

Policy Gradient Methods in Repeated Games

Galit Ashkenazi-Golan

Joint with D. Mergoni & E. Plumb

London School of Economics

July 2024

WLiG, Toulouse

Learning to Collude and the Folk Theorem

Calvano, E., Calzolari, G., Denicolo, V., and Pastorello, S. (2020). *Artificial intelligence, algorithmic pricing, and collusion*. American Economic Review

Assad, S., Clark, R., Ershov, D., and Xu, L. (2024). *Algorithmic pricing and competition: Empirical evidence from the German retail gasoline market*. Journal of Political Economy

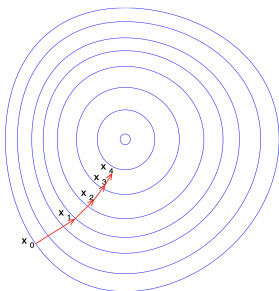
Giannou, A. Lotidis, K., Mertikopoulos, P. and EV Vlatakis-Gkaragkounis (2022). *On the convergence of policy gradient methods to Nash equilibria in general stochastic games*, Advances in Neural Information Processing Systems

RL Example - Gradient Ascent

For some function $f : \Theta \rightarrow \mathbb{R}$, we consider repeatedly updating the parameter θ by:

$$\theta^{\text{new}} = \theta^{\text{old}} + \gamma \nabla f(\theta^{\text{old}})$$

where $\gamma > 0$ is a step size.



The One Shot Game

A game is a triple $G = (N, (A_i)_{i \in N}, (u_i)_{i \in N})$ where:

- N is a finite set of players,
- $(A_i)_{i \in N}$ where A_i is the finite set of actions available to player $i \in N$,
- $(u_i)_{i \in N}$ where $u_i : \prod_{j \in N} \Delta(A_j) \rightarrow \mathbb{R}$ is the payoff function of player $i \in N$.

A strategy (or policy) for player $i \in N$ is a distribution over player i 's actions, denoted $\sigma_i \in \Delta(A_i)$.

A strategy profile is denoted $\sigma = (\sigma_1, \dots, \sigma_{|N|})$.

Definition: (Strict) Nash Equilibrium

A strategy profile σ^* is a (strict) Nash equilibrium if for any player $i \in N$ and any possible strategy $\pi_i \in \Delta(A_i) \setminus \{\pi_i^*\}$, we have $u_i(\pi_i^*, \pi_{-i}^*)(>) \geq u_i(\pi_i, \pi_{-i}^*)$.

Projected Gradient Dynamics

What do we get when we apply gradient ascent to games?

Starting from the initial condition σ^0 , we consider:

Project Gradient Dynamics

Given the strategy profile in episode n , σ^n , player i 's strategy in episode $n + 1$ is

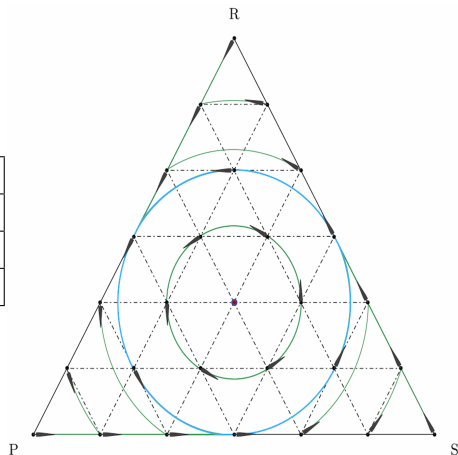
$$\sigma_i^{n+1} = \text{proj}_{\Delta(A_i)} (\sigma_i^n + \gamma_i^n \nabla_i u_i(\sigma^n))$$

where:

- $\nabla_i u_i(\sigma^n) = (u_i(a, \sigma_{-i}^n))_{a \in A_i}$ is the gradient of player i 's utility with respect to their strategy,
- $\gamma_i^n > 0$ is the step size of player i in episode n ,
- $\text{proj}_{\Delta(A_i)} : \mathbb{R}^{|A_i|} \rightarrow \Delta(A_i)$ is the Euclidean projection onto $\Delta(A_i)$.

Example: Rock, Paper, Scissors

	Rock	Paper	Scissors
Rock	0, 0	-1, 1	1, -1
Paper	1, -1	0, 0	-1, 1
Scissors	-1, 1	1, -1	0, 0



Generalised Gradients: The q -gradient

In this work, we consider a generalised form the gradient, which we call the q -gradient for $q \geq 0$.

q -Gradient

$$v_{i,j}^q(\sigma) = \sigma_{i,j}^q \left(u_i(a_j, \sigma_{-i}) - \frac{\sum_k \sigma_{i,k}^q u_i(a_k, \sigma_{-i})}{\sum_k \sigma_{i,k}^q} \right).$$

The term in the parenthesis is the surplus of utility that player i obtains by playing action a_j against a weighted average of the other pure actions. For $q = 0$, we obtain a “normalised” gradient.

The q -replicator Dynamics

q -Gradient

$$v_{i,j}^q(\sigma) = \sigma_{i,j}^q \left(u_i(a_j, \sigma_{-i}) - \frac{\sum_k \sigma_{i,k}^q u_i(a_k, \sigma_{-i})}{\sum_k \sigma_{i,k}^q} \right).$$

The q -replicator Dynamics

Given the strategy profile in episode n , σ^n , player i 's strategy in episode $n + 1$ is

$$\sigma_i^{n+1} = \text{proj}_{\Delta(A_i)} (\sigma_i^n + \gamma_i^n v_i^q(\sigma^n))$$

For $q = 0$, we obtain the projected gradient dynamics.

For $q = 1$, we obtain the replicator dynamics.

For $q = 2$, we obtain the log-barrier dynamics.

The q -replicator Dynamics

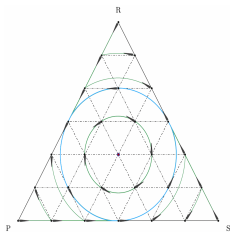


Figure 1: $q = 0$

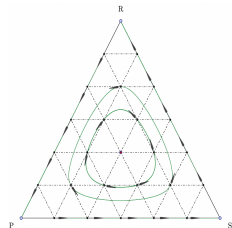


Figure 3: $q = 1$

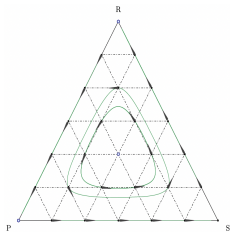


Figure 2: $q = \frac{3}{2}$

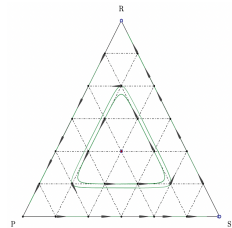


Figure 4: $q = 5$

Repeated Games with Finite Recall

In each period, each player i condition their action upon their private history of the last ℓ_i periods - own actions and own signals.

- $Z = \prod_{i \in N} Z_i$ - the set of signal profiles,
- $q : A \rightarrow \Delta(Z)$ - the joint distribution of signals

This induces a realised history $h = (a^t, z^t)_{t=0}^{\infty}$.

Repeated Games with Finite Recall

$\Pi_i^{\ell_i}$ - the set of mixed ℓ_i -recall strategies of player i .

A strategy profile $\pi \in \Pi^\ell := \prod_{i \in N} \Pi_i^{\ell_i}$ generates a distribution over the set of realisations of the repeated game.

Given a strategy profile $\pi \in \Pi^\ell$, we denote the (normalised) expected utility of player $i \in N$ as:

$$V_i(\pi) := (1 - \delta) \mathbb{E}_{h \sim \pi} \left[\sum_{t=0}^{\infty} \delta^t u_i(a^t) \right]$$

Equilibria

Definition: ℓ -recall equilibrium

A strategy profile $\pi^* \in \Pi^\ell$ is an ℓ -recall equilibrium if for any player $i \in N$ and any ℓ_i -recall strategy $\pi_i \in \Pi_i^{\ell_i}$, we have $V_i(\pi^*) \geq V_i(\pi_i, \pi_{-i}^*)$.

For a strategy profile $\pi \in \Pi^\ell$, let $S(\pi)$ be the set of strategy profiles that induce the same distribution over histories as π .

Definition: ℓ -recall strict equilibrium

A strategy profile $\pi^* \in \Pi^\ell$ is an ℓ -recall strict equilibrium if for any player i and any strategy $\pi_i \in \Pi_i^{\ell_i}$, we have $V_i(\pi^*) > V_i(\pi_i, \pi_{-i}^*)$ or $(\pi_i, \pi_{-i}^*) \in S(\pi^*)$.

The Equivalence Class $S(\pi^*)$

For a strategy profile $\pi \in \Pi^\ell$, we defined $S(\pi)$ to be the set of strategy profiles that induce the same distribution over histories as π .

Properties of $S(\pi)$

- $S(\pi)$ is an equivalence class.
- For π^* an ℓ -recall strict equilibrium, $S(\pi^*)$ may include strategy profiles that are not equilibria.
- For π^* an ℓ -recall strict equilibrium, some strategy profiles in $S(\pi^*)$ close enough to π^* are also strict equilibria.

The q -Replicator Dynamics for Repeated Games

q -Gradient

$$v_{i,\alpha}^q(\pi_i, \pi_{-i}) = \pi_{i,\alpha}^q \left(V_i(e_\alpha, \pi_{-i}) - \frac{\sum_{\beta} \pi_{i,\beta}^q V_i(e_\beta, \pi_{-i})}{\sum_{\beta} \pi_{i,\beta}^q} \right).$$

where e_α is the pure strategy associated to the α^{th} component of π_i .

Episode = Repeated Game

The q -replicator Dynamics

Given the strategy profile in episode n , π^n , player i 's strategy in episode $n + 1$ is

$$\pi_i^{n+1} = \text{proj}_{\Pi_i^{\ell_i}}(\pi_i^n + \gamma_i^n v_i^q(\pi^n))$$

Main Result

Theorem: Local Convergence to Strict Equilibrium

Let $\pi^* \in \Pi^\ell$ be an ℓ -recall strict equilibrium. There exists a neighbourhood \mathcal{U} of π^* in Π^ℓ such that, for any $\eta > 0$, for any $\pi^0 \in \mathcal{U}$, any $p \in (\frac{1}{2}, 1]$, and any positive m , there are $(\gamma_i)_{i \in N}$ small enough such that we have the following: let $(\pi^n)_{n \in \mathbb{N}}$ be the sequence of play generated by q -replicator learning dynamics with step sizes $\gamma_i^n = \frac{\gamma_i}{(n+m)^p}$. Then,

$$\mathbb{P}(\pi^n \rightarrow S(\pi^*) \text{ as } n \rightarrow \infty) \geq 1 - \eta.$$

This result holds with estimators, if noise is sufficiently small.

Main Steps of the Proof

Define $D_n(\pi) = \frac{1}{2} \|\pi - \pi^*\|^2$,

St 1: $D(\pi^{n+1}) \leq D(\pi^n) + \langle v(\pi), \pi^n - \pi^* \rangle + \text{error}$

St 2: If the initial strategy profile is close to π^* , then with high probability all π^n remain close to π^* and sum of error terms is bounded

St 3: If all π^n remain close to π^* then π^n admits a sub-sequence that converges to $S(\pi^*)$

St 4: There exists a random variable D_∞ such that $D(\pi^n)$ converges to D_∞ if all π^n remain close to π^*

Conclusion: Folk Theorem

Strict equilibria with finite recall can be learned.

From Barlo, Carmona and Sabourian (2016): any payoff that is feasible and (strictly) individually rational (with respect to pure minmax) can be obtained as a subgame perfect equilibrium of the infinitely repeated games, if full dimension or two players, the recall of the players is sufficiently long, and they are sufficiently patient.

In fact, for three players or more, we have the folk theorem with mixed minmax defining the individually rational level as well.

Generalisations - Imperfect Monitoring

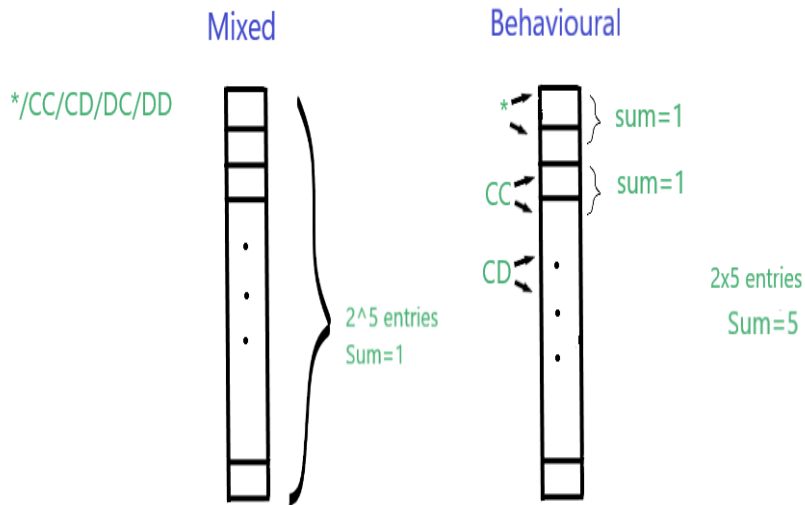
Using REINFORCE, the players obtain an estimator of the q -gradient. Similarly, if rather than perfectly observing past actions of the opponents, the players observe a signal from which they can infer their realised payoff - they still can compute such an estimator.

Imperfect Monitoring - Solution Concepts

If the signal is public, we obtain that any perfect public equilibrium with finite recall can be learned.

If the signal is private - we obtain convergence to "equilibria" that are not necessarily sequential equilibria.

Mixed vs. Behavioural Strategies



Behavioural Strategies

We recover the "traditional" definition of strictness.

A player updates simultaneously behaviour after multiple histories, with improvement "while keeping the others fixed".

Convergence to strict subgame perfect equilibrium with finite recall (same folk theorems).

Reason - the continues to take place even for counterfactual histories.

How Coordinated Do the Players Need to Be?

The players need to all use q -replicator with the same q .

The players may have different recall length.

The step-size can be individual, but some parameters should be the same:

$$\gamma_{i,n} = \frac{\gamma_i}{(n+m_i)^p} \text{ for } \gamma_i, m > 0 \text{ and } p \in (0, \frac{1}{2}].$$

Open Questions

- Different parameters of step-size
- Different q
- Basin of attraction
- Not self-play
- Speed of convergence

THANK YOU

Policy Gradient Literature

- [Leonardos et al., 2022]. *Global convergence of multi-agent policy gradient in Markov potential games.*
- [Daskalakis et al., 2020]. *Independent policy gradient methods for competitive reinforcement learning.*
- [Mertikopoulos and Sandholm, 2016]. *Learning in games via reinforcement and regularization.*
- [Mertikopoulos and Sandholm, 2018]. *Riemannian Game Dynamics*
- [Giannou et al., 2021]. *On the convergence of policy gradient methods to Nash equilibria in general stochastic games.*

Replicator Dynamics Literature

- [Hofbauer et al., 1996]. *Evolutionary Selection against Dominated Strategies.*
- [Viossat, 2007]. *The replicator dynamic does not lead to correlated equilibria.*
- [Viossat, 2008]. *Evolutionary dynamics may eliminate all strategies used in correlated equilibrium.*
- [Viossat, 2015]. *Evolutionary dynamics and dominated strategies.*
- [Sandholm et al., 2008]. *The projection dynamic and the replicator dynamic.*

Equilibria and the q -gradient

Lemma: Equivalence of Strict NE to Variational Inequalities

Let $q \geq 0$. Then, a strategy profile σ^* is a strict Nash equilibrium if and only if the following two conditions are satisfied:

- C(i) For any strategy profile σ , we have $\langle v^q(\sigma^*), \sigma - \sigma^* \rangle \leq 0$.
- C(ii) There is $\varepsilon > 0$ such that for any strategy profile $\sigma \in \prod_{i \in N} \Delta(A_i) \setminus \{\sigma^*\}$ at distance at most ε from σ^* , we have $\langle v^q(\sigma), \sigma - \sigma^* \rangle < 0$.

C(i) is sometimes known as first order stationarity. For $q = 0$, this is equivalent to the strategy being a Nash equilibrium. Whereas, for $q > 0$, this is equivalent to a 'selection equilibrium'.

Approximating the q -gradient

In general, the estimator is a random variable.

For simplicity, we denote this random variable as $\hat{v}_i^n := \hat{v}_i(\pi^n)$.

How good does this estimator need to be?

In this setting, the q -replicator dynamics π^n is a stochastic process.

We write $\mathcal{F}^n := \mathcal{F}(\pi^0, \dots, \pi^n)$ for the filtration of the probability space up to and including episode n . We define

$$b^n = \mathbb{E}[\hat{v}^n | \mathcal{F}^{n-1}] - v^q(\pi^n) \quad \text{and} \quad U^n = \hat{v}^n - \mathbb{E}[\hat{v}^n | \mathcal{F}^{n-1}]$$

We assume that b^n and U^n are bounded such that:

$$\mathbb{E}[\|b^n\| | \mathcal{F}^{n-1}] \leq B^n \quad \text{and} \quad \mathbb{E}[\|U^n\|^2 | \mathcal{F}^{n-1}] \leq (\sigma^n)^2$$

where $B^n = \mathcal{O}(n^{-l_b})$ and $\sigma^n = \mathcal{O}(n^{l_\sigma})$ for $l_b > 0$ and $l_\sigma \in (0, \frac{1}{2})$.

Main Result

For estimators \hat{v}_i^n that satisfy the above conditions, we obtain our main result:

Theorem: Local Convergence under Noisy Dynamics

Let $\pi^* \in \Pi^\ell$ be an ℓ -recall strict equilibrium. There exists a neighbourhood \mathcal{U} of π^* in Π^ℓ such that, for any $\eta > 0$, for any $\pi^0 \in \mathcal{U}$, any $p \in (\frac{1}{2}, 1]$, and any positive m , there are $(\gamma_i)_{i \in N}$ small enough such that we have the following: let $(\pi^n)_{n \in \mathbb{N}}$ be the sequence of play generated by q -replicator learning dynamics with step sizes $\gamma_i^n = \frac{\gamma_i}{(n+m)^p}$ and q -replicator estimates $\hat{v}_i^n(\pi^n)$ such that $p + l_b > 1$ and $p - l_\sigma > 1/2$. Then,

$$\mathbb{P}(\pi^n \rightarrow S(\pi^*) \text{ as } n \rightarrow \infty) \geq 1 - \eta.$$

Equilibria of the Repeated Game and Stability

For a strategy profile $\pi^* \in \Pi^\ell$, we consider the following conditions:

$$C' \text{ (i)} \quad \langle v(\pi^*), \pi - \pi^* \rangle \leq 0, \quad \forall \pi \in \Pi^\ell$$

$$C' \text{ (ii)} \quad \langle v(\pi), \pi - \pi^* \rangle < 0, \quad \forall \pi \in \Pi^\ell \setminus \mathcal{S}(\pi^*) \text{ close enough}$$

Lemma 2: Variational Inequalities and Repeated Game Equilibria

For $q = 0$

(a) Condition $C'(i)$ is equivalent to π^* being a Nash equilibrium.

(b) Strict equilibrium implies $C'(ii)$.

For $q > 0$

(a) Condition $C'(i)$ is equivalent to the following condition: for all $i \in N$, for all e_j in the support of π_i^* , $V_i(e_j, \pi_{-i}^*) \leq V_i(\pi^*)$.

(b) Strict equilibrium implies $C'(ii)$.

The Estimator - ε -Greedy Learning

Algorithm 1 ε -GREEDY q -REPLICATOR

```
1: Input:  $\pi^0 \in \Pi^\ell$ ,  $\{\gamma_i^n\}_{i \in N, n \in \mathbb{N}}$ ,  $\varepsilon \in (0, 1)$ 
2: for  $n = 1, 2, \dots$  do
3:    $\hat{\pi}^n \leftarrow (1 - \varepsilon)\pi^n + \varepsilon \text{Uni}$ 
4:   Sample  $h \sim \hat{\pi}^n$ 
5:   for  $i \in N$  do
6:     Compute  $R_i(h)$ 
7:      $\Lambda_i(h) \leftarrow \sum_{t=0}^{\tau(h)} \nabla_i(\log(\hat{\pi}_i(a_i^t | \hat{h}_i^{\ell_i})))$ 
8:      $\hat{v}_i^n \leftarrow \text{REINFORCE}(R_i(h), \Lambda_i(h))$ 
9:      $\pi_i^{n+1} \leftarrow \text{proj}_{\Pi_i}(\pi_i^n + \gamma_i^n \hat{v}_i^n)$ 
10:  end for
11: end for
```

The Estimator - REINFORCE

Algorithm 2 REINFORCE

- 1: **Input:** $R_i(h), \Lambda_i(h)$
 - 2: $\hat{w}_i \leftarrow R_i(h) \cdot \Lambda_i(h)$
 - 3: $\hat{v}_i \leftarrow \hat{\pi}_{i,j}^q \left(\hat{w}_i(e_j) - \frac{\sum_k \hat{\pi}_{i,k}^q \hat{w}_i(e_k)}{\sum_k \hat{\pi}_{i,k}^q} \right)$
 - 4: **return** \hat{v}_i
-

Estimating $\Lambda_i(h)$, the "log trick"

\hat{w}_i - an unbiased estimator for $\nabla_i V_i(\pi)$.

$$\nabla_i V_i(\pi) = \nabla_i \left(\sum_{h \in H^\infty} \mathbb{P}_\pi(h) R_i(h) \right) = \sum_{h \in H^\infty} (R_i(h) \nabla_i \mathbb{P}_\pi(h))$$

$$\nabla_i \mathbb{P}_\pi(h) = \mathbb{P}_\pi(h) \nabla_i \log(\mathbb{P}_\pi(h))$$




$$\nabla_i V_i(\pi) = \sum_{h \in H^\infty} (R_i(h) \mathbb{P}^\pi(h) \nabla_i \log(\mathbb{P}^\pi(h)))$$

$$\nabla_i V_i(\pi) = \sum_{h \in H^\infty} \left[R_i(h) \mathbb{P}^\pi(h) \sum_t \left(\nabla_i \log(\pi_i(a_i^t(h) | \hat{h}_i^{\ell_i}(t))) \right) \right]$$

$$\Lambda_i(h) := \sum_t \left(\nabla_i \log(\pi_i(a_i^t(h) | \hat{h}_i^{\ell_i}(t))) \right)$$

$$\hat{w}_i \leftarrow R_i(h) \cdot \Lambda_i(h)$$

Bibliography I

-  Daskalakis, C., J. Foster, D., and Golowich, N. (2020).
Independent policy gradient methods for competitive reinforcement learning.
Advances in Neural Information Processing Systems, 33.
-  Giannou, A., Lotidis, K., Mertikopoulos, P., and Vlatakis-Gkaragkounis, E. (2021).
On the convergence of policy gradient methods to nash equilibria in general stochastic games.
arXiv preprint: 2106.01969.
-  Hofbauer, J., Rgen, J., and Weibull, W. (1996).
Evolutionary selection against dominated strategies.

Bibliography II

-  Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. (2022). Global convergence of multi-agent policy gradient in markov potential games.
In International Conference on Learning Representations.
-  Mertikopoulos, P. and Sandholm, W. H. (2016). Learning in games via reinforcement and regularization.
Mathematics of Operations Research, 41(4):1297–1324.
-  Mertikopoulos, P. and Sandholm, W. H. (2018). Riemannian game dynamics.
Journal of Economic Theory, 177:315–364.
-  Sandholm, W. H., Dokumaci, E., and Lahkar, R. (2008). The projection dynamic and the replicator dynamic.
Games and Economic Behavior, 64:666–683.

Bibliography III



Viossat, Y. (2007).

The replicator dynamic does not lead to correlated equilibria.

Games and Economic Behavior, 59:397–407.



Viossat, Y. (2008).

Evolutionary dynamics may eliminate all strategies used in correlated equilibrium.

Mathematical social sciences, 56:27–43.



Viossat, Y. (2015).

Evolutionary dynamics and dominated strategies.

Economic Theory Bulletin, 3:91–113.