EPSRC
Pioneering research and skills

BAE SYSTEMS

Convergent Learning in Unknown Hypergraphical Games

Dr Archie Chapman, Dr David Leslie,
Dr Alex Rogers and Prof Nick Jennings

School of Mathematics, University of Bristol and
School of Electronics and Computer Science
University of Southampton

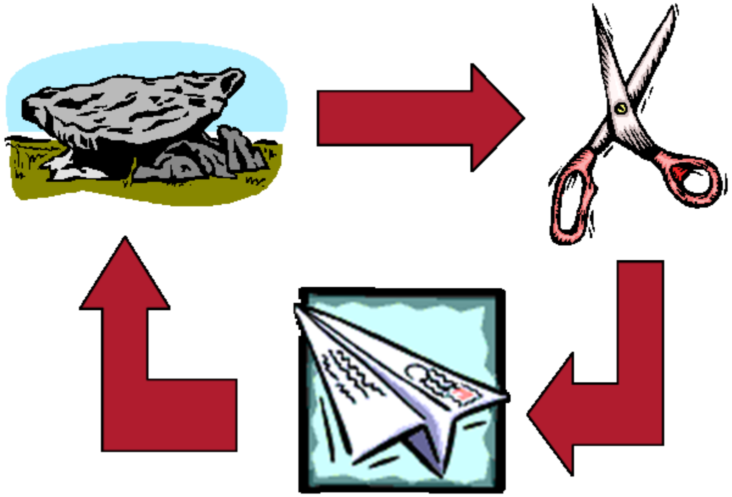
david.leslie@bristol.ac.uk

ALADDIN
autonomous learning agents for decentralised data and information networks

University of BRISTOL

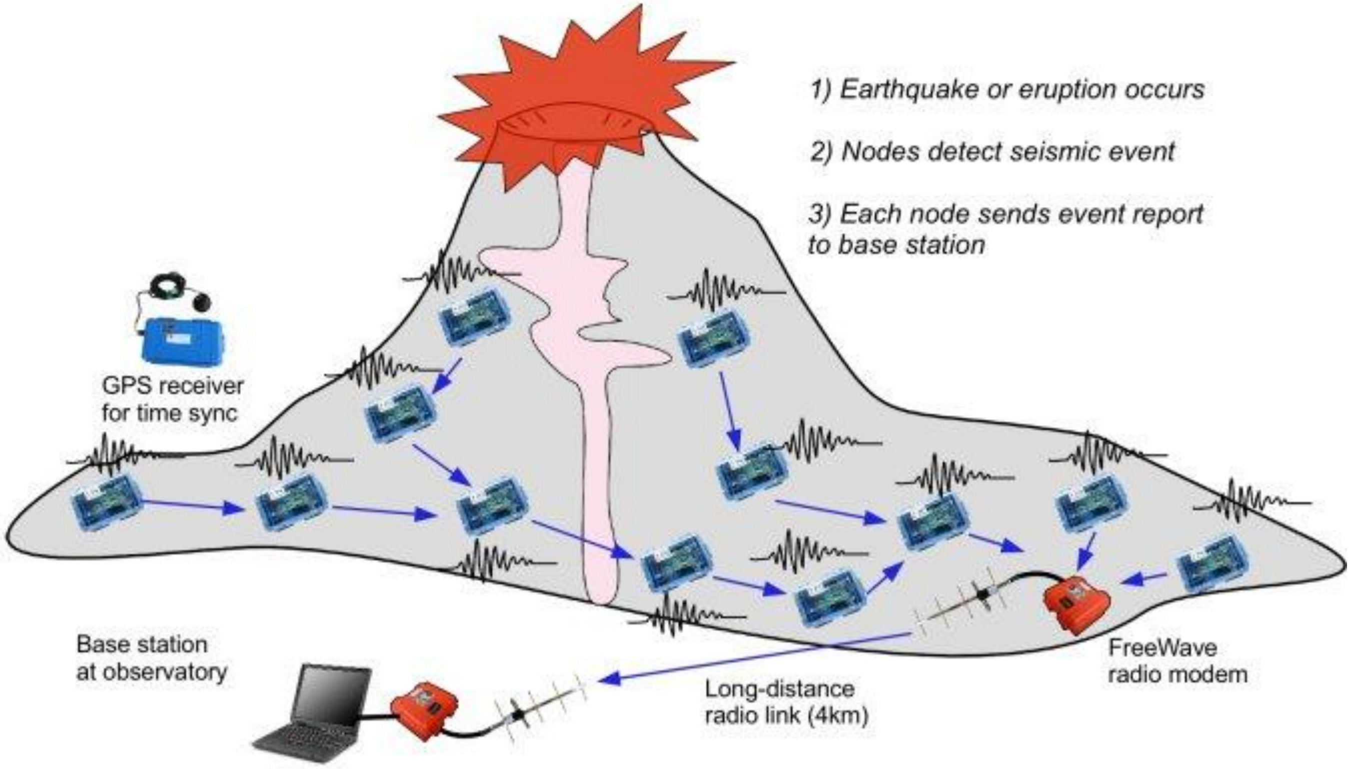
Playing games?

Playing games?



$$\begin{array}{cc} & \begin{array}{ccc}\text{Rock} & \text{Scissors} & \text{Paper}\end{array} \\ \begin{array}{c}\text{Rock} \\ \text{Scissors} \\ \text{Paper}\end{array} & \begin{pmatrix} (0,0) & (1,-1) & (-1,1) \\ (-1,1) & (0,0) & (1,-1) \\ (1,-1) & (-1,1) & (0,0) \end{pmatrix}\end{array}$$

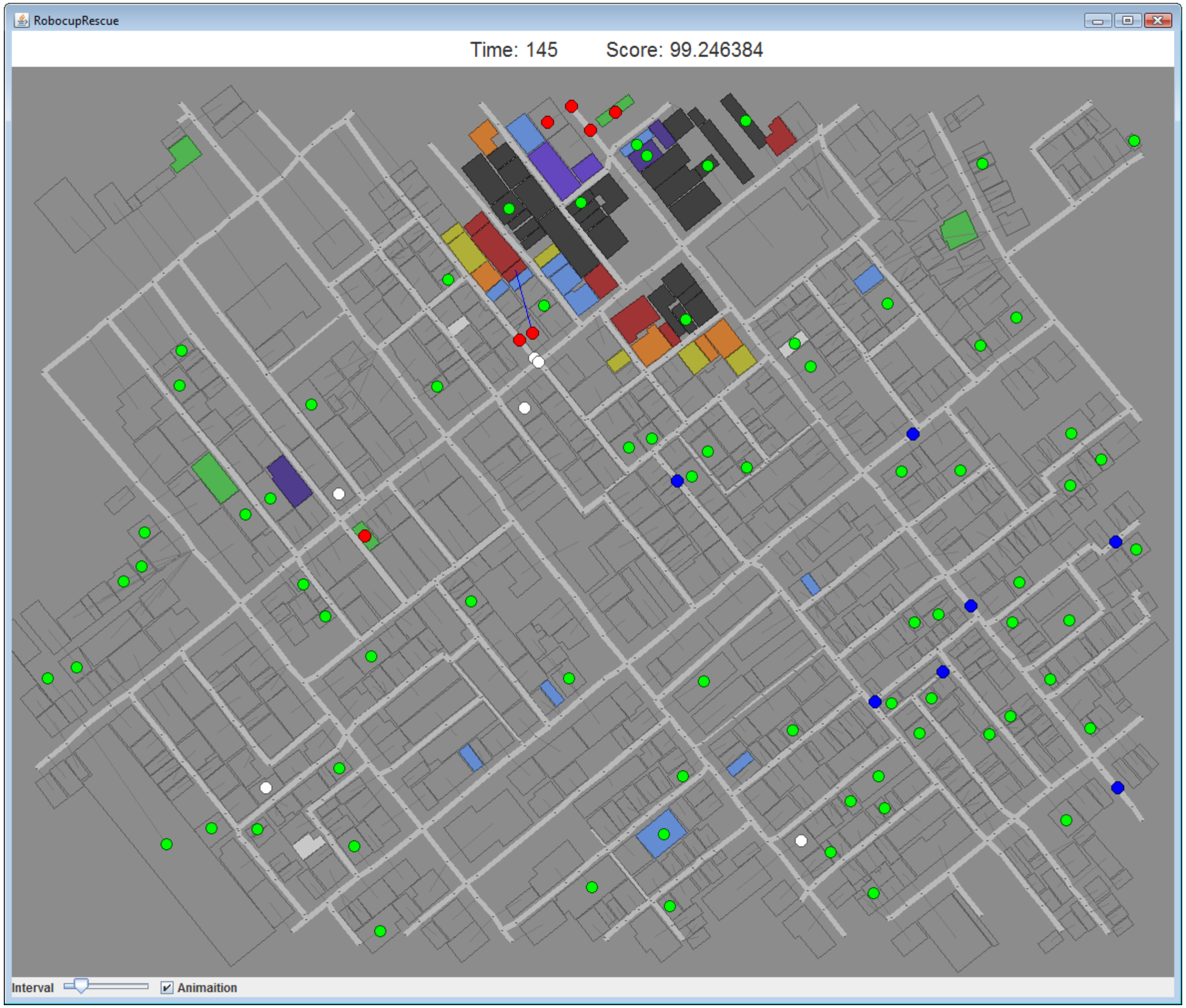
Playing games?


1) Earthquake or eruption occurs
2) Nodes detect seismic event
3) Each node sends event report to base station
GPS receiver for time sync
Base station at observatory
Long-distance radio link (4km)
FreeWave radio modem

EPSRC
Pioneering research and skills

BAE SYSTEMS

Playing games?



RobocupRescue

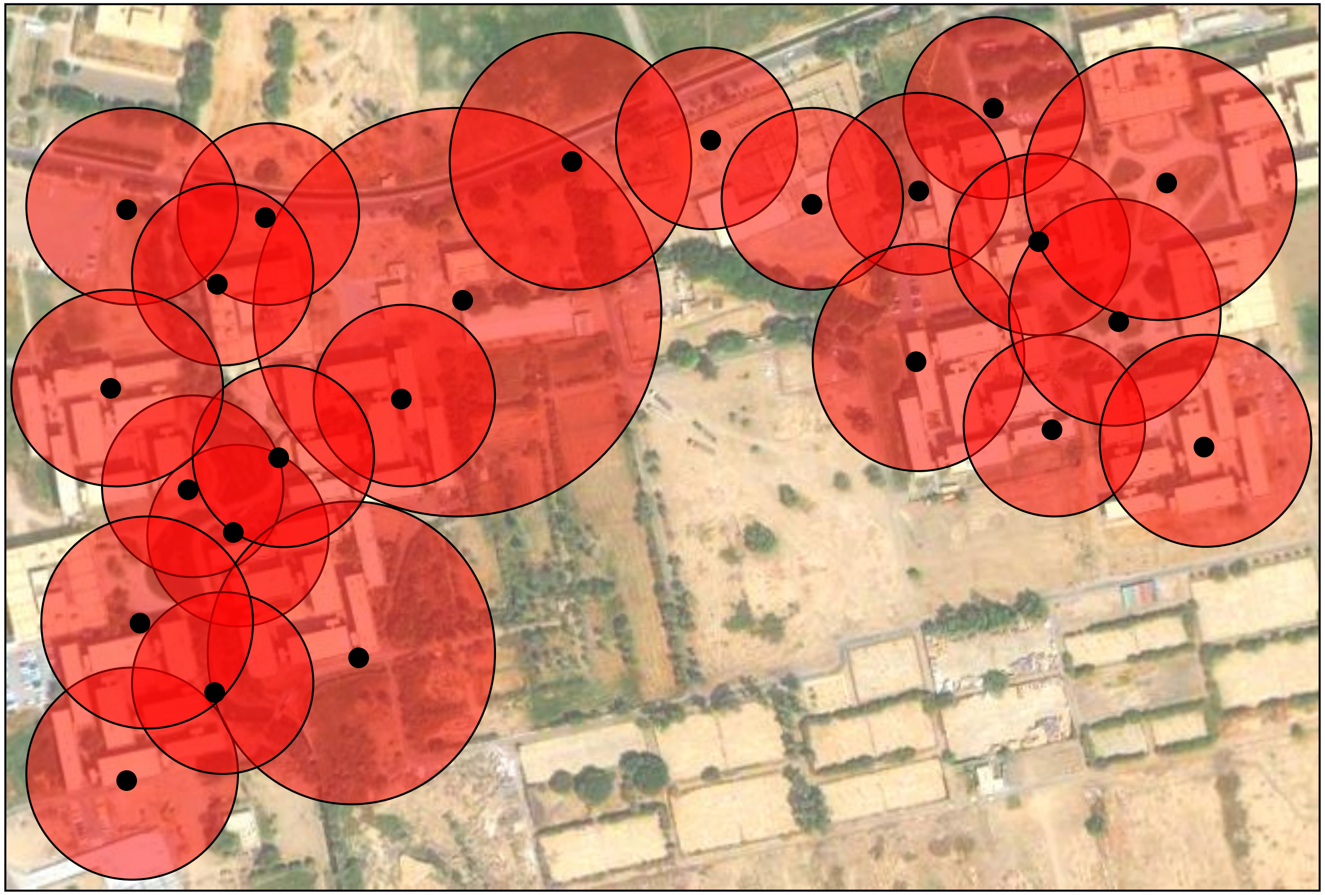
Time: 145 Score: 99.246384

ALADDIN
autonomous learning agents for decentralised data and information networks

University of BRISTOL

Playing games?

Dense deployment of sensors to detect pedestrian and vehicle activity within an urban environment.



Learning in games

- Adapt to observations of past play
- Hope to converge to something “good”
- Why?!
  - – Bounded rationality justification of equilibrium
  - – Robust to behaviour of “opponents”
  - – Language to describe distributed optimisation

Notation

- Players $i \in \{1,\ldots,N\}$
- Discrete action sets A_i
- Joint action set $A = A_1 \times \cdots \times A_N$
- Reward functions $r_i : A \rightarrow \mathbf{R}$

- Mixed strategies $\pi_i \in \Delta_i = \Delta(A_i)$
- Joint mixed strategy space $\Delta = \Delta_1 \times \cdots \times \Delta_N$
- Reward functions extend to $r_i : \Delta \rightarrow \mathbf{R}$

Best response / Equilibrium

- Mixed strategies of all players other than *i* is

$$\pi_{-i}$$

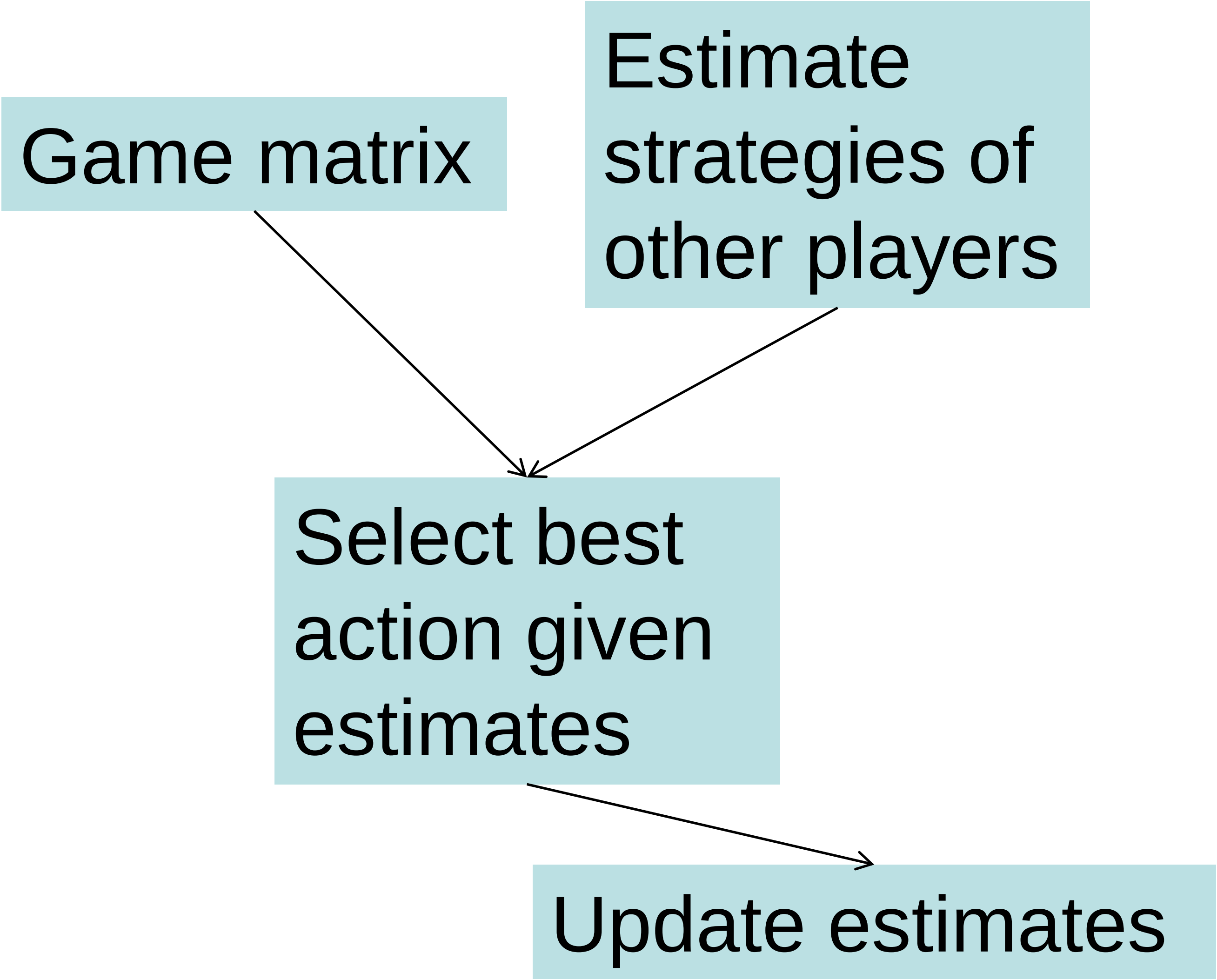
- Best response of player *i* is

$$b_i(\pi_{-i}) = \underset{\pi_i \in \Delta_i}{\operatorname{argmax}} r_i(\pi_i, \pi_{-i})$$

- An equilibrium is a π satisfying, for all *i*,

$$\pi_i \in b_i(\pi_{-i})$$

Fictitious play



Game matrix

Estimate strategies of other players

Select best action given estimates

Update estimates

Belief updates

- Belief about strategy of player *i* is the MLE

$$\sigma_i^t(a_i) = \frac{\kappa_i^t(a_i)}{t}$$

- Online updating

$$\sigma^t \in \sigma^{t-1} + \tfrac{1}{t}\left(b(\sigma^{t-1}) - \sigma^{t-1}\right)$$

Stochastic approximation

- Processes of the form

$$X_{t+1} \in X_t + \lambda_{t+1} \left[F(X_t) + M_{t+1} + e_{t+1} \right]$$

where $\mathbf{E}(M_{t+1} \mid X_t) = 0$ and $e_t \rightarrow 0$

- *F* is set-valued (convex and u.s.c.)
- Limit points are chain-recurrent sets of the differential inclusion

$$\dot{X} \in F(X)$$

Best-response dynamics

- Fictitious play has *M* and *e* identically 0, and $\lambda_t = \frac{1}{t}$
- Limit points are limit points of the best-response differential inclusion

$$\dot{\pi} \in b(\pi)$$

- In potential games (and zero-sum games and some others) the limit points must be Nash equilibria

Generalised weakened fictitious play

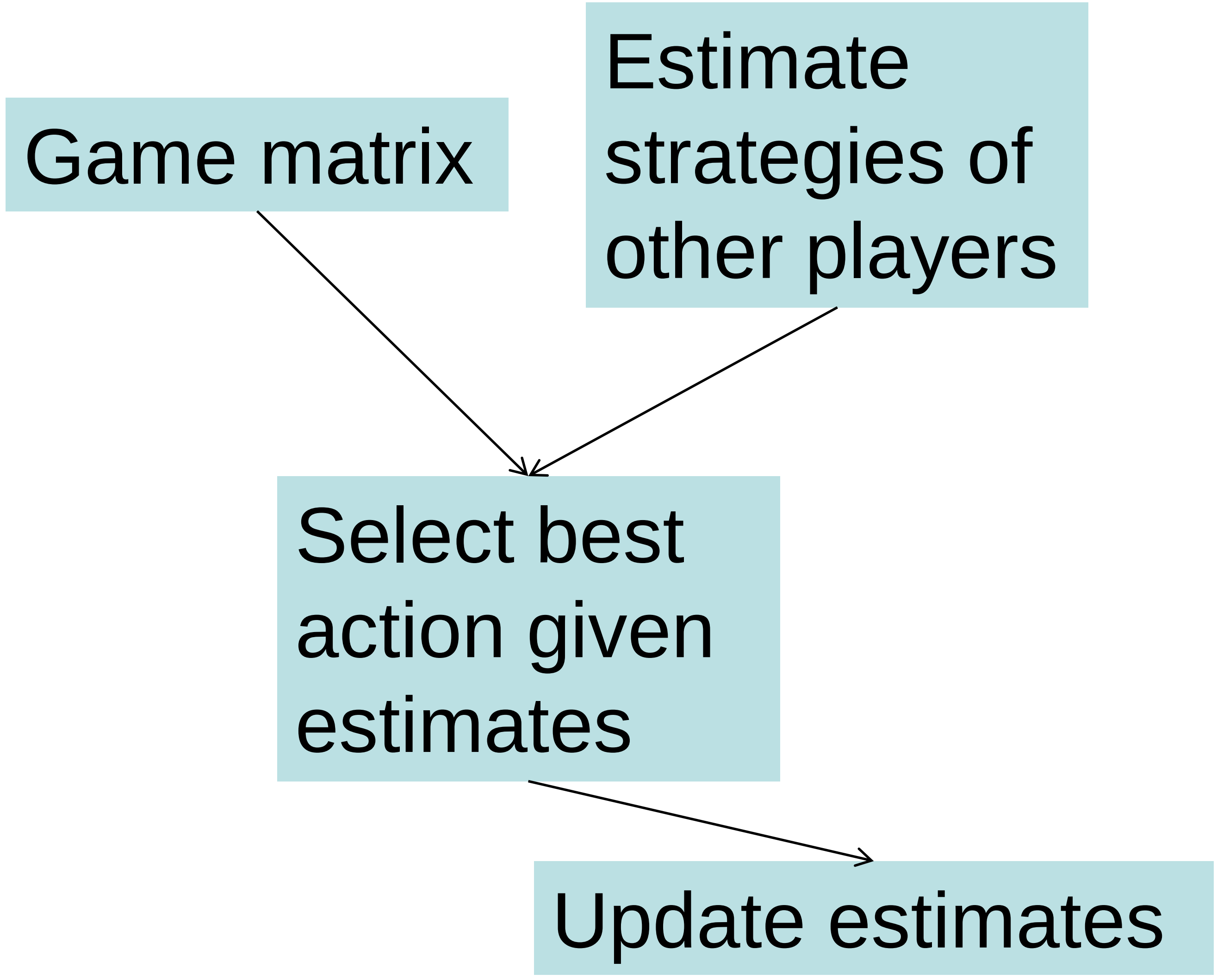
- Bring back non-zero *M* and *e*
- Any process such that

$$\sigma^{t} \in \sigma^{t-1} + \lambda^{t} \Big\{ b^{\varepsilon^{t}}(\sigma^{t-1}) - \sigma^{t-1} + M^{t}$$

where $\varepsilon^{t} \to 0,\ \lambda^{t} \to 0$ and $\sum \lambda^{t} = \infty$

and also an interplay between λ and *M.*

Fictitious play



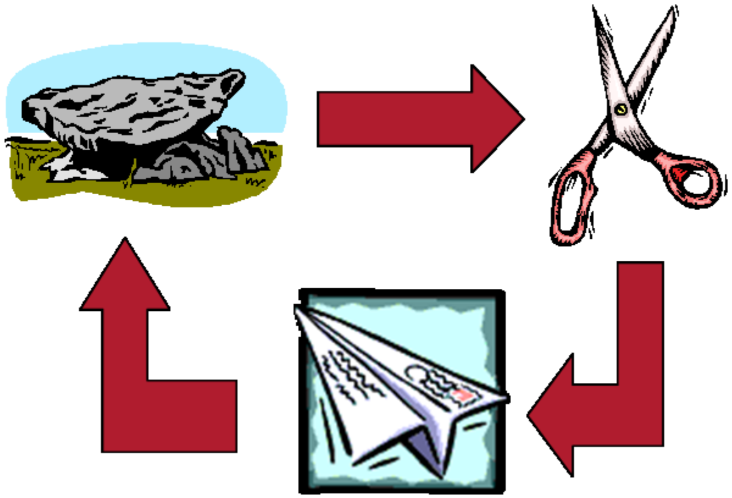
Game matrix

Estimate strategies of other players

Select best action given estimates

Update estimates

Learning the game



$$\begin{array}{c|ccc} & \text{Rock} & \text{Scissors} & \text{Paper} \\ \hline \text{Rock} & (?,\xcancel{?}) & (?,\xcancel{?}) & (?,\xcancel{?}) \\ \text{Scissors} & (?,\xcancel{?}) & (?,\xcancel{?}) & (?,\xcancel{?}) \\ \text{Paper} & (?,\xcancel{?}) & (?,\xcancel{?}) & (?,\xcancel{?}) \end{array}$$

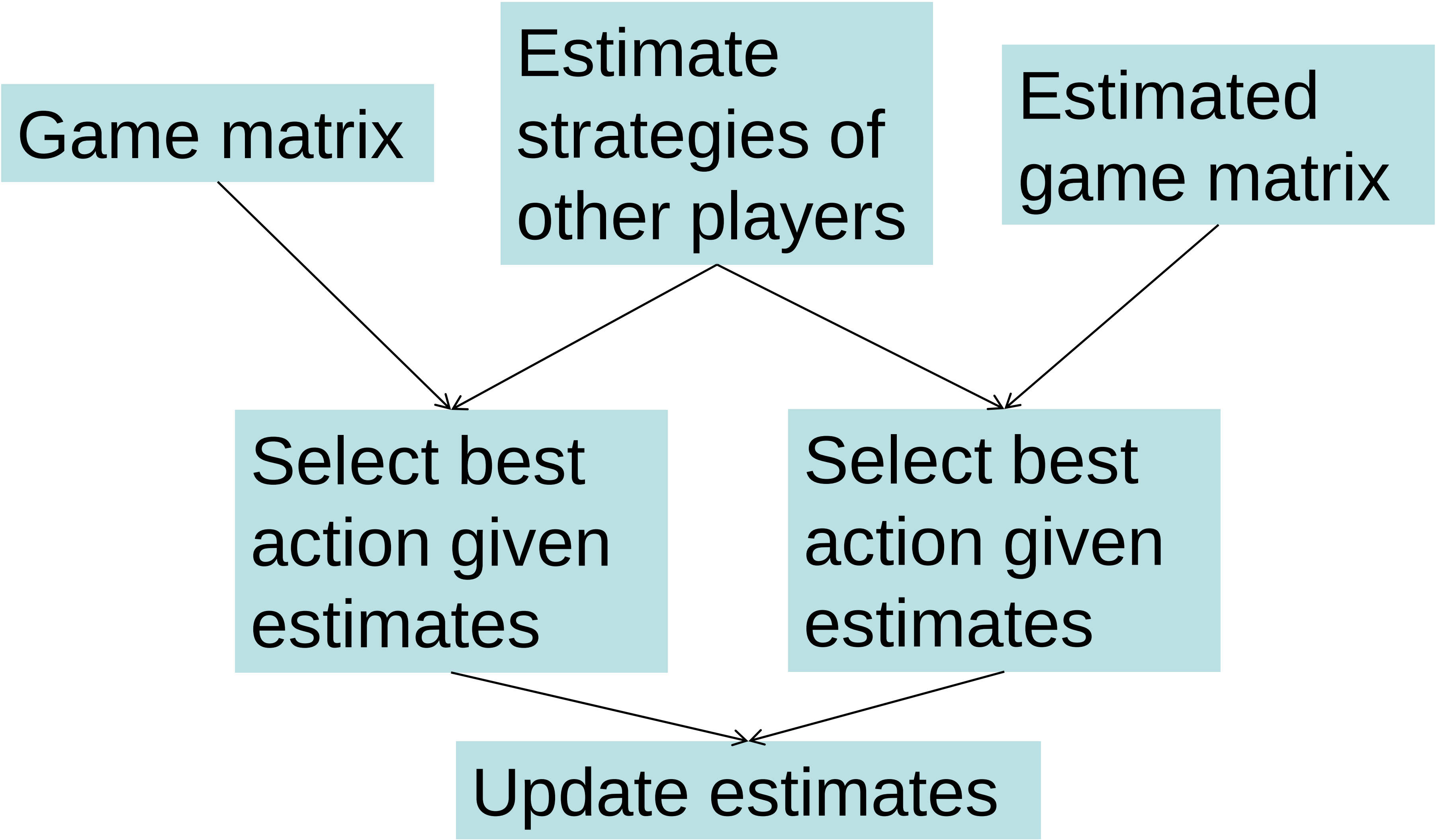
$$R_i^t = r_i(a^t) + e_i^t$$

Reinforcement learning

- Track the average reward for each joint action
- Play each joint action frequently enough
- Estimates will be close to the expected value

- Estimated game converges to the true game

Q-learned fictitious play



Game matrix

Estimate strategies of other players

Estimated game matrix

Select best action given estimates

Select best action given estimates

Update estimates

Theoretical result

Theorem – If all joint actions are played infinitely often then beliefs follow a GWFP

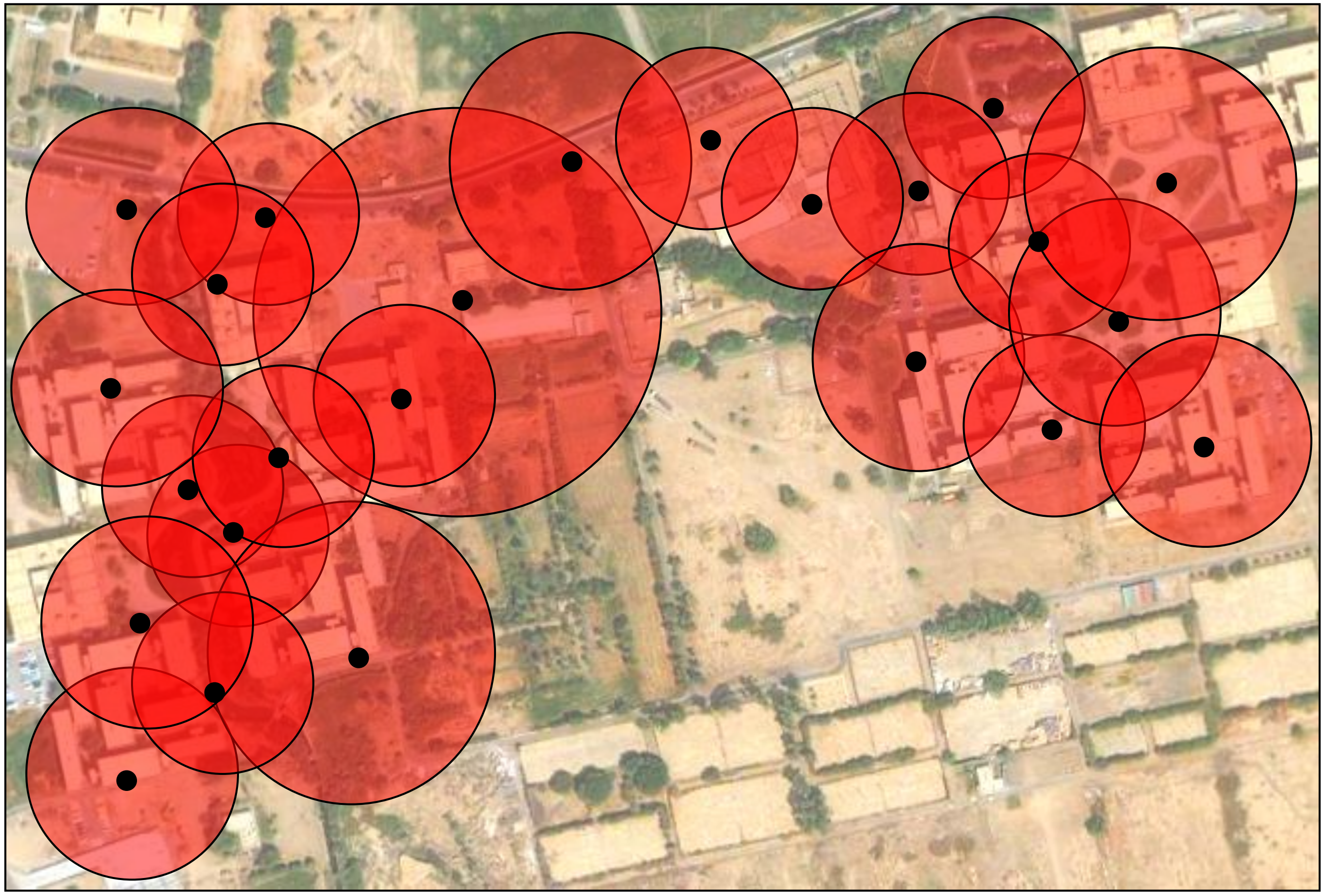
Proof: The estimated game converges to the true game, so selected strategies are ε-best responses.

Claus and Boutilier

- Claus and Boutilier (1998) state a similar result
- It is restricted to team games

Playing games?

Dense deployment of sensors to detect pedestrian and vehicle activity within an urban environment.



It’s impossible!

- N play each with A actions
- Game A^N entries to learn
- Each individual mus ategy of every other individual
- It’s **just not possible** for realistic game scenarios

Massive observational and computational requirement

Marginal contributions

- Marginal contribution of player *i* is

total system reward – system reward if *i* absent

- Maximised marginal contributions implies system is at a (local) optimum

- Marginal contribution might depend only on the actions of a small number of neighbours

Sensors – rewards

- Global reward for action *a* is

$$U_g(a) = \underset{\substack{\text{events } j \\ \text{and observation}}}{E}\left[\sum_j I_{j \text{ is observed}}\right] = \underset{\text{events}}{E}\left[\sum_j \left(1-\eta^{n_j(a)}\right)\right]$$

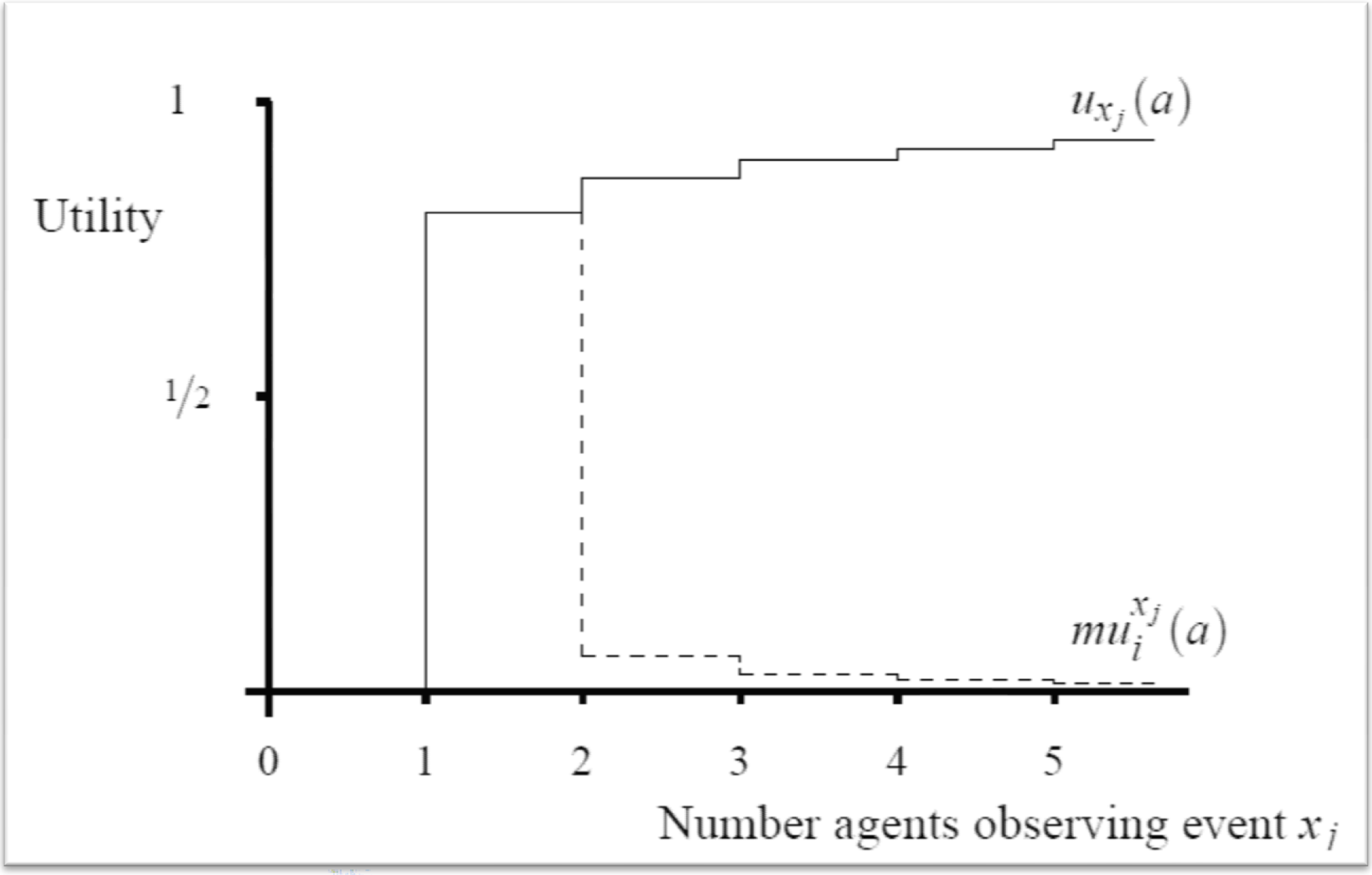
- Marginal reward for *i* is

$$r_i(a) = U_g(a) - U_g(a_{-i}) = \underset{\text{events}}{E}\left[\sum_{\substack{j \text{ observed} \\ \text{by } i}} \left(\eta^{n_j(a)-1} - \eta^{n_j(a)}\right)\right]$$

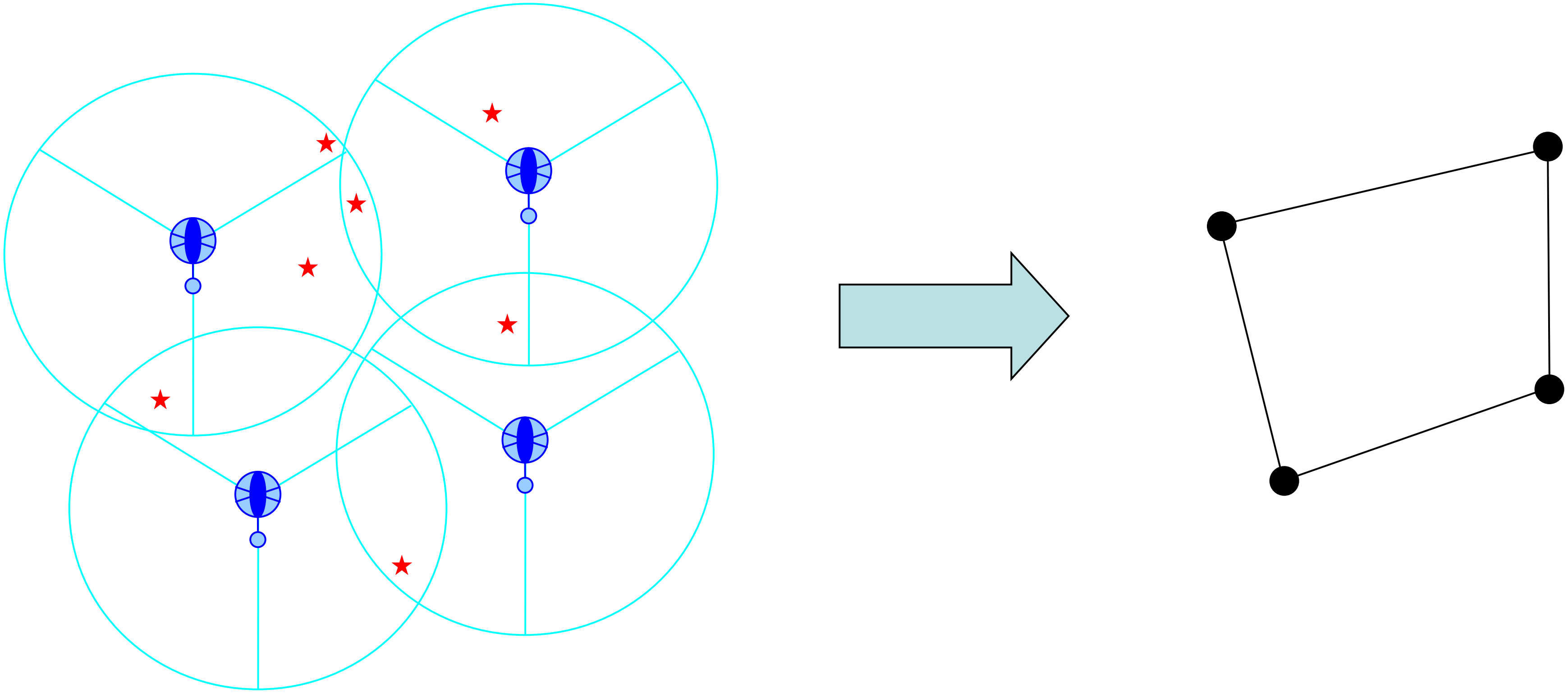
- Actually use

$$R_i^t = \sum_{\substack{j \text{ observed} \\ \text{by } i}} \left(\eta^{n_j(a^t)-1} - \eta^{n_j(a^t)}\right)$$

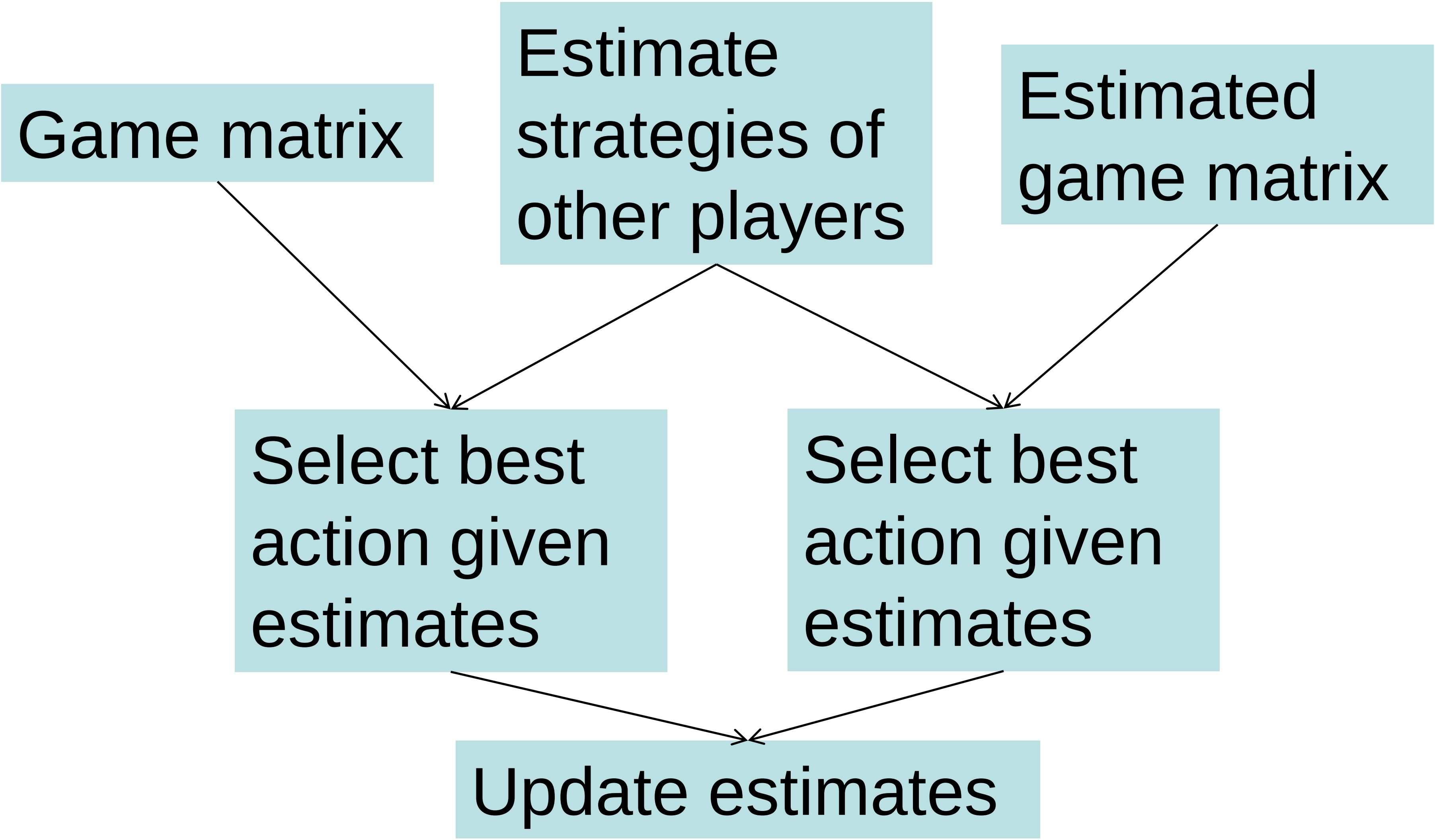
Marginal contributions



Graphical games



Local learning



Game matrix

Estimate strategies of other players

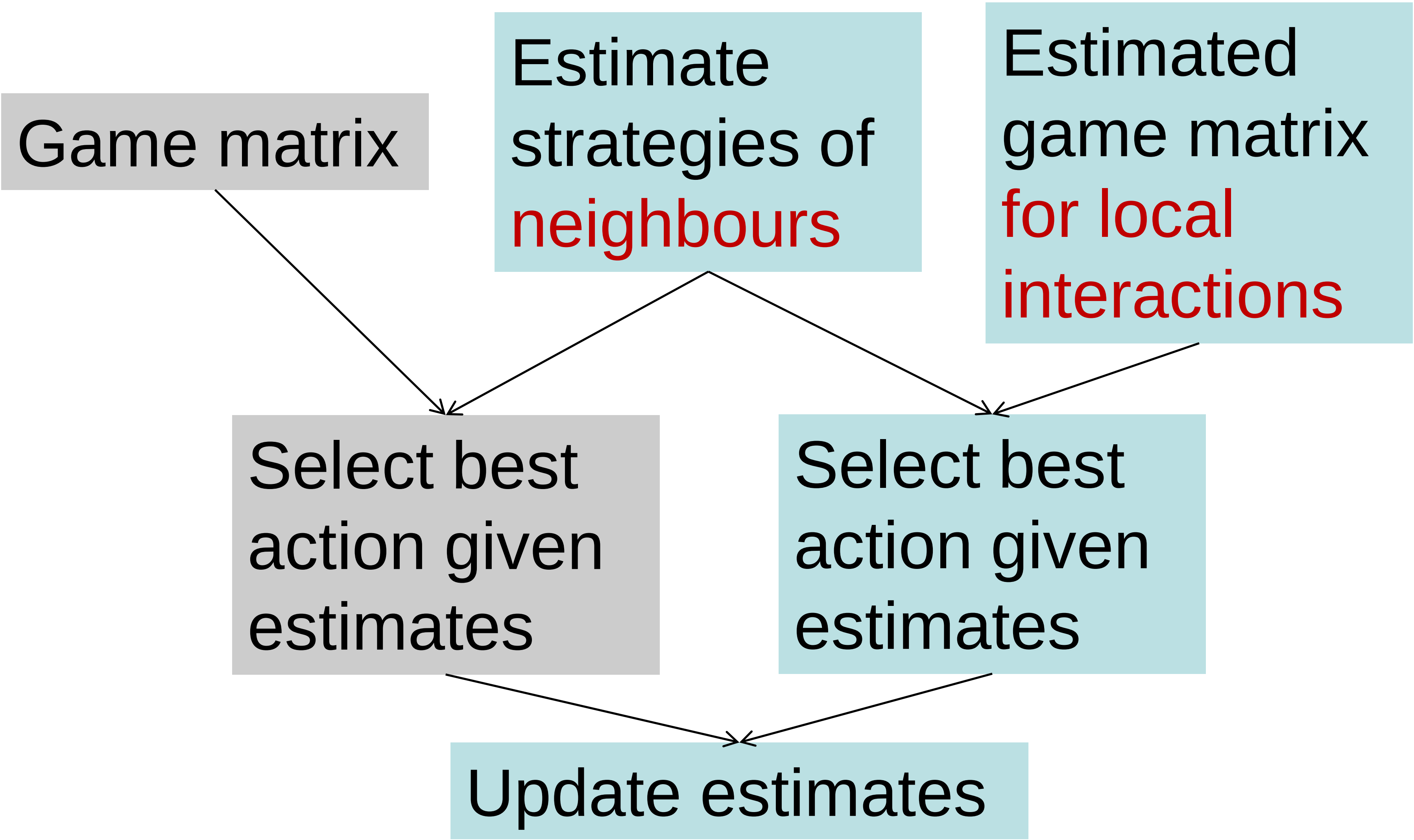
Estimated game matrix

Select best action given estimates

Select best action given estimates

Update estimates

Local learning



Game matrix

Estimate strategies of neighbours

Estimated game matrix for local interactions

Select best action given estimates

Select best action given estimates

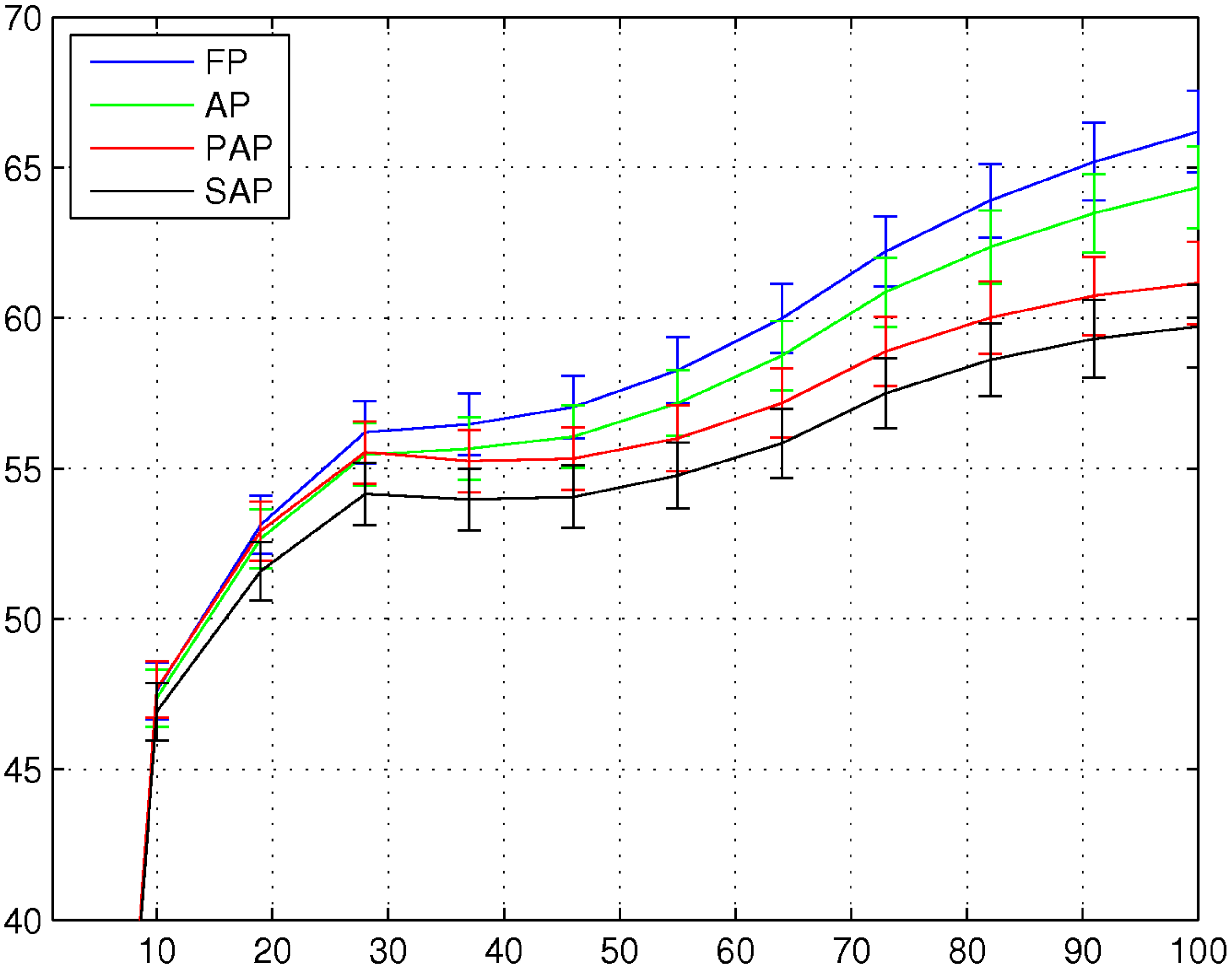
Update estimates

Theoretical result

Theorem – If all joint actions ***of local games*** are played infinitely often then beliefs follow a GWFP

Proof: The estimated game converges to the true game, so selected strategies are ε-best responses.

Sensing results



autonomous learning agents for decentralised data and information networks

So what?!

- Play converges to (local) optimum with only noisy information and local communication
- An individual always chooses an action to maximise expected reward given information
- If an individual doesn’t “play cricket”, the other individuals will reach an optimal point conditional on the behaviour of the itinerant

Summary

- Learning the game while playing is essential
- This can be accommodated within the GWFP framework
- Exploiting the neighbourhood structure of marginal contributions is essential for feasibility