Finite-sample Guarantees for Nash Q-learning with Linear Function Approximation

Pedro Cisneros-Velarde[1] **Sanmi Koyejo**[2]

[1]University of Illinois at Urbana-Champaign
[2]Stanford University, Google Research

Abstract

Nash Q-learning may be considered one of the first and most known algorithms in multi-agent reinforcement learning (MARL) for learning policies that constitute a Nash equilibrium of an underlying general-sum Markov game. Its original proof provided asymptotic guarantees and was for the tabular case. Recently, finite-sample guarantees have been provided using more modern RL techniques for the tabular case. Our work analyzes Nash Q-learning using linear function approximation – a representation regime introduced when the state space is large or continuous – and provides finite-sample guarantees that indicate its sample efficiency. We find that the obtained performance nearly matches an existing efficient result for single-agent RL under the same representation and has a polynomial gap when compared to the best-known result for the tabular case.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) has been successfully applied to a diversity of problems, such as solving the games of Go (Silver et al., 2016, 2017) and Starcraft (Vinyals et al., 2019), coordination of unmanned aerial vehicles (Pham et al., 2018), autonomous driving (Dinneweth et al., 2022), power systems (Foruzan et al., 2018), and management of water and energy resources (Ni et al., 2014; Yang et al., 2020). The theory and development of multi-agent reinforcement learning algorithms is currently a prolific area, as attested by various recent surveys on the field, e.g., (Zhang et al., 2021; Hernandez-Leal et al., 2019; Yang and Wang, 2021). In general, employing MARL to solve for a Nash equilibrium general-sum Markov game is computationally complex (Daskalakis et al., 2009). This motivated theoretical works to look for other weaker solution concepts (e.g., coarse-correlated equilibria), or, if looking for a Nash equilibrium, either: (i) leave the general-sum domain and focus on zero-sum games or fully cooperative games, or (ii) specify extra conditions for the underlying general-sum Markov game (MG) (Zhang et al., 2021). The seminal work (Hu and Wellman, 2003) introduced the *Nash Q-learning* algorithm in the context of infinite-horizon discounted Markov games. The idea of Nash Q-learning is that, at every time step, each agent needs to find a Nash equilibrium which solves some static game whose utilities or rewards are defined by the (estimates of the) Q-functions of all the agents – this is also called a *stage game*. Thus, a motivation for using Nash Q-learning is its algorithmic simplicity: it solves a static game where Q-learning (for classic single-agent RL) would otherwise solve for an optimum. In (Hu and Wellman, 2003), asymptotic learning guarantees are given when the chosen Nash equilibrium is consistent in all stage games and is either a *global optimal* or a *saddle* one. Despite this strong sufficient condition, Hu and Wellman (2003) presented numerical examples where Nash Q-learning solves games that do not satisfy such conditions. It is important to remark that there exist proven cases in which value-based methods – encompassing Nash Q-learning – cannot converge to a single Nash equilibrium of general-sum Markov games (Zinkevich et al., 2005). However, remarkably, Nash Q-learning stands as one of the few general-sum MARL algorithms and has elicited the development of algorithms specialized to other classes of Markov games or focused on other solution concepts. Further, it is still consistently cited in the applied literature (Hernandez-Leal et al., 2019).

The first formal proof for Nash Q-learning by Hu and Wellman (2003) only provided formal guarantees for asymptotic convergence in the tabular setting. However, recently, about two decades later, Liu et al. (2021) proposed a type of Nash Q-learning algorithm and used a modern approach from the theoretical reinforcement learning (RL) literature to establish finite-sample guarantees and thus guarantee the sample efficiency of learning in the tabular setting. Liu et al. (2021)

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

used regret as a performance metric, and thus it was of interest that the average performance of policies gets closer to the performance of a Nash equilibrium instead of an actual convergence to a single equilibrium.

In the modern RL literature, it is known that tabular approaches are not ideal in environments where the state space is large or continuous. This has motivated the development of *linear function approximation*, where, for example, the transition kernel and reward function of the underlying Markov decision process (MDP) are a linear function of a vector of features (Jin et al., 2020; Yang and Wang, 2020).

Taken together, these prior works motivate the central question of our paper:

Can we obtain finite-sample guarantees and sample efficiency for Nash Q-learning in the linear function approximation regime?

We answer this question positively by proposing a Nash Q-learning algorithm – called *Nash Q-learning with optimistic value iteration* (NQOVI) – and providing its finite sample guarantees under a regret performance metric. Interestingly, we find that the sample efficiency of our algorithm nearly matches the one reported in (Jin et al., 2020) for (single-agent) RL in the same approximation regime.

In general, our central question is also motivated from the fact that an increasing number of works providing sample efficient guarantees for (single-agent) RL problems has appeared in recent years. The works (Jin et al., 2020; Yang and Wang, 2020) started providing such guarantees in the linear function approximation domain using the principle of *optimism* under uncertainty for *online* RL – later, other works have applied it to *reward-free* RL (e.g. (Wang et al., 2020)) and have even applied a counterpart principle, called *pessimism*, to *offline* RL (Jin et al., 2021). Optimism consists of adding a bonus so that the estimated optimistic Q-function rewards more those state-action pairs that have been less explored. Pessimism basically does the opposite by subtracting a bonus value. However, when it comes to (online) MARL, to the best of our knowledge, the simultaneous application of optimism and pessimism to achieve sample efficiency for learning Nash equilibria has mainly been limited to two-player zero-sum games in the linear function approximation case (Qiu et al., 2021), and to general-sum games in the tabular case (Liu et al., 2021). In this work, we show that the principle of optimism can easily be applied to Nash Q-learning in general-sum games.

Contributions We summarize our contributions.

- We provide the first sample efficient guarantees for a Nash Q-learning algorithm in the linear function approximation regime for general-sum games – obtaining a regret bound $\widetilde{\mathcal{O}}(\sqrt{Kd^3H^5})$, with K being the number of episodes, H the episode length, and d the dimension of the feature vector of the linear function approximation.
- To prove our guarantees, we propose the Nash Q-learning with optimistic value iteration (NQOVI) algorithm. The original Nash Q-learning proposed by Hu and Wellman (2003) was in the context of tabular and discounted MGs, and considered convergence to a Nash equilibrium as a performance metric. In contrast, we consider episodic MGs with regret performance, and do no need the existence of special Nash equilibria on the stage games as in Hu and Wellman (2003).
- When directly transforming it to the tabular case, our performance bound has a polynomial gap on all factors except for the number of episodes K compared to the best-known result by Liu et al. (2021).
- In the single agent case, our NQOVI algorithm collapses to the model-free RL algorithm proposed by Jin et al. (2020) (instead of taking a (mixed) Nash equilibrium at each stage game, the agent takes the optimal greedy action). Remarkably, we show that our algorithm's sample efficiency differs only by a factor of H – the length of the episode – compared to the single agent one. To the best of our knowledge, this is the first time a general-sum MARL algorithm nearly matches the sample efficiency of an RL algorithm.

1.1 RELATED WORKS

Since our paper is of a theoretical nature, we limit ourselves to presenting prior work focused on theory.

Multi-agent RL (MARL). Although the applied MARL literature has been around for decades, theoretical works have been gaining more presence in recent years – we refer the reader to the recent surveys (Zhang et al., 2020; Hernandez-Leal et al., 2019; Yang and Wang, 2021). Importantly, we highlight that a large body of recent works have focused on the study of learning in the two-player zero-sum Markov game case – where one player tries to maximize the expected reward while the other tries to minimize it. One reason for its popularity is that it can be formulated as a minimax game and Nash equilibria are easily characterized (Zhang et al., 2021). Recent works have been done both in the tabular setting, e.g., (Kozuno et al., 2021; Zhang et al., 2020; Bai et al., 2020; Liu et al., 2021; Jin et al., 2022), and the linear function approximation setting, e.g., (Chen et al., 2022; Cisneros-Velarde et al., 2023; Qiu et al., 2021). In the case of general-sum Markov games, another large body of work has focused on providing guarantees for finding other solution concepts such as coarse correlated equilibria (CCE); e.g., (Liu et al., 2021; Jin et al., 2022; Mao and Başar, 2022). Minimax sample optimality has been shown – under certain assumptions – for finding CCE in general-sum games and Nash equilibria in zero-sum games for the tabular case (Li et al., 2022). In learning Nash equilibria,

Liu et al. (2021) proposed a Nash Q-learning algorithm for general-sum games in the tabular setting, with an underlying episodic MG – no extra conditions on the Nash equilibria are required. While writing our paper we found the recent preprint by Ni et al. (2022) who studied representation learning in general-sum games and whose proposed algorithms output a policy after a number of episodes. They focus on the harder problem of learning the feature vector of the linear approximation, whereas we assume it is given – we only focus on learning a good policy and not on learning a good representation. Thus our guarantees are not directly comparable. Finally, we remark that both Liu et al. (2021) and Ni et al. (2022) use the principle of optimism and pessimism, so they compute two Q-functions on their algorithms, while we compute just one optimistic Q-function. Two recent works (Cui et al., 2023; Wang et al., 2023) used function approximation and sought to avoid an exponential dependence on the size of the action spaces of the agents on the regret bounds when specialized to the tabular setting. While our results have such dependency when specialized to the tabular case, our setting is different than theirs. Cui et al. (2023) considered linear function approximation with each agent having its own feature vector encoding only its own action space, whereas we consider a feature vector that encodes the joint action space. Moreover, their work, unlike ours, restricted the underlying Markov game to be a potential Markov game when considering NE. The work by Wang et al. (2023), also considered independent feature vectors and was only concerned with CCE and correlated equilibria (CE) as solution concepts.

Linear function approximation in RL. The idea of using linear function approximation is ubiquitous in theoretical RL. The first works to combine it with optimism for sample efficient learning were (Jin et al., 2020; Yang and Wang, 2020) for online RL. Since then, such setting has been adapted to different RL problems, such as representation learning (of the feature vector of the linear function approximation), e.g. (Agarwal et al., 2020); parallel learning (multiple agents learning through independent MDPs but being able to communicate their experience), e.g., (Dubey and Pentland, 2021); deployment efficiency (RL algorithms when the number of times a policy can be updated is restricted), e.g., (Gao et al., 2021); reward-free RL (where exploration and exploitation are separated in different learning stages), e.g., (Wang et al., 2020; Wagenmaker et al., 2022). Some works have combined two of the aforementioned problems using linear function approximation; e.g. in the context of reward-free RL, Huang et al. (2021) studied deployment efficiency, whereas Cisneros-Velarde et al. (2023) studied the effect of parallel exploration. These works follow a similar skeleton in their algorithms since all of them have in common the use of optimism and value iteration – it is in this framework that we decided to propose an algorithm based on Nash Q-learning.

The paper is organized as follows. In Section 2, we formally introduce the setting. In Section 3, we introduce our Nash Q-learning algorithm and state our main result. In Section 4, we provide a sketch of the proof and some nuances of its formal analysis. Section 5 is the conclusion.

1.2 NOTATION

Let $\|\cdot\|$ be the Euclidean norm, and $\|v\|_A = \sqrt{v^T A v}$ for positive semidefinite matrix A. Let $[k] = \{1, 2, \ldots, k\}$ for a positive integer k. Let I_m be the $m \times m$ identity matrix. Let $\Delta(\mathcal{A})$ be the probability simplex defined on a given finite set $\mathcal{A}$. Given the big-O complexity notation $\mathcal{O}$, we use $\widetilde{\mathcal{O}}$ to hide polylogarithmic terms in the quantities of interest.

2 PRELIMINARIES

We consider an episodic Markov Game (MG) of the form $\mathcal{MG} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r,)$, with state space $\mathcal{S}$, action space $\mathcal{A} = \mathcal{A}_1 \times \ldots \mathcal{A}_n$ with $\mathcal{A}_i$ being the action space for agent $i \in [n]$, H is the number of steps per episode or episode length – we assume the non-bandit case $H \geq 2$, $\mathcal{P} = \{\mathcal{P}_h\}_{h\in[H]}$ are transition probability measures and $\mathcal{P}_h(\cdot \mid x, a)$ denotes the transition kernel over $h+1$ if all players take the action profile $a \in \mathcal{A}$ for state $x \in \mathcal{S}$. We denote agent i's reward function profile by $r_i = \{r_h^i\}_{h=1}^H$ with $r_h^i : \mathcal{S} \times \mathcal{A}_i \to [0, 1]$.[1] For any agent $i \in [n]$, its action taken at step h is denoted by $a_{i,h} \in \mathcal{A}_i$, and let $a_i = \{a_{i,h}\}_{h=1}^H$. We assume every agent has a finite action space, while the state space can be arbitrarily large or even continuous.

We denote agent i's policy by $\pi_i = \{\pi_{i,h}\}_{h=1}^H$ with $\pi_{i,h} : \mathcal{S} \to \Delta(\mathcal{A}_i)$. With some abuse of notation, we also let $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ denote the (joint) policy taken by the agents over the joint action space at time step $h \in [H]$ – the subindex k in π_k will be clear from the context whether it refers to an agent or a time step. Let π be the joint policy of all agents. We say π is a product policy (across agents) $\pi = \pi_1 \times \cdots \times \pi_n$ when, conditioned on the same state, the action of every agent can be sampled independently according to their own policy, i.e., $\pi_h(x) \in \Delta(\mathcal{A}_1) \times \cdots \times \Delta(\mathcal{A}_n)$ for every $x \in \mathcal{S}$, $h \in [H]$. For agent i, we define her value function $V_h^{i,\pi} : \mathcal{S} \to \mathbb{R}$ at the h-th step as $V_h^{i,\pi}(x) = \mathbb{E}_\pi[\sum_{h'=h}^H r_{h'}^i(s_{h'}, a_{h'}) \,|\, s_h = x]$ and her Q-function or action-value function $Q_h^{i,\pi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as $Q_h^{i,\pi}(x, a) := \mathbb{E}_\pi[\sum_{h'=h}^H r_{h'}^i(s_{h'}, a_{h'}) \,|\, s_h = x, a_h = a]$, where the expectation $\mathbb{E}_\pi$ is taken with respect to both the randomness in the transitions $\mathcal{P}$ and the randomness inherent in the policy π. If agent i has policy ν and the rest of the agents joint policy π_{-i}, we denote its associated value function at step h by the notation $V_h^{i,\nu,\pi_{-i}}$, i.e., by placing a superscript with i's policy before the (joint) policy of the

[1] We assume deterministic rewards for simplicity.

rest of the agents; consequently, $V_h^{i,\pi_i,\pi_{-i}} = V_h^{i,\pi}$.

We now define our solution concept of interest.

Definition 2.1 (Nash equilibrium for Markov games)**.** *Given an initial state $s_o \in \mathcal{S}$, a product policy profile π^* is called a* Nash equilibrium *(NE) if $V_1^{i,\pi_i^*,\pi_{-i}^*}(s_o) \geq V_1^{i,\pi_i,\pi_{-i}^*}(s_o)$ for any $i \in [n]$ and any policy π_i, and it is called an ϵ-*NE *if $V_1^{i,\pi_i^*,\pi_{-i}^*}(s_o) - V_1^{i,\pi_i,\pi_{-i}^*}(s_o) \leq \epsilon$.*

We say agent $i \in [n]$ plays a *best response* policy against the policy profile π_{-i} of the rest of the agents according to $\mathrm{br}_i(\pi_{-i}) \in \mathrm{argmax}_\nu V_h^{i,\nu,\pi_{-i}}(x)$ for any $(x,h) \in \mathcal{S} \times [H]$. Note that we can easily characterize a Nash equilibrium (Definition 2.1) using best-responses.

For any function $f : \mathcal{S} \to \mathbb{R}$, we define the transition operator as $(\mathbb{P}_h f)(x,a) = \mathbb{E}_{x' \sim \mathcal{P}_h(\cdot|x,a)}[f(x')]$. For any $i \in [n]$, the Bellman equation associated with a policy π is: $Q_h^{i,\pi}(x,a) = (r_h^i(x,a) + \mathbb{P}_h V_{h+1}^{i,\pi})(x,a)$, $V_h^{i,\pi}(x) = \mathbb{E}_{a\sim\pi_h(x)}[Q_h^{i,\pi}(x,a)]$, with $V_{H+1}^{i,\pi}(x) = 0$, for any $(x,a) \in \mathcal{S} \times \mathcal{A}$.

In this paper, we consider linear MGs.

Linear (function approximation in) Markov Games. Under a linear MG setting, there exists a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that for every $h \in [H]$, there exist d unknown (signed) measures $\mu_h = \left(\mu_h^{(1)}, \ldots \mu_h^{(d)}\right)$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that $\mathcal{P}_h(x'|x,a) = \langle \phi(x,a), \mu_h(x') \rangle$, $r_h(x,a) = \langle \phi(x,a), \theta_h \rangle$ for all $(x,a,x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. We assume the non-scalar case with $d \geq 2$ and that the feature map satisfies $\|\phi(x,a)\| \leq 1$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$ and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_h\|\} \leq \sqrt{d}$ at each step $h \in [H]$, where (with an abuse of notation) $\|\mu_h(\mathcal{S})\| = \int_{\mathcal{S}} \|\mu_h(x)\| dx$. Note that the transition kernel $\mathcal{P}_h(\cdot|x,a)$ may have infinite degrees of freedom since the measure μ_h is unknown.

Performance Metric. We consider that all agents are learning during a total of K episodes, starting at some initial state $s_o \in \mathcal{S}$ at the beginning of each episode. For a set of policies $\{\pi^k\}_{k\in[K]}$ provided by an online MARL algorithm, we use the following regret performance metric:

$$\mathrm{Regret}(K) = \sum_{k=1}^{K} \max_{i\in[n]} (V_1^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(s_o) - V_1^{i,\pi^k}(s_o)). \tag{2.1}$$

The idea behind such regret is that, at episode $k \in [K]$, $\max_{i\in[n]}(V_1^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(s_o) - V_1^{i,\pi^k}(s_o)) = 0$ iff (product) policy π^k is a Nash equilibrium for the Markov game.

Static games. We also consider that the n agents can play a static game, keeping their respective action spaces. Given that each agent has an associated reward function $g_i : \mathcal{A} \to \mathbb{R}$ in a static game, the game is defined by the tuple $(g_1, \ldots, g_n)$. Given $a \in \mathcal{A}$, we define a_{-i} as the respective element of $\mathcal{A}_{-i} = \mathcal{A}_1 \times \ldots \mathcal{A}_{i-1} \times \mathcal{A}_{i+1} \times \cdots \times \mathcal{A}_n$. We consider a tuple $\nu = (\nu_1, \ldots, \nu_n)$ with $\nu_i \in \Delta(\mathcal{A}_i)$, $i \in [n]$, to be a strategy profile; and let ν_{-i} be the tuple ν without its ith element. In this work, we consider strategy profiles $\nu \in \Delta(\mathcal{A})$ as product measures $\nu(a) = \prod_{i=1}^n \nu(a_i)$, a similar consideration follows for $\nu_{-i} \in \Delta(\mathcal{A}_{-i})$, $i \in [n]$. A strategy profile ν^* is a Nash equilibrium if $\nu_i^* \in \mathrm{argmax}_{\nu\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\substack{a_i\sim\nu \\ a_i\sim\nu_i^*}}[g_i(a)]$ for every $i \in [n]$.

Definition 2.2 (Global optimal and saddle Nash equilibria (Hu and Wellman, 2003))**.** *A strategy profile ν^* of the static game $(g_1, \ldots, g_n)$ is:*

(i) a global optimal *(Nash) equilibrium if $\mathbb{E}_{a\sim\nu^*}[g_i(a)] \geq \mathbb{E}_{a\sim\nu}[g_i(a)]$ for any strategy profile $\nu \in \Delta(\mathcal{A})$; and*

(ii) a saddle *Nash equilibrium if $\mathbb{E}_{a\sim\nu^*}[g_i(a)] \geq \mathbb{E}_{\substack{a_i\sim\nu_i \\ a_{-i}\sim\nu_{-i}^*}}[g_i(a)]$ for any $i \in [n]$ and any $\nu_i \in \Delta(\mathcal{A}_i)$, and $\mathbb{E}_{a\sim\nu^*}[g_i(a)] \leq \mathbb{E}_{\substack{a_i\sim\nu_i^* \\ a_{-i}\sim\nu_{-i}}}[g_i(a)]$ for any strategy profile $\nu_{-i} \in \Delta(\mathcal{A}_{-i})$.*

3 NASH Q-LEARNING AND ITS ANALYSIS

We propose a simple Nash Q-learning algorithm based on linear function approximation and optimism named *Nash Q-learning with optimistic value iteration* or NQOVI as described in Algorithm 1.

We provide an outline of the NQOVI algorithm. At each iteration $k \in [K]$ and step $h \in [H]$, the information of the explored state-action trajectories described by the agents in the game at the same step but up to the previous episode is collected in a covariance matrix Λ_h^k (line 6 of Algorithm 1). Then, all the agents participate in a static game described by some prior optimistic estimates of Q-functions and a Nash equilibrium is computed – this static game is called a *stage game* because it is solved in every episode and time-step (and depends on the current state of the Markov game). Then each agent, using its computed Nash policy from the stage game, computes a new *optimistic* estimate of the Q-function (line 10), using the optimism bonus $\beta(\phi(\cdot,\cdot)^\top(\Lambda_h^k)^{-1}\phi(\cdot,\cdot))^{1/2}$. Then, all the agents jointly explore the environment (lines 14-16) by taking actions coming from their respective policies computed from stage games. The resulting state-action trajectory across the episodes will then be collected and the whole process repeats.

Remark 3.1 (Computational aspects)**.** *Though a (mixed) Nash equilibrium (NE) is always guaranteed to exist for the static game defined by the optimistic Q-value function in lines 14 and 16 of Algorithm 1, solving for an (exact) NE is in general computationally intractable (Chen et al., 2009; Daskalakis et al., 2009).*

Algorithm 1 Nash Q-learning with optimistic value iteration (NQOVI)

Input: K, β, λ
for episode $k \in [K]$ **do**
  $x_1^k \leftarrow s_0$
  $Q_{H+1}^{i,k}(\cdot,\cdot) \leftarrow 0, i \in [n]$
  for $h = H, \ldots, 1$ **do**
    $\Lambda_h^k \leftarrow \lambda I_d + \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top$
    $\pi^* \leftarrow$ a Nash Equilibrium for the n-player game $(Q_{h+1}^{1,k}(x_{h+1}^k, \cdot), \ldots, Q_{h+1}^{n,k}(x_{h+1}^k, \cdot))$
    for $i \in [n]$ **do**
      $w_h^{i,k} \leftarrow (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\, [r_h^i(x_h^\tau, a_h^\tau) + \mathbb{E}_{a\sim\pi^*}[Q_{h+1}^{i,k}(x_{h+1}^\tau, a)]$
      $Q_h^{i,k}(\cdot,\cdot) \leftarrow \min\{(w_h^{i,k})^\top \phi(\cdot,\cdot) + \beta(\phi(\cdot,\cdot)^\top (\Lambda_h^k)^{-1}\phi(\cdot,\cdot))^{1/2}, H\}$
    end for
  end for
  for $h \in [H]$ **do**
    $\pi_h^k(x_h^k) \leftarrow$ a Nash Equilibrium for the n-player game $(Q_h^{1,k}(x_h^k, \cdot), \ldots, Q_h^{n,k}(x_h^k, \cdot))$
    Take $a_h^k \sim \pi_h^k(x_h^k)$
    Observe x_{h+1}^k
  end for
end for

Remark 3.2 (About information access in NQOVI)**.** *In this paper, we are primarily concerned with analyzing NQOVI as a solver for the policies of the underlying Markov game, e.g, as done in the recent work (Liu et al., 2021). One could think of relaxing some implementation details such as the information each agent has access to across episodes, but this is beyond the scope of the paper. For example, at each step $h \in [H]$ and iteration $k \in [K]$, one could make the optimistic Q-functions of every agent $i \in [n]$, $Q_h^{i,k}$, be private information to the rest of the agents. Then, each agent would try to estimate the Q-functions of the rest of the agents based on the observation of past rewards – an idea already outlined in (Hu and Wellman, 2003).*

We now present the paper's main result.

Theorem 3.1 (Performance of the NQOVI algorithm)**.** *There exists an absolute constant $c_\beta > 0$ such that, for any fixed $\delta \in (0,1)$, if we set $\lambda = 1$ and $\beta = c_\beta dH\sqrt{\iota}$, with $\iota := \log(dKH(n+2)/\delta)$, then, with probability at least $1-\delta$,*

$$\text{Regret}(K) \leq \mathcal{O}\left(\sqrt{K}\sqrt{d^3H^5\iota^2}\right). \tag{3.1}$$

Sample efficiency. Our regret bound is sublinear in the number of episodes K and – ignoring logarithmic terms – polynomial on the parameters d and H, i.e., there is learning with sample efficiency. Our finite-sample guarantee states that $K = \widetilde{\mathcal{O}}\left(\frac{d^3H^5}{\epsilon^2}\right)$ episodes are needed in order to achieve an average regret less or equal than ϵ, i.e., for the policies across the episodes to perform on average as an ϵ-Nash equilibrium.

About the number of agents. Our bound has a logarithmic dependence on the number of agents n. However, we remark that the feature dimension d of the linear MG *might* hide dependencies on n depending on how the feature vector ϕ is constructed (more on this below, when discussing the tabular case). In any case, the larger the number of agents, the more samples are needed to achieve the same average regret performance. Intuitively, this makes sense, since increasing the number of agents increases the number of possible decision makers and thus the complexity of the state-action space to be sampled. This is in stark contrast with other works in the single-agent RL case where multiple agents can be deployed to explore the *same* state-action space of the MDP, in which case their performance measure improves with the number of agents (Cisneros-Velarde et al., 2023).

Comparison with (single-agent) RL. For the classic single-agent RL case ($n = 1$), Jin et al. (2020) obtained, with the regret metric with respect to the optimal policy of the underlying MDP, the bound $\widetilde{\mathcal{O}}(\sqrt{K}\sqrt{d^3H^4})$. Thus, our result is essentially larger by a factor H – thus nearly-matching the sample efficiency. Having to learn a Nash equilibrium of an MG thus requires more samples than what would be necessary for an MDP. It is important to highlight that though the single-agent case requires taking an action that maximizes the optimistic Q-function (see (Jin et al., 2020, LSVI-UCB Algorithm)), NQOVI requires solving for Nash equilibrium and thus is computationally more complex.

Comparison with (Hu and Wellman, 2003). The original Nash Q-learning proposed by Hu and Wellman (2003) has as its performance metric the convergence to a Nash equilibrium of the underlying discounted MG. In order to ensure such convergence, they assumed the existence of either global optimal or saddle Nash equilibria uniformly on every stage game – see Definition 2.2. In contrast, since we use regret in the context of episodic MGs, we are interested in the average performance of the computed policies across iterations, with the expectation that it will approximate a Nash equilibrium performance. Therefore, we are not strictly interested in convergence to a *single* Nash equilibrium. For this reason, our proof makes no use of the assumptions across stage games by Hu and Wellman (2003). Their work and ours, though being model-free, use completely different proof techniques.

Comparison with (Liu et al., 2021). The first Nash Q-learning algorithm in (Hu and Wellman, 2003) was designed and analyzed for tabular RL. Motivated by concerns of large or continuous state spaces, we decided to opt for the function approximation regime. As it is known in the literature, a direct translation of the NQOVI algorithm to the tabular

case can be done by letting the feature vector ϕ capture $d = |\mathcal{S}||\mathcal{A}| = |\mathcal{S}| \prod_{i=1}^n |\mathcal{A}_i|$, which would give our regret bound a complexity of $\widetilde{\mathcal{O}}(\sqrt{H^5|\mathcal{S}|^3|(\prod_{i=1}^n |\mathcal{A}_i|)^3 K})$. In the tabular case, Liu et al. (2021) proposed the *Multi-Nash-VI* algorithm which obtains $\widetilde{\mathcal{O}}(\sqrt{H^4|\mathcal{S}|^2|(\prod_{i=1}^n |\mathcal{A}_i|)K})$ – tighter in horizon H and both sizes of the state and action spaces of the agents. Interestingly, in the tabular case, both NQOVI and Multi-Nash-VI are of different nature, since the former is model-free and the latter model-based. Interestingly as well, Multi-Nash-VI requires the computation of two Q-functions based on the constructed model – one using optimism and another using pessimism –, whereas NQOVI requires only the computation of an optimistic Q-function. As generally expected in general-sum MGs, both suffer from the *curse of multi-agents* in the tabular case since the sample bounds have exponential dependence on the number of agents (through the product of the cardinality of the agents' action spaces) (Song et al., 2022).

4 PROVING THE MAIN RESULT

We first present two lemmas that make use of the fact that we solve for Nash equilibria in the stage games. All missing proofs and results are in the supplementary material.

Lemma 4.1 (Bounding the covering number)**.** *Let $i \in [n]$, and let $\bar{w}_i \in \mathbb{R}^d$ be such that $\|\bar{w}_i\| \leq L$, $\bar{\Lambda} \in \mathbb{R}^{d\times d}$ be such that its minimum eigenvalue is greater or equal than λ, and, for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, let $\phi(x, a) \in \mathbb{R}^d$ be such that $\|\phi(x, a)\| \leq 1$, and let $\beta > 0$. Define the function class*

$$\begin{aligned}\mathcal{V}_i = \Big\{ V : \mathcal{S} \to \mathbb{R} \;\Big|\; V(\cdot) \\ = \max_{\nu \in \Delta(\mathcal{A}_i)} \mathbb{E}_{\substack{a_i \sim \nu \\ a_{-i} \sim \pi_{-i}(\cdot)}} \Big[\min \Big\{ \bar{w}_i^\top \phi(\cdot, a) \\ + \beta \sqrt{\phi(\cdot, a)^\top \bar{\Lambda}^{-1} \phi(\cdot, a)}, H \Big\} \Big] \Big\}, \quad (4.1)\end{aligned}$$

where $\pi_{-i}(\cdot) \in \Delta(\mathcal{A}_{-i})$. Let $\mathcal{N}_{\epsilon_i}$ be the ϵ_i-covering number of $\mathcal{V}_i$ with respect to the distance $\mathrm{dist}(V, V') = \sup_{x\in\mathcal{S}} |V(x) - V'(x)|$. Then,

$$\log \mathcal{N}_{\epsilon_i} \leq d \log(1 + 4L/\epsilon) + d^2 \log\big[1 + 8d^{1/2}\beta^2/(\lambda\epsilon^2)\big].$$

In Lemma B.2 of the supplementary material, we introduce an event $\mathcal{E}_i$ which defines a concentration bound over a cumulative quantity of the value function associated to agent $i \in [n]$ across iterations. We use this event in the lemma below.

Lemma 4.2 (Optimism bounds)**.** *Consider the setting of Theorem 3.1. Given the event $\mathcal{E}_i$ defined in Lemma B.2, we have for all $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ that*

$$Q_h^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(x, a) \leq Q_h^{i,k}(x, a) \quad \textit{and} \quad V_h^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(x) \leq V_h^{i,k}(x).$$

The importance of Lemma 4.1 and Lemma 4.2. Lemma 4.1 defines a function class $\mathcal{V}_i$ to which the function $V_h^{i,k}(\cdot) = \mathbb{E}_{a\sim\pi_h^k(\cdot)}[Q_h^{i,k}(\cdot, a)]$ belongs. Indeed, the characterization of $\mathcal{V}_i$ includes the one of a Nash equilibrium for a static game; however, we remark that $\pi_{-i}(\cdot) \in \Delta(\mathcal{A}_{-i})$ in the statement of Lemma 4.1 does not need to be a product measure. Using a covering number argument, Lemma 4.1 would be used to prove a series of results that would end up being used by Lemma 4.2. Fundamentally, Lemma 4.2 makes use of (i) the optimism bonus at each episode – the factor starting with β in line 10 of NQOVI – and (ii) the selection of Nash equilibria across all stage games. Bounding the best-response value functions across agents in Lemma 4.2 is important because it upper bounds one of the terms of the regret, see (2.1). Finally, we end our discussion by pointing out that these lemmas are the only two places in the proof of Theorem 3.1 which makes direct use of the notion of Nash equilibria.

4.1 PROOF SKETCH OF THEOREM 3.1

We present the proof sketch of our main result. A more detailed full version of the proof along with all auxiliary results and necessary proofs are found in the supplementary material.

Let us first condition on the event $\bigcap_{i=1}^n \mathcal{E}_i$ where $\mathcal{E}_i$ is defined in Lemma B.2. Since $\mathbb{P}[\text{not } \mathcal{E}_i] \leq \delta$, applying union bound let us conclude that $\mathbb{P}[\bigcap_{i\in[n]} \mathcal{E}_i] \geq 1 - n\delta$. Conditioning on this event allows us to use Lemma 4.2 for every $i \in [n]$.

For any $k \in [K]$, given the policy $\pi^k = \{\pi_i^k\}_{i\in[n]}$ defined by NQOVI, we define the functions $\widehat{Q}_h^k$ and $\widehat{V}_h^k$ recursively as: $\widehat{V}_{H+1}^k(x) = \widehat{Q}_{H+1}^k(x) = 0$ and

$$\begin{aligned}\widehat{Q}_h^k(x, a) &= \mathbb{P}_h \widehat{V}_{h+1}^k(x, a) + 2\beta\sqrt{(\phi_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k}, \\ \widehat{V}_h^k(x) &= \mathbb{E}_{a\sim\pi_h^k(x)}[\widehat{Q}_h^k(x, a)]\end{aligned}$$

for any $h = H, \ldots, 1$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$. Notice that since $2\beta\sqrt{(\phi_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k} \leq 2\beta\sqrt{(\phi_h^k)^\top \phi_h^k} = 2\beta \left\|\phi_h^k\right\| \leq 2\beta$, we have that $\widehat{Q}_h^k$ and $\widehat{V}_h^k$ are nonnegative with maximum value $2\beta H$.

Let $k \in [K]$. We can show that for any $(h, x, a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}\max_{i\in[n]} (Q_h^{i,k}(x, a) - Q_h^{i,\pi^k}(x, a)) \leq \widehat{Q}_h^k(x, a), \text{ and} \\ \max_{i\in[n]} (V_h^{i,k}(x) - V_h^{i,\pi^k}(x)) \leq \widehat{V}_h^k(x).\end{aligned} \quad (4.2)$$

We now introduce the following notation: $\delta_h^k := \mathbb{E}_{a\sim\pi_h^k(x_h^k)}[\widehat{Q}_h^k(x_h^k, a)] - \widehat{Q}_h^k(x_h^k, a_h^k)$, and $\xi_{h+1}^k := \mathbb{P}_h \widehat{V}_{h+1}^k(x_h^k, a_h^k) - \widehat{V}_{h+1}^k(x_{h+1}^k)$ with $\xi_1^k := 0$. Then, for

any $(h,k) \in [H] \times [K]$, we can show that

$$\widehat{V}_h^k(x_h^k) = \delta_h^k + \xi_{h+1}^k + 2\beta\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k} + \widehat{V}_{h+1}^k(x_{h+1}^k).$$

Now, let us focus on the regret performance metric.

$$\begin{aligned}\text{Regret}(K) &= \sum_{k=1}^K \max_{i\in[n]}(V_1^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(s_o) - V_1^{i,\pi^k}(s_o)) \\ &\overset{(a)}{\leq} \sum_{k=1}^K \max_{i\in[n]}(V_1^{i,k}(s_o) - V_1^{i,\pi^k}(s_o)) \\ &\overset{(b)}{\leq} \sum_{k=1}^K \widehat{V}_1^k(s_o) \\ &= \underbrace{\sum_{k=1}^K\sum_{h=1}^H \xi_h^k}_{\text{(I)}} + \underbrace{\sum_{k=1}^K\sum_{h=1}^H \delta_h^k}_{\text{(II)}} \\ &\quad + \underbrace{2\beta\sum_{k=1}^K\sum_{h=1}^H \sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k}}_{\text{(III)}},\end{aligned} \tag{4.3}$$

where (a) follows from Lemma 4.2 and the fact that we are conditioned on the event $\bigcap_{i=1}^n \mathcal{E}_i$; (b) follows from (4.2).

We first analyze the term (I) from (4.3). By defining an appropriate infinite sequence of tuples $\mathcal{L}^\star \subset \mathbb{Z}_{\geq 1} \times [H]$, we can show that $\{\xi_h^k\}_{(k,h)\in\mathcal{L}^\star}$ is a martingale difference sequence. Therefore, we can use the Azuma-Hoeffding inequality to conclude that, for any $\epsilon > 0$,

$$\Pr\left(\sum_{k=1}^K\sum_{h=1}^H \xi_h^k > \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{(KH)(16\beta^2H^2)}\right).$$

We choose $\epsilon = \sqrt{8KH^3\beta^2\log\left(\frac{1}{\delta}\right)}$. Then, with probability at least $1-\delta$,

$$\text{(I)} = \sum_{k=1}^K\sum_{h=1}^H \xi_h^k \leq \sqrt{8KH^3\beta^2\log\left(\frac{1}{\delta}\right)} \leq 8\beta H\sqrt{KH\iota}, \tag{4.4}$$

when setting $\iota = \log\left(\frac{dKH}{\delta}\right)$. We call $\bar{\mathcal{E}}$ the event such that (4.4) holds.

The term (II) can be analyzed in a very similar way as in (I) to show that $\{\delta_h^k\}_{(k,h)\in\mathcal{L}^*}$ is a martingale difference sequence, and thus obtain that with probability at least $1-\delta$,

$$\text{(II)} = \sum_{k=1}^K\sum_{h=1}^H \delta_h^k \leq 8\beta H\sqrt{KH\iota}. \tag{4.5}$$

We call $\widetilde{\mathcal{E}}$ the event such that (4.5) holds.

We now analyze the term (III) from (4.3).

$$\begin{aligned}\text{(III)} &= 2\beta\sum_{h=1}^H\sum_{k=1}^K \sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k} \\ &\overset{(a)}{\leq} 2\beta\sum_{h=1}^H \sqrt{K}\sqrt{\sum_{k=1}^K(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k} \\ &\overset{(b)}{\leq} 2\beta H\sqrt{2dK\iota},\end{aligned} \tag{4.6}$$

where (a) follows from the Cauchy-Schwartz inequality, and we can show (b) by using the so-called elliptical potential lemma (Abbasi-yadkori et al., 2011, Lemma 11).

Now, using the results in (4.4), (4.5), and (4.6) back in (4.3), we conclude that,

$$\begin{aligned}\text{Regret}(K) &\leq 8\beta H\sqrt{KH\iota} + 8\beta H\sqrt{KH\iota} + 2\beta H\sqrt{dK\iota} \\ &= 16c_\beta\sqrt{d^2KH^5\iota^2} + 2c_\beta\sqrt{d^3KH^4\iota^2} \\ &\overset{(a)}{\leq} 18c_\beta\sqrt{d^3KH^5\iota^2},\end{aligned} \tag{4.7}$$

where (a) follows from $\sqrt{\iota} \leq \iota$.

Finally, applying union bound let us conclude that $\mathbb{P}[\bigcap_{i\in[n]}\mathcal{E}_i \cap \bar{\mathcal{E}} \cap \widetilde{\mathcal{E}}] \geq 1-(n+2)\delta$, i.e., our final result holds with probability at least $1-(n+2)\delta$. We set the change of variables $\delta' := (n+2)\delta$ so that all results hold with probability at least $1-\delta'$ and the regret bound has now a logarithmic dependence $\iota = \log\left(\frac{dKH(n+2)}{\delta'}\right)$. This finishes the proof of Theorem 3.1.

5 CONCLUSION

We have shown the sample-efficiency of Nash Q-learning under linear function approximation – ideal for large state spaces or continuous ones – by making use of the principle of optimism in the face of uncertainty – largely exploited in the modern RL literature. We also compared our result to the sample complexity obtained for single-agent RL with linear function approximation and for general-sum MARL on the tabular case. We hope our work may open the path to the future analysis of a more diverse set of MARL algorithms.

One future research direction is obtaining sample performance lower bounds to analyze the (closeness to) minimax optimality of general-sum MARL algorithms such as Nash Q-learning. Moreover, though most modern theoretical work in RL (including this paper) mostly focus on sample efficiency, it is relevant to propose and study algorithms that are also computational efficient – for which other weaker solutions to MGs such as CE and CCE are important. Finally, another future direction would be to expand the analysis of Nash Q-learning to nonlinear function approximators such as neural networks.

Acknowledgements

We are grateful to all the reviewers and the meta-reviewer for their time and their comments to improve our paper. This work is partially supported by NSF III 2046795, CCF 1934986, NIH 1R01MH116226, NIFA award 2020-67021-32799, the Alfred P. Sloan Foundation, and Google Inc.

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Settling the complexity of computing two-player nash equilibria. *Journal of the ACM*, 56(3), 2009.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture Markov games. In *International Conference on Algorithmic Learning Theory*, pages 227–261. PMLR, 2022.

Pedro Cisneros-Velarde, Boxiang Lyu, Sanmi Koyejo, and Mladen Kolar. One policy is enough: Parallel exploration with a single policy is near-optimal for reward-free reinforcement learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 1965–2001. PMLR, 2023.

Qiwen Cui, Kaiqing Zhang, and Simon S. Du. Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. *arXiv preprint arXiv:2302.03673*, 2023.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009. doi: 10.1137/070699652.

Joris Dinneweth, Abderrahmane Boubezoul, René Mandiau, and Stéphane Espié. Multi-agent reinforcement learning for autonomous vehicles: a survey. *Autonomous Intelligent Systems*, 2(27), 2022.

Abhimanyu Dubey and Alex Pentland. Provably efficient cooperative multi-agent reinforcement learning with function approximation. *arXiv preprint arXiv:2103.04972*, 2021.

Elham Foruzan, Leen-Kiat Soh, and Sohrab Asgarpoor. Reinforcement learning approach for optimal distributed energy management in a microgrid. *IEEE Transactions on Power Systems*, 33(5):5749–5758, 2018. doi: 10.1109/TPWRS.2018.2823641.

Minbo Gao, Tianle Xie, Simon S. Du, and Lin F. Yang. A provably efficient algorithm for linear Markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz de Cote. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2019.

Junling Hu and Michael P. Wellman. Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, (4):1039–1069, 2003.

Jiawei Huang, Jinglin Chen, Li Zhao, Tao Qin, Nan Jiang, and Tie-Yan Liu. Towards deployment-efficient reinforcement learning: Lower bound and optimality. In *International Conference on Learning Representations*, 2021.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning – a simple, efficient, decentralized algorithm for multiagent RL. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, 2021.

Tadashi Kozuno, Pierre Ménard, Remi Munos, and Michal Valko. Learning in two-player zero-sum partially observable Markov games with perfect recall. *Advances in Neural Information Processing Systems*, 34, 2021.

Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent RL in markov games with a generative model. In *Advances in Neural Information Processing Systems*, 2022.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, 2022.

Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Chi Jin, and Mengdi Wang. Representation learning for general-sum low-rank markov games. *arXiv preprint arXiv:2210.16976*, 2022.

Jianjun Ni, Minghua Liu, Li Ren, and Simon X. Yang. A multiagent q-learning-based optimal allocation approach for urban water resource management system. *IEEE Transactions on Automation Science and Engineering*, 11 (1):204–214, 2014. doi: 10.1109/TASE.2012.2229978.

Huy Xuan Pham, Hung Manh La, David Feil-Seifer, and Aria Nefian. Cooperative and distributed reinforcement learning of drones for field coverage. *arXiv preprint arXiv:1803.07250*, 2018.

Shuang Qiu, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. On reward-free RL with kernel and neural function approximations: Single-agent MDP and Markov game. In *International Conference on Machine Learning*, 2021.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 2017.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2022.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354, 2019.

Andrew Wagenmaker, Yifang Chen, Max Simchowitz, Simon S Du, and Kevin Jamieson. Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456. PMLR, 2022.

Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 33, pages 17816–17826, 2020.

Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. *arXiv preprint arXiv:2302.06606*, 2023.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10746–10756. PMLR, 13–18 Jul 2020.

Lingxiao Yang, Qiuye Sun, Dazhong Ma, and Qinglai Wei. Nash q-learning based equilibrium transfer for integrated energy management game with we-energy. *Neurocomputing*, 396:216–223, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.01.109.

Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2021.

Kaiqing Zhang, Sham Kakade, Tamer Basar, and Lin Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33:1166–1178, 2020.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

Martin Zinkevich, Amy Greenwald, and Michael Littman. Cyclic equilibria in markov games. In *Advances in Neural Information Processing Systems*, volume 18, 2005.