DeepMind

From Poincaré Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization

Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, Karl Tuyls
Join work between DeepMind and SUTD.

Game theory setup:

Basic Setup:

- Two-player zero-sum Games
- Actions : $a^i \in A,\ a = (a^1, a^2) = (a^i, a^{-i})$
- Policy : $\pi^i \in \Delta A,\ \pi = (\pi^1, \pi^2) = (\pi^i, \pi^{-i})$
- Reward : $r^i(a^1, a^2)$
- Q-function : $Q^i_\pi(a^i) = \mathbb{E}_{a^{-i}\sim\pi^{-i}}[r^i_\pi(a^i, a^{-i})]$
- Value Function : $V^i_\pi = \mathbb{E}_{a\sim\pi}[r^i_\pi(a)] = \mathbb{E}_{a^i\sim\pi^i}[Q^i_\pi(a^i)]$

Nash Equilibrium:

π^* is a **Nash equilibrium** if for all π and for all i we have $V^i_{\pi^i,\pi^{*-i}} - V^i_{\pi^*} \leq 0$

Related Methods to do model free Learning in Games

NFSP:

- Theoretically Founded on Fictitious Play,
- Rely on a best response subroutine,
- Need to get an average policy.

PSRO:

- Theoretically Founded on Double Oracle methods,
- Rely on a best response subroutine,
- Iteration will be as slow as the best response computation and the metagame building.

DeepCFR/DREAM/ARMAC:

- Theoretically Founded on CFR,
- Need to get an average policy.

LOLA:

- Theoretically Founded on Extragradient methods,
- The High variance slows down the convergence.

Learning with Regularization

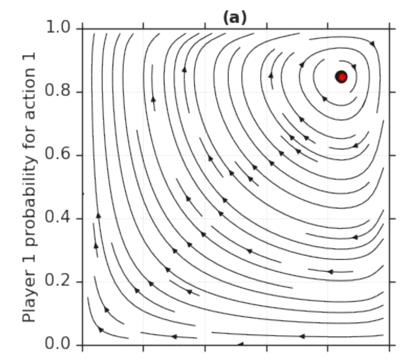
Follow The Regularized Leader:

$$y^i_t(a^i) = \int_0^t Q^i_{\pi_s}(a^i)ds \quad \text{and} \quad \pi^i_t = \text{argmax}_{p\in\Delta A}\Lambda^i(p, y^i_t)$$

With : $\Lambda^i(p, y) = \langle y, p\rangle - \phi_i(p)$ and $\phi_i(p)$ is a regularisation for the policy projection.

In zero-sum two-player games, the following quantity is preserved and the learning trajectory is recurrent:

$$J(y) = \sum_{i=1}^{2}[\phi^*_i(y_i) - \langle y_i, \pi^*_i\rangle]$$

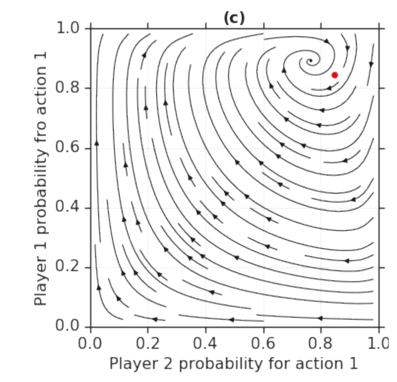


Adding a policy dependent term:

$$r^i_\pi(a) = r^i(a^i, a^{-i}) - \eta\log\frac{\pi^i(a^i)}{\mu^i(a^i)} + \eta\log\frac{\pi^{-i}(a^{-i})}{\mu^{-i}(a^{-i})}$$

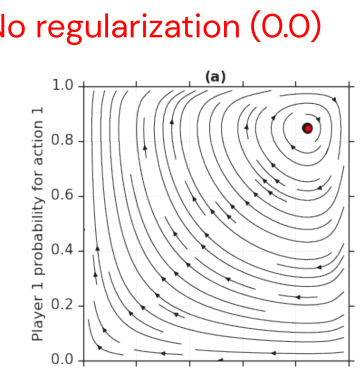
This policy depend term transforms a recurrent learning dynamic to a convergent one:

$$\frac{d}{dt}J(y) = \sum_{i=1}^{2}\underbrace{[V^i_{\pi^i_t,\pi^{*-i}} - V^i_{\pi^*}]}_{\leq 0 \text{ because } \pi^* \text{ is a Nash}} - \eta\sum_{i=1}^{2}KL(\pi^{*i}, \pi^i_t)$$

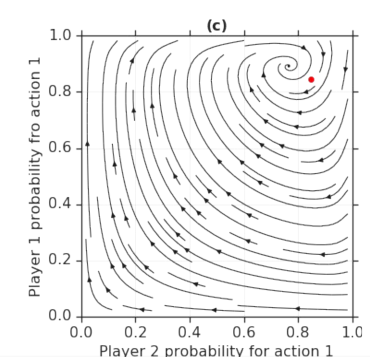


Increasing speed of convergence

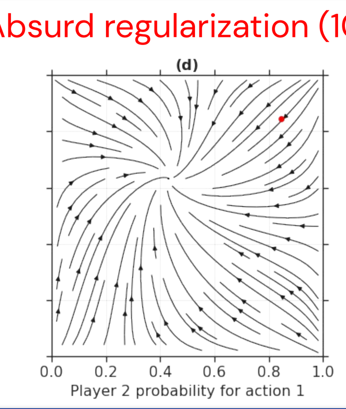
No regularization (0.0)



Small regularization (0.05)

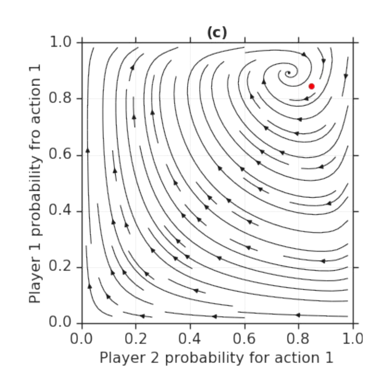


Absurd regularization (10.0)



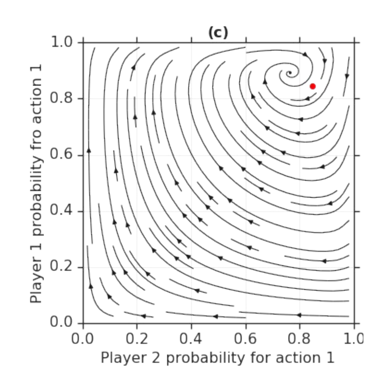
Increasing bias to the solution

Regularization centered around [⅓, ⅓, ⅓].



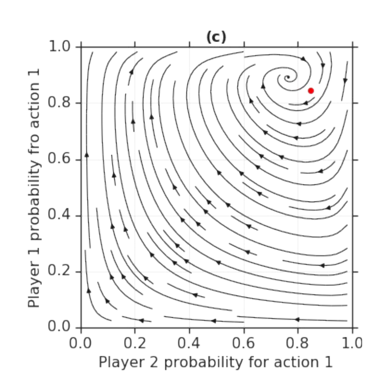
Solution : [0.38, 0.48, 0.12]

Regularization centered around [0.38, 0.48, 0.12].



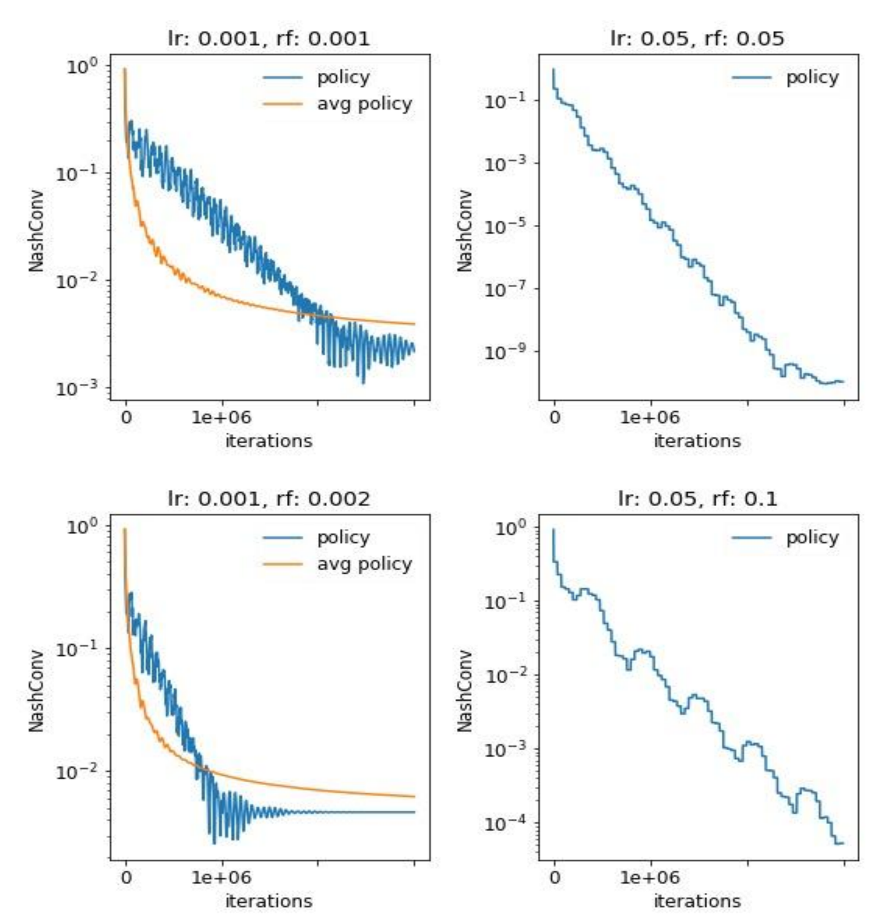
Solution : [0.29, 0.62, 0.07]

Regularization centered around [0.29, 0.62, 0.07].

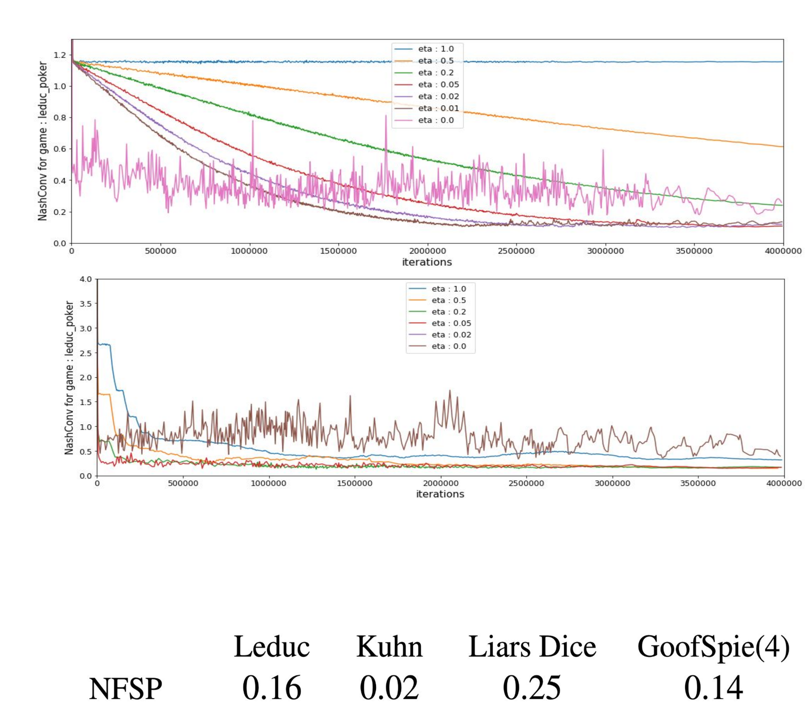


Solution : [0.19, 0.72, 0.07]

Convergence in Sequential Imperfect Information Games (Kuhn Tabular):



Convergence in Sequential Imperfect Information Games (Leduc with Neural Network and a NeuRD loss):



	Leduc	Kuhn	Liars Dice	GoofSpie(4)
NFSP	0.16	0.02	0.25	0.14
Deep CFR	0.23	0.009	**0.19**	0.25
Q-learning	2.44	0.33	0.94	2.0
PSRO	0.17	**0.002**	0.28	0.23
NeuRD	**0.10**	0.02	0.25	**0.22**

NashConv on a benchmark of small games.

References

- Omidshafiei, & al. *Neural replicator dynamics*. arXiv, 2019.
- Heinrich, J. and Silver, D. *Deep reinforcement learning from self-play in imperfect-information games*. arXiv, 2016.
- Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. *Cycles in adversarial regularized learning*. SODA, 2018.
- Lanctot, & al.. A unified game-theoretic approach to multiagent reinforcement learning. NIPS, 2017.

Conclusion:

- Our reward transform is a very simple modification of existing methods (NeuRD),
- Our method is very competitive in Imperfect information Games compared to other methods,
- The analysis covers a large class of general sum games.