (counterfactual)!

$$\begin{aligned} U_{s,i}(t+1) &= U_{s,i}(t) + \Gamma\left[1 - n_s(t)\right] & \text{if } s_i(t) = s \\ &= U_{s,i}(t) + \Gamma\left[1 - n_s(t) - 1/N\right] & \text{if } s_i(t) \neq s \end{aligned}$$

A small term with big consequences

- P{s_i(t)=+1}=p_i(t)

$$p_i = \frac{e^{\Gamma U_{+,i}}}{e^{\Gamma U_{+,i}} + e^{\Gamma U_{-,i}}}$$

- Learning

$$\dot{U}_{+,i} = \Gamma \left[1 - \frac{1}{N} \sum_{j \neq i} p_j \right]$$

- Lyapunov function:

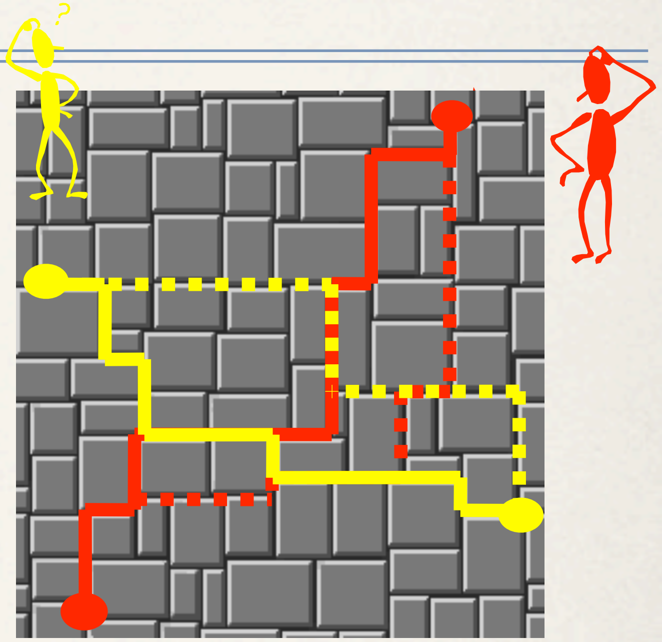
$$H_1 = \frac{1}{2} \left[\frac{1}{N} \sum_i (1 - 2p_i) \right]^2 - \frac{1}{2N^2} \sum_i (1 - 2p_i)^2$$

- Note: 2nd term gets relevant when $p_i \simeq \frac{1}{2} \pm \epsilon$ (late stage of dynamics)

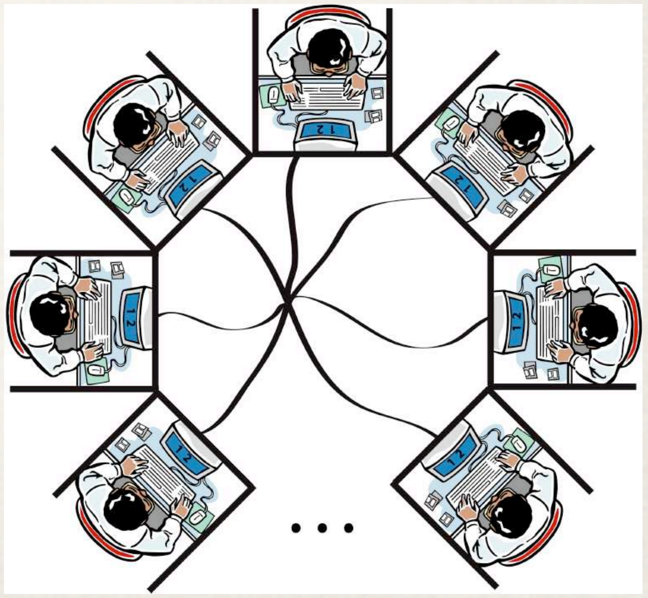
- Minima of H_1: p_i=0 for half of i's and 1 for the rest! $(\nabla^2 H_1 = 0)$

Extensions

- Holds for any N, even for N→∞
- For any number P of resources and heterogenous agents (N,P→∞)
  (Challet, MM, Zhang *Minority Games*, Oxford 2005)
- For any population playing a symmetric game with equilibria in mixed strategy
  (why are there bakers, plumbers, researchers, ...?)
  (Borkar, Jain, Rangarajan, *Complexity* **3**, 50-56, 1998)
- Verified in experiments of route choice games
  (Helbing, Shonhof and Kern, New J. Phys. 4, 33.1-33.16 2002)



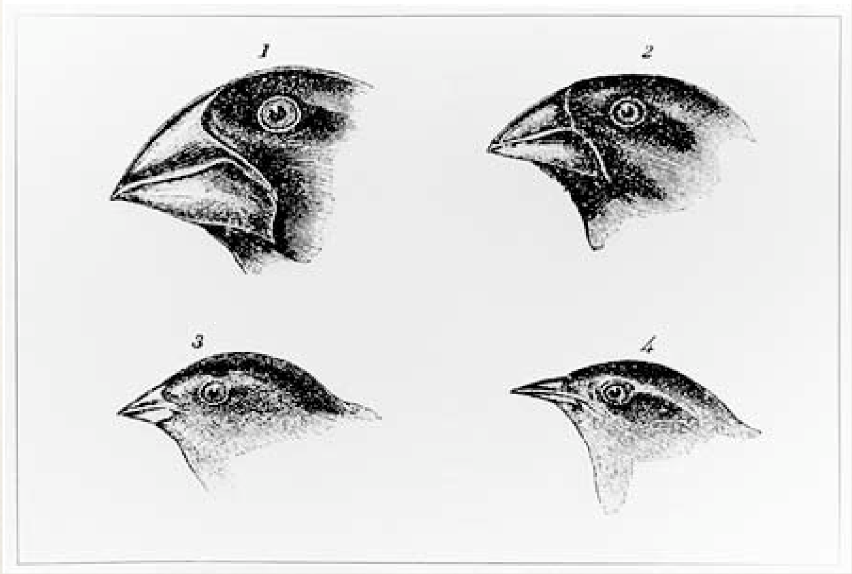
Dirk Helbing



possibly relevant for

- Price taking behavior in financial markets
  excess volatility
  (MM Challet Adv. Complex Sys. 3, 3-17 2001)
  Regularization in portfolio optimization
  (Caccioli, Still, MM, Kondor arxiv 2010)

- The beak of the finch
  (Weiner, The beak of the finch, Vintage books)



10 min break?

On the structure of preferences

- Background:
  Decentralized exploitation of resources by many agents
  Structure of preferences and degree of rationality
  Order-disorder transitions: Symmetry, luck and institutions
- Uncorrelated preferences
  Identical agents/resources: Route choice game
  Heterogeneous agents/resources: El Farol bar problem and Minority Game
- Aligned preferences
  Parking in Marseille
  A stylized model
- Conclusions

To specialize or not to specialize, this is the problem

Darwin's finches

Table 1. Occurrence Matrix for Darwin's Finch Data

Finch	*Island* A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*
Large ground finch	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Medium ground finch	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
Small ground finch	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Sharp-beaked ground finch	0	0	1	1	1	0	0	1	0	1	0	1	1	0	1	1	1
Cactus ground finch	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1	0	0
Large cactus ground finch	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
Large tree finch	0	0	1	1	1	1	1	1	1	0	0	1	0	1	1	0	0
Medium tree finch	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Small tree finch	0	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0
Vegetarian finch	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0
Woodpecker finch	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0
Mangrove finch	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Warbler finch	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

NOTE: Island name code: A = Seymour, B = Baltra, C = Isabella, D = Fernandina, E = Santiago, F = Rábida, G = Pinzón, H = Santa Cruz, I = Santa Fe, J = San Cristóbal, K = Española, L = Floreana, M = Genovesa, N = Marchena, O = Pinta, P = Darwin, Q = Wolf.

pure strategy

mixed strategy

Weiner, *The beak of the finch* (Vintage Books)

(courtesy of A. De Martino)

The setting

- N agents $\underset{\text{constraints}}{\overset{\text{preferences}}{\rightleftarrows}}$ P resources

  E.g. financial markets (speculators and information patterns), commodity markets (buyers and sellers), production economies (firms and goods' markets), urban traffic (vehicles and streets), internet traffic (surfers and Internet), ecosystems (species and resources), ...

  adaptive agents = preferences + bounded rationality

  noise = experimentation, environmental volatility, ...

- Question:

$$\text{collective behavior} = F\left(\text{structure of preferences},\ \text{degree of rationality},\ \ldots\right)$$

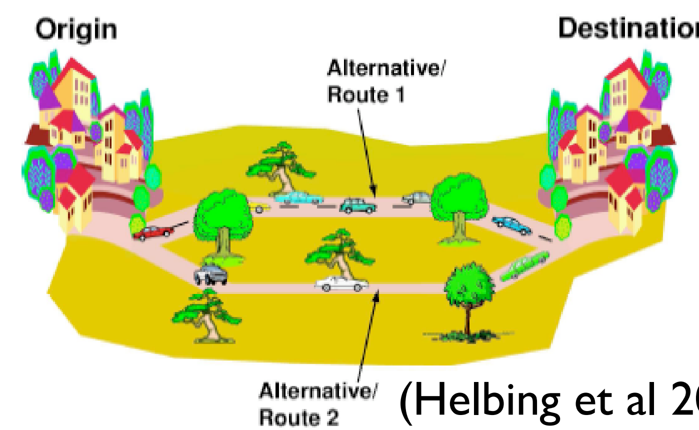
Order - disorder transition:
symmetric (mixed strat.)
asymmetric (pure strat.)

Identical/heterogeneous agents and/or resources

Learning, memory,...

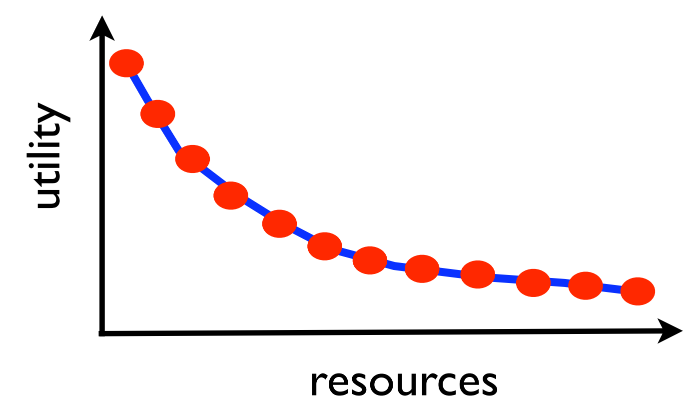
Uncorrelated or aligned preferences?

- Uncorrelated preferences:
  pure coordination



(Helbing et al 2005)

- Aligned preferences:
  coordination + competition



- Symmetric or asymmetric states?
  phase transitions
  social norms and institutions
  luck and property rights

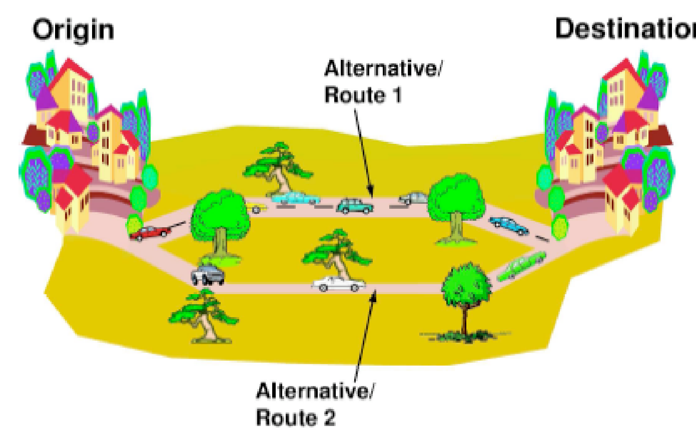
How much rationality?

- Zero intelligence: agents as automata
- Naive agents:
  play against Nature (e.g. price taking behavior)
- Sophisticated agents:
  account for feedback of own actions on Nature
- Strategic agents (players) and common knowledge of rationality

1. types of self-organization and degrees of rationality
2. degree of complexity and learnability

Uncorrelated preferences:
Identical agents/resources

- n agents, two actions: $a_i=\pm 1$,
  payoff: $u_i=-a_i A$, $A=\Sigma_j\ a_j$



- Symmetric NE: $P\{a_i=+1\}=1/2$ for all i
  Unique
  Large fluctuations $n_+ - n_- \sim \mathrm{sqrt}(n)$

- Asymmetric NE: $a_i=+1$ for $i<n/2$; $a_i=-1$ otherwise
  Exponentially many (in n)
  Small fluctuations $n_+ - n_- \sim O(1)$

- Naive reinforcement learning [$A>0 \rightarrow$ increase $P\{a_i=-1\}$]
  converges to symmetric equilibrium

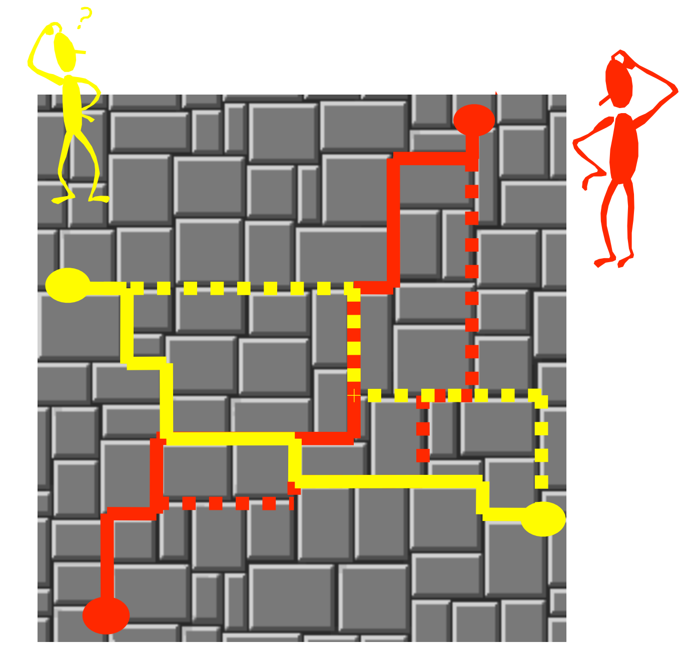
- + impact: $A - \varepsilon a_i>0 \rightarrow$ increase $P\{a_i=-1\}$
  converges to asymmetric equilibrium for all $\varepsilon>0$

Even for $n\to\infty$!

- Verified in experiments (Helbing et al. 2005)

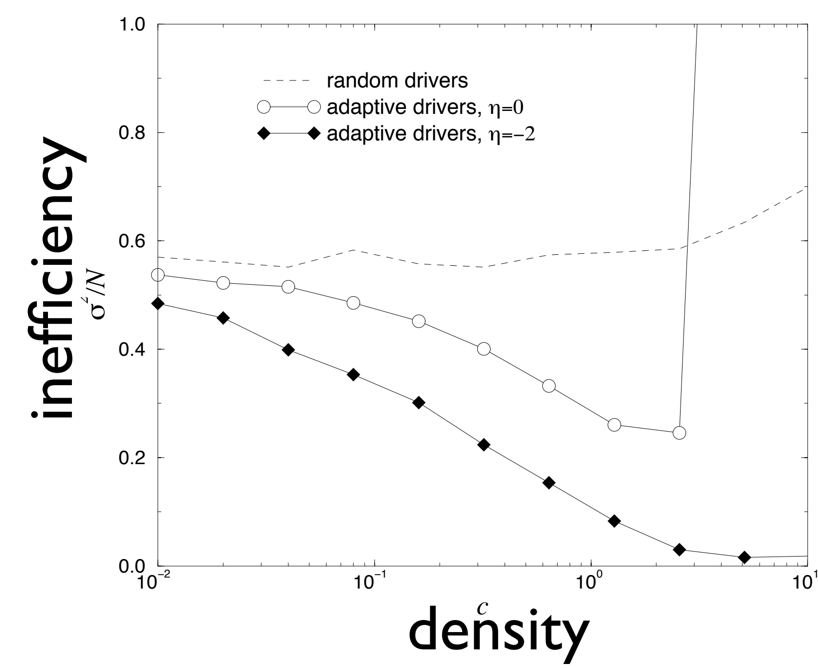
Uncorrelated preferences:
Heterogeneous agents/resources

(Minority Game)



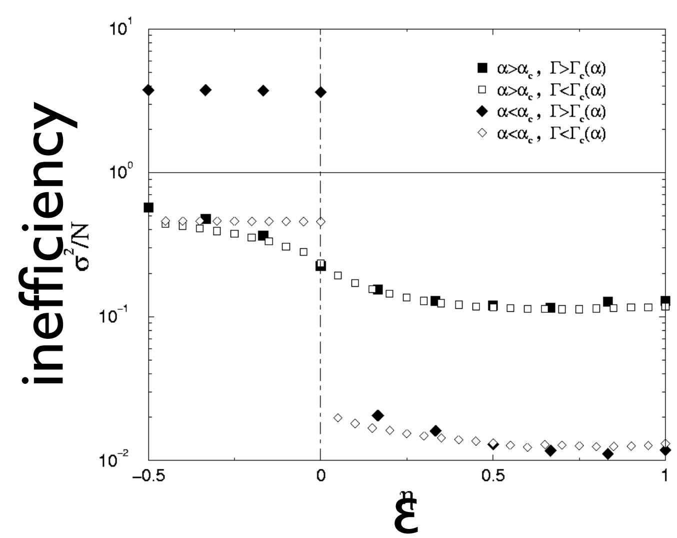
Naive agents → "mixed strategy" equilibrium:

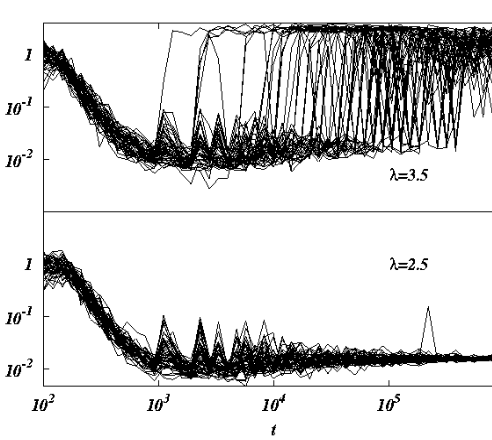
- unique
- even exploitation of resources
- large fluctuations
- inefficient
- easy to learn



Sophisticated agents → pure strategy equilibrium:

- large degeneracy
- minimal fluctuations
- most efficient
- hard to learn in a "volatile world"





(Challet, MM, Zhang, Minority Game, OUP 2005)

Aligned preferences

Examples:

- Looking for a parking
- Securing a territory or nesting sites, or establishing pecking order in animal populations
- Settlements and colonies
- Users and printers/CPU
- ...

Agents access resources when needed (volatility)

Parking in Marseille

(Kirman, Hanaki, MM, JEBO to appear)

- Population of n agents either
  - at home: going to work at rate η
  - at work: leaving the parking at rate 1
- n parking slots on one way street to office
  - payoff for parking at $s=1, \ldots, n$: $u(s) \downarrow s$
  - Strategy: go up to spot k and then park in first empty spot
  - if no empty slot found, payoff $= -L$
    (need to go all the way around to find parking)

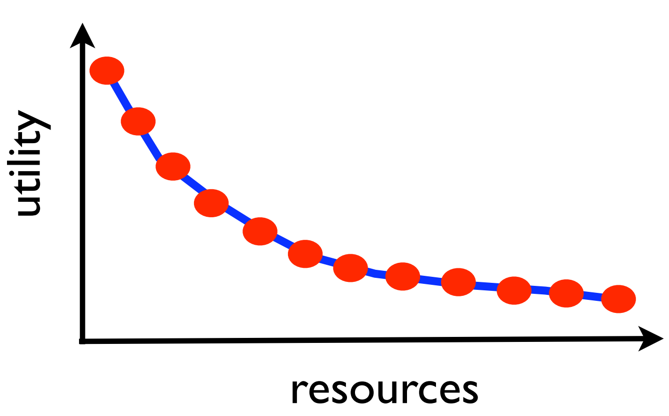
Symmetric and asymmetric equilibria and luck

- Take the parking example:
  - "Unlucky" people park at the first empty spot.
  - "Lucky" people keep going closer to the office and find an empty spot.
  - "Unlucky" people think, there will not be empty slots closer to the office because others take them.
  - But it is because "unlucky" people not trying to park closer, there are empty slots for "lucky" people closer to the office.
  - And because they have learned to behave in such ways, these outcomes repeat themselves.
- Lucky ones are not "born under a lucky star" but they have learned to be so.

(Kirman, Hanaki, MM, *Born under a lucky star?* preprint @ IDEAS, to appear on JEBO)

A simpler model

- n agents, n resources with exclusive use
- Agents on a resource leave it at rate 1
- Agents not on resource, look for a free resource at rate η
- Strategy:
  order $s_1, s_2, \ldots, s_n$ with which agents search for resources
- Payoff:
  Resource s=1,...,n
  utility $u(s) \downarrow s$ if free
  if occupied, agent pays cost c
  and needs to search further



Naive agents: symmetric equilibrium
(stationary state)

- i) agents know the probability p^m that resource m is free
- ii) agents adopt mixed strategy: P{go to resource m}=g^m

$$
\begin{aligned}
p^m W^m[0\to 1] &= (1-p^m)W^m[1\to 0] \\
W^m[1\to 0] &= 1 \\
W^m[0\to 1] &= (n-R)w^m[0\to 1] \\
w^m[0\to 1] &= \eta g^m + \left[\sum_{m'\neq m} g^{m'}(1-p^{m'})\right] w^m[0\to 1] \\
&= \frac{\eta g^m}{\bar{p}+(1-p^m)g^m}, \qquad \bar{p}=\sum_m g^m p^m
\end{aligned}
$$

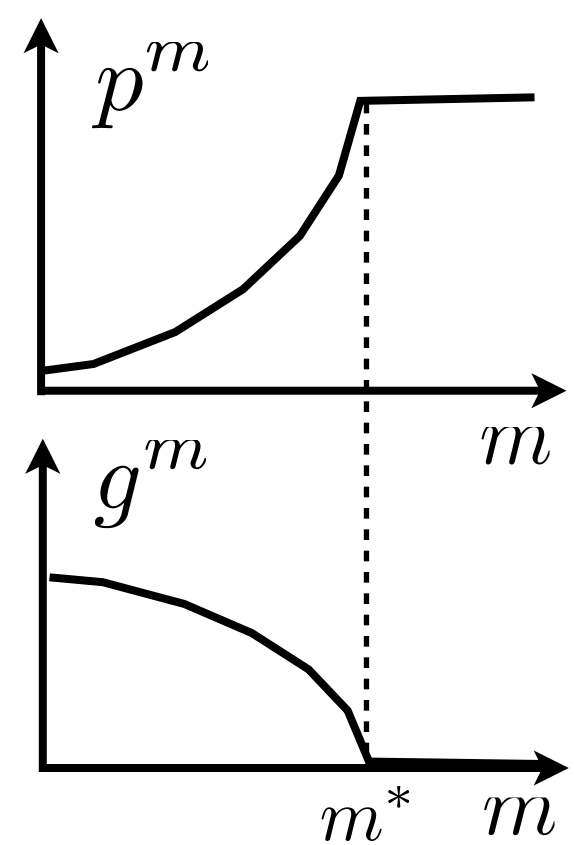
- R = number of occupied resources

$$
P\{R=r\} = \binom{n}{r}\frac{\eta^r}{(1+\eta)^n}
$$

Naive agents: II

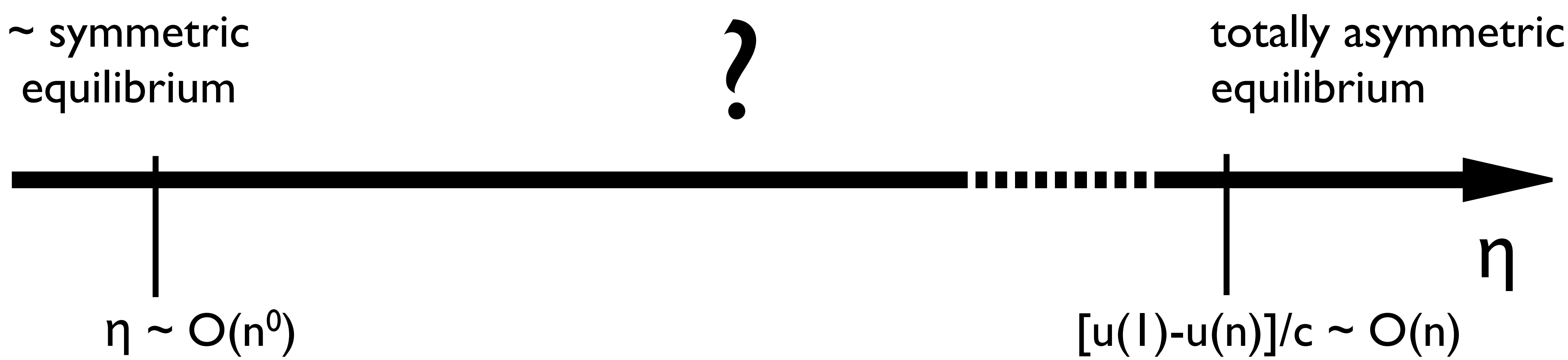
- This gives p^m as a function of g^m
- Expected utility: $$\begin{aligned} E[u(g)] &= \sum_m g^m \left\{ p^m u^m + (1-p^m)\left[E[u(g)] - c\right]\right\} \\ &= \frac{\overline{pu} - c(1-\bar{p})}{\bar{p}}, \qquad \overline{pu} = \sum_m g^m p^m u^m \end{aligned}$$
- Solve: $$\max_{g,\lambda} \left\{ E[u(g)] - \lambda \left(\sum_m g^m - 1 \right) \right\}$$
- Result: $$\begin{aligned} p^m &= \frac{c - \overline{pu} + \lambda \bar{p}^2}{\bar{p}u^m - \overline{pu} + c}, \qquad m \le m^*(\lambda^*) \\ &= 1, \qquad m > m^*(\lambda^*) \end{aligned}$$

$u^{m^*(\lambda^*)} = \lambda^* \bar{p}$ + normalization $\rightarrow \lambda^*$



Asymmetric NE

- Consider the equilibrium where agent i goes each time to resource i
- Is this a Nash equilibrium?
- Deviation: agent n tries to occupy resource 1:
  $E[u_{deviation}] = u(1)P\{1 \text{ free}\} + (1-P\{1 \text{ free}\})[u(n)-c] > u(n)$
  if $\eta > [u(1)-u(n)]/c$



Should agents remember where they came from?

- Agents have, in principle access to the information of past visited and attempted resources and the time elapsed
- If this information allows them to gain a higher payoff, wrt the symmetric equilibrium, it is evolutionarily advantageous to store it
- Consider the strategy $m_{last} \rightarrow g$ where agents first return to the last occupied site and, if this is occupied by another agent, play g
- **Proposition:** This strategy invades g

Example: 2 agents

Symmetric equilibrium with (temporarily) asymmetric allocation

Proposition 2.1. *A Nash equilibrium of the two player game, where agents remember which resource they last visited and when, is the following.*
Call $\Delta \equiv \frac{u_1 - u_2}{c}$. *If* $\Delta < \eta$ *then one of the two agents will always exploit resource* 1, *while the other will always exploit resource* 2.
If instead $\Delta \geq \eta$ *there is a time* $\tau > 0$ *such that both agents act according to the following strategy:*

a) if the last resource occupied was 1, *then first try resource* 1, *if* 1 *is occupied go to* 2;

b) if the last resource occupied was 2 *then*

1) *if the other agent was last seen* $t < \tau$ *time ago, then return to resource* 2;

2) *if the other agent was last seen* $t \geq \tau$ *time ago, then first try resource* 1, *if* 1 *is occupied go to* 2;

$$\tau \equiv \frac{1}{1+\eta} \log\left(\frac{\Delta + 1}{\Delta - \eta}\right)$$

$$\text{Idea: } P\{1 \text{ free}\} = \frac{1 - e^{-(1+\eta)t}}{1+\eta} \quad \nearrow \quad t$$

Generalizing to n agents

- Information:

$$\begin{aligned}\Im_i(t) &= \{\tau_i^m, z_i^m\}, \\ \tau_i^m \in R_+ &= \text{time of last visit of } i \text{ to } m \\ z_i^m \in \{0,1\} &= \text{occupation of } m \text{ at } \tau_i^m\end{aligned}$$

- Compute

$$q_i^m(t) = P\{m \text{ free} | \Im_i(t)\}$$

- **Proposition:** $q_i^m(t)$ determines the optimal search strategy of agent i at time t

If $u^m - c/q_i^m(t) > u^{m'} - c/q_i^{m'}(t)$ then $m >_i m'$

Proof:

$$\begin{aligned} q_i^m u_m + (1 - q_i^m)\left[-c + q_i^n u_n - (1 - q_i^n)(-c + \Pi)\right] &\geq \\ q_i^n u_n + (1 - q_i^n)\left[-c + q_i^m u_m - (1 - q_i^n)(-c + \Pi)\right] & \end{aligned}$$

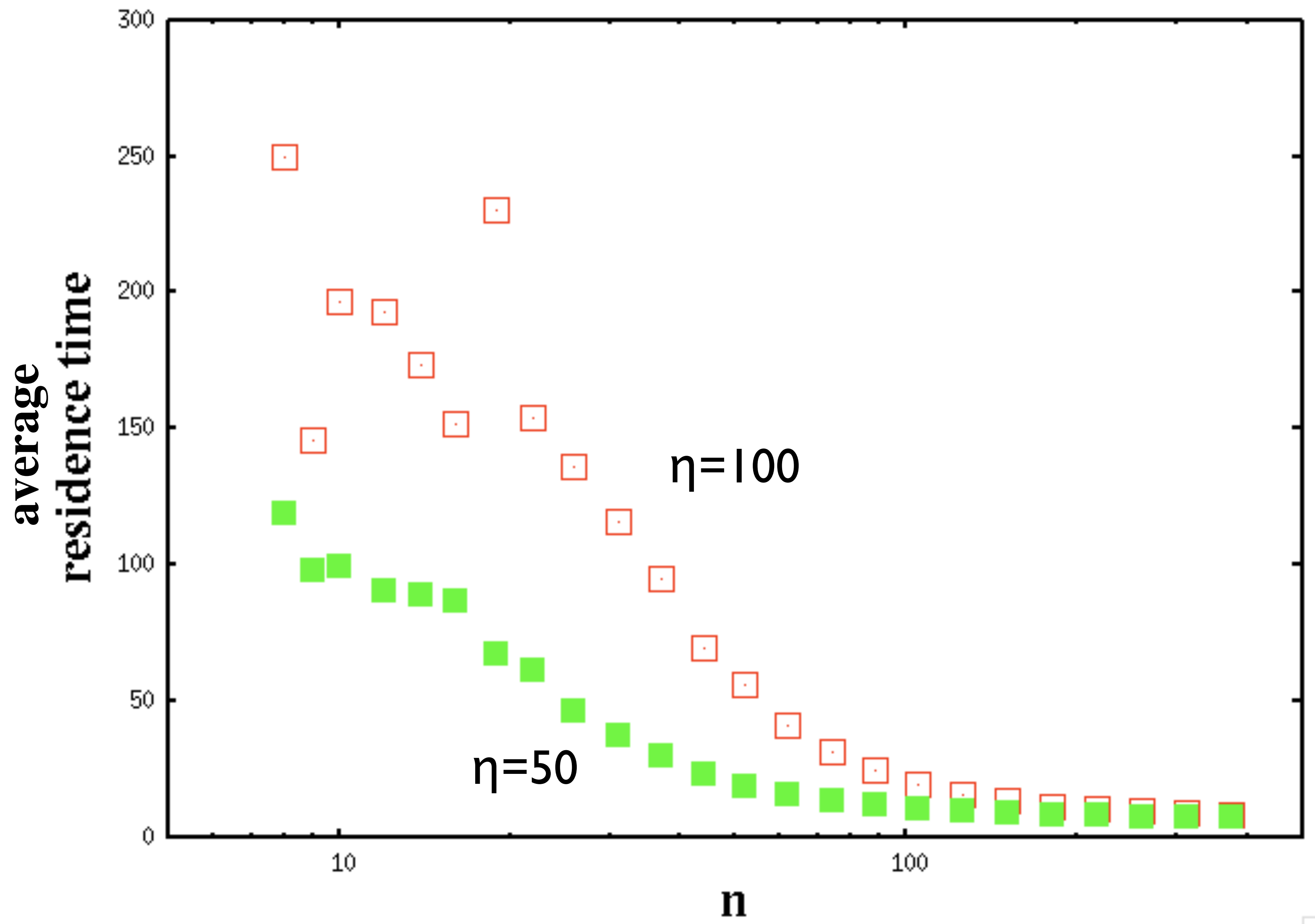
Simulations

- A minimal implementation:

$$q_i^m(t) \approx p^m \left(1 - e^{-(t-t_i^m)/p^m}\right) + (1 - z_i^m)\, e^{-(t-t_i^m)/p^m}$$

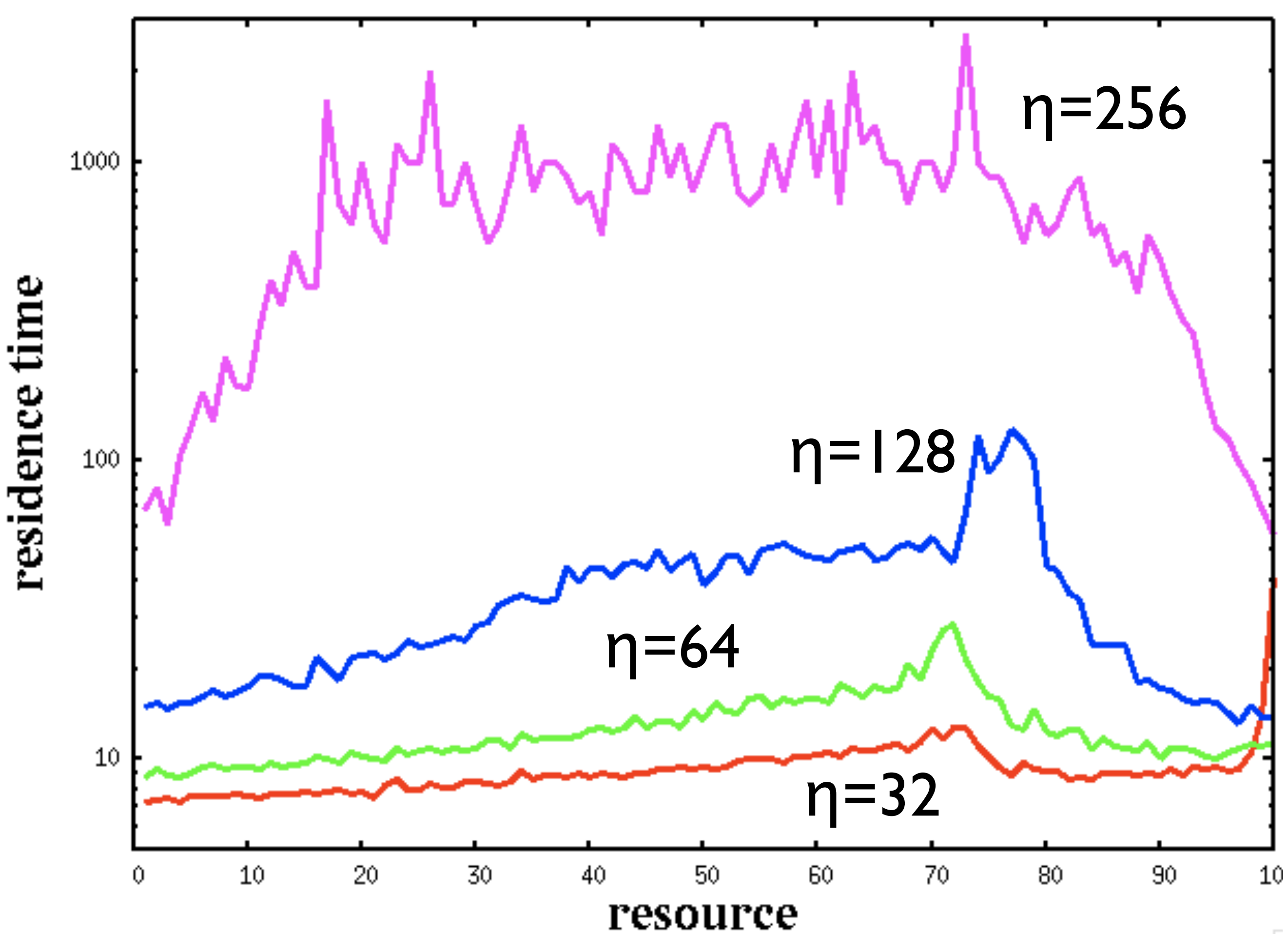
- Unconditional probability p^m assumed common knowledge
- Linear utility: u(s) = n-s
- Measure of residence time of resource m
  = average time spent by the same agent on resource m
- Simulations for t/n = 10^3, averages taken on last half period

Localization for n « η

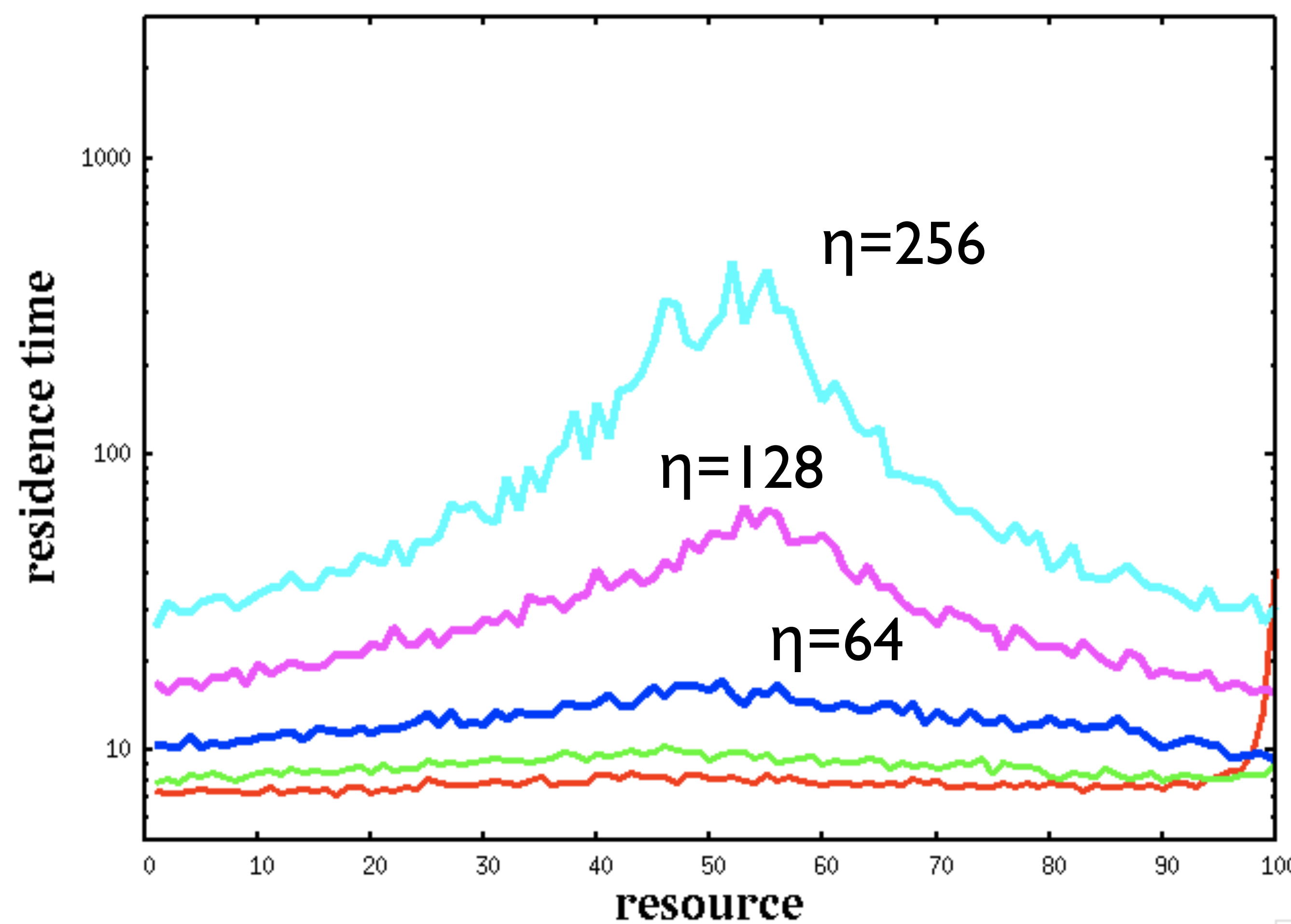


The safe middle property

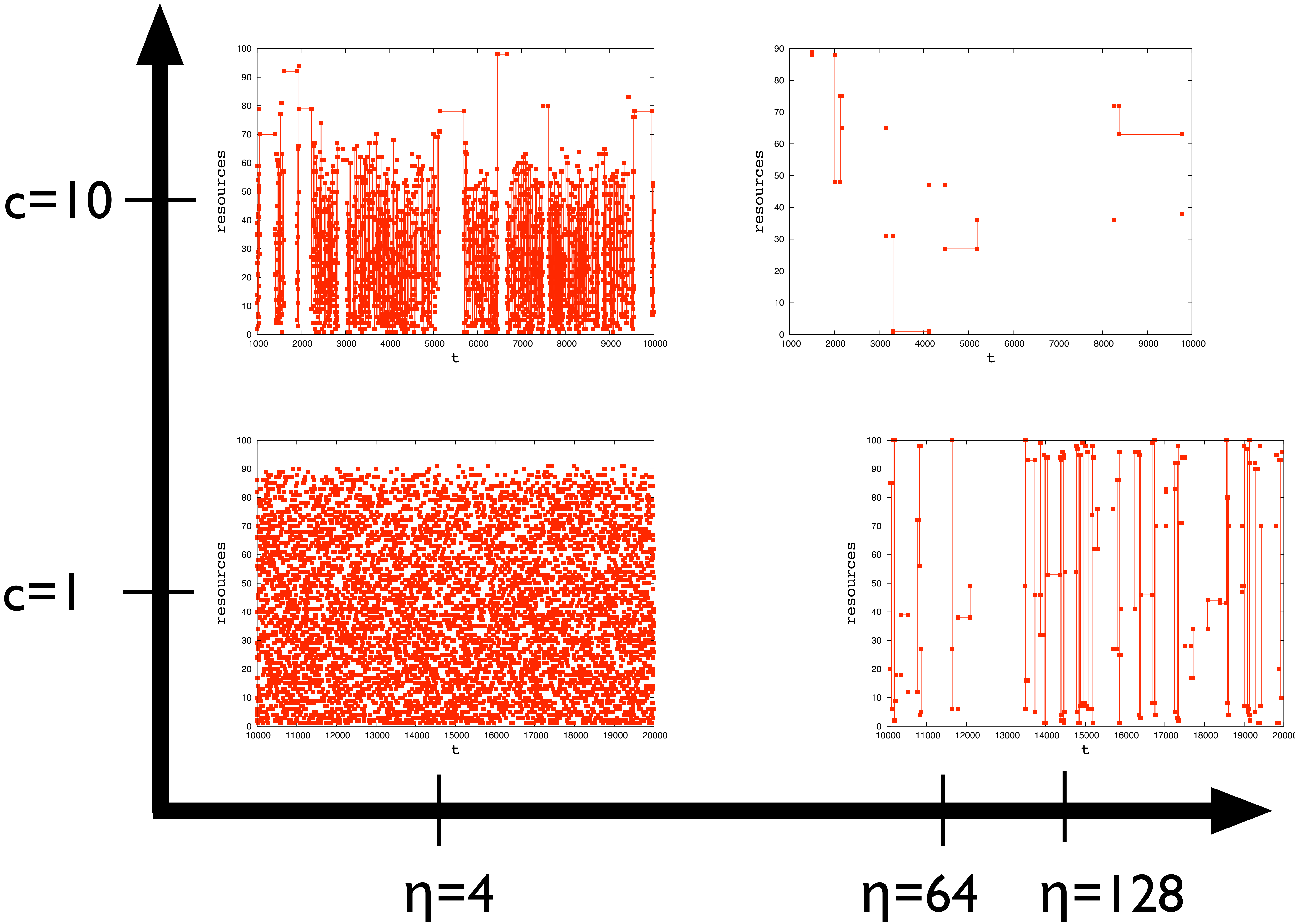
n=100, c=1



c=0.5



Intuition: agents on lousy resources will risk looking for other resources only if utility gain is large enough.



c=10
c=1
η=4
η=64
η=128
resources
t

Conclusions

- Different types of self-organization require different degrees of rationality
- Asymmetric states can be achieved when agents understand and account for their impact on "Nature"
- When preferences are aligned:
  - Memory is evolutionarily advantageous, specially in crowded contexts
  - Order nucleates from the middle

Thanks