Learning in Non-Cooperative Configurable Markov Decision Processes

Giorgia Ramponi*
ETH AI Center
Zurich, Switzerland
gramponi@ethz.ch

Alberto Maria Metelli
Politecnico di Milano
Milan, Italy
albertomaria.metelli@polimi.it

Alessandro Concetti
Politecnico di Milano
Milan, Italy
alessandro.concetti@mail.polimi.it

Marcello Restelli
Politecnico di Milano
Milan, Italy
marcello.restelli@polimi.it

Abstract

The Configurable Markov Decision Process framework includes two entities: a Reinforcement Learning agent and a configurator that can modify some environmental parameters to improve the agent's performance. This presupposes that the two actors have identical reward functions. What if the configurator does not have the same intentions as the agent? This paper introduces the Non-Cooperative Configurable Markov Decision Process, a framework that allows modeling two (possibly different) reward functions for the configurator and the agent. Then, we consider an online learning problem, where the configurator has to find the best among a finite set of possible configurations. We propose two learning algorithms to minimize the configurator's expected regret, which exploit the problem's structure, depending on the agent's feedback. While a naïve application of the UCB algorithm yields a regret that grows indefinitely over time, we show that our approach suffers only bounded regret. Furthermore, we empirically validate the performance of our algorithm in simulated domains.

1 Introduction

The standard Reinforcement Learning [RL, 40] framework involves an agent whose objective is to maximize the reward collected during its interaction with the environment. However, there exist real-world scenarios in which the agent itself or an external supervisor (configurator) can *partially* modify the environment. In a car racing problem, for example, it is possible to modify the car setup to better suit the driver's needs. Recently, the Configurable Markov Decision Processes [Conf-MDPs, 29] were introduced to model these scenarios and exploit the configuration opportunities. Solving a Conf-MDP consists of simultaneously optimizing a set of environmental parameters and the agent's policy to reach the maximum expected return. In many scenarios, however, the configurator does not know the agent's reward, and their intentions are different, leading to new forms of interaction between the two actors. For instance, imagine we are the owner of a supermarket, and we have to arrange the products on the shelves. Our objective is to increase the company's final profit; on the other hand, a customer aims to spend the shortest time possible inside the supermarket and buy the indispensable products only. Since we do not know the customer reward function, the only possibility is to try different dispositions and observe the customers' reactions. What if we knew what buyers

*Work done when Giorgia Ramponi was at Politecnico di Milano.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

are most interested in? In this case, we can *strategically* decide how to position other products close to the popular ones to induce the customer in a more profitable behavior for the supermarket owner.

In this paper, we model these scenarios introducing the Non-Cooperative Markov Decision Processes (NConf-MDP). This novel framework handles the possibility of having different reward functions for the agent and the configurator. While Conf-MDP assumes that the configurator acts to help the agent to optimize its expected reward, an NConf-MDP, instead, allows modeling a wider set of situations, including the cases in which agent and configurator display a non-cooperative behavior. Obviously, this setting cannot be addressed with straightforward application of the algorithms designed for *cooperative* Conf-MDP. In fact, if the configurator and the agent optimize separately different objectives, they might not converge to an equilibrium strategy [52, 12, 51, 13]. In this novel setting, we consider an online learning problem, where the configurator has to select a configuration, within a finite set of possible configurations, in order to maximize its own return. This framework can be seen as a *leader-follower* game, in which the *follower* (the agent) is selfish and optimizes its own reward function, and the *leader* (the configurator) has to decide the best configuration, based on its reward. Clearly, to adapt its decisions, the configurator has to receive some form of feedback related to the agent's behavior. We analyze two settings based on whether the configurator observes just the agent's actions or, in addition, a noisy version of the agent's reward.

Contributions In this paper, we extend the Configurable Markov Decision Process setting to deal with situations where the configurator and the agent have different reward functions. We call this new framework the Non-Cooperative Markov Decision Process (NConf-MDP, Section 3). Then, we formalize the problem of finding the best environment configuration, according to the configurator's reward, as a leader-follower game, in which the agent (follower) reacts to each presented configuration with its best response policy (Section 4). We provide a first algorithm, Action-feedback Optimistic Configuration Learning (AfOCL), to tackle this problem under the assumption that the configurator observes the agent's actions only (Section 5.1). We show AfOCL achieves finite expected regret, scaling linearly with the number of admissible configurations. As far as we know, this represents the *first problem-dependent* regret analysis in a Multi-Agent RL setting. Then, we introduce a second algorithm, Reward-feedback Optimistic Configuration Learning (RfOCL), that assumes the availability of a noisy version of the agent's reward, in addition to the agent's actions (Section 5.2). We prove that, under suitable conditions, RfOCL further exploits the *structure* underlying the decision process, removing the dependence on the number of configurations. The analysis use novel ideas, combining the *suboptimality gaps* of the configurator with those of the agent. Finally, we provide an experimental evaluation on benchmark domains, inspired by scenarios that motivate the NConf-MDPs framework (Section 7). The proofs of the results presented in the paper are reported in Appendix B. A preliminary version of this work was presented at "AAAI-21 Workshop on Reinforcement Learning in Games" [36].

2 Preliminaries

A *finite-horizon Markov Decision Process* [MDP, 35] is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \mu, r, H)$ where $\mathcal{S}$ is a finite state space ($S = |\mathcal{S}|$), $\mathcal{A}$ is a finite action space ($A = |\mathcal{A}|$), $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition model, which defines the density $p(s'|s, a)$ of state $s' \in \mathcal{S}$ when taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, $\mu : \mathcal{S} \to [0, 1]$ is the initial state distribution, $r : \mathcal{S} \to [0, 1]$ is the reward function, and $H \in \mathbb{N}_{\geq 1}$ is the horizon. A stochastic decision rule $\pi_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ with $h \in [H]$ prescribes the probability $\pi_h(a|s)$ of playing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. A stochastic policy $\pi = (\pi_1, \cdots, \pi_H) \in \Pi^H$ is a sequence of decision rules, where Π^H is the set of stochastic policies over the horizon H.

A *finite-horizon Configurable Markov Decision* Process [Conf-MDP, 29] is defined as $\mathcal{CM} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mu, r, H)$ and extends the MDP considering a configuration space $\mathcal{P}$ instead a single transition model p. The Q-value of a policy $\pi \in \Pi^H$ and configuration $p \in \mathcal{P}$ is the expected sum of the rewards starting from $(s, a) \in \mathcal{S} \times \mathcal{A}$ at step $h \in [H]$:

$$Q_h^{\pi,p}(s, a) = r(s) + \mathop{\mathbb{E}}_{s_{h'} \sim p,\pi} \left[\sum_{h'=h+1}^{H} r(s_{h'}) | s_h = s, a_h = a \right],$$

denoting with $\mathbb{E}_{s_{h'} \sim p,\pi}$ the expectation w.r.t. the state distribution induced by π and p at step h'. The value function is given by $V_h^{\pi,p}(s) = \mathbb{E}_{a \sim \pi_h(\cdot|s)}[Q_h^{\pi,p}(s, a)]$ and the expected return

is defined as $V^{\pi,p} = \mathbb{E}_{s\sim\mu}[V_1^{\pi,p}(s)]$. In a Conf-MDP the goal consists in finding a policy π^* together with an environment configuration p^* so as to maximize the expected return, i.e., $(\pi^*, p^*) \in \arg\max_{\pi\in\Pi^H, p\in\mathcal{P}} V^{\pi,p}$.

3 Non-Cooperative Conf-MDPs

The definition of Conf-MDP allows modeling scenarios in which agent and configurator share the same objective, encoded in a single reward function r. In this section, we introduce an extension of this framework to account for the presence of a configurator having interests that might differ from those of the agent.

Definition 3.1. *A Non-Cooperative Configurable Markov Decision Process (NConf-MDP) is defined by a tuple $\mathcal{NCM} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mu, r_c, r_o, H)$, where $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mu, H)$ is a Conf-MDP without reward and $r_c, r_o : \mathcal{S} \to [0, 1]$ are the configurator and agent (opponent) reward functions, respectively.*

Given a policy $\pi \in \Pi^H$ and a configuration $p \in \mathcal{P}$, for every $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$ we define the configurator and agent Q-values as:

$$Q_{c,h}^{\pi,p}(s,a) = r_c(s) + \mathop{\mathbb{E}}_{s_{h'}\sim p,\pi}\left[\sum_{h'=h+1}^{H} r_c(s_{h'})|s_h = s, a_h = a\right],$$

$$Q_{o,h}^{\pi,p}(s,a) = r_o(s) + \mathop{\mathbb{E}}_{s_{h'}\sim p,\pi}\left[\sum_{h'=h+1}^{H} r_o(s_{h'})|s_h = s, a_h = a\right].$$

We denote with $V_{c,h}^{\pi,p}(s) = \mathbb{E}_{a\sim\pi_h(s)}[Q_{c,h}^{\pi,p}(s,a)]$ and $V_{o,h}^{\pi,p} = \mathbb{E}_{a\sim\pi_h(s)}[Q_{o,h}^{\pi,p}(s,a)]$ the value functions and with $V_c^{\pi,p} = \mathbb{E}_{s\sim\mu}[V_{c,1}^{\pi,p}(s)]$ and $V_o^{\pi,p} = \mathbb{E}_{s\sim\mu}[V_{o,1}^{\pi,p}(s)]$ the expected returns for the configurator and the agent respectively.

4 Problem Formulation

While for classical Conf-MDPs [29, 27] a notion of optimality is straightforward as agent and configurator share the same objective, in an NConf-MDP, they can display possibly conflicting interests. We assume a *sequential* interaction between the configurator and the agent that resembles the leader-follower protocol [10, 6, 34, 38]. First, the configurator (leader) selects an environment configuration $p \in \mathcal{P}$, where $\mathcal{P}$ is a finite set made of M stochastic transition models $\mathcal{P} = \{p_1, \dots, p_M\}$. Then the agent (follower) plays a policy chosen by a *best response function* $f : \mathcal{P} \to \Pi^H$, such that: $f(p) \in \arg\max_{\pi\in\Pi^H} V_o^{\pi,p}$. The solution concept that we use is the well-known *Stackelberg equilibrium* [43, 15, 30, 32, 19]. It captures the outcome in which the configurator's transition model is optimal, under the assumption that the agent will always respond optimally [26]. However, this definition includes the possibility of ties, i.e., situations in which multiple agent optimal policies exist, with possibly different performance for the configurator. Therefore, it is necessary to employ a *tie-breaking rule*, i.e., a criterion to select *one* agent best response. Different tie-breaking rules lead to different Stackelberg equilibria, and the two prevailing solution concepts in the literature are the *Strong Stackelberg Equilibrium* (SSE) and the *Weak Stackelberg Equilibrium* (WSE). A policy-transition model pair (π^*, p^*) forms an SSE if ties are broken in favor of the configurator:

$$p^* \in \arg\max_{p\in\mathcal{P}} V_c^{f^S(p),p} \qquad \text{and} \qquad \pi^* := f^S(p) \in \arg\max_{\pi\in f(p)} V_o^{\pi,p}.$$

The WSE can be constructed by breaking the ties against the configurator. In the rest of the paper, we employ the concept of SSE; however, every result can be applied to any deterministic tie-breaking rule. We call π_p^* the application of the best response function f^S to a transition model p. Notice that the goal of the configurator is well-defined, whenever deciding the function f^S. From an online learning perspective, this goal is to minimize the expected regret:

$$\mathbb{E}[\text{Regret}(K)] = \mathbb{E}\left[\sum_{k\in[K]} \max_{p\in\mathcal{P}} V_c^{\pi_p,p} - V_c^{\pi_{p_k},p_k}\right]. \tag{1}$$

To lighten the notation, in the following, we will denote with π_i the agent's best response policy to the configuration p_i, i.e., $\pi_{p_i}^*$ and with V^i the configurator expected returned attained with configuration p_i and policy π_i, i.e., $V_c^{\pi_i,p_i}$. Finally, we denote with $V^* = \max_{i\in[M]} V^i$.

Agent's Feedback The configurator knows its reward r_c, but it does not know the agent reward r_o. At each episode $k \in [K]$, the configurator selects a configuration $p_{I_k} \in \mathcal{P}$ and observes a trajectory of H steps generated by the agent's best response policy π_{I_k}. We study two types of feedback:

- *Action-feedback* (Af). The configurator observes the states and the actions played by the agent $(s_1, a_1, \dots, s_{H-1}, a_{H-1}, s_H)$, where $a_h \sim \pi_{I_k,h}(s_h)$.
- *Reward-feedback* (Rf). The configurator observes the states, the actions played by the agent, and a noisy feedback of the agent reward function $(s_1, \widetilde{r}_1, a_1, \dots, s_{H-1}, \widetilde{r}_{H-1}, a_{H-1}, s_H, \widetilde{r}_H)$, where $a_h \sim \pi_{I_k,h}(s_h)$ and $\widetilde{r}_h$ is sampled from a distribution with mean $r_o(s)$ and support $[0, 1]$.[2]

The Rf models situations in which the agent's reward is known under uncertainty, or it is obtained in an approximate way through Inverse Reinforcement Learning [33].

Connections with Bandit Algorithms The online problem that we are facing can be seen as a stochastic multi-armed bandit [25], in which the arms are configurations, and the configurator receives a random realization of its expected return at every episode. Thus, in principle, it can be solved by standard algorithms for bandit problems, such as UCB1 [1]. These algorithms are computationally less demanding than those we will present in the next sections. On the other hand, they suffer regret that grows logarithmically, i.e., indefinitely, with the number of episodes. Indeed, they do not exploit either the information regarding the agent's policy or the structure induced by the agent's reward function. We will prove that, instead, the proposed algorithms, which use the problem structure, suffer bounded regret. Furthermore, our algorithms are combined with UCB1 confidence intervals, so their regret, at finite time, is never worse than the one of UCB1.

5 Optimistic Configuration Learning

In this section, we present two algorithms for the online learning problem introduced in Section 4. The first algorithm uses only the collected agent decisions to optimistically learn the best configuration (Section 5.1). In the second algorithm, we also use the noisy reward feedback to construct an algorithm that leverages the structure that links together all the transition probability models: the agent's reward function r_o (Section 5.2). In Appendix C, we provide some hints about the adversarial case to illustrate the additional complexities that arise. In the adversarial setting, the agent can play a different policy at each step, inside the set of possible policies that satisfy the SSE.

5.1 Action-feedback Optimistic Configuration Learning

We start with the action-feedback (Af) setting, in which the configurator observes the agent's actions only. The idea at the basis of the algorithm we propose, *Action-feedback Optimistic Configuration Learning* (AfOCL), is to maintain, for each configuration, a set of *plausible* policies that contains an agent's best response policy. The configurator plays the transition model that maximizes an optimistic approximation of its value function. Specifically, for every $i \in [M]$, $k \in [K]$, and $h \in [H]$ we denote with $\mathcal{A}^i_{k,h}(s) \subseteq \mathcal{A}$ the set of plausible actions in state s at step h for configuration p_i at the beginning of episode k. For every model p_i, the first time we visit an (s, h)-pair and observe the agent's action $a \sim \pi_{i,h}(\cdot|s)$, we set $\mathcal{A}^i_{k,h}(s) = \{a\}$. For the non-visited (s, h)-pairs, we leave $\mathcal{A}^i_{k,h}(s) = \mathcal{A}$. Based on this, we can compute an optimistic approximation $\widetilde{V}^i_{k,h}$ of the configurator value function V^i_h:

$$\widetilde{V}^i_{k,h}(s) = r_c(s) + \max_{a \in \mathcal{A}^i_{k,h}(s)} \sum_{s' \in \mathcal{S}} p_i(s'|s, a) \widetilde{V}^i_{k,h+1}(s'), \tag{2}$$

and $\widetilde{V}^i_{k,H}(s) = r_c(s)$. $\widetilde{V}^i_{k,h}$ can be computed applying a value-iteration-like algorithm [35] that employs the iterate as in Equation (2).[3] Clearly, if the agent is playing deterministically, it holds that $\mathcal{A}^i_{k,h}(s) = \{\pi_{i,h}(s)\}$ for all visited (s, h)-pairs and, consequently, $\widetilde{V}^i_{k,h}(s) \geq V^i_h(s)$. Instead, if the agent is playing stochastically, we possibly observe different actions whenever visiting (s, h) and we record *just* the first one. The following lemma shows that even for stochastic agents, if the SSE tie-breaking rule is employed, $\widetilde{V}^i_{k,h}$ is optimistic.

[2] Clearly, the results we present can be directly extended to subgaussian distributions on the reward.

[3] Notice that the computational complexity decreases as the number of visited states increases and, in any case, is bounded by that of value iteration $\mathcal{O}\left(HS^2A\right)$. Therefore, the time complexity of AfOCL is $\mathcal{O}\left(KMHS^2A\right)$.

Algorithm 1 Action-feedback Optimistic Configuration Learning (AfOCL).

Input: $\mathcal{S}$, $\mathcal{A}$, H, $\mathcal{P} = \{p_1, \dots, p_M\}$
Initialize $\mathcal{A}^i_{1,h}(s) = \mathcal{A}$ for all $s \in \mathcal{S}$, $h \in [H]$, and $i \in [M]$
for episodes $1, 2, \dots, K$ **do**
  Compute $\widetilde{V}^{i,\text{UCB}}_k$ for all $i \in [M]$
  Compute $\widetilde{V}^i_k$ for all $i \in [M]$
  Play p_{I_k} with $I_k \in \arg\max_{i \in [M]} \min\{\widetilde{V}^i_k, \widetilde{V}^{\text{UCB}}_k\}$
  Observe $(s_{k,1}, a_{k,1}, \dots, s_{k,H-1}, a_{k,H-1}, s_{k,H})$
  Compute the plausible actions for all $s \in \mathcal{S}$ and $h \in [H]$:

$$\mathcal{A}^i_{k+1,h}(s) = \begin{cases} \{a_{k,h}\} & \text{if } i = I_k \text{ and } s = s_{k,h} \text{ and } N_{k,h}(s) = 0 \\ \mathcal{A}^i_{k,h}(s) & \text{otherwise} \end{cases}$$

end for

Lemma 5.1. *The value function $\widetilde{V}^i_{k,h}(s)$ computed as in Equation* (2) *is such that $\widetilde{V}^i_{k,h}(s) \geq V^i_h(s)$ for all $s \in \mathcal{S}$, $h \in [H]$, and $i \in [M]$.*

In addition, we compute the confidence interval for UCB1 looking at the transition models as arms: $\widetilde{V}^{i,\text{UCB}}_k = \bar{V}^i_k + H\sqrt{2\log k / N_{i,k}}$, where $\bar{V}^i_k$ is the sample mean of the observed return for model p_i and $N_{i,k}$ is the number of times the algorithm plays model i up to episode k. Thus, at each episode $k \in [K]$ the configurator plays the transition model p_{I_k} maximizing the optimistic approximation:

$$I_k \in \arg\max_{i \in [M]} \min\{\widetilde{V}^i_k, \widetilde{V}^{i,\text{UCB}}_k\}.$$

The pseudocode of AfOCL is reported in Algorithm 1.

Regret Guarantees We now provide an expected regret bound for the AfOCL algorithm. If the agent's policy π_i is deterministic, it is not hard to get convinced that AfOCL suffers bounded regret since whenever an (s,h)-pair is visited under a p_i, the agent reveals its (deterministic) policy π_i. Thus, either an (s,h)-pair is visited with high probability, or it will impact only marginally on the performance. The main challenge arises when the agent is playing a stochastic policy π_i for some p_i. AfOCL just memorizes the first observed action for each (s,h), pretending the agent's policy to be deterministic. Let $\widehat{\pi}_i$ be the policy that plays the action memorized by AfOCL at the end of the K episodes, filled with the true agent's policy for the non-visited (s,h)-pairs. By construction, the support of $\widehat{\pi}_i$ is contained into the support of the true agent's policy π_i. Clearly, if π_i is optimal for the agent reward, $\widehat{\pi}_i$ is too. Furthermore, since the agent and the configurator are playing an SSE, $\widehat{\pi}_i$ will lead to the same configurator's performance as π_i. Indeed, if this were not the case, there would exist another deterministic policy optimal for the agent, leading to higher performance for the configurator, contradicting the definition of SSE. The following result shows that by switching π_i with $\widehat{\pi}_i$ changes the regret just by a multiplicative factor depending on the mismatch between the visitation distributions induced by the two policies, $d_{i,h}$ and $\widehat{d}_{i,h}$ respectively.

Theorem 5.1 (Regret of AfOCL)**.** *Let $\mathcal{NCM} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mu, r_c, r_o, H)$ with $\mathcal{P} = \{p_1, \dots, p_M\}$ be the M configurations. The expected regret of AfOCL at every episode $K > 0$ is bounded by:*

$$\mathbb{E}[\textit{Regret}(K)] \leq \mathcal{O}\left(\min\left\{ \underbrace{H^2 \sum_{i \in [M]: \Delta_i > 0} \frac{\log(K)}{\Delta_i}}_{\textit{UCB1 regret}}, \underbrace{MH^3S^2\rho}_{\textit{AfOCL regret}} \right\} \right), \tag{3}$$

where ρ is the $\max_{i \in [M]: \Delta_i > 0} \mathbb{E}\left[\max_{s \in \mathcal{S}} \max_{h \in [H]} \frac{\widehat{d}_{i,h}(s)}{d_{i,h}(s)}\right]$.

The result might be surprising as the regret is constant and independent of the suboptimality gaps between the configurations, i.e., $\Delta_i = V^* - V^i$ for every $i \in [M]$. As supported by intuition, we need to spend more time discarding MDPs that are more similar in performance to the optimal one. Formally, the maximum number of times a suboptimal configuration p_i is played is proportional to $1/\Delta_i$ (and not proportional to $1/\Delta_i^2$ as in standard bandits). This is because we *just* need *one* visit to

every reachable state. We underline that the term ρ , which indicates the expected ratio between the estimated policy's induced states distribution and real policy's induced states distribution, is equal to 1 when the agent plays a deterministic policy and bounded by SH in the worst case (see Lemma B.3). As far as we know, Theorem 5.1 is the *first problem-dependent* result for regret minimization for a multi-entity MDP. More details on the proof are given in the Appendix B.

5.2 Reward-feedback Optimistic Configuration Learning

The main drawback of AfOCL is that every transition model is treated separately, preventing from employing the underlying *structure* of the environment, which is represented by the agent reward function r_o. Indeed, if the configurator knew r_o, it could find the optimal configuration with no need for interaction by simply computing an agent's best response policies for the SSE.

The algorithm we propose in this section, *Reward-feedback Optimistic Configuration Learning* (RfOCL), employs the reward feedback (Rf), i.e., at every interaction, the configurator can see also a noisy version of the agent's reward function. The crucial point is that r_o is the same regardless of the chosen configuration, and, for this reason, it provides a link between them. Specifically, for every $k \in [K]$ and $s \in \mathcal{S}$, RfOCL maintains a confidence interval for the agent reward function $\mathcal{R}_k(s) = [\underline{r}_{o,k}(s), \overline{r}_{o,k}(s)]$ obtained using the samples collected up to episode $k-1$ *regardless* of the played configuration. We apply Höeffding's inequality to build the confidence interval: $\widehat{r}_{o,k}(s) \pm \sqrt{\frac{\log(2SHk^2)}{\max\{N_k(s),1\}}}$, where $N_k(s)$ is the number of visits of state s in the first $k-1$ episodes, and $\widehat{r}_{o,k}(s)$ is the sample mean of the observed rewards for state s up to episode k. Given the estimated reward, for every configuration $i \in [M]$, we can compute a confidence interval for the agent's Q-values $\mathcal{Q}_{k,h}(s,a) = [\underline{Q}^i_{o,k,h}(s,a), \overline{Q}^i_{o,k,h}(s,a)]$, by simply applying the Bellman equation:

$$\underline{Q}^i_{o,k,h}(s,a) = \underline{r}_{o,k}(s) + \sum_{s' \in \mathcal{S}} p_i(s'|s,a) \max_{a' \in \mathcal{A}} \underline{Q}^i_{o,k,h+1}(s',a'),$$

$$\overline{Q}^i_{o,k,h}(s,a) = \overline{r}_{o,k}(s) + \sum_{s' \in \mathcal{S}} p_i(s'|s,a) \max_{a' \in \mathcal{A}} \overline{Q}^i_{o,k,h+1}(s',a'),$$

and $\underline{Q}^i_{o,k,H}(s,a) = \underline{r}_{o,k}(s)$ and $\overline{Q}^i_{o,k,H}(s,a) = \overline{r}_{o,k}(s)$. If the true reward function belongs to the confidence interval, i.e., $r_o \in \mathcal{R}_k$, then the true Q-value belongs to the corresponding confidence interval, i.e., $Q^i_h \in \mathcal{Q}_{k,h}$. Consequently, we can use $\mathcal{Q}_{k,h}$ to restrict the set of plausible actions in a state *without* actually observing the agent playing the action in that state. Indeed, the plausible actions are those that have a Q-value upper bound larger than the maximum Q-value lower bound:

$$\widetilde{\mathcal{A}}^i_{k,h}(s) = \left\{ a \in \mathcal{A} : \overline{Q}^i_{o,k,h}(s,a) \geq \max_{a' \in \mathcal{A}} \underline{Q}^i_{o,k,h}(s,a') \right\}. \tag{4}$$

In other words, if the upper Q-value of an action is smaller than the largest lower Q-value, it cannot be the greedy action, and it is discarded. Clearly, if we observe, for the first time, the agent playing an action in (s,h) at episode k we can reduce the plausible actions to the singleton $a_{k,h}$, as in the action-feedback setting (Section 5.1). Based on this refined definition of plausible actions, we can compute the optimistic estimate $\widetilde{V}^i_{k,h}$ of the configurator value function V^i_h as in Equation (2) and proceed playing the optimistic configuration.

The pseudocode of RfOCL is reported in Algorithm 2. It is worth noting that we need to keep track of the states that have been already visited because for those, we know the agent's action, and there is no need to apply Equation (4). This is why we introduce the counts $N_{k,h}(s)$[4].

Regret Guarantees We now give a regret bound for the RfOCL algorithm. Obviously, the same arguments for AfOCL can also be applied for this extended version, and then the regret bound of Theorem 5.1 is valid for RfOCL. Moreover, for this algorithm, we prove that the regret, under the following assumption, does not depend on the number of configurations.

Assumption 1. *There exists $\epsilon > 0$ such that: $\min_{i \in [M]} \min_{s \in \mathcal{S}} \max_{h \in [H]} d^i_h(s) \geq \epsilon$, where $d^i_h(s)$ is the probability of visiting the state $s \in \mathcal{S}$ at time $h \in [H]$ in configuration p_i under the agent's best response policy π_i.*

[4]The value iteration dominates the computational complexity of an individual iteration of RfOCL (steps 5 and 9), leading, as for AfOCL, to $\mathcal{O}\left(KMHS^2A\right)$.

Algorithm 2 Reward-feedback Optimistic Configuration Learning (RfOCL)

Input: $\mathcal{S}$, $\mathcal{A}$, H, $\mathcal{P} = \{p_1, \dots, p_M\}$
Initialize $\mathcal{A}^i_{1,h}(s) = \mathcal{A}$ for all $s \in \mathcal{S}$, $h \in [H]$, and $i \in [M]$
Initialize $\overline{r}_{o,1}(s) = 1$, $\underline{r}_{o,1}(s) = 0$, and $N_{1,h}(s) = 0$ for all $s \in \mathcal{S}$ and $h \in [H]$
for episodes $1, 2, \dots, K$ **do**
  Compute $\widetilde{V}^{i,\text{UCB}}_k$ for all $i \in [M]$
  Compute $\widetilde{V}^i_k$ for all $i \in [M]$
  Play p_{I_k} with $I_k \in \arg\max_{i \in [M]} \min\{\widetilde{V}^i_k, \widetilde{V}^{\text{UCB}}_k\}$
  Observe $(s_{k,1}, \widetilde{r}_{k,1}, a_{k,1}, \dots, s_{k,H-1}, \widetilde{r}_{k,H-1}, a_{k,H-1}, s_{k,H}, \widetilde{r}_{k,H})$
  Compute $\overline{r}_{o,k+1}(s)$, $\underline{r}_{o,k+1}(s)$, and $N_{k+1,h}(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$ using $\widetilde{r}_{k,1} \cdots \widetilde{r}_{k,H}$
  Compute $\underline{Q}^i_{o,k+1,h}(s,a)$ and $\overline{Q}^i_{o,k+1,h}(s,a)$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, $h \in [H]$, and $i \in [M]$
  Compute the plausible actions for all $s \in \mathcal{S}$ and $h \in [H]$:

$$\mathcal{A}^i_{k+1,h}(s) = \begin{cases} \{a_{k,h}\} & \text{if } i = I_k \text{ and } s = s_{k,h} \text{ and } N_{k,h}(s) = 0 \\ \mathcal{A}^i_{k,h}(s) & \text{if } N_{k,h}(s) > 0 \\ \widetilde{\mathcal{A}}^i_{k+1,h}(s) & \text{otherwise} \end{cases}$$

  with $\widetilde{\mathcal{A}}^i_{k+1,h}(s)$ as in Equation (4).
end for

This assumption requires that in every model $p_i \in \mathcal{P}$ the agent has non-zero probability, in some step h, to visit every state s. This allows shrinking the confidence intervals for the reward of every state to estimate the agent's policy correctly, regardless of the played configuration. Notice that this assumption is less strict than requiring the well-known ergodicity of the Markov process induced by *any* policy, used in many algorithms [9, 21, 44].[5] Under Assumption 1 we prove the following regret guarantee.

Theorem 5.2 (Regret of RfOCL)**.** *Let $\mathcal{NCM} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mu, r_c, r_o, H)$ with $\mathcal{P} = \{p_1, \dots, p_M\}$ be the M configurations. Under Assumption 1, the expected regret of RfOCL at every episode $K > 0$ is bounded by:*

$$\mathbb{E}[\textit{Regret(K)}] \leq \mathcal{O}\left(\min\left\{ \underbrace{H^2 \sum_{i \in [M]: \Delta_i > 0} \frac{\log(K)}{\Delta_i}}_{\textit{UCB1 regret}} , \underbrace{MH^3S^2\rho}_{\textit{AfOCL regret}} , \underbrace{\overline{K}\Delta + \frac{\pi^2}{3}}_{\textit{RfOCL regret}} \right\} \right),$$

where ρ is defined as in Theorem 5.1, $\overline{K}$ is the smallest integer solution of the inequality $\overline{K} \geq 1 + \left(\frac{2H^2S^2 \log(2SH\overline{K}^2)}{2\Delta_Q^2} + \sqrt{\frac{\overline{K}-1}{2} \log(SH\overline{K}^2)} \right) \frac{1}{\epsilon}$, $\Delta = \max_{i \in [M]} \Delta_i$, i.e., the maximum suboptimality gap, and Δ_Q is the minimum positive gap of the agent's Q-values (see Appendix B).

The regret bound removes the dependence on the number of models M, as $\overline{K}$ is clearly independent of M, but it introduces, as expected, a dependence on the minimum visitation probability ϵ. The proof of the result is reported in Appendix B. Since RfOCL exploits additional information compared to AfOCL and the set of plausible actions $\mathcal{A}^i_{k,h}$ of RfOCL are subsets of those of AfOCL, the regret bound AfOCL (Theorem 5.1) also holds for RfOCL. Thus, we can take as regret bound for RfOCL the minimum between $\overline{K}\Delta + \frac{\pi^2}{3}$ and MH^3S^2. We underline that, as far as we know, this is the first proof that takes into consideration the *sub-optimality* gap of the uncontrollable entity, the agent, and the *sub-optimality* gap of the controllable entity, the configurator. This permits to derive a *problem dependent* regret bound. We think that similar techniques can also be of interest for Markov games.

6 Related Works

The idea of altering the environment dynamics to improve the agent's learning experience has been exploited before the introduction of Conf-MDPs. *Curriculum learning* [8] provides the agent with

[5] Moreover, the configurator can force this assumption since it has the *control* over the environmental transition model.

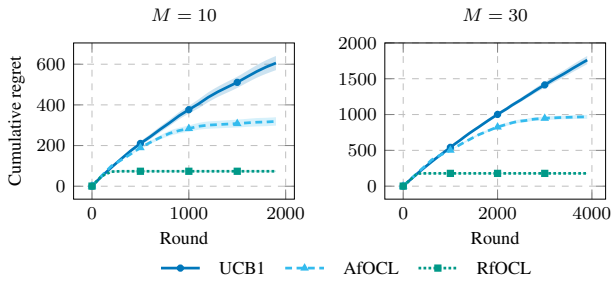


Figure 1: Cumulative regret for the Gridworld experiment. 50 runs, 98% c.i.

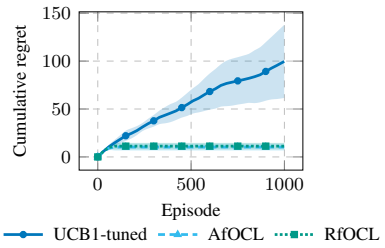


Figure 2: Cumulative regret for the Gridworld experiment without ergodicity. 50 runs, 98% c.i.

a sequence of environments, of increasing difficulty, to shape the learning process with possible benefits on the learning speed [e.g., 14, 16]. Although the learning process is carried out in a different environment, the configuration is typically performed in simulation only. The setting more similar to Non-Conf-MDP is the one presented in [47], where the configurator and the agent have opposite reward functions (similar to a zero-sum game).

In the Conf-MDP framework, instead, the configuration opportunities are an *intrinsic* property of the environment [29]. The initial approaches entitled the agent of the configuration activity and, consequently, this task was totally auxiliary to its learning experience [29, 39, 27]. More recently, it has been observed that environment configuration can be actuated even by an external entity, opening new opportunities for the application of environment configurability, including settings in which the configurator's interest conflicts with those of the agent. For instance, in [28] the configurator acts on the environment to induce the agent to reveal its capabilities in terms of perception and actuation. Instead, in [17] a threatener entity can change the transition probabilities either in a stochastic or adversarial manner. More generally, environment configuration carried out by an external entity has been studied in the field of planning as a form of *environment design* [48]. Thus, our NConf-MDP unifies these settings, allowing for arbitrary agent's and configurator's reward functions. An interesting connection is established with the *robust control* literature [31, 20]. Whenever the two reward functions are opposite, i.e., the interaction between the agent and the configuration is fully *competitive*, the resulting equilibrium corresponds to a robust policy. Indeed, while the agent tries to maximize its expected return, the configurator places the agent in the worst possible environment.

Configurable environments (cooperative and non-cooperative) share similarities with *environment design* [49]. At a high level, the two frameworks share analogous objectives: they both aim at determining an environment with a certain goal that can differ from that of the agent. However, there are some notable differences. In particular, the classical environment design formulation [49] assumes that the configurator (called "interested party") knows the agent's best response function, while in our approach, we learn it by interaction. Nevertheless, the general environment design makes no assumption about the underlying environment, that might not me an MDP. Instead, [22] limit to MDPs and considers a form of cooperative environment design in which the goal is to maximize the agent's performance. Interestingly, some works [22, 37] also account for a cost function to penalize expensive environment configurations.

The design of our approaches is based on the OFU principle used for stochastic multi-armed bandits [e.g., 23, 1, 18, 25] and MDPs [e.g., 2, 7, 3]. Moreover, our learning setting with reward feedback is related to structured bandits or bandits with correlated arms.[6] Interestingly, for certain structures, it is known that bounded regret is achievable [11, 24], a property that is enjoyed by both our algorithms. Our setting is also close to the Stochastic Games model, in which two or more agents act in an MDP to maximize their own reward functions. Recently, the stochastic game's framework gains growing interest [5, 4, 50], especially in the offline setting i.e., we can control all the agents. For this reason, these approaches do not apply to our setting, where we have the control of the configurator only.

[6] In our case, playing a single configuration provides information about the opponent's reward, which, in turn, provides information about the value of all configurations.

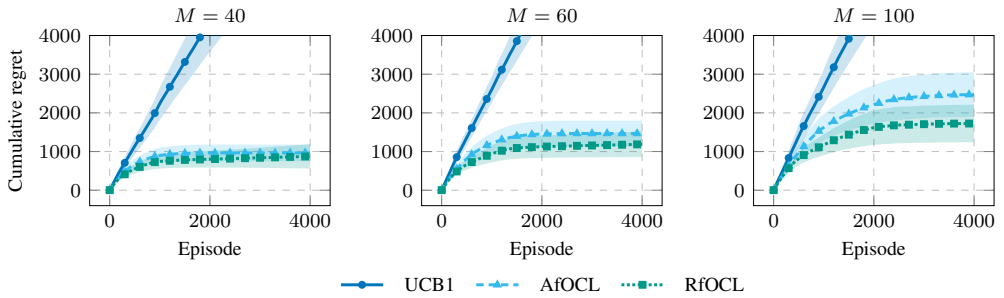


Figure 3: Cumulative regret as a function of the episodes for the Student-Teacher experiment. 50 runs, 98% c.i.

Although some works tackle the online setting [44, 45, 41], where we can control only one agent, all of these algorithms work in the zero-sum setting only.

7 Experiments

In this section, we provide the experimental evaluation of our algorithms in two different settings: when the policies are stochastic and when the policies are deterministic. For these experiments, we provide two novel environments: Configurable Gridworld and the Student-Teacher. We compare the algorithms with the standard (theoretical) implementation of UCB1 [1]. The environment description and additional results can be found in Appendix D.

Stochastic policies The Configurable Gridworld is a configurable version of a classic 3×3 Gridworld. The agent's starting state is in the cell $(0, 1)$, and its goal is to minimize the number of steps required to reach the exit located in the cell $(2, 1)$. The configurator takes reward 1 when the agent occupies the central cell $(1, 1)$ and 0 otherwise. In a classic Gridworld, the optimal policy would be trivial, as the agent would proceed straight to the exit. In this Configurable Gridworld, instead, the configurator can set the "power" p of a stochastic obstacle located in the cell $(1, 1)$. When the agent is in that cell and performs action "go right" to reach the exit, it will hit the obstacle, and will remain in the same position with probability p. The configurator's goal is to tune this probability to keep the agent in the central cell for the maximum number of steps.

The M configurations differ in the probability p and are obtained by a regular discretization of $[0, 1]$. In the first experiment (Figure 1), we considered 10 and 30 configurations with a number of episodes $K = 2000$ and $K = 4000$ and horizon $H = 10$. For this experiment, the agent plays *optimal stochastic policies*. We can see that AfOCL and RfOCL suffer constant regret, whereas UCB1 displays a logarithmic regret, as expected. Specifically, RfOCL outperforms AfOCL and stops playing suboptimal configuration in less than 500 episodes in both cases. This can be explained because, being Assumption 1 fulfilled (in fact, the agent has the probability 0.1 of failing its action), RfOCL is able to exploit the underlying structure of the problem more effectively.

Non-Ergodicity In Figure 2, we have only three configurations designed to induce an optimal agent's policy that generates a non-ergodic Markov chain. In this case, the optimal policies are deterministic, and we violate Assumption 1. For this reason, we observe that AfOCL and RfOCL display very similar behavior but still significantly better than UCB1.

Deterministic policies: Student-Teacher The Student-Teacher environment models a simple interaction between a student and a teacher. There is a set of exercises, with a different level of *teacher hardness* and *student hardness* each. The teacher has to decide the optimal sequence of exercises in order to make the student acquire as much knowledge as possible. The student's goal is to maximize the number of exercises and to reduce the *hardness* of the proposed exercises. At each timestep, the student decides whether to answer the exercise or not. If it answers, it receives a reward equal to the level of "correctness" of the exercise, the teacher receives a reward corresponding to the level of exercise's "teacher hardness", and they end up to the next exercise. If the student does not

answer, the student and the teacher will receive -1, and with a probability of 0.7, the next exercise will be easier to solve. In Figure 3, the results with $M \in \{40, 60, 100\}$ and horizon $H = 10$ are shown. The configurations represent the distribution over the next exercise, given a positive answer. In every run, we change the *student hardness* of the exercises. We observe that both AfOCL and RfOCL suffer significantly less regret compared to UCB1 and tend to converge to constant regret as expected. It is interesting to observe that, in line with our analysis, the gap between AfOCL and RfOCL appears more evident as the number of configurations grows.

8 Conclusions

In this paper, we have introduced an extension of the Conf-MDP framework to account for possible non-cooperative interaction between the agent and the configurator. We focused on an online learning problem in this new setting, proposing two regret minimization algorithms for identifying the best environment configuration within a finite set, based on the principle of optimism in the face of uncertainty. We proved that even when the agent's policy is stochastic, and the configurator observes the agent's actions, it is possible to achieve finite regret that depends linearly on the admissible number configurations. Furthermore, we illustrated that we can remove this dependence if the configurator observes a possibly noisy version of the agent's reward and under sufficient regularity conditions on the environment. This paper also gives interesting insights on the importance of properly exploiting the available *feedback* to construct efficient algorithms. Moreover, as far as we know, the ones we have presented are the first *problem-dependent* regret results for multi-entity MDPs. The experimental evaluation showed that our algorithms display a convergence speed significantly faster than UCB1, and RfOCL tends to outperform AfOCL thanks to the exploitation of the additional structure. Future research directions include a deeper analysis of the adversarial setting, as well as the application to inverse reinforcement learning.

Limitations and Societal Impact

Methods that incentive the manipulation of users' behavior can have, generally speaking, a negative societal impact, when used, for instance, in a marketing campaign. Nevertheless, our work is mainly theoretical and, at the present level, can hardly be used in a malevolent way. Another relevant aspect is the cost of environment configuration. We are aware that reconfiguring the environment is an activity that typically leads to higher costs compared with policy learning. However, we did not consider this aspect in the formalization of the Non-Cooperative Conf-MDP since it would possibly make the problem more complex (like, for instance, when considering bandits with switching costs).

References

[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002.

[2] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 89–96. Curran Associates, Inc., 2008.

[3] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Mach. Learn.*, 91(3):325–349, 2013.

[4] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 551–560. PMLR, 2020.

[5] Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[6] Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation (EC)*, pages 61–78. ACM, 2015.

[7] Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 35–42. AUAI Press, 2009.

[8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM, 2009.

[9] Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3:213–231, 2002.

[10] Michele Breton, Abderrahmane Alj, and Alain Haurie. Sequential stackelberg equilibria in two-person games. *Journal of Optimization Theory and Applications*, 59(1):71–97, 1988.

[11] Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *The 26th Annual Conference on Learning Theory (COLT)*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 122–134. JMLR.org, 2013.

[12] Lucian Busoniu, Robert Babuska, and Bart De Schutter. Multi-agent reinforcement learning: A survey. In *Ninth International Conference on Control, Automation, Robotics and Vision, ICARCV 2006, Singapore, 5-8 December 2006, Proceedings*, pages 1–6. IEEE, 2006.

[13] Benjamin Chasnov, Lillian J. Ratliff, Eric Mazumdar, and Samuel Burden. Convergence analysis of gradient-based learning in continuous games. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 115 of *Proceedings of Machine Learning Research*, pages 935–944. AUAI Press, 2019.

[14] Kamil Andrzej Ciosek and Shimon Whiteson. OFFER: off-environment reinforcement learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 1819–1825. AAAI Press, 2017.

[15] Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings 7th ACM Conference on Electronic Commerce (EC)*, pages 82–90. ACM, 2006.

[16] Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *1st Annual Conference on Robot Learning (CoRL)*, volume 78 of *Proceedings of Machine Learning Research*, pages 482–495. PMLR, 2017.

[17] Víctor Gallego, Roi Naveiro, and David Ríos Insua. Reinforcement learning under threats. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 9939–9940. AAAI Press, 2019.

[18] Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *The 24th Annual Conference on Learning Theory (COLT)*, volume 19 of *JMLR Proceedings*, pages 359–376. JMLR.org, 2011.

[19] Qingyu Guo, Jiarui Gan, Fei Fang, Long Tran-Thanh, Milind Tambe, and Bo An. On the inducibility of stackelberg equilibrium for security games. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 2020–2028. AAAI Press, 2019.

[20] Garud N. Iyengar. Robust dynamic programming. *Math. Oper. Res.*, 30(2):257–280, 2005.

[21] Michael J. Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynomial time. *Mach. Learn.*, 49(2-3):209–232, 2002.

[22] Sarah Keren, Luis Enrique Pineda, Avigdor Gal, Erez Karpas, and Shlomo Zilberstein. Equi-reward utility maximizing design in stochastic environments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4353–4360. ijcai.org, 2017.

[23] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[24] Tor Lattimore and Rémi Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 550–558, 2014.

[25] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[26] George Leitmann. On generalized stackelberg strategies. *Journal of optimization theory and applications*, 26(4):637–643, 1978.

[27] Alberto Maria Metelli, Emanuele Ghelfi, and Marcello Restelli. Reinforcement learning in configurable continuous environments. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 4546–4555. PMLR, 2019.

[28] Alberto Maria Metelli, Guglielmo Manneschi, and Marcello Restelli. Policy space identification in configurable environments. *Mach. Learn.*, 2021.

[29] Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. Configurable Markov decision processes. In *Proceedings of the 35th International Conference on Machine Learning, (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 3488–3497. PMLR, 2018.

[30] Thanh Hong Nguyen, Rong Yang, Amos Azaria, Sarit Kraus, and Milind Tambe. Analyzing the effectiveness of adversary modeling in security games. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press, 2013.

[31] Arnab Nilim and Laurent El Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. In *Advances in Neural Information Processing Systems 16 (NIPS)*, pages 839–846. MIT Press, 2003.

[32] Steven Okamoto, Noam Hazon, and Katia P. Sycara. Solving non-zero sum multiagent network flow security games with attack costs. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 879–888. IFAAMAS, 2012.

[33] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Found. Trends Robotics*, 7(1-2):1–179, 2018.

[34] Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning optimal strategies to commit to. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 2149–2156. AAAI Press, 2019.

[35] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

[36] Giorgia Ramponi, Alberto Maria Metelli, Alessandro Concetti, and Marcello Restelli. Online learning in non-cooperative configurable Markov decision process. *AAAI-21 Workshop on Reinforcement Learning in Games*, 2021.

[37] Sandhya Saisubramanian and Shlomo Zilberstein. Mitigating negative side effects via environment shaping. In *20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1640–1642. ACM, 2021.

[38] Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. Learning to play sequential games versus unknown opponents. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[39] Rui Silva, Francisco S. Melo, and Manuela Veloso. What if the world were different? gradient-based exploration for new optimal policies. In *4th Global Conference on Artificial Intelligence (GCAI)*, volume 55 of *EPiC Series in Computing*, pages 229–242. EasyChair, 2018.

[40] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[41] Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Provably efficient online agnostic learning in Markov games. *CoRR*, abs/2010.15020, 2020.

[42] Andrea Tirinzoni, Riccardo Poiani, and Marcello Restelli. Sequential transfer in reinforcement learning with a generative model. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 9481–9492. PMLR, 2020.

[43] Heinrich Von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media, 2010.

[44] Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 4987–4997, 2017.

[45] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 3674–3682. PMLR, 2020.

[46] Andrea Zanette, Mykel J. Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 5626–5635. 2019.

[47] Haifeng Zhang, Jun Wang, Zhiming Zhou, Weinan Zhang, Yin Wen, Yong Yu, and Wenxin Li. Learning to design games: Strategic environments in reinforcement learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3068–3074. ijcai.org, 2018.

[48] Haoqi Zhang, Yiling Chen, and David C. Parkes. A general approach to environment design with one agent. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2002–2014, 2009.

[49] Haoqi Zhang, Yiling Chen, and David C. Parkes. A general approach to environment design with one agent. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2002–2014, 2009.

[50] Kaiqing Zhang, Sham M. Kakade, Tamer Basar, and Lin F. Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[51] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *CoRR*, abs/1911.10635, 2019.

[52] Martin Zinkevich, Amy Greenwald, and Michael L. Littman. Cyclic equilibria in Markov games. In *Advances in Neural Information Processing Systems 18 (NIPS)*, pages 1641–1648, 2005.

A Notation

$\mathcal{S}$	State space
$\mathcal{A}$	Action space
$\mathcal{P}$	Configuration space
M	Configuration space size
r_o	Agent's reward function
r_c	Configurator's reward function
μ	Initial state distribution
H	Horizon
$Q_{c,h}^{\pi,p}(s,a)$	Configurator's Q-value with policy π and configuration p
$Q_{o,h}^{\pi,p}(s,a)$	Agent's Q-value with policy π and configuration p
$V_{c,h}^{\pi,p}(s)$	Configurator's value function with policy π and configuration p
$V_{o,h}^{\pi,p}(s)$	Agent's value function with policy π and configuration p
$V_c^{\pi,p}$	Configurator's expected return with policy π and configuration p
$V_o^{\pi,p}$	Agent's expected return with policy π and configuration p
$\pi_i = \pi_{p_i}^*$	Agent's best response to configuration p_i
$V^i = V_c^{\pi_{p_i}^*, p_i}$	Configurator's expected return with the agent's best response policy $\pi_{p_i}^*$ to configuration p_i
$V^* = V_c^{\pi_{p_{i^*}}^*, p_{i^*}}$	Configurator's expected return with the agent's best response policy $\pi_{p_i}^*$ to the best configuration p_{i^*}
$\widetilde{V}_k^i$	Optimistic configurator's expected return for configuration p_i at episode k
$\widetilde{\pi}_{i,k}$	Estimated agent's best response policy for configuration p_i at episode k
$\Delta_i = V^* - V^i$	Suboptimality gap of the configuration p_i
K	Number of episodes
N_i	Number of times the configuration p_i is played
$N_k(s)$	Number of visits of state s before episode k
$N_{k,h}^i(s)$	Number of visits of state s at step h before episode k with configuration p_i
$\underline{r}_{o,k}(s)$	Lower confidence value for the agent's reward
$\overline{r}_{o,k}(s)$	Upper confidence value for the agent's reward
$\widehat{r}_{o,k}(s)$	Sample mean of observed rewards
$\underline{Q}_{o,k,h}^i(s,a)$	Lower confidence value of the agent's Q-function with configuration p_i
$\overline{Q}_{o,k,h}^i(s,a)$	Upper confidence value of the agent's Q-function with configuration p_i
$\mathcal{A}_{k,h}^i(s)$	Set of agent's plausible actions in state s at step h up to episode k
$d_h^i(s)$	Visitation probability the state s at step h with configuration p_i under the agent's best response policy π_i
$\widetilde{d}_h^i(s)$	Visitation probability the state s at step h with configuration p_i under the estimated agent's best response policy $\widetilde{\pi}_{i,k}$

B Missing Proofs

In this appendix, we report the proofs of the results presented in the main paper.

B.1 Proofs of Section 5.1

Lemma 5.1. *The value function $\widetilde{V}^i_{k,h}(s)$ computed as in Equation (2) is such that $\widetilde{V}^i_{k,h}(s) \geq V^i_h(s)$ for all $s \in \mathcal{S}$, $h \in [H]$, and $i \in [M]$.*

Proof. We will prove the lemma by induction. We define $N^i_{k,h}(s')$ the number of times the state s is visited at step h with the configuration $p_i \in \mathcal{P}$ up to episode $k-1$.

Case base = $\widetilde{V}^i_{k,H}(s) \geq V^i_{k,H}(s)$ In this case is proven since $\widetilde{V}^i_{k,H}(s) = V^i_{k,H}(s) = r_c(s)$.

Induction step We assume that $\widetilde{V}^i_{k,h+1}(s) \geq V^i_{k,h+1}(s)$ and we will prove that $\widetilde{V}^i_{k,h}(s) \geq V^i_{k,h}(s)$.

$$\begin{aligned}\widetilde{V}^i_{k,h}(s) &= r_c(s) + \max_{a \in \mathcal{A}^i_{k,h}(s)} \sum_{s' \in S} p_i(s'|s,a)\widetilde{V}^i_{k,h+1}(s) \\ &\geq r_c(s) + \max_{a \in \mathcal{A}^i_{k,h}(s)} \sum_{s' \in S} p_i(s'|s,a)V^i_{k,h+1}(s),\end{aligned}$$

This is true for the induction hypothesis. Now there are two cases:

- $N^i_{k,h}(s) > 0$ i.e, we have already visited the state s at step h. In this case $\mathcal{A}^i_{k,h}(s) = a$ where a is the action that by the agent the last time we have seen state s at step h. If the policy is deterministic then:

  $$r_c(s) + \max_{a \in \mathcal{A}^i_{k,h}(s)} \sum_{s' \in S} p_i(s'|s,a)V^i_{k,h+1}(s) = r_c(s) + \sum_{s' \in S} p_i(s'|s,\pi_p(s))V^i_{k,h+1}(s).$$

  Instead, if the policy is stochastic:

  $$r_c(s) + \sum_{s' \in S} p_i(s'|s,a)V^i_{k,h+1}(s) = r_c(s) + \sum_{a' \in \mathcal{A}} \sum_{s' \in S} p_i(s'|s,\pi_p(s,a'))V^i_{k,h+1}(s),$$

  since, otherwise, the agent's policy played when we have seen the realization of a is not optimal and does not respect the SSE definition.

- $N^i_{k,h}(s) = 0$. In this case $\mathcal{A}^i_{k,h}(s) = \mathcal{A}$ then, clearly,:

  $$r_c(s) + \sum_{s' \in S} \max_{a \in \mathcal{A}^i_{k,h}(s)} p_i(s'|s,a)V^i_{k,h+1}(s) \geq r_c(s) + \sum_{a \in \mathcal{A}^i_{k,h}(s)} \sum_{s' \in S} p_i(s'|s,a)\pi_p(s,a)V^i_{k,h+1}(s).$$

  Then the result follows.

□

We denote with $\widetilde{d}^i_h(s)$ the visitation probability of visiting state s at step h under transition model p_i and playing the estimated agent's best response policy $\widetilde{\pi}_{i,k}$ (we will omit the subscript k in the following). Then we start by constructing the policy $\widehat{\pi}_i$ such that:

$$\widehat{\pi}_{i,h}(\cdot|s) = \begin{cases} \widetilde{\pi}_{i,h}(\cdot|s) & \text{if} \quad N^i_{K,h}(s) > 0 \\ \pi_{i,h}(\cdot|s) & \text{if} \quad N^i_{K,h}(s) = 0 \end{cases} \tag{5}$$

A simple extension of Lemma 5.1 proves that the policy is optimal $\widehat{\pi}_{i,h}$ (thanks to the SSE definition). We call $\widehat{V}^i$ the expected return of $\widehat{\pi}_i$, and, obviously, $\widehat{V}^i = V^i$.

The visitation probabilities satisfy the following equalities for all $h \geq 2$:

$$\begin{aligned} d^i_h(s) &= \sum_{s' \in \mathcal{S}} p_i(s|s',\cdot)^T \pi_{i,h}(\cdot|s') d^i_{h-1}(s') \\ \widehat{d}^i_h(s) &= \sum_{s' \in \mathcal{S}} p_i(s|s',\cdot)^T \widehat{\pi}_{i,h}(\cdot|s') \widehat{d}^i_{h-1}(s') \\ \widetilde{d}^i_h(s) &= \sum_{s' \in \mathcal{S}} p_i(s|s',\cdot)^T \widetilde{\pi}_{i,h}(\cdot|s') \widetilde{d}^i_{h-1}(s') = \sum_{s' \in \mathcal{S}} p_i(s|s',\widetilde{\pi}_{i,h}(s)) \widehat{d}^i_{h-1}(s'), \end{aligned} \tag{6}$$

where for $\widetilde{d}^i_h(s)$, we exploited the fact that $\widetilde{\pi}_{i,h}$ is deterministic, and $d^i_1(s) = \widetilde{d}^i_1(s) = \widehat{d}^i_1(s) = \mu(s)$.

Lemma B.1. *For every episode $k \in [K]$ and configuration $p_i \in \mathcal{P}$, the difference between the optimistic expected return $\widetilde{V}_k^i$ and the true expected return V^i is bounded by:*

$$\widetilde{V}_k^i - V^i \leq 2H\rho_i \sum_{s\in\mathcal{S}} \sum_{h=1}^{H-1} d_h^i(s) \mathbb{1}\left\{N_{k,h}^i(s) = 0\right\}. \tag{7}$$

where $N_{k,h}^i(s)$ is the number of times the state $s \in \mathcal{S}$ is visited at step $h \in [H]$ with the configuration $p_i \in \mathcal{P}$ up to episode $k-1$, where $\rho_i = \max_{s\in\mathcal{S}} \max_{h\in H} \frac{\widehat{d}_{i,h}(s)}{d_{i,h}(s)}$.

Proof. As observed above, π_i and $\widehat{\pi}_i$, given the definition of SSE, induce the same value function $V^i = \widehat{V}^i$. Thus, we have

$$\begin{aligned}
\widetilde{V}_k^i - V^i = \widetilde{V}_k^i - \widehat{V}^i &= \sum_{s\in\mathcal{S}} \left[\mu(s)r(s) - \mu(s)r(s) + \sum_{h=2}^{H} (\widetilde{d}_h^i(s) - \widehat{d}_h^i(s))r(s)\right] && \text{(P.1)}\\
&\leq \sum_{s\in\mathcal{S}} \sum_{h=2}^{H} \left|\widetilde{d}_h^i(s) - \widehat{d}_h^i(s)\right| && \text{(P.2)}\\
&= \sum_{s\in\mathcal{S}} \sum_{h=1}^{H-1} \left|\sum_{s'\in\mathcal{S}} \widetilde{d}_h^i(s') p_i(s|s', \widetilde{\pi}_{i,h}(s')) - \widehat{d}_h^i(s') p_i(s|s',\cdot)^T \widehat{\pi}_{i,h}(\cdot|s')\right| && \text{(P.3)}\\
&= \sum_{s\in\mathcal{S}} \sum_{h=1}^{H-1} \sum_{s'\in\mathcal{S}} \left|\widetilde{d}_h^i(s') - \widehat{d}_h^i(s')\right| p_i(s|s', \widetilde{\pi}_{i,h}(s')) + \widehat{d}_h^i(s') \left|p_i(s|s', \widetilde{\pi}_{i,h}(s')) - p_i(s|s',\cdot)^T \widehat{\pi}_{i,h}(\cdot|s')\right|\\
&= \sum_{s'\in\mathcal{S}} \sum_{h=2}^{H-1} \left|\widetilde{d}_h^i(s') - \widehat{d}_h^i(s')\right| + \sum_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} \sum_{h=1}^{H-1} \widehat{d}_h^i(s') \left|p_i(s|s', \widetilde{\pi}_{i,h}(s')) - p_i(s|s',\cdot)^T \widehat{\pi}_{i,h}(\cdot|s')\right| && \text{(P.4)}\\
&= \sum_{H'=2}^{H} \sum_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} \sum_{h=1}^{H'-1} \widehat{d}_h^i(s') \left|p_i(s|s', \widetilde{\pi}_{i,h}(s')) - p_i(s|s',\cdot)^T \widehat{\pi}_{i,h}(\cdot|s')\right| && \text{(P.5)}\\
&\leq H \sum_{s'\in\mathcal{S}} \sum_{h=1}^{H-1} \widehat{d}_h^i(s') \sum_{s\in\mathcal{S}} \left|p_i(s|s', \widetilde{\pi}_{i,h}(s')) - p_i(s|s',\cdot)^T \widehat{\pi}_{i,h}(\cdot|s')\right| && \text{(P.6)}\\
&\leq 2H \sum_{s'\in\mathcal{S}} \sum_{h=1}^{H-1} \mathbb{1}\left\{N_{k,h}^i(s) = 0\right\} \widehat{d}_h^i(s') && \text{(P.7)}\\
&= 2H \sum_{s'\in\mathcal{S}} \sum_{h=1}^{H-1} \mathbb{1}\left\{N_{k,h}^i(s) = 0\right\} d_h^i(s') \frac{\widehat{d}_h^i(s')}{d_h^i(s')}, && \text{(P.8)}
\end{aligned}$$

where in line (P.1) we use the definition of expected return. In line (P.2) we bound the value of every reward with its maximum value 1. In line (P.3) we expanded the probability distribution of visiting states using Equations (6). In line (P.4) we observe that $\widetilde{d}_1^i(s') - \widehat{d}_1^i(s') = \mu(s) - \mu(s) = 0$ to make the first summation starting from $h = 2$. In line (P.5), we apply the recursion with line (P.2). In line (P.6), we bound $H' \leq H$ and observe that the outer summation has less than H terms. Finally, in line (P.8) we upper bound the differences between the two probabilities with 2, and we use the fact that when we have seen a state s at step h with a configuration p_i the two policies are equal by construction. □

Lemma B.2. *A configuration $p_i \in \mathcal{P}$ is no longer played after episode $k \in [K]$ if for every state $s \in \mathcal{S}$ and $h \in [H]$, with $d_h^i(s) \geq \frac{\Delta_i - c}{2H^2 S\rho_i}$, we have $N_{k,h}^i(s) > 0$, where $c > 0$ is arbitrary and $\Delta_i = V^* - V^i$.*

Proof. It suffices to prove that the optimistic expected return satisfies $\widetilde{V}_k^i < V^*$, that, in turn, will satisfy $V^* \leq \widetilde{V}_k^{i^*}$ where $i^* \in \arg\max_{i\in[M]} V^i$ (this way configuration i will no longer be played):

$$\begin{aligned}
\widetilde{V}_k^i &= V^i + \widetilde{V}_k^i - V^i \\
&\leq V^i + 2H\rho_i \sum_{s\in\mathcal{S}} \sum_{h=1}^{H-2} d_h^i(s) \mathbb{1}\left\{N_{h,k}^i(s) = 0\right\} && \text{(P.9)} \\
&\leq V^i + 2H^2 S\rho_i \frac{\Delta_i - c}{2H^2 S\rho_i} && \text{(P.10)} \\
&= V^i + \Delta_i - c < V^*, && \text{(P.11)}
\end{aligned}$$

where in line (P.9) we apply Lemma B.1. In line (P.10) we bound the state visitation probabilities of the (s, h) pairs with $N_{h,k}^i(s) > 0$ with their maximum value, as in the statement hypothesis. In line (P.11) we use the fact that $\Delta_i = V^* - V_i$. □

Theorem 5.1 (Regret of AfOCL)**.** *Let $\mathcal{NCM} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mu, r_c, r_o, H)$ with $\mathcal{P} = \{p_1, \ldots, p_M\}$ be the M configurations. The expected regret of AfOCL at every episode $K > 0$ is bounded by:*

$$\mathbb{E}[\textit{Regret}(K)] \leq \mathcal{O}\left(\min\left\{ \underbrace{H^2 \sum_{i\in[M]:\Delta_i>0} \frac{\log(K)}{\Delta_i}}_{\textit{UCB1 regret}}, \underbrace{MH^3S^2\rho}_{\textit{AfOCL regret}} \right\}\right), \tag{3}$$

where ρ is the $\max_{i\in[M]:\Delta_i>0} \mathbb{E}\left[\max_{s\in\mathcal{S}} \max_{h\in[H]} \frac{\widehat{d}_{i,h}(s)}{d_{i,h}(s)}\right]$.

Proof. We start by dividing the analysis between the UCB1 algorithm and the proposed new algorithm. For the UCB1 algorithm the regret is straightforward from [1]:

$$\mathbb{E}[\text{Regret}(K)] \leq \mathcal{O}\left(H^2 \sum_{i\in[M]:\Delta_i>0} \frac{\log(K)}{\Delta_i} \right),$$

since the random variables V^i for each model $i \in [M]$ have their support in $[0, H]$.

Then, we analyze the regret of the proposed algorithm. We rephrase the regret as:

$$\mathbb{E}[\text{Regret}(K)] = \sum_{i\in[M]:\Delta_i>0} \Delta_i \, \mathbb{E}[N_i],$$

where N_i is the number of times that the algorithm plays model p_i which is not the optimal configuration p_{i^*}. We start bounding for every configuration p_i s.t. $\Delta_i > 0$ the expected value of N_i. We denote with k_l^i the round at which model i is selected for the l-th time:

$$\begin{aligned}
\mathbb{E}[N_i] &\leq \sum_{l=0}^{K} \Pr(N_i \geq l) \\
&\leq \sum_{l=0}^{\infty} \Pr(N_i \geq l) && \text{(P.12)} \\
&\leq \sum_{l=0}^{\infty} \Pr\left(\widetilde{V}_{k_l^i}^i - V^* \geq 0\right), && \text{(P.13)} \\
& && \text{(P.14)}
\end{aligned}$$

where in line (P.12) we extend the sum to ∞. In line (P.13) we exploit the fact that if configuration i is selected then it must be $\widetilde{V}_{k_l^i}^i \geq \widetilde{V}_{k_l^i}^{i^*}$ and, because of optimism $\widetilde{V}_{k_l^i}^{i^*} \geq V^*$. Then, we observe that for Lemma B.2, if configuration i is played at time k_l^i, then there must exists $s \in \mathcal{S}$ and $h \in [H]$ with $d_h^i(s) \geq \frac{\Delta_i - c}{2H^2 S}$ that is not played yet. Formally:

$$\mathbb{E}[N_i] \leq \sum_{l=0}^{\infty} \Pr\left(\widetilde{V}^i_{k^i_l} - V^* \geq 0\right)$$

$$\leq 1 + \sum_{l=1}^{\infty} \Pr\left(\exists s \in \mathcal{S},\, \exists h \in [H] \,:\, d^i_h(s) \geq \frac{\Delta_i - c}{2H^2 S \rho_i} \,\wedge\, N^i_{k^i_l,h}(s) = 0\right) \tag{P.15}$$

$$\leq 1 + \sum_{l=1}^{\infty} \sum_{s\in\mathcal{S}, h\in[H]} \Pr\left(d^i_h(s) \geq \frac{\Delta_i - c}{2H^2 S \rho_i} \,\wedge\, N^i_{k^i_l,h}(s) = 0\right) \tag{P.16}$$

$$\leq 1 + \sum_{l=1}^{\infty} \sum_{s\in\mathcal{S}, h\in[H]} \mathbb{E}\left[\Pr\left(N^i_{k^i_l,h}(s) = 0 \middle| d^i_h(s) \geq \frac{\Delta_i - c}{2H^2 S \rho_i}\right) \mathbb{1}\left\{d^i_h(s) \geq \frac{\Delta_i - c}{2H^2 S \rho_i}\right\}\right] \tag{P.17}$$

$$\leq 1 + \sum_{l=1}^{\infty} \sum_{s\in\mathcal{S}, h\in[H]} \mathbb{E}\left[\Pr\left(N^i_{k^i_l,h}(s) = 0 \middle| d^i_h(s) \geq \frac{\Delta_i - c}{2H^2 S \rho_i}\right)\right] \tag{P.18}$$

$$\leq 1 + SH\,\mathbb{E}\left[\sum_{l=1}^{\infty} \left(1 - \frac{\Delta_i - c}{2H^2 S \rho_i}\right)^{l-1}\right] \tag{P.19}$$

$$= 1 + SH\,\mathbb{E}\left[\frac{1}{\frac{\Delta_i - c}{2H^2 S \rho_i}}\right] \leq 1 + \frac{2H^3 S^2\,\mathbb{E}[\rho_i]}{\Delta_i - c}, \tag{P.20}$$

where, in line (P.15) we use Lemma B.2. In line (P.16) we use the union bound over the set employed for existential quantification. In line (P.17) we employed the definition of conditional probability and in line (P.18) we bounded the indicator with 1. In line (P.19) we bound the probability as $\Pr\left(N^i_{k^i_l,h}(s) = 0\right) = (1 - d^i_h(s))^{l-1}$, thanks to the independence of the rounds. In line (P.20) we use the geometric series properties.

So the expected regret is bounded by:

$$\mathbb{E}[\text{Regret}(K)] = \sum_{i\in[M]:\Delta_i>0} \Delta_i\,\mathbb{E}[N_i] \leq \sum_{i\in[M]:\Delta_i>0} \Delta_i \left(\frac{2H^3S^2\,\mathbb{E}[\rho_i]}{\Delta_i - c} + 1\right) \leq 3MH^3S^2\rho,$$

having taken the infimum over $c > 0$ and $\rho = \max_{i\in[M]:\Delta_i>0} \mathbb{E}[\rho_i]$. □

Lemma B.3. *The expected value of $\frac{\widehat{d}_{i,h}(s)}{d_{i,h}(s)}$, taken w.r.t. the randomness of the episodes, is 1.Moreover, the expectation of $\rho_i = \max_{s\in\mathcal{S}} \max_{h\in[H]} \frac{\widehat{d}_{i,h}(s)}{d_{i,h}(s)}$, taken w.r.t. the randomness of the episodes, is bounded by SH.*

Proof. First of all, we observe that, given its definition, for every $s, s' \in \mathcal{S}$ and $h, h' \in [H]$ such that $(s,h) \neq (s',h')$ we have that $\widehat{\pi}_h(\cdot|s)$ and $\widehat{\pi}_{h'}(\cdot|s')$ are independent. This is because $\widehat{\pi}$ is a policy that plays deterministically an action in each (s,h), selected by querying the true agent's policy π. Consequently, since actions played by the agent in different (s,h) are independent, also the policy entries $\widehat{\pi}_h(\cdot|s)$ are independent for different (s,h)-pairs. Moreover, $\mathbb{E}[\widehat{\pi}_h(\cdot|s)] = \pi_h(\cdot|s)$, where the expectation is taken w.r.t. the randomness of the episodes. We are going to prove by induction that $\mathbb{E}\left[\widehat{d}_{i,h}(s)\right] = d_{i,h}(s)$. Let us consider the case $h = 2$:

$$\mathbb{E}\left[\widehat{d}_{i,2}(s)\right] = \sum_{s'\in S} \mu(s') p(s|s',\cdot)^T\,\mathbb{E}[\widehat{\pi}_{i,1}(\cdot|s')] = d_{i,1}(s).$$

By induction, suppose that the statement hold for all $h' \leq h$, we prove it for $h+1$:

$$\begin{aligned}
\mathbb{E}\left[\widehat{d}_{i,h+1}(s)\right] &= \sum_{s'\in S} \mathbb{E}\left[\widehat{d}_{i,h}(s') p(s|s',\cdot)^T \widehat{\pi}_{i,h}(\cdot|s')\right] \\
&= \sum_{s'\in S} \mathbb{E}\left[\widehat{d}_{i,h}(s')\right] p(s|s',\cdot)^T\,\mathbb{E}\left[\widehat{\pi}_{i,h}(\cdot|s')\right] \\
&= \sum_{s'\in S} d_{i,h}(s') p(s|s',\cdot)^T \pi_{i,h}(\cdot|s') = d_{i,h+1}(s),
\end{aligned}$$

where the last but one line derives from the fact that $\widehat{d}_{i,h}$ and $\widehat{\pi}_{i,h}$ are independent. This is due to the fact that $\widehat{d}_{i,h}$ depends on the policies $\{\widehat{\pi}_{i,h'}\}$ for $h' < h$ only that, in turn, are independent from $\{\widehat{\pi}_{i,h'}\}$ as noted at the

beginning of the proof. The last line follows from the inductive hypothesis. For the second statement, we have:

$$\mathbb{E}[\rho_i] = \mathbb{E}\left[\max_{s\in\mathcal{S}}\max_{h\in[H]}\frac{\widehat{d}_{i,h}(s)}{d_{i,h}(s)}\right] \leq \sum_{s\in\mathcal{S}}\sum_{h\in[H]}\mathbb{E}\left[\frac{\widehat{d}_{i,h}(s)}{d_{i,h}(s)}\right] = SH,$$

having exploited the first statement. □

B.2 Proofs of Section 5.2

In this section, we are going to prove the regret bound RfOCL. In this second algorithm the configurator can observe at every episode also a realization of the agent's reward function. In the following we will show how the algorithm exploits this information under Assumption 1.

We start defining the good events G_k for $k \in [K]$:

$$G_k = \left\{\forall s \in \mathcal{S}, |\widehat{r}_{o,k}(s) - r_o(s)| \leq \sqrt{\frac{\log(2SHk^2)}{2N_k(s)}}\right\}$$

The event G_k means that, at episode $k \in [K]$, the estimated rewards of each state $s \in \mathcal{S}$ are inside the confidence intervals.

Lemma B.4. *For every configuration $p_i \in \mathcal{P}$ and state action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, the difference between the optimistic state-action value function $\overline{Q}^i_{o,k,1}(s,a)$ and the true optimal state-action value function $Q^i_{o,1}(s,a)$ is bounded by:*

$$\overline{Q}^i_{o,k,1}(s,a) - Q^i_{o,1}(s,a) \leq \overline{r}_{o,k}(s) - r_o(s) + \sum_{s'\in\mathcal{S}}\sum_{h=2}^{H}\overline{d}^i_{k,h}(s')\left(\overline{r}_{o,k}(s') - r_o(s')\right),$$

where $\overline{d}^i_{k,h}$ the visitation distribution induced by a greedy policy $\overline{\pi}_{i,k}$ w.r.t. $\overline{Q}^i_{o,k}$. Similarly, the difference between the true optimal state-action value function $Q^i_{o,1}(s,a)$ and the pessimistic state-action value function $\underline{Q}^i_{o,k,1}(s,a)$ is bounded by:

$$Q^i_{o,1}(s,a) - \underline{Q}^i_{o,k,1}(s,a) \leq r_o(s) - \underline{r}_{o,k}(s) + \sum_{s'\in\mathcal{S}}\sum_{h=2}^{H} d^i_{k,h}(s')\left(r_o(s') - \underline{r}_{o,k}(s')\right).$$

Proof. The proof is basically taken from [46, 3, 42]:

$$\begin{aligned}\overline{Q}^i_{o,k,1}(s,a) - Q^i_{o,1}(s,a) &\leq \overline{Q}^i_{o,k,1}(s,a) - Q^{\overline{\pi}_{i,k}}_{o,1}(s,a) && \text{(P.21)}\\ &= \overline{r}_{o,k}(s) - r_o(s) + \sum_{s'\in\mathcal{S}}\sum_{h=2}^{H}\overline{d}^i_{k,h}(s')\left(\overline{r}_{o,k}(s') - r_o(s')\right). && \text{(P.22)}\end{aligned}$$

where line (P.21) is due to $Q^i_{o,1}(s,a) \geq Q^{\overline{\pi}_{i,k}}_{o,1}(s,a)$, recalling that $Q^i_{o,1}$ is the optimal Q-value for the agent, under configuration p_i and the optimal agent's policy. Line (P.21) derives form the application of the simulation lemma since $\overline{Q}^i_{o,k,1}(s,a)$ and $Q^{\overline{\pi}_{i,k}}_{o,1}(s,a)$ are under the same policy $\overline{\pi}_{i,k}$. For the second statement, we proceed analogously by simply observing that $\underline{Q}^i_{o,k,1}(s,a) \geq \underline{Q}^{\pi_i}_{o,1}(s,a)$ where π_i is a greedy policy w.r.t. $Q^i_{o,k}(s,a)$. □

Lemma B.5. *If for all $k \in [K]$, the good events G_k hold, for all state-action pairs $(s,a) \in \mathcal{S}\times\mathcal{A}$, $h \in [H]$, and configuration $p_i \in \mathcal{P}$ it holds that:*

$$\overline{Q}^i_{o,k,1}(s,a) - Q^i_{o,1}(s,a) \leq SH\sqrt{\frac{\log(2SHk^2)}{2N_k(s)}},$$

$$Q^i_{o,1}(s,a) - \underline{Q}^i_{o,k,1}(s,a) \leq SH\sqrt{\frac{\log(2SHk^2)}{2N_k(s)}}.$$

Proof. We apply Lemma B.4, recall that $\overline{r}_{o,k}(s) = \widehat{r}_{o,k}(s) + \sqrt{\frac{\log(2SHk^2)}{2N_k(s)}}$ and $\underline{r}_{o,k}(s) = \widehat{r}_{o,k}(s) - \sqrt{\frac{\log(2SHk^2)}{2N_k(s)}}$, and make use of the definition of the events G_k. Then, we bound the visitation distribution with 1. □

Lemma B.6. *Let $s \in \mathcal{S}$ be a state with minimum visitation probability $d(s) := \min_{i\in[M]} \max_{h\in[H]} d_h^i(s) > 0$. Then, at episode $k \in [K]$, for every $\delta_k \in (0,1)$, with probability at least $1-\delta_k$ it holds that:*

$$N_k(s) \geq (k-1)d(s) - \sqrt{\frac{k-1}{2}\log\left(\frac{1}{\delta_k}\right)}.$$

Proof. First of all, we define the random variable $N_k^u(s)$ as the count of the visits to state s, where multiple visits in the same episode are considered just once:

$$N_k^u(s) = \sum_{i=1}^{k-1} \mathbb{1}\left\{\exists h \in [H] \,:\, s_{k,h} = s\right\}.$$

Clearly, $N_k^u(s) \leq N_k(s)$ and, consequently, $\mathbb{E}[N_k^u(s)] \leq \mathbb{E}[N_k(s)]$. The expectation of $\mathbb{E}[N_k^u(s)]$ can be bounded as:

$$\begin{aligned}
\mathbb{E}[N_k^u(s)] &= \mathbb{E}\left[\sum_{i=1}^{k-1} \mathbb{1}\left\{\exists h \in [H] \,:\, s_{k,h} = s\right\}\right] \\
&= \sum_{i=1}^{k-1} \Pr\left(\exists h \in [H] \,:\, s_{k,h} = s | p_{I_k}, \pi_{I_k}\right) && \text{(P.23)} \\
&= \sum_{i=1}^{k-1} \Pr\left(\bigcup_{h\in[H]} \{s_{k,h} = s\} | p_{I_k}, \pi_{I_k}\right) && \text{(P.24)} \\
&\geq \sum_{i=1}^{k-1} \max_{h\in[H]} \Pr\left(s_{k,h} = s | p_{I_k}, \pi_{I_k}\right) && \text{(P.25)} \\
&= \sum_{i=1}^{k-1} \max_{h\in[H]} d_h^{I_k}(s) && \text{(P.26)} \\
&\geq (k-1) \min_{i\in[M]} \max_{h\in[H]} d_h^i(s) = (k-1)d(s), && \text{(P.27)}
\end{aligned}$$

where line (P.23) and line (P.24) we simply rewrite the expectation as probability. In line (P.25) we bound the probability of the union with just one term. In line (P.26) we employ the definition of $d_h^{I_k}(s)$. Finally, in line (P.27), we take the minimum over I_k. Since $0 \leq N_k^u(s) \leq k-1$, by using Höeffding's inequality, we have that with probability at least $1-\delta_k$ it holds that:

$$N_k^u(s) \geq \mathbb{E}[N_k^u(s)] - \sqrt{\frac{k-1}{2}\log\frac{1}{\delta_k}} \geq (k-1)d(s) - \sqrt{\frac{k-1}{2}\log\frac{1}{\delta_k}},$$

having used the lower bound on $\mathbb{E}[N_k^u(s)]$. The result follows from recalling that$\mathbb{E}[N_k^u(s)] \leq \mathbb{E}[N_k(s)]$. □

Lemma B.7. *If for all $k \in [K]$, the good events G_k hold, and for all $s \in \mathcal{S}$ it holds that $\sqrt{\frac{\log(2SHk^2)}{2N_k(s)}} \leq \frac{\Delta_Q - c}{2SH}$, with arbitrary $c > 0$, then for every configuration $p_i \in \mathcal{P}$ we have that $\widetilde{\pi}_{i,k} = \pi_i$.*

Proof. Let Δ_Q be the minimum gap between the Q-function in the optimal action and a different action in all transition probabilities $p_i \in \mathcal{P}$:

$$\Delta_Q = \min_{i\in[M]} \min_{s\in\mathcal{S}} \min_{h\in[H]} \left\{ \max_{a\in\mathcal{A}} Q_{o,h}^i(s,a) - \max_{a'\in\mathcal{A}\setminus \arg\max_{a\in\mathcal{A}} Q_{o,h}^i(s,a)} Q_{o,h}^i(s,a') \right\}.$$

For all $s \in \mathcal{S}$ and $h \in [H]$, we denote with $a^* = \arg\max_{a \in \mathcal{A}} Q^i_{o,h}(s,a)$ and we have for all $a \in \mathcal{A} \setminus \{a^*\}$:

$$\begin{aligned}
\overline{Q}^i_{o,k,h}(s,a) - \underline{Q}^i_{o,k,h}(s,a^*) &= \overline{Q}^i_{o,k,h}(s,a) - \underline{Q}^i_{o,k,h}(s,a^*) \pm Q^i_{o,h}(s,a) \pm Q^i_{o,h}(s,a^*) \\
&= \underbrace{\overline{Q}^i_{o,k,h}(s,a) - Q^i_{o,h}(s,a)}_{(A)} + \underbrace{Q^i_{o,h}(s,a^*) - \underline{Q}^i_{o,k,h}(s,a^*)}_{(B)} \\
&\quad + \underbrace{Q^i_{o,h}(s,a) - Q^i_{o,h}(s,a^*)}_{(C)} \\
&\leq 2SH\sqrt{\frac{\log(2SHk^2)}{2N_k(s)}} - \Delta_Q \\
&\leq 2SH\frac{\Delta_Q - c}{2SH} - \Delta_Q \leq -c,
\end{aligned}$$

where for (A) and (B) we applied Lemma B.5 and for (C) we used the definition of Δ_Q. We have proved that the lower bound on the Q-value of the optimal action $\underline{Q}^i_{o,k,h}(s,a^*)$ falls above the upper bound on the Q-value of all other actions $\overline{Q}^i_{o,k,h}(s,a)$. Consequently, the greedy action will be properly identified and $\widetilde{\pi}_{i,k} = \pi_i$. □

Theorem 5.2 (Regret of RfOCL)**.** *Let $\mathcal{NCM} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mu, r_c, r_o, H)$ with $\mathcal{P} = \{p_1, \ldots, p_M\}$ be the M configurations. Under Assumption 1, the expected regret of RfOCL at every episode $K > 0$ is bounded by:*

$$\mathbb{E}[\textit{Regret(K)}] \leq \mathcal{O}\left(\min\left\{ \underbrace{H^2 \sum_{i \in [M]: \Delta_i > 0} \frac{\log(K)}{\Delta_i}}_{\textit{UCB1 regret}} , \underbrace{MH^3S^2\rho}_{\textit{AfOCL regret}} , \underbrace{\overline{K}\Delta + \frac{\pi^2}{3}}_{\textit{RfOCL regret}} \right\} \right),$$

where ρ is defined as in Theorem 5.1, $\overline{K}$ is the smallest integer solution of the inequality $\overline{K} \geq 1 + \left(\frac{2H^2S^2\log(2SH\overline{K}^2)}{2\Delta_Q^2} + \sqrt{\frac{\overline{K}-1}{2}\log(SH\overline{K}^2)} \right)\frac{1}{\epsilon}$, $\Delta = \max_{i \in [M]} \Delta_i$, i.e., the maximum suboptimality gap, and Δ_Q is the minimum positive gap of the agent's Q-values (see Appendix B).

Proof. We rewrite the expected regret as follows:

$$\begin{aligned}
\mathbb{E}[\text{Regret}(K)] &= \sum_{k=1}^{K} \left(\mathbb{E}[\Delta_{I_k} \mathbb{1}\{G_k\}] + \mathbb{E}[\Delta_{I_k} \mathbb{1}\{\neg G_k\}] \right) \\
&\leq \underbrace{\sum_{k=1}^{K} \mathbb{E}[\Delta_{I_k} | G_k]}_{(A)} + \underbrace{H \sum_{k=1}^{K} \Pr(\neg G_k)}_{(B)},
\end{aligned}$$

where we bounded $\Pr(G_k) \leq 1$ in term (A) and Δ_{I_k} with its maximum value H in term (B). We start bounding the (B) term:

$$\begin{aligned}
H \sum_{k=1}^{K} \Pr(\neg G_k) &= H \sum_{k=1}^{K} \Pr\left(\exists s \in \mathcal{S} \text{ s.t. } |\widehat{r}_{o,k}(s) - r(s)| > \sqrt{\frac{\log(2SHk^2)}{2N_k(s)}} \right) && \text{(P.28)} \\
&\leq H \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \Pr\left(|\widehat{r}_{o,k}(s) - r(s)| > \sqrt{\frac{\log(2SHk^2)}{2N_k(s)}} \right) && \text{(P.29)} \\
&\leq H \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \frac{1}{SHk^2} \leq \frac{\pi^2}{6}, && \text{(P.30)}
\end{aligned}$$

where line (P.28) follows from the definition of the good event G_k. Line (P.29) is a union bound on the states. Line (P.30) comes from Höeffding's inequality.

For the first term (A) we define the event E_k for all $k \in [K]$:

$$E_k = \left\{ \forall s \in \mathcal{S} \,:\, N_k(s) \geq (k-1)d(s) - \sqrt{\frac{k-1}{2}\log\left(SHk^2\right)} \right\}.$$

If this event holds then every state $s \in \mathcal{S}$ is visited at least $(k-1)d(s) - \sqrt{\frac{k}{2}\log(SHk^2)}$ times, where $d(s)$ is defined as in Lemma B.6.

Considering the term (A), we have:

$$\sum_{k=1}^{K} \mathbb{E}[\Delta_{I_k}|G_k] \leq \underbrace{\sum_{k=1}^{K} \mathbb{E}[\Delta_{I_k}|G_k, E_k]}_{\text{(C)}} + \underbrace{H \sum_{k=1}^{K} \Pr(\neg E_k)}_{\text{(D)}},$$

where we bound the in the second term $\Delta_{I_k} \leq H$.

We start bounding the second term (D). We apply Lemma B.6 after a union bound over the states:

$$\begin{aligned} H \sum_{k=1}^{K} \Pr(\neg E_k) &= H \sum_{k=1}^{K} \Pr\left(\exists s \in \mathcal{S} : N_k(s) < (k-1)d(s) - \sqrt{\frac{k-1}{2}\log(SHk^2)}\right) \\ &\leq H \sum_{s\in\mathcal{S}} \sum_{k=1}^{K} \Pr\left(N_k(s) < (k-1)d(s) - \sqrt{\frac{k-1}{2}\log(SHk^2)}\right) \\ &\leq H \sum_{s\in\mathcal{S}} \sum_{k=1}^{K} \frac{1}{SHk^2} \leq \frac{\pi^2}{6}. \end{aligned}$$

Now it remains to bound the term (C) that, using Lemma B.7, is zero whenever $\sqrt{\frac{\log(2SHk^2)}{2N_k(s)}} \leq \frac{\Delta_Q - c}{2SH}$. Thus, under the events E_k and recalling that under Assumption 1 we have $d(s) \geq \epsilon$, we obtain:

$$\sqrt{\frac{\log(2SHk^2)}{2N_k(s)}} \leq \sqrt{\frac{\log(2SHk^2)}{2(k-1)\epsilon - \sqrt{2(k-1)\log(SHk^2)}}}.$$

From which, we derive the condition:

$$\overline{K} \geq 1 + \left(\frac{2H^2S^2\log(2SH\overline{K}^2)}{2(\Delta_Q - c)^2} + \sqrt{\frac{\overline{K}-1}{2}\log(SH\overline{K}^2)}\right)\frac{1}{\epsilon}.$$

Then, we take the infimum over c. Thus, for the term (C), we consider the decomposition:

$$\sum_{k=1}^{K} \mathbb{E}[\Delta_{I_k}|G_k, E_k] \leq \sum_{k=1}^{\overline{K}} \mathbb{E}[\Delta_{I_k}|G_k, E_k] + \sum_{k=\overline{K}+1}^{\infty} \mathbb{E}[\Delta_{I_k}|G_k, E_k] = \overline{K}\Delta + 0,$$

where we bounded $\Delta_{I_k} \leq \Delta$ with $\Delta = \max_{i\in[M]} \Delta_i$. Then the total regret is given by:

$$\mathbb{E}[\text{Regret}(K)] \leq \overline{K}\Delta + \frac{\pi^2}{3}.$$

□

C Adversarial agent

In this paragraph, we provide some hints about the adversarial case, to illustrate the additional complexities that arise. In the adversarial setting, the agent can play a different policy at each step, inside the set of possible policies that satisfy the SSE, namely Π_i^{SSE}. For the Af setting, to have bounded regret, we have to add the following assumption.

Assumption 2. *For all $i \in [M]$, let $\pi, \pi' \in \Pi_i^{SSE}$, then $d_h^{i,\pi}(s) > 0$ if and only if $d_h^{i,\pi'}(s) > 0$.*

Under this assumption (less strict than Assumption 1), Theorem 5.1 continues to hold. Indeed, though the agent can adversarially change the policies, we can still define the policy $\widehat{\pi}$, since the policies in the set Π^{SSE} do not disconnect the reachable set of states. On the other hand, without this assumption, the algorithm needs some modifications, since the agent can stuck the configurator with actions that after some episodes will not be played any more; this behavior can lead to an estimated policy that visits unreachable states.

For the Rf, instead, under the following assumption (that is a natural extension of Assumption 1), Theorem 5.2 continues to apply.

Assumption 3. *There exists* $\epsilon > 0$ *such that:* $\min_{i \in [M]} \min_{s \in \mathcal{S}} \max_{h \in [H]} d_h^i(s) \geq \epsilon$, *where* $d_h^i(s)$ *is the probability of visiting the state* $s \in \mathcal{S}$ *at time* $h \in [H]$ *in configuration* p_i *under every agent's best response policy* $\pi_i \in \Pi_i^{SSE}$.

In this case the reward continues to give the structure to connect the policies and the models. However, we believe that to solve the adversarial case without these assumptions would require modifying the algorithm, and it is left to future work.

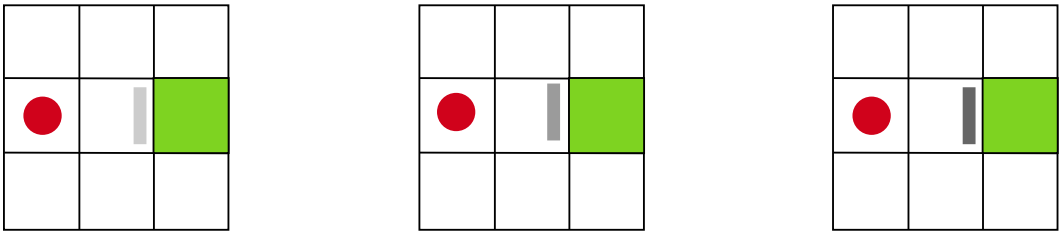
D Experimental Details

In this appendix, we report additional experimental details and results.

D.1 Configurable Gridworld

Description In Figure 4 the environment of the Configurable Gridworld is shown. The configurable Gridworld is a 3×3 gridworld with an obstacle in the cell $(2, 2)$, which with a probability p causes the agent action *right* not to be performed. The starting state is in every configuration $(1, 2)$ and the goal state is $(3, 2)$.

Additional Experiments We report additional experiments for the Configurable Gridworld environment. For the Configurable Gridworld with size 3×3, horizon 10, we perform 4 experiments with an increasing number of configurations. In this case the expert policy is deterministic. Figure 5 shows the results of the experiments. We can notice that with more than 100 configurations AfOCL does not achieve constant regret in 5000 steps, instead RfOCL converges in every experiment.



Configuration #1 Configuration #2 Configuration #3

Figure 4: Configurable Gridworld: from left to right the 3 configurations represent increasing "power" of the obstacle.

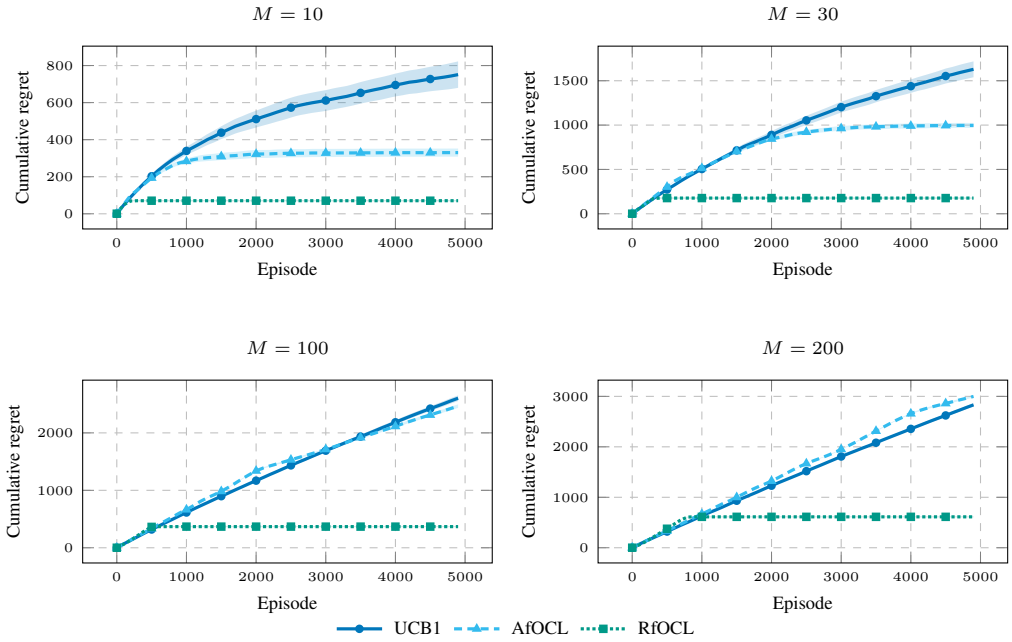


Figure 5: From left up to right down $10, 30, 100, 200$ configurations' number.

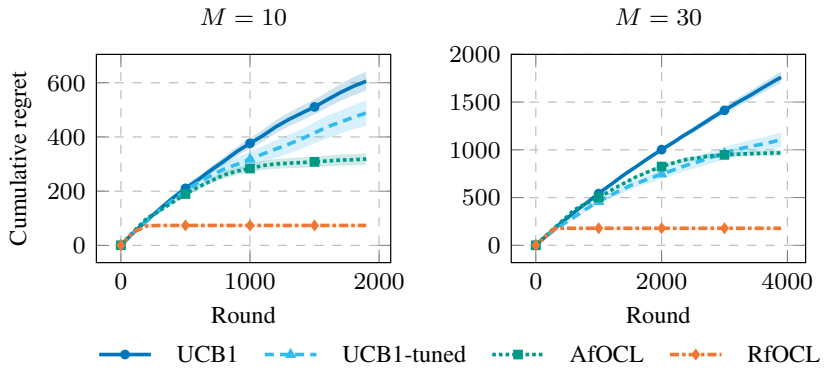


Figure 6: Cumulative regret for the Gridworld experiment. 50 runs, 98% c.i.

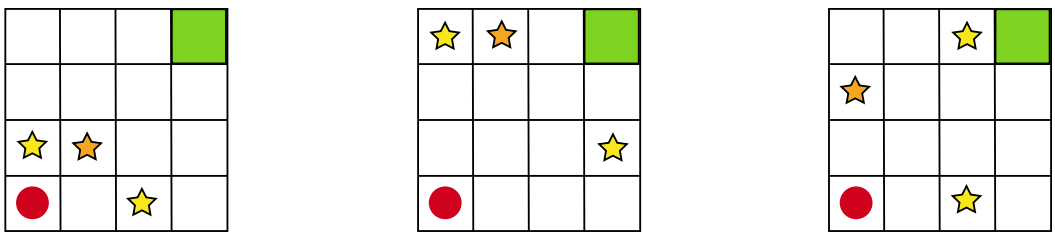
We report also the same experiment shown in the main paper with a tuned-version of UCB1 (see figure 6). We would like to underline that, also in this case, the two proposed algorithm AfOCL and RfOCL achieved a constant behavior while UCB1-tuned has a logarithmic behavior.

D.2 Market

A Configurable Market is a simplified model for a marketplace. The agent, namely the customer, wants to buy a given set of products Q_A in the minimum number of steps. Instead, the configurator has the role in placing all the products $Q \supset Q_A$ in the marketplace to maximize the market's revenue inducing the agent to buy other products in addition to those it would buy. The configurator's reward is 1 any time the agent passes over a state where a product is placed and 0 in all the other states. Whereas the agent's reward is -1 everywhere and gains a bonus of 0.9 when it passes over a state with a product in Q_A. In other words, the products remain fixed in the market, and the configurator can change the transition model within a set of random transition models. However, from an abstract point of view, this is equivalent to moving the products in the Gridworld.

In Figure 7 the market domain with 3 different configurations is shown. The market domains consists in $K \times K$ states, where every product is assigned to a specific state. The configurator can change the transition matrix for all the states except for the starting state and the "exit" state. Every different configuration can be thought as shuffling the cells of a gridworld.

In Figure 8, AfOCL and RfOCL are compared against UCB1. The number of configurations is 10, the horizon 15, and the Gridworld size is 4×4. In every run, we construct 10 different transition models, which specify the 10 configurations. Also, in this experiment, the trend is confirmed since AfOCL and RfOCL outperform UCB1. We observe that the two algorithms, in this environment, behave similarly, and this is due to the small number of configurations. However, we can notice RfOCL at the end of the considered episodes approaches the constant regret.



Configuration #1 Configuration #2 Configuration #3

Figure 7: Market: the figure shows a 5×5 market. The red state is the starting state, instead the green state is the "end" state. The stars are the product and the orange star is the only product the agent is interested in.

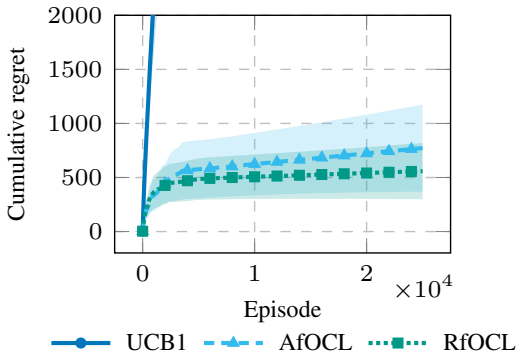
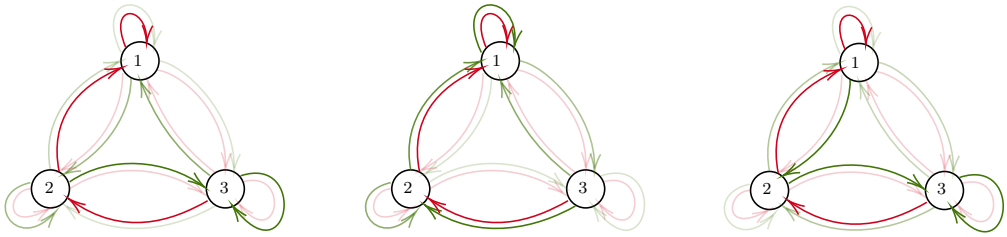


Figure 8: Cumulative regret as a function of the episodes for the Configurable Market experiment. 50 runs, 98% c.i.

D.3 Teacher Student

In Figure 9 an illustrative example of the Teacher-Student domain is reported. Right arrows correspond to answer No, and green arrows to answer Yes. The transparency is due to the level of probability of every transition. The configurator can change the transition matrix for the answer Yes, instead the transition matrix for action No is fixed for all the configurations.



Configuration #1 Configuration #2 Configuration #3

Figure 9: Teacher Student.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See conclusion.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] The paper is mostly a theoretical contribution and we think it will not have any potential negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions are all given in the main paper.
   (b) Did you include complete proofs of all theoretical results? [Yes] See appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See figures.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]