Learning Nash Equilibrium for General-Sum Markov Games from Batch Data

Julien Pérolat[(a)], Florian Strub[(a)], Bilal Piot[(a)], Olivier Pietquin[(a,b,c)]

[(a)] Univ. Lille, CNRS, Centrale Lille, Inria UMR 9189 - CRIStAL, F-59000 Lille, France
[(b)] Institut Universitaire de France (IUF), France, [(c)] now at DeepMind.

August 7, 2016

Motivation

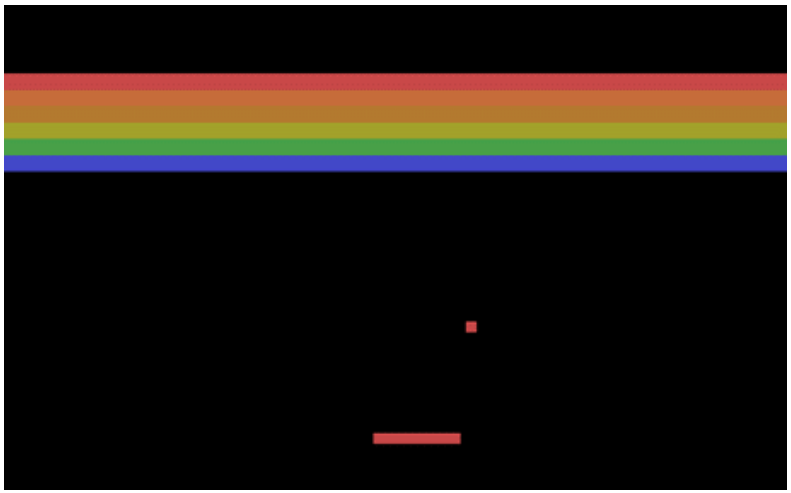


Figure : Breakout

Markov Decision Processes (MDPs):

Finding an optimal policy,

One agent maximizing his expected sum of rewards,

Motivation

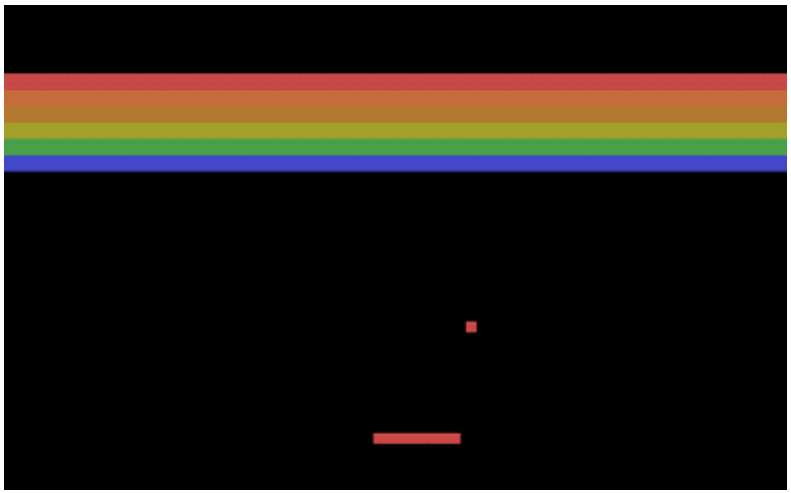


Figure : Breakout

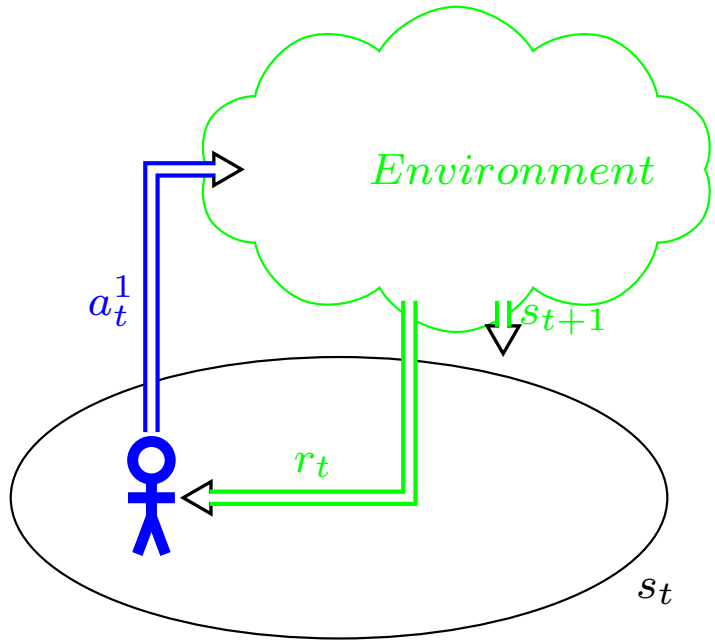


Figure : Markov Decision Process (MDP)

Markov Decision Processes (MDPs):

Finding an optimal policy,

One agent maximizing his expected sum of rewards,

Find the optimal Q-function $Q(s, a)$,

Act greedily according to the Q-function.

Motivation

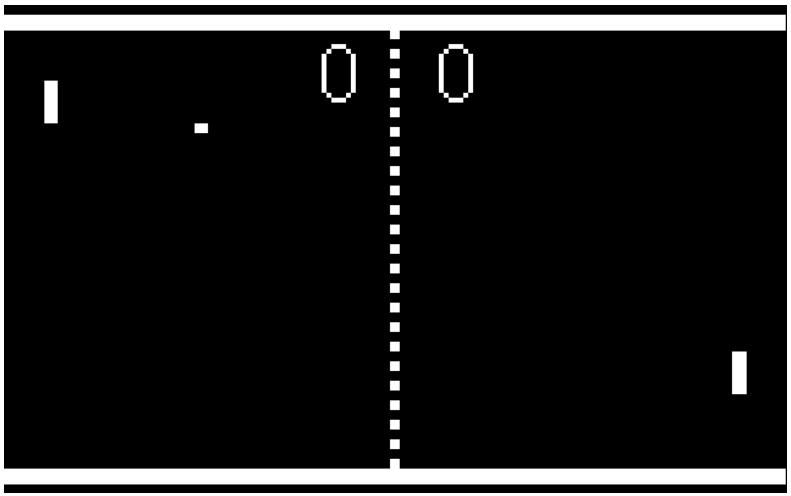


Figure : Pong

Motivation

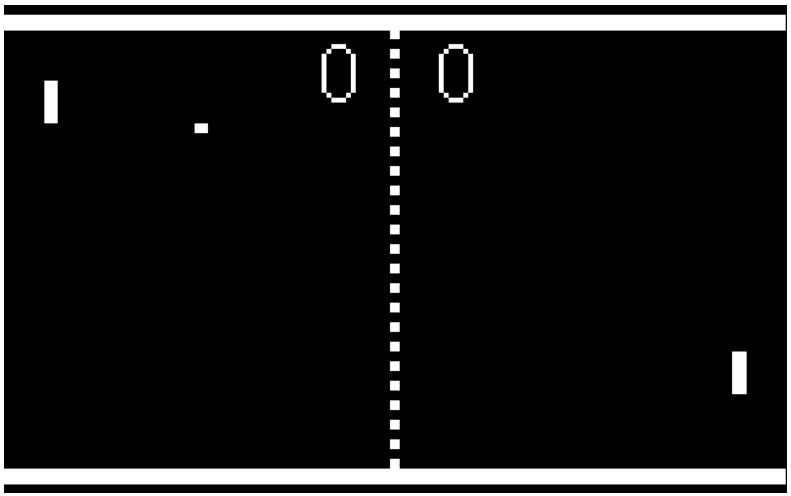


Figure : Pong

Zero-Sum Two-Player Markov Games:

Both player receive the same reward signal,

Finding a minimax policy.

Motivation

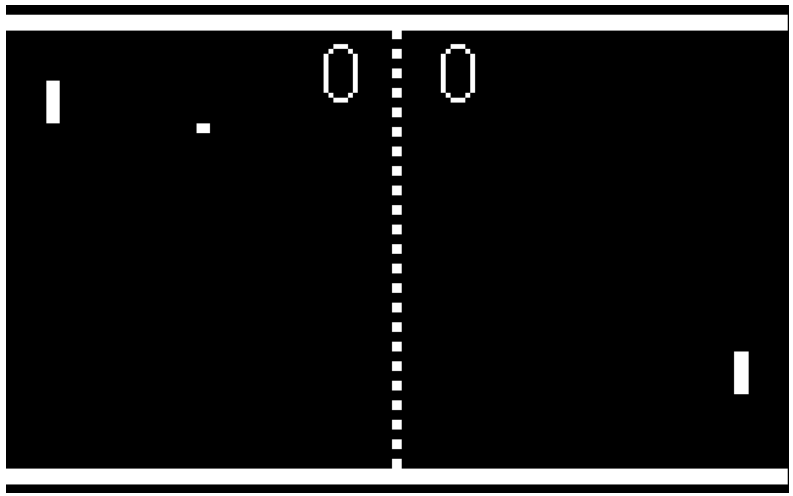


Figure : Pong

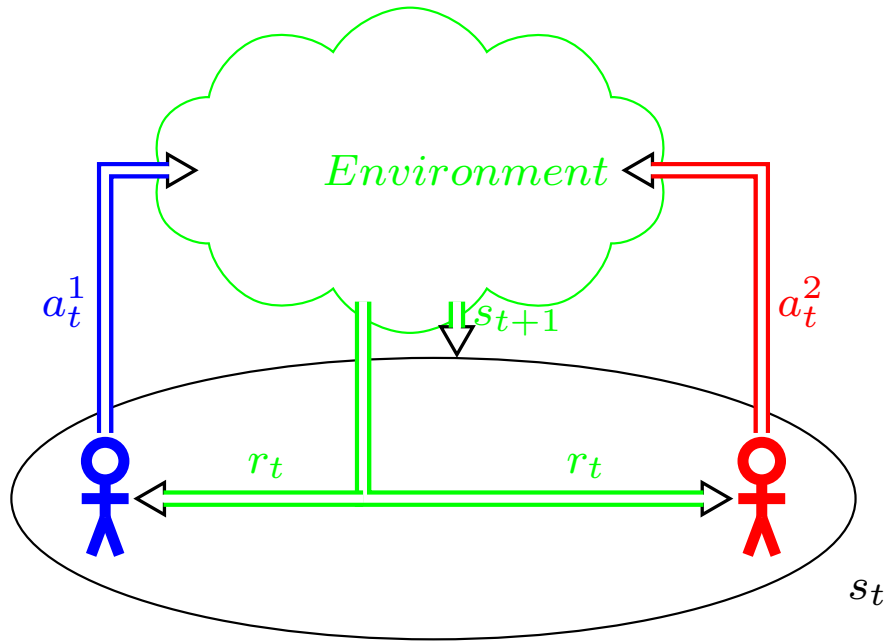


Figure : Zero-Sum Two-Player Markov Game

Zero-Sum Two-Player Markov Games:

Both player receive the same reward signal,

Finding a minimax policy.

Find an optimal Q-function ($Q(s, a^1, a^2)$),

Act greedily according to the Q-function.

Motivation

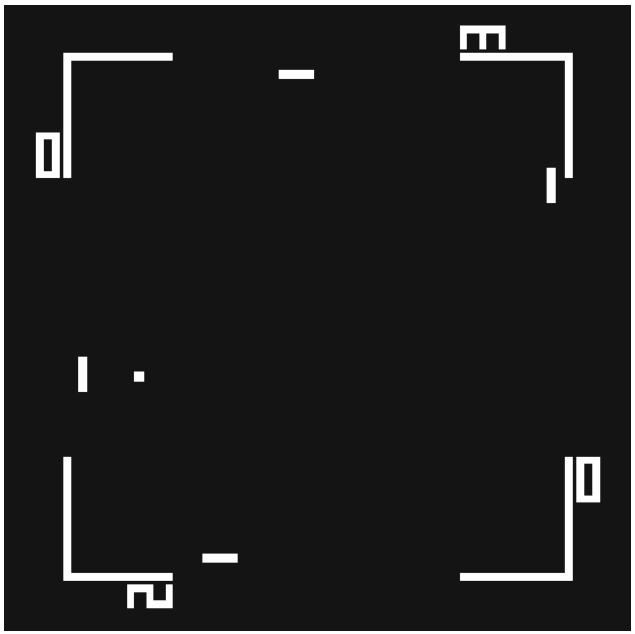


Figure : Pong

Motivation

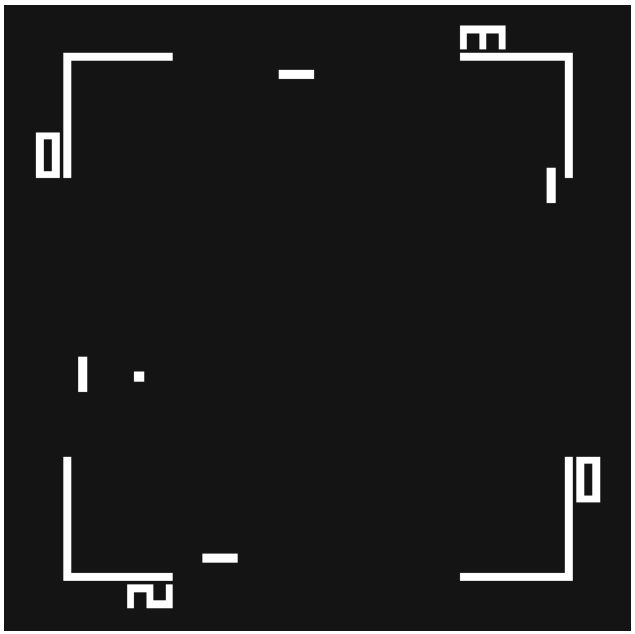


Figure : Pong

N-Player General-sum Markov Games:

Each player receives his own reward signal,

Finding a Nash equilibrium.

Motivation

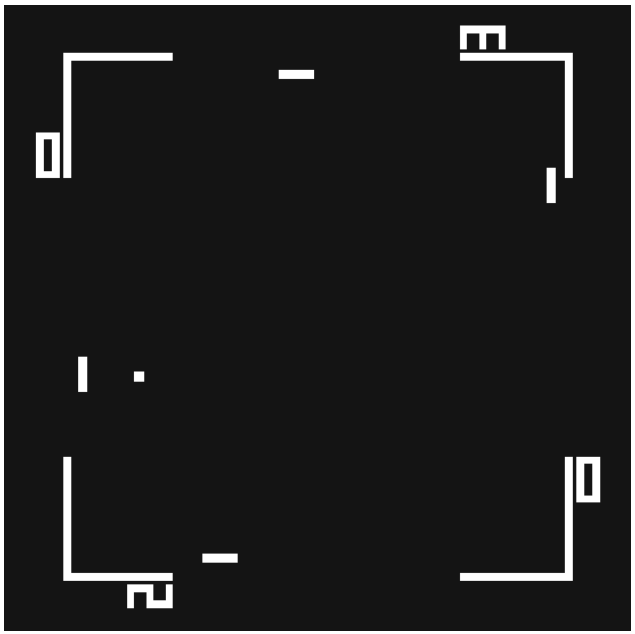


Figure : Pong

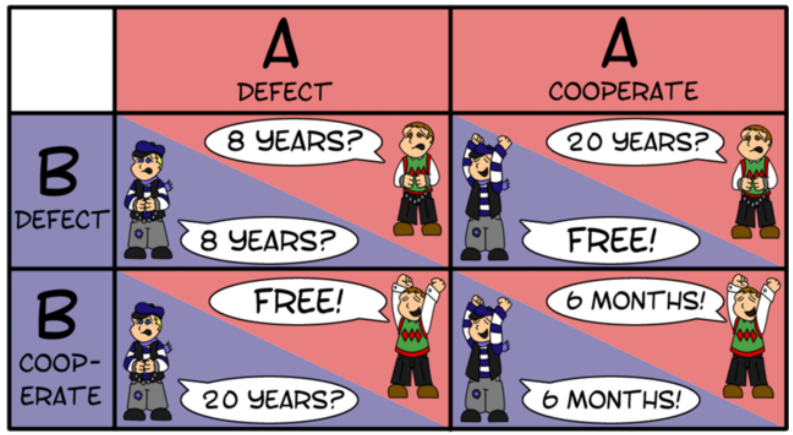


Figure : Nash Equilibrium: no player would benefit from modifying their current strategy.

N-Player General-sum Markov Games:

Each player receives his own reward signal,

Finding a Nash equilibrium.

Motivation

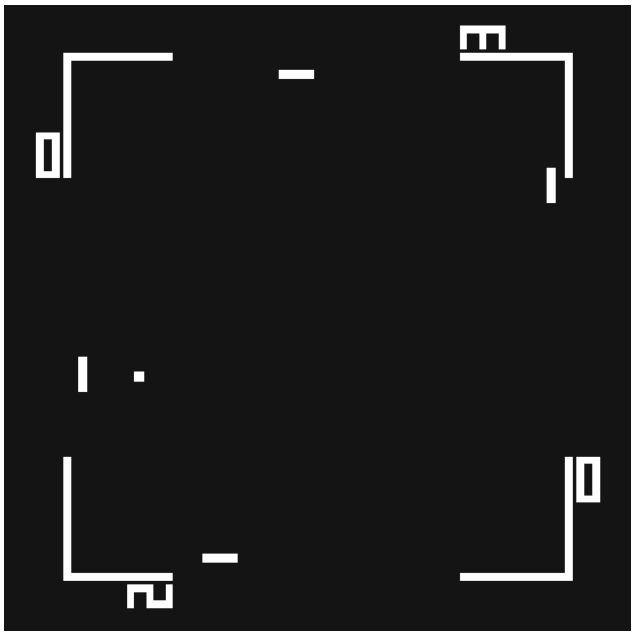


Figure : Pong

Can we only work on Q-functions?

N-Player General-sum Markov Games:

Each player receives his own reward signal,

Finding a Nash equilibrium.

Motivation

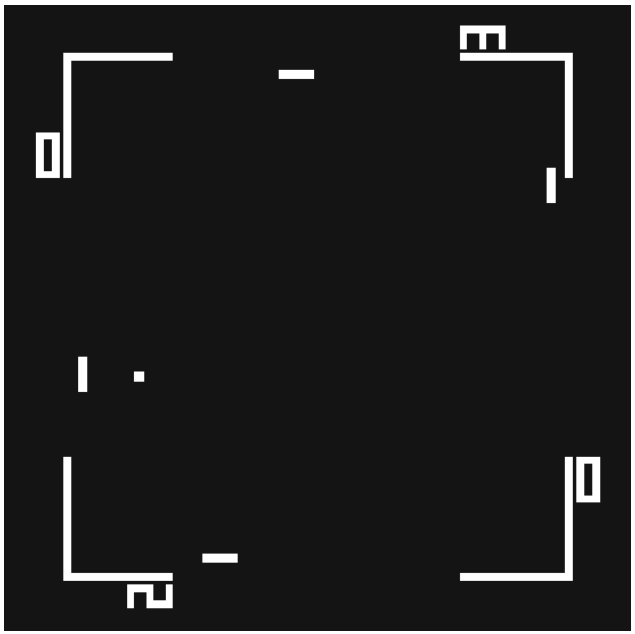


Figure : Pong

Can we only work on Q-functions?
No!

N-Player General-sum Markov Games:

Each player receives his own reward signal,

Finding a Nash equilibrium.

Motivation

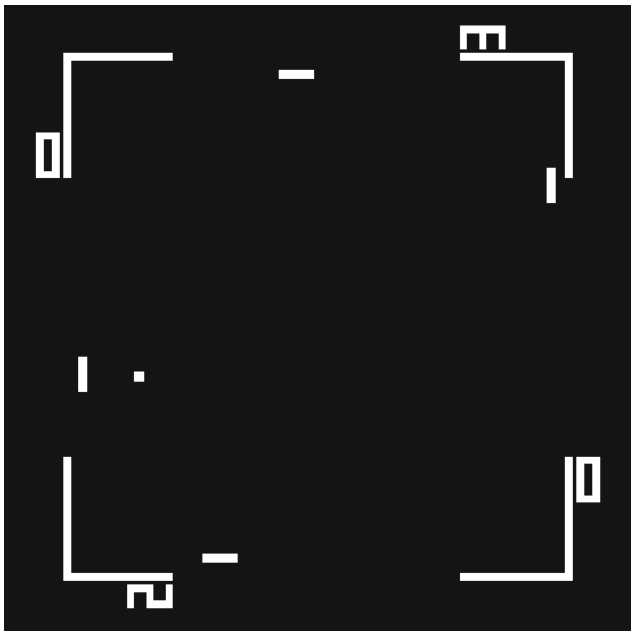


Figure : Pong

Can we only work on Q-functions?
No!

Problem:
Batch Reinforcement Learning algorithms are only based on the Q-Function.

N-Player General-sum Markov Games:

Each player receives his own reward signal,

Finding a Nash equilibrium.

Previous Work & Contributions

Our Goal:
Learning an ϵ**-Nash equilibrium** in N**-player general-sum Markov Games** from **historical data** with **function approximation**.

Previous work:
Learning from historical data:

Fitted-Q iteration on MDPs (Ernst & al [2005], Riedmiller [2005]),

LSPI (Lagoudakis & Parr [2002]),

Approximate dynamic programming on zero-sum two-player MGs (Pérolat & al [2015,2016]).

Learning in general sum Markov Games:

Stochastic approximation approaches (Prasad & al [2015]). Limited to the online case or to the model based case.

Contributions :

- Definition of a new (weak) ϵ-Nash equilibrium,
- Reduce the problem of learning a Nash equilibrium to the minimization of a surrogate loss (a sum of Bellman Residuals),
- Empirical evaluation of the method using Neural network.

Framework

A Markov Game is specified by:

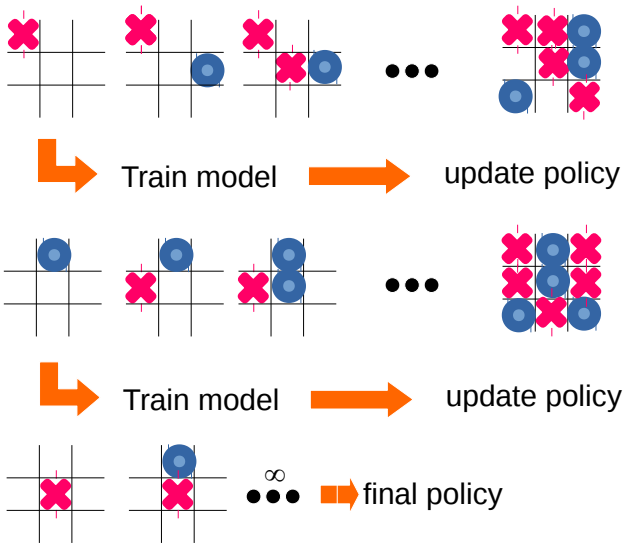
- a number of player N,
- a state space S,
- an action space per player $A^1, \ldots, A^N$,
- a transition kernel $p(s'|s, a^1, \ldots, a^N) = p(s'|s, a^i, a^{-i}) = p(s'|s, \boldsymbol{a})$,
- a reward signal per player $r^i(s, a^1, \ldots, a^N) = r^i(s, a^i, a^{-i}) = r^i(s, \boldsymbol{a})$,
- a discount factor γ.

Goal:

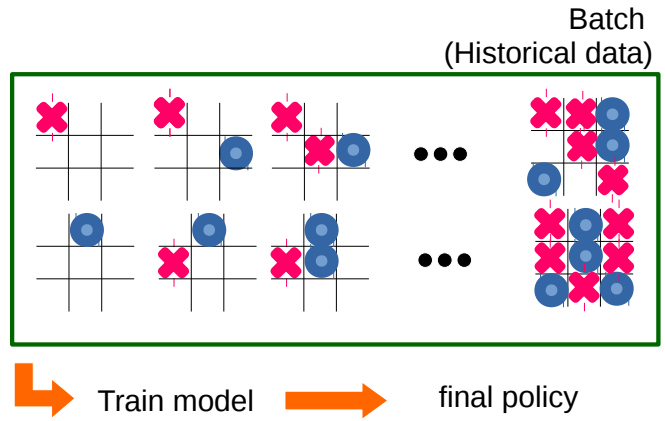
- Find a strategy $\pi^i(a^i|s)$ a strategy for each player $\boldsymbol{\pi} = (\pi^i, \pi^{-i}) = (\pi^1, \ldots, \pi^N)$,

The Batch Scenario

Online scenario:



Batch scenario:



Historical data required :

$$((s_j, a_j^1, \ldots, a_j^N), r_j^1, \ldots, r_j^N, s'_j)_{j=1,\ldots,k}$$

state-actions tuple $(s_j, a_j^1, \ldots, a_j^N)$,

a reward per player $r_j^1, ..., r_j^N$,

the next state $s'_j \sim p(.|s_j, \boldsymbol{a}_j)$.

Q-functions

- Q-functions :

$$Q^i_{\boldsymbol{\pi}}(s, \mathbf{a}) = E\left[\sum_{t=0}^{\infty} \gamma^t r^i_{\boldsymbol{\pi}}(s_t) | s_0 = s, \boldsymbol{a}_0 = \boldsymbol{a},\ s_{t+1} \sim P_{\boldsymbol{\pi}}(.|s_t)\right],$$

where $r_{\boldsymbol{\pi}}(s) = E_{a\sim\boldsymbol{\pi}}[r(s, \boldsymbol{a})]$ and $P_{\boldsymbol{\pi}}(s'|s) = E_{a\sim\boldsymbol{\pi}}[p(s'|s, \boldsymbol{a})]$,

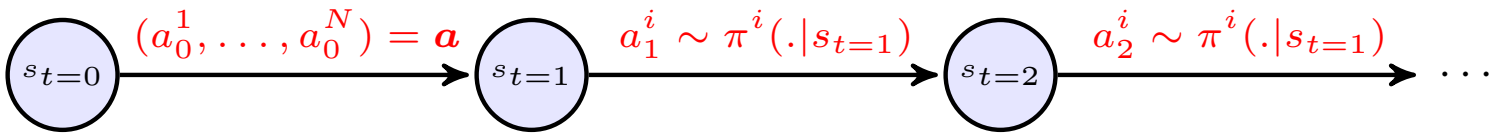


Figure : Q-function in MGs

- Bellman operator :

$$\mathcal{B}^i_{\boldsymbol{\pi}} Q = r^i(s, \mathbf{a}) + \sum_{s' \in S} p(s'|s, \mathbf{a}) E_{\mathbf{b}\sim\boldsymbol{\pi}}[Q(s', \mathbf{b})],$$

- Fixed point equation :

$$\mathcal{B}^i_{\boldsymbol{\pi}} Q^i_{\boldsymbol{\pi}} = Q^i_{\boldsymbol{\pi}}$$

,

Best Response

- Best Response of player i:

$$Q^{*i}_{\pi^{-i}}(s, \mathbf{a}) = \max_{\pi^i} Q^i_{\pi^i, \pi^{-i}}(s, \mathbf{a}),$$

- Bellman operator :

$$\mathcal{B}^{*i}_{\pi^{-i}} Q = r^i(s, \mathbf{a}) + \sum_{s' \in S} p(s'|s, \mathbf{a}) \max_{b^i} \left[E_{\mathbf{b^{-i}} \sim \pi^{-i}} [Q(s', b^i, b^{-i})] \right],$$

- Fixed point equation :

$$\mathcal{B}^{*i}_{\pi^{-i}} Q^{*i}_{\pi^{-i}} = Q^{*i}_{\pi^{-i}}.$$

,

Nash equilibrium, weak ϵ-Nash equilibrium

Nash equilibrium :
A strategy $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ is a Nash equilibrium if $\forall i$:

$$\underbrace{Q^{*i}_{\pi^{-i}}}_{\text{best response}} = \underbrace{Q^i_{\boldsymbol{\pi}}}_{\text{policy}}$$

Player i has no incentive to modify his current strategy.

A strategy $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ is a Nash equilibrium if there exists $Q^1, \dots, Q^N$ such as $\forall i$:

$$\underbrace{\mathcal{B}^i_{\boldsymbol{\pi}} Q^i = Q^i}_{Q^i = Q^i_{\boldsymbol{\pi}}} \text{ and } \underbrace{\mathcal{B}^{*i}_{\pi^{-i}} Q^i = Q^i}_{Q^i = Q^{*i}_{\pi^{-i}}}$$

Bellman Residual Approach

Weak ϵ-Nash equilibrium :
A strategy $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^N)$ is a weak ϵ-Nash equilibrium if $\forall i$:

$$\left\| \left\| Q^{*i}_{\pi^{-i}} - Q^i_{\boldsymbol{\pi}} \right\|_{\mu(s,\boldsymbol{a}),2} \right\|_{\rho(i),2} \leq \epsilon$$

Player i has no more than an ϵ incentive to modify his current strategy.

Bellman Residual Approach

Weak ϵ-Nash equilibrium :
A strategy $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ is a weak ϵ-Nash equilibrium if $\forall i$:

$$\left\| \left\| Q^{*i}_{\pi^{-i}} - Q^i_{\boldsymbol{\pi}} \right\|_{\mu(s,\boldsymbol{a}),2} \right\|_{\rho(i),2} \leq \epsilon$$

Player i has no more than an ϵ incentive to modify his current strategy.

The idea :
What if we can find $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ and $Q^1, \dots, Q^N$ such as $\forall i$:

$$\underbrace{\mathcal{B}^i_{\boldsymbol{\pi}} Q^i \simeq Q^i}_{Q^i \simeq Q^i_{\boldsymbol{\pi}}} \text{ and } \underbrace{\mathcal{B}^{*i}_{\pi^{-i}} Q^i \simeq Q^i}_{Q^i \simeq Q^{*i}_{\boldsymbol{\pi}}}$$

Bellman Residual Approach

Weak ϵ-Nash equilibrium :
A strategy $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ is a weak ϵ-Nash equilibrium if $\forall i$:

$$\left\| \left\| Q^{*i}_{\pi^{-i}} - Q^i_{\boldsymbol{\pi}} \right\|_{\mu(s,\boldsymbol{a}),2} \right\|_{\rho(i),2} \leq \epsilon$$

Player i has no more than an ϵ incentive to modify his current strategy.

The idea :
What if we can find $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ and $Q^1, \dots, Q^N$ such as $\forall i$:

$$\underbrace{\mathcal{B}^i_{\boldsymbol{\pi}} Q^i \simeq Q^i}_{Q^i \simeq Q^i_{\boldsymbol{\pi}}} \text{ and } \underbrace{\mathcal{B}^{*i}_{\pi^{-i}} Q^i \simeq Q^i}_{Q^i \simeq Q^{*i}_{\boldsymbol{\pi}}}$$

$$\forall i,\ \left\| \mathcal{B}^i_{\boldsymbol{\pi}} Q^i - Q^i \right\|_{\nu,2} \simeq 0 \text{ and } \left\| \mathcal{B}^{*i}_{\pi^{-i}} Q^i - Q^i \right\|_{\nu,2} \simeq 0$$

Bellman Residual Approach

Weak ϵ-Nash equilibrium :
A strategy $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ is a weak ϵ-Nash equilibrium if $\forall i$:

$$\left\| \left\| Q^{*i}_{\pi^{-i}} - Q^i_{\boldsymbol{\pi}} \right\|_{\mu(s,\boldsymbol{a}),2} \right\|_{\rho(i),2} \leq \epsilon$$

Player i has no more than an ϵ incentive to modify his current strategy.

The idea :
What if we can find $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ and $Q^1, \dots, Q^N$ such as $\forall i$:

$$\underbrace{\mathcal{B}^i_{\boldsymbol{\pi}} Q^i \simeq Q^i}_{Q^i \simeq Q^i_{\boldsymbol{\pi}}} \text{ and } \underbrace{\mathcal{B}^{*i}_{\pi^{-i}} Q^i \simeq Q^i}_{Q^i \simeq Q^{*i}_{\boldsymbol{\pi}}}$$

$$\forall i,\ \left\| \mathcal{B}^i_{\boldsymbol{\pi}} Q^i - Q^i \right\|_{\nu,2} \simeq 0 \text{ and } \left\| \mathcal{B}^{*i}_{\pi^{-i}} Q^i - Q^i \right\|_{\nu,2} \simeq 0$$

$$\forall i,\ \left\| \mathcal{B}^i_{\boldsymbol{\pi}} Q^i - Q^i \right\|_{\nu,2} + \left\| \mathcal{B}^{*i}_{\pi^{-i}} Q^i - Q^i \right\|_{\nu,2} \simeq 0$$

Bellman Residual Approach

Weak ϵ-Nash equilibrium :
A strategy $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ is a weak ϵ-Nash equilibrium if $\forall i$:

$$\left\| \left\| Q^{*i}_{\pi^{-i}} - Q^i_{\boldsymbol{\pi}} \right\|_{\mu(s,\boldsymbol{a}),2} \right\|_{\rho(i),2} \leq \epsilon$$

Player i has no more than an ϵ incentive to modify his current strategy.

The idea :
What if we can find $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N)$ and $Q^1, \dots, Q^N$ such as $\forall i$:

$$\underbrace{\mathcal{B}^i_{\boldsymbol{\pi}} Q^i \simeq Q^i}_{Q^i \simeq Q^i_{\boldsymbol{\pi}}} \text{ and } \underbrace{\mathcal{B}^{*i}_{\pi^{-i}} Q^i \simeq Q^i}_{Q^i \simeq Q^{*i}_{\boldsymbol{\pi}}}$$

$$\forall i,\ \left\| \mathcal{B}^i_{\boldsymbol{\pi}} Q^i - Q^i \right\|_{\nu,2} \simeq 0 \text{ and } \left\| \mathcal{B}^{*i}_{\pi^{-i}} Q^i - Q^i \right\|_{\nu,2} \simeq 0$$

$$\forall i,\ \left\| \mathcal{B}^i_{\boldsymbol{\pi}} Q^i - Q^i \right\|_{\nu,2} + \left\| \mathcal{B}^{*i}_{\pi^{-i}} Q^i - Q^i \right\|_{\nu,2} \simeq 0$$

$$\sum_{i=1}^{N} \left(\left\| \mathcal{B}^i_{\boldsymbol{\pi}} Q^i - Q^i \right\|_{\nu,2} + \left\| \mathcal{B}^{*i}_{\pi^{-i}} Q^i - Q^i \right\|_{\nu,2} \right) \simeq 0$$

What can we guarantee?

$$\underbrace{\left\| \left\| Q_{\boldsymbol{\pi}}^{i} - Q_{\boldsymbol{\pi}}^{*i} \right\|_{\mu(s,\boldsymbol{a}),2} \right\|_{i,2}}_{\text{Weak } \epsilon\text{-Nash equilibrium}} \leq \frac{C(\mu,\nu)}{1-\gamma} \underbrace{\left[\sum_{i=1}^{N} \left(\left\| \mathcal{B}_{\boldsymbol{\pi}}^{i} Q^{i} - Q^{i} \right\|_{\nu,2}^{2} + \left\| \mathcal{B}_{\pi^{-i}}^{*i} Q^{i} - Q^{i} \right\|_{\nu,2}^{2} \right) \right]^{\frac{1}{2}}}_{\text{Sum of Bellman Residuals}},$$

What can we guarantee?

$$\underbrace{\left\| \left\| Q^i_{\boldsymbol{\pi}} - Q^{*i}_{\boldsymbol{\pi}} \right\|_{\mu(s,\boldsymbol{a}),2} \right\|_{i,2}}_{\text{Weak } \epsilon\text{-Nash equilibrium}} \leq \frac{C(\mu,\nu)}{1-\gamma} \underbrace{\left[\sum_{i=1}^{N} \left(\left\| \mathcal{B}^i_{\boldsymbol{\pi}} Q^i - Q^i \right\|^2_{\nu,2} + \left\| \mathcal{B}^{*i}_{\pi^{-i}} Q^i - Q^i \right\|^2_{\nu,2} \right) \right]^{\frac{1}{2}}}_{\text{Sum of Bellman Residuals}},$$

The approach to learn from batch data:

Minimize an **empirical estimate** of the **sum of Bellman Residuals** with parametrized strategies $\pi^i_{\theta^i}$ and parametrized Q-function $Q^i_{\theta'^i}$ (meaning we use **function approximation**).

Estimators :

Estimation of the sum of Bellman Residuals :

$$\sum_{i=1}^{N} \left(\left\| \mathcal{B}^i_{\boldsymbol{\pi}} Q^i - Q^i \right\|^2_{\nu,2} + \left\| \mathcal{B}^{*i}_{\pi^{-i}} Q^i - Q^i \right\|^2_{\nu,2} \right)$$

With our batch dataset,

$$((s_j, a^1_j, \ldots, a^N_j), r^1_j, \ldots, r^N_j, s'_j)_{j=1,\ldots,k}$$

For each tuple $((s_j, \boldsymbol{a}_j), r^1_j, \ldots, r^N_j, s'_j)$**:**

$$\left| \mathcal{B}^i_{\boldsymbol{\pi}} Q^i(s_j, \mathbf{a}_j) - Q^i(s_j, \mathbf{a}_j) \right|^2 = \left| \underbrace{r^i_j}_{\text{target}} + \underbrace{\gamma E_{\mathbf{b} \sim \boldsymbol{\pi}}[Q^i(s'_j, \mathbf{b})] - Q^i(s_j, \mathbf{a}_j)}_{\text{estimator}} \right|^2$$

$$\left| \mathcal{B}^{*i}_{\pi^{-i}} Q^i(s_j, \mathbf{a}_j) - Q^i(s_j, \mathbf{a}_j) \right|^2 = \left| \underbrace{r^i_j}_{\text{target}} + \underbrace{\gamma \max_{b^i} \left[E_{b^{-i} \sim \pi^{-i}}[Q^i(s'_j, b^i, b^{-i})] \right] - Q^i(s_j, \mathbf{a}_j)}_{\text{estimator}} \right|^2$$

Learning Process and Experiment:

Learning Process :
Loss function with a parametric representation of the strategy $\pi^i_{\theta^i}$ and Q-function $Q^i_{\theta'^i}$ is:

$$\sum_{j=1}^{k} \underbrace{\psi_j(\boldsymbol{\theta}, \boldsymbol{\theta'})}_{\text{estimator of the sum of Bellman residuals}}$$

To learn parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta'}$, we use stochastic gradient descent.

Type of parametrization :

- linear,
- neural network.

Learning Process and Experiment:

Experiment on randomly generated turn-based MGs :

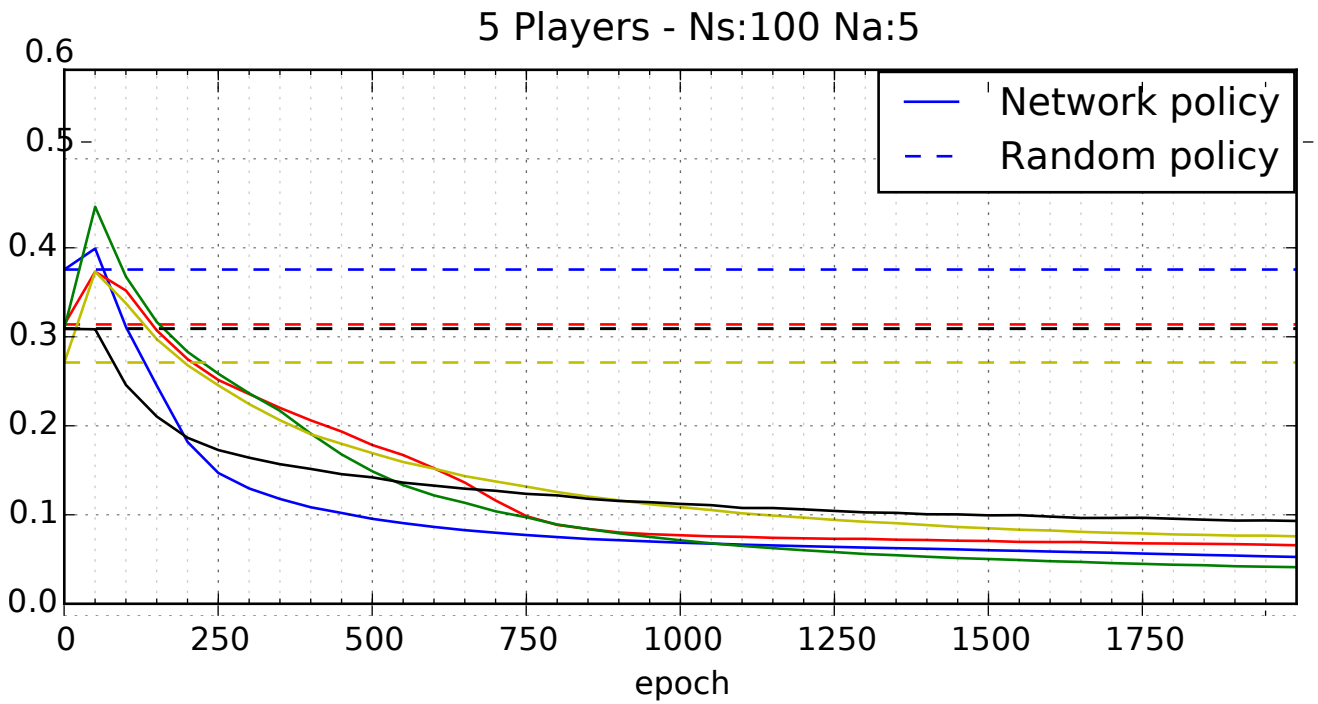


Figure : Error value of policy vs value of the best response for each player.

Conclusion

Contribution

New definition of an ϵ-Nash equilibrium,

Novel approach to learn Nash equilibrium from batch data based on Bellman residuals,

Empirical evaluation using neural network.

Future work

Extension of the experimental part on simultaneous games or multi-player large scale games,

Additional optimization methods could be studied,

We could study other class of function approximation such as trees.

`http://arxiv.org/abs/1606.08718`