Mathematics & Statistics | Lancaster University

# Multiple-timescales learning in games

February 2025
David S. Leslie

# Introduction

## . . . ... to me

Mathematics & Statistics | Lancaster University

- Prof. of Statistical Learning at Lancaster since 2014. Previously mathematics department at University of Bristol
- PI on EPSRC Data Science of the Natural Environment project (2018–2023)
- Researcher on EPSRC/BT Next Generation Converged Digital Infrastructure (2018–2023)
- Was consultant at Prowler.io (now Secondmind.ai), 2018–2020.

# Introduction



## . . . to mathematical sciences at Lancaster

### MARS (Maths for AI in Real-world Systems

- £15M investment to expand mathematical sciences at Lancaster (focus is AI especially with applications in health, environment, engineering, cybersecurity)
- 10 new permanent positions, 4 still to recruit (all levels)
- 8 post-doc positions, recruiting 2025/26

### ProbAI research hub

- £10M to build collaborations across multiple universities and industry, focusing on probabilistic techniques for AI
- Recruiting post-docs imminently

# Introduction

## . . . to the talk



- Introduction
- Stochastic fictitious play and stochastic approximation
- Two-timescales stochastic approximation
- Applications:
  - Actor–critic learning
  - Player-dependent learning rates
  - Learning in stochastic games
  - Noise reduction in gradient estimation

# Inner loops



Often we might want to run an inner loop between adaptations:

**Clinical trials** Several treatments. Experiment enough with each treatment. Adapt the set of treatments and repeat.

**Games** Fix the (mixed) strategies. Play long enough to learn the strategies. Adapt the strategies and repeat.

**Deep learning** Fix the weights. Gather enough observations with these weights. Adapt the weights and repeat.

# Inner loops



Often we might want to run an inner loop between adaptations:

**Clinical trials** Several treatments. Experiment enough with each treatment. Adapt the set of treatments and repeat.

**Games** Fix the (mixed) strategies. Play long enough to learn the strategies. Adapt the strategies and repeat.

**Deep learning** Fix the weights. Gather enough observations with these weights. Adapt the weights and repeat.

**Generally** A system has parameters $\theta$ and a performance gradient $v(\theta)$. If $v$ is not analytically available, fix $\theta$ for long enough to reliably estimate $v(\theta)$ on the basis of observations, update $\theta$ and repeat.

Two timescales helps to avoid “fix” and “enough”

# Normal form games



$$
\begin{array}{c} \\ R \\ S \\ P \end{array}
\begin{array}{c}
\begin{array}{ccc} \quad R \quad & \quad S \quad & \quad P \quad \end{array} \\
\left(\begin{array}{ccc}
(0,0) & (1,-1) & (-1,1) \\
(-1,1) & (0,0) & (1,-1) \\
(1,-1) & (-1,1) & (0,0)
\end{array}\right)
\end{array}
$$

- Finite set of players, labeled $i$
- Each player has an action space $A^i$; joint action space $A = A^1 \times \cdots \times A^N$
- Usually we consider mixed strategies $\pi^i \in \Delta(A^i)$; joint mixed strategies $\pi \in \Delta$
- Reward functions extend to $r^i : \Delta \to \mathbb{R}$

# Normal form games

## Equilibrium

Response to (beliefs about) other players becomes key.
Define the best response correspondence

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

$\{$Nash equilibria$\} := \{$fixed points $\pi^i \in b^i(\pi^{-i})\}$

# Normal form games

## Equilibrium



Response to (beliefs about) other players becomes key.
Define the best response correspondence

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

$\{$Nash equilibria$\} := \{$fixed points $\pi^i \in b^i(\pi^{-i})\}$

Often we want a continuous response. The canonical example is the smooth best response function $\beta^i_\tau(\pi^{-i})$ satisfying

$$\beta^i_\tau(\pi^{-i})(a^i) \propto \exp(r^i(a^i, \pi^{-i})/\tau)$$

$\{$fixed points $\pi^i = \beta^i_\tau(\pi^{-i})\} =: \{$smoothed Nash equilibrium$\}$

# Normal form games

## Fictitious play



Even if the game is fully known, things are non-trivial!

### Fictitious play

- Repeatedly play the game
- On iteration $n$, estimate $\pi^i$ by $\sigma^i_n$, the empirically observed distribution of opponent actions so far
- Play a best response $a^i_{n+1} \in b^i(\sigma^{-i}_n)$

$$\sigma^i_{n+1}(a^i) = \frac{1}{n+1}\sum_{m=1}^{n+1} \mathbb{I}_{\{a^i_m = a^i\}} = \sigma^i_n(a^i) + \frac{1}{n+1}\left[\mathbb{I}_{\{a^i_{n+1}=a^i\}} - \sigma^i_n(a^i)\right]$$

# Normal form games

## Fictitious play

Even if the game is fully known, things are non-trivial!

### Fictitious play

- Repeatedly play the game
- On iteration $n$, estimate $\pi^i$ by $\sigma^i_n$, the empirically observed distribution of opponent actions so far
- Play a best response $a^i_{n+1} \in b^i(\sigma^{-i}_n)$

$$\sigma^i_{n+1}(a^i) = \frac{1}{n+1} \sum_{m=1}^{n+1} \mathbb{I}_{\{a^i_m = a^i\}} = \sigma^i_n(a^i) + \frac{1}{n+1} \left[ \mathbb{I}_{\{a^i_{n+1} = a^i\}} - \sigma^i_n(a^i) \right]$$

$$\sigma^i_{n+1} = \sigma^i_n + \frac{1}{n+1} \left[ b^i(\sigma^{-i}_n) - \sigma^i_n \right]$$

# Normal form games

## Fictitious play

Even if the game is fully known, things are non-trivial!

### ~~Fictitious play~~ Stochastic fictitious play

- Repeatedly play the game
- On iteration $n$, estimate $\pi^i$ by $\sigma^i_n$, the empirically observed distribution of opponent actions so far
- ~~Play a best response $a^i_{n+1} \in b^i(\sigma^{-i}_n)$~~ Play $a^i_{n+1} \sim \beta^i(\sigma^{-i}_n)$

$$\sigma^i_{n+1}(a^i) = \frac{1}{n+1}\sum_{m=1}^{n+1} \mathbb{I}_{\{a^i_m = a^i\}} = \sigma^i_n(a^i) + \frac{1}{n+1}\left[\mathbb{I}_{\{a^i_{n+1}=a^i\}} - \sigma^i_n(a^i)\right]$$

$$\sigma^i_{n+1} = \sigma^i_n + \frac{1}{n+1}\left[\beta^i(\sigma^{-i}_n) - \sigma^i_n + M^i_{n+1}\right]$$

# Stochastic approximation



$$\theta_{t+1} = \theta_t + \alpha_{t+1} \left\{ F(\theta_t) + e_t + M_{t+1} \right\}$$

with $\alpha_t \to 0$, $e_t \to 0$ and $M_t$ a "Martingale difference sequence".

- Robbins–Monro
- Kiefer-Wolfowicz
- Ljung
- Kushner
- Benveniste, Metivier and Priouret
- Duflo
- Borkar
- Benaïm

# Stochastic approximation



$$\theta_{t+1} = \theta_t + \alpha_{t+1}\left\{F(\theta_t) + e_t + M_{t+1}\right\}$$

with $\alpha_t \to 0$, $e_t \to 0$ and $M_t$ a "Martingale difference sequence".

Rearrange:

$$\frac{\theta_{t+1} - \theta_t}{\alpha_t} = F(\theta_t) + e_t + M_{t+1}$$

# Stochastic approximation



$$\theta_{t+1} = \theta_t + \alpha_{t+1} \left\{ F(\theta_t) + e_t + M_{t+1} \right\}$$

with $\alpha_t \to 0$, $e_t \to 0$ and $M_t$ a "Martingale difference sequence".

Rearrange:

$$\frac{\theta_{t+1} - \theta_t}{\alpha_t} = F(\theta_t) + \cancel{e_t + M_{t+1}}$$

Looks like a discretisation of

$$\dot{\theta} = F(\theta).$$

# Stochastic approximation



$$\theta_{t+1} = \theta_t + \alpha_{t+1}\left\{F(\theta_t) + e_t + M_{t+1}\right\}$$

with $\alpha_t \to 0$, $e_t \to 0$ and $M_t$ a "Martingale difference sequence".

Rearrange:

$$\frac{\theta_{t+1} - \theta_t}{\alpha_t} = F(\theta_t) + \cancel{e_t + M_{t+1}}$$

Looks like a discretisation of

$$\dot{\theta} = F(\theta).$$

**Theorem (ish)**

If the ODE has a unique globally attracting fixed point $\theta^\star$ then the stochastic approximation iterates converge almost surely to $\theta^\star$

# Normal form games

## Smooth best response dynamics



Recall stochastic fictitious play (SFP):

$$\sigma_{t+1} = \sigma_t + \frac{1}{t+1} \{\beta(\sigma_t) - \sigma_t + M_{t+1}\}$$

This is a stochastic approximation with $F(\sigma_t) = \beta(\sigma_t) - \sigma_t$

Hence SFP converges if the smooth best response dynamics

$$\dot{\sigma} = \beta(\sigma) - \sigma$$

are globally convergent

Convergence in zero-sum-games, potential games, some other less obvious classes (Benaïm and Hirsch, Hofbauer, others)

# Normal form games

## Radically uncoupled



Suppose can't observe opponent actions and don't know the payoff matrix. Now what?!

# Normal form games

## Radically uncoupled



Suppose can't observe opponent actions and don't know the payoff matrix. Now what?!

Each player now faces a bandit problem $\Rightarrow$ Use RL in bandits approach $\Rightarrow$ Individual $Q$-learning (Leslie and Collins 2006)

# Normal form games

## Radically uncoupled

Suppose can't observe opponent actions and don't know the payoff matrix. Now what?!

Each player now faces a bandit problem $\Rightarrow$ Use RL in bandits approach $\Rightarrow$ Individual $Q$-learning (Leslie and Collins 2006)

Can mixed strategies behave like fictitious play beliefs

$$\pi_{t+1} = \pi_t + \alpha_{t+1} \{\beta(\pi_t) - \pi_t\}?$$

Yes, if each player can calculate $\beta^i(\pi_t^{-i}) \propto \exp(r^i(\cdot, \pi_t^{-i})/\tau)$

# Normal form games

## Estimating $r^i(\cdot, \pi^{-i})$

- "Wait everybody, don't move your $\pi^i$, we're all going to observe for a while"
- Play repeatedly and estimate $r^i(a^i, \pi^{-i})$ to be the average reward obtained with $i$ play action $a^i$
- When these have converged, everybody adjust $\pi^i$ a little bit

# Normal form games

## Estimating $r^i(\cdot, \pi^{-i})$

- "Wait everybody, don't move your $\pi^i$, we're all going to observe for a while"
- Play repeatedly and estimate $r^i(a^i, \pi^{-i})$ to be the average reward obtained with $i$ play action $a^i$
- When these have converged, everybody adjust $\pi^i$ a little bit

Like in fictitious play, the averages can be calculated "online":

$$Q^i_{n+1}(a^i) = Q^i_n(a^i) + \frac{\mathbb{I}_{\{a^i_n = a^i\}}}{\kappa^i_n(a^i)} \left\{ R^i_n - Q^i_n(a^i) \right\}$$

# Stochastic approximation



## Two timescales (Borkar 1997)

### Two SA processes, with $\alpha_n/\gamma_n \to 0$

$$\theta_{n+1} = \theta_n + \alpha_{n+1}\left\{F(\theta_n, \phi_n) + e_n + M_{n+1}\right\}$$
$$\phi_{n+1} = \phi_n + \gamma_{n+1}\left\{G(\theta_n, \phi_n) + h_n + N_{n+1}\right\}$$

# Stochastic approximation

Mathematics & Statistics | Lancaster University

## Two timescales (Borkar 1997)

**Two SA processes, with $\alpha_n/\gamma_n \to 0$**

$$\theta_{n+1} = \theta_n + \alpha_{n+1}\left\{F(\theta_n, \phi_n) + e_n + M_{n+1}\right\}$$
$$\phi_{n+1} = \phi_n + \gamma_{n+1}\left\{G(\theta_n, \phi_n) + h_n + N_{n+1}\right\}$$

Rewrite as a single SA, with learning parameters $\gamma_n$

$$\begin{pmatrix}\theta_{n+1}\\ \phi_{n+1}\end{pmatrix} = \begin{pmatrix}\theta_n\\ \phi_n\end{pmatrix} + \gamma_{n+1}\begin{pmatrix} & \frac{\alpha_{n+1}}{\gamma_{n+1}}\left\{F(\theta_n, \phi_n) + e_n + M_{n+1}\right\} \\ G(\theta_n, \phi_n) + & h_n + N_{n+1}\end{pmatrix}$$

# Stochastic approximation

Mathematics & Statistics | Lancaster University

## Two timescales (Borkar 1997)

**Two SA processes, with $\alpha_n/\gamma_n \to 0$**

$$\theta_{n+1} = \theta_n + \alpha_{n+1}\left\{F(\theta_n, \phi_n) + e_n + M_{n+1}\right\}$$
$$\phi_{n+1} = \phi_n + \gamma_{n+1}\left\{G(\theta_n, \phi_n) + h_n + N_{n+1}\right\}$$

Rewrite as a single SA, with learning parameters $\gamma_n$

$$\begin{pmatrix}\theta_{n+1}\\ \phi_{n+1}\end{pmatrix} = \begin{pmatrix}\theta_n\\ \phi_n\end{pmatrix} + \gamma_{n+1}\begin{pmatrix} 0+ & \tilde{e}_n \\ G(\theta_n, \phi_n)+ & h_n + N_{n+1}\end{pmatrix}$$

The approximated differential equation is

$$\begin{pmatrix}\dot{\theta}\\ \phi\end{pmatrix} = \begin{pmatrix}0\\ G(\theta, \phi)\end{pmatrix}$$

# Stochastic approximation



## Convergence of the fast timescale

$$\begin{pmatrix} \dot{\theta} \\ \dot{\phi} \end{pmatrix} = \begin{pmatrix} 0 \\ G(\theta, \phi) \end{pmatrix}$$

**Assumption:**

For each $\theta$ there is a unique, globally attracting, fixed point of the "fast ODE" $\dot{\phi} = G(\theta, \phi)$. Call this $\phi^{\star}(\theta)$.

Under this assumption, the set

$$\left\{ \begin{pmatrix} \theta \\ \phi^{\star}(\theta) \end{pmatrix} : \theta \in \Theta \right\}$$

is globally attracting; $\binom{\theta_n}{\phi_n}$ converges to this set

# Stochastic approximation



## Convergence of the slow timescale

We have shown that $\phi_n = \phi^\star(\theta_n) + \epsilon_n$. So

$$
\begin{aligned}
\theta_{n+1} &= \theta_n + \alpha_{n+1}\left\{F(\theta_n, \phi_n) + e_n + M_{n+1}\right\} \\
&= \theta_n + \alpha_{n+1}\left\{F(\theta_n, \phi^\star(\theta_n) + \epsilon_n) + e_n + M_{n+1}\right\} \\
&= \theta_n + \alpha_{n+1}\left\{F(\theta_n, \phi^\star(\theta_n)) + \eta_n + e_n + M_{n+1}\right\}
\end{aligned}
$$

**The "slow ODE" is**

$$\dot{\theta} = F(\theta, \phi^\star(\theta))$$

If the fast and slow ODEs both converge, then we're in business!

# Normal form games

## Actor–critic



Put the inner loop estimation of $r^i(\cdot, \pi^{-i})$ on the fast timescale:

$$\begin{aligned} \pi_{n+1} = \pi_n \quad & + \alpha_{n+1} \left\{ \beta(Q_n) - \pi_n + M_{n+1} \right\} \\ Q^i_{n+1}(a^i) = Q^i_n(a^i) \quad & + \gamma_{n+1} \mathbb{I}_{\{a^i_n = a^i\}} \left\{ R^i_{n+1} - Q^i_n(a^i) \right\} \end{aligned}$$

with learning parameters such that $\frac{\alpha_n}{\gamma_n} \to 0$

# Normal form games

## Actor–critic



Put the inner loop estimation of $r^i(\cdot, \pi^{-i})$ on the fast timescale:

$$\begin{aligned} \pi_{n+1} &= \pi_n && +\alpha_{n+1}\left\{\beta(Q_n) - \pi_n + M_{n+1}\right\} \\ Q^i_{n+1}(a^i) &= Q^i_n(a^i) && +\gamma_{n+1}\mathbb{I}_{\{a^i_n = a^i\}}\left\{R^i_{n+1} - Q^i_n(a^i)\right\} \end{aligned}$$

with learning parameters such that $\frac{\alpha_n}{\gamma_n} \to 0$

**Fast timescale: fix $\pi$**

$$\dot{Q}^i(a^i) = \pi^i(a^i)\left\{r^i(a^i, \pi^{-i}) - Q^i(a^i)\right\}$$

This ODE converges: $Q^i(a^i) \to r^i(a^i, \pi^{-i}) =: Q^{\star,i}(\pi)(a^i)$

Therefore $Q^i_n$ will be close to $Q^{\star,i}(\pi_n)$ for large $n$

# Normal form games

## Actor–critic

Put the inner loop estimation of $r^i(\cdot, \pi^{-i})$ on the fast timescale:

$$\begin{aligned} \pi_{n+1} &= \pi_n && +\alpha_{n+1}\left\{\beta(Q_n) - \pi_n + M_{n+1}\right\} \\ Q^i_{n+1}(a^i) &= Q^i_n(a^i) && +\gamma_{n+1}\mathbb{I}_{\{a^i_n = a^i\}}\left\{R^i_{n+1} - Q^i_n(a^i)\right\} \end{aligned}$$

with learning parameters such that $\frac{\alpha_n}{\gamma_n} \to 0$

**Slow timescale: analyse as if $Q^i(a^i) = r^i(a^i, \pi^{-i})$**

$$\dot{\pi}^i = \beta(Q^{\star,i}(\pi)) - \pi^i) = \beta(r^i(\cdot, \pi^{-i})) - \pi^i = \beta(\pi^{-i}) - \pi^i$$

which is the smooth best response dynamics

The actor–critic algorithm converges in the same games as stochastic fictitious play

# Player-dependent rates

## Leslie and Collins (2003)



Revert to stochastic fictitious play, two player games:

$$\sigma^1_{n+1} = \sigma^1_n + \frac{1}{n+1}\left\{\beta^1(\sigma^2_n) - \sigma^1_n + M^1_{n+1}\right\}$$
$$\sigma^2_{n+1} = \sigma^2_n + \frac{1}{n+1}\left\{\beta^2(\sigma^1_n) - \sigma^2_n + M^2_{n+1}\right\}$$

# Player-dependent rates

## Leslie and Collins (2003)

Revert to stochastic fictitious play, two player games:

$$\sigma^1_{n+1} = \sigma^1_n + \alpha_{n+1} \left\{ \beta^1(\sigma^2_n) - \sigma^1_n + M^1_{n+1} \right\}$$
$$\sigma^2_{n+1} = \sigma^2_n + \alpha_{n+1} \left\{ \beta^2(\sigma^1_n) - \sigma^2_n + M^2_{n+1} \right\}$$

# Player-dependent rates

## Leslie and Collins (2003)

Revert to stochastic fictitious play, two player games:

$$\sigma^1_{n+1} = \sigma^1_n + \alpha_{n+1}\left\{\beta^1(\sigma^2_n) - \sigma^1_n + M^1_{n+1}\right\}$$
$$\sigma^2_{n+1} = \sigma^2_n + \gamma_{n+1}\left\{\beta^2(\sigma^1_n) - \sigma^2_n + M^2_{n+1}\right\}$$

with $\alpha_n/\gamma_n \to 0$

# Player-dependent rates

## Leslie and Collins (2003)



$$\sigma^1_{n+1} = \sigma^1_n + \alpha_{n+1} \left\{ \beta^1(\sigma^2_n) - \sigma^1_n + M^1_{n+1} \right\}$$
$$\sigma^2_{n+1} = \sigma^2_n + \gamma_{n+1} \left\{ \beta^2(\sigma^1_n) - \sigma^2_n + M^2_{n+1} \right\}$$

**Fast timescale: fix $\sigma^1$**

$$\dot{\sigma}^2 = \beta^2(\sigma^1) - \sigma^2$$
$$\sigma^2 \to \beta^2(\sigma^1)$$

# Player-dependent rates

## Leslie and Collins (2003)

Mathematics & Statistics | Lancaster University

$$\sigma^1_{n+1} = \sigma^1_n + \alpha_{n+1} \left\{ \beta^1(\sigma^2_n) - \sigma^1_n + M^1_{n+1} \right\}$$
$$\sigma^2_{n+1} = \sigma^2_n + \gamma_{n+1} \left\{ \beta^2(\sigma^1_n) - \sigma^2_n + M^2_{n+1} \right\}$$

**Slow timescale: analyse as if $\sigma^2_n = \beta^2(\sigma^1_n)$**

$$\dot{\sigma}^1 = \beta^1(\beta^2(\sigma^1)) - \sigma^1$$

- This ODE has a globally attracting fixed point for zero-sum games, potential games and Shapley's game
- The ODE falls outside Hart and Mas-Colell's impossibility framework
- I have yet to find a game in which it does not converge

# Player-dependent rates

## Leslie and Collins (2003)



$$\sigma^1_{n+1} = \sigma^1_n + \alpha^1_{n+1} \left\{ \beta^1(\sigma^{-1}_n) - \sigma^1_n + M^1_{n+1} \right\}$$
$$\sigma^2_{n+1} = \sigma^2_n + \alpha^2_{n+1} \left\{ \beta^2(\sigma^{-2}_n) - \sigma^2_n + M^2_{n+1} \right\}$$
$$\sigma^3_{n+1} = \sigma^3_n + \alpha^3_{n+1} \left\{ \beta^3(\sigma^{-3}_n) - \sigma^3_n + M^3_{n+1} \right\}$$
$$\sigma^4_{n+1} = \sigma^4_n + \alpha^4_{n+1} \left\{ \beta^4(\sigma^{-4}_n) - \sigma^4_n + M^4_{n+1} \right\}$$

$\vdots$ with $\alpha^i_n / \alpha^{i+1}_n \to 0$

### Theorem-ish

If the fast strategies $\sigma^{>i}$ converge to a unique $\beta^{>i}(\sigma^{\leq i})$ for fixed $\sigma^{\leq i}$, for each $i$, then the system converges iff $\dot{\sigma}^1 = \beta^1(\beta^{>1}(\sigma^1)) - \sigma^1$ converges

# Stochastic games

## Setup



Stochastic game framework (Shapley 1953):

- Finite set of players $i \in \{1, \ldots, N\}$
- Finite set of states $s \in S$
- Finite set of actions $A^i(s)$ for each player $i$ in each state $s$
- Transitions $P_{s,s'}(a)$ and rewards $r^i(s, a)$ for $a = (a^1, \ldots, a^N)$
- Players attempt to maximise cumulative discounted reward

**Key concept: auxiliary games**

At each state $s$, all players choose actions, receive reward and move to next state. Next state has ‘continuation payoffs’ $V^i(s')$. Auxiliary game at $s$, with continuation payoffs $V$ has payoff matrix

$$q^i_{s,V}(a) = r^i(s, a) + \delta \sum_{s'} P_{ss'}(a) V^i(s')$$

# Learning in stochastic games

## Introduction

“Normal-forming” (Stochastic game strategy $\leftrightarrow$ Normal form action): Not v interesting! Finding a best response is solving an MDP. Mixed strategies are weird, except perhaps in evolutionary interpretation.

Per-state fictitious play: This can work (Sayin, Parise, Ozdaglar, SICON 2022 building on Leslie, Perkins, Xu, JET 2020). But can we do radically-uncoupled learning?

Simple $Q$-learning: Many hint this is solved. It is not!

# Learning in stochastic games

## Key idea

Challenge: There are many moving parts. State values we are yet to receive are affected by current strategies

Solution (ish): Fixing the “continuation payoffs” and learning in just the “auxiliary games” makes things much easier

Finishing off: If the auxiliary games are all played ‘at’ equilibrium, then the state values will converge

# Learning in stochastic games



## Reinforcement learning

$$V_n^i(s) = \max_{a^i} Q_n^i(s, a^i)$$

$$Q_{n+1}^i(s, a^i) = Q_n^i(s, a^i) + \gamma_n \mathbb{I}_{\{(s_n, a_n^i) = (s, a)\}} \left\{ r_n^i + \delta V_n^i(s_{n+1}) - Q_n^i(s, a) \right\}$$

# Learning in stochastic games

## Reinforcement learning

Mathematics & Statistics | Lancaster University

$$V_n^i(s) = \sum_b \pi_n^i(s,b) Q_n^i(s,b)$$

$$Q_{n+1}^i(s,a^i) = Q_n^i(s,a^i) + \frac{\gamma_n}{\pi_n^i(s,a)} \mathbb{I}_{\{(s_n,a_n^i)=(s,a)\}} \left\{ r_n^i + \delta V_n^i(s_{n+1}) - Q_n^i(s,a) \right\}$$

where $\pi_n^i(s,a) \propto \exp(Q_n^i(s,a)/\tau_n)$

# Learning in stochastic games



## Reinforcement learning

$$V^i_{n+1}(s) = V^i_n(s) + \alpha_n \mathbb{I}_{\{s_n = s\}} \left\{ \sum_b \pi^i_n(s,b) Q^i_n(s,b) - V^i_n(s) \right\}$$

$$Q^i_{n+1}(s,a^i) = Q^i_n(s,a^i) + \frac{\gamma_n}{\pi^i_n(s,a)} \mathbb{I}_{\{(s_n,a^i_n)=(s,a)\}} \left\{ r^i_n + \delta V^i_n(s_{n+1}) - Q^i_n(s,a) \right\}$$

where $\pi^i_n(s,a) \propto \exp(Q^i_n(s,a)/\tau_n)$

and $\alpha_n/\gamma_n \to 0$

# Learning in stochastic games

## Decoupling step

Two-timescale approach decouples the states:

$$\mathbb{E}\begin{bmatrix} Q^i_{n+1}(s,a) - Q^i_n(s,a) \\ V^i_{n+1}(s) - V^i_n(s) \\ \tau_{n+1} - \tau_n \end{bmatrix} = \alpha \begin{bmatrix} q^i_n(s,(a,\pi^{-i}_n(s))) - Q^i_n(s,a) \\ 0 \\ 0 \end{bmatrix} + e_{n+1}$$

where $q^i_n(s,\boldsymbol{a}) = r^i(s,\boldsymbol{a}) + \delta \sum_{s'} P_{ss'}(\boldsymbol{a}) V^i_n(s')$.

This fast timescale corresponds to considering “individual $Q$-learning” (Leslie and Collins 2005) in an arbitrary fixed auxiliary game with payoffs $q^i(s,\cdot)$.

# Learning in stochastic games



## New Lyapunov function (fast timescale)

- $\dot{Q}^i(a) = q^i(a, \pi^{-i}) - Q^i(a)$ with $\pi^i(a) \propto \exp(Q^i(a)/\tau)$
- Introduce auxiliary vars $\sigma^i$ defined by $\dot{\sigma}^i = \pi^i - \sigma^i$.
- New Lyapunov function:

$$L(Q^1, Q^2, \sigma^1, \sigma^2) = \left[ \sum_{i=1,2} \left\{ \pi^i \cdot Q^i + \tau v^i(\pi^i) \right\} - \lambda \zeta \right]_+ + \sum_{i=1,2} \| Q^i - q^i(\cdot, \sigma^{-i}) \|^2$$

where $\lambda \in (1, \gamma^{-1})$ and $\zeta = \| q^1 + (q^2)^T \|_{\max} + \tau \log(|A^1||A^2|)$.

# Learning in stochastic games



## New Lyapunov function (fast timescale)

$$L(Q^1, Q^2, \sigma^1, \sigma^2) = \left[ \sum_{i=1,2} \left\{ \pi^i \cdot Q^i + \tau v^i(\pi^i) \right\} - \lambda\zeta \right]_+ + \sum_{i=1,2} \| Q^i - q^i(\cdot, \sigma^{-i}) \|^2$$

- Start with standard Lyapunov function for smooth BR learning
- $\lambda\zeta$ term means we only make this small, not 0
- Second summation shows $Q$ are asymptotically belief based

# Learning in stochastic games

## New Lyapunov function (fast timescale)

$$L(Q^1, Q^2, \sigma^1, \sigma^2) = \left[\sum_{i=1,2}\left\{\pi^i \cdot Q^i + \tau v^i(\pi^i)\right\} - \lambda\zeta\right]_+ + \sum_{i=1,2}\|Q^i - q^i(\cdot, \sigma^{-i})\|^2$$

So, there exists $\epsilon_n \to 0$, such that

$$\sum_{i=1,2}\left\{\pi_n^i \cdot Q_n^i + \tau_n v^i(\pi_n^i)\right\} \leq \lambda\left\{\|q_n^1 + (q_n^2)^T\|_{\max} + \tau_n \log(|A^1||A^2|)\right\} + \epsilon_n$$

# Stochastic games

## Sketch proof

- For fixed continuation payoffs $V$, we have shown convergence (admittedly to a set)
- The two-timescales theory allows us to analyse $V$ as if the $Q$ values are always in this set
- Convergence follows in two-player zero-sum games

Refs:


- Leslie, Perkins, Xu, JET 2020
- Sayin, Parise, Ozdaglar, SICON 2022
- Sayin, Zhang, Leslie, Basar, Ozduglar, NeurIPS 2020

# Gradient noise reduction



In continuous games, we use a very different notation:

- Actions are $x = (x^1, \dots, x^N)$
- Payoffs are $u^i : \mathcal{X} \to \mathbb{R}$
- Individual payoff gradients are $v^i(x^i, x^{-i}) = \nabla_{x^i} u^i(x)$
- Pseudogradient is $v(x) = (v^1(x), \dots, v^N(x))$

Often players need to estimate $v^i(x)$. Estimates may have very high variance.

# Gradient noise reduction



Averaging several observations $v^i(x) + \epsilon_n^i$ would reduce the variance
So....

$$
\begin{aligned}
x_{n+1}^i &= x_n^i + \alpha_{n+1} V_n^i \\
V_{n+1}^i &= V_n^i + \gamma_{n+1} \left\{ v_n^i(x_n^i) + \epsilon_n^i - V_n^i \right\}
\end{aligned}
$$

For fixed $x$, the $V_n^i$ converge to $v_n^i$. Then the slow equation follows the gradient nicely.

# Gradient noise reduction



Rate of convergence is the elephant in the two-timescales room!

My student Miles Elvidge is working on some really cool ideas along these lines

Essentially, work out what the fast timescale analysis tells you, then plug that into a finite time analysis on the slow timescale

# Summary



- Whenever an inner loop would be useful, think about using two timescales
- Has been deployed in:
  - actor–critic learning
  - player-dependent learning rates
  - stochastic games
  - gradient smoothing
- Convergence rates are hard, but very recent work is getting there

Mathematics & Statistics | Lancaster University

February 2025
David S. Leslie