Learning with Bandit Feedback in Potential Games

Johanne Cohen
LRI-CNRS, Université Paris-Sud,Université Paris-Saclay, France
johanne.cohen@lri.fr

Amélie Héliou
LIX, Ecole Polytechnique,Université Paris-Saclay, France
amelie.heliou@polytechnique.edu

Panayotis Mertikopoulos
Univ. Grenoble Alpes, CNRS, Inria, LIG, F-38000, Grenoble, France
panayotis.mertikopoulos@imag.fr

Abstract

This paper examines the equilibrium convergence properties of no-regret learning with exponential weights in potential games. To establish convergence with minimal information requirements on the players' side, we focus on two low-information frameworks: the *semi-bandit case* (where players have access to a noisy estimate of their payoff vector, including strategies they did not play), and the *bandit case* (where players are only able to observe their in-game, realized payoffs). In the semi-bandit case, we show that the induced sequence of play converges almost surely to a Nash equilibrium at a quasi-exponential rate. In the bandit case, the same result holds for ε-approximations of Nash equilibria if we introduce a mixing factor $\varepsilon > 0$ that guarantees that action choice probabilities never fall below ε. In particular, if the algorithm is run with a suitably decreasing mixing factor, the sequence of play converges to a bona fide Nash equilibrium with probability 1.

1 Introduction

Given the manifest complexity of computing Nash equilibria in non-cooperative games, a central question that arises is whether such outcomes may be seen as the end result of a dynamic process where the participants accumulate and act on empirical information on their strategies' performance over time. Largely motivated by recent applications of game theory to networks and biology, this question becomes particularly important when the players' view of the game is obstructed by situational uncertainty and the "fog of war". For example, when deciding which route to take to work each morning, a commuter is typically unaware of how many other commuters there are at any given moment, what their possible strategies are, how to best respond to their choices, etc. In situations of this kind, players may not even know they are involved in a game so it does not seem reasonable to assume full rationality, common knowledge of rationality, flawless execution, etc. in order to justify the Nash equilibrium prediction.

A compelling alternative to this "rationalistic" viewpoint is provided by the framework of *online learning*, where players are treated as agnostic entities facing a repeated decision process with a priori unknown rules and outcomes. When the players have no Bayesian prior on their environment, the most widely used performance guarantee is that of regret minimization, a "worst-case" bound that was first introduced in game theory by Hannan [1] and which has given rise to a vigorous literature at the interface of optimization, statistics and theoretical computer science – for a survey, see [2, 3].

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Accordingly, our starting point in this paper is the following question: *If all players of a repeated game follow a no-regret learning algorithm, does play converge to a Nash equilibrium?*

For concreteness, we focus on the *exponential weights* (EW) algorithm [4, 5], one of the most popular and widely studied algorithms for no-regret learning. In a nutshell, the main idea of the scheme is that the optimizing agent maintains an auxiliary score vector that tallies the cumulative payoffs of each of their actions, and they then employ a pure strategy with probability proportional to the exponential of this score. Under this scheme, players are guaranteed a universal, $\min\max$-optimal $\mathcal{O}(\sqrt{T})$ regret bound (with T denoting the game's horizon), and their empirical frequency of play is known to converge to the game's set of *coarse correlated equilibria* (CCE) [6].

In this way, learning provides a positive partial answer to our original question: coarse correlated equilibria are indeed learnable if all players follow an exponential weights learning scheme. On the flip side however, this set is large enough to also contain highly non-rationalizable strategies, so the end prediction of empirical convergence to the set of coarse correlated equilibrium is fairly lax. For instance, in a recent paper, Viossat and Zapechelnyuk constructed a 4×4 variant of Rock-Paper-Scissors with a coarse correlated equilibrium that assigns positive weight *only* on strictly dominated strategies [7]. In view of such counterexamples, a more calibrated answer to this question is "not always": especially when the issue at hand is convergence to a Nash equilibrium (as opposed to e.g. a coarse correlated equilibrium), having "no regret" is a rather loose guarantee.

Paper outline and summary of results. To address the above limitations, we focus on two issues:

a) Convergence to Nash equilibrium (as opposed to some coarser solution concept).

b) The convergence of the actual sequence of play (as opposed to empirical frequencies).

The reason for focusing on the convergence of the actual sequence of play is that time-averages provide a fairly weak mode of convergence: a priori, a player could oscillate between two non-equilibrium strategies with suboptimal payoffs, but the time-average might still converge to equilibrium. On the other hand, convergence of the actual sequence of play both implies convergence of the time-averages and also guarantees that players will be playing a Nash equilibrium in the long run, so it is a much stronger requirement.

To establish convergence, we focus throughout on the class of *potential games* [8] that has found widespread applications in theoretical computer science, transportation networks [9], wireless communications [10], biology [11], and many other fields. We then focus on two different feedback models: in the *semi-bandit framework* (Section 3), players are assumed to have some (possibly imperfect) estimate of their payoff vectors at each stage, including strategies that they did not play; in the full *bandit framework* (Section 4), this assumption is relaxed and players are only assumed to observe their realized, in-game payoff at each stage.

Starting with the semi-bandit case, our main result is that under fairly mild conditions for the errors affecting the players' observations (zero-mean martingale noise with tame second-moment tails), learning with exponential weights converges to a Nash equilibrium of the game with probability 1 – or to an ε-equilibrium if the algorithm is implemented with a positive mixing factor $\varepsilon > 0$.[1] We also show that this convergence occurs at a quasi-exponential rate, i.e. much faster than the algorithm's $\mathcal{O}(\sqrt{T})$ regret minimization rate would suggest.

These conclusions also apply to the bandit case when the algorithm is run with a positive mixing factor $\varepsilon > 0$; hence, by choosing a sufficiently small mixing factor, the end state of the EW algorithm in potential games is arbitrarily close to a Nash equilibrium. Extending the stochastic approximation and martingale limit arguments that underlie the bandit analysis to the $\varepsilon = 0$ case is not straightforward. However, by letting the mixing factor go to zero at a suitable rate (similar to the temperature parameter of simulated annealing schemes), we are able to recover convergence precisely to the game's Nash set. We find this property particularly appealing for practical applications because it shows that equilibrium can be achieved in a wide class of games with minimal information requirements; in the supplement, we also show that this result extends to a much larger class of no-regret learning algorithms based on "following the regularized leader" (FoReL).

[1] Having a mixing factor of $\varepsilon > 0$ simply means here that action selection probabilities never fall below ε.

Related work. No-regret learning has given rise to a vast corpus of literature and several well-known families of algorithms have been proposed, the most popular being based on exponential/multiplicative weights and their variants [4, 12, 5], or online mirror descent (OMD) and FoReL [13]. In this context, [14] showed that the players' sum of payoffs approaches an approximate optimum, and there is convergence of time averages towards an equilibrium in two-player zero-sum games [15, 16, 12]. In all these examples, the players' average regret vanishes at a worst-case rate of $\mathcal{O}(1/\sqrt{T})$. This convergence rate to approximate efficiency and to coarse correlated equilibria was improved by Syrgkanis et al. [17] for a wide class of N-player normal form games using a natural class of regularized learning algorithms. This result was subsequently extended to a class of games known as *smooth games* [18] with good properties in terms of the game's price of anarchy [19, 17].

In the case of potential games, learning algorithms and dynamics have received signifcant attention over the last few years. For example, the HEDGE variant of the exponential weights algorithm was recently studied by Kleinberg et al. [20] who proved that, in a specific class of load balancing games, the dynamics' long-term limit is exponentially better than the worst correlated equilibrium and almost as good as that of the worst Equilibrium. Moreover, Kleinberg et al. [21] analyze a version of this algorithm in congestion games (which are themselves potential games): their main result is that the the empirical frequencies of play converge to the set of weakly stable equilibria (which, among others, contains Nash equilibrium in pure strategies). Their proof is based on Rosenthal's potential function and the fact that the algorithm can be approximated by an ordinary differential equation; however, it does not follow from the proof that the sequence of play actually converges to a Nash equilibrium. Krichene et al. [22] recently studied a similar question in the context of congestion games, and proved that a discounted variant of HEDGE converges to game's Nash Equilibrium set in the sense of Cesàro means, while stronger convergence results can be guaranteed with some additional conditions.

In the above works, the focus is on the empirical frequencies of play and it is assumed that players have full information on their payoff vectors. To obtain almost sure convergence results for the *actual* sequence of play, it is necessary to go beyond traditional regret-based techniques that rely on the averaging inherent to the empirical frequencies of play. To accomplish this, we rely on martingale limit theory and Benaïm's powerful stochastic approximation results for autonomous dynamical systems [23] which provide the mathematical foundations of our work.

2 The setup

2.1 Game-theoretic preliminaries

An N-player *game in normal form* consists of a (finite) set of *players* $\mathcal{N} = \{1, \dots, N\}$, each with a finite set of *actions* (or *strategies*) $\mathcal{S}_i$. The preferences of the i-th player for one action over another are then determined by an associated *payoff function* $u_i \colon \mathcal{S} \equiv \prod_i \mathcal{S}_i \to \mathbb{R}$ that maps the *profile* $(s_i; s_{-i})$ of all players' actions to the player's reward $u_i(s_i; s_{-i})$.[2] Putting all this together, a game will be written as a tuple $\Gamma \equiv \Gamma(\mathcal{N}, \mathcal{S}, u)$ with players, actions and payoffs defined as above.

Players can also use *mixed strategies* by playing a probability distribution $x_i = (x_{is_i})_{s_i \in \mathcal{S}_i} \in \Delta(\mathcal{S}_i)$ over their action sets $\mathcal{S}_i$. The resulting probability vector x_i is called a *mixed strategy* and we write $\mathcal{X}_i = \Delta(\mathcal{S}_i)$ for the mixed strategy space of player i. Aggregating over players, we also write $\mathcal{X} = \prod_i \mathcal{X}_i$ for the game's *strategy space*, i.e. the space of all strategy profiles $x = (x_i)_{i \in \mathcal{N}}$.

In this context (and in a slight abuse of notation), the expected payoff of the i-th player in the profile $x = (x_1, \dots, x_N)$ is

$$u_i(x) = \sum_{s_1 \in \mathcal{S}_1} \cdots \sum_{s_N \in \mathcal{S}_N} u_i(s_1, \dots, s_N)\, x_{1s_1} \cdots x_{Ns_N}. \tag{2.1}$$

To keep track of the payoff of each pure strategy, we also write $v_{is_i}(x) = u_i(s_i; x_{-i})$ for the payoff of strategy $s_i \in \mathcal{S}_i$ under the profile $x \in \mathcal{X}$ and $v_i(x) = (v_{is_i}(x))_{s_i \in \mathcal{S}_i}$ for the resulting *payoff vector* of player i. We then have

$$u_i(x) = \langle v_i(x), x_i \rangle = \sum_{s_i \in \mathcal{S}_i} x_{is_i} v_{is_i}(x), \tag{2.2}$$

[2] In the above $(s_i; s_{-i})$ is shorthand for $(s_1, \dots, s_i, \dots, s_N)$, used here to highlight the action of player i against that of all other players.

where $\langle v, x\rangle \equiv v^\top x$ denotes the ordinary pairing between v and x.

The most widely used solution concept in game theory is that of a *Nash equilibrium* (NE), i.e. a state $\hat{x} \in \mathcal{X}$ such that

$$u_i(\hat{x}_i; \hat{x}_{-i}) \geq u_i(x_i; \hat{x}_{-i}) \quad \text{for every deviation } x_i \in \mathcal{X}_i \text{ of player } i \text{ and all } i \in \mathcal{N}. \tag{NE}$$

Equivalently, writing $\operatorname{supp}(x_i) = \{s_i \in \mathcal{S}_i : x_i > 0\}$ for the support of $x_i \in \mathcal{X}_i$, we have the equivalent characterization:

$$v_{is_i}(\hat{x}) \geq v_{is_i'}(\hat{x}) \quad \text{for all } s_i \in \operatorname{supp}(\hat{x}_i) \text{ and all } s_i' \in \mathcal{S}_i,\ i \in \mathcal{N}. \tag{2.3}$$

A Nash equilibrium $\hat{x} \in \mathcal{X}$ is further said to be *pure* if $\operatorname{supp}(\hat{x}_i) = \{\hat{s}_i\}$ for some $\hat{s}_i \in \mathcal{S}_i$ and all $i \in \mathcal{N}$. In *generic games* (that is, games where small changes to any payoff do not introduce new Nash equilibria or destroy existing ones), every pure Nash equilibrium is also *strict* in the sense that (2.3) holds as a strict inequality for all $s_i \neq \hat{s}_i$.

In our analysis, it will be important to consider the following relaxations of the notion of a Nash equilibrium: First, weakening the inequality (NE) leads to the notion of a δ-*equilibrium*, defined here as any mixed strategy profile $\hat{x} \in \mathcal{X}$ such that

$$u_i(\hat{x}_i; \hat{x}_{-i}) + \delta \geq u_i(x_i; \hat{x}_{-i}) \quad \text{for every deviation } x_i \in \mathcal{X}_i \text{ and all } i \in \mathcal{N}. \tag{NE$_\delta$}$$

Finally, we say that $\hat{x}$ is a *restricted equilibrium* (RE) of Γ if

$$v_{is_i}(\hat{x}) \geq v_{is_i'}(\hat{x}) \quad \text{for all } s_i \in \operatorname{supp}(\hat{x}_i) \text{ and all } s_i' \in \mathcal{S}_i',\ i \in \mathcal{N}, \tag{RE}$$

where $\mathcal{S}_i'$ is some restricted subset of $\mathcal{S}_i$ containing $\operatorname{supp}(\hat{x}_i)$. In words, restricted equilibria are Nash equilibria of Γ restricted to subgames where only a subset of the players' pure strategies are available at any given moment. Clearly, Nash equilibria are restricted equilibria but the converse does not hold: for instance, every corner of $\mathcal{X}$ is a restricted equilibrium, but not necessarily a Nash equilibrium.

Throughout this paper, we will focus almost exclusively on the class of *potential games*, which have been studied extensively in the context of congestion, traffic networks, oligopoly markets, etc. Following Monderer and Shapley [8], Γ is a *potential game* if it admits a *potential function* $f\colon \prod_i \mathcal{S}_i \to \mathbb{R}$ such that

$$u_i(x_i; x_{-i}) - u_i(x_i'; x_{-i}) = f(x_i; x_{-i}) - f(x_i'; x_{-i}), \tag{2.4}$$

for all $x_i, x_i' \in \mathcal{X}_i$, $x_{-i} \in \mathcal{X}_{-i} \equiv \prod_{j\neq i} \mathcal{X}_i$, and all $i \in \mathcal{N}$. A simple differentiation of (2.1) then yields

$$v_i(x) = \nabla_{x_i} u_i(x) = \nabla_{x_i} f(x) \quad \text{for all } i \in \mathcal{N}. \tag{2.5}$$

Obviously, every local maximizer of f is a Nash equilibrium so potential games always admit Nash equilibria in pure strategies (which are also strict if the game is generic).

2.2 Learning with exponential weights

Our basic learning framework is as follows: At each stage $n = 1, 2, \dots$, every player $i \in \mathcal{N}$ selects an action $s_i(n) \in \mathcal{S}_i$ according to some mixed strategy $X_i(n) \in \mathcal{X}_i$. All players then receive some feedback on their chosen actions, they update their mixed strategies, and the process repeats.

A popular (and very widely studied) class of algorithms for no-regret learning in this setting is the *exponential weights* (EW) scheme introduced by Vovk [4] and Littlestone and Warmuth [5]. Somewhat informally, the main idea of the scheme is that each player maintains an auxiliary score vector $Y_i \in \mathbb{R}^{\mathcal{S}_i}$ that tallies the cumulative payoffs of each of their actions, and then employs a pure strategy $s_i \in \mathcal{S}_i$ with probability roughly proportional to the exponential of this score. Focusing on the so-called "ε-HEDGE" variant of the EW algorithm [24], this process can be described in pseudocode form as follows:

Algorithm 1 ε-HEDGE with generic feedback

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

for $n = 1, 2, \dots$ **do**
  for every player $i \in \mathcal{N}$ **do**
    set mixed strategy: $X_i \leftarrow \varepsilon/|\mathcal{S}_i| + (1-\varepsilon)\,\Lambda_i(Y_i)$;
    choose action $s_i \sim X_i$;
    acquire estimate $\hat{v}_i$ of realized payoff vector $v_i(s_i; s_{-i})$;
    update scores: $Y_i \leftarrow Y_i + \gamma_n \hat{v}_i$;
  end for
end for

Mathematically, Algorithm 1 represents the recursion

$$\begin{aligned} X_i(n) &= \varepsilon/|\mathcal{S}_i|\,\mathbf{1} + (1-\varepsilon)\,\Lambda_i(Y_i(n)), \\ Y_i(n+1) &= Y_i(n) + \gamma_n \hat{v}_i(n), \end{aligned} \tag{ε-Hedge}$$

where $\mathbf{1} = (1, \dots, 1)$ is a vector of ones of the appropriate dimension, while $\Lambda_i \colon \mathbb{R}^{\mathcal{S}_i} \to \mathcal{X}_i$ denotes the *logit choice map*

$$\Lambda_i(y_i) = \frac{(\exp(y_{is_i}))_{s_i \in \mathcal{S}_i}}{\sum_{s_i \in \mathcal{S}_i} \exp(y_{is_i})}, \tag{2.6}$$

which assigns exponentially higher probability to pure strategies with higher scores. Thus, action selection probabilities under (ε-Hedge) are a convex combination of uniform mixing (with total weight ε) and exponential weights (with total weight $1-\varepsilon$).[3] As a result, for $\varepsilon \approx 1$, action selection is essentially uniform; at the other extreme, when $\varepsilon = 0$, we obtain the original Hedge algorithm of [24] with feedback sequence $\hat{v}(n)$ and no uniform mixing.

The no-regret properties of (ε-Hedge) have been extensively studied in the literature as a function of the algorithm's step-size sequence γ_n, mixing factor ε, and the statistical properties of the payoff estimates $\hat{v}(n)$ – for a survey, we refer the reader to [2, 3]. In our convergence analysis, we examine the role of each of these factors in detail, focusing in particular on the distinction between "*semibandit feedback*" (when it is possible to estimate the payoff of pure strategies that were not played) and "*bandit feedback*" (when players only observe the payoff of their chosen action). For now, the only additional remark worth making is that (ε-Hedge) is itself a special case of a much broader class of no-regret learning methods based on "following the regularized leader" (FoReL); we explore the convergence properties of this class of algorithms in the supplement.

3 Learning with semi-bandit information

3.1 The model

We begin with the *semi-bandit framework*, i.e. the case where each player has access to a possibly imperfect estimate of their entire payoff vector at stage n. More precisely, we assume here that the feedback sequence $\hat{v}_i(n)$ to Algorithm 1 is of the general form

$$\hat{v}_i(n) = v_i(s_i(n); s_{-i}(n)) + \xi_i(n), \tag{3.1}$$

where $(\xi_i(n))_{i \in \mathcal{N}}$ is a martingale noise process representing the players' estimation error and satisfying the following statistical hypotheses:

1. *Zero-mean:*

$$\mathbb{E}[\xi_i(n) \,|\, \mathcal{F}_{n-1}] = 0 \quad \text{for all } n = 1, 2, \dots \text{ (a.s.).} \tag{H1}$$

2. *Tame tails:*

$$\mathbb{P}(\|\xi_i(n)\|_\infty^2 \geq z \,|\, \mathcal{F}_{n-1}) \leq A/z^q \quad \text{for some } q > 2,\ A > 0, \text{ and all } n = 1, 2, \dots \text{ (a.s.).} \tag{H2}$$

[3] Of course, the mixing factor ε could also be player-dependent. For simplicity, we state all our results here with the same ε for all players, and we discuss this more general case in the appendix.

In the above, the expectation $\mathbb{E}[\,\cdot\,]$ is taken with respect to some underlying filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ which serves as a stochastic basis for the process $(Y(n), X(n), s(n), \hat{v}(n))_{n\geq 1}$. In words, Hypothesis (H1) simply means that the players' feedback sequence $\hat{v}(n)$ is *conditionally unbiased* with respect to the history of play, i.e.[4]

$$\mathbb{E}[\hat{v}_i(n) \,|\, \mathcal{F}_{n-1}] = v_i(X(n)), \quad \text{for all } n = 1, 2, \dots \text{ (a.s.)}. \tag{3.2a}$$

As for Hypothesis (H2), it readily implies that the mean squared error of the estimator $\hat{v}$ is conditionally bounded, i.e.

$$\mathbb{E}[\|\hat{v}(n) - v(X(n))\|_\infty^2 \,|\, \mathcal{F}_{n-1}] \leq \sigma^2 \quad \text{for all } n = 1, 2, \dots \text{ (a.s.)}. \tag{3.2b}$$

Remark 1. By Chebyshev's inequality, an estimator with finite mean squared error enjoys the tail bound $\mathbb{P}(\|\xi_i(n)\|_\infty \geq z \,|\, \mathcal{F}_{n-1}) = \mathcal{O}(1/z^2)$. At the expense of working with slightly more conservative step-size policies (see below), much of our analysis goes through with this weaker requirement for the tails of ξ. However, given that the extra control provided by the $\mathcal{O}(1/z^q)$ tail bound simplifies the presentation considerably, we do not consider this relaxation here. In any event, Hypothesis (H2) is satisfied by a broad range of error noise distributions (including all compactly supported, sub-Gaussian and sub-exponential distributions), so the loss in generality is small compared to the gain in clarity and concision.

3.2 Convergence analysis

With all this at hand, our main result for the convergence of (ε-Hedge) with feedback of the form (3.1) is as follows:

Theorem 1. *Let Γ be a generic potential game and suppose that Algorithm 1 is run with i) semi-bandit feedback satisfying (H1) and (H2); ii) a nonnegative mixing factor $\varepsilon \geq 0$; and iii) a step-size sequence of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/q, 1]$. Then:*

1. *$X(n)$ converges (a.s.) to a δ-equilibrium of Γ with $\delta \equiv \delta(\varepsilon) \to 0$ as $\varepsilon \to 0$.*
2. *If $\lim_{n\to\infty} X(n)$ is an ε-pure state of the form $\hat{x}_i = \varepsilon/|\mathcal{S}_i|\mathbf{1} + (1-\varepsilon)e_{\hat{s}_i}$ for some $\hat{s} \in \mathcal{S}$, then $\hat{s}$ is a.s. a strict equilibrium of Γ and convergence occurs at a quasi-exponential rate:*

$$X_{i\hat{s}_i}(n) \geq 1 - \varepsilon - be^{-c\sum_{k=1}^n \gamma_k} \quad \text{for some positive } b, c > 0. \tag{3.3}$$

Corollary 2. *If Algorithm 1 is run with assumptions as above and no mixing ($\varepsilon = 0$), $X(n)$ converges to a Nash equilibrium with probability 1. Moreover, if the limit of $X(n)$ is pure and $\beta < 1$, we have*

$$X_{i\hat{s}_i}(n) \geq 1 - be^{-cn^{1-\beta}} \quad \text{for some positive } b, c > 0. \tag{3.4}$$

Sketch of the proof. The proof of Theorem 1 is fairly convoluted, so we relegate the details to the paper's technical appendix and only present here a short sketch thereof.

Our main tool is the so-called *ordinary differential equation* (ODE) method, a powerful stochastic approximation scheme due to Benaïm and Hirsch [25, 23]. The key observation is that the mixed strategy sequence $X(n)$ generated by Algorithm 1 can be viewed as a "Robbins–Monro approximation" (an *asymptotic pseudotrajectory* to be precise) of the ε-perturbed exponential learning dynamics

$$\begin{aligned} \dot{y}_i &= v_i(x), \\ x_i &= \varepsilon/|\mathcal{S}_i|\mathbf{1} + (1-\varepsilon)\,\Lambda_i(y_i), \end{aligned} \tag{XL$_\varepsilon$}$$

By differentiating, it follows that $x_i(t)$ evolves according to the *ε-perturbed replicator dynamics*

$$\dot{x}_{is} = \big(x_{is} - |\mathcal{S}_i|^{-1}\varepsilon\big)\Big[v_{is}(x) - (1-\varepsilon)^{-1}\sum\nolimits_{s'\in\mathcal{S}_i}(x_{is'} - |\mathcal{S}_i|^{-1}\varepsilon)v_{is'}(x)\Big], \tag{RD$_\varepsilon$}$$

which, for $\varepsilon = 0$, boil down to the ordinary *replicator dynamics* of Taylor and Jonker [26]:

$$\dot{x}_{is} = x_{is}[v_{is}(x) - \langle v_i(x), x_i\rangle], \tag{RD}$$

A key property of the replicator dynamics that readily extends to the ε-perturbed variant (RD$_\varepsilon$) is that the game's potential f is a strict Lyapunov function – i.e. $f(x(t))$ is increasing under (RD$_\varepsilon$)

[4] (a.s.) means *almost surely*.

unless $x(t)$ is stationary. By a standard result of [23], this implies that the discrete-time process $X(n)$ converges (a.s.) to a connected set of rest points of (RD$_\varepsilon$), which are themselves approximate restricted equilibria of Γ.

Now, since every ε-pure point of the form $(\varepsilon/|\mathcal{S}_i|\mathbf{1} + (1-\varepsilon)e_{s_i})_{i\in\mathcal{N}}$ is also stationary under (RD$_\varepsilon$), the above does not imply that the limit of $X(n)$ is an approximate Nash equilibrium of Γ. To rule out such outcomes, we first note that the set of rest points of (RD$_\varepsilon$) is finite (by genericity), so $X(n)$ must converge to a point. Then, the final step of our convergence proof is provided by a martingale recurrence argument which shows that when $X(n)$ converges to a point, this limit must be an approximate equilibrium of Γ. Finally, the rate of convergence (3.3) is obtained by comparing the payoff of a player's equilibrium strategy to that of the player's other strategies, and then "inverting" the logit choice map to translate this into an exponential decay rate for $\|X_{i\hat{s}_i}(n) - \hat{x}\|$. □

We close this section with two remarks on Theorem 1. First, we note that there is an inverse relationship between the tail exponent q in (H2) and the decay rate β of the algorithm's step-size sequence $\gamma_n \propto n^{-\beta}$. Specifically, higher values of q imply that the noise in the players' observations is smaller (on average and with high probability), so players can be more aggressive in their choice of step-size. This is reflected in the lower bound $1/q$ for β and the fact that the players' rate of convergence to Nash equilibrium increases with smaller β; in particular, (3.3) shows that Algorithm 1 enjoys a convergence bound which is just shy of $\mathcal{O}(\exp(-n^{1-1/q}))$. Thus, if the noise process ξ is sub-Gaussian/sub-exponential (so q can be taken arbitrarily large), a near-constant step-size sequence (small β) yields an almost linear convergence rate.

Second, if the noise process ξ is "isotropic" in the sense of [23, Thm. 9.1], the instability of non-pure Nash equilibria under the replicator dynamics can be used to show that the limit of $X(n)$ is pure with probability 1.[5] When this is the case, the quasi-exponential convergence rate (3.3) becomes universal in that it holds with probability 1 (as opposed to conditioning on $\lim_{n\to\infty} X(n)$ being pure). We find this property particularly appealing for practical applications because it shows that equilibrium is reached *exponentially faster* than the $\mathcal{O}(1/\sqrt{n})$ worst-case regret bound of (ε-Hedge) would suggest.

4 Payoff-based learning: the bandit case

We now turn to the *bandit framework*, a minimal-information setting where, at each stage of the process, players only observe their realized payoffs

$$\hat{u}_i(n) = u_i(s_i(n); s_{-i}(n)). \tag{4.1}$$

In this case, players have no clue about the payoffs of strategies that were not chosen, so they must *construct* an estimator for their payoff vector, including its missing components.

A standard way to do this is the bandit estimator

$$\hat{v}_{is_i}(n) = \frac{\mathbb{1}(s_i(n) = s_i)}{\mathbb{P}(s_i(n) = s_i \,|\, \mathcal{F}_{n-1})} \cdot \hat{u}_i(n) = \begin{cases} \hat{u}_i(n)/X_{is_i}(n) & \text{if } s_i = s_i(n), \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

Indeed, a straightforward calculation shows that

$$\begin{aligned} \mathbb{E}[\hat{v}_{is_i}(n) \,|\, \mathcal{F}_{n-1}] &= \sum_{s_{-i}\in\mathcal{S}_{-i}} X_{-i,s_{-i}}(n) \sum_{s_i'\in\mathcal{S}_i} X_{is_i'}(n) \frac{\mathbb{1}(s_i = s_i')}{X_{is_i}(n)} u_i(s_i'; s_{-i}) = u_i(s_i; X_{-i}(n)) \\ &= v_{is_i}(X(n)), \end{aligned} \tag{4.3}$$

so the estimator (4.2) is unbiased in the sense of (H1)/(3.2a). On the other hand, a similar calculation shows that the variance of $\hat{v}_{is_i}(n)$ grows as $\mathcal{O}(1/X_{is_i}(n))$, implying that (H2)/(3.2b) may fail if $X_{is_i}(n)$ becomes arbitrarily small.

Importantly, this can never happen if (ε-Hedge) is run with a strictly positive mixing factor $\varepsilon > 0$. In that case, we can show that the bandit estimator (4.2) satisfies both (H1) and (H2), leading to the following result:

[5] Specifically, we refer here to the so-called "folk theorem" of evolutionary game theory which states that $\hat{x}$ is asymptotically stable under (RD) if and only if it is a strict Nash equilibrium of Γ [11]. The extension of this result to the ε-replicator system (RD$_\varepsilon$) is immediate.

Theorem 3. *Let Γ be a generic potential game and suppose that Algorithm 1 is run with i) the bandit estimator (4.2); ii) a strictly positive mixing factor $\varepsilon > 0$; and iii) a step-size sequence of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (0, 1]$. Then:*

1. *$X(n)$ converges (a.s.) to a δ-equilibrium of Γ with $\delta \equiv \delta(\varepsilon) \to 0$ as $\varepsilon \to 0$.*
2. *If $\lim_{n\to\infty} X(n)$ is an ε-pure state of the form $\hat{x}_i = \varepsilon/|\mathcal{S}_i|\mathbf{1} + (1-\varepsilon)e_{\hat{s}_i}$ for some $\hat{s} \in \mathcal{S}$, then $\hat{s}$ is a.s. a strict equilibrium of Γ and convergence occurs at a quasi-exponential rate:*

$$X_{i\hat{s}_i}(n) \geq 1 - \varepsilon - be^{-c\sum_{k=1}^{n}\gamma_k} \quad \text{for some positive } b, c > 0. \tag{4.4}$$

Proof. Under Algorithm 1, the estimator (4.2) gives

$$\|\hat{v}_i(n)\| = \frac{|\hat{u}_i(n)|}{X_{is_i(n)}(n)} \leq \frac{|u_i(s_i(n); s_{-i}(n))|}{\varepsilon} \leq \frac{u_{\max}}{\varepsilon}, \tag{4.5}$$

where $u_{\max} = \max_{i\in\mathcal{N}} \max_{s_1\in\mathcal{S}_1} \cdots \max_{s_N\in\mathcal{S}_N} u_i(s_1, \dots, s_N)$ denotes the absolute maximum payoff in Γ. This implies that (H2) holds true for all $q > 2$, so our claim follows from Theorem 1. □

Theorem 3 shows that the limit of Algorithm 1 is closer to the Nash set of the game if the mixing factor ε is taken as small as possible. On the other hand, the crucial limitation of this result is that it does not apply to the case $\varepsilon = 0$ which corresponds to the game's bona fide Nash equilibrium. As we discussed above, the reason for this is that the variance of $\hat{v}(n)$ may grow without bound if action choice probabilities can become arbitrarily small, in which case the main components of our proof break down.

With this "bias-variance" trade-off in mind, we introduce below a modified version of Algorithm 1 with an "annealing" schedule for the method's mixing factor:

Algorithm 2 Exponential weights with annealing

Require: step-size sequence $\gamma_n > 0$, variable mixing factor $\varepsilon_n > 0$, initial scores $Y_i \in \mathbb{R}^{\mathcal{S}_i}$.

for $n = 1, 2, \dots$ **do**
  for every player $i \in \mathcal{N}$ **do**
    set mixed strategy: $X_i \leftarrow \varepsilon_n/|\mathcal{S}_i| + (1-\varepsilon_n)\,\Lambda_i(Y_i)$;
    choose action $s_i \sim X_i$ and receive payoff $\hat{u}_i \leftarrow u_i(s_i; s_{-i})$;
    set $\hat{v}_{is_i} \leftarrow \hat{u}_i/X_{is_i}$ and $\hat{v}_{is_i'} \leftarrow 0$ for $s_i' \neq s_i$;
    update scores: $Y_i \leftarrow Y_i + \gamma_n \hat{v}_i$;
  end for
end for

Of course, the convergence of Algorithm 2 depends heavily on the rate at which ε_n decays to 0 relative to the algorithm's step-size sequence γ_n. This can be seen clearly in our next result:

Theorem 4. *Let Γ be a generic potential game and suppose that Algorithm 1 is run with i) the bandit estimator (4.2); ii) a step-size sequence of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/2, 1]$; and iii) a decreasing mixing factor $\varepsilon_n \downarrow 0$ such that*

$$\lim_{n\to\infty} \frac{\gamma_n}{\varepsilon_n^2} = 0 \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{\gamma_n^2}{\varepsilon_n^2} < \infty. \tag{4.6}$$

Then, $X(n)$ converges (a.s.) to a Nash equilibrium of Γ.

The main challenge in proving Theorem 4 is that, unless the "innovation term" $U_i(n) = \hat{v}_i(n) - v_i(X(n))$ has bounded variance, Benaïm's general theory does not imply that $X(n)$ forms an asymptotic pseudotrajectory of the underlying mean dynamics – here, the unperturbed replicator system (RD). Nevertheless, under the summability condition (4.6), it is possible to show precisely this by a martingale limit argument based on Burkholder's inequality. Furthermore, under the stated conditions, it is also possible to show that, if $X(n)$ converges, its limit is necessarily a Nash equilibrium of Γ. Our proof then follows in roughly the same way as in the case of Theorem 1; for the details, we refer the reader to the appendix.

We close this section by noting that the summability condition (4.6) imposes a lower bound on the step-size exponent β that is different from the lower bound in Theorem 3. In particular, if $\beta = 1/2$, (4.6) cannot hold for any vanishing sequence of mixing factors $\varepsilon_n \downarrow 0$. Given that the innovation term U_i is bounded, we conjecture that this sufficient condition is not tight and can be relaxed further. We intend to address this issue in future work.

5 Conclusion and perspectives

The results of the previous sections show that no-regret learning via exponential weights enjoys appealing convergence properties in generic potential games. Specifically, in the semi-bandit case, the sequence of play converges to a Nash equilibrium with probability 1, and convergence to pure equilibria occurs at a quasi-exponential rate. In the bandit case, the same holds true for $\mathcal{O}(\varepsilon)$-equilibria if the algorithm is run with a positive mixing factor $\varepsilon > 0$; and if the algorithm is run with a decreasing mixing schedule, the sequence of play converges to an actual Nash equilibrium (again, with probability 1). In future work, we intend to examine the algorithm's convergence properties in other classes of games (such as smooth games), to explore the use of different regularizer functions to boost the algorithm's rate of convergence (some preliminary results are included in the supplement), and to examine the impact of asynchronicities and delays in the players' feedback/update cycles.

References

[1] James Hannan. Approximation to Bayes risk in repeated play. In Melvin Dresher, Albert William Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games, Volume III*, volume 39 of *Annals of Mathematics Studies*, pages 97–139. Princeton University Press, Princeton, NJ, 1957.

[2] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[3] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[4] Volodimir G. Vovk. Aggregating strategies. In *COLT '90: Proceedings of the 3rd Workshop on Computational Learning Theory*, pages 371–383, 1990.

[5] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

[6] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, September 2000.

[7] Yannick Viossat and Andriy Zapechelnyuk. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, March 2013.

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

[9] Noam Nisan, Tim Roughgarden, Eva Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[10] Samson Lasaulce and Hamidou Tembine. *Game Theory and Learning for Wireless Networks: Fundamentals and Applications*. Academic Press, Elsevier, 2010.

[11] Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40(4):479–519, July 2003.

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

[13] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[14] Avrim Blum, Mohammad Taghi Hajiaghayi, Katrina Ligett, and Aaron Roth. Regret minimization and the price of total anarchy. In *STOC '08: Proceedings of the 40th annual ACM symposium on the Theory of Computing*, pages 373–382. ACM, 2008.

[15] Avrim Blum and Yishay Mansour. Learning, regret minimization, and equilibria. In Noam Nisan, Tim Roughgarden, Eva Tardos, and V. V. Vazirani, editors, *Algorithmic Game Theory*, chapter 4. Cambridge University Press, 2007.

[16] Dean Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1):40–55, October 1997.

[17] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.

[18] Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)*, 62(5):32, 2015.

[19] Dylan J Foster, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems*, pages 4727–4735, 2016.

[20] Robert Kleinberg, Georgios Piliouras, and Éva Tardos. Load balancing without regret in the bulletin board model. *Distributed Computing*, 24(1):21–29, 2011.

[21] 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.

[22] Walid Krichene, Benjamin Drighès, and Alexandre M Bayen. Learning nash equilibria in congestion games. *arXiv preprint arXiv:1408.0017*, 2014.

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

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

[25] Michel Benaïm and Morris W. Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176, 1996.

[26] Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156, 1978.

[27] P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Probability and Mathematical Statistics. Academic Press, New York, 1980.

A Proofs on properties ε-HEDGE Algorithm

This section is devoted to the proof of Theorem 1. In Section 3, the recursion equation of Algorithm 1 depends on ε parameter. In this section, each player i can have its own mixing factor ε_i. This choice is done in order to simplify the notation in Sections 3 and 4. Thus the recursion equation of Algorithm 1 can be rewritten as

$$\begin{aligned} X_i(n) &= \varepsilon_i/|\mathcal{S}_i|\,\mathbf{1} + (1-\varepsilon_i)\,\Lambda_i(Y_i(n)), \\ Y_i(n+1) &= Y_i(n) + \gamma_n \hat{v}_i(n), \end{aligned} \tag{ε-Hedge}$$

To perform this, we introduce some other discrete-time process $(X_i^{\Lambda}(n))_{n\in\mathbb{N}}$ to help with this proof.

$$X_i^{\Lambda}(n) = \frac{1}{(1-\varepsilon_i)}(X_i(n) - \varepsilon_i/|\mathcal{S}_i|\mathbf{1})$$

By definition, we have $X_i(n) = \varepsilon_i/|\mathcal{S}_i|\mathbf{1} + (1-\varepsilon_i)X_i^{\Lambda}(n)$. Besides,

Observation 5. *For any $n \in \mathbb{N}$, $\|X(n) - X^{\Lambda}(n)\|_2 \leq N\epsilon$, where $\epsilon = max_{i\in\mathbb{N}}\frac{\epsilon_i}{\sqrt{|\mathcal{S}_i|}}$.*

Moreover, we can establish an other relation between these two discrete-time processes:

Lemma 6. *If $(X^{\Lambda}(n))_{n\in\mathbb{N}}$ converges to a Nash equilibrium x^*, then $(X(n))_{n\in\mathbb{N}}$, defined in (ε-Hedge), converges to a $\delta(\epsilon)$-Nash equilibrium with $\epsilon = \max_{i\in\mathcal{N}}\frac{\epsilon_i}{|\mathcal{S}_i|}$ and $\delta(\epsilon)$ such that if ϵ tends to 0, then $\delta(\epsilon)$ also tends to 0.*

Proof. Assume that $(X^{\Lambda}(n))_{n\in\mathbb{N}}$ converges to a point x^*. Then $(X(n))_{n\in\mathbb{N}}$ also converges to a limit $\hat{x}$. Observation 5 gives us that $\|X(n) - X^{\Lambda}(n)\|_2 \leq N\epsilon$, by continuity, $\|x^* - \hat{x}\|_2 \leq N\epsilon$.

We remain to show that if the two mixed strategy profiles p and p' are such that $\|p - p'\|_2 \leq \epsilon$ then $\|u_i(p) - u_i(p')\|_2 \leq \frac{\delta(\epsilon)}{2}$, for any player $i \in \mathcal{N}$.

$$\begin{aligned} |u_i(p) - u_i(p')| &= |\sum_{s_1\in\mathcal{S}_1}\cdots\sum_{s_N\in\mathcal{S}_N} u_i(s_1,\dots,s_N)(p_{1s_1}\cdots p_{Ns_N} - p'_{1s_1}\cdots p'_{Ns_N})| \\ &\leq \sum_{s_1\in\mathcal{S}_1}\cdots\sum_{s_N\in\mathcal{S}_N} u_i(s_1,\dots,s_N)|(p_{1s_1}\cdots p_{Ns_N} - p'_{1s_1}\cdots p'_{Ns_N})| \\ &\leq \sum_{s_1\in\mathcal{S}_1}\cdots\sum_{s_N\in\mathcal{S}_N} u_i(s_1,\dots,s_N)|(p_{1s_1}\cdots p_{Ns_N} - (p_{1s_1}-\epsilon)\cdots(p_{Ns_N}-\epsilon))| \\ &\leq \sum_{s_1\in\mathcal{S}_1}\cdots\sum_{s_N\in\mathcal{S}_N} u_i(s_1,\dots,s_N)\sum_{k=1}^{N}(-\epsilon)^k\binom{N}{k} \\ &= \frac{\delta(\epsilon)}{2} \end{aligned} \tag{A.1}$$

Moreover, observe that if $\epsilon \to 0$, then $\frac{\delta(\epsilon)}{2} \to 0$, and $|u_i(p) - u_i(p')| \to 0$.

Assume that $(X^{\Lambda}(n))_{n\in\mathbb{N}}$ converges to a Nash equilibrium x^*. By definition, for all $x_i \in X_i$, $i \in \mathbb{N}$ $u_i(x_i^*; x_{-i}^*) \geq u_i(x_i; x_{-i}^*)$. In addition, by combining the previous relations, we have $\|x^* - \hat{x}\|_2 \leq N\epsilon$ and $\|(x_i, x_{-i}^*) - (x_i^*, x_{-i}^*)\|_2 \leq N\epsilon \quad \forall x_i \in X_i$.

$$\begin{aligned} &u_i(\hat{x}_i;\hat{x}_{-i}) + \frac{\delta(\epsilon)}{2} \geq u_i(x_i^*; x_{-i}^*) \geq u_i(x_i; x_{-i}^*) \geq u_i(x_i;\hat{x}_{-i}) - \frac{\delta(\epsilon)}{2} \\ &u_i(\hat{x}_i;\hat{x}_{-i}) + \delta(\epsilon) \geq u_i(x_i;\hat{x}_{-i}) \text{ for all } x_i^* \in \mathcal{X}_i, i \in \mathcal{N} \end{aligned} \tag{A.2}$$

So, the latter equation corresponds to the definition of $delta(\epsilon)$-equilibrium. We can conclude that $(X(n))_{n\in\mathbb{N}}$ converges to a $\delta(\epsilon)$-Nash equilibrium with $\delta(\epsilon)$. □

The remainder of this section is devoted to prove the convergence results of ε-HEDGE Algorithm, and it is split into two parts according to the feedback assumption. The proof is based on Benaim's study of stochastic approximations [23]. We follow 4 points:

1. Show that X is an asymptotic pseudo trajectory of a continuous dynamics,
2. Show that the potential function of the game is strict Lyapunov function of the dynamics,
3. Show that X converges toward a rest point of the dynamics,
4. Show that if X converges toward a point it is a Nash Equilibrium.

For the first step, we focus on the sequences $(X_i(n))_{n\in\mathbb{N}}$ and its linear interpolation $(x_i(t))$. We also define the continuous variables, $x_{is}^{\Lambda}(t) = \frac{x_{is}(t)-\epsilon_i/|\mathcal{S}|_i}{1-\epsilon_i}$ and $y_{is}(t)$ such that $x_{is}^{\Lambda}(t) = \Lambda_i(y_{is}(t))$, with the same relations as the corresponding discret processes. We will prove that the sequences $(X_i(n))_{n\in\mathbb{N}}$ correspond to some classical family of stochastic approximation algorithms so-called *Approximate Robbins-Monro algorithm*.

Definition 7 (Approximate Robbins-Monro conditions)**.** The general stochastic approximation algorithm $x(n+1) = x(n) + \gamma_n(F(x(n)) + U_n + \beta_n)$ is said to be an *approximate Robbins-Monro algorithm* if:

- $F : \mathbb{R}^m \to \mathbb{R}^m$ is a continuous map;
- $U_n \in \mathbb{R}^m$ are perturbations and U_n is a martingale difference noise;
- $\{\gamma_n\}_{n\geq 1}$ is a given sequence of nonnegative numbers such that $\sum_k \gamma_k = \infty$ and $lim_{n\to\infty}\gamma_n = 0$;
- $lim_{n\to\infty} b_n = 0$ almost surely.

Then, we will prove that the interpolated process of the sequences $(X_i(n))_{n\in\mathbb{N}}$ is an asymptotic pseudo trajectory of the solutions of the following ordinary differential equation.

$$\begin{aligned}\dot{x}_{is} &= \big(x_{is} - |\mathcal{S}_i|^{-1}\varepsilon_i\big)\Big[v_{is}(x) - (1-\varepsilon_i)^{-1}\sum\nolimits_{s'\in\mathcal{S}_i}(x_{is'} - |\mathcal{S}_i|^{-1}\varepsilon_i)v_{is'}(x)\Big] \\ &= \big(1-\varepsilon_i\big)x_{is}^{\Lambda}\Big[v_{is}(x) - \sum\nolimits_{s'\in\mathcal{S}_i} x_{is'}^{\Lambda}v_{is'}(x)\Big]\end{aligned} \tag{A.3}$$

To proceed, recall the definition of the asymptotic pseudo-trajectory.

Definition 8 (Asymptotic Pseudo-trajectories)**.** Given a flow $\phi : \mathbb{R}\times M \to M, (n,x) \to \phi(n,x) = \phi_n(x)$ such that $\phi_0 =$ Identity and $\phi_{n+s} = \phi_n \circ \phi_s$, a continuous function $X : \mathbb{R} \to M$ is an asymptotic pseudo-trajectory if

$$lim_{n\to\infty} sup_{0\leq k\leq T} d((X(n+k), \phi_h(X(n)) = 0 \text{ for any } T > 0 \tag{A.4}$$

Proposition 9. *Suppose that Algorithm 1 is run with i) semi-bandit feedback satisfying (H1) and (H2); ii) a nonnegative mixing factor $\varepsilon \geq 0$; and iii) a step-size sequence of the form $\gamma_n \propto 1/n^{\beta}$ for some $\beta \in (1/q, 1]$. The interpolated process of the sequences $(X_i(n))_{n\in\mathcal{N}}$ is an asymptotic pseudo trajectory of the solutions of the following ordinary differential equation*

$$\dot{x}_{is} = \big(x_{is} - |\mathcal{S}_i|^{-1}\varepsilon_i\big)\Big[v_{is}(x) - (1-\varepsilon_i)^{-1}\sum\nolimits_{s'\in\mathcal{S}_i}(x_{is'} - |\mathcal{S}_i|^{-1}\varepsilon_i)v_{is'}(x)\Big], \tag{RD$_\varepsilon$}$$

Proof. The proof of Proposition 9 can be split into two parts. First, we will prove that the stochastic process $\{X_i(n)\}_{n\in\mathbb{N}}$ given by Algorithm 1 is an approximate Robbins-Monro algorithm. Second, we will conclude by applying some classical results in [23].

Observe that for any $i \in \mathcal{N}$, for any $s, s', s'' \in \mathcal{S}_i$, $x_{is}^{\Lambda} = \Lambda_{is}(y_i) = \frac{\exp(y_{is})}{\sum_{s\in\mathcal{S}_i}\exp(y_{is})}$, we have

$\frac{\partial\,\Lambda_{is}(y_i)}{\partial y_{is'}} = x_{is}^{\Lambda}(\mathbb{1}_{s=s'} - x_{is'}^{\Lambda})$, and $\frac{\partial^2\,\Lambda_{is}(y_i)}{\partial y_{is'}\partial y_{is''}} = x_{is}^{\Lambda}\left(\mathbb{1}_{s=s'=s''} - \mathbb{1}_{s=s'}x_{is''}^{\Lambda} - x_{is'}^{\Lambda}(\mathbb{1}_{s=s''} + \mathbb{1}_{s'=s''} - 2x_{is''}^{\Lambda})\right)$.

Using Taylor's Remainder Theorem, let us rewrite the equation $X_{is}(n+1) = \frac{\epsilon_i}{|\mathcal{S}_i|} + (1-\epsilon_i)\,\Lambda_{is}(Y_i(n+1))$ as

$$\begin{aligned} X_{is}(n+1) &= \frac{\epsilon_i}{|\mathcal{S}_i|} + (1-\epsilon_i)\,\Lambda_{is}(Y_i(n) + \gamma_n \hat{v}_i(n)) \\ &= \frac{\epsilon_i}{|\mathcal{S}_i|} + (1-\epsilon_i)\,\Lambda_{is}(Y_i(n)) + (1-\epsilon_i)\gamma_n \left(\nabla\,\Lambda_{is}^{\mathrm{T}}(Y_i(n))\hat{v}_i(n) + \frac{1}{2}\gamma_n \hat{v}_i^{\mathrm{T}}(n)\,\mathrm{Hess}\,\Lambda_{is}(\psi_i(n))\hat{v}_i(n)\right) \\ &= X_{is}(n) + (1-\epsilon_i)\gamma_n \left(\nabla\,\Lambda_{is}^{\mathrm{T}}(Y_i(n))\hat{v}_i(n) + \frac{\gamma_n}{2}\hat{v}_i^{\mathrm{T}}(n)\,\mathrm{Hess}\,\Lambda_{is}(\psi_i(n))\hat{v}_i(n)\right) \end{aligned} \tag{A.5}$$

where $\nabla\,\Lambda_{is}$ is the gradient vector of Λ_{is}, $\nabla\,\Lambda_{is}^{\mathrm{T}}$ is its transposed, $\mathrm{Hess}\,\Lambda_{is}$ is the Hessian matrix of Λ_{is}, and $\psi_i(n)$ is in the line segment going out from $Y_i(n)$ to the point $Y_i(n+1)$.

Since $\frac{\partial\,\Lambda_{is}(y_i)}{\partial y_{is'}} = x_{is}^{\Lambda}(\mathbb{1}_{s=s'} - x_{is'}^{\Lambda})$, this equality could be written as

$$\begin{aligned} X_{is}(n+1) &= X_{is}(n) + (1-\epsilon_i)\gamma_n \left(\nabla\,\Lambda_{is}^{\mathrm{T}}(Y_i(n))\hat{v}_i(n) + \frac{\gamma_n}{2}\hat{v}_i^{\mathrm{T}}(n) Hess(\Lambda_{is})(\psi_i(n))\hat{v}_i(n)\right) \\ &= X_{is}(n) + (1-\epsilon_i)\gamma_n \left(\nabla\,\Lambda_{is}^{\mathrm{T}}(Y_i(n))v_i(X(n)) + \nabla\,\Lambda_{is}^{\mathrm{T}}(Y_i(n))(\hat{v}_i(n) - v_i(X(n))) + \gamma_n a_n\right) \end{aligned} \tag{A.6}$$

where $a_n = \frac{1}{2}\hat{v}_i^{\mathrm{T}}(n) Hess\,\Lambda_{is}(\psi_i(n))\hat{v}_i(n)$.

Next, we focus on $\gamma_n a_n$ (corresponding to parameter b_n in Definition 7). Since $\frac{\partial^2\,\Lambda_{is}(y_i)}{\partial y_{is'}\partial y_{is''}} = x_{is}^{\Lambda}\left(\mathbb{1}_{s=s'=s''} - \mathbb{1}_{s=s'}x_{is''}^{\Lambda} - x_{is'}^{\Lambda}(\mathbb{1}_{s=s''} + \mathbb{1}_{s'=s''} - 2x_{is''}^{\Lambda})\right)$, all components of $Hess\,\Lambda_{is}(\psi_i(n))$ are bounded. So, the limit of $\gamma_n a_n$ (when $n \to \infty$) depends on the limit of $||\hat{v}_{is}(n)||^2$.

Let $E_{is,n}$ be the event $||\hat{v}_{is}(n)||^2 \geq n^{\alpha}$ for $\frac{1}{q} < \alpha < \frac{1}{p}$, with q defined in Hypothesis (H2) and p such that $\gamma_n = \mathcal{O}(1/n^{\frac{1}{p}})$.

Hypothesis (H2) gives us that

$$\sum_{n=0}^{\infty} \mathbb{P}(E_{is,n}) = \sum_{n=0}^{\infty} \mathbb{P}(\|\hat{v}_{is}(n)\|^2 \geq n^{\alpha} \,|\, \mathcal{F}_{n-1}) = \sum_{n=0}^{\infty} \mathcal{O}(\frac{1}{n^{q\alpha}}) < \infty \tag{A.7}$$

The Borel-Cantelli lemma gives us that $E_{is,n}$ is true for only a finite number of $n \in \mathbb{N}$. Therefore for $n > max_{m\in\mathbb{N}}\{m; \exists i \in \mathcal{N}, s \in \mathcal{S}_i, E_{is,m} \text{ is true}\}$, $||\hat{v}_{is}(n)||^2 < n^{\alpha}$. By assumption, $\gamma_n = o(n^b)$ for any $b > -1/p$. In particular, $\gamma_n = o(n^{-\alpha})$ so $\lim_{n\to\infty} a_n\gamma_n = 0$.

$$\begin{aligned} X_{is}(n+1) = X_{is}(n) + (1-\epsilon_i)\gamma_n \left(X_{is}^{\Lambda}(n)\left(v_{is}(X(n)) - \sum_{s'\in\mathcal{S}_i} v_{is'}(X(n))X_{is'}^{\Lambda}(n)\right)\right) \\ + (1-\epsilon_i)\gamma_n \left(X_{is}^{\Lambda}(n)(\hat{v}_{is}(n) - v_{is}(X(n)) - \sum_{s'\in\mathcal{S}_i} X_{is'}^{\Lambda}(n)[\hat{v}_{is'}(n) - v_{is'}(X(n))]) + \gamma_n a_n\right) \end{aligned}$$

Let $U_{is,n} = (1-\epsilon_i)X_{is}^{\Lambda}(n)(\hat{v}_{is}(n) - v_{is}(X(n)) - \sum_{s'\in\mathcal{S}_i} X_{is'}^{\Lambda}(n)[\hat{v}_{is'}(n) - v_{is'}(X(n))])$.

Recall that (3.2a) and (3.2b) can be deduced from Hypotheses (H1) (H2). We get

1. $\mathbb{E}[U_{is,n}|\mathcal{F}_{n-1}] = 0$ for all n
2. $\mathbb{E}[\|U_{is,n}\|^2] < \infty$ for all n

So, $U_{is,n}$ is a martingale difference noise.

Since the function which to x_{is} associates $\big(1-\varepsilon\big)x_{is}^{\Lambda}\big[v_{is}(x) - \sum_{s'\in\mathcal{S}_i} x_{is'}^{\Lambda}v_{is'}(x)\big]$ is a continuous map, we can conclude that the stochastic process $\{X_i(n)\}_{n\in\mathbb{N}}$ is an approximate Robbins-Monro algorithm.

Second, Remark 4.5 and Propositions 4.2 and 4.1 of [23] allow us to conclude that the interpolated process of the sequences $(X_i(n))_{n\in\mathbb{N}}$ is an asymptotic pseudo trajectory of the solutions of ODE (RD$_\varepsilon$).

□

Dynamics (RD$_\varepsilon$) can be viewed as a ε-perturbed variant of the replicator dynamics. that readily extends to the ε-perturbed variant (RD$_\varepsilon$). Moreover, the replicator dynamics have some good property in potential games. The next theorem expresses this property in our context:

Proposition 10. *Let Γ be a generic potential game. The potential function f of Γ is a strict increasing Lyapunov function of the flow inducted by the dynamics* (RD$_\varepsilon$).

Proof. We consider the variation of f. We have

$$\begin{aligned}
\dot{f}(x) &= \sum_{i\in\mathcal{N}}\sum_{s\in\mathcal{S}_i}\frac{\partial f}{\partial x_{is}}(x)\dot{x}_{is} \\
&= \sum_{i\in\mathbb{N}} v_i^{\mathrm{T}}(x(t))\dot{x}_i(t) \\
&= \sum_{i\in\mathbb{N}}\sum_{s\in\mathcal{S}_i} v_{is}(x(t))\dot{x}_{is}(t) \\
&= \sum_{i\in\mathbb{N}}\sum_{s\in\mathcal{S}_i}(1-\varepsilon_i)v_{is}(x(t))x^\Lambda_{is}(t)\left(v_{is}(x(t)) - \sum_{s'\in\mathcal{S}_i} v_{is'}(x(t))x^\Lambda_{is'}(t)\right) \\
&= \sum_{i\in\mathbb{N}}\sum_{s\in\mathcal{S}_i}\sum_{s'\in\mathcal{S}_i,s'>s}(1-\varepsilon_i)x^\Lambda_{is}(t)x^\Lambda_{is'}(t)[v_{is}(x(t)) - v_{is'}(x(t))]^2
\end{aligned}$$

The second equation is obtained by definitions of potential game and dynamic (RD$_\varepsilon$) (definition of $\dot{x}_{is}$). From the latter equation, we can conclude that $\dot{f}(x) \geq 0$.

Now we show that the rest points x of dynamics (RD$_\varepsilon$) are such that $\dot{f}(x) = 0$. Since $\dot{f}(x) = \sum_{i\in\mathcal{N}}\sum_{s\in\mathcal{S}_i}\frac{\partial f}{\partial x_{is}}(x)\dot{x}_{is}$, $\dot{f}(x) = 0$ when x is a rest point ($\dot{x} = 0$).

Conversely, we will prove that $\dot{f}(x) = 0$ implies that x is a rest point of dynamics (RD$_\varepsilon$).

Observe that $\dot{f}(x) = 0$ implies that for all players i in $\mathcal{N}$, and pure strategies s and s' in $\mathcal{S}_i$ we have $x^\Lambda_{is} = 0$ or $x^\Lambda_{is'} = 0$ or $v_{is}(x) = v_{is'}(x)$.

Since the dynamics (RD$_\varepsilon$) is $\dot{x}_{is} = (1-\varepsilon_i)x^\Lambda_{is}\left(v_{is}(x) - \sum_{s'\in\mathcal{S}_i} v_{is'}(x)x^\Lambda_{is'}\right)$, if for all players i in $\mathcal{N}$, and pure strategies s and s' in $\mathcal{S}_i$ we have $x^\Lambda_{is} = 0$ or $x^\Lambda_{is'} = 0$, then we can deduce $\dot{x}_{is} = 0$. Otherwise, if $v_{is}(x) = v_{is'}(x)$ for all s and s' in $\mathcal{S}_i$ such that $x^\Lambda_{is} \neq 0$ and $x^\Lambda_{is'} \neq 0$, we have then $\dot{x}_{is} = 0$

To conclude, f is increasing, and its derivative is null if and only if it is evaluated on a rest point of the dynamics (RD$_\varepsilon$). f is a strict increasing Lyapunov function of the dynamics (RD$_\varepsilon$). □

Proposition 11. *Let Γ be a generic potential game. Suppose that Algorithm 1 is run with i) semi-bandit feedback satisfying (H1) and (H2); ii) a nonnegative mixing factor $\varepsilon \geq 0$; and iii) a step-size sequence of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/q, 1]$. The interpolated process of the sequences $(X_i(n))_{n\in\mathbb{N}}$ converges to a rest point of RD$_\varepsilon$.*

Proof. We showed that under theses assumptions, $(X_i(n))_{n\in\mathbb{N}}$ is a pseudo asymptotic trajectory of the flow induced by the dynamics RD$_\varepsilon$.

We now show that RD$_\varepsilon$ has a finite number of rest points. When $\forall i \in \mathcal{N}, \varepsilon_i = 0$, we have $x^\Lambda = x$. Thus, we obtain from the previous proof that x is a rest point of RD$_\varepsilon$ if and only if

$$\sum_{i\in\mathbb{N}}\sum_{s\in\mathcal{S}_i}\sum_{s'\in\mathcal{S}_i,s'>s}(1-\varepsilon_i)x_{is}(t)x_{is'}(t)[v_{is}(x(t)) - v_{is'}(x(t))]^2 = 0 \tag{A.8}$$

So x is a rest point of RD_ε if and only if $v_{is}(x(t)) = v_{is'}(x(t)) \quad \forall s, s' \in supp(x_i)$. Therefore x is a rest point of RD_ε if and only if it is a restricted equlibrium. The game is finite so they are a finite number of game restrictions. Each restricted game is a finite, generic, potential game so it has a finite number of Nash-equilibrium.

Now we address the other case : $\exists i \in \mathcal{N}, \varepsilon_i \neq 0$. Thus, x is a rest point of RD_ε if and only if $v_{is}(x(t)) = v_{is'}(x(t)) \quad \forall s, s' \in supp(x_i^\Lambda)$. Let $supp_i$ be the cardinal of $supp(x_i^\Lambda)$. For each player i, we have $supp_i - 1$ unknown coordinates corresponding to the x_i^Λ non-zero coordinates. Thus, $supp_i - 1$ independent equations corresponding to $v_{is}(x(t)) = v_{is'}(x(t)) \quad \forall s, s' \in supp(x_i^\Lambda)$ because the game is generic. Such a system has only an unique solution. Therefore for each possible support of x there is only one rest point, and dynamics RD_ε has a finite number of rest points.

Corollary 6.6 of [23] allows to conclude that the continous-time process $x_i(t)$ converges to a rest point of RD_ε. □

In order to prove Point 1 of Theorem 1, we need to have the following technical result. This lemma is about the properties on rationality properties (such as comparaisons among strategies corresponding to the elimination of dominated strategies).

Lemma 12. *Suppose that Algorithm 1 is run with i) semi-bandit feedback satisfying (H1) and (H2); ii) a nonnegative mixing factor $\varepsilon \geq 0$; and iii) a step-size sequence of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/q, 1]$. If there exists some $a > 0$ such that $v_{s'}(x) - v_s(x) \geq a$ for all $x \in \mathcal{X}$ then for all $c \in (0, a)$, there exists some n_0 such that $Y_{s'}(n+1) - Y_s(n+1) \geq c \sum_{k=1}^n \gamma_k \quad$ for all $n \geq n_0$ (a.s.).*

Proof. Let $\zeta_k = \hat{v}_{s'}(k) - v_{s'}(X(k)) - [\hat{v}_s(k) - v_s(X(k))]$. By assumption there exists $a > 0$ such that $v_{s'}(X) - v_s(X) \geq a$ for all $X \in \mathcal{X}$. Then,

$$\begin{aligned} Y_{s'}(n+1) - Y_s(n+1) &= Y_{s'}(1) - Y_s(1) + \sum_{k=1}^n \hat{v}_{s'}(k) - \hat{v}_s(k) \\ &= Y_{s'}(1) - Y_s(1) + \sum_{k=1}^n \gamma_k \left[v_{s'}(X(k)) - v_s(X(k))\right] + \sum_{k=1}^n \gamma_k \zeta_k \quad \text{(A.9)} \\ &\geq Y_{s'}(1) - Y_s(1) + \sum_{k=1}^n \gamma_k \left[a + \frac{\sum_{k=1}^n \gamma_k \zeta_k}{\sum_{k=1}^n \gamma_k}\right]. \end{aligned}$$

Now we will prove that $\frac{\sum_{k=1}^n \gamma_k \zeta_k}{\sum_{k=1}^n \gamma_k} \to 0$.

The reformulation of Hypothesis (H1) gives :

$$\begin{aligned} \mathbb{E}[\zeta_k | \mathcal{F}_{k-1}] &= \mathbb{E}[\xi_{s'}(k) + v_{s'}(s(k)) - v_{s'}(X(k)) - \xi_s(k) - v_s(s(k)) + v_s(X(k)) | \mathcal{F}_{k-1}] \\ &= \mathbb{E}[\xi_{s'}(k) - \xi_s(k) | \mathcal{F}_{k-1}] + \mathbb{E}[v_{s'}(s(k)) | \mathcal{F}_{k-1}] - v_{s'}(X(k)) - \mathbb{E}[v_s(s(k)) | \mathcal{F}_{k-1}] + v_s(X(k)) \\ &= 0 \end{aligned} \quad \text{(A.10)}$$

ζ_k is $\mathcal{F}_k$-measurable, meaning that it is fully determined by the information of $\mathcal{F}_k$.

With $S_n = \sum_{k=1}^n \gamma_k \zeta_k$, it follows that

$$\mathbb{E}[S_k | \mathcal{F}_{k-1}] = \gamma_k \mathbb{E}[\zeta_k | \mathcal{F}_{k-1}] + \mathbb{E}[S_{k-1} | \mathcal{F}_{k-1}] = S_{k-1} \quad \text{(A.11)}$$

Therefore $\{S_n = \sum_{k=1}^n \gamma_k \zeta_k, \mathcal{F}_n, n \geq 1\}$ is a martingale, in addition as γ_t depends only on n and it is positive, $U_n = \sum_{k=1}^n \gamma_k$ is a nondecreasing sequence of positive random variable such that U_n is $\mathcal{F}_{n-1}$-measurable for each n. In addition $\beta \leq 1$ gives us that $lim_{n \to \infty} U_n = \infty$.

We focus now on proving the last hypothesis of Theorem 2.18 of [27], i.e., $\sum_{k=1}^\infty \frac{\mathbb{E}(\|\gamma_k \zeta_k\|^2 | \mathcal{F}_{k-1})}{U_k^2} < \infty$. First we show that $\mathbb{E}[\|\zeta(k)\|^2 | \mathcal{F}_{k-1}] \leq 4\sigma^2$:

$$\begin{aligned} \mathbb{E}[\|\zeta(k)\|^2 | \mathcal{F}_{k-1}] &= \mathbb{E}[\|\hat{v}_{s'}(k) - v_{s'}(X(k)) - [\hat{v}_s(k) - v_s(X(k))]\|^2 | \mathcal{F}_{k-1}] \\ &= 2\mathbb{E}[\|\hat{v}_{s'}(k) - v_{s'}(X(k))\|^2 | \mathcal{F}_{k-1}] + 2\mathbb{E}[\|\hat{v}_s(k) - v_s(X(k))\|^2 | \mathcal{F}_{k-1}] \\ &\leq 4\sigma^2 \end{aligned} \quad \text{(A.12)}$$

according to Equation (3.2b). Second, γ_t is decreasing so $U_n = \sum_{k=1}^n \gamma_k \geq n\gamma_n$ and $U_n^{-2} \leq \frac{1}{n^2\gamma_n^2}$. Therefore

$$\sum_{k=1}^{\infty} \frac{\mathbb{E}(\|\gamma_k \zeta_k\|^2 | \mathcal{F}_{k-1})}{U_k^2} < \sum_{k=1}^{\infty} \frac{\gamma_k^2 4\sigma^2}{k^2\gamma_k^2} = 4\sigma^2 \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty \tag{A.13}$$

Therefore all hypotheses are fulfilled and

$$\lim_{n\to\infty} \frac{\sum_{k=1}^n \gamma_k \zeta_k}{\sum_{k=1}^n \gamma_k} = 0 \quad (a.s.) \tag{A.14}$$

So for all $c \in (0, a)$ there exist some t_0 such that $-\frac{\sum_{k=1}^n \gamma_k \zeta_k}{\sum_{k=1}^n \gamma_k} \leq \frac{Y_{s'}(1) - Y_s(1)}{\sum_{k=1}^n \gamma_k} + a - c$ for all $t > t_0$. Putting that in (A.9) we have :

$$Y_{s'}(n+1) - Y_s(n+1) \geq c \sum_{k=1}^{n} \gamma_k \quad \text{for all } t \geq t_0 \text{ (a.s.)} \tag{A.15}$$

□

Now, we will prove the convergence of the discrete-time process $(X_i(n))_{n\in\mathbb{N}}$.

Theorem 1 ($Part\ 1$)**.** *Let Γ be a generic potential game. Suppose that Algorithm 1 is run with i) semi-bandit feedback satisfying (H1) and (H2); ii) a nonnegative mixing factor $\varepsilon \geq 0$; and iii) a step-size sequence of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/q, 1]$. $X(n)$ converges (a.s.) to a δ-equilibrium of Γ with $\delta \equiv \delta(\varepsilon) \to 0$ as $\varepsilon \to 0$.*

Proof. When $X(n)$ converges, $X^\Lambda(n)$ also converges (by definition). Applying Lemma 6, it suffices to prove that if $X^\Lambda(n)$ converges to x^* then x^* is Nash Equilibrium. We prove by contradiction that $(X^\Lambda(n))_{n\in\mathbb{N}}$ converges to x^* a Nash Equilibrium. Assume that x^* is not a Nash Equilibrium. By definition, we have

$$\exists i \in \mathcal{N}, \exists s' \in \mathcal{S}_i, s' \notin supp(x_i^*), s.t., v_{is'}(x^*) > v_{is}(x^*), \forall s \in supp(x_i^*).$$

By continuity of utility u, there is a neighborhood U of x^* and $a > 0$ such that:

$\exists i \in \mathcal{N}, \exists s' \in \mathcal{S}_i, s' \notin supp(x_i^*), s.t., v_{is'}(X^\Lambda) - v_{is}(X^\Lambda) > a, \forall s \in supp(x_i^*), X^\Lambda \in U$ For ε small enough and for all n big enough, $X(n) \in U$ because $\|X(n) - x^*\| \leq \|X(n) - X^\Lambda(n)\| + \|X^\Lambda(n) - x^*\| \leq N\varepsilon + \|X^\Lambda(n) - x^*\|$. So, $\exists i \in \mathcal{N}, \exists s' \in \mathcal{S}_i, s' \notin supp(x_i^*), s.t., v_{is'}(X(n)) - v_{is}(X(n)) > a, \forall s \in supp(x_i^*)$ Using Lemma 12, for n_0 big enough and $n \geq n_0$:

$$Y_{is'}(n) - Y_{is}(n) \geq C + b \sum_{t=n_0}^{n-1} \gamma_n \to \infty \tag{A.16}$$

Thus $\frac{X_{is'}(n)}{X_{is}(n)} = \exp\left(Y_{is'}(n) - Y_{is}(n)\right) \to \infty$.

$X_{is}(n) \to 0$ and s is not in the support of x_i^* which is a contradiction. Therefore x^* is a Nash Equilibrium and $\hat{x}$ is a $\delta(\varepsilon)$-Nash Equilibrium. □

Finally, we will prove the convergence rate of the discrete-time process $(X_i(n))_{n\in\mathbb{N}}$ corresponding to Part 2 of Theorem 1.

Theorem 1 ($Part\ 2$)**.** *Let Γ be a generic potential game and suppose that Algorithm 1 is run with i) semi-bandit feedback satisfying (H1) and (H2); ii) a nonnegative mixing factor $\varepsilon \geq 0$; and iii) a step-size sequence of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/q, 1]$. . Then If $\lim_{n\to\infty} X(n)$ is an ε-pure state of the form $\hat{x}_i = \varepsilon/|\mathcal{S}_i|\mathbf{1} + (1-\varepsilon)e_{\hat{s}_i}$ for some $\hat{s} \in \mathcal{S}$, then $\hat{s}$ is a.s. a strict equilibrium of Γ and convergence occurs at a quasi-exponential rate:*

$$X_{i\hat{s}_i}(n) \geq 1 - \varepsilon - be^{-c\sum_{k=1}^n \gamma_k} \quad \textit{for some positive } b, c > 0. \tag{4.4}$$

Proof. We will focus on the sequence $(X_i^\Lambda(n))_{n\in\mathbb{N}}$ and its limit x^*. From the proof of Theorem 1 (Part 1), x^* is a Nash equilibrium. By continuity of u, there is a neighborhood U of x^* and $a' > 0$ such that:

$\forall i \in \mathcal{N}, \forall s' \in \mathcal{S}_i, s' \neq s_i^*, v_{is_i^*}(x) - v_{is'}(x) > a', x \in U.$

Therefore, for ϵ small enough and for all n big enough, $X(n) \in U$ because $\|X(n) - x^*\| \leq \|X(n) - X^\Lambda(n)\| + \|X^\Lambda(n) - x^*\| \leq N\epsilon + \|X(n) - x^*\|$. So, $\exists i \in \mathbb{N}, \exists s' \in \mathcal{S}_i, s' \notin supp(x_i^*), s.t., v_{is'}(X^\Lambda(n)) - v_{is}(X^\Lambda(n)) > a, \forall s \in supp(x_i^*)$ Using Lemma 12 for n_0 big enough:

$$Y_{is^*}(n) - Y_{is}(n) \geq C + b \sum_{k=n_0}^{n-1} \gamma_k$$

So, by computation, we can deduce that

$$\sum_{s \in \mathcal{S}_i, s \neq \hat{s}} \exp(Y_{is}(n) - Y_{i\hat{s}}(n)) \leq \sum_{s \in \mathcal{S}_i, s \neq \hat{s}} \exp(-C_{n_0} - \sum_{k=n_0}^{n-1} \gamma_k a) \leq C \exp(-\sum_{k=n_0}^{n-1} \gamma_k a)$$

where we set $C = |\mathcal{S}_i| \exp(-C_{t_0})$. Now, we will focus on $x_{\hat{s}}^\Lambda(n)$.

$$\begin{aligned} X_{i\hat{s}}^\Lambda(n) &= \frac{\exp(Y_{i\hat{s}}(n))}{\sum_{s \in \mathcal{S}} \exp(Y_{is}(n))} = \frac{1}{\sum_{s \in \mathcal{S}_i} \exp(Y_{is}(n) - Y_{i\hat{s}}(n))} \\ &= \frac{1}{1 + \sum_{s \in \mathcal{S}_i, s \neq \hat{s}} \exp(Y_{is}(n) - Y_{i\hat{s}}(n))} \\ &\geq \frac{1}{1 + C \exp(-\sum_{k=n_0}^{n} \gamma_k a)} \end{aligned}$$

The first equation is due to the relation $\frac{\exp a}{\exp b} = \frac{1}{\exp a - b}$. The second equation is obtained by the fact that $\exp(\exp(Y_{i\hat{s}}(n) - Y_{i\hat{s}}(n))) = 1$. Since for any $z > 0$, $\frac{1}{1+z} \geq 1 - z$, we obtain $1 - X_{i\hat{s}}^\Lambda(n) \leq C \exp(-\sum_{k=n_0}^{n-1} \gamma_k a)$. And, the theorem holds since $|X_{i\hat{s}}^\Lambda(n) - X_{i\hat{s}}(n)| \leq \varepsilon$. □

B Convergence with bandit feedback

By reviewing the proof of Theorem 1 in the previous appendix, there are two components that we need to establish for Theorem 4 and which do not hold for a feedback sequence with possibly unbounded variance:

Claim 1. The sequence of play $X(n)$ is an asymptotic pseudotrajectory (APT) of the replicator dynamics (RD).

Claim 2. If $X(n)$ converges, its limit is a Nash equilibrium (a.s.).

Proof of Claim 1. To show that $X(n)$ is an APT of (RD), let $a_n = \frac{1}{2} \hat{v}_i^\top(n) \operatorname{Hess} \Lambda_{is} \hat{v}_i(n)$ be as in the proof of Proposition 9, with $\hat{v}_i$ taken from the bandit estimator (4.2). By (4.2) and the proof of Theorem 3, it follows that $\gamma_n a_n = \mathcal{O}(\gamma_n / \varepsilon_n^2)$. Hence, by Remark 4.5 in [23, p. 17] and the decay rate assumption (4.6) which implies that $\gamma_n a_n = \mathcal{O}(\gamma_n / \varepsilon_n^2) \to 0$, it suffices to show that (A.6) constitutes an APT of (RD) when the term containing a_n is dropped.

To that end, by Proposition 4.1 in [23], it suffices to show that the innovation term $W_n = \nabla \Lambda_{is}^{\mathrm{T}}(\hat{v}_i(n) - v_i(X(n)))$ of (A.6) satisfies Benaïm's summability condition

$$\lim_{n \to \infty} \sup \left\{ \left\| \sum_{\ell=n}^{k-1} \gamma_\ell W_\ell \right\| : k = n+1, \dots, \sup\{\ell \geq 0 : \theta_n + T \geq \theta_\ell\} \right\} = 0 \tag{B.1}$$

for all $T > 0$. To proceed, let $m(t) = \sup \ell \geq 0 t \geq \theta_\ell$ and note that Burkholder's inequality gives:

$$\mathbb{E}\left[\sup_{n \leq k \leq m(\theta_n + T)} \left\| \sum_{\ell=n}^{k-1} \gamma_\ell W_\ell \right\|^2 \right] \leq C \, \mathbb{E}\left[\left(\sum_{\ell=n}^{m(\theta_n+T)} \gamma_\ell^2 \|W_\ell^2\| \right) \right], \tag{B.2}$$

for some universal constant $C > 0$. We then obtain the following string of inequalities:

$$\begin{aligned}\mathbb{E}\left[\sup_{n\leq k\leq m(\theta_n+T)}\left\|\sum_{\ell=n}^{k-1}\gamma_{\ell+1}W_{\ell+1}\right\|^2\right] &\leq C\,\mathbb{E}\left[\sum_{\ell=n}^{m(\theta_n+T)}\gamma_\ell\times\sum_{\ell=n}^{m(\theta_n+T)}\gamma_\ell^2\|W_\ell\|^2\right]\\ &\leq CT\,\mathbb{E}\left[\sum_{\ell=1}^{m(\theta_n+T)}\gamma_\ell^2\|W_\ell\|^2\right]\\ &\leq C'T\sum_{\ell=1}^{m(\theta_n+T)}\frac{\gamma_\ell^2}{\varepsilon_\ell^2},\end{aligned}\tag{B.3}$$

where C' is a positive constant and we used the fact that $W_\ell = \mathcal{O}(1/\varepsilon_\ell^2)$ by the definition of the bandit estimator (4.2). Since $\gamma_\ell/\varepsilon_\ell$ is assumed square summable, our claim follows as in the proof of Proposition 4.2 in [23]. □

Proof of Claim 2. Following the proof of Lemma 12, it suffices to show that the last hypothesis of Theorem 2.18 in [27] is satisfied, i.e. $\sum_{k=1}^{\infty}\frac{\mathbb{E}(\|\gamma_k\zeta_k\|^2|\mathcal{F}_{k-1})}{U_k^2} < \infty$. However, with $\|\hat{v}(k) - v(X(k))\| = \mathcal{O}(1/\varepsilon_k)$ and $\sum_{n=1}^{\infty}\gamma_n^2/\varepsilon_n^2$, our claim follows by repeating the same steps taken to establish (A.13). □

C Following the ε-regularized leader

In this section, we will focus on class of following the regularized leader algorithms called ε-FoReL algorithms corresponding to a generalization of ε-Hedge algorithm. The difference is the way to update the player's mixed strategies. At each round n, player i update its mixed strategy $X_i(n)$ as follows:

$$X_i(n) = \frac{\varepsilon_i}{|\mathcal{S}_i|} + (1-\varepsilon_i)Q_i(Y_i(n)) \tag{C.1}$$

where $Q_i(y) = arg\,max_{x\in\mathcal{X}_i}(\langle y, x\rangle - h_i(x))$.

The mixed strategy of players are updated in order to maximize the player's expected utility ($\langle y, x\rangle$) regularized by a *penalty* function. (corresponding to maps h_i).

We introduce *penality* functions $h_i : \mathcal{X}_i \to \mathbb{R}$. In order to have an unique solution, the penalty function needs to be convex, stronger conditions defined a penalty function.

Definition 13 (Penalty function)**.** Let $\mathcal{C}$ be a compact convex subset of a finite-dimensional normed space $\mathcal{V}$. We say that $h : \mathcal{C} \to \mathbb{R}$ is a *penalty function* (or *regularizer*) on $\mathcal{C}$ if:

1. h is continuous,
2. h is strongly convex, i.e., there exists some $K > 0$ such that

$$h(tx + (1-t)x') \leq th(x) + (1-t)h(x') - \frac{1}{2}Kt(1-t)\|x' - x\|^2 \tag{C.2}$$

for all $x, x' \in \mathcal{C}$ and all $t \in [0, 1]$.

Moreover, in order to prove the convergence property, we add four hypotheses that for each player $i \in \mathcal{N}$, function h_i needs to satisfy:

1. h_i is steep, i.e. $\|\nabla(h_i(x_i))\| \to \infty$ when $x_i \to \partial\mathcal{X}_i$ where $\partial\mathcal{X}_i$ is the boundary of $\mathcal{X}_i$,
2. h_i' is continuous on $(0, \infty)$,
3. $\|\mathrm{Hess}(Q_i)\|_\infty < \infty$, where $\mathrm{Hess}(Q_i)$ is the Hessian matrix of Q_i,
4. h_i is decomposable, i.e. $h_i(x_i) = \sum_{s_i\in\mathcal{S}_i}\theta_i(x_{is_i})$.

The first hypothesis makes sure that x is always in the interior of $\mathcal{X}$ although it can convergence to the boundary where pure equilibria belong. The third hypothesis also implies that $\|Jac(Q_i)\|_\infty < \infty$, where $Jac(Q_i)$ is the Jacobian matrix of Q_i.

The following the regularized leader algorithm can be written as follows:

Algorithm 3 ε-Regularized Leader Algorithm with generic feedback

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

for $n = 1, 2, \dots$ **do**
  for every player $i \in \mathcal{N}$ simultaneously **do**
    set mixed strategy: $X_i \leftarrow \frac{\varepsilon_i}{|\mathcal{S}_i|} + (1-\varepsilon_i) Q_i(Y_i)$
      where $Q_i(y) = arg\,max_{x \in \mathcal{X}_i}(\langle y, x\rangle - h_i(x))$.
    choose action $s_i \sim X_i$;
    acquire estimate $\hat{v}_i$ of realized payoff vector $v_i(s_i; s_{-i})$;
    update scores: $Y_i \leftarrow Y_i + \gamma_n \hat{v}_i$;
  end for
end for

Mathematically, Algorithm 3 represents the recursion

$$\begin{aligned} X_i(n) &= \varepsilon_i/|\mathcal{S}_i|\,\mathbf{1} + (1-\varepsilon_i) Q_i(Y_i(n)), \\ Y_i(n+1) &= Y_i(n) + \gamma_n \hat{v}_i(n), \end{aligned} \tag{ε-Reg}$$

We start by presenting a possible penalty function. A classical example is the *Gibbs entropy*:

$$h_i(x) = \sum_{s_i \in \mathcal{S}_i} x_{is_i} \log(x_{is_i}) \quad \text{for all } i \in \mathcal{N}. \tag{C.3}$$

Observe that $\{h_i\}_{i\in\mathcal{N}}$ verify all the conditions: they are continuous decomposable, strongly convex, and steep ($\nabla(h_i(x_i))_{s_i} = 1 + \log(x_{is_i})$). By classical computation, this corresponding choice map is the previously studied *logit map*:

$$Q_i(Y(n)) = \left(\frac{\exp(Y_{is}(n))}{\sum_{s' \in \mathcal{S}_i} \exp(Y_{is'}(n))} \right)_{s \in \mathcal{S}_i} = \Lambda_i(Y_i(n)) \tag{C.4}$$

We can adapt our main result (Theorem 1) for the convergence of (ε-Hedge) for Algorithm 3. In the context of semi-bandit feedback, Theorem 1 can be generalized to the following theorem:

Theorem 14. *Let Γ be a generic potential game and suppose that Algorithm 3 is run with i) semi-bandit feedback satisfying (H1) and (H2); ii) a nonnegative mixing factor $\varepsilon \geq 0$; and iii) a step-size sequence of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/q, 1]$. Then:*

1. *$X(n)$ converges (a.s.) to a δ-equilibrium of Γ with $\delta \equiv \delta(\varepsilon) \to 0$ as $\varepsilon \to 0$.*
2. *If $\lim_{n\to\infty} X(n)$ is an ε-pure state of the form $\hat{x}_i = \varepsilon/|\mathcal{S}_i|\mathbf{1} + (1-\varepsilon)e_{\hat{s}_i}$ for some $\hat{s} \in \mathcal{S}$, then $\hat{s}$ is a.s. a strict equilibrium of Γ and convergence occurs at a quasi-exponential rate:*

$$X_{i\hat{s}_i}(n) \geq 1 - \varepsilon - \sum_{s \in \mathcal{S}_i, s \neq \hat{s}_i} {\theta'_i}^{-1}(c(\sum_{k=n_0}^{n-1} \gamma_k)) \quad \textit{for some positive } c > 0. \tag{C.5}$$

Using the same arguments in Section 3, Theorem 14 allow us to deduce that:

Corollary 15. *If Algorithm 3 is run with assumptions as above and no mixing ($\varepsilon = 0$), $X(n)$ converges to a Nash equilibrium with probability 1. Moreover, if the limit of $X(n)$ is pure and $\beta < 1$, we have*

$$X_{i\hat{s}_i}(n) \geq 1 - \sum_{s \in \mathcal{S}_i, s \neq \hat{s}_i} {\theta'_i}^{-1}(c(\sum_{k=n_0}^{n-1} \gamma_k)) \quad \textit{for some positive } c > 0. \tag{C.6}$$

Moreover, the converge result for (ε-Hedge) in bandit framework can be generalized to ε-following the regularized leader algorithms. Thus,

Theorem 16. *Let Γ be a generic potential game and suppose that Algorithm 3 is run with i) the bandit estimator (4.2); ii) a strictly positive mixing factor $\varepsilon > 0$; and iii) a step-size sequence of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (0, 1]$. Then:*

1. *$X(n)$ converges (a.s.) to a δ-equilibrium of Γ with $\delta \equiv \delta(\varepsilon) \to 0$ as $\varepsilon \to 0$.*
2. *If $\lim_{n\to\infty} X(n)$ is an ε-pure state of the form $\hat{x}_i = \varepsilon/|\mathcal{S}_i|\mathbf{1} + (1-\varepsilon)e_{\hat{s}_i}$ for some $\hat{s} \in \mathcal{S}$, then $\hat{s}$ is a.s. a strict equilibrium of Γ and convergence occurs at a quasi-exponential rate:*

$$X_{i\hat{s}_i}(n) \geq 1 - \sum_{s\in\mathcal{S}_i, s\neq\hat{s}_i} {\theta'}_i^{-1}(c(\sum_{k=n_0}^{n-1}\gamma_k)) \quad \text{for some positive } c > 0.$$

The remainder of this section is devoted to prove Theorem 14. Our proof follows the same step as previously:

1. Show that X is an asymptotic pseudo trajectory of a continuous dynamic (Proposition 17);
2. Show that the potential function of the game is strict Lyapunov function of the continuous dynamic (Proposition 19 and 20);
3. Show that X converges towards a rest point of the continuous dynamic (Proposition 20);
4. Show that if X converges towards a point it is a Nash Equilibrium (Propositions 22 and 21).

The first step is to show that the linear interpolation $x(t)$ of $(X(n))_{n\in\mathbb{N}}$ is an asymptotic pseudotrajectory of a dynamics. We also define the continuous variables, $x^\Lambda_{is}(t) = \frac{x_{is}(t)-\epsilon_i/|\mathcal{S}|_i}{1-\epsilon_i}$ and $y_{is}(t)$ such that $x^\Lambda_{is}(t) = Q_i(y_{is}(t))$, with the same relations as the corresponding discret processes.

C.1 Asymptotic pseudo trajectory

Proposition 17. *Suppose that Algorithm 3 is run with semi-bandit feedback satisfying (H1) and (H2), a small enough mixing factor $\varepsilon \geq 0$, and a step-size of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/q, 1]$.*

The interpolated process of the sequences $(X_i(n))_{n\in\mathcal{N}}$ is an asymptotic pseudo trajectory of the solutions of the following ordinary differential equation

$$\begin{aligned} \dot{x}_{is}(t) &= (1-\epsilon_i)\nabla Q^T_{is}(y_i(t))v_i(x(t)) \\ \dot{x}_i(t) &= Jac(Q_i)(y_i(t))v_i(x(t)) \end{aligned} \tag{C.7}$$

where ∇Q_{is} is the gradient vector of Q_{is} and $Jac(Q_i)$ the Jacobian matrix of Q_i.

Proof. First, we will check that the stochastic process $X_{i(n)}$ given by (ε-Reg) is an approximate Robbins-Monro algorithm. Using Taylor's Remainder Theorem, we obtain:

$$\begin{aligned} X_{is}(n+1) &= \frac{\epsilon_i}{|\mathcal{S}_i|} + (1-\epsilon_i)Q_{is}(Y_i(n+1)) \\ &= \frac{\epsilon_i}{|\mathcal{S}_i|} + (1-\epsilon_i)Q_{is}(Y_i(n) + \gamma_n\hat{v}_i(n)) \\ &= \frac{\epsilon_i}{|\mathcal{S}_i|} + (1-\epsilon_i)Q_{is}(Y_i(n)) + (1-\epsilon_i)\gamma_n\left(\nabla Q^T_{is}(Y_i(n))\hat{v}_i(n) + \frac{1}{2}\gamma_n\hat{v}^T_i(n)\,\mathrm{Hess}(Q_{is})(\psi_i(n))\hat{v}_i(n)\right) \\ &= X_{is}(n) + (1-\epsilon_i)\gamma_n\left(\nabla Q^T_{is}(Y_i(n))\hat{v}_i(n) + \frac{\gamma_n}{2}\hat{v}^T_i(n)\,\mathrm{Hess}(Q_{is})(\psi_i(n))\hat{v}_i(n)\right) \\ &= X_{is}(n) + (1-\epsilon_i)\gamma_n\left(\nabla Q^T_{is}(Y_i(n))v_i(X(n)) + \nabla Q^T_{is}(Y_i(n))(\hat{v}_i(n) - v_i(X(n))) + \gamma_n a_n\right) \end{aligned} \tag{C.8}$$

where ∇Q_{is} is the gradient vector of Q_{is}, $\nabla Q^{\mathrm{T}}_{is}$ is its transposed, $\mathrm{Hess}\, Q_{is}$ is the Hessian matrix of Q_{is}, and $\psi_i(n)$ is in the line segment going out from $Y_i(n)$ to the point $Y_i(n+1)$.

We omit the rest of the proof because it is similar to proof of Theorem 9 by remplacing the map Λ by Q. □

We also have the same result with $(Y_i(n))_{n\in\mathbb{N}}$.

Lemma 18. *With the same assumptions, the interpolated process of the sequences $(Y_i(n))_{n\in\mathbb{N}}$ is an asymptotic pseudo trajectory of the solutions of the following ordinary differential equation*

$$\dot{y}_i(t) = v_i(x(t)) \tag{C.9}$$

Proof. Recall that

$$Y_i(n+1) = Y_i(n) + \gamma_n \hat{v}_i(n) = Y_i(n) + \gamma_n v_i(X(n))) + \gamma_n(\hat{v}_i(n) - v_i(X(n))) \tag{C.10}$$

Using Hypothese (H1) and (3.2b) in Propositions 4.2 and 4.1 of [23] allow us to conclude the proof. □

We prove that is converges to rest points of (RD_ε). To that end we start by proving that the dynamics (RD_ε) admits a strict increasing Lyapunov function. Before that we introduce some other discrete-time process $(X_i^\Lambda(n))_{n\in\mathbb{N}}$ to help with this proof

$$X_i^\Lambda(n) = \frac{1}{(1-\varepsilon_i)}(X_i(n) - \varepsilon_i/|\mathcal{S}_i|\mathbf{1})$$

Its linear interpolation is denoted by $x_i^\Lambda(t)$.

Proposition 19. *Let Γ be a generic potential game. Thus, we have The potential function f of Γ is a strict increasing Lyapunov function of the flow inducted by the dynamics (RD_ε).*

Proof. We consider the variation of f. We have

$$\begin{aligned}
\dot{f}(x(t)) &= \sum_{i\in\mathcal{N}} \frac{\partial f}{\partial x_i}(x(t))\dot{x}_i(t) \\
&= \sum_{i\in\mathcal{N}} v_i^T(x(t))\dot{x}_i(t) \\
&= \sum_{i\in\mathcal{N}} \sum_{s\in\mathcal{S}_i} v_{is}(x(t))\dot{x}_{is}(t)
\end{aligned} \tag{C.11}$$

We look for an other expression of $\dot{x}_i(t) = (1-\epsilon_i)\dot{x}_i^\Lambda(t)$, to that end we use the Lagrange multiplier. Recall that $Q_i(Y_i) = arg\,max_{x_i^\Lambda\in\mathcal{X}_i}(\langle y_i(t), x_i^\Lambda(t)\rangle - h_i(x_i^\Lambda(t)))$, so Q_i can be seen as an optimisation of $\langle y_i(t), x_i^\Lambda(t)\rangle - h_i(x_i^\Lambda(t))$ when $\sum_{s\in\mathcal{S}_i} x_{is}^\Lambda = 1$. For simplicity we drop the "(t)".

$$\mathcal{L}_i(x_i^\Lambda, \lambda_i) = \sum_{s\in\mathcal{S}_i}(x_{is}^\Lambda y_{is} - \theta_i(x_{is}^\Lambda)) + \lambda_i(\sum_{s\in\mathcal{S}_i} x_{is}^\Lambda - 1) \tag{C.12}$$

By derivation we obtain

$$\frac{\partial\mathcal{L}_i}{\partial\lambda_i} = 0 = \sum_{s\in\mathcal{S}_i} x_{is}^\Lambda - 1 \tag{C.13}$$

and

$$\frac{\partial\mathcal{L}_i}{\partial x_{is}^\Lambda} = 0 = y_{is} - \theta_i'(x_{is}^\Lambda) + \lambda_i, \quad \text{for all } s\in\mathcal{S}_i. \tag{C.14}$$

Taking the derivative again we get

$$\dot{y}_{is} = \theta_i''(x_{is}^\Lambda)\dot{x}_{is}^\Lambda - \dot{\lambda}_i, \quad \text{for all } s\in\mathcal{S}_i. \tag{C.15}$$

Let H_i be the Hessian matrix of h_i. As h is strongly convex, its Hessian matrix is strictly positive and invertible, in particular $\theta_i''(x_{is}^\Lambda) > 0$. Let $d_{is} = \theta_i''(x_{is}^\Lambda)$, we obtain:

$$\begin{aligned}
\dot{x}_{is}^\Lambda &= \frac{\dot{y}_{is}}{d_{is}} + \frac{\dot{\lambda}_i}{d_{is}}, \quad \text{for all } s\in\mathcal{S}_i. \\
0 &= \sum_{s\in\mathcal{S}_i} \dot{x}_{is}^\Lambda = \sum_{s\in\mathcal{S}_i}(\frac{\dot{y}_{is}}{d_{is}} + \frac{\dot{\lambda}_i}{d_{is}})
\end{aligned} \tag{C.16}$$

Because $\sum_{s\in\mathcal{S}_i} x^\Lambda_{is} = 1$, therefore we can express $\dot{\lambda}_i$ and $\dot{x}^\Lambda_{is}$.

$$\begin{aligned}\dot{\lambda}_i &= -\frac{\sum_{s\in\mathcal{S}_i}\frac{\dot{y}_{is}}{d_{is}}}{\sum_{s\in\mathcal{S}_i}\frac{1}{d_{is}}}\\ \dot{x}^\Lambda_{is} &= \frac{\dot{y}_{is}}{d_{is}} - \frac{1}{d_{is}}\frac{\sum_{s'\in\mathcal{S}_i}\frac{\dot{y}_{is'}}{d_{is'}}}{\sum_{s'\in\mathcal{S}_i}\frac{1}{d_{is'}}} = \sum_{s'\in\mathcal{S}_i}\frac{\dot{y}_{is} - \dot{y_{is'}}}{d_{is'}d_{is}\sum_{s'\in\mathcal{S}_i}\frac{1}{d_{is'}}}\end{aligned} \tag{C.17}$$

Injecting that into (C.11) and using $\dot{y}_{is}(t) = v_{is}(x(t))$ (Lemma (18)):

$$\begin{aligned}\dot{f}(x(t)) &= (1-\epsilon_i)\sum_{i\in\mathcal{N}}\sum_{s\in\mathcal{S}_i}\sum_{s'\in\mathcal{S}_i}\frac{v^2_{is}(x(t)) - v_{is'}(x(t))v_{is}(x(t))}{d_{is'}d_{is}\sum_{s'\in\mathcal{S}_i}\frac{1}{d_{is'}}}\\ \dot{f}(x(t)) &= (1-\epsilon_i)\sum_{i\in\mathcal{N}}\sum_{s\in\mathcal{S}_i}\sum_{s'\in\mathcal{S}_i, s'\neq s}\frac{(v_{is}(x(t)) - v_{is'}(x(t)))^2}{d_{is'}d_{is}\sum_{s'\in\mathcal{S}_i}\frac{1}{d_{is'}}}\\ \dot{f}(x(t)) &\geq 0\end{aligned} \tag{C.18}$$

It is important to notice here that whereas payoff v are evaluated on x, d_{is} correspond to the second derivation of h_{is} evaluated on x^Λ. This comes from the payoffs being evaluated on the probability used to draw strategies whereas d_{is} comes from the analysis of the choice map that determines x^Λ. The consequences are that d_{is} convergences to ∞ when $x^\Lambda_{is} \to 0$ because h is steep, but x_{is} does not converge to 0.

If $x(t)$ is an equilibrium, $\dot{x}_{is}(t) = \frac{\dot{x}^\Lambda_{is}(t)}{1-\epsilon_i} = 0$ for all $i \in \mathcal{N}, s \in \mathcal{S}_i$ and $\dot{f}(x(t)) = 0$ using the Equation (C.18). Let show the reverse, meaning that if $\dot{f}(x(t)) = 0$ then $x(t)$ is an equilibrium. First we remark that $\dot{f}(x(t)) = 0$ if each term of the sum is nul. At least one of three conditions needs to the true for a term to be null:

1. $\frac{1}{d_{is}} = 0$,
2. $\frac{1}{d_{is'}} = 0$,
3. $v_{is}(x(t)) = v_{is'}(x(t))$.

Recall that h is steep, is derivative goes to the infinity when reaching the boundary of $\mathcal{X}$. Because $\mathcal{X}$ is bounded, the second derivative of h does the same and $d_{is} \to_{x_{is}->0} \infty$.

So those three conditions can be gathered in one condition :

$$\dot{f}(x(t)) = 0 \Leftrightarrow \forall i \in \mathcal{N} v_{is}(x(t)) = v_{is'}(x(t)) \forall s, s' \in \mathcal{S}_i \text{ such that } \frac{1}{d_{is}} \neq 0 \text{ and } \frac{1}{d_{is'}} \neq 0 \tag{C.19}$$

Recall that $\dot{y_{is}}(t) = v_{is}(x(t))$ (Lemma (18)) so this condition applied to (C.17) gives us that $\dot{x_{is}}(t) = 0 \quad \forall i \in \mathcal{N}, \forall s, \in \mathcal{S}_i$ so $\dot{x^\epsilon_{is}}(t) = 0$. Therefore $\dot{f}(x(t)) = 0$ if and only if $(x(t))$ is an equilibrium of the dynamics (C.7).

□

C.2 Lyapunov function and Potential game

Proposition 20. *Let Γ be a generic potential game. Suppose that Algorithm 3 is run with semi-bandit feedback satisfying (H1) and (H2), a small enough mixing factor $\varepsilon \geq 0$, and a step-size of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/q, 1]$. The interpolated process of the sequences $(X_i(n))_{n\in\mathbb{N}}$ converges to a rest point of C.7.*

Proof. The proof is the same as in Proposition 11. □

C.3 Convergence towards a rest point of the continuous dynamic

This following proposition corresponds to Point 1 of Theorem 14.

Proposition 21. *Suppose that Algorithm 3 is run with semi-bandit feedback satisfying (H1) and (H2), a small enough mixing factor $\varepsilon \geq 0$, and a step-size of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/q, 1]$. If $X(n) \to \hat{x}, \quad \hat{x}$ is a ϵ-Nash Equilibrium and $X^\Lambda(n) \to x^*$ with x^* a Nash Equilibrium, given that ϵ is sufficiently small.*

Proof. When $X(n)$ converges, $X^\Lambda(n)$ converges too. Applying Lemma 6 it suffices to show that if $X^\Lambda(n) \to x^*$ then x^* is Nash Equilibrium. We show by contradiction that $(X^\Lambda(n))_{n\in\mathbb{N}}$ converges to x^* a Nash Equilibrium. Assume that x^* is not a Nash Equilibrium. By definition, we have

$$\exists i \in \mathcal{N}, \exists s' \in \mathcal{S}_i, s' \notin \text{supp}(x_i^*), s.t., v_{is'}(x^*) > v_{is}(x^*), \forall s \in \text{supp}(x_i^*).$$

By continuity of u, there is a neighborhood U of x^* and $a > 0$ such that

$$\exists i \in \mathcal{N}, \exists s' \in \mathcal{S}_i, s' \notin \text{supp}(x_i^*), s.t., v_{is'}(x) - v_{is}(x) > a, \forall s \in \text{supp}(x_i^*), x \in U. \tag{C.20}$$

For ϵ small enough and for all n big enough, we have $X(n) \in U$ because

$$\|X(n) - x^*\| \leq \|X(n) - X^\Lambda(n)\| + \|X^\Lambda(n) - x^*\| \leq N\epsilon + \|X^\Lambda(n) - x^*\|. \tag{C.21}$$

So, $\exists i \in \mathcal{N}, \exists s' \in \mathcal{S}_i, s' \notin \text{supp}(x_i^*), s.t., v_{is'}(X(n)) - v_{is}(X(n)) > a, \forall s \in \text{supp}(x_i^*)$ Using Lemma 12, for n_0 big enough and $n \geq n_0$:

$$Y_{is'}(n) - Y_{is}(n) \geq C + b\sum_{t=n_0}^{n-1} \gamma_t \to \infty \tag{C.22}$$

Thus, using (C.14) $Y_{is'}(n) - Y_{is}(n) = \theta_i'(X_{is'}^\Lambda(n)) - \theta_i'(X_{is}^\Lambda(n))$ we have

$$\theta_i'(X_{is'}^\Lambda(n)) - \theta_i'(X_{is}^\Lambda(n)) \geq C + b\sum_{t=n_0}^{n-1} \gamma_n \tag{C.23}$$

And

$$\theta_i'(X_{is'}^\Lambda(n)) - \theta_i'(X_{is}^\Lambda(n)) \to_{n\to\infty} \infty \tag{C.24}$$

That is a contradiction because $\theta_i'(X_{is'}^\Lambda(n)) \to -\infty$ because s' is not in the support of x^* and $X_{is'}^\Lambda(n) \to 0$ and $\theta_i'(X_{is}^\Lambda(n))$ is bounded.

Therefore x^* is a Nash Equilibrium and this concludes the proof. □

C.4 Convergence Rate

This following proposition corresponds to Point 2 of Theorem 14.

Proposition 22. *Suppose that Algorithm 3 is run with semi-bandit feedback satisfying (H1) and (H2), a small enough mixing factor $\varepsilon \geq 0$, and a step-size of the form $\gamma_n \propto 1/n^\beta$ for some $\beta \in (1/q, 1]$. If $X(n)$ converges to an ε-pure state $\hat{x}$ of the form $\hat{x}_i = \varepsilon/|\mathcal{S}_i|\mathbf{1} + (1-\varepsilon)e_{\hat{s}_i}$ for some $\hat{s} \in \mathcal{S}$, then $\hat{s}$ is a.s. a strict equilibrium of Γ and convergence occurs at a quasi-exponential rate:*

$$1 - X_{\hat{s}_i}^\Lambda(n) \leq \sum_{s\in\mathcal{S}_i, s\neq\hat{s}_i} {\theta_i'}^{-1}(\mathcal{O}(\sum_{k=n_0}^{n-1}\gamma_k))$$

Proof. By continuity of u, there is a neighborhood U of $\hat{s}$ and $a' > 0$ such that:

$\forall i \in \mathcal{N}, \forall s' \in \mathcal{S}_i, s' \neq \hat{s}_i, v_{i\hat{s}_i}(x) - v_{is'}(x) > a', x \in U$. Using Lemma 12 for n_0 big enough and $s \in \mathcal{S}_i, s \neq \hat{s}_i$:

$$Y_{i\hat{s}_i}(n) - Y_{is}(n) \geq C + b\sum_{t=n_0}^{n-1}\gamma_t$$

So, using (C.14) $Y_{i\hat{s}_i}(n) - Y_{is}(n) = \theta_i'(X^{\Lambda}_{i\hat{s}_i}(n)) - \theta_i'(X^{\Lambda}_{is}(n))$, and that $\theta_i'(1) < \infty$ because h' is continuous on $(0, \infty)$, we can deduce that

$$\theta_i'(X^{\Lambda}_{i\hat{s}_i}(n)) - \theta_i'(X^{\Lambda}_{is}(n)) \geq C + b \sum_{t=n_0}^{n-1} \gamma_t$$

$$\theta_i'(X^{\Lambda}_{is}(n)) \leq \theta_i'(X^{\Lambda}_{i\hat{s}_i}(n)) - C - b \sum_{t=n_0}^{n-1} \gamma_t$$

$$\theta_i'(X^{\Lambda}_{is}(n)) \leq \theta_i'(1) - C - b \sum_{t=n_0}^{n-1} \gamma_t$$

$$X^{\Lambda}_{is}(n) \leq {\theta'}_i^{-1} \left(\mathcal{O}(\sum_{t=n_0}^{n-1} \gamma_n) \right). \quad \square$$