Decentralized Model-Free Reinforcement Learning in Stochastic Games with Average-Reward Objective[*]

Romain Cravic
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG
38000 Grenoble, France
romain.cravic@inria.fr

Nicolas Gast
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG
38000 Grenoble, France
nicolas.gast@inria.fr

Bruno Gaujal
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG
38000 Grenoble, France
bruno.gaujal@inria.fr

ABSTRACT

We propose the first model-free algorithm that achieves low regret performance for decentralized learning in two-player zero-sum tabular stochastic games with infinite-horizon average-reward objective. In decentralized learning, the learning agent controls only one player and tries to achieve low regret performances against an arbitrary opponent. This contrasts with centralized learning where the agent tries to approximate the Nash equilibrium by controlling both players. In our infinite-horizon undiscounted setting, additional structure assumptions is needed to provide good behaviors of learning processes : here we assume for every strategy of the opponent, the agent has a way to go from any state to any other. This assumption is the analogous to the "communicating" assumption in the MDP setting. We show that our Decentralized Optimistic Nash Q-Learning (DONQ-learning) algorithm achieves both sublinear high probability regret of order $T^{3/4}$ and sublinear expected regret of order $T^{2/3}$. Moreover, our algorithm enjoys a low computational complexity and low memory space requirement compared to the previous works of [24] and [10] in the same setting.

KEYWORDS

Stochastic game; Q-learning; Regret; Two-player zero-sum game; Optimistic algorithm; Upper confidence bounds (UCB).

ACM Reference Format:
Romain Cravic, Nicolas Gast, and Bruno Gaujal. 2023. Decentralized Model-Free Reinforcement Learning in Stochastic Games with Average-Reward Objective. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 9 pages.

1 INTRODUCTION

Recently, reinforcement learning, combined with technological progress, has achieved superhuman performances in computing optimal behaviors in various decision processes. Particularly, there is a development of self-play algorithms to design strong artificial intelligence in games such as Go or Chess [20, 21]. These methods have the practical interest that they do not require supervision neither with huge data set nor with expert feedback to train efficiently. We focus here on Competitive Multi-Agent Reinforcement Learning (MARL) where several agents interact with an environment that may involve randomness, in order to maximize their own profit. One of the mostly used model to deal with MARL is stochastic game (a.k.a. Markov game in the literature) that was introduced in the 50s by Shapley [18]. Stochastic games (SGs) are more pertinent to handle MARL compared to Markov decision processes (MDPs) that are widely used for single-agent reinforcement learning.

Here we consider learning in infinite-horizon two-player zero-sum game with the average-reward criterion: There are two players that play indefinitely, and each player wins exactly what their opponent loses. At each time, the game has a current state and players choose and play an action simultaneously : they receive rewards and then the state of the environment changes (possibly randomly), this depending on the current state and the joint choice of actions. At the early stage of the game, the reward function and transition probabilities are unknown by the learning agent.

An important concept in stochastic games is the Nash equilibrium. In two-player zero-sum stochastic games, all Nash equilibria have the same value which represents a bound of the minimum expected average-reward that each player can obtain if they play optimally, regardless of the strategy of the opponent. When both agents play according to the Nash equilibrium, each of them can not hope a greater payoff by changing unilaterally his own strategy. Therefore, Nash equilibrium is often a benchmark for learning performances and a large set of works focus on this notion.

Two settings can be considered for learning in stochastic games: centralized and decentralized. In the first setting, a central learner tries to find the Nash equilibrium as quick as possible by controlling both players, *e.g.* [2, 3, 13, 19, 27]. In the second setting, the learner controls only one player and tries to be as efficient as possible against an arbitrary opponent, see [5, 10, 14–17, 22, 24, 26, 27] for different settings and objectives. This setting is quite natural but more challenging than the first one because the learner has no control on the opponent strategy that may change over time. In this paper, we consider the decentralized learning problem. Several definitions for the regret are possible in this case, that we will discuss in more detail in Section 2.3.

Learning algorithms are often split into two families called respectively model-based algorithms and model-free algorithms. In a model-based algorithm, the learner gathers observations of the game in order to estimate the parameters of the model (reward functions and transition probabilities), and then uses this estimation to compute a policy that is used for playing and gathering more information. In a model-free algorithm, the learner directly tries to estimate the values of the true model and plays optimally according to these value estimations. The two papers that are closest to ours are the algorithms UCSG in [24] and PSRL-ZSG in [10]. They both present a decentralized learning algorithm in average-reward SG

[*] An extended version of the paper with all the proofs is available at [6].

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

and are both model-based. To the best of our knowledge, we propose in this paper the first model-free algorithm for the stochastic games with infinite-horizon average-reward objective. Model-free algorithms have the advantage that they often need considerably less memory space to work, and additionally they often enjoy much better time complexity, which motivates our work. They can also be generalized to other settings, such as linear SGs. The drawbacks of model-free is that, in the literature, model-based algorithms often achieve better performances than model-free.

1.1 Our Contribution

We propose an algorithm, that we call decentralized optimistic Nash Q-learning (DONQ-learning). This algorithm is designed for infinite-horizon stochastic games with average-reward objective. Our paper uses Q-learning optimism techniques similar to the one introduced by Jin *et al.* in their seminal paper [11] for finite-horizon MDPs. This optimistic Q-learning has then been extended in two directions: (1) to infinite-horizon average-reward MDPs in [25] by using an artificial discounted setting, and (2) to finite horizon stochastic games in [3, 22]. In this paper we combine these two lines of work to obtain an efficient learning algorithm with sublinear regret in infinite-horizon average-reward SGs. We show that the expected regret of DONQ-learning is sublinear in time and is bounded by $O(D(SAB\log(2T^2))^{\frac{1}{3}}T^{\frac{2}{3}})$, where T is the time horizon, S is the number of states, A and B are the size of learner's and of the opponent's actions sets, and D a known upper bound of the span of the game. This result is similar to the bound obtained in [25] for the regret in infinite-horizon MDP. In addition, our algorithm enjoys a low computational complexity and low memory space requirement compared to the previous works of [24] and [10] in the same setting. We also provide a high-probability bound for the regret of our algorithm. Surprisingly, our analysis also shows a high probability upper bound for the regret of order $O(\sqrt{D\log(2/\delta)}T^{\frac{3}{4}})$, which has a dependency in T that is significantly higher than for the expected regret. This higher dependency in T is due to an additional term that appears because of our decentralized setting where the learner may play against a weak opponent. This additional term appears as a martingale difference sequence which disappears when the expected regret is considered.

Outline. The rest of the paper is organized as follows. We first introduce the model, the regret definition and some background on optimistic Q-learning in Section 2. We present our DONQ-learning algorithm in Section 3. In Section 4, we give the regret upper bounds (expected regret and regret with high probability), along with their proofs. We provide a proof of the most important technical lemmas in Section 5. The proofs of some other lemmas are delegated to the extended version of the article that is available at [6]. Finally, we conclude in Section 6.

1.2 Related Work

Q-learning in MDP : Q-learning algorithms appeared in 1989 in the seminal paper of Watkins [23]. The first provably-efficient Q-learning appears in [11], where the authors add an optimism term to Q-learning and prove that the algorithm achieves low regret performances in finite-horizon MDPs. The authors use a smart choice of learning rates that we also use in this paper. The same techniques are then adapted in [8] for the discounted setting to derive an algorithm with sample-complexity guaranties. The above two works are used in [25] to derive an optimistic Q-learning for average-reward MDPs that has a regret bound of order $T^{\frac{2}{3}}$. This algorithm only assumes that the MDP is weakly-communicating. Our work is inspired from [25] and extend their algorithm for the decentralized multi-agent setting.

Q-learning in SG : For stochastic games, a Q-learning with optimism principle first appears in [3] for finite-horizon problems. An adaptation to decentralized learning (still in finite-horizon SG) is given in [22]. For the infinite-horizon setting, the authors of [16] use Q-learning in discounted SG for decentralized learning but without optimism so they only establish convergence results. Our work differs from the above because we consider the infinite-horizon average-reward setting.

Other model-free approaches : Some papers propose other model-free algorithms, different from Q-learning, that also use optimism techniques. This includes the following:
V-learning : An unavoidable assumption to use Q-learning like algorithms is that the learner must observe the opponent's action at each time. The V-learning algorithm is proposed in [3] for centralized learning in finite-horizon SG, and then is adapted in [22] to design an efficient algorithm for decentralized learning when the agents do not observe the action of their opponent.
OOMD : For infinite-horizon average-reward MDP, the authors of [25] propose the OOMD algorithm that achieves a regret bound of order $\sqrt{T}$ when they assume the ergodicity assumption. They construct an adversarial multi-armed bandit instance for each state to design efficient policies.
Gradient descent : Gradient descent is another approach for decentralized learning in SG when the agent does not observe the action of its opponent. Some papers propose procedures based on the optimistic gradient descent to achieve Nash equilibrium identification guaranties after a finite time, for the finite-horizon setting in [7], and for the discounted setting in [26].

Model-based algorithms : To the best of our knowledge, all the algorithms that deal with learning in infinite-horizon average reward stochastic games are model-based. UCRL2 is introduced in [1] and uses optimism to achieve sublinear regret bounds in infinite-horizon average-reward MDP with finite diameter. In [24], UCRL2 is adapted into UCSG for SG setting, and two possible definitions for the diameter are proposed. These definitions extend the notions of "communicating" MDP and "ergodic" MDP for SG. Under the weakest assumption, UCSG enjoys a regret bound of order $T^{\frac{2}{3}}$. In this paper we assume the same diameter assumption. In a recent work, inspired from PSRL algorithm for MDP, PSRL-ZSG algorithm is proposed in [10] and achieves a Bayesian regret of order $\sqrt{T}$ under the same structure assumption. Our work does not improve the above regret bounds but the model-free property of our algorithm induces a lower computational complexity and memory space requirement than the previous model-based algorithms. Plus, we do not study Bayesian regret but worst-case expected regret or worst-case high probability regret.

2 PRELIMINARIES

2.1 Stochastic Game Model and Notations

We consider two-player zero-sum simultaneous stochastic games (also known as Markov games in the literature), which generalize the standard Markov decision processes (MDPs) into the two-player setting where both players try to maximize their own payoff. *Zero-sum* means that in this game, what a player wins corresponds exactly to what the other loses. We write the payoff for the point of the first player, that we call the *max-player.* This player tries to maximize the payoff whereas the second player, that we call the *min-player*, tries to minimize it. In this paper, the learner is the max-player and the opponent is the min-player. The term *simultaneous* means that the two players choose simultaneously how to interact with the environment at each time. Here we consider the infinite-horizon setting where players play indefinitely. At each time step, the game is in a current state known by the players and they play simultaneously an action. The players obtain a reward and the state of the game changes, both depending on the current state and the choices of actions.

Formally, an instance of an infinite-horizon stochastic game is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{B}, r, P)$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is the action set of the max-player, and $\mathcal{B}$ it the action set of the min-player. We assume that all the states and actions spaces are finite and of sizes S, A and B. For every $s, a, b \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, $r(s, a, b) \in [0, 1]$ is the reward that the max-player wins (and that the min-player loses) when the pair of actions (a, b) is played in state s. Then the next state of the game is randomly drawn with the probability distribution $P(.|s, a, b)$. At each time step t the players observe the states of the game s_t, they both choose simultaneously an action a_t and b_t that they play. They then observe the action of their opponent, receive the reward $r(s_t, a_t, b_t)$. The next state s_{t+1} is then drawn according to $P(.|s_h, a_h, b_h)$.

Markovian policies : A max-player Markovian policy $\mu : \mathcal{S} \longrightarrow \Delta_{\mathcal{A}}$ is a function that maps every state s with a probability distribution $\mu(.|s)$ on max-player's actions. Similarly, a min-player Markovian policy $\nu : \mathcal{S} \longrightarrow \Delta_{\mathcal{B}}$ maps every state s with a probability distribution $\nu(.|s)$ on min-player's actions. The pair of policies (μ, ν) determines the rewards and the dynamics of the game.

Notations : For a function $f : \mathcal{S} \longrightarrow \mathbb{R}$, we write $Pf(s, a, b)$ for $\sum_{s'} P(s'|s, a, b) f(s')$. We also write $sp(f)$ for $\max_s f(s) - \min_s f(s)$. For a function $g : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \longrightarrow \mathbb{R}$ and a pair of Markovian policies (μ, ν), we write $\mathbb{D}_{(\mu,\nu)}[g](s) = \mathbb{E}_{a\sim\mu(.|s),b\sim\nu(.|s)}[g(s, a, b)]$. To lighten notation, we sometimes write $g^{(\mu,\nu)}(s)$ for this quantity. In particular, for the reward and transition probability, we have:

$$r^{(\mu,\nu)}(s) = \mathbb{E}_{a\sim\mu(.|s),b\sim\nu(.|s)}[r(s, a, b)]$$
$$P^{(\mu,\nu)}(s'|s) = \mathbb{E}_{a\sim\mu(.|s),b\sim\nu(.|s)}[P(s'|s, a, b)].$$

2.2 Average-Reward and Game Structure

In this paper we establish low regret bounds for the undiscounted setting where players play indefinitely and try to maximize their average-reward over time. The average-reward of a pair of policies (μ, ν) is defined for every state s, whenever the limit exists, by

$$J^{(\mu,\nu)}(s) = \lim_{T\rightarrow+\infty} \frac{1}{T} \mathbb{E}_{(\mu,\nu)} \left[\sum_{t=1}^{T} r^{(\mu,\nu)}(s_t) \mid s_1 = s \right],$$

where $\mathbb{E}_{(\mu,\nu)}$ hides that the $(s_t)_t$ is a Markov chain whose transition kernel is $P^{(\mu,\nu)}$. This represents the expected average-reward for the max-player starting from state s if both players play policies μ and ν respectively.

Given a min-player policy ν, a best response against ν for the undiscounted setting is a max-player policy μ^* such that for all state s, $J^{(\mu^*,\nu)}(s) = \max_\mu J^{(\mu,\nu)}(s)$. Similarly, given a max-player policy μ, a best response against μ is a min-player policy ν^* such that for all state s, $J^{(\mu,\nu^*)}(s) = \min_\nu J^{(\mu,\nu)}(s)$.

To deal with the undiscounted setting, we assume the following structure property, which is known in the particular case of MDP, to be a minimal relevant framework for learning [4] :

Asumption 1. *For every pair (s, s') of states, and any min-player policy ν, there exists a max-player policy μ such that s' is accessible from s with positive probability under policies (μ, ν).*

This assumption is intuitively necessary to get low regret because if this assumption does not hold, it means that the opponent may have a way to lock the learning agent in "bad stage" while at the early stage of the learning process, the agent could not have enough information to avoid this. Note that for the MDP setting, the analog of this assumption would be to assume that the MDP is "communicating": here the SG is communicating for the max-player's point of view, *i.e.*, from the learner's point of view. In fact we may additionally assume the existence of a set of states that are transient under all Markovian policies, which would be the analogous property of "weakly communicating" for the MDP setting. This would not change our result but would complexify notations and proofs.

As a consequence of Assumption 1, if $T^{(\mu,\nu)}_{s\rightarrow s'}$ is the expected time to go from s to s' under policies (μ, ν), there exists a real number D such that

$$\max_{s,s'} \max_\nu \min_\mu T^{(\mu,\nu)}_{s\rightarrow s'} \le D$$

The minimal D is the diameter of the game [24]. In this paper, we denote by D an upper bound of the diameter and we suppose that this D is known by the learning agent, so D can be used to set some values of parameters.

We summarize important results that hold under Assumption 1 in the Theorem 1 below. This theorem justifies the relevance of the learning objective in this framework. For a proof of these results, we refer to Theorem E.1 of [24].

Theorem 1 (Theorem E.1 of [24]). *Under assumption 1, there exist a unique real number J^*, a function $h^* : \mathcal{S} \longrightarrow \mathbb{R}$, and a pair of Markovian policies (μ^*, ν^*) such that*

(1) $J^* = \max_\mu \min_\nu J^{(\mu,\nu)}(s) = \min_\nu \max_\mu J^{(\mu,\nu)}(s)$ *for every s : the game value exists and does not depend on s.*
(2) (μ^*, ν^*) *is a pair of Markovian policies such that $J^{(\mu^*,\nu^*)}(s) = J^*$ for all s : μ^* is a best response against ν^* and vice-versa (Nash equilibrium).*

(3) *$sp(h^*) \leq D$ and satisfies the following Bellman's optimality equations :*

$$\begin{aligned} J^* + h^*(s) &= \max_{\mu} \left(r^{(\mu,\nu^*)}(s) + P^{(\mu,\nu^*)} h^*(s) \right), \\ J^* + h^*(s) &= \min_{\nu} \left(r^{(\mu^*,\nu)}(s) + P^{(\mu^*,\nu)} h^*(s) \right). \end{aligned}$$

2.3 Learning Objective

In this paper, we consider a decentralized learning algorithm that has a low regret in a two-player zero-sum stochastic game. Nash equilibria in this setting provide to players an optimal way to play whatever their opponent does: If the max-player plays their Nash policy, it guarantees them an average-reward of at least J^* against any opponent and we want to measure the online performance of the learning agent compared to this Nash policy. Thus our regret is given by

$$\text{Reg}(T) = \sum_{t=1}^{T} \left(J^* - r(s_t, a_t, b_t) \right) \tag{1}$$

This definition is classical for the regret in decentralized learning [10, 22, 24], and generalizes the definition for the MDP setting [1, 25]. Note that in contrast with MDPs, the regret is not necessarily non-negative: If the opponent is weak, then the learning agent can achieve an average-reward greater than J^*. The definition in Equation (1) ensures that, whatever the opponent's strategy, the learner should learn a defensive strategy that works against it - but does not need to exploit the opponent's weakness - to achieve sublinear performances.

A stronger regret definition would be to consider that the learner competes with the best response against the opponent (potentially assuming this one uses only Markovian policies). This would lead to the following regret definition:

$$\sum_{t=1}^{T} \left(\max_{\mu} J^{(\mu,\nu_t)} - r(s_t, a_t, b_t) \right),$$

or a weaker version where the learner competes with the best fixed (over time) max-player policy, that is

$$\max_{\mu} \sum_{t=1}^{T} \left(J^{(\mu,\nu_t)} - r(s_t, a_t, b_t) \right).$$

Unfortunately, it is shown in [14] that for the finite-horizon setting, it is not possible to design an algorithm that has a sublinear regret with the two above learning objectives. Hence, and because the infinite-horizon setting is more general than the finite-horizon one, in this paper we focus on the regret definition given by (1).

2.4 Background on Discounted SG

As we will see later, our DONQ-learning algorithm, uses an artificial discounted setting to run. Hence, we recall here some facts and notations about the discounted setting.

In the discounted setting, what players get at time t is amortized by a factor γ^{t-1} where $\gamma < 1$ is called the discount factor. Given a discount factor γ and a pair of policies (μ, ν), the value function for the discounted setting is defined for every state s as

$$V_{\gamma}^{(\mu,\nu)}(s) = \mathbb{E}_{(\mu,\nu)} \left[\sum_{t=1}^{+\infty} \gamma^{t-1} r^{(\mu,\nu)}(s_t) \mid s_1 = s \right],$$

where $\mathbb{E}_{(\mu,\nu)}$ hides that the $(s_t)_t$ is a Markov chain whose transition kernel is $P^{(\mu,\nu)}$. This represents the expected discounted cumulative reward for the max-player starting from state s if both players play policies μ and ν respectively. Then, for every pair of actions (a, b), the Q-value function in (s, a, b) is defined as

$$Q_{\gamma}^{(\mu,\nu)}(s, a, b) = r(s, a, b) + \gamma P V_{\gamma}^{(\mu,\nu)}(s, a, b).$$

This represents the expected discounted cumulative reward for the max-player starting from state s where players first play a and b respectively, and then follow policies μ and ν respectively. Therefore, for every state s we have the relation

$$V_{\gamma}^{(\mu,\nu)}(s) = \mathbb{D}_{(\mu,\nu)}[Q_{\gamma}^{(\mu,\nu)}](s).$$

Given a min-player policy ν, a best response against ν for the γ-discounted setting is a max-player policy μ^* such that for all state s, $V_{\gamma}^{(\mu^*,\nu)}(s) = \max_{\mu} V_{\gamma}^{(\mu,\nu)}(s)$. Similarly, given a max-player policy μ, a best response against μ is a min-player policy ν^* such that for all state s, $V_{\gamma}^{(\mu,\nu^*)}(s) = \min_{\nu} V_{\gamma}^{(\mu,\nu)}(s)$.

It is known that there always exists a pair of policies $(\mu^{\gamma}, \nu^{\gamma})$ such that μ^{γ} is a best response against ν^{γ} and vice-verse for the γ-discounted setting. $(\mu^{\gamma}, \nu^{\gamma})$ is called a Nash equilibrium of the game. All Nash equilibria have the same value function, denoted by V_{γ}^*, which satisfies for every state s

$$V_{\gamma}^*(s) = \max_{\mu} \min_{\nu} V_{\gamma}^{(\mu,\nu)}(s) = \min_{\nu} \max_{\mu} V_{\gamma}^{(\mu,\nu)}(s).$$

We also denote by Q_{γ}^* the unique Q-value function of the Nash equilibria. V_{γ}^* and Q_{γ}^* satisfy the following Bellman's optimality equations for all (s, a, b) :

$$\begin{aligned} Q_{\gamma}^*(s, a, b) &= r(s, a, b) + \gamma P V_{\gamma}^*(s, a, b), \\ V_{\gamma}^*(s) &= \max_{\mu} \min_{\nu} \mathbb{D}_{(\mu,\nu)}[Q_{\gamma}^*](s). \end{aligned}$$

We recall (see e.g. [24] Lemma E.2) that under Assumption 1, if D is a bound on the diameter of the game as defined above, then for any $\gamma < 1$ we have

$$sp(V_{\gamma}^*) \leq D. \tag{2}$$

2.5 Learning Rate Notations and Properties

In our DONQ-learning algorithm, we use optimistic Q-learning updates to maintain estimators. At a given time t, such an algorithm observes the current state-action triplet (s, a, b) and updates the Q-value of this triplet as:

$$Q_{t+1}(s, a, b) \leftarrow (1 - \alpha_{\tau}) Q_t(s, a, b) + \alpha_{\tau}([\text{new sample}] + [\text{bonus}]),$$

where τ is the number of visits of the triplet (s, a, b) up to time t.

The quantity α_{τ} is called the learning rate and has to be well-chosen to ensure good learning performances. In the optimistic Q-learning algorithm for H-finite-horizon MDPs, the authors of [11] propose the learning rate $\alpha_{\tau} = \frac{H+1}{H+\tau}$ and they show that such a learning rate has good properties summarized in Lemma 1 below. For every $0 \leq \tau \leq T$, they define $\alpha_{\tau}^0 = \prod_{j=1}^{\tau}(1 - \alpha_j)$ and for all

$1 \leq i \leq \tau$, $\alpha_\tau^i = \alpha_i \prod_{j=i+1}^{\tau}(1-\alpha_j)$. Note that if $\tau \geq 1$, $\alpha_\tau^0 = 0$ and that $\sum_{i==0}^{\tau} \alpha_\tau^i = 1$. The sequence α_τ^i has the following properties.

Lemma 1 (Lemma 4.1 of [11]). *The sequence α_τ^i satisfy :*

(1) $\frac{1}{\sqrt{\tau}} \leq \sum_{i=1}^{\tau} \frac{\alpha_\tau^i}{\sqrt{i}} \leq \frac{2}{\sqrt{\tau}}$ *for every* $\tau \geq 1$.
(2) $\sum_{t=1}^{\tau} (\alpha_\tau^i)^2 \leq \frac{2H}{\tau}$ *for every* $\tau \geq 1$.
(3) $\sum_{\tau=i}^{\infty} \alpha_\tau^i \leq 1 + \frac{1}{H}$ *for every* $i \geq 1$

In this paper we use the same choice of learning rate for some carefully chosen parameter H.

3 THE DONQ-LEARNING ALGORITHM

The main idea of our decentralized optimistic Nash Q-learning (that we present in Algorithm 1) is to learn the game in a well-chosen discounted setting using optimistic Q-Learning techniques first introduced in [11]. The convenience of the γ-discounted setting is that the Q-values are well-defined and bounded by $1/(1-\gamma)$ which enables to build and maintain optimistic estimators $\overline{Q}_t$ for the Nash Q-value Q_γ^*.

Algorithm 1 DONQ-learning Algorithm

```
Parameters : $H > 1$, $\delta \in (0,1)$
Initialization : $\forall s, a, b$
$\overline{Q}_1(s,a,b) = \overline{Q}^{\downarrow}_1(s,a,b) = \overline{V}^{\downarrow}_1(s) = H$
$N_1(s,a,b) = 0$ ; $\mu_1(a|s) = 1/A$ ; $\gamma = 1 - 1/H$
$\forall \tau \geq 1, \quad \alpha_\tau = \frac{H+1}{H+\tau} \quad \beta_\tau = 2D\sqrt{\frac{H\log(2T/\delta)}{\tau}}$
Begin :
Observe $s_1$.
for $t = 1, 2, \ldots, T$ do
    Draw and play $a_t \sim \mu_t(.|s_t)$.
    Observe $b_t$ (drawn from the unknown opponent's policy).
    Observe $r(s_t, a_t, b_t)$ and $s_{t+1}$.
    Increment $N_{t+1}(s_t, a_t, b_t) \leftarrow N_t(s_t, a_t, b_t) + 1$
    Set $\tau := N_{t+1}(s_t, a_t, b_t)$ and update $Q$-estimators as:
        $\overline{Q}_{t+1}(s_t, a_t, b_t) \leftarrow (1-\alpha_\tau)\overline{Q}_t(s_t, a_t, b_t) +$
                $\alpha_\tau \left[ r(s_t, a_t, b_t) + \gamma \overline{V}^{\downarrow}_t(s_{t+1}) + \beta_\tau \right]$
        $\overline{Q}^{\downarrow}_{t+1}(s_t, a_t, b_t) \leftarrow \min\left( \overline{Q}^{\downarrow}_t(s_t, a_t, b_t), \overline{Q}_{t+1}(s_t, a_t, b_t) \right)$
    Update the policy
        $(\mu_{t+1}(.|s_t), \tilde{\nu}_{t+1}(.|s_t)) \leftarrow \text{GetNashPolicies}\left( \overline{Q}^{\downarrow}_{t+1}(s_t, ., .) \right)$
    Update $V$-estimator
        $\overline{V}^{\downarrow}_{t+1}(s_t) \leftarrow \mathbb{D}_{(\mu_{t+1}, \tilde{\nu}_{t+1})}[\overline{Q}^{\downarrow}_{t+1}](s_t)$
    $N_{t+1}, \overline{Q}_{t+1}, \overline{Q}^{\downarrow}_{t+1}, \overline{V}^{\downarrow}_{t+1}, \mu_{t+1}$ are kept equal to their values
      at the previous time step, $N_t, \overline{Q}_t, \overline{Q}^{\downarrow}_t, \overline{V}^{\downarrow}_t, \mu_t$, for all other $s, a, b$.
end for
```

The DONQ-learning algorithm takes as input a parameter H and defines a discount factor $\gamma = 1 - \frac{1}{H}$. The agent then operates as if it was in a discounted objective. The algorithm also takes as parameter a confidence level $\delta \in (0,1)$ which is set either to obtain a high probability regret bound, or to obtain a sublinear expected regret.

All optimistic estimators are initialized with H. At each time t, the learner plays an action a_t drawn from the current policy μ_t and observes the opponent action b_t, the reward $r(s_t, a_t, b_t)$ and the next state s_{t+1}. Then the optimistic estimator $\overline{Q}_{t+1}$ is updated for the current state-actions tuple (s_t, a_t, b_t), by using an optimistic Q-learning like operation with a learning rate $\alpha_\tau = \frac{H+1}{H+\tau}$, the current optimistic estimator of the next state value $\overline{V}^{\downarrow}_t(s_{t+1})$ and a bonus term β_τ that scales in $D\sqrt{\frac{H}{\tau}}$, where τ is the number of visits of state-actions tuple (s_t, a_t, b_t). For technical reasons, the algorithm maintains a non-increasing version of $\overline{Q}_t$, denoted by $\overline{Q}^{\downarrow}_t$, which is the one used to update the learner's policy and to compute the optimistic value estimators $\overline{V}^{\downarrow}_t$.

The learner's policy is updated for the current state $s_t : \mu_{t+1}(.|s_t)$ is chosen to be the max-player Nash policy of the matrix game $\overline{Q}^{\downarrow}_{t+1}(s_t, ., .)$. Note that if μ^* is an optimal Markovian Nash policy of the max-player, then for all s, $\mu^*(.|s)$ is a max-player Nash policy of the matrix game $Q^*(s, ., .)$. Then the algorithm also computes the min-player Nash policy $\tilde{\nu}_{t+1}(.|s_t)$ of the matrix game $\overline{Q}^{\downarrow}_{t+1}(s_t, ., .)$ in order to update $\overline{V}^{\downarrow}_{t+1}(s_t)$ as the expected value of $\overline{Q}^{\downarrow}_{t+1}(s_t, ., .)$ under policies μ_{t+1} and $\tilde{\nu}_{t+1}$. We emphasize that $\tilde{\nu}_{t+1}$ is not the policy used by the opponent at time $t+1$ (the opponent's policy is unknown to us and can be arbitrary).

4 THEORETICAL GUARANTEES

4.1 Main Results: Regret Guarantees

The key challenge is to correctly set the parameter H to obtain sublinear regret guarantees. On the one hand, a large value for H implies that the undiscounted objective is well approximated by the discounted case, but at the same time, the regret coming from the discounted analysis will be large. On the other hand, a small value of H means that the undiscounted objective is not well approximated by the discounted case, which may have bad consequences the regret.

We give here two regret bounds, with high probability and in expectation, both with a sublinear dependency in T obtained with a smart choice of the parameter H. First, Theorem 2 provides the high probability regret bound, depending on the parameter H.

Theorem 2. *For any $\delta \in (0,1)$, there exists an absolute*[1] *constant C such that with probability at least $1 - 3\delta$, the regret is bounded by*

$$\begin{aligned} Reg(T) \leq & D\frac{T}{H} + 2H\sqrt{2T\log(2/\delta)} \\ & + 12D\sqrt{SABHT\log(2T/\delta)} \\ & + C\left(SABH + D\sqrt{T\log(4/\delta)} \right). \end{aligned}$$

If the parameter H scales in $T^{\frac{1}{4}}$, this theorem shows a high probability regret bound of order $T^{\frac{3}{4}}$. Indeed, such a value of H balances the terms $\frac{T}{H}$ and $H\sqrt{T}$ that appear in our analysis.

[1] By absolute, we mean a constant that does not depend on any of the models parameters.

Corollary 1. *For any $\delta \in (0,1)$, we have that with probability at least $1-3\delta$, by setting the parameter $H = \sqrt{\frac{D}{\log(2/\delta)^{\frac{1}{2}}}}T^{\frac{1}{4}}$, for T sufficiently large the regret is upper bounded by*

$$Reg(T) \leq \sqrt{D\log(2/\delta)^{\frac{1}{2}}}T^{\frac{3}{4}} + o\left(T^{\frac{3}{4}}\right).$$

The second term of the regret bound of Theorem 2 has a $H\sqrt{T}$ factor. This term does not appear in the analysis of [25] for the MDP case. It is due to the fact that we were not able to bound $(\overline{Q}_t^{\downarrow} - Q_\gamma^*)(s_t, a_t, b_t)$ with a better bound than H (see Equation (**s4**) in the regret decomposition below). This difficulty appears because the learner plays against an opponent that may be possibly weak. More explanations are given in Appendix A.7 of the extended version of the paper [6]. This term appears because we simply use Azuma-Hoeffding's inequality for our high probability regret bound. However, since this term appears as terms of a martingale difference sequence, their sum can be bounded more efficiently when we consider the expected regret, as we see below.

Theorem 3 provides upper bounds for the expected regret, also depending on the parameter H. This bound is similar to the regret bound of Theorem 2 with $\delta = 1/T$ but is significantly smaller because the second term of the regret bound of Theorem 2 disappears.

Theorem 3. *There exists an absolute constant C such that*

$$\begin{aligned}\mathbb{E}[Reg(T)] \leq & D\frac{T}{H} + 12D\sqrt{SABHT\log(2T^2)} \\ & + C\left(SABH + D\sqrt{T}\log(4T)\right).\end{aligned}$$

By setting a parameter H that scales in $T^{\frac{1}{3}}$, we can balance the terms $\frac{T}{H}$ and $\sqrt{HT}$ (due to regret bound for discounted setting) and obtain a regret bound in expectation of order $T^{\frac{2}{3}}$. The bound that we obtain for the expected regret is then significantly smaller than the high probability bound of Corollary 1 because Theorem 3 does not has the term in $H\sqrt{T}$ as Theorem 2.

Corollary 2. *For T sufficiently large, by setting the parameter $H = \left(\frac{T}{SAB\log(2T^2)}\right)^{1/3}$, the expectation of the regret is bounded by*

$$\mathbb{E}[Reg(T)] \leq D(SAB\log(2T^2))^{\frac{1}{3}}T^{\frac{2}{3}} + o\left(T^{\frac{2}{3}}\right).$$

We highlight the fact that in the algorithm and our analysis, we only use that the diameter D is an upper bound of $sp(h^*)$ (Theorem 1) and $sp(V^*)$: any other known quantity that satisfies these two properties can replace D in the parameter setting and the theoretical guarantees.

4.2 Computational Complexity

An interesting property of Q-learning like algorithms is that only the values of the current state and state-action tuple are updated at each step. This contrasts with model-based approaches [10, 24] where some policies have to be entirely computed in the inner loop of the algorithm, or with value-iteration like algorithms [2, 27] where a sweep of the state space is needed at each step. The only non-constant time operation in DONQ-learning is the call to GetNashPolicies procedure, but it is known that we can compute the value and the Nash policy of a zero-sum matrix game of size $A \times B$ in a polynomial time via linear programming [9, 12], and this does depend on the size of the state space.

4.3 Proofs of Theorems 2 and 3

To show Theorem 2 and 3, we decompose the regret as follows:

$$\begin{aligned}\mathrm{Reg}(T) &= \sum_{t=1}^{T}(J^* - r(s_t, a_t, b_t)) \\ &= \sum_{t=1}^{T}(J^* - (1-\gamma)V_\gamma^*(s_t)) && \textbf{(s1)} \\ &\quad + \sum_{t=1}^{T}(V_\gamma^*(s_t) - \mathbb{D}_{(\mu_t,\nu_t)}[Q_\gamma^*](s_t) - \zeta_t) && \textbf{(s2)} \\ &\quad + \sum_{t=1}^{T}(\mathbb{D}_{(\mu_t,\nu_t)}[Q_\gamma^*](s_t) - r(s_t, a_t, b_t) - \gamma V_\gamma^*(s_t)) && \textbf{(s3)} \\ &\quad + \sum_{t=1}^{T}\zeta_t && \textbf{(s4)}\end{aligned}$$

where $\zeta_t = \mathbb{D}_{(\mu_t,\nu_t)}[\overline{Q}_t^{\downarrow} - Q_\gamma^*](s_t) - (\overline{Q}_t^{\downarrow} - Q_\gamma^*)(s_t, a_t, b_t)$. Note that $\{\zeta_t\}_t$ is a martingale difference sequence, bounded by $2H$ in absolute value, with respect to the σ-algebra generated by all random variables up to s_t since a_t and b_t are drawn with policies $\mu_t(.|s_t)$ and $\nu_t(.|s_t)$ respectively. We bound **(s1)**, **(s2)** and **(s3)** separately, sometimes under high probability events, and we finally deal with **(s4)** specifically for both types of regret bounds.

4.3.1 Analysis of the Term ***(s1)****.* The term (**s1**) represents the gap we pay because we learn in an artificial discounted setting. Each term of (**s1**) tends to 0 when γ is close to 1, that is when H is big. Concretely we show the following Lemma 2 :

Lemma 2. *For every $s \in \mathcal{S}$,*

$$\left|J^* - (1-\gamma)V_\gamma^*(s)\right| \leq (1-\gamma)D = \frac{D}{H}.$$

When summing over t in **(s1)**, a factor T appears but the factor $1/H$ enables to get a sublinear dependency in T in the regret with a well-chosen parameter H.

4.3.2 Analysis of the Term ***(s2)****.* Then, we first bound **(s2)** under an event $\mathcal{E}_\delta$ defined below for any $\delta \in (0,1)$ and that ensures in particular that $\overline{Q}_t^{\downarrow}$ and $\overline{V}_t^{\downarrow}$ are really upper bounds of Q_γ^* and V_γ^* respectively.

$\mathcal{E}_\delta :=$ For all (s,a,b) and time t, the following inequalities hold

$$(\overline{Q}_t^{\downarrow} - Q_\gamma^*)(s,a,b) \geq 0, \qquad (\overline{V}_t^{\downarrow} - V_\gamma^*)(s) \geq 0, \tag{3}$$

$$\begin{aligned}(\overline{Q}_{t+1}^{\downarrow} - Q_\gamma^*)(s,a,b) \leq & \alpha_\tau^0 H + 6D\sqrt{\frac{H}{\tau}\log(2T/\delta)} \\ & + \sum_{i=1}^{\tau}\gamma\alpha_\tau^i\left[(\overline{V}_{t_i}^{\downarrow} - V_\gamma^*)(s_{t_i+1})\right]\end{aligned} \tag{4}$$

where $\tau = N_{t+1}(s,a,b)$ and $t_1, t_2, \ldots, t_\tau$ the times where (s,a,b) was visited.

Lemma 3. *For any $\delta \in (0,1)$, $\mathcal{E}_\delta$ is true with probability at least $1-\delta$.*

A proof of Lemma 3 is given in Appendix A.1 of the extended version [6]. $\mathcal{E}_\delta$ holds with probability at least $1-\delta$ and under $\mathcal{E}_\delta$, Lemma 4 gives an upper bound for (**s2**) that scales in $\sqrt{HT}$. This is the main contribution of the regret due to performances of optimistic Q-learning in the discounted setting.

Lemma 4. *For any $\delta \in (0,1)$, under $\mathcal{E}_\delta$, there exists an absolute constant c such that*

$$\sum_{t=1}^{T}(V_\gamma^*(s_t) - \mathbb{D}_{(\mu_t,\nu_t)}[Q_\gamma^*](s_t) - \zeta_t) \le 12D\sqrt{SABHT\log(2T/\delta)} + cSABH.$$

*4.3.3 Analysis of the Term **(s3)**.* To bound (**s3**) we define an other event $\mathcal{E}'_\delta$ for any $\delta \in (0,1)$ that ensures that Azuma-Hoeffding's inequalities used on functions V_γ^* and Q_γ^* hold.

$\mathcal{E}'_\delta$: The two following inequalities hold

$$\left|\sum_{t=1}^{T}(\mathbb{D}_{(\mu_t,\nu_t)}[Q_\gamma^*](s_t) - Q_\gamma^*(s_t,a_t,b_t))\right| \le (1+\gamma D)\sqrt{2T\log(4/\delta)} \quad (5)$$

$$\left|\sum_{t=1}^{T}(PV_\gamma^*(s_t,a_t,b_t) - V_\gamma^*(s_{t+1}))\right| \le D\sqrt{2T\log(4/\delta)} \quad (6)$$

Lemma 5. *$\mathcal{E}'_\delta$ holds with probability at least $1-\delta$*

A proof of Lemma 5 is given in Appendix A.2 of the extended version [6]. In fact, the terms in (**s3**) are almost terms of Bellman's equations so, since $\mathcal{E}'_\delta$ holds with probability at least $1-\delta$, we bound (**s3**) with high probability as shown in Lemma 6 that is proved in Appendix A.3 of the extended version.

Lemma 6. *For any $\delta \in (0,1)$, under $\mathcal{E}'_\delta$, there exists an absolute constant c' such that*

$$\sum_{t=1}^{T}(\mathbb{D}_{(\mu_t,\nu_t)}[Q_\gamma^*](s_t) - r(s_t,a_t,b_t) - \gamma V_\gamma^*(s_t)) \le c'D\sqrt{T\log(4/\delta)}.$$

*4.3.4 Analysis of the Term **(s4)** and Conclusion.* We prove the high probability regret bound by compiling the previous high probability bounds on **(s1)**, **(s2)**, and **(s3)** (Lemmas 2, 4 and 6). Additionally we use Azuma-Hoeffding to bound $\sum_{t=1}^{T}\zeta_t$ with high probability. This makes appear a term that scales in $H\sqrt{T}$. We balance this term and the term $\frac{T}{H}$ of **(s1)** with a well-chosen parameter H of order $T^{\frac{1}{4}}$ to obtain an upper bound of order $T^{\frac{3}{4}}$ for the regret. More details are given in Appendix A.5 of the extended version [6].

We also establish the regret bound in expectation using the results of previous sections (Lemmas 2, 4 and 6). To deal with the remaining $\sum_{t=1}^{T}\zeta_t$, we use the fact that $\{\zeta_t\}$ is a martingale difference sequence (bounded sequence with a conditional expectation of 0 according to its filtration) to show that its weight is negligible especially in the case where the previous high probability events hold. We then set H smartly to balance the dominant term $\frac{T}{H}$ and $\sqrt{HT}$ with respect to T, and obtain the desired regret bound in expectation of order $T^{\frac{2}{3}}$. More details are given in the Appendix A.6 of the extended version [6].

5 PROOFS OF MAIN LEMMAS

Here we give the proofs of Lemma 2 and 4. All the other omited proofs are given in Appendix of the extended version [6]. Lemma 2 (resp. Lemma 4) induces the term that scales in $\frac{T}{H}$ (resp. $\sqrt{HT}$) in the statements of Theorems 2 and 3.

5.1 Proof of Lemma 2

We denote by (μ^*,ν^*) and h^* the Nash equilibrium for the undiscounted setting and the bounded function, both given by Theorem 1. We also denote by (μ^γ,ν^γ) a Nash equilibrium for the discounted setting with discount γ.

By sub-optimality of μ^* against ν^γ for the discounted setting (first inequality), and the sub-optimality of ν^γ against μ^* for the undiscounted setting (second inequality), we get

$$\begin{aligned}
V_\gamma^*(s) &\ge V_\gamma^{(\mu^*,\nu^\gamma)}(s)\\
&= \mathbb{E}_{\mu^*,\nu^\gamma}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}r^{(\mu^*,\nu^\gamma)}(\mathscr{s}_t) \quad |\mathscr{s}_1 = s\right]\\
&\ge \mathbb{E}_{\mu^*,\nu^\gamma}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}\left(J^* + h^*(\mathscr{s}_t) - P^{(\mu^*,\nu^\gamma)}h^*(\mathscr{s}_t)\right) \quad |\mathscr{s}_1 = s\right].
\end{aligned}$$

Since $\mathscr{s}_{t+1}$ is generated with transition probability $P^{(\mu^*,\nu^\gamma)}(.|\mathscr{s}_t)$ in the stochastic process under $\mathbb{E}_{\mu^*,\nu^\gamma}$, the last term in the expectation can be seen as $h^*(\mathscr{s}_{t+1})$. Therefore we get

$$\begin{aligned}
V_\gamma^*(s) &\ge \frac{J^*}{1-\gamma} + \mathbb{E}_{\mu^*,\nu^\gamma}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}(h^*(\mathscr{s}_t) - h^*(\mathscr{s}_{t+1})) \quad |\mathscr{s}_1 = s\right]\\
&= \frac{J^*}{1-\gamma} + h^*(\mathscr{s}_1) - \mathbb{E}_{\mu^*,\nu^\gamma}\left[\sum_{t=2}^{+\infty}(\gamma^{t-2} - \gamma^{t-1})h^*(\mathscr{s}_t) \quad |\mathscr{s}_1 = s\right]\\
&\ge \frac{J^*}{1-\gamma} + \min_s h^*(s) - \max_s h^*(s)\sum_{t=2}^{+\infty}(\gamma^{t-2} - \gamma^{t-1})\\
&\ge \frac{J^*}{1-\gamma} - sp(h^*)\\
&\ge \frac{J^*}{1-\gamma} - D.
\end{aligned}$$

By the sub-optimality of ν^* against μ^γ in the discounted setting (first inequality), and the sub-optimality of μ^γ against ν^* in the undiscounted setting (second inequality), we have

$$\begin{aligned}
V_\gamma^*(s) &\le V_\gamma^{(\mu^\gamma,\nu^*)}(s)\\
&= \mathbb{E}_{(\mu^\gamma,\nu^*)}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}r^{(\mu^\gamma,\nu^*)}(\mathscr{s}'_t) \quad |\mathscr{s}'_1 = s\right]\\
&\le \mathbb{E}_{(\mu^\gamma,\nu^*)}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}(J^* + h^*(\mathscr{s}'_t) - P^{(\mu^\gamma,\nu^*)}h^*(\mathscr{s}'_t)) \quad |\mathscr{s}'_1 = s\right].
\end{aligned}$$

Here, $\mathscr{s}'_{t+1}$ is generated with transition probability $P^{(\mu^\gamma,\nu^*)}(.|\mathscr{s}'_t)$ in the stochastic process under $\mathbb{E}_{(\mu^\gamma,\nu^*)}$, the last term in the expectation can be seen as $h^*(\mathscr{s}'_{t+1})$. Therefore we get

$$\begin{aligned} V^*_\gamma(s) &\leq \frac{J^*}{1-\gamma} + \mathbb{E}_{\mu^\gamma,\nu^*}\left[\sum_{t=1}^{+\infty} \gamma^{t-1}(h^*(\mathscr{s}'_t) - h^*(\mathscr{s}'_{t+1})) \quad |\mathscr{s}'_1 = s\right] \\ &= \frac{J^*}{1-\gamma} + h^*(\mathscr{s}'_1) - \mathbb{E}_{\mu^\gamma,\nu^*}\left[\sum_{t=2}^{+\infty}(\gamma^{t-2} - \gamma^{t-1})h^*(\mathscr{s}'_t) \quad |\mathscr{s}'_1 = s\right] \\ &\leq \frac{J^*}{1-\gamma} + \max_s h^*(s) - \min_s h^*(s) \sum_{t=2}^{+\infty}(\gamma^{t-2} - \gamma^{t-1}) \\ &\leq \frac{J^*}{1-\gamma} + sp(h^*) \\ &\leq \frac{J^*}{1-\gamma} + D. \end{aligned}$$

Putting both parts together and multiplying by $(1-\gamma)$ we obtain the desired statement.

5.2 Proof of Lemma 4

We simply write V^* and Q^* for V^*_γ and Q^*_γ respectively to lighten notations.

For every time t we have

$$V^*(s_t) - \mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t) = (V^* - \overline{V}^\downarrow_t)(s_t) + (\overline{V}^\downarrow_t - \mathbb{D}_{(\mu_t,\nu_t)}[Q^*])(s_t).$$

The second difference is upper bounded by

$$\begin{aligned} \overline{V}^\downarrow_t(s_t) - \mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t) &= \mathbb{D}_{\mu_t,\tilde{\nu}_t}[\overline{Q}^\downarrow_t](s_t) - \mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t) \\ &\leq \mathbb{D}_{(\mu_t,\nu_t)}[\overline{Q}^\downarrow_t - Q^*](s_t) \\ &= (\overline{Q}^\downarrow_t - Q^*)(s_t,a_t,b_t) + \zeta_t \\ &= (\overline{Q}^\downarrow_{t+1} - Q^*)(s_t,a_t,b_t) \\ &\quad + (\overline{Q}^\downarrow_t - \overline{Q}^\downarrow_{t+1})(s_t,a_t,b_t) + \zeta_t. \end{aligned}$$

The inequality holds by definition of $\tilde{\nu}_t$ as a best-response against μ_t for the matrix game $\overline{Q}^\downarrow_t(s_t,.,.)$. Under $\mathcal{E}_\delta$, inequality (4) bounds the term $(\overline{Q}^\downarrow_{t+1} - Q^*)(s_t,a_t,b_t)$, and therefore,

$$\begin{aligned} V^*(s_t) - \mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t) - \zeta_t \leq{}& (V^* - \overline{V}^\downarrow_t)(s_t) \\ &+ (\overline{Q}^\downarrow_t - \overline{Q}^\downarrow_{t+1})(s_t,a_t,b_t) \\ &+ 6D\sqrt{\frac{H}{n_t}\log(2T/\delta)} \\ &+ \sum_{i=1}^{n_t}\gamma\alpha^i_{n_t}\left[(\overline{V}^\downarrow_{t_i(s_t,a_t,b_t)} - V^*)(s_{t_i(s_t,a_t,b_t)+1})\right], \end{aligned}$$

where $n_t = N_{t+1}(s_t,a_t,b_t)$ and $t_i(s_t,a_t,b_t)$ is the i-th time when (s_t,a_t,b_t) has been visited. Note that $\alpha^0_{n_t}H = 0$ since $n_t = N_{t+1}(s_t,a_t,b_t) > 0$ by definition.

Now we sum over t to get **(s2)**. We need this technical Lemma :

Lemma 7. *Under $\mathcal{E}_\delta$, we have*

$$\gamma\sum_{t=1}^{T}\sum_{i=1}^{n_t}\alpha^i_{n_t}(\overline{V}^\downarrow_{t_i(s_t,a_t,b_t)} - V^*)(s_{t_i(s_t,a_t,b_t)+1}) \leq SH + \sum_{t=2}^{T+1}(\overline{V}^\downarrow_t - V^*)(s_t)$$

where $n_t = N_{t+1}(s_t,a_t,b_t)$ and $t_i(s_t,a_t,b_t)$ is the i-th time when (s_t,a_t,b_t) has been visited.

A proof of Lemma 7 is given in Appendix A.4 of the extended version [6].

To rest of the proof can then be separated in the following steps.

- By Lemma 7, it holds that

$$\begin{aligned} \sum_{t=1}^{T}\sum_{i=1}^{n_t}\gamma\alpha^i_{n_t}(\overline{V}^\downarrow_{t_i(s_t,a_t,b_t)} - V^*)(s_{t_i(s_t,a_t,b_t)+1}) \\ \leq SH + \sum_{t=2}^{T+1}(\overline{V}^\downarrow_t - V^*)(s_t). \end{aligned}$$

  This gives a telescoping sum with $\sum_{t=1}^{T}(V^* - \overline{V}^\downarrow_t)(s_t)$ and only $(\overline{V}^\downarrow_{T+1} - V^*)(s_{T+1}) - (\overline{V}^\downarrow_1 - V^*)(s_1) \leq 2H$ remains.
- Since $\overline{Q}^\downarrow_t$ decreases of at most H per tuple (s,a,b), we have

$$\sum_{t=1}^{T}(\overline{Q}^\downarrow_t - \overline{Q}^\downarrow_{t+1})(s_t,a_t,b_t) \leq SABH.$$

- Since $\sum_{s,a,b} N_{T+1}(s,a,b) = T$, we have

$$\begin{aligned} \sum_{t=1}^{T}\frac{1}{\sqrt{n_t}} &= \sum_{t=1}^{T}\sum_{s,a,b}\frac{\mathbf{1}(s_t = s, a_t = a, b_t = b)}{\sqrt{N_{t+1}(s,a,b)}} = \sum_{s,a,b}\sum_{j=1}^{N_{T+1}(s,a,b)}\frac{1}{\sqrt{j}} \\ &\leq \sum_{s,a,b} 2\sqrt{N_{T+1}(s,a,b)} \leq 2\sqrt{SAB\sum_{s,a,b}N_{T+1}(s,a,b)} \leq 2\sqrt{SABT}. \end{aligned}$$

Putting every thing together, we get the desired inequality for a certain absolute constant c.

6 CONCLUSION

In this work we propose the DONQ-learning algorithm that is the first model-free algorithm that deals with decentralized learning in infinite-horizon average-reward stochastic games. It uses an artificial discounted setting in order to use optimism techniques developed in previous works. This algorithm achieves sublinear regret performances both with high probability and in expectation. Surprisingly, our analysis gives a high probability upper bound of order $T^{\frac{3}{4}}$ whereas we show an upper bound of order $T^{\frac{2}{3}}$ for the expected regret. We are not sure if this difference is a proof artifact or an actual difference between the expected and high-probability regret. The high probability regret bound is obtained by choosing a parameter $H = T^{1/4}$ while the bound for the expected regret is obtained by choosing $H = T^{1/3}$. An open question is whether choosing $H = T^{1/3}$ leads to a high probability bound in $T^{2/3}$ or in $T^{5/6}$ which is the high probability bound given by our analysis. Another open question is to know if there exists a model-free algorithm that achieves the information-theoretical lower bound of order $T^{\frac{1}{2}}$ under our Assumption 1. For future work, the same setting (decentralized learning in infinite-horizon average-reward SG) where we additionally assume that the opponent's actions are unobservable may be studied. This has already been studied for the finite-horizon setting with for instance the V-learning algorithm. It would be interesting to see if such model-free algorithms can be adapted to the infinite-horizon setting.

REFERENCES

[1] Peter Auer, Thomas Jaksch, and Ronald Ortner. 2008. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems* 21 (2008).

[2] Yu Bai and Chi Jin. 2020. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*. PMLR, 551–560.

[3] Yu Bai, Chi Jin, and Tiancheng Yu. 2020. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems* 33 (2020), 2159–2170.

[4] Peter L Bartlett and Ambuj Tewari. 2012. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. *arXiv preprint arXiv:1205.2661* (2012).

[5] Ronen I Brafman and Moshe Tennenholtz. 2002. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* 3, Oct (2002), 213–231.

[6] Romain Cravic, Nicolas Gast, and Bruno Gaujal. 2023. Decentralized model-free reinforcement learning in stochastic games with average-reward objective. https://doi.org/10.48550/ARXIV.2301.05630

[7] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. 2020. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems* 33 (2020), 5527–5540.

[8] Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. 2019. Q-learning with ucb exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311* (2019).

[9] DM Gay, NK Karmarkar, and KG Ramakrishnan. 1986. The Karmarkar Algorithm-Adding Wings to Linear Programming. *Record - AT&T Bell Laboratories* 64, 2 (1986), 4–10.

[10] Mehdi Jafarnia-Jahromi, Rahul Jain, and Ashutosh Nayyar. 2021. Learning zero-sum stochastic games with posterior sampling. *arXiv preprint arXiv:2109.03396* (2021).

[11] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. 2018. Is Q-learning provably efficient? *Advances in neural information processing systems* 31 (2018).

[12] Leonid G Khachiyan. 1979. A polynomial algorithm in linear programming (english translation). In *Soviet Mathematics Doklady*, Vol. 20. 191–194.

[13] Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*. Elsevier, 157–163.

[14] Qinghua Liu, Yuanhao Wang, and Chi Jin. 2022. Learning Markov Games with Adversarial Opponents: Efficient Algorithms and Fundamental Limits. *arXiv preprint arXiv:2203.06803* (2022).

[15] Weichao Mao and Tamer Başar. 2022. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications* (2022), 1–22.

[16] Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. 2021. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems* 34 (2021), 18320–18334.

[17] Muhammed O Sayin, Francesca Parise, and Asuman Ozdaglar. 2022. Fictitious play in zero-sum stochastic games. *SIAM Journal on Control and Optimization* 60, 4 (2022), 2095–2114.

[18] Lloyd S Shapley. 1953. Stochastic games. *Proceedings of the national academy of sciences* 39, 10 (1953), 1095–1100.

[19] Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. 2020. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2992–3002.

[20] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[21] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.

[22] Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. 2021. Online learning in unknown markov games. In *International conference on machine learning*. PMLR, 10279–10288.

[23] Christopher John Cornish Hellaby Watkins. 1989. Learning from delayed rewards. (1989).

[24] Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. 2017. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems* 30 (2017).

[25] Chen-Yu Wei and Mehdi Jafarnia-Jahromi. 2020. Model-free Reinforcement Learning in Infinite-horizon Average-reward Markov Decision Processes", ICML (Int'l Conf. on Machine Learning) 2020. *Proceedings of Machine Learning Research* 119 (2020).

[26] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. 2021. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on learning theory*. PMLR, 4259–4299.

[27] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. 2020. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*. PMLR, 3674–3682.