

# Regret Bounds for Kernel-Based Reinforcement Learning

Omar D. Domingues<sup>1</sup> Pierre Ménard<sup>1</sup> Matteo Pirotta<sup>2</sup> Emilie Kaufmann<sup>1,3</sup> Michal Valko<sup>4</sup>

## Abstract

We consider the exploration-exploitation dilemma in finite-horizon reinforcement learning problems whose state-action space is endowed with a metric. We introduce Kernel-UCBVI, a model-based optimistic algorithm that leverages the smoothness of the MDP and a non-parametric kernel estimator of the rewards and transitions to efficiently balance exploration and exploitation. Unlike existing approaches with regret guarantees, it does not use *any kind of partitioning of the state-action space*. For problems with  $K$  episodes and horizon  $H$ , we provide a regret bound of  $O\left(H^3 K^{\max(\frac{1}{2}, \frac{2d}{2d+1})}\right)$ , where  $d$  is the covering dimension of the joint state-action space. We empirically validate Kernel-UCBVI on discrete and continuous MDPs.

## 1. Introduction

Reinforcement learning (RL) is a learning paradigm in which an agent interacts with an environment by taking actions and receiving rewards. At each time step  $t$ , the environment is characterized by a state variable  $x_t \in \mathcal{X}$ , which is observed by the agent and influenced by its actions  $a_t \in \mathcal{A}$ . In this work, we consider the online learning problem where the agent has to learn how to act optimally by interacting with an unknown environment. To learn efficiently, the agent has to trade-off exploration to gather information about the environment and exploitation to act optimally w.r.t. the current knowledge. The performance of the agent is measured by the *regret*, i.e., the difference between the rewards that would be gathered by an optimal agent and the rewards obtained by the agent.

This problem has been extensively studied for Markov Decision Processes (MDPs) with finite state-action space. *Optimism in the face of uncertainty* (OFU, Jaksch et al., 2010) and *Thompson Sampling* (Strens, 2000; Osband et al., 2013)

principles have been used to design algorithms with sublinear regret. However, the guarantees for these approaches cannot be naturally extended to an arbitrarily large state-action space since the regret depends (typically sublinearly) on the number of states and actions. When the state-action space is continuous, additional structure in MDP is required to efficiently solve the exploration-exploitation dilemma.

In this paper, we focus on the online learning problem in MDPs with large or continuous state-action spaces. We suppose that the state-action set  $\mathcal{X} \times \mathcal{A}$  is equipped with a known *metric*. For instance, this is typically the case in continuous control problems in which the state space is a subset of  $\mathbb{R}^d$  equipped with the Euclidean metric.

We propose an algorithm based on non-parametric kernel estimators of the reward and transition functions of the underlying MDP. One of the main advantages of this approach is that it applies to problems with possibly infinite state-action sets without relying on any kind of discretization. This is particularly useful when we have a way to assess the similarity of state-action pairs (by defining a metric), but we do not have prior information on the shape of the state-action space in order to construct a good discretization.

**Related work** The exploration-exploitation dilemma has been extensively studied both in model-based and model-free settings. While model-based algorithms use the estimated rewards and transitions to perform planning at each episode, model-free algorithms directly build an estimate of the optimal Q-function that is updated incrementally. In the tabular case, model-based algorithms are understood to be more efficient than model-free algorithms in solving the exploration-exploitation dilemma. Indeed, while model-based algorithms (e.g., UCBVI by Azar et al., 2017 or EU-LER by Zanette & Brunskill, 2019) enjoy regret bounds that are first-order optimal, model-free algorithms (e.g., optimistic Q-Learning (OptQL) by (Jin et al., 2018)) are sub-optimal by a factor  $\sqrt{H}$ . A clear understanding of the relationship between model-based and model-free algorithms in continuous MDPs is still missing.

For MDPs with continuous state-action space, the sample complexity (e.g., Kakade et al., 2003; Kearns & Singh, 2002; Lattimore et al., 2013; Papis & Parr, 2013) or regret have been studied under structural assumptions. Regarding regret minimization, a standard assumption is that

<sup>1</sup>Inria Lille, SequeL team <sup>2</sup>Facebook AI Research, Paris

<sup>3</sup>CNRS & Univ. Lille <sup>4</sup>DeepMind Paris. Correspondence to: Omar D. Domingues <omar.darwiche-domingues@inria.fr>.

rewards and transitions are Lipschitz continuous. Ortner & Ryabko (2013) studied this problem in average reward problems. They combined the ideas of UCRL2 (Jaksch et al., 2010) and uniform discretization, proving a regret bound of  $\tilde{O}\left(T^{\frac{2d+1}{2d+2}}\right)$  for a learning horizon  $T$  in  $d$ -dimensional state spaces. This work was later extended by Lakshmanan et al. (2015) to use a kernel density estimator instead of a frequency estimator for each region of the fixed discretization. For each *discrete* region  $I(x)$ , the density  $p(\cdot|I(x), a)$  of the transition kernel<sup>2</sup> is computed through kernel density estimation. The granularity of the discretization is selected in advance based on the properties of the MDP and the learning horizon  $T$ . As a result, they improve upon the bound of Ortner & Ryabko (2013), but require the transition kernels to have densities that are  $\kappa$  times differentiable.<sup>1</sup> However, these two algorithms rely on an intractable optimization problem for finding an optimistic MDP. Qian et al. (2019) solve this issue by providing an algorithm that uses exploration bonuses, but they still rely on a discretization of the state space. Ok et al. (2018) studied the asymptotic regret in Lipschitz MDPs with *finite* state and action spaces, providing a nearly asymptotically optimal algorithm. Their algorithm leverages ideas from asymptotic optimal algorithms in structured bandits (Combes et al., 2017) and tabular RL (Burnetas & Katehakis, 1997), but does not scale to continuous state-action spaces.

In the last few months, there has been a surge of research in exploration for finite-horizon MDP with continuous state-action space. While Yang et al. (2019) focused on deterministic MDPs with Lipschitz transitions and Q-function, Song & Sun (2019) only assume that the Q-function is Lipschitz continuous (the MDP can be stochastic). Song & Sun (2019) provided an efficient model-free algorithm by combining the ideas of tabular optimistic Q-learning (Jin et al., 2018) with uniform discretization, showing a regret bound of  $O(H^{5/2}K^{\frac{d+1}{d+2}})$  where  $d$  is the covering dimension of the state-action space. This approach was recently extended by Sinclair et al. (2019) to use adaptive partitioning where the discretization is adapted to better estimate regions that are frequently visited or with high reward, achieving the same regret bound.

Aside the work on Lipschitz RL, we also note that there many results for facing the exploration problem in continuous MDP with *parametric* structure, e.g., linear-quadratic systems (Abbasi-Yadkori & Szepesvári, 2011) or other linearity assumptions (Yang & Wang, 2019; Jin et al., 2019), which are outside the scope of our paper.

Finally, *kernels* in machine learning name a few different concepts. In this work, “kernel” refers to a smoothing func-

tion used in a non-parametric estimator<sup>2</sup>, and not to some underlying Gaussian process or reproducing kernel Hilbert space, as in (Chowdhury & Gopalan, 2019). In that sense, our work is close to the kernel-based RL proposed by Ormoneit & Sen (2002), who study similar estimators. However, Ormoneit & Sen (2002) propose an algorithm which learns a policy based on transitions generated from *independent* samples, with *asymptotic* convergence guarantees, whereas we propose an algorithm which collects data *online* and has *finite-time regret* guarantees.

**Contributions** The main contributions of this paper are: **1)** Unlike existing algorithms for metric spaces, our algorithm does not require any form of discretization. This approach is entirely *data-dependent*, and we can choose the kernel bandwidth to reflect our prior knowledge about the smoothness of the underlying MDP. To the best of our knowledge, we prove the first regret bound in this setting. **2)** Existing model-based algorithms assume that the transition kernels are Lipschitz continuous with respect to the total variation distance, which does not hold for deterministic MDPs. In this work, we construct upper confidence bounds for the value functions which are themselves Lipschitz. This allows us to have an assumption with respect to the Wasserstein distance, which holds for deterministic MDPs with Lipschitz transitions. **3)** Model-free tabular algorithms are worse than model-based when looking at the first-order term. However, they have a better dependence w.r.t. the horizon  $H$  and the number of states  $X$  in the second-order term. This second-order term does not depend on the number of episodes  $K$ , and can be neglected if  $K$  is large enough. In the continuous setting, we show that the second-order term also depends on  $K$  and on the state-action dimension  $d$ . Consequently, it cannot be neglected even for large  $K$ . Hence, model-based algorithms seem to suffer from a worse dependence on  $d$  than model-free ones. **4)** In order to derive our regret bound, we provide novel high-probability confidence intervals for weighted sums and for non-parametric kernel estimators that we believe to be of independent interest.

## 2. Setting

**Notation** For any  $j \in \mathbb{Z}_+$ , we define  $[j] \stackrel{\text{def}}{=} \{1, \dots, j\}$ . For a measure  $P$  and any function  $f$ , let  $Pf \stackrel{\text{def}}{=} \int f(y)dP(y)$ . If  $P(\cdot|x, a)$  is a measure for all  $(x, a)$ , we let  $Pf(x, a) = P(\cdot|x, a)f = \int f(y)dP(y|x, a)$ .

**Markov decision processes** Let  $\mathcal{X}$  and  $\mathcal{A}$  be the sets of states and actions, respectively. We assume that there exists a metric  $\rho : (\mathcal{X} \times \mathcal{A})^2 \rightarrow \mathbb{R}_+$  on the state-action space

<sup>1</sup>For instance, when  $d = 1$  and  $\kappa$  goes to infinity, their bound approaches  $T^{2/3}$ , improving the previous bound of  $T^{3/4}$ .

<sup>2</sup>For disambiguation, notice that we also use the term “transition kernel” when referring to Markov kernels in probability theory, which is **not** related to kernel smoothing functions (or kernel density estimates).

and that  $(\mathcal{X}, \mathcal{T}_{\mathcal{X}})$  is a measurable space with  $\sigma$ -algebra  $\mathcal{T}_{\mathcal{X}}$ . We consider an episodic Markov decision process (MDP), defined by the tuple  $\mathcal{M} \stackrel{\text{def}}{=} (\mathcal{X}, \mathcal{A}, H, P, r)$  where  $H \in \mathbb{Z}_+$  is the length of each episode,  $P = \{P_h\}_{h \in [H]}$  is a set of transition kernels<sup>2</sup> from  $(\mathcal{X} \times \mathcal{A}) \times \mathcal{T}_{\mathcal{X}}$  to  $\mathbb{R}_+$ , and  $r = \{r_h\}_{h \in [H]}$  is a set of reward functions from  $\mathcal{X} \times \mathcal{A}$  to  $[0, 1]$ . A policy  $\pi$  is a mapping from  $[H] \times \mathcal{X}$  to  $\mathcal{A}$ , such that  $\pi(h, x)$  is the action chosen by  $\pi$  in state  $x$  at step  $h$ . The Q-value of a policy  $\pi$  for state-action  $(x, a)$  at step  $h$  is the expected sum of rewards obtained by taking action  $a$  in state  $x$  at step  $h$  and then following the policy  $\pi$ , that is

$$Q_h^\pi(x, a) \stackrel{\text{def}}{=} \mathbb{E} \left[ \sum_{h'=h}^H r_{h'}(x_{h'}, a_{h'}) \middle| \substack{x_h=x, a_h=a \\ a_{h'}=\pi(h', x_{h'}) \forall h' > h} \right],$$

where the expectation is under transitions in the MDP:  $x_{h'+1} \sim P_h(\cdot | x_{h'}, a_{h'})$ . The value function of policy  $\pi$  at step  $h$  is  $V_h^\pi(x) = Q_h^\pi(x, \pi(h, x))$ . The optimal value functions, defined by  $V_h^*(x) \stackrel{\text{def}}{=} \sup_\pi V_h^\pi(x)$  for  $h \in [H]$ , satisfy the optimal Bellman equations (Puterman, 1994):

$$\begin{aligned} V_h^*(x) &= \max_{a \in \mathcal{A}} Q_h^*(x, a), \quad \text{where} \\ Q_h^*(x, a) &\stackrel{\text{def}}{=} r_h(x, a) + \int_{\mathcal{X}} V_{h+1}^*(y) dP_h(y | x, a) \end{aligned}$$

and, by definition,  $V_{H+1}^*(x) = 0$  for all  $x \in \mathcal{X}$ .

**Learning problem** A reinforcement learning agent interacts with  $\mathcal{M}$  in a sequence of episodes  $k \in [K]$  of fixed length  $H$  by playing a policy  $\pi_k$  in each episode. At each episode, the initial state  $x_1^k$  is chosen arbitrarily and revealed to the agent. The learning agent does not know  $P$  and  $r$  and it selects the policy  $\pi_k$  based on the samples observed over previous episodes. Its performance is measured by the regret  $\mathcal{R}(K) \stackrel{\text{def}}{=} \sum_{k=1}^K (V_1^*(x_1^k) - V_1^{\pi_k}(x_1^k))$ .

We make the following assumptions:

**Assumption 1.** The metric  $\rho$  is given to the learner. Also, there exists a metric  $\rho_{\mathcal{X}}$  on  $\mathcal{X}$  and a metric  $\rho_{\mathcal{A}}$  on  $\mathcal{A}$  such that, for all  $(x, x', a, a')$ ,

$$\rho[(x, a), (x', a')] = \rho_{\mathcal{X}}(x, x') + \rho_{\mathcal{A}}(a, a').$$

**Assumption 2.** The reward functions are  $\lambda_r$ -Lipschitz:  $\forall(x, a, x', a') \text{ and } \forall h \in [H]$ ,

$$|r_h(x, a) - r_h(x', a')| \leq \lambda_r \rho[(x, a), (x', a')]$$

**Assumption 3.** The transition kernels are  $\lambda_p$ -Lipschitz with respect to the 1-Wasserstein distance:  $\forall(x, a, x', a') \text{ and } \forall h \in [H]$ ,

$$W_1(P_h(\cdot | x, a), P_h(\cdot | x', a')) \leq \lambda_p \rho[(x, a), (x', a')]$$

where, for two measures  $\mu$  and  $\nu$ , we have

$$W_1(\mu, \nu) \stackrel{\text{def}}{=} \sup_{f: \text{Lip}(f) \leq 1} \int_{\mathcal{X}} f(y) (d\mu(y) - d\nu(y))$$

and where, for any Lipschitz function  $f: \mathcal{X} \rightarrow \mathbb{R}$  with respect to  $\rho_{\mathcal{X}}$ ,  $\text{Lip}(f)$  denotes its Lipschitz constant.

To assess the relevance of these assumptions, we show below that they apply to deterministic MDPs with Lipschitz reward and transition functions (whose transition kernels are *not* Lipschitz w.r.t. the total variation distance).

**Example 1** (Deterministic MDP in  $\mathbb{R}^d$ ). Consider an MDP  $\mathcal{M}$  with a finite action set, with a compact state space  $\mathcal{X} \subset \mathbb{R}^d$ , and deterministic transitions  $y = f(x, a)$ , i.e.,  $P_h(y | x, a) = \delta_{f(x, a)}(y)$ . Let  $\rho_{\mathcal{X}}$  be the Euclidean distance on  $\mathbb{R}^d$  and  $\rho_{\mathcal{A}}(a, a') = 0$  if  $a = a'$  and  $\infty$  otherwise. Then, if for all  $a \in \mathcal{A}$ ,  $x \mapsto r_h(x, a)$  and  $x \mapsto f(x, a)$  are Lipschitz,  $\mathcal{M}$  satisfies assumptions 1, 2 and 3.

The following lemma states that, under our assumptions, the Q function of any policy  $\pi$  is Lipschitz continuous.

**Lemma 1.** Let  $L_h \stackrel{\text{def}}{=} \sum_{h'=h}^H \lambda_r \lambda_p^{H-h'}$ . Under assumptions 2 and 3, for all  $(x, a, x', a')$  and for all  $h \in [H]$ ,  $|Q_h^\pi(x, a) - Q_h^\pi(x', a')| \leq L_h \rho[(x, a), (x', a')]$  for any  $\pi$ .

### 3. Algorithm

In this section, we present **Kernel-UCBVI**, a model-based algorithm for exploration in MDPs in metric spaces that employs *kernel smoothing* to estimate the rewards and transitions, for which we derive confidence intervals. **Kernel-UCBVI** uses exploration bonuses based on these confidence intervals to efficiently balance exploration and exploitation. Our algorithm requires the knowledge of the metric  $\rho$  on  $\mathcal{X} \times \mathcal{A}$  and of the Lipschitz constants of the rewards and transitions<sup>3</sup>.

#### 3.1. Kernel Function

We leverage the knowledge of the state-action space metric to define the kernel function. Let  $u, v \in \mathcal{X} \times \mathcal{A}$ . For some function  $g: \mathbb{R}_+ \rightarrow [0, 1]$ , we define the kernel function as

$$\psi_\sigma(u, v) \stackrel{\text{def}}{=} g\left(\frac{\rho[u, v]}{\sigma}\right)$$

where  $\sigma$  is the bandwidth parameter that controls the degree of “smoothing” of the kernel. In order to be able to construct valid confidence intervals, we require certain structural properties for  $g$ .

<sup>3</sup>Theoretically, we could replace the Lipschitz constants in each episode  $k$  by  $\log(k)$ , and our regret bounds would be valid for large enough  $k$  (e.g., Reeve et al., 2018). However, this theoretical trick would degrade the performance of the algorithm in practice.

**Assumption 4.** The function  $g : \mathbb{R}_+ \rightarrow [0, 1]$  is differentiable, non-increasing,  $g(1) > 0$ , and there exists two constants  $C_1^g, C_2^g > 0$  that depend only on  $g$  such that  $g(z) \leq C_1^g \exp(-z^2/2)$  and  $\sup_z |g'(z)| \leq C_2^g$ .

This assumption is trivially verified by the Gaussian kernel  $g(z) = \exp(-z^2/2)$  by taking with  $C_1^g = 1$  and  $C_2^g = e^{-1/2}$ . Another example is  $g(z) = \exp(-z^4/2)$  that satisfies the assumptions with  $C_1^g = \sqrt[8]{e}$  and  $C_2^g = 1.5$ .

### 3.2. Kernel Estimators and Optimism

At each episode  $k$ , **Kernel-UCBVI** computes an optimistic estimate  $Q_h^k$  for all  $h$ , which is an upper confidence bound on the optimal  $Q$  function  $Q_h^*$ , and plays the associated greedy policy. Let  $(x_h^s, a_h^s, x_{h+1}^s, r_h^s)$  be the random variables representing the state, the action, the next state and the reward at step  $h$  of episode  $s$ , respectively. We denote by  $\mathcal{D}_h = \{(x_h^s, a_h^s, x_{h+1}^s, r_h^s)\}_{s \in [k-1]}$  for  $h \in [H]$  the samples collected at step  $h$  before episode  $k$ .

For any state-action pair  $(x, a)$  and  $(s, h) \in [k-1] \times [H]$ , we define the *weights* and the *normalized weights* as

$$w_h^s(x, a) \stackrel{\text{def}}{=} \psi_\sigma((x, a), (x_h^s, a_h^s))$$

$$\text{and } \tilde{w}_h^s(x, a) \stackrel{\text{def}}{=} \frac{w_h^s(x, a)}{\beta + \sum_{l=1}^{k-1} w_h^l(x, a)}$$

where  $\beta > 0$  is a regularization term. These weights are used to compute an estimate of the reward and transition function for each state-action pair<sup>4</sup>:

$$\hat{r}_h^k(x, a) \stackrel{\text{def}}{=} \sum_{s=1}^{k-1} \tilde{w}_h^s(x, a) r_h^s,$$

$$\hat{P}_h^k(y|x, a) \stackrel{\text{def}}{=} \sum_{s=1}^{k-1} \tilde{w}_h^s(x, a) \delta_{x_{h+1}^s}(y).$$

As other algorithms using OFU, **Kernel-UCBVI** computes an optimistic Q-function  $\tilde{Q}_h^k$  through value iteration, a.k.a. backward induction:

$$\tilde{Q}_h^k(x, a) = \hat{r}_h^k(x, a) + \hat{P}_h^k V_{h+1}^k(x, a) + \mathbf{B}_h^k(x, a), \quad (1)$$

where  $V_{H+1}^k(x) = 0$  for all  $x \in \mathcal{X}$  and  $\mathbf{B}_h^k(x, a)$  is an exploration bonus described later. From Lemma 1, the true  $Q$  function  $Q_h^*$  is  $L_h$ -Lipschitz. Computing  $\tilde{Q}_h^k$  for all previously visited state action pairs  $(x_h^s, a_h^s)$  for  $s \in [k-1]$  permits to define a  $L_h$ -Lipschitz upper confidence bound and the associated value function:

$$Q_h^k(x, a) \stackrel{\text{def}}{=} \min_{s \in [k-1]} \left( \tilde{Q}_h^k(x_h^s, a_h^s) + L_h \rho[(x, a), (x_h^s, a_h^s)] \right)$$

$$\text{and } V_h^k(x) \stackrel{\text{def}}{=} \min \left( H - h + 1, \max_{a'} Q_h^k(x, a') \right).$$

<sup>4</sup>Here,  $\delta_x$  denotes the Dirac measure with mass at  $x$ .

---

#### Algorithm 1 **Kernel-UCBVI**

---

**Input:** global parameters  $K, H, \delta, \lambda_r, \lambda_p, \sigma, \beta$   
 initialize data lists  $\mathcal{D}_h = \emptyset$  for all  $h \in [H]$   
**for** episode  $k = 1, \dots, K$  **do**  
   get initial state  $x_1^k$   
    $Q_h^k = \text{optimisticQ}(k, \{\mathcal{D}_h\}_{h \in [H]})$   
   **for** step  $h = 1, \dots, H$  **do**  
     execute  $a_h^k = \arg\max_a Q_h^k(x_h^k, a)$   
     observe reward  $r_h^k$  and next state  $x_{h+1}^k$   
     add sample  $(x_h^k, a_h^k, x_{h+1}^k, r_h^k)$  to  $\mathcal{D}_h$   
   **end for**  
**end for**

---



---

#### Algorithm 2 **optimisticQ**

---

**Input:** episode  $k$ , data  $\{\mathcal{D}_h\}_{h \in [H]}$   
 Initialize  $V_{H+1}^k(x) = 0$  for all  $x$   
**for** step  $h = H, \dots, 1$  **do**  
    $M = \text{length}(\mathcal{D}_h)$   
   // Compute optimistic targets  
   **for**  $m = 1, \dots, M$  **do**  
      $\tilde{Q}_h^k(x_h^m, a_h^m) = \sum_{s=1}^M \tilde{w}_h^s(x_h^m, a_h^m) (r_h^s + V_{h+1}^k(x_{h+1}^s))$   
      $\tilde{Q}_h^k(x_h^m, a_h^m) = \tilde{Q}_h^k(x_h^m, a_h^m) + \mathbf{B}_h^k(x_h^m, a_h^m)$   
   **end for**  
   // Interpolate the Q function  
    $Q_h^k(x, a) = \min_{s \in [k-1]} \left( \tilde{Q}_h^k(x_h^s, a_h^s) + L_h \rho[(x, a), (x_h^s, a_h^s)] \right)$   
   **for**  $m = 1, \dots, M$  **do**  
      $V_h^k(x_h^m) = \min(H - h + 1, \max_{a \in \mathcal{A}} Q_h^k(x_h^m, a))$   
   **end for**  
**end for**  
**return**  $Q_h^k$

---

The policy  $\pi_k$  executed by **Kernel-UCBVI** is the greedy policy with respect to  $Q_h^k$  (see Alg. 1).

The exploration bonus is defined based on the uncertainties on the transition and reward estimates and takes the form

$$\mathbf{B}_h^k(x, a) = \sqrt{\frac{\mathbf{v}_{\text{rp}}(k, h)}{\mathbf{C}_h^k(x, a)}} + \frac{\beta}{\mathbf{C}_h^k(x, a)} + \sigma \mathbf{b}_{\text{rp}}(k, h)$$

where  $\mathbf{C}_h^k(x, a) \stackrel{\text{def}}{=} \beta + \sum_{s=1}^{k-1} w_h^s(x, a)$  are the *generalized counts*, which are a proxy for the number of visits to  $(x, a)$ ,  $\mathbf{v}_{\text{rp}}(k, h) = \tilde{\mathcal{O}}(H^2)$  and  $\mathbf{b}_{\text{rp}}(k, h)$  depend on the smoothness of the MDP and on kernel properties. Refer to Eq. 4 in App. B for an exact definition of  $\mathbf{B}_h^k$ .

**Remark 1.** If the transitions and/or rewards are known to be stationary, that is  $P_1 = \dots = P_H$  or  $r_1 = \dots = r_H$ , the corresponding kernel estimator can be modified to include a sum over all previous episodes and all time steps  $h \in [1, H]$ , and the bonus  $\mathbf{B}_h^k(x, a)$  becomes tighter (see App. E).

## 4. Theoretical Guarantees

The theorem below gives a high probability regret bound for **Kernel-UCBVI**. It features the  $\sigma$ -covering number of



the state-action space. The  $\sigma$ -covering number of a metric space, formally defined in Def. 1 (App. A), is roughly the number of  $\sigma$ -radius balls required to cover the entire space. The covering dimension of a space is the smallest number  $d$  such that its  $\sigma$ -covering number is  $\mathcal{O}(\sigma^{-d})$ . For instance, the covering number of a ball in  $\mathbb{R}^d$  with the Euclidean distance is  $\mathcal{O}(\sigma^{-d})$  and its covering dimension is  $d$ .

**Theorem 1.** *With probability at least  $1 - \delta$ , the regret of **Kernel-UCBVI** for a bandwidth  $\sigma$  is of order*

$$\mathcal{R}(K) \leq \tilde{\mathcal{O}} \left( H^2 \sqrt{|\mathcal{C}_\sigma| K} + HK\sigma + H^3 |\mathcal{C}_\sigma|^2 + H^2 |\mathcal{C}_\sigma| \right),$$

where  $|\mathcal{C}_\sigma|$  is the  $\sigma$ -covering number of  $\mathcal{X} \times \mathcal{A}$ .

*Proof.* Restatement of Theorem 4. See proof in App. D.  $\square$

**Corollary 1.** *By taking  $\sigma = (1/K)^{1/(2d+1)}$ , we have  $\mathcal{R}(K) = \tilde{\mathcal{O}} \left( H^3 K^{\max(\frac{1}{2}, \frac{2d}{2d+1})} \right)$ , where  $d$  is the covering dimension of the state-action space.*

*Proof.* Using Theorem 1 and the fact that the  $\sigma$ -covering number of  $(\mathcal{X} \times \mathcal{A}, \rho)$  is bounded by  $\mathcal{O}(\sigma^{-d})$ , we obtain  $\mathcal{R}(K) = \tilde{\mathcal{O}} \left( H^2 \sigma^{-d/2} \sqrt{K} + H^3 \sigma^{-2d} + HK\sigma \right)$ . Taking  $\sigma = (1/K)^{1/(2d+1)}$ , we see that the regret is  $\tilde{\mathcal{O}} \left( H^2 K^{\frac{3d+1}{4d+2}} + H^3 K^{\frac{2d}{2d+1}} \right)$ . The fact that  $(3d+1)/(4d+2) \leq 2d/(2d+1)$  for  $d \geq 1$  allows us to conclude.  $\square$

To the best of our knowledge, this is the first regret bound for an algorithm without discretization for stochastic continuous MDPs and it achieves the best dependence in  $d$  when compared to other *model-based* algorithms without further assumptions on the MDP. When  $d = 1$ , our bound has an optimal dependence in  $K$ , leading to a regret of order  $\tilde{\mathcal{O}}(H^3 K^{2/3})$ . This bound strictly improves the one derived in (Ortner & Ryabko, 2013). Under the stronger assumption that the transition kernels have densities that are  $\kappa$ -times differentiable<sup>5</sup>, the UCCRL-KD algorithm (Lakshmanan et al., 2015) achieve a regret of order  $T^{\frac{d+2}{d+3}}$ , which has a slightly better dependence in  $d$  (when  $d > 1$ ).

**Model-free vs. Model-based** An interesting remark comes from the comparison between our algorithm and recent model-free approaches in continuous MDPs (Song & Sun, 2019; Sinclair et al., 2019). These algorithms are based on optimistic Q-learning (Jin et al., 2018), to which we refer as OptQL, and achieve a regret of order  $\tilde{\mathcal{O}} \left( H^{\frac{5}{2}} K^{\frac{d+1}{2d+2}} \right)$ . This bound has an optimal dependence in  $K$  and  $d$ . While we achieve the same  $\tilde{\mathcal{O}}(K^{2/3})$  regret when  $d = 1$ , our

bound is slightly worse for  $d > 1$  and, to understand this gap, we look at the regret bound for tabular MDPs. Since our algorithm is inspired by UCBVI (Azar et al., 2017) with Chernoff-Hoeffding bonus, we compare it to OptQL, which is used by (Song & Sun, 2019; Sinclair et al., 2019), with the same kind of exploration bonus. Consider an MDP with  $X$  states and  $A$  actions and non-stationary transitions. UCBVI has a regret bound of  $\tilde{\mathcal{O}} \left( H^2 \sqrt{XAK} + H^3 X^2 A \right)$  while OptQL has  $\tilde{\mathcal{O}} \left( H^{5/2} \sqrt{XAK} + H^2 XA \right)$ . As we can see, OptQL is a  $\sqrt{H}$ -factor worse than UCBVI when comparing the first-order term, but it is  $HX$  times better in the second-order term. For large values of  $K$ , second-order terms can be neglected in the comparison of the algorithms in tabular MDPs, since they do not depend on  $K$ . However, they play an important role in continuous MDPs, where  $X$  and  $A$  are replaced by the  $\sigma$ -covering number of the state-action space, which is roughly  $1/\sigma^d$ . In tabular MDPs, the second-order term is constant (i.e., does not depend on  $K$ ). On the other hand, in continuous MDPs, the algorithms define the granularity of the representation of the state-action space based on the number of episodes, connecting the number of states  $X$  with  $K$ . For example, in (Song & Sun, 2019) the  $\epsilon$ -net used by the algorithm is tuned such that  $\epsilon = (HK)^{-1/(d+2)}$  (see also (Ortner & Ryabko, 2013; Lakshmanan et al., 2015; Qian et al., 2019)). Similarly, in our algorithm we have that  $\sigma = K^{-1/(2d+1)}$ . For this reason, the second-order term in UCBVI becomes the dominant term in our analysis, leading to a worse dependence in  $d$  compared to model-free algorithms, as highlighted in the proof sketch. For similar reasons, **Kernel-UCBVI** has an additional  $\sqrt{H}$  factor compared to model-free algorithms. This shows that the direction of achieving first-order optimal terms at the expense of higher second-order terms may not be justified outside the tabular case. Whether this is a flaw in the algorithm design or in the analysis is left as an open question.

**Remark 2.** *As for other model-based algorithms, the dependence on  $H$  can be improved if the transitions are stationary. In this case, the regret of **Kernel-UCBVI** becomes  $\tilde{\mathcal{O}} \left( H^2 K^{\frac{2d}{2d+1}} \right)$  due to a gain a factor of  $H$  in the second order term (see App. E). It is unclear whether one can obtain similar improvements for model-free algorithms.*

#### 4.1. Proof sketch

We now provide a sketch of the proof of Theorem 1. The complete proof can be found in the supplementary material. The analysis splits into three parts: (i) deriving confidence intervals for the reward and transition kernel estimators; (ii) proving that the algorithm is optimistic, i.e., that  $V_h^k(x) \geq V_h^*(x)$  for any  $(x, k, h)$  in a high probability event  $\mathcal{G}$ ; and (iii) proving an upper bound on the regret by using the fact that  $\mathcal{R}(K) = \sum_k (V_1^*(x_1^k) - V_1^{\pi_k}(x_1^k)) \leq$

<sup>5</sup>Our assumptions do not require densities to exist. For instance, the transition kernels in deterministic MDPs are Dirac measures, which do not have density.

$$\sum_k (V_1^k(x_1^k) - V_1^{\pi_k}(x_1^k)).$$

#### 4.1.1. CONCENTRATION

The most interesting part is the concentration of the transition kernel. Since  $\hat{P}_h^k(\cdot|x, a)$  are weighted sums of Dirac measures, we cannot bound the distance between  $P_h(\cdot|x, a)$  and  $\hat{P}_h^k(\cdot|x, a)$  directly. Instead, for  $V_{h+1}^*$  the optimal value function at step  $h+1$ , we bound the difference

$$\begin{aligned} & \left| (\hat{P}_h^k - P_h) V_{h+1}^*(x, a) \right| \\ &= \left| \sum_{s=1}^{k-1} \tilde{w}_s^h(x, a) V_{h+1}^*(x_{h+1}^s) - P_h V_{h+1}^*(x, a) \right| \\ &\leq \underbrace{\left| \sum_{s=1}^{k-1} \tilde{w}_s^h(x, a) (V_{h+1}^*(x_{h+1}^s) - P_h V_{h+1}^*(x_h^s, a_h^s)) \right|}_{(\mathbf{A})} \\ &\quad + \underbrace{\lambda_p L_{h+1} \sum_{s=1}^{k-1} \tilde{w}_s^h(x, a) \rho[(x, a), (x_h^s, a_h^s)]}_{(\mathbf{B})} + \underbrace{\frac{\beta \|V_{h+1}^*\|_\infty}{\mathbf{C}_h^k(x, a)}}_{(\mathbf{C})}. \end{aligned}$$

The term  $(\mathbf{A})$  is a weighted sum of a martingale difference sequence. To control it, we propose a new Hoeffding-type inequality, Lemma 2, that applies to weighted sums with random weights. The term  $(\mathbf{B})$  is a bias term that is obtained using the fact that  $V_{h+1}^*$  is  $L_{h+1}$ -Lipschitz and that the transition kernel is  $\lambda_p$ -Lipschitz, and can be shown to be proportional to the bandwidth  $\sigma$  under Assumption 4 (Lemma 7). The term  $(\mathbf{C})$  is the bias introduced by the regularization parameter  $\beta$ . Hence, for a fixed state-action pair  $(x, a)$ , we show that<sup>6</sup>, with high-probability,

$$\left| (\hat{P}_h^k - P_h) V_{h+1}^*(x, a) \right| \lesssim \frac{1}{\sqrt{\mathbf{C}_h^k(x, a)}} + \frac{1}{\mathbf{C}_h^k(x, a)} + \sigma$$

Then, we extend this bound to all  $(x, a)$  by leveraging the continuity of all the terms involving  $(x, a)$  and a covering argument. This continuity is a consequence of kernel smoothing, and it is a key point in avoiding a discretization of  $\mathcal{X} \times \mathcal{A}$  in the algorithm.

In Theorem 3, we define a favorable event  $\mathcal{G}$ , of probability larger than  $1 - \delta/2$ , in which (a more precise version of) the above inequality holds, the mean rewards belong to their confidence intervals, and we further control the deviations of  $(\hat{P}_h^k - P_h)f(x, a)$  for any  $2L_1$ -Lipschitz function  $f$ . This last part is obtained thanks to a new Bernstein-like concentration inequality for weighted sums (Lemma 3).

<sup>6</sup>Here,  $\lesssim$  means smaller than or equal up to constants, that can be logarithmic in  $k$ .

#### 4.1.2. OPTIMISM

To prove that the optimistic value function  $V_h^k$  is indeed an upper bound on  $V_h^*$ , we proceed by induction on  $h$  and we use the  $Q$  functions. When  $h = H + 1$ , we have  $Q_{H+1}^k(x, a) = Q_{H+1}^*(x, a) = 0$  for all  $(x, a)$ , by definition. Assuming that  $Q_{h+1}^k(x, a) \geq Q_{h+1}^*(x, a)$  for all  $(x, a)$ , we have  $V_{h+1}^k(x) \geq V_{h+1}^*(x)$  for all  $x$  and

$$\begin{aligned} & \tilde{Q}_h^k(x, a) - Q_h^*(x, a) \\ &= \underbrace{\tilde{r}_h^k(x, a) - r_h(x, a) + (\hat{P}_h^k - P_h) V_{h+1}^*(x, a) + \mathbf{B}_h^k(x, a)}_{\geq 0 \text{ in } \mathcal{G}} \\ &\quad + \underbrace{\hat{P}_h^k(V_{h+1}^k - V_{h+1}^*)(x, a)}_{\geq 0 \text{ by induction hypothesis}} \geq 0. \end{aligned}$$

for all  $(x, a)$ . In particular  $\tilde{Q}_h^k(x_h^s, a_h^s) - Q_h^*(x_h^s, a_h^s) \geq 0$  for all  $s \in [k-1]$ , which gives us

$$\begin{aligned} & \tilde{Q}_h^k(x_h^s, a_h^s) + L_h \rho[(x, a), (x_h^s, a_h^s)] \\ &\geq Q_h^*(x_h^s, a_h^s) + L_h \rho[(x, a), (x_h^s, a_h^s)] \geq Q_h^*(x, a) \end{aligned}$$

for all  $s \in [k-1]$ , since  $Q_h^*$  is  $L_h$ -Lipschitz. It follows from the definition of  $Q_h^k$  that  $Q_h^k(x, a) \geq Q_h^*(x, a)$ , which in turn implies that, for all  $x$ ,  $V_h^k(x) \geq V_h^*(x)$  in  $\mathcal{G}$ .

#### 4.1.3. BOUNDING THE REGRET

To provide an upper bound on the regret in the event  $\mathcal{G}$ , let  $\delta_h^k \stackrel{\text{def}}{=} V_h^k(x_h^k) - V_h^{\pi_k}(x_h^k)$ . The fact that  $V_h^k \geq V_h^*$  gives us  $\mathcal{R}(K) \leq \sum_k \delta_1^k$ . Introducing  $(\tilde{x}_h^k, \tilde{a}_h^k)$ , the state-action pair in the past data  $\mathcal{D}_h$  that is the closest to  $(x_h^k, a_h^k)$  and letting  $\square_h^k = \rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)]$ , we bound  $\delta_h^k$  using the following decomposition:

$$\begin{aligned} \delta_h^k &\leq Q_h^k(x_h^k, a_h^k) - Q_h^{\pi_k}(x_h^k, a_h^k) \\ &\leq \tilde{Q}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) - Q_h^{\pi_k}(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_h \square_h^k \\ &\leq \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_h(1 + \lambda_p) \square_h^k \\ &\quad \textcircled{1} + (\hat{P}_h^k - P_h) V_{h+1}^*(\tilde{x}_h^k, \tilde{a}_h^k) \\ &\quad \textcircled{2} + P_h (V_{h+1}^k - V_{h+1}^{\pi_k})(x_h^k, a_h^k) \\ &\quad \textcircled{3} + (\hat{P}_h^k - P_h) (V_{h+1}^k - V_{h+1}^*)(\tilde{x}_h^k, \tilde{a}_h^k) \end{aligned}$$

The term  $\textcircled{1}$  is shown to be smaller than  $\mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)$ , by definition of the bonus. The term  $\textcircled{2}$  can be rewritten as  $\delta_{h+1}^k$  plus a martingale difference sequence  $\xi_{h+1}^k$ . To bound the term  $\textcircled{3}$ , we use that  $V_{h+1}^k - V_{h+1}^*$  is  $2L_1$ -Lipschitz. The uniform deviations that hold on event  $\mathcal{G}$  yield

$$\textcircled{3} \lesssim \frac{1}{H} (\delta_{h+1}^k + \xi_{h+1}^k) + \frac{H^2 |C_\sigma|}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \square_h^k + \sigma.$$

Let  $C_\sigma$  be a  $\sigma$ -covering of the state-action space. If, for a given episode  $k$ , there exists  $h$  such that  $\square_h^k > \sigma$ , we bound

the regret of the entire episode by  $H$ . The number of times that this can happen is bounded by  $\mathcal{O}(H|C_\sigma|)$ . Hence, the regret due to this term is bounded by  $\mathcal{O}(H^2|C_\sigma|)$ . The sum of  $\xi_{h+1}^k$  over  $(k, h)$  is bounded by  $\tilde{\mathcal{O}}(H^{\frac{3}{2}}\sqrt{K})$  by Hoeffding-Azuma's inequality, on some event  $\mathcal{F}$  of probability larger than  $1 - \delta/2$ . Hence, we focus on the episodes where  $\square_h^k \leq \sigma$  and we omit the terms involving  $\xi_{h+1}^k$ . Using the definition of the bonus, we obtain

$$\delta_h^k \lesssim \left(1 + \frac{1}{H}\right) \delta_{h+1}^k + \frac{H}{\sqrt{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)}} + \frac{H^2|C_\sigma|}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \sigma.$$

Using the fact that  $(1 + 1/H)^H \leq e$ , we have, on  $\mathcal{G} \cap \mathcal{F}$ ,

$$\mathcal{R}(K) \lesssim \sum_{h,k} \left( \frac{H}{\sqrt{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)}} + \frac{H^2|C_\sigma|}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} \right) + KH\sigma.$$

The term in  $1/\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)$  is the *second order term* (in  $K$ ). In the tabular case, it is multiplied by the number of states. Here, it is multiplied by the covering number  $|C_\sigma|$ .

From there it remains to bound the sum of the first and second-order terms, and we specifically show that

$$\sum_{h,k} \frac{1}{\sqrt{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)}} \lesssim H\sqrt{|C_\sigma|K} \quad (2)$$

$$\text{and} \quad \sum_{h,k} \frac{1}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} \lesssim H|C_\sigma|\log K, \quad (3)$$

where we note that (3) has a worse dependency in  $|C_\sigma|$ . As mentioned before, unlike in the tabular case the sum of "second-order" terms will actually be the leading term.

Finally, we obtain that on  $\mathcal{G} \cap \mathcal{F}$  (of probability  $\geq 1 - \delta$ )

$$\mathcal{R}(K) \lesssim H^2\sqrt{|C_\sigma|K} + H^3|C_\sigma|^2 + KH\sigma + H^2|C_\sigma|,$$

where the extra  $H^2|C_\sigma|$  takes into account the episodes where  $\square_h^k > \sigma$ .

If the transitions kernels are stationary, i.e.,  $P_1 = \dots = P_H$ , the bounds (2) and (3) can be improved to  $\sqrt{|C_\sigma|KH}$  and  $|C_\sigma|\log(KH)$  respectively, thus improving the final scaling in  $H$ .<sup>7</sup> See App. E for details.

## 5. Improving the Computational Complexity

**Kernel-UCBVI** is a non-parametric model-based algorithm and, consequently, it inherits the weaknesses of these approaches. In order to be data adaptive, it needs to store all

<sup>7</sup>This is because, in the non-stationary case, we bound the sums over  $k$  and then multiply the resulting bound by  $H$ . In the stationary case, we can directly bound the sums over  $(k, h)$ .

---

### Algorithm 3 Greedy-Kernel-UCBVI

---

**Input:** global parameters  $K, H, \delta, \lambda_r, \lambda_p, \sigma, \beta$   
 initialize  $\mathcal{D}_h = \emptyset$  and  $V_h^1(x) = H - h + 1$ , for all  $h \in [H]$   
**for** episode  $k = 1, \dots, K$  **do**  
   get initial state  $x_1^k$   
   **for** step  $h = 1, \dots, H$  **do**  
     compute  $\tilde{Q}_h^k(x_h^k, a)$  for all  $a$  as in Eq. 1 using  $\mathcal{D}_h$   
     execute  $a_h^k = \operatorname{argmax}_a \tilde{Q}_h^k(x_h^k, a)$ , observe  $r_h^k$  and  $x_{h+1}^k$   
      $\tilde{V}_h^k(x_h^k) = \min(H - h + 1, \max_{a \in \mathcal{A}} \tilde{Q}_h^k(x_h^k, a))$   
     // Interpolate  
     define  $V_h^{k+1}$  for all  $x \in \mathcal{D}_h$  as  
        $V_h^{k+1}(x) = \min\left(\min_{s \in [k-1]} [V_h^k(x_h^s) + L_h \rho_{\mathcal{X}}(x, x_h^s)], \right.$   
        $\left. \tilde{V}_h^k(x_h^k) + L_h \rho_{\mathcal{X}}(x, x_h^k) \right)$   
     add sample  $(x_h^k, a_h^k, x_{h+1}^k, r_h^k)$  to  $\mathcal{D}_h$   
   **end for**  
**end for**

---

the samples  $(x_h^k, a_h^k, x_{h+1}^k, r_h^k)$  and their optimistic values  $\tilde{Q}_h^k$  and  $V_h^k$  for  $(k, h) \in [K] \times [H]$ , leading to a total memory complexity of  $\mathcal{O}(HK)$ . Like standard model-based algorithms, it needs to perform planning at each episode which gives a total runtime of  $\mathcal{O}(HAK^3)$ <sup>8</sup>, where the factor  $A$  takes into account the complexity of computing the maximum over actions. **Kernel-UCBVI** has the similar time and space complexity of recent approaches for low-rank MDPs (e.g., Jin et al., 2019; Zanette et al., 2019).

To alleviate the computational burden of **Kernel-UCBVI**, we leverage Real-Time Dynamic Programming (RTDP), see (Barto et al., 1995), to perform incremental planning. Similarly to OptQL, RTDP-like algorithms maintain an optimistic estimate of the optimal value function that is updated incrementally by interacting with the MDP. The main difference is that the update is done by using an estimate of the MDP (i.e., model-based) rather than the observed transition sample. At episode  $k$  and step  $h$ , our algorithm, named **Greedy-Kernel-UCBVI**, computes an upper bound  $\tilde{Q}_h^k(x_h^k, a)$  for each action  $a$  using the kernel estimate as in Eq. 1. Then, it executes the greedy action  $a_h^k = \operatorname{argmax}_{a \in \mathcal{A}} \tilde{Q}_h^k(x_h^k, a)$ . As a next step, it computes  $\tilde{V}_h^k(x_h^k) = \tilde{Q}_h^k(x_h^k, a_h^k)$  and refines the previous  $L_h$ -Lipschitz upper confidence bound on the value function

$$V_h^{k+1}(x) = \min(V_h^k(x), \tilde{V}_h^k(x_h^k) + L_h \rho_{\mathcal{X}}(x, x_h^k)).$$

The complete description of **Greedy-Kernel-UCBVI** is given in Alg. 3. The total runtime of this efficient version is  $\mathcal{O}(HAK^2)$  with total memory complexity of  $\mathcal{O}(HK)$ .

RTDP has been recently analyzed by Efroni et al. (2019) in tabular MDPs. Following their analysis, we prove the following theorem, which shows that

<sup>8</sup>Since the runtime of an episode  $k$  is  $\mathcal{O}(HAK^2)$ .

**Greedy-Kernel-UCBVI** achieves the same guarantees of **Kernel-UCBVI** with a large improvement in computational complexity.

**Theorem 2.** *With probability at least  $1 - \delta$ , the regret of **Greedy-Kernel-UCBVI** for a bandwidth  $\sigma$  is of order  $\mathcal{R}(K) = \tilde{\mathcal{O}}(\mathcal{R}(K, \text{Kernel-UCBVI}) + H^2 |\mathcal{C}'_\sigma|)$ , where  $|\mathcal{C}'_\sigma|$  is the  $\sigma$ -covering number of state space. Leading to a regret of  $\tilde{\mathcal{O}}(H^3 K^{2d/(2d+1)})$  when  $\sigma = (1/K)^{1/(2d+1)}$ .*

*Proof.* The complete proof is provided in App. F. The key properties for proving this regret bound are: *i*) optimism; *ii*) the fact that  $(V_h^k)$  are point-wise non-increasing:

**Proposition 1.** *Under the good event  $\mathcal{G}$ , we have that  $\forall x \in \mathcal{X}, V_h^k(x) \geq V_h^*(x)$  and  $V_h^k(x) \geq V_h^{k+1}(x)$ .*

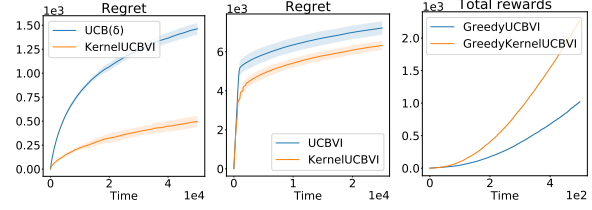
## 6. Experiments

To test the effectiveness of **Kernel-UCBVI**, we implemented it in three toy problems: a Lipschitz bandit problem (MDP with 1 state and  $H = 1$ ), a discrete  $8 \times 8$  GridWorld and a continuous version of a GridWorld, as described below. For the bandit problem, we compared it to a version of the  $\text{UCB}(\delta)$  (Abbasi-Yadkori et al., 2011). For the discrete MDP, we used  $\text{UCBVI}$  (Azar et al., 2017) as a baseline. For the continuous MDP, we implemented **Greedy-Kernel-UCBVI** and compared it to **Greedy-UCBVI** (Efroni et al., 2019) applied to a fixed discretization of the MDP. In all experiments, we used the Gaussian kernel  $g(z) = \exp(-z^2/2)$ . In both MDP experiments, the horizon was set to  $H = 20$ .

**Lipschitz bandit** We consider the 1-Lipschitz reward function  $r(a) = \max(a, 1 - a)$  for  $a \in [0, 1]$ . At each time  $k$ , the agent computes an optimistic reward function  $r_k$ , chooses the action  $a_k \in \arg\max_a r_k(a)$ , and observes  $r(a_k)$  plus noise. In order to solve this optimization problem, we choose 200 uniformly spaced points in  $[0, 1]$ . We chose a time-dependent kernel bandwidth in each episode as  $\sigma_k = 1/\sqrt{k}$ . For  $\text{UCB}(\delta)$ , we use the 200 points as arms.

**Discrete MDP** We consider a  $8 \times 8$  GridWorld whose states are a uniform grid of points in  $[0, 1]^2$  and 4 actions, left, right, up and down. When an agent takes an action, it goes to the corresponding direction with probability 0.9 and to any other neighbor state with probability 0.1. The agent starts at  $(0, 0)$  and the reward functions depend on the distance to the goal state  $(1, 1)$ . We chose a time-dependent kernel bandwidth in each episode as  $\sigma_k = \log k / \sqrt{k}$ , which allowed the agent to better exploit the smoothness of the MDP to quickly eliminate suboptimal actions in early episodes.

**Continuous MDP** We consider a variant of the previous environment having continuous state space  $\mathcal{X} = [0, 1]^2$ . When an agent takes an action (left, right, up or down) in a state  $x$ , its next state is  $x + \Delta x + \eta$ , where  $\Delta x$  is



**Figure 1.** **Left:** Regret of **Kernel-UCBVI** versus  $\text{UCB}(\delta)$  on a Lipschitz bandit (averaged over 20 runs). **Middle:** Regret of **Kernel-UCBVI** versus  $\text{UCBVI}$  on a  $8 \times 8$  GridWorld (averaged over 10 runs). **Right:** Total sum of rewards gathered by **Greedy-Kernel-UCBVI** in a continuous MDP versus **Greedy-UCBVI** in a discretized version of the MDP (averaged over 8 runs). The shaded regions represent  $\pm$  the standard deviation.

a displacement in the direction of the action and  $\eta$  is a noise. The agent starts at  $(0.1, 0.1)$  and the reward functions depend on the distance to the goal state  $(0.75, 0.75)$ . The bandwidth was fixed to  $\sigma = 0.1$ . For **Greedy-UCBVI**, we discretize the state-action space with a uniform grid with steps of size 0.1, matching the value of  $\sigma$ .

Figure 1 shows the performance of **Kernel-UCBVI** and its greedy version compared to the baselines described above. We see that **Kernel-UCBVI** has a better regret than  $\text{UCB}(\delta)$  and  $\text{UCBVI}$  in discrete environments. Also, in the continuous MDP, **Greedy-Kernel-UCBVI** outperforms **Greedy-UCBVI** applied in a uniform discretization, which shows that our algorithm exploits better the smoothness of the MDP. In Appendix I we provide more details about the experiments, in particular the choice of the exploration bonuses which were chosen to improve the learning speed, both for **Kernel-UCBVI** and for the baselines.

## 7. Conclusion

In this paper, we introduced **Kernel-UCBVI**, a model-based algorithm for finite-horizon reinforcement learning in metric spaces which employs kernel smoothing to estimate rewards and transitions. By providing new high-probability confidence intervals for weighted sums and non-parametric kernel estimators, we generalize the techniques introduced by Azar et al. (2017) in tabular MDPs to the continuous setting. We prove that the regret of **Kernel-UCBVI** is of order  $H^3 K^{\max(\frac{1}{2}, \frac{2d}{2d+1})}$ , which improves upon previous model-based algorithms under mild assumptions. In addition, we provide experiments illustrating the effectiveness of **Kernel-UCBVI** against baselines in discrete and continuous environments. As future work, we plan to extend our techniques to the model-free setting, and investigate further the gap that may exist between model-based and model-free methods in the continuous case.



## References

- Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In Kakade, S. M. and von Luxburg, U. (eds.), *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 1–26, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL <http://proceedings.mlr.press/v19/abbasi-yadkorilla.html>.
- Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved Algorithms for Linear Stochastic Bandits. *Advances in Neural Information Processing Systems (NIPS)*, pp. 2312—2320, 2011.
- Azar, M. G., Osband, I., and Munos, R. Minimax Regret Bounds for Reinforcement Learning. 2017.
- Barto, A. G., Bradtke, S. J., and Singh, S. P. Learning to act using real-time dynamic programming. *Artificial intelligence*, 72(1-2):81–138, 1995.
- Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- Burnetas, A. N. and Katehakis, M. N. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, feb 1997. ISSN 0364-765X.
- Chowdhury, S. R. and Gopalan, A. Online learning in kernelized markov decision processes. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3197–3205. PMLR, 16–18 Apr 2019. URL <http://proceedings.mlr.press/v89/chowdhury19a.html>.
- Combes, R., Magureanu, S., and Proutière, A. Minimal exploration in structured stochastic bandits. In *NIPS*, pp. 1763–1771, 2017.
- Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pp. 12203–12213, 2019.
- Gottlieb, L. A., Kontorovich, A., and Krauthgamer, R. Efficient Regression in Metric Spaces via Approximate Lipschitz Extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, 2017. ISSN 00189448. doi: 10.1109/TIT.2017.2713820.
- Jaksch, T., Ortner, R., and Auer, P. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 99:1563–1600, aug 2010. ISSN 1532-4435.
- Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning Provably Efficient? (NeurIPS), 2018. URL <http://arxiv.org/abs/1807.03765>.
- Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably Efficient Reinforcement Learning with Linear Function Approximation. pp. 1–28, 2019. URL <http://arxiv.org/abs/1907.05388>.
- Kakade, S. M., Kearns, M. J., and Langford, J. Exploration in Metric State Spaces. *ICML*, pp. 306–312, 2003. URL <http://www.aaai.org/Papers/ICML/2003/ICML03-042.pdf>.
- Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.
- Lakshmanan, K., Ortner, R., and Ryabko, D. Improved Regret Bounds for Undiscounted Continuous Reinforcement Learning. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015. URL <http://arxiv.org/abs/1302.2550>.
- Lattimore, T., Hutter, M., and Sunehag, P. The sample-complexity of general reinforcement learning. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 28–36. JMLR.org, 2013.
- Ok, J., Proutière, A., and Tranos, D. Exploration in structured reinforcement learning. In *NeurIPS*, pp. 8888–8896, 2018.
- Ormoneit, D. and Sen, S. Kernel-based reinforcement learning. *Machine Learning*, 49(2):161–178, Nov 2002. ISSN 1573-0565. doi: 10.1023/A:1017928328829. URL <https://doi.org/10.1023/A:1017928328829>.
- Ortner, R. and Ryabko, D. Online Regret Bounds for Undiscounted Continuous Reinforcement Learning. 2013. URL <http://arxiv.org/abs/1302.2550>.
- Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.
- Pazis, J. and Parr, R. PAC optimal exploration in continuous space markov decision processes. In *AAAI*. AAAI Press, 2013.
- Peña, V. H., Lai, T. L., and Shao, Q.-M. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.
- Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, Inc., New York, NY, USA, 1994. ISBN 0471619779.

- Qian, J., Fruit, R., Pirotta, M., and Lazaric, A. Exploration bonus for regret minimization in discrete and continuous average reward mdps. In *NeurIPS*, pp. 4891–4900, 2019.
- Reeve, H. W., Mellor, J., and Brown, G. The K-Nearest Neighbour UCB algorithm for multi-armed bandits with covariates. pp. 1–29, 2018. URL <http://arxiv.org/abs/1803.00316>.
- Sinclair, S. R., Banerjee, S., and Yu, C. L. Adaptive Discretization for Episodic Reinforcement Learning in Metric Spaces. pp. 1–46, 2019. URL <http://arxiv.org/abs/1910.08151>.
- Song, Z. and Sun, W. Efficient Model-free Reinforcement Learning in Metric Spaces. 2019. URL <http://arxiv.org/abs/1905.00475>.
- Strens, M. A bayesian framework for reinforcement learning. In *ICML*, volume 2000, pp. 943–950, 2000.
- Yang, L. F. and Wang, M. Reinforcement Learning in Feature Space: Matrix Bandit, Kernels, and Regret Bound. pp. 1–26, 2019. URL <http://arxiv.org/abs/1905.10389>.
- Yang, L. F., Ni, C., and Wang, M. Learning to Control in Metric Space with Optimal Regret. pp. 1–20, 2019. URL <http://arxiv.org/abs/1905.01576>.
- Zanette, A. and Brunskill, E. Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds. 2019. URL <http://arxiv.org/abs/1901.00210>.
- Zanette, A., Brandfonbrener, D., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. *CoRR*, abs/1911.00567, 2019.

# Appendices

## A. Notation and preliminaries

### A.1. Notation

Table 1. Table of notations

| Notation   | Meaning  |
|--|--|
| $\rho : (\mathcal{X} \times \mathcal{A})^2 \rightarrow \mathbb{R}_+$ | metric on the state-action space $\mathcal{X} \times \mathcal{A}$                        |
| $\psi_\sigma((x, a), (x', a'))$                                      | kernel function with bandwidth $\sigma$  |
| $g : \mathbb{R}_+ \rightarrow [0, 1]$                                | “mother” kernel function such that $\psi_\sigma(u, v) = g(\rho[u, v] / \sigma)$          |
| $C_1^g, C_2^g$   | positive constants that depend on $g$ (Assumption 4)                                     |
| $\mathcal{N}(\epsilon, \mathcal{X} \times \mathcal{A}, \rho)$        | $\epsilon$ -covering number of the metric space $(\mathcal{X} \times \mathcal{A}, \rho)$ |
| $\mathbf{b}_p(k, h), \mathbf{v}_p(k, h)$                             | bias and variance parameter of the transition bonus (see Cor. 3)                         |
| $\mathbf{b}_r(k, h), \mathbf{v}_r(k, h)$                             | bias and variance parameter of the reward bonus (see Cor. 3)                             |
| $\theta_b(k, h), \theta_v(k, h)$                                     | Bernstein bias and variance parameter (see Cor. 3)                                       |
| $\mathcal{G}$  | “good” event (see Corollary 3)   |
| $L_h, \text{ for } h \in [H]$  | Lipschitz constant of value functions (see Lemma 4)                                      |

### A.2. Preliminaries

Let  $\sigma > 0$ . We define the *weights* as

$$w_h^s(x, a) \stackrel{\text{def}}{=} \psi_\sigma((x, a), (x_h^s, a_h^s))$$

and the *normalized weights* as

$$\tilde{w}_h^s(x, a) \stackrel{\text{def}}{=} \frac{w_h^s(x, a)}{\beta + \sum_{l=1}^{k-1} w_h^l(x, a)}$$

where  $\beta > 0$  is a regularization parameter. We define the generalized count at  $(x, a)$  at time  $(k, h)$  as

$$\mathbf{C}_h^k(x, a) \stackrel{\text{def}}{=} \beta + \sum_{s=1}^{k-1} w_h^s(x, a).$$

We define the following estimator for the transition kernels  $\{P_h\}_{h \in [H]}$

$$\hat{P}_h^k(y|x, a) \stackrel{\text{def}}{=} \sum_{s=1}^{k-1} \tilde{w}_h^s(x, a) \delta_{x_{h+1}^s}(y)$$

and the following estimator for the reward functions  $\{r_h\}_{h \in [H]}$

$$\hat{r}_h^k(x, a) \stackrel{\text{def}}{=} \sum_{s=1}^{k-1} \tilde{w}_h^s(x, a) r_h^s.$$

For any function  $V : \mathbb{R} \rightarrow \mathbb{R}$ , we recall that

$$P_h V(x, a) = \int_{\mathcal{X}} V(y) dP_h(y|x, a) \quad \text{and} \quad \hat{P}_h^k V(x, a) = \sum_{s=1}^{k-1} \tilde{w}_h^s(x, a) V(x_{h+1}^s).$$

We will also using the notion of covering of metric spaces, according to the definition below.

**Definition 1** (covering of a metric space). Let  $(\mathcal{U}, \rho)$  be a metric space. For any  $u \in \mathcal{U}$ , let  $\mathcal{B}(u, \sigma) = \{v \in \mathcal{U} : \rho(u, v) \leq \sigma\}$ . We say that a set  $\mathcal{C}_\sigma \subset \mathcal{U}$  is a  $\sigma$ -covering of  $(\mathcal{U}, \rho)$  if  $\mathcal{U} \subset \bigcup_{u \in \mathcal{C}_\sigma} \mathcal{B}(u, \sigma)$ . In addition, we define the  $\sigma$ -covering number of  $(\mathcal{U}, \rho)$  as  $\mathcal{N}(\sigma, \mathcal{U}, \rho) \stackrel{\text{def}}{=} \min \{|\mathcal{C}_\sigma| : \mathcal{C}_\sigma \text{ is a } \sigma\text{-covering of } (\mathcal{U}, \rho)\}$ .

## B. Description of the algorithm

At the beginning of each episode  $k$ , the agent has observed the data  $\mathcal{D}_h = \{(x_h^s, a_h^s, x_{h+1}^s, r_h^s)\}_{s \in [k-1]}$  for  $h \in [H]$ . The number of data tuples in each  $\mathcal{D}_h$  is  $k-1$ .

At each step  $h$  of episode  $k$ , the agent has access to an optimistic value function at step  $h+1$ , denoted by  $V_{h+1}^k$ . Using this optimistic value function, the agent computes an upper bound for the  $Q$  function at each state-action pair in the data, denoted by  $\tilde{Q}_h^k(x_h^s, a_h^s)$  for  $s \in [k-1]$ , which we call *optimistic targets*. For any  $(x, a)$ , we can compute an optimistic target as

$$\tilde{Q}_h^k(x, a) = \hat{r}_h^k(x, a) + \hat{P}_h^k V_{h+1}^k(x, a) + \mathbf{B}_h^k(x, a)$$

where  $\mathbf{B}_h^k(x, a)$  is an exploration bonus for the pair  $(x, a)$ ,

$$\begin{aligned} \mathbf{B}_h^k(x, a) &= \mathbf{p}_h^k(x, a) + \mathbf{r}_h^k(x, a) \\ &= \underbrace{\left( \sqrt{\frac{\mathbf{v}_p(k, h)}{\mathbf{C}_h^k(x, a)}} + \frac{\beta}{\mathbf{C}_h^k(x, a)} + \mathbf{b}_p(k, h)\sigma \right)}_{\text{transition bonus}} + \underbrace{\left( \sqrt{\frac{\mathbf{v}_r(k, h)}{\mathbf{C}_h^k(x, a)}} + \frac{\beta}{\mathbf{C}_h^k(x, a)} + \mathbf{b}_r(k, h)\sigma \right)}_{\text{reward bonus}}, \end{aligned} \quad (4)$$

which represents the sum of uncertainties on the transitions and rewards estimates. The exact definition of  $\mathbf{v}_p(k, h)$ ,  $\mathbf{b}_p(k, h)$ ,  $\mathbf{v}_r(k, h)$  and  $\mathbf{b}_r(k, h)$  are given in Corollary 3.

Then, we build an optimistic  $Q$  function  $Q_h^k$  by interpolating the optimistic targets:

$$\forall (x, a), \quad Q_h^k(x, a) \stackrel{\text{def}}{=} \min_{s \in [k-1]} \left[ \tilde{Q}_h^k(x_h^s, a_h^s) + L_h \rho[(x, a), (x_h^s, a_h^s)] \right] \quad (5)$$

and the value function  $V_h^k$  is computed as

$$\forall x, \quad V_h^k(x) \stackrel{\text{def}}{=} \min \left( H - h + 1, \max_{a'} Q_h^k(x, a') \right).$$

We can check that  $(x, a) \mapsto Q_h^k(x, a)$  is  $L_h$ -Lipschitz with respect to  $\rho$  and that  $(x) \mapsto V_h^k(x)$  is  $L_h$ -Lipschitz with respect to  $\rho_{\mathcal{X}}$ .

## C. Concentration

The first step towards proving our regret bound is to derive confidence intervals for the rewards and transitions, which are presented in propositions 2 and 3, respectively.

In addition, we need Bernstein-type inequalities for the transition kernels, which are stated in propositions 4 and 5.

Finally, Corollary 3 defines a favorable event in which all the confidence intervals that we need to prove our regret bound are valid and we prove that this event happens with high probability.

### C.1. Confidence intervals for the reward functions

**Proposition 2.** *We have:*

$$\mathbb{P} \left[ \exists (k, h) \in [K] \times [H], \exists (x, a) : |\hat{r}_h^k(x, a) - r_h(x, a)| \geq \sqrt{\frac{\mathbf{v}_r(k, h)}{\mathbf{C}_h^k(x, a)}} + \frac{\beta}{\mathbf{C}_h^k(x, a)} + \mathbf{b}_r(k, h)\sigma \right] \leq H \mathcal{N} \left( \frac{\sigma^2}{K}, \mathcal{X} \times \mathcal{A}, \rho \right) \delta$$



where

$$\begin{aligned} \mathbf{v}_r(k, h) &= 2 \log \left( \frac{\sqrt{1 + k/\beta}}{\delta} \right) \\ \mathbf{b}_r(k, h) &= 2\lambda_r \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right) + \frac{2C_2^g}{\beta^{3/2}} \sqrt{2 \log \left( \frac{\sqrt{1 + k/\beta}}{\delta} \right)} + \frac{2C_2^g}{\beta} \end{aligned}$$

*Proof.* The proof is almost identical to the proof of Proposition 3. The main difference is that  $\mathbf{v}_r(k, h)$  is smaller than  $\mathbf{v}_p(k, h)$  by a factor of  $H^2$ , which is due to the fact that the reward function is bounded in  $[0, 1]$  whereas the value functions are bounded in  $[0, H]$ .  $\square$

## C.2. Confidence intervals for the transition kernels

**Proposition 3.** Let  $V : \mathcal{X} \rightarrow [0, H]$  be a deterministic function that is  $L_1$ -Lipschitz. Then,

$$\mathbb{P} \left[ \exists(k, h) \in [K] \times [H], \exists(x, a) : \left| \hat{P}_h^k V(x, a) - P_h V(x, a) \right| \geq \sqrt{\frac{\mathbf{v}_p(k, h)}{\mathbf{C}_h^k(x, a)}} + \frac{H\beta}{\mathbf{C}_h^k(x, a)} + \mathbf{b}_p(k, h)\sigma \right] \leq H\mathcal{N} \left( \frac{\sigma^2}{HK}, \mathcal{X} \times \mathcal{A}, \rho \right) \delta$$

where

$$\begin{aligned} \mathbf{v}_p(k, h) &= 2H^2 \log \left( \frac{\sqrt{1 + k/\beta}}{\delta} \right) \\ \mathbf{b}_p(k, h) &= 2\lambda_p L_1 \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right) + \frac{2C_2^g}{\beta^{3/2}} \sqrt{\log \left( \frac{\sqrt{1 + k/\beta}}{\delta} \right)} + \frac{2C_2^g}{\beta} \end{aligned}$$

*Proof.* We want to bound the probability of the failure event  $F$ , defined as

$$F \stackrel{\text{def}}{=} \left\{ \exists(k, h) \in [K] \times [H], \exists(x, a) : \left| \hat{P}_h^k V(x, a) - P_h V(x, a) \right| \geq \sqrt{\frac{\mathbf{v}_p(k, h)}{\mathbf{C}_h^k(x, a)}} + \frac{H\beta}{\mathbf{C}_h^k(x, a)} + \mathbf{b}_p(k, h)\sigma \right\}.$$

We proceed as follows:

- First, we fix a pair  $(x, a)$  and bound the probability of a failure at this pair, denoted by  $F(x, a)$ ,
- Second, we use the technical Lemma 8 to show that the upper confidence intervals, seen as a function of  $(x, a)$ , are Lipschitz continuous and, consequently, we can extend our bound to every  $(x, a)$  by using a covering argument (technical Lemma 6).

**Bounding the failure probability at a fixed  $(x, a)$**  Consider the failure event  $F(x, a)$  for a fixed state-action pair  $(x, a)$ :

$$F(x, a) \stackrel{\text{def}}{=} \left\{ \exists(k, h) : \left| \hat{P}_h^k V(x, a) - P_h V(x, a) \right| \geq \sqrt{\mathbf{v}_p(k, h)/\mathbf{C}_h^k(x, a)} + H\beta/\mathbf{C}_h^k(x, a) + 2\lambda_p L_1 \sigma \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right) \right\}$$

In  $F(x, a)$ , we have:

$$\begin{aligned} & \sqrt{\mathbf{v}_p(k, h)/\mathbf{C}_h^k(x, a)} + H\beta/\mathbf{C}_h^k(x, a) + 2\lambda_p L_1 \sigma \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right) \\ & \leq \left| \sum_{s=1}^{k-1} \tilde{w}_h^s(x, a) \underbrace{\left( V(x_{h+1}^s) - \int_{\mathcal{X}} V(y) dP_h(y|x, a) \right)}_{\text{(A)}} + \frac{\beta}{\beta + \sum_{l=1}^{k-1} w_h^l(x, a)} \underbrace{\int_{\mathcal{X}} V(y) dP_h(y|x, a)}_{\text{(B)}} \right|. \end{aligned}$$

The term **(B)** is bounded by  $H$  and the term **(A)** is bounded as follows

$$\begin{aligned}
 \mathbf{(A)} &= V(x_{h+1}^s) - \int_{\mathcal{X}} V(y) dP_h(y|x, a) \\
 &= V(x_{h+1}^s) - \int_{\mathcal{X}} V(y) dP_h(y|x_h^s, a_h^s) + \int_{\mathcal{X}} V(y) (dP_h(y|x_h^s, a_h^s) - dP_h(y|x, a)) \\
 &\leq V(x_{h+1}^s) - \int_{\mathcal{X}} V(y) dP_h(y|x_h^s, a_h^s) + L_{h+1} W_1 (P_h(\cdot|x_h^s, a_h^s), P_h(\cdot|x, a)) \\
 &\leq V(x_{h+1}^s) - \int_{\mathcal{X}} V(y) dP_h(y|x_h^s, a_h^s) + \lambda_p L_1 \rho [(x_h^s, a_h^s), (x, a)].
 \end{aligned}$$

Technical Lemma 7 gives us

$$\sum_{s=1}^{k-1} \tilde{w}_h^s(s, a) \rho [(x_h^s, a_h^s), (x, a)] \leq 2\sigma \left( 1 + \sqrt{\log(C_1^g k / \beta)} \right)$$

Hence,

$$\forall(x, a), \mathbb{P}[F(x, a)] \leq \mathbb{P} \left[ \exists(k, h) : \left| \frac{1}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} w_h^s(x, a) \left( V(x_{h+1}^s) - \int_{\mathcal{X}} V(y) dP_h(y|x_h^s, a_h^s) \right) \right| \geq \sqrt{\frac{\mathbf{v}_p(k, h)}{\mathbf{C}_h^k(x, a)}} \right]$$

Let  $W_s \stackrel{\text{def}}{=} V(x_{h+1}^s) - \int_{\mathcal{X}} V(y) dP_h(y|x_h^s, a_h^s)$  and consider the filtration  $\mathcal{F}_h^s$  generated by  $\{(x_h^t, a_h^t) : t \leq s+1\}$ . We have  $\mathbb{E}[W_s | \mathcal{F}_h^{s-1}] = 0$  and  $|W_s| \leq 2H$ . As a consequence of Lemma 2, which gives us a concentration inequality for weighted sums, we obtain

$$\forall(x, a), \mathbb{P}[F(x, a)] \leq H\delta.$$

by doing a union bound over  $h \in [H]$ .

**Extending the bound to every  $(x, a)$  by a covering argument** Now, we want to bound the probability  $\mathbb{P}[F]$  using our bound for  $\mathbb{P}[F(x, a)]$  for every  $(x, a)$ . We define

$$f_1(x, a) \stackrel{\text{def}}{=} \left| \frac{1}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} w_h^s(x, a) W_s \right| \text{ and } f_2(x, a) \stackrel{\text{def}}{=} \sqrt{\frac{\mathbf{v}_p(k, h)}{\mathbf{C}_h^k(x, a)}}$$

and we will use the following result:

**Claim 1.** *The functions  $f_1$  and  $f_2$  are Lipschitz continuous and their constants are bounded by  $\text{Lip}(f_1) = 2C_2^g Hk / (\beta\sigma)$  and  $\text{Lip}(f_2) = \left[ \sqrt{\mathbf{v}_p(k, h)} (C_2^g k / \sigma) \beta^{-3/2} \right]$ , respectively.*

*Proof.* This is a direct consequence of Lemma 8 and the fact that, for any  $L$ -Lipschitz function  $f$ , its absolute value  $|f|$  is also  $L$ -Lipschitz.  $\square$

Now, we can finish the proof of Proposition 3 by applying Technical Lemma 6, where we identify  $F = f_1$  and  $G = f_2$ , and we use a  $(\sigma^2 / (HK))$ -covering of  $(\mathcal{X} \times \mathcal{A}, \rho)$ . We have:

$$\begin{aligned}
 \mathbb{P}[F] &\leq \mathbb{P} \left[ \exists(x, a), \exists(k, h) : \left| \frac{1}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} w_h^s(x, a) W_s \right| \geq \sqrt{\frac{\mathbf{v}_p(k, h)}{\mathbf{C}_h^k(x, a)}} + (\text{Lip}(f_1) + \text{Lip}(f_2)) \frac{\sigma^2}{HK} \right] \\
 &\leq H\mathcal{N}(\sigma^2 / (HK), \mathcal{X} \times \mathcal{A}, \rho) \delta.
 \end{aligned}$$

$\square$

### C.3. A confidence interval for $P_h f$ uniformly over Lipschitz functions $f$

In the regret analysis, we will need to control quantities like  $(\hat{P}_h^k - P_h)(\hat{f}_h^k)$  for *random* Lipschitz functions  $\hat{f}_h^k$ , which motivate us to propose a deviation inequality for  $(\hat{P}_h^k - P_h)(f)$  which holds uniformly over  $f$  in a class of Lipschitz functions. We provide such a result in Proposition 5. But we first prove an alternative to Proposition 3 which rely on a uniform Bernstein inequality for weighted sums (Lemma 3) instead of its Hoeffding counterpart (Lemma 2).

**Proposition 4.** *Let  $f : \mathcal{X} \rightarrow \mathbb{R}$  be an  $L$ -Lipschitz function such that  $\|f\|_\infty \leq 2H$ . For  $(k, h) \in [K] \times [H]$ , let*

$$F = \left\{ \exists(k, h), \exists(x, a) \in \mathcal{X} \times \mathcal{A} : \left| \hat{P}_h^k f(x, a) - P_h f(x, a) \right| \geq \frac{1}{H} \int_{\mathcal{X}} |f(y)| P_h(dy|x, a) + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(x, a)} + \theta_b(k, h) \sigma^{1+d} \right\}$$

where

$$\begin{aligned} \theta_v(k, h) &\stackrel{\text{def}}{=} 12H^2 \log(4e(k+1)/\delta) \\ \theta_b(k, h) &\stackrel{\text{def}}{=} 2 \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right) \left( 1 + \frac{\lambda_p L}{H} \right) + H^{-2} \left( \frac{12H^2 C_2^g \log(4e(k+1)/\delta)}{\beta^2} + \frac{4C_2^g H}{\beta} + \frac{\lambda_p L \sigma}{K} \right) \end{aligned}$$

and where  $d \geq 0$  is any non-negative constant<sup>9</sup>.

Then,

$$\mathbb{P}[F] \leq H\mathcal{N} \left( \frac{\sigma^{2+d}}{H^2 K}, \mathcal{X} \times \mathcal{A}, \rho \right) \delta.$$

*Proof.* To bound the probability of the failure event  $F$ , we proceed as follows:

- First, we fix a pair  $(x, a)$  and bound the probability of a failure at this pair, denoted by  $F(x, a)$ ,
- Second, we use the technical Lemma 8 to show that the upper confidence intervals, seen as a function of  $(x, a)$ , are Lipschitz continuous and, consequently, we can extend our bound to every  $(x, a)$  by using a covering argument (technical Lemma 6).

**Bounding the failure probability at a fixed  $(x, a)$**  For a fixed pair  $(x, a) \in \mathcal{X} \times \mathcal{A}$ , let

$$F(x, a) = \left\{ \exists(k, h) : \left| \hat{P}_h^k f(x, a) - P_h f(x, a) \right| \geq 2\sigma \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right) + \frac{\overline{\text{var}}_h^k(f)}{2H^2} + \frac{12H^2 \log(4e(k+1)/\delta)}{\mathbf{C}_h^k(x, a)} \right\}$$

where

$$\overline{\text{var}}_h^k(f) \stackrel{\text{def}}{=} 2H \int_{\mathcal{X}} |f(y)| P_h(dy|x, a) + 4\sigma \lambda_p L H \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right)$$

represents an upper bound on the weighted sum of conditional variances  $\left( \sum_s \mathbb{V} \left[ w_h^s(x, a) f(x_h^s) \middle| \mathcal{F}_h^{s-1} \right] \right) / (\beta + \sum_s w_h^s(x, a))$ , where  $\mathcal{F}_h^s$  is the sigma-algebra generated by  $\{(x_h^t, a_h^t) : t \leq s+1\}$ .

In the proof of Proposition 3, we showed that

$$\left| \hat{P}_h^k f(x, a) - P_h f(x, a) \right| \leq 2\sigma \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right) + \underbrace{\left| \frac{1}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} w_h^s(x, a) W_s \right|}_{(*)}$$

<sup>9</sup>The constant  $d$  will be chosen to be the covering dimension of the state-action space.

where  $W_s \stackrel{\text{def}}{=} f(x_{h+1}^s) - \int_{\mathcal{X}} f(y) dP_h(y|x_h^s, a_h^s)$ . The process  $(W_s)_{s \in [k-1]}$  a martingale difference sequence with respect to the filtration  $\mathcal{F}_h^s$  generated by  $\{(x_h^t, a_h^t) : t \leq s+1\}$ , and such that  $|W_s| \leq 4H$ . We now apply Lemma 3 to the term  $(*)$  and, to do so, we first bound the conditional variances:

$$\begin{aligned} \mathbb{E} [W_s^2 | \mathcal{F}_{s-1}] &= \mathbb{E} [f(x_{h+1}^s)^2 | \mathcal{F}_{s-1}] - \left[ \int_{\mathcal{X}} f(y) P_h(dy | x_h^s, a_h^s) \right]^2 \\ &\leq 2H \mathbb{E} [|f(x_{h+1}^s)| | \mathcal{F}_{s-1}] = 2H \int_{\mathcal{X}} |f(y)| P_h(dy | x_h^s, a_h^s) \\ &\leq 2\lambda_p L H \rho[(x, a), (x_h^s, a_h^s)] + 2H \int_{\mathcal{X}} |f(y)| P_h(dy | x, a). \end{aligned}$$

Hence,

$$\begin{aligned} &\frac{1}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} w_h^s(x, a)^2 \mathbb{E} [W_s^2 | \mathcal{F}_{s-1}] \\ &\leq \frac{1}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} w_h^s(x, a) \mathbb{E} [W_s^2 | \mathcal{F}_{s-1}] \\ &\leq 2H \int_{\mathcal{X}} |f(y)| P_h(dy | x, a) + \frac{2\lambda_p L H}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} \psi_{\sigma}((x, a), (x_h^s, a_h^s)) \rho[(x, a), (x_h^s, a_h^s)] \\ &\leq 2H \int_{\mathcal{X}} |f(y)| P_h(dy | x, a) + 4\sigma \lambda_p L H \left( 1 + \sqrt{\log(C_1^g k / \beta)} \right) \\ &= \overline{\text{var}}_h^k(f). \end{aligned}$$

Lemma 3 and an union bound over  $h \in [H]$  give us

$$\begin{aligned} \forall(x, a), \mathbb{P} [F(x, a)] &= \mathbb{P} \left[ \exists(k, h) : \left| \frac{1}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} w_h^s(x, a) W_s \right| \geq \frac{\overline{\text{var}}_h^k(f)}{2H^2} + \frac{12H^2 \log(4e(k+1)/\delta)}{\mathbf{C}_h^k(x, a)} \right] \\ &\leq \mathbb{P} \left[ \exists(k, h) : \left| \frac{1}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} w_h^s(x, a) W_s \right| \geq \sqrt{2 \log(4e(k+1)/\delta)} \frac{\overline{\text{var}}_h^k(f) + (4H)^2}{\mathbf{C}_h^k(x, a)^2} + \frac{8H \log(4e(k+1)/\delta)}{3\mathbf{C}_h^k(x, a)} \right] \\ &\leq H\delta \end{aligned}$$

for  $(k, h) \in [K] \times H$ .

**Extending the bound to every  $(x, a)$  by a covering argument** Now, we bound the probability  $\mathbb{P} [F]$  using our bound for  $\mathbb{P} [F(x, a)]$  for every  $(x, a)$ . We define

$$\begin{aligned} f_1(x, a) &\stackrel{\text{def}}{=} \left| \frac{1}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} \psi_{\sigma}((x, a), (x_h^s, a_h^s)) W_s \right|, \quad f_2(x, a) \stackrel{\text{def}}{=} \frac{12H^2 \log(4e(k+1)/\delta)}{\mathbf{C}_h^k(x, a)}, \quad \text{and} \\ f_3(x, a) &\stackrel{\text{def}}{=} \frac{1}{H} \int_{\mathcal{X}} |f(y)| P_h(dy | x, a) \end{aligned}$$

and we will use the following result:

**Claim 2.** *The functions  $f_1$ ,  $f_2$  and  $f_3$  are Lipschitz continuous and their constants are bounded by  $\text{Lip}(f_1) = 4C_2^g H k / (\beta \sigma)$ ,  $\text{Lip}(f_2) = 12H^2 \log(4e(k+1)/\delta) C_2^g k / (\sigma \beta^2)$ , and  $\text{Lip}(f_3) = \lambda_p L / H$ , respectively.*

*Proof.* This is a direct consequence of technical Lemma 8, Assumption 3 and the fact that, for any  $L$ -Lipschitz function  $f$ , its absolute value  $|f|$  is also  $L$ -Lipschitz.  $\square$



Now, we can finish the proof of Proposition 4 by applying Technical Lemma 6, where we identify  $F = f_1$  and  $G = f_2 + f_3$ , and we use a  $(\sigma^{2+d}/(H^2K))$ -covering of  $(\mathcal{X} \times \mathcal{A}, \rho)$ . We have:

$$\begin{aligned} \mathbb{P}[F] &\leq \mathbb{P}\left[\exists(k, h), \exists(x, a) : \left| \frac{1}{\mathbf{C}_h^k(x, a)} \sum_{s=1}^{k-1} \psi_\sigma((x, a), (x_h^s, a_h^s)) W_s \right| \geq \right. \\ &\quad \left. \frac{1}{H} \int_{\mathcal{X}} |f(y)| P_h(dy|x, a) + \frac{2\sigma\lambda_p L}{H} \left(1 + \sqrt{\log(C_1^g k/\beta)}\right) + \frac{12H^2 \log(1/\delta)}{\mathbf{C}_h^k(x, a)} \right. \\ &\quad \left. + (\text{Lip}(f_1) + \text{Lip}(f_2) + \text{Lip}(f_3)) \frac{\sigma^{2+d}}{H^2 K} \right] \\ &\leq H\mathcal{N}\left(\frac{\sigma^{2+d}}{H^2 K}, \mathcal{X} \times \mathcal{A}, \rho\right) \delta. \end{aligned}$$

□

**Proposition 5.** Consider the following function space:

$$\mathcal{F}_L \stackrel{\text{def}}{=} \{f : \mathcal{X} \rightarrow [0, 2H] \text{ such that } f \text{ is } L\text{-Lipschitz}\}$$

and let

$$F = \left\{ \exists f \in \mathcal{F}_L, \exists(x, a), \exists(k, h) : \left| \hat{P}_h^k f(x, a) - P_h f(x, a) \right| \geq \frac{1}{H} \int_{\mathcal{X}} |f(y)| P_h(dy|x, a) + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(x, a)} + \theta_b(k, h)\sigma^{1+d} + 2L\sigma \right\}$$

where

$$\begin{aligned} \theta_v(k, h) &\stackrel{\text{def}}{=} 12H^2 \log(4e(k+1)/\delta) \\ \theta_b(k, h) &\stackrel{\text{def}}{=} 2 \left(1 + \sqrt{\log(C_1^g k/\beta)}\right) \left(1 + \frac{\lambda_p L}{H}\right) + H^{-2} \left(\frac{12H^2 C_2^g \log(4e(k+1)/\delta)}{\beta^2} + \frac{4C_2^g H}{\beta} + \frac{\lambda_p L \sigma}{K}\right) \end{aligned}$$

Then,

$$\mathbb{P}[F] \leq H\mathcal{N}\left(\frac{\sigma^{2+d}}{H^2 K}, \mathcal{X} \times \mathcal{A}, \rho\right) \left(\frac{16H}{\sigma}\right)^{\mathcal{N}(\sigma/4, \mathcal{X}, \rho_{\mathcal{X}})} \delta.$$

*Proof.* Let  $\mathcal{C}_{\mathcal{F}_L}(\sigma)$  be a covering of  $\mathcal{F}_L$  with respect to the norm  $\|\cdot\|_\infty$  such that:

$$\forall V \in \mathcal{F}_L, \exists \tilde{V} \in \mathcal{C}_{\mathcal{F}_L}(\sigma) \text{ such that } \|V - \tilde{V}\|_\infty \leq L\sigma$$

By Lemma 5, the covering number of  $\mathcal{F}_L$  is bounded as follows

$$\mathcal{N}(\sigma, \mathcal{F}_L, \|\cdot\|_\infty) \leq \left(\frac{16H}{\sigma}\right)^{\mathcal{N}(\sigma/4, \mathcal{X}, \rho_{\mathcal{X}})}$$

Now, let's bound the probability of  $F$ . Consider the function:

$$\begin{aligned} G : \mathcal{F}_L &\rightarrow \mathbb{R} \\ G : f &\mapsto \left| \hat{P}_h^k f(x, a) - P_h f(x, a) \right| \end{aligned}$$

We can check that, for all  $f_1, f_2 \in \mathcal{F}_L$ , we have  $|G(f_1) - G(f_2)| \leq 2\|f_1 - f_2\|_\infty$ . Hence, a union bound over  $\mathcal{C}_{\mathcal{F}_L}(\sigma)$  and over  $[K] \times [H]$  and Proposition 4 allow us to conclude.

□

#### C.4. Good event

**Theorem 3.** Consider the following event, on which  $P_h V_{h+1}(s, a)$  and  $r_h(x, a)$  are within their confidence intervals for all  $h, a, x$ , and the deviations of  $\hat{P}_h^k f(x, a)$  from  $P_h^k f(x, a)$  are bounded for any  $2L_1$ -Lipschitz function  $f$  :

$$\begin{aligned} \mathcal{G} \stackrel{\text{def}}{=} & \left\{ \forall(k, h), \forall(x, a), \left| \hat{P}_h^k V_{h+1}^*(x, a) - P_h V_{h+1}^*(x, a) \right| \leq \sqrt{\frac{\mathbf{v}_p(k, h)}{\mathbf{C}_h^k(x, a)}} + \frac{\beta}{\mathbf{C}_h^k(x, a)} + \mathbf{b}_p(k, h)\sigma \right\} \\ & \cap \left\{ \forall(k, h), \forall(x, a), \left| \hat{r}_h^k(x, a) - r_h(x, a) \right| \leq \sqrt{\frac{\mathbf{v}_r(k, h)}{\mathbf{C}_h^k(x, a)}} + \frac{\beta}{\mathbf{C}_h^k(x, a)} + \mathbf{b}_r(k, h)\sigma \right\} \\ & \cap \left\{ \forall f \in \mathcal{F}_{2L_1}, \forall(k, h), \forall(x, a), \left| \hat{P}_h^k f(x, a) - P_h f(x, a) \right| \leq \frac{1}{H} \int_{\mathcal{X}} |f(y)| P_h(dy|x, a) + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(x, a)} + \theta_b(k, h)\sigma^{1+d} + 4L_1\sigma \right\} \end{aligned}$$

where  $(k, h) \in [K] \times [H]$  and  $d \geq 0$ . For any  $\delta > 0$ , let

$$\delta_1 \stackrel{\text{def}}{=} \frac{\delta}{6HN \left( \frac{\sigma^2}{HK}, \mathcal{X} \times \mathcal{A}, \rho \right)} \quad \text{and} \quad \delta_2 \stackrel{\text{def}}{=} \frac{\delta}{6HN \left( \frac{\sigma^{2+d}}{H^2K}, \mathcal{X} \times \mathcal{A}, \rho \right) \left( \frac{16H}{\sigma} \right)^{\mathcal{N}(\sigma/8, \mathcal{X}, \rho_{\mathcal{X}})}}.$$

Then, if

$$\begin{aligned} \mathbf{v}_p(k, h) &= 2H^2 \log \left( \frac{\sqrt{1+k/\beta}}{\delta_1} \right), \\ \mathbf{b}_p(k, h) &= 2\lambda_p L_1 \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right) + \frac{2C_2^g}{\beta^{3/2}} \sqrt{2 \log \left( \frac{\sqrt{1+k/\beta}}{\delta_1} \right)} + \frac{2C_2^g}{\beta}, \\ \mathbf{v}_r(k, h) &= 2 \log \left( \frac{\sqrt{1+k/\beta}}{\delta_1} \right), \\ \mathbf{b}_r(k, h) &= 2\lambda_r \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right) + \frac{2C_2^g}{\beta^{3/2}} \sqrt{2 \log \left( \frac{\sqrt{1+k/\beta}}{\delta_1} \right)} + \frac{2C_2^g}{\beta}, \\ \theta_v(k, h) &\stackrel{\text{def}}{=} 12H^2 \log(4e(k+1)/\delta_2), \\ \theta_b(k, h) &\stackrel{\text{def}}{=} 2 \left( 1 + \sqrt{\log(C_1^g k/\beta)} \right) \left( 1 + \frac{\lambda_p L_1}{H} \right) + H^{-2} \left( \frac{12H^2 C_2^g \log(4e(k+1)/\delta_2)}{\beta^2} + \frac{4C_2^g H}{\beta} + \frac{\lambda_p L_1 \sigma}{K} \right) \end{aligned}$$

we have

$$\mathbb{P}[\mathcal{G}] \geq 1 - \delta/2.$$

*Proof.* The value functions  $V_h^k$ , computed by `optimisticQ`, are  $L_h$ -Lipschitz. Hence, they are also  $L_1$ -Lipschitz, since  $L_1 \geq L_2 \geq \dots L_H$ . The result follows from propositions 2, 3 and 5.  $\square$

#### D. Optimism and regret bound

**Proposition 6 (Optimism).** In the event  $\mathcal{G}$ , whose probability is greater than  $1 - \delta/2$ , we have:

$$\forall(x, a), \quad Q_h^k(x, a) \geq Q_h^*(x, a)$$

*Proof.* We proceed by induction.

**Initialization** When  $h = H + 1$ , we have  $Q_h^k(x, a) = Q_h^*(x, a) = 0$  for all  $(x, a)$ .

**Induction hypothesis** Assume that  $Q_{h+1}^k(x, a) \geq Q_{h+1}^*(x, a)$  for all  $(x, a)$ .

*Induction step* The induction hypothesis implies that  $V_{h+1}^k(x) \geq V_{h+1}^*(x)$  for all  $x$ . Hence, for all  $(x, a)$ , we have

$$\tilde{Q}_h^k(x, a) - Q_h^*(x, a) = \underbrace{(\hat{r}_h^k(x, a) - r_h(x, a)) + (\hat{P}_h^k - P_h)V_{h+1}^*(x, a) + \mathbf{B}_h^k(x, a)}_{\geq 0 \text{ in } \mathcal{G}} + \underbrace{\hat{P}_h^k(V_{h+1}^k - V_{h+1}^*)(x, a)}_{\geq 0 \text{ by induction hypothesis}} \geq 0.$$

In particular  $\tilde{Q}_h^k(x_h^s, a_h^s) - Q_h^*(x_h^s, a_h^s) \geq 0$  for all  $s \in [k-1]$ . This implies that

$$\tilde{Q}_h^k(x_h^s, a_h^s) + L_h \rho[(x, a), (x_h^s, a_h^s)] \geq Q_h^*(x_h^s, a_h^s) + L_h \rho[(x, a), (x_h^s, a_h^s)] \geq Q_h^*(x, a)$$

for all  $s \in [k-1]$ , since  $Q_h^*$  is  $L_h$ -Lipschitz. Finally, we obtain

$$\forall(x, a), \quad Q_h^k(x, a) = \min_{s \in [k-1]} [\tilde{Q}_h^k(x_h^s, a_h^s) + L_h \rho[(x, a), (x_h^s, a_h^s)]] \geq Q_h^*(x, a).$$

□

**Corollary 2.** Let  $\delta_h^k \stackrel{\text{def}}{=} V_h^k(x_h^k) - V_h^{\pi_k}(x_h^k)$ . Then, with probability at least  $1 - \delta/2$ ,  $\mathcal{R}(K) \leq \sum_{k=1}^K \delta_1^k$ .

*Proof.* Combining the definition of the regret with Proposition 6 easily yields, on the event  $\mathcal{G}$ ,

$$\begin{aligned} \mathcal{R}(K) &= \sum_{k=1}^K (V_1^*(x_1^k) - V_1^{\pi_k}(x_1^k)) = \sum_{k=1}^K \left( \max_a Q_1^*(x_1^k, a) - V_1^{\pi_k}(x_1^k) \right) \\ &\leq \sum_{k=1}^K \left( \min \left[ H - h + 1, \max_a Q_1^k(x_1^k, a) \right] - V_1^{\pi_k}(x_1^k) \right) = \sum_{k=1}^K (V_1^k(x_1^k, a) - V_1^{\pi_k}(x_1^k)), \end{aligned}$$

and the conclusions follows from the fact that  $\mathbb{P}[\mathcal{G}] \geq 1 - \delta/2$ . □

**Definition 2.** For any  $(k, h)$ , we define  $(\tilde{x}_h^k, \tilde{a}_h^k)$  as state-action pair in the past data  $\mathcal{D}_h$  that is the closest to  $(x_h^k, a_h^k)$ , that is

$$(\tilde{x}_h^k, \tilde{a}_h^k) \stackrel{\text{def}}{=} \underset{(x_h^s, a_h^s): s < k}{\operatorname{argmin}} \rho[(x_h^k, a_h^k), (x_h^s, a_h^s)].$$

**Proposition 7.** With probability  $1 - \delta$ , the regret of **Kernel-UCBVI** is bounded as follows

$$\begin{aligned} \mathcal{R}(K) &\leq H^2 \mathcal{N}(\sigma/4, \mathcal{X} \times \mathcal{A}, \rho) \\ &\quad + (L_1(1 + 2\lambda_p) + \theta_b(K, H)\sigma^d + 4L_1)KH\sigma \\ &\quad + \sum_{k=1}^K \sum_{h=1}^H \left(1 + \frac{1}{H}\right)^{H-h} \xi_{h+1}^k \\ &\quad + e \sum_{k=1}^K \sum_{h=1}^H \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} \mathbb{I} \{ \rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2 \} \\ &\quad + 2e \sum_{k=1}^K \sum_{h=1}^H \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) \mathbb{I} \{ \rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2 \}, \end{aligned}$$

where  $\xi_{h+1}^k$  is a martingale difference sequence with respect to the filtration  $\mathcal{F}_h^{k-1}$  generated by  $\{(x_h^s, a_h^s) : s < k, h \in [H]\} \cup \{(x_h^k, a_h^k)\}$  such that  $|\xi_{h+1}^k| \leq 2H$ .

*Proof.* On  $\mathcal{G}$ , we have

$$\begin{aligned}
 \delta_h^k &= V_h^k(x_h^k) - V_h^{\pi_k}(x_h^k) \\
 &\leq Q_h^k(x_h^k, a_h^k) - Q_h^{\pi_k}(x_h^k, a_h^k) \\
 &\leq Q_h^k(\tilde{x}_h^k, \tilde{a}_h^k) - Q_h^{\pi_k}(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_1\rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)], \quad \text{since } Q_h^k \text{ and } Q_h^{\pi_k} \text{ are } L_1\text{-Lipschitz} \\
 &\leq \tilde{Q}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) - Q_h^{\pi_k}(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_1\rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)], \quad \text{since } Q_h^k(\tilde{x}_h^k, \tilde{a}_h^k) \leq \tilde{Q}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) \text{ by definition of } Q_h^k \\
 &= \hat{r}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) - r_h(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_1\rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] + \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + \hat{P}_h^k V_{h+1}^k(\tilde{x}_h^k, \tilde{a}_h^k) - P_h V_{h+1}^{\pi_k}(\tilde{x}_h^k, \tilde{a}_h^k) \\
 &\leq \mathbf{r}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_1\rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] + \underbrace{\hat{P}_h^k V_{h+1}^k(\tilde{x}_h^k, \tilde{a}_h^k) - P_h V_{h+1}^{\pi_k}(\tilde{x}_h^k, \tilde{a}_h^k)}_{(\mathbf{A})}.
 \end{aligned}$$

The term  $(\mathbf{A})$  is bounded as follows:

$$(\mathbf{A}) = \underbrace{[\hat{P}_h^k - P_h] V_{h+1}^*(\tilde{x}_h^k, \tilde{a}_h^k)}_{(\mathbf{B})} + \underbrace{P_h (V_{h+1}^k - V_{h+1}^{\pi_k})(\tilde{x}_h^k, \tilde{a}_h^k)}_{(\mathbf{C})} + \underbrace{[\hat{P}_h^k - P_h] (V_{h+1}^k - V_{h+1}^*)(\tilde{x}_h^k, \tilde{a}_h^k)}_{(\mathbf{D})},$$

where

$(\mathbf{B}) \leq \mathbf{P}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)$ , by definition of the event  $\mathcal{G}$ ,

$(\mathbf{C}) \leq P_h (V_{h+1}^k - V_{h+1}^{\pi_k})(\tilde{x}_h^k, \tilde{a}_h^k) + 2\lambda_p L_1 \rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)]$ , by Assumption 3 and the fact that  $(V_{h+1}^k - V_{h+1}^{\pi_k})$  is  $2L_1$ -Lipschitz.

Now we give an upper bound for the term  $(\mathbf{D})$ . On  $\mathcal{G}$ , we have  $(V_{h+1}^k - V_{h+1}^*) \geq 0$  and, consequently,

$$\begin{aligned}
 &[\hat{P}_h^k - P_h] (V_{h+1}^k - V_{h+1}^*)(\tilde{x}_h^k, \tilde{a}_h^k) \\
 &\leq \frac{1}{H} \int_{\mathcal{X}} |(V_{h+1}^k - V_{h+1}^*)(y)| P_h(\mathrm{d}y | \tilde{x}_h^k, \tilde{a}_h^k) + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \theta_b(k, h) \sigma^{1+d} + 4L_1 \sigma \\
 &= \frac{1}{H} P_h (V_{h+1}^k - V_{h+1}^*)(\tilde{x}_h^k, \tilde{a}_h^k) + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \theta_b(k, h) \sigma^{1+d} + 4L_1 \sigma \\
 &\leq \frac{1}{H} P_h (V_{h+1}^k - V_{h+1}^*)(x_h^k, a_h^k) + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \theta_b(k, h) \sigma^{1+d} + 4L_1 \sigma + 2\lambda_p L_1 \rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)].
 \end{aligned}$$

Using the upper bound for  $(\mathbf{A})$ , and that  $V_{h+1}^* \geq V_{h+1}^{\pi_k}$ , we obtain

$$\begin{aligned}
 \delta_h^k &\leq 2 \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_1(1 + 2\lambda_p) \rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \theta_b(k, h) \sigma^{1+d} + 4L_1 \sigma \\
 &\quad + \frac{1}{H} P_h (V_{h+1}^k - V_{h+1}^*)(x_h^k, a_h^k) + P_h (V_{h+1}^k - V_{h+1}^{\pi_k})(x_h^k, a_h^k) \\
 &\leq \left(1 + \frac{1}{H}\right) P_h (V_{h+1}^k - V_{h+1}^{\pi_k})(x_h^k, a_h^k) + 2 \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) \\
 &\quad + 2L_1(1 + 2\lambda_p) \rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \theta_b(k, h) \sigma^{1+d} + 4L_1 \sigma \\
 &= \left(1 + \frac{1}{H}\right) (\delta_{h+1}^k + \xi_{h+1}^k) + 2 \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_1(1 + 2\lambda_p) \rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \theta_b(k, h) \sigma^{1+d} + 4L_1 \sigma,
 \end{aligned}$$

where  $\xi_{h+1}^k \stackrel{\text{def}}{=} P_h(V_{h+1}^k - V_{h+1}^{\pi_k})(x_h^k, a_h^k) - \delta_{h+1}^k$  is a martingale difference sequence with respect to the filtration  $\mathcal{F}_h^{k-1}$  generated by  $\{(x_h^s, a_h^s) : s < k, h \in [H]\} \cup \{(x_h^k, a_h^k)\}$ , that is,  $\mathbb{E}[\xi_{h+1}^k | \mathcal{F}_h^{k-1}] = 0$ .



This gives us, on  $\mathcal{G}$ ,

$$\begin{aligned}
 \mathcal{R}(K) &\leq \sum_{k=1}^K \delta_1^k \\
 &\leq \sum_{k=1}^K \sum_{h=1}^H \left(1 + \frac{1}{H}\right)^{H-h} \left( 2 \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_1(1 + 2\lambda_p)\rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \theta_b(k, h)\sigma^{1+d} + 4L_1\sigma \right) \\
 &\quad + \sum_{k=1}^K \sum_{h=1}^H \left(1 + \frac{1}{H}\right)^{H-h} \xi_{h+1}^k. \tag{6}
 \end{aligned}$$

Consider the definition below.

**Definition 3.** Let  $\mathcal{C}_\sigma$  be a  $\sigma/4$ -covering of  $(\mathcal{X} \times \mathcal{A}, \rho)$ . We write  $\mathcal{C}_\sigma \stackrel{\text{def}}{=} \{(x_j, a_j), j \in [|\mathcal{C}_\sigma|]\}$ . For each  $(x_j, a_j) \in \mathcal{C}_\sigma$ , we define the set  $B_j \subset \mathcal{X} \times \mathcal{A}$  as the set of state-action pairs whose nearest neighbor in  $\mathcal{C}_\sigma$  is  $(x_j, a_j)$ , with ties broken arbitrarily, such that  $\{B_j\}_{j \in [|\mathcal{C}_\sigma|]}$  form a partition of  $\mathcal{X} \times \mathcal{A}$ . For each  $(k, h)$ , we define  $B_h^k$  as the unique set in this partition such that  $(x_h^k, a_h^k) \in B_h^k$ . We say that  $B_h^k$  has been visited before  $(k, h)$  if there exists  $s \in [k-1]$  such that  $B_h^s = B_h^k$ .

Now, for each  $h$ , we distinguish two cases. In the first case, the set  $B_h^k$  has been visited before, and, in the second case, the set  $B_h^k$  has been visited for the first time at  $(k, h)$ .

In the first case, we have  $\rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2$ . In the second case, we upper bound  $\delta_1^k$  by  $H$ , and we note that this case can happen at most  $H\mathcal{N}(\sigma/4, \mathcal{X} \times \mathcal{A}, \rho)$  times.

Hence,

$$\begin{aligned}
 \sum_{k=1}^K \delta_1^k &\leq eH^2\mathcal{N}(\sigma/4, \mathcal{X} \times \mathcal{A}, \rho) \\
 &\quad + e(2L_1(1 + 2\lambda_p) + \theta_b(K, H)\sigma^d + 4L_1)KH\sigma \\
 &\quad + \sum_{k=1}^K \sum_{h=1}^H \left(1 + \frac{1}{H}\right)^{H-h} \xi_{h+1}^k \\
 &\quad + e \sum_{k=1}^K \sum_{h=1}^H \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} \mathbb{I}\{\rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2\} \\
 &\quad + 2e \sum_{k=1}^K \sum_{h=1}^H \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) \mathbb{I}\{\rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2\}
 \end{aligned}$$

where we use the fact that  $(1 + 1/H)^{H-h} \leq (1 + 1/H)^H \leq e$ .

□

**Proposition 8.** We have

$$\sum_{k=1}^K \sum_{h=1}^H \frac{1}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} \mathbb{I}\{\rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2\} \leq \beta^{-1}H|\mathcal{C}_\sigma| + \frac{H|\mathcal{C}_\sigma|}{g(1)} \log \left(1 + \frac{1 + g(1)\beta^{-1}K}{|\mathcal{C}_\sigma|}\right)$$

and

$$\sum_{k=1}^K \sum_{h=1}^H \frac{1}{\sqrt{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)}} \mathbb{I}\{\rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2\} \leq \left(\beta^{-1} + \frac{2}{g(1)}\right) \beta^{1/2}H|\mathcal{C}_\sigma| + \frac{2H}{g(1)} \sqrt{|\mathcal{C}_\sigma|g(1)K}. \tag{7}$$

*Proof.* First, we relate the generalized counts  $\mathbf{C}_h^k$  to the number of visits to the sets of the partition  $\{B_j\}_{j \in [|\mathcal{C}_\sigma|]}$  (see Definition 3).

**Relating the generalized counts to number of visits to the sets in the partition** Let  $B_j$  be a set in the partition of  $\mathcal{X} \times \mathcal{A}$  introduced in Definition 3. If  $(x_h^k, a_h^k) \in B_j$  and  $\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2$ , we have

$$\begin{aligned} \mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) &= \beta + \sum_{s=1}^{k-1} \psi_\sigma((\tilde{x}_h^k, \tilde{a}_h^k), (x_h^s, a_h^s)) \\ &= \beta + \sum_{s=1}^{k-1} g\left(\frac{\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^s, a_h^s)]}{\sigma}\right) \\ &\geq \beta + \sum_{s=1}^{k-1} g\left(\frac{\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^s, a_h^s)]}{\sigma}\right) \mathbb{I}\{(x_h^s, a_h^s) \in B_j\} \\ &\geq \beta + g(1) \sum_{s=1}^{k-1} \mathbb{I}\{(x_h^s, a_h^s) \in B_j\} \end{aligned}$$

since, if  $(x_h^s, a_h^s) \in B_j$ , we have  $\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^s, a_h^s)] \leq \sigma$  and we use the fact that  $g$  is non-increasing.

For each  $j$ , let

$$\mathbf{N}_h^k(B_j) \stackrel{\text{def}}{=} \sum_{s=1}^{k-1} \mathbb{I}\{(x_h^s, a_h^s) \in B_j\}$$

be the number of visits to the set  $B_j$  at step  $h$  before episode  $k$ . We proved that, if  $(x_h^k, a_h^k) \in B_j$  and  $\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2$ , then

$$\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) \geq \beta + g(1)\mathbf{N}_h^k(B_j) = \beta(1 + g(1)\beta^{-1}\mathbf{N}_h^k(B_j))$$

### Bounding the sum of the first order terms

$$\begin{aligned} &\sum_{k=1}^K \sum_{h=1}^H \sqrt{\frac{1}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)}} \mathbb{I}\{\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2\} \\ &= \sum_{k=1}^K \sum_{h=1}^H \sum_{j=1}^{|\mathcal{C}_\sigma|} \sqrt{\frac{1}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)}} \mathbb{I}\{\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2\} \mathbb{I}\{(x_h^k, a_h^k) \in B_j\} \\ &\leq \beta^{-1/2} \sum_{k=1}^K \sum_{h=1}^H \sum_{j=1}^{|\mathcal{C}_\sigma|} \frac{1}{\sqrt{1 + g(1)\beta^{-1}\mathbf{N}_h^k(B_j)}} \mathbb{I}\{\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2\} \mathbb{I}\{(x_h^k, a_h^k) \in B_j\} \\ &\leq \beta^{-1/2} \sum_{h=1}^H \sum_{j=1}^{|\mathcal{C}_\sigma|} \sum_{k=1}^K \frac{\mathbb{I}\{(x_h^k, a_h^k) \in B_j\}}{\sqrt{1 + g(1)\beta^{-1}\mathbf{N}_h^k(B_j)}} \leq \beta^{-1/2} \sum_{h=1}^H \sum_{j=1}^{|\mathcal{C}_\sigma|} \left(1 + \int_0^{\mathbf{N}_h^{K+1}(B_j)} \frac{dz}{\sqrt{1 + g(1)\beta^{-1}z}}\right) \quad \text{by Lemma 9} \\ &\leq \beta^{-1/2} H |\mathcal{C}_\sigma| + \frac{2\beta^{1/2}}{g(1)} \sum_{h=1}^H \sum_{j=1}^{|\mathcal{C}_\sigma|} \sqrt{1 + g(1)\beta^{-1}\mathbf{N}_h^{K+1}(B_j)} \\ &\leq \beta^{-1/2} H |\mathcal{C}_\sigma| + \frac{2\beta^{1/2}}{g(1)} \sum_{h=1}^H \sqrt{|\mathcal{C}_\sigma|} \sqrt{|\mathcal{C}_\sigma| + g(1)\beta^{-1}K} \quad \text{by Cauchy-Schwarz inequality} \\ &\leq H \left( \beta^{-1/2} + \frac{2\beta^{1/2}}{g(1)} \right) |\mathcal{C}_\sigma| + \frac{2H}{g(1)} \sqrt{g(1) |\mathcal{C}_\sigma| K}. \end{aligned}$$

## Bounding the sum of the second order terms

$$\begin{aligned}
 & \sum_{k=1}^K \sum_{h=1}^H \frac{1}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} \mathbb{I} \{ \rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2 \} \\
 &= \sum_{k=1}^K \sum_{h=1}^H \sum_{j=1}^{|\mathcal{C}_\sigma|} \frac{1}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} \mathbb{I} \{ \rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2 \} \mathbb{I} \{ (x_h^k, a_h^k) \in B_j \} \\
 &\leq \beta^{-1} \sum_{k=1}^K \sum_{h=1}^H \sum_{j=1}^{|\mathcal{C}_\sigma|} \frac{1}{1 + g(1)\beta^{-1}\mathbf{N}_h^k(B_j)} \mathbb{I} \{ \rho [(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \leq \sigma/2 \} \mathbb{I} \{ (x_h^k, a_h^k) \in B_j \} \\
 &\leq \beta^{-1} \sum_{h=1}^H \sum_{j=1}^{|\mathcal{C}_\sigma|} \sum_{k=1}^K \frac{\mathbb{I} \{ (x_h^k, a_h^k) \in B_j \}}{1 + g(1)\beta^{-1}\mathbf{N}_h^k(B_j)} \leq \beta^{-1} \sum_{h=1}^H \sum_{j=1}^{|\mathcal{C}_\sigma|} \left( 1 + \int_0^{\mathbf{N}_h^{K+1}(B_j)} \frac{dz}{1 + g(1)\beta^{-1}z} \right) \quad \text{by Lemma 9} \\
 &\leq \beta^{-1} H |\mathcal{C}_\sigma| + \frac{1}{g(1)} \sum_{h=1}^H \sum_{j=1}^{|\mathcal{C}_\sigma|} \log (1 + g(1)\beta^{-1}\mathbf{N}_h^{K+1}(B_j)) \\
 &\leq \beta^{-1} H |\mathcal{C}_\sigma| + \frac{1}{g(1)} \sum_{h=1}^H |\mathcal{C}_\sigma| \log \left( \frac{\sum_{j=1}^{|\mathcal{C}_\sigma|} (1 + g(1)\beta^{-1}\mathbf{N}_h^{K+1}(B_j))}{|\mathcal{C}_\sigma|} \right) \quad \text{by Jensen's inequality} \\
 &\leq \beta^{-1} H |\mathcal{C}_\sigma| + \frac{1}{g(1)} H |\mathcal{C}_\sigma| \log \left( 1 + \frac{1 + g(1)\beta^{-1}K}{|\mathcal{C}_\sigma|} \right).
 \end{aligned}$$

□

**Theorem 4.** With probability at least  $1 - \delta$ , the regret of *Kernel-UCBVI* is bounded as

$$\begin{aligned}
 \mathcal{R}(K) &\leq H^2 |\mathcal{C}_\sigma| \\
 &\quad + (L_1(1 + 2\lambda_p) + \theta_b(K, H)\sigma^d + 4L_1) KH\sigma \\
 &\quad + (\sqrt{8e^2 H^2 \log(1/\delta)}) \sqrt{KH} \\
 &\quad + (\theta_v(K, H) + 4\beta) eH |\mathcal{C}_\sigma| \left( \frac{1}{\beta} + \frac{1}{g(1)} \log \left( 1 + \frac{1 + g(1)\beta^{-1}K}{|\mathcal{C}_\sigma|} \right) \right) \\
 &\quad + 2e (\mathbf{b}_p(K, H) + \mathbf{b}_r(K, H)) KH\sigma \\
 &\quad + 2eH \left( \sqrt{\mathbf{v}_p(K, H)} + \sqrt{\mathbf{v}_r(K, H)} \right) \left( \left( \frac{1}{\beta^{1/2}} + \frac{2\beta^{1/2}}{g(1)} \right) |\mathcal{C}_\sigma| + \frac{2}{g(1)} \sqrt{|\mathcal{C}_\sigma| g(1)K} \right)
 \end{aligned}$$

where  $\mathcal{C}_\sigma$  is a  $\sigma/4$ -covering of  $(\mathcal{X} \times \mathcal{A}, \rho)$ .

Hence,

$$\mathcal{R}(K) = \tilde{\mathcal{O}} \left( H^2 \sqrt{|\mathcal{C}_\sigma| K} + L_1 H K \sigma + H^3 |\mathcal{C}_\sigma|^2 + H^2 |\mathcal{C}_\sigma| \right)$$

*Proof.* We have

$$\mathbf{B}_h^k(x, a) = \left( \sqrt{\frac{\mathbf{v}_p(k, h)}{\mathbf{C}_h^k(x, a)}} + \frac{\beta}{\mathbf{C}_h^k(x, a)} + \mathbf{b}_p(k, h)\sigma \right) + \left( \sqrt{\frac{\mathbf{v}_r(k, h)}{\mathbf{C}_h^k(x, a)}} + \frac{\beta}{\mathbf{C}_h^k(x, a)} + \mathbf{b}_r(k, h)\sigma \right).$$

The result follows from propositions 7 and 8 and from Hoeffding-Azuma's inequality, which ensures that the term  $\sum_{k=1}^K \sum_{h=1}^H (1 + 1/H)^{H-h} \xi_{h+1}^k$  is bounded by  $(\sqrt{8e^2 H^2 \log(2/\delta)}) \sqrt{KH}$  with probability at least  $1 - \delta/2$ .

Also, we use the fact that

$$\begin{aligned} \mathbf{v}_p(k, h) &= \tilde{\mathcal{O}}(H^2), \quad \mathbf{v}_r(k, h) = \tilde{\mathcal{O}}(1), \quad \theta_v(k, h) = \tilde{\mathcal{O}}(H^2 |\mathcal{C}_\sigma|) \\ \mathbf{b}_p(k, h) &= \tilde{\mathcal{O}}(L_1), \quad \mathbf{b}_r(k, h) = \tilde{\mathcal{O}}(L_1), \quad \theta_b(k, h) = \tilde{\mathcal{O}}(L_1) \end{aligned}$$

which comes from their definition in Theorem 3. We omit logarithmic terms and all constants except for terms involving  $K, H, L_1$  and  $|\mathcal{C}_\sigma|$ .  $\square$

## E. Remarks & Regret Bounds in Different Settings

### E.1. Improved regret for Stationary MDPs

The regret bound of **Kernel-UCBVI** can be improved if the MDP is stationary, i.e.,  $P_1 = \dots = P_H$  and  $r_1 = \dots = r_H$ . Let  $t = kh$  be the *total time* at step  $h$  of episode  $k$ , and now we index by  $t$  all the quantities that were indexed by  $(k, h)$ , e.g.,  $w_t(x, a) = w_h^k(x, a)$ . In the stationary case, the rewards and transitions estimates become

$$\hat{P}_t(y|x, a) \stackrel{\text{def}}{=} \frac{1}{\mathbf{C}_t(x, a)} \sum_{t'=1}^{t-1} w_{t'}(x, a) \delta_{x_{t'+1}}(y) \quad \text{and} \quad \hat{r}_t(x, a) \stackrel{\text{def}}{=} \frac{1}{\mathbf{C}_t(x, a)} \sum_{t'=1}^{t-1} w_{t'}(x, a) r_{t'},$$

respectively, where we redefine the generalized counts as

$$\mathbf{C}_t(x, a) \stackrel{\text{def}}{=} \beta + \sum_{t'=1}^{t-1} w_{t'}(x, a).$$

The proofs of the concentration results and of the regret bound remain valid, in particular Proposition 7, up to minor changes in the constants  $\mathbf{v}_p(k, h)$ ,  $\mathbf{b}_p(k, h)$ ,  $\mathbf{v}_r(k, h)$ ,  $\mathbf{b}_r(k, h)$ ,  $\theta_v(k, h)$  and  $\theta_b(k, h)$ . However, the bounds presented in Proposition 8 can be improved to obtain a better regret bound in terms of the horizon  $H$ . Consider the sets  $B_j$  introduced in Definition 3 and let

$$\mathbf{N}_t(B_j) \stackrel{\text{def}}{=} \sum_{t'=1}^{t-1} \mathbb{I}\{(x_{t'}, a_{t'}) \in B_j\}.$$

As we did in the proof Proposition 8, we can show that  $\mathbf{C}_t(\tilde{x}_t, \tilde{a}_t) \geq \beta + g(1)\mathbf{N}_t(B_j)$  if  $(x_t, a_t) \in B_j$  and  $\rho[(\tilde{x}_t, \tilde{a}_t), (x_t, a_t)] \leq \sigma/2$ . The sum of the first order terms  $\sum_t 1/\sqrt{\mathbf{C}_t(\tilde{x}_t, \tilde{a}_t)}$  is now bounded as

$$\begin{aligned} & \sum_{t=1}^{KH} \sqrt{\frac{1}{\mathbf{C}_t(\tilde{x}_t, \tilde{a}_t)}} \mathbb{I}\{\rho[(\tilde{x}_t, \tilde{a}_t), (x_t, a_t)] \leq \sigma/2\} \\ & \leq \beta^{-1} \sum_{j=1}^{|\mathcal{C}_\sigma|} \sum_{t=1}^{KH} \frac{\mathbb{I}\{(x_t, a_t) \in B_j\}}{\sqrt{1 + g(1)\beta^{-1}\mathbf{N}_t(B_j)}} \leq \beta^{-1} \sum_{j=1}^{|\mathcal{C}_\sigma|} \left( 1 + \int_0^{\mathbf{N}_{KH+1}(B_j)} \frac{dz}{\sqrt{1 + g(1)\beta^{-1}z}} \right) \quad \text{by Lemma 9} \\ & \leq \beta^{-1} |\mathcal{C}_\sigma| + \frac{2}{g(1)} \sum_{j=1}^{|\mathcal{C}_\sigma|} \sqrt{1 + g(1)\beta^{-1}\mathbf{N}_{KH+1}(B_j)} \\ & \leq \beta^{-1} |\mathcal{C}_\sigma| + \frac{2}{g(1)} \sqrt{|\mathcal{C}_\sigma|} \sqrt{|\mathcal{C}_\sigma| + g(1)\beta^{-1}KH} \quad \text{by Cauchy-Schwarz inequality} \\ & \leq \left( \beta^{-1} + \frac{2}{g(1)} \right) |\mathcal{C}_\sigma| + \frac{2}{g(1)} \sqrt{g(1)\beta^{-1}|\mathcal{C}_\sigma|HK} \\ & = \mathcal{O}\left(|\mathcal{C}_\sigma| + \sqrt{|\mathcal{C}_\sigma|HK}\right). \end{aligned}$$

When compared to the non-stationary case, where the corresponding sum is bounded by  $\mathcal{O}\left(H|\mathcal{C}_\sigma| + H\sqrt{|\mathcal{C}_\sigma|K}\right)$ , we gain a factor of  $\sqrt{H}$  in the term multiplying  $\sqrt{K}$  and a factor of  $H$  in the term multiplying  $|\mathcal{C}_\sigma|$ .



Similarly, the sum of the second order terms  $\sum_t 1/\mathbf{C}_t(\tilde{x}_t, \tilde{a}_t)$  is now bounded as

$$\begin{aligned} \sum_{t=1}^{KH} \frac{1}{\mathbf{C}_t(\tilde{x}_t, \tilde{a}_t)} \mathbb{I} \{ \rho[(\tilde{x}_t, \tilde{a}_t), (x_t, a_t)] \leq \sigma/2 \} &\leq \beta^{-1} |\mathcal{C}_\sigma| + \frac{1}{g(1)} |\mathcal{C}_\sigma| \log \left( 1 + \frac{1 + g(1)\beta^{-1}KH}{|\mathcal{C}_\sigma|} \right) \\ &= \tilde{\mathcal{O}}(|\mathcal{C}_\sigma|). \end{aligned}$$

In the non-stationary case, the corresponding sum is bounded by  $\tilde{\mathcal{O}}(H|\mathcal{C}_\sigma|)$ , thus we gain a factor of  $H$ .

Hence, if the MDP is stationary, we obtain a regret bound of

$$\mathcal{R}_{\text{stationary}}(K) = \tilde{\mathcal{O}} \left( H^{3/2} \sqrt{|\mathcal{C}_\sigma| K} + L_1 H K \sigma + H^2 |\mathcal{C}_\sigma|^2 \right)$$

which is  $\tilde{\mathcal{O}} \left( H^2 K^{\max(\frac{1}{2}, \frac{2d}{2d+1})} \right)$  by taking  $\sigma = (1/K)^{1/(2d+1)}$ .

#### E.1.1. IMPORTANT REMARK

Computationally, in order to achieve this improved regret for **Kernel-UCBVI**, every time a new transition and a new reward are observed at a step  $h$ , the estimates  $\hat{P}_t(y|x, a)$  and  $\hat{r}_t(x, a)$  need to be updated, and the optimistic  $Q$ -functions need to be recomputed through backward induction, which increases the computational complexity by a factor of  $H$ .

The UCBVI-CH algorithm of [Azar et al. \(2017\)](#) in the tabular setting for stationary MDPs also suffers from this problem. If the optimistic  $Q$ -function is not recomputed at every step  $h$ , its regret is  $\tilde{\mathcal{O}} \left( H^{3/2} \sqrt{XAK} + H^3 X^2 A \right)$  and not  $\tilde{\mathcal{O}} \left( H^{3/2} \sqrt{XAK} + H^2 X^2 A \right)$ , where  $X$  is the number of states, as claimed in their paper. To see why, let's analyze its second order term, which is  $\mathcal{O} \left( H^2 X \sum_{k,h} 1/N_k(x_h^k, a_h^k) \right)^{10}$ , where  $N_k(x, a)$  is the number of visits to  $(x, a)$  before episode  $k$ , i.e.,

$$N_k(x, a) = \max \left( 1, \sum_{h=1}^H \sum_{s=1}^{k-1} \mathbb{I} \{ (x_h^s, a_h^s) = (x, a) \} \right).$$

If  $KH \geq XA$ , and if  $N_k(x, a)$  is updated **only at the end** of each episode, we can show that there exists a sequence  $(x_h^k, a_h^k)$  such that the sum  $\sum_{k,h} 1/N_k(x_h^k, a_h^k)$  is greater than  $HXA$ . Let  $(x_k, a_k)_{k \in [XA]}$  be  $XA$  *distinct* state-action pairs, and take the sequence  $(x_h^k, a_h^k)_{h \in [H], k \in [XA]}$  such that  $(x_h^k, a_h^k) = (x_k, a_k)$ . That is, in each of the  $XA$  episodes, the algorithm visits, in each of the  $H$  steps, *only one* state-action pair that *has never been visited before*. Since  $N_k(x, a)$  is updated only at the end of the episodes, we have  $N_k(x_h^k, a_h^k) = 1$  for all  $h \in [H]$  and  $k \in [XA]$ , with this choice of  $(x_h^k, a_h^k)_{h,k}$ . Hence,

$$H^2 X \sum_{k=1}^{XA} \sum_{h=1}^H \frac{1}{N_k(x_h^k, a_h^k)} = H^2 X \sum_{k=1}^{XA} \sum_{h=1}^H 1 = H^3 X^2 A.$$

Consequently, the sum of second order term is lower bounded (in a worst case sense) by  $H^3 X^2 A$  and cannot be  $\tilde{\mathcal{O}}(H^2 X^2 A)$  as claimed in [Azar et al. \(2017\)](#), since their bound *must hold for any possible sequence*  $(x_h^k, a_h^k)_{h,k}$ . An application of Lemma 9 with  $c = H$  can be used to show that the second order term is indeed  $\tilde{\mathcal{O}}(H^3 X^2 A)$  when updates are done at the end of the episodes only.

To gain a factor of  $H$  (i.e., have  $\tilde{\mathcal{O}}(H^2 X^2 A)$  as second order term), one solution is to update the counts  $N_k(x_h^k, a_h^k)$  every time a new state-action pair is observed, and recompute the optimistic  $Q$ -function. Another solution is to recompute it every time the number of visits of the current state-action pair is *doubled*, as done by [Jaksch et al. \(2010\)](#) in the average-reward setting.

The efficient version of our algorithm, **Greedy-Kernel-UCBVI**, does not suffer from this increased computational complexity in the stationary case. This is due to the fact that the value functions are updated in real time, and there is no need to run

<sup>10</sup>See page 7 of [Azar et al. \(2017\)](#).

a backward induction every time a new transition is observed. Hence, in the stationary case, **Greedy-Kernel-UCBVI** has a regret bound that is  $H$  times smaller than in the non-stationary case, *without* an increase in the computational complexity.

## E.2. Dependency on the Lipschitz Constant & Regularity w.r.t. the Total Variation Distance

Notice that the regret bound of **Kernel-UCBVI** has a linear dependency on  $L_1$  that appears in the bias term  $L_1 H K \sigma$ :

$$\mathcal{R}(K) \leq \tilde{\mathcal{O}} \left( H^2 \sqrt{|\mathcal{C}_\sigma| K} + L_1 H K \sigma + H^3 |\mathcal{C}_\sigma|^2 + H^2 |\mathcal{C}_\sigma| \right).$$

As long as the Lipschitz constant  $L_1 = \sum_{h=1}^H \lambda_r \lambda_p^{H-h}$  is  $\mathcal{O}(H)$  or  $\mathcal{O}(H^2)$ , our regret bound has no additional dependency on  $H$ . However, if  $\lambda_p > 1$ , the constant  $L_1$  can be exponential in  $H$ . This issue is caused by the smoothness of the MDP and not by algorithmic design. With minor modifications to our proof, we could also consider that the transitions are Lipschitz with respect to the total variation distance, in which case  $L_1$  would always be  $\mathcal{O}(H)$  and the regret of **Kernel-UCBVI** would remain  $\tilde{\mathcal{O}} \left( H^3 K^{\max(\frac{1}{2}, \frac{2d}{2d+1})} \right)$  by taking  $\sigma = (1/K)^{1/(2d+1)}$ . The regret bounds of other algorithms for Lipschitz MDPs also depend on the Lipschitz constant, which always appears in a bias term (e.g., [Ortner & Ryabko \(2013\)](#)).

In addition, the value  $L_h = \sum_{h'=h}^H \lambda_r \lambda_p^{H-h'}$  represents simply an upper bound on the Lipschitz constant of the  $Q$ -function  $Q_h^*$ . If the functions  $Q_h^*$  for  $h \in [H]$  are  $\tilde{L}_h$ -Lipschitz with  $\tilde{L}_h$  known and such that  $\tilde{L}_h < L_h$ , **Kernel-UCBVI** could exploit the knowledge of  $\tilde{L}_h$  and use it instead of  $L_h$ , which would also improve the regret bound. For instance, if all rewards functions  $r_h$  are 0 except for  $r_H$ , we could use  $\tilde{L}_h = \lambda_r$ , the Lipschitz constant of  $r_H$ , which is independent of  $H$ .

## F. Efficient implementation

In this Appendix, following [Efroni et al. \(2019\)](#), we show that if we only apply the optimistic Bellman operator once instead of doing a complete value iteration we obtain almost the same guaranties as for Algorithm 1 but with a large improvement in computational complexity. Indeed, the time complexity of each episode  $k$  is reduced from  $\mathcal{O}(k^2)$  to  $\mathcal{O}(k)$ . This complexity is comparable to other model-based algorithm in structured MDPs, e.g., [Jin et al. \(2019\)](#).

The algorithm goes as follows. Assume we are at episode  $k$  at step  $h$  at state  $x_h^k$ . To compute the next action we will apply the optimistic Bellman operator to the previous value function. That is, for all  $a \in \mathcal{A}$  we compute the upper bounds on the  $Q$ -value based on a kernel estimator:

$$\tilde{Q}_h^k(x_h^k, a) = \hat{r}_h^k(x, a) + \hat{P}_h^k V_{h+1}^k(x, a) + \mathbf{B}_h^k(x, a).$$

Then we act greedily

$$a_h^k = \operatorname{argmax}_{a \in \mathcal{A}} \tilde{Q}_h^k(x_h^k, a),$$

and define a new optimistic target  $\tilde{V}_h^k(x_h^k) = \min(H - h + 1, \tilde{Q}_h^k(x_h^k, a_h^k))$  for the value function at state  $x_h^k$ . Then we build an optimistic value function  $V_h^k$  by interpolating the previous optimistic target and the new one we just defined

$$\forall x, V_h^{k+1}(x) = \min \left( \min_{s \in [k-1]} [V_h^k(x_h^s) + L_h \rho_{\mathcal{X}}(x, x_h^s)], \tilde{V}_h^k(x_h^k) + L_h \rho_{\mathcal{X}}(x, x_h^k) \right).$$

The complete procedure is detailed in Algorithm 3.

**Proposition 9 (Optimism).** *In the event  $\mathcal{G}$ , whose probability is greater than  $1 - \delta$ , we have:*

$$\forall (k, h), \forall x, V_h^k(x) \geq V_h^*(x) \text{ and } V_h^k(x) \geq V_h^{k+1}(x).$$

*Proof.* To show that  $V_h^k(x) \geq V_h^{k+1}(x)$ , notice that

$$\forall x, V_h^{k+1}(x) = \min \left( V_h^k(x), \tilde{V}_h^k(x_h^k) + L_h \rho_{\mathcal{X}}(x, x_h^k) \right) \leq V_h^k(x)$$

since, by definition,  $V_h^k(x) = \min_{s \in [k-1]} [V_h^k(x_h^s) + L_h \rho_{\mathcal{X}}(x, x_h^s)]$ .

To show that  $V_h^k(x) \geq V_h^*(x)$ , we proceed by induction on  $k$ . For  $k = 1$ ,  $V_h^k(x) = H - h \geq V_h^*(x)$  for all  $x$  and  $h$ .

Now, assume that  $V_h^{k-1} \geq V_h^*$  for all  $h$ . As in the proof of Proposition 6, we prove that  $V_h^k \geq V_h^*$  by induction on  $h$ . For  $h = H + 1$ ,  $V_h^k(x) = V_h^*(x) = 0$  for all  $x$ . Now, assume that  $V_{h+1}^k(x) \geq V_{h+1}^*(x)$  for all  $x$ . We have, for all  $(x, a)$ ,

$$\begin{aligned}\tilde{Q}_h^k(x, a) &= \hat{r}_h^k(x, a) + \hat{P}_h^k V_{h+1}^k(x, a) + \mathbf{B}_h^k(x, a) \\ &\geq \hat{r}_h^k(x, a) + \hat{P}_h^k V_{h+1}^*(x, a) + \mathbf{B}_h^k(x, a) \quad \text{by induction hypothesis on } h \\ &\geq r_h(x, a) + P_h V_{h+1}^*(x, a) = Q_h^*(x, a) \quad \text{in } \mathcal{G}\end{aligned}$$

which implies that  $\tilde{V}_h^k(x_h^k) \geq V_h^*(x_h^k)$  and, consequently,

$$\begin{aligned}\tilde{V}_h^k(x_h^k) + L_h \rho_{\mathcal{X}}(x, x_h^k) &\geq V_h^*(x_h^k) + L_h \rho_{\mathcal{X}}(x, x_h^k) \geq V_h^*(x) \\ \implies V_h^k(x) &= \min \left( V_h^{k-1}(x), \tilde{V}_h^k(x_h^k) + L_h \rho_{\mathcal{X}}(x, x_h^k) \right) \geq V_h^*(x) \quad \text{by induction hypothesis on } k\end{aligned}$$

and we used the fact that  $V_h^*$  is  $L_h$ -Lipschitz.  $\square$

**Proposition 10.** *With probability at least  $1 - \delta$ , the regret of **Greedy-Kernel-UCBVI** is bounded as*

$$\begin{aligned}\mathcal{R}(K) &\leq H^2 |\mathcal{C}_\sigma| \\ &\quad + (L_1(1 + 2\lambda_p) + \theta_b(K, H)\sigma^d + 4L_1) KH\sigma \\ &\quad + (\sqrt{8e^2 H^2 \log(1/\delta)}) \sqrt{KH} \\ &\quad + (\theta_v(K, H) + 4\beta) eH |\mathcal{C}_\sigma| \left( \frac{1}{\beta} + \frac{1}{g(1)} \log \left( 1 + \frac{1 + g(1)\beta^{-1}K}{|\mathcal{C}_\sigma|} \right) \right) \\ &\quad + 2e(\mathbf{b}_p(K, H) + \mathbf{b}_r(K, H)) KH\sigma \\ &\quad + 2eH \left( \sqrt{\mathbf{v}_p(K, H)} + \sqrt{\mathbf{v}_r(K, H)} \right) \left( \left( \frac{1}{\beta} + \frac{2}{g(1)} \right) |\mathcal{C}_\sigma| + \frac{2}{g(1)} \sqrt{|\mathcal{C}_\sigma| g(1)\beta^{-1}K} \right) \\ &\quad + eH^2 |\tilde{\mathcal{C}}_\sigma| + e\sigma L_1 HK\end{aligned}$$

where  $\mathcal{C}_\sigma$  is a  $\sigma/4$ -covering of  $(\mathcal{X} \times \mathcal{A}, \rho)$  and  $\tilde{\mathcal{C}}_\sigma$  is a  $\sigma$ -covering of  $(\mathcal{X}, \rho_{\mathcal{X}})$ .

Hence,

$$\mathcal{R}(K) = \tilde{\mathcal{O}} \left( H^2 \sqrt{|\mathcal{C}_\sigma| K} + L_1 HK\sigma + H^3 |\mathcal{C}_\sigma|^2 + H^2 |\mathcal{C}_\sigma| + H^2 |\tilde{\mathcal{C}}_\sigma| \right).$$

*Proof.* On  $\mathcal{G}$ , we have

$$\begin{aligned}\tilde{\delta}_h^k &\stackrel{\text{def}}{=} V_h^{k+1}(x_h^k) - V_h^{\pi_k}(x_h^k) \\ &\leq \tilde{V}_h^k(x_h^k) - V_h^{\pi_k}(x_h^k) \\ &\leq \tilde{Q}_h^k(x_h^k, a_h^k) - Q_h^{\pi_k}(x_h^k, a_h^k) \\ &= \hat{r}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) - r_h(\tilde{x}_h^k, \tilde{a}_h^k) + \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + \hat{P}_h^k V_{h+1}^k(\tilde{x}_h^k, \tilde{a}_h^k) - P_h V_{h+1}^{\pi_k}(\tilde{x}_h^k, \tilde{a}_h^k) \\ &\leq r_{\mathbf{B}_h^k}(\tilde{x}_h^k, \tilde{a}_h^k) + \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + \hat{P}_h^k V_{h+1}^k(\tilde{x}_h^k, \tilde{a}_h^k) - P_h V_{h+1}^{\pi_k}(\tilde{x}_h^k, \tilde{a}_h^k).\end{aligned}$$

From this point we can follow the proof of Proposition 7 to obtain

$$\begin{aligned}\tilde{\delta}_h^k &\leq \left( 1 + \frac{1}{H} \right) (\delta_{h+1}^k + \xi_{h+1}^k) + 2 \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_1(1 + 2\lambda_p)\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \theta_b(k, h)\sigma^{1+d} + 4L_1\sigma \\ &\leq \left( 1 + \frac{1}{H} \right) \left( \tilde{\delta}_{h+1}^k + (V_{h+1}^k - V_{h+1}^{k+1})(x_{h+1}^k) + \xi_{h+1}^k \right) + 2 \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_1(1 + 2\lambda_p)\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] \\ &\quad + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \theta_b(k, h)\sigma^{1+d} + 4L_1\sigma.\end{aligned}$$

This gives us, on  $\mathcal{G}$ , using that  $V_h^* \leq V_h^{k+1}$ ,

$$\begin{aligned} \mathcal{R}(K) &\leq \sum_{k=1}^K \tilde{\delta}_1^k \\ &\leq \sum_{k=1}^K \sum_{h=1}^H \left(1 + \frac{1}{H}\right)^{h-1} \left(2 \mathbf{B}_h^k(\tilde{x}_h^k, \tilde{a}_h^k) + 2L_1(1 + 2\lambda_p)\rho[(\tilde{x}_h^k, \tilde{a}_h^k), (x_h^k, a_h^k)] + \frac{\theta_v(k, h)}{\mathbf{C}_h^k(\tilde{x}_h^k, \tilde{a}_h^k)} + \theta_b(k, h)\sigma^{1+d} + 4L_1\sigma\right) \\ &\quad + \sum_{k=1}^K \sum_{h=1}^H \left(1 + \frac{1}{H}\right)^h \xi_{h+1}^k \\ &\quad + \sum_{k=1}^K \sum_{h=1}^H \left(1 + \frac{1}{H}\right)^h (V_{h+1}^k - V_{h+1}^{k+1})(x_{h+1}^k). \end{aligned}$$

This bound differs only by the last additive term above from bound (6) obtained in the proof of Proposition 7. Thus we just need to handle this sum and rely on the previous analysis to upper bound the other terms. We consider a partition of the state space analogous to the one of Definition 3

**Definition 4.** Let  $\tilde{\mathcal{C}}_\sigma$  be a  $\sigma$ -covering of  $\mathcal{X}$ . We write  $\tilde{\mathcal{C}}_\sigma \stackrel{\text{def}}{=} \{x_j, j \in [\|\mathcal{C}_\sigma\|]\}$ . For each  $x_j \in \tilde{\mathcal{C}}_\sigma$ , we define the set  $B_j \subset \mathcal{X}$  as the set of state whose nearest neighbor in  $\tilde{\mathcal{C}}_\sigma$  is  $x_j$ , with ties broken arbitrarily, such that  $\{B_j\}_{j \in [\|\mathcal{C}_\sigma\|]}$  form a partition of  $\mathcal{X}$ .

Using the fact that the  $V_h^k$  are point-wise non-increasing we can transform the last sum in the previous inequality in a telescopic sum

$$\begin{aligned} \sum_{k=1}^K \sum_{h=1}^H \left(1 + \frac{1}{H}\right)^h (V_{h+1}^k - V_{h+1}^{k+1})(x_{h+1}^k) &\leq e \sum_{k=1}^K \sum_{h=1}^H (V_{h+1}^k - V_{h+1}^{k+1})(x_{h+1}^k) \\ &\leq e \sum_{j=1}^{|\tilde{\mathcal{C}}_\sigma|} \sum_{k=1}^K \sum_{h=1}^H (V_{h+1}^k - V_{h+1}^{k+1})(x_{h+1}^k) \mathbb{I}\{x_{h+1}^k \in B_j\} \\ &\leq e \sum_{j=1}^{|\tilde{\mathcal{C}}_\sigma|} \sum_{k=1}^K \sum_{h=1}^H (V_{h+1}^k - V_{h+1}^{k+1})(x_j) \mathbb{I}\{x_{h+1}^k \in B_j\} + L_h \rho_{\mathcal{X}}(x_j, x_{h+1}^k) \mathbb{I}\{x_{h+1}^k \in B_j\} \\ &\leq e \sum_{j=1}^{|\tilde{\mathcal{C}}_\sigma|} \sum_{k=1}^K \sum_{h=1}^H (V_{h+1}^k - V_{h+1}^{k+1})(x_j) + eK \sum_{h=1}^H L_1 \sigma \\ &\leq eH^2 |\tilde{\mathcal{C}}_\sigma| + e\sigma L_1 H K, \end{aligned}$$

where in the third inequality, we used the fact that the function  $V_{h+1}^k - V_{h+1}^{k+1}$  is  $L_h$ -Lipschitz. Combining the previous inequalities and the proof of Theorem 4, as explained above, allows us to conclude.  $\square$

## G. New Concentration Inequalities

In this section we present two new concentration inequalities that control, uniformly over time, the deviation of weighted sums of zero-mean random variables. They both follow from the so-called method of mixtures (e.g., Peña et al. (2008)), and can have applications beyond the scope of this work.

**Lemma 2** (Hoeffding type inequality). *Consider the sequences of random variables  $(w_t)_{t \in \mathbb{N}^*}$  and  $(Y_t)_{t \in \mathbb{N}^*}$  adapted to a filtration  $(\mathcal{F}_t)_{t \in \mathbb{N}}$ . Assume that, for all  $t \geq 1$ ,  $w_t$  is  $\mathcal{F}_{t-1}$  measurable and  $\mathbb{E} \left[ \exp(\lambda Y_t) \middle| \mathcal{F}_{t-1} \right] \leq \exp(\lambda^2 c^2 / 2)$  for all  $\lambda > 0$ .*

Let

$$S_t \stackrel{\text{def}}{=} \sum_{s=1}^t w_s Y_s \quad \text{and} \quad V_t \stackrel{\text{def}}{=} \sum_{s=1}^t w_s^2.$$

Then, for any  $\beta > 0$ , with probability at least  $1 - \delta$ , for all  $t \geq 1$ ,

$$\frac{S_t}{\sum_{s=1}^t w_s + \beta} \leq \sqrt{2c^2 \left[ \log\left(\frac{1}{\delta}\right) + \frac{1}{2} \log\left(\frac{V_t + \beta}{\beta}\right) \right] \frac{V_t + \beta}{\left(\sum_{s=1}^t w_s + \beta\right)^2}}.$$

In addition, if  $w_s \leq 1$  almost surely for all  $s$ , we have  $V_t \leq \sum_{s=1}^t w_s \leq t$  and the above can be simplified to

$$\frac{S_t}{\sum_{s=1}^t w_s + \beta} \leq \sqrt{2c^2 \log\left(\frac{\sqrt{1+t/\beta}}{\delta}\right) \frac{1}{\sum_{s=1}^t w_s + \beta}}.$$

*Proof.* Let

$$M_t^\lambda = \exp\left(\lambda S_t - \frac{\lambda^2 c^2 V_t}{2}\right),$$

with the convention  $M_0^\lambda = 1$ . The process  $\{M_t^\lambda\}_{t \geq 0}$  is a supermartingale, since

$$\mathbb{E}\left[M_t^\lambda \middle| \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\exp\left(w_t Y_t - \frac{\lambda^2 c^2 w_t^2}{2}\right) \middle| \mathcal{F}_{t-1}\right] M_{t-1}^\lambda \leq M_{t-1}^\lambda, \quad (8)$$

which implies that  $\mathbb{E}[M_t^\lambda] \leq \mathbb{E}[M_0^\lambda] = 1$ . Now, we apply the method of mixtures, as in [Peña et al. \(2008\)](#) see also [Abbasi-Yadkori et al. \(2011\)](#). We define the supermartingale  $M_t$  as

$$M_t = \sqrt{\frac{\beta c^2}{2\pi}} \int_{\mathbb{R}} M_t^\lambda \exp\left(-\frac{\beta c^2 \lambda^2}{2}\right) d\lambda = \sqrt{\frac{\beta}{V_t + \beta}} \exp\left(\frac{S_t^2}{2(V_t + \beta)c^2}\right).$$

The maximal inequality for non-negative supermartingales gives us:

$$\mathbb{P}[\exists t \geq 0 : M_t \geq \delta^{-1}] \leq \delta \mathbb{E}[M_0] = \delta.$$

Hence, with probability at least  $1 - \delta$ , we have

$$\forall t \geq 0, \quad S_t \leq \sqrt{2c^2 [\log(1/\delta) + (1/2) \log((V_t + \beta)/\beta)] (V_t + \beta)}.$$

Dividing both sides by  $\sum_{s=1}^t w_s + \beta$  gives the result.  $\square$

**Lemma 3** (Bernstein type inequality). *Consider the sequences of random variables  $(w_t)_{t \in \mathbb{N}^*}$  and  $(Y_t)_{t \in \mathbb{N}^*}$  adapted to a filtration  $(\mathcal{F}_t)_{t \in \mathbb{N}}$ . Let*

$$S_t \stackrel{\text{def}}{=} \sum_{s=1}^t w_s Y_s, \quad V_t \stackrel{\text{def}}{=} \sum_{s=1}^t w_s^2 \mathbb{E}\left[Y_s^2 \middle| \mathcal{F}_{s-1}\right] \quad \text{and} \quad W_t \stackrel{\text{def}}{=} \sum_{s=1}^t w_s,$$

and  $h(x) = (x+1)h(x+1) - x$ . Assume that, for all  $t \geq 1$ ,

- $w_t$  is  $\mathcal{F}_{t-1}$  measurable,
- $\mathbb{E}[Y_t \middle| \mathcal{F}_{t-1}] = 0$ ,
- $w_t \in [0, 1]$  almost surely,

- there exists  $b > 0$  such that  $|Y_t| \leq b$  almost surely.

Then, we have

$$\mathbb{P} \left[ \exists t \geq 1, (V_t/b^2 + 1)h \left( \frac{b|S_t|}{V_t + b^2} \right) \geq \log(1/\delta) + \log(4e(2t+1)) \right] \leq \delta.$$

The previous inequality can be weakened to obtain a more explicit bound: for all  $\beta > 0$ , with probability at least  $1 - \delta$ , for all  $t \geq 1$ ,

$$\frac{|S_t|}{\beta + \sum_{s=1}^t w_s} \leq \sqrt{2 \log(4e(2t+1)/\delta) \frac{V_t + b^2}{\left(\beta + \sum_{s=1}^t w_s\right)^2}} + \frac{2b \log(4e(2t+1)/\delta)}{3 \beta + \sum_{s=1}^t w_s}.$$

*Proof.* By homogeneity we can assume that  $b = 1$  to prove the first part. First note that for all  $\lambda > 0$ ,

$$e^{\lambda w_t Y_t} - \lambda w_t Y_t - 1 \leq (w_t Y_t)^2 (e^\lambda - \lambda - 1),$$

because the function  $y \rightarrow (e^y - y - 1)/y$  (extended by continuity at zero) is non-decreasing. Taking the expectation yields

$$\mathbb{E} [e^{\lambda w_t Y_t} | \mathcal{F}_{t-1}] - 1 \leq w_t^2 \mathbb{E} [Y_t^2 | \mathcal{F}_{t-1}] (e^\lambda - \lambda - 1),$$

thus using  $y + 1 \leq e^y$  we get

$$\mathbb{E} [e^{\lambda(w_t Y_t)} | \mathcal{F}_{t-1}] \leq e^{w_t^2 \mathbb{E} [Y_t^2 | \mathcal{F}_{t-1}] (e^\lambda - \lambda - 1)}.$$

We just proved that the following quantity is a supermartingale with respect to the filtration  $(\mathcal{F}_t)_{t \geq 0}$ ,

$$M_t^{\lambda,+} = e^{\lambda(S_t + V_t) - V_t(e^\lambda - 1)}.$$

Similarly, using that the same inequality holds for  $-X_t$ , we have

$$\mathbb{E} [e^{-\lambda w_t Y_t} | \mathcal{F}_{t-1}] \leq e^{w_t^2 \mathbb{E} [Y_t^2 | \mathcal{F}_{t-1}] (e^\lambda - \lambda - 1)},$$

thus, we can also define the supermartingale

$$M_t^{\lambda,-} = e^{\lambda(-S_t + V_t) - V_t(e^\lambda - 1)}.$$

We now choose the prior over  $\lambda_x = \log(x+1)$  with  $x \sim \mathcal{E}(1)$ , and consider the (mixture) supermartingale

$$M_t = \frac{1}{2} \int_0^{+\infty} e^{\lambda_x(S_t + V_t) - V_t(e^{\lambda_x} - 1)} e^{-x} dx + \frac{1}{2} \int_0^{+\infty} e^{\lambda_x(-S_t + V_t) - V_t(e^{\lambda_x} - 1)} e^{-x} dx.$$

Note that by construction it holds  $\mathbb{E} [M_t] \leq 1$ . We will apply the method of mixtures to that super martingale thus we need to lower bound it with the quantity of interest. To this aim we will lower bound the integral by the one only around the maximum of the integrand. Using the change of variable  $\lambda = \log(1+x)$ , we obtain

$$\begin{aligned} M_t &\geq \frac{1}{2} \int_0^{+\infty} e^{\lambda_x(|S_t| + V_t) - V_t(e^{\lambda_x} - 1)} e^{-x} dx \geq \frac{1}{2} \int_0^{+\infty} e^{\lambda(|S_t| + V_t + 1) - (V_t + 1)(e^\lambda - 1)} d\lambda \\ &\geq \frac{1}{2} \int_{\log(|S_t|/(V_t+1)+1)}^{\log(|S_t|/(V_t+1)+1+1/(V_t+1))} e^{\lambda(|S_t| + V_t + 1) - (V_t + 1)(e^\lambda - 1)} d\lambda \\ &\geq \frac{1}{2} \int_{\log(|S_t|/(V_t+1)+1)}^{\log(|S_t|/(V_t+1)+1+1/(V_t+1))} e^{\log(|S_t|/(V_t+1)+1)(|S_t| + V_t + 1) - |S_t| - 1} d\lambda \\ &= \frac{1}{2e} e^{(V_t+1)h(|S_t|/(V_t+1))} \log \left( 1 + \frac{1}{|S_t| + V_t + 1} \right) \geq \frac{1}{4e(2t+1)} e^{(V_t+1)h(|S_t|/(V_t+1))}, \end{aligned}$$



where in the last line we used  $\log(1 + 1/x) \geq 1/(2x)$  for  $x \geq 1$  and the trivial bounds  $|S_t| \leq 1$ ,  $V_t \leq t$ . The method of mixtures, see [Peña et al. \(2008\)](#), allows us to conclude for the first inequality of the lemma. The second inequality is a straightforward consequence of the previous one. Indeed, using that (see Exercise 2.8 of [Boucheron et al. \(2013\)](#)) for  $x \geq 0$

$$h(x) \geq \frac{x^2}{2(1 + x/3)},$$

we get

$$\frac{|S_t|/b}{V_t/b^2 + 1} \leq \sqrt{\frac{2 \log(4e(2t+1)/\delta)}{V_t/b^2 + 1}} + \frac{2 \log(4e(2t+1)/\delta)}{3} \frac{1}{V_t/b^2 + 1}.$$

Dividing by  $\beta + \sum_{s=1}^t w_s$  and multiplying by  $b(V_t/b^2 + 1)$  the previous inequality allows us to conclude.  $\square$

## H. Auxiliary Results

### H.1. Proof of Lemma 1

In this section, we prove that the  $Q$ -value  $Q_h$  is  $L_h$ -Lipschitz.

**Lemma 4** (Value functions are Lipschitz continuous). *Under assumptions 2 and 3, we have:*

$$\forall(x, a, x', a'), \forall h \in [H], \quad |Q_h^*(x, a) - Q_h^*(x', a')| \leq L_h \rho[(x, a), (x', a')]$$

where  $L_h \stackrel{\text{def}}{=} \sum_{h'=h}^H \lambda_r \lambda_p^{H-h'}$ .

*Proof.* We proceed by induction. For  $h = H$ ,  $Q_h^*(x, a) = r(x, a)$  and the statement is true, since  $r$  is  $\lambda_r$ -Lipschitz. Now, assume that it is true for  $h + 1$  and let's prove it for  $h$ .

First, we note that  $V_{h+1}^*(x)$  is Lipschitz by the induction hypothesis:

$$\begin{aligned} V_{h+1}^*(x) - V_{h+1}^*(x') &= \max_a Q_{h+1}^*(x, a) - \max_a Q_{h+1}^*(x', a) \leq \max_a (Q_{h+1}^*(x, a) - Q_{h+1}^*(x', a)) \\ &\leq \max_a \sum_{h'=h+1}^H \lambda_r \lambda_p^{H-h'} \rho[(x, a), (x', a)] = \sum_{h'=h+1}^H \lambda_r \lambda_p^{H-h'} \rho_{\mathcal{X}}(x, x'). \end{aligned}$$

By applying the same argument and inverting the roles of  $x$  and  $x'$ , we obtain

$$|V_{h+1}^*(x) - V_{h+1}^*(x')| \leq \sum_{h'=h+1}^H \lambda_r \lambda_p^{H-h'} \rho_{\mathcal{X}}(x, x').$$

Now, we have

$$\begin{aligned} Q_h^*(x, a) - Q_h^*(x', a') &\leq \lambda_r \rho[(x, a), (x', a')] + \int_{\mathcal{X}} V_{h+1}^*(y) (P_h(dy|x, a) - P_h(dy|x', a')) \\ &\leq \lambda_r \rho[(x, a), (x', a')] + L_{h+1} \int_{\mathcal{X}} \frac{V_{h+1}^*(y)}{L_{h+1}} (P_h(dy|x, a) - P_h(dy|x', a')) \\ &\leq \left[ \lambda_r + \lambda_p \sum_{h'=h+1}^H \lambda_r \lambda_p^{H-h'} \right] \rho[(x, a), (x', a')] = \sum_{h'=h}^H \lambda_r \lambda_p^{H-h'} \rho[(x, a), (x', a')] \end{aligned}$$

where, in last inequality, we use fact that  $V_{h+1}^*/L_{h+1}$  is 1-Lipschitz, the definition of the 1-Wasserstein distance and Assumption 3.  $\square$

## H.2. Covering-related lemmas

**Lemma 5.** Let  $\mathcal{F}_L$  be the set of  $L$ -Lipschitz functions from the metric space  $(\mathcal{X}, \rho)$  to  $[0, H]$ . Then, its  $\epsilon$ -covering number with respect to the infinity norm is bounded as follows

$$\mathcal{N}(\epsilon, \mathcal{F}_L, \|\cdot\|_\infty) \leq \left( \frac{8H}{\epsilon} \right)^{\mathcal{N}(\epsilon/(4L), \mathcal{X}, \rho)}$$

*Proof.* Let's build an  $\epsilon$ -covering of  $\mathcal{F}_L$ . Let  $\mathcal{C}_\mathcal{X} = \{x_1, \dots, x_M\}$  be an  $\epsilon_1$ -covering of  $(\mathcal{X}, \rho)$  such that  $\rho(x_i, x_j) > \epsilon_1$  for all  $i, j \in [M]$  (i.e.,  $\mathcal{C}_\mathcal{X}$  is also an  $\epsilon_1$ -packing). Let  $\mathcal{C}_{[0, H]} = \{y_1, \dots, y_N\}$  be an  $\epsilon_2$ -covering of  $[0, H]$ . For any function  $p : [M] \rightarrow [N]$ , we build a  $2L$ -Lipschitz function  $\hat{f}_p : \mathcal{X} \rightarrow \mathbb{R}$  as follows

$$\hat{f}_p(x) = \min_{i \in [M]} [y_{p(i)} + 2L\rho(x, x_i)].$$

Let  $\epsilon_1 = \epsilon/(4L)$  and  $\epsilon_2 = \epsilon/8$ . We now show that the set  $\mathcal{C}_{\mathcal{F}_L} \stackrel{\text{def}}{=} \{\hat{f}_p : p \text{ is a function from } [M] \text{ to } [N]\}$  is an  $\epsilon$ -covering of  $\mathcal{F}_L$ . Take an arbitrary function  $f \in \mathcal{F}_L$ . Let  $p : [M] \rightarrow [N]$  be such that  $|f(x_i) - y_{p(i)}| \leq \epsilon_2$  for all  $i \in [M]$ . For any  $x \in \mathcal{X}$ , let  $j \in [M]$  be such that  $\rho(x, x_j) \leq \epsilon_1$ . We have

$$\begin{aligned} |f(x) - \hat{f}_p(x)| &\leq |f(x_j) - \hat{f}_p(x_j)| + |f(x) - f(x_j)| + |\hat{f}_p(x_j) - \hat{f}_p(x)| \\ &\leq |f(x_j) - \hat{f}_p(x_j)| + 3L\rho(x, x_j) \\ &\leq |f(x_j) - y_{p(j)}| + |y_{p(j)} - \hat{f}_p(x_j)| + 3L\epsilon_1 \\ &\leq |y_{p(j)} - \hat{f}_p(x_j)| + 3L\epsilon_1 + \epsilon_2. \end{aligned}$$

Now, let's prove that  $\hat{f}_p(x_j) = y_{p(j)}$ , which is true if and only if  $y_{p(j)} \leq y_{p(i)} + 2L\rho(x, x_i)$  for all  $i \in [M]$ . By definition of  $p$  and the fact that  $f$  is  $L$ -Lipschitz, we have  $y_{p(j)} \leq y_{p(i)} + L\rho(x_j, x_i) + 2\epsilon_2 \leq y_{p(i)} + 2L\rho(x_j, x_i)$  for all  $i \in [M]$ , since  $L\rho(x_j, x_i) > L\epsilon_1 = 2\epsilon_2$ . Consequently,

$$\forall x, |f(x) - \hat{f}_p(x)| \leq 3L\epsilon_1 + \epsilon_2 < \epsilon$$

which shows that  $\mathcal{C}_{\mathcal{F}_L}$  is indeed an  $\epsilon$ -covering of  $\mathcal{F}_L$  whose cardinality is bounded by  $N^M$ . To conclude, we take  $\mathcal{C}_{[0, H]} = \{0, \epsilon_2, \dots, N\epsilon_2\}$  for  $N = \lceil H/\epsilon_2 \rceil$  and  $\mathcal{C}_\mathcal{X}$  such that  $|\mathcal{C}_\mathcal{X}| = M = \mathcal{N}(\epsilon_1, \mathcal{X}, \rho)$ .

For  $H = 1$ , this result is also given by [Gottlieb et al. \(2017\)](#), Lemma 5.2.  $\square$

**Lemma 6.** Let  $(\mathcal{X} \times \mathcal{A}, \rho)$  be a metric space and  $(\Omega, \mathcal{T}, \mathbb{P})$  be a probability space. Let  $F$  and  $G$  be two functions from  $\mathcal{X} \times \mathcal{A} \times \Omega$  to  $\mathbb{R}$  such that  $\omega \rightarrow F(x, a, \omega)$  and  $\omega \rightarrow G(x, a, \omega)$  are random variables. Also, assume that  $(x, a) \rightarrow F(x, a, \omega)$  and  $(x, a) \rightarrow G(x, a, \omega)$  are  $L_F$  and  $L_G$ -Lipschitz, respectively, for all  $\omega \in \Omega$ . If

$$\forall (x, a), \quad \mathbb{P}[\omega \in \Omega : G(x, a, \omega) \geq F(x, a, \omega)] \leq \delta$$

then

$$\mathbb{P}[\omega \in \Omega : \exists (x, a), G(x, a, \omega) \geq F(x, a, \omega) + (L_G + L_F)\epsilon] \leq \delta \mathcal{N}(\epsilon, \mathcal{X} \times \mathcal{A}, \rho).$$

*Proof.* Let  $\mathcal{C}_\epsilon$  be an  $\epsilon$ -covering of  $(\mathcal{X} \times \mathcal{A}, \rho)$  and let

$$(x_\epsilon, a_\epsilon) \stackrel{\text{def}}{=} \underset{(x', a') \in \mathcal{C}_\epsilon}{\operatorname{argmin}} \rho[(x', a'), (x, a)].$$

Let  $E \stackrel{\text{def}}{=} \{\omega \in \Omega : \exists (x, a), G(x, a, \omega) \geq F(x, a, \omega) + (L_G + L_F)\epsilon\}$ . In  $E$ , we have, for some  $(x, a)$ ,

$$G(x^\epsilon, a^\epsilon, \omega) + L_G\epsilon \geq G(x, a, \omega) \geq F(x, a, \omega) + (L_G + L_F)\epsilon \geq F(x^\epsilon, a^\epsilon, \omega) + L_G\epsilon.$$

Hence, in  $E$ , there exists  $(x, a)$  such that:

$$G(x^\epsilon, a^\epsilon, \omega) \geq F(x^\epsilon, a^\epsilon, \omega)$$

and

$$\begin{aligned} \mathbb{P}[E] &\leq \mathbb{P}[\omega \in \Omega : \exists (x^\epsilon, a^\epsilon) \in C_\epsilon, G(x^\epsilon, a^\epsilon, \omega) \geq F(x^\epsilon, a^\epsilon, \omega)] \\ &\leq \sum_{(x^\epsilon, a^\epsilon) \in C_\epsilon} \mathbb{P}[\omega \in \Omega : G(x^\epsilon, a^\epsilon, \omega) \geq F(x^\epsilon, a^\epsilon, \omega)] \leq \sum_{(x^\epsilon, a^\epsilon) \in C_\epsilon} \delta \end{aligned}$$

which gives us  $\mathbb{P}[E] \leq \delta \mathcal{N}(\epsilon, \mathcal{X} \times \mathcal{A}, \rho)$ .  $\square$

### H.3. Technical lemmas

We state and prove three technical lemmas that help controlling some of the sums that appear in our regret analysis.

**Lemma 7.** Consider a sequence of non-negative real numbers  $\{z_s\}_{s=1}^t$  and let  $g : \mathbb{R}_+ \rightarrow [0, 1]$  satisfy Assumption 4. Let

$$w_s \stackrel{\text{def}}{=} g\left(\frac{z_s}{\sigma}\right) \text{ and } \tilde{w}_s \stackrel{\text{def}}{=} \frac{w_s}{\beta + \sum_{s'=1}^t w_{s'}}.$$

for  $\beta > 0$ . Then, for  $t \geq e\beta/C_1^g$ , we have

$$\sum_{s=1}^t \tilde{w}_s z_s \leq 2\sigma \left(1 + \sqrt{\log(C_1^g t / \beta)}\right).$$

*Proof.* We split the sum into two terms:

$$\sum_{s=1}^t \tilde{w}_s z_s = \sum_{s: z_s < c} \tilde{w}_s z_s + \sum_{s: z_s \geq c} \tilde{w}_s z_s \leq c + \sum_{s: z_s \geq c} \tilde{w}_s z_s$$

From Assumption 4, we have  $w_s \leq C_1^g \exp(-z_s^2/(2\sigma^2))$ . Hence,  $\tilde{w}_s \leq (C_1^g/\beta) \exp(-z_s^2/(2\sigma^2))$ , since  $\beta + \sum_{s'=1}^t w_{s'} \geq \beta$ .

We want to find  $c$  such that:

$$z_s \geq c \implies \frac{C_1^g}{\beta} \exp\left(-\frac{z_s^2}{2\sigma^2}\right) \leq \frac{1}{t} \frac{2\sigma^2}{z_s^2}$$

which implies, for  $z_s \geq c$ , that  $\tilde{w}_s \leq \frac{1}{t} \frac{2\sigma^2}{z_s^2}$ .

Let  $x = z_s^2/2\sigma^2$ . Reformulating, we want to find a value  $c'$  such that  $C_1^g \exp(-x) \leq \beta/(xt)$  for all  $x \geq c'$ . Let  $c' = 2\log(C_1^g t / \beta)$ . If  $x \geq c'$ , we have:

$$\frac{x}{2} \geq \log \frac{C_1^g t}{\beta} \implies x \geq \frac{x}{2} + \log \frac{C_1^g t}{\beta} \implies x \geq \log x + \log(C_1^g t / \beta) \implies (C_1^g / \beta) \exp(-x) \leq 1/(xt)$$

as we wanted. Hence, we choose  $c' = 2\log(C_1^g t / \beta)$ .

Now,  $x \geq c'$  is equivalent to  $z_s \geq \sqrt{2\sigma^2 c'} = 2\sigma \sqrt{\log(C_1^g t / \beta)}$ . Therefore, we take  $c = 2\sigma \sqrt{\log(C_1^g t / \beta)}$ , which gives us

$$\sum_{s: z_s \geq c} \tilde{w}_s z_s \leq \sum_{s: z_s \geq c} \frac{1}{t} \frac{2\sigma^2}{z_s^2} z_s \leq \frac{2\sigma^2}{t} \sum_{s: z_s \geq c} \frac{1}{z_s} \leq \frac{2\sigma^2}{c} \frac{|\{s : z_s \geq c\}|}{t} \leq \frac{2\sigma^2}{c}$$

Finally, we obtain, for  $t \geq e\beta/C_1^g$ :

$$\sum_{s=1}^t \tilde{w}_s z_s \leq c + \sum_{s: z_s \geq c} \tilde{w}_s z_s \leq c + \frac{2\sigma^2}{c} = 2\sigma \sqrt{\log(C_1^g t / \beta)} + \frac{\sigma}{\sqrt{\log(C_1^g t / \beta)}} \leq 2\sigma \left(1 + \sqrt{\log(C_1^g t / \beta)}\right)$$

$\square$

**Lemma 8.** Let  $\{y_s\}_{s=1}^t$  be a sequence of real numbers and let  $\sigma > 0$ . For  $z \in \mathbb{R}_+^t$ , let

$$f_1(z) \stackrel{\text{def}}{=} \frac{\sum_{s=1}^t g(z_s/\sigma) y_s}{\beta + \sum_{s=1}^t g(z_s/\sigma)}, \quad f_2(z) \stackrel{\text{def}}{=} \sqrt{\frac{1}{\beta + \sum_{s=1}^t g(z_s/\sigma)}} \quad \text{and} \quad f_3(z) \stackrel{\text{def}}{=} \frac{1}{\beta + \sum_{s=1}^t g(z_s/\sigma)}.$$

Then,  $f_1$  and  $f_2$  are Lipschitz continuous with respect to the norm  $\|\cdot\|_\infty$ . Also, the Lipschitz constant of  $f_1$  is bounded by  $2C_2^g t(\max_s |y_s|)/(\beta\sigma)$  and the Lipschitz constant of  $f_2$  is bounded by  $(C_2^g t \beta^{-2/3})/(2\sigma)$ .

*Proof.* Using Assumption 4 and Lemma 7, the partial derivatives of  $f_1$  and  $f_2$  are bounded as follows

$$\begin{aligned} \left| \frac{\partial f_1(z)}{\partial z_s} \right| &\leq \frac{1}{\sigma} \frac{|g'(z_s/\sigma)| |y_s|}{\beta + \sum_{s=1}^t g(z_s/\sigma)} + \frac{1}{\sigma} \frac{\sum_{s=1}^t g(z_s/\sigma) |y_s|}{\left(\beta + \sum_{s=1}^t g(z_s/\sigma)\right)^2} |g'(z_s/\sigma)| \leq \frac{2C_2^g}{\beta\sigma} \max_s |y_s| \\ \left| \frac{\partial f_2(z)}{\partial z_s} \right| &\leq \frac{1}{2\sigma} \frac{|g'(z_s/\sigma)|}{\left(\beta + \sum_{s=1}^t g(z_s/\sigma)\right)^{3/2}} \leq \frac{C_2^g}{2\sigma\beta^{3/2}} \\ \left| \frac{\partial f_3(z)}{\partial z_s} \right| &\leq \frac{1}{\sigma} \frac{|g'(z_s/\sigma)|}{\left(\beta + \sum_{s=1}^t g(z_s/\sigma)\right)^2} \leq \frac{C_2^g}{\sigma\beta^2} \end{aligned}$$

Therefore,

$$\|\nabla f_1(z)\|_1 \leq \frac{2C_2^g t(\max_s |y_s|)}{\beta\sigma}, \quad \|\nabla f_2(z)\|_1 \leq \frac{C_2^g t}{2\sigma\beta^{3/2}}, \quad \|\nabla f_3(z)\|_1 \leq \frac{C_2^g t}{\sigma\beta^2}$$

and the result follows from the fact that  $|f_i(z_1) - f_i(z_2)| \leq \sup_z \|\nabla f_i(z)\|_1 \|z_1 - z_2\|_\infty$  for  $i \in \{1, 2, 3\}$ .  $\square$

**Lemma 9.** Consider a sequence  $\{a_n\}_{n \geq 1}$  of non-negative numbers such that  $a_n \leq c$  for some constant  $c > 0$ . Let  $A_t = \sum_{n=1}^{t-1} a_n$ . Then, for any  $b > 0$  and any  $p > 0$ ,

$$\sum_{t=1}^T \frac{a_t}{(1 + bA_t)^p} \leq c + \int_0^{A_{T+1}-c} \frac{1}{(1 + bz)^p} dz$$

*Proof.* Let  $n \stackrel{\text{def}}{=} \max \{t : a_1 + \dots + a_{t-1} \leq c\}$ . We have  $\sum_{t=1}^{n-1} \frac{a_t}{(1 + bA_t)^p} \leq \sum_{t=1}^{n-1} a_t \leq c$  and, consequently,

$$\begin{aligned} \sum_{t=1}^T \frac{a_t}{(1 + bA_t)^p} &\leq c + \sum_{t=n}^T \frac{a_t}{(1 + bA_t)^p} = c + \sum_{t=n}^T \frac{A_{t+1} - A_t}{(1 + bA_t)^p} \\ &= c + \sum_{t=n}^T \frac{A_{t+1} - A_t}{(1 + bA_{t+1} - ba_t)^p} \leq c + \sum_{t=n}^T \frac{A_{t+1} - A_t}{(1 + b(A_{t+1} - c))^p} \\ &= c + \sum_{t=n}^T \int_{A_t}^{A_{t+1}} \frac{1}{(1 + b(A_{t+1} - c))^p} dz \leq c + \sum_{t=n}^T \int_{A_t}^{A_{t+1}} \frac{1}{(1 + b(z - c))^p} dz \\ &= c + \int_{A_n}^{A_{T+1}} \frac{1}{(1 + b(z - c))^p} dz \leq c + \int_c^{A_{T+1}} \frac{1}{(1 + b(z - c))^p} dz. \end{aligned}$$

$\square$

## I. Experiments

In this section, we provide details about the experiments described in Section 6.

### I.1. Lipschitz Bandits

We consider the 1-Lipschitz reward function  $r(a) = \max(a, 1 - a)$  for  $a \in [0, 1]$ . At each time  $k$ , the agent computes an optimistic reward function  $r_k$ , chooses the action  $a_k \in \operatorname{argmax}_a r_k(a)$ , and observes  $r(a_k)$  plus a Gaussian noise of variance  $c^2$ . In order to solve this optimization problem, we choose 200 uniformly spaced points in  $[0, 1]$ . We chose a time-dependent kernel bandwidth in each episode as  $\sigma_k = 1/\sqrt{k}$ . For  $\text{UCB}(\delta)$ , we use the 200 points as arms. Let  $\{a_i\}_{i=1}^{200}$  be the points in  $[0, 1]$  representing the arms.

For **Kernel-UCBVI**, we used the following upper bound on the reward function for each  $a_i$ :

$$r_k(a_i) = c \sqrt{2 \left( \log \left( \frac{1}{\delta} \right) + \frac{1}{2} \log \left( 1 + \frac{\mathbf{V}_k(a_i)}{\beta} \right) \right) (\mathbf{V}_k(a_i) + \beta) \frac{1}{\sqrt{\mathbf{C}_k(a_i)}} + \frac{\beta}{\mathbf{C}_k(a_i)} + \frac{1}{\mathbf{C}_k(a_i)} \sum_{s=1}^{k-1} w_{s,k}(a_i, a_s) |a_i - a_s|}$$

where

$$w_{s,k}(a_i, a_s) = \exp \left( \frac{|a_i - a_s|^2}{2\sigma_k} \right), \quad \mathbf{C}_k(a_i) = \beta + \sum_{s=1}^{k-1} w_{s,k}(a_i, a_s), \quad \mathbf{V}_k(a_i) = \sum_{s=1}^{k-1} w_{s,k}(a_i, a_s)^2.$$

This upper bound on  $r(a_i)$  comes directly from Lemma 2, and it is tighter than the one proposed in Theorem 3. Indeed, to prove this theorem, we replaced  $\mathbf{V}_k(a_i)$  and  $\frac{1}{\mathbf{C}_k(a_i)} \sum_{s=1}^{k-1} w_{s,k}(a_i, a_s) |a_i - a_s|$  by their upper bounds  $t$  and  $2\sigma_k \left( 1 + \sqrt{\log(t/\beta)} \right)$ <sup>11</sup>, respectively. Replacing these values by their upper bounds allowed us to simplify the proof of the regret bound, but can degrade the practical performance of the algorithm.

For the baseline,  $\text{UCB}(\delta)$ , we used the following upper bound:

$$r_k(a_i) = c \sqrt{2 \left( \log \left( \frac{1}{\delta} \right) + \frac{1}{2} \log \left( 1 + \frac{\mathbf{N}_k(a_i)}{\beta} \right) \right) (\mathbf{N}_k(a_i) + \beta) \frac{1}{\sqrt{\beta + \mathbf{N}_k(a_i)}} + \frac{\beta}{\beta + \mathbf{N}_k(a_i)}}$$

where  $\mathbf{N}_k(a_i) = \sum_{s=1}^{k-1} \mathbb{I}\{a_s = a_i\}$  is the number of pulls of the arm  $a_i$ . This is equivalent to the bonus used by **Kernel-UCBVI** when the bandwidth is  $\sigma_k = 0$ , and can be seen as a version of the  $\text{UCB}(\delta)$  algorithm proposed by Abbasi-Yadkori et al. (2011), which also has a high-probability regret guarantee.

For **Kernel-UCBVI**, the bandwidth decreased with time,  $\sigma_k = 1/\sqrt{k}$ . However, to improve the computational efficiency,  $\sigma_k$  was only updated every 200 rounds, to avoid the computation of  $\mathbf{C}_k(a_i)$  and  $\mathbf{V}_k(a_i)$  at every round: in the rounds where  $\sigma_k$  is kept constant, these values can be updated incrementally for each  $a_i$ . Also, when  $\sigma_k$  is updated and  $r_k$  is updated, we make sure that the upper bounds are non-increasing, i.e.,  $r_k(a_i) \leq r_{k'}(a_i)$  for every  $i$  and every  $k \geq k'$ . By doing this, we avoid re-exploration of sub-optimal arms, and there is no loss of theoretical guarantees, since the upper bounds remain valid.

The parameters used where  $c = 0.25$ ,  $\beta = 0.05$ ,  $\delta = 0.1/200$ .

### I.2. Discrete MDP

We consider a  $8 \times 8$  GridWorld whose states are a uniform grid of points in  $[0, 1]^2$  and 4 actions, left, right, up and down. When an agent takes an action, it goes to the corresponding direction with probability 0.9 and to any other neighbor state with probability 0.1. The agent starts at  $(0, 0)$  and the reward functions depend on the distance to the goal state  $(1, 1)$ :

$$\forall h \in [H], \quad r_h(x, a) = \exp \left( -\frac{1}{2} \frac{(x_1 - 1)^2 + (x_2 - 1)^2}{0.1^2} \right)$$

where  $x = (x_1, x_2) \in [0, 1]^2$ . The reward obtained at  $(x, a)$  is  $r_h(x, a)$  plus a Gaussian noise of variance  $c^2$ .

For **Kernel-UCBVI**, we used the following exploration bonus

$$\mathbf{B}_h^k(x, a) = \frac{1}{\sqrt{\mathbf{C}_k(x, a)}} + \frac{H - h + 1}{\mathbf{C}_k(x, a)} + \frac{2\beta}{\mathbf{C}_k(x, a)} + \sigma_k.$$

<sup>11</sup>See Lemma 7.

where

$$\mathbf{C}_k(x, a) = \beta + \sum_{h=1}^H \sum_{s=1}^{k-1} w_h^{s,k}(x, a), \quad \text{with} \quad w_h^{s,k}(x, a) = \mathbb{I}\{a_h^s = a\} \exp\left(-\frac{\|x_h^s - x\|_2^2}{2\sigma_k^2}\right)$$

and where sum over  $h$  is to exploit the fact that the MDP is stationary. To motivate this choice of bonus, we notice that the theoretical bonus comes from the concentration inequality used to bound  $(P_h - \hat{P}_h^k)V_{h+1}^*(x, a)$ . From a Bernstein-type inequality (Lemma 3), we have

$$(P_h - \hat{P}_h^k)V_{h+1}^*(x, a) \lesssim \sqrt{\frac{\mathbb{V}_{y \sim P_h(\cdot|x,a)}[V_{h+1}^*(y)]}{\mathbf{C}_k(x, a)}} + \frac{H - h + 1}{\mathbf{C}_k(x, a)}$$

where  $\mathbb{V}_{y \sim P_h(\cdot|x,a)}[V_{h+1}^*(y)]$  is the variance of the optimal value function at the next state, which is unknown. However, since the transition noise is small, we do the approximation  $\mathbb{V}_{y \sim P_h(\cdot|x,a)}[V_{h+1}^*(y)] \approx 1$ . In practice, using this heuristic bonus motivated by Bernstein's inequality increases learning speed. The extra term  $\frac{2\beta}{\mathbf{C}_k(x, a)} + \sigma_k$  takes into account the regularization bias introduced by  $\beta$  and the bias  $\sigma_k$  introduced by the kernel function.

For UCBVI, we used the following exploration bonus

$$\mathbf{B}_h^k(x, a) = \frac{1}{\sqrt{\beta + \mathbf{N}_k(x, a)}} + \frac{H - h + 1}{\beta + \mathbf{N}_k(x, a)} + \frac{2\beta}{\beta + \mathbf{N}_k(x, a)}$$

where  $\mathbf{N}_k(x, a) = \sum_{h=1}^H \sum_{s=1}^{k-1} \mathbb{I}\{x_h^s = x, a_h^s = a\}$  is the number of visits to the state-action pair  $(x, a)$ . This is equivalent to the bonus used for **Kernel-UCBVI** with  $\sigma_k = 0$ .

To improve the computational efficiency we performed value iteration every 25 episodes for **Kernel-UCBVI** and UCBVI. For **Kernel-UCBVI**, we chose a time-dependent kernel bandwidth  $\sigma_k = 0.1 \log(k/25) / \sqrt{(k/25)}$ , which was updated every 500 episodes, so that  $\mathbf{C}_k(x, a)$  could be updated incrementally for every  $(x, a)$  in the episodes where  $\sigma_k$  was kept constant. In addition, since the MDP is discrete, it was not necessary to perform the interpolation described in Equation 5.

The parameters used were  $c = 0.1$  (standard deviation of the reward noise),  $\beta = 0.01$  and  $H = 20$ .

### I.3. Continuous MDP

We consider a variant of the previous environment having continuous state space  $\mathcal{X} = [0, 1]^2$ . When an agent takes an action (left, right, up or down) in a state  $x$ , its next state is  $x + \Delta x + \eta$ , where  $\Delta x$  is a displacement in the direction of the action and  $\eta$  is a Gaussian noise with zero mean and covariance matrix  $c_p^2 I_{2 \times 2}$ . The table below shows the displacement for each action.

| Action                      | Left        | Right      | Up         | Down        |
|-----------------------------|-------------|------------|------------|-------------|
| Displacement ( $\Delta x$ ) | $(-0.1, 0)$ | $(0.1, 0)$ | $(0, 0.1)$ | $(0, -0.1)$ |

The agent starts at  $(0.1, 0.1)$  and the reward functions depend on the distance to the goal state  $(0.75, 0.75)$ ,

$$\forall h \in [H], \quad r_h(x, a) = \exp\left(-\frac{1}{2} \frac{(x_1 - 0.75)^2 + (x_2 - 0.75)^2}{0.25^2}\right).$$

The reward obtained at  $(x, a)$  is  $r_h(x, a)$  plus a Gaussian noise of variance  $c_r^2$ .

The bandwidth of **Greedy-Kernel-UCBVI** was fixed to  $\sigma = 0.1$ . For **Greedy-UCBVI**, we discretize the state-action space with a uniform grid with steps of size 0.1, matching the value of  $\sigma$ .

For **Greedy-Kernel-UCBVI**, we used the following exploration bonus

$$\mathbf{B}_h^k(x, a) = \frac{1}{\sqrt{\mathbf{C}_k(x, a)}} + \frac{H - h + 1}{\mathbf{C}_k(x, a)} + \frac{\beta}{\mathbf{C}_k(x, a)} + 0.05\sigma.$$



where

$$\mathbf{C}_k(x, a) = \beta + \sum_{h=1}^H \sum_{s=1}^{k-1} w_h^{s,k}(x, a), \quad \text{with} \quad w_h^{s,k}(x, a) = \mathbb{I}\{a_h^s = a\} \exp\left(-\frac{\|x_h^s - x\|_2^2}{2\sigma_k^2}\right)$$

For Greedy-UCBVI, we used the following exploration bonus

$$\mathbf{B}_h^k(x, a) = \frac{1}{\sqrt{\mathbf{N}_k(I(x), a)}} + \frac{H - h + 1}{\mathbf{N}_k(I(x), a)}$$

where  $I(x)$  is the index of the discrete state corresponding to the continuous state  $x$  and  $\mathbf{N}_k(I(x), a) = \max\left(1, \sum_{h=1}^H \sum_{s=1}^{k-1} \mathbb{I}\{I(x_h^s) = I(x), a_h^s = a\}\right)$ .

The parameters used were  $c_p = c_r = 0.01$  (standard deviation of transitions and rewards noise),  $\beta = 0.05$ ,  $\lambda_p = \lambda_r = 1$  (Lipschitz constants of transitions and rewards).

#### I.4. Continuous MDP - Comparison to Optimistic Q-Learning

We repeated the previous experiment and compared it to the Optimist Q-Learning (OptQL) algorithm of (Jin et al., 2018) applied on a discretization of the MDP. Since OptQL is designed for non-stationary MDPs, we implemented the non-stationary versions of Greedy-Kernel-UCBVI and Greedy-UCBVI, whose bonuses were adapted as described below. Figure 2 shows that Greedy-Kernel-UCBVI outperforms both baselines, and we also see that Greedy-UCBVI outperforms OptQL.

For the non-stationary version of Greedy-Kernel-UCBVI, we used the following exploration bonus

$$\mathbf{B}_h^k(x, a) = \frac{1}{\sqrt{\mathbf{C}_h^k(x, a)}} + \frac{H - h + 1}{\mathbf{C}_h^k(x, a)} + \frac{\beta}{\mathbf{C}_h^k(x, a)} + 0.05\sigma \quad \text{where} \quad \mathbf{C}_h^k(x, a) = \beta + \sum_{s=1}^{k-1} w_h^{s,k}(x, a).$$

and  $w_h^{s,k}(x, a)$  is the same as in the previous experiment.

For OptQL and the non-stationary version of Greedy-UCBVI, we used the following exploration bonus

$$\mathbf{B}_h^k(x, a) = \frac{1}{\sqrt{\mathbf{N}_h^k(I(x), a)}} + \frac{H - h + 1}{\mathbf{N}_h^k(I(x), a)} \quad \text{where} \quad \mathbf{N}_h^k(I(x), a) = \max\left(1, \sum_{s=1}^{k-1} \mathbb{I}\{I(x_h^s) = I(x), a_h^s = a\}\right)$$

and  $I(x)$  is the index of the discrete state corresponding to  $x$ .

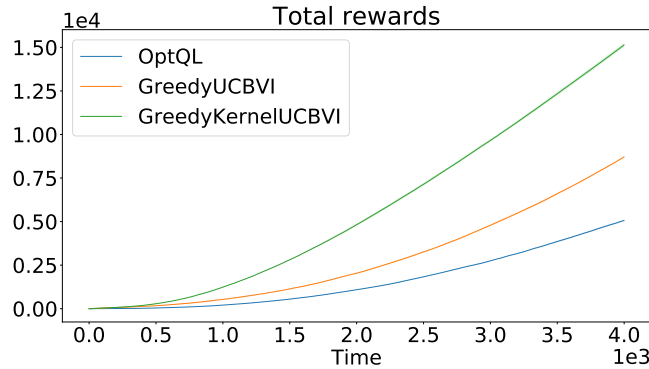


Figure 2. Total sum of rewards gathered by Greedy-Kernel-UCBVI in a continuous MDP versus Greedy-UCBVI and OptQL in a discretized version of the MDP (averaged over 8 runs). The shaded regions represent  $\pm$  the standard deviation.