# A Summary of Online Markov Decision Processes with Non-oblivious Strategic Adversary

JAAMAS Track

Le Cong Dinh
University of Southampton
United Kingdom

David Henry Mguni
Huawei R&D UK
United Kingdom

Long Tran-Thanh
University of Warwick
United Kingdom

Jun Wang
University College London
United Kingdom

Yaodong Yang
Institute for AI, Peking University
China

## 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 $O(\sqrt{T\log(L)}+\tau^2\sqrt{T\log(|A|)})$ 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 $O(\sqrt{T\log(L)}+\tau^2\sqrt{Tk\log(k)})$ 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 result to a NE. To our best knowledge, this is first work leading to the last iteration result in OMDPs.

## KEYWORDS

Non-oblivious adversary; online markov decision processes

**ACM Reference Format:**
Le Cong Dinh, David Henry Mguni, Long Tran-Thanh, Jun Wang, and Yaodong Yang. 2024. A Summary of Online Markov Decision Processes with Non-oblivious Strategic Adversary: JAAMAS Track. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

## 1 INTRODUCTION

Reinforcement Learning (RL) 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 [6]. In tackling non-stationary environments, we are interested in designing learning algorithms that can achieve no-regret guarantee [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 [4] or the bandit [7] 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) [9]. In a multi-agent system, since all agents are learning simultaneously, one agent's adaption on its strategy will make the environment *non-oblivious* from other agents' perspective. Therefore, to find the optimal strategy for each player, one must consider the strategic reactions of others rather than regarding 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.

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 [5], for example, an oligopoly with a dominant player, or ongoing interactions between industry players and an authority (e.g., a government that acts as an order-setting body). Under this setting, we study how the agent can achieve different goals such as no-policy regret and last-round convergence.

## 2 MDP-ONLINE ORACLE EXPERT ALGORITHM

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 history $\{\pi_1, \ldots, \pi_{t-1}\}$. Formally, we have OMDPs with finite state space $S$; finite action set at each state $A$; and a fixed transition model $P$. The agent's starting state, $x_1$, is distributed according to some distribution $\mu_0$ over $S$. At time $t$, given state $x_t \in S$, the agent chooses an action $a_t \in A$, then



*Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand.* © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

**Algorithm 1** MDP-Online Oracle Expert

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

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 $\Delta_L := \{\boldsymbol{l} \in \mathbb{R}^{|S||A|} | \boldsymbol{l} = \sum_{i=1}^{L} x_i \boldsymbol{l}_i,\ \sum_{i=1}^{L} x_i = 1,\ x_i \geq 0\ \forall i\}$ 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 [1, 4], meaning the agent can observe the loss function $\boldsymbol{l}_t$ after each round $t$.

The MDP-Online Oracle Expert (MDP-OOE) algorithm can be described as follows. MDP-OOE maintains a set of effective strategies $A_t^s$ 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 step 8 and 9 of Algorithm 1. 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., $T_i := \{t \mid |A_t^s| = i\}$. 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., $|A|^{|S|}$). MDP-OOE ignores the pure strategies with poor average performance and only considers ones with high average performance. The regret bound of MDP-OOE can be given as follows:

Theorem 2.1. *Suppose the agent uses Algorithm 1 in our online MDPs setting, then the policy regret with respect to the best fixed policy in hindsight can be bounded by:*

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

*where k is the number of time windows.*

In order to prove the above theorem, we first consider the regret with respect to the policy's stationary distribution. The full proof can be found in the main paper [2]. Notably, Algorithm 1 will not only reduce the regret bound in the case the number of strategy set $k$ is small, it also reduces the computational hardness of computing expert algorithm when the number of experts is prohibitively large.

**MDP-Online Oracle Algorithm with $\epsilon$-best response.** In Algorithm 1, 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 [8], an alternative way is to consider $\epsilon$-best response. That is, in each iteration in Algorithm 1, the agent can only access to a $\epsilon$-best response to the average loss function, where $\epsilon$ is a predefined parameter. In this situation, we provide the regret analysis for Algorithm 1 as follows.

Theorem 2.2. *Suppose the agent only accesses to $\epsilon$-best response in each iteration when following Algorithm 1. If the adversary follows a no-external regret algorithm then the average strategy of the agent and the adversary will converge to $\epsilon$-Nash equilibrium. Furthermore, the algorithm has $\epsilon$-regret.*

The full proof is given in Appendix A in [2]. Theorem 2.2 implies that by following MDP-OOE, the agent can optimise the accuracy level (in terms of $\epsilon$) based on the data that it receives to obtain the convergence rate and regret bound accordingly.

## 3 LAST-ROUND CONVERGENCE TO NE

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 Dinh et al. [3]). 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 [3].

The Last-Round Convergence OMDP algorithm can be described as follows. At each odd round, the agent follows the NE strategy $\pi^*$ so that in the next round, the strategy of the adversary will not deviating 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, the agent will choose a step size $\alpha_t$ such that the strategy of adversary will approach the NE. The full detail of Last-Round Convergence algorithm can be found in Algorithm 3 in [2]. The following theorem provides the convergence result for the algorithm:

Theorem 3.1. *Assume that the adversary follows the MWU algorithm with non-increasing step size $\mu_t$ 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 Last-Round Convergence (Algorithm 3 in [2]) 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 $lim_{t\to\infty} \pi_t = \pi^*$.*

The Last-Round Convergence algorithm also applies in situations where the adversary follows different learning dynamics such as Follow the Regularized Leader or linear MWU [3](i.e., see Section 6 in [2]). Since both the agent and the adversary converge to a NE, the NE is also the best fixed strategy in hindsight. Consequently, Last-Round Convergence algorithm is also a no-regret algorithm where the regret bound depends on the convergence rate to the NE.

## REFERENCES


[1] Travis Dick, Andras Gyorgy, and Csaba Szepesvari. 2014. Online learning in Markov decision processes with changing cost sequences. In *ICML*. 512–520.

[2] Le Cong Dinh, David Henry Mguni, Long Tran-Thanh, Jun Wang, and Yaodong Yang. 2023. Online Markov decision processes with non-oblivious strategic adversary. *Autonomous Agents and Multi-Agent Systems* 37, 1 (2023), 15.

[3] Le Cong Dinh, Tri-Dung Nguyen, Alain B Zemhoho, and Long Tran-Thanh. 2021. Last Round Convergence and No-Dynamic Regret in Asymmetric Repeated Games. In *Algorithmic Learning Theory*. PMLR, 553–577.

[4] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. 2009. Online Markov decision processes. *Mathematics of Operations Research* 34, 3 (2009), 726–736.

[5] Jerzy Filar and Koos Vrieze. 1997. Applications and Special Classes of Stochastic Games. In *Competitive Markov Decision Processes*. Springer, 301–341.

[6] Guillaume J Laurent, Laëtitia Matignon, Le Fort-Piat, et al. 2011. The world of independent learners is not Markovian. *International Journal of Knowledge-based and Intelligent Engineering Systems* 15, 1 (2011), 55–64.

[7] Gergely Neu and Julia Olkhovskaya. 2020. Online learning in MDPs with linear function approximation and bandit feedback. *arXiv e-prints* (2020), arXiv–2007.

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

[9] Yaodong Yang and Jun Wang. 2020. An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective. *arXiv preprint arXiv:2011.00583* (2020).