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

Game:
- We focus on potential games;
- $N$ players $\mathcal{N} = \{1, \ldots, N\}$;
- finite set of strategies per player $S_i$;
- mixed strategies $X_i = \Delta S_i$;
- payoff functions $u_i(x) = (v_i(x), x)$, with $v_i(x) = (u_i(s_i, x_i))_{s_i \in S_i}$.

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

Bandit estimator: $\hat{v}_i(n) = \left(\frac{1}{s_i(n) - s_i(n-1)}\right)_{s_i \in S_i}$.

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

Logit map: $A_i(y_i) = \frac{\exp(y_i)}{\sum_i \exp(y_i)}$.

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

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

\begin{algorithm}[H]
\caption{\textsc{\texttt{HEDGE with bandit feedback}}}

Require: step-size sequence $\gamma_n > 0$, initial strategy sequence $Y_i(0)$, initial strategies $\{s_i(n)\}_{n \geq 0}$.

1. for $n = 1, 2, \ldots$ do
2. for every player $i \in \mathcal{N}$ do
3. set strategy: $X_i(0) = \frac{s_i(n)}{N}$
4. choose action $s_i \sim X_i(0)$;
5. compute the bandit estimator $\hat{v}_i(n)$;
6. update scores: $Y_i(n) = Y_i(n-1) + \gamma_n \hat{v}_i(n)$;
7. end for
8. end for
\end{algorithm}

Experiment

Results
Convergence to $\delta$-NE with $\delta \rightarrow 0$ if $\epsilon_n$ is constant. And convergence to NE almost surely if the exploration factor $\epsilon_n$ decreases so that:

$$\limsup_{n \rightarrow \infty} \frac{\epsilon_n}{\gamma_n} = 0$$

Convergence rate

Semi-bandit $\hat{v}(n) = \left(\frac{1}{s_i(n) - s_i(n-1)}\right)_{s_i \in S_i}$.

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

- $\mathcal{P}(\|\xi(n)\|_q) \geq \frac{1}{2} A / z^q$
- $\mathbb{E}[\|\xi(n)\|_{F_{n-1}^{-1}}] = 0$

We obtain an exponential convergence rate:

$X_{ic}(n) \geq 1 - be^{-r \sum_{m=1}^n \gamma_m n}$ for some positive $b, c, r > 0$.

References

Contact Information
- Email: amelie.heliou@polytechnique.edu