Optimism in Face of a Context: Regret Guarantees for Stochastic Contextual MDP

Orin Levy[1], **Yishay Mansour**[1,2]
[1] Tel Aviv University
[2] Google Research, Tel Aviv
orinlevy@mail.tau.ac.il, mansour.yishay@gmail.com

Abstract

We present regret minimization algorithms for stochastic contextual MDPs under minimum reachability assumption, using an access to an offline least square regression oracle. We analyze three different settings: where the dynamics is known, where the dynamics is unknown but independent of the context and the most challenging setting where the dynamics is unknown and context-dependent. For the latter, our algorithm obtains regret bound of $\widetilde{O}((H + 1/p_{min})H|S|^{3/2}\sqrt{|A|T\log(\max\{|\mathcal{G}|,|\mathcal{P}|\}/\delta)})$ with probability $1-\delta$, where $\mathcal{P}$ and $\mathcal{G}$ are finite and realizable function classes used to approximate the dynamics and rewards respectively, p_{min} is the minimum reachability parameter, S is the set of states, A the set of actions, H the horizon, and T the number of episodes. To our knowledge, our approach is the first optimistic approach applied to contextual MDPs with general function approximation (i.e., without additional knowledge regarding the function class, such as it being linear and etc.). We present a lower bound of $\Omega(\sqrt{TH|S||A|\ln(|\mathcal{G}|)/\ln(|A|)})$, on the expected regret which holds even in the case of known dynamics. Lastly, we discuss an extension of our results to CMDPs without minimum reachability, that obtains $\widetilde{O}(T^{3/4})$ regret.

1 Introduction

Markov decision processes (MDPs) have been extensively studied, and are commonly used to describe dynamic environments. MDPs characterize a variety of real-life tasks and applications including: advertising, healthcare, games, robotics and more, where at each episode an agent interacts with the environment with the goal of maximizing her return. (See, e.g., Sutton and Barto (2018); Mannor, Mansour, and Tamar (2022).)

In many applications, in each episode, there are additional external factors that affect the environment, which we refer to as the *context*. One way to handle this is to extend the state space to include the context. This approach has the disadvantage of greatly increasing the state space, and, as a result, the complexity of learning and even the representation of a policy. An alternative approach, is to keep a small state space, and regard the context as an additional side-information. Contextual Markov Decision Process (CMDP) describes such a model, where for each context there is a potentially different optimal policy (Hallak, Di Castro, and Mannor 2015).

CMDPs are useful to model many user-driven applications, where the context is a user-related information which influences the optimal decision making. One natural application is in recommendation systems, where two different users might behave completely different from one another, hence, a single MDP can not describe them both. In those systems, users behavior can be described using a side information about them, such as age, gender, interest fields and hobbies. This information is referred to as the *context* which influences the environment. CMDP defines a mapping from context to a related MDP, and the optimal policy given a context is the optimal policy in the related MDP.

Our contributions. We present regret minimization algorithms for CMDP under three different settings: (1) known dynamics, (2) unknown context-independent dynamics and (3) unknown context-dependent dynamics, which is the most challenging. In all settings we assume an access a least square regression oracle, and finite function classes $\mathcal{G}$ and $\mathcal{P}$ used to approximate the rewards and dynamics, respectively. In addition, we assume minimum reachability, where any policy for any context has a probability of at least p_{min} to reach any state. For the known dynamics setting we obtain $\widetilde{O}((H+1/p_{min})\cdot|S|\sqrt{T|A|\log(|\mathcal{G}|/\delta)})$ regret. For the unknown context-independent dynamics we obtain $\widetilde{O}\Big(H^{1.5}|S|\sqrt{T|A|}\log(1/\delta) + (H+1/p_{min})\cdot|S|\sqrt{|A|T\log(|\mathcal{G}|/\delta)}\Big)$ regret. For the unknown context-dependent dynamics we obtain regret of $\widetilde{O}((H+1/p_{min})\cdot H|S|^{3/2}\sqrt{|A|T\log(\max\{|\mathcal{G}|,|\mathcal{P}|\}/\delta)})$. All of the bounds hold with high probability. We also show a lower bound of $\Omega(\sqrt{TH|S||A|\ln(|\mathcal{G}|)/\ln(|A|)})$ on the expected regret. Lastly, we discuss an extension of our results to CMDP without minimum reachability, that obtains $\widetilde{O}(T^{3/4})$ regret, in Section 7.

Our approach applies the "optimism in face of uncertainty" principle to CMDPs and achieves a sub-linear regret. Our algorithms and analysis were inspired by the optimistic approach of Xu and Zeevi (2020) for learning contextual multi armed bandits using least square regression oracle. We extended their approach to handle CMDPs and even a context-dependent dynamics.

Related Work

Contextual Reinforcement Leaning. CMDP was first introduce by Hallak, Di Castro, and Mannor (2015). Modi et al. (2018) gives a general framework for deriving generalization bounds for smooth CMDPs and finite contextual linear combination of MDPs. Modi and Tewari (2020) gives a regret bound of $\widetilde{O}(\sqrt{T})$ for Generalized Linear Models (GLMs). Our regret function approximation framework is more general than GLM.

Foster et al. (2021) present a new statistical complexity measure for interactive decision making, and show an application of it to obtain $\widetilde{O}(\sqrt{T})$ regret for Contextual RL. They assume an access to an online estimation oracle with regret guarantees, that maximizes over models and policies together. It is unclear when is this oracle implementable in polynomial time. In contrast, we make a significantly weaker and standard assumption regarding an offline regression oracle. Another difference is that we use an optimistic approach while they use inverse gap weighting. (More details later.)

Jiang et al. (2017) present OLIVE which is sample efficient for Contextual Decision Processes (CDP) with low Bellman rank. We do not make any assumptions on the Bellman rank.

Levy and Mansour (2022a) consider the sample complexity of learning CMDPs using function approximation. They provide the first general and efficient reduction from CMDP to offline supervised learning. Their sample complexity varies from $\widetilde{O}(1/\epsilon^2)$ to $\widetilde{O}(1/\epsilon^8)$, depending on the setting. We, in contrast, consider regret minimization and obtain $\widetilde{O}(\sqrt{T})$ regret under the minimum reachability assumption.

Contextual Bandits. Contextual bandits (CMAB) are a natural extension of the Multi-Arm Bandit (MAB), augmented by a context which influences the rewards (Slivkins 2019; Lattimore and Szepesvári 2020). Agarwal et al. (2014) use efficiently an optimization oracle to derive an optimal regret bound. Regression based approaches appear in Agarwal et al. (2012); Foster et al. (2018); Foster and Rakhlin (2020); Simchi-Levi and Xu (2021). We differ from CMAB, since our main challenge is the dynamics, and the need to optimize future rewards, which is the case in most RL settings.

Xu and Zeevi (2020) present the first optimistic algorithm for CMAB. They assume an access to a least-square regression oracle and achieve $\widetilde{O}(\sqrt{T|A|\log|\mathcal{F}|})$ regret, where $\mathcal{F}$ is a finite and realizable function class uses to approximate the rewards. Our algorithms and analysis are inspired by their optimistic approach and we extend it to CMDP.

Inverse Gap Weighting (IGW) technique. Foster and Rakhlin (2020); Simchi-Levi and Xu (2021) apply the IGW technique to CMAB and obtain $\widetilde{O}(\sqrt{T|A|\log|\mathcal{F}|})$ regret, assuming an access to a least square regression oracle. However, we do not see any straight-forward extension of their approach to CMDP which is both computationally efficient and has an optimal regret, under the same least-square oracle assumption (even when the dynamics is known to the learner). Foster et al. (2021) apply IGW to CMDP and obtain optimal regret. However they use the much strong online estimation oracle as discussed above.

Paper organization. Section 2 contains the notations we use, and our assumptions. Sections 3 to 5 contain an outline of our algorithms and regret analysis for each one of the settings. Section 6 presents our lower bound and Section 7 sketches an extension of our result to CMDPs without minimum reachability. We discuss our results in Section 8. The supplementary material of this work can be found in Levy and Mansour (2022b).

2 Preliminaries and Notations

Markov Decision Process (MDP) is a tuple (S, A, P, r, s_0, H), where (1) S is a finite state space, (2) A is a finite action space, (3) $s_0 \in S$ is the unique start state, (4) $P(\cdot|s,a)$ defines the transition probability function, i.e., $P(s'|s,a)$ is the probability that we reach state s' given that we are in state s and perform action a, (5) $R(s,a) \in [0,1]$ is a random variable for the reward of performing action a in state s, and $r(s,a)$ is its expectation, i.e., $r(s,a) = \mathbb{E}[R(s,a)|s,a]$, and (6) H is the finite horizon.

The state space is decomposed into $H+1$ disjoint subsets (layers) $S_0, S_1, \ldots, S_{H-1}, S_H$ such that transitions are only possible between consecutive layers (i.e., loop-free). There is a unique final state, i.e., $S_H = \{s_H\}$, with reward 0.

Policy. A *stochastic policy* π is a mapping from states to distribution over actions, i.e., $\pi : S \to \Delta(A)$. A *deterministic policy* π is a mapping from states to actions, i.e., $\pi : S \to A$.

Occupancy measure (see e.g., Puterman (2014); Zimin and Neu (2013)). Let $q_h(s,a|\pi,P)$ denote the probability of reaching state $s \in S$ and performing action $a \in A$ at time $h \in [H]$ of an episode generated using policy π and dynamics P. Let $q_h(s|\pi,P) = \sum_{a\in A} q_h(s,a|\pi,P)$ be the probability to visit state $s \in S$ at time h.

Episode and trajectory. At the start of each episode we select a policy π. The episode starts at the unique initial state s_0. In state $s_h \in S_h$, we play action $a_h \sim \pi(\cdot|s_h)$, observe a reward $r_h \sim R(s_h,a_h)$ and move to $s_{h+1} \sim P(\cdot|s_h,a_h)$. We generate a trajectory $\sigma_{H+1} = (s_0, a_0, r_0, s_1, \ldots, s_{H-1}, a_{H-1}, r_{H-1}, s_H)$ of length $H+1$.

Value functions. Given a policy π and a MDP $M = (S, A, P, r, s_0, H)$, the $h \in [H-1]$ stage value function of a state $s \in S_h$ is defined as $V^{\pi}_{M,h}(s) = \mathbb{E}_{\pi,M}[\sum_{k=h}^{H-1} r(s_k, \pi(s_k))|s_h = s]$ and for an action $a \in A$ we have $Q^{\pi}_{M,h}(s,a) = \mathbb{E}_{\pi,M}[\sum_{k=h}^{H-1} r(s_k,\pi(s_k))|s_h = s, a_h = a]$. When $h = 0$ we denote $V^{\pi}_{M,0}(s_0) := V^{\pi}_M(s_0)$.

Optimal policy $\pi^{\star}_M$ for MDP M satisfies, for every stage $h \in [H-1]$ and a state $s \in S_h$, $\pi^{\star}_{M,h}(s) \in \arg\max_{\pi}\{V^{\pi}_{M,h}(s)\}$, and w.l.o.g it is deterministic.

Planning. Given an MDP $M = (S, A, P, r, s_0, H)$ the algorithm `Planning`(M) returns an optimal policy $\pi^{\star}_M$ and its value $V^{\star}_M(s_0)$ and runs in time $O(|S|^2\,|A|\,H)$.

Contextual Markov Decision Process (CMDP) is a tuple $(\mathcal{C}, S, A, \mathcal{M})$ where $\mathcal{C} \subseteq \mathbb{R}^{d'}$ is the context space, S the state space and A the action space. The mapping $\mathcal{M}$ maps a context $c \in \mathcal{C}$ to a MDP $\mathcal{M}(c) = (S, A, P^c_{\star}, r^c_{\star}, s_0, H)$, where $r^c_{\star}(s,a) = \mathbb{E}[R^c_{\star}(s,a)|c,s,a]$, $R^c_{\star}(s,a) \sim \mathcal{D}_{c,s,a}$.

There is an unknown distribution $\mathcal{D}$ over the context space $\mathcal{C}$, and for each episode a context is sampled i.i.d. from $\mathcal{D}$. For mathematical convenience, we assume the context space

is finite (but potentially huge). Our results naturally extend to infinite contexts space.

Context-Independent and Context-Dependent dynamics. A CMDP has a *context-independent* dynamics when the context effects only the rewards function, while the dynamics are identical for all contexts, i.e., $P^c_\star = P$ for any context c. A *context-dependent* dynamics has a potentially different dynamics $P^c_\star$ for each context c. Hence, the partition of the states space into layers is also context-dependent. We denote by S^c_h the h layer of context c.

Context-dependent policies. A stochastic context-dependent policy $\pi = (\pi(c;\cdot) : S \to \Delta(A))_{c\in\mathcal{C}}$ maps a context $c \in \mathcal{C}$ to a stochastic policy $\pi(c;\cdot) : S \to \Delta(A)$. A deterministic context-dependent policy $\pi = (\pi(c;\cdot) : S \to A))_{c\in\mathcal{C}}$ maps a context $c \in \mathcal{C}$ to a policy $\pi(c;\cdot) : S \to A$. Let $\Pi_\mathcal{C}$ denote the class of all deterministic context-dependent policies. A context-dependent policy $\pi^\star \in \Pi_\mathcal{C}$ is *optimal* if for all $c \in \mathcal{C}$ it holds that $\pi^\star(c;\cdot) \in \arg\max_\pi V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)$ [1].

Minimum reachability. We assume that there exists $p_{min} \in (0, 1]$ such that for every $c \in \mathcal{C}$, $h \in [H-1]$ and $s_h \in S^c_h$, any context-dependent policy $\pi \in \Pi_\mathcal{C}$ satisfies $q_h(s_h|\pi(c;\cdot), P^c_\star) \geq p_{min}$. Let $q(s|\pi(c;\cdot), P^c_\star)$ denote the probability of visiting state s when playing π on the dynamics $P^c_\star$. When the dynamics is layered and loop-free, then $q(s|\pi(c;\cdot), P^c_\star) = q_h(s|\pi(c;\cdot), P^c_\star) \geq p_{min}$ iff $s \in S^c_h$. We remark that our minimum reachability assumption is more refined than that usually used in RL literature, that $P^c_\star(s'|s,a) \geq p_{min}$ (see, e.g., Wei et al. (2021)) for every context c and (s,a,s'). Clearly, this requirement implies our minimum reachability, but the other direction does not necessarily hold. An *example* for a large class of (non-layered) CMDPs that satisfies that assumption is as follows. (1) At the initial step, for all $c \in \mathcal{C}, a \in A, s' \in S : P^c_{\star,0}(s'|s_0,a) \geq p_{min}$. (2) For every step $h > 0$, the transition probability matrix $P^c_{\star,h}(\cdot|\cdot,a)$ is double stochastic for all $c \in \mathcal{C}$ and $a \in A$. This guarantees that for any policy π, the occupancy measure is at least p_{min}.

Interaction protocol. In each episode $t = 1, 2, ..., T$ the agent: (1) Observes context $c_t \in \mathcal{C}$. (2) Chooses a policy π_t (based on c_t and the observed history). (3) Observes a trajectory of π_t in $\mathcal{M}(c_t)$.

Trajectories and History. Each episode is of length H. A trajectory $\sigma = (c; s_0, a_0, r_0, \ldots, s_{h-1}, a_{h-1}, r_{h-1})$ is generated using the dynamics $P^c_\star$ and the played policy $\pi(c;\cdot)$. We denote the history up to time $t-1$ by $\mathbb{H}_{t-1} = (\sigma^1, \ldots, \sigma^{t-1})$ where σ^i is the trajectory observed in time $i \in [t-1]$, i.e., $\sigma^i = (c_i, s^i_0, a^i_0, r^i_0, \ldots, s^i_{H-1}, a^i_{H-1}, r^i_{H-1}, s^i_H)$.

Offline least square regression (LSR) oracle solves the optimization problem $\hat{f} \in \arg\min_{f\in\mathcal{F}} \sum_{i=1}^n (f(x_i) - y_i)^2$, given a data set $D = \{(x_i, y_i)\}_{i=1}^n$. We remark that there exist function classes for them the LSR oracle can be implemented efficiently. Clearly, this holds for linear functions.

Reward function approximation. We consider a finite function class $\mathcal{G} \subseteq (\mathcal{C} \times A \to [0,1])$ to approximate the context-dependent rewards function of each state $s \in S$. Many times it would be more convenient to consider a finite function class $\mathcal{F} = \mathcal{G}^S$ where $f \in \mathcal{F}$ are functions of the form $f(c,s,a) = g_s(c,a)$ where $g_s \in \mathcal{G}$. Note that, $\log(|\mathcal{F}|) = |S|\log(|\mathcal{G}|)$. Our algorithms get as input the finite function class $\mathcal{F} \subseteq (\mathcal{C} \times S \times A \to [0,1])$. Each function $f \in \mathcal{F}$ maps context $c \in \mathcal{C}$, state $s \in S$ and action $a \in A$ to a (approximate) reward $r \in [0,1]$. We use $\mathcal{F}$ to approximate the context-dependent rewards function using the LSR oracle under the following realizability assumption.

Assumption 2.1 (rewards realizability)**.** *We assume that $\mathcal{F}$ is realizable, meaning, there exists a function $f_\star \in \mathcal{F}$ such that $f_\star(c,s,a) = r^c_\star(s,a) = \mathbb{E}[R^c_\star(s,a)|c,s,a]$.*

For mathematical convenience, we state our algorithms and regret upper bounds in terms of the cardinality of $|\mathcal{F}|$, and use the cardinality of $|\mathcal{G}| = |S|^{-1}\log|\mathcal{F}|$ for our lower bound. We present a comparison between the bounds in Section 8.

Dynamics function approximation. For the unknown context-independent dynamics case we simply use a tabular approximation (see Section 4). For the unknown context-dependent case, our algorithm gets as input a finite function class $\mathcal{P} \subseteq (S \times (S \times A \times \mathcal{C}) \to [0,1])$, where every function $P \in \mathcal{P}$ satisfies $\sum_{s'\in S} P(s'|s,a,c) = 1$ for all $c \in \mathcal{C}$ and $(s,a) \in S \times A$. We use $\mathcal{P}$ to approximate the context-dependent dynamics using LSR oracle under the following realizability assumption. We denote $P^c(s'|s,a) = P(s'|s,a,c)$ for all $P \in \mathcal{P}$.

Assumption 2.2 (dynamics realizability)**.** *We assume that $\mathcal{P}$ is realizable, meaning, there exists a function $P_\star \in \mathcal{P}$ which is the true context-dependent dynamics.*

Learning goal. Our goal is to minimize the regret, relative to the optimal context-dependent policy $\pi^\star$, which defined as $\text{Regret}_T := \sum_{t=1}^T V^{\pi^\star(c_t;\cdot)}_{\mathcal{M}(c_t)}(s_0) - V^{\pi_t(c_t;\cdot)}_{\mathcal{M}(c_t)}(s_0)$, where $c_t \in \mathcal{C}$, $\pi^\star$ is an optimal context-dependent policy and $\pi_t \in \Pi_\mathcal{C}$ are the context and the selected policy at round t. We denote the expected regret as $\text{E.Regret}_T := \mathbb{E}[\text{Regret}_T]$ where the expectation is over the contexts, the randomization of the algorithm and the history.

3 Known Context-Dependent Dynamics

In this section, we present a regret minimization algorithm for contextual MDPs under the minimum reachability assumption, where the context-dependent dynamics $P^c_\star$ is known to the learner. We remark that the minimum reachability parameter p_{min} is unknown to the learner. This section sets the main building blocks of our approach, which we will later extend to handle the unknown dynamics cases.

Algorithm outline. For the first $|A|$ rounds, in each round $i \in \{1, 2, \ldots, |A|\}$ the agent plays the policy $\pi_i \in \Pi_\mathcal{C}$ that always selects action a_i, regardless of the context and the state. At every round $t > |A|$ we approximate the context-dependent rewards function using a least-square minimizer. Using it, we build an "optimistic in expectation" rewards function, and compute an optimal policy for that optimistic model. We run it to generate a trajectory and update the oracle. Here, we take an advantage of the ability to compute the optimal policy $\pi_k(c;\cdot)$ for every context $c \in \mathcal{C}$ separately,

[1] As for non-contextual MDP, there always exists a deterministic context-dependent policy that is optimal.

for all $k = |A|+1, \ldots, t$, to obtain computationally efficient algorithm. (We discuss this challenge later.)

Algorithm 1: Regret Minimization for CMDP with Known Dynamics (RM-KD)

inputs: MDP parameters: S, A, $P_\star$, s_0, H. Confidence $\delta > 0$ and tuning parameters $\{\beta_t\}_{t=1}^T$.
initialization: in round $i \leq |A|$ run $\pi_i(c;s) := a_i$
for round $t = |A|+1, \ldots, T$ **do**
  $\hat{f}_t \in \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}(f(c_i, s_h^i, a_h^i) - r_h^i)^2$ is being computed using the LSR oracle
  observe a fresh context $c_t \sim \mathcal{D}$
  for $k = |A|+1, |A|+2, \ldots, t$ **do**
    compute for all $(s,a) \in S \times A$: $\hat{r}_k^{c_t}(s,a) = \hat{f}_k(c_t, s, a) + \frac{\beta_k}{\sum_{i=1}^{k-1}\mathbb{I}[a=\pi_i(c_t;s)]q(s|\pi_i(c_t;\cdot),P_\star^{c_t})}$
    define the optimistic approximated MDP $\widehat{\mathcal{M}}_k(c_t) = (S, A, P_\star^{c_t}, \hat{r}_k^{c_t}, s_0, H)$
    compute $\pi_k(c_t;\cdot) \in \arg\max_{\pi\in S\to A} V^\pi_{\widehat{\mathcal{M}}_k(c_t)}(s_0)$ using a planning algorithm
  end for
  play $\pi_t(c_t;\cdot)$ and update oracle using $\sigma^t = (c_t, s_0^t, a_0^t, r_0^t, s_1^t, \ldots, s_{H-1}^t, a_{H-1}^t, r_{H-1}^t, s_H^t)$
end for

Remark 3.1. *Since the CMDP is layered, for every context $c \in \mathcal{C}$, layer $h \in [H]$ and state $s_h \in S_h^c$ we have $q_h(s_h|\pi_i(c;\cdot), P_\star^c) = q(s_h|\pi_i(c;\cdot), P_\star^c)$. For convenience, in Algorithm 1 we use q to compute the approximated rewards function, but in the regret analysis we use q_h.*

Analysis outline. Our analysis consists of four main steps.
Step 1: establish uniform convergence bound over any $t \geq 2$ and a fixed sequence of functions $f_2, f_3, \ldots \in \mathcal{F}$. Our bound implies for the least square minimizers sequence $\{\hat{f}_t\}_{t=|A|+1}^T$ and any $\delta \in (0,1)$, that with probability at least $1-\delta/2$ for all $t \geq 2$ it holds that

$$\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i, s_h^i, a_h^i}\left[(\hat{f}_t(c_i, s_h^i, a_h^i) - f_\star(c_i, s_h^i, a_h^i))^2 | \mathbb{H}_{i-1}\right] \leq 68H\log(4|\mathcal{F}|t^3/\delta).$$

Step 2: construct a confidence bound over the value of any given policy w.r.t the true rewards function $f_\star$ and the least square minimizer at round t, $\hat{f}_t$. The confidence bound holds with high probability, in expectation over the contexts (i.e., "optimism in expectation"). Formally, we show that with probability at least $1-\delta/2$ for all $t > |A|$ and any policy $\pi \in \Pi_\mathcal{C}$ it holds that

$$|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t, P_\star)}(c)}(s_0)]| \leq \sqrt{\phi_t(\pi)} \cdot \sqrt{68H\log(4|\mathcal{F}|t^3/\delta)},$$

where $\phi_t(\pi)$ is the contextual potential of π at round t, which is defined as $\phi_t(\pi) := \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]$ and $\mathcal{M}^{(f,P_\star)}(c) = (S, A, P_\star^c, f(c,\cdot,\cdot), s_0, H)$ for any $f \in \mathcal{F}$. The true MDP is $\mathcal{M}(c) = \mathcal{M}^{(f_\star, P_\star)}(c)$. Also, π_i is the selected policy at round i.
Step 3: relax the confidence bound of step 2 to be additive. We show that under the good event of step 2, for all $t > |A|$ and any policy $\pi \in \Pi_\mathcal{C}$, for $\beta_t = \sqrt{\frac{17t\log(4|\mathcal{F}|t^3/\delta)}{|S||A|}}$ we have

$$|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t,P_\star)}(c)}(s_0)]| \leq \beta_t \cdot (\phi_t(\pi) + H|S||A|/t).$$

Step 4: bound the cumulative contextual potential ϕ_t over every round $t = |A|+1, |A|+2, \ldots, T$. For the sequence of selected policies $\{\pi_t \in \Pi_\mathcal{C}\}_{t=1}^T$ it holds that

$$\sum_{t=|A|+1}^{T} \phi_t(\pi_t) \leq |S||A|p_{min}^{-1}(1+\log(T/|A|)).$$

By combining all the steps and applying Azuma's inequality, we obtain the following regret bound.

Theorem 3.2 (regret bound)**.** *For any $T > |A|$, finite function class $\mathcal{F}$ and $\delta \in (0,1)$, let $\beta_t = \sqrt{\frac{17t\log(4|\mathcal{F}|t^3/\delta)}{|S||A|}}$ for all $t \in [T]$. Then, with probability at least $1-\delta$ we have*

$$\textit{Regret}_T(\textit{RM-KD}) \leq \widetilde{O}((p_{min}^{-1}+H)\sqrt{T|S||A|\log|\mathcal{F}|/\delta}).$$

We remark that in all of our algorithms, for $T \leq |S||A|$ the regret is trivially bounded by $|S||A|H$.

Main Technical Challenges and Our Technique

Following steps 2 and 3, a natural "optimistic in expectation" strategy is to select at round t

$$\pi_t \in \arg\max_{\pi\in\Pi_\mathcal{C}}\left\{\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t,P_\star)}(c)}(s_0)] + \beta_t\cdot\phi_t(\pi)\right\} = \arg\max_{\pi\in\Pi_\mathcal{C}}\left\{\mathbb{E}_c[V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)]\right\}.$$

This approach has an obvious three major drawbacks.
(1) The distribution over the contexts, $\mathcal{D}$, is unknown. Hence, we cannot compute $\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right]$, for any policy π.
(2) Even when $\mathcal{D}$ is known, computing $\pi_t \in \Pi_\mathcal{C}$ is intractable when the context space $\mathcal{C}$ is large.
(3) The representation of a context-dependent policy π_t scales with the size of the context space $|\mathcal{C}|$, which can be huge.

We overcome these hurdles using two observations. The first observation is that

$$\max_{\pi\in\Pi_\mathcal{C}}\left\{\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right]\right\} = \mathbb{E}_c\left[\max_{\pi(c;\cdot)\in S\to A} V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right].$$

We conclude that to compute a context-dependent policy $\pi_t \in \Pi_\mathcal{C}$ which maximizes LHS, we can compute for each context $c \in \mathcal{C}$ separately, a policy $\pi_t(c;\cdot) : S \to A$ that is optimal for $\widehat{\mathcal{M}}_t(c)$. For each context $c \in \mathcal{C}$ separately, solving the maximization problem in RHS can be done efficiently using a standard planning algorithm.

The second observation is that in every round t, we do not have to know the full representation of π_k, for all $k \leq t$, but

only the mappings $\{\pi_k(c_i;\cdot)\}_{k=1}^t$ for the observed contexts $\{c_i\}_{i=1}^t$. By taking an advantage of these two observations, at every round $t \le T$, we compute $\pi_k(c_t;\cdot)$ for all $k \le t$, which can be done in $poly(|S|,|A|,H,t)$ using a planning algorithm. Using this, we obtain an efficient algorithm which is independent of $|\mathcal{C}|$.

4 Unknown Context-Independent Dynamics

In this section, we assume the dynamics is unknown to the learner, but is independent of the context. Meaning, for all $c \in \mathcal{C}$, $P_\star^c = P_\star$. We also assume the learner knows the (context-independent) partition of the states space to layers, $S = \{S_0, \ldots, S_H\}$, and the minimum reachability p_{min}.

Algorithm overview. Similarly to Algorithm 1, we define an optimistic-in-expectation rewards function, but, since the dynamics is unknown, we replace $q(s|\pi_i(c_t;\cdot), P_\star^{c_t})$ with its lower bound p_{min}. We denote by $N_t(s,a)$ and $N_t(s,a,s')$ the number of visits to (s,a) and (s,a,s'), respectively, up to round t. To approximate the dynamics, we use a tabular approximation and maintain the following confidence bounds over it, denote them by $\xi_t(s,a) = 2\sqrt{\frac{|S|+2\log(4|S||A|T^2/\delta)}{\max\{1,N_t(s,a)\}}}$, for all $(s,a) \in S \times A$. At round t, we compute an optimistic model w.r.t the rewards function $\widehat{r}_t^{c_t}$ and a deterministic optimal policy $\pi_t(c_t;\cdot)$ for it, under the constraints that the optimistic dynamics is within the confidence interval. (See Appendix C in Levy and Mansour (2022b) for an efficient implementation). We remark that the resulting optimistic approximated dynamics is context-dependent, since it was computed w.r.t the context-dependent approximated rewards function.

Algorithm 2: (sketch) Regret Minimization for Unknown Context Independent Dynamics (RM-UCID)

for round $t > |A|$ **do**
  $\hat{f}_t \in \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}(f(c_i,s_h^i,a_h^i) - r_h^i)^2$ is computed using the LSR oracle
  compute the empirical model for all (s,a,s') : $\bar{P}_t(s'|s,a) = \frac{N_t(s,a,s')}{\max\{1,N_t(s,a)\}}$
  observe a fresh context $c_t \sim \mathcal{D}$
  for $k = |A|+1, \ldots, t$ **do**
    compute for all $(s,a) \in S \times A$: $\hat{r}_k^{c_t}(s,a) = \hat{f}_k(c_t,s,a) + \frac{\beta_k}{p_{min}\sum_{i=1}^{k-1}\mathbb{I}[a=\pi_i(c_t;s)]}$
    compute an optimistic model $\widehat{\mathcal{M}}_k(c_t) = (S,A,\widehat{P}_k^{c_t},\hat{r}_k^{c_t},s_0,H)$ and policy $\pi_k(c_t;\cdot)$
  end for
  play $\pi_t(c_t;\cdot)$, observe trajectory σ^t and update oracle
end for

Analysis overview. We construct confidence intervals for both the dynamics and rewards. For the analysis, we define an intermediate CMDPs where for all $t > |A|$ and $c \in \mathcal{C}$: (1) $\mathcal{M}^{(f,\widehat{P}_t)}(c) = (S,A,\widehat{P}_t^c, f(c,\cdot,\cdot),s_0,H)$, $f \in \mathcal{F}$ and $\widehat{P}_t^c$ is the optimistic dynamics w.r.t $\widehat{r}_t^c$ defined in Algorithm 2. (2) $\mathcal{M}^{(f,P_\star)}(c) = (S,A,P_\star,f(c,\cdot,\cdot),s_0,H)$, $f \in \mathcal{F}$ and $P_\star$ is the true dynamics. (3) $\mathcal{M}^{(\widehat{r}_t,P_\star)}(c) = (S,A,P_\star,\widehat{r}_t^c,s_0,H)$. Let $\pi^\star \in \Pi_\mathcal{C}$ be an optimal policy of the true CMDP.

Analysing the error caused by the rewards approximation. Similar to the analysis for the known dynamics (Section 3), we show that with high probability, for all $t > |A|$, and any policy $\pi \in \Pi_\mathcal{C}$ the following holds:

$$|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t,P_\star)}(c)}(s_0)]| \le \beta_t\left(\phi_t(\pi) + H|S||A|/t\right),$$

where we abuse the contextual potential in round t as $\phi_t(\pi) := \mathbb{E}_c[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{q_h(s_h,\pi(c;s_h)|\pi(c;\cdot),P_\star)}{p_{min}\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}]$.

Analysing the error caused by the dynamics approximation. We show that with high probability the following good event holds. For all $t > |A|$ and $(s,a) \in S \times A$, we have that $\|\bar{P}_t(\cdot|s,a) - P_\star(\cdot|s,a)\|_1 \le \xi_t(s,a)$. Under this good event, our optimistic approximated model $\widehat{\mathcal{M}}_t(c) = (S,A,\widehat{P}_t^c,\widehat{r}_t^c,s_0,H)$ and the selected policy $\pi_t(c;\cdot)$ satisfy for all $c \in \mathcal{C}$ and $t > |A|$ that $V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0) \ge V^{\pi^\star(c;\cdot)}_{\mathcal{M}(\widehat{r}_t^c,P_\star)}(s_0)$. When combining the latter inequality with the confidence bounds over the rewards, we obtain for all $t > |A|$ that

$$\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] \le \beta_t \cdot H|S||A|/t.$$

Moreover, we show that under the good event of the dynamics approximation, for $T > |S||A|$ with high probability we have

$$\sum_{t=|A|+1}^{T}\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\hat{f}_t,\widehat{P}_t)(c)}] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\hat{f}_t,P_\star)(c)}] \le O(H^{1.5}|S|\sqrt{|A|T}\log(|S||A|T^2/\delta)).$$

Lastly, we bound $\sum_{t=|A|+1}^T \phi_t(\pi_t)$ similarly to Section 3. By combining all the above, and applying Azuma's inequality, we obtain the following regret bound

Theorem 4.1 (regret bound)**.** *For any $T > |S||A|$, finite function class $\mathcal{F}$ and $\delta \in (0,1)$, for the choice of $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$ for all t, with probability at least $1-\delta$,*

$$\textit{Regret}_T(\textit{RM-UCID}) \le \widetilde{O}\Big(H^{1.5}|S|\sqrt{T|A|}\log(1/\delta) + (H + p_{min}^{-1})\cdot\sqrt{|S||A|T\log(|\mathcal{F}|/\delta)}\Big).$$

5 Unknown Context-Dependent Dynamics

In this section, we consider the most challenging case, where the dynamics is unknown and context-dependent. We assume an access to a finite function class $\mathcal{P} \subseteq (S \times (S \times A \times \mathcal{C}) \to [0,1])$, for which every function $P \in \mathcal{P}$ satisfies $\sum_{s'\in S} P(s'|s,a,c) = 1, \ \ \forall (s,a,c) \in S \times A \times \mathcal{C}$. We use $\mathcal{P}$ to approximate the context-dependent dynamics under the dynamics realizability assumption (Assumption 2.2).

Algorithm outline. In Algorithm RM-UCDD (Algorithm 3), we approximate both the rewards and the dynamics using a LSR oracle. The first $|A|$ rounds are initialization rounds, as before. At round $t > |A|$, we compute the approximated rewards function for the context c_t as is done in previous

sections. For the dynamics approximation, we use the least square minimizer $\widehat{P}_t$. We define the approximated model for c_t, compute an optimal policy $\pi_t(c_t;\cdot)$ for it and run it to generate a trajectory and update the oracles. We feed the LSR oracle for the dynamics with samples of the form $((c_t, s_h^t, a_h^t, s'), \mathbb{I}[s' = s_{h+1}^t])$ for all $t \leq T$, $h \in [H-1]$ for every $s' \in S$, where $\mathbb{I}$ is an indicator function.

Algorithm 3: Regret Minimization for CMDP with Unknown Context-Dependent Dynamics

inputs: MDP parameters: S, A, H, s_0. Confidence $\delta > 0$ and tuning parameters $\{\beta_t, \gamma_t\}_{t=1}^T$. Minimum reachability parameter $p_{min} > 0$.
initialization: in round $i \leq |A|$ run $\pi_i(c;s) := a_i$
for round $t = |A|+1, \ldots, T$ **do**
  $\hat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} (f(c_i, s_h^i, a_h^i) - r_h^i)^2$ is computed using the LSR oracle
  also compute $\widehat{P}_t \in \arg\min_{\widetilde{P} \in \mathcal{P}} \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} \sum_{s' \in S} (\widetilde{P}^{c_i}(s'|s_h^i, a_h^i) - \mathbb{I}[s' = s_{h+1}^i])^2$ using the LSR oracle
  observe a fresh context $c_t \sim \mathcal{D}$
  for $k = |A|+1, |A|+2, \ldots, t$ **do**
    compute for all $(s,a) \in S \times A$: $\hat{r}_k^{c_t}(s,a) = \hat{f}_k(c_t, s, a) + \frac{\beta_k + H|S|\gamma_k}{p_{min} \sum_{i=1}^{k-1} \mathbb{I}[a = \pi_i(c_t;s)]}$
    define $\widehat{\mathcal{M}}_k(c_t) = (S, A, \widehat{P}_k^{c_t}, \hat{r}_k^{c_t}, s_0, H)$
    compute $\pi_k(c_t;\cdot) \in \arg\max_{\pi \in S \to A} V^{\pi}_{\widehat{\mathcal{M}}_k(c_t)}(s_0)$ using planning algorithm
  end for
  play $\pi_t(c_t,\cdot)$, observe trajectory σ^t and update oracles
end for

Analysis outline. In the analysis, we define the following intermediate MDPs for every $t > |A|$ and context $c \in \mathcal{C}$: (1) $\mathcal{M}^{(\hat{r}_t, P)}(c) = (S, A, P^c, \hat{r}_t^c, s_0, H)$ for context-dependent dynamics $P \in \mathcal{P}$, where $\hat{r}_t^c$ is the approximated rewards function in round t, which is defined in Algorithm 3. By definition, $\widehat{\mathcal{M}}_t(c) = \mathcal{M}^{(\hat{r}_t, \widehat{P}_t)}(c)$. (2) $\mathcal{M}^{(f,P)}(c) = (S, A, P^c, f(c,\cdot,\cdot), s_0, H)$ for any $f \in \mathcal{F}$ and $P \in \mathcal{P}$. By definition, $\mathcal{M}(c) = \mathcal{M}^{(f_\star, P_\star)}(c)$. We denote by $\psi_t(\pi)$ the contextual potential at round t, which is defined as $\psi_t(\pi) := \mathbb{E}_c[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi(c;\cdot), \widehat{P}_t^c)}{p_{min} \sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)]}]$.

Analysing the error caused by the rewards approximation. Similar to the known dynamics setting (Section 3), we show that with probability at least $1 - \delta/4$, for all $t > |A|$ and $\pi \in \Pi_{\mathcal{C}}$ the following holds.

$$|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t, \widehat{P}_t)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star, \widehat{P}_t)}(c)}(s_0)]| \leq \beta_t(\psi_t(\pi) + H|S||A|/t).$$

Analysing the error of the dynamics approximation.
Key observation. Let $\mathcal{B}$ be a random variable which generates the next state s_{h+1} given the true dynamics associated with c, $P_\star^c$, the state s_h and the action a_h. The random variable $\mathcal{B}(P_\star^c, s_h, a_h)$ is distributed $P_\star^c(\cdot|s_h, a_h)$. Our observation is that since the CMDP is layered, given the context c_t state s_h^t and action a_h^t, we have that the random variables $\mathcal{B}(P_\star^{c_t}, s_h^t, a_h^t)$ and $(s_0^t, a_0^t, s_1^t, \ldots, s_{h-1}^t, a_{h-1}^t)$ are independent random variables. Using that observation, we are able to extend our uniform convergence bound to the dynamics approximation. Hence, we can apply the four steps strategy above for the dynamics approximation as well.
Step 1: establish uniform convergence bound over any $t \geq 2$ and a fixed sequence of functions $P_2, P_3, \ldots \in \mathcal{P}$. The bound implies that for the least square minimizers sequence $\{\widehat{P}_t\}_{t=|A|+1}^T$ with high probability, for all $t > |A|$,

$$\sum_{i=1}^{t-1} \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i, s_h^i, a_h^i} \left[\|\widehat{P}_t^{c_i}(\cdot|s_h^i, a_h^i) - P_\star^{c_i}(\cdot|s_h^i, a_h^i)\|_2^2 \,\middle|\, \mathbb{H}_{i-1} \right] \leq 72H|S| \log(8|\mathcal{P}|t^3/\delta).$$

Step 2: construct a confidence bound over the value of any given policy w.r.t the approximated and true dynamics, where the rewards function is $f_\star$. The confidence bound holds with high probability, in expectation over the contexts.
Formally, we show that with probability at least $1 - \delta/4$, for all $t > |A|$ and any policy $\pi \in \Pi_{\mathcal{C}}$ it holds that

$$|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star, P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star, \widehat{P}_t)}(c)}(s_0)]| \leq \sqrt{H|S|\psi_t(\pi)} \cdot \sqrt{72H^2|S| \log(8|\mathcal{P}|t^3/\delta)}.$$

Step 3: relax the confidence bound in step 2 to be additive. we show that under the good event of step 2 for all $t > |A|$ and any policy $\pi \in \Pi_{\mathcal{C}}$, for $\gamma_t = \sqrt{\frac{18t \log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$,

$$|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star, P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star, \widehat{P}_t)}(c)}(s_0)]| \leq \gamma_t H|S| \left(\psi_t(\pi) + H|S||A|/t\right).$$

Step 4: bound the sum of contextual potential functions similarly to shown for the rewards, in previous sections.

Using all the above, we obtain the optimism lemma which states that under the good events of step 2 for both the dynamics and rewards approximation, for all $t > |A|$,

$$\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0) - V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] \leq \frac{(H|S|\gamma_t + \beta_t)H|S||A|}{t},$$

yielding the following regret bound.

Theorem 5.1 (regret bound)**.** *For any $\delta \in (0,1)$, $T > |A|$ and finite function classes $\mathcal{F}$ and $\mathcal{P}$, for the choice of $\beta_t = \sqrt{\frac{17t \log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$ and $\gamma_t = \sqrt{\frac{18t \log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$ for all t, with probability at least $1-\delta$ it holds that*

$$\textit{Regret}_T(\textit{RM-UCDD}) \leq \widetilde{O}((H + 1/p_{min})H|S|^{3/2}\sqrt{|A|T \log(\max\{|\mathcal{F}|, |\mathcal{P}|\}/\delta)}).$$

6 Lower bound

We present a lower bound for layered CMDP, where the dynamics is known and context-independent, which based on the lower bound for CMAB presented by Agarwal et al. (2012), in which $K = |A|$, $\mathcal{G} \subseteq (\mathcal{C} \times A \to [0,1])$ and $N \in \mathbb{N}$.

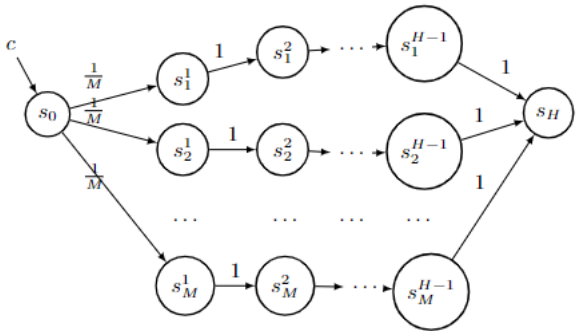

Figure 1: Lower bound illustration

Theorem 6.1 (Theorem 5.1, Agarwal et al. (2012))**.** *For every N and K such that $\ln N/\ln K \leq T$, and every algorithm $\mathfrak{A}$, there exist a functions class $\mathcal{G}$ of cardinality at most N and a distribution $D(c,r)$ for which the realizability assumption holds, but the expected regret of $\mathfrak{A}$ is $\Omega(\sqrt{KT\ln N/\ln K})$.*

Theorem 6.2 (Lower bound for CMDP)**.** *Let $\delta \in (0,1)$, horizon $H \geq 2$ and $M, N \in \mathbb{N}$. Let $T \geq 8M\log(|S|/\delta) + 2M\ln N/\ln|A|$ and consider a CMDP $(\mathcal{C}, S, A, \mathcal{M})$ for which $|S| = M\cdot(H-1)+2$.*

Then, for any algorithm $\mathfrak{A}$, there exist a base function class $\mathcal{G} \subseteq (\mathcal{C}\times A \to [0,1])$ of cardinality at most N, and a distribution $D(c,s,a,r)$ for which the realizability assumption holds for $\mathcal{F} = \mathcal{G}^S$ and, with probability at least $1-\delta$, the expected regret of $\mathfrak{A}$ is $\Omega(\sqrt{TH|S||A|\ln(N)/\ln(|A|)})$.

proof idea. Solving the CMDP illustrated in Figure 1 is equivalent to solving $M(H-1)+1$ CMAB problems. Hence, the theorem follows by Theorem 6.1.

7 Extension: Remove the Reachability

The minimum reachability assumption allows us to limit the exploration-exploitation trade-off only to the actions selection, since any state is reached with probability at least $p_{min} > 0$. In this section we sketch a derivation of $\widetilde{O}(T^{3/4})$ regret, without the reachability assumption.

A first step towards removing the reachability assumption is to consider a dynamics class that is mixed with the uniform distribution with probability $\rho > 0$. For every $P \in \mathcal{P}$ let $S_h^c(P^c)$ denote the h-layer defined by the transition matrix P^c. Using $\mathcal{P}$ we define $\mathcal{P}(\rho)$, where for each dynamics $P^c \in \mathcal{P}$ there is a dynamics $\widetilde{P}^c \in \mathcal{P}(\rho)$, where in time h with probability ρ we transition to a random state in S_h^c. Assume that we have an access to a LSR oracle that gets as inputs a parameter ρ and a realizable function class $\mathcal{P}$. Let $P_\star \in \mathcal{P}$ denote the true context-dependent dynamics, and $\widetilde{P}_\star \in \mathcal{P}(\rho)$ be the related dynamics in $\mathcal{P}(\rho)$.

Please notice the following observations:

(1) $\widetilde{P}_\star$ has the minimum reachability property for $p_{min} = \rho/|S|$, even if $P_\star$ does not have it.

(2) For every context $c \in \mathcal{C}$, layer $h \in [H-1]$ and state-action $(s,a) \in S_h^c \times A$, it holds that $\|P_\star^c(\cdot|s,a) - \widetilde{P}_\star^c(\cdot|s,a)\|_1 \leq 2\rho$. For $\rho < 1/2$ this also implies that the function class $\mathcal{P}(\rho)$ has an agnostic approximation error of at most 2ρ (w.r.t the square loss).

(3) Let any rewards function $r \in [0,1]$, context $c \in \mathcal{C}$ and policy π. Then, the value of π on the model defined by $(r, P_\star^c)$ and the value of π on the model defined by $(r, \widetilde{P}_\star^c)$ differ by at most $\widetilde{O}(\rho H^2)$. This implies that the optimal policy for one of them is near optimal for the other.

By (2), in this setting, our uniform convergence bound of Step 1 has an additional error of 2ρ, which yields (approximately) an additional term of $\widetilde{O}(\rho H^2|S|)$ in the additive confidence bound of a policy (Step 3) for the dynamics approximation. Hence, the overall regret bound is $\widetilde{O}(\rho H^2|S|T + H^2|S|^{3/2}\sqrt{T|A|\log(\max\{|\mathcal{F}|,|\mathcal{P}|\}/\delta)}|S|/\rho)$.

For $\rho \approx |S|^{3/4}T^{-1/4}$, we obtain a regret bound of $\widetilde{O}(H^2|S|^{7/4}T^{3/4}\sqrt{|A|\log(\max\{|\mathcal{F}|,|\mathcal{P}|\}/\delta)})$. Therefore, our approach yields a sub-linear regret bound that does not depend on the minimum reachability parameter p_{min}, which is now a tuned parameter.

8 Discussion

To the best of our knowledge, this work is the first that obtains sub-liner regret bounds using general function approximation (i.e., without additional structural assumption regarding the CMDP or the function classes) and to present an expected regret lower bound. Our results can be naturally extended to infinite function classes using covering numbers analysis (see e.g., Shalev-Shwartz and Ben-David (2014)). Our algorithms has $poly(|S|,|A|,H,T)$ running time and space complexity, assuming an efficient least-square regression oracle.

The main advantages of our technique: (1) We present a novel confidence interval for general function approximation in CMDPs. (2) We use an access to a standard offline least-square regression oracle, which we call only $O(T)$ times. (3) Our algorithms do no fully represent the selected context-dependent policy at each time step, as the representation of it scales linearly in the context space size $|\mathcal{C}|$, which can be huge, but rather compute it only for the observed contexts.

Tightness of our bounds. Consider our regret upper bounds in terms of the base class $\mathcal{G}$ cardinality, recalling that $\mathcal{F} = \mathcal{G}^S$. *Known context-dependent* dynamics: $\widetilde{O}((H + p_{min}^{-1})|S|\sqrt{T|A|\log(|\mathcal{G}|/\delta)})$. For the *Unknown context-independent* dynamics: $\widetilde{O}\Big(H^{1.5}|S|\sqrt{T|A|}\log(1/\delta) + (H + p_{min}^{-1})\cdot|S|\sqrt{|A|T\log(|\mathcal{G}|/\delta)}\Big)$. *Unknown context-dependent* dynamics: $\widetilde{O}((H+p_{min}^{-1})H|S|^{3/2}\sqrt{|A|T\log(\max\{|\mathcal{G}|,|\mathcal{P}|\}/\delta)})$. On the other hand, recall our lower bound is $\Omega(\sqrt{TH|S||A|\ln(|\mathcal{G}|)/\ln(|A|)})$. While our dependency in T, $|A|$ and $|\mathcal{G}|$ is near-optimal, bridging the gap in $|S|$, H and p_{min} is an important open question.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation(grant number 993/17), Tel Aviv University Center for AI and Data Science

(TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University.

References

Agarwal, A.; Dudík, M.; Kale, S.; Langford, J.; and Schapire, R. 2012. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, 19–26. PMLR.

Agarwal, A.; Hsu, D.; Kale, S.; Langford, J.; Li, L.; and Schapire, R. 2014. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, 1638–1646. PMLR.

Foster, D.; Agarwal, A.; Dudik, M.; Luo, H.; and Schapire, R. 2018. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, 1539–1548. PMLR.

Foster, D.; and Rakhlin, A. 2020. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, 3199–3210. PMLR.

Foster, D. J.; Kakade, S. M.; Qian, J.; and Rakhlin, A. 2021. The Statistical Complexity of Interactive Decision Making. *arXiv preprint arXiv:2112.13487*.

Hallak, A.; Di Castro, D.; and Mannor, S. 2015. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*.

Jiang, N.; Krishnamurthy, A.; Agarwal, A.; Langford, J.; and Schapire, R. E. 2017. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, 1704–1713. PMLR.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Levy, O.; and Mansour, Y. 2022a. Learning Efficiently Function Approximation for Contextual MDP. *arXiv preprint arXiv:2203.00995*.

Levy, O.; and Mansour, Y. 2022b. Optimism in Face of a Context: Regret Guarantees for Stochastic Contextual MDP. *arXiv preprint arXiv:2207.11126*.

Mannor, S.; Mansour, Y.; and Tamar, A. 2022. *Reinforcement Learning: Foundations*. Online manuscript; https://sites.google.com/view/rlfoundations/home. Accessed March-05-2023.

Modi, A.; Jiang, N.; Singh, S.; and Tewari, A. 2018. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, 597–618. PMLR.

Modi, A.; and Tewari, A. 2020. No-regret exploration in contextual reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, 829–838. PMLR.

Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Simchi-Levi, D.; and Xu, Y. 2021. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*.

Slivkins, A. 2019. Introduction to Multi-Armed Bandits. *Found. Trends Mach. Learn.*, 12(1-2): 1–286.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Wei, C.-Y.; Lee, C.-W.; Zhang, M.; and Luo, H. 2021. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on learning theory*, 4259–4299. PMLR.

Xu, Y.; and Zeevi, A. 2020. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*.

Zimin, A.; and Neu, G. 2013. Online learning in episodic Markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26.