# Provably Efficient Decentralized Communication for Multi-Agent RL

**Justin Lidard,**[1] **Udari Madhushani,**[1] **Naomi Ehrich Leonard** [1]

[1] Princeton University, 41 Olden Street, Princeton, NJ 08544, USA
{jlidard, udarim, naomi}@princeton.edu

## Abstract

A challenge in reinforcement learning (RL) is minimizing the cost of sampling associated with exploration. Distributed exploration reduces sampling complexity in multi-agent RL (MARL). We investigate the benefits to performance in MARL when exploration is fully decentralized. Specifically, we consider a class of online, episodic, tabular $Q$-learning problems under time-varying reward and transition dynamics, in which agents can communicate in a decentralized manner.We show that group performance, as measured by the bound on regret, can be significantly improved through communication when each agent uses a decentralized message-passing protocol, even when limited to sending information up to its $\gamma$-hop neighbors. We prove regret and sample complexity bounds that depend on the number of agents, communication network structure and $\gamma$. We show that incorporating more agents and more information sharing into the group learning scheme speeds up convergence to the optimal policy.

## Introduction

Multi-agent games capture the salient features of cooperation, competition and mixed motives arise in multi-agent systems. Similar to the single-agent setting, in MARL agents try to maximize their cumulative reward through estimation of *value function*. Typically what sets MARL systems apart from a group of single-agent RL decision makers is the consideration of the *joint action space*; that is, agents must coordinate their efforts in order to converge on the optimal joint policy. We measure this convergence in two ways: *sample complexity*, which bounds the number of reward samples needed to find an approximately optimal value function, and *regret*, which bounds the cumulative value function error over time. We denote an algorithm as *sample efficient* if it has near-optimal sample complexity.

In training MARL algorithms, allowing agents to share value function parameters or reward samples can lead to faster convergence, but it still remains an open question whether this information sharing should be centralized or decentralized. There has been empirical evidence (Rashid et al. 2020; Foerster et al. 2018; Wang et al. 2020) that training using a central controller to aggregate joint states and actions is effective even when the state and action spaces are large. However, as the number of agents increases, the number of joint states and actions becomes exponentially large, demanding more storage space to tabulate scenarios, increasing the *space complexity*. Due to the distributed nature of real-world MARL applications and the combinatorial complexity of MARL, there has been increased interest in developing theory for *decentralized* model-free algorithms that allow agents to train and to perform optimally with space complexity remaining polynomial in the number of agents (Dubey and Pentland 2021; Zhang et al. 2021b,a), particularly when the problem of optimizing the joint action factors as optimizing the individual agent actions.

A key underlying trade-off in RL is known as *exploration* versus *exploitation*; an efficient exploration strategy is always necessary to discover new scenarios while capitalizing on experience from prior scenarios. Upper-confidence bound algorithms, such as those utilized in episodic model-free $Q$-learning (Jin et al. 2018a; Zhang, Zhou, and Ji 2020) literature, choose a time-varying, Hoeffding-style exploration bonus leading to $\tilde{\mathcal{O}}(\sqrt{H^4SAT})$ regret[1]. Here, $S$ is the number of states, $A$ is the number of actions, $H$ is the number of steps per episode, and $T$ is the total number of reward samples. The exploration bonus is chosen to match the value function error up to a constant factor.

Multi-agent reinforcement learning changes the landscape of exploration because multiple agents interact with a shared environment simultaneously and share information. A naive application of the single-agent result running in parallel gives $\tilde{\mathcal{O}}(M\sqrt{H^4SAT})$ regret, where $M$ is the number of agents. In this paper, we show how tabular $Q$-learning can give $\tilde{\mathcal{O}}(\sqrt{MH^4SAT})$ regret by providing an optimal exploration strategy that considers communication among the agents to accelerate online learning.

Our results apply to tasks in which each agent in a network learns an optimal value function under the episodic Markov decision process (MDP) paradigm. We focus on the scenario where agents interact with their respective MDP in parallel, and there is no coupling between joint actions and the joint reward. Agents are allowed to perform one action and one round of message passing per time step. To this end, we investigate the principle of *exploration by proxy*.



[1] $\tilde{\mathcal{O}}$ ignores log terms.

Specifically, we ask the question: if agents operate in similar environments, can they share samples in such a way that enables discovery of the optimal policy faster than individual agents operating in parallel? To this end, we propose an algorithm that permits *full decentralization* of the learning process in which agents explore proportional to the amount of information they receive from other agents. Agents use a message-passing communication scheme where an agent can send and receive information up to its $\gamma$-hop neighbors.

**Contributions.** The key contributions of this paper are as follows: (1) we provide a novel multi-agent UCB $Q$-learning algorithm in which agents use message passing to share information, and (2) we show that group performance, as measured by the bound on regret, improves upon the Hoeffding-style exploration strategy in the single agent setting by a factor of $\sqrt{1/M}$. Moreover, our regret takes into consideration the network structure and communication threshold $\gamma$, suggesting that even mild communication leads to improvement in the regret. As far as we know, this paper provides the first multi-agent regret bound for decentralized tabular $Q$-learning for a general network.

## Related Work

**Episodic $Q$-learning.** Online $Q$-learning has become a popular approach when agents do not have access to a generator/simulator. Jin et al. (2018a) provide both $\mathcal{O}(\sqrt{H^4SAT})$ and $\mathcal{O}(\sqrt{H^3SAT})$ theoretical regret bounds for episodic $Q$-learning under time-varying dynamics, i.e. state transitions and reward structures. The extra $\sqrt{H}$ factor can be attributed to accurate estimation of the moments of the empirical value function. Zhang, Zhou, and Ji (2020) improve this regret bound to $\mathcal{O}(\sqrt{H^2SAT})$, which is proved in (Jin et al. 2018a; Jin, Liu, and Miryoosefi 2021) to be minimax-optimal in the single-agent case.

**Fully Decentralized Multi-Agent Reinforcement Learning.** Theoretical analysis of decentralized reinforcement learning is still a new and growing field. Several works provide a decentralized actor-critic approach to convergence guarantees in the cooperative setting (Zhang et al. 2018; Zhang, Yang, and Basar 2018), although no finite-time sample complexity results are provided. Zhang et al. (2021a) provide $\mathcal{O}(\varepsilon^{-2})$ sample complexity analysis for cooperative actor-critic under a general utility function by empirically estimating the state-action occupancy measures; our method differs by being value-based only and does not require differentiating the cumulative reward with respect to the occupancy measures. Zhang et al. (2021b) provide a finite-sample PAC bound in the cooperative and competitive batch (i.e. not online) settings. In contrast, Zhang et al. (2020) and Sayin et al. (2021) provide a finite sample analysis for MARL in zero-sum Markov games, but emphasize only the competitive setting. Arslan and Yüksel (2017) consider decentralized $Q$-learning, but also for competitive MARL only. For continuous state and action spaces, Asghari, Ouyang, and Nayyar (2020) provide a decentralized multi-agent regret bound for linear-quadratic systems for unknown dynamics and a one-directional communication from the agent controlling the unknown system to the other agents, suggesting that ideas from discrete $Q$-learning algorithms (such as ours) can be readily extended to continuous domains. Dubey and Pentland (2021) provide multi-agent regret bounds for cooperative RL for parallel MDPs (see Preliminaries). A linear function approximation and a central server are used to perform least-squares value iteration with shared transition samples. Our work complements these results by providing the first decentralized multi-agent regret bound for tabular $Q$-learning, with message passing eliminating the need for a central server. In the sequel, we show further that our bound matches centralized benchmarks.

## Mathematical Preliminaries

We consider *parallel MDPs* (Dubey and Pentland 2021; Bernstein et al. 2002), which are defined as a collection $\{\textbf{MDP}(\mathcal{S}_i, \mathcal{A}_i, \mathbb{P}_i, r_i, \gamma_d)\}_{i=1}^M$, where each agent $i \in [M]$ has access to identical state space $\mathcal{S}_i = \mathcal{S}_j$ and action space $\mathcal{A}_i = \mathcal{A}_j$, $\forall i, j \in [M]$. The joint state space is $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times ... \times \mathcal{S}_M$. The joint action space is similarly $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times ... \times \mathcal{A}_M$. Again we take $\gamma_d = 1$.The local reward functions $r_i$ and transitions $\mathbb{P}_i$ can be different but only depend on the local state and action information of each agent. Thus, each agent interacts *in parallel* with their corresponding MDP, and there is no coupling between the state and actions chosen by an agent and the reward received by any other agent. Here, we let $r = r_1 = ... = r_M$ and $\mathbb{P} = \mathbb{P}_1 = ... = \mathbb{P}_M$, such that every agent interacts with the same MDP.

Dubey and Pentland (2021) provide a multi-agent regret bound under heterogeneous reward and transition structure using estimation of the feature covariance and bias. In contrast, our framework permits direct sharing of transition-reward tuples where the $Q$-function can be tabulated directly. We let the $M$ agents communicate over a network $G$ with edge set $E$. In the episodic setting, each agent $m \in [M]$ gets a potentially time-varying local reward $r_h(x_{m,h}^k, a_{m,h}^k)$ for the state-action pair it chooses at step $h$ of episode $k$. In the Parallel MDP construction, we assume that average total reward and transition probability are decoupled:

$$r_{tot}(\mathbf{x}, \mathbf{a}) := \frac{1}{M} \sum_{i=1}^{M} r_{i,h}(x_{i,h}, a_{i,h})$$

$$P(\mathbf{x}'|\mathbf{x}, \mathbf{a}) := \prod_{i=1}^{M} P(x_i'|x_i, a_i).$$

The objective is then to maximize the global (average total) reward through optimization of local (individual) rewards. The cumulative value error is measured by the *group regret*:

$$\text{Regret}_G(K) := \sum_{m=1}^{M} \text{Regret}_m(K)$$

Here, we show that the local optimizations are provably accelerated through communication.

# MARL with Full Communication

## Algorithm

We consider a multi-agent extension to the online $Q$-learning algorithm provided by (Jin et al. 2018a). Consider $M$ agents operating in a parallel MDP for $K$ episodes of length $H$. In the single-agent case, each agent makes one update corresponding to the state-action pair $(x_h^k, a_h^k)$ visited at each step $h$ of episode $k$, for a total of $T = HK$ samples. In the multi-agent case, at each time $\tau$, each agent $m \in [M]$ sees a state-action pair $(x_{m,h}^k, a_{m,h}^k)$, reward $r_h(x_{m,h}^k, a_{m,h}^k)$, and next state $x_{m,h+1}^k$ and exchanges messages with its neighbors. We consider a message-passing protocol in which each agent $m \in M$ sends to its neighbors a message $m_h^k := \langle h, k, m, x_{m,h}^k, a_{m,h}^k, x_{m,h+1}^k, r_h^k \rangle$ containing step, episode, agent id, current state, current action, next state and current reward. Each neighbor then forwards the message to its neighbors. All messages older than $\gamma$ are excluded. Here $0 \leq \gamma \leq D(G)$, where $D(\cdot)$ is graph diameter. The message-passing protocol allows each agent to send information up to $\gamma$-hop neighbors. For $\gamma = 0$, we recover $M$ copies of the single-agent case with no communication and group regret $\mathcal{O}(M\sqrt{T})$.

We make several definitions for the $Q$-learning update.

**Definition 1.** (Number of state-action observations)

$$N_{m,h}^k(x,a) := \sum_{\ell=1}^{k} \sum_{j=1}^{M} \mathbb{1}[x_{j,h}^\ell = x, a_{j,h}^\ell = a] \mathbb{1}[(m,j) \in E], \tag{1}$$

Counting the state-action visitations is crucial for controlling the error due to exploration in UCB-style algorithms. For any set $S$ of agents, we take $N_{S,h}^k(x,a)$ to be the smallest number of samples available to any agent $m \in S$.

Consider the power graph $G_\gamma$ of $G$: nodes $m$ and $m'$ share an edge if and only if $d(m,m') \leq \gamma$. Let $G_\gamma(m)$ be the neighbors of $m$ in $G_\gamma$. Define set $\mathcal{V}_{m,h}^k(x,a) = \mathcal{V}$,

$$\mathcal{V} = \begin{cases} \left\{ \bigcup_{i \in G_\gamma(m)} (r_{i,h}^{k-1}, x_{i,h+1}^{k-1}) \right\} & \exists i : N_{i,h}^k(x,a) > 0 \\ \emptyset & \text{o.w.} \end{cases}$$

to be the set of all *new* reward and next-state tuples available to each agent $m$ for state-action pair $(x,a)$ for any agent across the network at step $h$ and episode $k$. Define $\mathcal{U}_{m,h}^k$ to be the set of reward and next-state tuples taken by and observed by agent $m$. Clearly, $\mathcal{U}_{m,h}^k(x,a) = \{(r_{m,h}^{k-1}, x_{m,h+1}^{k-1})\}$. The update rule for $\mathcal{V}_{m,h}^k(x,a)$ is

$$\mathcal{V}_{m,h}^k = \mathcal{U}_{m,h}^k \cup \{ \bigcup_{m' \in G} \mathcal{V}_{m',h-d(m,m')+1}^k \},$$

where $\mathcal{V}_{m,h}^k(x,a) = \emptyset$ if $N_{m,h}^k(x,a) = 0$ and $h < 0$.

Assume for fixed $(x,a,h,k)$, $|\mathcal{V}_{m,h}^k(x,a)| > 0$. Let $t_m = N_{m,h}^k(x,a)$. The $Q$-learning update is

$$Q_{m,h}^{k+1}(x,a) := \sum_{(r,x') \in \mathcal{V}_{m,h}^k(x,a)} (1-\alpha_{m,t}) Q_{m,h}^k(x,a) + \alpha_{m,t}[r + V_{m,h+1}^k(x') + b_{m,t}]$$

The online multi-agent $Q$-learning algorithm with Hoeffding-style upper confidence bound is provided in Algorithm 1, where $\mathcal{C}(m)$ denotes the clique size of agent $m$ in the clique cover of $G_\gamma$.

Let $t_m = N_{m,h}^k(x,a)$ and suppose $(x,a)$ was previously taken at step $h$ of episodes $k_1, ..., k_{t_m} < k$, for each agent $m \in [M]$. Let $t = \sup_{m \in [M]} t_m$ and $k_1, ..., k_t < k$ denote the episodes where $(x,a)$ was visited for any $m \in [M]$. Let $N_h^k(x,a) := \sum_{m \in [M]} N_{m,h}^k(x,a)$ represent the total number of times $(x,a)$ is seen by all agents. This makes the episode-wise $Q$-function determined as:

$$Q_{m,h}^k(x,a) := \alpha_{m,t}^0 H + \sum_{(r,x') \in \mathcal{V}_{m,h}^i(x,a)}^{t} \alpha_{m,t}^i [r + V_{m,h+1}^{k_i}(x') + b_{m,t}].$$

As in (Jin et al. 2018a), the $Q$-learning update is highly asynchronous. Thus, the regret bound depends on optimal choice of $\alpha_{m,t}$, $b_{m,t}$, network structure and $\gamma$.

## Convergence proof

Let $(x,a)$ be a state-action pair, $m$ an arbitrary agent (estimating the value function and $Q$-function) and $(h,k)$ a step and episode index, respectively. We briefly state some notation from (Jin et al. 2018b) extended to multiple agents.

**Definition 2.** (Estimated policy performance gap)

$$\delta_{m,h}^k := (V_{m,h}^k - V_{m,h}^{\pi_{m,k}})(x_{m,h}^k)$$

**Definition 3.** (Value estimation error)

$$\phi_{m,h}^k := (V_{m,h}^k - V_{m,h}^*)(x_{m,h}^k)$$

**Definition 4.** (Value gap due to modeling error)

$$\xi_{m,h}^k := (\mathbb{P}_h - \hat{\mathbb{P}}_h)(V_{m,h+1}^* - V_{m,h+1}^{\pi_{m,k}})(x_{m,h}^k, a_{m,h}^k)$$

Note that $\xi_{m,h}^k$ is a martingale difference sequence. The key idea is to bound the estimated performance gap $\delta_{m,h}^k$ recursively for a given $h$ in terms of $\delta_{m,h+1}^k$, $\phi_{m,h+1}^k$, and $\xi_{m,h+1}^k$.

**Assumption 1.** (Episodic length bounds message life). Assume $0 \leq \gamma \leq \min(D(G), H)$.

Lemmas 1 and 2 are reproduced from (Jin et al. 2018a).

**Lemma 1.** *Let $\alpha_{m,t} := \frac{H+1}{H+t}$, where $t$ denotes the number of times a state-action pair has been sampled, i.e. $t := N_{m,h}^k(x,a)$. Let $\alpha_{m,t}^0 := \prod_{i=1}^t (1-\alpha_{m,i})$ and $\alpha_{m,t}^i := \alpha_{m,i} \prod_{j=i+1}^t (1-\alpha_{m,j})$. The following hold for $\alpha_{m,t}^i$:*

*(a)* $\frac{1}{\sqrt{t}} \leq \sum_{i=1}^t \frac{\alpha_{m,t}^i}{\sqrt{i}} \leq \frac{2}{\sqrt{t}}$.

*(b)* $\max_{i \in [t]} \alpha_{m,t}^i \leq \frac{2H}{t}$ *and* $\sum_{i=1}^t (\alpha_{m,t}^i)^2 \leq \frac{2H}{t}$ *for every* $t \geq 1$.

*(c)* $\sum_{t=i}^\infty \alpha_{m,t}^i = 1 + \frac{1}{H}$ *for every* $i \geq 1$.

**Lemma 2.** (Recursion on $Q$) For any $(m,x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, episode $k \in [K]$, let $t = N_h^k(x,a)$, and $(x,a)$ be

Algorithm 1: Multi-Agent $Q$-learning with UCB-Hoeffding

**INPUT:** $Q_{m,h}(x,a) \leftarrow H$ and $N_{m,h}(x,a) \leftarrow 0$ for all $(x,a,h,m) \in \mathcal{S} \times A \times [H]$

**for** episode $k = 1, ..., K$ **do**
  receive vector $x_1$.
  **for** step $h = 1, ..., H$ **do**
    **for** agent $m = 1, ..., M$ **do**
      Take action $a^k_{m,h} \leftarrow \arg\max'_a Q^k_{m,h}(x^k_{m,h}, a')$ and observe $x^k_{m,h+1}$.
      Update $\mathcal{V}^k_{m,h}, \mathcal{U}^k_{m,h}$ via message passing.
      **for** each sample $(r, x')$ such that $\mathcal{V}^k_{m,h}(x,a) \neq \emptyset$ **do**
        $N^k_{m,h}(x,a) \leftarrow N^k_{m,h}(x,a) + 1$
        $t \leftarrow N^k_{m,h}(x,a)$
        $b_{m,t} \leftarrow c\sqrt{H^3\iota/(\mathcal{C}(m),t)}$
        $Q^k_{m,h}(x,a) \leftarrow (1-\alpha_{m,t})Q^k_{m,h}(x,a) + \alpha_{m,t}[r(x,a) + V^k_{m,h+1}(x') + b_{m,t}]$
        $V^k_{m,h}(x_h) \leftarrow \min\{H, \max_{a' \in \mathcal{A}} Q^k_{m,h}(x^k_{m,h}, a')\}$

observed by agent $m$ at episodes $k_1, ..., k_t < k$. Then

$$(Q^k_h - Q^*_h)(x,a) = \alpha^0_{m,t}(H - Q^*_h(x,a)) + \sum_{i=1}^{t} \alpha^i_{m,t}(V^{k_i}_{m,h+1} - V^*_{m,h+1})(x^k_{m,h}) + \sum_{i=1}^{t}[(\hat{\mathbb{P}}^{k_i}_h - \mathbb{P}_h)V^*_{m,h+1}](x^k_{m,h}), a^k_{m,h})) + b_{m,t}.$$

**Lemma 3.** (Bound on $\xi^k_{m,h}$ accumulation)

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H} \xi^k_{m,h} \leq \sqrt{2H^3MT\log\left(\frac{2}{p}\right)}$$

*Proof.* (Based on Dubey & Pentland (Dubey and Pentland 2021)). We know from Def. 4 that $\xi^k_{m,h}$ represents a martingale difference sequence. A straightforward application of the Azuma-Hoeffding inequality gives

$$\mathbb{P}\left(\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{h=1}^{H} \xi^k_{m,h} > \varepsilon\right) \leq \exp\left(\frac{-\varepsilon^2}{2MH^3K}\right).$$

Solving for $\varepsilon$ completes the proof. □

**Lemma 4.** (Clique bound on $Q$-error) Let $\mathbf{C}_\gamma$ be a partition of $G_\gamma$. If for every clique $\mathcal{C} \in \mathbf{C}_\gamma$ we assign exploration bonus $b_{m,t} = b_{\mathcal{C},t} = c\sqrt{H^3\iota/(|\mathcal{C}|t)}$ $\forall m \in \mathcal{C}$, we have bounded error rate $e_{m,t} = e_{\mathcal{C},t} = 2\sum_{m\in\mathcal{C}}\sum_{i=1}^{t_m}\alpha^i_{1,t}b_i \leq 4c\sqrt{|\mathcal{C}|H^3\iota/t}$, $\forall m \in \mathcal{C}$, where $t_m = N^k_{m,h}(x,a)$ is the number of times $(x,a)$ has been seen by step $h$ in episode $k$ by agent $m$. Further, there exists an absolute constant $c > 0$ such that, for any $p \in (0,1)$, with probability at least $1-p$, we have simultaneously for all $(x,a,h,k,m) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K] \times [M]$:

$$0 \leq \sum_{m\in\mathcal{C}}(Q^k_{m,h} - Q^*_{m,h})(x,a) \leq |\mathcal{C}|\alpha^0_{1,t}H + \sum_{m\in\mathcal{C}}\sum_{i=1}^{t_m}(V^{k_i}_{m,h+1} - V^*_{m,h+1})(x^{k_i}_{m,h+1}) + e_{\mathcal{C},t}.$$

*Proof.* First, consider a clique $\mathcal{C} \in \mathbf{C}$. Assume $\gamma > 0$. Fix $(x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Let $k_i$ be the episode where $(x,a)$ is seen for the $i$th time (by any agent), and $k_i = K+1$ if $(x,a)$ hasn't been seen for the $i$th time yet. Thus, $k_i$ is a random variable and a stopping time. Next, let $X_{m,i} = \mathbb{1}[k_i \leq K] \cdot [(\hat{\mathbb{P}}^{k_i}_h - \mathbb{P}_h)V^*_{m,h+1}](x,a)$. For filtration $\{\mathcal{F}_i\}_{i\geq 0}$, it can be seen that $\{X_i\}_{i\geq 0}$ is a martingale difference sequence by taking $\mathbb{E}[X_{m,i}|\mathcal{F}_{i-1}]\mathbb{1}[k_i \leq K] \cdot \mathbb{E}[(\hat{\mathbb{P}}^{k_i}_h - \mathbb{P}_h)V^*_{m,h+1}|\mathcal{F}_{i-1}] = \mathbb{1}[k_i \leq K] \cdot \mathbb{E}[V^*_{m,h+1}(x^{k_i}_{h+1}) - \mathbb{E}_{x'\in\mathbb{P}_h(\cdot|x,a)}V^*_{m,h+1}(x')|\mathcal{F}_{i-1}] = 0$.

Let $\tau_m$ represent the fixed total number of $Q$-updates made by each agent $m$, and $\tau_\mathcal{C}$ the number of samples generated by the clique. We proceed to bound the augmented trajectory length $\tau_m$ by the total samples generated by the clique. Consider an index set $I_\mathcal{C} \subset [\tau_m]$ for an agent $m \in \mathcal{C}$, which represents the total number of samples generated by the clique. Clearly, $|I_\mathcal{C}| \leq \tau_\mathcal{C}$ since an agent's clique is connected under $G_\gamma$. Let $I_{G/\mathcal{C}} := [\tau_m]/I_\mathcal{C}$, i.e. $|I_{G/\mathcal{C}}| = \tau_M - |I_\mathcal{C}|$. By the Azuma-Hoeffding inequality and a union bound, for fixed $(x,a,h)$, we have for any collection of *fixed* data-sets $\{\tau_m\}_{i=1}^M$ of size $\tau_m \in [0, MK]$ $\forall m$, with probability $1 - p/(MSAH)$:

$$\begin{aligned}
&\left|\sum_{m\in\mathcal{C}}\sum_{i=1}^{\tau_m}\alpha^i_{m,\tau_m}\cdot X_{m,i}\right| \\
&\leq \left|\sum_{m\in\mathcal{C}}\sum_{i\in I_\mathcal{C}}\alpha^i_{m,\tau_m}\cdot X_{m,i}\right| + \left|\sum_{m\in\mathcal{C}}\sum_{i\in I_{G/\mathcal{C}}}\alpha^i_{m,\tau_m}\cdot X_{m,i}\right| \\
&\leq \left|\sum_{m\in\mathcal{C}}\sum_{i=1}^{\tau_\mathcal{C}}\alpha^i_{m,\tau_\mathcal{C}}\cdot X_{m,i}\right| + \left|\sum_{m\in\mathcal{C}}\sum_{i=1}^{\tau_m-\tau_\mathcal{C}}\alpha^i_{m,\tau_m-\tau_\mathcal{C}}X_{m,i}\right| \\
&\leq \sqrt{\frac{|\mathcal{C}|H^3\iota}{\tau_\mathcal{C}}} + \sqrt{\frac{|\mathcal{C}|H^3\iota}{\tau_m - \tau_\mathcal{C}}}
\end{aligned}$$

where $\iota = p/(MSAH)$. This follows since $\{\alpha^i_{m,M}\}_{i=1}^M$ is dominated pointwise by $\{\alpha^i_{m,N}\}_{i=1}^N$ for any $M > N$. Since $\tau \in [0, |\mathcal{C}| \cdot K]$, the above bound also holds for

$\tau_{\mathcal{C}} = t := N^k_{m,h}(x,a)$, which is a random variable. Since $\mathbb{1}[k_i \le K]$ for any $i \le N^k_{m,h}(x,a)$, we have that with probability $1-p$, for all $(m,x,a,h,k) \in [M]\times\mathcal{S}\times\mathcal{A}\times[H]\times[K]$, $|\sum_{m\in\mathcal{C}}\sum_{i=1}^{t}\alpha^i_{\tau_m}\cdot X_{m,i}| \le c\sqrt{|\mathcal{C}|\frac{H^3\iota}{t}}$. Thus, picking exploration bonus $b_{\mathcal{C},t} := c\sqrt{\frac{H^3\iota}{t|\mathcal{C}|}}$,

$$
\begin{aligned}
&\sum_{m\in\mathcal{C}}(Q^k_{m,h} - Q^*_{m,h})(x,a)\\
\le&|\mathcal{C}|\alpha^0_t H + \sum_{m\in\mathcal{C}}\sum_{i=1}^{t_m}\alpha^i_{m,t}(V^{k_i}_{m,h+1} - V^*_{m,h+1})(x^{k_i}_{h+1})\\
&+ 2\sum_{m\in\mathcal{C}}\sum_{i=1}^{t_m}\alpha^i_{m,t}b_{m,t}\\
\le&|\mathcal{C}|\alpha^0_t H + \sum_{m\in\mathcal{C}}\sum_{i=1}^{t_m}\alpha^i_{m,t}(V^{k_i}_{m,h+1} - V^*_{m,h+1})(x^{k_i}_{h+1})\\
&+ 4c\sqrt{|\mathcal{C}|\frac{H^3\iota}{t}}\\
\le&|\mathcal{C}|\alpha^0_t H + \sum_{m\in\mathcal{C}}\sum_{i=1}^{t_m}\alpha^i_{m,t}(V^{k_i}_{m,h+1} - V^*_{m,h+1})(x^{k_i}_{h+1})\\
&+ e_{\mathcal{C},t},
\end{aligned}
$$

where $\alpha^0_t$ denotes the default initial learning rate. Thus, exploration error accumulation $e_{m,t} = e_{\mathcal{C},t} \le 4c\sqrt{|\mathcal{C}|H^3\iota/t}$, $\forall m \in \mathcal{C}$. If the clique structure is not known, we can bound $e_{m,t} \le \sqrt{\mathcal{N}(m)H^3\iota/t} \le e_{\mathcal{C},t}$, where $\mathcal{N}(m)$ represents neighbors of node $m$ in $G$. □

**Remark 1.** This analysis allows exploration *by proxy*. That is, all agents in a clique $\mathcal{C}$ are able to explore proportional to the size of the clique, and share the resulting samples. If $d^*_{\mathcal{C}} \le |\mathcal{C}|$, then $\tau_m - \tau_{\mathcal{C}} = 0$. So assume $d^*_{\mathcal{C}} > |\mathcal{C}|$. We show that the number of samples used by clique $\mathcal{C}$ is approximately generated by $\mathcal{C}$, and the concentration is bounded by the left hand term.

**Lemma 5.** *(Single agent cumulative value error). For fixed $h$, we have*

$$
\sum_{k=1}^{K}\delta^K_{m,1} \le \mathcal{O}\Big(H^2SA + \sum_{h=1}^{H}\sum_{k=1}^{K}(e_{n^k_{m,h}} + \xi^k_{m,h+1})\Big).
$$

*Proof.* See equation 4.8 of (Jin et al. 2018a). □

**Theorem 1.** (Hoeffding regret bound for parallel MDP with communication) There exists an absolute constant $c > 0$ such that, for any $p \in (0,1)$, if we choose $b_{m,t} = b_{\mathcal{C},t} = c\sqrt{H^3\iota/(|\mathcal{C}|t)}$, then with probability $1-p$, the group regret of multi-agent $Q$-learning with Algorithm 1 is at most $\mathcal{O}(\sqrt{MH^4SAT\iota})$, where $\iota := log(SATM/(p))$.

*Proof.* Expanding the group regret defined earlier, we have

$$
\begin{aligned}
\sum_{m=1}^{M}\text{Regret}_m(K) &= \sum_{m=1}^{M}\sum_{k=1}^{K}[V^*_{m,1}(x^k_1) - V^{\pi_{m,k}}_{m,1}(x^k_1)]\\
&\le \sum_{m=1}^{M}\sum_{k=1}^{K}[V^k_{m,1}(x^k_1) - V^{\pi_{m,k}}_{m,1}(x^k_1)]\\
&= \sum_{m=1}^{M}\sum_{k=1}^{K}\delta^k_{m,1}
\end{aligned}
$$

We know from single-agent analysis that the regret at each step $h$ and episode $k$ is bounded recursively as

$$
\begin{aligned}
\delta^k_{m,h} \le (V^k_h - V^{\pi_{m,k}}_h)(x^k_{m,h})\alpha^0_{m,t}H + \sum_{i=1}^{t}\Big\{&\alpha^i_{m,t}\phi^{k_i}_{m,h+1}\\
&+ e_{m,t}\Big\} - \phi^k_{m,h+1} + \delta^k_{m,h+1} + \xi^k_{m,h+1}
\end{aligned}
$$

for $t := n^k_{m,h}(x^k_{m,h}, a^k_{m,h})$ and applying Lemma 3; see (Jin et al. 2018a) for the complete derivation. Extending to a clique performance gap is straightforward:

$$
\begin{aligned}
\sum_{m\in\mathcal{C}}\delta^k_{m,h} &\le \sum_{m\in\mathcal{C}}(V^k_{m,h} - V^{\pi_{m,k}}_{m,h})(x^k_{m,h})\\
&\le |\mathcal{C}|\alpha^0_t H + \sum_{m\in\mathcal{C}}\sum_{\substack{i=1\\(r,x')}}\alpha^i_{m,t}\phi^{k_i}_{m,h+1} + e_{\mathcal{C},t}\\
&\quad + \sum_{m\in\mathcal{C}}\delta^k_{m,h+1} - \phi^k_{m,h+1} + \xi^k_{m,h+1}
\end{aligned}
$$

for $t = N^k_{\mathcal{C},m}(x,a)$. Let $\mathbf{C}_\gamma$ denote a clique covering of $G_\gamma$. From Lemma 5,

$$
\begin{aligned}
&\sum_{m=1}^{M}\sum_{k=1}^{K}\delta^k_{m,1} \le \sum_{\mathcal{C}\in\mathbf{C}_\gamma}\sum_{m\in\mathcal{C}}\sum_{k=1}^{K}\delta^k_{m,1}\\
&\le \mathcal{O}(1)\sum_{\mathcal{C}\in\mathbf{C}_\gamma}\sum_{m\in\mathcal{C}}\sum_{k=1}^{K}e_{m,n^k_{m,h}(x^k_{m,h},a^k_{m,h})}\\
&\le \mathcal{O}(1)\sum_{\mathcal{C}\in\mathbf{C}_\gamma}\sum_{m\in\mathcal{C}}\sum_{k=1}^{K}\left(\sqrt{\frac{|\mathcal{C}|H^3\iota}{N^k_{m,h}(x^k_{m,h},a^k_{m,h})}}\right.\\
&\quad\left.+\sqrt{\frac{|\mathcal{C}|H^3\iota}{N^{G/\mathcal{C},k}_{m,h}(x^k_{m,h},a^k_{m,h})}}\right)\\
&\overset{(1)}{\le} \mathcal{O}(1)\sum_{\mathcal{C}\in\mathbf{C}_\gamma}\sqrt{\frac{KH^3\iota}{d^*_{\mathcal{C}} - |\mathcal{C}|}}\\
&\quad + \mathcal{O}(1)\sum_{\mathcal{C}\in\mathbf{C}_\gamma}\sum_{k=1}^{K}|\mathcal{C}|\sup_{\substack{(x,a):\,\exists m\in\mathcal{C}:\\(x^k_{m,h},a^k_{m,h})\\=(x,a)}}\sqrt{\frac{|\mathcal{C}|H^3\iota}{N^k_{\mathcal{C},h}(x,a)}}\Bigg)\\
&\overset{(2)}{\le} \mathcal{O}(1)\sqrt{\frac{\bar{\chi}(G_\gamma)KH^3\iota}{d^{\text{eff}}_{\mathbf{G}_\gamma}}} + \mathcal{O}(1)\sum_{\mathcal{C}\in\mathbf{C}_\gamma}\sqrt{|\mathcal{C}|H^3SAK\iota}
\end{aligned}
$$

$$\leq \mathcal{O}(1)\sqrt{\frac{\bar{\chi}(G_\gamma)KH^3\iota}{d^{\text{eff}}_{\mathbf{G}_\gamma}}} + \mathcal{O}(1)\sqrt{MH^3SAK\iota}$$
$$= \mathcal{O}(1)\sqrt{\frac{\bar{\chi}(G_\gamma)KH^3\iota}{d^{\text{eff}}_{\mathbf{G}_\gamma}}} + \mathcal{O}(1)\sqrt{MH^2SAT\iota}$$

where (1) comes from the fact that for any $(x, a, k, h)$, $N^{G/\mathcal{C},k}_{\mathcal{C},h}(x, a) \leq k|\mathcal{C}|(d^*_{\mathcal{C}} - |\mathcal{C}|)$, where $d^*_{\mathcal{C}}$ is the max degree in the clique. Expression (2) is due to Cauchy-Schwartz, where $d^{\text{eff}}_{\mathbf{G}_\gamma} := (\sum_{C\in\mathbf{C}_\gamma}(d^*_{\mathcal{C}} - |\mathcal{C}|)^{-1})^{-1}$ and $\bar{\chi}(G_\gamma)$ denotes the clique covering number of the communication power graph $G_\gamma$. In the worst case $N^k_{\mathcal{C},h} = |\mathcal{C}|^2k/SA$. Finally, the total multi-agent regret is bounded as

$$\sum_{m=1}^{M}\text{Regret}_m(K) \leq \sum_{m=1}^{M}\sum_{k=1}^{K}\delta^K_{1,\mathcal{C}}$$
$$\leq \mathcal{O}\Big(MH^2SA + \sum_{m=1}^{M}\sum_{h=1}^{H}\sum_{k=1}^{K}\Big\{e_{n^k_{m,h}}\xi^k_{m,h+1}\Big\}\Big)$$
$$\leq \mathcal{O}\Big(MH^2SA\Big) + \tilde{\mathcal{O}}\Big(\sqrt{2H^3MT}\Big)$$
$$+ \sum_{h=1}^{H}\left\{\mathcal{O}(1)\sqrt{\frac{\bar{\chi}(G_\gamma)KH^3\iota}{d^{\text{eff}}_{\mathbf{G}_\gamma}}} + \mathcal{O}(1)\sqrt{MH^2SAT\iota}\right\}$$
$$= \tilde{\mathcal{O}}\left(\sqrt{\frac{\bar{\chi}(G_\gamma)H^4KT\iota}{d^{\text{eff}}_{\mathbf{G}_\gamma}}} + \sqrt{MH^4SAT\iota} + \sqrt{H^3MT}\right) \quad (2)$$

with probability $1 - p$. □

**Remark 2.** (On sample complexity). Note that Algorithm 1 achieves a high probability regret of $\sum_{m=1}^{M}\sum_{k=1}^{K}[V^*_{m,1}(x^k_{m,1}) - V^{\pi_{m,k}}_{m,1}(x^k_{m,1})] \leq \tilde{\mathcal{O}}(\sqrt{MH^4SAT\iota})$. Suppose that we add the assumption also that $\mathbf{x}^k_1 \leftarrow \mathbf{x}_{\text{nom}}\ \forall k$ for some nominal initial state $x_{\text{nom}}$. Then, by dividing by $M$ and $K$, we see that for any agent $m$ and episode $k$, $V^*_{m,1}(x^k_{m,1}) - \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}V^{\pi_{m,k}}_{m,1}(x^k_{m,1}) \leq \tilde{\mathcal{O}}(\sqrt{H^5SA/\iota(MT)})$. Then, with probability $2/3$, we randomly select an agent $m \in [M]$ and policy $\pi \in \{\pi_{m,1}, ..., \pi_{m,k}\}$ such that $V^\pi_{m,1}(x^k_{m,1}) - V^\pi_{m,1}(x^k_{m,1}) \leq 3\tilde{\mathcal{O}}(\sqrt{H^5SA\iota/(MT)}) < \varepsilon$ under policy $\pi$. Hence, the total number of samples required are $T = \mathcal{O}(\varepsilon^{-2}M^{-1/2}))$. Note that this is an $\mathcal{O}(M^{-1/2})$ reduction in sample complexity from the single-agent case (Jin et al. 2018a). If we do not place the restriction on $\mathbf{x}^k_1 \leftarrow \mathbf{x}_{\text{nom}}\ \forall k$, then we cannot make such a sample complexity assertion, because there is no guarantee $\mathbf{x}^k_1$ has been seen again (Dann, Lattimore, and Brunskill 2017).

**Remark 3.** (Comparison with regret bounds in centralized online learning). Our regret bound matches the group regret bound $\mathcal{O}(\sqrt{MT})$ for centralized communication (Dubey and Pentland 2021).

## Experimental Results

In this section we provide numerical simulations to illustrate results and validate theoretical bounds. We consider a cooperative game where the goal of each agent is navigating to a specific landmark. This is a modified version of Cooperative Navigation task (Lowe et al. 2020; Terry et al. 2021).

Here we consider $M$ agents and $M$ landmarks. Agents are modeled with double-integrator dynamics; the (continuous) state space is 4-dimensional and consists of a planar position and a planar velocity. In our modified environment, agents are assigned a landmark and rewarded based on their distance to the landmark. We discretize the continuous state space into a $10 \times 10 \times 10 \times 10$ grid on the box $\mathcal{S} = [-2, 2]^4$, giving $10^4$ total states. The agents' are initialized to lie within $[-1, 1]^4$; states outside of $\mathcal{S}$ are mapped to the nearest state in $S$. The action space is discrete and consists of movements `left, right, down, up` and `no-op`. The reward assigned to each agent is given as the euclidean distance to their assigned landmark. Note that since each agent is rewarded locally based on their distance to their landmark (instead of by interacting with other agents), the optimal joint policy is attained through convergence to the optimal local policy for each agent. We highlight how communication accelerates convergence to the optimal local policy, including when agents must operate in under-explored regions of the state space. Training is performed according to Algorithm 1, and test performance is given as the average reward accrued over a rollout of length $H = 10$ under the implied policy $\pi^k_{m,h}(s) := \arg\max_{a\in\mathcal{A}} Q^k_{m,h}(s, a)$ for any agent $m$, state $s$, step $h$, and episode $k$. Figures show average performance over 10 trials.

### Simple Communication with 2 Agents

Figure 1: Simple communication scenario

In this scenario, $M = 2$ agents (Red and Blue) must navigate to assigned landmarks that are roughly at the same location (at the origin), as shown in Fig. 1. During training, the initial vector state $x_1$ is chosen such that both agents are at the same position every time: Red starts at position $[-1, 0]$ and Blue at position $[1, 0]$. Therefore, Red and Blue explore regions of the state space that are roughly disjoint, since any optimal trajectory will keep Red on the left half plane and Blue on the right half plane. During testing, the scenario is flipped: Red starts on the right and Blue on the left. Without communication, Red and Blue will exhibit poor rollout performance since for every step $h \in H$, $Q^k_{m,h}(\cdot, a)$ will be at the default value of $H$ and $\pi^k_{m,h}(\cdot)$ is uniform for $m \in \{\texttt{Red}, \texttt{Blue}\}$ . Fig. 2 plots the average reward per rollout versus the training episode.

### Message Passing Communication with 4 Agents

In this scenario, $M = 4$ agents (Red, Green, Magenta, and Blue) must navigate to assigned landmarks that are roughly at the same location (also at the origin). During training, Red

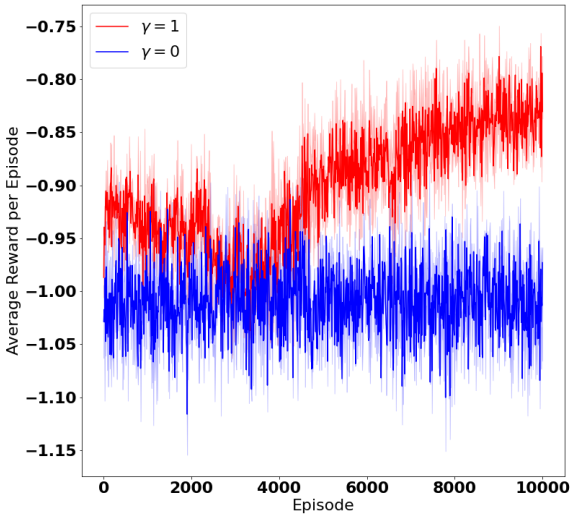


Figure 2: Average group rollout reward achieved per episode for the 2-agent scenario. When the initial states are swapped, communicating agents perform better in the new scenario, while non-communicating agents show little improvement.

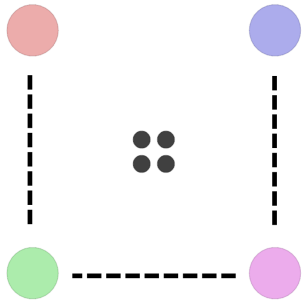

Figure 3: Message passing communication scenario with $M = 4$ agents with the communication network shown as a dashed line.

is assigned initial position $[-1, 1]/\sqrt{2}$, Green $[-1, -1]/\sqrt{2}$, Magenta $[1, -1]/\sqrt{2}$, and Blue $[1, 1]/\sqrt{2}$. The communication network is a line graph with Red connected to Green, Green to Magenta, and Magenta to Blue, shown in Fig. 3 We take the message life parameter to be the diameter of the graph, i.e. $\gamma = 3$. Similar to the first scenario, at test time the positions of Red and Blue are switched. Since Red and Blue do not share an edge, they must communicate state, action and reward samples through *message passing*. Since each episode is $H = 10 > \gamma$, samples are fully propagated through the network one episode (i.e. $H = 10$ steps) later. If $\gamma = 3$, Red and Blue show improvement when their initial condition is swapped. Fig. 4 plots the average reward per rollout versus the training episode.

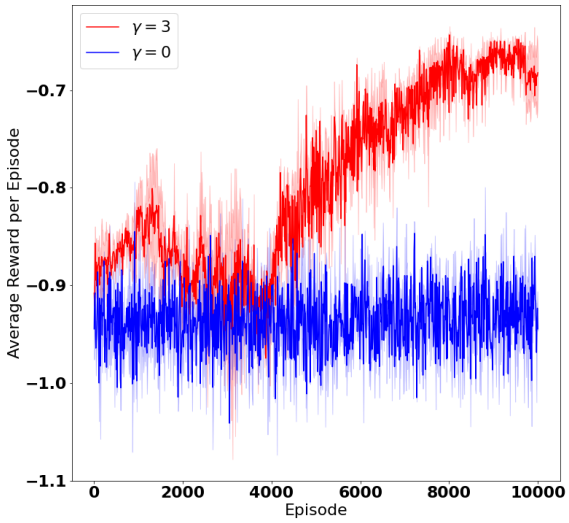


Figure 4: Average rollout reward achieved per episode for the 4-agent scenario for different $\gamma$. The reward is averaged over the Red and Blue agents. When the initial states of Red and Blue, which do not share an edge, are swapped, both are able to learn off-policy from message passing data.

## Conclusion

Optimal exploration in online reinforcement learning is a key consideration that impacts sample complexity. We investigate the benefits of fully decentralized exploration in MARL using regret as a metric. We provide a multi-agent extension of the UCB-Hoeffding algorithm provided in (Jin et al. 2018a) where the agents are equipped with a message passing scheme. We prove a regret bound that is $\tilde{\mathcal{O}}(\sqrt{MH^4SAT}$, an $\tilde{\mathcal{O}}(M^{-1/2})$ improvement over the single-agent setting Specifically, we consider general network $G$ and show that the regret also depends on the clique structure of the power graph $G_\gamma$, in addition to the number of agents. This key result suggests that the dense network structure, higher message life $\gamma$, and higher number of agents $M$ all reduce the average regret incurred by each agent, as shown in the simulations. While the assumption of time-varying dynamics demands samples that are polynomial in the episode length $H$, cooperative estimation of the optimal value functions allow parallel experience generation and *exploration by proxy*, or off-policy learning using communication. When the initial state is fixed, our regret bound corresponds to $\mathcal{O}(\varepsilon^{-2}M^{-1/2})$ sample complexity. Further work may involve providing a multi-agent minimax-optimal regret bound structure using a Bernstein-style or similar UCB bonus as shown in (Jin et al. 2018a; Zhang, Zhou, and Ji 2020). Further, the message life $\gamma$ should be optimized with respect to the number of agents and network structure by considering communication cost and privacy. We hope to extend our tabular results to deep $Q$-learning.

## References


Arslan, G.; and Yüksel, S. 2017. Decentralized Q-Learning for Stochastic Teams and Games. *IEEE Transactions on Automatic Control*, 62(4): 1545–1558.

Asghari, S. M.; Ouyang, Y.; and Nayyar, A. 2020. Regret Bounds for Decentralized Learning in Cooperative Multi-Agent Dynamical Systems. *arXiv:2001.10122 [cs, math, stat]*. ArXiv: 2001.10122.

Bernstein, D. S.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The Complexity of Decentralized Control of Markov Decision Processes. *Mathematics of Operations Research*, 27(4): 819–840.

Dann, C.; Lattimore, T.; and Brunskill, E. 2017. Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning. *Advances in Neural Information Processing Systems 31*, 11.

Dubey, A.; and Pentland, A. 2021. Provably Efficient Cooperative Multi-Agent Reinforcement Learning with Function Approximation. *arXiv:2103.04972 [cs, stat]*. ArXiv: 2103.04972.

Foerster, J. N.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual Multi-Agent Policy Gradients. *The Thirty-Second AAAI Conference on Artificial Intelligence*, 9.

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018a. Is Q-Learning Provably Efficient? *Advances in Neural Information Processing Systems 31*, 11.

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018b. Is Q-Learning Provably Efficient? In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jin, C.; Liu, Q.; and Miryoosefi, S. 2021. Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms. *arXiv:2102.00815 [cs, stat]*. ArXiv: 2102.00815.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2020. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *arXiv:1706.02275 [cs]*. ArXiv: 1706.02275.

Rashid, T.; Farquhar, G.; Peng, B.; and Whiteson, S. 2020. Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems 33*, 12.

Sayin, M. O.; Zhang, K.; Leslie, D. S.; Basar, T.; and Ozdaglar, A. 2021. Decentralized Q-Learning in Zero-sum Markov Games. *arXiv:2106.02748 [cs, math]*. ArXiv: 2106.02748.

Terry, J. K.; Black, B.; Grammel, N.; Jayakumar, M.; Hari, A.; Sullivan, R.; Santos, L.; Perez, R.; Horsch, C.; Dieffendahl, C.; Williams, N. L.; Lokesh, Y.; and Ravi, P. 2021. PettingZoo: Gym for Multi-Agent Reinforcement Learning. *arXiv:2009.14471 [cs, stat]*. ArXiv: 2009.14471.

Wang, T.; Dong, H.; Lesser, V.; and Zhang, C. 2020. ROMA: Multi-Agent Reinforcement Learning with Emergent Roles. *Proceedings of Machine Learning Research*, 119: 11.

Zhang, J.; Bedi, A. S.; Wang, M.; and Koppel, A. 2021a. MARL with General Utilities via Decentralized Shadow Reward Actor-Critic. *arXiv:2106.00543 [cs, math, stat]*. ArXiv: 2106.00543.

Zhang, K.; Kakade, S.; Basar, T.; and Yang, L. 2020. Model-Based Multi-Agent RL in Zero-Sum Markov Games with Near-Optimal Sample Complexity. In *Advances in Neural Information Processing Systems*, volume 33, 1166–1178.

Zhang, K.; Yang, Z.; and Basar, T. 2018. Networked Multi-Agent Reinforcement Learning in Continuous Spaces. In *2018 IEEE Conference on Decision and Control (CDC)*, 2771–2776. ISSN: 2576-2370.

Zhang, K.; Yang, Z.; Liu, H.; Zhang, T.; and Basar, T. 2018. Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents. *Proceedings of the 35th International Conference on Machine Learning*, 10.

Zhang, K.; Yang, Z.; Liu, H.; Zhang, T.; and Basar, T. 2021b. Finite-Sample Analysis For Decentralized Batch Multi-Agent Reinforcement Learning With Networked Agents. *IEEE Transactions on Automatic Control*, 1–1.

Zhang, Z.; Zhou, Y.; and Ji, X. 2020. Almost Optimal Model-Free Reinforcement Learning via Reference-Advantage Decomposition. *34th Conference on Neural Information Processing Systems*, 10.