FINITE-SAMPLE GUARANTEES FOR LEARNING DYNAMICS IN ZERO-SUM POLYMATRIX GAMES[*]

FATHIMA ZARIN FAIZAL[†], ASUMAN OZDAGLAR[†], AND MARTIN J. WAINWRIGHT[†][‡]

Abstract. We study best-response type learning dynamics for zero-sum polymatrix games under two information settings. The two settings are distinguished by the type of information that each player has about the game and their opponent's strategy. The first setting is the full information case, in which each player knows their own and their opponents' payoff matrices and observes everyone's mixed strategies. The second setting is the minimal information case, where players do not observe their opponents' strategies and are not aware of any payoff matrices (instead they only observe their realized payoffs). For this setting, also known as the radically uncoupled case in the learning in games literature, we study a two-timescale learning dynamics that combine smoothed best-response type updates for strategy estimates with a TD-learning update to estimate a local payoff function. For these dynamics, without additional exploration, we provide polynomial-time finite-sample guarantees for convergence to an ϵ-Nash equilibrium.

1. Introduction.

Game theory is used to model interactions between two or more players, with the assumption that each player behaves rationally so as to maximize their individual payoffs subject to available information. A standard solution concept is that of a Nash equilibrium: it can be justified as the steady-state outcome of the learning process of players acting in their self-interest. This motivates the study of various learning dynamics by which players choose an action while learning the strategy of their opponents and adjust their play accordingly. Key questions associated with a given set of learning dynamics are stability, convergence to equilibria, and in a more refined analysis, the rate of convergence to an equilibrium.

Among the most natural and well-studied forms of learning dynamics is fictitious play (FP), first introduced by Brown [8, 9]. It is a simple and interpretable form of dynamics: players use information from previous rounds of play to estimate the opponent's strategy, and then play a best-response action based on this estimate. For two-player zero-sum games, Robinson [34] established the asymptotic convergence of FP. This was followed by a number of papers that studied its convergence properties for different classes of games; see Subsection 1.2 below for further details. Classical FP assumes that each player observes their opponent's actions and knows the payoff function, which can be viewed as a full information setting. While this is a useful benchmark, players may have limited information about their opponents' play or even their own payoff functions. In this paper, in addition to studying the full information setting, we also study the case where players do not observe their opponents' play; they only have access to their realized payoffs at each step, a case which we refer to as the minimal information setting.

We examine these issues within the class of zero-sum polymatrix games, which is a generalization of the class of two-player zero-sum games to the multiplayer setting. Players interact with each other in a pairwise manner, and an underlying interaction graph captures each player's payoff's dependencies on the other players. A pairwise matrix game is played on each edge and players choose a single strategy for each of the pairwise games they play in their neighborhood. Each player's payoff is the sum

[*] Submitted to the editors DATE.

[†] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA (fathima@mit.edu,asuman@mit.edu).

[‡] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA (mjwain@mit.edu).

of the payoffs they receive from each of the pairwise games they play. The zero-sum constraint ensures that the sum of the payoffs of all the players equals zero, i.e., there is no total flux of payoffs in or out of the system. Note that the matrix game played on each edge *need not* be zero-sum; each edge-based game being zero-sum is a special case of the more general class of zero-sum polymatrix games.

Zero-sum polymatrix games are used to model scenarios where pairwise interactions are dominant. For instance, consider a competitive resource allocation problem where multiple countries attempt to allocate their defense budget in order to obtain maximum control of shared resources. The existence of an edge between two countries in an underlying interaction graph $(\mathcal{N}, \mathcal{E})$ indicates whether or not those countries share a border. The action taken by each country in the polymatrix game would be to divide their defense budget among each of the pairwise interactions in which they are involved. On a particular edge $(i, j) \in \mathcal{E}$ between two countries i and j, the country that allocates more budget for edge (i, j) receives a payoff of $r_{i,j}$ which represents the utility obtained from controlling the shared resource on that edge; the country that allocates less budget for edge (i, j) receives a payoff of 0. While not a zero-sum game, this is a constant-sum polymatrix game, i.e., the total sum of payoffs across all countries for each possible budget allocation that all countries choose is equal to $\sum_{(i,j)\in\mathcal{E}} r_{i,j}$. By subtracting this quantity from the payoff matrices of each player, this can be transformed into a zero-sum polymatrix game.

Despite much work on (asymptotic) convergence guarantees for normal-form games, there are relatively few results on the iteration complexity of best-response dynamics. The iteration complexity, for a given tolerance level $\epsilon > 0$ refers to the number of rounds $K(\epsilon)$ required to obtain strategies that form an ϵ-optimal Nash equilibrium. Of particular relevance is the recent work of Chen et al. [15], which considered a smoothed best-response type dynamics for two-player zero-sum normal-form and Markov games. They established explicit bounds on the iteration complexity, but in the absence of an additional exploration device—such as mixing with the uniform distribution—their bounds on $K(\epsilon)$ scale exponentially in $(1/\epsilon)$.

1.1. Our contributions.

In this paper, we study smoothed best-response type dynamics for zero-sum polymatrix games. We analyze their behavior in both a *full information* setting, as well as in the *minimal information* setting. In the latter setting, also referred to as the radically uncoupled setting or the bandit setting, each player observes only their own realized payoff at each round, and has no other information about either the game or their opponents' play. Focusing on smoothed best-response dynamics without any modifications to encourage additional exploration, we prove that the number of iterations $K(\epsilon)$ required to converge to ϵ-optimal Nash equilibrium scales polynomially in the ratio $1/\epsilon$ for both information settings. To the best of our knowledge, this is the first known polynomial-time guarantee for best-response type dynamics in the minimal information setting without the introduction of additional exploration in the learning dynamics.

To be clear, our focus in this paper is on the properties of natural best-response type dynamics, and *not* on developing alternative—and possibly more efficient—algorithms for computing Nash equilibria. By searching more broadly in the space of updates, it can be possible to obtain faster iteration complexities (e.g., see the papers [3, 10] for results of this type). However, best response dynamics are a classical and arguably natural form of interaction between a collection of agents. By providing convergence guarantees for best response, we give evidence for the emergence of Nash equilibria as efficiently achievable steady-state behavior of a collection of agents who each act greedily at each step while playing a zero-sum polymatrix game.

In more detail, our first main result applies to the simpler case of full information, where players can perform smoothed best-response updates based on knowledge of the opponents' mixed strategy. We introduce a modified version of Lyapunov functions used in previous work [15, 20], and use it to prove that the dynamics converge to an ϵ-Nash equilibrium in $K(\epsilon) \asymp (1/\epsilon^2)$ iterations. In contrast to previous analysis [15]—which led to exponential dependence in $(1/\epsilon)$—our modification has desirable smoothness properties that allow us to establish the claimed polynomial scaling in $1/\epsilon$.

We then turn to the minimal information setting, in which players do not observe their opponents' play, but instead only observe their realized payoffs at each round. For this more challenging setting, we analyze a simple best-response type learning dynamics involving updates on two timescales. First, on the faster timescale, each player maintains and updates an estimate of their local payoff function or q-value—i.e., the average payoff as a function of their actions. The local payoff function carries information about the opponents' strategy. At the same time, on a slower timescale, the players also modify their strategies by forming a smoothed best-response based on the estimated q-values. The updates of the q-values are in the spirit of TD learning [39], and involve an adaptive learning rate [27] that ensures unbiased estimates of the local payoff function. When recast as a form of stochastic approximation [5], this unbiased property means that the estimated q-values are updated using zero-mean noisy estimates of the underlying payoff function.

However, a major challenge is that the noise variance explodes as the player's strategies approach the boundary of the probability simplex. We overcome this variance explosion—without any modifications to the updates—by explicitly tracking how quickly the strategies approach the boundary. Combined with careful design of Lyapunov functions with favorable smoothness properties, we prove an upper bound on the iteration complexity that scales as $(1/\epsilon)^{8+\nu}$, where the offset $\nu > 0$ can be chosen arbitrarily close to zero at the price of growth in constant pre-factors.

Finally, for both information settings, our bounds on the iteration complexity show a dependence on the underlying graph and the pairwise games. In the worst case, it scales with the maximum degree of the graph but depending on the graph structure and the nature of the edge-wise games, it can exhibit a much milder dependence.

1.2. Related work. So as to put our contributions in context, we now turn to a discussion of past work on best-response type dynamics and the minimal information setting. To be clear, this overview is far from comprehensive: we limit our discussion to the literature most closely related to our work.

There is a long line of work on the convergence of fictitious play (FP) for various classes of finite normal-form games; all of the classical work assumes knowledge of the payoff matrices. In more detail, following the introduction of fictitious play [8, 9], Robinson [34] established the convergence of both the payoffs and strategies of discrete-time FP applied to two-player zero-sum games. Later work by Shapiro [36] proved the payoffs converge at least as quickly as $k^{-1/(|\mathcal{A}^1|+|\mathcal{A}^2|-2)}$, where $\mathcal{A}^1, \mathcal{A}^2$ are the action sets of each player. Karlin [24] conjectured that the actual convergence rate was $k^{-1/2}$, but it was later shown that this conjecture does not hold for general tie-breaking rules [1, 17]. Miyasawa [28] established FP convergence for two-person non-zero-sum games in which players have at most two actions and satisfy a notion of non-degeneracy; other classes known to have convergence properties under FP dynamics include common interest games [29] and weighted-potential games [30]. On the negative side, Shapley [37] provided a two-person non-zero-sum game with three actions for which FP fails to converge; moreover, other counterexamples have been given for certain coordination games [18].

Harris [20] studied the continuous-time version of fictitious play, and used a Lyapunov function argument to show that it converges at the rate $(1/t)$ for two-player zero-sum games. Our analysis makes use of a modification of the Lyapunov function from this paper. Smoothed versions of fictitious play—in which actual best-responses are replaced by smoothed versions—have been studied in various papers [19,22]; we also analyze a smoothed version in this paper. In our work, we focus on best-response type dynamics based on fictitious play since they are a well-studied model of learning for myopic agents. We note that recent literature has studied various other dynamics to compute equilibria of different classes of games, including gradient descent, mirror descent and its variants (e.g., [38,42]); regret matching [21]; extragradient versions of multiplicative updates [12]; as well as various regret-based and online learning methods (e.g., [13,14,16,33,40]).

There is also a more recent and evolving line of work on the minimal information setting for the class of two-player zero-sum games. Leslie and Collins [27] introduced the use of adaptive stepsizes to estimate the average payoffs of each player, and established its asymptotic convergence. Our algorithm includes updates of this type, and the adaptive stepsizes ensure a key unbiasedness property. In addition, inspired by the papers [15,26], we also make use of a *doubly-smoothed* best-response to update the strategies where in addition to a smoothed best-response, we use a learning rate to mix with the current strategy estimate. Our set-up has connections with the entropy-regularized algorithms studied in the paper [4]. Their analysis is predicated upon availability of the exact payoff functions, as well as the opponent's actions, whereas by contrast, we assume only availability of the random payoff resulting from the unobserved actions. The paper [15] also analyzes this form of minimal information; while their main focus is on stochastic games, they also establish a bound on iteration complexity for two player zero-sum matrix games. However, they do not make use of adaptive stepsizes [27], and provide guarantees relative to ϵ-Nash equilibrium up to a smoothing bias. Incorporating the smoothing bias means that the end-to-end guarantees grow exponentially in the inverse tolerance $(1/\epsilon)$, whereas one of our main contributions is to provide schemes (and analysis) with polynomial scaling.

Some recent work [11,32] has also provided guarantees with polynomial dependence on $(1/\epsilon)$ for the minimal information setting, but for schemes that depart from the usual best-response learning dynamics. In the paper [11], exploration is enforced by explicitly limiting how quickly strategies are allowed to approach the boundary of the probability simplex; the work [15] also notes that mixing with a uniform distribution can alleviate these issues. On the other hand, the analysis in the paper [32] regularizes using the Tsallis entropy, resulting in a non-standard strategy update.

It is also possible to exploit linear programming methods so as to compute the equilibria of zero-sum polymatrix games (cf. the papers [6,7,10,23]). Leonardos et al. [25] studies the continuous-time version of our full information dynamics. In other related work in the full information setting, Ao et al. [3] provide finite-time guarantees for zero-sum polymatrix games, in particular by analyzing the finite-time convergence of the optimistic multiplicative weights (OMWU) method. They show that the rate of convergence of their algorithm scales with the maximum degree of the graph.

2. Background and problem set-up. We now provide background along with a more precise set-up of the problem. We begin in Subsection 2.1 with notation and the basic formalism of zero-sum polymatrix games. In Subsection 2.2, we discuss various types of learning dynamics, including the two sets of updates (for full and minimal information respectively) that we study in this paper.

2.1. Zero-sum polymatrix games. A zero-sum polymatrix game with N players is defined by an underlying graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with $\mathcal{N} = \{1, \dots, N\}$. Each node $i \in \mathcal{N}$ corresponds to a player who has access to a finite set $\mathcal{A}^i$. Let $\boldsymbol{\mathcal{A}} = \otimes_{i=1}^{n} \mathcal{A}^i$ represent the joint action set and let $A_{\max} = \max\limits_{i \in \mathcal{N}} |\mathcal{A}^i|$ be the maximum number of actions across all the players. Each edge (i,j) of the graph is associated with a pairwise game between players i and j, with the matrix $R^{(i,j)} \in \mathbb{R}^{|\mathcal{A}_i| \times |\mathcal{A}_j|}$ representing the payoff matrix of player i in the matrix game played by players i and j. If there is no edge between these two players, i.e., $(i,j) \notin \mathcal{E}$, then $R^{(i,j)}$ is the zero matrix of appropriate dimensions. We assume throughout that the payoff matrices are normalized so that $\max_{a^i, a^j} |R^{(i,j)}(a^i, a^j)| \leq 1$. We compile the pairwise payoff matrices into a block matrix $\boldsymbol{R}$ whose $(i,j)^{\text{th}}$ block is given by $R^{(i,j)}$.

At each discrete time instant, all players simultaneously choose a strategy to apply across all of their pairwise interactions. Players may randomize their choice of actions, so for each player $i \in \mathcal{N}$, we denote their set of possible strategies by Π^i, corresponding to the probability simplex supported on $\mathcal{A}^i$. We use $\mathbf{\Pi} := \otimes_{i=1}^{n} \Pi^i$ to denote the set of all possible joint strategies that can be played at each time instant. For any $i \in \mathcal{N}$, we use π^{-i} to denote the collection of all strategies indexed by players $j \in \mathcal{N} \backslash \{i\}$; it belongs to the Cartesian product of simplices denoted by Π^{-i}. For any strategy profile $\pi^{-i} \in \Pi^{-i}$, we denote the vector of expected payoffs of player i for each action they play by $q^i(\pi^{-i}) \in \mathbb{R}^{|\mathcal{A}^i|}$, and note the relationship

$$q^i(\pi^{-i}) = \sum_{j \in \mathcal{N} \backslash \{i\}} R^{(i,j)} \pi^j = \sum_{j \in \mathcal{N}} R^{(i,j)} \pi^j,$$

i.e., the sum of the expected payoffs from the matrix games that i plays on each of the edges connected to i. At times, we use $\boldsymbol{q}(\boldsymbol{\pi}) = (q^i(\pi^{-i}))_{i \in \mathcal{N}}$ to refer to the collection of average payoffs of each player under $\boldsymbol{\pi}$. The zero-sum constraint on the polymatrix game means that the sum of the payoffs across all players equals zero—that is

$$\sum_{i \in \mathcal{N}} (\pi^i)^\top q^i(\pi^{-i}) = \boldsymbol{\pi}^\top \boldsymbol{R} \boldsymbol{\pi} = 0 \qquad \text{for any } \boldsymbol{\pi} \in \mathbf{\Pi}. \tag{2.1}$$

A joint strategy $\overline{\boldsymbol{\pi}} \in \mathbf{\Pi}$ is said to be a *Nash equilibrium* if

$$(\overline{\pi}^i)^\top q^i(\overline{\pi}^{-i}) \geq (\pi^i)^\top q^i(\overline{\pi}^{-i}), \quad \text{for all players } i \in \mathcal{N}, \text{ and } \pi^i \in \Pi^i. \tag{2.2}$$

Nash's existence theorem [31] ensures the existence of a Nash equilibrium in any finite N-player game. When $N = 2$, the class of zero-sum polymatrix games reduces to the class of two-player zero-sum games, in which case the existence of a Nash equilibrium also follows from Von Neumann's minimax theorem [41]. If for some $\epsilon > 0$ the inequalities (2.2) hold up to an ϵ-additive relaxation for some $\overline{\boldsymbol{\pi}}_\epsilon$, then $\overline{\boldsymbol{\pi}}_\epsilon$ is called an *ϵ-Nash equilibrium.*

Nash gap. Let us now introduce a measure of the closeness to a Nash equilibrium (or NE for short). For any strategy $\boldsymbol{\pi} \in \boldsymbol{\Delta}$, the *Nash gap* is defined as

$$\text{NG}(\boldsymbol{\pi}) = \sum_{i \in \mathcal{N}} \left\{ \max_{\hat{\pi} \in \Pi^i} \left(\hat{\pi} - \pi^i\right)^\top q^i(\pi^{-i}) \right\}. \tag{2.3}$$

Observe that any mixed strategy $\boldsymbol{\pi} \in \boldsymbol{\Delta}$ with Nash gap bounded as $\mathrm{NG}(\boldsymbol{\pi}) \leq \epsilon$ is guaranteed to be an *ϵ-Nash equilibrium*, in the sense that it satisfies the NE inequalities (2.3) up to an additive offset of ϵ. Our goal is to show that natural best-response type dynamics that have been studied in the literature can be used to recover an ϵ-Nash equilibrium in time polynomial in $1/\epsilon$ even when players cannot observe the strategies used by the other players.

2.2. Learning dynamics. We now describe and provide intuition for the different learning dynamics we analyze in this paper. We begin with the smoothed best-response and then discuss a form of best-response dynamics for the full and minimal information settings (see 2.1 and Algorithm 2.2).

2.2.1. The smoothed best-response. Recall the definition of $q^i(\pi^{-i})$ as the vector of expected payoffs of player i when all the players except for player i choose their strategy to be π^{-i}. The *best-response function* for player i maps a strategy profile π^{-i} of players other than i to a strategy for player i via

$$\mathrm{br}^i(\pi^{-i}) \in \arg\max_{\hat{\pi}\in\Pi^i} \; \hat{\pi}^\top q^i(\pi^{-i}). \tag{2.4a}$$

In this paper, we study a smoothed variant of the best-response that is indexed by a *regularization parameter* $\tau > 0$. For a strategy $\pi^i \in \Pi^i$, we define the *Shannon entropy* function

$$H(\pi^i) := -\sum_{a^i\in\mathcal{A}^i} \pi^i(a^i)\log \pi^i(a^i). \tag{2.4b}$$

Now suppose that rather than the pure best-response, player i instead plays the *τ-regularized best-response*

$$\sigma^i_\tau(\pi^{-i}) := \arg\max_{\hat{\pi}\in\Pi^i} \big\{\hat{\pi}^\top q^i(\pi^{-i}) + \tau H(\hat{\pi})\big\}, \tag{2.4c}$$

where $\tau > 0$ is a smoothing parameter. In discrete choice theory, for a given π^{-i}, $\sigma^i_\tau(\pi^{-i})$ is the strategy that player i would play if their payoffs had been perturbed by Gumbel-distributed noise before choosing the action with the maximum payoff. Such a perturbation to the payoffs can be used to model an outsider's uncertainty about player i's unobserved preferences [2]. Also called the logit choice model, the τ-regularized best response (2.4c) has been studied extensively in the learning-in-games literature, particularly in connection with the fictitious play paradigm, where players iteratively update their strategies based on estimated opponent behavior [19, 22].

From an algorithmic perspective, there are several benefits associated with using such a smoothed best-response as well. The update rule is unique: the optimization problem (2.4c) admits the closed-form expression

$$\sigma^i_\tau(\pi^{-i})(a^i) = \frac{e^{q^i(\pi^{-i})(a^i)/\tau}}{\sum_{a\in\mathcal{A}^i} e^{q^i(\pi^{-i})(a)/\tau}} \qquad \text{for each } a^i \in \mathcal{A}^i. \tag{2.4d}$$

Consequently, for any $\tau > 0$, the strategy assigns strictly positive mass to each action, thereby ensuring exploration. Moreover, as $\tau \to 0^+$, the smoothed best-response $\sigma^i_\tau(\pi^{-i})$ converges to the best-response $\mathrm{br}(\pi^{-i})$.

Using a smoothed best-response leads to a natural relaxation of a Nash equilibrium. We say that a mixed strategy $\boldsymbol{\pi}$ is a *τ-regularized Nash equilibrium* if

$$\pi^i = \sigma^i_\tau(\pi^{-i}), \; i \in \mathcal{N}. \tag{2.5}$$

The existence of a τ-regularized NE follows from the Brouwer fixed-point theorem.[1] Moreover, it can be shown that a τ-regularized equilibrium is an approximate Nash equilibrium in a precise sense: given any tolerance $\epsilon > 0$, setting $\tau \leq \epsilon/(N \log A_{\max})$ ensures that any τ-regularized Nash equilibrium has Nash gap (2.3) at most ϵ, i.e., a τ-regularized Nash equilibrium is an ϵ-Nash equilibrium. This notion of a τ-regularized equilibrium plays a central role in our dynamics and analysis.

We now turn to the two classes of smoothed best-response dynamics that we study.

2.2.2. Dynamics for full information.

We begin by analyzing smoothed best-response updates in the simpler full information setting. In this case, players have access to the payoff matrices on every edge, and observe the other players' mixed strategies at all times. The classical best-response dynamics in continuous time, given by $\dot{\pi}^i_t \in \mathrm{br}(\pi^{-i}_t) - \pi^i_t$ for each $i \in \mathcal{N}$. The set of Nash equilibria of the underlying game coincides with the equilibrium points of these dynamics. Since each player has access to their opponents' previous strategies in this information setting, player i employs the best-response type update $\pi^i_{k+1} = \pi^i_k + \frac{1}{k+1}\big(\mathrm{br}^i(q^i(\pi^{-i}_k)) - {\pi^i}_k\big)$, where $\{\beta_k\}_{k\geq 1} \subset (0,1)$ is a sequence of stepsizes. This can be interpreted as a discretization of the continuous-time best-response dynamics.

For the reasons outlined in Subsection 2.2.1, we study the closely related τ-smoothed best response dynamics given by

$$\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k + \beta_k \left(\sigma_\tau(\boldsymbol{\pi}_k) - \boldsymbol{\pi}_k\right) \qquad \text{where } \sigma_\tau(\boldsymbol{\pi}_k) = (\sigma^i_\tau(\pi^{-i}_k))_{i\in\mathcal{N}}, \tag{2.6}$$

where $\tau > 0$ is an algorithmic parameter. These updates correspond to a damped version of the usual operator power method[2] attempting to find a τ-regularized Nash equilibrium (2.5) and hence can be considered an approximation of a greedy best-response. These updates also have an interpretation involving players' beliefs as in stochastic fictitious play (see the paper [22] for more details). We give a pseudocode specification of the full information dynamics in Algorithm 2.1.

Algorithm 2.1 Learning dynamics for full information

```
Input: $K$, ${\pi^i}_1 \sim \mathrm{Unif}(\mathcal{A}^i)$, temperature $\tau$ and stepsizes $\{\beta_i\}_{i=1}^{\infty}$
  for $k = 1, \ldots, K$ do
      $\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k + \beta_k(\sigma_\tau(\boldsymbol{\pi}_k) - \boldsymbol{\pi}_k)$
  end for
Output: $\boldsymbol{\pi}_{K+1}$
```

2.2.3. Dynamics for minimal information.

In the minimal information setting, players only have access to the realized payoffs of their own actions. Notably, they do not observe the opponents' strategies or actions and are not privy to any of the payoff matrices. If each player j chooses action $a^j \in \mathcal{A}^j$, player i observes $\sum_{j\in\mathcal{N}} R^{(i,j)}(a^i, a^j)$, the sum of the payoffs from each of the pairwise games that they

[1] In fact, for each $\tau > 0$, there is a unique strategy satisfying condition (2.5); see Proposition SM6.1 in the sequel for details.

[2] For sufficiently large values of τ, the operator σ_τ is a contraction, in which case convergence of the update $\boldsymbol{\pi}_{k+1} = \sigma_\tau(\boldsymbol{\pi}_k)$ follows easily. However, given that we use τ-regularized NE as an approximation to NE, our focus is the more challenging small τ-regime, in which this contractive properly need not hold.

play. Consequently, it is not possible to directly compute the exact expected payoff vector for every joint strategy.

In view of these restrictions, a natural solution (e.g., see the papers [15, 27, 35]) involves each player estimating the expected payoff vector for the strategy chosen by their opponents in the previous round. In particular, if action $\boldsymbol{a}$ is chosen according to the joint strategy $\boldsymbol{\pi}$, then (conditionally on $\boldsymbol{\pi}$) the random variable $\frac{\mathbb{I}(a^i=a)}{\pi^i(a^i)} \sum_{j\in\mathcal{N}} R^{(i,j)}(a^i, a^j)$ is an unbiased estimate of the average payoff $q^i(\pi^{-i})(a)$ for each $a \in \mathcal{A}^i$. We use this idea to sequentially build approximations of the expected payoff vectors $\boldsymbol{q}(\boldsymbol{\pi}_k)$ for the joint strategy $\boldsymbol{\pi}_k$ employed in round k. Note that conditioned on past history, each player plays an independent strategy at each step in this setting.

For each player $i \in \mathcal{N}$ and action $a^i \in \mathcal{A}^i$, let $e^i(a^i) \in \mathbb{R}^{|\mathcal{A}^i|}$ be the standard basis vector with a one in the position indexed by $a^i \in \mathcal{A}^i$. Consider the updates

$$q^i_{k+1} = q^i_k - \alpha_k \, \frac{e^i(a^i_k)}{\pi^i_k(a^i_k)} \big(\sum_{j\in\mathcal{N}} R^{(i,j)}(a^i_k, a^j_k) - q^i_k(a^i_k) \big) \qquad \text{for } k = 1, 2, \ldots, \tag{2.7}$$

where $\alpha_k > 0$ denotes a positive stepsize and a^j_k is the action chosen by player $j \in \mathcal{N}$ according to their strategy π^j_k. This is an asynchronous update rule where only the component corresponding to the action a^i_k is updated. Note that in order to compute this update, the only additional information required by player i is the payoff $\sum_{j\in\mathcal{N}} R^{(i,j)}(a^i_k, a^j_k)$ received by playing action a^i_k. We re-iterate that player i does *not* require knowledge of the payoff matrices nor the actions of their opponent.

Equivalently, the update (2.7) can be rewritten as

$$q^i_{k+1} \;=\; q^i_k - \alpha_k \, \frac{E^i(a^i_k)}{\pi^i_k(a^i_k)} \Big\{ \sum_{j\in\mathcal{N}} R^{(i,j)} \boldsymbol{e}^j(a^j_k) - q^i_k \Big\},$$

where $E^i(a^i_k) := e^i(a^i_k) e^i(a^i_k)^\top$, i.e., the matrix with zeros in all entries except for a one in diagonal entry (a^i_k, a^i_k). By concatenating the iterates of all players as $\boldsymbol{q}_k = (q^i_k)_{i\in\mathcal{N}}$, we have the combined update rule

$$\boldsymbol{q}_{k+1} = \boldsymbol{q}_k - \alpha_k \, \frac{\boldsymbol{E}(\boldsymbol{a}_k)}{\boldsymbol{\pi}_k(\boldsymbol{a}_k)} \big(\boldsymbol{R} \, \boldsymbol{e}(\boldsymbol{a}_k) - \boldsymbol{q}_k \big) \qquad \text{for } k = 1, 2, \ldots, \tag{2.8a}$$

where $\boldsymbol{a}_k = (a^i_k)_{i\in\mathcal{N}}$ denotes the action profile sampled from $\boldsymbol{\pi}_k$, i.e., for each $i \in \mathcal{N}$, a^i_k is sampled according to π^i_k, $\boldsymbol{e}(\boldsymbol{a}_k) = (e^i(a^i_k))_{i\in\mathcal{N}}$ is the concatenation of the action vectors of all players and $\boldsymbol{E}(\boldsymbol{a}_k)/\boldsymbol{\pi}_k(\boldsymbol{a}_k)$ is a diagonal block matrix whose i^{th} diagonal block is $E^i(a^i_k)/\pi^i_k(a^i_k)$ for every $i \in \mathcal{N}$.

Since the update (2.8a) is in the spirit of TD Learning [15, 39], we refer to it as a TD update. The rescaling of the stepsize by $\pi^i(a^i_k)$ for each player i in this update was originally proposed by Leslie and Collins [27]. It can be understood as a form of importance reweighting designed to ensure that $\mathbb{E}\big[q^i_{k+1} - q^i_k \mid \boldsymbol{\pi}_k, q^i_k\big] = \alpha_k(q^i_k - q^i(\pi^{-i}_k))$, so that in expectation, the updates are synchronous, and for a fixed constant strategy $\boldsymbol{\pi}_k \equiv \tilde{\boldsymbol{\pi}}$, the TD updates have $\boldsymbol{q}(\tilde{\boldsymbol{\pi}})$ as a fixed point. We often refer to the iterates of the TD updates as q-values.

In parallel with the TD updates for the expected payoff vectors, the players update their strategies according to the smoothed best-response dynamics

$$\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k + \beta_k \big(\sigma_\tau(\boldsymbol{q}_k) - \boldsymbol{\pi}_k \big) \qquad \text{where } \sigma^i_\tau(\boldsymbol{q}_k) = (\sigma^i_\tau(q^i_k))_{i\in\mathcal{N}}. \tag{2.8b}$$

Algorithm 2.2 Learning dynamics for the minimal information setting

```
Input: $K$, $\boldsymbol{q}_1 = \mathbf{0} \in \otimes_{i\in\mathcal{N}} \mathbb{R}^{|\mathcal{A}^i|}$, $\pi_1^i \sim \mathrm{Unif}(\mathcal{A}^i), \tau, c_{\alpha,\beta}, \{\beta_i\}_{i=1}^{\infty}$
  for $k = 1, \ldots, K$ do
    Play $a_k^i \sim \pi_k^i$ independently for $i \in \mathcal{N}$
    $\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k + \beta_k \left(\sigma_\tau(\boldsymbol{q}_k) - \boldsymbol{\pi}_k\right)$
    $\boldsymbol{q}_{k+1} = \boldsymbol{q}_k - \alpha_k \frac{\boldsymbol{E}(\boldsymbol{a}_k)}{\boldsymbol{\pi}_k(\boldsymbol{a}_k)}\big(\boldsymbol{R}\,\boldsymbol{e}(\boldsymbol{a}_k) - \boldsymbol{q}_k\big)$
  end for
Output: $\boldsymbol{\pi}_{K+1}$
```

The smoothed best-response plays an important role here: it ensures that each action is played infinitely often as is required for TD convergence [39]. As noted earlier, smoothed best-response updates ensure that this property holds. See Algorithm 2.2 for a pseudocode description of our dynamics for the minimal information setting.

We note that the TD updates combined with the smoothed best-response dynamics form a coupled two-timescale stochastic approximation scheme (e.g., [5]). In such schemes, the iterates of the update with the larger stepsize (often referred to as the "faster" timescale) has a fixed point that depends on the current value of the iterates on the "slower" timescale, which evolves with a smaller stepsize. The larger stepsize enables the faster-timescale iterates to effectively "track" their moving fixed point. Similarly, the strategies are updated slower than the q-values so that the TD updates can track their fixed point. We achieve this timescale separation by setting $\beta_k = c_{\alpha,\beta}\alpha_k$ for some scalar $c_{\alpha,\beta} \in (0,1)$ to be specified in our analysis.

2.3. Lyapunov function. In order to analyze the learning dynamics in Algorithm 2.1 and Algorithm 2.2, we make use of a novel Lyapunov function for the strategies, given by

$$\mathcal{V}(\boldsymbol{\pi}) := \sum_{i=1}^{N} \max_{\hat{\pi}\in\Pi^i} \left\{ \hat{\pi}^\top q^i(\pi^{-i}) + \tau H(\hat{\pi}) \right\}, \tag{2.9}$$

where H denotes the Shannon entropy function (2.4b). Note that the Lyapunov function $\mathcal{V}$ has a natural game-theoretic interpretation: it corresponds to the sum of the average payoffs when each player chooses the best-response to the entropy-regularized payoffs. It can be viewed as a regularized version of the function used by Harris [20] to prove finite-time guarantees for continuous-time zero-sum games. The main purpose of our Lyapunov function (2.9) is to provide a way to bound the Nash gap from above in our proofs; more precisely, they are related by the inequality

$$\mathrm{NG}(\boldsymbol{\pi}) \leq \mathcal{V}(\boldsymbol{\pi}). \tag{2.10}$$

The Lyapunov function $\mathcal{V}$ differs from other Lyapunov functions used in the literature [15, 20] in some crucial ways. Its smoothness properties enable us to prove a polynomial-time finite-sample guarantee (as opposed to the exponential-time guarantee in Chen et al. [15]). We discuss this issue in further detail in the supplementary Section SM6.

In addition, our proof also involves the *q-Lyapunov function*

$$\mathcal{W}(\boldsymbol{\pi}, \boldsymbol{q}) := \sum_{i\in\mathcal{N}} \left\| q^i - q^i(\pi^{-i}) \right\|_2^2, \tag{2.11}$$

which we use to track the quality of q^i as an estimate of $q^i(\pi^{-i})$. The *total Lyapunov function* $\mathcal{T}$ for our analysis is given by the sum

$$\mathcal{T}(\boldsymbol{\pi}, \boldsymbol{q}) := \mathcal{V}(\boldsymbol{\pi}) + \mathcal{W}(\boldsymbol{\pi}, \boldsymbol{q}). \tag{2.12}$$

3. Main results for zero-sum polymatrix games.

In this section, we state our results for zero-sum polymatrix games for both information settings. Subsection 3.1 provides guarantees for the full information dynamics (cf. Algorithm 2.1) whereas in Subsection 3.2, we study the minimal information two-timescale dynamics (cf. Algorithm 2.2). In stating and proving these results, al variables of the form $c_i, i \in \mathbb{N}$ denote numerical constants that are independent of both the game and the learning dynamics. Our results can also be specialized for the class of two-player zero-sum games, with more detail given in the supplementary Section SM5.1.

3.1. Full information.

In this section, we provide finite-sample guarantees for the full information dynamics specified in Algorithm 2.1. Theorem 3.1 provides an upper bound on the Nash gap for two choices of stepsize schedules. In stating these claims, we make use of the shorthand $V_1 := \mathcal{V}(\boldsymbol{\pi}_1)$ for the initial value of the Lyapunov function $\mathcal{V}$ from equation (2.9), where $\boldsymbol{\pi}_1$ is the initial set of mixed strategies.

Theorem 3.1 (Nash gap finite-sample guarantees). *For any $\tau > 0$, consider the full information dynamics (Algorithm 2.1) initialized with $\boldsymbol{\pi}_1$. Then the Nash gap after K iterations is bounded as:*
(a) For a constant stepsize sequence $\beta_k \equiv \beta \in (0, 1)$, we have

$$\mathrm{NG}(\boldsymbol{\pi}_{K+1}) \leq (1-\beta)^K V_1 + \tau N \log A_{\max} + \frac{N \|\boldsymbol{R}\|_2^2 \beta}{\tau}.$$

(b) For the inverse linear stepsize sequence $\beta_k = \frac{\beta}{k}$ for some $\beta \in (1, 2]$, we have

$$\mathrm{NG}(\boldsymbol{\pi}_{K+1}) \leq \frac{V_1}{(K+1)^\beta} + 8N\tau \log A_{\max} + \frac{4N\|\boldsymbol{R}\|_2^2 \beta^2}{\tau(\beta-1)K}.$$

For both stepsizes, the first term in the upper bound involves the initial Lyapunov value V_1, and so reflects the rate at which the algorithm “forgets” its initialization as it converges. The second term in each bound scales linearly in τ, and corresponds to a form of bias introduced by the players using a τ-regularized best-response instead of an actual best-response. The third term in each upper bound scales with $(1/\tau)$, which is a measure of how smooth the τ-regularized best-response is.

For the purposes of interpretation, it is useful to derive bounds on on the iteration complexity of the procedure. For a given level $\epsilon > 0$, the *iteration complexity* $K(\epsilon)$ is the minimum number of iterations required to ensure that $\mathrm{NG}(\boldsymbol{\pi}_{K(\epsilon)+1}) \leq \epsilon$. In order to obtain explicit bounds on $K(\epsilon)$, we choose the temperature parameter τ and stepsizes so as to ensure that after $K(\epsilon)$ rounds, each of the three terms in the upper bound in Theorem 3.1 is at most $\frac{\epsilon}{3}$. By doing so, we obtain the following:

Corollary 3.2. *Consider the full information dynamics (Algorithm 2.1) initialized with $\tau = c_1\epsilon/(N \log A_{\max})$. Then the iteration complexity $K(\epsilon)$ is bounded as follows:*

(a) For the constant stepsizes $\beta_k \equiv \beta := \epsilon^2/(c_2 N^2 \|\boldsymbol{R}\|_2^2 \log A_{\max})$, we have

$$K(\epsilon) \leq \frac{c_3 N^2 \|\boldsymbol{R}\|_2^2 \log A_{\max}}{\epsilon^2} \log\big(\frac{V_1}{\epsilon}\big).$$

(b) For the inverse linear decay $\beta_k = \beta/k$ for some $\beta \in (1, 2]$, we have

$$K(\epsilon) \leq \frac{c_4 N^2 \|\boldsymbol{R}\|_2^2 \beta^2 \log A_{\max} V_1}{(\beta - 1)\epsilon^2}.$$

Focusing on the triple $(N, \|\boldsymbol{R}\|_2, \epsilon)$, our theory guarantees that smooth best-response dynamics has iteration complexity bounded as $\mathcal{O}(\|\boldsymbol{R}\|_2^2 N^2/\epsilon^2)$. As mentioned previously, if one allows for different types of algorithms, it can be possible to obtain different scalings of the iteration complexity. For example, Ao et al. [3] studied updates based on a multiplicative weights (MW) update, and analyzed its behavior in terms of the maximum error over all players, as opposed to the sum of errors (2.3) in our analysis. For this error metric and algorithm, they obtained an iteration complexity scaling as as $\mathcal{O}(d_{\max}(\mathcal{G})\|\boldsymbol{R}\|_{\max}/\epsilon)$, where $\|\boldsymbol{R}\|_{\max}$ is the maximum absolute element of the matrix $\boldsymbol{R}$, and $d_{\max}(\mathcal{G})$ is the maximum degree of the graph. Since $d_{\max}(\mathcal{G}) \leq N$ and $\|\boldsymbol{R}\|_{\max} \leq \|\boldsymbol{R}\|_2$, this MW iteration complexity—albeit for a different error metric—represents a quadratic improvement over that of smoothed best-response. Apart from the error metrics (which differ by a factor of N in an extreme case), we attribute this gap to the linear nature of our best-response updates as opposed to their multiplicative updates, which are known to work well with KL divergence-based Lyapunov functions.

It is notable that our bounds scale (quadratically) in the spectral norm $\|\boldsymbol{R}\|_2$, a quantity which depends on subtle ways on both the graph structure and the pairwise games on each edge. In this way, our analysis affords some insight into the type of zero-sum polymatrix games for which it is easier to converge to Nash equilibria. We discuss the structural properties of the parameter $\|\boldsymbol{R}\|_2$ in further detail in Subsection 3.3. Additionally, we also discuss an inverse polynomial stepsize schedule in Subsection SM1.1.

3.2. Minimal information. We now turn to the analysis of the updates in Algorithm 2.2 that apply to the minimal information setting. Since this is a stochastic algorithm, our bounds apply to the *iteration complexity* $K(\epsilon)$ defined by the minimum number of rounds required to ensure that $\mathbb{E}\big[\,\mathrm{NG}(\boldsymbol{\pi}_{K(\epsilon)+1})\big] \leq \epsilon$. We again show that the iteration complexity scales polynomially in $(1/\epsilon)$–in this case, we can guarantee a scaling of the order $(1/\epsilon)^{8+\nu}$ for an exponent parameter $\nu > 0$ that can be chosen arbitrarily close to zero. The price of taking $\nu \to 0^+$ manifests in the growth of certain pre-factors; we use functions of the form $g(\nu)$ and variants thereof to indicate terms of this type. We make use of the shorthand T_1 for the initial value of the Lyapunov function $\mathcal{T}(\boldsymbol{\pi}_1, \boldsymbol{q}_1)$ from equation (2.12), where $\boldsymbol{\pi}_1$ is the initial set of mixed strategies and $\boldsymbol{q}_1$ is the initial estimate of $\boldsymbol{q}(\boldsymbol{\pi}_1)$.

Our result applies to Algorithm 2.2 where the temperature and the timescale separation constant are set as

$$\tau = \frac{g_\tau(\nu)\epsilon}{N \log A_{\max}} \quad \text{and} \quad c_{\alpha,\beta} = \frac{g_{\alpha,\beta}(\nu)\tau^3}{\|\boldsymbol{R}\|_2^2} \quad \text{, respectively,} \tag{3.2}$$

and the function $g_{\alpha,\beta}$ satisfies the scaling $g_{\alpha,\beta}(\nu) \to 0^+$ as $\nu \to 0^+$. Our guarantees holds in terms of a triple of functions (g_1, g_2, g_3) such that $\max_{j=1,2,3} g_j(\nu) \to 0^+$ as $\nu \to 0^+$.

Theorem 3.3. *Consider the minimal information dynamics (Algorithm* 2.2*) initialized with the parameters* (3.2)*. Then for any $\nu > 0$ and in each of the following cases, the iteration complexity $K(\epsilon)$ satisfies the following upper bounds:*

(a) For the constant stepsize $\beta_k \equiv \beta := \frac{g_1(\nu)\epsilon^{8+\nu}}{A_{\max}^6 N^8 \|\boldsymbol{R}\|_2^6}$*, we have*

$$K(\epsilon) \le \frac{A_{\max}^6 N^8 \|\boldsymbol{R}\|_2^6}{g_2(\nu)\epsilon^{8+\nu}} \log\big(\frac{3T_1}{\epsilon}\big).$$

(b) For the inverse polynomial stepsize $\beta_k = \frac{\beta}{(k+k_0)^\eta}$ *for some exponent* $\eta \in (0,1)$*, offset* $k_0 = \lceil (\frac{2\eta}{\beta})^{1/(1-\eta)} \rceil$*, and* $\beta = \frac{g_1(\nu)\epsilon^{8+\nu}}{A_{\max}^6 N^8 \|\boldsymbol{R}\|_2^6}$*, we have*

$$K(\epsilon) \le \Big\{ \frac{(1-\eta) A_{\max}^6 N^8 \|\boldsymbol{R}\|_2^6}{g_3(\nu)\epsilon^{8+\nu}} \log\big(\frac{3T_1}{\epsilon}\big) \Big\}^{\frac{1}{1-\eta}}.$$

As discussed previously, it is key that all the guarantees in Theorem 3.3 scale polynomially in $(1/\epsilon)$—in particular, as $(1/\epsilon)^{8+\nu}$ along with additional logarithmic factors. To the best of our knowledge, these are the first known polynomial guarantees for the standard best-response dynamics in this setting. One of the main challenges in establishing this polynomial scaling is controlling the variance of the q-updates. This variance depends on the probability of the actions chosen in each round, and choosing actions based on softmax response leads to probabilities that are *exponentially small* in τ. Consequently, a naive approach yields an exponential dependence on $1/\tau$, and hence—since our choice of τ is proportional to the target accuracy ϵ—an exponential dependence on $(1/\epsilon)$. Notably, the analysis in some past work [15] exhibits this type of exponential growth. In our analysis, we resolve this issue by initializing the dynamics away from the boundary, and then choosing the stepsize parameter in a way that allows us to control how quickly the iterates approach the boundary of the probability simplex. See Subsection 4.2 for the details of this argument.

The stepsize choices in Theorem 3.3 depend on the functions g_1 and g_2 of ν, which have explicit expressions given in the proof of Theorem 3.3 in Subsection 4.2. The stepsizes also depend on the action sizes, number of players, and the spectral norm $\|\boldsymbol{R}\|_2$. The latter parameter is global in nature, requiring each player to be aware of the full payoff structure of the polymatrix game. This requirement can be relaxed by bounding $\|\boldsymbol{R}\|_2$ from above: for example, since the block matrix $\boldsymbol{R}$ is formed by concatenating the payoff matrices $R^{(i,j)}$ for game (i,j), we have the bound

$$\|\boldsymbol{R}\|_2 \le \max_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} \|R^{(i,j)}\|_2 \;\le\; d_{\max}(\mathcal{G}) \max_{i,j} \|R^{(i,j)}\|_2 \;\le\; A_{\max} d_{\max}(\mathcal{G}) \|\boldsymbol{R}\|_{\max},$$

where $\|\boldsymbol{R}\|_{\max}$ denotes the maximum absolute entry of the matrix $\boldsymbol{R}$.

3.3. Studying the spectral norm. The spectral norm $\|\boldsymbol{R}\|_2$ captures the dependence of the rate of convergence on the underlying graph and the pairwise games. Studying the parameter $\|\boldsymbol{R}\|_2$ can reveal answers to various qualitative questions of interest. For instance, what types of graphs and pairwise games admit the fastest convergence to Nash equilibria for best-response dynamics?

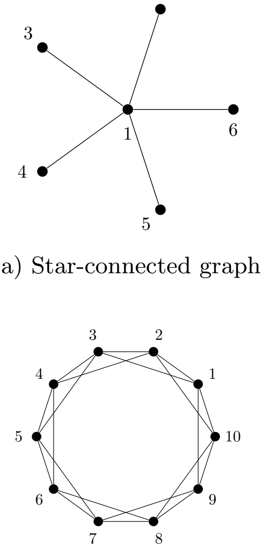
In order to address this question, let us consider the special class of zero-sum polymatrix games in which each edge is associated with the *same* two-player zero-sum game. Given a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, define a skew-symmetric weighted adjacency matrix $\mathbf{A} \in \{-1,0,1\}^{N\times N}$ where the $(i,j)^{\text{th}}$ entry is non-zero only if $(i,j) \in \mathcal{E}$. Letting $\bar{R} \in \mathbb{R}^{m\times m}$ be a symmetric matrix of payoffs, we associate with edge (i,j) the zero-sum game defined by the payoff matrix $A_{ij}\bar{R}$ for player i, and $-A_{ij}\bar{R}$ for player j.

A useful property of this zero-sum polymatrix game is that the block matrix $\boldsymbol{R}$ can be written as $\boldsymbol{R} = \mathbf{A} \otimes \bar{R}$, where $\otimes$ is the Kronecker product. It follows that

$\|\boldsymbol{R}\|_2 = \|\mathbf{A}\|_2\|\bar{R}\|_2$, so that this decomposition separates out the dependency on the game payoff $\bar{R}$ and the graph structure $\mathbf{A}$. We now discuss these two factors in turn.

3.3.1. Graph dependence. Recall the matrix norms $\|\mathbf{A}\|_\infty := \|\mathbf{A}^\top\|_1$ and $\|\mathbf{A}\|_1 := \max_{j\in\mathcal{N}} \sum_{i\in\mathcal{N}} |A_{ij}|$. We always have the elementary upper bound $\|\mathbf{A}\|_2 \leq \max\left\{\|\mathbf{A}\|_1,\ \|\mathbf{A}\|_\infty\right\} \leq d_{\max}(\mathcal{G})$. For graphs with particular structure, this bound can be tightened: we now derive bounds $\|\mathbf{A}\|_2$ that reveal its exact dependence on $d_{\max}(\mathcal{G})$ for two sub-classes of graphs.

Star-connected graph. In this type of graph, a central node—say node 1—is connected to all the other nodes, yielding a graph with maximum degree $d_{\max}(\mathcal{G}) = N-1$. With node 1 as the central node, only the first row and first column of A have non-zero entries. Figure 1a provides an illustration of a star graph with $N=6$ nodes. Introducing the shorthand $\mathbf{M} := \mathbf{A}^\top\mathbf{A}$, we have $\mathbf{M}_{11} = N-1$ and $\mathbf{M}_{1j} = \mathbf{M}_{j1} = 0$ for all $j \neq 1$. Therefore, the matrix $\mathbf{M}$ has $N-1$ as an eigenvalue. The submatrix of $\mathbf{M}$ formed by excluding the first row and column has rank 1 with the only non-zero eigenvalue being equal to $N-1$. It follows that $\|\mathbf{A}\|_2 = \sqrt{N-1}$, i.e., $\|\boldsymbol{R}\|_2$ scales as $\sqrt{d_{\max}(\mathcal{G})}$.



(a) Star-connected graph

(b) 4-regular ring graph

Fig. 1: Graph structures.

k-regular ring graph. In a k-regular ring graph (see Figure 1b for an example), every node is connected to k other nodes. The weighted adjacency matrix $\mathbf{A}$ in this case is circulant, i.e., the entire matrix can be generated by shifting the first row. It is possible to explicitly write down the imaginary eigenvalues of $\mathbf{A}$ and show that $\|\mathbf{A}\|_2$ grows linearly with k, i.e., grows linearly with $d_{\max}(\mathcal{G})$ (see Section SM4 for details).

To summarize, we see that for graphs with a fixed number of players N, the rate of convergence to approximate Nash equilibria of the best-response type dynamics we analyze does scale with the maximum degree of the graph $d_{\max}(\mathcal{G})$. A similar trend is documented in the paper [3] using an algorithm based on the multiplicative weights method; they observed that the iteration complexity for their algorithm grows with $d_{\max}(\mathcal{G})$. Our bounds exhibit a finer dependence on $d_{\max}(\mathcal{G})$ through the quantity $\|\boldsymbol{R}\|_2$ in the sense that it enables us to compare different types of graphs. For instance, based on the preceding analysis, our bounds indicate that for a large enough number of players N, a k-regular graph with $k > \sqrt{N-1}$ converges more slowly than a star-connected graph with the same number of nodes, which in turn converges more slowly than a k-regular graph with $k < \sqrt{N-1}$.

3.3.2. Dependence on pairwise games. Finally, we turn to the effect of the underlying pairwise games on the rate of convergence. Let $\mathbf{A}$ be the skew-symmetric adjacency matrix of a k-regular ring graph, and let $\bar{R}$ be a fixed payoff matrix. Each node $i \in \mathcal{N}$ is associated with k payoff matrices: for some scalar $\rho \in [0,1]$, suppose that at most ρk of these payoff matrices are set equal to $\bar{R}$, whereas the remaining payoffs are set equal to $\omega\bar{R}$ for a weight parameter $\omega \in [0,1]$.

The parameter ρ controls the number of "strong" pairwise matrix games, whereas the parameter ω is a measure of the ratio between weak and strong games. Using the techniques described previously, the spectral norm $\vert\!\vert\!\vert \boldsymbol{R} \vert\!\vert\!\vert_2$ of the block matrix $\boldsymbol{R}$ can be bounded as

$$\vert\!\vert\!\vert \boldsymbol{R} \vert\!\vert\!\vert_2 \leq (\rho k) \vert\!\vert\!\vert \bar{R} \vert\!\vert\!\vert_2 + k(1-\rho) \vert\!\vert\!\vert \omega \bar{R} \vert\!\vert\!\vert_2 \; = \; k \vert\!\vert\!\vert \bar{R} \vert\!\vert\!\vert_2 \big\{ \rho + (1-\rho)\omega \big\}.$$

Thus, the dependence of $\vert\!\vert\!\vert \boldsymbol{R} \vert\!\vert\!\vert_2$ on the maximum degree k can be made arbitrarily small by suitable choices of the two parameters (ρ, ω).

4. Proofs of main results.

This section is to the proofs of our main results, with Subsections 4.1 and 4.2 devoted to the proofs of Theorems 3.1 and 3.3, corresponding to full information and minimal information cases, respectively. Due to space constraints, some technical details are given in the online supplementary material.

4.1. Proof of Theorem 3.1.

It is convenient to introduce the shorthand $V_k := \mathcal{V}(\boldsymbol{\pi}_k)$ for the value of the Lyapunov function at iteration k. All three sub-claims in Theorem 3.1 are derived via a drift inequality for $\mathcal{V}$: in particular, we claim that

$$V_{k+1} \leq (1-\beta_k) V_k + \beta_k \tau N \log A_{\max} + \tfrac{N \vert\!\vert\!\vert \boldsymbol{R} \vert\!\vert\!\vert_2^2 \beta_k^2}{\tau}, \qquad \text{for } k = 1, 2, \ldots. \tag{4.1}$$

See Subsection 4.1.2 for the proof of this claim.

4.1.1. From drift inequality to Nash gap.

Solving the inequality (4.1) for each type of stepsize gives an upper bound on V_{K+1}. Via inequality (2.10), these bounds translate to upper bounds on the Nash gap, and hence the iteration complexity. Here we provide this argument for the constant stepsizes; see Section SM1.2 for the remaining cases.

In the constant stepsize case, we have $\beta_k \equiv \beta \in (0,1)$ for all iterations $k = 1, 2, \ldots$. For this choice, iterating the bound (4.1) K times yields

$$V_{K+1} \leq (1-\beta)^K V_1 + \tau N \log A_{\max} + \frac{4N \vert\!\vert\!\vert \boldsymbol{R} \vert\!\vert\!\vert_2^2 \beta}{\tau}.$$

For a target error $\epsilon \in (0,1)$, setting $\tau = \vert\!\vert\!\vert \boldsymbol{R} \vert\!\vert\!\vert_2 \sqrt{\beta / \log A_{\max}}$ yields the bound

$$V_{K+1} \leq (1-\beta)^K V_1 + 2N \vert\!\vert\!\vert \boldsymbol{R} \vert\!\vert\!\vert_2 (\log A_{\max})^{3/2} \sqrt{\beta}.$$

Thus, if we set $\beta = \epsilon^2 / (16 N^2 \vert\!\vert\!\vert \boldsymbol{R} \vert\!\vert\!\vert_2^2 (\log A_{\max})^3)$, we can conclude that it suffices to take $K(\epsilon) = \frac{16 N^2 \vert\!\vert\!\vert \boldsymbol{R} \vert\!\vert\!\vert_2^2 (\log A_{\max})^3}{\epsilon^2} \log\big(\frac{V_1}{\epsilon}\big)$ iterations in order to ensure that $V_{K(\epsilon)+1} \leq \epsilon$, which in turn results in the Nash gap being less than ϵ.

4.1.2. Proof of drift inequality (4.1).

The following auxiliary result plays a key role in this proof (as well as that of Theorem 3.3).

Lemma 4.1. *For the Lyapunov function $\mathcal{V}$ from equation* (2.9)*:*

(a) It is twice continuously differentiable, and the ℓ_2-operator norm $\vert\!\vert\!\vert \cdot \vert\!\vert\!\vert_2$ of its Hessian is uniformly bounded as

$$\vert\!\vert\!\vert \nabla^2 \mathcal{V}(\boldsymbol{\pi}) \vert\!\vert\!\vert_2 \leq L := \frac{\vert\!\vert\!\vert \boldsymbol{R} \vert\!\vert\!\vert_2^2}{\tau} \qquad \textit{for all } \boldsymbol{\pi} \in \boldsymbol{\Delta}. \tag{4.2a}$$

(b) For all $\boldsymbol{\pi} \in \boldsymbol{\Delta}$, we have the bound

$$\sum_{i=1}^{N} \big\langle \nabla_{\pi^i} \mathcal{V}(\boldsymbol{\pi}), \sigma_\tau(q^i(\pi^{-i})) - \pi^i \big\rangle \leq -\mathcal{V}(\boldsymbol{\pi}) + N \tau \log A_{\max}. \tag{4.2b}$$

See Subsection SM2.1 for the proof of this claim.

Using this auxiliary result, we now prove the claim (4.1). From the Hessian bound (4.2a) on $\mathcal{V}$, we have

$$\underbrace{\mathcal{V}(\boldsymbol{\pi}_{k+1})}_{V_{k+1}} \leq \underbrace{\mathcal{V}(\boldsymbol{\pi}_k)}_{V_k} + \langle \nabla \mathcal{V}(\boldsymbol{\pi}_k),\, \boldsymbol{\pi}_{k+1} - \boldsymbol{\pi}_k \rangle + \frac{L}{2} \|\boldsymbol{\pi}_{k+1} - \boldsymbol{\pi}_k\|_2^2.$$

Using Hölder's inequality, we can upper bound $\|\boldsymbol{\pi}_{k+1} - \boldsymbol{\pi}_k\|_2^2$ by

$$\beta_k^2 \sum_{i=1}^N \|\sigma_\tau(q^i(\pi_k^{-i})) - {\pi^i}_k\|_2^2 \;\leq\; \beta_k^2 \sum_{i=1}^N \|\sigma_\tau(q^i(\pi_k^{-i})) - {\pi^i}_k\|_1^2 \|\sigma_\tau(q^i(\pi_k^{-i})) - {\pi^i}_k\|_\infty^2. \tag{4.3}$$

Since both $\sigma_\tau(R^i \pi_k^{-i})$ and ${\pi^i}_k$ belong to the probability simplex, we have $\|\sigma_\tau(R^i \pi_k^{-i}) - {\pi^i}_k\|_\infty \leq 1$ and $\|\sigma_\tau(R^i \pi_k^{-i}) - {\pi^i}_k\|_1 \leq 2$, whence $\|\boldsymbol{\pi}_{k+1} - \boldsymbol{\pi}_k\|_2^2 \leq 4N\beta_k^2$. Recalling that $\boldsymbol{\pi}_{k+1} - \boldsymbol{\pi}_k = \beta_k\big(\sigma_\tau(\boldsymbol{q}(\boldsymbol{\pi}_k)) - \boldsymbol{\pi}_k\big)$, observe that we have $\langle \nabla \mathcal{V}(\boldsymbol{\pi}_k),\, \boldsymbol{\pi}_{k+1} - \boldsymbol{\pi}_k \rangle = \beta_k \Big\{ \sum_{i=1}^N \big\langle \nabla_{\pi^i} \mathcal{V}(\boldsymbol{\pi}_k), \sigma_\tau(q^i(\pi_k^{-i})) - {\pi^i}_k \big\rangle \Big\}$. Putting together the pieces yields the bound

$$V_{k+1} \leq V_k + \beta_k \Big\{ \sum_{i=1}^N \big\langle \nabla_{\pi^i} \mathcal{V}(\boldsymbol{\pi}), \sigma_\tau(q^i(\pi^{-i})) - \pi^i \big\rangle \Big\} + \frac{4N \|\!|\boldsymbol{R}|\!\|_2^2}{\tau} \beta_k^2. \tag{4.4}$$

It remains to bound the first-order term on the right-hand side of inequality (4.4), and we do so using inequality (4.2b) from Lemma 4.1. Substituting this bound into inequality (4.4) and re-arranging yields the claimed drift inequality (4.1).

4.2. Proof of Theorem 3.3. We now turn to the proof of Theorem 3.3 that applies to the minimal information setting. We proceed in two parts: first, we establish a drift inequality for the Lyapunov functions defined in Subsection 2.3 subject to a condition on how far away the strategy vectors $\boldsymbol{\pi}_k$ are from the boundary of the simplex. This is done to control the variance of the q-updates. Solving this drift inequality gives us a bound on the iteration complexity subject to the aforementioned condition. In the second part of the proof, we show that by choosing certain quantities related to the dynamics appropriately, this condition will be satisfied.

Recall the definition (2.9) of the strategy-tracking Lyapunov function $\mathcal{V}$, the q-Lyapunov function $\mathcal{W}$ (2.11) and the total Lyapunov function $\mathcal{T}$ (2.12). Since our goal is to bound the Nash gap, it is useful to make note of the sandwich relation

$$\mathrm{NG}(\boldsymbol{\pi}) \overset{(i)}{\leq} \mathcal{V}(\boldsymbol{\pi}) \overset{(ii)}{\leq} \mathcal{T}(\boldsymbol{\pi}, \boldsymbol{q}), \tag{4.5}$$

where inequality (i) follows from the bound (2.10) and inequality (ii) follows from the definition of the total Lyapunov function and the non-negativity of $\mathcal{W}(\boldsymbol{\pi}, \boldsymbol{q})$. Thus, in order to bound the Nash gap, it suffices to upper bound $\mathcal{T}(\boldsymbol{\pi}, \boldsymbol{q})$. Also recall that our algorithm generates a sequence $\{\boldsymbol{\pi}_q, \boldsymbol{q}_k\}_{k\geq 1}$ of strategy and q-value pairs. For $k = 1, 2, \ldots$, we introduce the shorthand notation $V_k := \mathcal{V}(\boldsymbol{\pi}_k)$, $W_k := \mathcal{W}(\boldsymbol{\pi}_k, \boldsymbol{q}_k)$, and $T_k := \mathcal{T}(\boldsymbol{\pi}_k, \boldsymbol{q}_k)$ for the values of the three Lyapunov functions as a function of the iteration number k.

4.2.1. Drift inequality for total Lyapunov function $\mathcal{T}$. The first step in proving Theorem 3.3 is to establish a drift inequality (or recursive bound) on the expected values $\mathbb{E}T_k$ of the total Lyapunov function $\mathcal{T}$ over iterations k. The main challenge in this step is that the variance of the q-updates explodes exponentially in $(1/\tau)$ as the strategy vectors $\boldsymbol{\pi}_k$ approach the boundary of the probability simplex. So as to avoid this explosion, we need to track carefully the rate at which the iterates approach the boundary. For a tolerance parameter $\delta \in (0, 1)$, we say that the iterates $\{\boldsymbol{\pi}_k\}_{k\geq 1}$ are δ-*good* up to time K if we have

$$\min_{a^i \in \mathcal{A}^i} \boldsymbol{\pi}_k^i(a^i) \geq \delta \qquad \text{for all iterations } k = 1, \ldots, K, \text{ and nodes } i \in \mathcal{N}. \tag{4.6}$$

LEMMA 4.2. *Suppose that the iterates $\{\boldsymbol{\pi}_k\}_{k\geq 1}$ are δ-good up to time K. Then for each $k = 1, \ldots, K$, we have*

$$\begin{aligned} \mathbb{E}T_{k+1} \leq \big(1 - \beta_k\big(1 - 2r\big)\big)\mathbb{E}V_k + \Big(1 - \alpha_k + \alpha_k^2 \tfrac{A_{\max} N}{\delta} + \beta_k \tfrac{3|\!|\!|\boldsymbol{R}|\!|\!|_2^2}{r\tau^3}\Big)\mathbb{E}W_k \\ + 2N\tau\beta_k \log A_{\max} + \tfrac{8N|\!|\!|\boldsymbol{R}|\!|\!|_2^2}{\tau}\beta_k^2 + \tfrac{4N^2 A_{\max}|\!|\!|\boldsymbol{R}|\!|\!|_2^2}{\delta}\alpha_k^2. \end{aligned} \tag{4.7}$$

for any choice of scalar $r \in (0, \frac{1}{2})$.

See Section SM2.2 for the proof.

A few remarks are in order: note that inequality (4.7) contains a term that grows as $(1/\delta)$ in terms of the distance to the boundary. This term arises because the variance of the TD updates explodes at the boundary. For this reason, we need to track the distance to the boundary as a function of the iteration number. Note that the bounds also involve a parameter $r \in (0, 1/2)$, and we use the freedom in choosing this quantity in a later part of the argument.

Our next step is to use Lemma 4.2 to derive a recursive bound on the expected value $\mathbb{E}T_k$. To control the error terms on the right-hand side of equation (4.7), our analysis involves the temperature and stepsize parameter specifications

$$\tau(r) := \frac{(1 - 2r)\epsilon}{6N \log A_{\max}}, \quad \text{and} \quad \beta(\epsilon, r, \delta) := \frac{(1 - 2r)c_{\alpha,\beta}(r, \tau(r))^2 \delta\epsilon}{15 A_{\max}^2 N^2 |\!|\!|\boldsymbol{R}|\!|\!|_2^2}. \tag{4.8a}$$

The right-hand side of the bound (4.7) contains contraction factors in front of $\mathbb{E}V_k$ and $\mathbb{E}W_k$; we would like to relate these contraction factors so as to form a single term involving the sum $\mathbb{E}T_k = \mathbb{E}V_k + \mathbb{E}W_k$. We do so via a careful choice of the timescale separation constant that relates the two stepsizes via $\frac{\beta_k}{\alpha_k} = c_{\alpha,\beta}$:

$$c_{\alpha,\beta}(r, \tau) := \frac{r(1 - r)\,\tau^3}{4|\!|\!|\boldsymbol{R}|\!|\!|_2^2}. \tag{4.8b}$$

This choice enables us to match the contraction terms for $\mathbb{E}V_k$ and $\mathbb{E}W_k$ (see Section SM2.3). From our choice (4.8b), we are also guaranteed to have

$$\tfrac{8N|\!|\!|\boldsymbol{R}|\!|\!|_2^2}{\tau}\beta_k^2 \leq \tfrac{N^2 A_{\max}|\!|\!|\boldsymbol{R}|\!|\!|_2^2}{\delta}\alpha_k^2.$$

Combining these relations with the bound (4.7) yields

$$\mathbb{E}T_{k+1} \leq \big(1 - \beta_k(1 - 2r)\big)\mathbb{E}T_k + 2N\tau\beta_k \log A_{\max} + \tfrac{5N^2 A_{\max}|\!|\!|\boldsymbol{R}|\!|\!|_2^2}{\delta}\alpha_k^2. \tag{4.9}$$

Note that this drift inequality is valid as long as the iterates are δ-good (cf. equation (4.6)).

Given the drift inequality (4.9), our next step is to bound the iteration complexity. Concretely, for each $\epsilon > 0$, we let $K(\epsilon, \delta, r)$ denote the number of δ-good iterations of Algorithm 2.2 that are required to ensure that $\mathbb{E}\big[\mathrm{NG}(\boldsymbol{\pi}_{K(\epsilon,\delta,r)+1}\big] \leq \epsilon$. The following result, proved in Section SM2.4, gives upper bounds on this iteration complexity:

Lemma 4.3. *For any $r \in (0, \frac{1}{2})$, initialize Algorithm* 2.2 *with the timescale separation constant and the temperature and stepsize parameters from equations* (4.8a) *and* (4.8b). *Then the iteration complexity $K(\epsilon, \delta, r)$ is bounded as follows:*

(a) For the constant stepsize $\beta_k \equiv \beta(\epsilon, r, \delta)$, we have

$$K(\epsilon, \delta, r) \leq \Big\lceil \frac{\log\big(\frac{\epsilon}{3T_1}\big)}{\log\big(1 - \beta(\epsilon, r, \delta)\big(1 - 2r\big)\big)} \Big\rceil. \tag{4.10a}$$

(b) For the inverse polynomial stepsize $\beta_k = \frac{\beta(\epsilon,r,\delta)}{(k+k_0)^\eta}$ for some exponent $\eta \in (0, 1)$, and the offset $k_0 = \big\lceil\big(\frac{2\eta}{\beta(\epsilon,r,\delta)}\big)^{1/(1-\eta)}\big\rceil$, we have

(4.10b)
$$K(\epsilon, \delta, r) \leq \Big\lceil \Big(\frac{1 - \eta}{(1 - 2r)\beta(\epsilon, r, \delta)} \log \frac{3T_1}{\epsilon} + (1 + k_0)^{(1-\eta)} \Big)^{1/(1-\eta)} - k_0 - 1 \Big\rceil.$$

4.2.2. Controlling the distance to the boundary.

The delicacy with the iteration complexity in Lemma 4.3 is that it is valid for the actual iterates of our process *only when*[3] they are δ-good at least up to iteration $K(\epsilon, \delta, r)$. Moreover, note that the stepsize parameter β from equation (4.8a) depends on the target accuracy ϵ, as well as the distance δ to the boundary. As a degree of freedom, our results so far allow us to choose $\delta \in (0, 1)$ and $r \in (0, 1/2)$, and we do so carefully in the following analysis.

Define $K_{\text{good}}(\beta, \delta)$ to be the maximum number of iterations of Algorithm 2.2 with the stepsize parameter $\beta \in (0, 1)$ that can be taken while ensuring that the iterates remain δ-good (cf. equation (4.6)), when the initial mixed strategies $\boldsymbol{\pi}_1$ are uniform.

For a given target accuracy $\epsilon \in (0, 1)$ and convergence exponent $\nu > 0$, our proof strategy consists of the following steps:

(i) We first establish a lower bound on $K_{\text{good}}(\beta, \delta)$ for any $\beta, \delta \in (0, 1)$. We apply this lower bound for the $\beta(\epsilon, r, \delta)$ specified by Lemma 4.3.

(ii) Our next step is to make careful choice of $\delta(\epsilon, \nu)$ and $r(\epsilon, \nu)$ as a function of ϵ and ν—in particular, see Lemma 4.4 to follow—such that the lower bound on the number of "good" iterations $K_{\text{good}}\big(\beta(\epsilon, r(\epsilon, \nu), \delta(\epsilon, \nu)), \delta(\epsilon, \nu)\big)$ is larger than the upper bound on the iteration complexity $K\big(\epsilon, \delta(\epsilon, \nu), r(\epsilon, \nu)\big)$ shown in Lemma 4.3. This inequality ensures that the iterates $\{\boldsymbol{\pi}_k\}_{k\geq 1}$ remain $\delta(\epsilon, \nu)$-good until time $K\big(\epsilon, \delta(\epsilon, \nu), r(\epsilon, \nu)\big)$.

(iii) In the final step, we can then apply Lemma 4.3 to argue that

$$\mathbb{E}[\mathrm{NG}(\boldsymbol{\pi}_{K\big(\epsilon, \delta(\epsilon,\nu), r(\epsilon,\nu)\big)+1}] \leq \epsilon.$$

We now state the lemma that provides a lower bound on the number of "good" iterations (cf. step (iii) above). In proving this result, we choose the free parameters

[3] To be explicit, the proof of Lemma 4.3 exploits the drift inequality (4.9), which is only valid for our process when the iterates are δ-good.

r and δ as a function of the exponent ν that appears in our final guarantee; see the proof in Subsection 4.2.3 for details.

LEMMA 4.4. *For any $\nu > 0$, define*

$$\delta(\epsilon,\nu) := \begin{cases} \left(\frac{\epsilon}{3T_1}\right)^{1+\nu}\frac{1}{A_{\max}} & \text{for the constant stepsize case, and} \\ \left(\frac{\epsilon}{3T_1}\right)^{1+\nu}\frac{\exp(1)}{A_{\max}} & \text{for the inverse polynomial case.} \end{cases} \tag{4.11}$$

There exists a scalar $r(\epsilon,\nu) \in (0,\frac{1}{2})$ such that the maximum number of "good" iterations $K_{good}\big(\beta(\epsilon, r(\epsilon,\nu), \delta(\epsilon,\nu)), \delta(\epsilon,\nu)\big)$ is lower bounded by the upper bounds in Lemma 4.3 for both stepsizes.

4.2.3. Proof of Lemma 4.4. We give the proof for the constant stepsize case here; the proofs for the other cases are analogous, and given in the supplementary Section SM2.5. In order to prove Lemma 4.4, it suffices to show that for any $\xi(\nu) \in (1, 1+\nu)$, there exists a choice $r(\nu) \in (0,\frac{1}{2})$ such that

$$K_{\text{good}}\big(\beta(\epsilon, r(\epsilon,\nu), \delta(\epsilon,\nu)), \delta(\epsilon,\nu)\big) \geq \frac{\xi(\nu)\log\left(\frac{\epsilon}{3T_1}\right)}{\log\big(1-(1-2r(\nu))\beta(\epsilon, r(\nu), \delta(\epsilon,\nu))\big)}, \tag{4.12a}$$

where $\delta(\epsilon,\nu) := \left(\frac{\epsilon}{3T_1}\right)^{1+\nu}\frac{1}{A_{\max}}$. Here the quantity $\xi(\nu) > 1$ is the factor by which $K_{\text{good}}(\beta,\delta)$ is greater than $K(\epsilon,\delta,r)$ (cf. Lemma 4.3), for the given choices of δ and r.

In order to prove the bound (4.12a), we first show that the iterates remain δ-good for all iterates K such that

$$K \leq \frac{\log(A_{\max}\delta)}{\log(1-\beta)}. \tag{4.12b}$$

Indeed, from the strategy update dynamics (2.8b), we have the elementwise inequality ${\pi^i}_{k+1} \succeq {\pi^i}_k(1-\beta_k)$. Iterating this inequality for $k = 1,\ldots,K+1$ yields ${\pi^i}_{K+1} \succeq \frac{1}{A_{\max}}\prod_{j=1}^{K}(1-\beta_j)$, where we have made use of the lower bound $\boldsymbol{\pi}_1 \succeq \frac{1}{A_{\max}}\boldsymbol{e}$, as guaranteed by our uniform initialization. Therefore, in order to ensure that ${\pi^i}_k \succeq \delta$ for all $k = 1,\ldots,K+1$, it suffices to have $\frac{1}{A_{\max}}\prod_{k=1}^{K}(1-\beta_k) \geq \delta$, using the fact that $\beta_k \in (0,1)$ for all k. For the constant stepsize $\beta_k \equiv \beta \in (0,1)$, this condition translates to the bound $(1-\beta)^K \geq A_{\max}\delta$, or equivalently $K \leq \frac{\log(A_{\max}\delta)}{\log(1-\beta)}$.

We now use the bound (4.12b) to prove the claim (4.12a). Note that the latter bound is a lower bound on $K_{\text{good}}(\beta,\delta)$. Now we use the specified choices of δ and r as a function of ϵ and ν to ensure that these lower bounds on $K_{\text{good}}(\beta,\delta) \geq K(\epsilon,\delta,r)$. For any $\xi \in (1, 1+\nu)$, by choosing $r(\nu) \in (0,\frac{1}{2})$ small enough, we can ensure that

$$\frac{\log(1-\beta)}{\log\big(1-\beta\big(1-2r(\nu)\big)\big)} \leq \frac{1+\nu}{\xi}. \tag{4.13}$$

Applying this bound yields

$$\frac{\log(A_{\max}\ \delta(\epsilon,\nu))}{\log(1-\beta)} = (1+\nu)\frac{\log\left(\frac{\epsilon}{3T_1}\right)}{\log(1-\beta)} \geq \frac{\xi\log\left(\frac{\epsilon}{3T_1}\right)}{\log\big(1-\beta\big(1-2r(\nu)\big)\big)}.$$

Therefore, the iterates $\{\boldsymbol{\pi}_k\}_{k\geq 1}$ remain $\delta(\epsilon,\nu)$-good up until the claimed time (4.12a).

4.2.4. Completing the proof. In order to complete the proof of Theorem 3.1, it suffices to combine the previous pieces. Summarizing, we have shown that:

- For any pair of scalars $\delta \in (0,1)$ and $r \in (0, \frac{1}{2})$, Lemma 4.3 specifies the number of iterations $K(\epsilon, \delta, r)$ sufficient to ensure that $\mathbb{E}[\mathrm{NG}(\boldsymbol{\pi}_{K(\epsilon,\delta,r)+1})] \leq \epsilon$. (This guarantee is predicated upon the iterates $\{\boldsymbol{\pi}_k\}_{k\geq 1}$ of Algorithm 2.2 being δ-good until time $K(\epsilon, \delta, r)$, and that Algorithm 2.2 is initialized according to equations (4.8a) and (4.8b)).
- For a given target Nash gap $\epsilon \in (0,1)$ and $\nu > 0$, Lemma 4.4 specifies the pair (δ, r) as a function of the pair (ϵ, ν), so that we may write $\delta(\epsilon, \nu)$ and $r(\epsilon, \nu)$ to indicate this dependence. Note that Lemma 4.3 applies to these choices of $\delta(\epsilon, \nu)$ and $r(\epsilon, \nu)$ as well.
- For these choices of (δ, r), Lemma 4.4 shows that the iterates $\{\boldsymbol{\pi}_k\}_{k\geq 1}$ of Algorithm 2.2 are $\delta(\epsilon, \nu)$-good for at least $K^\star(\epsilon, \nu)$ iterations, where $K^\star(\epsilon, \nu)$ is the upper bound on $K\big(\epsilon, \delta(\epsilon, \nu), r(\epsilon, \nu)\big)$ from Lemma 4.3.

It follows from the results of Lemmas 4.3 and 4.4 that the Nash gap evaluated at time $K^\star(\epsilon, \nu) + 1$ is at most ϵ. The upper bound on $K(\epsilon)$ given in Theorem 3.3 follows by bounding $K^\star(\epsilon, \nu)$ from above.

Remarks. In Lemma 4.3, the stepsize parameter is chosen as a function of δ. It could be desirable to obtain an anytime version of Theorem 3.3, for which the Nash gap continues to be bounded by ϵ for all time steps greater than $K^\star(\epsilon, \nu)$. Such a guarantee can be achieved by running $K^\star(\epsilon, \nu)$ iterations, and then annealing the stepsize parameter as δ goes to zero, to ensure that the iterates continue to be δ-good and Lemma 4.3 holds.

5. Discussion.

In this paper, we studied best-response type learning dynamics that arise from the discretization of continuous-time best-response dynamics applied to zero-sum polymatrix games. We analyzed these dynamics in both the full information setting as well as the more challenging setting of minimal information, in which each player observes only their random payoff. In the latter context, we provided the first polynomial-scaling finite-sample guarantees for these best-response type dynamics in the minimal information setting without additional exploration. Our results also exhibited an interesting dependence on the underlying polymatrix game through the parameter $\|\boldsymbol{R}\|_2$. The analysis involved some new ideas, including careful tracking of variance of the stochastic updates as they approach the boundary of the probability simplex.

In terms of open questions, we suspect that our current results are not sharp in terms of dependence on ϵ and $A_{\max}$ for last-iterate convergence to a Nash equilibrium. For instance, in the simpler setting of full information, the optimal rate is known to scale as $\mathcal{O}(\frac{1}{\epsilon})$; we are not yet aware if our dynamics can achieve this rate, or if the $\mathcal{O}(\frac{1}{\epsilon^2})$ guarantees that we have provided are, in fact, unimprovable. In terms of the dependence on the underlying polymatrix game, it would be interesting to see if a finer dependence on the underlying pairwise games can be elicited. Additionally, our analysis in this paper was based upon a fixed temperature parameter τ. Studying dynamics with a time-varying temperature $\{\tau_k\}_{k\geq 1}$ might help eliminate the constant smoothing bias in the current bounds on the Nash gap.

Acknowledgments. MJW was partially supported by ONR grant N00014-21-1-2842 and NSF DMS-2311072, AO was supported by ONR grant N00014-25-1-2296 and FZF was supported by the MIT EECS Thriving Stars Energy Fellowship.

REFERENCES

[1] J. Abernethy, K. A. Lai, and A. Wibisono, *Fast convergence of fictitious play for diagonal payoff matrices*, in Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA), SIAM, 2021, pp. 1387–1404.

[2] S. P. Anderson, A. De Palma, and J.-F. Thisse, *Discrete choice theory of product differentiation*, MIT press, 1992.

[3] R. Ao, S. Cen, and Y. Chi, *Asynchronous gradient play in zero-sum multi-agent games*, arXiv preprint arXiv:2211.08980, (2022).

[4] A. Bakhtin, D. J. Wu, A. Lerer, J. Gray, A. P. Jacob, G. Farina, A. H. Miller, and N. Brown, *Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning*, arXiv preprint arXiv:2210.05492, (2022).

[5] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint (2nd ed.)*, Hindustan Book Agency and Springer Nature, 2022.

[6] L. Bregman and I. Fokin, *Methods of determining equilibrium situations in zero-sum polymatrix games*, Optimizatsia, 40 (1987), pp. 70–82.

[7] L. Bregman and I. Fokin, *On separable non-cooperative zero-sum games*, Optimization, 44 (1998), pp. 69–84.

[8] G. W. Brown, *Some notes on computation of games solutions*, Tech. Report P-78, The Rand Corporation, 1949.

[9] G. W. Brown, *Iterative solution of games by fictitious play*, in Activity Analysis of Production and Allocation, T. C. Koopmans, ed., Wiley, New York, 1951.

[10] Y. Cai, O. Candogan, C. Daskalakis, and C. Papadimitriou, *Zero-sum polymatrix games: A generalization of minmax*, Mathematics of Operations Research, 41 (2016), pp. 648–655.

[11] Y. Cai, H. Luo, C.-Y. Wei, and W. Zheng, *Uncoupled and convergent learning in two-player zero-sum Markov games with bandit feedback*, in Advances in Neural Information Processing Systems, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds., vol. 36, 2023, pp. 36364–36406.

[12] S. Cen, Y. Wei, and Y. Chi, *Fast policy extragradient methods for competitive games with entropy regularization*, Advances in Neural Information Processing Systems, 34 (2021), pp. 27952–27964.

[13] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, Cambridge University Press, 2006.

[14] X. Chen and B. Peng, *Hedging in games: Faster convergence of external and swap regrets*, Advances in Neural Information Processing Systems, 33 (2020), pp. 18990–18999.

[15] Z. Chen, K. Zhang, E. Mazumdar, A. Ozdaglar, and A. Wierman, *A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games*, arXiv preprint arXiv:2303.03100, (2023).

[16] C. Daskalakis, A. Deckelbaum, and A. Kim, *Near-optimal no-regret algorithms for zero-sum games*, in Proceedings of the 2011 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2011, pp. 235–254.

[17] C. Daskalakis and Q. Pan, *A counter-example to Karlin's strong conjecture for fictitious play*, in 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, IEEE, 2014, pp. 11–20.

[18] D. P. Foster and H. P. Young, *On the nonconvergence of fictitious play in coordination games*, Games and Economic Behavior, 25 (1998), pp. 79–96.

[19] D. Fudenberg and D. M. Kreps, *Learning mixed equilibria*, Games and Economic Behavior, 5 (1993), pp. 320–367.

[20] C. Harris, *On the rate of convergence of continuous-time fictitious play*, Games and Economic Behavior, 22 (1998), pp. 238–259.

[21] S. Hart and A. Mas-Colell, *A simple adaptive procedure leading to correlated equilibrium*, Econometrica, 68 (2000), pp. 1127–1150.

[22] J. Hofbauer and W. H. Sandholm, *On the global convergence of stochastic fictitious play*, Econometrica, 70 (2002), pp. 2265–2294, http://www.jstor.org/stable/3081987 (accessed 2024-04-09).

[23] E. Janovskaja, *Equilibrium points in polymatrix games*, Lithuanian Mathematical Journal, 8 (1968), p. 381–384.

[24] S. Karlin, *Mathematical Methods and Theory in Games, Programming and Economics. Vol. 2: The Theory of Infinite Games*, Addison-Wesley Publishing Company, 1959.

[25] S. Leonardos, G. Piliouras, and K. Spendlove, *Exploration-exploitation in multi-agent competition: convergence with bounded rationality*, Advances in Neural Information Processing Systems, 34 (2021), pp. 26318–26331.

[26] D. S. Leslie and E. J. Collins, *Convergent multiple-timescales reinforcement learning algorithms in normal form games*, The Annals of Applied Probability, 13 (2003), pp. 1231–1251.

[27] D. S. Leslie and E. J. Collins, *Individual Q-learning in normal-form games*, SIAM Journal on Control and Optimization, 44 (2005), pp. 495–514.

[28] K. Miyasawa, *On the convergence of the learning process in a* 2×2 *non-zero-sum two-person game*, Princeton University Princeton, 1961.

[29] D. Monderer and L. S. Shapley, *Fictitious play property for games with identical interests*, Journal of economic theory, 68 (1996), pp. 258–265.

[30] D. Monderer and L. S. Shapley, *Potential games*, Games and economic behavior, 14 (1996), pp. 124–143.

[31] J. F. Nash Jr, *Equilibrium points in n-person games*, Proceedings of the National Academy of Sciences, 36 (1950), pp. 48–49.

[32] R. Ouhamma and M. Kamgarpour, *Learning Nash equilibria in zero-sum Markov games: A single time-scale algorithm under weak reachability*, arXiv preprint arXiv:2312.08008, (2023).

[33] S. Rakhlin and K. Sridharan, *Optimization, learning, and games with predictable sequences*, Advances in Neural Information Processing Systems, 26 (2013).

[34] J. Robinson, *An iterative method of solving a game*, Annals of Mathematics, 54 (1951), pp. 296–301, http://www.jstor.org/stable/1969530 (accessed 2024-04-09).

[35] M. Sayin, K. Zhang, D. Leslie, T. Basar, and A. Ozdaglar, *Decentralized Q-learning in zero-sum Markov games*, Advances in Neural Information Processing Systems, 34 (2021), pp. 18320–18334.

[36] H. N. Shapiro, *Note on a computation method in the theory of games*, Communications on Pure and Applied Mathematics, 11 (1958), pp. 587–593.

[37] L. S. Shapley, *Some topics in two-person games*, in Advances in Game Theory, M. Dresher, L. S. Shapley, and A. W. Tucker, eds., vol. 52 of Annals of Mathematics Studies, Princeton University Press, Princeton, NJ, 1963, pp. 1–29.

[38] S. Sokota, R. D'Orazio, J. Z. Kolter, N. Loizou, M. Lanctot, I. Mitliagkas, N. Brown, and C. Kroer, *A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games*, arXiv preprint arXiv:2206.05825, (2022).

[39] R. S. Sutton, *Learning to predict by the methods of temporal differences*, Machine learning, 3 (1988), pp. 9–44.

[40] V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire, *Fast convergence of regularized learning in games*, Advances in Neural Information Processing Systems, 28 (2015).

[41] J. von Neumann, *Zur theorie der gesellschaftsspiele*, in Mathematische Annalen, vol. 100, 1928, pp. 295–320.

[42] S. Zeng, T. Doan, and J. Romberg, *Regularized gradient descent ascent for two-player zero-sum Markov games*, Advances in Neural Information Processing Systems, 35 (2022), pp. 34546–34558.