Generalised weakened fictitious play and random belief learning

David S. Leslie

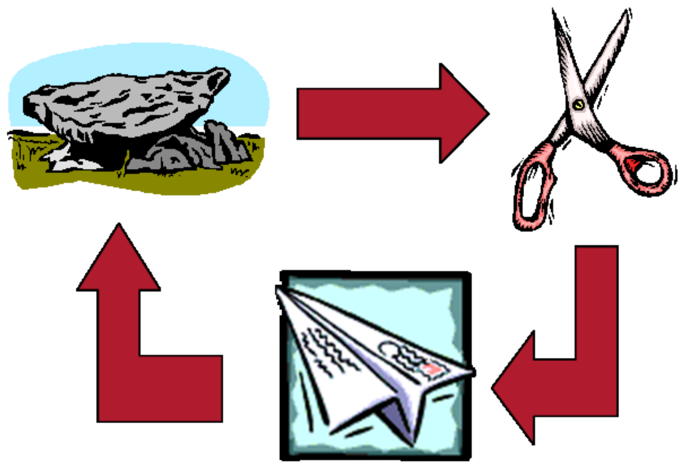
12 April 2010

Collaborators: **Sean Collins, Claudio Mezzetti, Archie Chapman**

Overview

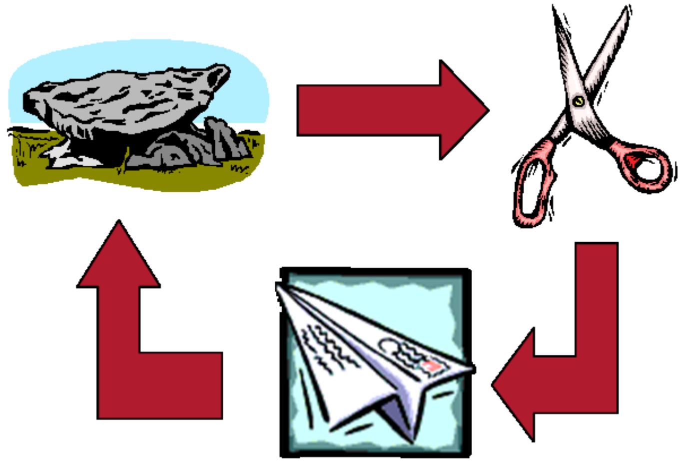
- Learning in games
- Stochastic approximation
- Generalised weakened fictitious play
  - Random belief learning
  - Oblivious learners

Normal form games



- Players $i = 1, \ldots, N$
- Action sets A^i
- Reward functions $r^i : A^1 \times \cdots \times A^N \to \mathbb{R}$

Mixed strategies



- Mixed strategies $\pi^i \in \Delta^i$
- Joint mixed strategy $\pi = (\pi^1, \ldots, \pi^N)$
- Reward function extended so that $r^i(\pi) = \mathbb{E}_\pi[r^i(\boldsymbol{a})]$

Best responses

Assume other players use mixed strategy π^{-i}.

Player i should choose a mixed strategy in the best response set

$$b^i(\pi^{-i}) = \operatorname*{argmax}_{\tilde{\pi}^i \in \Delta^i} r^i(\tilde{\pi}^i, \pi^{-i})$$

Best responses

Assume other players use mixed strategy π^{-i}.

Player i should choose a mixed strategy in the best response set

$$b^i(\pi^{-i}) = \operatorname*{argmax}_{\tilde{\pi}^i \in \Delta^i} r^i(\tilde{\pi}^i, \pi^{-i})$$

A Nash equilibrium is a fixed point of the best response map:

$$\pi^i \in b^i(\pi^{-i}) \quad \text{for all } i$$

A problem with Nash

Consider the game

$$\begin{pmatrix} (2,0) & (0,1) \\ (0,2) & (1,0) \end{pmatrix}$$

with unique Nash equilibrium

$$\pi^1 = (2/3, 1/3), \quad \pi^2 = (1/3, 2/3)$$

A problem with Nash

Consider the game

$$\begin{pmatrix} (2,0) & (0,1) \\ (0,2) & (1,0) \end{pmatrix}$$

with unique Nash equilibrium

$$\pi^1 = (2/3, 1/3), \quad \pi^2 = (1/3, 2/3)$$

- $r^i(a^i, \pi^{-i}) = 2/3$ for each i, a^i
- How does Player 1 know to use $\pi^1 = (2/3, 1/3)$?
- Player 2 to use $\pi^2 = (1/3, 2/3)$?

Learning in games

- Attempts to justify equilibrium play as the end point of a learning process

- Generally assumes pretty stupid players!

- Related to evolutionary game theory

Multi-armed bandits

At time n, choose action a_n, and receive reward R_n

Multi-armed bandits

Estimate after time n of the expected reward for action $a \in A$ is:

$$Q_n(a) = \frac{\displaystyle\sum_{m \leq n \,:\, a_m = a} R_m}{\kappa_n(a)}$$

where $\kappa_n(a) = \sum_{m=1}^{n} \mathbb{I}\{a_m = a\}$

Multi-armed bandits

If $a_n \neq a$, $\kappa_n(a) = \kappa_{n-1}(a)$ and:

$$Q_n(a) = \frac{\left(\sum_{m=1}^{n-1} \mathbb{I}\{a_m = a\} R_m\right) + 0}{\kappa_{n-1}(a)} = Q_{n-1}(a)$$

Multi-armed bandits

if $a_n = a$,

$$\begin{aligned} Q_n(a) &= \frac{\left(\sum_{m=1}^{n-1} \mathbb{I}\{a_m = a\} R_m\right) + R_n}{\kappa_n(a)} \\ &= \left(1 - \frac{1}{\kappa_n(a)}\right) Q_{n-1}(a) + \frac{1}{\kappa_n(a)} R_n \end{aligned}$$

Multi-armed bandits

Update estimates using

$$Q_n(a) = \begin{cases} Q_{n-1}(a) + \frac{1}{\kappa_n(a)}\{R_n - Q_{n-1}(a)\} & \text{if } a_n = a \\ Q_{n-1}(a) & \text{if } a_n \neq a \end{cases}$$

At time $n+1$ use Q_n to choose an action a_{n+1}

Fictitious play

At iteration $n+1$, player i:

- forms beliefs $\sigma_n^{-i} \in \Delta^{-i}$ about the other players' strategies
- chooses an action in $b^i(\sigma_n^{-i})$

Belief formation

The beliefs about player j are simply the MLE:

$$\sigma_n^j(a^j) = \frac{\kappa_n^j(a^j)}{n} \qquad \text{where } \kappa_n^j(a^j) = \sum_{m=1}^{n} \mathbb{I}\{a_m^j = a^j\}$$

Belief formation

The beliefs about player j are simply the MLE:

$$\sigma_n^j(a^j) = \frac{\kappa_n^j(a^j)}{n} \qquad \text{where } \kappa_n^j(a^j) = \sum_{m=1}^{n} \mathbb{I}\{a_m^j = a^j\}$$

Recursive update:

$$\sigma_{n+1}^j(a^j) = \frac{\kappa_{n+1}^j(a^j)}{n+1} = \frac{\kappa_n^j(a^j) + \mathbb{I}\{a_{n+1}^j = a^j\}}{n+1}$$

Belief formation

The beliefs about player j are simply the MLE:

$$\sigma_n^j(a^j) = \frac{\kappa_n^j(a^j)}{n} \qquad \text{where } \kappa_n^j(a^j) = \sum_{m=1}^n \mathbb{I}\{a_m^j = a^j\}$$

Recursive update:

$$\sigma_{n+1}^j(a^j) = \frac{\kappa_{n+1}^j(a^j)}{n+1} = \frac{\kappa_n^j(a^j)+\mathbb{I}\{a_{n+1}^j=a^j\}}{n+1} = \frac{n}{n+1}\frac{\kappa_n^j(a^j)}{n}+\frac{\mathbb{I}\{a_{n+1}^j=a^j\}}{n+1}$$

Belief formation

The beliefs about player j are simply the MLE:

$$\sigma_n^j(a^j) = \frac{\kappa_n^j(a^j)}{n} \qquad \text{where } \kappa_n^j(a^j) = \sum_{m=1}^{n} \mathbb{I}\{a_m^j = a^j\}$$

Recursive update:

$$\sigma_{n+1}^j(a^j) = \left(1 - \frac{1}{n+1}\right) \sigma_n^j(a^j) + \frac{1}{n+1} \mathbb{I}\{a_{n+1}^j = a^j\}$$

Belief formation

The beliefs about player j are simply the MLE:

$$\sigma_n^j(a^j) = \frac{\kappa_n^j(a^j)}{n} \qquad \text{where } \kappa_n^j(a^j) = \sum_{m=1}^{n} \mathbb{I}\{a_m^j = a^j\}$$

Recursive update:

$$\sigma_{n+1}^j = \left(1 - \tfrac{1}{n+1}\right) \sigma_n^j + \tfrac{1}{n+1} \boldsymbol{e}_{a_{n+1}^j}$$

Belief formation

The beliefs about player j are simply the MLE:

$$\sigma_n^j(a^j) = \frac{\kappa_n^j(a^j)}{n} \qquad \text{where } \kappa_n^j(a^j) = \sum_{m=1}^n \mathbb{I}\{a_m^j = a^j\}$$

Recursive update:

$$\sigma_{n+1}^j \qquad = \left(1 - \tfrac{1}{n+1}\right)\sigma_n^j \qquad + \tfrac{1}{n+1}\boldsymbol{e}_{a_{n+1}^j}$$

In terms of best responses:

$$\sigma_{n+1}^j \qquad \in \left(1 - \tfrac{1}{n+1}\right)\sigma_n^j \qquad + \tfrac{1}{n+1}b^j(\sigma_n^{-j})$$

Belief formation

The beliefs about player j are simply the MLE:

$$\sigma_n^j(a^j) = \frac{\kappa_n^j(a^j)}{n} \qquad \text{where } \kappa_n^j(a^j) = \textstyle\sum_{m=1}^n \mathbb{I}\{a_m^j = a^j\}$$

Recursive update:

$$\sigma_{n+1}^j \qquad = \left(1 - \tfrac{1}{n+1}\right)\sigma_n^j \qquad + \tfrac{1}{n+1}\boldsymbol{e}_{a_{n+1}^j}$$

In terms of best responses:

$$\sigma_{n+1} \qquad \in \left(1 - \tfrac{1}{n+1}\right)\sigma_n \qquad + \tfrac{1}{n+1}b\ (\sigma_n\ )$$

Stochastic approximation

Stochastic approximation

$$\theta_{n+1} \in \theta_n + \alpha_{n+1}\{F(\theta_n) + M_{n+1}\}$$

Stochastic approximation

$$\boxed{\theta_{n+1} \in \theta_n + \alpha_{n+1}\left\{F(\theta_n) + M_{n+1}\right\}}$$

- $F : \Theta \to \Theta$ is a (bounded u.s.c.) set-valued map
- $\alpha_n \to 0$, $\sum_n \alpha_n = \infty$
- For any $T > 0$,

$$\lim_{n\to\infty} \sup_{k>n \,:\, \sum_{i=n}^{k-1} \alpha_{i+1} \le T} \left\| \sum_{i=n}^{k-1} \alpha_{i+1} M_{i+1} \right\| = 0$$

The last is implied by: $\sum_n (\alpha_n)^2 < \infty$, $\mathbb{E}[M_{n+1} \,|\, \theta_n] \to 0$, and $\mathrm{Var}[M_{n+1} \,|\, \theta_n] < C$ almost surely.

Stochastic approximation

$$\boxed{\theta_{n+1} \in \theta_n + \alpha_{n+1}\{F(\theta_n) + M_{n+1}\}}$$

$$\frac{\theta_{n+1} - \theta_n}{\alpha_n} \in F(\theta_n) + M_{n+1}$$

$$\downarrow$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta \in F(\theta),$$

a differential inclusion

(Benaïm, Hofbauer and Sorin, 2005)

Stochastic approximation

$$\boxed{\theta_{n+1} \in \theta_n + \alpha_{n+1} \{F(\theta_n) + M_{n+1}\}}$$

In fictitious play:

$$\sigma_{n+1} \in \sigma_n + \tfrac{1}{n+1} \{b(\sigma_n) - \sigma_n\}$$

$$\downarrow$$

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\sigma \in b(\sigma) - \sigma,$$

the best response differential inclusion.

Hence σ_n converges to the set of Nash equilibria in zero-sum games, potential games, and generic $2 \times m$ games.

Generalised weakened fictitious play

Weakened fictitious play

- Van der Genugten (2000) showed that the convergence rate of fictitious play can be improved if players use ϵ_n-best responses. (For 2-player zero-sum games, and a very specific choice of ϵ_n)

- $\pi \in b_{\epsilon_n}(\sigma_n) \quad \Rightarrow \quad \pi \in b(\sigma_n) + M_{n+1}$
  where $M_n \to 0$ as $\epsilon_n \to 0$ (by continuity properties of b and boundedness of r)

- For general games and general $\epsilon_n \to 0$ this fits into the stochastic approximation framework

Generalised weakened fictitious play

Theorem: Any process such that

$$\sigma_{n+1} \in \sigma_n + \alpha_{n+1} \{ b_{\epsilon_n}(\sigma_n) - \sigma_n + M_{n+1} \}$$

where

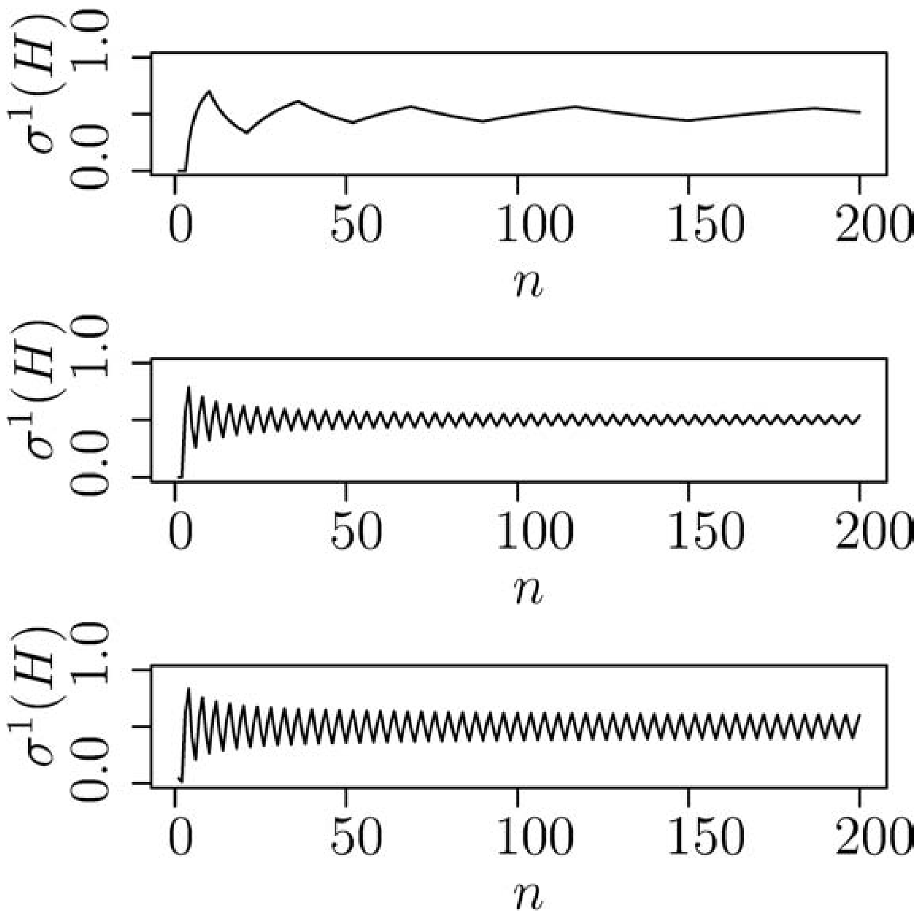
- $\epsilon_n \to 0$ as $n \to \infty$
- $\alpha_n \to 0$ as $n \to \infty$
- $\lim_{n\to\infty} \sup_{k>n : \sum_{i=n}^{k-1} \alpha_{i+1} \leq T} \left\| \sum_{i=n}^{k-1} \alpha_{i+1} M_{i+1} \right\| = 0$

converges to the set of Nash equilibria for zero-sum games, potential games and generic $2 \times m$ games.

Recency

- For classical fictitious play $\alpha_n = \frac{1}{n}$, $\epsilon_n \equiv 0$ and $M_n \equiv 0$
- For any $\alpha_n \to 0$ the conditions are met (since $M_n \equiv 0$)
- How about $\alpha_n = \frac{1}{\sqrt{n}}$, or even $\alpha_n = \frac{1}{\log n}$?

Recency



Belief that Player 1 plays Heads over 200 plays of the two-player matching pennies game under classical fictitious play (top), under a modified fictitious play with $\alpha_n = \frac{1}{\sqrt{n}}$ (middle), and with $\alpha_n = \frac{1}{\log n}$ (bottom)

Stochastic fictitious play

In fictitious play, players always choose pure actions

$\Rightarrow$ strategies never converge to mixed strategies

(beliefs do, but played strategies do not)

Stochastic fictitious play

Instead consider smooth best responses:

$$\beta_\tau^i(\sigma^{-i}) = \underset{\pi^i \in \Delta^i}{\operatorname{argmax}} \left\{ r^i(\pi^i, \sigma^{-i}) + \tau v(\pi^i) \right\}$$

For example $\beta_\tau^i(\sigma^{-i})(a^i) = \frac{\exp\{r^i(a^i,\sigma^{-i})/\tau\}}{\sum_{a \in A^i} \exp\{r^i(a,\sigma^{-i})/\tau\}}$

Stochastic fictitious play

Instead consider smooth best responses:

$$\beta^i_\tau(\sigma^{-i}) = \operatorname*{argmax}_{\pi^i \in \Delta^i} \left\{ r^i(\pi^i, \sigma^{-i}) + \tau v(\pi^i) \right\}$$

For example $\beta^i_\tau(\sigma^{-i})(a^i) = \frac{\exp\left\{r^i(a^i,\sigma^{-i})/\tau\right\}}{\sum_{a \in A^i} \exp\left\{r^i(a,\sigma^{-i})/\tau\right\}}$

Strategies evolve according to

$$\sigma_{n+1} = \sigma_n + \frac{1}{n+1}\{\beta_\tau(\sigma_n) + M_{n+1} - \sigma_n\} \quad \text{where } \mathbb{E}[M_{n+1} \,|\, \sigma_n] = 0$$

Convergence

$$\sigma_{n+1} = \sigma_n + \frac{1}{n+1}\{\beta_\tau(\sigma_n) - \sigma_n + M_{n+1}\}$$

Convergence

$$\begin{aligned} \sigma_{n+1} &= \sigma_n + \tfrac{1}{n+1}\{\beta_\tau(\sigma_n) - \sigma_n + M_{n+1}\} \\ &\in \sigma_n + \tfrac{1}{n+1}\{\, b_\epsilon(\sigma_n) - \sigma_n + M_{n+1}\} \end{aligned}$$

Convergence

$$\begin{aligned}\sigma_{n+1} &= \sigma_n + \tfrac{1}{n+1}\{\beta_\tau(\sigma_n) - \sigma_n + M_{n+1}\} \\ &\in \sigma_n + \tfrac{1}{n+1}\{\, b_\epsilon(\sigma_n) - \sigma_n + M_{n+1}\}\end{aligned}$$

But can now consider the effect of using smooth best response β_{τ_n} with $\tau_n \to 0 \ldots$

. . . it means that $\epsilon_n \to 0$, resulting in a GWFP!

Random belief learning

Random beliefs

(Friedman and Mezzetti 2005)

Best response 'assumes' complete confidence in:

- knowledge of the reward functions
- beliefs σ about opponent strategy

Random beliefs

(Friedman and Mezzetti 2005)

Best response ‘assumes’ complete confidence in:

- knowledge of the reward functions
- beliefs σ about opponent strategy

Random beliefs

(Friedman and Mezzetti 2005)

Best response 'assumes' complete confidence in:

- knowledge of the reward functions
- beliefs σ about opponent strategy

Uncertainty in the beliefs $\sigma_n \longleftrightarrow$ distribution on belief space

Belief distributions

- The belief about player j is that $\pi^j \sim \mu^j$
- $\mathbb{E}_{\mu^j}[\pi^j] = \sigma^j$, the focus of μ^j.

Belief distributions

- The belief about player j is that $\pi^j \sim \mu^j$
- $\mathbb{E}_{\mu^j}[\pi^j] = \sigma^j$, the focus of μ^j.

Response to random beliefs:

sample $\pi^{-i} \sim \mu^{-i}$ and play $a^i \in b^i(\pi^{-i})$

Let $\tilde{b}^i(\mu^{-i})$ be the resulting mixed strategy

Random belief equilibrium

A random belief equilibrium is a set of belief distributions μ^i such that the focus of μ^i is equal to the mixed strategy played by i:

$$\mathbb{E}_{\mu^i}[\pi^i] = \tilde{b}^i(\mu^{-i})$$

A refinement of Nash equilibria when μ^i depends on ϵ and $\mathrm{Var}_{\mu^j_\epsilon}(\pi^j) \to 0$ as $\epsilon \to 0$.

Inference

- In fictitious play, σ_n^j is the MLE of π^j

Inference

- In fictitious play, σ_n^j is the MLE of π^j

- Fudenberg and Levine (1998): if the prior is Dirichlet($\boldsymbol{\alpha}$), then the posterior is Dirichlet($\boldsymbol{\alpha} + \boldsymbol{\kappa}$)

$$\Downarrow$$

Fictitious play is doing Bayesian learning, with best replies taken with respect to the expected opponent strategy

Random belief learning

- Start with priors μ_0^j

Random belief learning

- Start with priors μ_0^j

- After observing actions for n steps, have posteriors μ_n^j

Random belief learning

- Start with priors μ_0^j
- After observing actions for n steps, have posteriors μ_n^j
- Select actions using response to random beliefs (i.e. mixed strategy $\tilde{b}^i(\mu_n^{-i})$)

Convergence

Can show:

- $\tilde{b}^i(\mu_n^{-i}) \in b_{\epsilon_n}(\sigma_n^{-i})$

- So the beliefs follow a GWFP process

Unfortunately it is the beliefs, not the strategies.

Learning the game

Best response 'assumes' complete confidence in:

- knowledge of the reward functions
- beliefs σ about opponent strategy

Learning the game

Best response 'assumes' complete confidence in:

- knowledge of the reward functions
- beliefs σ about opponent strategy

Learn reward matrices using reinforcement learning ideas:

- at iteration n, observe joint action $\boldsymbol{a}_n$ and reward $R^i(\boldsymbol{a}_n) = r^i(\boldsymbol{a}_n) + \epsilon_n$
- update estimates σ^{-i} of opponent strategies
- update estimate $Q^i(\boldsymbol{a}_n)$ of $r^i(\boldsymbol{a}_n)$

Convergence

Assume all joint actions $\boldsymbol{a}$ are played infinitely often. Can show:

- $Q_n^i(\boldsymbol{a}) \to r^i(\boldsymbol{a})$ for all $\boldsymbol{a}$
- Best responses with to σ_n^{-i} with respect to Q_n^i are ϵ_n-best responses with respect to r^i
- So the beliefs follow a GWFP process

Potentially very useful in DCOP games (Chapman, Rogers, Jennings and Leslie 2008)

Oblivious learners

Oblivious learners

What if players are oblivious to opponents?

Each individual treats the problem a multi-armed bandit

Can we expect equilibrium play?

Best response/inertia

Suppose individuals (somehow by magic) actually know $Q^i(a^i) = r^i(a^i, \pi_n^{-i})$

They can adjust their own strategy towards a best response:

$$\pi_{n+1}^i = (1 - \alpha_{n+1})\pi_n^i + \alpha_{n+1} b^i(\pi^{-i})$$

Strategies converge, not just beliefs

But it's just not possible

If π^{-i} were fixed. . .

- Player i actually faces a multi-armed bandit

- So can learn $Q^i(a^i)$ by playing all actions infinitely often

- Then adjust π^i

Actor–critic learning

$$Q^i_{n+1}(a^i_{n+1}) = Q^i_n(a^i_{n+1}) + \lambda_{n+1}\left\{R_{n+1} - Q_n(a^i_{n+1})\right\}$$

$$\pi^i_{n+1} = \pi^i_n + \alpha_n\left\{b^i(Q^i_n) - \pi^i_n\right\}$$

Actor–critic learning

$$Q^i_{n+1}(a^i_{n+1}) = Q^i_n(a^i_{n+1}) + \lambda_{n+1}\left\{R_{n+1} - Q_n(a^i_{n+1})\right\}$$

$$\pi^i_{n+1} = \pi^i_n + \alpha_n\left\{b^i(Q^i_n) - \pi^i_n\right\}$$

With all players adjusting simultaneously, need to be careful

If $\frac{\alpha_n}{\lambda_n} \to 0$, the system can be analysed as if all players have accurate Q values.

Convergence

- Can show that $|Q_n^i(a^i) - r^i(a^i, \pi_n^{-i})| \to 0$
- So best responses with respect to the Q^i's are ϵ-best responses to π_n^{-i}
- So the π_n follow a GWFP process

We have a process under which played strategy converges to Nash equilibrium

Conclusions

- Generalised weakened fictitious play is a class that is closely related to the best response dynamics
- All GWFP processes converge to Nash equilibrium in zero-sum games, potential games, and generic $2 \times m$ games
- GWFP encompasses numerous models of learning in games:
  - Fictitious play with greater weight on recent observations
  - Stochastic fictitious play with vanishing smoothing
  - Random belief learning
  - Fictitious play while learning the reward matrices
  - An oblivious actor–critic process