Generalized Individual Q-learning for Polymatrix Games with Partial Observations

Ahmed Said Donmez and Muhammed O. Sayin

Abstract— This paper addresses the challenge of limited observations in non-cooperative multi-agent systems where agents can have partial access to other agents' actions. We present the generalized individual Q-learning dynamics that combine belief-based and payoff-based learning for the networked interconnections of more than two self-interested agents. This approach leverages access to opponents' actions whenever possible, demonstrably achieving a faster (guaranteed) convergence to quantal response equilibrium in multi-agent zero-sum and potential polymatrix games. Notably, the dynamics reduce to the well-studied smoothed fictitious play and individual Q-learning under full and no access to opponent actions, respectively. We further quantify the improvement in convergence rate due to observing opponents' actions through numerical simulations.

I. INTRODUCTION

Limited access to other agents' actions is a significant challenge for multi-agent systems, particularly for networked interconnections [1]. In the absence of observations, the literature often focuses on the extreme case of no observation, i.e., payoff-based dynamics [2], [3]. However, payoff-based dynamics ignoring partial observations available can suffer from slow convergence to equilibrium though asymptotic convergence guarantees might still be possible.

This paper addresses the limited access to opponent actions by proposing novel learning dynamics that combine belief-based and payoff-based learning for self-interested agents in non-cooperative environments. To this end, we focus on polymatrix games [4], where agents have network separable *pairwise* interconnections while they can only observe the actions of the agents in proximity, as illustrated in Fig. 1 over two separate layers. We present the *generalized* individual Q-learning dynamics, in which agents leverage opponents' actions that they can observe so that they can accelerate their learning process. The dynamics reduce to the well-studied smoothed fictitious play [5] and individual Q-learning [2], in scenarios with access to the actions of every or none of the other agents, respectively.

This paper is related to the broad literature of *learning in games* studying whether simplistic learning or adaptation rules of self-interested agents can reach equilibrium, e.g., see the overview [5]. Such behavioral rules are generally based on the simplifying assumption that the opponents play according to some stationary strategies and the agents adapt
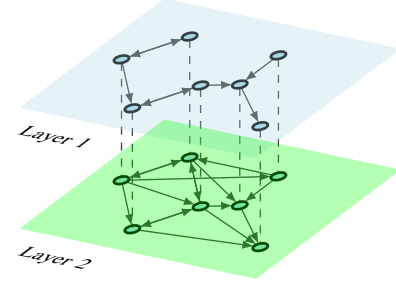
Fig. 1: An illustration of observability and interconnectedness over two layers. Nodes connected by dashed lines across layers represent an agent. In Layer 1, observability is depicted, where directed edges indicate that agents can observe the actions of others, forming the observability graph. In Layer 2, agents interact and can affect each other's payoffs, as represented by the directed edges, forming the interaction graph.

their act either based on beliefs, as in the smoothed fictitious play [5], or payoffs, as in the individual Q-learning [2]. The smoothed fictitious play is known to reach equilibrium in two-agent zero-sum and multi-agent potential games [5]. On the other hand, the individual Q-learning dynamics are known to reach equilibrium in two-agent zero-sum and *two-agent* potential games [2]. Here, we show that they indeed also reach equilibrium in *multi-agent* zero-sum and *multi-agent* potential polymatrix games.

Notably, the experience-weighted attraction (EWA) learning unifies the belief-based and payoff-based learning for normal-form games with some parameter controlling the weight of experiences (similar to the beliefs formed) and attractions (based on the payoffs received) in the actions taken [6]. However, agents still need to have access to opponent actions unless the attractions have the full weight, i.e., it is fully payoff-based. On the other hand, partial access to opponent actions can also lead to heterogeneous dynamics, as studied in [7]. They only study two-agent zero-sum games where belief-based and payoff-based dynamics can play against each other.

In this paper, we demonstrate that the generalized individual Q-learning dynamics converge to quantal response equilibrium in multi-agent zero-sum and potential polymatrix games almost surely. Furthermore, we quantify the significant improvement in convergence rate achieved through access to opponents' actions using numerical simulations.

The rest of the paper is organized as follows. We describe polymatrix games in Section II. We present the new dy-

namics in Section III. We provide analytical and numerical results, resp., in Sections IV and V. We conclude the paper with some remarks in Section VI.

II. PRELIMINARY: POLYMATRIX GAMES

Consider n agents having pairwise interactions with each other. We can represent their interactions via a graph (depicted in Figure 1, Layer 2) $G = (I, E)$, where I is the set of vertices and $E \subset I \times I$ is the set of directed edges, such that vertices and edges, resp., refer to agents and the pairwise interactions. Then, we can model their interactions via a polymatrix game characterized by the tuple $\mathscr{P} = \langle G, (A^i, u^i)_{i \in I}^n \rangle$, where A^i and $u^i : \prod_{j \in I} A^j \to \mathbb{R}$, resp., denote agent i's *finite* action set and payoff function. Furthermore, the payoff function $u^i(\cdot)$ can be written as

$$u^i((a^t)_{t \in I}) = \sum_{j:(i,j) \in E} u^{ij}(a^i, a^j) \quad \forall (a^t \in A^t)_{t \in I}, \quad (1)$$

for some sub-payoff function $u^{ij} : A^i \times A^j \to \mathbb{R}$ depending on the actions of the pair (i, j) only. Indeed, the separability of the payoff function in (1) is the defining characteristic of the polymatrix games [8]. Without loss of generality, we let $u^{ij}(a^i, a^j) = 0$ for all (a^i, a^j) if $(i, j) \notin E$. Furthermore, we let each agent i randomize their actions based on some mixed strategy $\pi^i \in \Delta^i$ independently, where Δ^i denotes the probability simplex over A^i.

For example, we say that $\mathscr{P}$ is a *zero-sum polymatrix game* if $\sum_{i \in I} u^i(a) = 0$ for all $a \in A := \prod_{i \in I} A^i$. Correspondingly, we say that $\mathscr{P}$ is a *potential polymatrix game* if there exists some potential function $\phi : A \to \mathbb{R}$ satisfying

$$\phi(\tilde{a}^i, a^{-i}) - \phi(a^i, a^{-i}) = u^i(\tilde{a}^i, a^{-i}) - u^i(a^i, a^{-i}), \quad (2)$$

for all $a^i, \tilde{a}^i \in A^i$, $a^{-i} \in A^{-i} := \Pi_{j \neq i} A^j$, and $i \in I$.

We define agent i's smoothed best response $\mathrm{br}^i : \mathbb{R}^{|A^i|} \to \Delta^i$ by

$$\mathrm{br}^i(q) := \arg\max_{\mu \in \Delta^i} \left\{ \mu^T q + \tau H(\mu) \right\} \quad \forall q \in \mathbb{R}^{|A^i|}, \quad (3)$$

where $\tau > 0$ is the temperature parameter controlling the exploration level and $H(\mu) = -\sum_a \mu(a) \log \mu(a)$ is the entropy regularization. There exists a unique smoothed best response due to the strict concavity of the regularization. Indeed, the smoothed best response $\mathrm{br}^i(q)$ can also be written as [9]

$$\mathrm{br}^i(q)(a) = \frac{\exp(q(a)/\tau)}{\sum_{\tilde{a} \in A^i} \exp(q(\tilde{a})/\tau)} \quad \forall a \in A^i. \quad (4)$$

Note also that the smoothed best response has an inherent exploration property since every action have positive probability, i.e., $\mathrm{br}^i(q)(a) > 0$ for all $a \in A^i$.

Based on the smoothed best response, we focus on quantal response equilibrium (QRE) [10] as a solution concept for $\mathscr{P}$. We say that a strategy profile $\pi = (\pi^i \in \Delta^i)_{i \in I}$ for $\mathscr{P}$ is QRE provided that

$$\pi^i = \mathrm{br}^i \left(u^i(\cdot, \pi^{-i}) \right) \quad \forall i \in I, \quad (5)$$

where $\pi^{-i} = (\pi^j)_{j \neq i}$.[1] Correspondingly, the QRE-gap of any

---

[1] Given any function $f : A \to \mathbb{R}$, we let $f(\mu) = \mathrm{E}_{a \sim \mu}[f(a)]$ for any probability distribution μ over A, with a slight abuse of notation for simpler expressions.

strategy profile π is defined by

$$\mathrm{QRE}(\pi) := \sum_{i \in I} \left| u^i(\pi^i, \pi^{-i}) - u^i \left(\mathrm{br}^i(u^i(\cdot, \pi^{-i})), \pi^{-i} \right) \right|. \quad (6)$$

III. GENERALIZED INDIVIDUAL-Q LEARNING

We first describe the individual Q-learning dynamics [2] for the repeated play of *two-agent* normal-form games where agents do not have access to opponent actions. Each agent i assumes that the opponent $j \neq i$ plays according to some stationary mixed strategy $\pi^j \in \Delta^j$ in the repeated play of the underlying game. Then, the value of action $a^i \in A^i$, called the *Q-function* and denoted by $\underline{q}^i : A^i \to \mathbb{R}$, is given by

$$\underline{q}^i(a^i) = u^i(a^i, \pi^j) \quad \forall a^i \in A^i. \quad (7)$$

Since the agents do not observe the opponent actions, they estimate the Q-function based on the payoff received (different from the belief-based dynamics such as fictitious play). Let $q_k^i : A^i \to \mathbb{R}$ be agent i's estimate at the kth repetition. Then, agent i can update the estimate q_k^i according to

$$q_{k+1}^i(a^i) = q_k^i(a^i) + \overline{\alpha}_k^i(a^i) \cdot \left(u_k^i - q_k^i(a^i) \right), \quad (8)$$

where $u_k^i = u^i(a_k^i, a_k^j)$ is the payoff received when the action profile (a_k^i, a_k^j) is played and $\overline{\alpha}_k^i(a^i) \in [0, 1]$ is some step size.

Observe that agents only receive the payoff for the current action profile (a_k^i, a_k^j). Therefore, we have $\overline{\alpha}_k^i(a^i) = 0$ for all $a^i \neq a_k^i$. Correspondingly, the estimates for the actions taken frequently get updated frequently. However, this leads to asynchronous update of the Q-functions for different actions. The individual Q-learning dynamics address this issue via the step sizes normalized by the probability of the current action gets played, i.e., $\mathrm{br}^i(q_k^i)(a^i)$, so that the Q-functions of every action get updated synchronously in the expectation [2]. Particularly, the step size $\overline{\alpha}_k^i$ is given by

$$\overline{\alpha}_k^i(a^i) = \mathbb{I}_{\{a^i = a_k^i\}} \min \left\{ 1, \frac{\alpha_k}{\mathrm{br}^i(q_k^i)(a^i)} \right\} \quad \forall a^i \in A^i, \quad (9)$$

where $\alpha_k \in (0, 1)$ is some reference step size decaying to zero and $\mathbb{I}_{\{\cdot\}}$ is the indicator function. We further threshold the ratio $\alpha_k / \mathrm{br}^i(q_k^i)(a^i)$ from above by 1. The threshold ensures that the step size $\overline{\alpha}_k^i(a^i) \in [0, 1]$ for all a^i so that the update (8) is the convex combination of the current estimate and the payoff received, and therefore, the estimates remain bounded as the payoffs are bounded if the initialization is from zero.

Due to the threshold in (9), the estimates for every action do not necessarily get updated synchronously in the expectation. However, the bounded estimates ensure that the denominator $\mathrm{br}^i(q_k^i)(a^i)$ is uniformly bounded from below by some $\varepsilon > 0$. Therefore, the decaying $\{a_k\}_{k \geq 0}$ ensures that eventually the threshold does not have any impact on the step size $\overline{\alpha}_k^i$, i.e., there exists some K (not depending on the trajectory) such that $\alpha_k / \mathrm{br}^i(q_k^i)(a^i) < 1$ for all $k > K$.

Next, we generalize the individual Q-learning to the repeated play of multi-agent polymatrix games where agents have limited access to opponent actions. To this end, we introduce the observability graph (depicted in Figure 1, Layer

1) $\overline{G} = (I, \overline{E})$ where $\overline{E} \subset I \times I$ is the set of directed edges such that $(i, j) \in \overline{E}$ if agent i can observe agent j's actions.

When agent i has access to agent j's action, agent i can form a belief $\pi_k^j \in \Delta^j$ about agent j's strategy (assumed to be stationary) based on the action history. For example, agent i can take the weighted empirical average of the past actions, as in the widely studied fictitious play dynamics. Hence, we let agent i use the reference step size $\alpha_k \in (0, 1)$ in the update of the belief π_k^j as

$$\pi_{k+1}^j = \pi_k^j + \alpha_k \cdot (a_k^j - \pi_k^j) \quad (10)$$

for all $j : (i, j) \in \overline{E}$, where we view the action a_k^j as a degenerate mixed strategy giving probability one to the associated action.

Similar to (7), agent i's Q-function is now given by

$$\underline{q}^i(a^i) = u^i(a^i, \pi^{-i}) = \sum_{j \neq i} u^{ij}(a^i, \pi^j), \quad (11)$$

where the second equality follows from (1). Then, based on the observability graph $\overline{G}$, we can write the Q-function as

$$\underline{q}^i(a^i) = \sum_{j:(i,j)\in\overline{E}} u^{ij}(a^i, \pi^j) + \sum_{j:(i,j)\notin\overline{E}} u^{ij}(a^i, \pi^j). \quad (12)$$

Agent i can estimate the first summation on the right-hand side as

$$\breve{q}_k^i(a^i) := \sum_{j:(i,j)\in\overline{E}} u^{ij}(a^i, \pi_k^j) \quad (13)$$

based on the beliefs π_k^j formed for all j such that $(i, j) \in \overline{E}$. Indeed, (10) yields that $\breve{q}_k^i(\cdot)$ evolves according to

$$\breve{q}_{k+1}^i(a^i) = \breve{q}_k^i(a^i) + \alpha_k \cdot (\breve{u}_k^i(a^i) - \breve{q}_k^i(a^i)) \quad \forall a^i \in A^i, \quad (14)$$

where

$$\breve{u}_k^i(a^i) := \sum_{j:(i,j)\in\overline{E}} u^{ij}(a^i, a_k^j) \quad \forall a^i \in A^i. \quad (15)$$

On the other hand, agent i does not form a belief about agents whose actions cannot be observed. Instead, agent i can estimate the second summation based on the payoff $u_k^i = u^i(a_k^i, a_k^{-i})$ received as in the individual Q-learning dynamics (8). However, the payoff received depends also on the pairwise interactions with the agents whose actions can be observed. Agent i can still isolate the payoffs coming from the pairwise interactions with the agents whose actions cannot be observed by

$$\widehat{u}_k^i := u_k^i - \breve{u}_k^i(a_k^i) \in \mathbb{R}. \quad (16)$$

Then, the estimation for the second term on the right-hand side of (12), denoted by $\widehat{q}_k^i : A^i \to \mathbb{R}$, can be updated according to

$$\widehat{q}_{k+1}^i(a^i) = \widehat{q}_k^i(a^i) + \overline{\alpha}_k^i(a^i) \cdot (\widehat{u}_k^i - \widehat{q}_k^i(a^i)) \quad \forall a^i \in A^i, \quad (17)$$

where the step size $\overline{\alpha}_k^i(a^i) \in [0, 1]$ is as described in (9).

Given the estimates $\breve{q}_k^i$ and $\widehat{q}_k^i$, the Q-function estimate is given by

$$q_k^i(a^i) := \breve{q}_k^i(a^i) + \widehat{q}_k^i(a^i) \quad \forall a^i \in A^i. \quad (18)$$

Algorithm 1 Generalized Individual Q-learning Dynamics

initialize: the estimates $\breve{q}_0^i, \widehat{q}_0^i \in \mathbb{R}^{A^i}$ arbitrarily
for each stage $k = 0, 1, \dots$ **do**
  set $q_k^i = \breve{q}_k^i + \widehat{q}_k^i$
  play $a_k^i \sim \mathrm{br}^i(q_k^i)$
  receive the payoff $u_k^i = u^i(a_k^i, a_k^{-i})$
  observe the opponent actions $(a_k^j)_{j:(i,j)\in\overline{E}}$ partially
  set $\breve{u}_k^i(a^i) = \sum_{j:(i,j)\in\overline{E}} u^{ij}(a^i, a_k^j)$ for all $a^i \in A^i$
  set $\widehat{u}_k^i = u_k^i - \breve{u}_k^i(a_k^i)$
  update the estimates

$$\breve{q}_{k+1}^i(a^i) = \breve{q}_k^i(a^i) + \alpha_k \cdot (\breve{u}_k^i(a^i) - \breve{q}_k^i(a^i)) \quad \forall a^i \in A^i$$
$$\widehat{q}_{k+1}^i(a^i) = \widehat{q}_k^i(a^i) + \overline{\alpha}_k^i(a^i) \cdot (\widehat{u}_k^i - \widehat{q}_k^i(a^i)) \quad \forall a^i \in A^i$$

  with the step sizes $\alpha_k \in (0, 1)$ and

$$\overline{\alpha}_k^i(a^i) = \mathbb{I}_{\{a^i = a_k^i\}} \min \left\{ 1, \frac{\alpha_k}{\mathrm{br}^i(q_k^i)(a^i)} \right\} \quad \forall a^i \in A^i$$

end for

Therefore, agent i can play $a_k^i \sim \mathrm{br}^i(q_k^i)$, as in the individual Q-learning dynamics. Algorithm 1 is a description of the generalized individual Q-learning dynamics for agent i. Note that the dynamics presented reduces to the smoothed fictitious play and individual Q-learning if $\overline{E} = E$ or $\overline{E} = \varnothing$, respectively.

Remark 1. We highlight the resemblance between (14) and (17). They differ in the step sizes $\alpha_k \in [0, 1]$ vs $\overline{\alpha}_k^i \in [0, 1]^{A^i}$ and the targets $\breve{u}_k^i \in \mathbb{R}^{A^i}$ vs $\widehat{u}_k^i \in \mathbb{R}$. The distinct target values arise from the need to account for counterfactual payoffs without directly observing opponents' actions. While agent i knows what $\breve{u}_k^i$ is for other possible actions, $\widehat{u}_k^i$ is only known for the current action taken by agent i. However, as shown later, both update have identical structures in the expectation conditioned on the play history due to the normalization in (9) (and for sufficiently large k due to the threshold in (9)).

IV. CONVERGENCE RESULTS

In this section, we characterize Algorithm 1's convergence properties for any observability graph structure. Agents can have full, partial, or no access to observations. They can even have heterogeneous accesses, as in [7] yet beyond two agents.

We make the following assumption on the step sizes used.

Assumption 1. *The step size $\alpha_k \in (0, 1)$ satisfies the following standard conditions:*

$$\lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad and \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty \quad (19)$$

for stochastic approximation.

The following theorem shows that Algorithm 1 reaches equilibrium in the sense that the weighted empirical averages of the actions taken, π_k^i's, converges to QRE almost surely

in multi-agent zero-sum and potential polymatrix games. Furthermore, the Q-function estimates, q_k^i's, track the Q-functions associated with those empirical averages.

Theorem 1 (Main Result) *Consider a polymatrix game characterized by $\mathscr{P} = \langle G, (A^i, u^i)_{i \in I} \rangle$. Let every agent follow Algorithm 1 and Assumption 1 holds. If $\mathscr{P}$ is a zero-sum or potential polymatrix game, then the weighted empirical averages of the actions $(\pi_k^i)_{i \in I}$ evolving according to (10) converge to QRE, i.e.,*

$$\text{QRE}(\pi_k) \to 0 \qquad (20)$$

as $k \to \infty$ almost surely. Furthermore, the Q-function estimates are asymptotically belief-based in the sense that

$$\|q_k^i - u^i(\cdot, \pi_k^{-i})\|_2 \to 0 \quad \forall i \in I, \qquad (21)$$

as $k \to \infty$ almost surely.[2]

Proof: The proof methodology is based on the stochastic approximation techniques [11]. The result follows from the stochastic approximation of Algorithm 1 based on new Lyapunov function formulations. We first focus on approximating discrete-time updates in Algorithm 1 via their limiting ordinary differential equations (o.d.e.). For example, we can write (10) as

$$\pi_{k+1}^i = \pi_k^i + \alpha_k(\text{br}^i(q_k^i) - \pi_k^i + \omega_k^i) \quad \forall i \in I, \qquad (22)$$

where the stochastic approximation error is defined by

$$\omega_k^i := a_k^i - \text{br}^i(q_k^i). \qquad (23)$$

Correspondingly, the update (14) can be written as

$$\breve{q}_{k+1}^i \equiv \breve{q}_k^i + \alpha_k \left(\sum_{j:(i,j)\in\overline{E}} u^{ij}(\cdot, \text{br}^j(q_k^j)) - \breve{q}_k^i + \breve{w}_k^i \right), \quad (24)$$

where the stochastic approximation error is defined by

$$\breve{w}_k^i(a^i) := \sum_{j:(i,j)\in\overline{E}} (u^{ij}(\cdot, a_k^j) - u^{ij}(\cdot, \text{br}^j(q_k^j))) \quad \forall a^i \in A^i. \quad (25)$$

Recall the discussion that the threshold in the step size $\overline{\alpha}_k^i(a^i)$ is ineffective, i.e., $\overline{\alpha}_k^i(a^i) < 1$, for sufficiently large k. Therefore, for sufficiently large k, we can write (17) as

$$\widehat{q}_{k+1}^i \equiv \widehat{q}_k^i + \alpha_k \left(\sum_{j:(i,j)\notin\overline{E}} u^{ij}(\cdot, \text{br}^j(q_k^j)) - \widehat{q}_k^i + \widehat{w}_k^i \right), \quad (26)$$

where the stochastic approximation error is defined by

$$\widehat{w}_k^i(a^i) := \frac{\mathbb{I}_{\{a^i = a_k^i\}}}{\text{br}^i(q_k^i)(a_k^i)} \left(\widehat{u}_k^i - \widehat{q}_k^i(a^i) \right)$$
$$- \left(\sum_{j:(i,j)\notin\overline{E}} u^{ij}(a^i, \text{br}^j(q_k^j)) - \widehat{q}_k^i(a^i) \right) \quad \forall a^i \in A^i. \quad (27)$$

Let $x_k = [\breve{q}_k^i; \widehat{q}_k^i; \pi_k^i]_{i \in I} \in \mathbb{R}^d$, where $d := 3\prod_i |A^i|$, and $w_k = [\breve{w}_k^i; \widehat{w}_k^i; \omega_k^i]_{i \in I}$. Given the filtration $\mathscr{F}_k := \sigma(x_l, w_l; l \le k)$, the

[2]We can view finite dimensional functions (such as $q_k^i : A^i \to \mathbb{R}$) as vectors (in $\mathbb{R}^{|A^i|}$ since A^i is a finite set).

stochastic approximation errors $(w_k)_{k \ge 0}$ form a Martingale difference sequence with $\mathrm{E}[\breve{w}_k^i \mid \mathscr{F}_k] \equiv 0$ and $\mathrm{E}[\widehat{w}_k^i \mid \mathscr{F}_k] \equiv 0$, where the expectation is taken with respect to the randomness induced by randomized actions. This yields that the updates (24) and (26) have identical structures in the expectation conditioned on the history $\mathscr{F}_k$, as discussed in Remark 1.

To have a simpler expression for the evolution of x_k, we define the function $F(\cdot)$ by

$$F \left(\begin{bmatrix} \breve{q}_k^i \\ \widehat{q}_k^i \\ \pi_k^i \end{bmatrix}_{i \in I} \right) = \begin{bmatrix} \sum_{j:(i,j)\in\overline{E}} u^{ij}(\cdot, \text{br}^j(\breve{q}_k^j + \widehat{q}_k^j)) \\ \sum_{j:(i,j)\notin\overline{E}} u^{ij}(\cdot, \text{br}^j(\breve{q}_k^j + \widehat{q}_k^j)) \\ \text{br}^i(\breve{q}_k^j + \widehat{q}_k^j) \end{bmatrix}_{i \in I}.$$

Since $\text{br}^i(\cdot)$ is Lipschitz continuous [12], we can show that $F(\cdot)$ is also Lipschitz continuous. Then, for sufficiently large k, we can write (22), (24), and (26) as

$$x_{k+1} = x_k + \alpha_k(F(x_k) - x_k + w_k). \qquad (28)$$

We can also show that the iterates x_k remain bounded since $\alpha_k \in (0,1)$ and payoffs are bounded, as the updated values are simply the convex combination of previous x_k and the bounded $F(x_k)$. Since the step sizes satisfy Assumption 1, we can apply stochastic approximation methods and obtain the limiting o.d.e. of (28) as $\dot{x} = F(x) - x$ [11]. More explicitly, we have

$$\frac{d\breve{q}^i}{dt}(\cdot) = \sum_{j:(i,j)\in\overline{E}} u^{ij}(\cdot, \text{br}^j(\breve{q}^j + \widehat{q}^j)) - \breve{q}^i(\cdot) \qquad (29a)$$

$$\frac{d\widehat{q}^i}{dt}(\cdot) = \sum_{j:(i,j)\notin\overline{E}} u^{ij}(\cdot, \text{br}^j(\breve{q}^j + \widehat{q}^j)) - \widehat{q}^i(\cdot) \qquad (29b)$$

$$\frac{d\pi^i}{dt} = \text{br}^i(\breve{q}^i + \widehat{q}^i) - \pi^i \qquad (29c)$$

for the continuous-time functions $\breve{q}^i : [0,\infty) \to \mathbb{R}^{A^i}$, $\widehat{q}^i : [0,\infty) \to \mathbb{R}^{A^i}$, and $\pi^i : [0,\infty) \to \Delta^i$ for all $i \in I$.

The discrete-time iterates x_k almost surely converges to a compact connected internally chain transitive set of the o.d.e. (29). To characterize the limit behavior of (28) through (29), we follow the approach in [11, Section 6.2] and call a continuously differentiable function $V : \mathbb{R}^d \to \mathbb{R}$ by a *Lyapunov function* for some compact invariant set $\Lambda \subset \mathbb{R}^d$ provided that for any solution to (29), we have
- $\dot{V}(x(t)) < 0$ for all $x(t) \notin \Lambda$
- $\dot{V}(x(t)) = 0$ for all $x(t) \in \Lambda$.

Then, [11, Proposition 6.4] says that every internally chain transitive set of (29) is contained in Λ provided that $V(\Lambda) := \{V(x) : x \in \Lambda\} \subset \mathbb{R}$ has empty interior.

Next, we formulate Lyapunov functions for (29). We first focus on zero-sum polymatrix games and introduce the candidate $V_z(\cdot)$ defined by

$$V_z(\breve{q}, \widehat{q}, \pi) := \sum_{i \in I} L^i(\breve{q}^i + \widehat{q}^i, \pi), \qquad (30)$$

where for each $i \in I$, the auxiliary $L^i(\cdot)$ is defined by

$$L^i(q^i, \pi) := \max_{\mu^i \in \Delta^i} \left\{ q^i(\mu^i) + \tau H(\mu^i) \right\} - \left(u^i(\pi) + \tau H(\pi^i) \right)$$
$$+ \|q^i - u^i(\cdot, \pi^{-i})\|_2. \qquad (31)$$
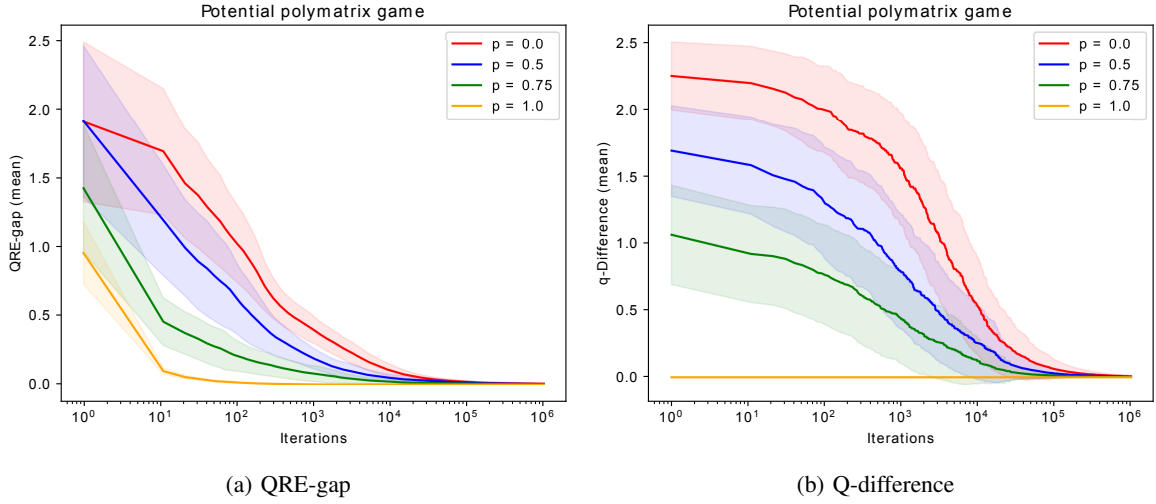
(a) QRE-gap   (b) Q-difference

Fig. 2: Convergence of the QRE-gap (left), and q_{diff} (right) for different values of edge-connection probability p, in the potential polymatrix games. The solid curves represent the mean over 50 independent trials and the shaded areas are ± 0.5 standard deviations from the mean values.

Remark 2. The Lyapunov function $V_z(\cdot)$ reduces to the one presented in [13] (and [9]) if agents have full access to observations (and there are only two agents).

The following proposition shows that the candidate $V_z(\cdot)$ is a Lyapunov function for (29), and (20) and (21) hold for zero-sum polymatrix games.

Proposition 1. *Given a zero-sum polymatrix game $\mathscr{P}$ and the o.d.e. (29), the function $V_z(\cdot)$ is a Lyapunov function for*

$$\Lambda := \{(\breve{q}, \widehat{q}, \pi) : \pi^i = \mathrm{br}^i(\breve{q}^i + \widehat{q}^i) \text{ and } \breve{q}^i + \widehat{q}^i \equiv u^i(\cdot, \pi^{-i})\}$$

and $V_z(\Lambda) = 0$.[3]

We next focus on potential polymatrix games with the potential function $\phi(\cdot) : A \to \mathbb{R}$ and introduce the candidate $V_p(\cdot)$ defined by

$$V_p(\breve{q}, \widehat{q}, \pi) := -\phi(\pi) - \sum_{i \in I} \tau H(\pi^i)$$
$$+ 2 \sum_{i \in I} \|q^i - u^i(\cdot, \pi^{-i})\|_2. \quad (32)$$

Remark 3. The Lyapunov function $V_p(\cdot)$ reduces to the one in [9] if agents have full access to observations.

The following proposition shows that the candidate $V_p(\cdot)$ is a Lyapunov function for (29) in potential polymatrix games.

Proposition 2. *Given a potential polymatrix game $\mathscr{P}$ and the o.d.e. (29), the function $V_p(\cdot)$ is a Lyapunov function for Λ, which is defined in Proposition 1 yet $V_p(\Lambda)$ is not necessarily singleton.*[3]

Note that V_p is continuously differentiable. Therefore, Sard's Lemma yields that $V_p(\Lambda)$ has empty interior, e.g., see [14, Corollary 3.28]. This implies that (20) and (21) also hold for potential polymatrix games. This completes the proof of Theorem 1. ∎

V. ILLUSTRATIVE EXAMPLES

In this section, we present simulation results. We generate fully-connected polymatrix games with uniformly distributed edge-payoffs u^{ij} for each edge in $(i, j) : i < j$. For zero-sum games, we set $u^{ji} = -u^{ij}$, and for potential games, $u^{ji} = u^{ij}$. In zero-sum games, payoffs sum to zero, $\sum_{i \in I} u^i(a) = 0$, while in potential games, $\phi(a) = \frac{1}{2} \sum_{i \in I} u^i(a)$.

We focus on $(n = 4)$-agent games where each agent has 3 actions, i.e., $|A^i| = 3$ for all i. We set the temperature parameter $\tau = 0.25$. We generate the observability graphs based on the Erdős Rényi model with probabilities $p \in \{0, 0.5, 0.75, 1\}$ such that $p = 0$ and $p = 1$ correspond, resp., to the no and full access to observations. We run 50 independent trials for each p and at each trial, we generate a random observability graph according to this model. We also run our simulations over 10^6 repetitions of the underlying game.

To measure the performance of Algorithm 1, we use two metrics: QRE-gap, as described in (6), and

$$q_{\mathrm{diff}}(q_k, \pi_k) := \frac{1}{n} \sum_{i \in I} \|q_k^i(\cdot) - u^i(\cdot, \pi_k^{-i})\|_2 \quad (33)$$

measuring the tracking error between q_k^i and $u^i(\cdot, \pi_k^{-i})$ averaged across agents, similar to (21). We plot the evolutions of these metrics for potential polymatrix and zero-sum polymatrix games, resp., in Figs. 2 and 3. We observe that these metrics decay to zero, as suggested by Theorem 1.

We achieve asymptotic convergence for any edge-connection probability p, with the rate improving as p increases. Convergence is slowest with no observations $p = 0$, and fastest with full observations $p = 1$. We observe that even though the QRE-gap decays at similar rates for the potential and zero-sum games under full observations, i.e., $p = 1$, the QRE-gap decays slower for the potential games than the zero-sum games under partial/no observations, i.e., $p < 1$. Indeed, the q_{diff} metrics also decay to zero slower for the
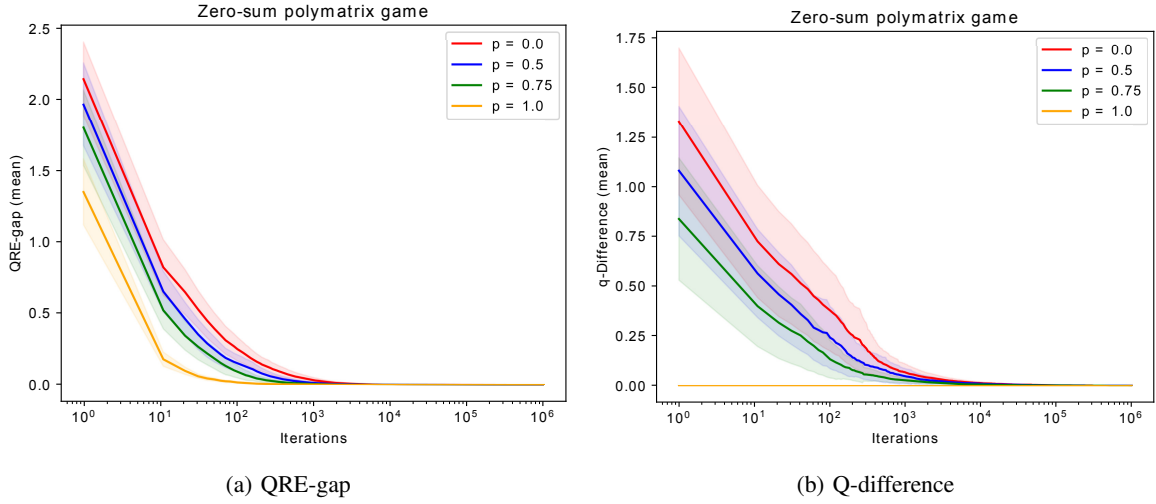
(a) QRE-gap

(b) Q-difference

Fig. 3: Convergence of the QRE-gap (left), and q_{diff} (right) for different values of edge-connection probability p, in the zero-sum polymatrix games. The solid curves represent the mean over 50 independent trials and the shaded areas are ± 0.5 standard deviations from the mean values.
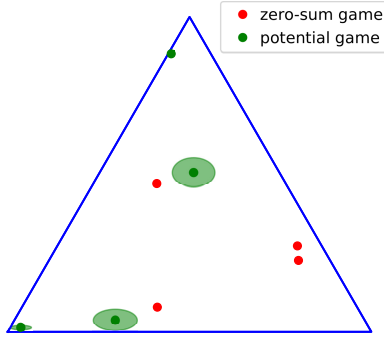

Fig. 4: An illustration of π_k^i's over the probability simplex $\Delta^i \subset \mathbb{R}^3$ for each agent i at the end of 10^6 iterations and over 50 independent trials for $p = 1$. Green and red dots are, resp., for the potential and zero-sum games. The shaded areas represent the standard deviations from the mean values.

potential games and this can slow down the QRE-gap's decay under partial/no observations. Exploration plays an important role in minimizing the tracking error. The equilibrium for the zero-sum game turns out to be more mixed than the potential game, as illustrated in Fig. 4.

VI. CONCLUSION

We introduced the generalized individual Q-learning dynamics to address the challenge of limited observations in non-cooperative multi-agent systems. By combining belief-based and payoff-based learning, this approach demonstrates faster convergence to quantal response equilibrium in networked interconnections of more than two self-interested agents. The proposed dynamics offer a significant improvement over existing methods, establishing theoretical convergence guarantees in multi-agent zero-sum and potential polymatrix games. Through numerical simulations, we validate the effectiveness of our approach and experimentally observe the benefits of observing opponents' actions.

We can list some of the future research directions as follows: (*i*) generalizing other payoff-based dynamics to partial observation settings, (*ii*) addressing limited access issue for learning in stochastic games, and (*iii*) quantifying the impact of the observations on the learning rate analytically.

REFERENCES

[1] S. Yüksel and T. Başar, *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*, ser. Systems & Control. Birkhäuser, 2013.

[2] D. S. Leslie and E. J. Collins, "Individual Q-learning in normal form games," *SIAM J. Control Optim.*, vol. 44, no. 2, pp. 495–514, 2005.

[3] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff-based dynamics for multiplayer weakly acyclic games," *SIAM J. Control and Optim.*, vol. 48, no. 1, pp. 373–396, 2009.

[4] Y. Cai, O. Candogan, C. Daskalakis, and C. Papadimitriou, "Zero-sum polymatrix games: A generalization of minmax," *Math. Oper. Res*, vol. 41, no. 2, pp. 648–655, 2016.

[5] D. Fudenberg and D. K. Levine, "Learning and equilibrium," *The Annual Rev. Econ.*, vol. 1, pp. 385–419, 2009.

[6] C. Camerer and T. Ho, "Experience-weighted attraction learning in normal form games," *Econometrica*, vol. 67, no. 4, pp. 827–874, 1999.

[7] Y. Arslantas, E. Yuceel, Y. Yalin, and M. O. Sayin, "Convergence of heterogeneous learning dynamics in zero-sum stochastic games," *arXiv preprint arXiv:2311.00778*, 2023.

[8] J. T. Howson Jr, "Equilibria of polymatrix games," *Management Science*, vol. 18, no. 5-part-1, pp. 312–318, 1972.

[9] J. Hofbauer and E. Hopkins, "Learning in perturbed asymmetric games," *Games Econom. Behav.*, vol. 52, pp. 133–152, 2005.

[10] R. McKelvey and T. Palfrey, "Quantal response equilibria for normal form games," *Games Econom. Behav.*, vol. 10, pp. 6–38, 1995.

[11] M. Benaim, "Dynamics of stochastic approximation algorithms," in *Le Seminaire de Probabilites, Lecture Notes in Math. 1709*. Berlin: Springer-Verlag, 1999, pp. 1–68.

[12] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *arXiv preprint arXiv:1704.00805*, 2017.

[13] C. Park, K. Zhang, and A. Ozdaglar, "Multi-player zero-sum Markov games with networked separable interactions," in *Proc. Int. Conf. Neural Information Processing Systems (NeurIPS)*, 2023, pp. 37 354–37 369.

[14] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions," *SIAM J. Control Optim.*, vol. 44, no. 1, pp. 328–348, 2005.