Learning with Bandit Feedback in Potential Games

Johanne Cohen, Amélie Héliou, Panayotis Mertikopoulos

LRI-CNRS, LIX, Université Paris Saclay, Univ.Grenoble Alpes, CNRS, Inria, LIG.

Objectives

We show the convergence towards Nash equilibria of the HEDGE algorithm in generic potential games. We focus on the **bandit case**, where players only observe their realized payoffs.

Introduction

Motivated by **current challenges** (network, biology,...) we study algorithms that can be applied to a large number of players that only have a **limited knowledge** of the game. In such games, **no-regret** algorithms are broadly used. **Nash equilibria (NE)**, have the desirable property that no player would benefit from changing alone her strategy. Recent studies [1] show that the long-term limit of play of certain no-regret algorithms is arbitrarily close to a NE with probability close to 1.

*Can **Nash equilibria (NE)** almost surely be the limit result of a no-regret learning algorithm?*

We positively answered this question focusing on the Hedge algorithm [2] that has the property of no-regret. We studied a **low-information framework** where players have only access to a estimate of the pure strategy they played (bandit). We show that when HEDGE is applied to generic potential games [3], **the induced sequence of play converges towards NE** regardless of initialization.

Method

Steps of the proof based on the dynamics of stochastic approximation algorithms:

1. X is an asymptotic pseudo trajectory of the replicator dynamics [4];
2. The potential function is a strict Lyapunov function of the dynamics;
3. X converges toward a rest point of the dynamics [4];
4. If X converges it converges to a NE.

Setup

Game:

- We focus on potential games;
- N players $\mathcal{N} = \{1, \dots, N\}$;
- finite set of strategies per player $\mathcal{S}_i$;
- **mixed strategies** $\mathcal{X}_i = \Delta\mathcal{S}_i$;
- payoff functions $u_i(x) = \langle v_i(x), x\rangle$, with $v_i(x) = (u_i(s_i, x_{-i}))_{s_i \in \mathcal{S}_i}$.

Payoff information: $u_i(s(n))$.

Bandit estimator: $\hat{v}_i(n) = \left(1_{s_i(n)=s_i} \frac{u_i(s_i, s(n)_{-i})}{X_{i,s_i}(n-1)}\right)_{s_i \in \mathcal{S}_i}$.

Step size: $\gamma_n \propto \frac{1}{n^\beta}$ for some $\beta \in (\frac{1}{2}, 1]$.

Logit map: $\Lambda_i(y_i) = \frac{(\exp(y_{is_i}))_{s_i \in \mathcal{S}_i}}{\sum_{s_i \in \mathcal{S}_i} \exp(y_{is_i})}$.

Algorithm

A variant of the **Exponential Weights [2]**, with:

Algorithm 1 ϵ-HEDGE with bandit feedback

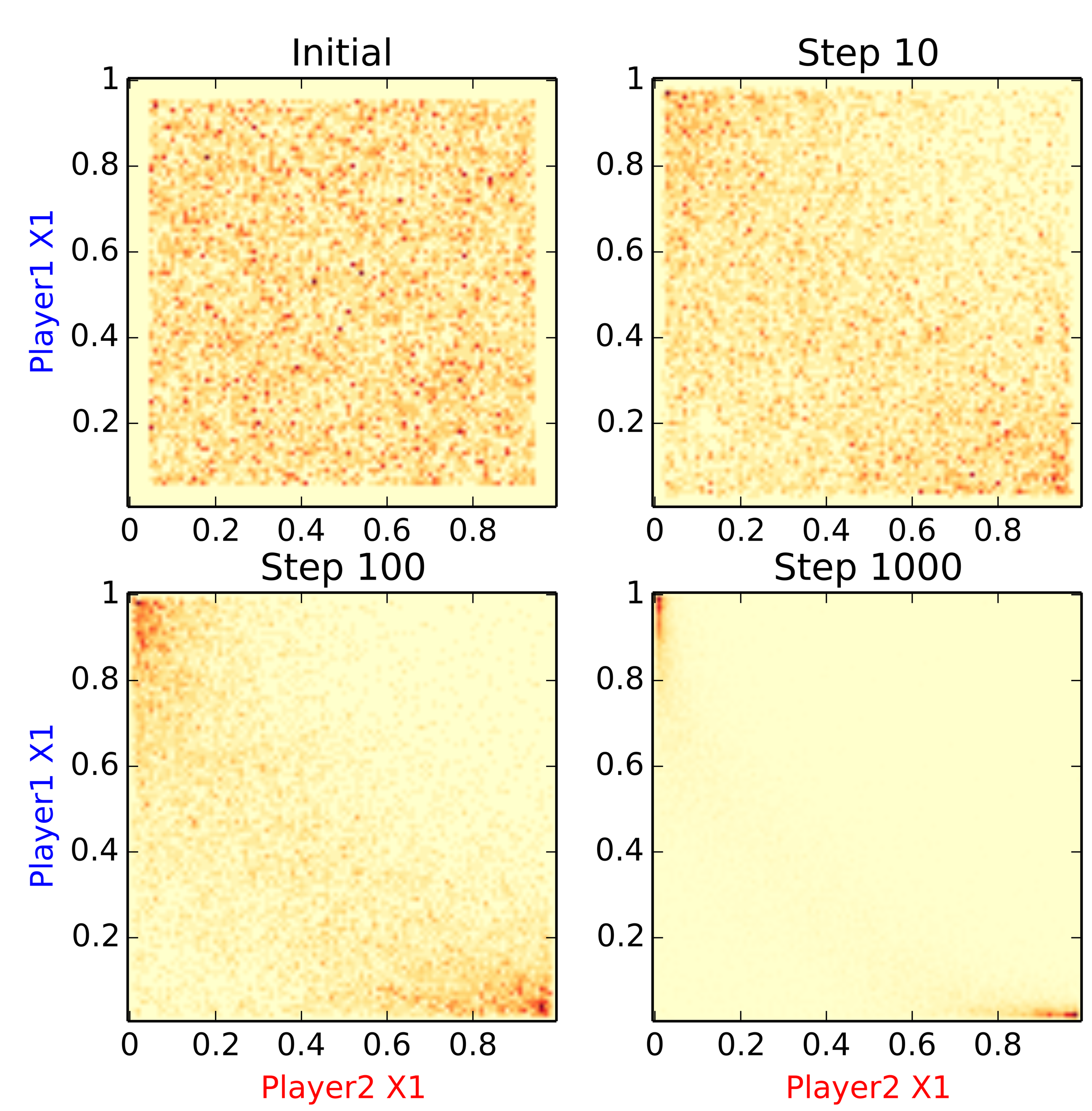
Require: step-size sequence $\gamma_n > 0$, exploration factor sequence $\epsilon_n \in [0, 1]$, initial scores $Y_i \in \mathbb{R}^{S_i}$, $i \in \mathcal{N}$.

```
for n = 1, 2, ... do
  for every player i ∈ N do
    set strategy: X_i ← ε_n/|S_i| + (1 − ε_n)Λ_i(Y_i);
    choose action s_i ∼ X_i;
    compute the bandit estimator v̂_i(n);
    update scores: Y_i ← Y_i + γ_n v̂_i;
  end for
end for
```

Main Result

With an adapted exploration factor, the sequence of play **converges to a Nash equilibrium (a.s.)**.

Experiment



Example Game:

		Player 2: S1	Player 2: S2
Player 1	S1	0 (P1), 0 (P2)	4 (P1), 3 (P2)
Player 1	S2	3 (P1), 4 (P2)	3 (P1), 3 (P2)

Experimentation:

- 10,000 runs of 1,000 steps;
- random initial strategies;
- $\gamma_n = 0.05\frac{1}{n^{2/3}}$;
- $\epsilon_n = 0.1\frac{1}{n^{1/4}}$.

Results

Convergence to δ-NE with $\delta \to_{\epsilon \to 0} 0$ if ϵ_n is constant.

And **convergence to NE almost surely** if the exploration factor ϵ_n decreases so that:

$$\lim_{n\to\infty} \frac{\gamma_n}{\epsilon_n^2} = 0\ ,\ \sum_{n=1}^{\infty} \frac{\gamma_n^2}{\epsilon_n} < \infty \text{ and } \lim_{n\to\infty} \frac{\epsilon_n - \epsilon_{n+1}}{\gamma_n} = 0.$$

Convergence rate

Semi-bandit $\hat{v}_i(n) = (u_i(s_i, s(n)_{-i}) + \xi_n)_{s_i \in \mathcal{S}_i}$.

Noise hypotheses: for some $q > 2$, $A > 0$, and for all $n = 1, 2, \dots$ (a.s.):

- $\mathcal{P}(\|\xi_i(n)\|_\infty^2 \geq z | \mathcal{F}_{n-1}) \leq A/z^q$;
- $\mathbb{E}[\xi_i(n)|\mathcal{F}_{n-1}] = 0$.

We obtain an exponential convergence rate :

$X_{is_i^*}(n) \geq 1 - be^{-c\sum_{i=1}^n \gamma_i}$ for some positive $b, c > 0$.

References

[1] Robert Kleinberg, Georgios Piliouras, and Eva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 533–542. ACM, 2009.

[2] Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1):79–103, 1999.

[3] Dov Monderer and Lloyd S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124 – 143, 1996.

[4] Michel Benaïm. Dynamics of stochastic approximation algorithms. *Séminaire de probabilités de Strasbourg*, 33, 1999.

Contact Information

- Email: amelie.heliou@polytechnique.edu