Online Markov Decision Processes with Non-oblivious Strategic Adversary

Le Cong Dinh[1], David Henry Mguni[2], Long Tran-Thanh[3], Jun Wang[4] and Yaodong Yang[5*]

[1]University of Southampton.
[2]Huawei R&D U.K.
[3]University of Warwick.
[4]University College London.
[5]Peking University.

*Corresponding author(s). E-mail(s): yaodong.yang@pku.edu.cn;
Contributing authors: l.c.dinh@soton.ac.uk;
davidmguni@hotmail.com; long.tran-thanh@warwick.ac.uk;
jun.wang@cs.ucl.ac.uk;

Abstract

We study a novel setting in Online Markov Decision Processes (OMDPs) where the loss function is chosen by a *non-oblivious* strategic adversary who follows a no-external regret algorithm. In this setting, we first demonstrate that MDP-Expert, an existing algorithm that works well with oblivious adversaries can still apply and achieve a policy regret bound of $\mathcal{O}(\sqrt{T\log(L)} + \tau^2\sqrt{T\log(|A|)})$ where L is the size of adversary's pure strategy set and $|A|$ denotes the size of agent's action space. Considering real-world games where the support size of a NE is small, we further propose a new algorithm: *MDP-Online Oracle Expert* (MDP-OOE), that achieves a policy regret bound of $\mathcal{O}(\sqrt{T\log(L)} + \tau^2\sqrt{Tk\log(k)})$ where k depends only on the support size of the NE. MDP-OOE leverages the key benefit of Double Oracle in game theory and thus can solve games with prohibitively large action space. Finally, to better understand the learning dynamics of no-regret methods, under the same setting of no-external regret adversary in OMDPs, we introduce an algorithm that achieves last-round convergence to a NE result. To our best knowledge, this is the first work leading to the last iteration result in OMDPs.

Keywords: Multi-agent system, Game Theory, Online Learning, Online Markov Decision Processes, Non-oblivious Adversary, Last Round Convergence

1 Introduction

Reinforcement Learning (RL) [1] provides a general solution framework for optimal decision making under uncertainty, where the agent aims to minimise its cumulative loss while interacting with the environment. While RL algorithms have shown empirical and theoretical successes in stationary environments, it is an open challenge to deal with non-stationary environments in which the loss function and/or the transition dynamics change over time [2]. In tackling non-stationary environments, we are interested in designing learning algorithms that can achieve a no-regret guarantee [3, 4], where the regret is defined as the difference between the accumulated total loss and the total loss of the best fixed stationary policy in hindsight.

There are online learning algorithms that can achieve no-external regret property with changing loss function (but not changing transition dynamics), either in the full-information [3, 4] or the bandit [5, 6] settings. However, most existing solutions are established based on the key assumption that the adversary is *oblivious*, meaning the changes in loss functions do not depend on the historical trajectories of the agent. This crucial assumption limits the applicability of no-regret algorithms to many RL fields, particularly multi-agent reinforcement learning (MARL) [7]. In a multi-agent system, since all agents are learning simultaneously, one agent's adaption of its strategy will make the environment *non-oblivious* from other agents' perspectives. Therefore, to find the optimal strategy for each player, one must consider the strategic reactions from others rather than regard them as purely oblivious. As such, studying no-regret algorithms against a non-oblivious adversary is a pivotal step in adapting existing online learning techniques into MARL settings.

Another challenge in online learning is the non-convergence dynamics in a system. When agents apply no-regret algorithms such as Multiplicative Weights Update (MWU) [8] or Follow the Regularized Leader (FTRL) [9] to play against each other, the system demonstrates behaviours that are *Poincaré recurrent* [10], meaning the last-round convergence can never be achieved [11]. Recent works [12, 13] have focused on different learning dynamics in normal-form games that can lead to last-round convergence to a Nash equilibrium (NE) while maintaining the no-regret property. Yet, when it comes to OMDPs, it still remains an open challenge of how the no-regret property and the last-round convergence can be both achieved, especially against the strategic adversary. The focus of OMDPs is often on regret bound analysis against oblivious adversary [3–5], in which last-round convergence property is impossible to achieve due to the adversary's fixed behaviour. When a non-oblivious adversary is considered, the focus is on finding stationary points of the system [14, 15] rather than analysing the dynamic leading to the last round convergence to a NE.

Markov decision processes (MDPs) provide a popular tool to formulate stochastic optimization problems [1], yet it is often that only a relaxation of real models can satisfy the Markovian assumption. In situations where the reward function can change over time and thus the Markovian assumption is not satisfied, OMDPs offer a general solution by applying existing experts' algorithms to more adversarial MDPs [3]. OMDPs algorithms provide the agent with a performance guarantee under the assumption that the adversary is oblivious [4, 16], thus limiting its application in settings where the adversary is also a learning agent.

In this paper, we relax the assumption of the oblivious adversary in OMDPs and study a new setting where the loss function is chosen by a strategic agent that follows a no-external regret algorithm. This setting can be used in applications within economics to model systems and firms [17], for example, an oligopoly with a dominant player, or ongoing interactions between industry players and authority (e.g., a government that acts as an order-setting body). Another motivating example is the stochastic inventory control problem [18]. In each period, based on the current inventory, the store manager needs to decide the number of items to order from the supplier. The manager faces the dilemma: having too many items will increase the inventory cost while running out of items will lead to revenue loss. Since both the item price and the inventory cost can change over time, the problem can be considered as OMDPs. Furthermore, the supplier can decide the item price based on the total demand of the item as well as its capacity to maximise its profit, thus making it a non-oblivious strategic adversary.

Under this setting, we study how the agent can achieve different goals such as no-policy regret and last-round convergence.

Our contributions are at three folds:

- We prove that the well-known MDP-Expert (MDP-E) algorithm [3] can still apply by achieving a policy regret bound of $\mathcal{O}(\sqrt{T\log(L)}+\tau^2\sqrt{T\log(|A|)})$, and the average strategies of the agents will converge to a NE of the game.
- For many real-world applications where the support size of NE is small [19, 20], we introduce an efficient no-regret algorithm, *MDP-Online Oracle Expert (MDP-OOE)*, which achieves the policy regret bound of $\mathcal{O}(\tau^2\sqrt{Tk\log(k)} + \sqrt{T\log(L)})$ against non-oblivious adversary, where k depends on the support size of the NE. MDP-OOE inherits the key benefits of both Double Oracle [19] and MDP-E [3]; it can solve games with large action space while maintaining the no-regret property.
- To achieve last-round convergence guarantee for no-external regret algorithms, we introduce the algorithm of *Last-Round Convergence in OMDPs (LRC-OMDPs)* such that in cases where the adversary follows a no-external regret algorithm, the dynamics will lead to the last-round convergence to a NE. To the best of our knowledge, this is the first last-iteration convergence result in OMDPs.

Table 1: The scope of our contribution in this work.

	Non-oblivious adversary within a two-player game framework	Oblivious adversary in Markov Decision Processes
Regret *w.r.t* best policy in hindsight	MDP-OOE (our contribution) $\mathcal{O}(\tau^2\sqrt{Tk\log(k)}+\sqrt{T\log(L)})$	OMDPs: (MDP-E) [3] $\mathrm{Reg}_T=\mathcal{O}(\tau^2\sqrt{T\log(\lvert A\rvert)})$
Regret *w.r.t.* value of the game	SGs:(UCSG) [21] $\mathrm{Reg}_T=\tilde{\mathcal{O}}(D^3\mathsf{S}^5\lvert A\rvert+D\mathsf{S}\sqrt{\lvert A\rvert T})$	OMDPs

2 Related Work

The setting of OMDPs with no-external regret adversary, though novel, shares certain aspects in common with existing literature in online learning and stochastic game domains. Here we review each of them.

Many researchers have considered OMDPs with an oblivious environment, where the loss function can be set arbitrarily. The performance of the algorithm is measured by external regret: the difference between the total loss and the best stationary policy in hindsight. In this setting with stationary transition dynamics, MDP-E [3] proved that if the agent bounds the "local" regret in each state, then the "global" regret will be bounded. Neu et al. [5, 16] considered the same problem with the bandit reward feedback and provided no-external regret algorithms in this setting. Dick et al. [4] studied a new approach for OMDPs where the problem can be transformed into an online linear optimization form, from which no-external regret algorithms can be derived. Cheung et al. [22] proposed a no-external regret algorithm in the case of non-stationary transition distribution, given that the variation of the loss and transition distributions do not exceed certain variation budgets.

In a non-oblivious environment, Yu et al.[23] provided an example demonstrating that no algorithms can guarantee sublinear external regret against a non-oblivious adversary. Thus, in OMDPs with non-oblivious opponents (e.g., agents using adaptive algorithms), the focus is often on finding stationary points of the system rather than finding a no-external regret algorithm [14]. In this paper, we study cases where the adversary follows an adaptive no-regret algorithm, and tackle the hardness result of non-oblivious environments in OMDPs.

The problem of the non-oblivious adversary has also been studied in the multi-armed bandit setting, a special case of OMDPs. In this setting, Arora et al. [24] considered m-memory bounded adversary and provided an algorithm with a policy regret bound that depends linearly on m, where the policy regret includes the adversary's adaptive behaviour (i.e., see Equation (1)). Compared to their work, our paper considers strategic adversary which turns out to be ∞-memory bounded adversary. Thus the algorithm suggested in [24] can not be applied. Recently, Dinh et al. [12] studied the same strategic adversary in full information normal-form setting and provided an algorithm that leads to last round convergence. However, both of the above works only studied the simplified version of OMDPs, thus they do not capture the complexity of the problem. We argue that since strategic adversary setting has many applications due to the popularity of no-regret algorithms [25–27], it is important to study no-regret methods in more practical settings such as OMDPs.

Stochastic games (SGs) [28, 29] offer a multi-player game framework where agents jointly decide the loss and the state transition. Compared to OMDPs, the main difference is that SGs allow each player to have a representation of states, actions and rewards, thus players can learn the representations over time and find the NE of the stochastic games [21, 30]. The performance in SGs is often measured by the difference between the average loss and the value of the game (i.e., the value when both players play a NE), which is a weaker notion of regret compared to the best fixed policy in hindsight in OMDPs. Intuitively, the player can learn the structure of the game (i.e., transition model, reward function) over time, thus on average, the player can calculate and compete with the value of the game. In non-episodic settings, the Upper Confidence Stochastic Game algorithm (UCSG) [21] guarantees the regret of $\text{Reg}_T = \tilde{\mathcal{O}}(D^3\mathsf{S}^5|A| + D\mathsf{S}\sqrt{|A|T})$ with high probability, given that the opponent's action is observable. However, to compete with the best stationary policy, knowing the game structure does not guarantee a good performance (i.e., the performance will heavily depend on the strategic behaviour of opponents). Tian et al. [30] proved that in the SG setting, achieving no regret with respect to the best stationary policy in hindsight is statistically hard. Our settings can be considered as a sub-class of SGs where only the agent controls the transition model (i.e., single controller SGs), based on this, we try to overcome the above challenge.

We summarise the difference between our setting and OMDPs and SGs in Table 1. Compared to OMDPs, we relax the assumption about the oblivious environment and study a non-oblivious counterpart with a strategic adversary. Compared to SGs, we relax the assumption of knowing the opponent's action in a non-episodic setting and our results only require observing the loss functions. Furthermore, the performance measurement is with respect to the best stationary policy in hindsight, which is proved to be statistically hard in SGs [30]. Intuitively, since we consider the problem of single controller SGs, it can overcome the hardness result. Guan et al. [15] studied a similar setting to our paper, where only one player affects the transition kernel of the game. By viewing the game as an online linear optimisation, it can derive the minimax equilibrium of the game. There are two main challenges of the algorithm. Firstly, it requires both players to pre-calculate the minimax equilibrium of the game and fixes to this strategy during the repeated game. Thus, in the situation where the adversary is an independent agent (i.e., it follows a different learning dynamic), the proposed algorithm can not be applied. Secondly, and most importantly, the no regret analysis is not provided for the algorithm in [15], thus the algorithm can not be applied in an adversary environment. We fully address both challenges in this paper.

3 Problem Formulations & Preliminaries

We consider OMDPs where at each round $t \in \mathbb{N}$, an adversary can choose the loss function $\boldsymbol{l}_t$ based on the agent's policy history $\{\pi_1, \pi_2, \dots, \pi_{t-1}\}$. Formally, we have OMDPs with finite state space S; finite action set for the agent at each state A; and a fixed transition model P. The agent's starting state, x_1, is distributed according to some distribution μ_0 over S. At time t, given state $x_t \in S$, the agent chooses an action $a_t \in A$, then the agent moves to a new random state x_{t+1} which is determined by the fixed transition model $P(x_{t+1}|x_t, a_t)$. Simultaneously, the agent receives an immediate loss $\boldsymbol{l}_t(x_t, a_t)$, in which the loss function $\boldsymbol{l}_t : S \times A \to R$ is bounded in $[0,1]^{|A|\times|S|}$ and chosen by the adversary from a simplex $\Delta_L := \{\boldsymbol{l} \in \mathbb{R}^{|A|\times|S|} | \boldsymbol{l} = \sum_{i=1}^{L} x_i \boldsymbol{l}_i,\ \sum_{i=1}^{L} x_i = 1,\ x_i \geq 0\ \forall i\}$ where $\{\boldsymbol{l}_1, \boldsymbol{l}_2, \dots, \boldsymbol{l}_L\}$ are the loss vectors of the adversary. We assume zero-sum game setting where the adversary receives the loss of $-\boldsymbol{l}_t(x_t, a_t)$ at round t and consider popular full information feedback [3, 4], meaning the agent can observe the loss function $\boldsymbol{l}_t$ after each round t.

Against the strategic adversary, the formal definition of no-external regret becomes inadequate since the adversary is allowed to adapt to the agent's action. In this paper, we adopt the same approach in [24] and consider policy regret. Formally, the goal of the agent is to have minimum policy regret with respect to the best fixed policy in hindsight:

$$R_T(\pi) = \mathbb{E}_{X,A}\left[\sum_{t=1}^{T} \boldsymbol{l}_t^{\pi_t}(X_t, A_t)\right] - \mathbb{E}_{X,A}\left[\sum_{t=1}^{T} \boldsymbol{l}_t^{\pi}(X_t^{\pi}, A_t^{\pi})\right], \tag{1}$$

where $\boldsymbol{l}_t^{\pi_t}$ denotes the loss function at time t while the agent follows $\pi_1, \dots, \pi_T$ and $\boldsymbol{l}_t^{\pi}$ is the adaptive loss function against the fixed policy π of the agent. We say that the agent achieves sublinear policy regret (i.e., no-policy regret property) with respect to the best fixed strategy in hindsight if $R_T(\pi)$ satisfies:

$$\lim_{T\to\infty} \max_{\pi} \frac{R_T(\pi)}{T} = 0.$$

In a general non-oblivious adversary, we prove by a counter example that it is impossible to achieve an algorithm with a sublinear policy regret [1]. Suppose the agent faces an adversary such that it gives a very low loss for the agent if the action in the first round of the agent is a specific action (i.e., by fixing the loss function to $\mathbf{0}$), otherwise the adversary will give a high loss (i.e., by fixing the loss function to $\mathbf{1}$). Against this type of adversary, without knowing the specific action, the agent's policy regret in Equation (1) will be $\mathcal{O}(T)$. Thus, in general non-oblivious adversary cases, we will have a hardness result in policy regret. To resolve the hardness result, we study the strategic adversary in OMDPs.

[1] In the multi-armed bandit setting, it is also impossible to achieve sublinear policy regret against all adaptive adversaries (see Theorem 1 in [24]).

Assumption 1 (Strategic Adversary) *The adversary flows a no-external regret algorithm such as for any sequence of π_t:*

$$\lim_{T\to\infty} \max_{\boldsymbol{l}} \frac{R_T(\boldsymbol{l})}{T} = 0, \text{ where } R_T(\boldsymbol{l}) = \mathbb{E}_{X,A}\left[\sum_{t=1}^{T} \boldsymbol{l}(X_t, A_t)\right] - \mathbb{E}_{X,A}\left[\sum_{t=1}^{T} \boldsymbol{l}_t^{\pi_t}(X_t, A_t)\right].$$

The rationale of Assumption 1 comes from the vanilla property of no-external algorithms: without prior information, the adversary will not do worse than the best-fixed strategy in hindsight [12]. Thus, without the priority knowledge about the agent, the adversary will have an incentive to follow a no-external regret algorithm. In the same way as the full information feedback assumption for the agent, we assume that after each round t, the adversary observes the agent's stationary policy distribution $\boldsymbol{d}_{\pi_t}$.

For every policy π, we define $P(\pi)$ the state transition matrix induced by π such that $P(\pi)_{s,s'} = \sum_{a\in A} \pi(a|s) P^a_{s,s'}$. We assume through the paper that we have the mixing time assumption, which is a common assumption in OMDPs [3, 4, 16]:

Assumption 2 (Mixing time) *There exists a constant $\tau > 0$ such that for all distributions $\boldsymbol{d}$ and $\boldsymbol{d}'$ over the state space, any policy π,*

$$\left\|\boldsymbol{d}P(\pi) - \boldsymbol{d}'P(\pi)\right\|_1 \le e^{-1/\tau}\left\|\boldsymbol{d} - \boldsymbol{d}'\right\|_1,$$

where $\|\boldsymbol{x}\|_1$ denotes the l_1 norm of a vector $\boldsymbol{x}$.

Denote $\boldsymbol{v}_t^{\pi}(x, a)$ the probability of (state, action) pair (x, a) at time step t by following policy π with initial state x_1. Following Assumption 2, for any initial states, $\boldsymbol{v}_t^{\pi}$ will converge to a stationary distribution $\boldsymbol{d}_{\pi}$ as t goes to infinity. Denote $\boldsymbol{d}_{\Pi}$ the stationary distribution set from all agent's deterministic policies. With a slight abuse of notation, when an agent follows an algorithm A with use $\pi_1, \pi_2, \ldots$ at each time step, we denote $\boldsymbol{v}_t(x, a) = \mathbb{P}\left[X_t = x, A_t = a\right], \quad \boldsymbol{d}_t = \boldsymbol{d}_{\pi_t}$. Thus, the regret in Equation (1) can be expressed as

$$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{v}_t \right\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \boldsymbol{l}_t^{\pi}, \boldsymbol{v}_t^{\pi} \right\rangle\right].$$

Assumption 2 allows us to define the average loss of policy π in an online MDP with a loss $\boldsymbol{l}$ as $\eta_{\boldsymbol{l}}(\pi) = \langle \boldsymbol{l}, \boldsymbol{d}_{\pi} \rangle$ and the accumulated loss $Q_{\pi,\boldsymbol{l}}(s, a)$ is defined as

$$Q_{\pi,\boldsymbol{l}}(s, a) = \mathbb{E}\left[\sum_{t=1}^{\infty} \left(\boldsymbol{l}(s_t, a_t) - \eta_{\boldsymbol{l}}(\pi)\right) \middle| s_1 = s, a_1 = a, \pi\right].$$

As the dynamic between the agent and adversary is zero-sum, we can apply the minimax theorem [31]:

Algorithm 1 MDP-Expert (MDP-E)

```
Input: Expert algorithm $B_s$ (i.e., MWU) for each state
for $t = 1$ to $\infty$ do
    Using algorithm $B_s$ with set of expert $A$ and the feedback $Q_{\pi_t,\boldsymbol{l}_t}(s,.)$ for each state $s$
    Output $\pi_{t+1}$ and observe $\boldsymbol{l}_{t+1}$
end for
```

$$\min_{\boldsymbol{d}_\pi \in \Delta_{\boldsymbol{d}_\Pi}} \max_{\boldsymbol{l} \in \Delta_L} \langle \boldsymbol{l}, \boldsymbol{d}_\pi \rangle = \max_{\boldsymbol{l} \in \Delta_L} \min_{\boldsymbol{d}_\pi \in \Delta_{\boldsymbol{d}_\Pi}} \langle \boldsymbol{l}, \boldsymbol{d}_\pi \rangle = v. \tag{2}$$

The saddle point $(\boldsymbol{l}, \boldsymbol{d}_\pi)$ that satisfies Equation (2) is the NE of the game [32] and v is the called the value of the game. Our work is based on no-external regret algorithms in normal-form games such as Multiplicative Weights Update [8], which is described as

Definition 1 (Multiplicative Weights Update) Let $\boldsymbol{k}_1, \boldsymbol{k}_2, ...$ be a sequence of feedback received by the agent. The agent is said to follow the MWU if strategy $\tilde{\boldsymbol{\pi}}_{t+1}$ is updated as follows

$$\tilde{\boldsymbol{\pi}}_{t+1}(i) = \tilde{\boldsymbol{\pi}}_t(i) \frac{\exp(-\mu_t \boldsymbol{k}_t(\boldsymbol{a}^i))}{\sum_{i=1}^n \tilde{\boldsymbol{\pi}}_t(i) \exp(-\mu_t \boldsymbol{k}_t(\boldsymbol{a}^i))}, \forall i \in [n], \tag{3}$$

where $\mu_t > 0$ is a parameter, n is the number of pure strategies (i.e., experts) and $\tilde{\boldsymbol{\pi}}_0 = [1/n, \dots, 1/n]$.

We also consider ϵ-Nash equilibrium of the game:

Definition 2 (ϵ-Nash equilibrium) Assume $\epsilon > 0$. We call a point $(\boldsymbol{l}, \boldsymbol{d}_\pi) \in \Delta_L \times \Delta_{\boldsymbol{d}_\Pi}$ ϵ-NE if:

$$\max_{\boldsymbol{l} \in \Delta_L} \langle \boldsymbol{l}, \boldsymbol{d}_\pi \rangle - \epsilon \le \langle \boldsymbol{l}, \boldsymbol{d}_\pi \rangle \le \min_{\boldsymbol{d}_\pi \in \Delta_{\boldsymbol{d}_\Pi}} \langle \boldsymbol{l}, \boldsymbol{d}_\pi \rangle + \epsilon.$$

Under the setting of OMDPs against the strategic adversary who aims to minimise the external regret (i.e., Assumption 1), we study several properties that the agent can achieve such as no-policy regret and last round convergence.

4 MDP-Expert against Strategic Adversary

When the agent plays against a non-oblivious opponent, one challenge is that the best fixed policy π is not based on the current loss sequence $[\boldsymbol{l}_1, \boldsymbol{l}_2, \dots]$ of the agent but a different loss sequence $[\boldsymbol{l}_1^\pi, \boldsymbol{l}_2^\pi \dots]$ induced by the policy π. Thus, to measure the regret in the case of a non-oblivious opponent, we need information on how the opponent will play against a fixed policy π. Under Assumption 1, we prove that existing MDP-E [3] method, which is designed for the oblivious adversary, will have no- policy regret property against the non-oblivious strategic adversary in our setting. Intuitively, MDP-E maintains a no-external regret algorithm (i.e., MWU) in each state to bound the local regret, thus the global regret can be bounded accordingly. The pseudocode of MDP-E is given in Algorithm 1. The following lemma links the relationship between the external-regret of the adversary and the regret with respect to the policy stationary distribution:

Lemma 1 *Under MDP-E played by the agent, the external-regret of the adversary in Assumption 1 can be expressed as:*

$$\begin{aligned} R_T(\boldsymbol{l}) &= \mathbb{E}_{X,A}\left[\sum_{t=1}^{T} \boldsymbol{l}(X_t, A_t)\right] - \mathbb{E}_{X,A}\left[\sum_{t=1}^{T} \boldsymbol{l}_t^{\pi_t}(X_t, A_t)\right] \\ &= \sum_{t=1}^{T}\langle \boldsymbol{l}, \boldsymbol{d}_{\pi_t}\rangle - \sum_{t=1}^{T}\langle \boldsymbol{l}_t, \boldsymbol{d}_{\pi_t}\rangle + \mathcal{O}\big(\tau^2\sqrt{T\log(|A|)}\big). \end{aligned}$$

Proof It is sufficient to show that for any sequence of $\boldsymbol{l}_t$

$$\mathbb{E}_{X,A}\left[\sum_{t=1}^{T} \boldsymbol{l}_t(X_t, A_t)\right] - \sum_{t=1}^{T}\langle \boldsymbol{l}_t, \boldsymbol{d}_{\pi_t}\rangle = \mathcal{O}(\tau^2\sqrt{T\log(|A|)}),$$

where $\boldsymbol{l}_t$ denotes the loss vector of the adversary when the agent follows $\pi_1, \pi_2, \ldots$ (i.e., the same as $\boldsymbol{l}_t^{\pi_t}$).

Using the consequence of Lemma 5.2 in [3] [2], for any sequence of $\boldsymbol{l}_t$ we have:

$$\begin{aligned} &\mathbb{E}_{X,A}\left[\sum_{t=1}^{T} \boldsymbol{l}_t(X_t, A_t)\right] - \sum_{t=1}^{T}\langle \boldsymbol{l}_t, \boldsymbol{d}_{\pi_t}\rangle \\ &= \sum_{t=1}^{T}\langle \boldsymbol{l}_t, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\rangle \leq \sum_{t=1}^{T}|\langle \boldsymbol{l}_t, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\rangle| \leq \sum_{t=1}^{T}\|\boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\|_1 \\ &\leq \sum_{t=1}^{T} 2\tau^2\sqrt{\frac{\log(|A|)}{t}} + 2e^{-t/\tau} \\ &\leq 4\tau^2\sqrt{T\log(|A|)} + 2(1+\tau) = \mathcal{O}\big(\tau^2\sqrt{T\log(|A|)}\big). \end{aligned} \tag{4}$$

The proof is complete. □

Based on Lemma 1, we can tell that the sublinear regret will hold if and only if the adversary maintains a sublinear regret with respect to the agent's policy stationary distribution. As we assume that after each time t, the adversary can observe the stationary distribution $\boldsymbol{d}_{\pi_t}$, then by applying standard no-external regret algorithm for online linear optimization against the feedback $\boldsymbol{d}_{\pi_t}$ (i.e., MWU), the adversary can guarantee good performance for himself. Thus, the Assumption 1 for the adversary is justifiable.

In the rest of the paper, without loss of generality, we will study the case where the external-regret of the adversary with respect to the agent's policy

[2] For the completeness of the paper, we provide the lemma in Appendix A.

stationary distribution has the following bound (i.e., the adversary follows optimal no-external regret algorithms such as MWU, FTRL with respect to policy stationary distribution of the agent [3]):

$$\max_{\boldsymbol{l}\in\Delta_L}\left(\sum_{t=1}^{T}\langle \boldsymbol{l}, \boldsymbol{d}_{\pi_t}\rangle - \sum_{t=1}^{T}\langle \boldsymbol{l}_t, \boldsymbol{d}_{\pi_t}\rangle\right) = \sqrt{\frac{T\log(L)}{2}}.$$

The next lemma provides a lower bound for the performance of a fixed policy of the agent against a strategic adversary.

Lemma 2 *Suppose the agent follows a fixed stationary strategy π, then the adversary will converge to the best response to the fixed stationary strategy and*

$$\sum_{t=1}^{T}\langle \boldsymbol{l}_t^{\pi}, \boldsymbol{d}_{\pi}\rangle \geq Tv - \sqrt{\frac{T\log(L)}{2}}.$$

Proof From Lemma 1, if the adversary follows a no-regret algorithm to achieve good performance in Assumption 1, then the adversary must follow a no-external regret algorithm with respect to the policy's stationary distribution. Without loss of generality, we can assume that the adversary follows the Multiplicative Weight Update with respect to the policy's stationary distribution $\boldsymbol{d}_\pi$. Then following the property of Multiplicative Weight Update in an online linear problem, we have:

$$\max_{\boldsymbol{l}\in L}\langle \boldsymbol{l}, \boldsymbol{d}_{\pi}\rangle - \frac{1}{T}\sum_{t=1}^{T}\langle \boldsymbol{l}_t^{\pi}, \boldsymbol{d}_{\pi}\rangle \leq \sqrt{\frac{\log(L)}{2T}}.$$

From the famous minimax theorem [31] we also have:

$$\max_{\boldsymbol{l}\in L}\langle \boldsymbol{l}, \boldsymbol{d}_{\pi}\rangle \geq \min_{\boldsymbol{d}_\pi\in\boldsymbol{d}_\Pi}\max_{\boldsymbol{l}\in L}\langle \boldsymbol{l}, \boldsymbol{d}_{\pi}\rangle = v.$$

Thus we have:

$$\begin{aligned}\sum_{t=1}^{T}\langle \boldsymbol{l}_t^{\pi}, \boldsymbol{d}_{\pi}\rangle &\geq T\max_{\boldsymbol{l}\in L}\langle \boldsymbol{l}, \boldsymbol{d}_{\pi}\rangle - \sqrt{\frac{T\log(L)}{2}} \\ &\geq Tv - \sqrt{\frac{T\log(L)}{2}}.\end{aligned} \tag{5}$$

□

From Lemma 2, we can prove the following theorem:

Theorem 1 *Suppose the agent follows MDP-E Algorithm 1, then the regret with respect to the stationary distribution will be bounded by*

$$\sum_{t=1}^{T}\left\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\right\rangle - \sum_{t=1}^{T}\left\langle \boldsymbol{l}_t^{\pi}, \boldsymbol{d}_{\pi}\right\rangle \leq \sqrt{\frac{T\log(L)}{2}} + 3\tau\sqrt{\frac{T\log(|A|)}{2}}.$$

[3] If the adversary does not follow the optimal bound (i.e., irrational), then regret bound of the agent will change accordingly.

Proof From Lemma 2, it is sufficient to show that

$$\sum_{t=1}^{T} \left\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \right\rangle \leq Tv + 3\tau\sqrt{\frac{T\log(|A|)}{2}}.$$

Since the agent uses a no-regret algorithm with respect to the stationary distribution (i.e., MDP-E), following the same argument in Theorem 5.3 in [3] we have:

$$\sum_{t=1}^{T} \left\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \right\rangle \leq T \min_{\boldsymbol{d}_\pi} \left\langle \hat{\boldsymbol{l}}, \boldsymbol{d}_\pi \right\rangle + 3\tau\sqrt{\frac{T\log(|A|)}{2}},$$

where $\hat{\boldsymbol{l}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{l}_t^{\pi_t}$. From the minimax equilibrium, we also have

$$\min_{\boldsymbol{d}_\pi} \langle \hat{\boldsymbol{l}}, \boldsymbol{d}_\pi \rangle \leq \max_{\boldsymbol{l}\in\Delta_L} \min_{\boldsymbol{d}_\pi \in \boldsymbol{d}_\Pi} \langle \boldsymbol{l}, \boldsymbol{d}_\pi \rangle = v.$$

Thus, the proof is complete. □

Now, we can make the link between the stationary regret and the regret of the agent in Equation (1).

Theorem 2 *Suppose the agent follows MDP-E Algorithm 1, then the agent's regret in Equation (1) will be bounded by*

$$R_T(\pi) = \mathcal{O}(\sqrt{T\log(L)} + \tau^2\sqrt{T\log(|A|)}).$$

Proof Using the consequence of Lemma 5.2 in [3], for any sequence of $\boldsymbol{l}_t$ we have:

$$\begin{aligned} \sum_{t=1}^{T} \langle \boldsymbol{l}_t, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t} \rangle &\leq \sum_{t=1}^{T} |\langle \boldsymbol{l}_t, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t} \rangle| \leq \sum_{t=1}^{T} \|\boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\|_1 \\ &\leq \sum_{t=1}^{T} 2\tau^2 \sqrt{\frac{\log(|A|)}{t}} + 2e^{-t/\tau} \\ &\leq 4\tau^2\sqrt{T\log(|A|)} + 2(1+\tau) = \mathcal{O}\big(\tau^2\sqrt{T\log(|A|)}\big). \end{aligned} \tag{6}$$

Thus we have

$$\sum_{t=1}^{T} |\langle \boldsymbol{l}_t, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t} \rangle| \leq 2(1+\tau) + 4\tau^2\sqrt{T\log(|A|)}. \tag{7}$$

Furthermore, if the agent uses a fixed policy π then by Lemma 2, we have:

$$|\sum_{t=1}^{T} \langle \boldsymbol{l}_t, \boldsymbol{d}_\pi - \boldsymbol{v}_t^\pi \rangle| \leq 2\tau + 2.$$

Since the agent uses MDP-E, a no-external regret algorithm, following the same argument in Theorem 4.1 in [3] we have:

$$\sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle \leq T \min_{\boldsymbol{d}_\pi} \langle \hat{\boldsymbol{l}}, \boldsymbol{d}_\pi \rangle + 3\tau\sqrt{\frac{T\log(|A|)}{2}} \leq Tv + 3\tau\sqrt{\frac{T\log(|A|)}{2}}.$$

Along with Lemma 2, we have:

$$\begin{aligned}\sum_{t=1}^{T}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \sum_{t=1}^{T}\langle \boldsymbol{l}_t^{\pi}, \boldsymbol{d}_{\pi}\rangle &\leq \left(Tv + 3\tau\sqrt{\frac{T\log(|A|)}{2}}\right) - \left(Tv - \sqrt{\frac{T\log(L)}{2}}\right)\\ &= 3\tau\sqrt{\frac{T\log(|A|)}{2}} + \sqrt{\frac{T\log(L)}{2}}.\end{aligned}$$

Using the above two inequalities, we can bound the regret of the agent with respect to the regret of the policy's stationary distribution:

$$\begin{aligned}R_T(\pi) &= \mathbb{E}_{x,a}\left[\sum_{t=1}^{T}\boldsymbol{l}_t^{\pi_t}(x_t, a_t)\right] - \mathbb{E}_{x,a}\left[\sum_{t=1}^{T}\boldsymbol{l}_t^{\pi}(x_t^{\pi}, a_t^{\pi})\right]\\ &= \sum_{t=1}^{T}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{v}_t\rangle - \sum_{t=1}^{T}\langle \boldsymbol{l}_t^{\pi}, \boldsymbol{v}_t^{\pi}\rangle\\ &\leq \sum_{t=1}^{T}\left(\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle + |\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\rangle|\right) - \sum_{t=1}^{T}\left(\langle \boldsymbol{l}_t^{\pi}, \boldsymbol{d}_{\pi}\rangle - |\langle \boldsymbol{l}_t^{\pi}, \boldsymbol{v}_t^{\pi} - \boldsymbol{d}_{\pi}\rangle|\right) \qquad (8)\\ &\leq \sum_{t=1}^{T}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \sum_{t=1}^{T}\langle \boldsymbol{l}_t^{\pi}, \boldsymbol{d}_{\pi}\rangle + 2(1+\tau) + 4\tau^2\sqrt{T\log(|A|)} + 2 + 2\tau\\ &\leq \sqrt{\frac{T\log(L)}{2}} + 3\tau\sqrt{\frac{T\log(|A|)}{2}} + 4(1+\tau) + 4\tau^2\sqrt{T\log(|A|)}\\ &= \mathcal{O}(\sqrt{T\log(L)} + \tau^2\sqrt{T\log(|A|)}).\end{aligned}$$

The proof is complete. □

We note that Theorem 2 will hold true for a larger set of adversaries outside Assumption 1 (e.g., FP [33]) satisfying the following property: for every fixed policy of the agent, the adversary's policy converges to the best response with respect to this fixed policy. With this property, we can bound the performance of the agent's fixed policy in Lemma 2 and thus derive the regret bound of the algorithm. Note that the regret bound in Theorem 2 will depend on the rate of convergence to the best response against the agent's fixed policy.

As we have shown in previous theorems, the dynamic of playing a no-regret algorithm in OMDPs against a strategic adversary can be interpreted as a two-player zero-sum game setting with the corresponding stationary distribution. From the classical saddle point theorem [8], if both players follows a no-regret algorithm then the average strategies will converge to the saddle point (i.e., a NE).

Theorem 3 *Suppose the agent follows MDP-E, then the average strategies of both the agent and the adversary will converge to the ϵ_t-Nash equilibrium of the game with:*

$$\epsilon_T = \sqrt{\frac{\log(L)}{2T}} + 3\tau\sqrt{\frac{\log(|A|)}{2T}}.$$

Proof Since the agent and the adversary use no-regret algorithms with respect to the policy's stationary distribution, we can use the property of regret bound in a normal-form game to apply. Thus we have:

$$\max_{\boldsymbol{l} \in L} \langle \boldsymbol{l}, \hat{\boldsymbol{d}}_\pi \rangle - \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle \leq \sqrt{\frac{\log(L)}{2T}},$$

$$\frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - \min_{\boldsymbol{d}_\pi} \langle \hat{\boldsymbol{l}}, \boldsymbol{d}_\pi \rangle \leq 3\tau \sqrt{\frac{\log(|A|)}{2T}},$$

where $\hat{\boldsymbol{d}}_\pi = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{d}_{\pi_t}$ and $\hat{\boldsymbol{l}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{l}_t^{\pi_t}$. From this, we can prove that

$$\begin{aligned} \langle \hat{\boldsymbol{l}}, \hat{\boldsymbol{d}}_\pi \rangle &\geq \min_{\boldsymbol{d}_\pi} \langle \hat{\boldsymbol{l}}, \boldsymbol{d}_\pi \rangle \geq \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - 3\tau \sqrt{\frac{\log(|A|)}{2T}} \\ &\geq \max_{\boldsymbol{l} \in L} \langle \boldsymbol{l}, \hat{\boldsymbol{d}}_\pi \rangle - \sqrt{\frac{\log(L)}{2T}} - 3\tau \sqrt{\frac{\log(|A|)}{2T}}, \end{aligned}$$

and,

$$\begin{aligned} \langle \hat{\boldsymbol{l}}, \hat{\boldsymbol{d}}_\pi \rangle &\leq \max_{\boldsymbol{l} \in L} \langle \boldsymbol{l}, \hat{\boldsymbol{d}}_\pi \rangle \leq \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle + \sqrt{\frac{\log(L)}{2T}} \\ &\leq \min_{\boldsymbol{d}_\pi} \langle \hat{\boldsymbol{l}}, \boldsymbol{d}_\pi \rangle + 3\tau \sqrt{\frac{\log(|A|)}{2T}} + \sqrt{\frac{\log(L)}{2T}}. \end{aligned}$$

Thus, with $\epsilon_t = \sqrt{\frac{\log(L)}{2T}} + 3\tau \sqrt{\frac{\log(|A|)}{2T}}$, we derive

$$\max_{\boldsymbol{l} \in L} \langle \boldsymbol{l}, \hat{\boldsymbol{d}}_\pi \rangle - \epsilon_t \leq \langle \hat{\boldsymbol{l}}, \hat{\boldsymbol{d}}_\pi \rangle \leq \min_{\boldsymbol{d}_\pi} \langle \hat{\boldsymbol{l}}, \boldsymbol{d}_\pi \rangle + \epsilon_t.$$

By definition, $(\hat{\boldsymbol{l}}, \hat{\boldsymbol{d}}_\pi)$ is ϵ_t-Nash equilibrium. □

With the sublinear convergence rate to an NE, the dynamic between MDP-E and no-regret adversary (i.e., MWU) provides an efficient method to solve the single-controller SGs.

5 MDP-Online Oracle Expert Algorithm

As shown in the previous section, we can bound the regret in Equation (1) by bounding the regret with respect to the stationary distribution. In MDP-E, the regret bound (i.e., $\mathcal{O}\big(\sqrt{T \log(L)} + \tau^2 \sqrt{T \log(|A|)}\big)$) depends on the size of pure strategy set (i.e., $|A|$) thus it becomes less efficient when the agent has a prohibitively large pure strategy set.

Interestingly, a recent paper by Dinh et al. [20] suggested that in normal-form games, it is possible to achieve a better regret bound where it only depends on the support size of NE rather than $|A|$. Unfortunately, extending this finding for OMDPs is highly non-trivial. The method in [20] is designed for normal-form games only; in the worst scenario, its regret bound will depend on the size of the pure strategy set, which is huge under our settings (i.e., $|A|^{\mathsf{S}}$).

In this section, we provide a no-policy regret algorithm: MDP-Online Oracle Expert (MDP-OOE). It achieves the regret bound that only depends on the size of NE support rather than the size of the game. We start by presenting the small NE support size assumption.

Assumption 3 (Small Support Size of NE) *Let $(\boldsymbol{d}_{\boldsymbol{\pi}^*}, \boldsymbol{l}^*)$ be a Nash equilibrium of the game of size $|A|^{\mathsf{S}} \times L$. We assume the support size of $(\boldsymbol{d}_{\boldsymbol{\pi}^*}, \boldsymbol{l}^*)$ is smaller than the game size:* $\max\big(|\operatorname{supp}(\boldsymbol{d}_{\boldsymbol{\pi}^*})|, |\operatorname{supp}(\boldsymbol{l}^*)|\big) < \min(|A|^{\mathsf{S}}, L)$.

Note that the assumption of the small support size of NE holds in many real-world games [20, 34–37]. In addition, we prove that such an assumption also holds in cases where the loss vectors $[\boldsymbol{l}_1, ..., \boldsymbol{l}_L]$ are sampled from a continuous distribution and the size of the loss vector set L is small compared to the agent's pure strategy set, that is, $|A|^{\mathsf{S}} \gg L$, thus further justifying the generality of this assumption.

Lemma 3 *Suppose that all loss functions are sampled from a continuous distribution and the size of the loss function set is small compared to the agent's pure strategy set (i.e., $|A|^{\mathsf{S}} \gg L$). Let $(\boldsymbol{d}_{\boldsymbol{\pi}^*}, \boldsymbol{l}^*)$ be a Nash equilibrium of the game of size $|A|^{\mathsf{S}} \times L$. Then we have:*

$$\max\big(|\operatorname{supp}(\boldsymbol{d}_{\boldsymbol{\pi}^*})|, |\operatorname{supp}(\boldsymbol{l}^*)|\big) \leq L.$$

Proof Within the set of all zero-sum games, the set of zero-sum games with non-unique equilibrium has Lebesgue measure zero [11]. Thus, if the loss function's entries are sampled from a continuous distribution, then with probability one, the game has a unique NE. Following the Theorem 1 in [38] for games with unique NE, we have:

$$|\operatorname{supp}(\boldsymbol{d}_{\boldsymbol{\pi}^*})| = |\operatorname{supp}(\boldsymbol{l}^*)|.$$

We also note that the support size of the NE can not exceed the size of the game:

$$|\operatorname{supp}(\boldsymbol{d}_{\boldsymbol{\pi}^*})| \leq |A|^{\mathsf{S}}; \quad |\operatorname{supp}(\boldsymbol{l}^*)| \leq L.$$

Thus we have:

$$\max\big(|\operatorname{supp}(\boldsymbol{d}_{\boldsymbol{\pi}^*})|, |\operatorname{supp}(\boldsymbol{l}^*)|\big) = |\operatorname{supp}(\boldsymbol{l}^*)| \leq L.$$

□

Since the pure strategy set of the adversary L is much smaller compared to the pure strategy set of the agent $|A|^{|S|}$, the support size of NE will highly likely be smaller compared to the size of the agent's strategy set. Thus the agent can exploit this extra information to achieve better performance.

We now present the MDP-Online Oracle Expert (MDP-OOE) algorithm as follows. MDP-OOE maintains a set of effective strategy A_t^s in each state. In each iteration, the best response with respect to the average loss function will be calculated. If all the actions in the best response are included in the current effective strategy set A_t^s for each state, then the algorithm continues with the current set A_t^s in each state. Otherwise, the algorithm updates the set of effective strategies in steps 8 and 9 of Algorithm 2. We define the period of consecutive iterations as one *time window* T_i in which the set of effective strategy A_t^s stays fixed, i.e., $T_i := \left\{ t \mid |A_t^s| = i \right\}$. Intuitively, since both the agent and the adversary use a no-regret algorithm to play, the average strategy of both players will converge to the NE of the game. Under the small NE support size assumption, the size of the agent's effective strategy set is also small compared to the whole pure strategy set (i.e., $|A|^{\mathsf{S}}$). MDP-OOE ignores the pure strategies with poor average performance and only considers ones with high average performance. The regret bound with respect to the agent's stationary distribution is given as follows:

Algorithm 2 MDP-Online Oracle Expert

```
Initialise: Sets $A_0^1, \dots A_0^S$ of effective strategy set in each state
for $t = 1$ to $\infty$ do
    $\pi_t = BR(\bar{\boldsymbol{l}})$
    if $\pi_t(s, .) \in A_{t-1}^s$ for all $s$ then
        $A_t^s = A_{t-1}^s$ for all $s$
        Using the expert algorithm $B_s$ with effective strategy set $A_t^s$ and the feedback $Q_{\pi_t, \boldsymbol{l}_t}(s, .)$
    else if there exists $\pi_t(s, .) \notin A_{t-1}^s$ then
        $A_t^s = A_{t-1}^s \cup \pi_t(s, .)$   if $\pi_t(s, .) \notin A_{t-1}^s$
        $A_t^s = A_{t-1}^s \cup a$   if $\pi_t(s, .) \in A_{t-1}^s$ where a is randomly selected from the set $A/A_{t-1}^s$.
        Reset the expert algorithm $B_s$ with effective strategy set $A_t^s$ and the feedback $Q_{\pi_t, \boldsymbol{l}_t}(s, .)$
    end if
    $\bar{\boldsymbol{l}} = \sum_{i=\bar{T}_i}^{T} \boldsymbol{l}_t$
end for
```

Theorem 4 *Suppose the learning agent uses Algorithm 2, then the regret with respect to the stationary distribution will be bounded by:*

$$\sum_{t=1}^{T} \left\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \right\rangle - \left\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi} \right\rangle \leq 3\tau \left(\sqrt{2Tk\log(k)} + \frac{k\log(k)}{8} \right),$$

where k is the number of time windows.

Proof We first have:

$$\begin{aligned} \mathrm{E}_{s\sim\boldsymbol{d}_\pi}[Q_{\pi_t,\boldsymbol{l}_t}(s,\pi)] &= \mathrm{E}_{s\sim\boldsymbol{d}_\pi, a\sim\pi}[Q_{\pi_t,\boldsymbol{l}_t}(s,a)] \\ &= \mathrm{E}_{s\sim\boldsymbol{d}_\pi, a\sim\pi}[\boldsymbol{l}_t(s,a) - \eta_{\boldsymbol{l}_t}(\pi_t) + \mathrm{E}_{s'\sim P_{s,a}}[Q_{\pi_t,\boldsymbol{l}_t}(s',\pi_t)]] \\ &= \mathrm{E}_{s\sim\boldsymbol{d}_\pi, a\sim\pi}[\boldsymbol{l}_t(s,a)] - \eta_{\boldsymbol{l}_t}(\pi_t) + \mathrm{E}_{s\sim\boldsymbol{d}_\pi}[Q_{\pi_t,\boldsymbol{l}_t}(s,\pi_t)] \\ &= \eta_{\boldsymbol{l}_t}(\pi) - \eta_{\boldsymbol{l}_t}(\pi_t) + \mathrm{E}_{s\sim\boldsymbol{d}_\pi}[Q_{\pi_t,\boldsymbol{l}_t}(s,\pi_t)]. \end{aligned}$$

Thus we have:

$$\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle = \sum_{s \in S} \boldsymbol{d}_\pi(s) \left(Q_{\pi_t, \boldsymbol{l}_t}(s, \pi) - Q_{\pi_t, \boldsymbol{l}_t}(s, \pi_t) \right). \tag{9}$$

Let $T_1, T_2, ..., T_k$ be the time window that the $\mathrm{BR}(\bar{\boldsymbol{l}})$ does not change. Then in that time window, the best response to the current $\bar{\boldsymbol{l}}$ is inside the current pure strategies set in each state. In each time window, following Equation (9) we have:

$$\sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle = \sum_{s \in S} \boldsymbol{d}_\pi(s) \sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \left(Q_{\pi_t, \boldsymbol{l}_t}(s, \pi) - Q_{\pi_t, \boldsymbol{l}_t}(s, \pi_t) \right). \tag{10}$$

Since during each time window, the pure strategies A_t^s does not change, thus we have:

$$\min_{\pi \in \Pi} \sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle = \min_{\pi \in A^s_{|\bar{T}_i|}} \sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle.$$

Thus, in each state s of a time window, the agent only needs to minimize the loss with respect to the action in $A^s_{|\bar{T}_i|}$. Put it differently, the expert algorithm in each state does not need to consider all pure action in each state, but just the current effective strategy set. For a time window T_i, if the agent uses a no-regret algorithm with the current effective action set and the learning rate $\mu_t = \sqrt{8 \log(i)/t}$, then the regret in each state will be bounded by [25]:

$$3\tau \left(\sqrt{2|T_i| \log(A_t^s)} + \frac{\log(A_t^s)}{8} \right) \leq 3\tau \left(\sqrt{\text{2T\_i} \log(i)} + \frac{\log(i)}{8} \right).$$

Thus, the regret in this time interval will also be bounded by:

$$\sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle \leq 3\tau \left(\sqrt{2|T_i| \log(i)} + \frac{\log(i)}{8} \right). \tag{11}$$

Sum up from $i = 1$ to k in Inequality (11) we have:

$$\begin{aligned} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle &= \sum_{i=1}^{k} \sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle \\ \leq \sum_{i=1}^{k} 3\tau \left(\sqrt{2|T_i| \log(i)} + \frac{\log(i)}{8} \right) &\leq 3\tau \left(\sqrt{2Tk \log(k)} + \frac{k \log(k)}{8} \right). \end{aligned} \tag{12}$$

The proof is complete. □

In Algorithm 2, each time the agent updates the effective strategy set A_t^s at state s, exactly one new pure strategy is added into the effective strategy set for each state, thus the number k will be at most $|A|$. Therefore, we have the regret w.r.t the stationary distribution in the worst case will be:

$$3\tau \left(\sqrt{2T|A| \log(|A|)} + \frac{|A| \log(|A|)}{8} \right).$$

However, as shown in [20, Figure 1], the number of iterations in DO method (respectively the number of time windows in our setting) is linearly dependent on the support size of the NE, thus with Assumption 3, Algorithm 2 will be highly efficient.

Remark 1 *The regret bound in Theorem 4 will still hold in the case we consider the total average lost instead of the average lost in each time window when calculating the best response in Algorithm 2.*

Proof We prove by induction that

$$\min_{\pi\in\Pi}\sum_{t=1}^{\bar{T}_k}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi}\rangle \leq \sum_{j=1}^{k}\left[\sum_{t=\bar{T}_{j-1}+1}^{\bar{T}_j}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_j}\rangle\right],$$

where $\boldsymbol{d}_{\pi_j}$ denotes the best response in the interval $[1, \bar{T}_j]$.

For $k = 1$, the claim is obvious. Suppose the claim is true k. We then have:

$$\begin{aligned}
&\min_{\pi\in\Pi}\sum_{t=1}^{\bar{T}_{k+1}}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi}\rangle = \sum_{t=1}^{\bar{T}_{k+1}}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_{k+1}}\rangle \\
&= \sum_{t=1}^{\bar{T}_k}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_{k+1}}\rangle + \sum_{t=\bar{T}_k+1}^{\bar{T}_{k+1}}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_{k+1}}\rangle \\
&\leq \min_{\pi\in\Pi}\sum_{t=1}^{\bar{T}_k}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi}\rangle + \sum_{t=\bar{T}_k+1}^{\bar{T}_{k+1}}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_{k+1}}\rangle \\
&\leq \sum_{j=1}^{k}\left[\sum_{t=\bar{T}_{j-1}+1}^{\bar{T}_j}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_j}\rangle\right] + \sum_{t=\bar{T}_k+1}^{\bar{T}_{k+1}}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_{k+1}}\rangle \qquad (13a)\\
&= \sum_{j=1}^{k+1}\left[\sum_{t=\bar{T}_{j-1}+1}^{\bar{T}_j}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_j}\rangle\right],
\end{aligned}$$

where the inequality (13a) dues to the induction assumption. Thus, for all k we have:

$$\min_{\pi\in\Pi}\sum_{t=1}^{\bar{T}_k}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi}\rangle \leq \sum_{j=1}^{k}\left[\sum_{t=\bar{T}_{j-1}+1}^{\bar{T}_j}\langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_j}\rangle\right].$$

In other words, the Algorithm 2 will have a similar regret bound when using the best response with respect to the total average strategy of the adversary. □

Given the regret with respect to policy's stationary distribution in Theorem 4, we can now derive the regret bound of Algorithm 2 with respect to the true performance:

Theorem 5 *Suppose the agent uses Algorithm 2 in our online MDPs setting, then the regret in Equation (1) can be bounded by:*

$$R_T(\pi) = \mathcal{O}(\tau^2\sqrt{Tk\log(k)} + \sqrt{T\log(L)}).$$

The full proof is given in Appendix A. Notably, Algorithm 2 will not only reduce the regret bound in the case the number of strategies set k is small, but it also reduces the computational hardness of computing expert algorithm when the number of experts is prohibitively large.

MDP-Online Oracle Algorithm with ϵ-best response. In Algorithm 2, in each iteration the agent needs to calculate the exact best response to the average loss function $\bar{\boldsymbol{l}}$. Since calculating the exact best response is computationally hard and maybe infeasible in many situations [39], an alternative way is to consider ϵ-best response. That is, in each iteration in Algorithm 2, the agent can only access to a ϵ-best response to the average loss function, where ϵ is a predefined parameter. In this situation, we provide the regret analysis for Algorithm 2 as follows.

Theorem 6 *Suppose the agent only accesses to ϵ-best response in each iteration when following Algorithm 2. If the adversary follows a no-external regret algorithm then the average strategy of the agent and the adversary will converge to ϵ-Nash equilibrium. Furthermore, the algorithm has ϵ-regret.*

The full proof is given in Appendix A. Theorem 6 implies that by following MDP-OOE, the agent can optimise the accuracy level (in terms of ϵ) based on the data that it receives to obtain the convergence rate and regret bound accordingly.

6 Last-Round Convergence to NE in OMDPs

In this section, we investigate OMDPs where the agent not only aims to minimize the regret but also stabilize the strategies. This is motivated by the fact that changing strategies through repeated games may be undesirable (e.g., see [12, 40]). In online learning literature, minimizing regret and achieving the system's stability are often two conflict goals. That is, if all player in a system follows a no-regret algorithm (e.g., MWU, FTRL) to minimise the regret, then the dynamic of the system will become chaotic and the strategies of players will not converge in the last round [10, 12].

To achieve the goal, we start by studying the scenarios where the agent knows its NE of the game π^*. We then propose an algorithm: Last-Round Convergence in OMDPs (LRC-OMDP) that leads to last-round convergence to NE of the game in our setting. This is the first algorithm to our knowledge that achieves last-round convergence in OMDPs where only the learning agent knows the NE of the game. Notably, this goal is non-trivial to achieve. For example, if the agent keeps following the same strategy (i.e., the NE), then while the system might be stabilised (i.e., the adversary converges to the best response), yet this is still not a no-regret algorithm. Moreover, we notice that understanding the learning dynamics even when the NE is known is still challenging in the multi-agent learning domain. The AWESOME [41] and CMLeS [42] algorithms make significant efforts to achieve convergence to NE under the assumption that each agent has access to a precomputed NE strategy. Compared to these algorithms, LRC-OMDP enjoys the key benefit that it does not require the adversary to know its NE. Importantly, the adversary in our setting can be any type of strategic agent who observes the history and applies a no-regret algorithm to play, rather than being a restricted opponent such as a stationary opponent in AWESOME or a memory-bounded opponent in CMLeS.

Algorithm 3 Last-Round Convergence in OMDPs

```
Input: Current iteration $t$
Output: Strategy $\pi_t$ for the agent
for $t = 1, 2, \ldots, T$ do
    if $t = 2k - 1, k \in \mathbb{N}$ then
        $\pi_t = \pi^*$
    else if $t = 2k, k \in \mathbb{N}$ then
        $\hat{\pi}_t(s) = \text{argmin}_{a \in A} Q_{\pi^*, \boldsymbol{l}_t}(s, a) \; \forall s \in S$
        $\alpha_t = \frac{v - \eta_{\boldsymbol{l}_{t-1}}(\hat{\pi}_t)}{\beta}; \quad \boldsymbol{d}_{\pi_t} = (1 - \alpha_t)\boldsymbol{d}_{\pi^*} + \alpha_t \boldsymbol{d}_{\hat{\pi}_t}$
        Output $\pi_t$ via $\boldsymbol{d}_{\pi_t}$
    end if
end for
```

The LRC-OMDP algorithm can be described as follow. At each odd round, the agent follows the NE strategy π^* so that in the next round, the strategy of the adversary will not deviate from the current strategy. Then, at the following even round, the agent chooses a strategy such that $\boldsymbol{d}_{\pi_t}$ is a direction towards the NE strategy of the adversary. Depending on the distance between the current strategy of the adversary and its NE (which is measured by $v - \eta_{\boldsymbol{l}_{t-1}}(\hat{\pi}_t)$), the agent will choose a step size α_t such that the strategy of the adversary will approach the NE. Note here that β is a constant parameter and depends on the specific no-regret algorithm adversary follows, there is a different optimal value for β. In case where the adversary follows the MWU algorithm, we can set $\beta = 1$.

We first introduce the condition in which the system achieves stability through the following lemma:

Lemma 4 *Let π^* be the NE strategy of the agent. Then, $\boldsymbol{l}$ is the Nash Equilibrium of the adversary if the two following conditions hold:*

$$Q_{\pi^*, \boldsymbol{l}}(s, \pi^*) = \underset{\pi \in \Pi}{\text{argmin}}\, Q_{\pi^*, \boldsymbol{l}}(s, \pi) \;\; \forall s \in S \;\; \textit{and} \;\; \eta_l(\pi^*) = v.$$

Proof Using the definition of accumulated loss function Q we have

$$\begin{aligned}
&\mathbb{E}_{s \in \boldsymbol{d}_\pi}[Q_{\pi^*, \boldsymbol{l}}(s, \pi)] = \mathbb{E}_{s \in \boldsymbol{d}_\pi, a \in \pi}[Q_{\pi^*, \boldsymbol{l}}(s, a)] \\
&= \mathbb{E}_{s \in \boldsymbol{d}_\pi, a \in \pi}[\boldsymbol{l}(s, a) - \eta_l(\pi^*) + \mathbb{E}_{s' \sim P_{sa}}[Q_{\pi^*, \boldsymbol{l}}(s', \pi^*)]] \\
&= \mathbb{E}_{s \in \boldsymbol{d}_\pi, a \in \pi}[\boldsymbol{l}(s, a) - \eta_l(\pi^*)] + \mathbb{E}_{s \in \boldsymbol{d}_\pi}[Q_{\pi^*, \boldsymbol{l}}(s, \pi^*)] \\
&= \eta_l(\pi) - \eta_l(\pi^*) + \mathbb{E}_{s \in \boldsymbol{d}_\pi}[Q_{\pi^*, \boldsymbol{l}}(s, \pi^*)].
\end{aligned} \tag{14}$$

Thus we have

$$\eta_l(\pi) - \eta_l(\pi^*) = \mathbb{E}_{s \in \boldsymbol{d}_\pi}[Q_{\pi^*,\boldsymbol{l}}(s,\pi) - Q_{\pi^*,\boldsymbol{l}}(s,\pi^*).] \tag{15}$$

Since we assume that

$$\mathbb{Q}_{\pi^*,\boldsymbol{l}}(s,\pi^*) = \operatorname*{argmin}_{\pi \in \Pi} \mathbb{Q}_{\pi^*,\boldsymbol{l}}(s,\pi) \;\; \forall s \in S,$$

we have

$$\mathbb{Q}_{\pi^*,\boldsymbol{l}}(s,\pi) \geq \mathbb{Q}_{\pi^*,\boldsymbol{l}}(s,\pi^*) \;\; \forall s \in S, \pi \in \Pi. \tag{16}$$

It implies that

$$\mathbb{E}_{s \in \boldsymbol{d}_\pi}[Q_{\pi^*,\boldsymbol{l}}(s,\pi) - Q_{\pi^*,\boldsymbol{l}}(s,\pi^*)] \geq 0 \;\; \forall \pi \in \Pi. \tag{17}$$

Therefore we have

$$\eta_l(\pi) \geq \eta_l(\pi^*) \;\; \forall \pi \in \Pi. \tag{18}$$

Along with the assumption $\eta_l(\pi^*) = v$, we have the following relationship:

$$\operatorname*{argmin}_{\pi \in \Pi} \eta_l(\pi) = \eta_l(\pi^*) = v. \tag{19}$$

Now we prove that for the loss function $\boldsymbol{l}$ that satisfies Equation (19), then $\boldsymbol{l}$ is NE for the adversary. Let $(\pi^*, \boldsymbol{l}^*)$ be one of the NE of the game. Since the game we are considering is a zero-sum game, $(\pi^*, \boldsymbol{l}^*)$ satisfies the famous minimax theorem:

$$\min_{\pi \in \Pi} \max_{\boldsymbol{l}_1 \in L} \langle \boldsymbol{l}_1, \boldsymbol{d}_\pi \rangle = \max_{\boldsymbol{l}_1 \in L} \min_{\pi \in \Pi} \langle \boldsymbol{l}_1, \boldsymbol{d}_\pi \rangle = v \;\; \text{where } \langle \boldsymbol{l}, \boldsymbol{d}_\pi \rangle = \eta_l(\pi). \tag{20}$$

From Equation (19) we have

$$v = \min_{\pi \in \Pi} \langle \boldsymbol{l}, \boldsymbol{d}_\pi \rangle \leq \langle \boldsymbol{l}, \boldsymbol{d}_{\pi^*} \rangle. \tag{21}$$

Further, since $\boldsymbol{l}^*$ is the NE of the game, then we have

$$v = \langle \boldsymbol{l}^*, \boldsymbol{d}_{\pi^*} \rangle = \max_{\boldsymbol{l}_1 \in L} \langle \boldsymbol{l}_1, \boldsymbol{d}_{\pi^*} \rangle \geq \langle \boldsymbol{l}, \boldsymbol{d}_{\pi^*} \rangle. \tag{22}$$

From Inequalities (21) and (22) we have

$$v = \langle \boldsymbol{l}, \boldsymbol{d}_{\pi^*} \rangle = \min_{\pi \in \Pi} \langle \boldsymbol{l}, \boldsymbol{d}_\pi \rangle = \max_{\boldsymbol{l}_1 \in L} \langle \boldsymbol{l}_1, \boldsymbol{d}_{\pi^*} \rangle. \tag{23}$$

Thus, by definition $(\boldsymbol{l}, \pi^*)$ is the Nash equilibrium of the game. In other words, the loss function $\boldsymbol{l}$ satisfies the above assumption is the NE of the adversary. □

The above lemma implies that if there is no improvement in the Q-value function for every state and the value of the current loss function equals the value of the game, then there is last-round convergence to the NE. In situations where there is an improvement in one state, the following lemma bounds the value of a new strategy:

Lemma 5 *Assume that* $\forall \pi \in \Pi$, $\boldsymbol{d}_\pi(s) > 0$. *Then if there exists* $s \in S$ *such that*

$$Q_{\pi^*,\boldsymbol{l}_t}(s,\pi^*) > \operatorname*{argmin}_{\pi \in \Pi} Q_{\pi^*,\boldsymbol{l}_t}(s,\pi),$$

then for $\pi_{t+1}(s) = \operatorname{argmin}_{a \in A} Q_{\pi^*,\boldsymbol{l}_t}(s,a) \; \forall s \in S$:

$$\eta_{\boldsymbol{l}_t}(\pi_{t+1}) < v.$$

Proof From the minimax theorem, we have:

$$\eta_{\boldsymbol{l}_t}(\pi^*) \le \eta_{\boldsymbol{l}^*}(\pi^*) = v \ \ \forall \boldsymbol{l} \in L.$$

From the proof of Lemma 4 we have:

$$\eta_{\boldsymbol{l}_t}(\pi) - \eta_{\boldsymbol{l}_t}(\pi^*) = \mathbb{E}_{s\in \boldsymbol{d}_\pi}[Q_{\pi^*,\boldsymbol{l}_t}(s,\pi) - Q_{\pi^*,\boldsymbol{l}_t}(s,\pi^*)] \ \forall \pi \in \Pi.$$

Since the construction of the new strategy π_{t+1} we have:

$$\mathbb{E}_{s\in \boldsymbol{d}_{\pi_{t+1}}}[Q_{\pi^*,\boldsymbol{l}_t}(s,\pi_{t+1}) - Q_{\pi^*,\boldsymbol{l}_t}(s,\pi^*)] < 0,$$

thus we have:

$$\eta_{\boldsymbol{l}_t}(\pi) < \eta_{\boldsymbol{l}_t}(\pi^*) \le 0.$$

The proof is complete. □

Based on the above lemmas, we can bound the relative entropy distance between the current strategy of the adversary and a Nash equilibrium:

Lemma 6 *Assume that the adversary follows the MWU algorithm with non-increasing step size μ_t such that $\lim_{T\to\infty}\sum_{t=1}^{T}\mu_t = \infty$ and there exists $t' \in \mathbb{N}$ with $\mu_{t'} \le \frac{1}{3}$. Then we have:*

$$RE\left(\boldsymbol{l}^* \| \boldsymbol{l}_{2k-1}\right) - RE\left(\boldsymbol{l}^* \| \boldsymbol{l}_{2k+1}\right) \ge \frac{1}{2}\mu_{2k}\alpha_{2k}(v - \eta_{\boldsymbol{l}_{2k-1}}(\hat{\pi}_{2k})) \ \ \forall k \in \mathbb{N}: \ \ 2k \ge t'.$$

Proof Using the definition of relative entropy we have:

$$\begin{aligned}
&\mathrm{RE}\left(\boldsymbol{l}^* \| \boldsymbol{l}_{2k-1}\right) - \mathrm{RE}\left(\boldsymbol{l}^* \| \boldsymbol{l}_{2k+1}\right) \\
&= \left(\mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k+1}) - \mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k})\right) + \left(\mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k}) - \mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k-1})\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{l}^*(i)\log\left(\frac{\boldsymbol{l}^*(i)}{\boldsymbol{l}_{2k+1}(i)}\right) - \sum_{i=1}^{n} \boldsymbol{l}^*(i)\log\left(\frac{\boldsymbol{l}^*(i)}{\boldsymbol{l}_{2k}(i)}\right)\right) + \\
&\quad \left(\sum_{i=1}^{n} \boldsymbol{l}^*(i)\log\left(\frac{\boldsymbol{l}^*(i)}{\boldsymbol{l}_{2k}(i)}\right) - \sum_{i=1}^{n} \boldsymbol{l}^*(i)\log\left(\frac{\boldsymbol{l}^*(i)}{\boldsymbol{l}_{2k-1}(i)}\right)\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{l}^*(i)\log\left(\frac{\boldsymbol{l}_{2k}(i)}{\boldsymbol{l}_{2k+1}(i)}\right)\right) + \left(\sum_{i=1}^{n} \boldsymbol{l}^*(i)\log\left(\frac{\boldsymbol{l}_{2k-1}(i)}{\boldsymbol{l}_{2k}(i)}\right)\right).
\end{aligned}$$

Following the update rule of the Multiplicative Weights Update algorithm we have:

$$\begin{aligned}
&\mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k+1}) - \mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k-1}) \\
&= \left(-\mu_{2k}\langle \boldsymbol{l}^*, \boldsymbol{d}_{\pi_{2k}}\rangle + \log(Z_{2k})\right) + \left(-\mu_{2k-1}\langle \boldsymbol{l}^*, \boldsymbol{d}_{\pi_{2k}}\rangle + \log(Z_{2k-1})\right) \\
&\le \left(-\mu_{2k}v + \log\left(\sum_{i=1}^{n} \boldsymbol{l}_{2k}(i)e^{\mu_{2k}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\pi_{2k}}\rangle}\right)\right) + \left(-\mu_{2k-1}v + \log(Z_{2k-1})\right) \qquad \text{(24a)} \\
&= \left(-\mu_{2k}v + \log\left(\sum_{i=1}^{n} \boldsymbol{l}_{2k-1}(i)e^{\mu_{2k-1}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\pi_{2k-1}}\rangle}e^{\mu_{2k}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\pi_{2k}}\rangle}\right) - \log(Z_{2k-1})\right) \\
&+ \left(-\mu_{2k-1}v + \log(Z_{2k-1})\right),
\end{aligned}$$

where Inequality (24a) is due to the fact that $\langle \boldsymbol{l}^*, \boldsymbol{d}_\pi\rangle \ge v \ \forall \pi$. Thus,

$$\mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k+1}) - \mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k-1})$$

$$
\begin{aligned}
&\le \left(-\mu_{2k} v + \log\left(\sum_{i=1}^{n} \boldsymbol{l}_{2k-1}(i) e^{\mu_{2k-1}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\pi_{2k-1}}\rangle} e^{\mu_{2k}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\pi_{2k}}\rangle}\right)\right) - \mu_{2k-1} v \\
&\le \left(-\mu_{2k} v + \log\left(\sum_{i=1}^{n} \boldsymbol{l}_{2k-1}(i) e^{\mu_{2k-1} v} e^{\mu_{2k}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\pi_{2k}}\rangle}\right)\right) - \mu_{2k-1} v \qquad (25a) \\
&= -\mu_{2k} v + \log\left(\sum_{i=1}^{n} \boldsymbol{l}_{2k-1}(i) e^{\mu_{2k}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\pi_{2k}}\rangle}\right),
\end{aligned}
$$

where Inequality (25a) is the result of the inequality:

$$\langle \boldsymbol{l}, \boldsymbol{d}_{\pi^*}\rangle \le v \;\; \forall \boldsymbol{l}.$$

Now, using the update rule of Algorithm 3

$$\boldsymbol{d}_{\pi_{2k}} = (1-\alpha_{2k})\boldsymbol{d}_{\pi^*} + \alpha_{2k}\boldsymbol{d}_{\hat{\pi}_{2k}},$$

we have

$$
\begin{aligned}
&\mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k+1}) - \mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k-1}) \\
&\le -\mu_{2k} v + \log\left(\sum_{i=1}^{n} \boldsymbol{l}_{2k-1}(i) e^{\mu_{2k}((1-\alpha_{2k})\langle \boldsymbol{e}_i, \boldsymbol{d}_{\pi^*}\rangle + \alpha_{2k}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\hat{\pi}_{2k}}\rangle)}\right) \\
&\le -\mu_{2k}\alpha_{2k} v + \log\left(\sum_{i=1}^{n} \boldsymbol{l}_{2k-1}(i) e^{\mu_{2k}\alpha_{2k}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\hat{\pi}_{2k}}\rangle}\right).
\end{aligned}
$$

Denote $f(\boldsymbol{l}_{2k-1}) = \langle \boldsymbol{l}_{2k-1}, \boldsymbol{d}_{\hat{\pi}_{2k}}\rangle$, we then have

$$
\begin{aligned}
&\mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k+1}) - \mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k-1}) \\
&\le -\mu_{2k}\alpha_{2k} v + \log\left(\sum_{i=1}^{n} \boldsymbol{l}_{2k-1}(i) e^{\mu_{2k}\alpha_{2k}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\hat{\pi}_{2k}}\rangle}\right) \\
&= \mu_{2k}\alpha_{2k}(1-v) + \log\left(\sum_{i=1}^{n} \boldsymbol{l}_{2k-1}(i) e^{-\mu_{2k}\alpha_{2k}(1-\langle \boldsymbol{e}_i, \boldsymbol{d}_{\hat{\pi}_{2k}}\rangle)}\right) \qquad (27a) \\
&\le \mu_{2k}\alpha_{2k}(1-v) + \log\left(\sum_{i=1}^{n} \boldsymbol{l}_{2k-1}(i)(1-(1-e^{-\mu_{2k}\alpha_{2k}})(1-\langle \boldsymbol{e}_i, \boldsymbol{d}_{\hat{\pi}_{2k}}\rangle))\right) \qquad (27b) \\
&= \mu_{2k}\alpha_{2k}(1-v) + \log\left(1-(1-e^{-\mu_{2k}\alpha_{2k}})(1-\langle \boldsymbol{l}_{2k-1}, \boldsymbol{d}_{\hat{\pi}_{2k}}\rangle)\right) \\
&\le \mu_{2k}\alpha_{2k}(1-v) - (1-e^{-\mu_{2k}\alpha_{2k}})(1-\langle \boldsymbol{l}_{2k-1}, \boldsymbol{d}_{\hat{\pi}_{2k}}\rangle) \qquad (27c) \\
&= \mu_{2k}\alpha_{2k}(1-v) - (1-e^{-\mu_{2k}\alpha_{2k}})(1-f(\boldsymbol{l}_{2k-1})),
\end{aligned}
$$

Equation (27a) is created by adding and subtracting $\mu_{2k}\alpha_{2k}$ on the first and second terms.

Inequalities ($27b$, $27c$) are due to

$$\beta^x \le 1-(1-\beta)x \quad \forall \beta \ge 0 \; \boldsymbol{l} \in [0,1] \text{ and } \log(1-x) \le -x \;\; \forall x < 1.$$

We can develop Inequality (27c) further as

$$
\begin{aligned}
&\mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k+1}) - \mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k-1}) \\
&\le \mu_{2k}\alpha_{2k}(1-v) - \left(1-e^{-\mu_{2k}\alpha_{2k}}\right)(1-f(\boldsymbol{l}_{2k-1})) \\
&\le \mu_{2k}\alpha_{2k}(1-v) - \left(1-\left(1-\mu_{2k}\alpha_{2k}+\frac{1}{2}(\mu_{2k}\alpha_{2k})^2\right)\right)(1-f(\boldsymbol{l}_{2k-1})) \qquad (28a)
\end{aligned}
$$

$$
\begin{aligned}
&= \mu_{2k}\alpha_{2k}(f(\boldsymbol{l}_{2k-1}) - v) + \frac{1}{2}(\mu_{2k}\alpha_{2k})^2(1 - f(\boldsymbol{l}_{2k-1})) \\
&\leq \mu_{2k}\alpha_{2k}(f(\boldsymbol{l}_{2k-1}) - v) + \frac{1}{2}\mu_{2k}\alpha_{2k}\mu_{2k}\frac{v - f(\boldsymbol{l}_{2k-1})}{\beta}(1 - f(\boldsymbol{l}_{2k-1})) \qquad (28b) \\
&\leq \mu_{2k}\alpha_{2k}(f(\boldsymbol{l}_{2k-1}) - v) + \frac{1}{2}\mu_{2k}\alpha_{2k}\ (v - f(\boldsymbol{l}_{2k-1})) \qquad (28c) \\
&= -\frac{1}{2}\mu_{2k}\alpha_{2k}(v - f(\boldsymbol{l}_{2k-1})) \leq 0.
\end{aligned}
$$

Here, Inequality (28*a*) is due to $e^x \leq 1 + x + \frac{1}{2}x^2 \quad \forall \boldsymbol{l} \in [-\infty, 0]$, Inequality (28b) comes from the definition of α_t:

$$\alpha_t = \frac{v - f(\boldsymbol{l}_{2k-1})}{\beta},\ \beta \geq 1 - f(\boldsymbol{l}),\ f(\boldsymbol{l}_{2k-1}) \leq 1.$$

Finally, Inequality (28*c*) comes from the choice of k at the beginning of the proof, i.e., $\mu_{2k} \leq 1$. □

we finally reach the last-round convergence of LRC-MDP in Algorithm 3.

Theorem 7 *Assume that the adversary follows the MWU algorithm with non-increasing step size μ_t such that $\lim_{T\to\infty}\sum_{t=1}^{T}\mu_t = \infty$ and there exists $t' \in \mathbb{N}$ with $\mu_{t'} \leq \frac{1}{3}$. If the agent follows Algorithm 3 then there exists a Nash equilibrium $\boldsymbol{l}^*$ for the adversary such that $lim_{t\to\infty}\boldsymbol{l}_t = \boldsymbol{l}^*$ almost everywhere and $lim_{t\to\infty}\pi_t = \pi^*$.*

Proof We focus on the regret analysis with respect to the stationary distribution $\boldsymbol{d}_{\pi_t}$. Let $\boldsymbol{l}^*$ be a minimax equilibrium strategy of the adversary ($\boldsymbol{l}^*$ may not be unique). Following the above Lemma, for all $k \in \mathbb{N}$ such that $2k \geq t'$, we have

$$\mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k+1}) - \mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k-1}) \leq -\frac{1}{2}\mu_{2k}\alpha_{2k}(v - f(\boldsymbol{l}_{2k-1})), \qquad (29)$$

where we denote $f(\boldsymbol{l}_{2k-1}) = \langle \boldsymbol{l}_{2k-1}, \boldsymbol{d}_{\hat{\pi}_{2k}} \rangle$. Thus, the sequence of relative entropy $\mathrm{RE}(\boldsymbol{l}^* \| \boldsymbol{l}_{2k-1})$ is non-increasing for all $k \geq \frac{t'}{2}$. As the sequence is bounded below by 0, it has a limit for any minimax equilibrium strategy $\boldsymbol{l}^*$. Since t' is a finite number and $\sum_{t=1}^{\infty}\mu_t = \infty$, we have $\sum_{t=t'}^{\infty}\mu_t = \infty$. Thus,

$$\lim_{T\to\infty}\sum_{k=\left\lceil \frac{t'}{2} \right\rceil}^{T}\mu_{2k} = \infty.$$

We will prove that $\forall \epsilon > 0,\ \exists h \in \mathbb{N}$ such that when the agent follows Algorithm 3 and the adversary follows MWU algorithm, the adversary will play strategy $\boldsymbol{l}_h$ at round h and $v - f(\boldsymbol{l}_h) \leq \epsilon$. In particular, we prove this by contradiction. That is, suppose that $\exists \epsilon > 0$ such that $\forall h \in \mathbb{N},\ v - f(\boldsymbol{l}_h) > \epsilon$. Then $\forall k \in \mathbb{N}$,

$$\alpha_{2k}(v - f(\boldsymbol{l}_{2k-1})) = \frac{(v - f(\boldsymbol{l}_{2k-1}))^2}{\beta} > \frac{\epsilon^2}{\beta}.$$

Let k vary from $\left\lceil \frac{t'}{2} \right\rceil$ to T in Equation (29). By summing over k, we obtain:

$$\begin{aligned}\mathrm{RE}(\boldsymbol{l}^*\|\boldsymbol{l}_{2T+1}) &\leq \mathrm{RE}(\boldsymbol{l}^*\|\boldsymbol{l}_{t'}) - \frac{1}{2}\sum_{k=\left\lceil \frac{t'}{2} \right\rceil}^{T} \mu_{2k}\alpha_{2k}(v - f(\boldsymbol{l}_{2k-1})) \\ &\leq \mathrm{RE}(\boldsymbol{l}^*\|\boldsymbol{l}_{t'}) - \frac{1}{2}\frac{e^2}{\beta}\sum_{k=\left\lceil \frac{t'}{2} \right\rceil}^{T} \mu_{2k}.\end{aligned}$$

Since $\lim_{T\to\infty}\sum_{k=\left\lceil \frac{t'}{2} \right\rceil}^{T}\mu_{2k} = \infty$ and $\mathrm{RE}(\boldsymbol{l}^*\|\boldsymbol{l}_{T+1}) \geq 0$, it contradicts our assumption about $\forall h \in \mathbb{N},\ v - f(\boldsymbol{l}_h) > \epsilon$.

Now, we take a sequence of $\epsilon_k > 0$ such that $\lim_{k\to\infty}\epsilon_k = 0$. Then for each k, there exists $\boldsymbol{l}_{t_k} \in \Delta_n$ such that $v - \epsilon_k \leq f(\boldsymbol{l}_{t_k}) \leq v$. As Δ_n is a compact set and $\boldsymbol{l}_{t_k}$ is bounded then following the Bolzano-Weierstrass theorem, there is a convergence subsequence $\boldsymbol{l}_{\bar{t}_k}$. The limit of that sequence, $\bar{\boldsymbol{l}}^*$, is a minimax equilibrium strategy of the row player (since $f(\bar{\boldsymbol{l}}^*) = f(\lim_{k\to\infty}\boldsymbol{l}_{\bar{t}_k}) = \lim_{k\to\infty} f(\boldsymbol{l}_{\bar{t}_k}) = v$). Combining with the fact that $\mathrm{RE}(\bar{\boldsymbol{l}}^*\|\boldsymbol{l}_{2k-1})$ is non-increasing for $k \geq \left\lceil \frac{t'}{2} \right\rceil$ and $\mathrm{RE}(\bar{\boldsymbol{l}}^*\|\bar{\boldsymbol{l}}^*) = 0$, we have $\lim_{k\to\infty}\mathrm{RE}(\bar{\boldsymbol{l}}^*\|\boldsymbol{l}_{2k-1}) = 0$. We also note that

$$\begin{aligned}\mathrm{RE}(\bar{\boldsymbol{l}}^*\|\boldsymbol{l}_{2k}) - \mathrm{RE}(\bar{\boldsymbol{l}}^*\|\boldsymbol{l}_{2k-1}) &= -\mu_{2k-1}\langle \bar{\boldsymbol{l}}^*, \boldsymbol{d}_{\pi_{2k-1}}\rangle + \log\left(\sum_{i=1}^{n}\boldsymbol{l}_{2k-1}(i)e^{\mu_{2k-1}\langle \boldsymbol{e}_i, \boldsymbol{d}_{\pi^*}\rangle}\right) \\ &\leq -\mu_{2k-1}v + \log\left(\sum_{i=1}^{n}\boldsymbol{l}_{2k-1}(i)e^{\mu_{2k-1}v}\right) = 0,\end{aligned}$$

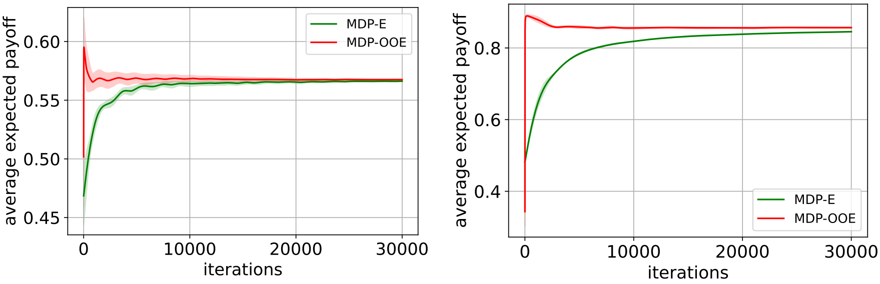
following the fact that $\langle \bar{\boldsymbol{l}}^*, \boldsymbol{d}_\pi\rangle \geq v$ for all $\pi \in \Pi$ and $\langle \boldsymbol{l}, \boldsymbol{d}_{\pi^*}\rangle \leq v$ for all $\boldsymbol{l}$. Thus, we have $\lim_{k\to\infty}\mathrm{RE}(\bar{\boldsymbol{l}}^*\|\boldsymbol{l}_{2k}) = 0$ as well. Subsequently, $\lim_{t\to\infty}\mathrm{RE}(\bar{\boldsymbol{l}}^*\|\boldsymbol{l}_t) = 0$, which concludes the proof. □

The Algorithm 3 also applies in the situations where the adversary follows different learning dynamics such as Follow the Regularized Leader or linear MWU [12]. In these situations, Algorithm 3 requires adapting the constant parameter β so that the convergence result still holds. Since both the agent and the adversary converge to a NE, the NE is also the best fixed strategy in hindsight. Consequently, LRC-OMDP is also a no-regret algorithm where the regret bound depends on the convergence rate to the NE.

7 Experiment

In this section, we aim to demonstrate the effectiveness of our practical use algorithm MDP-OOE compared to the well-known MDP-E algorithm [3].

We consider random games in which the entries of the transition matrix are first sampled from a uniform distribution $\mathbf{U}(0,1)$, then follow the normalization. Similarly, the entries of the loss vectors from the adversary $\boldsymbol{l}_t$ are also sampled from a uniform distribution $\mathbf{U}(0,1)$. Following Lemma 3, by fixing a small number of loss vectors L, we can bound the size of the Nash support of our games. Thus, we fix the number of loss vectors $L = 3$ and consider different games with the number of actions in each state in the set $[3, 100, 500]$.



(a) 3 actions in each state

(b) 100 actions in each state

Fig. 1: Performance comparisons in average payoff in random games

We then run MDP-E and MDP-OOE against the same opponent following a no-regret MWU algorithm and measure the average payoff of the two algorithms [4]. For each setting, we run 5 seeds where each seed considers an MWU adversary with a different starting strategy.

As we can see in Figure 1, MDP-OOE outperforms MDP-E in all games we consider. The difference in performance between the MDP-OOE and MDP-E becomes more significant when a larger action set is considered (See Figure B2 in the Appendix). Intuitively, since the performance of MDP-OOE only depends on the support size of the NE, a large size of the action set will not affect its performance. In contrast, a large action set will significantly affect the performance of MDP-E as it considers the whole action set in the strategy update. We observe a similar performance in other settings with a different number of loss vectors as shown in Figure B1 in the Appendix B. The advantage of MDP-OOE in term of average payoff over MDP-E match our expectation as the support size of the NEs in these games are much smaller than the action set by design. Interestingly, even when the action set is small (i.e., $\mathsf{A} = 3$), MDP-OOE still outperforms MDP-E in our experiments.

Note here that since we consider two-player zero-sum games and both the agent and the opponent follow no-regret algorithms, the average payoff of MDP-OOE and MDP-E will eventually converge to the value of the game, as shown in Figure 1.

8 Conclusion

In this paper, we have studied a novel setting in Online Markov Decision Processes where the loss function is chosen by a non-oblivious strategic adversary who follows a no-external regret algorithm. In this new setting, we then revisited the MDP-E algorithm and provided a sublinear regret bound $\mathcal{O}(\sqrt{T\log(L)} + \tau^2\sqrt{T\log(|A|)})$. We suggested a new algorithm of MDP-OOE

[4] W.l.o.g, we consider the payoff (i.e., -the loss) for the agent in our experiments so that the agent aims to maximize the payoff.

that achieves the policy regret of $\mathcal{O}(\sqrt{T\log(L)} + \tau^2\sqrt{Tk\log(k)})$ where the regret does not depend the size of strategy set $|A|$ but the effective strategy set k. Finally, in tackling non-convergence property of no-regret algorithms in self-plays, we provided the LRC-OMDP algorithm for the agent that leads to the first-known result of the last-round convergence to a NE against the strategic adversary.

Our paper offers several interesting directions for future research. Firstly, while MDP-OOE achieves better performance both in theory (when k is small) and in experiments compared to MDP-E, it still requires the calculation of best response oracles in each iteration, thus increasing its time complexity. Even though Theorem 6 provides an alternative to using ϵ-best response, further research can be done to improve the efficiency of MDP-OOE algorithm with regards to the best response oracle. Secondly, LRC-OMDP provides the first last-round convergence to a NE against a strategic adversary, yet it requires a strong assumption of knowing the agent's NE before playing. While this assumption is common in literature [41, 42], relaxing this assumption could further enhance the application of LRC-OMDP algorithm in practical situations. Thirdly, since our experiment section only serves as a validation test for the performance between MDP-E and MDP-OOE, further experiments on real-world and large-size games are needed to demonstrate the efficiency of both MDP-E and MDP-OOE algorithms against the strategic adversary. Finally, our paper provides a new direction to tackle the hardness result of playing against the non-oblivious adversary. In the future, apart from the strategic adversary, other important types of adversary should be considered to study relevant regret bound and convergence properties.

Appendix A Proofs

We provide the following lemmas and proposition:

Lemma 7 (Lemma 3.3 in [3]) *For all loss function $\boldsymbol{l}$ in $[0,1]$ and policies π, $Q_{\boldsymbol{l},\pi}(s,a) \leq 3\tau$.*

Lemma 8 (Lemma 1 from [16]) *Consider a uniformly ergodic OMDPs with mixing time τ with losses $\boldsymbol{l}_t \in [0,1]^{\boldsymbol{d}}$. Then, for any $T > 1$ and policy π with stationary distribution $\boldsymbol{d}_\pi$, it holds that*

$$\sum_{t=1}^{T} |\langle \boldsymbol{l}_t, \boldsymbol{d}_\pi - \boldsymbol{v}_t^\pi \rangle| \leq 2\tau + 2.$$

This lemma guarantees that the performance of a policy's stationary distribution is similar to the actual performance of the policy in the case of a fixed policy.

In the other case of non-fixed policy, the following lemma bound the performance of policy's stationary distribution of algorithm A with the actual performance:

Lemma 9 (Lemma 5.2 in [3]) *Let $\pi_1, \pi_2, \ldots$ be the policies played by MDP-E algorithm $\mathcal{A}$ and let $\tilde{\boldsymbol{d}}_{\mathcal{A},t}$, $\tilde{\boldsymbol{d}}_{\pi_t} \in [0,1]^{\mathsf{S}}$ be the stationary state distribution. Then,*

$$\|\tilde{\boldsymbol{d}}_{\mathcal{A},t} - \tilde{\boldsymbol{d}}_{\pi_t}\|_1 \leq 2\tau^2 \sqrt{\frac{\log(|A|)}{t}} + 2e^{-t/\tau}.$$

From the above lemma, since the policy's stationary distribution is a combination of stationary state distribution and the policy's action in each state, it is easy to show that:

$$\|\boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\|_1 \leq \|\tilde{\boldsymbol{d}}_{\mathcal{A},t} - \tilde{\boldsymbol{d}}_{\pi_t}\|_1 \leq 2\tau^2 \sqrt{\frac{\log(|A|)}{t}} + 2e^{-t/\tau}.$$

Proposition 8 *For the MWU algorithm [8] with appropriate μ_t, we have:*

$$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{l}_t(\pi_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{l}_t(\pi)\right] \leq M\sqrt{\frac{T\log(n)}{2}},$$

where $\|\boldsymbol{l}_t(.)\| \leq M$. Furthermore, the strategy $\boldsymbol{\pi}_t$ does not change quickly: $\|\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t+1}\| \leq \sqrt{\frac{\log(n)}{t}}$.

Proof For a fixed T, if the loss function satisfies $\boldsymbol{l}_t(.)\| \leq 1$ then by setting $\mu_t = \sqrt{\frac{8\log(n)}{T}}$, following Theorem 2.2 in [25] we have:

$$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{l}_t(\pi_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{l}_t(\pi)\right] \leq 1\sqrt{\frac{T\log(n)}{2}}. \tag{A1}$$

Thus, in the case where $\boldsymbol{l}_t(.)\| \leq M$, by scaling up both sides by M in Equation (A1) we have the first result of the Proposition 8. For the second part, follow the updating rule of MWU we have:

$$\begin{aligned}\pi_{t+1}(i) - \pi_t(i) &= \pi_t(i)\left(\frac{\exp(-\mu_t \boldsymbol{l}_t(\boldsymbol{a}^i))}{\sum_{i=1}^n \boldsymbol{\pi}_t(i)\exp(-\mu_t \boldsymbol{l}_t(\boldsymbol{a}^i))} - 1\right) \\ &\approx \pi_t(i)\left(\frac{1-\mu_t \boldsymbol{l}_t(\boldsymbol{a}^i)}{1-\mu_t \boldsymbol{l}_t(\pi_t)} - 1\right) \qquad \text{(A2a)} \\ &= \mu_t \pi_t(i)\frac{\boldsymbol{l}_t(\pi_t) - \boldsymbol{l}_t(\boldsymbol{a}^i)}{1-\mu_t \boldsymbol{l}_t(\pi_t)} = \mathcal{O}(\mu_t),\end{aligned}$$

where we use the approximation $e^x \approx 1 + x$ for small x in Equation (A2a). Thus, the difference in two consecutive strategies π_t will be proportional to the learning rate μ_t, which is set to be $\mathcal{O}\big(\sqrt{\frac{\log(n)}{t}}\big)$. Similar result can be found in Proposition 1 in [3]. □

Theorem (Theorem 5) *Suppose the agent uses Algorithm 2 in our online MDPs setting, then the regret in Equation (1) can be bounded by:*

$$R_T(\pi) = \mathcal{O}(\tau^2\sqrt{Tk\log(k)} + \sqrt{T\log(L)}).$$

Proof First we bound the difference between the true loss and the loss with respect to the policy's stationary distribution. Following the Algorithm 2, at the start of each time interval T_i (i.e., the time interval in which the effective strategy set does not change), the learning rate needs to restart to $\mathcal{O}(\sqrt{\log(i)/t_i})$, where i denotes the number of pure strategies in the effective strategy set in the time interval T_i and t_i is relative position of the current round in that interval. Thus, following Lemma 5.2 in [3], in each time interval T_i, the difference between the true loss and the loss with respect to the policy's stationary distribution will be:

$$\begin{aligned}&\sum_{t=t_{i-1}+1}^{t_i} |\langle \boldsymbol{l}_t, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\rangle| \leq \sum_{t=t_{i-1}+1}^{t_i} \|\boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\|_1 \\ &\leq \sum_{t=1}^{T_i} 2\tau^2\sqrt{\frac{\log(i)}{t}} + 2e^{-t/\tau} \\ &\leq 4\tau^2\sqrt{T_i\log(i)} + 2(1+\tau).\end{aligned}$$

From this we have:

$$\begin{aligned}\sum_{t=1}^{T} |\langle \boldsymbol{l}_t, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\rangle| &= \sum_{i=1}^{k} \sum_{t=t_{i-1}+1}^{t_i} |\langle \boldsymbol{l}_t, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\rangle| \\ &\leq \sum_{i=1}^{k} \left(4\tau^2 \sqrt{T_i \log(i)} + 2(1+\tau)\right) \\ &\leq 4\tau^2 \sqrt{Tk \log(k)} + 2k(1+\tau).\end{aligned}$$

Following Lemma 1 from [16], we also have:

$$\sum_{t=1}^{T} |\langle \boldsymbol{l}_t, \boldsymbol{d}_\pi - \boldsymbol{v}_t^\pi \rangle| \leq 2\tau + 2.$$

Thus the regret in Equation (1) can be bounded by:

$$\begin{aligned}&R_T(\pi) \leq \left(\sum_{t=1}^{T} \langle \boldsymbol{d}_{\pi_t}, \boldsymbol{l}_t\rangle + \sum_{t=1}^{T} |\langle \boldsymbol{l}_t, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\rangle|\right) - \left(\sum_{t=1}^{T} \langle \boldsymbol{l}_t^\pi, \boldsymbol{d}_\pi\rangle - \sum_{t=1}^{T} |\langle \boldsymbol{l}_t, \boldsymbol{d}_\pi - \boldsymbol{v}_t^\pi\rangle|\right) \\ &= \left(\sum_{t=1}^{T} \langle \boldsymbol{d}_{\pi_t}, \boldsymbol{l}_t\rangle - \sum_{t=1}^{T} \langle \boldsymbol{l}_t^\pi, \boldsymbol{d}_\pi\rangle\right) + \sum_{t=1}^{T} |\langle \boldsymbol{l}_t, \boldsymbol{v}_t - \boldsymbol{d}_{\pi_t}\rangle + \sum_{t=1}^{T} |\langle \boldsymbol{l}_t, \boldsymbol{d}_\pi - \boldsymbol{v}_t^\pi\rangle| \\ &\leq 3\tau \left(\sqrt{2Tk \log(k)} + \frac{k \log(k)}{8}\right) + \frac{\sqrt{T \log(L)}}{\sqrt{2}} + 4\tau^2 \sqrt{Tk \log(k)} + 2k(1+\tau) + 2\tau + 2 \\ &= \mathcal{O}(\tau^2 \sqrt{Tk \log(k)} + \sqrt{T \log(L)}).\end{aligned} \tag{A3}$$

The proof is complete. □

Theorem (Theorem 6) *Suppose the agent only accesses to ϵ-best response in each iteration when following Algorithm 2. If the adversary follows a no-external regret algorithm then the average strategy of the agent and the adversary will converge to ϵ-Nash equilibrium. Furthermore, the algorithm has ϵ-regret.*

Proof Suppose that the player uses the Multiplicative Weights Update in Algorithm 2 with ϵ-best response. Let $T_1, T_2, \dots, T_k$ be the time window that the players does not add up a new strategy. Since we have a finite set of strategies A then k is finite. Furthermore,

$$\sum_{i=1}^{k} T_k = T.$$

In a time window T_i, the regret with respect to the best strategy in the set of strategy at time T_i is:

$$\sum_{t=\bar{T}_i}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t}\rangle - \min_{\pi \in A_{\bar{T}_i+1}} \sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi\rangle \leq 3\tau \left(\sqrt{2T_i \log(i)} + \frac{\log(i)}{8}\right), \tag{A4}$$

where $\bar{T}_i = \sum_{j=1}^{i-1} T_j$. Since in the time window T_i, the ϵ-best response strategy stays in $\Pi_{\bar{T}_i+1}$ and therefore we have:

$$\min_{\pi \in A_{\bar{T}_i+1}} \sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi\rangle - \min_{\pi \in \Pi} \sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi\rangle \leq \epsilon T_i.$$

Then, from the Equation (A4) we have:

$$\sum_{t=\bar{T}_i}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - \min_{\pi \in \Pi} \sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle \leq 3\tau \left(\sqrt{2T_i \log(i)} + \frac{\log(i)}{8} \right) + \epsilon T_i. \tag{A5}$$

Sum up the Equation (A5) for $i = 1, \dots k$ we have:

$$\sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - \sum_{i=1}^{k} \min_{\pi \in \Pi} \sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle \leq \sum_{i=1}^{k} 3\tau \left(\sqrt{2T_i \log(i)} + \frac{\log(i)}{8} \right) + \epsilon T_i$$

$$\implies \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - \min_{\pi \in \Pi} \sum_{i=1}^{k} \sum_{t=|\bar{T}_i|}^{\bar{T}_{i+1}} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle \leq \epsilon T + \sum_{i=1}^{k} 3\tau \left(\sqrt{2T_i \log(i)} + \frac{\log(i)}{8} \right) \tag{A6a}$$

$$\implies \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - \min_{\pi \in \Pi} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle \leq \epsilon T + \sum_{i=1}^{k} 3\tau \left(\sqrt{2T_i \log(i)} + \frac{\log(i)}{8} \right)$$

$$\implies \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - \min_{\pi \in \Pi} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_\pi \rangle \leq \epsilon T + 3\tau \left(\sqrt{2Tk \log(k)} + \frac{k \log(k)}{8} \right). \tag{A6b}$$

Inequality (A6a) is due to $\sum \min \leq \min \sum$. Inequality (A6b) comes from Cauchy-Schwarz inequality and Stirling' approximation. Using Inequality (A6b), we have:

$$\min_{\pi \in \Pi} \langle \bar{\boldsymbol{l}}, \boldsymbol{d}_\pi \rangle \geq \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - 3\tau \left(\sqrt{\frac{2k \log(k)}{T}} + \frac{k \log(k)}{8T} \right) - \epsilon. \tag{A7}$$

Since the adversary follows a no-regret algorithm, we have:

$$\begin{aligned} \max_{\boldsymbol{l} \in \Delta_L} \sum_{t=1}^{T} \langle \boldsymbol{l}, \boldsymbol{d}_{\pi_t} \rangle - \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle &\leq \sqrt{\frac{T}{2}} \sqrt{\log(L)} \\ \implies \max_{\boldsymbol{l} \in \Delta_L} \sum_{t=1}^{T} \langle \boldsymbol{l}, \bar{\boldsymbol{d}_\pi} \rangle &\leq \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle + \sqrt{\frac{\log(L)}{2T}}. \end{aligned} \tag{A8}$$

Using the Inequalities (A7) and (A8) we have:

$$\begin{aligned} \langle \bar{\boldsymbol{l}}, \bar{\boldsymbol{d}_\pi} \rangle \geq \min_{\pi \in \Pi} \langle \bar{\boldsymbol{l}}, \boldsymbol{d}_\pi \rangle &\geq \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle - 3\tau \left(\sqrt{\frac{2k \log(k)}{T}} + \frac{k \log(k)}{8T} \right) - \epsilon \\ &\geq \max_{\boldsymbol{l} \in \Delta_L} \sum_{t=1}^{T} \langle \boldsymbol{l}, \bar{\boldsymbol{d}_\pi} \rangle - \sqrt{\frac{\log(L)}{2T}} - 3\tau \left(\sqrt{\frac{2k \log(k)}{T}} + \frac{k \log(k)}{8T} \right) - \epsilon. \end{aligned}$$

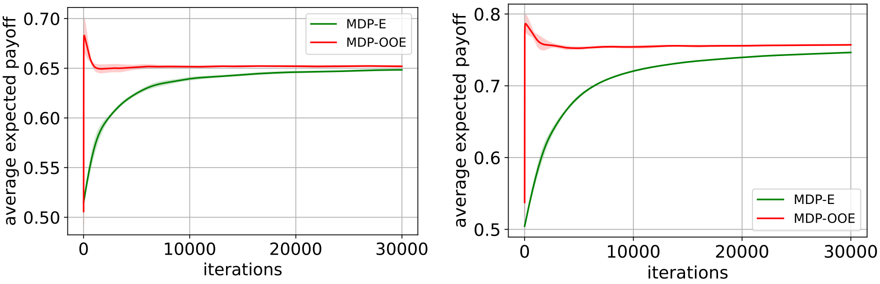
Similarly, we also have:

$$\begin{aligned} \langle \bar{\boldsymbol{l}}, \bar{\boldsymbol{d}_\pi} \rangle \leq \max_{\boldsymbol{l} \in \Delta_L} \sum_{t=1}^{T} \langle \boldsymbol{l}, \bar{\boldsymbol{d}_\pi} \rangle &\leq \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{l}_t^{\pi_t}, \boldsymbol{d}_{\pi_t} \rangle + \sqrt{\frac{\log(L)}{2T}} \\ &\leq \min_{\pi \in \Pi} \langle \bar{\boldsymbol{l}}, \boldsymbol{d}_\pi \rangle + 3\tau \left(\sqrt{\frac{2k \log(k)}{T}} + \frac{k \log(k)}{8T} \right) + \epsilon. \end{aligned}$$

Take the limit $T \to \infty$, we then have:

$$\max_{\boldsymbol{l} \in \Delta_L} \sum_{t=1}^{T} \langle \boldsymbol{l}, \bar{\boldsymbol{d}_\pi} \rangle - \epsilon \leq \langle \bar{\boldsymbol{l}}, \bar{\boldsymbol{d}_\pi} \rangle \leq \min_{\pi \in \Pi} \langle \bar{\boldsymbol{l}}, \boldsymbol{d}_\pi \rangle + \epsilon.$$

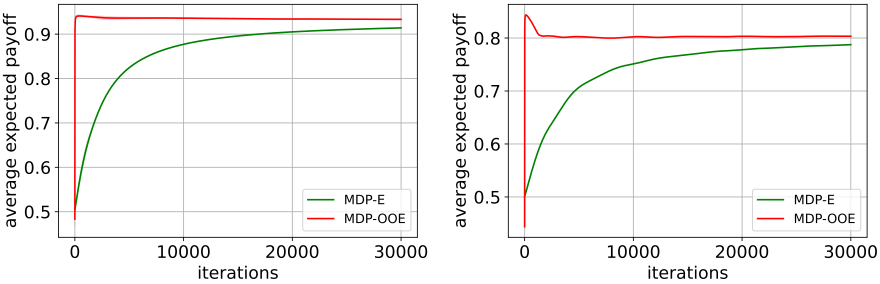
Thus $(\bar{\boldsymbol{l}}, \bar{\boldsymbol{d}_\pi})$ is the ϵ-Nash equilibrium of the game. □



(a) 7 actions in each state

(b) 100 actions in each state

Fig. B1: Performance comparisons in average payoff in random games with $L = 7$



(a) 500 actions in each state with opponent's pure strategies $L = 3$

(b) 500 actions in each state with opponent' pure strategies $L = 7$

Fig. B2: Performance comparisons in average payoff in random games

Appendix B Experiments

We provide further experiment results to demonstrate the performance of MDP-OOE and MDP-E.

In Figure B1, by considering the different number of loss vectors ($L = 7$), we test whether the performance difference between MDP-OOE and MDP-E is consistent with regard to the number of loss vectors. As we can see in Figure B1, MDP-OOE also outperforms MDP-E with the number of loss functions $L = 7$. The result further validates the advantage of MDP-OOE over MDP-E in the setting of a small support size of the NE.

In Figure B2, we consider a larger set of agent's action in each state ($A = 500$). As we can see in Figure B2, the difference in performance between MDP-OOE and MDP-E becomes more significant when a larger action set is considered in both cases when $L = 3$ and $L = 7$, as expected by our theoretical results.

References

[1] Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT press, Massachusetts (2018)

[2] Laurent, G.J., Matignon, L., Fort-Piat, L., *et al.*: The world of independent learners is not markovian. International Journal of Knowledge-based and Intelligent Engineering Systems **15**(1), 55–64 (2011)

[3] Even-Dar, E., Kakade, S.M., Mansour, Y.: Online markov decision processes. Mathematics of Operations Research **34**(3), 726–736 (2009)

[4] Dick, T., Gyorgy, A., Szepesvari, C.: Online learning in markov decision processes with changing cost sequences. In: ICML, pp. 512–520 (2014)

[5] Neu, G., Antos, A., György, A., Szepesvári, C.: Online markov decision processes under bandit feedback. In: NeurIPS, pp. 1804–1812 (2010)

[6] Neu, G., Olkhovskaya, J.: Online learning in mdps with linear function approximation and bandit feedback. arXiv e-prints, 2007 (2020)

[7] Yang, Y., Wang, J.: An overview of multi-agent reinforcement learning from game theoretical perspective. arXiv preprint arXiv:2011.00583 (2020)

[8] Freund, Y., Schapire, R.E.: Adaptive game playing using multiplicative weights. Games and Economic Behavior **29**(1-2), 79–103 (1999)

[9] Shalev-Shwartz, S., *et al.*: Online learning and online convex optimization. Foundations and trends in Machine Learning **4**(2), 107–194 (2011)

[10] Mertikopoulos, P., Papadimitriou, C., Piliouras, G.: Cycles in adversarial regularized learning. In: Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 2703–2717 (2018). SIAM

[11] Bailey, J.P., Piliouras, G.: Multiplicative weights update in zero-sum games. In: Proceedings of the 2018 ACM Conference on Economics and Computation, pp. 321–338 (2018)

[12] Dinh, L.C., Nguyen, T.-D., Zemhoho, A.B., Tran-Thanh, L.: Last round convergence and no-dynamic regret in asymmetric repeated games. In: Algorithmic Learning Theory, pp. 553–577 (2021). PMLR

[13] Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., Piliouras, G.: Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In: ICLR 2019-7th International Conference on Learning Representations, pp. 1–23 (2019)

[14] Leslie, D.S., Perkins, S., Xu, Z.: Best-response dynamics in zero-sum stochastic games. Journal of Economic Theory **189**, 105095 (2020)

[15] Guan, P., Raginsky, M., Willett, R., Zois, D.-S.: Regret minimization algorithms for single-controller zero-sum stochastic games. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 7075–7080 (2016). IEEE

[16] Neu, G., György, A., Szepesvári, C., Antos, A.: Online markov decision processes under bandit feedback. IEEE Transactions on Automatic Control **59**(3), 676–691 (2013)

[17] Filar, J., Vrieze, K.: Applications and special classes of stochastic games. In: Competitive Markov Decision Processes, pp. 301–341. Springer, New York (1997)

[18] Puterman, M.L.: Markov decision processes. Handbooks in operations research and management science **2**, 331–434 (1990)

[19] McMahan, H.B., Gordon, G.J., Blum, A.: Planning in the presence of cost functions controlled by an adversary. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 536–543 (2003)

[20] Dinh, L.C., Yang, Y., Tian, Z., Nieves, N.P., Slumbers, O., Mguni, D.H., Wang, J.: Online double oracle. arXiv preprint arXiv:2103.07780 (2021)

[21] Wei, C.-Y., Hong, Y.-T., Lu, C.-J.: Online reinforcement learning in stochastic games. arXiv preprint arXiv:1712.00579 (2017)

[22] Cheung, W.C., Simchi-Levi, D., Zhu, R.: Non-stationary reinforcement learning: The blessing of (more) optimism. Available at SSRN 3397818 (2019)

[23] Yu, J.Y., Mannor, S., Shimkin, N.: Markov decision processes with arbitrary reward processes. Mathematics of Operations Research **34**(3), 737–757 (2009)

[24] Arora, R., Dekel, O., Tewari, A.: Online bandit learning against an adaptive adversary: from regret to policy regret. arXiv preprint arXiv:1206.6400 (2012)

[25] Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge university press, Cambridge (2006)

[26] Zinkevich, M., Johanson, M., Bowling, M., Piccione, C.: Regret minimization in games with incomplete information. Advances in neural

information processing systems **20**, 1729–1736 (2007)

[27] Daskalakis, C., Ilyas, A., Syrgkanis, V., Zeng, H.: Training gans with optimism. arXiv preprint arXiv:1711.00141 (2017)

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

[29] Deng, X., Li, Y., Mguni, D.H., Wang, J., Yang, Y.: On the complexity of computing markov perfect equilibrium in general-sum stochastic games. arXiv preprint arXiv:2109.01795 (2021)

[30] Tian, Y., Wang, Y., Yu, T., Sra, S.: Online learning in unknown markov games. (2020)

[31] Neumann, J.v.: Zur theorie der gesellschaftsspiele. Mathematische annalen **100**(1), 295–320 (1928)

[32] Nash, J.F., *et al.*: Equilibrium points in n-person games. Proceedings of the national academy of sciences **36**(1), 48–49 (1950)

[33] Brown, G.W.: Iterative solution of games by fictitious play. Activity analysis of production and allocation **13**(1), 374–376 (1951)

[34] Czarnecki, W.M., Gidel, G., Tracey, B., Tuyls, K., Omidshafiei, S., Balduzzi, D., Jaderberg, M.: Real world games look like spinning tops. arXiv preprint arXiv:2004.09468 (2020)

[35] Perez-Nieves, N., Yang, Y., Slumbers, O., Mguni, D.H., Wen, Y., Wang, J.: Modelling behavioural diversity for learning in open-ended games. In: International Conference on Machine Learning, pp. 8514–8524 (2021). PMLR

[36] Liu, X., Jia, H., Wen, Y., Yang, Y., Hu, Y., Chen, Y., Fan, C., Hu, Z.: Unifying behavioral and response diversity for open-ended learning in zero-sum games. arXiv preprint arXiv:2106.04958 (2021)

[37] Yang, Y., Luo, J., Wen, Y., Slumbers, O., Graves, D., Bou Ammar, H., Wang, J., Taylor, M.E.: Diverse auto-curriculum is critical for successful real-world multiagent learning systems. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, pp. 51–56 (2021)

[38] Bohnenblust, H., Karlin, S., Shapley, L.: Solutions of discrete, two-person games. Contributions to the Theory of Games **1**, 51–72 (1950)

[39] Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., *et al.*:

Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature **575**(7782), 350–354 (2019)

[40] Daskalakis, C., Panageas, I.: Last-iterate convergence: Zero-sum games and constrained min-max optimization. 10th Innovations in Theoretical Computer Science (2019)

[41] Conitzer, V., Sandholm, T.: Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. Machine Learning **67**(1-2), 23–43 (2007)

[42] Chakraborty, D., Stone, P.: Multiagent learning in the presence of memory-bounded agents. Autonomous agents and multi-agent systems **28**(2), 182–213 (2014)