A kernel-based approach to non-stationary reinforcement learning in metric spaces

Omar D Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, Michal Valko

A Kernel-Based Approach to Non-Stationary Reinforcement Learning in Metric Spaces

Omar D. Domingues[1,2] **Pierre Ménard**[3] **Matteo Pirotta**[4] **Emilie Kaufmann**[1,2,5] **Michal Valko**[1,6]
[1]Inria Lille [2]Université de Lille [3]OvGU [4]Facebook AI Research [5]CNRS [6]DeepMind Paris

Abstract

In this work, we propose KeRNS: an algorithm for episodic reinforcement learning in non-stationary Markov Decision Processes (MDPs) whose state-action set is endowed with a metric. Using a non-parametric model of the MDP built with time-dependent kernels, we prove a regret bound that scales with the covering dimension of the state-action space and the total variation of the MDP with time, which quantifies its level of non-stationarity. Our method generalizes previous approaches based on sliding windows and exponential discounting used to handle changing environments. We further propose a practical implementation of KeRNS, we analyze its regret and validate it experimentally.

1 Introduction

In reinforcement learning (RL), an agent interacts with an environment by sequentially taking actions, receiving rewards and observing state transitions. One of the main challenges in RL is the trade-off between exploration, the act of gathering information about the environment, and exploitation, the act of using the current knowledge to maximize the sum of rewards. In non-stationary environments, handling this trade-off becomes much harder: what has been learned in the past may no longer be valid in the present. Therefore, the agent needs to constantly re-explore previously known parts of the environment to discover possible changes. In this work, we propose KeRNS,[1] an algorithm that handles this problem by acting optimistically and by forgetting data that are far in the past, which naturally causes the agent to keep exploring to discover changes. KeRNS relies on non-parametric kernel estimators of the MDP, and the non-stationarity is handled by using time-dependent kernels.

The regret of an algorithm, defined as the difference between the rewards obtained by an optimal agent and the ones obtained by the algorithm, allows us to quantify how well an agent balances exploration and exploitation. We prove a regret bound for KeRNS that holds in a challenging setting, where the state-action space can be continuous and the environment can change in every episode, as long as the cumulative changes remain small when compared to the total number of episodes.

Related work Regret bounds for RL in stationary environments have been extensively studied in finite (tabular) MDPs (Jaksch et al., 2010; Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Zanette and Brunskill, 2019), and also in metric spaces under Lipschitz continuity assumptions (Ortner and Ryabko, 2012; Song and Sun, 2019; Sinclair et al., 2019; Domingues et al., 2020; Sinclair et al., 2020). Recent works provide algorithms with regret bounds for non-stationary RL in the tabular setting (Gajane et al., 2018; Ortner et al., 2019; Cheung et al., 2020). These algorithms estimate the transitions and the rewards in an episode k using the data observed up to episode $k-1$. However, since the MDP can change from one episode to another, these estimators are *biased*. If nothing is done to handle this bias, the algorithms will suffer a linear regret (Ortner et al., 2019) that depends on the magnitude of the bias. To deal with this issue, different approaches have been proposed: Gajane et al. (2018) and Cheung et al. (2020) use sliding windows to compute estimators that use only the most recently observed transitions, whereas Ortner et al. (2019) restart the algorithm periodically and, after each restart, new estimators are build and past data are discarded. In the multi-armed bandit literature, in addition to sliding windows, exponential discounting has also been used as a mean to give more importance to recent data (Kocsis and Szepesvári, 2006;

[1]meaning Kernel-based Reinforcement Learning in Non-Stationary environments.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Garivier and Moulines, 2011; Russac et al., 2019). In this paper, we study the *dynamic regret* of the algorithm, where, in each episode k, we compare the learner to the optimal policy of the MDP in episode k. A related approach consists in comparing the performance of the learner to the best stationary policy in hindsight, e.g., (Even-Dar et al., 2009; Yu and Mannor, 2009; Neu et al., 2013; Dick et al., 2014), which is less suited to non-stationary environments, since the performance of any fixed policy can be very bad. Non-stationary RL has also been studied outside the regret minimization framework, without, however, tackling the issue of exploration. For instance, Choi et al. (2000) propose a model where the MDP varies according to a sequence of tasks whose changes form a Markov chain. Szita et al. (2002) and Csáji and Monostori (2008) study the convergence of Q-learning when the environment changes but remain close to a fixed MDP. Assuming full knowledge of the MDP at each time step, but with unknown evolution, Lecarpentier and Rachelson (2019) introduce a risk-averse approach to planning in slowly changing environments. In a related setting, Lykouris et al. (2019) study episodic RL problems where the MDP can be corrupted by an adversary and provide regret bounds in this case.

Contributions We provide the first regret bound for non-stationary RL in continuous environments. More precisely, we show that the `Kernel-UCBVI` algorithm of Domingues et al. (2020), based on non-parametric kernel smoothing, can be modified to tackle non-stationary environments by using appropriate time- and space-dependent kernels. We analyze the resulting algorithm, `KeRNS`, under mild assumptions on the kernel, which in particular recover previously studied forgetting mechanisms to tackle non-stationarity in bandits and RL: sliding windows (Gajane et al., 2018) and exponential discounting (Kocsis and Szepesvári, 2006; Garivier and Moulines, 2011; Russac et al., 2019), and allow for combinations between those. On the practical side, kernel-based approaches can be very computationally demanding since their complexity grows with the number of data points. Building on the notion of representative states, promoted in previous work on practical kernel-based RL (Kveton and Theocharous, 2012; Barreto et al., 2016) we propose an efficient version of `KeRNS`, called `RS-KeRNS`, which has constant runtime per episode. We analyze the regret of `RS-KeRNS`, showing that it enables a trade-off between regret and runtime, and we validate this algorithm empirically.

2 Setting

Notation For any $n \in \mathbb{N}^*$, let $[n] \stackrel{\text{def}}{=} \{1, \dots, n\}$. If μ and $P(\cdot|x,a)$ are measures for any (x,a) and f is an arbitrary function, we define $\mu f \stackrel{\text{def}}{=} \int f(y) \mathrm{d}\mu(y)$ and $Pf(x,a) \stackrel{\text{def}}{=} \int f(y) \mathrm{d}P(y|x,a)$.[2]

Non-stationary MDPs We consider an episodic RL setting where, in each episode $k \in [K]$, an agent interacts with the environment for $H \in \mathbb{N}^*$ time steps. The time is indexed by (k,h), where k represents an episode and h the time step within the episode. The environment is modeled as a non-stationary MDP, defined by the tuple $(\mathcal{X}, \mathcal{A}, r, \mathrm{P})$, where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, $r = \left\{r_h^k\right\}_{k,h}$ and $\mathrm{P} = \left\{\mathrm{P}_h^k\right\}_{k,h}$ are sets of reward functions and transition kernels, respectively. More precisely, when taking action a in state x at time (k,h), the agent observes a random reward $\widetilde{r}_h^k \in [0,1]$ with mean $r_h^k(x,a)$ and makes a transition to the next state according to the probability measure $\mathrm{P}_h^k(\cdot|x,a)$. A deterministic policy π is a mapping from $[H] \times \mathcal{X}$ to $\mathcal{A}$, and we denote by $\pi(h,x)$ the action chosen in state x at step h. The action-value function of a policy π in step h of episode k is defined as

$$\mathrm{Q}_{k,h}^{\pi}(x,a) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}^k(x_{h'}, a_{h'}) \Big| x_h = x, a_h = a\right]$$

where $x_{h'+1} \sim \mathrm{P}_{h'}^k(\cdot|x_{h'}, a_{h'})$, $a_{h'} = \pi(h', x)$, and its value function is defined by $\mathrm{V}_{k,h}^{\pi}(x) = \mathrm{Q}_{k,h}^{\pi}(x, \pi(h,x))$. The optimal value functions, $\mathrm{V}_{k,h}^{*}(x) \stackrel{\text{def}}{=} \sup_\pi \mathrm{V}_{k,h}^{\pi}(x)$ satisfy the Bellman equations (Puterman, 2014)

$$\mathrm{V}_{k,h}^{*}(x) = \max_{a \in \mathcal{A}} \mathrm{Q}_{k,h}^{*}(x,a), \text{ where}$$
$$\mathrm{Q}_{k,h}^{*}(x,a) \stackrel{\text{def}}{=} r_h^k(x,a) + \mathrm{P}_h^k \mathrm{V}_{k,h+1}^{*}(x,a)$$

and where $\mathrm{V}_{k,H+1}^{*} = 0$ by definition.

Dynamic regret The agent interacts with the environment in a sequence of episodes and, in each episode k, it uses a policy π_k that can be chosen based on its observations from previous episodes. We measure its performance by the dynamic regret, defined as the sum over all episodes of the difference between the optimal value function in episode k and the value of π_k:

$$\mathcal{R}(K) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \left(\mathrm{V}_{k,1}^{*}(x_1^k) - \mathrm{V}_{k,1}^{\pi_k}(x_1^k)\right)$$

where x_1^k is the starting state in each episode, which is chosen arbitrarily and given to the learner.

Assumptions Since regret lower bounds scale with the number of states and actions (Jaksch et al., 2010), structural assumptions are needed in order to enable

[2] See also Table 2 in Appendix A summarizing the main notations used in the paper and in the proofs.

learning in continuous MDPs. A common assumption is that rewards and transitions are Lipschitz continuous with respect to some known metric (Ortner and Ryabko, 2012; Song and Sun, 2019; Domingues et al., 2020; Sinclair et al., 2020), which is the approach that we follow in this work. We make no assumptions regarding how the MDP changes, and our regret bounds will be expressed in terms of its total variation over time.

Assumption 1. *The state-action space $\mathcal{X} \times \mathcal{A}$ is equipped with a metric $\rho : (\mathcal{X} \times \mathcal{A})^2 \to \mathbb{R}_+$, which is given to the learner. Also, we assume that there exists a metric $\rho_{\mathcal{X}}$ on $\mathcal{X}$ such that, for all (x, x', a), $\rho[(x, a), (x', a)] \leq \rho_{\mathcal{X}}(x, x')$.*[3]

Assumption 2. *The reward functions are L_{r}-Lipschitz and the transition kernels are L_{p}-Lipschitz with respect to the 1-Wasserstein distance: $\forall (x, a, x', a')$ and $\forall (k, h) \in [K] \times [H]$,*

$$\left|r_h^k(x, a) - r_h^k(x', a')\right| \leq L_{\mathrm{r}} \rho[(x, a), (x', a')], \text{ and}$$
$$\mathbb{W}_1\left(\mathrm{P}_h^k(\cdot|x, a), \mathrm{P}_h^k(\cdot|x', a')\right) \leq L_{\mathrm{p}} \rho[(x, a), (x', a')]$$

where, for two measures μ and ν, we have $\mathbb{W}_1(\mu, \nu) \stackrel{\text{def}}{=} \sup_{f:\mathrm{Lip}(f)\leq 1} \int_{\mathcal{X}} f(y)(\mathrm{d}\mu(y) - \mathrm{d}\nu(y))$ and where, for any Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ with respect to $\rho_{\mathcal{X}}$, $\mathrm{Lip}(f)$ denotes its Lipschitz constant.

Assumption 3. *For any (k, h), the optimal Q-function $\mathrm{Q}_{k,h}^*$ is L-Lipschitz with respect to ρ. Assumptions 1 and 2 imply that $L \leq \sum_{h=1}^{H} L_{\mathrm{r}} {L_{\mathrm{p}}}^{H-h}$ (Lemma 26 in the Appendix).*

3 An Algorithm for Kernel-Based RL in Non-Stationary Environments

In this section, we introduce KeRNS, a model-based RL algorithm for learning in non-stationary MDPs. In each episode k, we estimate the transitions and the rewards using the data observed up to episode $k - 1$. Using exploration bonuses that represent the uncertainty in the estimated model, KeRNS builds a Q-function Q_h^k, and plays the greedy policy with respect to it. KeRNS generalizes sliding-window and exponential discounting approaches by considering time-dependent kernel functions, which also allow us to handle exploration in continuous environments (Domingues et al., 2020).

3.1 Kernel-Based Estimators for Changing MDPs

Let $\Gamma : \mathbb{N} \times (\mathcal{X} \times \mathcal{A})^2 \to [0, 1]$ be a *non-stationary kernel function*, where $\Gamma(t, u, v)$ represents the similarity between two state action pairs u, v in $\mathcal{X} \times \mathcal{A}$ visited at an interval t.

Definition 1 (kernel weights)**.** *Let (x_h^s, a_h^s) be the state-action pair visited at time (s, h). For any $(x, a) \in \mathcal{X} \times \mathcal{A}$ and $s < k$, we define the weights and the normalized weights at time (k, h) as*

$$w_h^{k,s}(x, a) \stackrel{\text{def}}{=} \Gamma\left(k - s - 1, (x, a), (x_h^s, a_h^s)\right)$$

and $\widetilde{w}_h^{k,s}(x, a) \stackrel{\text{def}}{=} w_h^{k,s}(x, a) / \mathbf{C}_h^k(x, a)$, where $\mathbf{C}_h^k(x, a) \stackrel{\text{def}}{=} \beta + \sum_{s=1}^{k-1} w_h^{k,s}(x, a)$ and $\beta > 0$ is a regularization parameter.

Using the kernel function Γ and past data, KeRNS builds estimators $\widehat{r}_h^k$ of the reward function and $\widehat{P}_h^k$ of the transitions at time (k, h), which are defined below.

Definition 2 (empirical MDP)**.** *At time $(s, h) \in [K] \times [H]$, let $(x_h^s, a_h^s, x_{h+1}^s, \widetilde{r}_h^s)$ represent the state, the action, the next state and the reward observed by the algorithm. Before each episode k, KeRNS estimates the rewards and transitions using the data observed up to episode $k - 1$:*

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

where δ_x is the Dirac measure at x. Let $\widehat{\mathcal{M}}_k$ be the MDP whose rewards and transitions at step h are $\widehat{r}_h^k(x, a)$ and $\widehat{P}_h^k(y|x, a)$.[4]

The weights $w_h^{k,s}(x, a)$ measure the influence that the transitions and rewards observed at time (s, h) will have on the estimators for the state-action pair (x, a) at time (k, h). Their sum, $\mathbf{C}_h^k(x, a)$, is a proxy for the number of visits to (x, a). Intuitively, the kernel function Γ must be designed in order to ensure that $w_h^{k,s}(x, a)$ is small when (x, a) is very far from (x_h^s, a_h^s), with respect to the distance ρ. It must also be small when $k - s - 1$ is large, which means that the sample (x_h^s, a_h^s) was collected too far in the past and should have a small impact on the estimators. For our theoretical analysis, we will need the assumptions below on the kernel function Γ.

Assumption 4 (kernel properties)**.** *Let $\sigma > 0$, $\eta \in]0, 1[$ and $W \in \mathbb{N}$ be the kernel parameters. For each set of parameters, we assume that we have access to a base kernel function $\overline{\Gamma}_{(\eta,W)} : \mathbb{N} \times \mathbb{R} \to [0, 1]$ and we define, for any $t, u, v \in \mathbb{N}^* \times \mathcal{X} \times \mathcal{A}$,*

$$\Gamma(t, u, v) = \overline{\Gamma}_{(\eta,W)}\left(t, \rho[u, v] / \sigma\right).$$

[3] If $(\mathcal{A}, \rho_{\mathcal{A}})$ is also a metric space, we can take $\rho[(x, a), (x', a')] = \rho_{\mathcal{X}}(x, x') + \rho_{\mathcal{A}}(a, a')$, for instance. See Section 2.3 of Sinclair et al. (2019) for more examples and a discussion.

[4] Since the normalized weights do not sum to 1, $\widehat{P}_h^k$ is not a probability kernel. In this case, we suffer a bias of order β and the property that $\widehat{P}_h^k$ is a sub-probability measure is enough for the analysis.

We assume that $z \mapsto \overline{\Gamma}_{(\eta,W)}(t,z)$ is non-increasing for any $t \in \mathbb{N}$. Additionally, we assume that there exists positive constants C_1, C_2, a constant $C_3 \geq 0$ and an arbitrary function $G : \mathbb{R} \to \mathbb{R}_{\geq 0}$ that satisfies $G(4) > 0$ such that

(1) $\forall (t,z),\ \overline{\Gamma}_{(\eta,W)}(t,z) \leq C_1 \exp\left(-z^2/2\right)$

(2) $\forall (t,y,z),\ \left|\overline{\Gamma}_{(\eta,W)}(t,y) - \overline{\Gamma}_{(\eta,W)}(t,z)\right| \leq C_2 \left|y - z\right|$

(3) $\forall z,\ \overline{\Gamma}_{(\eta,W)}(t,z) \leq C_3 \eta^t, \quad$ *for all* $t \geq W$

(4) $\forall z,\ \overline{\Gamma}_{(\eta,W)}(t,z) \geq G(z)\eta^t, \quad$ *for all* $t < W$.

We now provide some justification for these conditions. (1) ensures that the bias due to kernel smoothing remains bounded by $\widetilde{\mathcal{O}}(\sigma)$ (Lemma 23); (2) ensures smoothness conditions that are needed to provide concentration inequalities for the rewards and transitions (Lemma 24); (3) and (4) allow us to control the bias and the variance due to non-stationarity, respectively (Lemmas 2 and 16). Intuitively, (3) says the algorithm should forget data further than W episodes in the past, and (4) says that recent data in the W most recent episodes must have a minimum weight. The condition $G(4) > 0$ is mostly technical: it is used to ensure that $\mathbf{C}_h^k(x,a)$ is not too small in a 4σ-neighborhood of (x,a) (see lemmas 15 and 16). The kernels in the example below satisfy our conditions, and show that they indeed generalize sliding-window and exponential discounting approaches:

Example 1 (sliding-window and exponential discount)**.** *The kernels $\overline{\Gamma}_{(\eta,W)}(t,z) = \mathbb{I}\{t < W\} \exp(-|z|^p/2)$ (sliding-window) and $\overline{\Gamma}_{(\eta,W)}(t,z) = \eta^t \exp(-|z|^p/2)$ (exponential discount) satisfy Assumption 4 for $p \geq 2$.*

The conditions in Assumption 4 are needed to prove our regret bounds. However, if one has further knowledge about the MDP and its changes, this information can also be integrated to the kernel function Γ. For example, if the MDP only changes in certain region of the state-action space, the kernel can be designed to forget past data only in that region. Also, the kernel Γ can be designed to enforce restarts, as proposed by Ortner et al. (2019) for finite MDPs, by setting $\Gamma(t,u,v)$ to zero every time t exceeds a certain threshold. Although this would require a separate analysis, our proof could be combined to the one of (Ortner et al., 2019) to obtain a regret bound in this case.

3.2 Algorithm

KeRNS is presented in Algorithm 1. At time (k,h), let $\mathtt{B}_h^k(x,a)$ be the exploration bonus at (x,a) representing the uncertainty of $\widehat{\mathcal{M}}_k$ with respect to the true MDP:

$$\mathtt{B}_h^k(x,a) = \widetilde{\mathcal{O}}\left(\frac{H}{\sqrt{\mathbf{C}_h^k(x,a)}} + \frac{\beta H}{\mathbf{C}_h^k(x,a)} + L\sigma\right) \quad (1)$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic terms. The exact expression for the bonuses is given in Def. 5 in Appendix A. Before starting episode k, KeRNS computes, for all $h \in [H]$, the values Q_h^k by running backward induction on $\widehat{\mathcal{M}}_k$, with the bonus $\mathtt{B}_h^k(x,a)$ added to the rewards, followed by an interpolation step:

$$\widetilde{Q}_h^k(x,a) = \widehat{r}_h^k(x,a) + \widehat{P}_h^k V_{h+1}^k(x,a) + \mathtt{B}_h^k(x,a)$$
$$Q_h^k(x,a) = \min_{s \in [k-1]} \left(\widetilde{Q}_h^k(x_h^s, a_h^s) + L\rho\left[(x,a),(x_h^s,a_h^s)\right]\right)$$
$$V_h^k(x) = \min\left(H - h + 1, \max_a Q_h^k(x,a)\right)$$

where $V_{H+1}^k \stackrel{\text{def}}{=} 0$. The interpolation is needed to ensure that Q_h^k and V_h^k are L-Lipschitz. This procedure is defined in detail in Algorithm 3 in Appendix A, which is the same kind of backward induction used by Kernel-UCBVI (Domingues et al., 2020). Once Q_h^k is computed, KeRNS plays the greedy policy associated to it. Notice that, although $Q_h^k(x,a)$ and $V_h^k(x)$ are defined for all (x,a), they only need to be computed for the states and actions observed by the algorithm up to episode k.

Algorithm 1 KeRNS

Input: K, H, L, L_{r}, L_{p}, β, δ, d, σ, η, W.
Initialize history: $\mathcal{T}_h = \emptyset$ for all $h \in [H]$.
for episode $k = 1, \ldots, K$ **do**
  get initial state x_1^k
  // Run kernel backward induction
  compute $(Q_h^k)_h$ using $(\mathcal{T}_h)_h$ and Algorithm 3.
  for $h = 1, \ldots, H$ **do**
    execute $a_h^k = \mathrm{argmax}_a\, Q_h^k(x_h^k, a)$
    observe reward $\widetilde{r}_h^k$ and next state x_{h+1}^k
    store transition $\mathcal{T}_h = \mathcal{T}_h \cup \{x_h^k, a_h^k, x_{h+1}^k, \widetilde{r}_h^k\}$
  end for
end for

3.3 Theoretical guarantees

We introduce Δ, the total variation of the MDP in K episodes:

Definition 3 (MDP variation)**.** *We define $\Delta = \Delta^{\mathrm{r}} + L\Delta^{\mathrm{p}}$, where*

$$\Delta^{\mathrm{r}} \stackrel{\text{def}}{=} \sum_{i=1}^{K}\sum_{h=1}^{H} \sup_{x,a} \left|r_h^i(x,a) - r_h^{i+1}(x,a)\right|,$$
$$\Delta^{\mathrm{p}} \stackrel{\text{def}}{=} \sum_{i=1}^{K}\sum_{h=1}^{H} \sup_{x,a} \mathbb{W}_1\left(\mathrm{P}_h^i(\cdot|x,a), \mathrm{P}_h^{i+1}(\cdot|x,a)\right)$$

A similar notion has been introduced, for instance, by Ortner et al. (2019); Li and Li (2019) for MDPs and by Besbes et al. (2014) for multi-armed bandits. Here, the difference is that we use the Wasserstein distance to define the variation of the transitions, instead of the total variation (TV) distance $\|\mathrm{P}_h^i(\cdot|x,a) - \mathrm{P}_h^{i+1}(\cdot|x,a)\|_1$. This choice was made in order to take into account the metric ρ when measuring changes in the environment: our results would be analogous if we had chosen the TV distance.[5]

Using the same algorithm, we provide two regret bounds for KeRNS, which are given below. The notation $\lesssim$ omits constants and logarithmic terms (see Definition 4 in Appendix A).

Theorem 1. *The regret of KeRNS is bounded as* $\mathcal{R}^{\texttt{KeRNS}}(K) \lesssim \min\left(\mathcal{R}_1(K), \mathcal{R}_2(K)\right) + \mathbf{bias}(\sigma, \eta, W, \Delta)$, *where*

$$\mathcal{R}_1(K) = H^2 K \sqrt{\log\frac{1}{\eta}} \sqrt{|\mathcal{C}'_\sigma|\,|\mathcal{C}_\sigma|} + H^2\,|\mathcal{C}_\sigma|\,K\log\frac{1}{\eta}$$

$$\mathcal{R}_2(K) = H^2 K \sqrt{\log\frac{1}{\eta}} \sqrt{|\mathcal{C}_\sigma|} + H^3\,|\mathcal{C}_\sigma|\,|\mathcal{C}'_\sigma|\,K\log\frac{1}{\eta}$$

$$\mathbf{bias}(\sigma, \eta, W, \Delta) = W\Delta H + \frac{\eta^W}{1-\eta} K H^3 + LKH\sigma$$

with probability at least $1-\delta$. *Here,* $|\mathcal{C}'_\sigma|$ *and* $|\mathcal{C}_\sigma|$ *are the* σ*-covering numbers of* $(\mathcal{X}, \rho_\mathcal{X})$ *and* $(\mathcal{X}\times\mathcal{A}, \rho)$ *respectively,* (σ, η, W) *are the kernel parameters.*

Proof. This result comes from combining theorems 3 and 4 in Appendix F. See Section 5 for a proof outline.

As discussed below, after optimizing the kernel parameters (Table 1), the bound $\mathcal{R}_1$ has a worse dependence on K, and a better dependence on Δ. On the other hand, $\mathcal{R}_2$ is better with respect to K, but worse in Δ. Concretely, this trade-off may give hints on how to choose the kernel parameters according to the amount of variation that we expect to see in the environment. Technically, the difference comes from how we handle the concentration of the transitions in the proof. To obtain $\mathcal{R}_1$, we use concentration inequalities on the term $|(\widehat{P}_h^k - \mathrm{P}_h^k)f|$ for *all* functions f that are bounded and Lipschitz continuous. To obtain $\mathcal{R}_2$, the concentration is done only for $f = \mathrm{V}^*_{k,h+1}$, but this results in larger second-order terms, as in (Azar et al., 2017; Domingues et al., 2020).

Corollary 1. *Let* d *be the covering dimension of* $(\mathcal{X}\times\mathcal{A}, \rho)$. *By optimizing the kernel parameters, we obtain the regret bounds in Table 1. Table 3 in Appendix B.2 gives the values of* (σ, η, W) *that yield these bounds.*

Proof. Assuming that $|\mathcal{C}'_\sigma| \le |\mathcal{C}_\sigma|$, we have that $|\mathcal{C}_\sigma|$ and $|\mathcal{C}'_\sigma|$ are $\mathcal{O}\left(1/\sigma^d\right)$. Then, the bounds follow from Theorem 1. The general case, handling separately the covering dimensions of $(\mathcal{X}\times\mathcal{A}, \rho)$ and $(\mathcal{X}, \rho_\mathcal{X})$, is stated in corollaries 6 and 9 in Appendix F.

Discussion We now discuss regret bounds for optimized kernel parameters, according to the covering dimension of $(\mathcal{X}\times\mathcal{A}, \rho)$, denoted by d. Roughly, the covering dimension is the smallest number $d \ge 0$ such that the σ-covering number $|\mathcal{C}_\sigma|$ is $\mathcal{O}\left(1/\sigma^d\right)$.[6] We consider two cases: the tabular (finite MDP) case, where the covering dimension of $(\mathcal{X}\times\mathcal{A}, \rho)$ is $d = 0$, and the continuous case, where $d > 0$.

Tabular case Let $X = |\mathcal{X}|$ and $A = |\mathcal{A}|$. By taking $\sigma = 0$, we have $|\mathcal{C}'_\sigma| = X$ and $|\mathcal{C}_\sigma| = XA$. As shown in Table 1, the $\mathcal{R}_1$ bound states that the regret of KeRNS is $\widetilde{\mathcal{O}}\left(H^2 X\sqrt{A}\Delta^{\frac{1}{3}}K^{\frac{2}{3}}\right)$. This bound matches the one proved by Ortner et al. (2019) for the average reward setting using restarts, up to a factor of $H^{\frac{2}{3}}$ coming from our episodic setting, where the transitions P_h^k depend on h. The $\mathcal{R}_2$ bound states that the regret of KeRNS can be improved to $\widetilde{\mathcal{O}}\left(H^2\sqrt{XA}\Delta^{\frac{1}{3}}K^{\frac{2}{3}}\right)$, up to second-order terms. In the bandit case ($H = 1$), these bounds are *optimal* in terms of K and Δ (Besbes et al., 2014).

Continuous case For $d > 0$, we prove the first dynamic regret bounds in our setting, which are of order $H^2\Delta^{\frac{1}{3}}K^{\frac{2d+2}{2d+3}}$ (better in Δ) or $H^2\Delta^{\frac{1}{2}}K^{\frac{2d+1}{2d+2}}$ (better in K) for two different tunings of the kernel. Deriving a lower bound in the non-stationary case for $d > 0$ is an open problem, even for multi-armed bandits. As a sanity-check, we note that in stationary MDPs, for which $\Delta = 0$, we recover the regret bound of Kernel-UCBVI[7] (Domingues et al., 2020) of $H^3K^{\frac{2d}{2d+1}}$ from the bound $\mathcal{R}_2$ with $\log(1/\eta) = 1/K$, $W \to \infty$ and $\sigma = K^{-\frac{1}{2d+1}}$, which is optimal for $d = 1$ in the (stationary) bandit case (Bubeck et al., 2011).

In tabular MDPs, we may achieve sub-linear regret as long as $\Delta < K$.[8] In the continuous case however, our bounds show that we might need $\Delta < K^{\frac{3}{2d+3}}$ (for the $\mathcal{R}_1$ bound) or $\Delta < K^{\frac{1}{d+1}}$ (for the $\mathcal{R}_2$ bound) in order to avoid a linear regret, which is an immediate consequence of the bounds in Table 1.

Knowledge of Δ To optimally choose the kernel parameters, KeRNS requires an upper bound on the

[5] More precisely, in the proof of Corollary 2, the Wasserstein distance could be replaced by the TV distance.

[6] For more details about covering numbers and covering dimension, see Section 3 of Kleinberg et al. (2019) and Section 2.2 of Sinclair et al. (2019).

[7] Another choice of η might allow us to avoid the dependence on H^3 of Kernel-UCBVI and get H^2 instead.

[8] Notice that, if Δ scales linearly with the number of episodes K, we cannot expect to learn. Indeed, according to the lower bound (Besbes et al., 2014), the regret is necessarily linear in this case.

Table 1: Regret for optimized kernel parameters.

	bound	regret
$d = 0$	$\mathcal{R}_1$	$H^2 X \sqrt{A} \Delta^{\frac{1}{3}} K^{\frac{2}{3}}$
	$\mathcal{R}_2$	$H^2 \sqrt{XA} \Delta^{\frac{1}{3}} K^{\frac{2}{3}} + H^3 X^2 A \Delta^{\frac{2}{3}} K^{\frac{1}{3}}$
$d > 0$	$\mathcal{R}_1$	$H^2 \Delta^{\frac{1}{3}} K^{\frac{2d+2}{2d+3}}$
	$\mathcal{R}_2$	$H^2 \Delta^{\frac{1}{2}} K^{\frac{2d+1}{2d+2}} + H^{\frac{3}{2}} \Delta^{\frac{1}{4}} K^{\frac{3}{4}}$

variation Δ. Recent work has started to tackle this issue in bandit algorithms (Chen et al., 2019; Auer et al., 2019), and finite MDPs using sliding windows (Cheung et al., 2020). Their extension to continuous MDPs is left to future work.

4 Efficient Implementation

Since KeRNS uses non-parametric kernel estimators, its computational complexity scales with the number of observed transitions. Let τ_A be the time required to compute the maximum of $a \mapsto Q_h^k(x,a)$. Similarly to Kernel-UCBVI, its total space complexity is $\mathcal{O}(KH)$ and its time complexity per episode k is $\mathcal{O}\left(Hk^2 + H\tau_A k\right)$, resulting in a total runtime of $\mathcal{O}\left(HK^3 + H\tau_A K^2\right)$. This runtime is very prohibitive in practice, especially in changing environments, where we might need to run the algorithm for a very long time. Domingues et al. (2020) propose a version of Kernel-UCBVI with improved per-episode time complexity of $\mathcal{O}(H\tau_A k)$ based on real-time dynamic programming (RTDP) (Barto et al., 1995; Efroni et al., 2019). However, this requires the upper bounds V_h^k to be non-increasing, which is not the case in KeRNS, since V_h^k increases in regions that were not visited recently. This property is necessary to promote extra exploration and adapt to possible changes. Additionally, the RTDP-based approach of Domingues et al. (2020) still has a time complexity that scales with time, which can be a considerable issue in practice. Here, we propose an alternative to run KeRNS in *constant* time per episode, while controlling the impact of this speed-up on the regret.

4.1 Using Representative States and Actions

As proposed by Kveton and Theocharous (2012) and Barreto et al. (2016), we take an approach based on using *representative states* to construct an algorithm called RS-KeRNS (for KeRNS on Representative States). In each episode k, RS-KeRNS keeps and updates sets of representative states $\bar{\mathcal{X}}_h$, actions $\bar{\mathcal{A}}_h$ and next-states $\bar{\mathcal{Y}}_h$, for each h, whose cardinalities are denoted by $\bar{X}_h$, $\bar{A}_h$ and $\bar{Y}_h$, respectively. For simplicity, we omit the dependence on k of these sets and their cardinalities. Every time a new transition $\left\{x_h^k, a_h^k, x_{h+1}^k, \tilde{r}_h^k\right\}$ is observed, the representative sets are updated using Algorithm 2, which ensures that any two representative state-action pairs are at a distance greater than ε from each other. Similarly, it ensures that any pair of representative next-states are at a distance greater than $\varepsilon_{\mathcal{X}}$ from each other. Then, (x_h^k, a_h^k) and x_{h+1}^k are mapped to their nearest neighbors in $\bar{\mathcal{X}}_h \times \bar{\mathcal{A}}_h$ and $\bar{\mathcal{Y}}_h$, respectively, and the estimators of the rewards and transitions are updated. Consequently, we build a finite MDP, denoted by $\widetilde{\mathcal{M}}_k$, with $\bar{X}_h$ states, $\bar{A}_h$ actions and $\bar{Y}_h$ next-states, *per stage* h. The rewards and transitions of $\widetilde{\mathcal{M}}_k$ can be stored in arrays of size $\bar{X}_h \bar{A}_h$ and $\bar{X}_h \bar{A}_h \bar{Y}_h$, for each h.

RS-KeRNS is described precisely in Algorithm 4 in Appendix G. It computes a Q-function for all $(\overline{x}, \overline{a}) \in \cup_h \bar{\mathcal{X}}_h \times \bar{\mathcal{A}}_h$ by running backward induction in $\widetilde{\mathcal{M}}_k$, which is then extended to any $(x,a) \in \mathcal{X} \times \mathcal{A}$ by performing an interpolation step, as in KeRNS. In Appendix G, we explain how the rewards and transitions estimators of $\widetilde{\mathcal{M}}_k$ can be updated online. Below, we provide regret and runtime guaranties for this efficient implementation.

Algorithm 2 Update Representative Sets

Input: k, h, $\bar{\mathcal{X}}_h$, $\bar{\mathcal{A}}_h$, $\bar{\mathcal{Y}}_h$, $\left\{x_h^k, a_h^k, x_{h+1}^k\right\}$, ε, $\varepsilon_{\mathcal{X}}$.
if $\min_{(\overline{x},\overline{a}) \in \bar{\mathcal{X}}_h \times \bar{\mathcal{A}}_h} \rho\left[(\overline{x},\overline{a}), (x_h^k, a_h^k)\right] > \varepsilon$ **then**
  $\bar{\mathcal{X}}_h = \bar{\mathcal{X}}_h \cup \left\{x_h^k\right\}, \quad \bar{\mathcal{A}}_h = \bar{\mathcal{A}}_h \cup \left\{a_h^k\right\}$
end if
if $\min_{\overline{y} \in \bar{\mathcal{Y}}_h} \rho_{\mathcal{X}}\left(\overline{x}, x_{h+1}^k\right) > \varepsilon_{\mathcal{X}}$ **then**
  $\bar{\mathcal{Y}}_h = \bar{\mathcal{Y}}_h \cup \left\{x_{h+1}^k\right\}$
end if

4.2 Theoretical Guarantees & Runtime

Theorem 2 states that RS-KeRNS enjoys the same regret bounds as KeRNS plus a bias term that can be controlled by ε and $\varepsilon_{\mathcal{X}}$, as long as we use a Gaussian kernel.

Theorem 2 (regret of RS-KeRNS). *Let $\chi_{(\eta,W)} : \mathbb{N} \to [0,1]$, $u, v \in \mathcal{X} \times \mathcal{A}$, and consider the kernel*

$$\Gamma(t,u,v) = \chi_{(\eta,W)}(t) \exp\left(-\rho\left[u,v\right]^2 / (2\sigma^2)\right)$$

assumed to satisfy Assumption 4. With this choice of kernel, the regret of RS-KeRNS is bounded by

$$\mathcal{R}(K) \lesssim \mathcal{R}^{\text{KeRNS}}(K) + L(\varepsilon + \varepsilon_{\mathcal{X}}) KH^2 + \frac{\varepsilon}{\sigma} KH^3$$

with probability at least $1-\delta$, where $\mathcal{R}^{\text{KeRNS}}$ is regret bound of KeRNS given in Theorem 1.

Proof. This result comes from theorems 5 and 6 in Appendix G. See Appendix B for a proof outline.

Lemma 1 (runtime of RS-KeRNS). *Consider the kernel defined in Theorem 2, and let $(\alpha_i)_{i\geq 1}$ be a sequence of real numbers in $[0,1]$. If we take $\chi_{(\eta,W)}(t) = \prod_{i=1}^{t} \alpha_i$, the per-episode runtime of RS-KeRNS is bounded by*

$$\mathcal{O}\left(H \min\left(k^2, |\mathcal{C}_\varepsilon| \left|\mathcal{C}'_{\varepsilon_\mathcal{X}}\right|\right) + H \min\left(k, \left|\mathcal{C}'_{\varepsilon_\mathcal{X}}\right|\right) \tau_A\right),$$

where $|\mathcal{C}_\varepsilon|$ is the ε-covering number of $(\mathcal{X}\times\mathcal{A}, \rho)$, $\left|\mathcal{C}'_{\varepsilon_\mathcal{X}}\right|$ is the $\varepsilon_\mathcal{X}$-covering number of $(\mathcal{X}, \rho)$, and τ_A is the time required to compute the maximum of $a \mapsto Q_h^k(x,a)$. In particular, we can take $\alpha_i = \eta$ for all i, which gives an exponential-discount strategy for handling non-stationarity.

Proof. By construction, in any episode k, we have $\bar{X}_h \bar{A}_h \leq \min\left(k, |\mathcal{C}_\varepsilon|\right)$ and $\bar{Y}_h \leq \min\left(k, \left|\mathcal{C}'_{\varepsilon_\mathcal{X}}\right|\right)$. Backward induction (Algorithm 5) is performed in $\mathcal{O}\left(\sum_h \bar{X}_h \bar{A}_h \bar{Y}_h + \tau_A \sum_h \bar{Y}_h\right)$ time, and the choice of $\chi_{(\eta,W)}(t)$ implies that the model updates can be done in $\mathcal{O}\left(\sum_h \bar{X}_h \bar{A}_h \bar{Y}_h\right)$ time, as detailed in Appendix G.2.

Consequently, the constants ε and $\varepsilon_\mathcal{X}$ provide a trade-off between regret and computational complexity. Since $|\mathcal{C}_\varepsilon| = \mathcal{O}\left(\varepsilon^{-d_1}\right)$ and $\left|\mathcal{C}'_{\varepsilon_\mathcal{X}}\right| = \mathcal{O}\left(\varepsilon^{-d_2}\right)$, increasing $(\varepsilon, \varepsilon_\mathcal{X})$ may reduce *exponentially* the runtime of RS-KeRNS, while having only a *linear* increase in its regret.

Kveton and Theocharous (2012) and Barreto et al. (2016) studied the idea of using representative states to accelerate kernel-based RL (KBRL), but we provide the first regret bounds in this setting. More precisely, our result improves previous work in the following aspects: (i) Kveton and Theocharous (2012) and Barreto et al. (2016) do not tackle exploration and do not have finite–time analyses: they provide approximate versions of the KBRL algorithm of Ormoneit and Sen (2002) which has asymptotic guarantees assuming that transitions are generated from *independent* samples; (ii) The error bounds of Kveton and Theocharous (2012) scale with $\exp(1/\sigma^2)$. In our online setting, σ can be chosen as a function of the horizon K, and their bound could result in an error that scales exponentially with K, instead of linearly, as ours. Our result comes from an improved analysis of the smoothness of kernel estimators, that leverages the regularization constant β (Lemma 25); (iii) Barreto et al. (2016) propose an algorithm that also builds a set of representative states in an online way. However, their theoretical guarantees only hold when this set is *fixed*, i.e., cannot be updated during exploration, whereas our bounds hold in this case; (iv) unlike (Kveton and Theocharous, 2012; Barreto et al., 2016), our theoretical results also hold in continuous action spaces.

4.3 Numerical Validation

To illustrate the behavior of RS-KeRNS, we consider a continuous MDP whose state-space is the unit ball in $\mathbb{R}^2$ with four actions, representing a move to the right, left, up or down. The agent starts at $(0,0)$. Let $b_i^k \in \{0, 0.25, 0.5, 0.75, 1\}$ and $x_i \in \{(0.8, 0.0), (0.0, 0.8), (-0.8, 0.0), (0.0, -0.8)\}$. We consider the following mean reward function:

$$r_h^k(x,a) = \sum_{i=1}^{4} b_i^k \max\left(0, 1 - \frac{\|x - x_i\|_2}{0.5}\right)$$

which do not depend on h. Every N episodes, the coefficients b_i^k are changed, which impact the optimal policy.

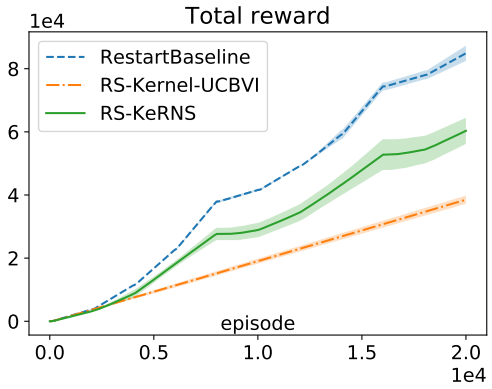

Figure 1: Performance of RS-KeRNS compared to baselines for $N = 2000$. Average over 4 runs.

Taking $\eta = \exp(-(1/N)^{2/3})$, we used the kernel $\Gamma(t,u,v) = \eta^t \exp\left(-(\rho[u,v]/\sigma)^4/2\right)$. We set $\sigma = 0.05$, $\varepsilon = \varepsilon_\mathcal{X} = 0.1$, $\beta = 0.01$, $H = 15$ and ran the algorithm for 2×10^4 episodes. KeRNS was compared to two baselines: (i) `Kernel-UCBVI` combined with representative states, that we call `RS-Kernel-UCBVI`, which is designed for stationary environments. This corresponds to RS-KeRNS with $\chi(t) = 1$, that is, $\eta = 1$; (ii) A restart-based algorithm, called `RestartBaseline`, which is implemented as `RS-Kernel-UCBVI`, but it has *full information* about when the environment changes, and, at every change, it restarts its reward estimator and its bonuses. We can see that, as expected, RS-KeRNS outperforms `RS-Kernel-UCBVI`, which was not designed for non-stationary environments, and is able to "track" the behavior of the restart-based algorithm which has full information about how the environment changes. In Appendix I, we give more details about the experimental setup and provide more experiments, varying the period N of changes in the MDP and the kernel function.

5 Proof Outline

We now outline the proof of Theorem 1 assuming, for simplicity, that the rewards are known.

Bias due to non-stationarity To bound the bias, we introduce an average MDP with transitions $\overline{P}_h^k$:

$$\overline{P}_h^k(y|x,a) \stackrel{\text{def}}{=} \sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) \mathrm{P}_h^s(y|x,a) + \frac{\beta\, \mathrm{P}_h^k(y|x,a)}{\mathbf{C}_h^k(x,a)},$$

where $(\mathrm{P}_h^s)_{s,h}$ are the true transitions at time (s,h). We prove that, for any L-Lipschitz function f bounded by H, (Corollary 2):

$$\left|\left(\mathrm{P}_h^k - \overline{P}_h^k\right) f(x,a)\right| \le \mathbf{bias_p}(k,h)$$

where the term $\mathbf{bias_p}(k,h)$ is defined as

$$L \sum_{i=1\vee(k-W)}^{k-1} \sup_{x,a} \mathbb{W}_1\left(\mathrm{P}_h^i(\cdot|x,a), \mathrm{P}_h^{i+1}(\cdot|x,a)\right) + \frac{2C_3 H}{\beta} \frac{\eta^W}{1-\eta}.$$

Concentration Using concentration inequalities for weighted sums, we prove that $\widehat{P}_h^k$ is close to the average transition $\overline{P}_h^k$ using Hoeffding- and Bernstein-type inequalities (lemmas 5, 6, 7, and 8), and define an event $\mathcal{G}$ where our confidence sets hold (Lemma 9), such that $\mathbb{P}[\mathcal{G}] \ge 1 - \delta/2$. For instance, Lemma 5 gives us

$$\left|(\widehat{P}_h^k - \overline{P}_h^k)\mathrm{V}_{k,h+1}^*(x,a)\right| \lesssim \sqrt{\frac{H^2}{\mathbf{C}_h^k(x,a)}} + \frac{\beta H}{\mathbf{C}_h^k(x,a)} + L\sigma,$$

which explains the form of the exploration bonuses.

Upper bound on the true value function On the event $\mathcal{G}$, we show that (Lemma 10):

$$Q_h^k(x,a) + \sum_{h'=h}^{H} \mathbf{bias}(k,h) \ge \mathrm{Q}_{k,h}^*(x,a)$$

where the term $\mathbf{bias}(k,h)$ is the sum of $\mathbf{bias_p}(k,h)$ defined above, and a similar term representing the bias in the reward estimation.

Regret bounds Let $(\widetilde{x}_h^k, \widetilde{a}_h^k)$ be the state-action pair among the previously visited ones that is the closest to (x_h^k, a_h^k):

$$(\widetilde{x}_h^k, \widetilde{a}_h^k) \stackrel{\text{def}}{=} \operatorname*{argmin}_{(x_h^s, a_h^s): s<k, h\in[H]} \rho\left[(x_h^k, a_h^k), (x_h^s, a_h^s)\right].$$

We show that (see proof of Lemma 11):

$$H \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{I}\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] > 2\sigma\right\} \le H^2 |\mathcal{C}_\sigma|.$$

Thus, to simplify the outline, for all (k,h), we assume that $\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \le 2\sigma$ and add $H^2|\mathcal{C}_\sigma|$ to the final regret bound. On the event $\mathcal{G}$, we prove that the regret of KeRNS is bounded by (lemmas 11 and 12):

$$\mathcal{R}(K) \lesssim \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\frac{H}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}\right) + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k,h) + LKH\sigma + H^2|\mathcal{C}_\sigma|$$

where we omitted factors involving $|\mathcal{C}_\sigma|$ and $|\mathcal{C}'_\sigma|$ (which depend on the type of bound considered, $\mathcal{R}_1$ or $\mathcal{R}_2$), and martingale terms (which are bounded by $\approx H^{3/2}\sqrt{K}$ with probability at least $1-\delta/2$).

Using the properties of the kernel Γ (Assumption 4), we prove that (Lemma 16):

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{1}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} \lesssim HK \log\frac{1}{\eta}\left(|\mathcal{C}_\sigma| + \sqrt{\frac{|\mathcal{C}_\sigma|}{\log(1/\eta)}}\right)$$

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{1}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} \lesssim H|\mathcal{C}_\sigma| K \log\frac{1}{\eta}$$

Finally, in Corollary 5, we prove that the bias $\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(h,k)$ is bounded by

$$2W\left(\Delta^{\mathrm{r}} + L\Delta^{\mathrm{p}}\right) + \frac{2C_3(H+1)KH}{\beta}\frac{\eta^W}{1-\eta}.$$

Putting these bounds together, we prove Theorem 1. The proof of Theorem 2 is outlined in Appendix B.

6 Conclusion

In this paper, we introduced and analyzed KeRNS, the first algorithm for continuous MDPs with dynamic regret guarantees in changing environments. Building upon previous work on using representative states for kernel-based RL, we proposed RS-KeRNS, a practical version of KeRNS that runs in constant time per episode. Moreover, we provide the first analysis that quantifies the trade-off between the regret and the computational complexity of this approach. In discrete environments, our regret bound matches the existing lower bound for multi-armed bandits in terms of the number of episodes and the variation of MDP, whereas finding a lower bound in continuous environments remains an open problem.

We believe that the ideas introduced in this paper might be useful for large-scale problems. Indeed, we provide stronger online guarantees for practical kernel-based RL, which has already been shown to be empirically

successful in medium-scale environments ($d \approx 10$) (Kveton and Theocharous, 2012; Barreto et al., 2016), and we show that kernel-based RL is naturally suited to tackle non-stationarity. In larger dimension, kernel-based exploration bonuses have been recently shown to enhance exploration in RL for Atari games (Badia et al., 2020), and our approach might inspire the design of bonuses for high-dimensional non-stationary environments.

Acknowledgements

The research presented was supported by European CHIST-ERA project DELTA, French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council, French National Research Agency project BOLD (ANR19-CE23-0026-04), FMJH PGMO project 2018-0045, and the SFI Sachsen-Anhalt for the project RE-BCI ZS/2019/10/102024 by the Investitionsbank Sachsen-Anhalt.

References

Auer, P., Gajane, P., and Ortner, R. (2019). Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org.

Badia, A. P., Sprechmann, P., Vitvitskyi, A., Guo, D., Piot, B., Kapturowski, S., Tieleman, O., Arjovsky, M., Pritzel, A., Bolt, A., and Blundell, C. (2020). Never give up: Learning directed exploration strategies. In *International Conference on Learning Representations*.

Barreto, A. M., Precup, D., and Pineau, J. (2016). Practical kernel-based reinforcement learning. *The Journal of Machine Learning Research*, 17(1):2372–2441.

Barto, A. G., Bradtke, S. J., and Singh, S. P. (1995). Learning to act using real-time dynamic programming. *Artificial intelligence*, 72(1-2):81–138.

Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pages 199–207.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2011). X-armed bandits. *Journal of Machine Learning Research*, 12:1587–1627.

Chen, Y., Lee, C.-W., Luo, H., and Wei, C.-Y. (2019). A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pages 696–726. PMLR.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2020). Reinforcement learning for non-stationary Markov decision processes: The blessing of (More) optimism. In *Proceedings of the 37th International Conference on Machine Learning*.

Choi, S. P., Yeung, D.-Y., and Zhang, N. L. (2000). Hidden-mode markov decision processes for nonstationary sequential decision making. In *Sequence Learning*, pages 264–287. Springer.

Csáji, B. C. and Monostori, L. (2008). Value function based reinforcement learning in changing markovian environments. *Journal of Machine Learning Research*, 9(Aug):1679–1709.

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723.

Dick, T., Gyorgy, A., and Szepesvari, C. (2014). Online learning in markov decision processes with changing cost sequences. In *International Conference on Machine Learning*, pages 512–520. PMLR.

Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. (2020). Regret Bounds for Kernel-Based Reinforcement Learning. *arXiv e-prints*, page arXiv:2004.05599.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. (2019). Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pages 12203–12213.

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

Gajane, P., Ortner, R., and Auer, P. (2018). A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066*.

Garivier, A. and Moulines, E. (2011). On Upper-Confidence Bound Policies For Switching Bandit Problems. In *Algorithmic Learning Theory (ALT)*, pages 174–188. PMLR.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

Kleinberg, R., Slivkins, A., and Upfal, E. (2019). Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77.

Kocsis, L. and Szepesvári, C. (2006). Discounted UCB. In *2nd PASCAL Challenges Workshop*.

Kveton, B. and Theocharous, G. (2012). Kernel-based reinforcement learning on representative states. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Lecarpentier, E. and Rachelson, E. (2019). Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 7214–7223.

Li, Y. and Li, N. (2019). Online learning for markov decision processes in nonstationary environments: A dynamic regret analysis. In *2019 American Control Conference (ACC)*, pages 1232–1237. IEEE.

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. (2019). Corruption robust exploration in episodic reinforcement learning. *arXiv preprint arXiv:1911.08689*.

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

Ormoneit, D. and Sen, Ś. (2002). Kernel-based reinforcement learning. *Machine Learning*, 49(2):161–178.

Ortner, R., Gajane, P., and Auer, P. (2019). Variational regret bounds for reinforcement learning. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*.

Ortner, R. and Ryabko, D. (2012). Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1763–1771.

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

Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 12017–12026.

Sinclair, S., Wang, T., Jain, G., Banerjee, S., and Yu, C. (2020). Adaptive discretization for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3858–3871.

Sinclair, S. R., Banerjee, S., and Yu, C. L. (2019). Adaptive discretization for episodic reinforcement learning in metric spaces. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–44.

Song, Z. and Sun, W. (2019). Efficient model-free reinforcement learning in metric spaces. *arXiv preprint arXiv:1905.00475*.

Szita, I., Takács, B., and Lörincz, A. (2002). ε-mdps: Learning in varying environments. *Journal of Machine Learning Research*, 3(Aug):145–174.

Yu, J. Y. and Mannor, S. (2009). Online learning in markov decision processes with arbitrarily changing rewards and transitions. In *2009 international conference on game theory for networks*, pages 314–322. IEEE.

Zanette, A. and Brunskill, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning*.

Appendix

Table of Contents

A Preliminaries

A.1 Notation

Throughout the proof, we use the following notation when omitting constants and logarithmic terms:

Definition 4.

$$A \lesssim B \iff A \leq B \times \text{polynomial}\left(d_1, d_2, \log(K), \log(1/\delta), \beta, 1/\beta, L_{\mathrm{r}}, L_{\mathrm{p}}\right).$$

Table 2 summarizes the main notations used in the paper and in the proofs.

Table 2: Table of notations

Notation	Meaning
$\rho : (\mathcal{X} \times \mathcal{A})^2 \to \mathbb{R}_+$	metric on the state-action space $\mathcal{X} \times \mathcal{A}$
$\rho_{\mathcal{X}} : \mathcal{X}^2 \to \mathbb{R}_+$	metric on the state space $\mathcal{X}$
$\mathcal{N}(\epsilon, \mathcal{X} \times \mathcal{A}, \rho)$	ϵ-covering number of the metric space $(\mathcal{X} \times \mathcal{A}, \rho)$
K	number of episodes
H	horizon, length of each episode
δ	confidence parameter
σ	kernel bandwidth parameter
η	kernel temporal decay parameter
W	kernel temporal window parameter
β	regularization parameter
$\lvert\mathcal{C}_\sigma\rvert$, $\lvert\mathcal{C}'_\sigma\rvert$	σ-covering numbers of $(\mathcal{X} \times \mathcal{A}, \rho)$ and $(\mathcal{X}, \rho_{\mathcal{X}})$, respectively
$L_{\mathrm{r}}, L_{\mathrm{p}}, L$	Lipschitz constants of the rewards, transitions and value functions
Γ	kernel function from $\mathbb{N}^* \times (\mathcal{X} \times \mathcal{A})^2$ to $[0,1]$
$\overline{\Gamma}_{(\eta,W)}$	parameterized kernel function from $\mathbb{N}^* \times \mathbb{R}_+$ to $[0,1]$
C_1, C_2, C_3	constants related to kernel properties, see Assumption 4
$G : \mathbb{R}_+ \to \mathbb{R}_+$	function related to kernel properties, see Assumption 4
$\mathcal{M}_k$	true MDP at episode k, with rewards r_h^k and transitions P_h^k
$\widehat{\mathcal{M}}_k$	empirical MDP built by `KeRNS` in episode k
$w_h^{k,s}(x,a)$	weight at (x,a) in at time (k,h) w.r.t the sample (x_h^s, a_h^s) (Def. 1)
$\widetilde{w}_h^{k,s}(x,a)$	normalized version of $w_h^{k,s}(x,a)$ (Def. 1)
$(\mathrm{Q}^*_{k,h}, \mathrm{V}^*_{k,h})_{h\in[H]}$	true value functions in episode k
$(Q_h^k, V_h^k)_{h\in[H]}$	value functions computed by `KeRNS` in episode k
$(\widetilde{Q}_h^k, \widetilde{V}_h^k)_{h\in[H]}$	value functions computed by `RS-KeRNS` in episode k
$\bar{\mathcal{X}}_h^k, \bar{\mathcal{A}}_h^k, \bar{\mathcal{Y}}_h^k$	sets of representative states, actions and next states, at stage h of episode k
ζ_h^k	mapping from $\mathcal{X} \times \mathcal{A}$ to $\bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k$
$\bar{\zeta}_h^k$	mapping from $\mathcal{X}$ to $\bar{\mathcal{Y}}_h^k$
$\Delta^{\mathrm{r}}, \Delta^{\mathrm{p}}$	temporal variation of the rewards and transitions (Def. 3)
Δ	temporal variation of the MDP $= \Delta^{\mathrm{r}} + L\Delta^{\mathrm{p}}$
d $(= d_1)$	covering dimension of $(\mathcal{X} \times \mathcal{A}, \rho)$
d_2	covering dimension of $(\mathcal{X}, \rho_{\mathcal{X}})$
ε	threshold distance to add a new representative state-action pair
$\varepsilon_{\mathcal{X}}$	threshold distance to add a new representative state
$\mathcal{L}(L, H)$	set of L-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}$ bounded by H
$\mathbf{bias_p}(k,h)$	bias in transition estimation at time (k,h), (Def. 6)
$\mathbf{bias_r}(k,h)$	bias in reward estimation at time (k,h) (Def. 6)
$\mathbf{bias}(k,h)$	sum of biases $\mathbf{bias_r}(k,h) + \mathbf{bias_p}(k,h)$ (Def. 6)
$\mathcal{G}$	good event, on which confidence intervals hold (Lemma 9)
$\log^+(z)$	equal to $\log(z+e)$ for any $z \in \mathbb{R}$

A.2 Probabilistic model

The interaction between the algorithm and the MDP defines a stochastic process $(x_h^s, a_h^s, x_{h+1}^s, \widetilde{r}_h^s)$ for $h \in [H]$ and $s \in \mathbb{N}^*$, representing the state, the action, the next state and the reward at step h of episode s. Let $\mathcal{H}_h^s \stackrel{\text{def}}{=} \left\{x_{h'}^{s'}, a_{h'}^{s'}, x_{h'+1}^{s'}, \widetilde{r}_{h'}^{s'}\right\}_{s'<s, h' \in [H]} \cup \left\{x_{h'}^{s}, a_{h'}^{s}, x_{h'+1}^{s}, \widetilde{r}_{h'}^{s}\right\}_{h'<h}$ be the history of the process up to time (s,h).

We define $\mathcal{F}_h^s$ as the σ-algebra generated by $\mathcal{H}_h^s$, and denote its corresponding filtration by $(\mathcal{F}_h^s)_{s,h}$.

A.3 Exploration Bonuses and Kernel Backward Induction

A reinforcement learning algorithm can be seen as a mapping from the set of possible histories $\bigcup_{h \in [H], k \in \mathbb{N}^*} (\mathcal{X} \times \mathcal{A} \times \mathcal{X} \times [0,1])^{kh-1}$ to the set of actions $\mathcal{A}$.

For a time (k,h), KeRNS performs this mapping in the following way:

1. Build $\widehat{r}_h^k$ and $\widehat{P}_h^k$ as in Definition 2, which are $\mathcal{F}_h^{k-1}$-measurable.

2. For each $h \in [H]$, with $V_{H+1}^k = 0$,

   - Compute

$$\widetilde{Q}_h^k(x,a) = \widehat{r}_h^k(x,a) + \widehat{P}_h^k V_{h+1}^k(x,a) + \mathtt{B}_h^k(x,a) \quad \text{for all } (x,a) \in \{(x_h^s, a_h^s)\}_{s<k}$$

   - Define, for any (x,a),

$$Q_h^k(x,a) = \min_{s \in [k-1]} \left[\widetilde{Q}_h^k(x_h^s, a_h^s) + L\rho\left[(x,a), (x_h^s, a_h^s)\right]\right]$$
$$V_h^k(x) = \min\left(H - h + 1, \max_{a'} Q_h^k(x,a')\right)$$

3. Choose the action $a_h^k \in \operatorname{argmax}_a Q_h^k(x_h^k, a)$.

Notice the algorithmic structure of KeRNS is the same as `Kernel-UCBVI` (Domingues et al., 2020). However, KeRNS uses non-stationary kernels to be able to adapt to changing environments.

It can be checked that Algorithm 3 returns the functions Q_h^k described above.

The exploration bonus is defined below:

Definition 5 (exploration bonuses). *The exploration bonus in (x,a) at time (k,h) is defined as*

$$\mathtt{B}_h^k(x,a) \stackrel{\text{def}}{=} {}^{\mathrm{r}}\mathtt{B}_h^k(x,a) + {}^{\mathrm{p}}\mathtt{B}_h^k(x,a), \quad \textit{where}$$

$${}^{\mathrm{r}}\mathtt{B}_h^k(x,a) \stackrel{\text{def}}{=} \sqrt{\frac{2\square_1^{\mathrm{r}}(k,\delta/8)}{\mathbf{C}_h^k(x,a)}} + \frac{\beta}{\mathbf{C}_h^k(x,a)} + \mathbf{b}_{\mathrm{r}}(k,\delta/8)\sigma, \quad \textit{and}$$

$${}^{\mathrm{p}}\mathtt{B}_h^k(x,a) \stackrel{\text{def}}{=} \sqrt{\frac{2H^2\square_1^{\mathrm{p}}(k,\delta/8)}{\mathbf{C}_h^k(x,a)}} + \frac{\beta H}{\mathbf{C}_h^k(x,a)} + \mathbf{b}_{\mathrm{p}}(k,\delta/8)\sigma$$

where

$$\square_1^{\mathrm{r}}(k,\delta) = \widetilde{\mathcal{O}}\left(d_1\right) = \log\left(\frac{\mathcal{N}\left(\sigma^2/K, \mathcal{X}\times\mathcal{A}, \rho\right)\sqrt{1+k/\beta}}{\delta}\right)$$

$$\mathbf{b}_{\mathrm{r}}(k,\delta) = \widetilde{\mathcal{O}}\left(L+\sqrt{d_1}\right) = \left(\frac{C_2}{2\beta^{3/2}}\sqrt{2\square_1^{\mathrm{r}}(k,\delta)} + \frac{4C_2}{\beta}\right) + 2L_{\mathrm{r}}L\left(1+\sqrt{\log^+(C_1k/\beta)}\right)$$

$$\square_1^{\mathrm{p}}(k,\delta) = \widetilde{\mathcal{O}}\left(d_1\right) = \log\left(\frac{H\mathcal{N}\left(\sigma^2/KH, \mathcal{X}\times\mathcal{A}, \rho\right)\sqrt{1+k/\beta}}{\delta}\right)$$

$$\mathbf{b}_{\mathrm{p}}(k,\delta) = \widetilde{\mathcal{O}}\left(L+\sqrt{d_1}\right) = \left(\frac{C_2}{2\beta^{3/2}}\sqrt{2\square_1^{\mathrm{p}}(k,\delta)} + \frac{4C_2}{\beta}\right) + 2L_{\mathrm{p}}L\left(1+\sqrt{\log^+(C_1k/\beta)}\right).$$

and where d_1 is the covering dimension of $(\mathcal{X}\times\mathcal{A},\rho)$ and, for any $z\in\mathbb{R}$, $\log^+(z) = \log(z+e)$.

Algorithm 3 Kernel Backward Induction with Exploration Bonuses

Input: k, H, Γ, L, β , transitions $(x_h^s, a_h^s, x_{h+1}^s, \widetilde{r}_h^s)_{s=1}^{k-1}$ for all $h\in[H]$.
Initialization: $V_{H+1}(x) = 0$ for all $x\in\mathcal{X}$
for $h = H,\dots,1$ **do**
  for $m = 1,\dots,k-1$ **do**
    // Using weights given in Def.1 and bonus given in Def. 5, compute:
    $\widetilde{Q}_h(x_h^m,a_h^m) = \sum_{s=1}^{k-1}\widetilde{w}_h^{k,s}(x_h^m,a_h^m)\left(\widetilde{r}_h^s + V_{h+1}(x_{h+1}^s)\right) + \mathtt{B}_h^k(x_h^m,a_h^m)$
  end for
  // Interpolated Q-function. Defined, but not computed, for all (x,a)
  $Q_h(x,a) = \min_{m\in[k-1]}\left(\widetilde{Q}_h(x_h^m,a_h^m) + L\rho\left[(x,a),(x_h^m,a_h^m)\right]\right)$
  for $m = 1,\dots,k-1$ **do**
    $V_h(x_h^m) = \min\left(H-h+1, \max_a Q_h(x_h^m,a)\right)$
  end for
end for
Return: $(Q_h)_{h\in[H]}$

B Proof Outline

In this section, we outline the proof of the regret bound of RS-KeRNS (Theorem 2).

B.1 Theorem 2

To prove the regret bound in Theorem 2 for RS-KeRNS, we consider the kernel:

$$\Gamma(t,u,v) = \chi_{(\eta,W)}(t)\phi(u,v), \text{ where } \quad \phi(u,v) \stackrel{\text{def}}{=} \exp\left(-\rho\left[u,v\right]^2/(2\sigma^2)\right),$$

for a given function $\chi_{(\eta,W)} : \mathbb{N} \to [0,1]$. In each episode k, RS-KeRNS has build representative sets of states $\bar{\mathcal{X}}_h^k$, actions $\bar{\mathcal{A}}_h^k$ and next states $\bar{\mathcal{Y}}_h^k$, for each $h \in [H]$. We define of the projections:

$$\zeta_h^k(x,a) \stackrel{\text{def}}{=} \operatorname*{argmin}_{(\overline{x},\overline{a})\in\bar{\mathcal{X}}_h^k\times\bar{\mathcal{A}}_h^k} \rho\left[(x,a),(\overline{x},\overline{a})\right], \quad \overline{\zeta}_h^k(y) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\overline{y}\in\bar{\mathcal{Y}}_h^k} \rho_{\mathcal{X}}(y,\overline{y}).$$

from any (x,a,y) to their representatives.

Let $\breve{W}_h^{k+1}(x,a) = \sum_{s=1}^k \chi(k-s)\phi\left(\zeta_h^{k+1}(x,a),\zeta_h^{s+1}(x_h^s,a_h^s)\right)$. In episode $k+1$, RS-KeRNS computes the following estimate of the rewards

$$\breve{r}_h^{k+1}(x,a) = \frac{1}{\beta + \breve{W}_h^{k+1}(x,a)} \sum_{s=1}^k \chi(k-s)\phi\left(\zeta_h^{k+1}(x,a),\zeta_h^{s+1}(x_h^s,a_h^s)\right)\widetilde{r}_h^s$$

and the following estimate of the transitions

$$\breve{P}_h^{k+1}(y|x,a) = \frac{1}{\beta + \breve{W}_h^{k+1}(x,a)} \sum_{s=1}^k \chi(k-s)\phi\left(\zeta_h^{k+1}(x,a),\zeta_h^{s+1}(x_h^s,a_h^s)\right)\delta_{\overline{\zeta}_h^{s+1}(x_{h+1}^s)}(y).$$

which are similar to the estimates that would be computed by KeRNS, but using the projections ζ and $\overline{\zeta}$ to the representative states and actions. The values of $\breve{r}_h^{k+1}(x,a)$ and $\breve{P}_h^{k+1}(y|x,a)$ are defined for all $(x,a,y) \in \mathcal{X}\times\mathcal{A}\times\mathcal{X}$, but they only need to be stored for $(x,a,y) \in \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k \times \bar{\mathcal{Y}}_h^k$, which corresponds to storing a *finite* representation of the MDP. The exploration bonuses of RS-KeRNS are defined similarly:

$$\breve{\mathrm{B}}_h^{k+1}(x,a) \stackrel{\text{def}}{=} \widetilde{\mathcal{O}}\left(\frac{H}{\sqrt{\beta + \breve{W}_h^{k+1}(x,a)}} + \frac{\beta H}{\beta + \breve{W}_h^{k+1}(x,a)} + L\sigma\right)$$

We prove that the estimates used by RS-KeRNS are close to the ones used by KeRNS up to bias terms. Then, this result is used to prove that the regret bound of RS-KeRNS is the same as KeRNS, but adding a bias term multiplied by the number of episodes. For any (x_h^s,a_h^s) with $s<k$ and $h\in[H]$, we show that (consequence of Lemma 18):

$$\left|\left(\widehat{P}_h^k - \breve{P}_h^k\right)V(x_h^s,a_h^s)\right| \lesssim L\varepsilon_{\mathcal{X}} + 8H\frac{\varepsilon}{\sigma}$$

and similar bounds are obtained for the rewards $\breve{r}_h^k(x,a)$ (Lemma 19) and the exploration bonuses (Lemma 20). This allows us to prove that the regret of RS-KeRNS is bounded as (theorems 5 and 6)

$$\mathcal{R}^{\text{RS}-\text{KeRNS}}(K) \lesssim \mathcal{R}^{\text{KeRNS}}(K) + L(\varepsilon+\varepsilon_{\mathcal{X}})KH^2 + \frac{\varepsilon}{\sigma}KH^3.$$

If we choose $\chi_{(\eta,W)}(t) = \prod_{i=1}^t \alpha_i$, the estimators used by RS-KeRNS can be updated online. Indeed, as detailed in Appendix G, we can related the estimates at time $(k+1,h)$ to the ones at time (k,h):

$$\begin{aligned}
\breve{W}_h^{k+1}(\overline{x},\overline{a}) &= \phi\left((\overline{x},\overline{a}),\zeta_h^{k+1}(x_h^k,a_h^k)\right) + \alpha_{k-s}\breve{W}_h^k(\overline{x},\overline{a}),\\
\breve{r}_h^{k+1}(\overline{x},\overline{a}) &= \frac{\phi\left((\overline{x},\overline{a}),\zeta_h^{k+1}(x_h^k,a_h^k)\right)}{\beta + \breve{W}_h^{k+1}(\overline{x},\overline{a})}\widetilde{r}_h^k + \alpha_{k-s}\left(\frac{\beta + \breve{W}_h^k(\overline{x},\overline{a})}{\beta + \breve{W}_h^{k+1}(\overline{x},\overline{a})}\right)\breve{r}_h^k(\overline{x},\overline{a}), \quad \text{and}\\
\breve{P}_h^{k+1}(y|\overline{x},\overline{a}) &= \frac{\phi\left((\overline{x},\overline{a}),\zeta_h^{k+1}(x_h^k,a_h^k)\right)}{\beta + \breve{W}_h^{k+1}(\overline{x},\overline{a})}\delta_{\overline{\zeta}_h^{k+1}(x_{h+1}^k)}(y) + \alpha_{k-s}\left(\frac{\beta + \breve{W}_h^k(\overline{x},\overline{a})}{\beta + \breve{W}_h^{k+1}(\overline{x},\overline{a})}\right)\breve{P}_h^k(y|\overline{x},\overline{a}).
\end{aligned}$$

One issue that we need to solve is that $\breve{W}_h^k(\overline{x},\overline{a})$, $\breve{r}_h^k(\overline{x},\overline{a})$ and $\breve{P}_h^k(y|\overline{x},\overline{a})$ were not necessarily computed before episode $k+1$. This happens when $(\overline{x},\overline{a})$ is a new representative state-action pair added in episode k. In Section G.2, we show that this can be easily handled by defining some auxiliary quantities that can be updated online and that can be used to initialize the values $\breve{W}_h^k(\overline{x},\overline{a})$, $\breve{r}_h^k(\overline{x},\overline{a})$ and $\breve{P}_h^k(y|\overline{x},\overline{a})$ when necessary, with little overhead to the runtime of the algorithm.

B.2 Optimized Kernel Parameters and Regret Bounds

Table 3: Regret bound for optimized kernel parameters, for $W = \log_\eta\left((1-\eta)/K\right)$.

	σ	$\log\left(\frac{1}{\eta}\right)$	condition	bound	regret
$d=0$	0	$\Delta^{\frac{2}{3}}K^{-\frac{2}{3}}$	$\Delta < K$	$\mathcal{R}_1$	$H^2 X\sqrt{A}\Delta^{\frac{1}{3}}K^{\frac{2}{3}}$
	0	$\Delta^{\frac{2}{3}}K^{-\frac{2}{3}}$	$\Delta < K$	$\mathcal{R}_2$	$H^2\sqrt{XA}\Delta^{\frac{1}{3}}K^{\frac{2}{3}} + H^3X^2A\Delta^{\frac{2}{3}}K^{\frac{1}{3}}$
$d>0$	$\left(\frac{1}{K}\right)^{\frac{1}{2d+3}}$	$\Delta^{\frac{2}{3}}K^{-\frac{2d+2}{2d+3}}$	$\Delta < K^{\frac{3}{2d+3}}$	$\mathcal{R}_1$	$H^2\Delta^{\frac{1}{3}}K^{\frac{2d+2}{2d+3}}$
	$\left(\frac{1}{K}\right)^{\frac{1}{2d+2}}$	$\frac{\Delta^{\frac{1}{2}}}{H}K^{-\frac{2d+1}{2d+2}}$	$\Delta < K^{\frac{1}{d+1}}$	$\mathcal{R}_2$	$H^2\Delta^{\frac{1}{2}}K^{\frac{2d+1}{2d+2}} + H^{\frac{3}{2}}\Delta^{\frac{1}{4}}K^{\frac{3}{4}}$

C Handling the bias due to non-stationarity

Lemma 2 (temporal bias). *Let $(F_h^s)_{h,s}$ be an arbitrary sequence of functions from $\mathcal{X} \times \mathcal{A}$ to $\mathbb{R}$ bounded by M. Then,*

$$\left| \frac{1}{\mathbf{C}_h^k(x,a)} \sum_{s=1}^{k-1} w_h^{k,s}(x,a) \left(F_h^s(x,a) - F_h^k(x,a) \right) \right| \leq \sum_{i=1 \vee (k-W)}^{k-1} \left| F_h^i(x,a) - F_h^{i+1}(x,a) \right| + \frac{2MC_3}{\beta} \frac{\eta^W}{1-\eta} .$$

Proof. The result is straightforward when $k \leq W$. Assuming $k > W$, we have

$$\begin{aligned}
&\frac{1}{\mathbf{C}_h^k(x,a)} \sum_{s=1}^{k-1} w_h^{k,s}(x,a) \left(F_h^s(x,a) - F_h^k(x,a) \right) \\
&= \frac{1}{\mathbf{C}_h^k(x,a)} \sum_{s=k-W}^{k-1} w_h^{k,s}(x,a) \left(F_h^s(x,a) - F_h^k(x,a) \right) + \frac{1}{\mathbf{C}_h^k(x,a)} \sum_{s=1}^{k-W-1} w_h^{k,s}(x,a) \left(F_h^s(x,a) - F_h^k(x,a) \right) \\
&\leq \frac{1}{\mathbf{C}_h^k(x,a)} \sum_{s=k-W}^{k-1} w_h^{k,s}(x,a) \sum_{i=s}^{k-1} \left(F_h^i(x,a) - F_h^{i+1}(x,a) \right) + \frac{2MC_3}{\beta} \sum_{s=1}^{k-W-1} \eta^{k-1-s} \\
&\leq \frac{1}{\mathbf{C}_h^k(x,a)} \sum_{i=k-W}^{k-1} \left(\sum_{s=k-W}^{i} w_h^{k,s}(x,a) \right) \left(F_h^i(x,a) - F_h^{i+1}(x,a) \right) + \frac{2MC_3}{\beta} \frac{\eta^W - \eta^{k-1}}{1-\eta} \\
&\leq \sum_{i=k-W}^{k-1} \left| F_h^i(x,a) - F_h^{i+1}(x,a) \right| + \frac{2MC_3}{\beta} \frac{\eta^W}{1-\eta} ,
\end{aligned}$$

where in the first inequality we used by Assumption 4 that

$$\begin{aligned}
w_h^{k,s}(x,a) &= \Gamma\left(k-s-1, (x,a), (x_h^s, a_h^s)\right) \\
&= \overline{\Gamma}_{(\eta,W)}(k-s-1, \rho\left[(x,a), (x_h^s, a_h^s)\right]) \\
&\leq C_3 \eta^{k-s-1} .
\end{aligned}$$

By symmetry, we obtain

$$\frac{1}{\mathbf{C}_h^k(x,a)} \sum_{s=1}^{k-1} w_h^{k,s}(x,a) \left(F_h^k(x,a) - F_h^s(x,a) \right) \leq \sum_{i=k-W}^{k-1} \left| F_h^i(x,a) - F_h^{i+1}(x,a) \right| + \frac{2MC_3}{\beta} \frac{\eta^W}{1-\eta}.$$

which concludes the proof. □

Definition 6 (temporal bias of the MDP). *The temporal bias at time (k,h) is defined by*

$$\mathbf{bias}(k,h) = \mathbf{bias_r}(k,h) + \mathbf{bias_p}(k,h)$$

where

$$\mathbf{bias_r}(k,h) = \sum_{i=1 \vee (k-W)}^{k-1} \sup_{x,a} \left| r_h^i(x,a) - r_h^{i+1}(x,a) \right| + \frac{2C_3}{\beta} \frac{\eta^W}{1-\eta}$$

$$\mathbf{bias_p}(k,h) = L \sum_{i=1 \vee (k-W)}^{k-1} \sup_{x,a} \mathbb{W}_1 \left(\mathrm{P}_h^i(\cdot|x,a), \mathrm{P}_h^{i+1}(\cdot|x,a) \right) + \frac{2C_3 H}{\beta} \frac{\eta^W}{1-\eta}$$

Definition 7 (Average MDP at episode k)**.** *Let*

$$\overline{r}_h^k(x,a) \stackrel{\text{def}}{=} \sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) r_h^s(x,a) + \frac{\beta}{\mathbf{C}_h^k(x,a)} r_h^k(y|x,a)$$

$$\overline{P}_h^k(y|x,a) \stackrel{\text{def}}{=} \sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) \mathrm{P}_h^s(y|x,a) + \frac{\beta}{\mathbf{C}_h^k(x,a)} \mathrm{P}_h^k(y|x,a)$$

and let $\mathcal{M}_k^{\text{av}}$ *be the MDP with transitions* $\left\{\overline{P}_h^k\right\}_h$ *and rewards* $\left\{\overline{r}_h^k\right\}_h$.

Corollary 2. *Let* $\mathcal{L}(L,H)$ *be the class of* L*-Lipschitz functions from* $\mathcal{X}$ *to* $\mathbb{R}$ *bounded by* H*. Then,*

$$\sup_{x,a} \left|r_h^k(x,a) - \overline{r}_h^k(x,a)\right| \le \mathbf{bias_r}(k,h)$$

$$\sup_{f \in \mathcal{L}(L,H)} \left|\left(\mathrm{P}_h^k - \overline{P}_h^k\right) f(x,a)\right| \le \mathbf{bias_p}(k,h)$$

Proof. For the reward term, we have from Lemma 2:

$$\begin{aligned}
\left|r_h^k(x,a) - \overline{r}_h^k(x,a)\right| &= \left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) \left(r_h^k(x,a) - \overline{r}_h^k(x,a)\right)\right| \\
&\le \sum_{i=1\vee(k-W)}^{k-1} \left|r_h^i(x,a) - r_h^{i+1}(x,a)\right| + \frac{2C_3}{\beta} \frac{\eta^W}{1-\eta} \\
&\le \mathbf{bias_r}(k,h).
\end{aligned}$$

For the transitions term, we also apply Lemma 2 and the definition of the 1-Wasserstein distance:

$$\begin{aligned}
\left|\left(\mathrm{P}_h^k - \overline{P}_h^k\right) f(x,a)\right| &= \left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) \left(\mathrm{P}_h^k f(x,a) - \overline{P}_h^k f(x,a)\right)\right| \\
&\le \sum_{i=1\vee(k-W)}^{k-1} \left|\mathrm{P}_h^i f(x,a) - \mathrm{P}_h^{i+1} f(x,a)\right| + \frac{2C_3 H}{\beta} \frac{\eta^W}{1-\eta} \\
&\le L \sum_{i=1\vee(k-W)}^{k-1} \mathbb{W}_1\left(\mathrm{P}_h^i(\cdot|x,a), \mathrm{P}_h^{i+1}(\cdot|x,a)\right) + \frac{2C_3 H}{\beta} \frac{\eta^W}{1-\eta} \\
&\le \mathbf{bias_p}(k,h)\,.
\end{aligned}$$

Remark 1. *Since the functions in* $\mathcal{L}(L,H)$ *are bounded, the 1-Wasserstein distance could be replaced by the total variation (TV) distance* $\|\mathrm{P}_h^i(\cdot|x,a) - \mathrm{P}_h^{i+1}(\cdot|x,a)\|_1$.

□

D Concentration

In this Section, we provide confidence intervals that will be used to prove our regret bounds. The main concentration results are presented in Lemma 9, which defines an event $\mathcal{G}$ where all the confidence intervals hold, and we show that $\mathbb{P}[\mathcal{G}] \geq 1 - \delta/2$.

D.1 Concentration inequalities for weighted sums

We reproduce here the concentration inequalities for weighted sums proved by Domingues et al. (2020), which we will need.

Lemma 3 (Hoeffding type inequality (Domingues et al., 2020)). *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 $S_t \stackrel{\text{def}}{=} \sum_{s=1}^t w_s Y_s$ and $V_t \stackrel{\text{def}}{=} \sum_{s=1}^t w_s^2$, and assume $w_s \leq 1$ almost surely for all s. Then,for any $\beta > 0$, with probability at least $1-\delta$, for all $t \geq 1$,

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

Proof. See Lemma 2 of Domingues et al. (2020). □

Lemma 4 (Bernstein type inequality (Domingues et al., 2020)). *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*

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

Assume that, for all $t \geq 1$, (i) w_t is $\mathcal{F}_{t-1}$ measurable, (ii) $\mathbb{E}\left[Y_t\middle|\mathcal{F}_{t-1}\right] = 0$, (iii) $w_t \in [0,1]$ almost surely, (iv) there exists $b > 0$ such that $|Y_t| \leq b$ almost surely. Then, for all $\beta > 0$, with probability at least $1-\delta$, for all $t \geq 1$,

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

Proof. See Lemma 3 of Domingues et al. (2020). □

D.2 Hoeffding-type concentration inequalities

Lemma 5. *For all $(x,a,k,h) \in \mathcal{X}\times\mathcal{A}\times[K]\times[H]$, we have*

$$\left|(\widehat{P}_h^k - \overline{P}_h^k)\mathrm{V}_{k,h+1}^*(x,a)\right| \leq \sqrt{\frac{2H^2\square_1^{\mathrm{p}}(k,\delta)}{\mathbf{C}_h^k(x,a)}} + \frac{\beta H}{\mathbf{C}_h^k(x,a)} + \mathbf{b}_{\mathrm{p}}(k,\delta)\sigma$$

with probability at least $1-\delta$, where

$$\square_1^{\mathrm{p}}(k,\delta) = \widetilde{\mathcal{O}}\left(d_1\right) = \log\left(\frac{KH\mathcal{N}\left(\sigma^2/(KH), \mathcal{X}\times\mathcal{A}, \rho\right)\sqrt{1+k/\beta}}{\delta}\right)$$

$$\mathbf{b}_{\mathrm{p}}(k,\delta) = \widetilde{\mathcal{O}}\left(L + \sqrt{d_1}\right) = \left(\frac{C_2}{2\beta^{3/2}}\sqrt{2\square_1^{\mathrm{p}}(k,\delta)} + \frac{4C_2}{\beta}\right) + 2L_{\mathrm{p}}L\left(1 + \sqrt{\log(C_1 k/\beta)}\right)$$

and where d_1 is the covering dimension of $(\mathcal{X}\times\mathcal{A}, \rho)$.

Proof. Let $V = \mathrm{V}^*_{k,h+1}$. For fixed (x, a, h), we have

$$\begin{aligned}
&\left|(\widehat{P}^k_h - \overline{P}^k_h)\mathrm{V}^*_{k,h+1}(x,a)\right| \\
&= \left|\sum_{s=1}^{k-1} \widetilde{w}^{k,s}_h(x,a)\left(V(x^s_{h+1}) - \int_{\mathcal{X}} V(y)\mathrm{dP}^s_h(y|x,a)\right) - \frac{\beta}{\mathbf{C}^k_h(x,a)}\int_{\mathcal{X}} V(y)\mathrm{dP}^k_h(y|x,a)\right| \\
&\le \underbrace{\left|\sum_{s=1}^{k-1} \widetilde{w}^{k,s}_h(x,a)\left(V(x^s_{h+1}) - \int_{\mathcal{X}} V(y)\mathrm{dP}^s_h(y|x^s_h,a^s_h)\right)\right|}_{①} \\
&\quad + \underbrace{\left|\sum_{s=1}^{k-1} \widetilde{w}^{k,s}_h(x,a)\left(\int_{\mathcal{X}} V(y)\mathrm{dP}^s_h(y|x^s_h,a^s_h) - \int_{\mathcal{X}} V(y)\mathrm{dP}^s_h(y|x,a)\right)\right|}_{②} \\
&\quad + \frac{\beta H}{\mathbf{C}^k_h(x,a)}\,.
\end{aligned}$$

Bounding ① (martingale term) Let $Y_s = V(x^s_{h+1}) - \mathrm{P}^s_h V(x^s_h, a^s_h)$. From Lemma 3, we have, for a fixed tuple (x, a, k, h),

$$① = \left|\sum_{s=1}^{k-1} \widetilde{w}^{k,s}_h(x,a) Y_s\right| \le \sqrt{2H^2 \log\left(\frac{\sqrt{1+k/\beta}}{\delta}\right)\frac{1}{\mathbf{C}^k_h(x,a)}}$$

with probability at least $1-\delta$, since $(Y_s)_s$ is a martingale difference sequence with respect to $(\mathcal{F}^s_h)_s$.

From Lemma 24, we verify that the functions

$$(x,a) \mapsto \sqrt{1/\mathbf{C}^k_h(x,a)} \quad \text{and} \quad (x,a) \mapsto \sum_{s=1}^{k-1} \widetilde{w}^{k,s}_h(x,a) Y_s$$

are Lipschitz continuous, with Lipschitz constants bounded by $C_2 k/(2\sigma\beta^{3/2})$ and $4HC_2k/(\beta\sigma)$, respectively. Let $\mathcal{C}_{\mathcal{X}\times\mathcal{A}}\left(\sigma^2/KH\right)$ be a (σ^2/KH)-covering of $\mathcal{X}\times\mathcal{A}$. Using the Lipschitz continuity of the functions above and a union bound over $\mathcal{C}_{\mathcal{X}\times\mathcal{A}}\left(\sigma^2/(KH)\right)$ and over $h \in [H]$, we have

$$\begin{aligned}
① = \left|\sum_{s=1}^{k-1} \widetilde{w}^{k,s}_h(x,a) Y_s\right| &\le \sqrt{2H^2 \log\left(\frac{\sqrt{1+k/\beta}}{\delta}\right)\frac{1}{\mathbf{C}^k_h(x,a)}} \\
&\quad + \left(\frac{C_2 k}{2\sigma\beta^{3/2}}\sqrt{2H^2\log\left(\frac{\sqrt{1+k/\beta}}{\delta}\right)} + \frac{4HC_2k}{\beta\sigma}\right)\frac{\sigma^2}{KH}
\end{aligned}$$

for all (x, a, k, h) with probability at least $1 - \delta KH\mathcal{N}\left(\sigma^2/(KH), \mathcal{X}\times\mathcal{A}, \rho\right)$.

Bounding ② (spatial bias term) We have

$$\begin{aligned}
② &= \left|\sum_{s=1}^{k-1} \widetilde{w}^{k,s}_h(x,a)\left(\int_{\mathcal{X}} V(y)\mathrm{dP}^s_h(y|x^s_h,a^s_h) - \int_{\mathcal{X}} V(y)\mathrm{dP}^s_h(y|x,a)\right)\right| \\
&\le L\sum_{s=1}^{k-1} \widetilde{w}^{k,s}_h(x,a)\mathbb{W}_1\left(\mathrm{P}^s_h(\cdot|x^s_h,a^s_h), \mathrm{P}^s_h(\cdot|x,a)\right) \quad \text{by the definition of } \mathbb{W}_1(\cdot,\cdot) \\
&\le L_{\mathrm{p}} L\sum_{s=1}^{k-1} \widetilde{w}^{k,s}_h(x,a)\rho\left[(x^s_h,a^s_h),(x,a)\right] \quad \text{by Assumption 2} \\
&\le 2\sigma L_{\mathrm{p}} L\left(1 + \sqrt{\log^+(C_1 k/\beta)}\right) \quad \text{by Lemma 23.}
\end{aligned}$$

Putting together the bounds for ① and ② concludes the proof. □

Lemma 6. *For all $(x,a,k,h) \in \mathcal{X}\times\mathcal{A}\times[K]\times[H]$, we have*

$$\left|\widehat{r}_h^k(x,a) - \overline{r}_h^k(x,a)\right| \le \sqrt{\frac{2\square_1^{\mathrm{r}}(k,\delta)}{\mathbf{C}_h^k(x,a)}} + \frac{\beta}{\mathbf{C}_h^k(x,a)} + \mathbf{b}_{\mathrm{r}}(k,\delta)\sigma$$

with probability at least $1-\delta$, where

$$\square_1^{\mathrm{r}}(k,\delta) = \widetilde{\mathcal{O}}\left(d_1\right) = \log\left(\frac{\mathcal{N}\left(\sigma^2/K, \mathcal{X}\times\mathcal{A}, \rho\right)\sqrt{1+k/\beta}}{\delta}\right)$$

$$\mathbf{b}_{\mathrm{r}}(k,\delta) = \widetilde{\mathcal{O}}\left(L + \sqrt{d_1}\right) = \left(\frac{C_2}{2\beta^{3/2}}\sqrt{2\square_1^{\mathrm{r}}(k,\delta)} + \frac{4C_2}{\beta}\right) + 2L_{\mathrm{r}}L\left(1 + \sqrt{\log(C_1 k/\beta)}\right)$$

and where d_1 is the covering dimension of $(\mathcal{X}\times\mathcal{A},\rho)$.

Proof. Almost identical to the proof of Lemma 5, except for the fact that the rewards are bounded by 1 instead of H. □

Lemma 7. *Let $\mathcal{L}(2L,2H)$ be the class of $2L$-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}$ bounded by $2H$. With probability at least $1-\delta$, for all $(x,a,h,k)\in\mathcal{X}\times\mathcal{A}\times[K]\times[H]$ and for all $f\in\mathcal{L}(2L,2H)$, we have*

$$\left|(\widehat{P}_h^k - \overline{P}_h^k)f(x,a)\right| \le \sqrt{\frac{8H^2\square_2^{\mathrm{p}}(k,\delta)}{\mathbf{C}_h^k(x,a)}} + \frac{2\beta H}{\mathbf{C}_h^k(x,a)} + \theta_{\mathrm{b}}^1(k,\delta)\sigma^{1+d_2/2} + \theta_{\mathrm{b}}^2(k,\delta)\sigma$$

where

$$\square_2^{\mathrm{p}}(k,\delta) = \widetilde{\mathcal{O}}\left(|\mathcal{C}'_\sigma| + d_1 d_2\right) = \log\left(\frac{KH\mathcal{N}\left(\sigma^{2+d_2/2}/KH, \mathcal{X}\times\mathcal{A}, \rho\right)\sqrt{1+k/\beta}}{\delta}\left(\frac{2H}{L\sigma}\right)^{\mathcal{N}(\sigma,\mathcal{X},\rho_{\mathcal{X}})}\right)$$

$$\theta_{\mathrm{b}}^1(k,\delta) = \widetilde{\mathcal{O}}\left(\sqrt{|\mathcal{C}'_\sigma|} + \sqrt{d_1 d_2}\right) = \frac{4C_2}{\beta} + \frac{C_2}{2\beta^{3/2}}\sqrt{8\square_2^{\mathrm{p}}(k,\delta)}$$

$$\theta_{\mathrm{b}}^2(k,\delta) = \widetilde{\mathcal{O}}\left(L\right) = 4L_{\mathrm{p}}L\left(1 + \sqrt{\log^+(C_1 k/\beta)}\right) + 32L$$

and where d_1 is the covering dimension of $(\mathcal{X}\times\mathcal{A},\rho)$, d_2 is the covering dimension of $(\mathcal{X},\rho_{\mathcal{X}})$ and $|\mathcal{C}'_\sigma| = \mathcal{N}(\sigma,\mathcal{X},\rho_{\mathcal{X}})$.

Proof. Fix a function $f\in\mathcal{L}(2L,2H)$. Proceeding as in the proof of Lemma 5, we have

$$\begin{aligned}
&\left|(\widehat{P}_h^k - \overline{P}_h^k)f(x,a)\right| \\
&= \left|\sum_{s=1}^{k-1}\widetilde{w}_h^{k,s}(x,a)\left(f(x_{h+1}^s) - \int_{\mathcal{X}} f(y)\mathrm{dP}_h^s(y|x,a)\right) - \frac{\beta}{\mathbf{C}_h^k(x,a)}\int_{\mathcal{X}} f(y)\mathrm{dP}_h^k(y|x,a)\right| \\
&\le \left|\sum_{s=1}^{k-1}\widetilde{w}_h^{k,s}(x,a)Y_s(f)\right| + 4\sigma L_{\mathrm{p}}L\left(1 + \sqrt{\log^+(C_1 k/\beta)}\right) + \frac{2\beta H}{\mathbf{C}_h^k(x,a)}.
\end{aligned}$$

where $Y_s(f) = f(x_{h+1}^s) - \mathrm{P}_h^s f(x_h^s, a_h^s)$.

Now, for fixed (x, a, k, h, f), Lemma 3 gives us

$$\left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) Y_s(f)\right| \leq \sqrt{8H^2 \log\left(\frac{\sqrt{1+k/\beta}}{\delta}\right) \frac{1}{\mathbf{C}_h^k(x,a)}}$$

with probability at least $1-\delta$, since $(Y_s(f))_s$ is a martingale difference sequence with respect to $(\mathcal{F}_h^s)_s$ bounded by $4H$.

Covering $\mathcal{L}(2L, 2H)$ Now let $\mathcal{C}_\mathcal{L}$ be a $8L\sigma$-covering of $(\mathcal{L}(2L,2H), \|\cdot\|_\infty)$. Using the fact that the function $f \mapsto Y_s(f)$ is 2-Lipschitz with respect to $\|\cdot\|_\infty$, we do a union bound over $\mathcal{C}_\mathcal{L}$ to obtain

$$\left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) Y_s(f)\right| \leq \sqrt{8H^2 \log\left(\frac{\sqrt{1+k/\beta}}{\delta}\right) \frac{1}{\mathbf{C}_h^k(x,a)}} + 32L\sigma$$

for all k and all $f \in \mathcal{L}(2L, 2H)$, with probability at least $1 - \delta \left(\frac{2H}{L\sigma}\right)^{\mathcal{N}(\sigma, \mathcal{X}, \rho_\mathcal{X})}$, since the $8L\sigma$-covering number of $\mathcal{C}_\mathcal{L}$ is bounded by $\left(\frac{2H}{L\sigma}\right)^{\mathcal{N}(\sigma, \mathcal{X}, \rho_\mathcal{X})}$, by Lemma 5 of Domingues et al. (2020).

Covering $(\mathcal{X} \times \mathcal{A}, \rho)$ By Lemma 24, the functions

$$(x,a) \mapsto \left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) Y_s(f)\right| \quad \text{and} \quad (x,a) \mapsto \sqrt{\frac{1}{\mathbf{C}_h^k(x,a)}}$$

are $4HC_2k/(\beta\sigma)$-Lipschitz and $C_2k/(2\beta^{3/2}\sigma)$, respectively, with respect to the distance ρ. Let $\mathcal{C}_{\mathcal{X}\times\mathcal{A}}$ be a $\sigma^{2+d_2/2}/KH$ covering of $(\mathcal{X} \times \mathcal{A}, \rho)$. Using the continuity of the functions above, a union bound over $\mathcal{C}_{\mathcal{X}\times\mathcal{A}}$ gives us[9]

$$\begin{aligned}\left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) Y_s(f)\right| \leq{}& \sqrt{8H^2 \log\left(\frac{\sqrt{1+k/\beta}}{\delta}\right) \frac{1}{\mathbf{C}_h^k(x,a)}} + 32L\sigma \\ &+ \frac{\sigma^{2+d_2/2}}{KH}\left(\frac{4HC_2k}{\beta\sigma} + \frac{C_2k}{2\beta^{3/2}\sigma}\sqrt{8H^2 \log\left(\frac{\sqrt{1+k/\beta}}{\delta}\right)}\right)\end{aligned}$$

for all k, all $f \in \mathcal{L}(2L, 2H)$ and all $(x,a) \in \mathcal{X} \times \mathcal{A}$, with probability at least

$$1 - \delta \left(\frac{2H}{L\sigma}\right)^{\mathcal{N}(\sigma, \mathcal{X}, \rho_\mathcal{X})} \mathcal{N}\left(\sigma^{2+d_2/2}/KH, \mathcal{X} \times \mathcal{A}, \rho\right)$$

and a union bound over $(k,h) \in [K] \times [H]$ concludes the proof. □

[9] see, for instance, Lemma 6 of Domingues et al. (2020).

D.3 Bernstein-type concentration inequality

Lemma 8. *Let $\mathcal{L}(2L, 2H)$ be the class of $2L$-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}$ bounded by $2H$. With probability at least $1-\delta$, for all $(x,a,h,k) \in \mathcal{X}\times\mathcal{A}\times[K]\times[H]$ and for all $f \in \mathcal{L}(2L,2H)$, we have*

$$\left|(\widehat{P}_h^k - \overline{P}_h^k) f(x,a)\right| \le \frac{1}{H}\mathrm{P}_h^k|f|(x,a) + \frac{14H^2C_2\square_3(k,\delta)+2\beta H}{\mathbf{C}_h^k(x,a)} + \theta_{\mathrm{b}}^3(k,\delta)\sigma^{1+d_2} + \theta_{\mathrm{b}}^4(k,\delta)\sigma + \frac{2}{H}\,\mathbf{bias_p}(k,h)$$

where d_1 is the covering dimension of $(\mathcal{X}\times\mathcal{A},\rho)$, d_2 is the covering dimension of $(\mathcal{X},\rho_{\mathcal{X}})$ and

$$\square_3(k,\delta) = \widetilde{\mathcal{O}}\left(|\mathcal{C}'_\sigma| + d_1 d_2\right) = \log\left(\frac{4e(2k+1)}{\delta} KH\mathcal{N}\left(\frac{\sigma^{2+d_2}}{H^2K}, \mathcal{X}\times\mathcal{A}, \rho\right)\left(\frac{2H}{L\sigma}\right)^{\mathcal{N}(\sigma,\mathcal{X},\rho_{\mathcal{X}})}\right)$$

$$\theta_{\mathrm{b}}^3(k,\delta) = \widetilde{\mathcal{O}}\left(|\mathcal{C}'_\sigma| + d_1 d_2 + L\sigma\right) = \frac{2L_{\mathrm{p}}L\sigma}{H^2K} + \frac{4C_2}{H\beta} + \frac{14\square_3(k,\delta)C_2}{\beta^2}$$

$$\theta_{\mathrm{b}}^4(k,\delta) = \widetilde{\mathcal{O}}\left(L\right) = 32L + 6L_{\mathrm{p}}L\left(1+\sqrt{\log^+(C_1k/\beta)}\right)$$

where $|\mathcal{C}'_\sigma| = \mathcal{O}\left(1/\sigma^{d_2}\right)$ is the σ-covering number of $(\mathcal{X},\rho_{\mathcal{X}})$.

Proof. We have

$$\begin{aligned}
&\left|(\widehat{P}_h^k - \overline{P}_h^k) f(x,a)\right| \\
&= \left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a)\left(f(x_{h+1}^s) - \mathrm{P}_h^s f(x,a)\right) - \frac{\beta \mathrm{P}_h^k f(x,a)}{\mathbf{C}_h^k(x,a)}\right| \\
&\le \underbrace{\left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a)\left(f(x_{h+1}^s) - \int_{\mathcal{X}} f(y)\mathrm{d}P_h^s(y|x_h^s,a_h^s)\right)\right|}_{①} \\
&\quad + \underbrace{\left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a)\left(\int_{\mathcal{X}} f(y)\mathrm{d}P_h^s(y|x_h^s,a_h^s) - \int_{\mathcal{X}} f(y)\mathrm{d}P_h^s(y|x,a)\right)\right|}_{②} + \frac{2\beta H}{\mathbf{C}_h^k(x,a)}\,.
\end{aligned}$$

Bounding ② (spatial bias term) As in the proof of Lemma 5, we can show that

$$② = \left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a)\left(\int_{\mathcal{X}} f(y)\mathrm{dP}_h^s(y|x_h^s,a_h^s) - \int_{\mathcal{X}} f(y)\mathrm{dP}_h^s(y|x,a)\right)\right| \le 4\sigma L_{\mathrm{p}}L\left(1+\sqrt{\log^+(C_1k/\beta)}\right)$$

Bounding the martingale term (①) with a Bernstein-type inequality Notice that $(x,a) \mapsto \int_{\mathcal{X}} f(y)\mathrm{dP}_h^k(y|x,a)$ is bounded by $2H$ and

$$\mathbb{E}\left[f(x_{h+1}^s)|\mathcal{F}_h^s\right] = \int_{\mathcal{X}} f(y)\mathrm{dP}_h^s(y|x_h^s,a_h^s).$$

The conditional variance of $f(x_{h+1}^s)$ is bounded as follows

$$\begin{aligned}
\mathbb{V}\left[f(x_{h+1}^s)|\mathcal{F}_h^s\right] &= \mathbb{E}\left[f(x_{h+1}^s)^2|\mathcal{F}_h^s\right] - \left(\int_{\mathcal{X}} f(y)\mathrm{dP}_h^s(y|x_h^s,a_h^s)\right)^2 \\
&\le 2H\mathbb{E}\left[\left|f(x_{h+1}^s)\right||\mathcal{F}_h^s\right] \\
&= 2H\int_{\mathcal{X}} |f(y)|\,\mathrm{dP}_h^s(y|x_h^s,a_h^s)
\end{aligned}$$

which we use to bound its weighted average

$$\begin{aligned}
&\frac{1}{\mathbf{C}_h^k(x,a)}\sum_{s=1}^{k-1} w_h^{k,s}(x,a)^2 \mathbb{V}\left[f(x_{h+1}^s)|\mathcal{F}_h^s\right] \\
&\leq \frac{1}{\mathbf{C}_h^k(x,a)}\sum_{s=1}^{k-1} w_h^{k,s}(x,a) \mathbb{V}\left[f(x_{h+1}^s)|\mathcal{F}_h^s\right] \\
&\leq \frac{2H}{\mathbf{C}_h^k(x,a)}\sum_{s=1}^{k-1} w_h^{k,s}(x,a) \int_{\mathcal{X}} |f(y)| \,\mathrm{dP}_h^s(y|x_h^s,a_h^s) \\
&= \frac{2H}{\mathbf{C}_h^k(x,a)}\sum_{s=1}^{k-1} w_h^{k,s}(x,a) \mathrm{P}_h^s |f|(x,a) + \frac{2H}{\mathbf{C}_h^k(x,a)}\sum_{s=1}^{k-1} w_h^{k,s}(x,a) \left(\mathrm{P}_h^s |f|(x_h^s,a_h^s) - \mathrm{P}_h^s |f|(x,a)\right) \\
&\leq 2H\left(\overline{P}_h^k |f|(x,a) - \frac{\beta \mathrm{P}_h^k |f|(x,a)}{\mathbf{C}_h^k(x,a)}\right) + \frac{4HL_{\mathrm{p}}L}{\mathbf{C}_h^k(x,a)}\sum_{s=1}^{k-1} w_h^{k,s}(x,a)\rho\left[(x_h^s,a_h^s),(x,a)\right] \\
&\leq 2H\overline{P}_h^k |f|(x,a) + 8HL_{\mathrm{p}}L\sigma\left(1+\sqrt{\log^+(C_1 k/\beta)}\right)
\end{aligned}$$

where, in the last inequality, we used Lemma 23.

Let $\triangle(k,\delta) = \log\left(4e(2k+1)/\delta\right)$. Let $Y_s(f) = f(x_{h+1}^s) - \mathrm{P}_h^s f(x_h^s,a_h^s)$. By Lemma 4, we have

$$① = \left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) Y_s(f)\right| \leq \sqrt{2\triangle(k,\delta)\frac{\sum_{s=1}^{k-1} w_h^{k,s}(x,a)^2 \mathbb{V}\left[f(x_{h+1}^s)|\mathcal{F}_h^s\right]}{\mathbf{C}_h^k(x,a)^2}} + \frac{10H\triangle(k,\delta)}{\mathbf{C}_h^k(x,a)}$$

with probability at least $1-\delta$, since, for a fixed f, $(Y_s(f))_s$ is a martingale difference sequence with respect to $(\mathcal{F}_h^s)_s$. Using the fact that $\sqrt{uv} \leq (u+v)/2$ for all $u,v>0$,

$$\begin{aligned}
\left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) Y_s(f)\right| &\leq \frac{4H^2\triangle(k,\delta)}{\mathbf{C}_h^k(x,a)} + \frac{1}{4H^2}\frac{\sum_{s=1}^{k-1} w_h^{k,s}(x,a)^2 \mathbb{V}\left[f(x_{h+1}^s)|\mathcal{F}_h^s\right]}{\mathbf{C}_h^k(x,a)} + \frac{10H\triangle(k,\delta)}{\mathbf{C}_h^k(x,a)} \\
&\leq \frac{1}{H}\int_{\mathcal{X}} |f(y)| \,\mathrm{d}\overline{P}_h^k(y|x,a) + \frac{(4H^2+10H)\triangle(k,\delta)}{\mathbf{C}_h^k(x,a)} + \frac{2L_{\mathrm{p}}L\sigma}{H}\left(1+\sqrt{\log^+(C_1 t/\beta)}\right)
\end{aligned}$$

From Corollary 2, we have

$$\int_{\mathcal{X}} |f(y)| \,\mathrm{d}\overline{P}_h^k(y|x,a) = (\overline{P}_h^k - \mathrm{P}_h^k)|f(y)|(x,a) + \mathrm{P}_h^k |f(y)|(x,a) \leq 2\,\mathbf{bias_p}(k,h) + \mathrm{P}_h^k |f(y)|(x,a)$$

which gives us

$$\begin{aligned}
\left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) Y_s(f)\right| &\leq \frac{1}{H}\mathrm{P}_h^k |f(y)|(x,a) + \frac{(4H^2+10H)\triangle(k,\delta)}{\mathbf{C}_h^k(x,a)} \\
&\quad + \frac{2}{H}\,\mathbf{bias_p}(k,h) + \frac{2L_{\mathrm{p}}L\sigma}{H}\left(1+\sqrt{\log^+(C_1 t/\beta)}\right)
\end{aligned}$$

with probability $1-\delta$.

Covering of $\mathcal{X}\times\mathcal{A}$ As a consequence of Assumption 2, the function $(x,a) \mapsto (1/H)\mathrm{P}_h^k |f(y)|(x,a)$ is $2L_{\mathrm{p}}L$-Lipschitz . Also, the functions

$$(x,a) \mapsto \left|\sum_{s=1}^{k-1} \widetilde{w}_h^{k,s}(x,a) Y_s(f)\right| \quad \text{and} \quad (x,a) \mapsto \frac{1}{\mathbf{C}_h^k(x,a)}$$

are $4HC_2k/(\beta\sigma)$-Lipschitz and $C_2k/(\beta^2\sigma)$, respectively, by Lemma 24. Consequently, a union bound over a $(\sigma^{2+d_2}/(H^2K))$-covering of $(\mathcal{X}\times\mathcal{A},\rho)$ and over $[H]$ gives us

$$\begin{aligned}\left|\sum_{s=1}^{k-1}\widetilde{w}_h^{k,s}(x,a)Y_s(f)\right| &\le \frac{1}{H}\mathrm{P}_h^k\left|f(y)\right|(x,a)+\frac{(4H^2+10H)\triangle(k,\delta)}{\mathbf{C}_h^k(x,a)}\\ &+\frac{2}{H}\,\mathbf{bias_p}(k,h)+\frac{2L_{\mathrm{p}}L\sigma}{H}\left(1+\sqrt{\log^+(C_1t/\beta)}\right)\\ &+\left(2L_{\mathrm{p}}L+\frac{4HC_2k}{\beta\sigma}+\frac{(4H^2+10H)\triangle(k,\delta)C_2k}{\beta^2\sigma}\right)\frac{\sigma^{2+d_2}}{H^2K}\end{aligned}$$

for all (x,a,h,k) with probability at least $1-\delta KH\mathcal{N}\left(\frac{\sigma^{2+d_2}}{H^2K},\mathcal{X}\times\mathcal{A},\rho\right)$.

Covering of $\mathcal{L}(2L,2H)$ The bounds for ① and ② give us

$$\begin{aligned}\left|(\widehat{P}_h^k-\overline{P}_h^k)f(x,a)\right| &\le \frac{1}{H}\mathrm{P}_h^k\left|f(y)\right|(x,a)+\frac{(4H^2+10H)\triangle(k,\delta)}{\mathbf{C}_h^k(x,a)}\\ &+\frac{2}{H}\,\mathbf{bias_p}(k,h)+\left(2L_{\mathrm{p}}L+\frac{4HC_2k}{\beta\sigma}+\frac{(4H^2+10H)\triangle(k,\delta)C_2k}{\beta^2\sigma}\right)\frac{\sigma^{2+d_2}}{H^2K}\\ &+6\sigma L_{\mathrm{p}}L\left(1+\sqrt{\log^+(C_1k/\beta)}\right)+\frac{2\beta H}{\mathbf{C}_h^k(x,a)}\,.\end{aligned}$$

The $8L\sigma$-covering number of $\mathcal{L}(2L,2H)$ with respect to the infinity norm is bounded by $(2H/(L\sigma))^{\mathcal{N}(\sigma,\mathcal{X},\rho_{\mathcal{X}})}$, by Lemma 5 of Domingues et al. (2020). The functions $f\mapsto\left|(\mathrm{P}_h^k-\widehat{P}_h^k)f(x,a)\right|$ and $f\mapsto\frac{1}{H}\int_{\mathcal{X}}\left|f(y)\right|\mathrm{d}\overline{P}_h^k(y|x,a)$ are 2-Lipschitz with respect to $\|\cdot\|_\infty$. Consequently, with probability at least

$$1-\delta KH\mathcal{N}\left(\frac{\sigma^{2+d_2}}{H^2K},\mathcal{X}\times\mathcal{A},\rho\right)\left(\frac{2H}{L\sigma}\right)^{\mathcal{N}(\sigma,\mathcal{X},\rho_{\mathcal{X}})},$$

for all $\mathcal{L}(2L,2H)$ and for all (x,a,h,k), we have

$$\begin{aligned}\left|(\widehat{P}_h^k-\overline{P}_h^k)f(x,a)\right| &\le \frac{1}{H}\mathrm{P}_h^k\left|f(y)\right|(x,a)+\frac{(4H^2+10H)\triangle(k,\delta)}{\mathbf{C}_h^k(x,a)}\\ &+\frac{2}{H}\,\mathbf{bias_p}(k,h)+\left(2L_{\mathrm{p}}L+\frac{4HC_2k}{\beta\sigma}+\frac{(4H^2+10H)\triangle(k,\delta)C_2k}{\beta^2\sigma}\right)\frac{\sigma^{2+d_2}}{H^2K}\\ &+6\sigma L_{\mathrm{p}}L\left(1+\sqrt{\log^+(C_1k/\beta)}\right)+\frac{2\beta H}{\mathbf{C}_h^k(x,a)}+32L\sigma\end{aligned}$$

which concludes the proof. □

D.4 Good event

Lemma 9. *Let* $\mathcal{G} = \mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_3 \cap \mathcal{G}_4$, *where*

$$\mathcal{G}_1 \stackrel{\text{def}}{=} \left\{ \forall (x,a,k,h),\ \left|\widehat{r}_h^k(x,a) - \overline{r}_h^k(x,a)\right| \le \sqrt{\frac{2\square_1^{\mathrm{r}}(k,\delta/8)}{\mathbf{C}_h^k(x,a)}} + \frac{\beta}{\mathbf{C}_h^k(x,a)} + \mathbf{b}_{\mathrm{r}}(k,\delta/8)\sigma \right\}$$

$$\mathcal{G}_2 \stackrel{\text{def}}{=} \left\{ \forall (x,a,k,h),\ \left|(\widehat{P}_h^k - \overline{P}_h^k)\mathrm{V}_{k,h+1}^*(x,a)\right| \le \sqrt{\frac{2H^2\square_1^{\mathrm{p}}(k,\delta/8)}{\mathbf{C}_h^k(x,a)}} + \frac{\beta H}{\mathbf{C}_h^k(x,a)} + \mathbf{b}_{\mathrm{p}}(k,\delta/8)\sigma \right\}$$

$$\mathcal{G}_3 \stackrel{\text{def}}{=} \left\{ \forall (x,a,k,h,f),\ \left|(\widehat{P}_h^k - \overline{P}_h^k)f(x,a)\right| \le \sqrt{\frac{2H^2\square_2^{\mathrm{p}}(k,\delta/8)}{\mathbf{C}_h^k(x,a)}} + \frac{\beta H}{\mathbf{C}_h^k(x,a)} + \theta_{\mathrm{b}}^1(k,\delta/8)\sigma^{1+d_2/2} + \theta_{\mathrm{b}}^2(k,\delta/8)\sigma \right\}$$

$$\mathcal{G}_4 \stackrel{\text{def}}{=} \left\{ \forall (x,a,k,h,f),\ \left|(\widehat{P}_h^k - \overline{P}_h^k)f(x,a)\right| \le \frac{1}{H}\mathrm{P}_h^k\left|f(y)\right|(x,a) + \frac{14H^2C_2\square_3(k,\delta/8) + 2\beta H}{\mathbf{C}_h^k(x,a)} + \theta_{\mathrm{b}}^3(k,\delta/8)\sigma^{1+d_2} + \theta_{\mathrm{b}}^4(k,\delta/8)\sigma + \frac{2}{H}\,\mathbf{bias}_{\mathbf{p}}(k,h) \right\}$$

for $(x,a,k,h) \in \mathcal{X}\times\mathcal{A}\times[K]\times[H]$ *and* $f \in \mathcal{L}(2L,2H)$, *and where*

$$\square_1^{\mathrm{p}}(k,\delta) = \widetilde{\mathcal{O}}(d_1), \quad \mathbf{b}_{\mathrm{p}}(k,\delta) = \widetilde{\mathcal{O}}\left(L + \sqrt{d_1}\right), \quad \square_1^{\mathrm{r}}(k,\delta) = \widetilde{\mathcal{O}}(d_1), \quad \mathbf{b}_{\mathrm{r}}(k,\delta) = \widetilde{\mathcal{O}}\left(L + \sqrt{d_1}\right)$$

$$\square_2^{\mathrm{p}}(k,\delta) = \widetilde{\mathcal{O}}(|\mathcal{C}'_\sigma| + d_1d_2), \quad \theta_{\mathrm{b}}^1(k,\delta) = \widetilde{\mathcal{O}}\left(\sqrt{|\mathcal{C}'_\sigma|} + \sqrt{d_1d_2}\right), \quad \theta_{\mathrm{b}}^2(k,\delta) = \widetilde{\mathcal{O}}(L)$$

$$\square_3(k,\delta) = \widetilde{\mathcal{O}}(|\mathcal{C}'_\sigma| + d_1d_2), \quad \theta_{\mathrm{b}}^3(k,\delta) = \widetilde{\mathcal{O}}(|\mathcal{C}'_\sigma| + d_1d_2 + L\sigma), \quad \theta_{\mathrm{b}}^4(k,\delta) = \widetilde{\mathcal{O}}(L)$$

are defined in Lemmas 5, 6, 7 and 8, respectively. Then,

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

Proof. Immediate consequence of Lemmas 5, 6, 7 and 8. □

E Upper bound on true value function

In this section, we show that the true value functions can be upper bounded by the value functions computed by KeRNS plus a bias term. This result will be used to upper bound the regret in the next section.

Lemma 10 (upper bound on Q functions). *On $\mathcal{G}$, for all $(x,a,k,h) \in \mathcal{X} \times \mathcal{A} \times [K] \times [H]$, we have*

$$Q_h^k(x,a) + \sum_{h'=h}^{H} \mathbf{bias}(k,h) \geq \mathrm{Q}_{k,h}^*(x,a)$$

where $\mathbf{bias}(k,h) = \mathbf{bias_r}(k,h) + \mathbf{bias_p}(k,h)$ is the temporal bias of the algorithm at time (k,h) (see Definition 6).

Proof. We proceed by induction on h. For $h = H+1$, both quantities are zero, so the inequality is trivially verified. Now, assume that it is true for $h+1$ and let's prove it for h.

From the induction hypothesis, we have $V_{h+1}^k(x) + \sum_{h'=h+1}^{H} \mathbf{bias}(k,h) \geq \mathrm{V}_{k,h+1}^*(x)$. Indeed,

$$\max_a Q_{h+1}^k(x,a) + \sum_{h'=h+1}^{H} \mathbf{bias}(k,h) \geq \max_a \mathrm{Q}_{k,h}^*(x,a) = \mathrm{V}_{k,h+1}^*(x)$$

and, since $\mathrm{V}_{k,h+1}^*(x) \leq H - h$, we have

$$V_{h+1}^k(x) + \sum_{h'=h+1}^{H} \mathbf{bias}(k,h) = \min\left(H-h, \max_a Q_{h+1}^k(x,a)\right) + \sum_{h'=h+1}^{H} \mathbf{bias}(k,h) \geq \mathrm{V}_{k,h+1}^*(x).$$

From the definition of the algorithm, we have

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

where $\widetilde{Q}_h^k(x,a) = \widehat{r}_h^k(x,a) + \widehat{P}_h^k V_{h+1}^k(x,a) + \mathtt{B}_h^k(x,a)$. Hence,

$$\begin{aligned}
&\widetilde{Q}_h^k(x,a) - \mathrm{Q}_{k,h}^*(x,a) \\
&= \underbrace{\widehat{r}_h^k(x,a) - r_h^k(x,a) + {}^{r}\mathtt{B}_h^k(x,a)}_{(\mathbf{A})} + \underbrace{\widehat{P}_h^k V_{h+1}^k(x,a) - \mathrm{P}_h^k \mathrm{V}_{k,h+1}^*(x,a) + {}^{p}\mathtt{B}_h^k(x,a)}_{(\mathbf{B})}.
\end{aligned}$$

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

$$(\mathbf{A}) = \widehat{r}_h^k(x,a) - \overline{r}_h^k(x,a) + {}^{r}\mathtt{B}_h^k(x,a) + \overline{r}_h^k(x,a) - r_h^k(x,a) \geq -\mathbf{bias_r}(k,h)$$

by Corollary 2 and the fact that $\widehat{r}_h^k(x,a) - \overline{r}_h^k(x,a) + {}^{r}\mathtt{B}_h^k(x,a) \geq 0$ on $\mathcal{G}$.

Similarly, for the term $(\mathbf{B})$, we have

$$\begin{aligned}
(\mathbf{B}) &= \widehat{P}_h^k\left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(x,a) + \left(\widehat{P}_h^k - \overline{P}_h^k\right)\mathrm{V}_{k,h+1}^*(x,a) + \left(\overline{P}_h^k - \mathrm{P}_h^k\right)\mathrm{V}_{k,h+1}^*(x,a) + {}^{p}\mathtt{B}_h^k(x,a) \\
&\geq \widehat{P}_h^k\left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(x,a) - \mathbf{bias_p}(k,h)
\end{aligned}$$

which gives us

$$\begin{aligned}
&\widetilde{Q}_h^k(x,a) - \mathrm{Q}_{k,h}^*(x,a) \\
&\geq \widehat{P}_h^k \left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(x,a) - (\ \mathbf{bias_r}(k,h) + \ \mathbf{bias_p}(k,h)) \\
&= \widehat{P}_h^k \left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(x,a) + \sum_{h'=h+1}^{H} \mathbf{bias}(k,h) - \sum_{h'=h}^{H} \mathbf{bias}(k,h) \\
&\geq \widehat{P}_h^k \left(V_{h+1}^k + \sum_{h'=h+1}^{H} \mathbf{bias}(k,h) - \mathrm{V}_{k,h+1}^*\right)(x,a) - \sum_{h'=h}^{H} \mathbf{bias}(k,h) \\
&\geq -\sum_{h'=h}^{H} \mathbf{bias}(k,h) \quad \text{by the induction hypothesis.}
\end{aligned}$$

Consequently, for all (x,a) and all $s \in [k-1]$, we have

$$\begin{aligned}
\mathrm{Q}_{k,h}^*(x,a) - \sum_{h'=h}^{H} \mathbf{bias}(k,h) &\leq \mathrm{Q}_{k,h}^*(x_h^s,a_h^s) + L\rho\left[(x,a),(x_h^s,a_h^s)\right] - \sum_{h'=h}^{H} \mathbf{bias}(k,h) \\
&\leq \widetilde{Q}_h^k(x_h^s,a_h^s) + L\rho\left[(x,a),(x_h^s,a_h^s)\right]
\end{aligned}$$

since $\mathrm{Q}_{k,h}^*$ is L-Lipschitz . Which implies the result

$$\mathrm{Q}_{k,h}^*(x,a) - \sum_{h'=h}^{H} \mathbf{bias}(k,h) \leq \min_{s\in[k-1]} \left[\widetilde{Q}_h^k(x_h^s,a_h^s) + L\rho\left[(x,a),(x_h^s,a_h^s)\right]\right] = Q_h^k(x,a).$$

□

Corollary 3. *Let $Q_{k,h}^+$ and $V_{k,h}^+$ be defined as as*

$$Q_{k,h}^+(x,a) \stackrel{\text{def}}{=} Q_h^k(x,a) + \sum_{h'=h}^{H} \mathbf{bias}(k,h'), \quad V_{k,h}^+(x) \stackrel{\text{def}}{=} \min\left(H, \max_a Q_{k,h}^+(x,a)\right)$$

Then,

$$\sup_{x\in\mathcal{X}} \left|V_h^k(x) - V_{k,h}^+(x)\right| \leq \sum_{h'=h}^{H} \mathbf{bias}(k,h')$$

and, by Lemma 10, we have $V_{k,h}^+ \geq \mathrm{V}_{k,h}^*$ on the event $\mathcal{G}$.

Proof. For any $x \in \mathcal{X}$,

$$\begin{aligned}
\left|V_h^k(x) - V_{k,h}^+(x)\right| &= \left|\min\left(H, \max_a Q_h^k(x,a)\right) - \min\left(H, \max_a Q_{k,h}^+(x,a)\right)\right| \\
&\leq \left|\max_a Q_h^k(x,a) - \max_a Q_{k,h}^+(x,a)\right| \\
&\leq \max_a \left|Q_h^k(x,a) - Q_{k,h}^+(x,a)\right| \\
&= \sum_{h'=h}^{H} \mathbf{bias}(k,h').
\end{aligned}$$

where we used the fact that, for any $a,b,c \in \mathbb{R}$, we have $|\min(a,b) - \min(a,c)| \leq |b-c|$. □

F Regret bounds

Using the results proved in the previous sections, we are now ready to prove our regret bounds. We first start by proving that the regret is bounded by sums involving $\sqrt{1/\mathbf{C}_h^k(x,a)}$, $1/\mathbf{C}_h^k(x,a)$ and bias terms. Then, we provide upper bounds for these sums, which result in the final regret bounds.

In Theorem 1, we prove two regret bounds, $\mathcal{R}_1$ and $\mathcal{R}_2$. Here, we refer to $\mathcal{R}_1$ as a UCRL-type regret bound and to $\mathcal{R}_2$ as a UCBVI-type bound, due to the technique used to bound the difference between $\widehat{P}_h^k$ and P_h^k. Making an analogy with finite MDPs, in UCRL (Jaksch et al., 2010), a term analogous to $\|\widehat{P}_h - \mathrm{P}_h^k\|_1$ is bounded (as in Lemma 7), whereas in UCBVI (Azar et al., 2017), the term $|(\widehat{P}_h - \mathrm{P}_h^k)\mathrm{V}_{k,h+1}^*|$ is bounded (as in Lemma 5).

Corollary 4. *Let* $\delta_h^k \overset{\text{def}}{=} V_h^k(x_h^k) - \mathrm{V}_{k,h}^{\pi_k}(x_h^k)$. *Then, on* $\mathcal{G}$

$$\mathcal{R}(K) \le \sum_{k=1}^K \delta_1^k + \sum_{k=1}^K \sum_{h=1}^H \mathbf{bias}(k,h).$$

Proof. It follows directly from Lemma 10:

$$\mathcal{R}(K) = \sum_{k=1}^K \mathrm{V}_{k,1}^*(x_1^k) - \mathrm{V}_{k,h}^{\pi_k}(x_1^k) \le \sum_{k=1}^K \left(V_1^k(x_1^k) + \sum_{h=1}^H \mathbf{bias}(k,h) - \mathrm{V}_{k,h}^{\pi_k}(x_1^k) \right).$$

□

Definition 8. *For any* (k,h), *let* $(\widetilde{x}_h^k, \widetilde{a}_h^k)$ *be defined as*

$$(\widetilde{x}_h^k, \widetilde{a}_h^k) \overset{\text{def}}{=} \operatorname*{argmin}_{(x_h^s, a_h^s): s<k} \rho\left[(x_h^k, a_h^k), (x_h^s, a_h^s)\right]$$

that is, the state-action pair in the history that is the closest to (x_h^k, a_h^k).

F.1 Regret bound in terms of the sum of exploration bonuses (UCRL-type)

Lemma 11 (UCRL-type bound with sum of bonuses)**.** *On the event* $\mathcal{G}$, *the regret of* `KeRNS` *is bounded by*

$$\mathcal{R}(K) \lesssim \sum_{k=1}^K \sum_{h=1}^H \left(\frac{H\sqrt{|\mathcal{C}'_\sigma|}}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} \right) \mathbb{I}\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \le 2\sigma\right\} + H^2 |\mathcal{C}_\sigma|$$
$$+ \sum_{k=1}^K \sum_{h=1}^H \widetilde{\xi}_{h+1}^k + H \sum_{k=1}^K \sum_{h=1}^H \mathbf{bias}(k,h) + LKH\sigma$$

where $|\mathcal{C}_\sigma|$ *is the* σ*-covering number of* $(\mathcal{X} \times \mathcal{A}, \rho)$, $|\mathcal{C}'_\sigma|$ *is the* σ*-covering number of* $(\mathcal{X}, \rho_{\mathcal{X}})$ *and* $(\widetilde{\xi}_{h+1}^k)_{k,h}$ *is a martingale difference sequence with respect to* $(\mathcal{F}_h^k)_{k,h}$ *bounded by* $4H$.

Proof. **Regret decomposition** On $\mathcal{G}$, we upper bound δ_h^k using the following decomposition:

$$
\begin{aligned}
\delta_h^k &= V_h^k(x_h^k) - \mathrm{V}_{k,h}^{\pi_k}(x_h^k) \\
&\le Q_h^k(x_h^k, a_h^k) - \mathrm{Q}_{k,h}^{\pi_k}(x_h^k, a_h^k) \\
&\le Q_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathrm{Q}_{k,h}^{\pi_k}(x_h^k, a_h^k) + L\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right], \quad \text{since } Q_h^k \text{ is } L\text{-Lipschitz} \\
&\le \widetilde{Q}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathrm{Q}_{k,h}^{\pi_k}(x_h^k, a_h^k) + L\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right], \quad \text{since } Q_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) \le \widetilde{Q}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) \\
&= \widehat{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(x_h^k, a_h^k) + \widehat{P}_h^k V_{h+1}^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathrm{P}_h^k \mathrm{V}_{k,h+1}^{\pi_k}(x_h^k, a_h^k) + \mathtt{B}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + L\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right] \\
&= \underbrace{\widehat{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(x_h^k, a_h^k)}_{(\mathbf{A})} + \underbrace{\left[\widehat{P}_h^k - \mathrm{P}_h^k\right] \mathrm{V}_{k,h+1}^*(\widetilde{x}_h^k, \widetilde{a}_h^k)}_{(\mathbf{B})} + \underbrace{\left[\widehat{P}_h^k - \mathrm{P}_h^k\right] \left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k)}_{(\mathbf{C})} \\
&\quad + \underbrace{\mathrm{P}_h^k V_{h+1}^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathrm{P}_h^k \mathrm{V}_{k,h+1}^{\pi_k}(x_h^k, a_h^k)}_{(\mathbf{D})} + \mathtt{B}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + 2L\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right].
\end{aligned}
$$

Now, we bound each term (**A**)-(**D**) separately.

Term (**A**):

$$
\begin{aligned}
(\mathbf{A}) &= \widehat{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + r_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(x_h^k, a_h^k) \\
&\le \widehat{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + L_{\mathrm{r}}\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right] \\
&= \widehat{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \overline{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + \overline{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + L_{\mathrm{r}}\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right] \\
&\le {}^{r}\mathtt{B}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + \mathbf{bias_r}(k,h) + L_{\mathrm{r}}\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right]
\end{aligned}
$$

by the definition of $\mathcal{G}$ and Corollary 2.

Term (**B**):

$$
\begin{aligned}
(\mathbf{B}) &= \left[\widehat{P}_h^k - \mathrm{P}_h^k\right] \mathrm{V}_{k,h+1}^*(\widetilde{x}_h^k, \widetilde{a}_h^k) = \left[\widehat{P}_h^k - \overline{P}_h^k\right] \mathrm{V}_{k,h+1}^*(\widetilde{x}_h^k, \widetilde{a}_h^k) + \left[\overline{P}_h^k - \mathrm{P}_h^k\right] \mathrm{V}_{k,h+1}^*(\widetilde{x}_h^k, \widetilde{a}_h^k) \\
&\le {}^{p}\mathtt{B}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + \mathbf{bias_p}(k,h).
\end{aligned}
$$

Term (**C**): Using Corollary 2, we obtain

$$
\begin{aligned}
(\mathbf{C}) &= \left[\widehat{P}_h^k - \mathrm{P}_h^k\right] \left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) \\
&\le \left[\widehat{P}_h^k - \overline{P}_h^k\right] \left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) + 2\,\mathbf{bias_p}(k,h) \\
&\le \sqrt{\frac{8H^2 \square_2^{\mathrm{p}}(k, \delta/8)}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{2\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + \theta_{\mathrm{b}}^1(k, \delta/8)\sigma^{1+d_2/2} + \theta_{\mathrm{b}}^2(k, \delta/8)\sigma + 2\,\mathbf{bias_p}(k,h) \\
&\lesssim \sqrt{\frac{H^2 \left|\mathcal{C}_\sigma'\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + L\sigma + 2\,\mathbf{bias_p}(k,h)
\end{aligned}
$$

by the definition of $\mathcal{G}$.

Term (**D**): From Assumption 2, for any L-Lipschitz function, the mapping $(x,a) \mapsto \mathrm{P}_h^k f(x,a)$ is $L_{\mathrm{p}}L$-Lipschitz . Consequently,

$$
\begin{aligned}
(\mathbf{D}) &= \mathrm{P}_h^k V_{h+1}^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathrm{P}_h^k \mathrm{V}_{k,h+1}^{\pi_k}(x_h^k, a_h^k) \\
&\le \mathrm{P}_h^k V_{h+1}^k(x_h^k, a_h^k) - \mathrm{P}_h^k \mathrm{V}_{k,h+1}^{\pi_k}(x_h^k, a_h^k) + L_{\mathrm{p}}L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \\
&= \mathrm{P}_h^k \left(V_{h+1}^k - \mathrm{V}_{k,h+1}^{\pi_k}\right)(x_h^k, a_h^k) + L_{\mathrm{p}}L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \\
&= \delta_{h+1}^k + \xi_{h+1}^k + L_{\mathrm{p}}L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right]
\end{aligned}
$$

where

$$\xi_{h+1}^k \stackrel{\text{def}}{=} \mathrm{P}_h^k \left(V_{h+1}^k - \mathrm{V}_{k,h+1}^{\pi_k}\right)(x_h^k, a_h^k) - \delta_{h+1}^k$$

is a martingale difference sequence with respect to $(\mathcal{F}_h^k)_{k,h}$ bounded by $4H$.

Putting together the bounds for (**A**)-(**D**) and using the definition of the bonuses $\mathtt{B}_h^k$, we obtain

$$\delta_h^k \lesssim \delta_{h+1}^k + \xi_{h+1}^k + L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] + \sqrt{\frac{H^2\left|\mathcal{C}'_\sigma\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + L\sigma + \mathbf{bias}(k,h)$$

where the constant in front of δ_{h+1}^k is *exact* (i.e., not omitted by $\lesssim$).

Let $E_h^k \stackrel{\text{def}}{=} \left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \le 2\sigma\right\}$. The inequality above implies

$$\mathbb{I}\left\{E_h^k\right\}\delta_h^k \lesssim \mathbb{I}\left\{E_h^k\right\}\delta_{h+1}^k + \mathbb{I}\left\{E_h^k\right\}\left(\xi_{h+1}^k + \sqrt{\frac{H^2\left|\mathcal{C}'_\sigma\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}\right) + 3L\sigma + \mathbf{bias}(k,h). \tag{2}$$

Now, we bound $\mathbb{I}\left\{E_h^k\right\}\delta_{h+1}^k$ in terms of δ_{h+1}^k, which will be later used to bound δ_h^k in terms of δ_{h+1}^k. On $\mathcal{G}$, we have

$$\begin{aligned}
\mathbb{I}\left\{E_h^k\right\}\delta_{h+1}^k &= \mathbb{I}\left\{E_h^k\right\}\left(V_{h+1}^k(x_{h+1}^k) - \mathrm{V}_{k,h+1}^{\pi_k}(x_{h+1}^k)\right) \\
&= \mathbb{I}\left\{E_h^k\right\}\left(\underbrace{V_{h+1}^k(x_{h+1}^k) + \sum_{h'=h+1}^H \mathbf{bias}(k,h') - \mathrm{V}_{k,h+1}^{\pi_k}(x_{h+1}^k)}_{\ge 0 \text{ by Lemma 10}} - \sum_{h'=h+1}^H \mathbf{bias}(k,h')\right) \\
&\le V_{h+1}^k(x_{h+1}^k) + \sum_{h'=h+1}^H \mathbf{bias}(k,h') - \mathrm{V}_{k,h+1}^{\pi_k}(x_{h+1}^k) - \mathbb{I}\left\{E_h^k\right\}\sum_{h'=h+1}^H \mathbf{bias}(k,h') \\
&= \delta_{h+1}^k + \sum_{h'=h+1}^H \mathbf{bias}(k,h') - \mathbb{I}\left\{E_h^k\right\}\sum_{h'=h+1}^H \mathbf{bias}(k,h') \\
&\le \delta_{h+1}^k + \sum_{h'=h+1}^H \mathbf{bias}(k,h').
\end{aligned}$$

The inequality above, combined with (2) yields

$$\mathbb{I}\left\{E_h^k\right\}\delta_h^k \lesssim \delta_{h+1}^k + \sum_{h'=h}^H \mathbf{bias}(k,h') + \mathbb{I}\left\{E_h^k\right\}\left(\xi_{h+1}^k + \sqrt{\frac{H^2\left|\mathcal{C}'_\sigma\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}\right) + 3L\sigma.$$

Let $\overline{E}_h^k$ be the complement of E_h^k. Since $\delta_h^k \le H$, we have

$$\begin{aligned}
\delta_h^k &= \mathbb{I}\left\{\overline{E}_h^k\right\}\delta_h^k + \mathbb{I}\left\{E_h^k\right\}\delta_h^k \\
&\le H\mathbb{I}\left\{\overline{E}_h^k\right\} + \mathbb{I}\left\{E_h^k\right\}\delta_h^k \\
&\lesssim H\mathbb{I}\left\{\overline{E}_h^k\right\} + \delta_{h+1}^k + \sum_{h'=h}^H \mathbf{bias}(k,h') + \mathbb{I}\left\{E_h^k\right\}\left(\xi_{h+1}^k + \sqrt{\frac{H^2\left|\mathcal{C}'_\sigma\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}\right) + L\sigma.
\end{aligned}$$

This yields

$$\delta_1^k \lesssim \sum_{h=1}^{H} \mathbb{I}\left\{E_h^k\right\} \sqrt{\frac{H^2 \left|\mathcal{C}'_\sigma\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \sum_{h=1}^{H} \mathbb{I}\left\{E_h^k\right\} \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + \sum_{h=1}^{H} \mathbb{I}\left\{E_h^k\right\} \xi_{h+1}^k + H \sum_{h=1}^{H} \mathbf{bias}(k,h) + H \sum_{h=1}^{H} \mathbb{I}\left\{\overline{E}_h^k\right\} + HL\sigma$$

Using Corollary 4, we obtain

$$\begin{aligned}\mathcal{R}(K) &\leq \sum_{k=1}^{K} \delta_1^k + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k,h) \\ &\lesssim \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{I}\left\{E_h^k\right\} \sqrt{\frac{H^2 \left|\mathcal{C}'_\sigma\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{I}\left\{E_h^k\right\} \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} \\ &+ \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{I}\left\{E_h^k\right\} \xi_{h+1}^k + H \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k,h) + H \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{I}\left\{\overline{E}_h^k\right\} + KHL\sigma\,.\end{aligned}$$

For each h, the number of episodes k where the event $\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] > 2\sigma\right\}$ occurs is bounded by $|\mathcal{C}_\sigma|$. Hence, we can bound the sum

$$H \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{I}\left\{\overline{E}_h^k\right\} = H \sum_{h=1}^{H}\sum_{k=1}^{K} \mathbb{I}\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] > 2\sigma\right\} \leq H^2 \left|\mathcal{C}_\sigma\right|.$$

We conclude the proof by recalling the definition $E_h^k \stackrel{\text{def}}{=} \left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \leq 2\sigma\right\}$ and using the fact that $\widetilde{\xi}_{h+1}^k \stackrel{\text{def}}{=} \mathbb{I}\left\{E_h^k\right\} \xi_{h+1}^k$ is a martingale difference sequence with respect to $(\mathcal{F}_h^k)_{k,h}$ bounded by $4H$. □

F.2 Regret bound in terms of the sum of exploration bonuses (UCBVI-type)

Lemma 12 (UCBVI-type bound with sum of bonuses)**.** *In the event $\mathcal{G}$, the regret of KeRNS is bounded by*

$$\begin{aligned}\mathcal{R}(K) \lesssim &\sum_{k=1}^{K}\sum_{h=1}^{H} \left(\frac{H}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{H^2 \left|\mathcal{C}'_\sigma\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}\right) \mathbb{I}\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \leq 2\sigma\right\} + H^2 \left|\mathcal{C}_\sigma\right| \\ &+ \sum_{k=1}^{K}\sum_{h=1}^{H} \left(1 + \frac{1}{H}\right)^h \widetilde{\xi}_{h+1}^k + H \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k,h) + LKH\sigma\end{aligned}$$

where $|\mathcal{C}_\sigma|$ is the σ-covering number of $(\mathcal{X} \times \mathcal{A}, \rho)$, $|\mathcal{C}'_\sigma|$ is the σ-covering number of $(\mathcal{X}, \rho_\mathcal{X})$ and $(\widetilde{\xi}_{h+1}^k)_{k,h}$ is a martingale difference sequence with respect to $(\mathcal{F}_h^k)_{k,h}$ bounded by $4H$.

Proof. The proof follows the one of Proposition 5 of Domingues et al. (2020). The key difference is that we need to handle the temporal bias. In particular, V_h^k is not an upper bound on $\mathrm{V}_{k,h+1}^*$ due to the temporal bias, which makes our proof slightly more technical by introducing $V_{k,h}^+$ (see Cor. 3) when applying the Bernstein-type concentration of Lemma 8.

Regret decomposition We use the same regret decomposition as in the proof of Lemma 11. The terms (**A**), (**B**) and (**D**) are bounded in the same way, but we handle the term (**C**) differently.

Term (**C**): To bound this term, we use corollaries 2 and 3:

$$
\begin{aligned}
(\mathbf{C}) &= \left[\widehat{P}_h^k - \mathrm{P}_h^k\right]\left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) \\
&= \left[\widehat{P}_h^k - \mathrm{P}_h^k\right]\left(V_{k,h+1}^+ - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) + \left[\widehat{P}_h^k - \mathrm{P}_h^k\right]\left(V_{h+1}^k - V_{k,h+1}^+\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) \\
&\le \left[\widehat{P}_h^k - \mathrm{P}_h^k\right]\left(V_{k,h+1}^+ - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) + 2\sum_{h'=h}^{H} \mathbf{bias}(k,h'), \quad \text{by Cor. 3} \\
&= \left[\widehat{P}_h^k - \overline{P}_h^k\right]\left(V_{k,h+1}^+ - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) + \left[\overline{P}_h^k - \mathrm{P}_h^k\right]\left(V_{k,h+1}^+ - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) + 2\sum_{h'=h}^{H} \mathbf{bias}(k,h') \\
&\le \left[\widehat{P}_h^k - \overline{P}_h^k\right]\left(V_{k,h+1}^+ - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) + 2\,\mathbf{bias_p}(k,h) + 2\sum_{h'=h}^{H} \mathbf{bias}(k,h'), \quad \text{by Cor. 2} \\
&\le \frac{1}{H}\mathrm{P}_h^k\left(V_{k,h+1}^+ - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) + \frac{14H^2C_2\square_3(k,\delta/8) + 2\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} \\
&\quad + \theta_{\mathrm{b}}^3(k,\delta/8)\sigma^{1+d_2} + \theta_{\mathrm{b}}^4(k,\delta/8)\sigma + \frac{2}{H}\,\mathbf{bias_p}(k,h) + 2\,\mathbf{bias_p}(k,h) + 2\sum_{h'=h}^{H} \mathbf{bias}(k,h') \\
&\le \frac{1}{H}\mathrm{P}_h^k\left(V_{k,h+1}^+ - \mathrm{V}_{k,h+1}^*\right)(x_h^k, a_h^k) + \frac{14H^2C_2\square_3(k,\delta/8) + 2\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + \frac{2L_{\mathrm{p}}L}{H}\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \\
&\quad + \theta_{\mathrm{b}}^3(k,\delta/8)\sigma^{1+d_2} + \theta_{\mathrm{b}}^4(k,\delta/8)\sigma + \frac{2}{H}\,\mathbf{bias_p}(k,h) + 2\,\mathbf{bias_p}(k,h) + 2\sum_{h'=h}^{H} \mathbf{bias}(k,h'),
\end{aligned}
$$

where we also used the definition of $\mathcal{G}$ and the fact that the function $(x,a) \mapsto \mathrm{P}_h^k\left(V_{k,h+1}^+ - \mathrm{V}_{k,h+1}^*\right)(x,a)$ is $2L_{\mathrm{p}}L$-Lipschitz , from Assumption 2. Now, since

$$
\square_3(k,\delta) = \widetilde{\mathcal{O}}\left(|\mathcal{C}'_\sigma| + d_1 d_2\right), \quad \theta_{\mathrm{b}}^3(k,\delta) = \widetilde{\mathcal{O}}\left(|\mathcal{C}'_\sigma| + d_1 d_2 + L\sigma\right), \quad \theta_{\mathrm{b}}^4(k,\delta) = \widetilde{\mathcal{O}}\left(L\right)
$$

and $|\mathcal{C}'_\sigma| = \mathcal{O}\left(1/\sigma^{d_2}\right)$, we have

$$
\begin{aligned}
(\mathbf{C}) \lesssim & \frac{1}{H}\mathrm{P}_h^k\left(V_{k,h+1}^+ - \mathrm{V}_{k,h+1}^*\right)(x_h^k, a_h^k) + \frac{H^2\,|\mathcal{C}'_\sigma|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + L\sigma \\
& + \left(2 + \frac{2}{H}\right)\,\mathbf{bias_p}(k,h) + 2\sum_{h'=h}^{H} \mathbf{bias}(k,h') + \frac{2L_{\mathrm{p}}L}{H}\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right].
\end{aligned}
$$

Using again Corollary 3, we have

$$
\frac{1}{H}\mathrm{P}_h^k\left(V_{k,h+1}^+ - \mathrm{V}_{k,h+1}^*\right)(x_h^k, a_h^k) \le \frac{1}{H}\mathrm{P}_h^k\left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(x_h^k, a_h^k) + \frac{1}{H}\sum_{h'=h}^{H} \mathbf{bias}(k,h')
$$

which gives us, since $\mathrm{V}_{k,h+1}^{\pi_k} \le \mathrm{V}_{k,h+1}^*$,

$$
\begin{aligned}
(\mathbf{C}) &\lesssim \frac{1}{H}\mathrm{P}_h^k\left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(x_h^k, a_h^k) + \frac{H^2\,|\mathcal{C}'_\sigma|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + L\sigma + \sum_{h'=h}^{H} \mathbf{bias}(k,h') + L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \\
&\lesssim \frac{1}{H}\mathrm{P}_h^k\left(V_{h+1}^k - \mathrm{V}_{k,h+1}^{\pi_k}\right)(x_h^k, a_h^k) + \frac{H^2\,|\mathcal{C}'_\sigma|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + L\sigma + \sum_{h'=h}^{H} \mathbf{bias}(k,h') + L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right]
\end{aligned}
$$

where we omit constants. Notice, however, that there are no constants omitted in the term $\frac{1}{H}\mathrm{P}_h^k\left(V_{h+1}^k - \mathrm{V}_{k,h+1}^{\pi_k}\right)(x_h^k, a_h^k)$.

Putting together the bounds for (**A**)-(**D**), we obtain

$$\delta_h^k \lesssim \left(1+\frac{1}{H}\right)\left(\delta_{h+1}^k+\xi_{h+1}^k\right)+L\rho\left[(x_h^k,a_h^k),(\widetilde{x}_h^k,\widetilde{a}_h^k)\right]+2\,\mathtt{B}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)+\frac{H^2\,|\mathcal{C}'_\sigma|}{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}+L\sigma+\sum_{h'=h}^{H}\mathbf{bias}(k,h')$$

where the constant in front of δ_{h+1}^k is *exact* (i.e., not omitted by $\lesssim$).

Let $E_h^k \stackrel{\text{def}}{=} \left\{\rho\left[(x_h^k,a_h^k),(\widetilde{x}_h^k,\widetilde{a}_h^k)\right]\le 2\sigma\right\}$. Using the definition of the bonus

$$\mathtt{B}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)\lesssim\frac{H}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}}+\frac{\beta H}{\mathbf{C}_h(\widetilde{x}_h^k,\widetilde{a}_h^k)}+L\sigma\,,$$

and the same argument as in the proof of Lemma 11, we obtain

$$\begin{aligned}\mathcal{R}(K)&\le\sum_{k=1}^{K}\delta_1^k+\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{bias}(k,h)\\&\lesssim\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}\left\{E_h^k\right\}\sqrt{\frac{H^2}{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}}+\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}\left\{E_h^k\right\}\frac{H^2\,|\mathcal{C}'_\sigma|}{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}\\&+\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}\left\{E_h^k\right\}\xi_{h+1}^k+H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{bias}(k,h)+H^2\,|\mathcal{C}_\sigma|+KHL\sigma.\end{aligned}$$

As in Lemma 11, we conclude the proof by recalling the definition $E_h^k \stackrel{\text{def}}{=} \left\{\rho\left[(x_h^k,a_h^k),(\widetilde{x}_h^k,\widetilde{a}_h^k)\right]\le 2\sigma\right\}$ and using the fact that $\widetilde{\xi}_{h+1}^k \stackrel{\text{def}}{=} \mathbb{I}\left\{E_h^k\right\}\xi_{h+1}^k$ is a martingale difference sequence with respect to $(\mathcal{F}_h^k)_{k,h}$ bounded by $4H$. □

F.3 Bounding the sum of bonuses and bias

Lemma 13. *Let $(\mu_i)_{i\ge}$ be a sequence of non-negative numbers. Then,*

$$\sum_{k=1}^{K}\sum_{i=1\vee(k-W)}^{k-1}\mu_i\le 2W\sum_{i=1}^{K}\mu_i.$$

Proof. We have

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1\vee(k-W)}^{k-1}\mu_i&=\sum_{k=1}^{W}\sum_{i=1}^{k-1}\mu_i+\sum_{k=W+1}^{K}\sum_{i=k-W}^{k-1}\mu_i\le W\sum_{i=1}^{K}\mu_i+\sum_{i=1}^{K-1}\sum_{k=i+1}^{i+W}\mu_i\\&\le W\sum_{i=1}^{K}\mu_i+W\sum_{i=1}^{K}\mu_i=2W\sum_{i=1}^{K}\mu_i.\end{aligned}$$

□

Corollary 5 (bound on the temporal bias)**.** *Let Δ^{r} and Δ^{p} be the variation of the MDP over KH time steps,*

$$\Delta^{\mathrm{r}}\stackrel{\text{def}}{=}\sum_{i=1}^{K}\sum_{h=1}^{H}\sup_{x,a}\left|r_h^i(x,a)-r_h^{i+1}(x,a)\right|,\quad\Delta^{\mathrm{p}}\stackrel{\text{def}}{=}\sum_{i=1}^{K}\sum_{h=1}^{H}\sup_{x,a}\mathbb{W}_1\left(\mathrm{P}_h^i(\cdot|x,a),\mathrm{P}_h^{i+1}(\cdot|x,a)\right).$$

Then,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{bias}(h,k)\le 2W\left(\Delta^{\mathrm{r}}+L\Delta^{\mathrm{p}}\right)+\frac{2C_3(H+1)KH}{\beta}\frac{\eta^W}{1-\eta}\,.$$

Proof. From Lemma 13, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{bias_p}(h,k) = \sum_{k=1}^{K}\sum_{h=1}^{H} L \sum_{i=1\vee(k-W)}^{k-1} \sup_{x,a} \mathbb{W}_1\left(\mathrm{P}_h^i(\cdot|x,a), \mathrm{P}_h^{i+1}(\cdot|x,a)\right) + \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{2C_3H}{\beta}\frac{\eta^W}{1-\eta}$$
$$\leq 2WL\sum_{h=1}^{H}\sum_{i=1}^{K}\sup_{x,a}\mathbb{W}_1\left(\mathrm{P}_h^i(\cdot|x,a), \mathrm{P}_h^{i+1}(\cdot|x,a)\right) + \frac{2C_3KH^2}{\beta}\frac{\eta^W}{1-\eta}.$$

The sum $\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{bias_r}(h,k)$ is bounded in the same way, which concludes the proof. □

Lemma 14 (bounding sum on sliding window)**.** *Let $\{a_n\}_{n\geq 1}$ be a sequence of real numbers such that $0 \leq a \leq c$ for some constant $c > 0$. Let $A_t = \sum_{n=1\vee(t-W)}^{t-1} a_n$. Then, for any $p, b > 0$,*

$$\sum_{t=1}^{T}\frac{a_t}{(1+bA_t)^p} \leq \sum_{n=1}^{\lceil T/W\rceil}\left(c + \int_0^{A_{nW+1}-c}\frac{1}{(1+bz)^p}\mathrm{d}z\right).$$

Proof. We have

$$\sum_{t=1}^{T}\frac{a_t}{(1+bA_t)^p} = \underbrace{\sum_{t=1}^{W}\frac{a_t}{(1+bA_t)^p}}_{①} + \underbrace{\sum_{t=W+1}^{T}\frac{a_t}{(1+bA_t)^p}}_{②}.$$

By Lemma 9 of Domingues et al. (2020), we have

$$① = \sum_{t=1}^{W}\frac{a_t}{(1+bA_t)^p} \leq c + \int_0^{A_{W+1}-c}\frac{1}{(1+bz)^p}\mathrm{d}z.$$

Now, we handle ②:

$$② \leq \sum_{n=1}^{\lceil T/W\rceil-1}\sum_{t=nW+1}^{(n+1)W}\frac{a_t}{(1+bA_t)^p} = \sum_{n=1}^{\lceil T/W\rceil-1}\sum_{l=1}^{W}\frac{a_{l+nW}}{(1+bA_{l+nW})^p}$$
$$\leq \sum_{n=1}^{\lceil T/W\rceil-1}\left(c + \int_0^{A_{(n+1)W+1}-c}\frac{1}{(1+bz)^p}\mathrm{d}z\right).$$

□

Definition 9. *Consider a σ-covering of $(\mathcal{X}\times\mathcal{A}, \rho)$, $\mathcal{C}_\sigma = \{(x_j, a_j) \in \mathcal{X}\times\mathcal{A},\ j = 1, \ldots, |\mathcal{C}_\sigma|\}$. We define a partition $\{B_j\}_{j\in[|\mathcal{C}_\sigma|]}$ such that*

$$B_j = \left\{(x,a)\in\mathcal{X}\times\mathcal{A} : (x_j,a_j) = \operatorname*{argmin}_{(x_i,a_i)\in\mathcal{C}_\sigma}\rho\left[(x,a),(x_i,a_i)\right]\right\}$$

with ties broken arbitrarily.

Lemma 15. *Let $U_\eta \overset{\text{def}}{=} \lceil 1/\log(1/\eta)\rceil$ and*

$$\mathbf{N}_h^k(B_j, U_\eta) \overset{\text{def}}{=} \sum_{s=1\vee(k-U_\eta)}^{k-1}\mathbb{I}\left\{(x_h^s, a_h^s)\in B_j\right\}.$$

If $U_\eta \leq W$, $\rho\left[(x_h^k,a_h^k),(\widetilde{x}_h^k,\widetilde{a}_h^k)\right] \leq 2\sigma$ and $(x_h^k,a_h^k)\in B_j$ then

$$\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k) \geq \beta + G(4)e^{-1}\mathbf{N}_h^k(B_j,U_\eta).$$

Proof. This result is based on the proof of Proposition 6 of Domingues et al. (2020), which we generalize to the case where the kernel is time-dependent. From Assumption 4, we have $\overline{\Gamma}_{(\sigma,\eta,W)}(k-1-s,z) \geq G(z)\eta^{k-1-s}$ for all $s \geq k-W$. Consequently, if $U_\eta \leq W$, for all $s \geq k-U_\eta$:

$$\overline{\Gamma}_{(\sigma,\eta,W)}(k-1-s,z) \geq G(z)\eta^{k-1-s} \geq G(z)\eta^{U_\eta} \geq G(z)\exp(-1). \tag{3}$$

Also, if $\rho\left[(x_h^k,a_h^k),(\widetilde{x}_h^k,\widetilde{a}_h^k)\right] \leq 2\sigma$, $(x_h^k,a_h^k) \in B_j$, and $(x_h^s,a_h^s) \in B_j$, we have

$$\begin{aligned}\rho\left[(\widetilde{x}_h^k,\widetilde{a}_h^k),(x_h^s,a_h^s)\right] &\leq \rho\left[(\widetilde{x}_h^k,\widetilde{a}_h^k),(x_h^k,a_h^k)\right] + \rho\left[(x_h^k,a_h^k),(x_h^s,a_h^s)\right] \\ &\leq 2\sigma + \rho\left[(x_h^k,a_h^k),(x_j,a_j)\right] + \rho\left[(x_j,a_j),(x_h^s,a_h^s)\right] \leq 4\sigma.\end{aligned} \tag{4}$$

By Assumption 4, the function $z \mapsto \overline{\Gamma}_{(\eta,W)}(t,z)$ is non-increasing. Together with (3) and (4), this yields:

$$\begin{aligned}\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k) &= \beta + \sum_{s=1}^{k-1}\overline{\Gamma}_{(\sigma,\eta,W)}\left(k-1-s,\frac{\rho\left[(\widetilde{x}_h^k,\widetilde{a}_h^k),(x_h^s,a_h^s)\right]}{\sigma}\right) \\ &\geq \beta + \sum_{s=1}^{k-1}\overline{\Gamma}_{(\sigma,\eta,W)}\left(k-1-s,\frac{\rho\left[(\widetilde{x}_h^k,\widetilde{a}_h^k),(x_h^s,a_h^s)\right]}{\sigma}\right)\mathbb{I}\left\{(x_h^s,a_h^s)\in B_j\right\} \\ &\geq \beta + \sum_{s=1\vee(k-U_\eta)}^{k-1}\overline{\Gamma}_{(\sigma,\eta,W)}\left(k-1-s,4\right)\mathbb{I}\left\{(x_h^s,a_h^s)\in B_j\right\} \\ &\geq \beta + G(4)e^{-1}\sum_{s=k-U_\eta}^{k-1}\mathbb{I}\left\{(x_h^s,a_h^s)\in B_j\right\},\end{aligned}$$

which concludes the proof. □

Lemma 16. *Let $U_\eta = \lceil 1/\log(1/\eta)\rceil$. If $U_\eta \leq W$, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}}\mathbb{I}\left\{\rho\left[(x_h^k,a_h^k),(\widetilde{x}_h^k,\widetilde{a}_h^k)\right]\leq 2\sigma\right\} \lesssim H\left\lceil\frac{K}{U_\eta}\right\rceil\left(|\mathcal{C}_\sigma| + \sqrt{|\mathcal{C}_\sigma|\,U_\eta}\right)$$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}\mathbb{I}\left\{\rho\left[(x_h^k,a_h^k),(\widetilde{x}_h^k,\widetilde{a}_h^k)\right]\leq 2\sigma\right\} \lesssim H\,|\mathcal{C}_\sigma|\left\lceil\frac{K}{U_\eta}\right\rceil.$$

Proof. The proof relies on Lemmas 14 and 15.

Here, we define the constant c as $c = G(4)\beta^{-1}e^{-1} > 0$, since $G(4) > 0$ by Assumption 4.

Bounding the sum $\sum_k 1/\sqrt{\mathbf{C}_h^k}$

$$
\begin{aligned}
&\sum_{k=1}^{K} \frac{1}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} \mathbb{I}\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \leq 2\sigma\right\} \\
&= \sum_{j=1}^{|\mathcal{C}_\sigma|} \sum_{k=1}^{K} \frac{1}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} \mathbb{I}\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \leq 2\sigma\right\} \mathbb{I}\left\{(x_h^k, a_h^k) \in B_j\right\} \\
&\leq \beta^{-1/2} \sum_{j=1}^{|\mathcal{C}_\sigma|} \sum_{k=1}^{K} \frac{\mathbb{I}\left\{(x_h^k, a_h^k) \in B_j\right\}}{\sqrt{1 + c\mathbf{N}_h^k(B_j, U_\eta)}}, \quad \text{by Lemma 15} \\
&\leq \beta^{-1/2} \sum_{j=1}^{|\mathcal{C}_\sigma|} \sum_{n=1}^{\lceil K/U_\eta \rceil} \left(1 + \int_0^{\mathbf{N}_h^{nU_\eta+1}(B_j, U_\eta)} \frac{1}{\sqrt{1+cz}} \mathrm{d}z\right), \quad \text{by Lemma 14} \\
&= \beta^{-1/2} |\mathcal{C}_\sigma| \left\lceil \frac{K}{U_\eta} \right\rceil + \frac{2\beta^{-1/2}}{c} \sum_{n=1}^{\lceil K/U_\eta \rceil} \sum_{j=1}^{|\mathcal{C}_\sigma|} \sqrt{1 + c\mathbf{N}_h^{nU_\eta+1}(B_j, U_\eta)} \\
&\leq \beta^{-1/2} |\mathcal{C}_\sigma| \left\lceil \frac{K}{U_\eta} \right\rceil + \frac{2\beta^{-1/2}}{c} \sum_{n=1}^{\lceil K/U_\eta \rceil} \sqrt{|\mathcal{C}_\sigma|} \sqrt{|\mathcal{C}_\sigma| + c \sum_{j=1}^{|\mathcal{C}_\sigma|} \mathbf{N}_h^{nU_\eta+1}(B_j, U_\eta)}, \quad \text{by Cauchy-Schwarz inequality} \\
&\leq \beta^{-1/2} |\mathcal{C}_\sigma| \left\lceil \frac{K}{U_\eta} \right\rceil + \frac{2\beta^{-1/2}}{c} \sum_{n=1}^{\lceil K/U_\eta \rceil} \sqrt{|\mathcal{C}_\sigma|} \sqrt{|\mathcal{C}_\sigma| + cU_\eta} \lesssim |\mathcal{C}_\sigma| \left\lceil \frac{K}{U_\eta} \right\rceil + \sqrt{|\mathcal{C}_\sigma| \, U_\eta} \left\lceil \frac{K}{U_\eta} \right\rceil .
\end{aligned}
$$

Bounding the sum $\sum_k 1/\mathbf{C}_h^k$

$$
\begin{aligned}
&\sum_{k=1}^{K} \frac{1}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} \mathbb{I}\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \leq 2\sigma\right\} \\
&= \sum_{j=1}^{|\mathcal{C}_\sigma|} \sum_{k=1}^{K} \frac{1}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} \mathbb{I}\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \leq 2\sigma\right\} \mathbb{I}\left\{(x_h^k, a_h^k) \in B_j\right\} \\
&\leq \beta^{-1} \sum_{j=1}^{|\mathcal{C}_\sigma|} \sum_{k=1}^{K} \frac{\mathbb{I}\left\{(x_h^k, a_h^k) \in B_j\right\}}{1 + c\mathbf{N}_h^k(B_j, U_\eta)}, \quad \text{by Lemma 15} \\
&\leq \beta^{-1} \sum_{j=1}^{|\mathcal{C}_\sigma|} \sum_{n=1}^{\lceil K/U_\eta \rceil} \left(1 + \int_0^{\mathbf{N}_h^{nU_\eta+1}(B_j, U_\eta)} \frac{1}{1+cz} \mathrm{d}z\right), \quad \text{by Lemma 14} \\
&= \beta^{-1} |\mathcal{C}_\sigma| \left\lceil \frac{K}{U_\eta} \right\rceil + \frac{\beta^{-1} |\mathcal{C}_\sigma|}{c} \sum_{n=1}^{\lceil K/U_\eta \rceil} \sum_{j=1}^{|\mathcal{C}_\sigma|} \frac{1}{|\mathcal{C}_\sigma|} \log\left(1 + c\mathbf{N}_h^{nU_\eta+1}(B_j, U_\eta)\right) \\
&\leq \beta^{-1} |\mathcal{C}_\sigma| \left\lceil \frac{K}{U_\eta} \right\rceil + \frac{\beta^{-1} |\mathcal{C}_\sigma|}{c} \sum_{n=1}^{\lceil K/U_\eta \rceil} \log\left(1 + \frac{c}{|\mathcal{C}_\sigma|} \sum_{j=1}^{|\mathcal{C}_\sigma|} \mathbf{N}_h^{nU_\eta+1}(B_j, U_\eta)\right), \quad \text{by Jensen's inequality} \\
&\leq \beta^{-1} |\mathcal{C}_\sigma| \left\lceil \frac{K}{U_\eta} \right\rceil + \frac{\beta^{-1} |\mathcal{C}_\sigma|}{c} \left\lceil \frac{K}{U_\eta} \right\rceil \log\left(1 + c\frac{U_\eta}{|\mathcal{C}_\sigma|}\right) \lesssim |\mathcal{C}_\sigma| \left\lceil \frac{K}{U_\eta} \right\rceil .
\end{aligned}
$$

□

F.4 Final regret bounds

F.4.1 UCRL-type regret bounds

Theorem 3 (UCRL-type regret bound). *If $U_\eta = \lceil 1/\log(1/\eta) \rceil \le W$, the regret of* KeRNS *is bounded by*

$$\begin{aligned}\mathcal{R}(K) \lesssim & H^2 \left\lceil \frac{K}{U_\eta} \right\rceil \sqrt{|\mathcal{C}'_\sigma|} \left(|\mathcal{C}_\sigma| + \sqrt{|\mathcal{C}_\sigma| U_\eta} \right) + H^2 |\mathcal{C}_\sigma| \left\lceil \frac{K}{U_\eta} \right\rceil + H^{3/2}\sqrt{K} \\ & + W \left(\Delta^{\mathrm{r}} + L\Delta^{\mathrm{p}} \right) H + \frac{\eta^W}{1-\eta} K H^3 \\ & + H^2 |\mathcal{C}_\sigma| + LKH\sigma\end{aligned}$$

with probability at least $1-\delta$, where $|\mathcal{C}_\sigma|$ and $|\mathcal{C}'_\sigma|$ are the σ-covering numbers of $(\mathcal{X}\times\mathcal{A}, \rho)$ and $(\mathcal{X}, \rho_{\mathcal{X}})$, respectively, and

$$\Delta^{\mathrm{r}} \stackrel{\mathrm{def}}{=} \sum_{i=1}^{K}\sum_{h=1}^{H} \sup_{x,a} \left| r_h^i(x,a) - r_h^{i+1}(x,a) \right|, \quad \Delta^{\mathrm{p}} \stackrel{\mathrm{def}}{=} \sum_{i=1}^{K}\sum_{h=1}^{H} \sup_{x,a} \mathbb{W}_1 \left(\mathrm{P}_h^i(\cdot|x,a), \mathrm{P}_h^{i+1}(\cdot|x,a) \right)$$

represent the variation of the rewards and transitions, respectively.

Proof. We apply Lemma 11, Lemma 16 and Corollary 5 and the fact that $\mathbb{P}[\mathcal{G}] \ge 1-\delta/2$ by Lemma 9. To conclude, notice that $\sum_{k=1}^{K}\sum_{h=1}^{H} \widetilde{\xi}_{h+1}^k \lesssim H\sqrt{KH}$ with probability at least $1-\delta/2$ by Hoeffding-Azuma's inequality. □

Corollary 6. *Let d_1 and d_2 be the covering dimensions of $(\mathcal{X}\times\mathcal{A}, \rho)$ and $(\mathcal{X}, \rho_{\mathcal{X}})$, respectively. Let $\alpha = \frac{1}{d_1+d_2+3}$, $\Delta = \Delta^{\mathrm{r}} + L\Delta^{\mathrm{p}}$ and*

$$\sigma = K^{-\alpha}, \quad \log\left(\frac{1}{\eta}\right) = \left(\frac{\Delta}{K^{1+\alpha(d_1+d_2)/2}} \right)^{2/3}, \quad W = \frac{\log\left(K/(1-\eta)\right)}{\log\left(1/\eta\right)}$$

Since $W \ge U_\eta = \lceil 1/\log(1/\eta) \rceil$, we have, with probability at least $1-\delta$,

$$\begin{aligned}\mathcal{R}(K) \lesssim & H^2 \Delta^{\frac{2}{3}} K^{\frac{d_1+d_2/6+1}{d_1+d_2+3}} + H^2 \Delta^{\frac{1}{3}} K^{\frac{d_1+d_2+2}{d_1+d_2+3}} \\ & + H^2 \Delta^{\frac{2}{3}} K^{\frac{d_1+d_2/6+1}{d_1+d_2+3}} + H^{\frac{3}{2}}\sqrt{K} + H\Delta^{\frac{1}{3}} K^{\frac{d_1+d_2+2}{d_1+d_2+3}} \log\left(\frac{HK}{1-\eta}\right) + H^3 \\ & + H^2 K^{\frac{d_1}{d_1+d_2+3}} + LHK^{\frac{d_1+d_2+2}{d_1+d_2+3}}\end{aligned}$$

that is,

$$\mathcal{R}(K) = \widetilde{\mathcal{O}}\left(\Delta^{\frac{1}{3}} K^{\frac{d_1+d_2+2}{d_1+d_2+3}} \left(H^2 + H\log\left(\frac{HK}{1-\eta}\right)\right)\right).$$

Furthermore, if $\limsup_{K\to\infty} \Delta/K = 0$, we have

$$\eta = \exp\left(-\Delta^{2/3}/K^{\frac{d_1+d_2+2}{d_1+d_2+3}}\right) \underset{K\to\infty}{\sim} 1 - \Delta^{2/3}/K^{\frac{d_1+d_2+2}{d_1+d_2+3}}$$

which implies

$$\mathcal{R}(K) = \widetilde{\mathcal{O}}\left(H^2 \Delta^{\frac{1}{3}} K^{\frac{d_1+d_2+2}{d_1+d_2+3}} \right).$$

Proof. Immediate consequence of Theorem 3 and the fact that $|\mathcal{C}_\sigma| = \mathcal{O}\left(\sigma^{-d_1}\right)$ and $|\mathcal{C}'_\sigma| = \mathcal{O}\left(\sigma^{-d_2}\right)$. □

Corollary 7 (UCRL-type regret bound in discrete case)**.** *If $\mathcal{X} \times \mathcal{A}$ is finite, we can take $\sigma = 0$ and $|\mathcal{C}_\sigma| = XA$, $|\mathcal{C}'_\sigma| = X$, where $X = |\mathcal{X}|$ and $A = |\mathcal{A}|$. In this case, Theorem 3 and Corollary 6 give us*

$$\mathcal{R}(K) = \widetilde{\mathcal{O}}\left(H^2 X \sqrt{A} \Delta^{\frac{1}{3}} K^{\frac{2}{3}}\right).$$

F.4.2 UCBVI-type regret bounds

Theorem 4 (UCBVI-type regret bound)**.** *If $U_\eta = \lceil 1/\log(1/\eta) \rceil \leq W$, the regret of KeRNS is bounded by*

$$\begin{aligned}\mathcal{R}(K) \lesssim & H^2 \left\lceil \frac{K}{U_\eta} \right\rceil \left(|\mathcal{C}_\sigma| + \sqrt{|\mathcal{C}_\sigma| U_\eta} \right) + H^3 |\mathcal{C}_\sigma| |\mathcal{C}'_\sigma| \left\lceil \frac{K}{U_\eta} \right\rceil + H^{3/2} \sqrt{K} \\ & + W \left(\Delta^{\mathrm{r}} + L \Delta^{\mathrm{p}} \right) H + \frac{\eta^W}{1-\eta} K H^3 \\ & + H |\mathcal{C}_\sigma| + LKH\sigma\end{aligned}$$

with probability at least $1 - \delta$, where $|\mathcal{C}_\sigma|$ and $|\mathcal{C}'_\sigma|$ are the σ-covering numbers of $(\mathcal{X} \times \mathcal{A}, \rho)$ and $(\mathcal{X}, \rho_{\mathcal{X}})$, respectively, and

$$\Delta^{\mathrm{r}} \stackrel{\text{def}}{=} \sum_{i=1}^{K} \sum_{h=1}^{H} \sup_{x,a} \left| r_h^i(x,a) - r_h^{i+1}(x,a) \right|, \quad \Delta^{\mathrm{p}} \stackrel{\text{def}}{=} \sum_{i=1}^{K} \sum_{h=1}^{H} \sup_{x,a} \mathbb{W}_1 \left(\mathrm{P}_h^i(\cdot|x,a), \mathrm{P}_h^{i+1}(\cdot|x,a) \right)$$

represent the variation of the rewards and transitions, respectively.

Proof. We apply Lemma 12, Lemma 16 and Corollary 5 and the fact that $\mathbb{P}[\mathcal{G}] \geq 1 - \delta/2$ by Lemma 9. To conclude, notice that $\sum_{k=1}^{K} \sum_{h=1}^{H} \left(1 + \frac{1}{H}\right)^h \widetilde{\xi}_{h+1}^k \lesssim H\sqrt{KH}$ with probability at least $1 - \delta/2$ by Hoeffding-Azuma's inequality. □

Corollary 8 (UCBVI-type regret bound in discrete case)**.** *If $\mathcal{X} \times \mathcal{A}$ is finite, we can take $\sigma = 0$ and $|\mathcal{C}_\sigma| = XA$, $|\mathcal{C}'_\sigma| = X$, where $X = |\mathcal{X}|$ and $A = |\mathcal{A}|$. In this case, Theorem 4 gives us*

$$\mathcal{R}(K) \lesssim H^2 \left\lceil \frac{K}{U_\eta} \right\rceil \left(XA + \sqrt{XAU_\eta} \right) + H^3 X^2 A \left\lceil \frac{K}{U_\eta} \right\rceil + H^{3/2}\sqrt{K} + W\left(\Delta^{\mathrm{r}} + L\Delta^{\mathrm{p}}\right) H + \frac{\eta^W}{1-\eta} K H^3 + H^2 XA$$

Let $\Delta = \Delta^{\mathrm{r}} + L\Delta^{\mathrm{p}}$. By choosing

$$\log\left(\frac{1}{\eta}\right) = \left(\frac{\Delta}{K}\right)^{2/3}, \quad W = \frac{\log\left(K/(1-\eta)\right)}{\log\left(1/\eta\right)}$$

we obtain

$$\mathcal{R}(K) \lesssim H^2 XA\Delta^{\frac{2}{3}} K^{\frac{1}{3}} + H^2\sqrt{XA}\Delta^{\frac{1}{3}} K^{\frac{2}{3}} + H^3 X^2 A\Delta^{\frac{2}{3}} K^{\frac{1}{3}} + H^{3/2}\sqrt{K} + \log\left(\frac{K}{1-\eta}\right) H\Delta^{\frac{1}{3}} K^{\frac{2}{3}} + H^3 + HXA.$$

since $W \geq U_\eta = \lceil 1/\log(1/\eta) \rceil$. Furthermore, if $\limsup_{K\to\infty} \Delta/K = 0$, we have

$$\eta = \exp\left(-\Delta^{\frac{2}{3}}/K^{\frac{2}{3}}\right) \underset{K\to\infty}{\sim} 1 - \Delta^{\frac{2}{3}}/K^{\frac{2}{3}}$$

which implies

$$\mathcal{R}(K) = \widetilde{\mathcal{O}}\left(H^2\sqrt{XA}\Delta^{\frac{1}{3}} K^{\frac{2}{3}} + H^3 X^2 A\Delta^{\frac{2}{3}} K^{\frac{1}{3}}\right).$$

Corollary 9. *Let d_1 and d_2 be the covering dimensions of $(\mathcal{X} \times \mathcal{A}, \rho)$ and $(\mathcal{X}, \rho_{\mathcal{X}})$, respectively. Let $\alpha = \frac{1}{d_1+d_2+2}$, $\Delta = \Delta^{\mathrm{r}} + L\Delta^{\mathrm{p}}$ and*

$$\sigma = K^\alpha, \quad \log\left(\frac{1}{\eta}\right) = \left(\frac{\Delta}{HK^{1+\alpha(d_1+d_2)}}\right)^{1/2}, \quad W = \left\lceil \frac{\log\left(K/(1-\eta)\right)}{\log(1/\eta)} \right\rceil$$

Since $W \geq U_\eta = \lceil 1/\log(1/\eta) \rceil$, we have, with probability at least $1-\delta$,

$$\mathcal{R}(K) \lesssim H^2 \left(1 + \log\left(\frac{K}{1-\eta}\right)\right) \Delta^{\frac{1}{2}} K^{\frac{d_1+d_2+1}{d_1+d_2+2}} + H^{\frac{3}{2}}\Delta^{\frac{1}{4}} K^{\frac{3}{4}} + LHK^{\frac{d_1+d_2+1}{d_1+d_2+2}} + H^2.$$

Furthermore, if $\limsup_{K\to\infty} \Delta/K = 0$, we have

$$\eta = \exp\left(-\Delta^{1/2}/K^{\frac{d_1+d_2+1}{d_1+d_2+2}}\right) \underset{K\to\infty}{\sim} 1 - \Delta^{1/2}/K^{\frac{d_1+d_2+1}{d_1+d_2+2}}$$

which implies

$$\mathcal{R}(K) = \widetilde{\mathcal{O}}\left(H^2\Delta^{\frac{1}{2}} K^{\frac{d_1+d_2+1}{d_1+d_2+2}} + H^{\frac{3}{2}}\Delta^{\frac{1}{4}} K^{\frac{3}{4}}\right).$$

Proof. Immediate consequence of Theorem 4 and the fact that $|\mathcal{C}_\sigma| = \mathcal{O}\left(\sigma^{-d_1}\right)$ and $|\mathcal{C}'_\sigma| = \mathcal{O}\left(\sigma^{-d_2}\right)$. □

G RS-KeRNS: An efficient version of KeRNS using representative states

RS-KeRNS is described in Algorithm 4, which uses a backward induction on representative states (Algorithm 5) and updates the model online (algorithms 6 and 7). In this section, we introduce the main definitions used by RS-KeRNS, and we analyze its runtime and regret.

G.1 Definitions

In each episode k and for each h, RS-KeRNS keeps and updates sets of representative states $\bar{\mathcal{X}}_h^k$, actions $\bar{\mathcal{A}}_h^k$, and next-states $\bar{\mathcal{Y}}_h^k$, with cardinalities $\bar{X}_h^k, \bar{A}_h^k$ and $\bar{Y}_h^k$, respectively. These sets are built using the data observed up to episode $k-1$. We define the following projections:

$$\zeta_h^{k+1}(x,a) \overset{\text{def}}{=} \operatorname*{argmin}_{(\overline{x},\overline{a})\in\bar{\mathcal{X}}_h^k\times\bar{\mathcal{A}}_h^k} \rho\left[(x,a),(\overline{x},\overline{a})\right], \quad \overline{\zeta}_h^{k+1}(y) \overset{\text{def}}{=} \operatorname*{argmin}_{\overline{y}\in\bar{\mathcal{Y}}_h^k} \rho_{\mathcal{X}}\left(y,\overline{y}\right).$$

where we also assume to have access to the metric $\rho_{\mathcal{X}}$. The definitions below introduce the kernel function and the estimated MDP used by RS-KeRNS.

Definition 10 (kernel function for RS-KeRNS)**.** *Let $(\alpha_i)_{i\geq 1}$ be a sequence of numbers in $[0,1]$. RS-KeRNS uses a kernel of the form $\Gamma(t,u,v) = \chi(t)\phi(u,v)$, where*

$$\chi(t) \overset{\text{def}}{=} \prod_{i=1}^{t} \alpha_i, \quad \phi(u,v) \overset{\text{def}}{=} \exp\left(-\rho\left[u,v\right]^2/(2\sigma^2)\right),$$

and, by convention, $\chi(0)=1$.

Definition 11 (empirical MDP for RS-KeRNS)**.** *Let*

$$\breve{W}_h^{k+1}(x,a) = \sum_{s=1}^{k} \chi(k-s)\phi\left(\zeta_h^{k+1}(x,a), \zeta_h^{s+1}(x_h^s,a_h^s)\right).$$

In episode $k+1$, RS-KeRNS uses the following estimate of the reward function

$$\breve{r}_h^{k+1}(x,a) = \frac{1}{\beta + \breve{W}_h^{k+1}(x,a)} \sum_{s=1}^{k} \chi(k-s)\phi\left(\zeta_h^{k+1}(x,a), \zeta_h^{s+1}(x_h^s,a_h^s)\right) \widetilde{r}_h^s$$

and the follow estimate of the transitions

$$\breve{P}_h^{k+1}(y|x,a) = \frac{1}{\beta + \breve{W}_h^{k+1}(x,a)} \sum_{s=1}^{k} \chi(k-s)\phi\left(\zeta_h^{k+1}(x,a), \zeta_h^{s+1}(x_h^s,a_h^s)\right) \delta_{\overline{\zeta}_h^{s+1}(x_{h+1}^s)}(y).$$

Also, its exploration bonuses are computed as

$$\breve{\mathtt{B}}_h^{k+1}(x,a) \overset{\text{def}}{=} \widetilde{\mathcal{O}}\left(\frac{H}{\sqrt{\beta + \breve{W}_h^{k+1}(x,a)}} + \frac{\beta H}{\beta + \breve{W}_h^{k+1}(x,a)} + L\sigma\right)$$

where the factors hidden by $\widetilde{\mathcal{O}}(\cdot)$ are the same as in Definition 5.

At step h, RS-KeRNS needs to store the quantities in Def. 11 *only for the representatives* (x,a) *in* $\bar{\mathcal{X}}_h^{k+1}\times\bar{\mathcal{A}}_h^{k+1}$ *and* $y\in\bar{\mathcal{Y}}_h^{k+1}$. We will show that, using the auxiliary quantities defined below, the values of $\breve{W}_h^k$, $\breve{r}_h^k$ and $\breve{P}_h^k$ can be updated online in $\mathcal{O}\left(\sum_h \bar{X}_h^k \bar{A}_h^k \bar{Y}_h^k\right)$ time per episode k.

Definition 12 (auxiliary quantities for online updates)**.** *For any (h, x, a), we define*

$$\breve{N}_h^{k+1}(x,a,y) \stackrel{\text{def}}{=} \sum_{s=1}^{k} \chi(k-s)\mathbb{I}\left\{\zeta_h^{s+1}(x_h^s,a_h^s)=(x,a)\right\}\delta_{\bar{\zeta}_h^{s+1}(x_{h+1}^s)}(y)$$

$$\breve{N}_h^{k+1}(x,a) \stackrel{\text{def}}{=} \sum_{s=1}^{k} \chi(k-s)\mathbb{I}\left\{\zeta_h^{s+1}(x_h^s,a_h^s)=(x,a)\right\}$$

$$\breve{S}_h^{k+1}(x,a) \stackrel{\text{def}}{=} \sum_{s=1}^{k} \chi(k-s)\mathbb{I}\left\{\zeta_h^{s+1}(x_h^s,a_h^s)=(x,a)\right\}\widetilde{r}_h^s.$$

Notice that, if $(x,a) \notin \bar{\mathcal{X}}_h^{k+1} \times \bar{\mathcal{A}}_h^{k+1}$, the quantities above are equal to zero.

The following Lemma will be necessary in order to derive online updates.

Lemma 17. *The empirical MDP used by RS-KeRNS can be computed as*

$$\breve{r}_h^{k+1}(x,a) = \frac{\sum_{(\overline{x},\overline{a})}\phi\left(\zeta_h^{k+1}(x,a),(\overline{x},\overline{a})\right)\breve{S}_h^{k+1}(\overline{x},\overline{a})}{\beta+\sum_{(\overline{x},\overline{a})}\phi\left(\zeta_h^{k+1}(x,a),(\overline{x},\overline{a})\right)\breve{N}_h^{k+1}(\overline{x},\overline{a})} \tag{5}$$

$$\breve{P}_h^{k+1}(y|x,a) = \frac{\sum_{(\overline{x},\overline{a})}\phi\left(\zeta_h^{k+1}(x,a),(\overline{x},\overline{a})\right)\breve{N}_h^{k+1}(\overline{x},\overline{a},y)}{\beta+\sum_{(\overline{x},\overline{a})}\phi\left(\zeta_h^{k+1}(x,a),(\overline{x},\overline{a})\right)\breve{N}_h^{k+1}(\overline{x},\overline{a})} \tag{6}$$

$$\breve{W}_h^{k+1}(x,a) = \sum_{(\overline{x},\overline{a})}\phi\left(\zeta_h^{k+1}(x,a),(\overline{x},\overline{a})\right)\breve{N}_h^{k+1}(\overline{x},\overline{a}) \tag{7}$$

where the sums are over $(\overline{x},\overline{a}) \in \bar{\mathcal{X}}_h^{k+1} \times \bar{\mathcal{A}}_h^{k+1}$.

Proof. It is an immediate consequence of the definitions. For instance,

$$\begin{aligned}
\breve{W}_h^{k+1}(x,a) &= \sum_{s=1}^{k}\chi(k-s)\phi\left(\zeta_h^{k+1}(x,a),\zeta_h^{s+1}(x_h^s,a_h^s)\right)\\
&= \sum_{s=1}^{k}\chi(k-s)\phi\left(\zeta_h^{k+1}(x,a),\zeta_h^{s+1}(x_h^s,a_h^s)\right)\sum_{(\overline{x},\overline{a})\in\bar{\mathcal{X}}_h^{k+1}\times\bar{\mathcal{A}}_h^{k+1}}\mathbb{I}\left\{\zeta_h^{s+1}(x_h^s,a_h^s)=(\overline{x},\overline{a})\right\}\\
&= \sum_{(\overline{x},\overline{a})}\sum_{s=1}^{k}\chi(k-s)\phi\left(\zeta_h^{k+1}(x,a),(\overline{x},\overline{a})\right)\mathbb{I}\left\{\zeta_h^{s+1}(x_h^s,a_h^s)=(\overline{x},\overline{a})\right\}\\
&= \sum_{(\overline{x},\overline{a})}\phi\left(\zeta_h^{k+1}(x,a),(\overline{x},\overline{a})\right)\sum_{s=1}^{k}\chi(k-s)\mathbb{I}\left\{\zeta_h^{s+1}(x_h^s,a_h^s)=(\overline{x},\overline{a})\right\}\\
&= \sum_{(\overline{x},\overline{a})}\phi\left(\zeta_h^{k+1}(x,a),(\overline{x},\overline{a})\right)\breve{N}_h^{k+1}(\overline{x},\overline{a}).
\end{aligned}$$

□

Algorithm 4 RS-KeRNS

Input: global parameters K, H, L, L_{r} , L_{p}, β, δ, d, σ, η, W, $\varepsilon_{\mathcal{X}}$, ε.
Initialize representative states, actions and next states: $\bar{\mathcal{X}}_h = \emptyset$, $\bar{\mathcal{A}}_h = \emptyset$, $\bar{\mathcal{Y}}_h = \emptyset$, for $h \in [H]$.
for episode $k = 1, \ldots, K$ **do**
  get initial state x_1^k
  compute $(\breve{Q}_h^k)_h$ using kernel backward induction on the representative sets (Alg. 5).
  for $h = 1, \ldots, H$ **do**
    execute $a_h^k = \operatorname{argmax}_a \breve{Q}_h^k(x_h^k, a)$, observe reward $\tilde{r}_h^k$ and next state x_{h+1}^k
    update representatives $\bar{\mathcal{X}}_h$, $\bar{\mathcal{A}}_h$, $\bar{\mathcal{Y}}_h$ using $\{x_h^k, a_h^k, x_{h+1}^k\}$ with Alg. 6
    update model using $x_h^k, a_h^k, x_{h+1}^k, \tilde{r}_h^k$ with Alg. 7
  end for
end for

Algorithm 5 Kernel Backward Induction on Representative States

Input: $\breve{r}_h^k(\overline{x}, \overline{a})$, $\breve{P}_h^k(\overline{y}|\overline{x}, \overline{a})$, $\breve{\mathtt{B}}_h^k(\overline{x}, \overline{a})$ for all $(\overline{x}, \overline{a}, \overline{y}) \in \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k \times \bar{\mathcal{Y}}_h^k$ and all $h \in [H]$.
Initialization: $\breve{V}_{H+1}(x) = 0$ for all $x \in \mathcal{X}$
for $h = H, \ldots, 1$ **do**
  for $(\overline{x}, \overline{a}) \in \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k$ **do**
    $\widetilde{Q}_{h,\zeta}^k(\overline{x}, \overline{a}) = \breve{r}_h^k(\overline{x}, \overline{a}) + \breve{P}_h^k \breve{V}_{h+1}(\overline{x}, \overline{a}) + \breve{\mathtt{B}}_h^k(\overline{x}, \overline{a})$
  end for
  // Interpolated Q-function. Defined, but not computed for all (x, a)
  $\breve{Q}_h^k(x, a) = \min_{(\overline{x}, \overline{a}) \in \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k} \left(\widetilde{Q}_{h,\zeta}^k(\overline{x}, \overline{a}) + L\rho\left[(x, a), (\overline{x}, \overline{a})\right] \right)$
  if $h > 1$ **then**
    // Compute V-function at the next states for the stage $h - 1$
    for $\overline{y} \in \bar{\mathcal{Y}}_{h-1}^k$ **do**
      $\breve{V}_h^k(\overline{y}) = \min\left(H - h + 1, \max_a \breve{Q}_h^k(\overline{y}, a)\right)$
    end for
  end if
end for
Return: $(\breve{Q}_h^k)_{h \in [H]}$

Algorithm 6 Update Representative Sets

Input: $\bar{\mathcal{X}}_h^k$, $\bar{\mathcal{A}}_h^k$, $\bar{\mathcal{Y}}_h^k$, $\{x_h^k, a_h^k, x_{h+1}^k\}$, ε, $\varepsilon_{\mathcal{X}}$.
if $\min_{(\overline{x}, \overline{a}) \in \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k} \rho\left[(\overline{x}, \overline{a}), (x_h^k, a_h^k)\right] > \varepsilon$ **then**
  $\bar{\mathcal{X}}_h^{k+1} = \bar{\mathcal{X}}_h^k \cup \{x_h^k\}$, $\bar{\mathcal{A}}_h^{k+1} = \bar{\mathcal{A}}_h^k \cup \{a_h^k\}$
else
  $\bar{\mathcal{X}}_h^{k+1} = \bar{\mathcal{X}}_h^k$, $\bar{\mathcal{A}}_h^{k+1} = \bar{\mathcal{A}}_h^k$
end if
if $\min_{\overline{y} \in \bar{\mathcal{Y}}_h^k} \rho_{\mathcal{X}}\left(\overline{x}, x_{h+1}^k\right) > \varepsilon_{\mathcal{X}}$ **then**
  $\bar{\mathcal{Y}}_h^{k+1} = \bar{\mathcal{Y}}_h^k \cup \{x_{h+1}^k\}$
else
  $\bar{\mathcal{Y}}_h^{k+1} = \bar{\mathcal{Y}}_h^k$
end if

Algorithm 7 Online update of RS-KeRNS Model

Input: $k, h, x_h^k, a_h^k, x_{h+1}^k, \widetilde{r}_h^k$.
// Map to representatives
Map $(\widetilde{x}, \widetilde{a}) = \varsigma_h^{k+1}(x_h^k, a_h^k)$ and $\widetilde{y} = \overline{\varsigma}_h^{k+1}(x_{h+1}^k)$
// Update auxiliary quantities
$\breve{N}_h^{k+1}(\widetilde{x}, \widetilde{a}, \widetilde{y}) = 1 + \alpha_{k-s}\breve{N}_h^k(\widetilde{x}, \widetilde{a}, \widetilde{y})$
$\breve{N}_h^{k+1}(\widetilde{x}, \widetilde{a}) = 1 + \alpha_{k-s}\breve{N}_h^k(\widetilde{x}, \widetilde{a})$
$\breve{S}_h^{k+1}(\widetilde{x}, \widetilde{a}) = \widetilde{r}_h^k + \alpha_{k-s}\breve{S}_h^k(\widetilde{x}, \widetilde{a})$
// Update empirical MDP
for $(\overline{x}, \overline{a}) \in \bar{\mathcal{X}}_h^{k+1} \times \bar{\mathcal{A}}_h^{k+1}$ **do**
  if $(\overline{x}, \overline{a}) \in \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k$ **then**
    // $(\overline{x}, \overline{a})$ was added before episode k
    $\breve{W}_h^{k+1}(\overline{x}, \overline{a}) = \phi\left((\overline{x}, \overline{a}), (\widetilde{x}, \widetilde{a})\right) + \alpha_{k-s}\breve{W}_h^k(\overline{x}, \overline{a})$
    $\breve{r}_h^{k+1}(\overline{x}, \overline{a}) = \frac{\phi((\overline{x}, \overline{a}), (\widetilde{x}, \widetilde{a}))}{\beta + \breve{W}_h^{k+1}(\overline{x}, \overline{a})}\widetilde{r}_h^k + \alpha_{k-s}\left(\frac{\beta + \breve{W}_h^k(\overline{x}, \overline{a})}{\beta + \breve{W}_h^{k+1}(\overline{x}, \overline{a})}\right)\breve{r}_h^k(\overline{x}, \overline{a})$
    for $y \in \bar{\mathcal{Y}}_h^{k+1}$ **do**
      $\breve{P}_h^{k+1}(y|\overline{x}, \overline{a}) = \frac{\phi((\overline{x}, \overline{a}), (\widetilde{x}, \widetilde{a}))}{\beta + \breve{W}_h^{k+1}(\overline{x}, \overline{a})}\delta_{\widetilde{y}}(y) + \alpha_{k-s}\left(\frac{\beta + \breve{W}_h^k(\overline{x}, \overline{a})}{\beta + \breve{W}_h^{k+1}(\overline{x}, \overline{a})}\right)\breve{P}_h^k(y|\overline{x}, \overline{a})$
    end for
  else
    // $(\overline{x}, \overline{a})$ was added in episode k
    Initialize $\breve{r}_h^{k+1}(\overline{x}, \overline{a}), \breve{P}_h^{k+1}(\cdot|\overline{x}, \overline{a}), \breve{W}_h^{k+1}(\overline{x}, \overline{a})$ using equations (5), (6) and (7)
  end if
end for

G.2 Online updates & runtime

Assume that we observed a transition $\left\{x_h^k, a_h^k, x_{h+1}^k, \widetilde{r}_h^k\right\}$ at time (k, h), updated the representative sets, and mapped the transition to the representatives $(\widetilde{x}, \widetilde{a}, \widetilde{y}) \in \bar{\mathcal{X}}_h^{k+1} \times \bar{\mathcal{A}}_h^{k+1} \times \bar{\mathcal{Y}}_h^{k+1}$. We wish to update the estimated MDP given in Def. 11, which, at step h, are only stored for (x, a) in $\bar{\mathcal{X}}_h^{k+1} \times \bar{\mathcal{A}}_h^{k+1}$ and $y \in \bar{\mathcal{Y}}_h^{k+1}$.

The auxiliary quantities (Def. 12) are updated as:

$$\begin{aligned}
\breve{N}_h^{k+1}(\widetilde{x}, \widetilde{a}, \widetilde{y}) &= 1 + \alpha_{k-s}\breve{N}_h^k(\widetilde{x}, \widetilde{a}, \widetilde{y}) \\
\breve{N}_h^{k+1}(\widetilde{x}, \widetilde{a}) &= 1 + \alpha_{k-s}\breve{N}_h^k(\widetilde{x}, \widetilde{a}) \\
\breve{S}_h^{k+1}(\widetilde{x}, \widetilde{a}) &= \widetilde{r}_h^k + \alpha_{k-s}\breve{S}_h^k(\widetilde{x}, \widetilde{a}).
\end{aligned}$$

We need to update $\breve{W}_h^k$, $\breve{r}_h^k$ and $\breve{P}_h^k$ for all $(\overline{x}, \overline{a}, \overline{y}) \in \bar{\mathcal{X}}_h^{k+1} \times \bar{\mathcal{A}}_h^{k+1} \times \bar{\mathcal{Y}}_h^{k+1}$. The update rule will depend on whether the $(\overline{x}, \overline{a})$ is a *new* representative state-action pair (included in episode k) or it was *visited before episode* k. These two cases are studied below.

Case 1: $(\overline{x}, \overline{a}) \in \bar{\mathcal{X}}_h^{k+1} \times \bar{\mathcal{A}}_h^{k+1}$ **and** $(\overline{x}, \overline{a}) \notin \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k$ This means that the representative state-action pair $(\overline{x}, \overline{a})$ was added at time (k, h). In this case, for all $y \in \bar{\mathcal{Y}}_h^{k+1}$, the quantities $\breve{r}_h^{k+1}(\overline{x}, \overline{a})$, $\breve{P}_h^{k+1}(y|\overline{x}, \overline{a})$ and $\breve{W}_h^{k+1}(\overline{x}, \overline{a})$ can be initialized using equations (5), (6) and (7). This is done in $\mathcal{O}\left(\bar{X}_h^{k+1}\bar{A}_h^{k+1}\bar{Y}_h^{k+1}\right)$ time and can happen, at most, for one pair $(\overline{x}, \overline{a})$: the one that was newly added. Therefore, we have a total per-episode runtime of $\mathcal{O}\left(\sum_{h=1}^H \bar{X}_h^{k+1}\bar{A}_h^{k+1}\bar{Y}_h^{k+1}\right)$ taking this case into account.

Case 2: $(\overline{x}, \overline{a}) \in \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k$ This means that the representative state-action pair $(\overline{x}, \overline{a})$ was added *before* episode k, which implies that $\zeta_h^{k+1}(\overline{x}, \overline{a}) = \zeta_h^k(\overline{x}, \overline{a}) = (\overline{x}, \overline{a})$. Hence,

$$\begin{aligned}
\breve{W}_h^{k+1}(\overline{x}, \overline{a}) &= \sum_{s=1}^{k} \chi(k-s)\phi\left(\zeta_h^{k+1}(\overline{x}, \overline{a}), \zeta_h^{s+1}(x_h^s, a_h^s)\right) \\
&= \phi\left(\zeta_h^{k+1}(\overline{x}, \overline{a}), \zeta_h^{k+1}(x_h^k, a_h^k)\right) + \sum_{s=1}^{k-1} \left(\prod_{i=1}^{k-s} \alpha_i\right) \phi\left(\zeta_h^k(\overline{x}, \overline{a}), \zeta_h^{s+1}(x_h^s, a_h^s)\right) \\
&= \phi\left((\overline{x}, \overline{a}), \zeta_h^{k+1}(x_h^k, a_h^k)\right) + \alpha_{k-s} \sum_{s=1}^{k-1} \left(\prod_{i=1}^{k-s-1} \alpha_i\right) \phi\left(\zeta_h^k(\overline{x}, \overline{a}), \zeta_h^{s+1}(x_h^s, a_h^s)\right) \\
&= \phi\left((\overline{x}, \overline{a}), \zeta_h^{k+1}(x_h^k, a_h^k)\right) + \alpha_{k-s} \breve{W}_h^k(\overline{x}, \overline{a}),
\end{aligned}$$

where we used the convention $\chi(0) = 1$. This implies that, for a fixed $(\overline{x}, \overline{a})$, the quantity $\breve{W}_h^{k+1}(\overline{x}, \overline{a})$ can be updated in $\mathcal{O}(1)$ time, assuming that the mapping $\zeta_h^{k+1}(x_h^k, a_h^k)$ was previously computed (this mapping is only computed *once* for all the updates, and takes $\mathcal{O}\left(\bar{X}_h^{k+1} \times \bar{A}_h^{k+1}\right)$ time).

Now, notice that

$$\begin{aligned}
\breve{r}_h^{k+1}(\overline{x}, \overline{a}) &= \frac{\sum_{s=1}^{k} \chi(k-s)\phi\left(\zeta_h^{k+1}(\overline{x}, \overline{a}), \zeta_h^{s+1}(x_h^s, a_h^s)\right) \widetilde{r}_h^s}{\beta + \breve{W}_h^{k+1}(\overline{x}, \overline{a})} \\
&= \frac{\phi\left((\overline{x}, \overline{a}), \zeta_h^{k+1}(x_h^k, a_h^k)\right)}{\beta + \breve{W}_h^{k+1}(\overline{x}, \overline{a})} \widetilde{r}_h^k + \alpha_{k-s} \left(\frac{\beta + \breve{W}_h^k(\overline{x}, \overline{a})}{\beta + \breve{W}_h^{k+1}(\overline{x}, \overline{a})}\right) \breve{r}_h^k(\overline{x}, \overline{a})
\end{aligned}$$

where we used again the fact that, in this case, $\zeta_h^{k+1}(\overline{x}, \overline{a}) = \zeta_h^k(\overline{x}, \overline{a})$. Hence, similarly to $\breve{W}_h^{k+1}(\overline{x}, \overline{a})$, the quantity $\breve{r}_h^{k+1}(\overline{x}, \overline{a})$ can be updated in $\mathcal{O}(1)$ time. A similar reasoning shows that $\breve{P}_h^{k+1}(y, \overline{x}, \overline{a})$ can be updated, for all $y \in \bar{\mathcal{Y}}_h^{k+1}$, in $\mathcal{O}\left(\bar{Y}_h^{k+1}\right)$ time:

$$\breve{P}_h^{k+1}(y|\overline{x}, \overline{a}) = \frac{\phi\left((\overline{x}, \overline{a}), \zeta_h^{k+1}(x_h^k, a_h^k)\right)}{\beta + \breve{W}_h^{k+1}(\overline{x}, \overline{a})} \delta_{\zeta_h^{k+1}(x_{h+1}^k)}(y) + \alpha_{k-s} \left(\frac{\beta + \breve{W}_h^k(\overline{x}, \overline{a})}{\beta + \breve{W}_h^{k+1}(\overline{x}, \overline{a})}\right) \breve{P}_h^k(y|\overline{x}, \overline{a}).$$

Summary Every time a new transition is observed at time (k, h), the estimators for all $(x, a, y) \in \bar{\mathcal{X}}_h^{k+1} \times \bar{\mathcal{A}}_h^{k+1} \times \bar{\mathcal{Y}}_h^{k+1}$ must be updated. For a given representative (x, a), the updates can be done in $\mathcal{O}\left(\bar{Y}_h^{k+1}\right)$ time if it has been observed before episode k (case 2). This results in a total runtime, per episode, of $\mathcal{O}\left(\sum_h \bar{X}_h^{k+1} \bar{A}_h^{k+1} \bar{Y}_h^{k+1}\right)$ for all the representatives observed before episode k. If the representative (x, a) has not been observed before episode k (case 1), the updates require $\mathcal{O}\left(\bar{X}_h^{k+1} \bar{A}_h^{k+1} \bar{Y}_h^{k+1}\right)$ time, and this can happen, at most, for one state-action pair at each time (k, h). Hence, the total runtime required for the updates is $\mathcal{O}\left(\sum_h \bar{X}_h^{k+1} \bar{A}_h^{k+1} \bar{Y}_h^{k+1}\right)$ per episode.

G.3 Regret analysis

The regret analysis of RS-KeRNS is based on the following result, which is a corollary of Lemma 25, and is used to bound the bias introduced by using representative states.

Corollary 10. *Let $\chi_{(\eta,W)} : \mathbb{N} \to [0,1]$, consider the following kernel*

$$\Gamma(t,u,v) = \chi_{(\eta,W)}(t) \exp\left(-\rho\left[u,v\right]^2 / (2\sigma^2)\right)$$

where $u, v \in \mathcal{X} \times \mathcal{A}$. For $s < k$, let

$$w_h^{k,s}(x,a) = \Gamma(k-s-1,(x,a),(x_h^s,a_h^s)),\ w_{h,\zeta}^{k,s}(x,a) \stackrel{\text{def}}{=} \Gamma\left(k-s-1,\zeta_h^k(x,a),\zeta_h^k(x_h^s,a_h^s)\right)$$

and consider the functions

$$g_1(x,a) = \frac{\sum_{s=1}^{k-1} w_h^{k,s}(x,a) Y_s}{\beta + \sum_{s=1}^{k-1} w_h^{k,s}(x,a)}, \quad g_1^\zeta(x,a) = \frac{\sum_{s=1}^{k-1} w_{h,\zeta}^{k,s}(x,a) Y_s}{\beta + \sum_{s=1}^{k-1} w_{h,\zeta}^{k,s}(x,a)},$$
$$g_2(x,a) = \sqrt{\frac{1}{\beta + \sum_{s=1}^{k-1} w_h^{k,s}(x,a)}}, \quad g_2^\zeta(x,a) = \sqrt{\frac{1}{\beta + \sum_{s=1}^{k-1} w_{h,\zeta}^{k,s}(x,a)}},$$
$$g_3(x,a) = \frac{1}{\beta + \sum_{s=1}^{k-1} w_h^{k,s}(x,a)}, \quad g_3^\zeta(x,a) = \frac{1}{\beta + \sum_{s=1}^{k-1} w_{h,\zeta}^{k,s}(x,a)}.$$

where $(Y_s)_{s=1}^{k-1}$ is an arbitrary sequence. Then, g_1, g_2 and g_3 are Lipschitz continuous, whose Lipschitz constants are bounded by L_1, L_2 and L_3, respectively, with

$$L_1 = \frac{4\max_s |Y_s|}{\sigma}\left(1 + \sqrt{\log^+(k/\beta)}\right), \quad L_2 = \frac{1 + \sqrt{\log^+(k/\beta)}}{2\beta^{1/2}\sigma}, \quad L_3 = \frac{1 + \sqrt{\log^+(k/\beta)}}{\beta\sigma}$$

Furthermore, for any (x,a) and for $i \in \{1,2,3\}$,

$$\left|g_i^\zeta(x,a) - g_i(\zeta_h^k(x,a))\right| \le L_i \max_s \rho\left[(x_h^s,a_h^s),\zeta_h^k(x_h^s,a_h^s)\right].$$

Proof. First, let's prove that g_1, g_2 and g_3 are Lipschitz continuous. From Lemma 25, taking $z = (\rho\left[(x,a),(x_h^s,a_h^s)\right])_{s=1}^{k-1}$ and $y = (\rho\left[(x',a'),(x_h^s,a_h^s)\right])_{s=1}^{k-1}$ we have

$$|g_1(x,a) - g_1(x',a')| \le \frac{4\max_s |Y_s|}{\sigma}\left(1 + \sqrt{\log^+(k/\beta)}\right) \max_{s\in[k-1]} |z_s - y_s|.$$

By the triangle inequality, for all s,

$$|z_s - y_s| = |\rho\left[(x,a),(x_h^s,a_h^s)\right] - \rho\left[(x',a'),(x_h^s,a_h^s)\right]| \le \rho\left[(x,a),(x',a')\right]$$

which implies

$$|g_1(x,a) - g_1(x',a')| \le \frac{4\max_s |Y_s|}{\sigma}\left(1 + \sqrt{\log^+(k/\beta)}\right) \rho\left[(x,a),(x',a')\right].$$

giving the Lipschitz constant of g_1. The Lipschitz constants of g_2 and g_3 follow similarly from Lemma 25.

Now, let's bound the differences $\left|g_i^\zeta(x,a) - g_i(\zeta_h^k(x,a))\right|$. Let $\chi(s) = \chi_{(\eta,W)}(k-s-1)$. For $i=1$, and applying

again Lemma 25, we obtain

$$\begin{aligned}
&\left|g_1^{\zeta}(x,a) - g_1(\zeta_h^k(x,a))\right| \\
&= \left| \frac{\sum_{s=1}^{k-1} \chi(s) \exp\left(-\frac{\rho\left[\zeta_h^k(x,a),\zeta_h^k(x_h^s,a_h^s)\right]^2}{2\sigma^2}\right) Y_s}{\beta + \sum_{s=1}^{k-1} \chi(s) \exp\left(-\frac{\rho\left[\zeta_h^k(x,a),\zeta_h^k(x_h^s,a_h^s)\right]^2}{2\sigma^2}\right)} - \frac{\sum_{s=1}^{k-1} \chi(s) \exp\left(-\frac{\rho\left[\zeta_h^k(x,a),(x_h^s,a_h^s)\right]^2}{2\sigma^2}\right) Y_s}{\beta + \sum_{s=1}^{k-1} \chi(s) \exp\left(-\frac{\rho\left[\zeta_h^k(x,a),(x_h^s,a_h^s)\right]^2}{2\sigma^2}\right)} \right| \\
&\le \frac{4 \max_s |Y_s|}{\sigma} \left(1 + \sqrt{\log^+(k/\beta)}\right) \max_{s\in[k-1]} \left|\rho\left[\zeta_h^k(x,a),\zeta_h^k(x_h^s,a_h^s)\right] - \rho\left[\zeta_h^k(x,a),(x_h^s,a_h^s)\right]\right| \\
&\le \frac{4 \max_s |Y_s|}{\sigma} \left(1 + \sqrt{\log^+(k/\beta)}\right) \max_{s\in[k-1]} \rho\left[\zeta_h^k(x_h^s,a_h^s),(x_h^s,a_h^s)\right],
\end{aligned}$$

where, in the last line, we used the triangle inequality. The proof for $i \in \{2,3\}$ also follow from Lemma 25. □

We use Corollary 10 to bound the difference between the estimators and the bonuses of KeRNS, $(\widehat{r}_h^k, \widehat{P}_h^k, \mathtt{B}_h^k)$, and the ones of RS-KeRNS, $(\breve{r}_h^k, \breve{P}_h^k, \breve{\mathtt{B}}_h^k)$.

Lemma 18. *Let V be an arbitrary L-Lipschitz function bounded by H. Then, for any (x,a),*

$$\begin{aligned}
\left|\left(\widehat{P}_h^k - \breve{P}_h^k\right) V(x,a)\right| \le& \frac{4H}{\sigma}\left(1 + \sqrt{\log^+(k/\beta)}\right)\left(\rho\left[(x,a),\zeta_h^k(x,a)\right] + \varepsilon\right) \\
&+ 3L\varepsilon_{\mathcal{X}} + 8H\left(1 + \sqrt{\log^+(k/\beta)}\right)\frac{\varepsilon}{\sigma}
\end{aligned}$$

Proof. To simplify the notations, let $f(s) \stackrel{\text{def}}{=} \chi(k-s-1)$ for $s \in [k-1]$. We have

$$\begin{aligned}
&\left|\left(\widehat{P}_h^k - \breve{P}_h^k\right) V(x,a)\right| \\
&= \left| \frac{\sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right) V(x_{h+1}^s)}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right)} - \frac{\sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),\zeta_h^{s+1}(x_h^s,a_h^s)\right) V(\overline{\zeta}_h^{s+1}(x_{h+1}^s))}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),\zeta_h^{s+1}(x_h^s,a_h^s)\right)} \right| \\
&\le ① + ②
\end{aligned}$$

where ① and ② are defined and bounded below. First,

$$\begin{aligned}
① &= \left| \frac{\sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right) V(x_{h+1}^s)}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right)} - \frac{\sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),\zeta_h^k(x_h^s,a_h^s)\right) V(\overline{\zeta}_h^k(x_{h+1}^s))}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),\zeta_h^k(x_h^s,a_h^s)\right)} \right| \\
&\le \left| \frac{\sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right) V(x_{h+1}^s)}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right)} - \frac{\sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right) V(\overline{\zeta}_h^k(x_{h+1}^s))}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right)} \right| \\
&+ \left| \frac{\sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right) V(\overline{\zeta}_h^k(x_{h+1}^s))}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right)} - \frac{\sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),(x_h^s,a_h^s)\right) V(\overline{\zeta}_h^k(x_{h+1}^s))}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),(x_h^s,a_h^s)\right)} \right| \\
&+ \left| \frac{\sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),(x_h^s,a_h^s)\right) V(\overline{\zeta}_h^k(x_{h+1}^s))}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),(x_h^s,a_h^s)\right)} - \frac{\sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),\zeta_h^k(x_h^s,a_h^s)\right) V(\overline{\zeta}_h^k(x_{h+1}^s))}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),\zeta_h^k(x_h^s,a_h^s)\right)} \right| \\
&\le L \max_{s\in[k-1]} \rho_{\mathcal{X}}\left(x_{h+1}^s, \overline{\zeta}_h^k(x_{h+1}^s)\right) \\
&+ \frac{4H}{\sigma}\left(1 + \sqrt{\log^+(k/\beta)}\right)\rho\left[(x,a),\zeta_h^k(x,a)\right] \\
&+ \frac{4H}{\sigma}\left(1 + \sqrt{\log^+(k/\beta)}\right)\max_s \rho\left[(x_h^s,a_h^s),\zeta_h^k(x_h^s,a_h^s)\right]
\end{aligned}$$

by using the fact that V is L-Lipschitz and Corollary 10.

To bound ②, let $z_s = \rho\left[\zeta^k(x,a), \zeta^{s+1}(x_h^s, a_h^s)\right]$ and $y_s = \rho\left[\zeta^k(x,a), \zeta^k(x_h^s, a_h^s)\right]$, for $s \in [k-1]$. Using again the fact that V is L-Lipschitz and Corollary 10, we obtain

$$
\begin{aligned}
② &= \left| \frac{\sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a), \zeta_h^k(x_h^s, a_h^s)\right) V(\overline{\zeta}_h^k(x_{h+1}^s))}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a), \zeta_h^k(x_h^s, a_h^s)\right)} - \frac{\sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a), \zeta_h^{s+1}(x_h^s, a_h^s)\right) V(\overline{\zeta}_h^{s+1}(x_{h+1}^s))}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a), \zeta_h^{s+1}(x_h^s, a_h^s)\right)} \right| \\
&\le L \max_s \rho_{\mathcal{X}}\left(\overline{\zeta}^{s+1}(x_{h+1}^s), \overline{\zeta}^k(x_{h+1}^s)\right) + \frac{4H}{\sigma}\left(1 + \sqrt{\log^+(k/\beta)}\right) \max_s |z_s - y_s| \\
&\le 2L\varepsilon_{\mathcal{X}} + \frac{4H}{\sigma}\left(1 + \sqrt{\log^+(k/\beta)}\right) \max_s \left|\rho\left[\zeta^k(x_h^s, a_h^s), \zeta^{s+1}(x_h^s, a_h^s)\right]\right| \\
&\le 2L\varepsilon_{\mathcal{X}} + 8H\left(1 + \sqrt{\log^+(k/\beta)}\right)\frac{\varepsilon}{\sigma}.
\end{aligned}
$$

By the construction of the algorithm, $\rho_{\mathcal{X}}\left(x_{h+1}^s, \overline{\zeta}_h^k(x_{h+1}^s)\right) \le \varepsilon_{\mathcal{X}}$, $\rho\left[(x_h^s, a_h^s), \zeta_h^k(x_h^s, a_h^s)\right] \le \varepsilon$, $\rho_{\mathcal{X}}\left(\overline{\zeta}^{s+1}(x_{h+1}^s), \overline{\zeta}^k(x_{h+1}^s)\right) \le 2\varepsilon_{\mathcal{X}}$ and $\rho\left[\zeta^k(x_h^s, a_h^s), \zeta^{s+1}(x_h^s, a_h^s)\right] \le 2\varepsilon$, which concludes the proof. □

Lemma 19. *For any (x,a),*

$$
\begin{aligned}
\left|\widehat{r}_h^k(x,a) - \breve{r}_h^k(x,a)\right| \le & \frac{4}{\sigma}\left(1 + \sqrt{\log^+(k/\beta)}\right)\left(\rho\left[(x,a), \zeta_h^k(x,a)\right] + \varepsilon\right) \\
& + 8\left(1 + \sqrt{\log^+(k/\beta)}\right)\frac{\varepsilon}{\sigma}
\end{aligned}
$$

Proof. It follows from a similar decomposition as in the proof of Lemma 18, and from Corollary 10. □

Lemma 20. *For any (x,a), we have,*

$$
\left|\mathtt{B}_h^k(x,a) - \breve{\mathtt{B}}_h^k(x,a)\right| \lesssim \frac{H}{\sigma}\left(1 + \frac{1}{2\beta^{1/2}}\right)\left(\rho\left[(x,a), \zeta_h^k(x,a)\right] + \varepsilon\right) + H\left(1 + \frac{1}{\beta^{1/2}}\right)\frac{\varepsilon}{\sigma}
$$

Proof. To simplify the notations, let $f(s) \overset{\text{def}}{=} \chi(k-s-1)$ for $s \in [k-1]$. Using definitions 5, 10, and 11, we have

$$
\begin{aligned}
&\left|\mathtt{B}_h^k(x,a) - \breve{\mathtt{B}}_h^k(x,a)\right| \\
&\lesssim H \underbrace{\left| \sqrt{\frac{1}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right)}} - \sqrt{\frac{1}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a), \zeta_h^{s+1}(x_h^s,a_h^s)\right)}} \right|}_{①} \\
&+ \beta H \underbrace{\left| \frac{1}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right)} - \frac{1}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a), \zeta_h^{s+1}(x_h^s,a_h^s)\right)} \right|}_{②}.
\end{aligned}
$$

Using Corollary 10, we obtain

$$
\begin{aligned}
① \leq & \left| \sqrt{\frac{1}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left((x,a),(x_h^s,a_h^s)\right)}} - \sqrt{\frac{1}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),(x_h^s,a_h^s)\right)}} \right| \\
& + \left| \sqrt{\frac{1}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),(x_h^s,a_h^s)\right)}} - \sqrt{\frac{1}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),\zeta_h^k(x_h^s,a_h^s)\right)}} \right| \\
& + \left| \sqrt{\frac{1}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),\zeta_h^k(x_h^s,a_h^s)\right)}} - \sqrt{\frac{1}{\beta + \sum_{s=1}^{k-1} f(s)\phi\left(\zeta_h^k(x,a),\zeta_h^{s+1}(x_h^s,a_h^s)\right)}} \right| \\
& \leq \left(\frac{1 + \sqrt{\log^+(k/\beta)}}{2\beta^{1/2}\sigma} \right) \left(\rho\left[(x,a),\zeta_h^k(x,a)\right] + \max_{s\in[k-1]} \rho\left[(x_h^s,a_h^s),\zeta_h^k(x,a)\right] \right. \\
& \qquad \left. + \max_{s\in[k-1]} \rho\left[\zeta_h^k(x_h^s,a_h^s),\zeta_h^{s+1}(x,a)\right] \right).
\end{aligned}
$$

Similarly, Corollary 10 yields

$$
\begin{aligned}
② \leq & \left(\frac{1 + \sqrt{\log^+(k/\beta)}}{\beta\sigma} \right) \left(\rho\left[(x,a),\zeta_h^k(x,a)\right] + \max_{s\in[k-1]} \rho\left[(x_h^s,a_h^s),\zeta_h^k(x,a)\right] \right. \\
& \qquad \left. + \max_{s\in[k-1]} \rho\left[\zeta_h^k(x_h^s,a_h^s),\zeta_h^{s+1}(x,a)\right] \right).
\end{aligned}
$$

By the construction of the algorithm, $\rho\left[(x_h^s,a_h^s),\zeta_h^k(x,a)\right] \leq \varepsilon$ and $\rho\left[\zeta^k(x_h^s,a_h^s),\zeta^{s+1}(x_h^s,a_h^s)\right] \leq 2\varepsilon$, which concludes the proof. □

Lemma 21. *Let $\breve{Q}_h^k$ and $\widetilde{Q}_{h,\zeta}^k$ be the Q-functions defined in Algorithm 5. Then,*

$$
\begin{aligned}
\breve{Q}_h^k(x,a) &\overset{\text{def}}{=} \min_{(\overline{x},\overline{a})\in\bar{\mathcal{X}}_h^k\times\bar{\mathcal{A}}_h^k} \left(\widetilde{Q}_{h,\zeta}^k(\overline{x},\overline{a}) + L\rho\left[(x,a),(\overline{x},\overline{a})\right] \right) \\
&= \min_{s\in[k-1]} \left(\widetilde{Q}_{h,\zeta}^k(x_h^s,a_h^s) + L\rho\left[(x,a),\zeta_h^k(x_h^s,a_h^s)\right] \right).
\end{aligned}
$$

Proof. Notice that, although $\widetilde{Q}_{h,\zeta}^k$ is only computed for the representative state-action pairs, it is defined for any (x,a) as

$$
\widetilde{Q}_{h,\zeta}^k(x,a) = \breve{r}_h^k(x,a) + \breve{P}_h^k \breve{V}_{h+1}(x,a) + \breve{\mathtt{B}}_h^k(x,a).
$$

We claim that

$$
\underbrace{\left\{\zeta_h^k(x_h^s,a_h^s) : s\in[k-1]\right\}}_{A} = \underbrace{\bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k}_{B}.
$$

First, $A \subset B$, since $\forall(s,h)$, we have $\zeta_h^k(x_h^s,a_h^s) \in \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k$. Second, for any $(\overline{x},\overline{a}) \in \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k$, there exists (s,h) such that $(\overline{x},\overline{a}) = (x_h^s,a_h^s) = \zeta_h^{s+1}(x_h^s,a_h^s) = \zeta_h^k(x_h^s,a_h^s) \in A$, which implies that $B \subset A$.

Together with the fact that $\widetilde{Q}_{h,\zeta}^k(x_h^s,a_h^s) = \widetilde{Q}_{h,\zeta}^k(\zeta_h^k(x_h^s,a_h^s))$, which holds by Definition 11 and Algorithm 5, this concludes the proof. □

Lemma 22. *The difference between the Q-values computed by `KeRNS` and `RS-KeRNS` are bounded as follows*

$$
\sup_{x,a} \left| Q_h^k(x,a) - \breve{Q}_h^k(x,a) \right| \lesssim \left(L(\varepsilon + \varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma} H \right) (H - h + 1).
$$

Proof. We proceed by induction on h. Let

$$\epsilon_h \stackrel{\text{def}}{=} \left(L(\varepsilon + \varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma} H\right)(H - h + 1).$$

For $h = H + 1$, $\epsilon_{H+1} = 0$, $Q^k_{H+1} = \breve{Q}^k_{H+1} = 0$ and the claim holds.

Now, assume that the claim is true for $h+1$. In this case, we have, for any x,

$$\begin{aligned}
&\left|V^k_{h+1}(x) - \breve{V}^k_{h+1}(x)\right| \\
&= \left|\min\left(H - h, \max_a Q^k_{h+1}(x,a)\right) - \min\left(H - h, \max_a \breve{Q}^k_{h+1}(x,a)\right)\right| \\
&\le \left|\max_a Q^k_{h+1}(x,a) - \max_a \breve{Q}^k_{h+1}(x,a)\right| \\
&\le \max_a \left|Q^k_{h+1}(x,a) - \breve{Q}^k_{h+1}(x,a)\right| \le \epsilon_{h+1},
\end{aligned}$$

where we used induction hypothesis and the fact that, for any $a, b, c \in \mathbb{R}$, we have $|\min(a,b) - \min(a,c)| \le |b - c|$ (Fact 1).

For any (x^s_h, a^s_h) with $s \in [k-1]$, we have

$$\begin{aligned}
&\left|\widetilde{Q}^k_h(x^s_h, a^s_h) - \widetilde{Q}^k_{h,\zeta}(x^s_h, a^s_h)\right| \\
&\le \left|\widehat{r}^k_h(x^s_h, a^s_h) - \breve{r}^k_h(x^s_h, a^s_h)\right| + \left|\,\mathtt{B}^k_h(x^s_h, a^s_h) - \breve{\mathtt{B}}^k_h(x^s_h, a^s_h)\right| \\
&\quad + \left|\left(\widehat{P}^k_h - \breve{P}^k_h\right) V^k_{h+1}(x^s_h, a^s_h) + \breve{P}^k_h\left(V^k_{h+1} - \breve{V}^k_{h+1}\right)(x^s_h, a^s_h)\right| \\
&\lesssim \epsilon_{h+1} + \frac{H}{\sigma}\left(\rho\left[(x^s_h, a^s_h), \zeta^k_h(x^s_h, a^s_h)\right] + \varepsilon\right) + L\varepsilon_{\mathcal{X}} + \frac{\varepsilon}{\sigma} H
\end{aligned}$$

where, in the last line, we used the induction hypothesis and lemmas 18, 19 and 20.

By the construction of RS-KeRNS, we have $\max_{s' \in [k-1]} \rho\left[(x^{s'}_h, a^{s'}_h), \zeta^k_h(x^{s'}_h, a^{s'}_h)\right] \le \varepsilon$. Consequently,

$$\left|\widetilde{Q}^k_h(x^s_h, a^s_h) - \widetilde{Q}^k_{h,\zeta}(x^s_h, a^s_h)\right| \lesssim \epsilon_{h+1} + L\varepsilon_{\mathcal{X}} + \frac{\varepsilon}{\sigma} H.$$

Now, take an arbitrary (x, a). By Lemma 21,

$$\begin{aligned}
\breve{Q}^k_h(x,a) &= \min_{(\overline{x},\overline{a}) \in \bar{\mathcal{X}}^k_h \times \bar{\mathcal{A}}^k_h} \left(\widetilde{Q}^k_{h,\zeta}(\overline{x},\overline{a}) + L\rho\left[(x,a),(\overline{x},\overline{a})\right]\right) \\
&= \min_{s \in [k-1]} \left(\widetilde{Q}^k_{h,\zeta}(x^s_h, a^s_h) + L\rho\left[(x,a), \zeta^k_h(x^s_h, a^s_h)\right]\right)
\end{aligned}$$

and, by definition,

$$Q^k_h(x,a) = \min_{s \in [k-1]} \left(\widetilde{Q}^k_h(x^s_h, a^s_h) + L\rho\left[(x,a),(x^s_h, a^s_h)\right]\right),$$

we obtain, for any (x, a),

$$\begin{aligned}
&\left|Q^k_h(x,a) - \breve{Q}^k_h(x,a)\right| \\
&\lesssim \min_{s \in [k-1]} \left|\widetilde{Q}^k_h(x^s_h, a^s_h) - \widetilde{Q}^k_{h,\zeta}(x^s_h, a^s_h)\right| + L \min_{s \in [k-1]} \left|\rho\left[(x,a),(x^s_h, a^s_h)\right] - \rho\left[(x,a), \zeta^k_h(x^s_h, a^s_h)\right]\right| \\
&\le \epsilon_{h+1} + L\varepsilon_{\mathcal{X}} + \frac{\varepsilon}{\sigma} H + L \max_{s \in [k-1]} \rho\left[(x^s_h, a^s_h), \zeta^k_h(x^s_h, a^s_h)\right] \\
&\le \epsilon_{h+1} + L(\varepsilon + \varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma} H = \epsilon_h.
\end{aligned}$$

which concludes the proof. □

Theorem 5 (UCRL-type regret bound for RS-KeRNS)**.** *With probability at least* $1-\delta$*, the regret of RS-KeRNS is bounded as follows*

$$\mathcal{R}^{\texttt{RS-KeRNS}}(K) \lesssim \mathcal{R}_1^{\texttt{KeRNS}}(K) + L(\varepsilon + \varepsilon_{\mathcal{X}})KH^2 + \frac{\varepsilon}{\sigma}KH^3.$$

where $\mathcal{R}_1^{\texttt{KeRNS}}(K)$ *is the UCRL-type regret bound given in Theorem 3 for KeRNS.*

Proof. Let $\breve{\pi}_k$ be the policy followed by RS-KeRNS in episode k and let $\breve{\delta}_h^k \stackrel{\text{def}}{=} \breve{V}_h^k(x_h^k) - \mathrm{V}_{k,h}^{\breve{\pi}_k}(x_h^k)$

Regret decomposition Consider the following decomposition, also used in the proof of Lemma 11:

$$\begin{aligned}
\breve{\delta}_h^k &= \breve{V}_h^k(x_h^k) - \mathrm{V}_{k,h}^{\breve{\pi}_k}(x_h^k) \\
&\le \breve{Q}_h^k(x_h^k, a_h^k) - \mathrm{Q}_{k,h}^{\breve{\pi}_k}(x_h^k, a_h^k) \\
&\le \breve{Q}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathrm{Q}_{k,h}^{\breve{\pi}_k}(x_h^k, a_h^k) + L\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right], \quad \text{since } \breve{Q}_h^k \text{ is } L\text{-Lipschitz} \\
&\le \widetilde{Q}_{h,\zeta}^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathrm{Q}_{k,h}^{\breve{\pi}_k}(x_h^k, a_h^k) + L\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right], \quad \text{since } \breve{Q}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) \le \widetilde{Q}_{h,\zeta}^k(\widetilde{x}_h^k, \widetilde{a}_h^k) \\
&= \breve{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(x_h^k, a_h^k) + \breve{P}_h^k V_{h+1}^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathrm{P}_h^k \mathrm{V}_{k,h+1}^{\breve{\pi}_k}(x_h^k, a_h^k) + \breve{\mathtt{B}}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + L\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right] \\
&= \underbrace{\breve{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(x_h^k, a_h^k)}_{(\mathbf{A})} + \underbrace{\left[\breve{P}_h^k - \mathrm{P}_h^k\right] \mathrm{V}_{k,h+1}^*(\widetilde{x}_h^k, \widetilde{a}_h^k)}_{(\mathbf{B})} + \underbrace{\left[\breve{P}_h^k - \mathrm{P}_h^k\right]\left(\breve{V}_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k)}_{(\mathbf{C})} \\
&\quad + \underbrace{\mathrm{P}_h^k \breve{V}_{h+1}^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathrm{P}_h^k \mathrm{V}_{k,h+1}^{\breve{\pi}_k}(x_h^k, a_h^k)}_{(\mathbf{D})} + \breve{\mathtt{B}}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + 2L\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right].
\end{aligned}$$

We will use the following results, which are a consequence of Lemmas 18, 19 and 20 and the fact that $\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), \zeta(\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \le \varepsilon$:

$$\left|\widehat{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \breve{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)\right| \lesssim \frac{4}{\sigma}\left(\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), \zeta_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)\right] + \varepsilon\right) + \frac{8\varepsilon}{\sigma} \lesssim \frac{\varepsilon}{\sigma}.$$

$$\begin{aligned}
\breve{\mathtt{B}}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) &= \mathtt{B}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + \breve{\mathtt{B}}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathtt{B}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) \\
&\lesssim \mathtt{B}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + \frac{H}{\sigma}\left(\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), \zeta_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)\right] + \varepsilon\right) + \frac{H\varepsilon}{\sigma} \\
&\lesssim \mathtt{B}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + \frac{\varepsilon}{\sigma}H.
\end{aligned}$$

and, for any function f that is L-Lipschitz and bounded by H,

$$\left|\left(\widehat{P}_h^k - \breve{P}_h^k\right) f(\widetilde{x}_h^k, \widetilde{a}_h^k)\right| \lesssim L\varepsilon_{\mathcal{X}} + \frac{4H}{\sigma}\left(\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), \zeta(\widetilde{x}_h^k, \widetilde{a}_h^k)\right] + \varepsilon\right) + \frac{H\varepsilon}{\sigma} \lesssim L\varepsilon_{\mathcal{X}} + \frac{\varepsilon}{\sigma}H.$$

Now, we bound each term $(\mathbf{A}) - (\mathbf{D})$.

Term $(\mathbf{A})$: by Lemma 19, the definition of $\mathcal{G}$ and Corollary 2, we have

$$\begin{aligned}
(\mathbf{A}) &= \breve{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \widehat{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + \widehat{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + r_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(x_h^k, a_h^k) \\
&\lesssim \frac{\varepsilon}{\sigma} + \widehat{r}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + r_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - r_h^k(x_h^k, a_h^k) \\
&\le \frac{\varepsilon}{\sigma} + {}^{\mathrm{r}}\mathtt{B}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + \mathbf{bias}_{\mathrm{r}}(k,h) + L_{\mathrm{r}}\rho\left[(\widetilde{x}_h^k, \widetilde{a}_h^k), (x_h^k, a_h^k)\right].
\end{aligned}$$

Term (**B**):

$$\begin{aligned}(\mathbf{B}) &= \left[\breve{P}_h^k - \widehat{P}_h^k\right] \mathrm{V}^*_{k,h+1}(\widetilde{x}_h^k, \widetilde{a}_h^k) + \left[\widehat{P}_h^k - \mathrm{P}_h^k\right] \mathrm{V}^*_{k,h+1}(\widetilde{x}_h^k, \widetilde{a}_h^k) \\ &\lesssim L\varepsilon_{\mathcal{X}} + \frac{\varepsilon}{\sigma} H + \left[\widehat{P}_h^k - \overline{P}_h^k\right] \mathrm{V}^*_{k,h+1}(\widetilde{x}_h^k, \widetilde{a}_h^k) + \left[\overline{P}_h^k - \mathrm{P}_h^k\right] \mathrm{V}^*_{k,h+1}(\widetilde{x}_h^k, \widetilde{a}_h^k) \\ &\leq L\varepsilon_{\mathcal{X}} + \frac{\varepsilon}{\sigma} H + {}^{p}\mathtt{B}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k) + \mathbf{bias_p}(k,h)\end{aligned}$$

Term (**C**): Using Corollary 2, we obtain

$$\begin{aligned}(\mathbf{C}) &= \left[\breve{P}_h^k - \mathrm{P}_h^k\right] \left(\breve{V}_{h+1}^k - \mathrm{V}^*_{k,h+1}\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) \\ &= \left[\breve{P}_h^k - \widehat{P}_h^k\right] \left(\breve{V}_{h+1}^k - \mathrm{V}^*_{k,h+1}\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) + \left[\widehat{P}_h^k - \mathrm{P}_h^k\right] \left(\breve{V}_{h+1}^k - \mathrm{V}^*_{k,h+1}\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) \\ &\lesssim L\varepsilon_{\mathcal{X}} + \frac{\varepsilon}{\sigma} H + \left[\widehat{P}_h^k - \overline{P}_h^k\right] \left(\breve{V}_{h+1}^k - \mathrm{V}^*_{k,h+1}\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) + \mathbf{bias_p}(k,h) \\ &\lesssim L\varepsilon_{\mathcal{X}} + \frac{\varepsilon}{\sigma} H + \sqrt{\frac{H^2\,|\mathcal{C}'_\sigma|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + L\sigma + \mathbf{bias_p}(k,h)\end{aligned}$$

by the definition of $\mathcal{G}$.

Term (**D**): From Assumption 2, for any L-Lipschitz function, the mapping $(x,a) \mapsto \mathrm{P}_h^k f(x,a)$ is $L_{\mathrm{p}}L$-Lipschitz . Consequently,

$$\begin{aligned}(\mathbf{D}) &= \mathrm{P}_h^k \breve{V}_{h+1}^k(\widetilde{x}_h^k, \widetilde{a}_h^k) - \mathrm{P}_h^k \mathrm{V}^{\breve{\pi}_k}_{k,h+1}(x_h^k, a_h^k) \\ &\leq \mathrm{P}_h^k \breve{V}_{h+1}^k(x_h^k, a_h^k) - \mathrm{P}_h^k \mathrm{V}^{\breve{\pi}_k}_{k,h+1}(x_h^k, a_h^k) + L_{\mathrm{p}}L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \\ &= \mathrm{P}_h^k \left(\breve{V}_{h+1}^k - \mathrm{V}^{\breve{\pi}_k}_{k,h+1}\right)(x_h^k, a_h^k) + L_{\mathrm{p}}L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \\ &= \breve{\delta}_{h+1}^k + \xi_{h+1}^k + L_{\mathrm{p}}L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right].\end{aligned}$$

where

$$\xi_{h+1}^k \stackrel{\mathrm{def}}{=} \mathrm{P}_h^k \left(\breve{V}_{h+1}^k - \mathrm{V}^{\breve{\pi}_k}_{k,h+1}\right)(x_h^k, a_h^k) - \breve{\delta}_{h+1}^k$$

is a martingale difference sequence with respect to $(\mathcal{F}_h^k)_{k,h}$ bounded by $4H$.

Putting together the bounds for (**A**)-(**D**) and using the definition of the bonuses $\mathtt{B}_h^k$, we obtain

$$\begin{aligned}\breve{\delta}_h^k \lesssim \breve{\delta}_{h+1}^k + \xi_{h+1}^k + L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] + \sqrt{\frac{H^2\,|\mathcal{C}'_\sigma|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} \\ + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + L\sigma + \mathbf{bias}(k,h') + L\varepsilon_{\mathcal{X}} + \frac{\varepsilon}{\sigma} H\end{aligned}$$

where the constant in front of $\breve{\delta}_{h+1}^k$ is *exact* (i.e., not omitted by $\lesssim$).

Now, we follow the same arguments as in the proof of Lemma 11. Consider the event $E_h^k \stackrel{\mathrm{def}}{=} \left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \leq 2\sigma\right\}$. The inequality above gives us

$$\begin{aligned}\mathbb{I}\left\{E_h^k\right\} \breve{\delta}_{h,}^k \lesssim \mathbb{I}\left\{E_h^k\right\} \breve{\delta}_{h+1}^k + \mathbb{I}\left\{E_h^k\right\} \xi_{h+1}^k + \mathbb{I}\left\{E_h^k\right\} \sqrt{\frac{H^2\,|\mathcal{C}'_\sigma|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} \\ + \mathbb{I}\left\{E_h^k\right\} \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} + 3L\sigma + \mathbf{bias}(k,h') + L\varepsilon_{\mathcal{X}} + \frac{\varepsilon}{\sigma} H. \qquad (8)\end{aligned}$$

Using Lemmas 10 and 22, we upper bound $\mathbb{I}\{E_h^k\}\,\breve{\delta}_{h+1}^k$ in terms of $\breve{\delta}_{h+1}^k$:

$$
\begin{aligned}
\mathbb{I}\{E_h^k\}\,\breve{\delta}_{h+1}^k &= \mathbb{I}\{E_h^k\}\left(\breve{V}_{h+1}^k(x_{h+1}^k) - \mathrm{V}_{k,h}^{\breve{\pi}_k}(x_{h+1}^k)\right)\\
&\lesssim \mathbb{I}\{E_h^k\}\left(HL(\varepsilon+\varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma}H^2 + V_{h+1}^k(x_{h+1}^k) - \mathrm{V}_{k,h+1}^{\breve{\pi}_k}(x_{h+1}^k)\right)\\
&\le HL(\varepsilon+\varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma}H^2 + \mathbb{I}\{E_h^k\}\underbrace{\left(V_{h+1}^k(x_{h+1}^k) + \sum_{h'=h+1}^{H}\mathbf{bias}(k,h) - \mathrm{V}_{k,h+1}^{\breve{\pi}_k}(x_{h+1}^k)\right)}_{\ge 0 \text{ by Lemma 10}}\\
&\le HL(\varepsilon+\varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma}H^2 + V_{h+1}^k(x_{h+1}^k) - \mathrm{V}_{k,h+1}^{\breve{\pi}_k}(x_{h+1}^k) + \sum_{h'=h+1}^{H}\mathbf{bias}(k,h)\\
&\le 2HL(\varepsilon+\varepsilon_{\mathcal{X}}) + 2\frac{\varepsilon}{\sigma}H^2 + \breve{V}_{h+1}^k(x_{h+1}^k) - \mathrm{V}_{k,h+1}^{\breve{\pi}_k}(x_{h+1}^k) + \sum_{h'=h+1}^{H}\mathbf{bias}(k,h)\\
&= \breve{\delta}_{h+1}^k + 2HL(\varepsilon+\varepsilon_{\mathcal{X}}) + 2\frac{\varepsilon}{\sigma}H^2 + \sum_{h'=h+1}^{H}\mathbf{bias}(k,h) \qquad (9)
\end{aligned}
$$

Let $\overline{E}_h^k$ be the complement of E_h^k. Using the inequality above combined with (8), and the fact that $\breve{\delta}_h^k \le H$, we obtain

$$
\begin{aligned}
\breve{\delta}_h^k &= \mathbb{I}\left\{\overline{E}_h^k\right\}\breve{\delta}_h^k + \mathbb{I}\{E_h^k\}\,\breve{\delta}_h^k \qquad (10)\\
&\le H\mathbb{I}\left\{\overline{E}_h^k\right\} + \mathbb{I}\{E_h^k\}\,\breve{\delta}_h^k\\
&\lesssim H\mathbb{I}\left\{\overline{E}_h^k\right\} + \breve{\delta}_{h+1}^k + HL(\varepsilon+\varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma}H^2 + \sum_{h'=h}^{H}\mathbf{bias}(k,h) + \mathbb{I}\{E_h^k\}\,\xi_{h+1}^k\\
&\quad + \mathbb{I}\{E_h^k\}\sqrt{\frac{H^2\,|\mathcal{C}_\sigma'|}{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}} + \mathbb{I}\{E_h^k\}\,\frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)} + L\sigma
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\sum_{k=1}^{K}\breve{\delta}_1^k \lesssim \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\frac{H\sqrt{|\mathcal{C}_\sigma'|}}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}} + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}\right)\mathbb{I}\{E_h^k\} + H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}\left\{\overline{E}_h^k\right\}\\
+ \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}\{E_h^k\}\,\xi_{h+1}^k + H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbf{bias}(k,h) + L(\varepsilon+\varepsilon_{\mathcal{X}})KH^2 + \frac{\varepsilon}{\sigma}KH^3
\end{aligned}
$$

Now, as in the proof of Lemma 11, we show that

$$
H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}\left\{\overline{E}_h^k\right\} \le H^2\,|\mathcal{C}_\sigma|
$$

and, from lemmas 10 and 22, we have:

$$\begin{aligned}
\mathcal{R}^{\texttt{RS-KeRNS}}(K) &= \sum_{k=1}^{K}\left(\mathrm{V}^{*}_{k,1}(x_1^k) - \mathrm{V}^{\breve{\pi}_k}_{k,h}(x_1^k)\right) \\
&\le \sum_{k=1}^{K}\left(V_1^k(x_1^k) - \mathrm{V}^{\breve{\pi}_k}_{k,h}(x_1^k)\right) + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k,h) \\
&\lesssim \sum_{k=1}^{K}\left(\breve{V}_1^k(x_1^k) - \mathrm{V}^{\breve{\pi}_k}_{k,h}(x_1^k)\right) + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k,h) + LKH(\varepsilon+\varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma}KH^2 \\
&= \sum_{k=1}^{K}\breve{\delta}_1^k + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k,h) + LKH(\varepsilon+\varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma}KH^2 \\
&\lesssim \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\frac{H\sqrt{|\mathcal{C}'_\sigma|}}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}} + \frac{\beta H}{\mathbf{C}_h^k(\widetilde{x}_h^k,\widetilde{a}_h^k)}\right)\mathbb{I}\left\{E_h^k\right\} + H^2\,|\mathcal{C}_\sigma| \\
&+ \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}\left\{E_h^k\right\}\xi_{h+1}^k + H\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k,h) + L(\varepsilon+\varepsilon_{\mathcal{X}})KH^2 + \frac{\varepsilon}{\sigma}KH^3.
\end{aligned}$$

Recall the definition $E_h^k \overset{\text{def}}{=} \left\{\rho\left[(x_h^k,a_h^k),(\widetilde{x}_h^k,\widetilde{a}_h^k)\right] \le 2\sigma\right\}$ and the fact that $\widetilde{\xi}_{h+1}^k \overset{\text{def}}{=} \mathbb{I}\left\{E_h^k\right\}\xi_{h+1}^k$ is a martingale difference sequence with respect to $(\mathcal{F}_h^k)_{k,h}$ bounded by $4H$. Then, as in the proof of Theorem 3, we obtain

$$\mathcal{R}^{\texttt{RS-KeRNS}}(K) \lesssim \mathcal{R}_1^{\texttt{KeRNS}}(K) + LKH^2(\varepsilon+\varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma}KH^3$$

with probability at least $1-\delta$. □

Theorem 6 (UCBVI-type regret bound for RS-KeRNS)**.** *With probability at least $1-\delta$, the regret of RS-KeRNS is bounded as follows*

$$\mathcal{R}^{\textit{RS-KeRNS}}(K) \lesssim \mathcal{R}_2^{\textit{KeRNS}}(K) + L(\varepsilon+\varepsilon_{\mathcal{X}})KH^2 + \frac{\varepsilon}{\sigma}KH^3.$$

where $\mathcal{R}_2^{\textit{KeRNS}}(K)$ is the UCBVI-type regret bound given in Theorem 4 for KeRNS.

Proof. We use the same regret decomposition as in the proof of Theorem 5, but the term (**C**) is bounded differently (as in Lemma 12).

Using Lemma 18, Lemma 22, and the same arguments as in the proof of Lemma 12, we have

$$
\begin{aligned}
(\mathbf{C}) &= \left[\breve{P}_h^k - \mathrm{P}_h^k\right]\left(\breve{V}_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) \\
&= \left[\breve{P}_h^k - \widehat{P}_h^k\right]\left(\breve{V}_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) + \left[\widehat{P}_h^k - \mathrm{P}_h^k\right]\left(\breve{V}_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) \\
&\lesssim L(\varepsilon + \varepsilon_{\mathcal{X}})H + \frac{\varepsilon}{\sigma}H^2 + \left[\widehat{P}_h^k - \mathrm{P}_h^k\right]\left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(\widetilde{x}_h^k, \widetilde{a}_h^k) \quad \text{(by lemmas 18 and 22)} \\
&\lesssim L(\varepsilon + \varepsilon_{\mathcal{X}})H + \frac{\varepsilon}{\sigma}H^2 + \frac{1}{H}\mathrm{P}_h^k\left(V_{h+1}^k - \mathrm{V}_{k,h+1}^*\right)(x_h^k, a_h^k) + \frac{H^2\left|\mathcal{C}_\sigma'\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} \\
&\quad + L\sigma + \sum_{h'=h}^{H} \mathbf{bias}(k, h') + L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \quad \text{(following the proof of Lemma 12)} \\
&\lesssim L(\varepsilon + \varepsilon_{\mathcal{X}})H + \frac{\varepsilon}{\sigma}H^2 + \frac{1}{H}\mathrm{P}_h^k\left(\breve{V}_{h+1}^k - \mathrm{V}_{k,h+1}^{\breve{\pi}_k}\right)(x_h^k, a_h^k) + \frac{H^2\left|\mathcal{C}_\sigma'\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)} \\
&\quad + L\sigma + \sum_{h'=h}^{H} \mathbf{bias}(k, h') + L\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right]
\end{aligned}
$$

where, in the last line, we used the fact that $\mathrm{V}_{k,h+1}^{\breve{\pi}_k} \le \mathrm{V}_{k,h+1}^*$ and that $V_{k+1}^k \le \breve{V}_{k+1}^k + L(\varepsilon + \varepsilon_{\mathcal{X}})H + (\varepsilon/\sigma)H^2$ by Lemma 22.

Putting together the bounds for $(\mathbf{A}) - (\mathbf{D})$ and using the same arguments as in the proof of Theorem 5, especially the inequalities (8), (9) and (10), we obtain

$$
\begin{aligned}
\sum_{k=1}^{K} \breve{\delta}_1^k \lesssim \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\frac{H}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{H^2\left|\mathcal{C}_\sigma'\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}\right)\mathbb{I}\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \le 2\sigma\right\} + H^2\left|\mathcal{C}_\sigma\right| \\
+ \sum_{k=1}^{K}\sum_{h=1}^{H}\left(1 + \frac{1}{H}\right)^{H-h+1}\widetilde{\xi}_{h+1}^k + H\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k, h) + LKH\sigma + L(\varepsilon + \varepsilon_{\mathcal{X}})H^2 + \frac{\varepsilon}{\sigma}H^3
\end{aligned}
$$

where $\widetilde{\xi}_{h+1}^k$ is a martingale difference sequence with respect to $(\mathcal{F}_h^k)_{k,h}$ bounded by $4H$.

Now, from lemmas 10 and 22 , we have:

$$
\begin{aligned}
\mathcal{R}^{\mathtt{RS-KeRNS}}(K) &= \sum_{k=1}^{K}\left(\mathrm{V}_{k,1}^*(x_1^k) - \mathrm{V}_{k,h}^{\breve{\pi}_k}(x_1^k)\right) \le \sum_{k=1}^{K}\left(V_1^k(x_1^k) - \mathrm{V}_{k,h}^{\breve{\pi}_k}(x_1^k)\right) + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k, h) \\
&\lesssim \sum_{k=1}^{K}\left(\breve{V}_1^k(x_1^k) - \mathrm{V}_{k,h}^{\breve{\pi}_k}(x_1^k)\right) + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k, h) + LKH(\varepsilon + \varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma}KH^2 \\
&= \sum_{k=1}^{K} \breve{\delta}_1^k + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k, h) + LKH(\varepsilon + \varepsilon_{\mathcal{X}}) + \frac{\varepsilon}{\sigma}KH^2 \\
&\lesssim \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\frac{H}{\sqrt{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}} + \frac{H^2\left|\mathcal{C}_\sigma'\right|}{\mathbf{C}_h^k(\widetilde{x}_h^k, \widetilde{a}_h^k)}\right)\mathbb{I}\left\{\rho\left[(x_h^k, a_h^k), (\widetilde{x}_h^k, \widetilde{a}_h^k)\right] \le 2\sigma\right\} + H^2\left|\mathcal{C}_\sigma\right| \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}\left(1 + \frac{1}{H}\right)^{h}\widetilde{\xi}_{h+1}^k + H\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{bias}(k, h) + LKH\sigma + L(\varepsilon + \varepsilon_{\mathcal{X}})KH^2 + \frac{\varepsilon}{\sigma}KH^3
\end{aligned}
$$

Then, as in the proof of Theorem 4, we obtain

$$
\mathcal{R}^{\mathtt{RS-KeRNS}}(K) \lesssim \mathcal{R}_2^{\mathtt{KeRNS}}(K) + L(\varepsilon + \varepsilon_{\mathcal{X}})KH^2 + \frac{\varepsilon}{\sigma}KH^3
$$

with probability at least $1 - \delta$. □

H Technical Lemmas

Lemma 23 (adapted from Domingues et al. (2020)). *Consider a sequence of non-negative real numbers* $\{z_s\}_{s=1}^t$ *and let* $\overline{\Gamma}_{(\eta,W)} : \mathbb{R}_+ \to [0,1]$ *satisfy Assumption 4. For a given* t*, let*

$$w_s \stackrel{\text{def}}{=} \overline{\Gamma}_{(\eta,W)}\left(t-s-1, \frac{z_s}{\sigma}\right) \quad \text{and} \quad \widetilde{w}_s \stackrel{\text{def}}{=} \frac{w_s}{\beta + \sum_{s'=1}^t w_{s'}}$$

for $\beta > 0$*. Then, we have*

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

Proof. For completeness, we reproduce here the proof of Lemma 7 of Domingues et al. (2020), which also applies to our setting. We split the sum into two terms:

$$\sum_{s=1}^t \widetilde{w}_s z_s = \sum_{s:z_s<c} \widetilde{w}_s z_s + \sum_{s:z_s\ge c} \widetilde{w}_s z_s \le c + \sum_{s:z_s\ge c} \widetilde{w}_s .$$

From Assumption 4, we have $w_s \le C_1 \exp\left(-z_s^2/(2\sigma^2)\right)$. Hence, $\widetilde{w}_s \le (C_1/\beta)\exp\left(-z_s^2/(2\sigma^2)\right)$, since $\beta + \sum_{s'=1}^t w_{s'} \ge \beta$.

We want to find c such that:

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

which implies, for $z_s \ge c$, that $\widetilde{w}_s \le \frac{1}{t}\frac{2\sigma^2}{z_s^2}$.

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

$$\begin{aligned}\frac{x}{2} \ge \log\left(\frac{C_1 t}{\beta} + e\right) \implies x \ge \frac{x}{2} + \log\left(\frac{C_1 t}{\beta} + e\right) &\implies x \ge \log x + \log(C_1 t/\beta + e) \\ &\implies (C_1/\beta)\exp(-x) \le 1/(xt)\end{aligned}$$

as we wanted.

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

$$\sum_{s:z_s\ge c} \widetilde{w}_s z_s \le \sum_{s:z_s\ge c} \frac{1}{t}\frac{2\sigma^2}{z_s^2} z_s \le \frac{2\sigma^2}{t}\sum_{s:z_s\ge c}\frac{1}{z_s} \le \frac{2\sigma^2}{c}\frac{|\{s : z_s \ge c\}|}{t} \le \frac{2\sigma^2}{c}.$$

Finally, we obtain:

$$\begin{aligned}\sum_{s=1}^t \widetilde{w}_s z_s &\le c + \sum_{s:z_s\ge c} \widetilde{w}_s z_s \le c + \frac{2\sigma^2}{c} \\ &= 2\sigma\sqrt{\log(C_1 t/\beta + e)} + \frac{\sigma}{\sqrt{\log(C_1 t/\beta + e)}} \le 2\sigma\left(1 + \sqrt{\log(C_1 t/\beta)}\right).\end{aligned}$$

□

Lemma 24. *Let $\overline{\Gamma}_{(\eta,W)} : \mathbb{R}_+ \to [0,1]$ be a kernel that satisfies Assumption 4. Let $a \in \mathbb{R}_+^t$ and f_1, f_2, f_3 be functions from $\mathbb{R}_+^t$ to $\mathbb{R}$ defined as*

$$f_1(z) = \frac{\sum_{s=1}^t \overline{\Gamma}_{(\eta,W)}(t-s-1, z_s/\sigma)\, a_s}{\beta + \sum_{s=1}^t \overline{\Gamma}_{(\eta,W)}(t-s-1, z_s/\sigma)},$$
$$f_2(z) = \sqrt{\frac{1}{\beta + \sum_{s=1}^t \overline{\Gamma}_{(\eta,W)}(t-s-1, z_s/\sigma)}},$$
$$f_3(z) = \frac{1}{\beta + \sum_{s=1}^t \overline{\Gamma}_{(\eta,W)}(t-s-1, z_s/\sigma)}$$

Then, for any $y, z \in \mathbb{R}_+$, we have

$$|f_1(z) - f_1(y)| \le \frac{2C_2 \|a\|_\infty t}{\beta\sigma} \|z-y\|_\infty$$
$$|f_2(z) - f_2(y)| \le \frac{C_2 t}{2\beta^{3/2}\sigma} \|z-y\|_\infty$$
$$|f_3(z) - f_3(y)| \le \frac{C_2 t}{\beta^2\sigma} \|z-y\|_\infty$$

Proof. From Assumption 4, the function $z \mapsto \overline{\Gamma}_{(\eta,W)}(t-s-1, z)$ is C_2-Lipschitz , which yields

$$\begin{aligned}
&|f_1(z) - f_1(y)| \\
&\le \left| \frac{\sum_{s=1}^t \left(\overline{\Gamma}_{(\eta,W)}(t-s-1, z_s/\sigma) - \overline{\Gamma}_{(\eta,W)}(t-s-1, y_s/\sigma)\right) a_s}{\beta + \sum_{s=1}^t \overline{\Gamma}_{(\eta,W)}(t-s-1, z_s/\sigma)} \right| \\
&+ \left| \frac{\sum_{s=1}^t \overline{\Gamma}_{(\eta,W)}(t-s-1, y_s/\sigma)\, a_s}{\beta + \sum_{s=1}^t \overline{\Gamma}_{(\eta,W)}(t-s-1, y_s/\sigma)} \right| \left| \frac{\sum_{s=1}^t \left(\overline{\Gamma}_{(\eta,W)}(t-s-1, z_s/\sigma) - \overline{\Gamma}_{(\eta,W)}(t-s-1, y_s/\sigma)\right)}{\beta + \sum_{s=1}^t \overline{\Gamma}_{(\eta,W)}(t-s-1, z_s/\sigma)} \right| \\
&\le \frac{C_2 \sum_{s=1}^t (1/\sigma)\, |z_s - y_s|\, a_s}{\beta + \sum_{s=1}^t \overline{\Gamma}_{(\eta,W)}(t-s-1, z_s/\sigma)} + \|a\|_\infty \frac{C_2 \sum_{s=1}^t (1/\sigma)\, |z_s - y_s|}{\beta + \sum_{s=1}^t \overline{\Gamma}_{(\eta,W)}(t-s-1, z_s/\sigma)} \\
&\le \frac{2C_2 \|a\|_\infty t}{\beta\sigma} \|z-y\|_\infty .
\end{aligned}$$

The proofs for f_2 and f_3 are analogous. For f_2, we also use the fact that the function $x \mapsto (1/\sqrt{\beta + x})$ is $1/(2\beta^{3/2})$-Lipschitz . □

Lemma 25. *Let $a \in \mathbb{R}_+^t$ and f_1, f_2, f_3 be functions from $\mathbb{R}_+^t$ to $\mathbb{R}$ defined as*

$$f_1(z) = \frac{\sum_{s=1}^t g(s) \exp\left(-z_s^2/(2\sigma^2)\right) a_s}{\beta + \sum_{s=1}^t g(s) \exp\left(-z_s^2/(2\sigma^2)\right)},$$
$$f_2(z) = \sqrt{\frac{1}{\beta + \sum_{s=1}^t g(s) \exp\left(-z_s^2/(2\sigma^2)\right)}},$$
$$f_3(z) = \frac{1}{\beta + \sum_{s=1}^t g(s) \exp\left(-z_s^2/(2\sigma^2)\right)}$$

where $g : \mathbb{N}^* \mapsto [0,1]$ is an arbitrary function bounded by 1. Then, for any $y, z \in \mathbb{R}_+$, we have

$$|f_1(z) - f_1(y)| \le \frac{4\|a\|_\infty}{\sigma} \left(1 + \sqrt{\log^+(t/\beta)}\right) \|z-y\|_\infty$$
$$|f_2(z) - f_2(y)| \le \frac{1}{2\beta^{1/2}\sigma} \left(1 + \sqrt{\log^+(t/\beta)}\right) \|z-y\|_\infty$$
$$|f_3(z) - f_3(y)| \le \frac{1}{\beta\sigma} \left(1 + \sqrt{\log^+(t/\beta)}\right) \|z-y\|_\infty$$

Proof. We use the fact that, for any differentiable $f : \mathbb{R}_+^t \to \mathbb{R}$,

$$|f(x) - f(y)| \leq \sup_{z \in \mathbb{R}_+^t} \|\nabla f(z)\|_1 \|x - y\|_\infty .$$

We have

$$\left| \frac{\partial f_1(z)}{\partial z_i} \right| \leq \frac{2\|a\|_\infty}{\sigma^2} \frac{g(i) z_i \exp(-z_i^2/(2\sigma^2))}{\beta + \sum_{s=1}^t g(s) z_s \exp(-z_s^2/(2\sigma^2))}$$

which implies, by Lemma 23,

$$\begin{aligned}
\|\nabla f_1(z)\|_1 &\leq \frac{2\|a\|_\infty}{\sigma^2} \frac{\sum_{i=1}^t g(i) z_i \exp(-z_i^2/(2\sigma^2))}{\beta + \sum_{s=1}^t g(s) z_s \exp(-z_s^2/(2\sigma^2))} \\
&\leq \frac{2\|a\|_\infty}{\sigma^2} 2\sigma \left(1 + \sqrt{\log^+(t/\beta)}\right) = \frac{4\|a\|_\infty}{\sigma} \left(1 + \sqrt{\log^+(t/\beta)}\right).
\end{aligned}$$

The proofs for f_2 and f_3 are analogous. □

Lemma 26 (value functions are Lipschitz continuous)**.** *Under Assumptions 1 and 2, for all (k, h), the functions $\mathrm{V}_{k,h}^*$ and $\mathrm{Q}_{k,h}^*$ are L_h-Lipschitz , where $L_h \stackrel{\text{def}}{=} \sum_{h'=h}^H L_\mathrm{r} {L_\mathrm{p}}^{H-h'}$.*

Proof. This fact is proved in Lemma 4 of Domingues et al. (2020) and also in Proposition 2.5 of Sinclair et al. (2019). For completeness, we also present a proof here.

We proceed by induction. For $h = H$, $\mathrm{Q}_{k,H}^*(x, a) = r_H^k(x, a)$ which is L_r-Lipschitz by Assumption 2. Also,

$$\begin{aligned}
\mathrm{V}_{k,H}^*(x) - \mathrm{V}_{k,H}^*(y) &= \max_a \mathrm{Q}_{k,H}^*(x, a) - \max_a \mathrm{Q}_{k,H}^*(y, a) \leq \max_a \left(\mathrm{Q}_{k,H}^*(x, a) - \mathrm{Q}_{k,H}^*(y, a)\right) && (11) \\
&\leq \max_a L_H \rho\left[(x, a), (y, a)\right] \leq L_H \rho_{\mathcal{X}}(x, y), \quad \text{by Assumption 1} && (12)
\end{aligned}$$

which verifies the induction hypothesis for $h = H$, since we can invert the roles of x and y to obtain $\left|\mathrm{V}_{k,H}^*(x) - \mathrm{V}_{k,H}^*(y)\right| \leq L_H \rho_{\mathcal{X}}(x, y)$.

Now, assume that the hypothesis is true for $h + 1$, i.e., that $\mathrm{V}_{k,h+1}^*$ and $\mathrm{Q}_{k,h+1}^*$ are L_{h+1}-Lipschitz . We have

$$\begin{aligned}
\mathrm{Q}_{k,h}^*(x, a) - \mathrm{Q}_{k,h}^*(x', a') &\leq L_\mathrm{r} \rho\left[(x, a), (x', a')\right] + \int_{\mathcal{X}} \mathrm{V}_{k,h+1}^*(y)(P_h(\mathrm{d}y|x, a) - P_h(\mathrm{d}y|x', a')) \\
&\leq L_\mathrm{r} \rho\left[(x, a), (x', a')\right] + L_{h+1} \int_{\mathcal{X}} \frac{\mathrm{V}_{k,h+1}^*(y)}{L_{h+1}} (P_h(\mathrm{d}y|x, a) - P_h(\mathrm{d}y|x', a')) \\
&\leq \left[L_\mathrm{r} + L_\mathrm{p} \sum_{h'=h+1}^H L_\mathrm{r} {L_\mathrm{p}}^{H-h'}\right] \rho\left[(x, a), (x', a')\right] \\
&= \sum_{h'=h}^H L_\mathrm{r} {L_\mathrm{p}}^{H-h'} \rho\left[(x, a), (x', a')\right]
\end{aligned}$$

where, in last inequality, we use fact that $\mathrm{V}_{k,h+1}^*/L_{h+1}$ is 1-Lipschitz, the definition of the 1-Wasserstein distance and Assumption 2. The same argument used in Eq. 11 shows that $\left|\mathrm{V}_{k,h}^*(x) - \mathrm{V}_{k,h}^*(y)\right| \leq L_h \rho_{\mathcal{X}}(x, y)$, which concludes the proof. □

Fact 1 (small useful result)**.**

$$|\min(a, b) - \min(a, c)| \leq |b - c|$$

Proof. Since $\min(x, y) = (x + y)/2 - |x - y|/2$, we have

$$\begin{aligned}
\min(a, b) - \min(a, c) &= \frac{a+b}{2} - \frac{|a-b|}{2} - \frac{a+c}{2} + \frac{|a-c|}{2} \\
&= \frac{b-c}{2} + \frac{|a-c| - |a-b|}{2} \\
&\leq \frac{b-c}{2} + \frac{|b-c|}{2} \\
&\leq |b-c|
\end{aligned}$$

where we used the fact that $|a - c| \leq |a - b| + |b - c|$. By symmetry, $\min(a, c) - \min(a, b) \leq |b - c|$. which gives us the result. □

I Experiments

I.1 Setup

We consider a continuous MDP whose state-space is the unit ball in $\mathbb{R}^2$ with four actions, representing a move to the right, left, up or down. Each action results in a displacement of 0.1 in the corresponding direction, plus a Gaussian noise, in both directions, of standard variation 0.01. The agent starts at $(0, 0)$. Let $b_i^k \in \{0, 0.25, 0.5, 0.75, 1\}$ and $x_i \in \{(0.8, 0.0), (0.0, 0.8), (-0.8, 0.0), (0.0, -0.8)\}$. We consider the following mean reward function:

$$r_h^k(x, a) = \sum_{i=1}^{4} b_i^k \max\left(0, 1 - \frac{\|x - x_i\|_2}{0.5}\right)$$

Every N episodes, the coefficients b_i^k are changed, which impact the optimal policy.

Taking $\eta = \exp(-(1/N)^{2/3})$, we tested the Gaussian kernel $\Gamma(t, u, v) = \eta^t \exp\left(-\rho\,[u, v]^2 / (2\sigma^2)\right)$ and a higher-order kernel $\Gamma(t, u, v) = \eta^t \exp\left(-(\rho\,[u, v]/\sigma)^4/2\right)$. We set $\sigma = 0.05$, $\varepsilon = \varepsilon_{\mathcal{X}} = 0.1$, $\beta = 0.01$, $H = 15$.

We ran the experiment with horizon $H = 15$ for 2×10^4 episodes. Every N episodes, the coefficients b_i^k were changed, according to Table 4.

Table 4: Value of b_i^k according to x_i and the episode k

episode / x_i	$(0.8, 0.0)$	$(0.0, 0.8)$	$(-0.8, 0.0)$	$(0.0, -0.8)$
$k \mod N = 0$	1/4	0	0	0
$k \mod N = 1$	1/4	1/2	0	0
$k \mod N = 2$	1/4	1/2	3/4	0
$k \mod N = 3$	1/4	1/2	3/4	1

We took $\beta = 0.01$ and used the following simplified exploration bonuses:

$$\mathtt{B}_h^k(x, a) = \frac{0.1}{\sqrt{\mathbf{C}_h^k(x, a)}} + \frac{\beta H}{\mathbf{C}_h^k(x, a)} \tag{13}$$

where the factor 0.1 was chosen in order to ensure that the baseline is able to learn a good policy in less than 1000 episodes, i.e., before there is a change in the environment.

Additionally, to take into account the fact that the Lipschitz constant is rarely known in practical problems, we replaced the interpolation step (line 8 of Alg. 5) by a nearest-neighbor search in the representative states:

$$\breve{Q}_h^k(x, a) = \widetilde{Q}_{h,\zeta}^k(x', a'), \quad \text{where} \quad (x', a') = \operatorname*{argmin}_{(\overline{x}, \overline{a}) \in \bar{\mathcal{X}}_h^k \times \bar{\mathcal{A}}_h^k} \rho\left[(x, a), (\overline{x}, \overline{a})\right].$$

I.2 Results

Figures 2 and 3 show the total reward and the regret of `RS-KeRNS` compared to baselines for the two choices of kernel function (Gaussian and 4-th order kernel), for 3 different values of Δ, which is determined by the period N of changes in the MDP (the reward changes every N episodes).

In all experiments we observe that `Kernel-UCBVI` is not able to adapt to the changes in the environment, whereas `RS-KeRNS` is able to track the behavior of the baseline `RestartBaseline` which knows when the changes happen and resets the reward estimator when there is a change.

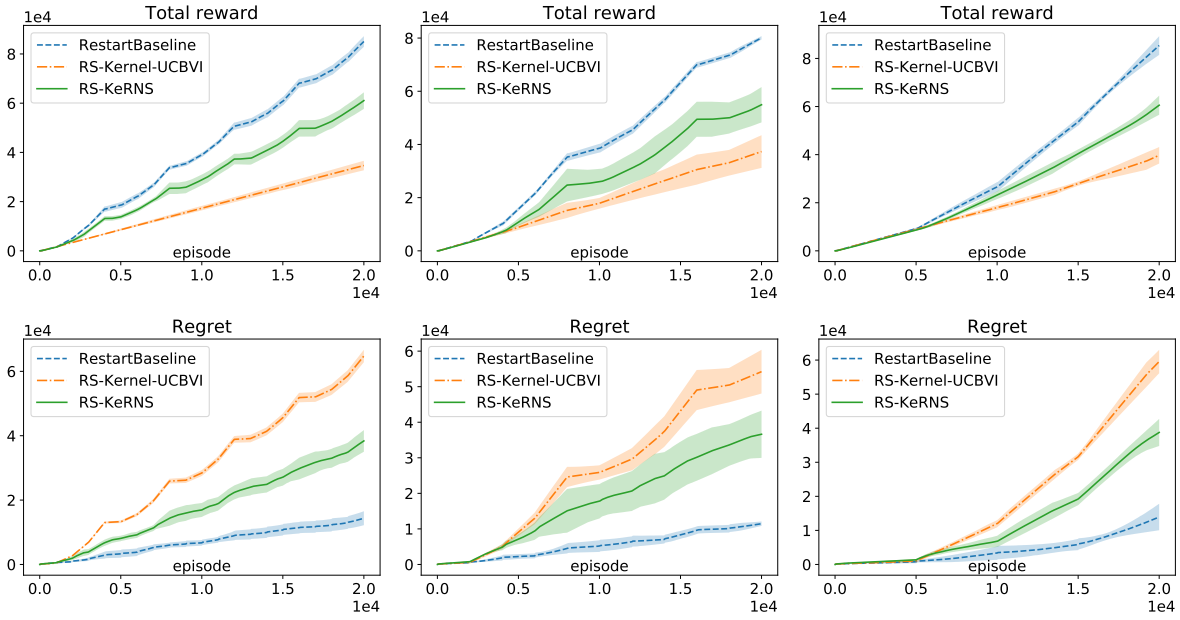

Figure 2: Total reward (top row) and regret(bottom row) of RS-KeRNS compared to baselines, using the Gaussian kernel $\Gamma(t,u,v) = \eta^t \exp\left(-\rho\left[u,v\right]^2/(2\sigma^2)\right)$. The figures on the left, in the middle, and on the right correspond to $N = 1000$, $N = 2000$ and $N = 5000$, respectively, where N is the period of the changes in the MDP. Average over 4 runs.

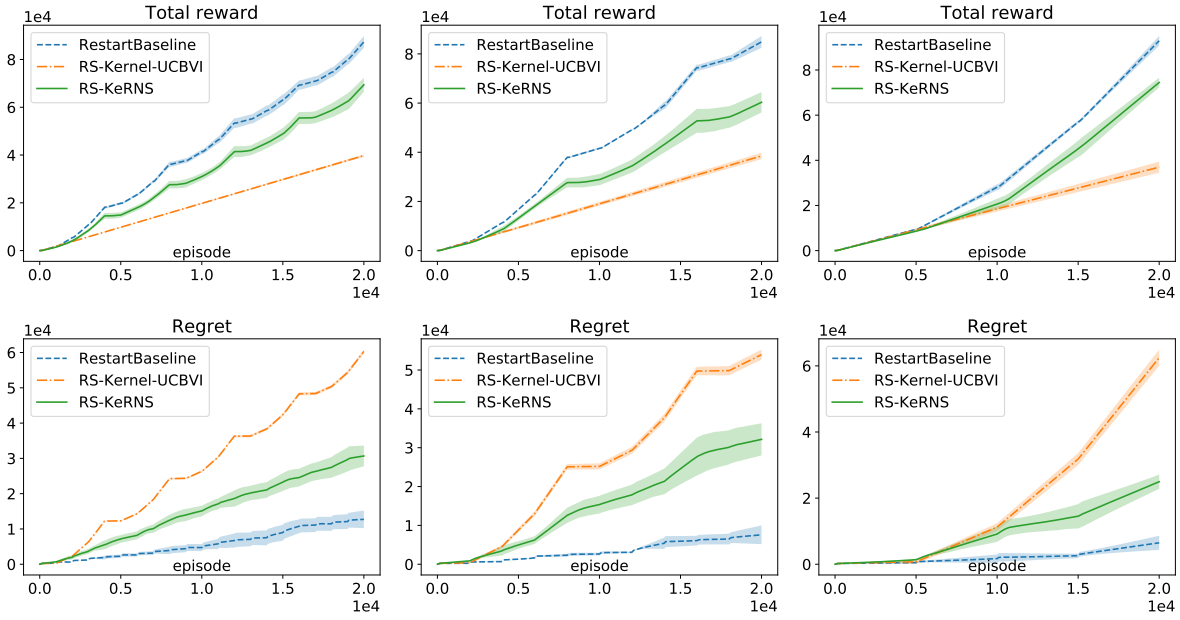

Figure 3: Total reward (top row) and regret(bottom row) of RS-KeRNS compared to baselines, using the kernel $\Gamma(t,u,v) = \eta^t \exp\left(-(\rho\left[u,v\right]/\sigma)^4/2\right)$. The figures on the left, in the middle, and on the right correspond to $N = 1000$, $N = 2000$ and $N = 5000$, respectively, where N is the period of the changes in the MDP. Average over 4 runs.